INTEGRATED FRAMEWORKS FOR KNOWLEDGE DISCOVERY IN HUMAN-MACHINE COMPLEX SYSTEMS USING MULTIPLE DATA STREAMS

By

NASIM ARBABZADEH

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Industrial And Systems Engineering

Written under the direction of

Mohsen A. Jafari

And approved by

____

____

____

____

New Brunswick, New Jersey

January, 2017

ABSTRACT OF THE DISSERTATION

INTEGRATED FRAMEWORKS FOR KNOWLEDGE DISCOVERY IN HUMAN-MACHINE COMPLEX SYSTEMS USING MULTIPLE DATA STREAMS

by NASIM ARBABZADEH

Dissertation Director:

Mohsen A. Jafari

Complex human-machine systems where human plays controlling roles are highly dynamic and complicated making the traditional models and methodologies less effective. The operability of such a complex system is affected by the performance and inter-relationships of a wide range of both internal and exogenous variables. The dynamic nature of such systems makes it necessary to apply probabilistic and stochastic models to capture the system variability. In this study, we propose integrated frameworks for two such systems, transportation and healthcare, by applying advanced data analytics, statistical and stochastics models and machine learning methods to extract important knowledge for either prediction or causal analysis. The results can be used for both off-line design of better targeted countermeasures and corrective actions or on-line monitoring for situational awareness which can in turn assist with well-informed control actions.

For the transportation system, we present a novel approach to formulate the real-time traffic safety risk of individual drivers and present data-driven frameworks to predict the drivers' individualized safety risks. In particular, the models take advantage of near-crashes in addition to crashes and is capable of handling different types of variables. We first used the VTTI's 100-car Naturalistic Driving Study (NDS) data to develop an ensemble classifier to classify driving events into the crash and near-crash. We have then extended our methodology and developed a model for the Second Strategic Highway Research Program (SHRP-2) NDS data which is a more comprehensive study with more safety-related variables. Extensive data preparation and feature engineering were necessary to make data ready for model building. For the traffic safety risk prediction, we have used a weighted regularized regression model, to classify the trichotomous driving outcomes in relation to multi-stream safety data. We have further improved the resolution of the classes of driving outcomes by decomposing the class of normal driving. The developed prediction models can be used in advanced driver assistance systems to warn drivers of critical traffic incidents. We have also proposed a hybrid physics/data-driven approach to be used in a personalized kinematic-based Forward Collision Warning (FCW) system. In particular, we have used a hierarchical regularized regression model to estimate the driver's reaction time in relation to his/her individual characteristics, driving behavior and surrounding driving conditions. This personalized reaction time will be then plugged into the Brill's one-dimensional car-following model. We have also developed a simple rule-based algorithm to decide when to use the predicted values in a conservative FCW system.

For the healthcare system, we also develop a quantitative framework to identify the main sources of variation in patient flow. Since 1983, under Health Care Financing Administration (HCFA)'s system each hospital inpatient is classified into predefined Diagnosis-Related Groups (DRGs), and the hospital is paid the amount that HCFA has assigned to each DRG. In other words, irrespective of what the hospital charges for, it will be paid only a fixed price for each DRG through major reimbursement plans. Therefore, it is logical to expect that by reducing the within DRG discrepancies, hospitals can cut cost and improve patient safety and satisfaction. In order to reach this goal, the first step is to identify the main sources of variations. We have used a mixture of first-order n-step Markov models to cluster patients into similar groups and then applied the well-known random forest classifier to identify significant factors affecting the patient sequence among tens or hundreds of potential factors including patient profile and hospital-related variables. We illustrated the applicability of our proposed approach by using a simulated data based on a real-life case study.

ACKNOWLEDGEMENTS

The completion of this dissertation would not have been possible without the support of my family, professors and friends. Thank you so much to everyone who has helped me along my journey. So many people have been so kind. There are far too many to name but there are some people whom I would like to especially thank.

First and foremost, I would like to express my special appreciation and gratitude to my advisor Professor Mohsen Jafari, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on both research as well as on my career have been invaluable. You and Roxana are true friends and I am so grateful to have you both in my life!

I would also like to sincerely thank the members of my Ph.D. committee, Professor Myong K. Jeong, Dr. Kang Li and Dr. Jing Jin for taking the time to review my dissertation and for providing valuable and constructive suggestions. I would also like to thank the Department of Industrial and Systems Engineering and Center for Advanced Infrastructure and Transportation (CAIT) for their continuous support during my Ph.D. years. I would also like to thank Kian Seyed and Dr. Al Maghazehe for their support and insight during my internship at Capital Health Systems.

I gratefully acknowledge partial financial supports by a grant from the U.S. Department of Transportation, Office of the Secretary of Transportation (OST), Office of the Assistant Secretary for Research and Technology under Grant no. DTRT12-G-UTC16; and a grant from Qatar National Research Fund (QNRF) under Grant no. NPRP8-910-2-387.

Last but not least, special thanks to my beloved parents and my wonderful sister

Shabnam. Words cannot express how grateful I am to my family for their encouragement, patience and unconditional love. I am forever indebted to you for giving me the opportunities and experiences that have made me who I am today. You have selflessly encouraged me to explore new directions in life to find happiness and success. This journey would not have been possible without your love and support! I thank you and love you all!

TABLE OF CONTENTS

1 INTRODUCTION

Complex human-machine systems where human plays controlling roles are highly dynamic and complicated making the traditional models and methodologies less effective. The operability of such a complex system is affected by the performance and inter-relationships of a wide range of both internal and exogenous variables. The dynamic nature of such systems makes it necessary to apply probabilistic and stochastic models to capture the system variability. These variables or sources of variations can mainly be of three types: (i) known controllable variables, (ii) known uncontrollable variables and (iii) nuisance factors (sometime called lurking variables). Nuisance factors refer to un-assignable causes, which are unknown and therefore uncontrollable. Any significant reduction in uncontrollable variations will increase system capability and improve the process performance, which can be achieved by building a strong model given the values of known variables and controlling the controllable variables.

One example of such a complex system where humans have real-time control is transportation. In this system, multiple sources of variation can affect the driving outcome, categorized here into crash, near-crash, and normal driving. Driver behavior is known to be an essential safety factor in this system. In addition to driver behavior, roadway characteristics, vehicle condition, time of day, surrounding externalities (such as other roadway users), incidents (accidents, work-zones, etc.), and environmental conditions are other potentially contributing variables.

Another example of human-machine systems is the patient care process in hospitals. In these systems, apart from the fact that humans are subjects receiving the service, the physicians, nurses, lab technicians, and also hospital administrative staff, each has some level of control over the process, and their behavior must be considered in the models. In addition to the human factor, there are generally two sets of variables in patient care processes: (a) patient profile which are mainly uncontrollable variables such as patient's age, gender, medical history, and (semi-) controllable factors such as medication; (b) hospital related variables which are mostly controllable such as test turn-around times, physician practices, nurse level of expertise, etc.

Another commonality of the above-mentioned systems is the high risk of human errors or faults. In transportation, the risk of errors can be a crash with consequences ranging from property damage to fatality; similar to the healthcare system where the consequences range from loss of time and money due to unnecessary tests to loss of life. The ultimate goal is to reduce the risk of unfavorable events in these systems to save lives and reduce costs. One way to achieve this is through modeling and learning the relationships between potentially contributive factors and the process output. The main objective is to construct integrated frameworks by applying advanced data analytics, statistical and stochastics models and machine learning methods to extract important knowledge from complex human-machine systems for either prediction or causal analysis. The results can be used for both off-line design of better targeted countermeasures and corrective actions or on-line monitoring for situational awareness which can in turn assist with well-informed control actions.

1.1 Motivation

Our motivations behind this work are two-fold:

1) For the transportation system:

During the past decade, the demand for transportation services has increased remarkably due to steady increase in population coupled with strong economic growth and this increasing trend is likely to continue over the next 25 years [1]. About 1.24 million people die each year on the world's roads and between 20 and 50 million sustain non-fatal injuries. Studies show that road traffic injuries remain an important public health problem despite progress in a number of countries [2]. Among Americans aged 1 to 34, motor vehicle crashes are the leading cause of death. According to the National Highway Traffic Safety Administration (NHTSA), U.S. motor vehicle crashes in 2010 cost almost $1 trillion in loss of productivity and loss of life [3]. The report cites several behavioral factors, including drunk driving, speeding, distraction, and seat-belt use, as contributing to the huge price-tag of roadway crashes based on the 32,999 fatalities, 3.9 million non-fatal injuries, and 24 million damaged vehicles that took place in 2010 [4]. According to 2015 data released by the National Safety Council (NSC), the one-year percentage increase of the death toll in America reached its highest in half a century. Despite tremendous efforts to mitigate the risk of roadway crashes, the US is falling behind peer nations in traffic safety.

The good news is that technology is changing traffic safety and with that vehicle safety is progressing beyond basic seatbelts and lighting, to high-tech safety features that can help drivers avoid accidents altogether. Three distinct but related streams of technological change and development are occurring simultaneously:

- In-vehicle crash avoidance systems that provide warnings and/or limited automated control of safety functions, such as automated emergency braking systems, lane-departure and forward collision warning systems, and electronic stability control

- Connected vehicle technologies—vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications that support various crash avoidance applications; and
- Fully automated and self-driving vehicle technology.

The Insurance Institute for Highway Safety have estimated that if all vehicles had forward collision and lane departure warning, blind spot assist, and adaptive headlights, about 1 in 3 fatal crashes and 1 in 5 injury crashes could be prevented[1]. Automated driving innovations could dramatically decrease the number of crashes tied to human choices and behavior. Experts optimistically estimate that advanced vehicle technology can reduce the number of crashes by up to 90% by eliminating the primary cause of crashes that is the human error[2].

Although there will be a significant growth in the number of autonomous vehicles by 2030, non-autonomous cars will make at least 85% of the traffic mix[3]. Furthermore, in vehicles with less than full automation, the system can only drive the car under specific conditions, and still the human driver needs to be ready to take back control of the vehicle when necessary and drive under difficult conditions. Last but not least, combining autonomous and non-autonomous vehicles in a single traffic network will bring about

[1] The Insurance Institute for Highway Safety, New estimates of benefits of crash avoidance features on passenger vehicles, available from http://www.iihs.org/iihs/sr/statusreport/article/45/5/2

[2] Ten ways autonomous driving could redefine the automotive world, McKinsey & Company Podcast, June 2015; available from http://www.mckinsey.com/industries/automotive-and-assembly/our-insights/ten-ways-autonomous-driving-could-redefine-the-automotive-world.

[3] Self-driving Cars and The Future of the Auto Sector, McKinsey & Company Podcast, August 2016; available from http://www.mckinsey.com/industries/automotive-and-assembly/our-insights/self-driving-cars-and-the-future-of-the-auto-sector.

unimaginable traffic safety challenges and the most difficult time is expected to be the transition period, while all kinds of cars will share the road before self-driving ones predominate. *Therefore, it is imperative to enhance the performance of the present Driver Assistance Systems for the lower classes of vehicles to ensure a safe and smooth transition to the future of transportation.*

Naturalistic driving studies (NDS) are recent research projects intended to observe and record drivers' driving behavior as events happen in real time. The collected data from Naturalistic Driving Study projects provide interesting and useful informational data about driver behavior, road, vehicle, and weather and traffic conditions in case of a crash, near-crash or under normal driving conditions. The 2nd Strategic Highway Research Program (SHRP2)'s Naturalistic Driving Study is the largest of its kind whose data was released in 2015. We had the opportunity to obtain a portion of this dataset through a grant from the U.S. Department of Transportation, Office of the Secretary of Transportation (OST), Office of the Assistant Secretary for Research and Technology under Grant no. DTRT12-G-UTC16.

With this background in mind, we are motivated to propose novel frameworks to quantify and predict the real-time individualized traffic safety risk of drivers. We have used NDS data to illustrate the applicability of our proposed methodologies. In chapter 2, we propose a data-driven and learning-based approach to predict the likelihood of crash and near-crash events. The proposed prediction models can be used in an Advanced Driver Assistance System for situational awareness to primarily alert drivers of critical traffic incidents and unsafe situations. With the emerging trends in smart transportation and infrastructure, the widespread use of advanced technologies such as sensors, radars, cameras, smartphones, and on-board vehicular devices and advances in big data storage and analytics, recording and processing of the required data will be readily available.

According to the National Highway Traffic Safety Administration (NHTSA), rear-end collisions account for approximately 23% of all motor vehicle crashes. In 2012 alone, more than 1.7 million rear-end crashes occurred on US roadways, resulting in more than 1,700 fatalities and 500,000 injured people. The National Transportation Safety Board (NTSB) estimated that 80% of the deaths and injuries resulting from rear-end collisions could be prevented by collision avoidance systems.

An effective ADAS is expected to give a safety alert sometime before the driver realizes the presence of a rear-end collision's risk in the hope of shortening the response time and evading a crash. Therefore, the use of a personalized reaction time instead of an average value for all drivers and under any driving conditions will enhance the performance of the ADAS in issuing more timely alerts.

This has motivated us in chapter 3, to propose a hybrid physics/data-driven approach to be used in a kinematic-based Forward Collision Warning system. We propose a framework which can be used to customize the FCWS according to individual characteristics of drivers such as their demographics, cognitive abilities and risk taking/perception behavior. In particular, we have modeled the driver's *reaction time* in relation to his/her individual characteristics, driving behavior and surrounding driving conditions. This personalized estimated reaction time will be then plugged into the kinematics model to issue a collision warning.

2) For the healthcare system:

Since 1983, under Health Care Financing Administration (HCFA)'s system, generally referred to as the Prospective Payment System (PPS), each hospital inpatient is classified into one of around 500 Diagnosis-Related Groups (DRGs), and the hospital is paid the amount that HCFA has assigned to each DRG. Thus, hospitals will be paid the same amount for patients within a particular DRG. One limitation to this methodology is that individual DRG categories often combine subgroups of patients with predictably different expected resource costs. HCFA has repeatedly improved the DRG definitions since 1984 but these enhancements, while necessary, do not fully account for differences in illness severity associated with substantial disparities in providers' costs.

During their hospital stay, patients may experience redundant steps and procedures that may lead to unnecessary excessive expenses, lower Quality of Care (QoC) and customer dissatisfaction. The excessive costs are often covered by hospitals or paid by individual patients, since insurance companies have standard payment plans ranging from the infamous charge master or fee-for-service (FFS) price list to bundled payment systems such as diagnosis-related groups (DRGs) with various forms of "discounts off charges" and "per diems" somewhere in between. Regardless of who pays for these excessive and unnecessary expenses, the adverse societal impacts and negative business consequences are immense.

On the other hand, renewed focus on quality measurement and improvement and on medical-error reduction has heightened interest in paying for performance, rather than just reimbursing providers for services rendered. Private Pay for Performance (P4P) programs for hospitals usually pays bonuses as an incentive above the agreed-upon reimbursement rate. A more rational reimbursement system, which rewards quality of care rather than simply doing more to patients, is the short-term goal of paying for performance. The longer-term goal is also to make the health care system more efficient. It has become clear that under existing reimbursement structures, current market forces are insufficient to ensure either higher-quality or more-cost-effective care [5]. P4P programs can be seen as additional incentives for hospitals to seek to improve their patient flow processes, which can be attained through our variation reduction framework. These facts have motivated many researchers and practitioners to pay much attention into the development of novel technologies and methods for improving patient flow processes. In this study, we are motivated to develop a quantitative framework to identify the main sources of variation in patient flow using advanced stochastic models and machine learning methods.

1.2 Synopsis of Contributions

1.2.1 A Data-Driven Approach To Traffic Safety Risk Prediction (Chapter 2)

In chapter 2, the main objective is to build a traffic safety risk prediction model in relation to traffic safety factors. To do this, we propose building an integrated framework which uses data from multiple sources, extract relevant features and/or build new features to be used in an advanced statistical model or a machine learning algorithm for real-time traffic safety risk prediction. The application of the proposed platform is multifold:

1- The individualized traffic safety risk can be used in an Advanced Driver Assistance System (ADAS) for situational awareness to primarily alert drivers of critical traffic incidents and unsafe situations.

2- The aggregated traffic safety risks of a cohort of drivers over time and location can be integrated into a navigation system to help drivers make informed decisions by planning their trips through the safest routes.

3- It can be used in an advanced decision support system for roadway network owners to do network screening and hot-spot analysis using risk based measures.

In this study, we focus on the first application of this platform. We have used real-world datasets from two different Naturalistic Driving Studies (NDS), namely VTTI's 100-car[4] and The Second Strategic Highway Research Program (SHRP-2)[5] NDS data to illustrate the use of our proposed data-driven approach. In particular:

In **Part I**, we have used 100-car data to develop an ensemble of Breiman's random forest [6] and a newly proposed Multivariate Time Series Random Forest [7] to classify driving events into crash and near-crash classes in relation to a set of safety factors. The replicated k-fold cross validation is used to evaluate the models.

In **Part II**, we have extended our methodology and developed a model which can better fit SHRP-2 NDS data, a more comprehensive naturalistic driving study, with more extensive data fields (variables). First, data preparation and feature engineering steps were necessary to make the data ready for model building. For traffic safety risk prediction, we have used a weighted regularized multinomial regression model [8], to classify the driving outcomes in relation to multi-stream safety data. Our selection of this methodology is mainly motivated by its built-in mechanism for variable selection and ability for bias/variance trade-off. We have

[4] http://forums.vtti.vt.edu/index.php?/files/category/3-100-car-data/

[5] https://insight.shrp2nds.us/

further improved the resolution of the original trichotomous driving classes by decomposing the normal driving state according to driver behavior and secondary task involvement. The proposed prediction models can be used in a Basic or a Conservative driver assistance system, termed according to their sensitivity to critical events and unsafe driving situations. The former system alerts drivers of crashes and near-crashes while the latter system warns of unsafe and distracted driving situations as well.

1.2.2 A Hybrid Physics/Data-Driven Approach for a Personalized Forward Collision Warning System (Chapter 3)

In paper 3, we propose a hybrid physics/data-driven approach to be used in a kinematic-based Forward Collision Warning system. Our proposed approach utilizes both the laws of physics governing moving objects and the supplemental data explaining driver and his/her surrounding conditions to assess traffic safety risks. In particular, we have focused on an FCW system which uses Brill's one-dimensional car-following model to calculate the critical distance to issue a collision warning. The driver's reaction time is one of the main parameters of the Brill's model whose value defines whether a critical event would turn into a crash. It has been a common practice to use a nominal value, the mean or the 95th percentile of the reaction time distribution of participants in experimental studies or driving simulators.

In reality, it is well known that individuals react quite differently to the road events. There are many factors affecting a driver's reaction time yet unexplored by the driver modeling literature due to the lack of sufficient observational data. To close this gap in the literature, we propose building a hierarchical regression model on top of the kinematic model, which can capture the variations attributed to driver characteristics and driving behavior. We use SHRP-2's Naturalistic Driving Study (NDS) data [9], the largest and most comprehensive study of its kind, to model the driver's brake-to-stop response time. The results show that the inclusion of driver characteristics improves the predictive performance of the reaction time model. The explained variation by the driver's intrinsic characteristics and driving behavior supports the necessity for developing personalized Advanced Driver Assistance Systems to enhance the performance and increase their acceptance by drivers. We have also proposed a simple rule-based algorithm to decide when to use the predicted values by our proposed reaction time prediction model in a conservative FCW system.

1.2.3 Modeling And Clustering Patient Pathways (Chapter 4)

In this chapter, we propose a novel framework to model patient flow and relate them to system covariates for the purpose of process improvement. To do so, we have used a mixture of first-order Markov models to cluster patients into similar groups. Next, we applied the well-known random forest classifier to identify significant factors affecting the patient sequence among tens or hundreds of potential factors including patient profile and hospital-related variables. The idea is that by monitoring and controlling the important factors we will be able to control the variation in patient pathways which is interpreted as a process improvement. We will illustrate the applicability of our proposed approach by using a simulated data based on a real-life case study. The DRG under study was chest pain and the collected data includes patient pathways and their related variables. Due to limitations in data collection, we used this sample data to generate more simulated patients and used them as inputs to our method.

2 A DATA-DRIVEN APPROACH TO TRAFFIC SAFETY RISK PREDICTION

2.1 Introduction

With the emerging trends in smart transportation and infrastructure, the widespread use of advanced technology such as sensors, radars, cameras, and on-board vehicular devices and advances in big data storage and analytics, onboard recording and processing of real-time driving data will be readily available. It will soon be possible to aggregate the traffic safety related data from these sources, overlaid over time and location for a specific driver, in order to create a high-resolution insight into the driving events. Furthermore, these real-time driving data can be merged with supplementary real-time network and weather data, and slow-changing data on driver, vehicle and roadway to build a holistic view of events and consequential behavioral patterns and safety factors.

In this chapter, we define the real-time individualized traffic safety risk as the likelihood of a crash or near-crash and model its relationship to safety factors using advanced statistical and/or machine learning methods. The proposed approach is a data-driven and learning-based algorithm and its performance is expected to improve as the sample size increases and quality of data improves. The prediction model can be used in an Advanced Driver Assistance System for situational awareness to primarily alert drivers of critical traffic incidents and unsafe situations. We have used real-world datasets from two different Naturalistic Driving Studies (NDS), namely VTTI's 100-car[6] and The Second Strategic Highway Research Program (SHRP-2)[7] NDS data to illustrate the use of our

[6] http://forums.vtti.vt.edu/index.php?/files/category/3-100-car-data/

[7] https://insight.shrp2nds.us/

proposed approach.

The organization of this chapter is as follows: In section 2.2, we will first present the background and literature review of the traffic safety risk prediction models. Next, the general problem statement will be presented in section 2.3. Then, the chapter will be divided into two major parts according to the use of the above-mentioned NDS datasets. Each part will contain a particular problem formulation, model evaluation and numerical results, and a conclusion.

In Part I, we have used 100-car data to develop an ensemble of Breiman's random forest [6] and a newly proposed Multivariate Time Series Random Forest [7] to classify driving events into crash and near-crash classes in relation to a set of safety factors. The replicated k-fold cross validation is used to evaluate the models. This is a relatively short part which presents some preliminary results using a small dataset.

In Part II, we have extended our methodology and developed a model which can better fit SHRP-2 NDS data which is a more comprehensive naturalistic driving study with many more data fields (variables). First, data preparation and feature engineering steps were necessary to make the data ready for model building. For traffic safety risk prediction, we have used the elastic net regularized regression model [8], to classify the driving outcomes in relation to multi-stream safety data. Our selection of this methodology is mainly motivated by its built-in mechanism for variable selection and ability for bias/variance trade-off. We have also improved the resolution of the classes of the driving outcome by decomposing the class of normal driving. The results can be used in a Basic or a Conservative alerting system, termed according to their sensitivity to critical events and unsafe driving situations. The former system alerts drivers of crashes and near-crashes while the latter system warns of unsafe and distracted driving situations as well. Finally, this part is concluded with a summary of numerical results and a discussion of possible model improvements and future works.

2.2 Background and Literature Review

According to the National Highway Traffic Safety Administration (NHTSA), U.S. motor vehicle crashes in 2010 cost almost $1 trillion in loss of productivity and loss of life [3]. The report cites several behavioral factors, including: drunk driving, speeding, distraction, and seat-belt use, as main contributors to the huge price-tag of roadway crashes based on the 32,999 fatalities, 3.9 million non-fatal injuries, and 24 million damaged vehicles that took place in 2010 [4]. Tremendous efforts have been taken to mitigate the risk of roadway conflicts in order to alleviate the negative socio-economic impacts of roadway crashes including: traditional reactive and systematic approach to safety planning; adopting proactive countermeasures such as safe corridors, stricter laws for alcohol and under age driving; new strategies such as variable speed limits (VSL) due to advances in real-time data collection capabilities; and designing safe cars with different crash-avoidance warning systems. With the enormous advances in connected vehicles technology and the Internet of Things (IoT), new game changer solutions are appearing, and the opportunities for more advanced safety techniques are becoming more realizable.

In order to enhance the overall safety in roadway networks safety management approaches must focus more on individual driver's behavior [10]. Driver's behavior is a major contributor to traffic safety risks. A number of studies in the US report that approximately 90% of the light-vehicle crashes involved same type of human error such as impaired conditions, inadvertent errors and risky driving behavior [11], [12]. Driver error is also a main reason for approximately 87% of all commercial vehicles crashes [13]. Similar studies in other countries, including Japan, report similar conclusions; for instance, 40% of accidents in Japan are attributed to judgment errors, 47% to cognitive errors and 13% to operation errors [14]. Clearly, safety risks mitigation strategies can only be effective if driver behavior along with external conditions and factors, including weather, roadway conditions, time of day, traffic flow and density, together with their interactions are all accounted for. Advanced technology in image processing and IoT can certainly play a major role in such a holistic risk assessment. On another note, real-time crash risk prediction models using traffic data collected from loop detector stations have been proposed for dynamic safety management systems aimed at improving traffic safety through application of proactive countermeasures. The premise of the proactive traffic management is that there are certain freeway traffic patterns that are associated with a high likelihood of crash occurrence and that they may be detectable in the loop detector data [15]. Traffic detectors, singly or in combination, can be used to measure real-time variables such as presence, volume, speed, and occupancy.

Here we group traffic safety models into the following categories: (i) Systematic models that use historical crash data (mainly produced from accident reports) in conjunction with roadway information data such as the New Jersey Department of Transportation's Straight Line Diagrams (SLD) or similar legacy databases. These models range from simple crash frequency models ([16], [17]**,** [18]) to more advanced Poisson regression Poisson-Gamma or Poisson-lognormal/Negative Binomial ([19], [20], [21]) and Poisson and Negative Binomial Zero-inflated models [22]. A list of potential problems and methodological issues are available in [23] and [24]. From a practical point of view, Safety Performance Function (SPF) [25], which uses expected average crash frequency, is widely used in the US along with the appropriate adjusting Crash Modification Factors (*CMF*). (ii) Qualitative risk based on systemic safety models which focus on similar geometric features of roadway segments in which specific crash types have occurred [26]. The main advantage of this approach is revealing the site features which are directly associated with high crash risk and implementing countermeasures before experiencing several crashes. (iii) Advanced risk based models aiming at more rigorous and proactive safety mitigation strategies, and fueled by advances in sensing and data collection. In this category we are particularly interested in those models that use real time onboard data combined with data from other sources. In [27], the authors calculate risk of visual distractions by calculating the rate of gaze in some specific spots of the road (mainly road center) using a simple eye tracker and a mono-camera system. A study conducted in Germany [28] concludes that the main causes for personal injury crashes may significantly be different for different ranges of age. For example, turning errors are the most contributing causes for elderly people's crashes, while inappropriate speed is the most important cause for personal injury crashes of younger drivers. Their study suggests using age-specific safety assistant devices for different ranges of people. An Australian team conducts a simulator-based study to investigate the effect of driver inattention in increasing the driver safety risk. Their system is able to identify multiple "at risk" mental stats such as daydreaming and fatigue [29].

Naturalistic driving studies (NDS) are recent research projects intended to observe and record drivers' driving behavior as events happen in real time. The 100-Car Naturalistic

Driving Study, sponsored by the National Highway Traffic Safety Administration (NHTSA) and the Virginia Department of Transportation (VDOT), was the first instrumented-vehicle study undertaken with the primary purpose of collecting large-scale, naturalistic driving data in the US. The 100-Car Study was followed by a larger and more comprehensive study, the Strategic Highway Research Program 2 (SHRP2) conducted from 2006 to 2015.

In these studies, drivers were given no special instructions, no experimenter was present, and the data collection instrumentation was unobtrusive. The collected data from Naturalistic Driving Study projects provide interesting and useful informational data about driver behavior, road, vehicle, and weather and traffic conditions in case of either crash or near-crash events. These studies are our opportunity to better understand crash causality by supplementing crash observations with a much larger number of near-crash events.

A near-crash can be defined as a conflict situation requiring a rapid, severe evasive maneuver to avoid a crash [30]. In the 100-car study, the near-crashes were detected through a two-step data reduction process. First, events were identified using predefined trigger criteria values that resulted in a low miss-rate and a high false alarm rate to avoid missing valid events. The rule was if the value of at least one of the trigger criteria, namely, *Lateral Acceleration*, *Longitudinal Acceleration*, forward *Time-To-Collision (TTC)*, rear *Time-To-Collision (TTC)* or *Yaw Rate*, violated a threshold, or the *Event Button* activated by the driver, then a near-crash was detected. Reference [30] presents detailed information about each of the above mentioned trigger criteria, their definitions, descriptions and threshold values, in different Naturalistic Driving Studies. Second, the video data for all the identified events were reviewed by data reductionists to validate the event, determine severity, and code the event for a data reduction dictionary [31].

The analysis of NDS data reveals correlations between driver behaviors, roadway segment and weather conditions in either crash/near or normal situation ([32], [33], [34], [35]). According to FHWA's HSIP, the frequency of traffic conflicts is sometimes used as a rough proxy for safety. Data from the NDS [34] have been used to evaluate the causes of these conflicts and their relationship with actual crashes. One study finds strong relationship between the frequencies of contributing factors for crashes and for near-crashes, and that the combined crash and near-crash data increase the precision of the estimates [36]. In a SHRP2 study [37] the authors attempt to determine if crash surrogates can be related to actual crashes and use a Bayesian Seemingly Unrelated Regression (SUR) to capture the correlation structure between crashes and crash surrogates and estimate relative risk (RR) between the two. The model can identify the safety factors, which have the same impact on both crashes and near-crashes. This analysis was not exhaustive, and was conducted as an exemplar of the method.

Some has also suggested prediction of real-time risk of crashes using loop detector data. These works estimate the likelihood of crash occurrence for a given freeway segment over a short time period without taking into account the driver's personalized safety factors contributing to a crash. They rather warn roadway drivers, entering a specific highway segment, about a high risk of a potential roadway conflict by using traffic flow data. Some of the models include a limited number of roadway characteristics but they lack the driver behavior data and target vehicle information ([38], [39], [40], [41], [42] and [43]). The use of sequential logit model to link the likelihood of crash occurrences at different severity levels to various traffic flow characteristics derived from detector data was presented in [30].

In this chapter, we present a novel approach to formulate the real-time traffic safety risk of individual drivers and data-driven models to predict the individualized safety risks. In contrast to the traditional traffic safety models, our approach can potentially include different types of safety factors as mentioned above. In particular, it takes advantage of near-crashes in addition to the traditional crashes and is capable of handling different types of variables. The details of the model formulation will be explained in Section 2.3 and two different prediction models, namely the ensemble classifier and the elastic net, presented in Part I and II, respectively.

2.3 Problem Formulation

We denote the driving outcome by y_{ijlt} for driver, i on his/her trip j at location l, at time $t \in \mathbb{R}$. As it can be seen, y_{ijlt} has both time and location dimensions and is driver specific. We split the driving outcome's spectrum into discrete categories and assume that Y_{ijlt} has a categorical distribution. A categorical distribution, also known as a generalized Bernoulli distribution, is a probability distribution that describes the possible results of a random event that can take on one of the C possible outcomes, with the probability of each outcome separately specified. We denote these class probabilities by $\pi_1, \pi_2, \dots, \pi_C$. These probabilities are constrained only by the fact that each must be in the range zero to one, and all must sum to one. An example would be a trichotomous Y_{ijlt} with classes of normal driving, near-crash and crash and class labels of 1, 2 and 3, respectively:

$$Y_{ijlt} = \begin{cases} 1, & \textit{If the state is normal driving} \\ 2, & \textit{If the state is a near crash} \\ 3, & \textit{If the state is a crash} \end{cases} \tag{2.1}$$

Later in this chapter, we will improve the state resolution by further decomposing the class of normal driving. In general, one can define a near-continuum spectrum of colors as shown in Figure 2.1, where each general state category is associated with many sub-states or colors. This spectrum ranges from the safest mode (a near-zero chance of conflicts) to the riskiest mode of driving (a major fatal crash), and in the mid-range there will be mild to significant chances of near-crashes.

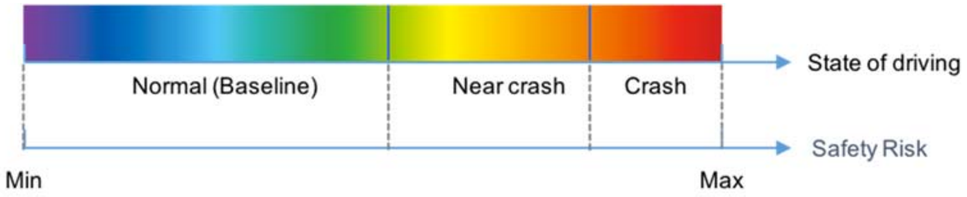


Figure 2.1 Spectral driving outcome.

There can be different approaches to map the driving outcome to the safety risk. In this chapter, we will define the safety risk as the likelihood of critical events such as near-crashes and crashes. We denote the vector of independent safety factors of length p by $\boldsymbol{z}_{ijlt}$ and define it as follows:

$$z_{ijlt} = (xD_{ij\circ\circ}, xV_{ij\circ\circ}, x_{ijlt}, u_{ijlt}, xC_{ijlt}) \tag{2.2}$$

where $\boldsymbol{x}D_{ij\circ\circ}$ and $\boldsymbol{x}V_{j\circ\circ\circ}$ are the static safety factors related to the characteristics of the target driver and vehicle, respectively. In this study, we use the term *static* to refer to both the invariant factors such as a driver's gender or a vehicle's make and model; and to the low-frequency changing factors, such as a driver's driving experience or a vehicle's maintenance condition which remains constant during a certain trip but may change during longer time intervals of a month, season or year and possibly from one trip to another. Vectors $\boldsymbol{x}_{ijlt}$ and $\boldsymbol{u}_{ijlt}$ contain the vehicle's kinematics and control variables similar to the state and input vectors in the state-space representation of physical systems, respectively. Finally, $\boldsymbol{xC}_{ijlt}$ contains all the other variables defining the context of driving, i.e. driver's dynamic behavior, engineering and dynamic features of the roadway network, weather, time and historical crash data. It is worth noting that an engineering feature of a roadway network (for example a traffic sign) is a static or very low-frequency changing factor from the view of a network owner but is apparently a dynamic real-time factor from the view of an individual driver traveling on this network (the sign appears to the driver at a certain location and time during a trip).

