12-2019

Machine Learning Applications to Predict Road Crash and Soccer Game Outcomes

Lu Bai

THE UNIVERSITY OF NEW HAVEN

Machine Learning Applications to Predict Road Crash and Soccer Game Outcomes

A THESIS

submitted in partial fulfillment of the requirements for the degree of Master of Science in Industrial Engineering

BY

Lu Bai

University of New Haven
West Haven, Connecticut
December 2019

Machine Learning Applications to Predict Road Crash and Soccer Game Outcomes

APPROVED BY

Gökhan Eğilmez, Ph.D.
Thesis Co-Advisor

Rıdvan Gedik, Ph.D.
Thesis Co-Advisor

Nadiye O. Erdil, Ph.D.
Committee Member

Ceyda Mumcu, Ph.D.
Committee Member

M. Ali Montazer, Ph.D.
Program Coordinator

Ronald S. Harichandran, Ph.D.
Dean / Vice Provost for Research

Mario T. Gaboury, J.D. Ph.D.
Interim Provost and Senior Vice
President for Academic Affairs

ACKNOWLEDGEMENT

I would like to express my sincerest gratitude to my thesis advisors, Dr. Gokhan Egilmez and Dr. Ridvan Gedik. I was fortunate to have two excellent professors as my thesis advisors. Dr. Gedik guided my thesis to the right track, with his abundant knowledge and experience. He was always there to support and help me. Under his encouragement, I kept making progress and improving myself. He did not only help me with the research in this thesis but also taught me how to do research and think about questions like a scientist. He inspired my interest in doing research. Because I was an international student, Dr. Gedik also tried his best to make sure I can quickly adapt to the new environment. Dr. Egilmez, with his rich experience and knowledge, provided many insightful suggestions and comments on my thesis. Without his help, many of the exciting results of my research would not have been possible. He was always there for me when my research came across a bottleneck. I am deeply grateful to his kindest help and encouragement when I encountered a health issue in the middle of writing up the thesis.

I would like to thank Dr. Nadiye O. Erdil and Dr. Ceyda Mumcu for being part of my thesis committee, reviewing my thesis, and providing helpful feedback. I would also like to thank my program coordinator, Dr. M. Ali Montazer, for encouraging me and stimulating my interest in research. I would like to thank my classmates and all the professors at UNH who helped me. They were so kind and friendly that I did not feel lonely and helpless. Last but not least, I would like to express my gratitude and love to my parents and my husband. Although my parents were far away in China, their love and care were always by my side. My husband, as an excellent scientist, always encouraged me and gave me lots of good advice. He was always there, taking care of me and supporting me. I couldn't have done it without his love.

ABSTRACT

Machine learning has become a cutting-edge and widely studied data science field of study in recent years across many industries and disciplines. In this thesis, two problems (1- crash severity prediction, 2- soccer game outcome prediction.) were investigated by using a set of machine learning approaches, namely: Ridge regression, Lasso Regression, Support Vector Machine (SVM), Neural Network (NN), Random Forest (RF).

The first study is focused on investigating the critical factors affecting crash severity on a comprehensive time-series state-wide traffic crash data. The dataset covers crashes occurred in the state of Connecticut between 1995 and 2014. Traffic crashes are an increasing cause of death and injury in the world. The overall purposes of the first study were to propose, develop, and implement machine learning approaches in predicting the severity levels of human beings involved in the crashes and investigating the important crash predictors contributing to the injury severity. The predictor variables included road and vehicle conditions, characteristics of drivers and passengers, and environmental conditions. Results indicate that RF provided the best prediction accuracy of 73.85% in correctly classifying a crash based on its severity: fatal, injury, or property damage only. In addition to the overall comparison of proposed machine learning approaches in terms of accuracy, the prediction results were combined with the economic loss of each severity level to provide managerial insights on estimating the financial consequences of traffic crashes. RF provided the importance of each predictor in affecting the severity levels of involved human beings. The ejection status of the driver or passenger was found to be as the most crucial factor leading to the most severe injuries. Besides, a time series analysis of the 20-years crash data was conducted. The analysis results demonstrated that the prediction accuracy of RF increased with period, and the importance of some predictors also changed. From the perspective of policy

making, strict inspection on drunk driving and drug use could lead to substantial road safety improvement. Ejection status is the essential risk factors that affect fatal and incapacitating severity level. The use of seat belts significantly reduces the risk of passengers being ejected out of the vehicle when the crash occurred.