Suppose that we have a sample dataset of pairs of $(\boldsymbol{z}_{ijlt}, \boldsymbol{y}_{ijlt})$. We want to find a function, f, that can best model the relationship between the vector of safety predictors, $\boldsymbol{z}_{ijlt}$, and the outcome of driving, $\boldsymbol{y}_{ijlt}$:

$$f : z_{ijlt} \rightarrow y_{ijlt}$$

For simplicity purposes and in practice for storage capacity considerations, the continuous time t, can be replaced by kT_s, where k is greater than 0 and T_s is a constant time step (sample time). Some of the safety factors such as a vehicle's kinematic variables (for example speed and acceleration) are high-frequency time-varying variables while other factors such as weather, number of travel lanes, traffic signs or traffic flow changes less frequently and can be assumed to be invariant during a small time horizon, NT_S, where N is the number of time steps in the horizon or length of the time-series. Thus, according to the frequency of real-time changes, we divide safety variables into *time-series* and *event* variables and denote them as follows:

$$T_{ij(l,L),(k,N)} = \left(\begin{pmatrix} \vec{x}_{ij,l-L,k-N} \\ \vec{u}_{ij,l-L,k-N} \end{pmatrix}'_{1\times p_1}, \cdots, \begin{pmatrix} \vec{x}_{ijlt} \\ \vec{u}_{ijlt} \end{pmatrix}'_{1\times p_1} \right)_{N\times p_1} \tag{2.3}$$

$$ev_{ij(l,L),(k,N)} = xC_{ij(l,L),(k,N)1\times p_2} \tag{2.4}$$

Given the above discrete-time representation and the new categorization of safety factors into time series and event variables, we redefine the vector of safety predictors, $\boldsymbol{z}_{ijlt}$, to an array of vectors and scalars, $\boldsymbol{z}_{ij(l,L),(k,N)}$, as follows:

$$\begin{gathered} z_{ij(l,L),(k,N)} = \\ \left(xD_{ij\circ\circ}, xV_{ij\circ\circ}, \begin{pmatrix} x_{ij,l-L,k-N} \\ \vdots \\ x_{ijlt} \end{pmatrix}, \cdots, \begin{pmatrix} u_{ij,l-L,k-N} \\ \vdots \\ u_{ijlt} \end{pmatrix}, xC_{ij(l,L),(k,N)} \right) = \\ \left(xD_{ij\circ\circ}, xV_{ij\circ\circ}, T_{ij(l,L),(k,N)}, ev_{ij(l,L),(k,N)} \right) \end{gathered} \tag{2.5}$$

As mentioned before, we define the traffic safety risk as the likelihood of an unfavorable outcome and compute it for driver i, at trip j, location *l* and time step k from the following conditional probability:

$$P_a = \Pr(y_{ijkl} = a \mid z_{ij(l,L),(k,N)}) = f(z_{ij(l,L),(k,N)}); \ for \ a \in S_{Cr} \tag{2.6}$$

where S_{Cr} is the set of critical outcomes. In this chapter, f is a data-driven function. The predicted driving outcome given $\boldsymbol{z}_{ij(l,L),(k,N)}$ can be found as follows:

$$\hat{y}_{ijkl} = \underset{a=1,2,\cdots,C}{Arg\max}\, P_a \tag{2.7}$$

As we mentioned in section 2.1, from this point forward, the chapter is divided into two parts each presenting a data-driven prediction model to estimate the real-time traffic safety risk of individual drivers. In Part I, the ensemble classifier for the VTTI’s 100-car data will be presented followed by the elastic net model for the SHRP-2 NDS data in Part II.

Part I- Traffic Safety Risk Prediction Using 100-car NDS data

In this part, we explain the details of the ensemble classifier to calculate $\hat{y}_{ijkl}$, the predicted class of the driving outcome, in Equation (2.7). The 100-car NDS data is used to illustrate the applicability of the proposed model.

2.4 Ensemble classifiers

Ensemble learning is primarily used to improve the (classification, prediction, function approximation, etc.) performance of a model, or reduce the likelihood of an unfortunate selection of a poor one. An ensemble-based system is obtained by combining diverse models (henceforth classifiers). In order to fully and practically appreciate the importance of using multiple classifier systems, it is perhaps instructive to look at a psychological backdrop to this otherwise statistically sound argument: we use such an approach routinely in our daily lives by asking the opinions of several experts before making a decision. For example, we typically ask the opinions of several doctors before agreeing to a medical procedure, we read user reviews before purchasing an item (particularly big ticket items), we evaluate future employees by checking their references, etc. In each case, a final decision is made by combining the individual decisions of several experts [44].

There are several scenarios where using an ensemble-based system makes statistical sense; data fusion is one of them [44]. In many applications that call for automated decision-making, it is not unusual to receive data obtained from different sources that may provide complementary information. A suitable combination of such information is known as data or information fusion, and can lead to improved accuracy of the classification decision compared to a decision based on any of the individual data sources alone. Sometimes, heterogeneous features, such as the *time-series* and *event* variables in Equations (2.3) and (2.4), cannot be used all together to train a single classifier (and even if they could such a training is unlikely to be successful). In such cases, an ensemble of classifiers can be used where a separate classifier is trained on each of the feature sets independently. The decisions made by each classifier can then be combined by a combination rule.

Two different types of combining rules exist:

1. Algebraic (or fixed),
2. Trained.

Algebraic rules combine the continuous-valued output of classifiers through an algebraic expression, such as minimum, maximum, mean, product, median, majority vote, etc. In each case, the final ensemble decision is class *j* that receives the largest support after the algebraic expression is applied to individual supports obtained by each class. Following our earlier notations, P_{ia} the conditional probability of output j obtained by using classifier M_i can be shown as follows:

$$P_{ia} = \Pr\{Y_{ijkl} = a \mid \boldsymbol{T}_{\boldsymbol{ij(l,L),(k,N)}}, \boldsymbol{ev}_{\boldsymbol{ij(l,L),(k,N)}}, M_i\};\ \mathrm{a}\ \in \mathrm{S}_{Cr}\ \text{and i} = 1, 2, \dots, \mathrm{m} \qquad (2.8)$$

where M_i is classifier *i*, and *m* is the number of classifiers. Then, P_a in Equation (2.6) can be calculated as follows:

$$P_a = \Pr\{Y_{ijkl} = a \mid \boldsymbol{T}_{\boldsymbol{ij(l,L),(k,N)}}, \boldsymbol{ev}_{\boldsymbol{ij(l,L),(k,N)}}\} = \frac{\underset{i}{\mathrm{rule}}\, P_{ia}}{\sum_c \underset{i}{\mathrm{rule}}\, P_{ic}} \qquad (2.9)$$

where $\underset{i}{rule}\, P_{ia}$ gives the combined value of continuous-valued output a over all methods (classifiers). Then, the class of event scenario $\{\boldsymbol{T}_{\boldsymbol{ij(l,L),(k,N)}}, \boldsymbol{ev}_{\boldsymbol{ij(l,L),(k,N)}}\}$ can be calculated from Equation (2.7). On the other hand, one can train an arbitrary classifier using the values of P_{ia} (for all i and a) as features in the intermediate space. Then the combining rule is called a trained rule. It is a point of discussion whether it is wise to use the posterior probabilities directly for building the intermediate feature space, and it is beyond the scope of this study to investigate it. We have only used fixed combing rules in this study.

2.5 Classification Models

As we explained in section 2.3, we have two different types of safety variables in according to the frequency of real-time changes:

- *Time-series* variables, $T_{ij(l,L),(k,N)}$,
- *Event* variables, $ev_{ij(l,L),(k,N)}$.

In a recent study, Jafari et al. [45] used multinomial logistic regression (MLR) on *Event* variables of VTTI's 100-car NDS data to classify the driving scenarios. In this study, we propose using Breiman's random forest (RF) [6] to classify *Event* variables. We have conducted a replicated cross validation to test and compare the performances of MLR and RF methods in classifying *Event* data. Furthermore, we have used a recent generalization of random forests for multivariate time series by Baydogan et al. [7] to classify the *time-series* variables.

2.5.1 Classification of Event Data

First, we briefly introduce the two classification methods: 1-MLR, 2- RF for classifying *Event* data, and then present the performance comparison results.

Multinomial Logistic Regression

Logistic regression technique is designed to estimate the parameters of a multivariate explanatory model where the dependent variable is dichotomous (binary), and the independent variables are continuous or categorical [46]. Multinomial logistic regression is an extension of logistic regression to multiple outcome categories [47]. This fits to our trichotomous driving outcome, namely normal-driving (Baseline), near-crash and crash. The predicted values from the analysis can be interpreted as probabilities of membership to the target groups.

Multinomial logistic regression uses a linear predictor function to predict the probability that observation i has outcome k. It has the following form:

$$\text{logit}\,\Pr\left\{Y_{ijlt} = a\,|\boldsymbol{ev}_{\boldsymbol{ij}(\boldsymbol{l},\boldsymbol{L}),(\boldsymbol{k},\boldsymbol{N})}\right\} = {\beta_0}^{a} + \beta^{a}.\,ev_{\boldsymbol{ij}(\boldsymbol{l},\boldsymbol{L}),(\boldsymbol{k},\boldsymbol{N})} \qquad (2.10)$$

where $({\beta_0}^{a}, \beta^{a})$ is the vector of regression coefficients associated with outcome a. For C possible outcomes, C-1 independent binary logistic regression models were built, so that one outcome is chosen as a pivot (reference category) and the other C-1 outcomes are separately regressed against the pivot outcome.

When applying an MLR model, there is no need for the independent variables to have specific probability distributions or to be statistically independent from each other; however, collinearity is assumed to be relatively low, as it becomes difficult to differentiate between the impacts of several variables if they are highly correlated. Furthermore, MLR classifiers provide linear decision boundaries. Therefore, in cases where the above-mentioned assumptions violate, the performance of MLR classifier deteriorates.

Random Forest

Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees [48]. Regression trees assume a model of the form:

$$f(x) = \sum_{m=1}^{M} c_m . 1(X \in R_m) \tag{2.11}$$

Where $R_1, \dots, R_M$ represent a partition of feature space, depicted in Figure 2.2-a and Figure 2.2-b. Trees are invariant under scaling and various other transformations of feature values, are robust to inclusion of irrelevant features, and produces inspectable models. However, they are seldom accurate. In particular, trees that are grown very deep tend to learn highly irregular patterns: they overfit their training sets, because they have low bias, but very high variance.

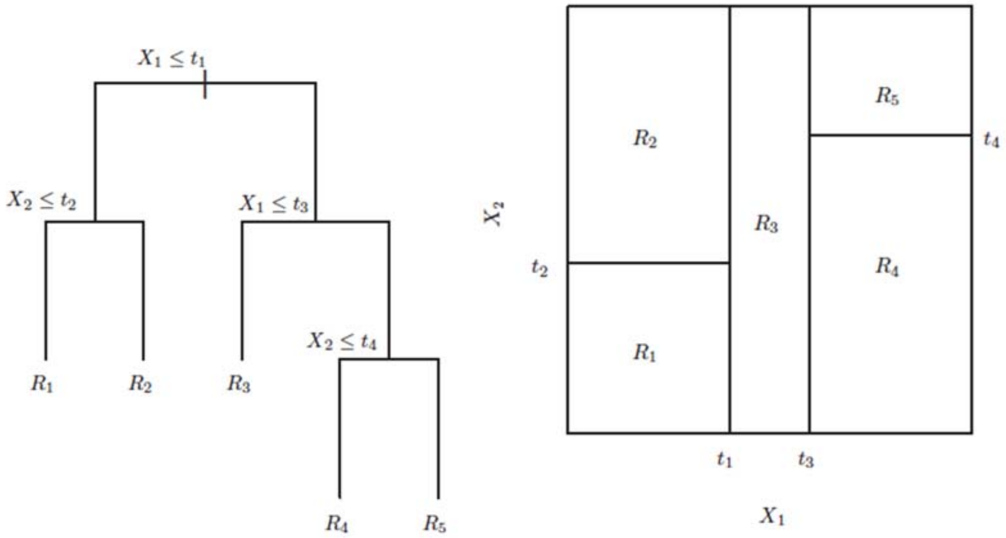


Figure 2.2 a: A tree corresponding to the partition of two-dimensional feature space. b: The partition of the two-dimensional example in (a) **[48]**.

Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally it greatly boosts the performance of the final model. Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

Some important features of Random Forests are as follows:

- It is unexcelled in accuracy among current algorithms.
- It runs efficiently on large databases.
- It can handle thousands of input variables without variable deletion.
- It gives estimates of what variables are important in the classification.
- It generates an internal unbiased estimate of the generalization error as the forest building progresses.
- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.
- It has methods for balancing error in class population unbalanced data sets.
- Generated forests can be saved for future use on other data.
- Prototypes are computed that give information about the relation between the variables and the classification.
- It computes proximities between pairs of cases that can be used in clustering, locating outliers, or (by scaling) give interesting views of the data.

- The capabilities of the above can be extended to unlabeled data, leading to unsupervised clustering, data views and outlier detection.
- It offers an experimental method for detecting variable interactions.

Method Selection Using Cross-Validation

We used k-fold Cross Validation (CV) to assess the prediction performance of MLR and RF classifiers on *Event* data. Cross-validation is a way to predict the fit of a model to a hypothetical validation set when an explicit validation set is not available which is true in our problem. In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k - 1$ subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged to produce a single estimation. k is an unfixed parameter, and its best value can be determined through experiments.

As we mentioned before, crashes are rare events and even near-crashes occur less frequently compared to the baseline (normal-driving) events. As a result, the traffic safety data is highly imbalanced, i.e. the classification categories are not represented approximately equally. For example, in the 100-car NDB data, the proportions of classification categories are 68 crashes to 760 near-crashes to about 19,000 baseline events. In order to take this into account in performing the CV tests, we used stratified k-fold cross-validation. In stratified k-fold cross-validation, the folds are selected so that the mean response value is approximately equal in all the folds. In the case of a classification problem, this means that each fold contains roughly the same proportions of each class labels. We performed both k-fold CV and stratified k-fold CV to test the performance of MLR and RF on NB's *event* data and presented the results in section 2.6.

2.5.2 Classification Of Time-Series Data

Multivariate time series (MTS) classification has gained importance with the increase in the number of temporal datasets in different domains (such as medicine, finance, multimedia, etc.) [7]. Similarity-based approaches, such as nearest-neighbor classifiers, are often used for univariate time series, but MTS are characterized not only by individual attributes, but also by their relationships. In this study, we use a Multivariate Time Series Random Forest to classify events according to time series variables.

Multivariate Time Series Random Forest (MTS-RF)

Baydogan and Runger (2014) provide a classifier based on a new symbolic representation for MTS (denoted as SMTS). SMTS considers all attributes of MTS simultaneously, rather than separately, to extract information contained in the relationships. Here, an equivalent formulation of our time series classification problem according to [7] follows: $\boldsymbol{T}_{\boldsymbol{ij(l,L),(k,N)}}$is a p_1-attribute time series each of which has N observations where t_m^n is the mth attribute (safety factor) of series n and $t_m^n(k)$ denotes the observation at time step k. Time series can be of different sizes and MTS-RF handles this situation, but for the purpose of illustration here we assume time series to be of the same length. MTS example $\boldsymbol{T^n}$ is represented by a $N \times p_1$ matrix as:

$$\boldsymbol{T}^n = \left[t_1^n, t_1^n, \dots, t_m^n, \dots, t_{p_1}^n\right] \tag{2.12}$$

where

$$t_m^n = [t_m^n(k-N), \dots, t_m^n(k)] \tag{2.13}$$

is the time series in column m. There are N' training MTS, each of which is associated with a class label $Y_n \in \{1, 2, \dots, C\}$ for n = 1, 2, . . . , N'. Given a set of unlabeled MTS, the task is to map each MTS to one of the predefined classes. Instead of extracting features from each time series, each row of $\boldsymbol{T}^n$ is considered to be an instance. This is achieved by creating a matrix of instances $D_{N'N \times p_1}$:

$$D_{N'N \times p_1} = \begin{bmatrix} t_1^1 & t_2^1 & \cdots & t_{p_1}^1 \\ t_1^2 & t_2^2 & & t_{p_1}^2 \\ \vdots & & \ddots & \vdots \\ t_1^{N'} & t_2^{N'} & \cdots & t_{p_1}^{N'} \end{bmatrix} \tag{2.14}$$

Equation (2.14) is basically the concatenation of training examples $\boldsymbol{T}^n$ in Equation (2.12). We assign the label of each instance to be the same as the time series. Then, $D_{N\boldsymbol{N} \times p_1}$ is mapped to the feature space $\Phi_{N'\boldsymbol{N} \times (2p_1+1)}$ that adds the following columns: time index, first differences for each numerical attribute. The row of Φ for series n at time step k is

$$\left[k, t_1^n(k), t_1^n(k) - t_1^n(k-1), \dots, t_{p_1}^n(k), t_{p_1}^n(k) - t_{p_1}^n(k-1)\,\right] \tag{2.15}$$

The differences provide trend information. A tree learner can capture this information if it relates to the class. If an attribute is nominal, first differences are not included. A RF tree learner is trained on Φ assuming that each instance has the same class label as its time series. Each tree of RFins (RF applied to the instances) provides a symbolic representation for the time series. Because time is used as a predictor variable, and because RFs can effectively handle interactions, complex regions in two-dimensional signal space (S) where one class dominates can be detected. In this sense, the time ordering of the data is used. RF Ensemble provides a symbolic representation, which includes different views of the same time series mapping them to the high-dimensional space of terminal nodes (that correspond to regions in S where one class dominates).

After the symbolic representation is generated from the trees in RFins, a bag-of-words (BoW) approach is used to classify the time series. Each symbol is simply considered to be a word and the relative frequency vector of the symbols from each tree are concatenated and used to classify the time series by the second RF ensemble. This frequency vector from each tree is normalized by the number of instances in the time series to obtain the relative frequency vector. We refer the interested reader to [7] for more details of the method. The authors have made the codes accessible for researchers through: http://www.mustafabaydogan.com/multivariate-time-series-discretization-for-classification.html.

2.6 Numerical Results

In order to examine the performance of our methodology, we used VTTI's naturalistic driver behavior data (http://www.vtti.vt.edu/). VTTI data has been collected over a course of 18-month period. The data collection effort resulted in approximately 2,000,000 vehicle miles of driving, almost 43,000 hours of data, 241 primary and secondary driver participants, 12 to 13-month data collection period for each vehicle, five channels of video and many vehicle state and kinematic variables. Two databases were created: the *event* database, and the *baseline* database. The former database consists of crash and near-crash events while the latter one consists of normal driving incidents. The number of epochs selected per vehicle in the baseline database is proportional to the number of vehicle involvement in crashes or near-crashes.

NDB database includes two major types of data: *time series* data and *event* or *video-*

reduced data. *Time series* data include direct readings from on board devices, such as radars, sensors, and accelerometers. This data was available for 68 crashes and 760 near-crashes. For each driving event, the dataset contains time series variables, for example gas pedal position and speed vehicle composite, spanning 30s before and 10s after an event. *Video reduced* data contains detailed event, driver state, and driving environment information derived from video reduction for the same 68 crashes and 760 near-crashes. Time series variables are not yet available for baseline events. Therefore, we present the numerical results of our general model for the dichotomous problem of crash and near-crashes.

Crash and near-crash events in naturalistic driving are typically identified through the detection of unusual vehicle kinematics recorded electronically through accelerometers and gyroscopic sensors. The driver may also highlight a driving event by using an "event" button located in the vehicle for this purpose. Kinematics measures such as forward and rear Time To Collision (TTC) can be used with vehicle kinematics (including measurements of a target vehicle) to identify additional events. Once identified kinematically, the events are reviewed through use of forward and face video. They are retained if verified as safety-related events and discarded if not. Within each event, factors that precipitated the event, that contributed to the event, and that were associated with the event are grouped into pre-event maneuvers, precipitating factors, contributing factors, associated factors, and avoidance maneuvers. The event begins at the onset of the precipitating factors and ends after the evasive maneuvers. Data for the period shortly before, during and shortly after the event are then preserved.

In addition to the kinematic variables discussed above, there are three other sets of data routinely collected in naturalistic driving studies:

1. Context variables – these are descriptors of the physical features, such as road and environment, at the time of the event including geometric alignment and environmental factors (e.g. rain or snow; day or night). Some geometric features may be obtained by linking on-board GPS to existing geographic information systems (e.g. roadway inventory systems maintained by most state highway departments).

2. Event attributes - attributes of the event occurring immediately prior to and during event occurrence. Examples include the occurrence of driver distraction (sometimes identified by type of distraction) and presence of fatigue.

3. Driver attributes - typically obtained during subject intake to the study and may include age, stated prior driving record, propensity to take risks when driving and physiological conditions such as vision and reaction time.

Table 2.1 presents the list of the 25 (p=25) variables included in our model. From this set, fifteen factors are *event* variables (p_1=15) including 5 driver-related, 2 environmental-conditions, 6 roadway-characteristics, and 2 surrounding-externalities variables. We also considered 10 time-series variables (p_2=10) in the model, which are all driver-related safety factors except for *Lighting* explaining an externality. This table also shows the source and type of each variable.

Table 2.1 Input variables to the traffic safety risk model.

Variable Name	Time Dependency	Group	Source	Variable Type
Distraction	Snapshot	Driver	Internal	Categorical
Driver Behavior	Snapshot	Driver	Internal	Categorical
Driver Seatbelt Use	Snapshot	Driver	Internal	Binomial
Subject age	Snapshot	Driver	Internal	Categorical
Subject gender	Snapshot	Driver	Internal	Binomial
Lighting	Snapshot	Environmental conditions	External	Categorical
Weather	Snapshot	Environmental conditions	External	Categorical
Alignment	Snapshot	Roadway-characteristics	External	Categorical
Locality	Snapshot	Roadway-characteristics	External	Categorical
Relation to Junction	Snapshot	Roadway-characteristics	External	Categorical
Surface Conditions	Snapshot	Roadway-characteristics	External	Categorical
Traffic Control	Snapshot	Roadway-characteristics	External	Categorical
Travel Lanes	Snapshot	Roadway-characteristics	External	Integer
Traffic Density	Snapshot	Surrounding externalities	External	Categorical
Traffic Flow	Snapshot	Surrounding externalities	External	Categorical
Gas pedal position	Time Series	Driver	Internal	Continuous
Speed Vehicle Composite	Time Series	Driver	Internal	Continuous
Speed GPS horizontal	Time Series	Driver	Internal	Continuous
Yaw rate	Time Series	Driver	Internal	Continuous
Heading GPS	Time Series	Driver	Internal	Continuous
Lateral acceleration	Time Series	Driver	Internal	Continuous
Longitudinal acceleration	Time Series	Driver	Internal	Continuous
Brake on/off	Time Series	Driver	Internal	Binomial
Turn signal state	Time Series	Driver	Internal	Categorical
Lighting	Time Series	Environmental conditions	External	Continuous

Following our proposed methodology, we first used cross validation to compare the prediction performances of MLR and RF classifiers on *event* data. Figure 2.3 and Figure 2.4 show the results of k-fold CV and stratified k-fold CV for k=2, 3, …, 10, respectively. We also performed replicated CV with n=10 to smooth the error-rate values over k. These figures suggest that RF classifier performs better on the feature space of the dichotomous *event* data. It also has a pretty robust performance excelling MLR model over all values of k with an error rate close to %8. Therefore, we select the RF classifier over the MLR model to classify *event* data. We used RF to calculate the driver's risk, i.e. the probabilities P_{ia}'s in Equation (2.8), and set i=1 for the RF classifier given only the *event* data.

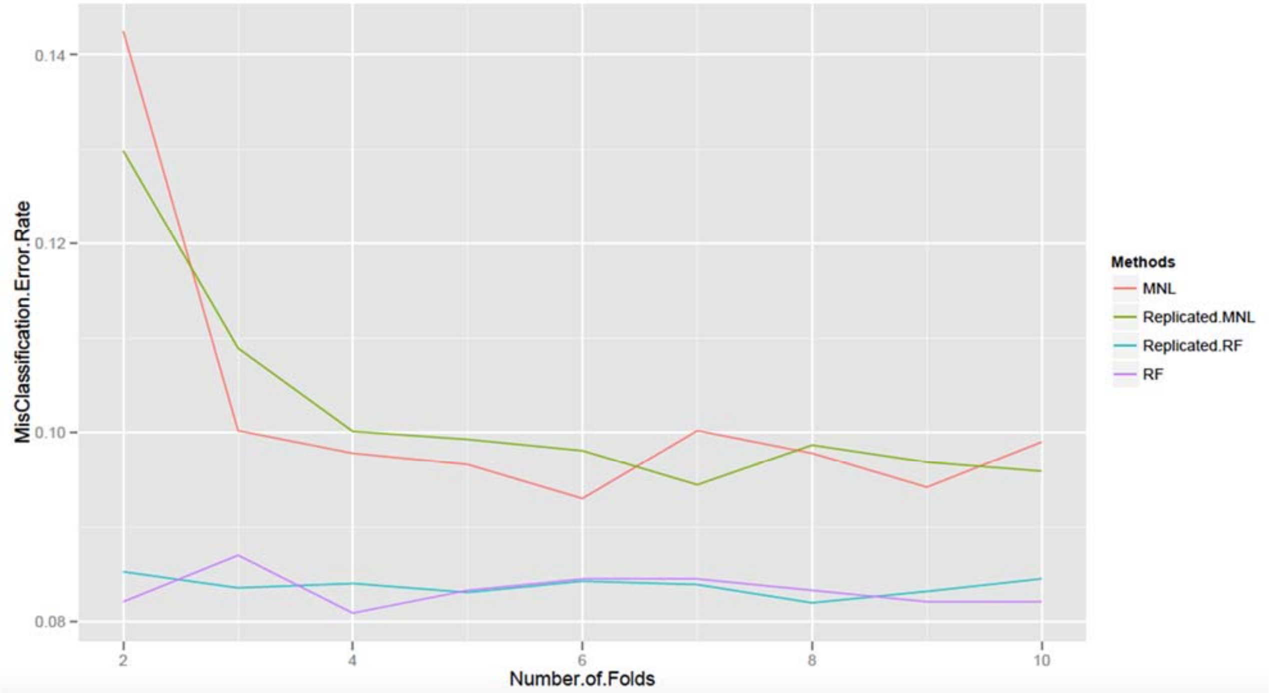


Figure 2.3 Cross validation (including single and replicated runs) for the case of crash and near-crash.

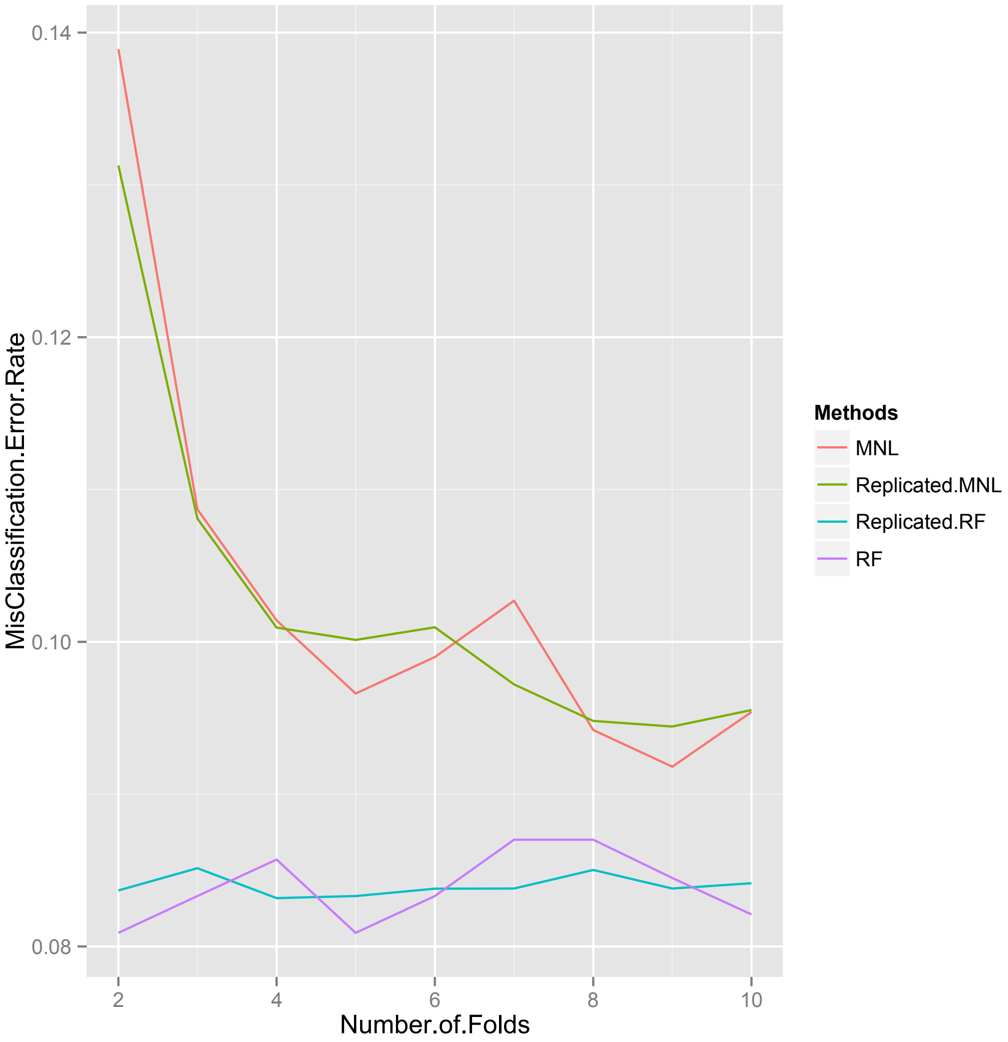


Figure 2.4 Stratified cross validation (including single and replicated runs) for the case of crash and near-crash.

Next, we used MTS-RF to classify the dichotomous output of crash and near-crash events on the time-series feature space. The best values of number of trees and number of nodes, best in term of OOB error rate, were used to grow the forest. After 10 replications of MTS-RF, the average OOB error rate was calculated to be and the test error rate to be Table 2.2 shows an example of the confusion matrix on a random split of data into %70 training and %30 test sets while the ration of the number of crashes to near-crashes was kept equal to the original ratio. As it can be seen, the OOB error rate of the model is very small equal to 2.42% while the error rate for the class of crashes was 29.8% and the near-crashes 0%.

Table 2.2 Confusion matrix for the training dataset

		Predicted Classes			
		Crash	Near-Crash	Total	Class Error
Actual Classes	Crash	33	14	47	0.298
	Near-Crash	0	532	532	0

Then, we run the model on the test data set in order to evaluate the prediction performance of our classification model for a new unobserved data point. The total accuracy of the model is 0.956. Also, Cohen's kappa statistic which compares the accuracy of the system to the accuracy of a random system was calculated to be 0.6247. According to Landis and Koch [49], it falls between 0.61-0.80 and shows a substantial agreement in classification. Table 2.3 shows the confusion matrix and class errors of the test data set. The error rate of the class of crashes has increased to 52.4% for the test set.

Table 2.3 Confusion matrix for the test dataset

		Predicted Classes			
		Crash	Near_Crash	Total	Class Error
Actual Classes	Crash	10	11	21	0.524
	Near_Crash	0	228	228	0

Table 2.2 and Table 2.3 were presented to show how the overall error rate was distributed between classes. In the next step, we are going to combine the results of RF classifier on *event* feature space (P_{1j}'s) with the results of MTS-RF on time-series feature space (P_{2j}'s). In doing so, we designed stratified k folds of cases to train RF and MTS-RF separately on k-1 folds and calculate P_{1j}'s and P_{2j}'s on the left-out kth fold. Then, by

applying one of the fixed combining rules of ensemble classifiers introduced in section 2.4, we calculated the final P_j's, i.e. the driving safety risks.

Figure 2.5, Figure 2.6 and Figure 2.7 show the results of stratified k-fold CV using RF on *event* data and MTS-RF on *time-series* data and ensemble classifiers with combining rules of minimum, maximum, mean, and product for the final decision fusion. For illustration purposes, we are only showing the results for k=5 since the same interpretations were concluded from all values of k=2,…,10. As it can be seen from these figures, MTS-RF has the best performance with the lowest error rate of.

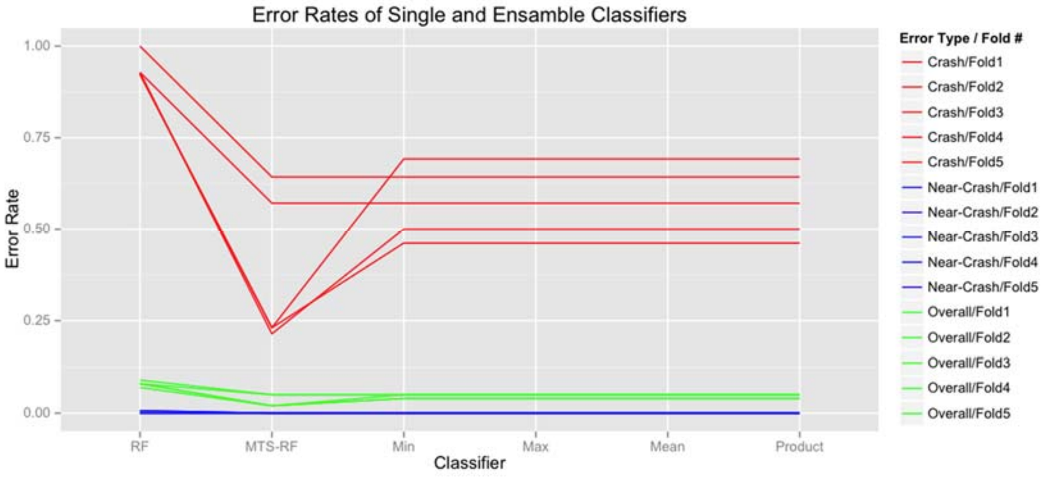


Figure 2.5 Error rates of single and ensemble classifiers per fold, Nfolds=5.