In the second study, recent five-season game data of three major leagues were scraped from whoscore.com. The Leagues were two top European leagues, Spanish La Liga, English Premier League (EPL), and one US League, Major League Soccer (MLS). The purpose of the study was to develop a statistically credible machine learning approaches to predict a soccer game outcome and investigate the significance of predictors (game statistics). Different from previous closely-related studies, the proposed machine learning models were not only applied to the combined dataset of the three leagues but also were studied separately on each league to compare the prediction performance and important predictors. The best prediction performance was achieved by NN with an accuracy of 85.71% (+/- 0.73%) of the combined dataset. For each league, RF had the best performance. RF also provided the importance of each predictor. The results presented that the home-field advantage was more evident in the MLS games than in the other two Europe leagues. The home team or away team factor was the most critical predictor that affected the MLS games. Although it was also an important predictor for La Liga and EPL games, the most influential predictor was the difference in the number of shots on target between the home team and away team. For the three leagues, the number of crosses was the most significant pass type, and the difference in the rate of card per foul was the most crucial card situation. The referee primarily determines the difference in the rate of card per foul. For the Europe leagues, the difference in the number of counter attacks and open plays were consequential attempt types affecting a game result

in La Liga and EPL, while in the MLS, the difference in the number of set-piece was the most crucial predictor variable.

Overall, the results of the two studies indicated that the proposed machine learning approaches yielded effective prediction performance for crash severity and soccer outcomes' prediction. RF had slightly superior prediction performance among the five machine learning models for both studies. Even though the two problem domains were from different industries or policy making area, the proposed machine learning approaches effectively dealt with the complexity of the data in terms of dimensionality and time-series nature.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Machine learning has been used as an effective and practical analytical modeling approach to various problems. In contrast to statistical modeling, which focuses on drawing population inferences from a sample, machine learning focuses on finding the most accurate generalizable predictive patterns among the variables of a dataset (Bzdok et al., 2018). Machine learning models are widely used in various fields for prediction to keep improving prediction accuracy. In this thesis, five machine learning approaches (Ridge regression, Lasso regression, Support Vector Machine (SVM), Neural Network (NN), Random Forest (RF)) were applied on two problem domains, namely: traffic crashes and soccer games. These five models were adopted to both problems since they have been widely used in previous studies in both areas and have been proved to have good prediction performance.

In this thesis, the prediction accuracy results of the five models were generated, compared and analyzed from different perspectives. In addition, potentially useful information was extracted from the data. The findings of the experimentation with machine learning models on crash data revealed implications about which variables to be focused on to most effectively reduce the negative outcomes of traffic accidents for policy making. Insurance companies, safety planners, hospitals, and emergency management centers could use the results to evaluate the economic cost and predict the injury severities of involved human beings. In terms of the soccer data, the betting companies could use the results to more specifically select the optimal prediction model to calculate the odds. The results of feature importance could be used for the reference of coaches of each league to increase the winning probability of soccer teams. The rest of this thesis is organized as follows. The first study “Predict Severity of Traffic Crashes in Connecticut” is provided in

section two. The second study “Predict the Outcome of Soccer Games” is provided in section three. The overall conclusion and future work are provided in section four.

2. PREDICTING SEVERITY OF TRAFFIC CRASHES IN CONNECTICUT

2.1 Introduction

Road traffic crashes is a huge threat to the modern society. According to a recent report by the U.S. Department of Health & Human Service, each year traffic crashes kill 1.35 million people and cause $518 billion economic damage worldwide (https://www.cdc.gov/injury/features/global-road-safety/index.html). Immediate actions are needed to improve road traffic safety and reduce casualties and economic losses. A large body of past research has applied data mining and statistical analysis methods to crash data to gain insight into the pattern of traffic accidents and the significant risk factors associated with the severity of crashes (Chen et al., 2016; Delen et al., 2017; Khattak et al., 2002; Prato et al., 2012). Other studies in this field have mainly focused on the comparison of machine learning models in predicting the severity of the traffic crashes (e.g., Chang and Wang, 2006; Iranitalab and Khattak, 2017; Jeong et al., 2018; Singh et al., 2018; Ye and Lord, 2014; Zhang et al., 2018).

The Connecticut Crash Data Repository (CTCDR), a data repository collected by local and state police, contains detailed information about instances of crashes that happened in Connecticut from 1995 to 2014. It includes information about crash-related, traffic unit related, and involved person associated features. The objective of this research is twofold: 1) comparing the performance of different models and sampling approaches in predicting the severity of involved human beings in traffic crashes; 2) understanding the significant crash-related, traffic unit related, and involved person related features that affect the severity of traffic hazards for the involved person (i.e., driver and/or passengers). In this study, five different machine learning models: Ridge and Lasso regression, Neural Network (NN), Support Vector Machine (SVM), and Random Forest (RF) were

applied to the CTCDR. Due to the heterogeneity of the dataset, over and under-sampling approaches were employed to improve the performance of the algorithms. Economic analysis was conducted to provide managerial insights into a better way to estimate financial consequences of crash accidents. Time-series analysis was conducted to investigate the behavior of data over time.