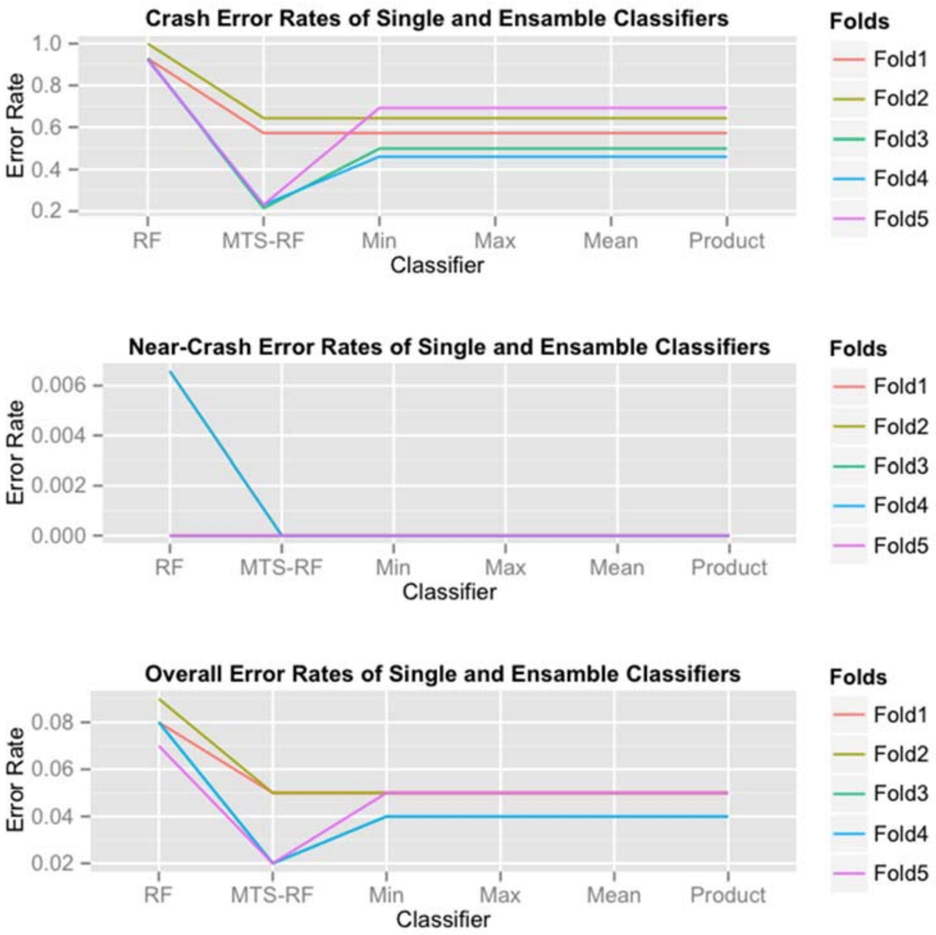


Figure 2.6 Error rates of single and ensemble classifiers in separate views for crash, near-crashes, and overall errors, Nfolds=5.

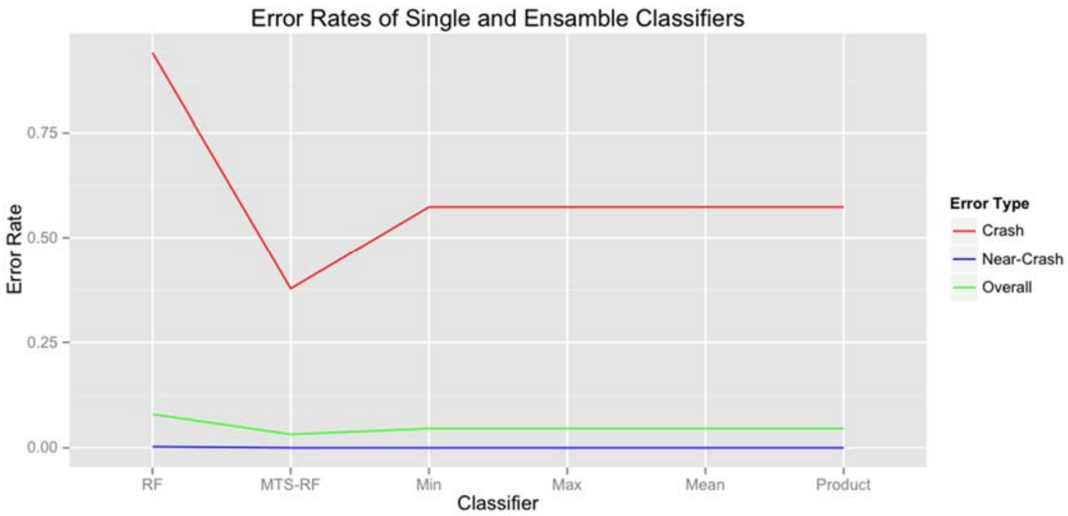


Figure 2.7 Average CV error rates of single and ensemble classifiers for different type of error rates, Nfolds=5.

Furthermore, random forests can be used to rank the importance of variables [6]. There are two criteria based on which the Breiman's random forest calculates the importance of variables: *Gini importance* and *permutation accuracy importance*. The variable importance plot gives a relative ranking of significant features. Table 2.4 shows the results of variable importance plot for time-series data in a tabular format.

Table 2.4 Variable Importance List.

Rank	Variable	Type	Mean Decrease Gini	Percentage
1	Travel Lanes	Video reduced	1738.11	0.16
2	Lighting	Time Series	1521.16	0.14
3	Gas pedal position	Time Series	1467.87	0.14
4	Driver Behavior	Video reduced	972.76	0.09
5	Traffic Density	Video reduced	959.37	0.09
6	Traffic Flow	Video reduced	439.40	0.04
7	Traffic Control	Video reduced	402.65	0.04
8	Locality	Video reduced	363.03	0.03
9	Speed GPS horizontal	Time Series	337.86	0.03
10	Lighting	Video reduced	326.48	0.03
11	Relation To Junction	Video reduced	275.13	0.03
12	Subject age	Video reduced	246.63	0.02
13	Speed .Vehicle composite	Time Series	237.54	0.02
14	Distraction	Video reduced	216.00	0.02

15	Heading GPS	Time Series	205.44	0.02
16	Alignment	Video reduced	203.03	0.02
17	Weather	Video reduced	166.43	0.02
18	Surface Conditions	Video reduced	95.83	0.01
19	Brake on off	Time Series	84.61	0.01
20	Yaw rate	Time Series	71.47	0.01
21	Subject gender	Video reduced	60.75	0.01
22	Lateral acceleration	Time Series	57.81	0.01
23	Driver Seatbelt Use	Video reduced	55.46	0.01
24	Longitudinal acceleration	Time Series	46.66	0
25	Turn signal state	Time Series	16.52	0

In order to build an active safety model, we will feed the updated state vector at each time step into our classification model to predict the real time crash risk of an individual driver. It can be presented to the driver similar to the on-board safety warning system, such as blind spot warning or forward collision warning, through the smart cars' terminal notifying the driver of possible risks of engagement in a near-crash or crash event. For convenience, we have color-coded the three safety states as follows:

State	**Color Code**
Crash	🟥
Near-Crash	🟨
Safe	🟩

Figure 2.8 schematically illustrates how the active safety model dynamically compute the safety risk (probability of crash/near-crash/normal driving) as the driver travels through the network and his state vector changes.

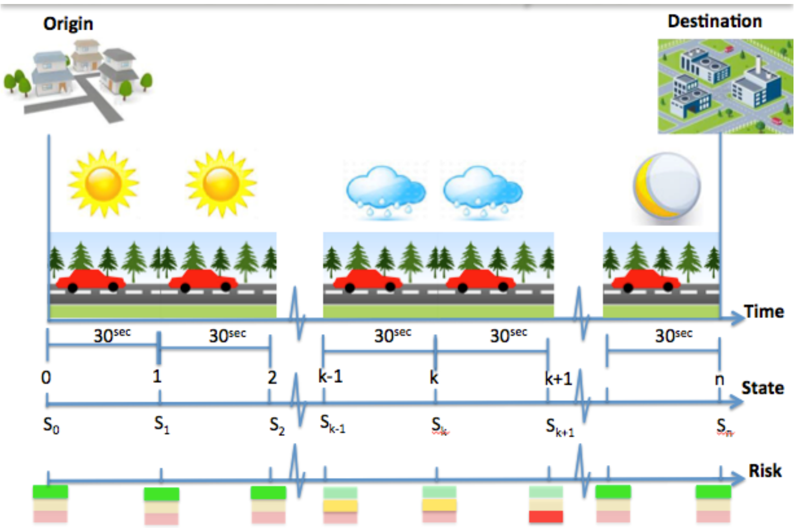


Figure 2.8 schematic illustration of our active safety model

Part II- Traffic Safety Risk Prediction Using SHRP-2 NDS data

2.7 Prediction Models

In this section, we present our methodology to calculate the real-time traffic safety risk of a driver at a specific location and time during a certain trip for the SHRP-2 NDS data. We propose using a weighted regularized multinomial logistic regression for classification of driving outcomes. In this chapter, we use the elastic net, a regularized regression technique, that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces [1].

We use a linear logistic function to model the log-likelihood ratio of driving outcomes, $Y_{ijlk} \in \{1, 2, \dots, C\}$, as a linear combination of independent variables, i.e. the vector of safety factors $\boldsymbol{ev}_{ij(l,L),(k,N)}$, as follows:

$$\log \frac{p(Y_{ijkl} = a \mid ev_{ij(l,L),(k,N)})}{p(Y_{ijkl} = C \mid ev_{ij(l,L),(k,N)})} = \beta_0{}^{a} + ev^{T}{}_{ij(l,L),(k,N)}\beta^{a} \tag{2.16}$$

where $a = 1, \cdots, C-1$ and outcome C is chosen as the pivot. Inverting this transformation yields an expression for the conditional probability:

$$p(Y_{ijkl} = a \mid ev_{ij(l,L),(k,N)}) = \frac{\exp(\beta_0{}^{a} + ev^{T}{}_{ij(l,L),(k,N)}\beta^{a})}{\sum_{c=1}^{C}\exp(\beta_0{}^{c} + ev^{T}{}_{ij(l,L),(k,N)}\beta^{c})} \tag{2.17}$$

To estimate the parameters of the above model, $\{\beta_0{}^{c}, \boldsymbol{\beta}^{c}\}_1^C$, we applied Tibshirani and Hastie’s proposed regularization model, called elastic net. This model solves the following problem:

$$\max_{\{\beta_0{}^{c},\beta^{c}\}_1^C} \left[\frac{1}{N} \sum_{i,j,l,k} w_{ijkl} \log p(Y_{ijkl} = c \mid ev_{ij(l,L),(k,N)}, \{\beta_0{}^{c},\beta^{c}\}_1^C) - \lambda \sum_{c=1}^{C} P_\alpha(\beta^{c}) \right] \tag{2.18}$$

where $\lambda \geq 0$ is a tuning parameter, w_{ijkl} is the weight of the $ijkl$th instance and N is the total number of instances used for the parameter estimation. α is the elastic-net parameter providing a mix between ridge regression and the lasso (least absolute shrinkage and selection operator). Equation (2.18) trades off two different criteria. The first part seeks coefficient estimates that fit the data well by maximizing the likelihood function, while the second term, called a *shrinkage penalty*, shrinks the coefficient estimates towards zero. The intercepts $\beta_0{}^{a}$ need not be regularized. The tuning parameter, λ, serves to control the relative impact of these two terms on the regression coefficient estimates.

$P_\alpha(\boldsymbol{\beta}^{a})$ is the elastic net penalty and can be computed from Equation (2.19). $||\boldsymbol{\beta}^{a}||_1$ and $||\boldsymbol{\beta}^{a}||_2$ are l_1 and l_2 norms, also called Manhattan and Euclidian norms, and can be calculated from (2.20) and (2.21), respectively. The elastic net penalty is a compromise between the ridge regression penalty ($\alpha = 0$) and the lasso penalty ($\alpha = 1$). It is particularly useful in $N \gg P$ situations, or any situation where there are many correlated predictor variables.

$$\begin{aligned} P\alpha\,(\beta^{a}) &= (1-\alpha)\frac{1}{2}(\|\beta^{a}\|_2)^2 + \alpha\|\beta^{a}\|_1 \\ &= \Sigma_{r=1}^{p_2}\left[\frac{1}{2}(1-\alpha)(\beta_r^{a})^2 + \alpha|\beta_r^{a}|\right] \end{aligned} \tag{2.19}$$

$$\|\beta^{a}\|_2 = \left(\Sigma_{r=1}^{p_2}(\beta_r^{a})^2\right)^{1/2} \tag{2.20}$$

$$\|\beta^{a}\|_1 = \Sigma_{r=1}^{p_2}|\beta_r^{a}| \tag{2.21}$$

Ridge regression shrinks the coefficients of correlated predictors towards each other while Lasso is somewhat indifferent to very correlated predictors, and will tend to pick one and ignore the rest. The lasso penalty corresponds to a Laplace prior, which expects many coefficients to be close to zero, and a small subset to be larger and nonzero. Thus, lasso can be used for variable selection.

The elastic net with $\alpha = 1 - \varepsilon$ for some small $\varepsilon > 0$ performs much like the lasso, but removes any degeneracies and wild behavior caused by extreme correlations. More generally, the entire family P_α creates a useful compromise between ridge and lasso.

The first part of (2.18) is simply the log likelihood function of the multinomial logistic model and can be written as follows:

$$\begin{aligned} l(\{\beta_0{}^{c}, \beta^{c}\}_1^{C}) = \frac{1}{N}\sum_{i,j,l,k} w_{ijkl}\Bigg[&\sum_{c=1}^{C}\Big(I(Y_{ijlk} = c) \times \\ &(\beta_0{}^{c} + ev^{T}{}_{ij(l,L),(k,N)}\beta^{c})\Big) \\ &- \log\sum_{c=1}^{C}\exp(\beta_0{}^{c} + ev^{T}{}_{ij(l,L),(k,N)}\beta^{c})\Bigg] \end{aligned} \tag{2.22}$$

To solve the maximization problem in (2.22) for $\{\beta_0{}^{c}, \boldsymbol{\beta}^{c}\}_1^{C}$, we have used the R package, *glmnet* [1]. The *glmnet* algorithms use cyclical coordinate descent, which successively optimizes the objective function over each parameter with others fixed, and cycles repeatedly until convergence.

The use of the elastic net in our problem has the following advantages:

- Has a built-in mechanism for variable selection.
- Provides better test prediction via bias/variance trade-off with its continuous shrinkage and variable selection.
- Accepts both numerical and categorical inputs.
- Offers cost-sensitive learning by applying class-specific weights in the loss function, which is paramount for imbalanced data.
- Deals with highly correlated variables, particularly in a high dimensional problem, via grouped variable selection and shrinkage.

Next, we will discuss the methods for evaluation of the classification model of imbalanced traffic safety data.

2.8 Model Evaluation

A confusion matrix, also known as a contingency table is a popular tool that allows visualization of the performance of a supervised learning algorithm. It is a square matrix of size *c* which is the number of classes of a categorical response variable, where element (i,j) is the count of instances in class j, predicted by the algorithm to belong to class i. Table 2.5 shows the confusion matrix for the classification of the 3-class driving outcome.

Table 2.5- The Confusion Matrix of The 3-Class Driving output.

			Actual	
		Baseline	**Near-Crash**	**Crash**
Predicted	**Baseline**	True Baseline T_B	False Baseline $FB_{\mid NC}$	False Baseline $FB_{\mid C}$
	Near-Crash	False Near-Crash $FNC_{\mid B}$	True Near-Crash T_{NC}	False Near-Crash $FNC_{\mid C}$
	Crash	False Crash $FC_{\mid B}$	False Near-Crash $FC_{\mid NC}$	True Crash T_C

Each diagonal element of the confusion matrix represents a true classification, for example T_C, read *True Crash*, is the proportion of *Crash* events that were correctly classified as such while an off-diagonal element represents a misclassification, for example $FB_{|C}$, read *False Baseline given Crash*, is the proportion of *Crash* events which were wrongly classified as *Normal Driving (Baseline)*. Equations 2.23 and 2.24 show the calculations of T_C and $FB_{|C}$, respectively.

$$T_C = \sum_j I(\hat{y}_j = C \mid y_j = C) \tag{2.23}$$

$$FB_{|C} = \sum_j I(\hat{y}_j = B \mid y_j = C) \tag{2.24}$$

I is an indicator function, and $I(event\ A)$ is an indicator variable defined as follows:

$$I(event\ \ A) = \begin{cases} 1 & event\ \ A\, happens, \\ 0 & \quad otherwise. \end{cases}$$

For example, $I(\hat{y}_j = C|y_j = C)$ is an *indicator variable* that equals one if $\hat{y}_j = C|y_j = C$ and zero otherwise. Other elements of the confusion matrix can be calculated in a similar fashion.

2.8.1 Misclassification Error

The most common approach for quantifying the accuracy of a classifier is the misclassification error rate (here denoted by *MCER*) or the proportion of mistakes. This is actually the average of the off diagonal elements of the confusion matrix. We can apply the trained classifier to predict the observations in the *train* set and calculate the error rate, called *training error rate* denoted by $MCER_{tr}$ computed from (2.25). But, we are usually interested in the error rates that result from applying the classifier to test observations that were not used in training the model. Equation (2.26) shows the formula to compute the *test error rate*, $MCER_{ts}$.

$$MCER_{tr} = \frac{1}{N_{tr}} \sum_{j \in train} I(\hat{y}_j \neq y_j) \tag{2.25}$$

$$MCER_{ts} = \frac{1}{N_{ts}} \sum_{j \in test} I(\hat{y}_j \neq y_j) \tag{2.26}$$

where N_{tr} and N_{ts} are the number of instances in the train and test datasets, respectively.

2.8.2 Type I and Type II Errors

The misclassification error rate gives us some information about the overall performance of a classifier while we are most often interested in the distribution of errors over the classes. Especially, in an imbalanced classification setting where classes have unequal frequencies and the class of interest is of a lower frequency; the main objective of detecting the minority class is more challenging. Furthermore, in these unbalanced settings, usually the associated costs of errors over different classes are different. For example, in a driving scenario, the cost of an error in which the ADAS improperly indicates no presence of a critical condition when in reality it is present, is higher than the cost of an error in which the ADAS improperly alerts while it is actually a normal driving situation. The cost of the latter, called type I error, is the driver's annoyance or discomfort while the cost of the former error, called type II error, ranges from a property damage to loss of lives.

A type I error, also known as an error of the first kind, occurs when the null hypothesis (H_0) is true, but is rejected. It is asserting something that is absent, a false hit. The type I error rate or significance level is the probability of rejecting the null hypothesis given that it is true. On the other hand, a type II error, also known as an error of the second kind, occurs when the null hypothesis is false, but erroneously fails to be rejected. It is failing to assert what is present, a miss. What we actually call type I or type II errors depends directly on the null hypothesis. In what follows we present the null and alternative hypotheses for our problem.

In designing an ADAS, it is accepted to have a higher type-I error rate (false alarm) in exchange of a lower type-II error rate, i.e. missing a true crash. It is worth noting that, in long term, an unreasonable high rate of nuisance alert can lead in drivers' mistrust in the system and a potential passive behavior toward an upcoming safety alert [50]. Since, false alarms go up with attempts to detect higher percentages of true objects, the success of the classification model is a trade-off between type I and II errors.

In order to compute the type I and II error rates for our multi-class problem, we first need to define three sets, denoted by S, S_{Cr} and S_{NCr}, to represent all the possible driving outcomes, critical outcomes and non-critical outcomes, respectively. We propose an Advanced Driver Assistance System (ADAS), which alerts critical outcomes in S_{Cr}. For example, for a trichotomous driving outcome, these three sets are as follows:

$$\begin{aligned} S &= \{Baseline, Near-Crash, Crash\} \\ S_{Cr} &= \{Near-Crash, Crash\} \\ S_{NCr} &= \{Baseline\} \end{aligned}$$

The above-mentioned ADAS alerts of *Crash* and *Near-Crash* outcomes. Table 2.6 shows the alert modes of this system. As it can be seen, the ADAS will give different visual and voice warnings to the driver. The proposed ADAS in this chapter uses the basic information about the driver characteristics, i.e. Age, Gender and Years of Driving, and combine it with the driver's real-time behavior such as Speeding and Impaired Driving; the engineering roadway data, such as Locality, Road Alignment, Relation To Junction, Traffic Flow; and real-time network data, such as Traffic Density and Weather, to predict the outcome of driving.

Table 2.6- Alert modes of the proposed ADAS for the trichotomous driving outcome.

Class	**Status**	**Visual color code**	**Voice alert**
Crash	Crash	Red	Yes
Near-	Near-	Orange	Yes
Baseline	Safe	Green	No

For the proposed ADAS, the hypothesis test is as follows:

$$\begin{cases} H_0: & \textit{The driving outcome is not critical,} \\ H_1: & \textit{otherwise.} \end{cases}$$

Then, the type I and II error rates are defined as follows:

$$\begin{aligned} \textit{Type I error rate} &= \Pr\{H_0 \textit{ is rejected} \mid H_0 \textit{ is true}\} \\ &= \Pr\{\hat{y}_{ijkl} \in S_{Cr} \mid y_{ijkl} \in S_{NCr}\} \end{aligned} \tag{2.27}$$

$$\begin{aligned} \textit{Type II error rate} &= \Pr\{H_0 \textit{ is not rejected} \mid H_1 \textit{ is true}\} \\ &= \Pr\{\hat{y}_{ijkl} \in S_{NCr} \mid y_{ijkl} \in S_{Cr}\} \end{aligned} \tag{2.28}$$

Equations (2.29) and (2.30) show the type I and II error rates of detecting a crash in the 3-class classification problem of driving outcomes using the confusion matrix in Table 2.5.

$$Type\ I\_ER = \frac{FNC_{|B} + FC_{|B}}{T_B + FNC_{|B} + FC_{|B}} \tag{2.29}$$

$$Type\ II\_ER = \frac{FB_{|NC} + FB_{|C}}{\left(FB_{|NC} + T_{NC} + FC_{|NC}\right) + \left(FB_{|C} + FNC_{|C} + T_C\right)} \tag{2.30}$$

To further evaluate the performance of our multiclass classification problem, we introduce two new measures, namely the off-diagonal upper triangular error rate and off-diagonal lower triangular error rate. Figure 2.9 shows the general structure of the confusion matrix for classifying the multiclass driving outcomes. As it can be seen, rows represent predicted classes and columns represent the actual classes. Furthermore, the safety risk of classes from left to right, and top to bottom is increasing.

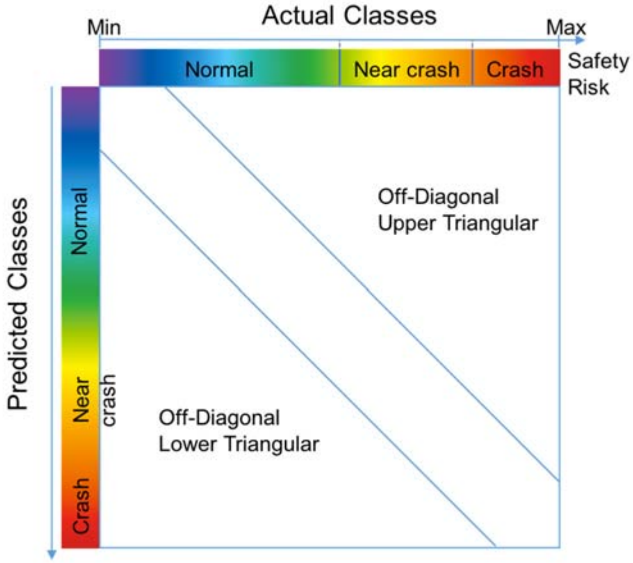


Figure 2.9 The General structure of the confusion matrix for the classification of spectral driving outcomes.

Equations (2.31) and (2.32) show the formula to calculate these two measures:

$$Off\ \_\ diagonal\ Lower.\ Triangular\ Error\ Rate(OLT\ \_\ ER) = \frac{FC_{|NC} + FC_{|B} + FNC_{|B}}{N} \tag{2.31}$$

$$Off\ \_\ diagonal\ Upper.\ Triangular\ Error\ Rate(OUT\ \_\ ER) = \frac{FB_{|NC} + FB_{|C} + FNC_{|C}}{N} \tag{2.32}$$

where N is the total number of classified instances. OLT_ER measures the average error rate of misclassifying an instance into a higher-risk class while OUT_ER measures the average error rate of misclassifying an instance into a lower-risk class. OLT_ER and OUT_ER are general forms of type I and II errors and for that reason lower values of OUT_ER are preferable in exchange of higher values of OLT_ER.

2.8.3 k-fold Cross Validation

In order to select the best model, we use the cross-validation method. Considering a bias-variance trade-off, performing k -fold cross-validation using k = 5 or k = 10 are recommended as these values have been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance [48]. The k-fold cross-validated misclassification error rate takes the form:

$$kCV\ \_\ MCER = \frac{1}{k}\sum_{i=1}^{k} MCER_i = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{N_i}\sum_{j\in Fold\ i} I(\hat{y}_j \neq y_j) \tag{2.33}$$

where N_i is the number of instances in fold i. Table 2.7 shows a list of 5 different error rates that we will report on each model in the cross-validated model selection. In the next two sections, we will first prepare the data to be used in a regression model and then fit the elastic net models with different parameters and select the best model using cross-validation.

Table 2.7- A Summary list of the reported error rates for the cross-validated model selection.

Type of Error Rate	k-fold Formula
Mis-Classification	$kCV\_MCER = \frac{1}{k}\sum_{i=1}^{k} MCER_i$
Type I	$kCV\_Type\ I = \frac{1}{k}\sum_{i=1}^{k} Type\ I\_ER_i$
Type II	$kCV\_Type\ II = \frac{1}{k}\sum_{i=1}^{k} Type\ II\_ER_i$
Off-diagonal Lower Triangular	$kCV\_OLT\_ER = \frac{1}{k}\sum_{i=1}^{k} OLT\_ER_i$
Off-diagonal Upper Triangular	$kCV\_OUT\_ER = \frac{1}{k}\sum_{i=1}^{k} OUT\_ER_i$

2.9 Data Preparation and Feature Engineering

The Second Strategic Highway Research Program (SHRP-2) Naturalistic Driving Study (NDS) was the largest and most comprehensive study of its kind. The study included a three-year data-collection effort that produced driving data of about 5.5 million trips for over 3,000 drivers, including over 1,500 crashes and nearly 3,000 near-crashes in six states throughout the United States. It has been the largest study of its kind to investigate the role of driver performance and behavior in traffic safety. Detailed information about the NDS data is available at [51]. A complete list of publications and projects on SHRP-2 safety data can be found at [52].

The Data Acquisition System (DAS) in the participant vehicles included a forward radar; four video cameras, including one forward-facing, color, wide-angle view; accelerometers (x, y, and z axes); rate sensors (x, y, and z axes); illuminance sensor; passive cabin alcohol presence sensor; incident pushbutton; turn signal state; vehicle network information; Geographic Positioning System; onboard computer vision lane tracking, plus other computer vision algorithms; and data storage capability. Data from the DAS are recorded continuously while the participant's vehicle is operating. This continuous recording allowed for an exposure-based approach and was central to the SHRP-2 *safety* focus area.

In addition to the real-time driving- and vehicle-related data collected via the installed data acquisition equipment, a variety of non-DAS data, about the driver and vehicle characteristics, was also procured. Driver data include basic demographic information, functional ability relative to driving safety and risk, vision tests, cognitive assessments, physical ability metrics, vehicle information and post hoc crash investigations. These non-DAS data streams were obtained through a variety of instruments, including questionnaires; assessments of physical acumen, cognitive capacity, and visual acuity; and participant interviews.

For this research work, we have obtained a subset of SHRP-2 data, which required a data sharing agreement with Virginia Tech Transportation Institute and an IRB approval from The Rutgers University's Office of Research and Regulatory Affairs. All the analyses in this study are performed in R 3.1.2 [53] and the main package for data visualization was ggplot2 [54]. The dataset was limited to thirty percent of the complete SHRP-2 data containing a total of 8131 events from which 1217 were *crashes* (15%), 2644 *near-crashes* (33%) and 4270 sampled *baseline* (53%) epochs. The sampled *baseline* epochs come from a pool of 20,000 baselines stratified per driving time with the driver's speed not dropping below five miles per hour for more than two seconds. The additional criterion to select the *baselines* was to select samples whose drivers were also present in the set of the *crash* and *near-crash* events. Overall, 1250 drivers and 1299 vehicles were included in this dataset. The bar plots in Figure 2.10 shows the distributions of crash severity (top) and crash type (bottom). As it can be seen, only 18 percent of crashes were severe and/or police reportable and 41 percent of crashes were low risk tire strikes for example clipping a curb during a tight turn. The most prevalent type of crashes was the road departure with 68 percentages. The conflict with a lead vehicle, and conflict with a following vehicle with 7 and 6 percentages are the next two frequent crash types. As we discussed, crashes are rare events meaning time is needed to record enough crashes of any specific type to analyze. The use of surrogates for collisions, such as near-collisions, critical incidents, or traffic conflicts, would greatly increase the power of the field studies, because the surrogate events occur much more frequently than crashes and without any severe consequences [55].

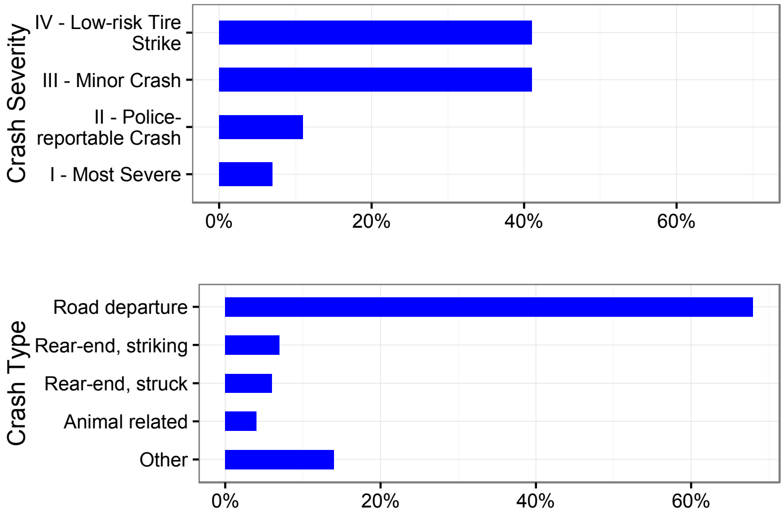


Figure 2.10 Severity (top) and type (bottom) distributions of crashes.

SHRP-2 safety data contain four major categories of data: 1- event-detailed, 2- time-series, 3- driver, and 4- vehicle data. In the SHRP-2 data, the typical length of the time-series data is 30 seconds for the crash and near-crash events and 21 seconds for the baseline epochs.

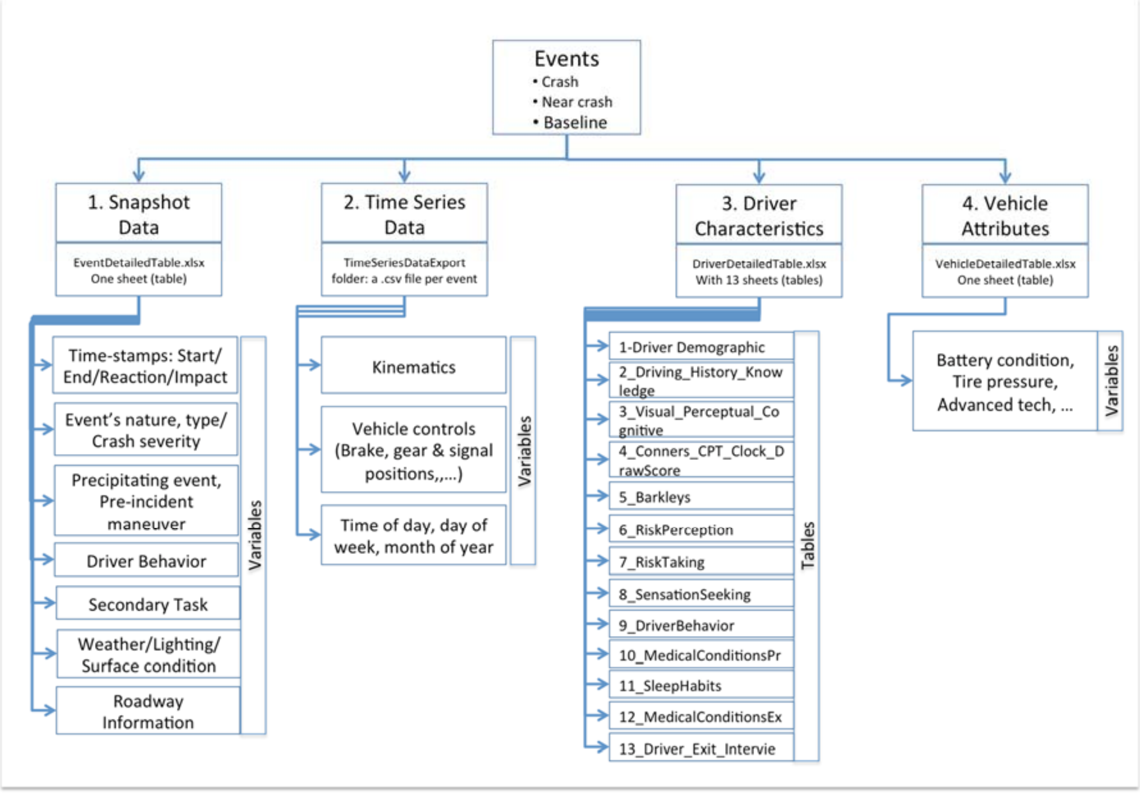


Figure 2.11 A Tree structure of SHRP-2 NDS safety data.

Figure 2.11 shows a tree structure of the main data tables with either their fields or sub-tables in our subset of SHRP-2 safety data. As it can be seen, *event detailed* data mainly include time stamps; event nature, type and severity; precipitating event, pre-incident maneuver, maneuver judgment; driver behavior; driver's secondary task (if any); weather, lighting, and surface condition; and finally roadway information data. *Time-series* data include vehicle kinematics such as speed and acceleration; status of vehicle controls such as brakes, steering wheel, gear and signal position; and time and date information.

Driver data contain detailed information about the driver characteristics including demographics, pre- and exit medical conditions, sleep habits, driving history and knowledge, visual and cognitive tests, Conner Continuous Performance Test (a test for assessment of attention disorders and neurological functioning), clock drawing score (a test score for assessment of dementia or other neurological disorders), Barkley's ADHD screening test, risk perception, risk taking, sensation seeking, past driver behavior. Finally, the vehicle data include vehicle types (car, truck, van, etc.), ages, condition and technologies and equipment.

Appendix-1 shows a list of twenty-one variables from the *event-detailed* and *driver* tables that we included in our model as potential predictors of traffic safety risk. The data we received were already pre-processed to be used by researchers but still it needed some extra preparation steps specific to our problem. The very first step to prepare the data is to make the data readable in R. The structure of data is tested to check if the type, values and the range of each variable matches the targets explained in the SHRP-2 data dictionaries. Data munging and wrangling are also required to clean the data, remove the unnecessary punctuations and to map the semi-raw data to meaningful values according to the data dictionaries.

Data preparation continues by properly handling missing data. Some algorithms such as Breiman's Random Forest [6] have built-in procedures to handle missing values. In fact, Random forests has two ways of replacing missing values: (i) A fast way is to replace a missing value with the median for numeric variables and with the most frequent level (breaking ties at random) for factor variables; (ii) A computationally more expensive approach with a better performance is to use missing value proximities to iteratively compute weighted measures to replace the missing values. There are more computationally expensive methods to impute missing values such as Multivariate Imputation by Chained Equations (MICE) [56]. In this study, we use the fast approach to replace the missing values using the randomForest package [57].

The next step after the preliminary data preparations is feature engineering. Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data. As it can be seen in Appendix 1, the type of all the included variables in our model is either categorical or binary requiring especial treatments to enter the predictive model.

Build New Binary Variables From Levels of a Categorical Variable with Many Levels

For each crash or near-crash event, there are three variables of *Driver_Behavior_1*, *Driver_Behavior_2*, *and Driver_Behavior_3* that record up to three of the most critical driver behaviors i.e. behaviors that most directly caused or contributed to the corresponding crash or near-crash event. These categorical variables can accept 58 distinct values. Similarly, the variables *Secondary_Task_1*, *Secondary_Task_2* and *Secondary_Task_3*, store up to three of the most critical observable driver engagements in any secondary task during the event. Distractions include non-driving related glances away from the direction of vehicle movement and do not include tasks that are critical to the driving task, such as speedometer checks, mirror/blind spot checks, activating wipers/headlights, or shifting gears. Each of the Secondary_Task variables has 64 levels.

Since these variables have many rare levels, which make them uninformative, we focus on some of the most prevalent levels and re-define them as new binary variables. For example, the binary variables of $Speeding$ and Low_Speed in Table 2.8 are engineered features created from the levels of three categorical variables of *Driver_Behavior*. Table 2.8 and Table 2.9 show the mapping of the speeding- and low-speed-related levels of *Driver_Behavior* variables to the new binary variables of $Speeding$ and Low_Speed, respectively.

Table 2.8- Mapping of the speeding-related levels of *Driver_Behavior* variables to the new binary variable of *Speeding*.

Driver_Behavior level	**New Binary Variable, level**
Exceeded safe speed but not speed limit	Speeding, Yes
Exceeded speed limit	Speeding, Yes
Stop sign violation, intentionally ran stop sign at speed	Speeding, Yes

Table 2.9- Mapping of the low-speed-related levels of *Driver_Behavior* variables to the new binary variable of *Low-Speed*.

Driver_Behavior level	**New Binary Variable, level**
Driving slowly in relation to other traffic: not below speed limit	Low Speed, Yes
Driving slowly: below speed limit	Low Speed, Yes

Similarly, the binary variables of $Passenger\_In\_Adjacent\_Seat$, and $Cellphone\_Use$ were created from the levels of three categorical variables of *Secondary_Task_1*, *Secondary_Task_2* and *Secondar_Task_3*, which can store up to three distinct secondary tasks during a single event. The rest of the variables were readily available either from the *event-detailed* or *driver* data table.

Combine Rare Levels

To avoid redundant levels in a categorical variable and to deal with rare levels, we can aggregate levels. There are various methods of combining levels, two common ones are: (i) Using business logic, i.e. combining similar levels into similar groups based on domain or business experience, (ii) Using frequency or response rate, i.e. combining levels by considering the frequency distribution or response rate. To combine levels using their frequency, we first look at the frequency distribution of each level and combine levels having frequency less than 5% of total observation (5% is standard but you can change it based on distribution). This is an effective method to deal with rare levels. One can actually look at both frequency and domain knowledge to combine levels more effectively. Table 2.10 shows the combined levels of *Traffic_Density* according to both criteria.

Table 2.10- Combine Levels of The Traffic Density

Traffic Density: Original Level	Traffic Density: New Level
Level-of-service D: Unstable flow - temporary restrictions substantially slow driver	Unstable flow
Level-of-service E: Flow is unstable, vehicles are unable to pass, temporary stoppages, etc.	Unstable flow
Level-of-service F: Forced traffic flow condition with low speeds and traffic volumes that are below	Unstable flow

Dummy coding

Finally, in order to include any of these categorical variables with c levels in a multiple regression prediction model, $c-1$ dichotomous variables, also called dummy variables, must be created. Therefore, after dummy coding of all the categorical predictors in Table 1, the dimensionality of the problem increases and the total number of binary predictors becomes sixty-six (66).

2.10 Prediction Models

In order to select the best model, we run the 10-fold cross-validation to calculate mean and standard deviation of the error rates, i.e. MCER, conservative and non-conservative type I and type II error rates, Off-diagonal Lower Triangular Error Rate (OLT_ER), and Off-diagonal Upper Triangular Error Rate (OUT_ER). We do this for ten cases: to select the elastic-net parameter $\alpha = \{0,0.25,0.5,0.75,1\}$ and the class weights $\boldsymbol{w} = \{(1,1,1),(1,3,5)\}$. We call α and $\boldsymbol{w}$ hyper-parameters. The weight vector $\boldsymbol{w}=(1,1,1)$ represents equally-weighted classes while $\boldsymbol{w} = (1,3,5)$ assigns weights of one, three and five to the *Baseline*, *Near-Crash* and *Crash* instances, respectively. The assignment of smaller weights to the majority class cases and larger weights to the minority class cases is one way of dealing with the imbalanced data. The weights were initially set to be inversely proportional to the fraction of cases of the corresponding class ($\boldsymbol{w} = (2,3,7)$). But through trial and error of neighboring values, we found $\boldsymbol{w} = (1,3,5)$ to result in a better prediction performance. Table 2.11 shows the values of the hyper-parameters of 10 cases for model selection.

Table 2.11- Candidate-model parameters to select the best model through cross validation.

Case	**1**	**2**	**3**	**4**	**5**
Weights	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)
Alpha	0	0.25	0.5	0.75	1

Case	**6**	**7**	**8**	**9**	**10**
Weights	(1,3,5)	(1,3,5)	(1,3,5)	(1,3,5)	(1,3,5)
Alpha	0	0.25	0.5	0.75	1

For each case in Table 8, the glmnet algorithm automatically generates a sequence of 100 values for the tuning parameter, λ. Then, it estimates the regression coefficients and computes MCER at each λ. We are interested in two values of the tuning parameter: λ_{min} which gives the minimum mean cross-validated MCER and λ_{1se} which gives the most regularized model such that MCER is within one standard error of the minimum. The value of λ_{1se} gives a more regularized (sparse) model than λ_{min}. Therefore, to select the best model, we first find λ_{min} (or λ_{1se}) for each case and then compare the performance of these ten best models to select the hyper-parameters (i.e. α and $\boldsymbol{w}$)

Since all of the predictor variables in our problem are categorical variables, adding an interaction term can significantly increase the size of the design matrix. For example, adding an interaction term for "*Years Driving*" and "*Subject Age*", with the highest observed correlation, increases the dimension from 66 to 94 (additional 28 binary variables). In general, adding interaction terms to the model may initially decrease the bias but eventually will increase the variance of the prediction model. Furthermore, the elastic net has a way of dealing with highly correlated variables. It simultaneously does automatic variable selection and continuous shrinkage, and selects groups of correlated variables. Elastic net with strict convexity guarantees the grouping effect in the extreme situation with identical predictors (the case of perfect multi-collinearity). We therefore decided not to include any interaction term into our model.

Figure 2.12 shows the ten-fold cross-validated error rates for ten cases in Table 2.11. The x-axis shows the number of classes of the driving outcome, whether the instances were weighted or not, and the alpha value, each separated with a dash. Table 2.12 shows the same results in a tabular format. Recall that the main objective of an ADAS is to identify the critical events properly, that is, to have a smaller type II error in exchange for an inevitable larger type I error. As it can be seen, the best weighted model is the elastic net with $\alpha = 0.5$ with 62 variables. This model gives a Type I error of 0.175, Type II error of 0.5 and has an overall MCER of 0.384. By introducing the weights into the model, we select the weighted elastic net with $\alpha = 0.75$ with 63 variables. Comparing to the equally-weighted model, this model has slightly a larger MCER (0.419) but it has reduced the type II error to 0.106. Since type I and II error rates is a trade off at a fixed sample size, the type II error rate has increased to 0.636. As it can be seen there is always a tradeoff between type I and II errors, and between OLT_ER and OUT_ER. For any given case, the effort to reduce one type of error generally results in increasing the other type of error. Therefore, since the main objective of the ADAS is to identify the critical events, we select the weighted elastic net model with $\alpha = 0.75$.

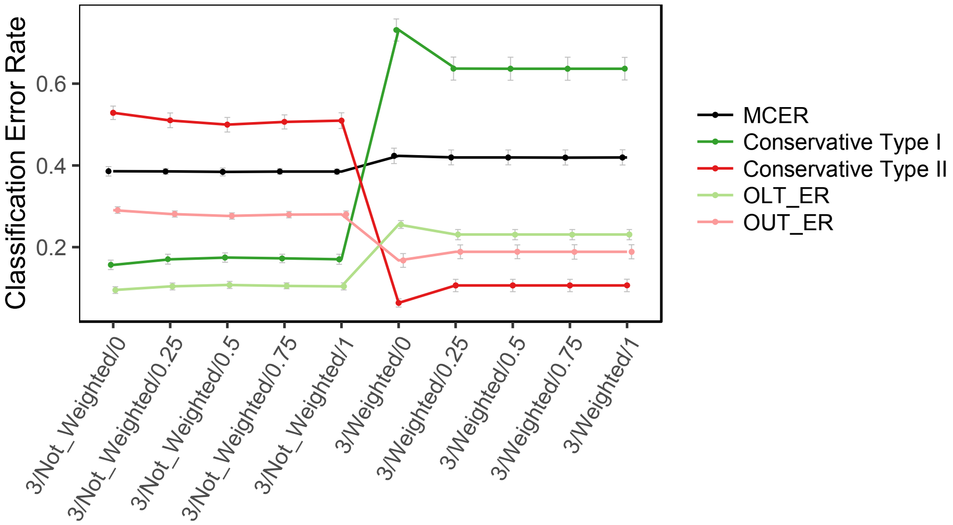


Figure 2.12 Ten-fold cross-validated estimates of the Miss-Classification Error Rate (MCER), conservative Type I and Type II error rates, Off-diagonal Lower Triangular Error Rate (OLT_ER), and Off-diagonal Upper Triangular Error Rate (OUT_ER), calculated at $\boldsymbol{\lambda_{min}}$, for the ten candidate models in Table 2.11 for the 3-class driving outcome.

Table 2.12- Ten-fold mean cross-validated error rates and their standard deviations in parenthesis calculated at $\boldsymbol{\lambda_{min}}$ for the 3-class driving outcome.

Weights	alpha	MCER	Conservative Type I	Conservative Type II
(1,1,1)	0	0.386	0.157	0.529
		(0.012)	(0.011)	(0.016)
(1,1,1)	0.25	0.385	0.171	0.510
		(0.007)	(0.012)	(0.018)
(1,1,1)	0.5	**0.384**	**0.175**	**0.500**
		(0.010)	(0.011)	(0.018)
(1,1,1)	0.75	0.385	0.173	0.507
		(0.007)	(0.010)	(0.017)
(1,1,1)	1	0.385	0.170	0.509
		(0.007)	(0.012)	(0.019)
(1,3,5)	0	0.423	0.731	0.064
		(0.019)	(0.027)	(0.010)
(1,3,5)	0.25	0.420	0.637	0.106
		(0.018)	(0.028)	(0.015)
(1,3,5)	0.5	0.420	0.636	0.106
		(0.018)	(0.028)	(0.015)
(1,3,5)	0.75	**0.419**	**0.636**	**0.106**
		(0.018)	(0.028)	(0.015)
(1,3,5)	1	0.419	0.637	0.107
		(0.019)	(0.028)	(0.015)

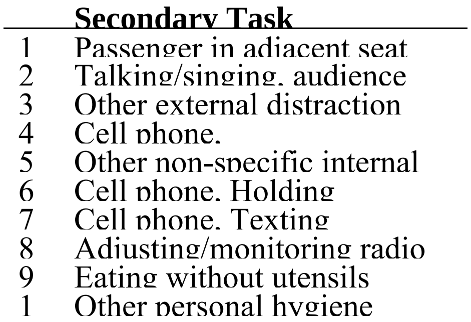
2.10.1 Re-defining The Classes of Driving Outcome

In order to improve the performance of our model, we re-define the classes of driving outcome. In the SHRP-2 NDS data, we are only provided with three tags for the sample epochs: “crash”, “near-crash” or “baseline”. If an epoch is not related to a crash or near-crash event, it is tagged with a “baseline” label, i.e. no significant road incident was observed. But, as we discussed it earlier, the driving outcomes and their associated risks belong to a spectrum. This led us to break down the class of “baseline” events into three new classes according to the driver’s unsafe behavior or involvement in secondary tasks. To do so, we used the levels of six categorical variables of *Driver_Behavior_1*, *Driver_Behavior_2* and *Driver_Behavior_3*; and Secondary*_Task_1*, *Secondary_Task_2* and *Secondar_Task_3* and recorded them to up to six new unsafe driving behaviors and non-driving tasks or distractions during a specific epoch. The variable “driver behavior” has 53 and “Secondary task” 57 distinct levels. Table 2.13 and Table 2.14 show the top 10 prevalent unsafe driving behavior and secondary tasks, respectively. Figure 2.13 shows the breakdown of baseline events into three new classes.

Table 2.13- Top ten prevalent unsafe driving behavior.

	Unsafe Driving Behavior
1	Exceeded speed limit
2	Drowsy, sleepy, asleep, fatigued
3	Failed to signal
4	Stop sign violation, "rolling stop"
5	Driving slowly in relation to other traffic: not below speed limit
6	Exceeded safe speed but not speed limit
7	Driving slowly: below speed limit
8	Avoiding other vehicle
9	Improper turn, cut corner on left
10	Wrong side of road, not overtaking

Table 2.14- Top ten prevalent secondary tasks.



	Secondary Task
1	Passenger in adjacent seat
2	Talking/singing, audience
3	Other external distraction
4	Cell phone,
5	Other non-specific internal
6	Cell phone, Holding
7	Cell phone, Texting
8	Adjusting/monitoring radio
9	Eating without utensils
1	Other personal hygiene

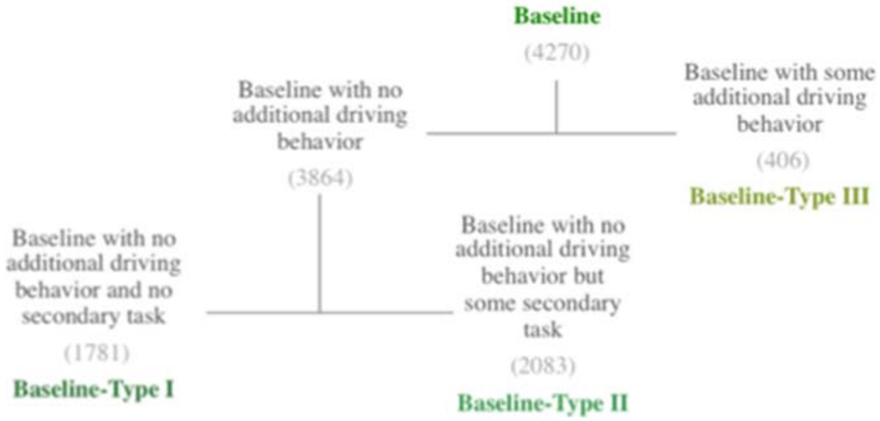

Baseline
(4270)
Baseline with no additional driving behavior
(3864)
Baseline with some additional driving behavior
(406)
Baseline-Type III
Baseline with no additional driving behavior and no secondary task
(1781)
Baseline-Type I
Baseline with no additional driving behavior but some secondary task
(2083)
Baseline-Type II

Figure 2.13 Break down of baseline events according to unsafe driving behavior and secondary task involvement.

The definition of Y_{ijlt} with five classes of driving outcomes is as follows:

$$Y_{ijlt} = \begin{cases} 1 & \textit{If the outcome is of mi.nimum risk}, \\ 2 & \textit{If the theoutcome is self hazardous}, \\ 3 & \textit{If the outcome is hazardous to self and others}, \\ 4 & \textit{If the outcome is a near crash}, \\ 5 & \textit{If the outcome is a crash}. \end{cases} \tag{2.34}$$

Figure 2.14 shows the distribution of event types after the break down of Normal Driving outcome into three subgroups. Table 2.15 shows the elements of the confusion matrix for the 5-class driving outcome.

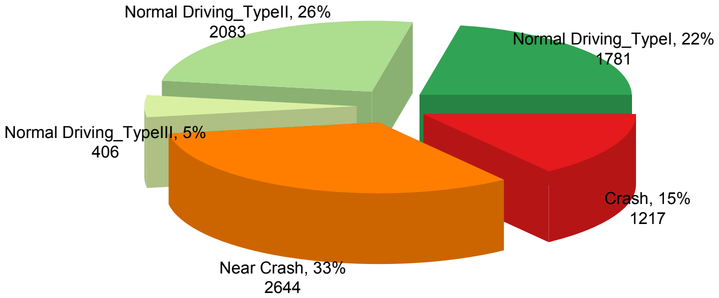


Figure 2.14 Distribution of event types after breaking down the class of normal driving.

Table 2.15 The confusion matrix of the 5-class driving output.

		Actual				
		Baseline I	**Baseline II**	**Baseline III**	**Near-Crash**	**Crash**
Predicted	**Baseline I**	True Baseline I T_{B1}	False Baseline I $FB1_{\vert B2}$	False Baseline I $FB1_{\vert B3}$	False Baseline I $FB1_{\vert NC}$	False Baseline I $FB1_{\vert C}$
	Baseline II	False Baseline II $FB2_{\vert B1}$	True Baseline II T_{B2}	False Baseline II $FB2_{\vert B3}$	False Baseline II $FB2_{\vert NC}$	False Baseline II $FB2_{\vert C}$
	Baseline III	False Baseline III $FB3_{\vert B1}$	False Baseline III $FB3_{\vert B2}$	True Baseline III T_{B3}	False Baseline III $FB3_{\vert NC}$	False Baseline III $FB3_{\vert C}$
	Near-Crash	False Near-Crash $FNC_{\vert B1}$	False Near-Crash $FNC_{\vert B2}$	False Near-Crash $FNC_{\vert B3}$	True Near-Crash T_{NC}	False Near-Crash $FNC_{\vert C}$
	Crash	False Crash $FC_{\vert B1}$	False Crash $FC_{\vert B2}$	False Crash $FC_{\vert B3}$	False Crash $FC_{\vert NC}$	True Crash T_C

We propose two Advanced Driver Assistance Systems (ADAS):

1. Basic ADAS: this system alerts a driver of crash or near-crash events. This is similar to the 3-class ADAS system.

2. Conservative ADAS: in addition to the Basic ADAS's alarms, this system also gives warnings for any un-safe or distracted driving behavior (as causes of potential crashes or near-crashes).

Table 2.16 shows the alert modes for the Conservative ADAS. The conservative ADAS, as its name implies, may not be favorable for more advanced or aggressive drivers. But, parents with teenage drivers or elderly drivers with limited abilities can benefit from the timely alerts of un-safe or distracted driving modes.

Table 2.16 Alert modes for the proposed ADAS.

Classes	**Status**	**Color-code**	**Voice Alert**
Crash	Crash	Red	Yes
Near-crash	Near-crash	Orange	Yes
Baseline_3	Un-safe Driving	Amber	Yes
Baseline_2	Distracted Driving	Yellow	Yes
Baseline_1	Safe	Green	No

To investigate the performance of the 5-class ADAS, We run ten cases with the elastic-net parameter set $\alpha = \{0,0.25,0.5,0.75,1\}$ and the class weights $\boldsymbol{w} = \{(1,1,1,1,1),(1,1,1,3,5)\}$. The weight vector $\boldsymbol{w} = (1,1,1,1,1)$ represents the equally-weighted case while $\boldsymbol{w} = (1,1,1,3,5)$ assigns weights of one, one, one, three and five to the *Baseline-1*, *Baseline-2*, *Baseline-3*, *Near-Crash* and *Crash* instances, respectively. Table 2.17 shows the values of the hyper-parameters of ten cases for model selection of 5-class driving outcome. Figure 2.15 and Table 2.18 show the ten-fold cross-validated error rates for ten cases in Table 2.17.

Table 2.17 Candidate-model parameters to select the best model through cross validation for the 5-class driving outcome.

Case	**1**	**2**	**3**	**4**	**5**
Weights	(1,1,1,1,1)	(1,1,1,1,1)	(1,1,1,1,1)	(1,1,1,1,1)	(1,1,1,1,1)
Alpha	0	0.25	0.5	0.75	1

Case	**6**	**7**	**8**	**9**	**10**
Weights	(1,1,1,3,5)	(1,1,1,3,5)	(1,1,1,3,5)	(1,1,1,3,5)	(1,1,1,3,5)
Alpha	0	0.25	0.5	0.75	1

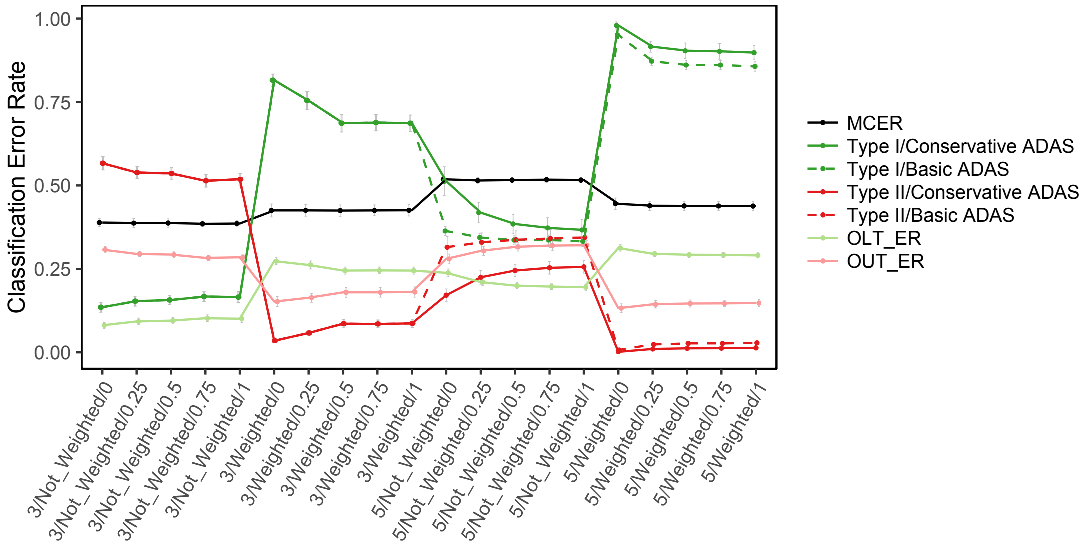


Figure 2.15 Ten-fold cross-validated estimates of the Miss-Classification Error Rate (MCER), Type I and Type II error rates for the Basic and Conservative ADAS's, Off-diagonal Lower Triangular Error Rate (OLT_ER), and Off-diagonal Upper Triangular Error Rate (OUT_ER), calculated at $\boldsymbol{\lambda_{min}}$, for the ten candidate models in **Table 2.17** of the 5-class driving outcome.

Table 2.18 Ten-fold mean cross-validated error rates and their standard deviations in gray font, calculated at $\boldsymbol{\lambda_{min}}$, for the ten candidate models in **Table 2.17** of the 5-class driving outcome.

				Conservative ADAS		Basic ADAS	
Case	Weights	alpha	MCER	Type I	Type II	Type I	Type II
1	(1,1,1,1,1)	0	0.508	0.451	0.199	0.336	0.328
			0.009	0.029	0.014	0.011	0.029
2	(1,1,1,1,1)	0.25	**0.504**	**0.399**	**0.225**	**0.310**	**0.353**
			0.014	0.030	0.016	0.010	0.027

3	(1,1,1,1,1)	0.5	0.506	0.392	0.228	0.312	0.355
			0.013	0.026	0.017	0.011	0.029
4	(1,1,1,1,1)	0.75	0.505	0.393	0.227	0.305	0.358
			0.014	0.034	0.017	0.012	0.028
5	(1,1,1,1,1)	1	0.505	0.393	0.227	0.305	0.358
			0.014	0.034	0.017	0.012	0.028
6	(1,1,1,3,5)	0	0.441	0.964	0.004	0.926	0.011
			0.013	0.017	0.002	0.013	0.004
7	(1,1,1,3,5)	0.25	0.436	0.873	0.016	0.825	0.034
			0.015	0.023	0.004	0.014	0.005
8	(1,1,1,3,5)	0.5	0.436	0.873	0.016	0.828	0.033
			0.016	0.021	0.004	0.012	0.005
9	(1,1,1,3,5)	0.75	**0.436**	**0.872**	**0.016**	**0.828**	**0.033**
			0.016	0.019	0.004	0.011	0.005
10	(1,1,1,3,5)	1	0.437	0.872	0.016	0.828	0.034
			0.016	0.017	0.004	0.011	0.005

The best equally-weighted model is the elastic net with $\alpha = 0.25$ and the overall MCER of 0.504, for the Conservative ADAS: Type I and II error rates are 0.399 and 0.225; while for the Basic model these rates are 0.310 and 0.353. Since, the Conservative ADAS is more sensitive and reacts to any unsafe driving situations, its Type II error which is at the expense of a larger Type I error. For the weighted model, the elastic net models with $\alpha = 0.25, 0.5,0.75$ performs similarly. Among them, the model with $\alpha = 0.75$ is selected since it is the most regularized model with a slightly lower type I error. The overall MCER is 0.436 (a decrease from 0.504), for the Conservative ADAS: Type I and II error rates are 0.872 and 0.016; while for the Basic model these rates are 0.828 and 0.033. The best model, a model which can best detect the unsafe driving situations, is the weighted elastic net for the 5-class driving outcome.

Eventually, the complete dataset was used to estimate the coefficients of the best model ($\alpha = 0.75, \boldsymbol{w} = (1,1,1,3,5)$ and number of classes=5). Figure 2.16 shows the regularized regression coefficient paths for this model. There are five coefficient-paths plots in this figure each of which relates to one of the five classes of driving outcome. The vertical dashed lines indicate the point at which the cross-validation MCER is the smallest. The colored paths in these plots represent the top twenty predictors that are selected through the elastic net algorithm.

Table 2.19 shows the confusion matrix of this best model. As it can be seen, the majority of the misclassifications have populated the lower triangular of the matrix. This means that false alarm rate is high but that sixty one (61) percent of crashes and seventy six percent (76) of near-crashes were identified correctly.

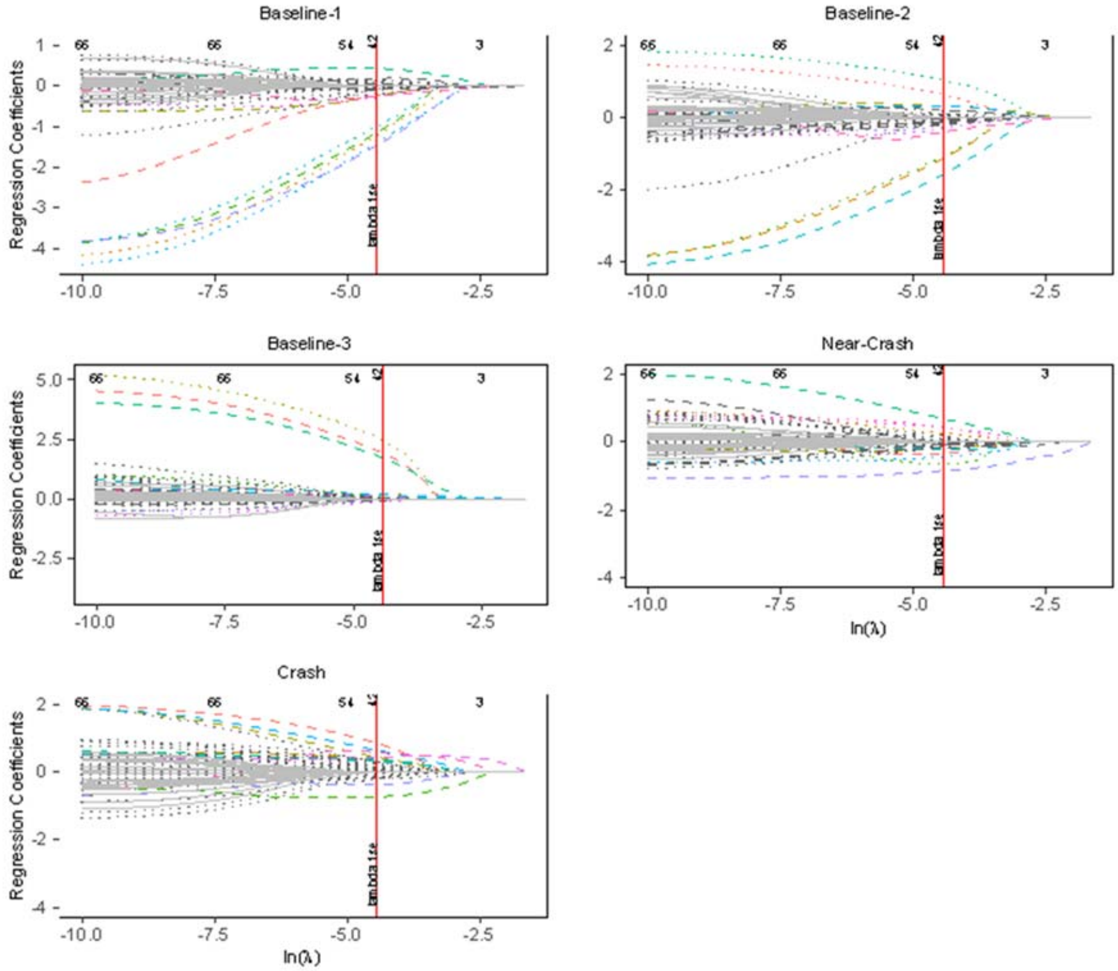


Figure 2.16 Coefficient paths for the best classification model of the 5-class driving outcome. Each plot shows the estimated regression coefficients (paths) of one of the five classes in the multinomial logistic model.

Table 2.19 Confusion matrix of the best 5-class classification model with $\boldsymbol{\alpha = 0.75}$, $\boldsymbol{w = (1, 1, 1, 3, 5)}$.

		Actual				
		Baseline I	Baseline II	Baseline III	Near Crash	Crash
Predicted	Baseline I	228	135	1	28	9
	Baseline II	5	332	6	48	34
	Baseline III	0	0	29	6	2
	Near-crash	1029	1050	185	2000	430
	Crash	519	566	185	562	742
	Sum	1781	2083	406	2644	1217

2.11 Conclusion

In this chapter, we presented an integrated traffic safety platform to develop a real-time individualized risk prediction model to be used in an Advanced Driver Assistance System. We first introduced our methodology to calculate the likelihood of adverse driving events. We proposed using the elastic net regularized regression model with a built-in variable selection and shrinkage mechanism and a cost-sensitive loss function for imbalanced data. We introduced five measures of goodness (lack of goodness) to evaluate the performance of the prediction model, namely miss-classification, Type I, Type II, Upper Off-diagonal Triangular and Lower Off-diagonal Triangular error rates to take into account the sensitivity and specificity of the classifier in identifying cases in minority classes. We used 10-fold cross validation to evaluate the prediction performance of the trained models on the testing data.

We used a subset of SHRP-2 NDS safety data to show the applicability of our platform. We presented a detailed explanation of our data preparation and feature engineering to prepare this dataset for the training of the proposed prediction model. The prediction model was used in an ADAS to warn drivers of critical event and/or unsafe driving situations. In particular, we designed two distinct ADAS's, called 1-Basic ADAS and 2-Conservative ADAS, according to their sensitivities to critical driving events. The Basic ADAS used the trichotomous driving outcome as its response variable and alerted drivers of crashes and near-crashes. The best weighted elastic net model for the Basic ADAS, with $\alpha = 0.75$ and W=(1,3,5), resulted in an overall MCER of 0.419, and Type I and II error rates of 0.636 and 0.106 respectively.

The Conservative ADAS further broke down the baseline outcomes into three new levels according to their incurred risk severities. This system warned drivers of unsafe and distracted driving situations in addition to the crash and near-crash events. The Conservative ADAS system has a higher sensitivity toward critical events and unsafe driving situations. The best weighted elastic net model for the Conservative ADAS, with $\alpha = 0.75$ and W=(1,3,5), resulted in an overall MCER of 0.436, and Type I and II error rates of 0.872 and 0.016 respectively. The Conservative system may not be favorable by more experienced or aggressive due to its higher false alarm rates but parents with teenage drivers or elderly drivers with limited abilities can benefit from the timely alerts of additional un-safe or distracted driving modes.

For any given case, the effort to reduce one type of error generally results in increasing the other type of error. Since the main objective of the ADAS is to identify the critical events, we may the increased type I error. One approach to simultaneously reduce type I and II errors of the classification model in an imbalanced data setting is to collect more data to increase the sample sizes of minority classes. In fact, in this study we have only used one third of the SHRP-2 data. Including more crash and near-crash cases most likely improve the prediction performance of the prediction model. Another approach to tackle this problem is to use re-sampling methods such as bootstrapping which can be the subject of a future work.

Disclaimer

The contents of this chapter reflect the views of the author, who is responsible for the facts and the accuracy of the information presented herein. The U.S. Government assumes no liability for the contents or use thereof.