The rest of our paper is organized as follows. A literature review of the most relevant studies on crash analysis and prediction is summarized in Section 2.2. Section 2.3 describes the preparation process and the features of the dataset. Section 2.4 explains the details the methodology adopted in this study, including the data processing techniques, machine learning models that have been employed, and the performance assessment methods. Section 2.5 introduces the best tuning parameters for each model. Section 2.6 presents predicting and feature importance results, and also compares the performance of the five models with two additional data balancing approaches. Section 2.6 also introduces two analysis approaches based on the results. Finally, Section 2.7 provides the conclusion and discussion.

2.2 Literature Review

Most of the studies on traffic crashes focus on two main aspects. Some studies use feature selection to identify significant factors affecting the severity of traffic accidents and/or to analyze the pattern of different types of crashes (Chen et al., 2016; Delen et al., 2017; Khattak et al., 2002; Prato et al., 2012). Other studies fit and compare various machine learning models in predicting the severity traffic crashes (Chang and Wang, 2006; Iranitalab and Khattak, 2017; Jeong et al., 2018; Singh et al., 2018; Ye and Lord, 2014; Zhang et al., 2018). Some of the studies (Delen et al., 2017; Jeong et al., 2018) used two severity levels, namely the crashes with or without fatalities, while others adopt three (Chen et al., 2016; Jeong et al., 2018), four (Iranitalab and Khattak, 2017; Singh et al., 2018) or five (Zhang et al., 2018) levels of severity.

Regression models are widely used to analyze the features that affect the injury severity. For instance, Khattak et al. (2002) used ordered probit modeling technique to investigate potential factors that contribute to injury severity of older drivers (aged 65 years and above) involved in traffic crashes occurred in Iowa, United States between 1990 to 1999. Ye and Lord (2014) examined the effects of sample size on the three most commonly used models: multinomial logit (MNL), ordered probit, and mixed logit. Additionally, machine learning models have recently been widely used in the literature. For instance, Chang and Wang (2006) employed a classification and regression tree (CART) modeling approach on the 2001 crash data for Taipei. The dataset contains 20 risk factors and among them the vehicle type was the most critical factor associated with injury severity. Prato et al. (2012) applied Kohonen NN for clustering analysis to identify and study the impact of the contributing factors of the pedestrian fatal accidents between 2003 and 2006 in Israel. Five patterns were extracted from the data, which involved the location, circumstances and demographic characteristics of pedestrian accidents.

Several prediction models have been broadly used in the studies on predicting injury severity level, such as NN, SVM and Decision Tree (DT). Iranitalab and Khattak (2017) compared the performance of four methods, including MNL, Nearest Neighbor Classification (NNC), SVM and RF, in predicting the severity levels of the accidents in a dataset that includes 68,448 two-vehicle crashes from 2012 to 2015 in Nebraska, United States. The response variable, namely the severity of the crashes, consists of five categories in the original dataset, with fewer number of observations in the disabling injury and fatal crash categories. To handle the imbalanced data, the authors combined the observations in these two categories and used four categories as the four classes of dependent variable. Two clustering methods, K-mean clustering (KC) and Latent class clustering (LCC), were also implemented along with each machine learning model to tackle the existence of

unobserved heterogeneity in the dataset. Among the four models, MNL performed best with 64.17% overall accuracy, followed by SVM with 61.52%, RF with 59.43% and NNC with 54.74%. The clustering methods did not improve the overall accuracy. The authors also proposed an approach based on crash costs to investigate the overall prediction cost error (OPE) of the models. With this comparison approach, they found that although clustering did not affect the accuracy of the machine learning models, KC and LCC improved the OPE results of MNL, NNC and RF. NNC with KC clustering obtained the best OPE of 26.05%.