Furthermore, the findings and conclusions are those of the author and do not necessarily represent the views of the VTTI, SHRP 2, the Transportation Research Board, or the National Academies.

3 A HYBRID PHYSICS/DATA-DRIVEN APPROACH FOR A PERSONALIZED FORWARD COLLISION WARNING SYSTEM

3.1 Introduction

In this chapter, we propose a hybrid physics/data-driven approach which utilizes both the laws of physics governing moving objects and the supplemental data explaining driver and his/her surrounding conditions to assess traffic safety risks. For each type of traffic conflicts, there exist a physics-based model, which explains the relationships among vehicle's kinematic and dynamic variables. On the other hand, the parameters of the physical model, such as speed, acceleration or the driver's reaction time are affected by the context variables, such as weather, surface condition, daylight; and driver's intrinsic characteristics such as driver's cognitive abilities and demographics. A driver's demographics, including his/her age and sex can result in different evasive maneuvers to avoid a specific type of crash. In particular, we focus on a Forward Collision Warning (FCW) technique that uses Brill's one-dimensional stop-to-break model [58]. The parameters of this model are speed and acceleration of the following and lead vehicles and the following car's temporal headway and reaction time. The time of issuing the warning alert can be determined from these parameters. The challenge is that these parameters are not deterministic and in real-world traffic scenarios, the surrounding conditions, vehicle's condition, driver characteristics and driving behavior confound their values. Stochastic, probabilistic and/or statistical modeling techniques can be used to capture the nondeterministic nature of these parameters according to the contributing factors.

The proposed methodology can be used to enhannce the perfomance of Advanced Driver Assistance Systems by customizing the alerts according to driver intrinsic characteristics and driving behavior. An effective ADAS is expected to give a safety alert sometime before the driver realizes the presence of a rear-end collision's risk in the hope of shortening the response time and evading a crash. Therefore, the use of a personalized reaction time instead of an average value for all drivers and under any driving conditions will enhance the performance of the ADAS in issuing more timely alerts. One can also use our model to design impactful countermeasures or mitigation tools focusing on human behavioral characteristics. For example, educational and enforcement campaigns, for programs targeting drunk driving or seatbelt use, can be modified to frame the accepted beliefs about safety within a specific area or region. Ultimately advanced knowledge in this area can influence the effectiveness of both behavioral, enforcement and infrastructure safety programs. Information can be provided to drivers that improve their situational awareness while driving and allow them to make driving decisions based on safety risk. Although, the immediate benefit of such real-time information system has to be studied in another experiment before implementation.

There are many factors affecting a driver's reaction time yet unexplored by the driver modeling literature due to the lack of sufficient observational data. For a long time, it has been a common practice to use a nominal value, the mean or the 95th percentile of the reaction time distribution of participants in either an experimental study or a driving simulator. This research is an effort to investigate the effects of driver characteristics and driving behavior on the driver's stop-to-brake reaction time in real-world driving scenarios. In particular, we propose building a hierarchical regression model, which can capture the variations attributed to driver characteristics and driving behavior. We use SHRP-2's Naturalistic Driving Study (NDS) data [9], the largest and most comprehensive study of its kind, to model the driver's brake-to-stop response time. The results show that the inclusion of driver characteristics decreases the cross-validated mean squared error of the reaction time prediction model by an average of 24%. It also increased the precision of the model in correctly predicting the longer reaction times (>2.5 seconds) by an average of 27%. The explained variation by the driver's intrinsic characteristics and driving behavior supports the necessity of developing personalized Advanced Driver Assistance Systems to enhance the performance and increase their acceptance by users.

The organization of this chapter is as follows: In section 4.2, the background and literature review of the problem in hand is presented. In section 4.3, the problem statement is presented. Model formulation, including Brill's model of the car following behavior and the statistical model, which predicts the parameters of the kinematic model, is presented in section 4.4. Section 4.5 presents the numerical results using SHRP-2 NDS data. Finally, conclusions and directions for future research are presented in section 4.6.

3.2 Background and Literature Review

According to the National Highway Traffic Safety Administration (NHTSA), rear-end collisions account for approximately 23 percent of all motor vehicle crashes [59]. In 2012 alone, more than 1.7 million rear-end crashes occurred on US roadways, resulting in more than 1,700 fatalities and 500,000 injured people. The National Transportation Safety Board (NTSB) estimated that 80 percent of the deaths and injuries resulting from rear-end collisions could be prevented by collision avoidance systems. The first demonstration of a forward collision avoidance system dates back to 1995 by a team of scientists and engineers at Hughes Research Laboratories in Malibu, California. While primarily a warning system with various feedbacks, the system did have only a minor control of the brakes, which were pulsed to begin a braking action in the event of a potential collision, making it also the beginning of avoidance systems. It took almost 20 years for this technology to reach the consumer marketplace. Since then, these systems have evolved significantly from a mere warning system to smart automated braking systems. Integrated safety systems for rear-end crashes can be broadly divided into three categories [60]:

1. Forward collision warning (FCW): sensors detect a potential collision and warn the driver.
2. Collision mitigation braking systems (CMBS): sensors detect a potential collision but take no immediate action to avoid it. Once the sensing system has detected that the collision has become inevitable regardless of braking or steering actions then emergency braking is automatically applied (independent of driver action) to reduce the collision speed, and hence injury severity, of the collision.
3. Collision avoidance: Sensors detect a potential collision and take action to avoid it entirely, taking control away from the driver.

The demand for the Advanced Driver Assistance Systems (ADAS), including the forward collision mitigation systems, is expected to increase substantially in the coming years.

The Insurance Institute for Highway Safety have estimated that if all vehicles had forward collision and lane departure warning, blind spot assist, and adaptive headlights, about 1 in 3 fatal crashes and 1 in 5 injury crashes could be prevented[8].

The Society of Automotive Engineers' (SAE) vehicle standards committee has defined six levels of driving automation levels, namely L0 to L5 spanning from no automation (as in regular cars) to full automation [61]. Automated driving innovations could dramatically decrease the number of crashes tied to human choices and behavior through technologies that corrects for human mistakes or takes over the full driving responsibility. Experts optimistically estimate that advanced vehicle technology can reduce the number of crashes by up to 90 percent by eliminating the primary cause of or the contributing factor to crashes that is the human error[9]. Although there will be a significant growth in the number of autonomous vehicles by 2030, non-autonomous cars will make at least 85% of the traffic mix[10]. Furthermore, in vehicles with less than full automation (i.e., L1 to L4), the system can only drive the car under specific conditions, and still the human driver needs to be ready to take back control of the vehicle when necessary and drive under difficult conditions. Last but not least, combining autonomous and non-autonomous vehicles in a single traffic network will bring about unimaginable traffic safety challenges and the most difficult time is expected to be the transition period, while all kinds of cars will share the road before self-driving ones predominate. Therefore, it is necessary to enhance the performance of the present Driver Assistance Systems for the

[8] The Insurance Institute for Highway Safety, New estimates of benefits of crash avoidance features on passenger vehicles, available from http://www.iihs.org/iihs/sr/statusreport/article/45/5/2

[9] Ten ways autonomous driving could redefine the automotive world, McKinsey & Company Podcast, June 2015; available from http://www.mckinsey.com/industries/automotive-and-assembly/our-insights/ten-ways-autonomous-driving-could-redefine-the-automotive-world.

[10] Self-driving Cars and The Future of the Auto Sector, McKinsey & Company Podcast, August 2016; available from http://www.mckinsey.com/industries/automotive-and-assembly/our-insights/self-driving-cars-and-the-future-of-the-auto-sector.

lower classes of vehicles to ensure a safe and smooth transition to the future of transportation.

According to a study [50] conducted by Delphi Electronics & Safety collision warning algorithms use one of the following criteria: *time-headway*, *time-to-contact*, or the *underlying kinematic constraints*. Although *time-headway* (also called temporal headway) algorithms offer simplicity and are consistent with current driving-manual recommendations for safe driving, they are insensitive to relative velocity. *Time-to-contact* (also called Time-To-Collision) algorithms (e.g. [62] and [63]) are based on D. Lee's theory of direct time-to-contact perception [64], and are sensitive to relative velocity. Algorithms based on kinematic constraints offer increased accuracy by calculating the moment that the driver must initiate braking, given an assumed reaction time and host-vehicle deceleration response. Because this class of algorithms considers both reaction time and the capacity of the host vehicle to decelerate, it offers a more comprehensive model than the other two categories. Algorithms of this class are highly dependent on assumptions about driver *reaction time* and braking rate.

A driver's *reaction time*, sometimes called *response time*, consists of two elements: Perception Reaction Time (PRT) and Maneuver Time (MT). Perception Reaction Time is the time it takes for the driver to realize that a reaction is needed due to a road condition, decide what maneuver is appropriate (in the case of rear-end collision, stopping the vehicle), and start the maneuver (taking the foot off the accelerator and depressing the brake pedal) [65]. Maneuver Time, also called Movement Time, is the time it takes to complete the maneuver (decelerating and coming to a stop). Figure 3.1 shows the elements of *reaction time* as the sequence of events take place in a rear-end collision scenario. In this scenario, the risk becomes present when a lead vehicle starts slowing down or coming to a stop. An effective ADAS is expected to give a safety alert sometime before the driver realizes the presence of a rear-end collision's risk in the hope of shortening the response time and evading a crash.

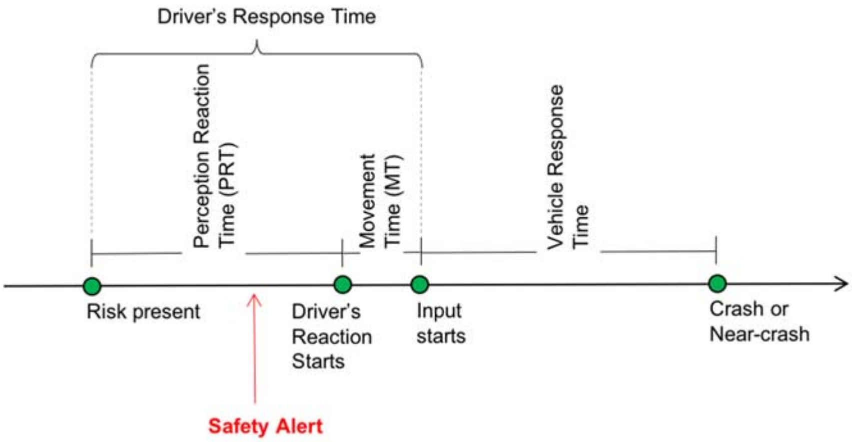


Figure 3.1 Sequence of events during a crash or near-crash **[66]**.

Although *response time* vary significantly according to different safety factors, it has been a common practice to use a nominal value, usually the mean or the 90th percentile, in traffic models especially for accident reconstruction and causal analysis. For example, the design standards of the American Association of State Highway and Transportation Officials (AASHTO) allow 1.5 seconds for perception time and 1.0 second for maneuver time [67]. In fact, many researchers have used the 2.5-second standard for the reaction time value. These fixed values are calculated in experimental studies, which are unable to capture the effects of driving behavior in most of the real world scenarios and varying contexts. Furthermore, despite the fact that response times are skewed to the right, measured response times are often reported in the literature as mean values, making it difficult to estimate values away from the mean. A few authors have presented response times as a distribution. For example, Taoka [67] describes a distribution of brake reaction times based on work by Sivak et al. [68], and Eberhard et al. [69] provide a summary of different distributions.

Most studies have estimated the reaction time based on indoor experiments and driving simulators [70]. For example, in the study by Johansson and Rumer [71], 321 subjects were instructed to brake pedal as soon as they heard a sound. The estimated reaction time varied from 0.4 second to 2.7 seconds with a mean, and standard deviation of 1.01, and 0.37 seconds. Since the drivers were informed that they were participating in a brake reaction study and the use of sound as stimulus, these values may be biased. A recent study using both a real driving environment and a simulator [72] shows that the reaction time of drivers to an anticipated danger in a real environment has a mean value of 0.42 seconds and a standard deviation of 0.14 seconds. The same study also shows that the mean value of the reaction time distribution to an unanticipated danger by extreme braking is about 1.1 seconds and that in a simulator it is about 0.9 seconds. In real traffic, the driver reaction to expected and unexpected stimuli are also different [73]. Fambro reported that the mean reaction times for unexpected and expected stimuli are 1.3 seconds and 0.7 seconds, respectively. Ranjitkar et al. [74] applied the graphical method in stability analysis of car-following behaviors, and based on car-following data collected on a test track, they estimated that the average driver reaction time for individual drivers ranged from 1.27 to 1.55 seconds.

Chandler et al. [75] developed a linear car-following model using eight male drivers. Their estimate of the reaction time was approximately 1.5 seconds. Gipps [76] did not estimate the individual reaction time, but instead used constant reaction time of 2/3 seconds for all drivers. Lerner et al. estimated the reaction time distribution from a sample of 56 drivers in real traffic scenarios [77]. To estimate the brake reaction time for unexpected situations (to mimic real driving conditions), subjects were not informed that they were participating in a brake reaction time study. When a subject reached the test site at 40 mph speed, a large yellow highway crash barrel was released approximately 200 feet in front of the vehicle. The brake reaction time varied from 0.7 to 2.5 seconds with a median, mean, and standard deviation of 1.44, 1.51, and 0.39 seconds, respectively. In overall, the current state of the art of reaction time modeling lacks thorough studies that relate reaction time variations to individual drivers while accounting for driving context factors.

The present commercial products are designed for an average driver which can be too conservative for a more experienced or aggressive driver, or ineffective for a more vulnerable driver such an elderly or a young inexperienced driver. This may lead to a higher rate of false alarms and consequently a driver's mistrust in the system. Over the past decade, there has been significant research effort dedicated to the enhancement of forward collision mitigation systems, intended to improve safety by monitoring the driver and the on-road environment. The most recent advances in the collision warning and avoidance technologies are the cooperative and predictive driver assistance systems which fuses data from additional sources, such as the driver, near-by vehicles or infrastructure, in order to enhance the performance of their risk assessment under different conditions [78]. For example, one way to enhance a driver assistance system is to take into account the characteristics and dynamic behavior of each individual driver for a more impactful and personalized warning system. There has recently been a handful of researches trying to address this problem. For example, Butakov and Ioannou [79] developed a methodology that learns the characteristics of an individual driver/vehicle response before and during lane changes and under different driving environments. They have developed a two-layer model to describe maneuver kinematics. The lower layer describes lane change as a kinematic model. The higher layer model establishes the kinematic model parameter values for the particular driver and represents their dependence on the configuration of the surrounding vehicles.

Our work benefits from a recent large-scale observational SHRP-2 study of driving behavior. We build a statistical model to estimate the brake-to-stop reaction time in rear-end conflict scenarios in relation to driver's intrinsic characteristics and other additional context variables. We show that by including driver characteristics, we can explain some of the variations in the driver's reaction time attributed to individual differences. This approach requires drivers to provide information about their demographics, sleep habits, driving history and knowledge and cognitive, visual and ADHD test results. We demonstrate that our model significantly results in a more realistic estimation of driver's reaction time, which could in turn lead to the design of more effective personalized ADAS.

3.3 Problem Formulation and Preliminaries

The problem of interest is to build a hybrid physics/data-driven traffic model in relation to roadway and driver's intrinsic characteristics and environmental factors. Figure 3.2 shows the framework for our proposed model for a personalized collision warning system.

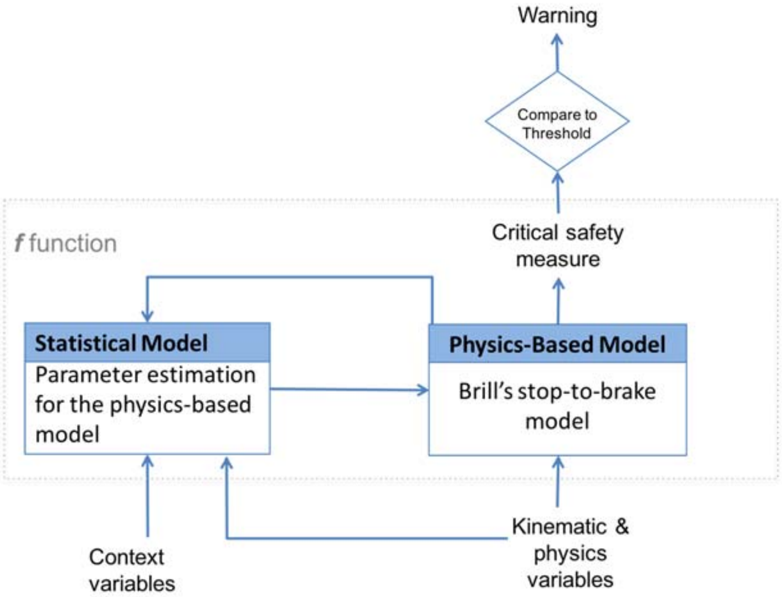


Figure 3.2 A hybrid physics/data-driven model for a personalized collision warning system.

Our model takes advantage of a causal model (Figure 3.3) that explains the relationships among safety factors and driving outcome [80]. In this figure endogenous variables are grouped into observed and unobserved variables, where observed (manifest) variables are shown inside rectangles, and unobserved variables (latent) inside ellipses. Unobserved variables are those variables that are not measured directly but they are rather created as constructs of observed variables. For instance, a driver's dynamic behavior is a product of his/her individual intrinsic characteristics in conjunction with his/her interactions with the surrounding environment. We can also have relationships among exogenous variables, shown by curved arrows, such as the impact of weather and time on roadway's dynamic conditions and other road user's behavior.

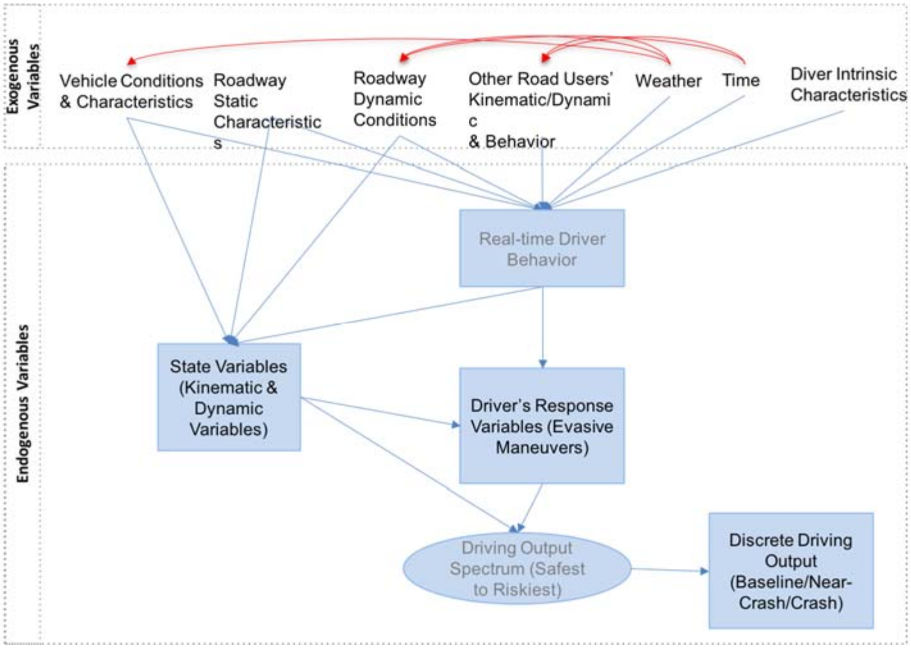

Exogenous Variables
Vehicle Conditions & Characteristics
Roadway Static Characteristics
Roadway Dynamic Conditions
Other Road Users' Kinematic/Dynamic & Behavior
Weather
Time
Diver Intrinsic Characteristics
Endogenous Variables
Real-time Driver Behavior
State Variables (Kinematic & Dynamic Variables)
Driver's Response Variables (Evasive Maneuvers)
Driving Output Spectrum (Safest to Riskiest)
Discrete Driving Output (Baseline/Near-Crash/Crash)

Figure 3.3 Causal model of traffic safety incidents.

Some of the exogenous variables have direct impacts on state variables such as vehicle make and model and its maintenance condition, grade of roadway, or surface condition (friction); while others such as misty weather, a crossing pedestrian, rush hour, or a driver's sleepiness affect the state variables through changing driver's behavior in response to these factors (more cautious driving => slowing down in response to misty weather). State variables together with real-time driver behavior can precipitate a potential traffic conflict. At this point, the driver may take some evasive actions, i.e. braking, accelerating and steering or a combination of the three, to avoid the conflict. We refer to these evasive actions as driver's response variables. The results of driver response in conjunction with state variables will define the driving outcome. Driving outcomes have been traditionally classified as either crash or no crash in the past. Naturalistic driving studies have made it possible to add another class of driving outcome as the class of near-crash. But the reality is that the driving outcome can be seen as a risk spectrum ranging from the safest mode (a near-zero chance of conflicts) to the riskiest mode of driving (a major fatal crash), and in the mid-range there will be mild to significant chances of near-crashes.

Referring to our causal model, state and driver response variables can be modeled in relation to driver behavior making them customized per individual driver. For example, driver's reaction time is one of the parameters of the kinematic model and space state model for trajectory reconstruction. This parameter itself can be regressed against the values of the real-time driver's behavior and his/her intrinsic characteristics under varying driving conditions. For example, according to [23], brake reaction to an unexpected condition is faster in older drivers than younger drivers. Therefore, depending on the age group of the driver, his/her reaction time can assume different values. We will use regularized regression models to estimate the parameters of the physical model in relation to the context variables. Then, the estimated parameters will be used in the physics-based model.

To illustrate the hybrid physics/data-driven approach, we consider the simple one-dimensional trajectory model of a rear-end collision (striking) scenario with braking as the only evasive maneuver. We will use Brill's brake-to-stop model originally proposed in 1972 [58]. Figure 3.4 shows the parameters of this model. A crash occurs when the available distance to stop for the follower vehicle is less than the distance needed to stop without striking the lead vehicle [81], i.e.:

$$v_2(t_2).r_2 + \frac{v_2(t_2)^2}{2a_2} < v_2(t_2).h_2 + \frac{v_1(t_1)^2}{2a_1} \tag{3.1}$$

where t_1 and t_2 are the time epochs at which the lead vehicle and the following vehicle push the brakes, respectively. Furthermore, $v_1(t_1)$, and a_1 are the initial speed and braking deceleration of the lead vehicle; $v_2(t_2)$ and a_2 the initial speed and braking deceleration of the following vehicle; and h_2 and r_2 are the following driver's temporal headway and braking reaction time, respectively. The variables $v_1(t_1)$, a_1, $v_2(t_2)$, a_2, h_2 and r_2 are referred to as the Brill elements.

Here is a list of Brill's elements:

t_1: time epoch when the lead vehicle's driver brakes,

t_2: time epoch when the following vehicle's driver brakes,

t_3: time epoch when the lead vehicle stops,

t_4: time epoch when the following vehicle stops,

$v_1(t_1)$: speed of the lead vehicle when braking begins,

$v_2(t_2)$: speed of the following vehicle when braking begins,

h_2: the following vehicle's temporal headway when the lead vehicle brakes,

r_2: the following vehicle's braking reaction time,

a_1: braking deceleration used by the lead vehicle (stopping deceleration),

a_2: braking deceleration used by the following vehicle (stopping deceleration).

From this point forward and in accordance with the related literature on reaction time, we use the notation, τ, to refer to the driver's reaction time of the following car.

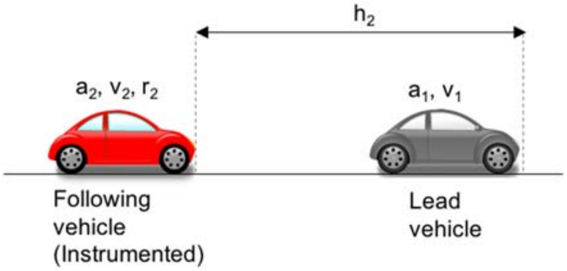


Figure 3.4 Rear-end collision scenario: parameters of the simple brake-to-stop model.

We define the driving outcome, based on a quantitative criterion, d, which is calculated according to the laws of physics of moving objects [82]. At a higher level, the parameters of this physics-based model can be estimated according to the additional information obtained from the external surrounding conditions, driver's characteristics and behavior or vehicle's specifications and conditions. Equation (3.2) shows the calculation of d.

$$d = v_2(t_2).h_2 + \frac{{v_1}^2(t_1)}{2a_1} - v_2(t_2).\tau - \frac{{v_2}^2(t_2)}{2a_2} \tag{3.2}$$

When d is equal to 0, a rear-end crash happens, and small values of d can suggest a near rear-end crash though it should be interpreted in relation to the relative speed of the two vehicles and traffic flow. Then, similar to Equation (2.1) in chapter 2, the driving outcome at time t and location l of driver i in his or her trip j can be defined as follows:

$$Y_{ijlt} = \begin{cases} Normal\ driving, & If\ d_{ijlt} > Tr_{NC}, \\ Near-crash, & If\ Tr_C < d_{ijlt} \leq Tr_{NC}\ , \\ Crash, & If\ d_{ijlt} \leq Tr_C. \end{cases} \tag{3.3}$$

Where Tr_{NC} and Tr_C are the thresholds for near-crash and crash (low-level and high-level severity) warnings. By computing d_{ijlt} in real time and comparing it to the critical thresholds, the driving outcome can be determined at any given time and location for a specific driver during his/her trips. To apply this kinematic model in a warning system, a method to grading the severity of vehicle interactions needs to be developed to define the thresholds and rules. For example, Smith et al. has developed such a method to support evaluation of collision warning systems [83]. Using driver behavior data observed in driving simulators and test-track experiments, the parameters' curves were then used to partition the set of possible values into subsets reflecting crash, near-crash, conflict, and low-risk situations. Similarly, the values of Tr_{NC} and Tr_C can be defined according to the distribution of d at the onset of critical events compared to values at normal driving conditions. This approach can be extended to other types of traffic conflicts, such as a collision with an adjacent car or conflict with a following vehicle (struck versus striking); and also to more detailed models such as a 2-dimensional trajectory model that can account for both braking/accelerating and steering maneuvers. The design of the FCW system, i.e. the determination of threshold values and the Human-Machine Interface, and the user acceptability are beyond the scope of this work. Our contribution is to the use of additional safety-related data, i.e. the individual driver's characteristics and surrounding driving conditions, to enhance and personalize a kinematic-based FCW system.

Figure 3.5 shows the use of additional safety data in our proposed hybrid physics/data-driven model to enhance the collision warning timing. In the Brill's model, there are two components defining whether a critical event would turn into a crash, namely, the driver's *response time* and the *braking deceleration* of the following vehicle. A critical event where the *response time* of the driver is longer than his or her following headway would lead to a rear-end crash unless his or her braking deceleration is greater than the lead vehicle's. In addition to the individual characteristics, a driver's reaction time may depend on some roadway characteristics such as road's grade; network condition such as traffic density and environmental factors such as lighting, weather and surface condition.

Last but not least, a driver's real-time behavior such as speeding or his/her involvement in a secondary task can affect the length of reaction time. The focus of this study is on the modeling and prediction of the driver's *response time* according to individual characteristics and real-time behavior in varying driving contexts. Similarly, a vehicle's response time, varies according to the vehicle and roadway characteristics and conditions and affect the actual *braking acceleration* which can be modeled accordingly and is beyond the scope of this work.

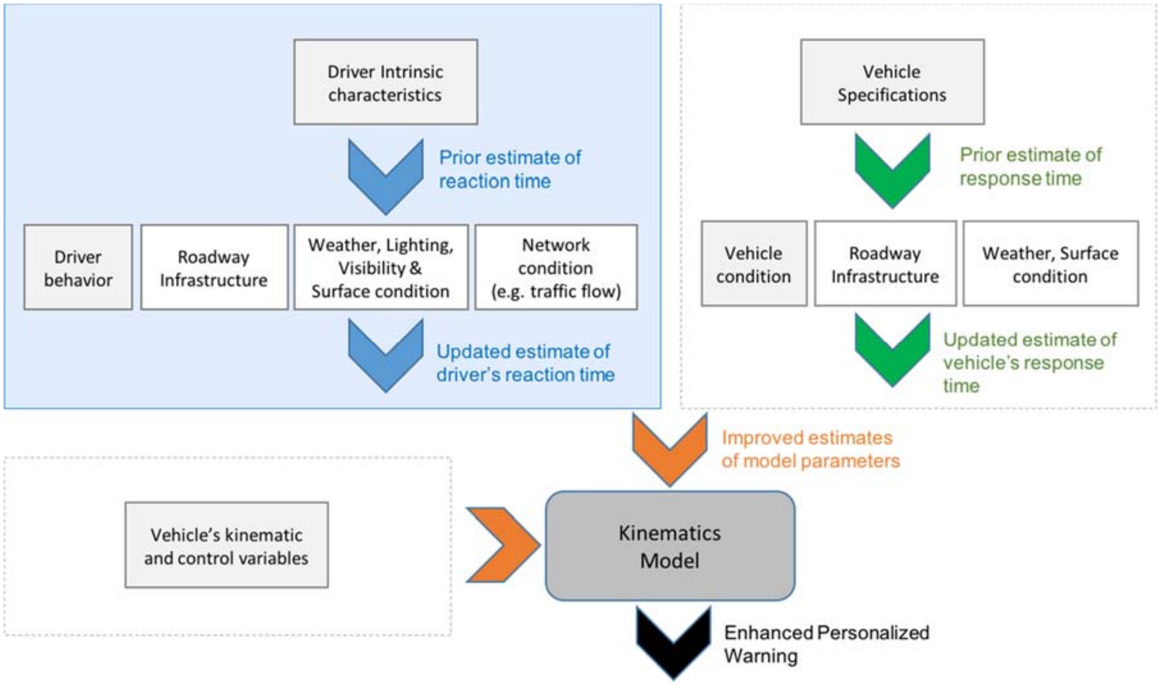

Driver Intrinsic characteristics
Prior estimate of reaction time
Driver behavior
Roadway Infrastructure
Weather, Lighting, Visibility & Surface condition
Network condition (e.g. traffic flow)
Updated estimate of driver's reaction time
Vehicle Specifications
Prior estimate of response time
Vehicle condition
Roadway Infrastructure
Weather, Surface condition
Updated estimate of vehicle's response time
Improved estimates of model parameters
Vehicle's kinematic and control variables
Kinematics Model
Enhanced Personalized Warning

Figure 3.5 Breakdown of the elements of a hybrid data-driven/kinematics ADAS.

It is well known that people vary in all sorts of ways. In predicting the drivers' reaction time, it is reasonable to assume that distributions will vary across individuals. This assumption, however, is violated by aggregating data across individuals. It has been repeatedly demonstrated that aggregating data across people or items may distort the estimate of a functional relationship [84]. There are three approaches to deal with the individual-level variation. The first approach is to use a multi-level model in which data are structured in groups and coefficients can vary by group. This means that the reaction time model can have different coefficients for each driver. Equation (3.4) shows a varying-intercept, varying-slope multi-level model:

$$\tau_i = \beta_{0j[i]} + \boldsymbol{xC}_i^T.\boldsymbol{\beta}_{\boldsymbol{j}[\boldsymbol{i}]} + \varepsilon_i \tag{3.4}$$

where $j[i]$ indexes the driver for event i. For example, if j[35] = 4, then the 35th event in the data (i = 35) belongs to driver 4. The first requirement to use this approach is to have enough number of drivers and replications per individual driver to build a multi-level model. Otherwise, the multi-level regression will reduce to a classical single-level regression model. The second approach is the inclusion of categorical predictors using indicator variables. This means that if a cohort of J drivers uses the proposed personalized ADAS, the model will choose one of the drivers as the baseline and include indicator variables for other J-1 drivers. The coefficient for each driver then represents its comparison to the baseline individual. Yet again, this approach requires replications per individual drivers to correctly estimate the individual-level coefficients of indicator variables. The third approach is to include a set of predictors to the model, which can collectively explain the behavior of an individual driver. Inclusion of these driver-specific variables, if significant, will explain the variation in reaction time due to individual-level differences. In this study, we apply the third approach due to the characteristics and limitations of the SHRP-2 NDS data. The main limitation is that there are not statistically sufficient critical events per individual driver, in particular for the rear-end crash and near-crash events. For the rear-end critical events, about 83% of drivers have only one event, about 12% have two events and only 5% have more than two and still less than 5 events. On the other hand, SHRP-2 NDS data provides very comprehensive driver-specific variables, which allows us to capture the individual-level variations.

In the next section, SHRP-2 NDS safety data will be used to model and predict the brake-to-stop reaction time of drivers in rear-end collision critical events, namely crashes and near-crashes through a hierarchical regression modeling, the importance of including both the driver and context variables in predicting the reaction time. Then, a preliminary result and a simple rule-based approach are presented to demonstrate how applying this method in practice can give a driver an extra time to respond to a present rear-end collision risk.

3.4 Data Preparation and Reaction Time Modeling

Our SHRP-2 NDS data subset includes a total of 3,861 critical events including 1,217 crashes and 2,644 near-crashes. Rear-end crash and near-crashes were extracted by filtering the event-detailed variables of INCIDENT TYPE and EVENT NATURE to only include “rear-end, striking” and “conflict with a lead vehicle” instances, respectively. This immediately reduced the sample size to 1239 consisting of 86 crashes and 1153 near-crashes. In order to make this subset as uniform as possible in terms of the vehicle’s physical movements, we applied two more filters on the variable’s PRE-INCIDENT MANEUVER and EVASIVE MANUEVER. PRE-INCIDENT MANEUVER is a vehicle kinematic measure based on what the vehicle does (movement and position of the vehicle), not on what the driver is doing inside the vehicle. As it can be seen in Table 3.1, only the four types of PRE-INCIDENT MANEUVER, “Going straight, constant speed”, “Decelerating in traffic lane”, “Starting in traffic lane” and “Going straight, accelerating” were retained. Similarly, for the EVASIVE MANUEVER, i.e. the subject driver's reaction or avoidance maneuver in response to the event or incident, only the events with “Braked (lockup)” and “Braked (no lockup)” were considered and the rest were discarded (Table 3.2). After applying these two filters the number of events reduced to 55 crashes 765 near-crashes to study the brake-to-stop following behaviors.

Table 3.1 Levels of the PRE-INCIDENT MANEUVER in rear-end critical events.

Pre-incident Maneuver	Number of Near-crashes	Number of Crashes
Going straight, constant speed	458	26
Decelerating in traffic lane	226	21
Going straight, accelerating	219	11
Changing lanes	68	4
Starting in traffic lane	59	16
Negotiating a curve	53	2
Turning left	24	1
Merging	20	3
Turning right	15	2
Passing or overtaking another vehicle	6	0
Stopped in traffic lane	4	0
Making U-turn	1	0
	1153	86

Table 3.2 Levels of the EVASIVE MANUEVER in rear-end critical events.

Evasive Maneuver	Number of Near-crashes	Number of Crashes
Braked (no lockup)	757	41
Braked (lockup)	154	24
Braked and steered right	138	3
Braked and steered left	81	2
Accelerated and steered left	6	0
Steered to right	5	0
No reaction	4	16
Steered to left	3	0
Other actions	2	0

Accelerated	1	0
Accelerated and steered right	1	0
Braked (lockup unknown)	1	0
Total	1153	86

In order to calculate the reaction time, two time stamps in the EVENT DETAILED database were used: *EVENT SATRT* and *SUBJECT REACTION START*. The variable, *EVENT SATRT*, is defined as the time stamp, in milliseconds, at which the precipitating event begins, that is, the point in the video when the sequence of events defining the occurrence of the incident, near-crash, or crash begins. In the case of a rear-end collision this is the time when the lead vehicle starts decelerating or slowing down to stop. Table 3.3 shows the distribution of the PRECIPITATING EVENT in the rear-end crashes. In order to adhere to the Brill's car-following behavior model and to not further complicate the human response phenomena, only the first three levels of the PRECIPITATING EVENT in Table 3.3 where the lead vehicle either *decelerated*, *slowed down to stop* or *stopped* were retained in the data. After applying this last filter, the sample size decreased to 776 including 53 crashes and 723 near-crashes.

Table 3.3 Levels of the PRECIPITATING EVENT in the rear-end critical events.

Precipitating Event	Number of Near-crashes	Number of Crashes
Other vehicle ahead - decelerating	705	18
Other vehicle ahead - slowed and stopped 2 seconds or less	244	34
Other vehicle ahead - stopped on roadway more than 2 seconds	99	31
Other vehicle lane change - right in front of subject	44	0
Other vehicle lane change - left in front of subject	32	1
Other vehicle ahead - at a slower constant speed	18	1
Subject lane change - right behind	2	0

vehicle		
This vehicle lost control - other cause	2	0
Object in roadway	1	0
Other event not attributed to subject vehicle	1	0
Other vehicle - making U-turn	1	0
Other vehicle ahead - accelerating	1	0
Other vehicle lane change - left other	1	1
Subject in intersection - turning right	1	0
This vehicle lost control - excessive speed	1	0
This vehicle lost control - poor road conditions	0	1
Total	1153	87

SUBJECT REACTION START is the timestamp, in milliseconds after the start of the event, when the driver is first seen to recognize and begin to react to the safety critical incidents occurring. It is defined as the first change in facial expression to one of alarm or surprise or the first movement of a body part in a way that indicates awareness and/or the start of an evasive maneuver, whichever occurs first. After applying the necessary filters and using the previously explained timestamps, the brake-to-stop reaction time in seconds can be computed from Equation (3.5).

$$\tau = (SUBJECT\ REACTION\ START - EVENT\ SATRT)/1000 \qquad (3.5)$$

Table 3.4lists SHRP-2 NDS driver basic demographic information, functional ability relative to driving safety and risk, vision tests, cognitive assessments, and physical ability metrics data. For detailed information about each of these data tables, variables and descriptions, data dictionaries are available on https://insight.shrp2nds.us/.

Table 3.4 Summary table of driver characteristics data.

	Data Table	Total Number of Variables	Number of Numeric Variables	Number of Categorical Variables	Number of Required Dummy Variables	Total Number of Numeric and Required Dummy Variables
1	Demographics	47	11	36	150	161
2	Medical Conditions	34	3	31	236	239
3	Sleep Habits	37	6	31	303	309
4	Driving History & Knowledge	18	2	16	358	360
5	Visual Cognitive Test	21	13	8	57	70
6	Conner CPT Clock & Draw Score	14	13	1	6	19
7	Barkley	7	1	6	24	25
8	Risk Perception	32	0	32	224	224
9	Risk Taking	31	0	31	123	123
10	Sensation Seeking	5	5	0	0	5
11	Driver Behavior	24	0	24	121	121
	Total	270	54	216	1602	1656

As it can be seen in Table 3.4, there are eleven main groups of driver variables. The total number of these variables is 270 with 51 numeric and 216 categorical variables. But, in order to include a categorical variable with c levels in a multiple regression prediction model, it needs to be recoded to $c - 1$ dichotomous variables called dummy variables. After the dummy coding step, the number of required dichotomous variables becomes 1602, and the dimensionality increases from 270 to 1656. Since the number N of instances is 776 and the number of potential predictors, p, is 1656, we have a problem where the dimension is significantly larger than the sample size. To solve this problem, Tibshirani and Hastie's proposed model for regularization, called *elastic net,* is used. The elastic net solves the following problem:

$$\min_{\{\beta_0{}^{a},\boldsymbol{\beta}^{a}\}_1^{C}} \left[\frac{1}{2N} \sum_{i=1}^{N} (\tau_i - \beta_0 - x^T\,\boldsymbol{\beta})^2 + \lambda P_\alpha(\boldsymbol{\beta}) \right] \tag{3.6}$$

where $P_{\alpha}(\boldsymbol{\beta})$ is:

$$P_{\alpha}(\boldsymbol{\beta}) = (1-\alpha)\frac{1}{2}(\|\boldsymbol{\beta}\|_2)^2 + \alpha\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p}\left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|\right] \tag{3.7}$$

$\lambda \geq 0$ is a tuning parameter and N is the total number of instances used for the parameter estimation. $P_{\alpha}(\boldsymbol{\beta})$ is the elastic-net penalty that is a compromise between the ridge regression penalty ($\alpha = 0$) and the lasso penalty ($\alpha = 1$). $\|\boldsymbol{\beta}^a\|_1$ and $\|\boldsymbol{\beta}^a\|_2$ are l_1 and l_2 norms, also called Manhattan and Euclidian norms. It is particularly useful in $P \gg N$ situations, or any situation where there are many correlated predictor variables. The ridge regression shrinks the coefficients of correlated predictors towards each other while Lasso is somewhat indifferent to very correlated predictors, and will tend to pick one and ignore the rest. The lasso penalty corresponds to a Laplace prior, which expects many coefficients to be close to zero, and a small subset to be larger and nonzero. Thus, lasso can be used for variable selection. The elastic net with $\alpha = 1 - \varepsilon$ for some small $\varepsilon > 0$ performs much like the lasso, but removes any degeneracies and wild behavior caused by extreme correlations. More generally, the entire family P_{α} creates a useful compromise between ridge and lasso.

To estimate the parameters of the model, i.e. $\beta_0, \boldsymbol{\beta}$, we have used the R package glmenet [8], [85]. The *glmnet* algorithms use cyclical coordinate descent, which successively optimizes the objective function over each parameter with others fixed, and cycles repeatedly until convergence. The Gaussian family for linear regression is applied to fit the model.

Since a response to a collision risk precedes the outcome (crash or near-crash) and the fact that near-crashes are more frequent than crashes, we decided to jointly use the crash and near-crash events to model the brake-to-stop response times. In order to do so, it is necessary to first check if their distributions are statistically identical. Figure 3.6 shows a boxplot of reaction times between crash and near-crash groups.

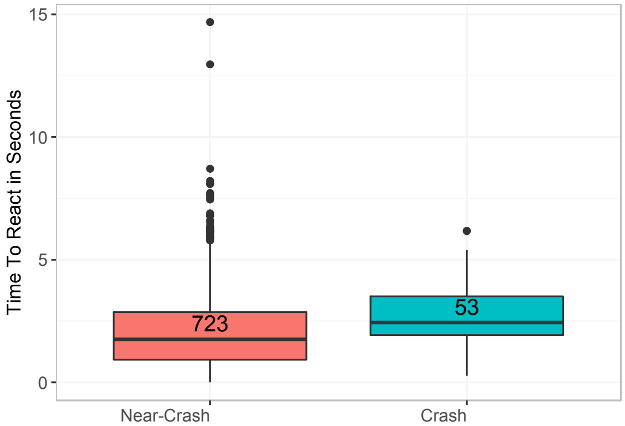


Figure 3.6 Box plot of reaction time between groups of crash and near-crash events**.**

At first glance, there seem to be quite a few outliers and a couple of extreme values. But, it is known that human response times are right-skewed [66]. Therefore, those points that may look to be outliers actually come from the natural distribution of reaction times. The only events that were removed from this dataset were three near-crashes with extreme values for reaction times of 0, 12.96 and 14.69 seconds. After removing the extreme values, the Shapiro-Wilk normality test was applied to test the normality of reaction times. The test resulted in a p-value less than $2.2\times e^{-16}$ concluding that the normality assumption is rejected. Since, the reaction times do not follow normal distribution and is not symmetric, non-parametric hypothesis tests were used to test whether the distributions of reaction times are identical.

To test the difference in the central tendency, the Wilcoxon–Mann–Whitney two-sample rank-sum test was used. Assuming that the rear-end events were independent, the test resulted in a p-value of $0.000112 < 0.05$ meaning that the distributions in the two groups are not identical. After further exploring the data, it was discovered that the reaction times distributions are different according to their PRECIPITATING EVENT. Figure 3.7 shows the distribution of reaction times according to the PRECIPITATING EVENT in only crash, only near-crash and crash and near-crash events together. As it can be seen, the distribution of reaction times to detect a decelerating lead vehicle is quite different from the distribution of reaction times to detect a stopped vehicle. The Wilcoxon–Mann–Whitney test further confirmed this finding resulting in a p-value of $2.2\times e^{-16} < 0.05$. Therefore, the decision was made to build a separate model for the events in which the driver reacted to a stopped vehicle rather than a decelerating car. The Wilcoxon–Mann–Whitney test resulted in a p-value of $0.07431 > 0.05$ concluding that the distributions of reaction times are identical in the retained groups of PRECIPITATING EVENT. The Shapiro Wilk test still shows that the reaction times are not normally distributed (p-value= 2.801e-10).

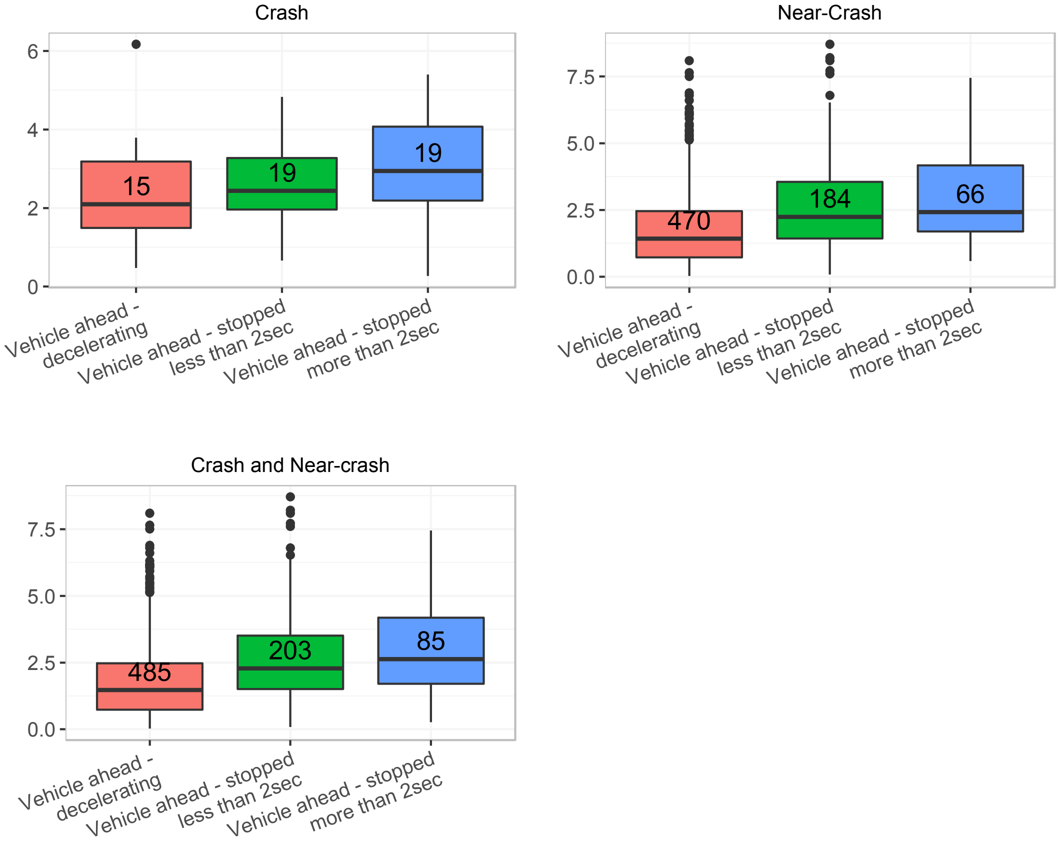


Figure 3.7 Box plot of reaction times according to the Precipitating Event.

Finally, we test whether the crash and near-crash reaction times have identical distributions in the retained data. Levene's test, a non-parametric two-sample test was used to assess the equality of variances of reaction times between crash and near-crash groups. The Levene's test resulted in a p-value of 0.1491, which is greater than the significance level 0.05. Therefore, it is concluded that the assumption of equal variances of reaction times cannot be rejected. The Wilcoxon–Mann–Whitney test resulted in a p-value of 0.1974 < 0.05 meaning that the distributions in the two groups did not differ significantly (Mann–Whitney U =4133, n(Crash) = 250, n(Near-Crash) = 38, P > 0.05 two-tailed).

The above pre-processing of reaction times resulted in a smaller sample of 288 rear-end events with 250 near-crashes and 38 crashes. In order to use the elastic net model with Gaussian errors, the data needs to follow a Normal distribution. To meet this requirement, we use the cubic root transformation to normalize the right-skewed reaction times. The Shapiro-Wilk normality test on the transformed data resulted in a p-value of 0.576 > 0.05 concluding that the normality test cannot be rejected.

Next, preprocessing the explanatory variables was necessary to make them ready for the elastic net regression model. The first step was to impute the missing values. We used the built-in Breiman's Random Forest algorithm to handle the missing values. It replaces a missing value with the median for numeric variables and with the most frequent level (breaking ties at random) for factor variables.

After the missing value imputation, a very comprehensive feature engineering was performed on the driver characteristics variables. In section 2.9, we explained the necessary feature engineering treatments to the categorical variables, namely 1- Build new binary variables from levels of a categorical variable with many levels; 2- Combine rare levels; and 3- Dummy coding. We used these methods on the driving context variables in the SHRP-2's *Event Detailed* data table such as roadway features, weather and driver dynamic behavior. In this chapter, the same treatments are applied to prepare the variables in the SHRP-2's *Driver* data tables (Table 3.4).

In addition to the above-mentioned treatments, additional steps were also required due to the special nature of driver variables. There were categorical predictors with only a single value in the whole dataset. These variables are called zero variance predictors and can easily be discarded. Not only they have no information but also some models such as linear regression would find them problematic and is likely to cause an error in the computations. The next and most time-consuming step was to treat the survey data. Data wrangling for example removing the punctuations, and mapping the codes to actual survey answers were necessary. The levels of most of these categorical variables were stored in alphabetic or numeric codes. We used the SHRP-2's online data dictionaries to map the codes to their original descriptions for each variable.

We used the Kruskal-Wallis (on the original reaction time) [86] and one-way ANOVA (on the transformed cubic root reaction times) for a preliminary feature screening and also for combining the rare levels. The Kruskal–Wallis statistical test is a non-parametric test that makes no assumptions about the distribution of the data (e.g., normality) and is an alternative to the independent group ANOVA, when the assumption of normality or equality of variance may not apply. The larger the test statistic H, the weaker the null hypothesis becomes, indicating that the feature under consideration has a high discriminating power. The p-value were used as a soft threshold to recode and merge the levels of variables. For example, most of the variables in the risk-perception and risk taking tables became significant or their p-values decreased (test statistics increased) by the following recoding:

Original level	**Recoded Level**
1	Low
2	
3	Medium
4	
5	
6	High
7	

After engineering the features, the number of numeric and categorical driver variables reduced to 113 (from 270) and more notably the total number of numeric and dummy variables reduced to 216 (from 1656). There are potential advantages to reducing the dimension of data prior to modeling. For one thing, fewer predictors mean decreased computational time and complexity. Furthermore, as it was mentioned a regression model would find un-informative variables problematic and its performance may deteriorate.

After building the feature vector, we are ready to fit the reaction time prediction model using SHRP-2 data. The 10-fold cross validation method was run 100 times to calculate the mean and standard deviation of measures of goodness in order to select the best model, i.e. the non-zero variables and their coefficients in the elastic net model. Two measures of goodness were considered:

1- Mean Squared Error (MSE): It measures the average of the squares of the prediction errors or deviations. The cross-validated MSE is a means of measuring the actual predictive capability of the selected regression model. The k-fold cross-validated MSE is calculated according to Equation (3.8):

$$CV\_MSE = \frac{1}{k}\sum_{j=1}^{k}\frac{\sum_{i=1}^{n_j}(\tau_i - \hat{\tau}_i)^2}{n_j} \tag{3.8}$$

2- Precision of predicting the reaction times greater than 2.5 seconds: It measures the fraction of instances that was truly predicted to be greater than 2.5 seconds. Equation (3.9) shows the calculation of this measure:

$$CV\_Precision = \frac{1}{k}\sum_{j=1}^{k}\frac{\sum_{i=1}^{n_j} I(\hat{\tau}_i > 2.5 \;\&\; \tau_i > 2.5)}{\sum_{i=1}^{n_j} I(\hat{\tau}_i > 2.5)} \tag{3.9}$$

> As it was mentioned in section 3.2, 2.5-second is the allowed reaction time for a typical driver in a brake-to-stop scenario. The main purpose to define this measure was to compare the performance of an FCW System which uses the proposed reaction time model to an FCW System using the standard 2.5-second for all drivers. A kinematics-based FCW with a fixed 2.5-second reaction time is not going to be effective for drivers and driving conditions where the actual reaction time of drivers would be longer than 2.5 seconds. Instead, using a reaction time prediction model with high precision in predicting longer reaction times renders a potential safety benefit in that it will give an additional time to slower-moving drivers or in more complex collision scenarios.

Table 3.5 shows the MSE and Precision values of the prediction models. These values are the mean of 100 replications of running the 10-fold cross validation. As it can be seen in Table 3.5, we started with the null model, i.e. the model with only the intercept, and then added the groups of potential safety factors hierarchically. The main objective of this study is to predict the value of reaction time in real driving scenarios. Therefore, we only compare the mean squared prediction errors through cross validation. At each level of the hierarchy, the percentage changes in the MSE and Precision are reported as well. The largest improvement, in terms of MSE decrease and precision increase, was achieved by introducing the driver characteristics into the previously built model. By average, the MSE decreased 24% and the Precision was increased 27%. Once again, by using the elastic net regularized regression model, we have benefited from its built-in mechanism for variable selection, ability for bias/variance trade-off, ability to deal with highly correlated variables and high dimensional problems, and handling of both numerical and categorical input variables. Overall, the results show that including the driving context variables together with the driver characteristics in the full model decreased the MSE by 36% and increased the precision by 57%.

Table 3.5 MSE and Precision of the reaction time prediction models.

	MSE	**Precision**
Nominal value 2.5 sec	0.0802	-
Null model	0.0800	47%
Percentage Change	-0.2%	-
Roadway and Traffic	0.0740	56%
Percentage Change	-8%	20%
Roadway and Traffic & Environment	0.0680	57%
Percentage Change	-8%	1%
Driver Dynamic Behavior & Roadway and Traffic & Environment	0.0665	58%
Percentage Change	-2%	2%
Full Model (All the above + Driver characteristics)	0.0508	73%
Percentage Change	-24%	27%

Based on the numerical results and in order to not compromise safety for comfort, a simple rule-based algorithm for a conservative FCW system is proposed that uses the *reaction time* prediction model. Figure 3.8 shows the steps of the algorithm which are as follows:

- Calculate the driver's reaction time according to his/her characteristics and real-time driving context using the elastic net model.
- If the predicted reaction time is greater than 2.5 seconds, set the reaction time value to its elastic net estimate, otherwise use the standard 2.5 seconds.
- Pass the reaction time value to the kinematics model to calculate the critical distance, *d*.

- Compare the critical distance, *d*, to the thresholds in Equation (3.2), and decide whether to issue a Forward Collision Warning.

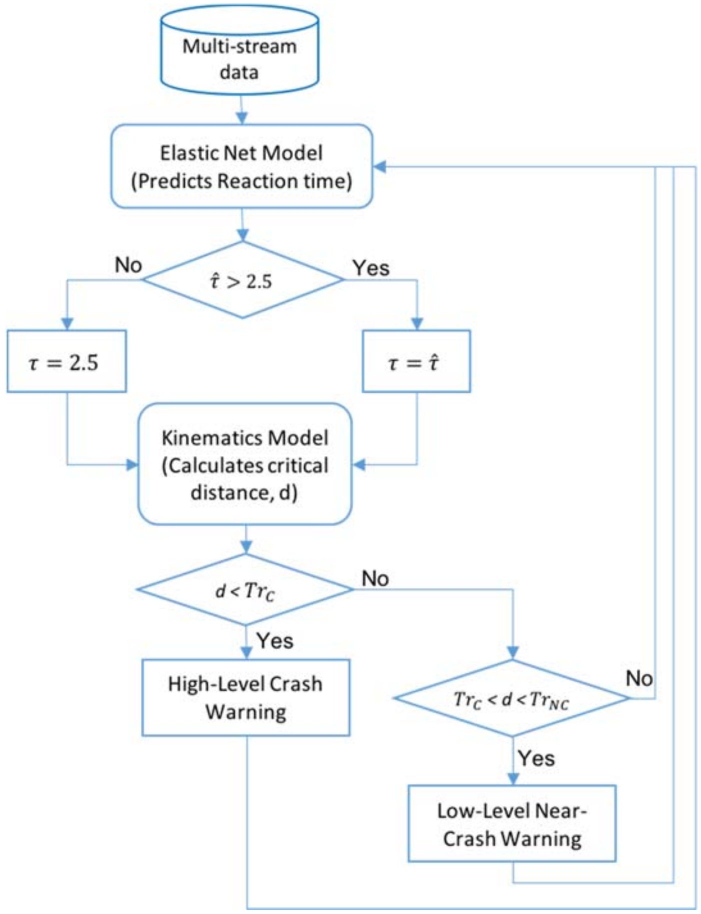


Figure 3.8 Proposed hybrid algorithm for a conservative Forward Collision Warning System.

The success of the proposed FCW system directly depends on the amount and quality of the trained data to build the prediction model. With our limited data of 776 crash and near-crash events, the highest achieved precision in predicting reaction times greater than 2.5 seconds were 73%. It is expected that with more high quality data, the precision of the predictions increases and therefore the rate of false alarms will decrease.

Next, the process of selecting the best model, i.e. the elastic net non-zero variables and their coefficients, for the full model is presented. Figure 3.9 shows the coefficient paths versus the logarithm of λ for the full model. The solutions were computed at 100 values of λ, uniformly spaced on the log scale. The values of λ are decreasing from right to left, i.e. the far right λ corresponds to the largest penalty. In another word, the null model corresponds to the λ at the far right and the full model to the λ at the far left of the plot. As λ decreases more new variables enter the model. As it can be seen, due to the correlation among predictors, a variable that has entered the model may go out at a later phase. The red vertical line in Figure 3.9 crosses the values of coefficient at λ_{min}, i.e. the value of λ for which the previously estimated cross-validated error was minimum. Figure 3.9 shows the 21 selected variables by the best model and their estimated regression coefficients.

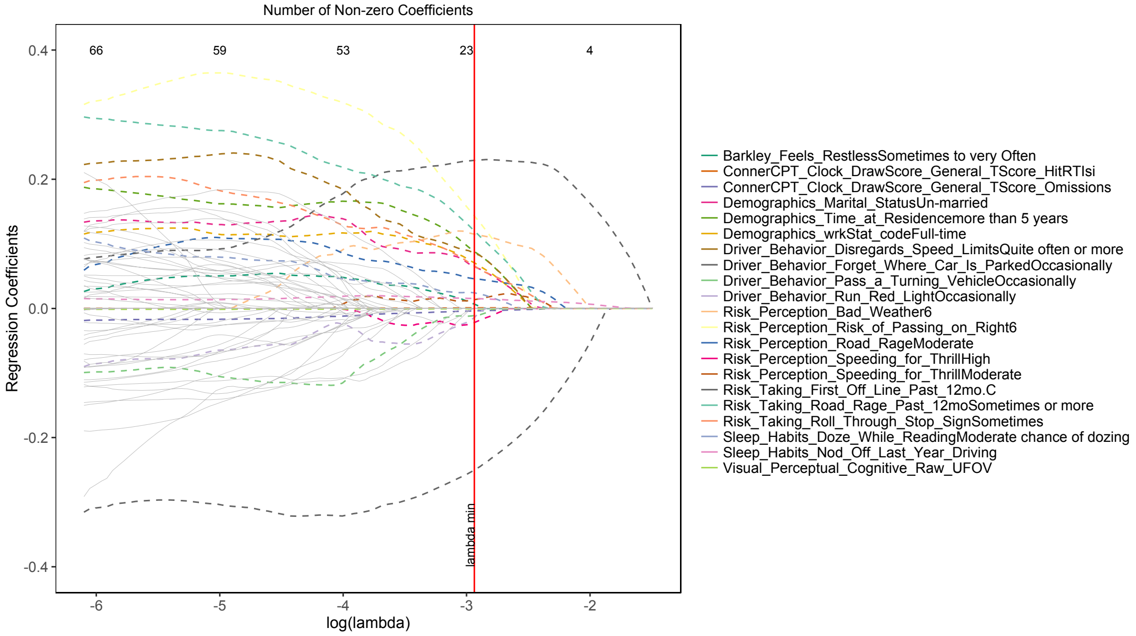


Figure 3.9 Coefficient paths for the lasso model for the reaction time in rear end crash and near-crashes.

Table 3.6 Estimated regression coefficients of the best model with the minimum cross-validated error.

Predictor	Coefficient Estimate
• Demographics_wrkStat_code: Full-time	0.0705
• Demographics_Time_at_Residence: more than 5 years	0.08804
• Sleep_Habits_Doze_While_Reading: Moderate chance of dozing	0.02419
• Demographics_Marital_Status: Un-married	0.06462
• Barkley_Feels_Restless: Sometimes to very Often	0.00645
• Driver_Behavior_Pass_a_Turning_Vehicle: Occasionally	-0.01118
• Driver_Behavior_Forget_Where_Car_Is_Parked: Occasionally	-0.25013
• Driver_Behavior_Run_Red_Light: Occasionally	-0.01408
• Driver_Behavior_Disregards_Speed_Limits: Quite often or more	0.0858
• Risk_Taking_First_Off_Line_Past_12month: Sometime or more	0.22918
• Risk_Perception_Road_Rage: Moderate	0.04577
• Risk_Taking_Roll_Through_Stop_Sign: Sometimes	0.06574
• Risk_Perception_Speeding_for_Thrill: Moderate	0.01365
• Risk_Perception_Speeding_for_Thrill: High	-0.0219
• Risk_Perception_Bad_Weather: level 6	0.11493
• Risk_Perception_Risk_of_Passing_on_Right: level 6	0.14325
• Risk_Taking_Road_Rage_Past_12month: Sometimes or more	0.12018
• Sleep_Habits_Nod_Off_Last_Year_Driving (numeric)	0.01525
• Visual_Perceptual_Cognitive_Raw_UFOV (numeric)	-0.00049
• ConnerCPT_Clock_DrawScore_General_TScore_Omissions (numeric)	-0.0039
• ConnerCPT_Clock_DrawScore_General_TScore_HitRTIsi (numeric)	0.00087

Furthermore, Figure 3.10 to Figure 3.13 show the boxplots to better visualize the effects of each categorical variable and the impact direction (negative or positive signs of estimated coefficients) of each variable’s levels on reaction time. For example, Figure 3.10, top-right panel shows that there is a positive trend between the driver’s sleeping habit of dozing while reading: the reaction time increases as the chances of dozing increases.

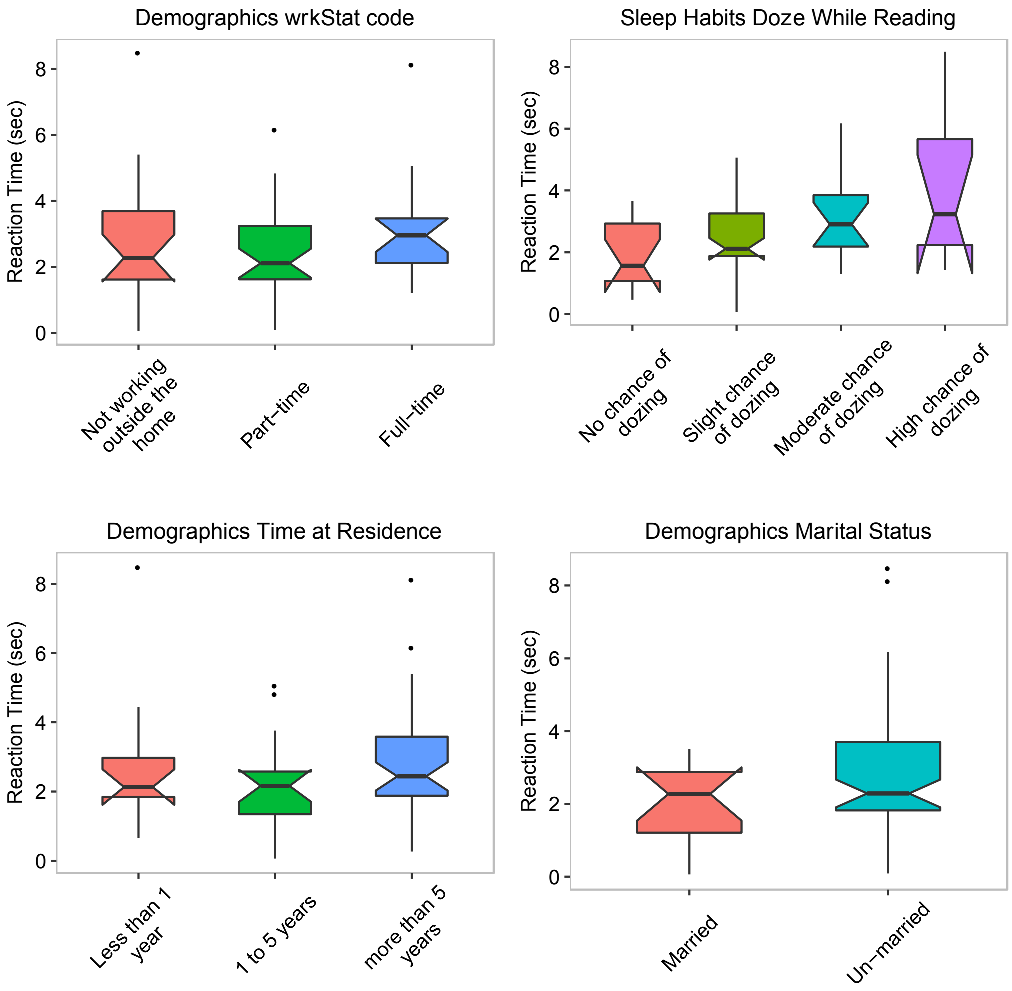


Figure 3.10 The relationship between reaction time and four categorical driver characteristics of: work status (top-left), chance of dozing while driving (top-right), time lived at current residence (bottom-left) and marital status (bottom-right)

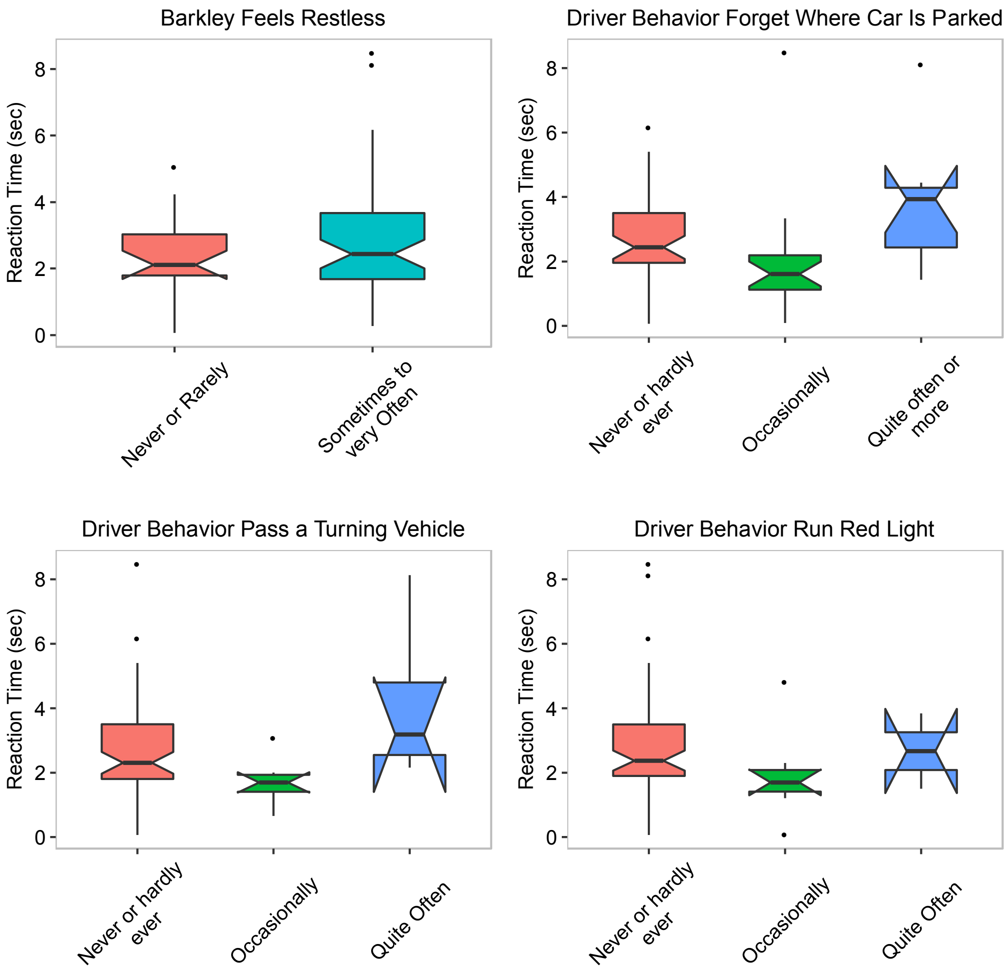


Figure 3.11 The relationship between reaction time and four categorical driver characteristics of: Results of Barkley screening test of feeling restless (top-left), How often the driver forgets where the car is parked (top-right), How often the driver passes a turning car (bottom-left) and How often the driver runs a red light (bottom-right)

An interesting result from Figure 3.11 is that driver behaviors of $Forget\ Where\ Car\ Is\ Parked$, $Pass\ a\ Turning\ Vehicle$ and $Run\ the\ Red\ Light$ have similar effects on reaction time that is drivers who are $Occatsionally$ involved in these behavior react faster than other drivers but committing these behaviors quite often or more increases the reaction time of drivers.

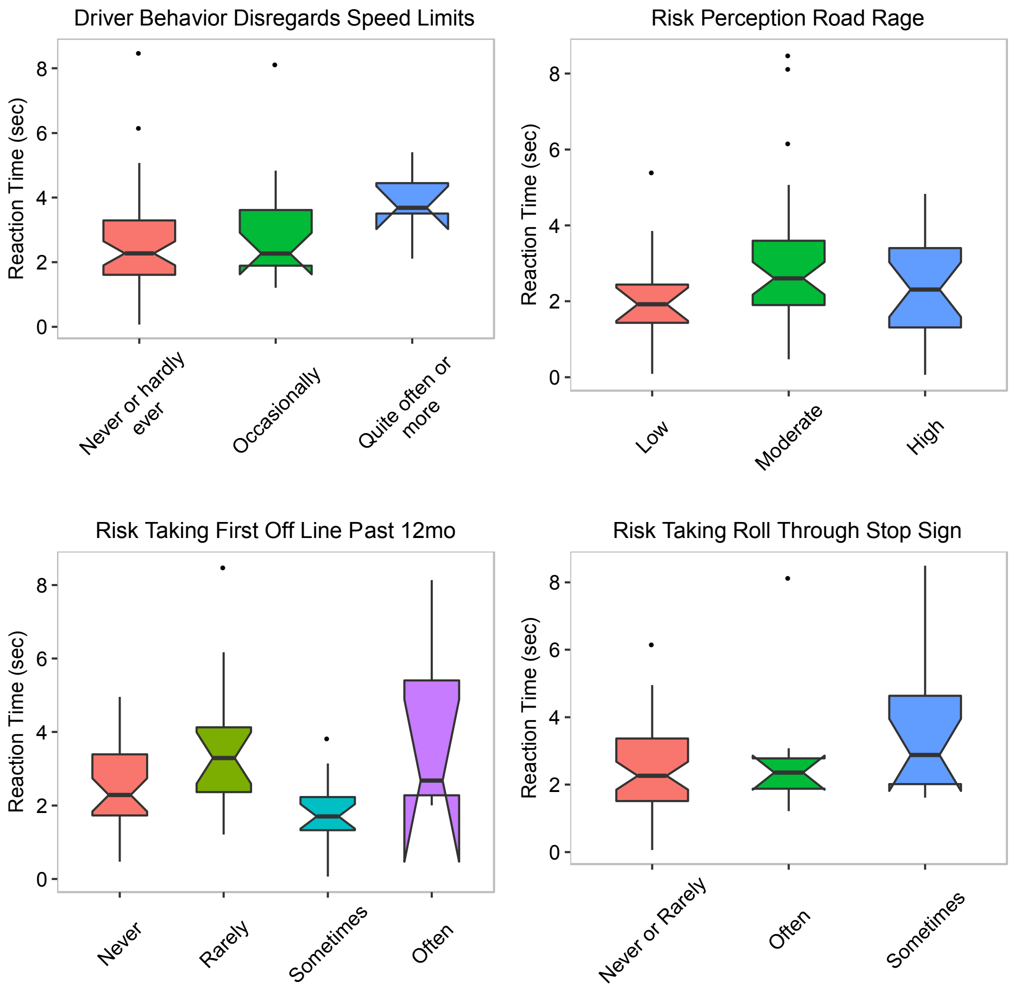


Figure 3.12 The relationship between reaction time and four categorical driver characteristics of: how often the driver disregards the speed limit (top-left), risk perception of road rage (top-right), risk taking to be the first off the line during past 12 months (bottom-left) and risk taking to roll through stop sign (bottom-right)

Another interesting observation from Figure 3.12 and Figure 3.13 is that those drivers whose risk perception of risky behavior are moderate, they have longer reaction times comparing to other drivers, and as expected those who have high perception of risk have reacted faster to road incidents.

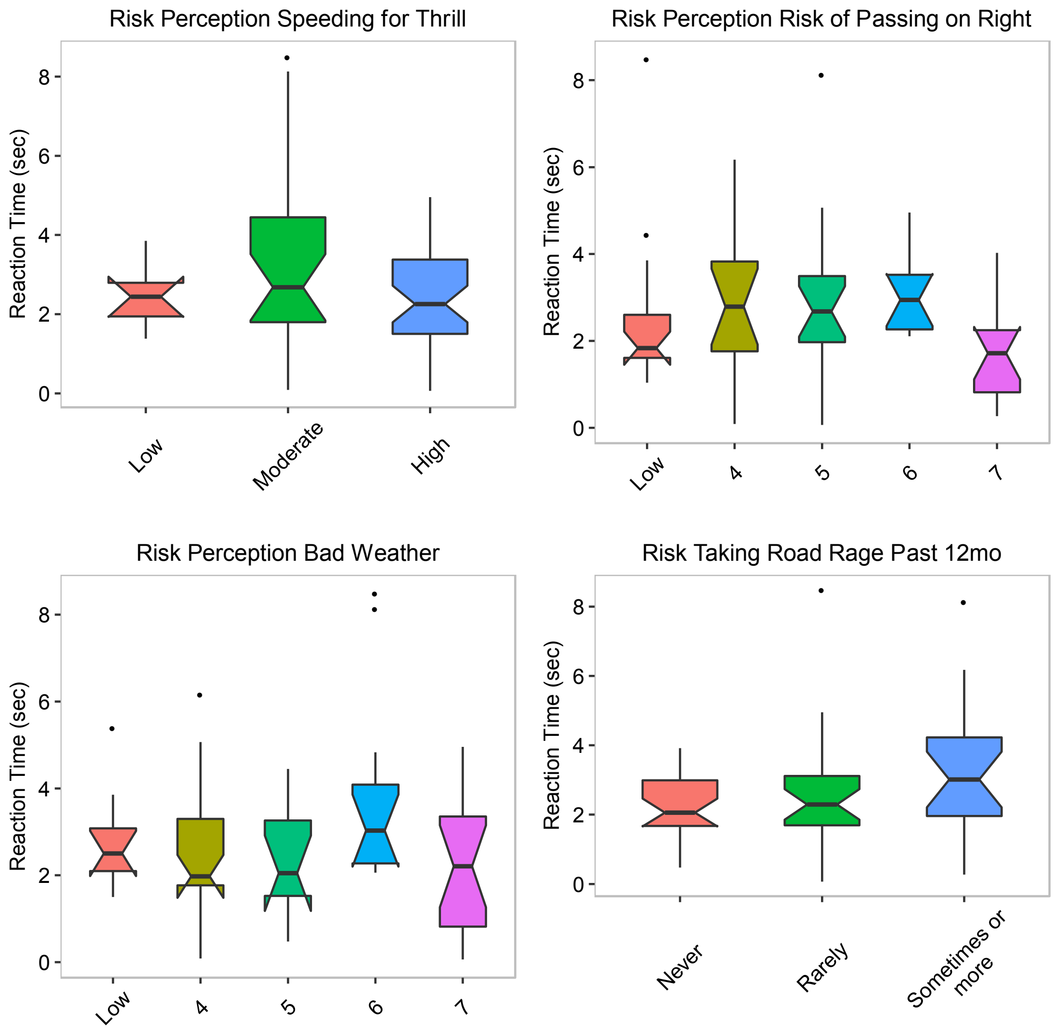


Figure 3.13 The relationship between reaction time and four categorical driver characteristics of: risk perception of speeding for thrill (top-left), risk perception of passing or right (top-right), risk perception of bad weather (bottom-left) and risk taking of road rage during the past 12 months (bottom-right).

Finally, from Figure 3.12 and Figure 3.13, when risk taking behavior of road rage and roll through stop sign increases the reaction time of the driver increases. Similar positive relationships were observed in other risk taking behaviors not reported here. The data shows that in general getting involved in risky behavior would increase the average reaction time of drivers.

3.5 Conclusion

In this chapter, we presented a hybrid physics/data-driven approach to design an Advanced Driver Assistance System in particular a Forward Collision Warning System. We focused on a simple one dimensional brake-to-stop physics model for rear-end collision scenarios and estimated the driver-intrinsic parameter of this model by using the SHRP-2 NDS rich data on driver characteristics. The data analysis showed that contrary to the common practice of using an average reaction time value in accident reconstruction and driver assistance systems, the driver's brake-to-stop reaction time in real-world forward collision scenarios spanned from .065 to 8 seconds. We used a regularized regression model to estimate the reaction time of a driver according to his/her intrinsic characteristics and the driving context. The results showed that introducing the driver characteristics decreased the mean squared prediction error of the reaction time prediction model by 24%. More importantly, the precision of the model in truly predict reaction time greater than 2.5 second reached the level of 74%. In another words, using the elastic net estimates instead of the prefixed 2.5 second on average gives the system 1.34 additional seconds to warn those slower-reacting drivers. The explained variation by driver's intrinsic characteristics and driving context supports the necessity for developing personalized Advanced Driver Assistance Systems to enhance the performance and increase their acceptance by users. Using this estimated reaction times instead of the common fixed values for all drivers, can enhance the performance of safety warning systems taking into account the differences in human ability to react to road incidents.

Overall, the most significant limitation to this study was the sample size and representation of different drivers and driving conditions. Only one-third of the full SHRP-2 NDS data was made accessible at the time of the data request. It is expected that by including a larger dataset of naturalistic driving data the performance of the reaction time prediction would increase. This work can be extended in two different ways: 1- to use more complex physics model such as 2 dimensional models which considers both braking and steering as evasive maneuvers and 2- to use vehicle and roadway characteristics data to model the vehicle response time which is out of the control of the driver.

Disclaimer

The findings and conclusions of this chapter are those of the author and do not necessarily represent the views of the VTTI, SHRP 2, the Transportation Research Board, or the National Academies.

4 A WHOLE-SYSTEM APPROACH TO IDENTIFY THE SOURCES OF VARIATION IN PATIENT FLOW

4.1 Introduction

During their hospital stay, patients may experience redundant steps and procedures that may lead to unnecessary excessive expenses, lower Quality of Care (QoC) and customer dissatisfaction. The excessive costs are often covered by hospitals or paid by individual patients, since insurance companies have standard payment plans ranging from the infamous charge master or fee-for-service (FFS) price list to bundled payment systems such as diagnosis-related groups (DRGs) with various forms of "discounts off charges" and "per diems" somewhere in between. Regardless of who pays for these excessive and unnecessary expenses, the adverse societal impacts and negative business consequences are immense. In this chapter, we focus on the patient flow process in a hospital with DRG based payment system for its inpatient claims.

Renewed focus on quality measurement and improvement and on medical-error reduction has heightened interest in paying for performance, rather than just reimbursing providers for services rendered. Private Pay for Performance (P4P) programs for hospitals usually pays bonuses as an incentive above the agreed-upon reimbursement rate. A more rational reimbursement system, which rewards quality of care rather than simply doing more to patients, is the short-term goal of paying for performance. The longer-term goal is also to make the health care system more efficient. It has become clear that under existing reimbursement structures, current market forces are insufficient to ensure either higher-quality or more-cost-effective care [5]. P4P programs can be seen as additional incentives for hospitals to seek to improve their patient flow processes, which can be attained through our variation reduction framework.

Since 1983, under Health Care Financing Administration (HCFA)'s system, generally referred to as the Prospective Payment System (PPS), each hospital inpatient is classified into one of around 500 Diagnosis-Related Groups (DRGs), and the hospital is paid the amount that HCFA has assigned to each DRG. Thus, hospitals will be paid the same amount for patients within a particular DRG. One limitation to this methodology is that individual DRG categories often combine subgroups of patients with predictably different expected resource costs. HCFA has repeatedly improved the DRG definitions since 1984; in fact a new DRG system, called Medicare Severity DRGs (MS-DRGs), was adopted in October 1, 2007, which replaced 538 DRG system with 745 new MS-DRGs [87]. This enhancement, while necessary, does not fully account for differences in illness severity associated with substantial disparities in providers' costs.

The fact is that only a part of these disparities is attributed to the uncontrollable factors in patient profile including his/her demographics, medical history, medication, physical exams, and so on. There are also controllable factors that influence patient's experience from hospital admission to discharge. These include, but not be limited to, the order of treatments patient receives, medical procedures, current medications, received resources including physicians, nurses, technicians, transporters, and administrative work. These sources of variability could severely impact patient safety, QoC, professional satisfaction, and hospital revenue. The potential reduction in costs and increase in QoC and patient safety and satisfaction will be too rewarding to ignore. All these tools become handier especially when the regular normal operation of hospital is affected by an external incident varying from highway crashes to earthquakes and terrorist attacks. It is in such situations that having a managed patient flow can be of great value to the hospital management to increase patient care and lower the number of fatalities.

This chapter is organized as follows. Section 2 presents the literature survey. In section 3 we present the formulation of our problem. The data to test our procedure and the results of applying our methodology are discussed in Section 4. Conclusions are presented in the final section.

4.2 Literature Survey

A number of researchers have used queueing models to study various aspects of the patient flow process. McClean et al. (2005) use phase-type distributions to carry out model-based clustering of patients using the time spent by the patients in hospital [88]. They cluster patients into classes on the basis of the number of phases involved. Cadez et al. (2003) presented a new methodology for exploring and analyzing navigation patterns on a web site [89]. They partition site users into clusters such that users with similar navigation paths through the site are placed into the same cluster. Their proposed method clusters users by learning a mixture of first-order Markov models using the Expectation-Maximization algorithm. In this chapter, we base our methodology on their results.

4.3 Technical Approach

Patient flow is not a single datum but a pattern or a sequence of steps. Unlike classical statistics where singular or array of data is used, we need to work with flow patterns and ordered data. In this thesis, we use a mixture of first-order Markov models to describe patient flow. Each patient is admitted to an inpatient floor with an initial diagnosis determined by the admitting physician. After patient is discharged, her chart is reviewed by coders and a DRG is assigned primarily based on the definitive final diagnosis and other diagnoses together with treatments, resources and procedures utilized towards treating patient's condition during her stay. For each DRG certain level of resources (treatments, diagnostic tests, procedures, etc.) are assigned and required. From admission to discharge, a patient goes through a sequence of steps both in terms of her condition and the utilized resources, treatments and procedures. Throughout this chapter we will refer to this sequence of steps as *patient flow* vector and denote it by $\vec{S}^m$, defined as follows:

$$\vec{S}^m = \left[S_1^m, S_2^m, \cdots, S_t^m, \cdots, S_{n_m}^m\right]', m = 1,2,\cdots,M \; and \; t = 1,2,\cdots,n_m \tag{4.1}$$

where $\vec{S}^m$ is an n_m×1 ordered vector with the t^th^ element, S_t^m, as the state of patient *m* at step t (t=1, 2, …, n_m). n_m is the length of the m^th^ sequence and can be different for each patient (n_m=1,2, …, N). S_t^m takes on values (s_t^m) from among K possible patient states ($s_t^m \in \{1,2,\cdots,K\}$). Therefore the sequence $\left[S_1^m, S_2^m, \cdots, S_t^m, \cdots, S_{n_m}^m\right]'$ indicates that patient m first was at state s_1^m, then s_2^m, and so on. In our model, the last state is always K, which denotes the *discharged* state. The nature and definition of these states can be different according to the level of granularity of the problem, i.e. the level of detail at which patient flow is observed. They can be as aggregated as generic states that any patient may go through during a hospital stay (like the admission, inpatient floor stay, and discharge), or they can be very detailed including all the steps in each of the above mentioned high level states.

As we mentioned earlier there can be several sources of variability that are intrinsic to all healthcare delivery systems. We have categorized these sources into three groups:

(i) Unique characteristics of each patient (patient profile), including demographics, medical history and other health conditions upon admission. $\vec{X}^m$ defines these characteristics:

$$\vec{X}^m = \left[X_1^m, X_2^m, \cdots, X_p^m, \cdots, X_P^m\right]', m = 1,2,\cdots,M \text{ } and \text{ } p = 1,2,\cdots,P \tag{4.2}$$

where $\vec{X}^m$ is a P×1 vector whose pth element, X_p^m, denotes the pth explanatory variable quantifying a characteristic of patient m.

(ii) Hospital resources, including medical and non-medical (overhead) staff {direct (nurse, tech, doctor) and indirect (unit secretary, housekeeping) labor and overhead labor}, major equipment, units and their functionalities (hospital factor). We denote these characteristics by $\vec{Z}^m$:

$$\vec{Z}^m = \left[Z_1^m, Z_2^m, \cdots, Z_q^m, \cdots, Z_Q^m\right]', m = 1,2,\cdots,M \text{ } and \, q = 1,2,\cdots,Q \tag{4.3}$$

where $\vec{Z}^m$ is a Q×1 vector whose qth element, Z_q^m, represents the qth explanatory variable quantifying the l^{qh} hospital resource on patient m. Depending on the attribute which they quantify, X_p^m and Z_q^m can each be mixture of continuous or categorical explanatory variables.

The covariates of both patient profile and hospital resources can remain fixed during a patient stay, for example the patient's gender; age; or medical history, or can be state-dependent, which is often the case for hospital resources, for example, the variables defined on imaging tests, like an echo test's turnaround time, accept values only if the patient receives the test and after his leaving the corresponding test unit (state).

(iii) Random noise denoted by ε_m are assumed to be i.i.d. random variables with mean zero and standard deviation σ_m. There are always un-assignable causes, which are usually grouped under random noise. Since random noise is statistically un-controllable, it is imperative to reduce its effect as much as possible. Any significant reduction in un-controllable variations will increase "process capability" and improve the process, which will in turn lead to significant cost reductions.

Furthermore, we assume that reentry of patient *m* to the hospital is a new admission with an updated $\vec{X}_i$ vector due to the new set of treatments that may be required. Then a historical data set of size *M*, containing *M* vectors of $\vec{S}$, $\vec{X}$, and $\vec{Z}$ defines patient pathways, patient characteristics and hospital resources of *M* observed patients categorized under a specific DRG during a given time interval. We intend to find clusters of similar patients in term of their pathways, i.e. the vectors of $\vec{S}$'s. We assume that the number of clusters is known. We also intend to link $\vec{X}$, and $\vec{Z}$ to $\vec{S}$ in order to determine significant factors that lead to clusters within a DRG. Finally, by controlling the important attributes and reducing their variation we expect to see a reduction in the variations inherent in the patient flow process. Sections 3.1 to 3.5 explain the steps of our algorithms in details.

The data collection was performed in a 500-bed community hospital with a level II trauma center in New Jersey. This general hospital provides a wide range of services with a total number of 383 DRGs in 2012. Having met with groups of hospital experts, including physicians, head nurses, directors of financial department, and case management, we decided to focus on chest pain DRG since it was the second most frequent DRG with a very low contribution margin during 2012. Different pieces of patient information, both medical and personal, had to be mined and mended together to obtain $\vec{X}$, $\vec{Z}$ and $\vec{S}$ for each patient as explained previously in the technical approach. The data had been scattered in different databases either with some time lags or near real time. Several commercial software packages were used to obtain patient flow (the locations that the patient visited chronologically). Another software NTT contains both medical (physicians, nurses, tests with their time stamps), and personal information. PACS and Xcelera applications were also used for more detailed information, especially the exact time and duration, of radiology, CT scan and other imaging test results. The next step was to prepare data for analysis, including handling of missing data. For this purpose, we used the fully conditional specification (FCS) method, also known as multivariate imputation by chained equations (MICE) [56]. The joint distribution of covariates $\vec{X}$ and $\vec{Z}$ is not known involving both categorical and continuous variable, thus FCS is a better suited alternative to Joint Modeling method which requires a known multivariate distribution for data.

Expert opinion was sought to extract, filter, and transform data into meaningful quantifiable variables that we further fed into our statistical engine. For this purpose, we built a multidisciplinary team whose members brought different perspectives and knowledge about the problem [90]. The core team included physicians, head nurses, directors of financial department, and case management who had direct contact with the process. The team was brought together in brainstorming sessions for two important tasks:

1. Define the state space of patient flow vector ($\vec{S}$): Medical judgment was used to construct states, which both exhibit the necessary independence and make sense in

terms of the delivery of care. A state space was constructed in a manner that resulted in state definitions, which are mutually exclusive and collectively exhaustive [91]. This was essential to ensuring that Markov modeling of patient flow is valid.

2. Quantify vectors of patient profile and hospital resources ($\vec{X}$, and $\vec{Z}$): To perform this task, we identified as many potential variables as possible according to the historical data. We used fishbone diagram, also known as cause-and-effect diagram, to identify the potential causes of variation [92]. Causes were grouped into major categories to identify sources of variation.

We translated the potential causes into quantifiable random variables of either continuous or categorical types. For example, patient's gender was defined as a Bernoulli variable with classes of "male" and "female". The complete list of covariates can be found in Appendix 1.

For sequence clustering, we apply a mixture of first-order Markov models to model patient flow sequences. We assume that the flow of each patient in the data set, $\vec{s}^m$, is generated independently (the traditional i.i.d. assumption). Statisticians refer to such a model as a mixture model with R components (R is the number of clusters). We apply Expected Maximization (EM) method to train our model. Once the model is trained, we can use it to assign each patient to a cluster or fractionally to the set of clusters. A mixture model for $\vec{S}$ with R components has the form:

$$p(\vec{S} \mid \theta) = \sum_{r=1}^{R} p(c_r \mid \theta).p_r(\vec{S} \mid c_r, \theta) \tag{4.4}$$

where c_r is the cluster assignment for a given patient, $p(c_r \mid \theta)$ is the marginal probability of the r^th cluster ($\sum_r p(c_r \mid \theta) = 1$) and $p_r(\vec{S} \mid c_r, \theta)$ is the statistical model describing the distribution for the variables for patients in the r^th cluster, and θ denotes the parameters of the model. We further assume that each model component is a first-order Markov model capturing the sequence of steps taken by a patient to some degree. Then, the EM method is used to train the parameters of the mixture model with known number of components R, given training data $d_{train} = \{\vec{S}^1, \vec{S}^2, \cdots, \vec{S}^M\}$ such that the following equation holds:

$$\theta^{ML} = \arg\max_{\theta} p(d_{train} \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{M} p(\vec{S}_i \mid \theta) \tag{4.5}$$

where θ^{ML} are the maximum likelihood or ML estimates of the model parameters.

In this study, we have used Microsoft Sequence Clustering algorithm (SQL Server Analysis Services or SSAS) to carry out the sequence analysis. Microsoft SQL Server provides us with the membership assignment of each patient. Therefore, having a training data set of size M, we can run the sequence-clustering algorithm and obtain the *vector of class memberships*, denoted by $\vec{Y}$, as follows:

$$\vec{Y} = [Y_1, Y_2, \cdots, Y_M]' \tag{4.6}$$

where Y_i is the class membership of patient i, and can accept values of 1, 2, …, R. Later, we will feed this vector into the Variable Selection module.

In this step, we will use a well-known classifier, namely random forest, to identify the significant variables which affect the patient flow sequences. Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees [6]. The data-set used for training comes in records of the form ($\vec{D},\vec{Y}$) for each data-point, where $\vec{Q}$ denotes a vector of observed characteristics (also referred as features or factors) and $\vec{Y}$ denotes a group label (also called target variable). In our application, $\vec{D}$ is a (p+q)×1 vector of $\begin{bmatrix} \vec{X}_{p\times 1} \\ \vec{Z}_{q\times 1} \end{bmatrix}$ which contains the information of patient profile and hospital resources, i.e. the explanatory variables, and $\vec{Y}$ is the *vector of class memberships*, i.e., the output of the sequence clustering algorithm.

In order to perform the classification task we will use the *randomForest* package available in R software [57]. The input to the software will be feature vector $\vec{D}_{(p+q)\times 1} = \begin{bmatrix} \vec{X}_{p\times 1} \\ \vec{Z}_{q\times 1} \end{bmatrix}$ and vector of *class memberships* $\vec{Y}$.

Random forests can be used to rank the importance of variables. There are two criteria based on which the Breiman's random forest calculates the importance of variables: *Gini importance* which calculates the mean Gini gain produced by $\vec{D}^m$ over all trees, and *permutation accuracy importance* which is the mean decrease in classification accuracy after permuting $\vec{D}^m$ over all trees. The variable importance plot gives a relative ranking of significant features, and absolute values of the importance scores should not be interpreted or compared over different studies. We consider, the first B variables as the most important variables where B < P+Q. We will refer to the vector of important variables as $\vec{D}'_{B\times 1} = \begin{bmatrix} \vec{X}'_{p'\times 1} \\ \vec{Z}'_{q'\times 1} \end{bmatrix}$, and define $\vec{X}'^{m}$, and $\vec{Z}'_i$ as follows:

$$\vec{X}'^{m} = \left[X_1^m, X_2^m, \cdots, X_{p'}^m, \cdots, X_{P'}^m\right], m = 1,2,\cdots,M \textit{ and } p' = 1,2,\cdots,P' \tag{4.7}$$

$$\vec{Z}'^{m} = \left[Z_{1}^{m}, Z_{2}^{m}, \cdots, Z_{q'}^{m}, \cdots, Z_{Q'}^{m}\right], m = 1,2,\cdots, M \text{ } and \text{ } q' = 1,2,\cdots, Q' \tag{4.8}$$

where p ≤ p’, and q ≤ q’.

Monitoring and controlling of important variables is the last step in our model. In the previous steps we established a relationship between patient flow sequences and process attributes, and identified those attributes that affect the patient flow process significantly. In this step we investigate how and why these attributes affected patient flow. For this purpose, questions must be asked to find the assignable causes of variations and then a proper corrective action must be taken to eliminate them. To maintain the gained improvement and be able to detect future assignable variations, advanced statistical tools such as single-variable or multivariate control charts can be used. Using control charts is an ongoing activity over time to bring continuous improvements to the process.

4.4 Numerical Experimentation

In this section, we illustrate the performance of our algorithm using both real and simulated data. The data simulation will closely mimic the true real life process. As explained in section , we collected $\vec{X}$, $\vec{Z}$ and $\vec{S}$ for 87 patients with primary DRG of chest pain admitted during 2012 and 2013. All these patients initially came to the emergency room ambulatory, or dropped by a friend/relative or by ambulance. They then went through different states according to their personal and medical needs. Fifteen distinct states were observed in our sample data which are shown in Table 4.1. We discarded patients who left against medical advice. The discharge statuses of all the 87 patients were discharged to home or another hospital unit. Neither of these patients was readmitted for the same cause within 30 days of their discharges.

Table 4.1 Distinct States of Patient Flow.

Code	Description
1	ED
2	GS
3	Telemetry Unit
4	Cardiology
5	Nuclear Medicine
6	Radiology
7	Ultrasound
8	Vascular Lab
9	Pulmonary Function Lab
10	Endoscopy
11	MRI
12	Dialysis
13	CT-Scan
14	Operation Room
15	Discharge

According to our collected sample data, patient flow sequences of chest pain DRG can at most have fifteen steps (N=15). Thirty-four factors have been identified as the potential causes of variation, nine of which are patient profile-related attributes (P=9), and the rest are hospital resources (Q=25). The definitions of these variables can be found in Appendix 2. Since our sample size is relatively small (87), comparing to the total number of variables and the covariates' levels, we used this data to simulate additional cases for our model verification and validation. To generate a simulated data set, the initial step is to use our real sample data, $d^{(\mathrm{Re}\,al)} = \{\vec{S}_{87\times n_m}, \vec{X}_{87\times 9}, \vec{Z}_{87\times 25}\}$, to generate the covariates, $\vec{X}, \vec{Z}$, from their empirical distributions. In our case study, the types of all the covariates are categorical although in general the model can accept both types of continuous and categorical variables. If Xp follows a categorical distribution, also called generalized Bernoulli distribution, then its probability mass function *f* is:

$$f(x_p) = \prod_{i=1}^{K_p} \pi_i^{I(x=i)} \qquad (4.9)$$

where I(.) is the indictor function, and πi is p(xp=i). We have used the MLE parameter estimator, i.e. the empirical fraction $\hat{\pi}_i = n_i / n$, to estimate the parameters of the distribution, namely $\hat{\pi}_p = (\hat{\pi}_1, \hat{\pi}_2, \cdots, \hat{\pi}_{k_p})$. Kp is the number of categories of the pth covariate. Furthermore, in order for the simulated covariates of virtual patients to have physiologically reasonable covariate distributions resembling the real patients, we used the continuous covariate simulation method proposed in reference [93]. In this proposed Continuous method, the parameters of a single multivariate normal distribution (MVND) are estimated by treating all categorical covariates as if they are continuous values, a procedure seen commonly in statistical simulation. In order to constrain all covariates to be positive, typically a lognormal multivariate distribution is assumed. Thus, the MVND variance–covariance matrix is defined in terms of the logarithms of the covariate values. First, all the categorical variables must be coded to possess positive values. Complete patient covariate vectors are then sampled from a single MVND; because the sampled values are logarithmic, each component of the vector is then exponentiated to obtain the true covariate values [93]. These continuous values are then mapped to discrete categorical values, based on a continuous critical value (CrV), according to the following equation:

$$CrV(\mu, \sigma, p_i) = e^{\mu + \sigma . NORMINV(\sum_{j=1}^{i} \pi_j)} \tag{4.10}$$

where μ = mean(ln(X_p)), σ = SD(ln(X_p)), and NORMINV is the invers of the standard normal distribution. We then used the Kolmogorov-Smirnov test to test if the distribution of the simulated covariates is the same as the distribution of the covariates of real patients. We have also conducted the Jennrich test to test if the covariance matrices of the simulated data are statistically equal to the covariance matrix of the real data. Next, in order to generate the patient pathways, we need to calculate the initial values of transition probabilities, i.e. P_{ij}'s. We assumed that each row of the transition probability matrix of patient pathways follows a mixed distribution as follows:

$$P_{ij} = \Pr(S_{t+1}^m = j \mid S_t^m = i) = \begin{cases} f(\vec{x}_{S_0:S_t}^m, \vec{z}_{S_0:S_t}^m); & j \in Set1 = \{1,2,\cdots,K'\} \\ \dfrac{\#\{S_t^m = i, S_{t+1}^m = j\}}{\sum_k \#\{S_t^m = i, S_{t+1}^m = k\}}; & j \in Set2 = \{1,2,\cdots,K\} - Set1 \end{cases} \tag{4.11}$$

$$\sum_{j \in Set1} P_{ij} + \sum_{j \in Set2} P_{ij} \neq 1 \tag{4.12}$$

$$C.\sum_{j \in Set1} P_{ij} + \sum_{j \in Set2} P_{ij} = 1 \tag{4.13}$$

$$C = \frac{1 - \sum_{j \in Set2} P_{ij}}{\sum_{j \in Set1} P_{ij}} = \frac{1 - \sum_{j \in Set2} P_{ij}}{1} \tag{4.14}$$

$$Adj.P_{ij} = C.P_{ij}, j \in Set1 \tag{4.15}$$

where $\Pr(S_{t+1}^m = j \mid S_t^m = i)$ is the probability that the m^{th} patient is in state j at step t+1, given that he was in state i at step t. This definition comes from our assumption that the patient transfer between states follows a first order Markov model. *Set 1* in Equation (4.11) is the set of states, j's, to which there have been enough number of transitions from state i so that a multinomial logistic model could be fitted. *f(.)* is a multinomial logit function regressing the transition probabilities on the value of covariates up until step t ($\vec{x}_{S_0:S_t}^m, \vec{z}_{S_0:S_t}^m$). The multinomial logit model is given by:

$$P_{ij} = \Pr(S_{t+1}^{m} = j \mid S_{t}^{m} = i, \vec{x}_{S_0:S_t}^{m}, \vec{z}_{S_0:S_t}^{m}) = \frac{e^{\beta_{j0} + \vec{\alpha}_j \vec{x}_{S_0:S_t}^{m} + \vec{\gamma}_j \vec{z}_{S_0:S_t}^{m}}}{1 + \sum_{k \in Set1} e^{\beta_{k0} + \vec{\alpha}_k \vec{x}_{S_0:S_t}^{m} + \vec{\gamma}_k \vec{z}_{S_0:S_t}^{m}}}, j \in Set1 - \{k'\} \qquad (4.16)$$

$$P_{ij} = \Pr(S_{t+1}^{m} = k' \mid S_{t}^{m} = i, \vec{x}_{S_0:S_t}^{m}, \vec{z}_{S_0:S_t}^{m}) = \frac{1}{1 + \sum_{k \in Set1} e^{\beta_{k0} + \vec{\alpha}_k \vec{x}_{S_0:S_t}^{m} + \vec{\gamma}_k \vec{z}_{S_0:S_t}^{m}}}, j = k' \qquad (4.17)$$

We estimated the parameters of the logit model ($\vec{\beta}_j = \left[\beta_{j0}, \vec{\alpha}_j, \vec{\gamma}_j\right]_{(P+Q+1)\times 1}$) using the multinom function available in the R package nnet. Set 2 in Equation (4.11) is the set of states, j's, to which there has not been enough number of transitions from state i to fit a regression model but the patient pathways still show that the transition from i to j is likely to occur. Therefore, in order to keep these less frequent transitions, we have used the empirical fractions (MLE estimates of the multinomial distribution's parameters) to calculate the Pij's for $j \in Set2$ as shown in Equation (4.11). Following the above-mentioned method, we first simulated new covariates of virtual patients. We then conducted the Kolmogorov-Smirnov test where the p-values of all the covariates were greater than alpha=0.05 showing that their distributions were the same as the distributions of real patients' covariates while preserving the covariance structure. Next, we used the Equations (4.11) to (4.17) to generate the patient sequences as a partial function of covariates. We kept generating new sequences until Then, we used the simulated data set $d^{(Simulated)} = \left\{\vec{S}_{1000\times n_m}, \vec{X}_{1000\times 9}, \vec{Z}_{1000\times 25}\right\}$ to test the performance of our variance reduction methodology.

Following the above approach, at each iteration a training data set of 1000 cases was generated and fed into the statistical engine. Figure 4.1 show the variable importance plots for real data. As it can be seen in this figure, among patient profile variables of gender, smoking, and BMI and among hospital resources Attending physician's group, number of CKMB tests and nuclear stress test's turnaround time are identified as the most important factors. After sharing the results with our team of hospital experts, the following concluding remarks were made:

- Strategic planning: Physician practices have always been a concern in this hospital. This study and the numerical results can be an incentive for physicians to more religiously follow the evidence-based medicine to reduce patient flow variations.
- Tactical planning: Contribution of nuclear stress test's turnaround time to patient flow variation was associated with unavailability of nuclear tests during weekends making it longer for patients staying almost idle on weekends. The management decided to extend staff-on-call plan to alleviate this problem.

In summary, by using the random forest classifier we have been able to identify the significant factors that truly impact patient flow. With this valuable information, the hospital management should focus their efforts and resources to improve these attributes, which can consequently improve and facilitate patient flow in the hospital. Finally, to maintain the acquired improvements, the use of multinomial or multiattribute control charts is suggested to constantly monitor and control the important attributes and be alerted if a disturbance occurs in the patient flow process [94].

Note that in our example all the important variables are hospital resource-related attributes. In case a patient profile attribute is identified as a significant variable one should use other alternative solutions to control the process. One solution would be the use of robust optimization methods to control such a process since we cannot control or change statistical distributions of patient profile attributes into our favor [95].

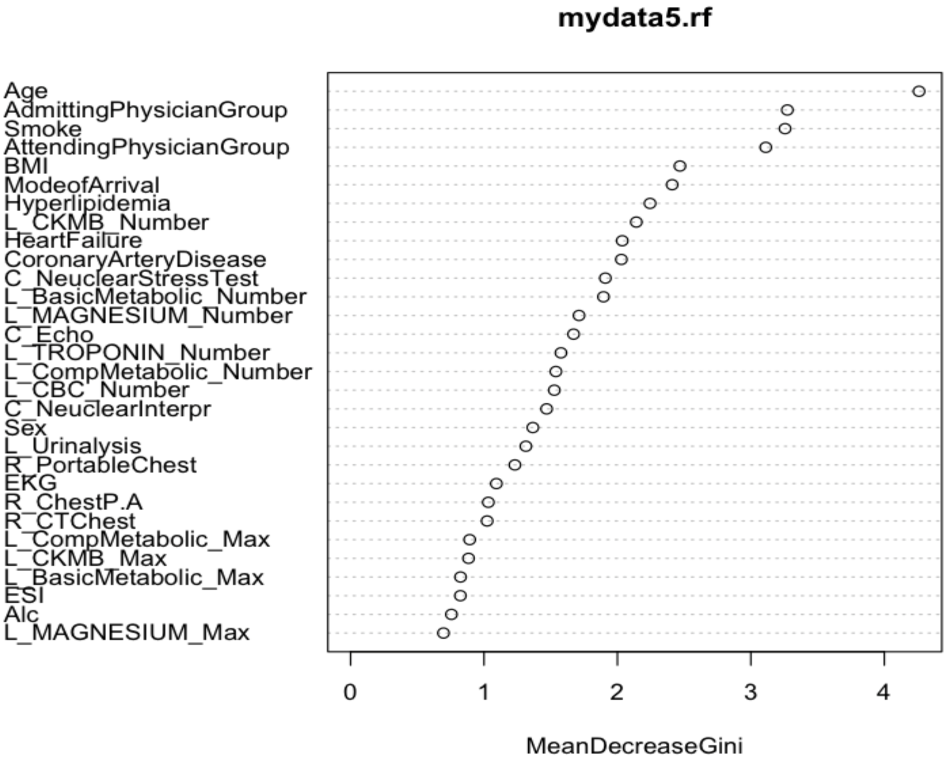


Figure 4.1 Variable Importance Plot, Real Data Set

4.5 Conclusion

In this chapter, we have proposed a novel framework to identify the sources of variations in the patient flow process. The main idea is that by reducing the variations of these single processes we will be able to reduce the variation of patient sequences. Our simulated results show that having a statistically large historical data set, the classifier can correctly determine the important variables, which truly had relationships with patient sequences. We further suggest the use of statistical control charts to maintain the gained improvements. The hospital management can use this valuable information to improve the quality of its patient flow process which consequently improve patient and staff satisfaction and results in a better cost management.

APPENDIX 1- LIST OF TRAFFIC SAFETY RISK PREDICTORS

	Variable Name	Data Table	Subgroup	Source	Variable Type	Number of Levels	Values
1	Age Group	Driver	Demographics	Internal	Categorical	8	Teen, Early20s, Late20s, 30s, 40s, 50s, 60s, 70Plus.
2	Gender	Driver	Demographics	Internal	Binary	2	Male, Female.
3	Years of Driving	Driver	Driving Knowledge	Internal	Categorical	5	[0.0, 3.5), [3.5, 7.0), [7.0,17.0), [17.0,42.0), [42.0,74.0].
4	Passenger In Adjacent Seat	Event-detailed	Driver-Secondary Task	Internal	Binary	2	No, Yes.
5	Cellphone Use	Event-detailed	Driver-Secondary Task	Internal	Binary	2	No, Yes.
6	Speeding	Event-detailed	Driver-behavior	Internal	Binary	2	No, Yes.
7	Low Speed	Event-detailed	Driver-behavior	Internal	Binary	2	No, Yes.
8	Drowsiness	Event-detailed	Driver-behavior	Internal	Binary	2	No, Yes.
9	Alcohol Drug Impairment	Event-detailed	Driver-behavior	Internal	Binary	2	No, Yes
10	Driver Seatbelt	Event-detailed	Driver-behavior	Internal	Binary	2	Yes, No.
11	Travel Lanes	Event-detailed	Roadway Data-Engineering	External	Categorical	5	0, 1, 2, 3, 4Plus.
12	Alignment	Event-detailed	Roadway Data-Engineering	External	Categorical	3	"Curve left", "Curve right", "Straight".
13	Grade	Event-detailed	Roadway Data-Engineering	External	Categorical	3	"Grade Down", "Grade Up", "Level".
14	Relation To Junction	Event-detailed	Roadway Data-Engineering	External	Categorical	8	"Entrance/Exit ramp", "Driveway, alley access", "Parking lot, Inside", "Interchange area", "Parking lot entrance/exit", "Intersection-related", "Intersection", "Non-junction".
15	Traffic Control	Event-detailed	Roadway Data-Engineering	External	Categorical	9	"School zone related sign", "Yield sign", "Slow or warning sign, other", "Construction signs/warnings", "Traffic lanes marked", "Stop sign", "Traffic signal", "No traffic control", "Other".
16	Traffic Flow	Event-detailed	Roadway Data-Engineering	External	Categorical	5	"One-way traffic", "Not divided - center 2-way left turn lane", "No lanes", "Divided (median strip or barrier)", "Not divided - simple 2-way traffic"
17	Locality	Event-detailed	Roadway Data-Engineering	External	Categorical	9	"Bypass/Divided Highway with traffic signals", "Church, "Open Residential, Urban", "School", "Interstate/Bypass/Divided", "Highway with no traffic signals", "Moderate Residential", "Business/Industrial", " Other".
18	Surface Conditions	Event-detailed	Roadway Data-Condition	External	Categorical	3	"Snowy/Icy", "Wet", "Dry".
19	Traffic Density	Event-detailed	Roadway Network	External	Categorical	5	"Flow with some restrictions", "Free flow, leading traffic present", "Free flow, no lead traffic", "Stable flow, restricted maneuverability", "Unstable Flow".
20	Weather	Event-detailed	Weather	External	Categorical	4	"Snow or Sleet or Fog", "Mist or Light Rain or Fog", "Rain or Sleet or Fog", "No Adverse Conditions".
21	Lighting	Event-detailed	Lighting	External	Categorical	4	"Darkness/not lighted", "Dawn or Dusk", "Darkness/lighted", "Daylight".

APPENDIX 2- LIST OF PATIENT PROFILE AND HOSPITAL RESOURCES

Patient profile variables are as follows:

Patient's age, $X_1 = \begin{cases} 1, & age < 45 \\ 2, & 45 \le age < 55 \\ 3, & 55 \le age < 65 \\ 4, & 65 \le age \end{cases}$

Patient's gender, $X_2 = \begin{cases} 1 & Female \\ 2 & Male \end{cases}$

BMI, $X_3 = \begin{cases} 1, & BMI < 18.5, Under-Weight \\ 2, & 18.5 \le BMI < 25, Normal-Range \\ 3, & 25 \le BMI < 30, Over-Weight \\ 4, & 30 \le BMI, Obese \end{cases}$

Alcohol, $X_4 = \begin{cases} 1, & No \quad drinking \\ 2, & Moderate \quad drinking \\ 3, & Heavy \quad drinking \end{cases}$

Smoking, $X_5 = \begin{cases} 1, & Non \quad smo\ker \\ 2, & Moderate \quad smo\ker \\ 3, & Heavy \quad smo\ker \end{cases}$

Drug, $X_6 = \begin{cases} 1, & Non \\ 2, & Moderate \quad use \\ 3, & Subs\tan ce \quad abuse \end{cases}$

Coronary Artery Disease, $X_7 = \begin{cases} 1 & Patient \quad has \quad a \quad history \\ 0 & o.w. \end{cases}$

Heart Failure, $X_8 = \begin{cases} 1 & Patient \quad has \quad a \quad history \\ 0 & o.w. \end{cases}$

Hyperlipidemia, $X_9 = \begin{cases} 1 & Patient \quad has \quad a \quad history \\ 0 & o.w. \end{cases}$

Hospital resource-related variables are:

Attending physician group: $Z_1 = \begin{cases} 1, & Hospital \quad associate \\ 2, & Hospitalist \\ 3, & Medical \quad Teaching \quad group \\ 4, & Other \end{cases}$

Test Turnaround time, $Z_2 \quad to \quad Z \ = \begin{cases} 1, & Within \quad acceptable \quad \lim its \\ 0, & o.w. \end{cases}$

REFERENCES

[1] U.S. Department of Transportation Bureau of Transportation Statistics, "The changing face of transportation," BTS00-007. Washington, DC, 2000.

[2] "10 facts on global road safety," [Online]. Available: http://www.who.int/features/factfiles/roadsafety/en/.

[3] L. J. Blincoe, T. R. Miller, E. Zaloshnja and B. A. Lawrence, "The economic and societal impact of motor vehicle crashes, 2010. (Revised)," NHTSA, Washington, DC, Rep. DOT HS 812 013, 2015.

[4] C. V. J. Oster, T. Bliss, W. A. Bronrott, T. E. Costales, K. L. Cravens and J. J. Cullerton, "Achieving traffic safety goals in the United States, lessons from other nations," TRB, Washington, DC. Rep. 300, 2011.

[5] L. M. Nichols and A. S. O'Malley, "Hospital payment systems: Will payers like the future better than the past?," *Health Aff.,* vol. 25, no. 1, pp. 81-93, 2006.

[6] L. Breiman, "Random forests," *Mach Learn,* vol. 45, no. 1, pp. 5-32, 2001.

[7] G. M. Baydogan and G. Runger, "Learning a symbolic representation for multivariate time series classification," *Data Mining and Knowledge Discovery,* vol. 29, no. 2, pp. 400-422, 2014.

[8] J. Friedman, T. Hastie and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software,* vol. 33, no. 1, pp. 1-22, 2010.

[9] Transportation Research Board of the National Academies of Science, *The 2nd strategic highway research program naturalistic driving study dataset,* 2013.

[10] D. C. Murray, B. Lantz and S. A. Keppler, "Predicting truck crash involvement: developing a commercial driver behavior-based model and recommended countermeasures," ATRI, VA, 2005.

[11] J. R. Treat, N. S. Tumbas, S. T. McDonald, D. Shinar, R. D. Hume and R. Mayer, "Tri-level study of the causes of traffic accidents: Volume I: Causal factor tabulations and assessment," NHTSA, Washington, DC. Rep. DOT HS-805 085, 1979.

[12] D. L. Hendricks, J. C. Fell and M. Freedman, "The relative frequency of unsafe driving acts in serious traffic crashes," NHTSA, Washington, DC. Rep. DTNH22-94-C-05020, 1999.

[13] FMCSA, "Report to Congress on the Large Truck Crash Causation Study," USDOT FMCSA, Washington, DC, Rep. MC-R/MC-RRA, 2006.

[14] ITARDA, "Hito wa don-na misu wo shite koutsuu-jiko wo okosunoka?; kiiwado wa'omoikomi," (in Japanese) ITARDA Information, no. 33, 2001.

[15] M. Abdel-Aty, A. Pande and L. Hsia, "The concept of proactive traffic management for enhancing freeway safety and operation," *ITE Journal,* vol. 80, no. 4, pp. 34-41, 2010.

[16] P. P. Jovanis and H. L. Chang, "Modeling the relationship of accidents to miles traveled," *Transport Res Rec,* vol. 1068, pp. 42-51, 1986.

[17] S. C. Joshua and N. J. Garber, "Estimating truck accident rate and involvements using

linear and Poisson regression models," vol. 15, no. 1, pp. 41-58, 1990.

[18] S. P. Miaou, J. J. Song and B. K. Mallick, "Roadway traffic crash mapping: a space-time modeling approach," *Journal of Transportation and Statistics,* vol. 6, no. 1, pp. 33-57, 2003.

[19] D. Lord, S. R. Geedipally and S. Guikema, "Extension of the application of Conway–Maxwell–Poisson models: analyzing traffic crash data exhibiting underdispersion," *Risk Analalysis,* vol. 30, no. 8, 2010.

[20] S. P. Miaou and J. J. Song, "Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion and spatial dependence," *Accident Analysis and Prevention,* vol. 37, no. 4, pp. 699-720, 2005.

[21] J. Aguero-Valverde and P. P. Jovanis, "Analysis of road crash frequency with spatial models," *TRR: Journal of TRB,* vol. 2061, pp. 55-63, 2008.

[22] S. P. Washington, M. G. Karlaftis and F. L. Mannering, Statistical and econometric methods for transportation data analysis, 2nd ed., Boca Raton: Chapman & Hall/ CRC, 2010.

[23] D. Lord and F. Mannering, "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives," *Transportation Research Part A: Policy Practice,* vol. 44, no. 5, pp. 291-305, 2010.

[24] S. Moutari, M. Herty, A. Klein, M. Oeser, B. Steinauer and V. Schleper, "Modeling road traffic accidents using macroscopic second-order models of traffic flow," *IMA J Appl Math,* vol. 78, no. 5, pp. 1087-1108, 2013.

[25] "Highway safety manual," 1st ed., vol. 2, AASHTO, Washington, DC, 2010.

[26] "A systemic approach to safety- using risk to drive action," US DOT, FHWA, [Online]. Available: http://safety.fhwa.dot.gov/systemic/why.htm.

[27] A. Yumoto, N. Nakano and S. Sano, "Driver visual distraction analysis using percent area of interest method," in *21st ITS World Cong.*, Detroit, MI, 2014.

[28] W. A., A. Glaser, F. Hartwich and F. Roßner, "ViFa 65plus – visual driver assistance systems for elderly drivers," in *21st ITS World Cong*, Detroit, MI, 2014.

[29] R. Taib, I. B. and K. M. G. G. G. Yu, "The future of driver assistance: driver mental state monitoring," in *21st ITS World Congr*, Detroit, MI, 2014.

[30] V. L. Neale, T. A. Dingus, S. G. Klauer, J. Sudweeks and M. Goodman, "An overview of the 100-car naturalistic driving study and findings," 2005. [Online]. Available: https://pdfs.semanticscholar.org/7b74/1bbe1a4da54c48e235b2cfd33c8df8f0b28b.pdf.

[31] H. Rakha, J. Du, S. Park, F. Guo, Z. Doerzaph and D. Viita, "Feasibility of using in-vehicle video data to explore how to modify driver behavior that causes non-recurring congestion," Transportation Research Board, SHRP-2 Report S2-L10-RR-1, Washington D.C., 2011.

[32] "SWOW fact sheet, naturalistic driving: observing everyday driving behavior," Institute for Road Safety Research, 2012. [Online]. Available: https://www.swov.nl/rapport/Factsheets/UK/FS_Naturalistic_driving_UK.pdf.

[33] S. Klauer, T. A. Dingus, V. L. Neale, J. Sudweeks and D. Ramsey, "The impact of driver inattention on near-crash/crash risk: an analysis using the 100-car naturalistic driving study data," NHTSA, Washington, DC. Rep. DOT HS 810 594, 2006.

[34] T. A. Dingus, S. Klauer, V. L. Neale, A. Petersen, S. E. Lee and J. Sudweeks, "The 100-car naturalistic driving study phase II – results of the 100-car field experiment," NHTSA, Washington, DC, Rep. DOT HS 810 593, 2006.

[35] R. J. Hanowski, R. L. Olson, J. S. Hickman and T. A. Dingus, "The 100-car naturalistic driving study: A descriptive analysis of light vehicle-heavy vehicle interactions from the light vehicle driver's perspective," USDOT FMCSA, Report No. FMCSA-RRR-06-004, 2006.

[36] F. Guo, . Klauer, . Hankey and . Dingus, "Near crashes as crash surrogate for naturalistic driving studies," *TRR: Journal of TRB,* vol. 2147, pp. 66-74, 2010.

[37] T. J. Gordon, L. P. Kostyniuk, P. E.Green, M. A. Barnes, D. F. Blower, S. E. Bogard, A. D. Blankespoor, D. J. LeBlanc, B. R. Cannon and S. B. McLaughlin, "A multivariate analysis of crash and naturalistic driving data in relation to highway factors," SHRP 2 Report S2-S01C-RW-1, Project Number: S01(C), 2013.

[38] C. Xua, A. P. Tarkob, W. Wanga and P. Liua, "Predicting crash likelihood and severity on freeways with real-time loop detector data," *Accident Anal Prev,* vol. 57, pp. 30-39, 2013.

[39] M. Hossain and Y. A. Muromachi, "Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways," *Accident Anal Prev,* vol. 45, pp. 373-381, 2012.

[40] C. Shew, A. Pande and C. Nuworsoo, "Transferability and robustness of real-time freeway crash risk assessment," *J Safety Res,* vol. 46, pp. 83-90, 2013.

[41] C. Xu, W. Wanga, P. Liu, R. Guo and Z. Li, "Using the Bayesian updating approach to improve the spatial and temporal transferability of real-time crash risk prediction models," *Transport Res C-Emer,* vol. 38, pp. 167-176, 2014.

[42] M. Ahmed and M. Abdel-Aty, "A data fusion framework for real-time risk assessment on freeways," *Transport Rese C-Emer,* vol. 26, pp. 203-213, 2013.

[43] R. Yua and M. Abdel-Atya, "Multi-level Bayesian analyses for single- and multi-vehicle freeway crashes," *Accident Anal Prev,* vol. 58, pp. 97-105, 2013.

[44] R. Polikar, "Ensemble learning," *Scholarpedia,* vol. 4, no. 1, p. 2776, 2009.

[45] M. A. Jafari, F. Farzan, K. N. M. N. Al-Khalifa and T. Gang, "Development of a risk-based model using naturalistic driver study," presented at 21st ITS World Congress, Detroit, Michigan, 2014.

[46] A. Dutta, G. Bandopadhyay and S. Sengupta, "Prediction of stock performance in the Indian stock market using logistic regression," *International Journal of Business and Information,* vol. 7, pp. 105-135, 2012.

[47] A. Gelman and J. Hill, Data analysis using regression and multilevel- hierarchical models, Cambridge, U.K.: Cambridge University Press, 2007.

[48] T. Hastie, R. Tibshirani and J. Friedman, The elements of statistical learning: data mining, inference, and prediction, 2nd ed., New York: Springer-Verlag, 2009.

[49] J. Landis and G. Koch, "The measurement of observer agreement for categorical data," *Biometrics,* vol. 33, no. 1, pp. 159-174, 1977.

[50] M. Smith and H. Zhang, "SAfety VEhicles using adaptive interface technology (Task 9): A literature review of safety warning countermeasures," 2004. [Online]. Available: https://www.volpe.dot.gov/sites/volpe.dot.gov/files/docs/SAVE-IT%20-

%20A%20Literature%20Review%20of%20Safety%20Warning%20Countermeasures.doc.

[51] Transportation Research Board of the National Academy of Sciences, "The 2nd Strategic Highway Research Program Naturalistic Driving Study Dataset," 2013. [Online]. Available: the SHRP 2 NDS InSight Data Dissemination web site: https://insight.shrp2nds.us.

[52] http://www.trb.org/Publications/PubsSHRP2ResearchReportsSafety.aspx/.

[53] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org.

[54] H. Wickham, ggplot2: Elegant graphics for data analysis, New York: Springer-Verlag, 2009.

[55] K. L. Campbell, "The SHRP-2 naturalistic driving study addressing driver performance and behavior in traffic safety," *TR News,* vol. 282, pp. 30-35, 2012.

[56] S. Buuren and K. Groothuis-Oudshoorn, "mice: multivariate imputation by chained equations in R," *J Stat Softw,* vol. 45, no. 3, 2011.

[57] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News,* vol. 2, no. 3, pp. 18-22, 2002.

[58] E. Brill, "A car-following model relating reaction times and temporal headways to accident frequency," *Transportation Science,* vol. 6, p. 343–353, 1972.

[59] The Texas Department of Insurance, "Driving and Tailgating FactSheet," [Online]. Available: http://www.tdi.texas.gov/pubs/videoresource/fsriskstailgati.pdf.

[60] C. Grover, I. Knight, F. Okoro, I. Simmons, G. Couper, P. Massie and B. Smith, "Automated emergency brake systems: technical requirement, costs and benefits (PPR227)," TRL for DG Enterprise, European Commission, 2008. [Online]. Available: https://circabc.europa.eu/sd/a/3ab87fdc-5715-4733-af50-c3608034ca56/report_aebs_en.pdf.

[61] *Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems, SAE Standard J-3016,* 2014.

[62] R. van der Horst and J. Hogema, "Time-to-collision and collision avoidance systems, Human Factors Research Institute TNO, Soesterberg (Netherlands) 13 p. Safety Evaluation of Traffic Systems: Traffic Conflicts and Other Measures," in *Proceedings of the 6th ICTCT Workshop, International Cooperation on Theories and Concepts in Traffic Safety*, Salzburg, pp. 109-121, 1993.

[63] R. Graham and S. Hirst, "The effect of a collision avoidance system on drivers' braking responses," in *Proceedings of the IVHS AMERICA 1994 Annual Meeting , IVHS AMERICA, Moving Toward Deployment*, Atlanta, Georgia, pp. 743-750, 1994.

[64] D. N. Lee, "A theory of visual control of braking based on information about time-to-collision," *Perception,* vol. 5, pp. 437-459, 1976.

[65] M. Green, "Perception-reaction time: Is Olson (& Sivak) all you need to know?," *Collision,* vol. 4, pp. 88-95, 2009.

[66] S. B. McLaughlin, J. M. Hankey and T. A. Dingus, "A method for evaluating collision avoidance systems using naturalistic driving data," *Accid. Anal. Prev.,* vol. 40, no. 1, pp. 8-16, 2008.

[67] G. T. Taoka, "Brake reaction times of unalerted drivers," *ITE Journal,* vol. 59, no. 3, p. 19–21, 1989.

[68] M. Sivak, P. Olson and K. Farmer, "Radar-measured reaction times of unalerted drivers to brake signals," *Perceptual Motor Skills,* vol. 55, p. 594, 1982.

[69] C. Eberhard, P. Moffa, S. Young and R. Allen, "Development of performance specifications for collision avoidance systems for lane change, merging and backing, Task 4 Interim Report: Development of Preliminary Performance Spec. (No. DOT HS 808 430)," National Highway Traffic Safety Administration, Washington, DC, 1995.

[70] I. Andréasson and X. Ma, "Estimation of driver reaction time from car-following data application in evaluation of general motor–type model," *Transportation Research Record: Journal of the Transportation Research Board,* vol. 1965, p. 130–141, 2006.

[71] G. Johansson and K. Rumar, "Driver brake reaction times," *Human Factors,* vol. 13, p. 23–27, 1972.

[72] T. Magister, R. Krulec, M. Batista and L. Bogdanović, "The driver reaction time measurement experiences," in *In Proceedings of Innovative Automotive Technology (IAT'05) conference*, Bled, 2005.

[73] D. B. Fambro, R. J. Koppa, D. L. Picha and K. Fitzpatrick, "Driver perception-brake response in stopping sight distance situations," *Transportation Research Record,* vol. 1628, p. 1–7., 1998.

[74] P. Ranjitkar, T. Nakatsuji, Y. Azuta and G. S. Gurusinghe, "Stability analysis based on instantaneous driving behavior using car-following data," *Transportation Research Record: Journal of the Transportation Research Board,* vol. 1852, p. 140–151, 2003.

[75] R. E. Chandler, H. R. and E. Montroll, "Traffic dynamics: studies in car following," *Operation Research,* vol. 6, no. 2, pp. 165-184, 1958.

[76] P. G. Gipps., "A behavioural carfollowing model for computer simulation," *Transport ResearchB,* vol. 15, p. 105–111, 1981.

[77] N. Lerner, R. Huey, H. McGee and A. Sullivan, "Older driver perception reaction time for intersection sight distance and object detection," *Report FHWA-RD-93-168, Federal Highway Administration, US DOT,* vol. 1, pp. 33-40, 1995.

[78] S. Sivaraman and M. M. Trivedi, "Towards cooperative, predictive driver assistance," in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, The Hague, 2013, pp. 1719-1724.

[79] V. A. Butakov and P. Ioannou, "Personalized driver/vehicle lane change models for ADAS," *IEEE Transactions On Vehicular Technology,* vol. 64, no. 10, 2015.

[80] J. Pearl, Causality: models, reasoning, and inference, Cambridge: Cambridge University Press, 2000.

[81] G. A. Davis, J. Hourdosb, H. Xionga and I. Chatterjee, "Outline for a causal model of traffic conflicts and crashes," *Accident Analysis and Prevention,* vol. 43, p. 1907– 1919, 2011.

[82] I. Chatterjee and G. A. Davis, "Use of naturalistic driving data to characterize driver behavior in freeway shockwaves," *Transportation Research Record: Journal of the Transportation Research Board,* vol. 2434, p. 9–17, 2014.

[83] D. Smith, W. Najm and R. Glassco, "Feasibility of driver judgment as basis for a crash

avoidance database," *Transportation Research Record: Journal of the Transportation Research Board,* vol. 1784, p. 9–16, 2002.

[84] J. N. Rouder, F. Tuerlinckx, P. Speckman, J. Lu and P. Gomez, "A hierarchical approach for fitting curves to response time measurements," *Psychonomic Bulletin & Review,* vol. 15, no. 6, pp. 1201-1208, 2008.

[85] T. Hastie and J. Qian, "Glmnet vignette," [Online]. Available: http://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html.

[86] A. Cord, C. Ambroise and J.-P. Cocquerez, "Feature selection in robust clustering based on Laplace mixture," *Pattern Recognition Letters,* vol. 27, p. 627–635, 2006.

[87] W. J. Lynk, "One DRG, one price? The effects of patient condition on price variation within DRGs and across hospitals," *International Journal of Health Economics and Management,* vol. 1, no. 2, pp. 111-137, 2001.

[88] S. McClean, M. Faddy and P. Millard, "Markov model-based clustering for efficient patient care," in *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05).*, Dublin, Ireland, 2005.

[89] I. Cadez, D. Heckerman, C. Meek, P. Smyth and S. White, "Model-based clustering and visualization of navigation," *Data Min. Knowl. Discov.,* vol. 7, no. 4, pp. 399-424, 2003.

[90] M. McHugh, K. V. Dyke, M. McClelland and D. Moss, "Improving patient flow and reducing emergency department crowding: a guide for hospitals," *AHRQ Publication.,* vol. 11, no. 12, p. 0094, 2011.

[91] E. N. Weiss and M. A. H. J. C. Cohen, "An iterative estimation and validation procedure for specification of semi-Markov models with application to hospital patient flow," *Oper. Res.,* vol. 30, no. 6, pp. 1082-1104, 1982.

[92] K. (. J. H. L. Ishikawa, Introduction to Quality Control, Productivity Press, 1990, p. 448.

[93] S. J. Tannenbaum, N. H. G. Holford, H. Lee, C. C. Peck and D. R. Mould, "Simulation of correlated continuous and categorical variables using a single multivariate distribution," *Journal of Pharmacokinetics and Pharmacodynamics,* vol. 33, no. 6, pp. 773-794, 2006.

[94] E. Topalidou and S. Psarakis, "Review of multinomialand multiattribute quality control charts," *Qual. Reliab. Engng. Int.,* vol. 25, pp. 773-804, 2009.

[95] C. Wu and M. Hamada, Experiments: planning, analysis and optimization, John Wiley and Sons, Inc., 2000.

[96] M. Dozza, "What factors influence drivers' response time for evasive maneuvers in real traffic?," *Accid Anal Prev.,* vol. 58, p. 299–308, 2013.

[97] K.-F. Wu and P. P. Jovanis, "Crashes and crash-surrogate events: Exploratory modeling with naturalistic driving data," *Accident Analysis and Prevention ,* vol. 45, p. 507– 516, 2012.

[98] J. Reat, N. Tumbas, S. McDonald, D. Shinar, R. Hume, R. Mayer, R. Stansifer and N. Catellan, "Tri-level study of the causes of traffic accidents: Volume I: Causal factor tabulations and assessment, Report No. DOT HS-805 085," National Highway Traffic Safety Administration, USDOT, Washington, D.C., 1979.

[99] S. Lefèvre, A. Carvalho, . Gao, H. E. Tseng and . Borrelli, "Driver models for personalized driving assistance," *Vehicle System Dynamics,* vol. 53, no. 12, pp. 1705-1720, 2015.

[100] H. Halmaoui, K. Joulan, A. C. Nicolas Hautière and R. Brémond, "Quantitative model of the driver’s reaction time during daytime fog – application to a head up display-based advanced driver assistance system," *IET Intelligent Transport Systems,* vol. 9, no. 4, pp. 375-381, 2015.

[101] G. Turghun and S. B. Kim, "Gower distance-based multivariate control charts for mixture and high-dimensional data," 2012. [Online]. Available: http://space.postech.ac.kr/cyber2012_fall/sessionB/B2-5.pdf.

[102] M. Mazzola and G. Schaaf, "Modeling and control design of a centralized adaptive cruise control system," *International Science Index,* vol. 8, no. 7, pp. 1109-1113, 2014.

[103] C. Xu, A. P. Tarko, W. Wang and P. Liu, "Predicting crash likelihood and severity on freeways with real-time loop detector data," *Accident Anal Prev,* vol. 57, pp. 30-39, 2013.

[104] E. S. Park and D. Lord, "Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity," *Transportation Research Record,* vol. 2019, pp. 1-6, 2007.

[105] "Rutgers Plan4safety," Center for Advanced Infrastructure and Transportation, Rutgers University, NJ, [Online]. Available: https://cait.rutgers.edu/tsrc/plan4safety.

[106] K. I. Ahmed, *Modeling drivers' acceleration and lane changing behavior, Ph.D. dissertation, Transportation Systems and Decision Sciences,* Massachusetts Institute of Technology, Cambridge, MA, 1999.

[107] M. M. Balas, V. E. Balas and J. Duplaix, "Optimizing the distance-gap between cars by constant time to collision planning ," in *IEEE International Symposium on Industrial Electronics 2007*, Vigo, Spain, 2007.

[108] A. Cameron and P. Trivedi, Regression analysis of count data, Cambridge, UK: Cambridge University Press, 1998.

[109] FHWA, "Reducing non-recurring congestion," 2012. [Online]. Available: http://ops.fhwa.dot.gov/program-areas/reduce-non-cong.htm.

[110] W. D. S. R. S.-S.-R.-1. [. A. Gary Davis and John Hourdos TRB, "Development of analysis methods using recent data," 2012.

[111] K. D. Kusano and H. Gabler, "Method for estimating time to collision at braking in real-world, lead vehicle stopped rear-end crashes for use in pre-crash system design," *SAE International Journal of Passenger Cars- Mechanical Systems,* vol. 4, no. 1, pp. 435-443, 2011.

[112] J. Li, F. Tsung and Z. C., "Multivariate binomial/multinomial control chart," *IIE Transactions,* vol. 46, no. 5, pp. 526-542, 2014.

[113] G. Maycock and R. Hall, "Accidents at 4-arm roundabouts," TRRL Laboratory Report 1120. Transportation and Road Research Laboratory, Crowthorne, UK, 1984.

[114] National Transportation Safety Board (NTSB), *The use of forward collision avoidance systems to prevent and mitigate rear-end crashes, special investigation report NTSB/SIR-15/01 PB2015-104098,* Washington, D.C., 2015.

[115] T. J. Triggs and W. G. Harris, *Reaction time of drivers to road stimuli,* Victoria, Australia: Human Factors Report No. HFR-12 Human Factors Group Department of Psychology Monash University, 1982.

[116] United States. Federal Transportation Advisory Group, "Vision 2050: An integrated national transportation system," Washington D.C. UNT Digital Library, [Online].

Available: http://digital.library.unt.edu/ark:/67531/metadc25998/. [Accessed 8 November 2016].

[117] M. J. Maher and I. Summersgill, "A comprehensive methodology for the fitting predictive accident models," *Accident Analysis and Prevention,* vol. 28, no. 3, pp. 281-296, 1996.