The dataset in Jeong et al. (2018) contained 297,113 crash records between 2016 and 2017 and was obtained from the Michigan Traffic Crash Facts (MTCF) database. The original dataset had five categories of severity levels. The authors clustered the five severity levels into two and three levels. The objective of the study was to classify the accidents severity. Five machine learning algorithms, two resampling methods (over-sampling and under-sampling) to tackle the imbalance data, and two ensemble methods (Bootstrap aggregating and majority voting) for the over- fitting problem were adopted. The five algorithms are Logistic regression (LR), DT, Gradient boosting model (GBM), NN and Naive Bayes classifier (NB). The highest performance for 5-class classification was obtained with Bootstrap aggregated decision trees and over-sampling treatment (G-mean=32.1%). The 3-class classification performed best when Bootstrap aggregating was used with decision trees and over-sampling (G-mean=55.4%). For the 2-class classification, the GBM and under-sampling conditions had the highest performance of 62.6% G-mean.

Chen et al. (2016) investigated driver injury severity patterns in 3,185 rollover crashes observed in New Mexico between 2010 and 2011. CART was utilized to identify the significant contributing factors and SVM was used to evaluate the performance in predicting the severity. To handle the influence of imbalanced data, they reduced the original five different severity levels into three

categories. Among a total of 22 predictor variables in the dataset, the results of CART demonstrated that driver seatbelt usage was the most critical factor leading to injury severity outcome in rollover crashes. Lighting condition and road grade were found to be insignificant. Later, 18 significant variables were used as inputs for an SVM learning algorithm. The SVM model performed best on the non-injury category with an accuracy of 58.77%, which is followed by the accuracy of 50.46% for non-incapacitating injury category. For incapacitating injury and fatality category, the model performed most poorly with an accuracy of 22.67%.

Delen et al. (2017) categorized the severity levels of the 279,470 crashes occurred in the U.S. between 2011 to 2012 into two categories: low-level injury and high-level injury. The main goal of Delen et al. (2017) was to identify the significant risk factors influencing the severity of the crashes. Four statistical and machine learning models (NN, SVM, DT and logistic regression (LR)) were adopted. The results showed that SVM model provided the most accurate classification with an overall accuracy of 90.41% in predicting the severity levels, followed by DT with an accuracy of 86.61%. The NN resulted in an accuracy of 85.77% which was better than the LR with an accuracy of 76.97%. In order to incorporate the prediction accuracy of the four models into the analysis of the importance of contributing factors, the prediction accuracy was used as the weights of the models. The weighted importance of each variable was calculated according to the importance of variable in each model and the weight of each model. In this way, they obtained a weighted variable importance value. The results revealed that the most significant variables related to the severity of crashes were the usage of restraining system such as seat belt, the type of collision, whether the driver was ejected from the car and the results of drug test.

Zhang et al. (2018) used a five-level severity crashes dataset that contains 5,538 crashes obtained from Florida, United States. The authors employed ordered probit, MNL, K-Nearest Neighbor

(KNN), DT, RF, and SVM to compare the performance of the six models. The results showed that RF produced the best performance with an overall accuracy of 53.9%, followed by the KNN and SVM with 52.9% and 52.6% accuracy levels, respectively. Singh et al. (2018) employed RF, DT, and MNL models on 2,664 crashes that occurred on Indian highways with four sensitivity levels. Forty-one percent of the dataset pertain to fatal crashes, while only 5.4% of the dataset fall into the property damage crashes. Two sampling methods, synthetic minority over-sampling technique (SMOTE) and randomize class balancing (RCB), were used to tackle the imbalanced data. The SMOTE and RCB improved the overall accuracy of the models. RF with RCB has the best accuracy of 81%.

The objective of the current study was to investigate the degree of influence of different variables that contribute to the severity of crash in the level of involved person (i.e., driver and/or passengers) and identify the best model in classifying the severity levels of involved person in traffic crashes. Our study is the first to apply five machine learning models (i.e., Ridge and Lasso regression, SVM, NN and RF) to the CTCDR data and investigate the significance of risk factors leading to different levels of severity at the same time. Compared with other similar studies, our dataset is more extensive in sample size and scope, covering the crash records in Connecticut, United States, for over 20 years. Because the dataset was extracted from three different sources, there were considerable amount of the missing and heterogeneous cases to be handled. This gave us the opportunity to compare the impacts of over-sampling and under-sampling methods in prediction performance via the proposed machine learning models. In terms of feature significance, our study included 30 independent variables that affect the injury severity of involved person, surpassing most of the aforementioned studies mentioned. Prato et al. (2012), for example,

included 30 independent variables but they only considered pedestrian fatal accidents. Khattak et al. (2002), Chen et al. (2016) and Delen et al. (2017) included 20, 22, 29 variables, respectively.

2.3 Data Preparation

The CTCDR query tool provides detailed information about the crash accidents that occurred in Connecticut between 1995 and 2014. The CTCDR divides the crash features into three independent datasets: (i) 1,723,858 crash records, (ii) 3,218,116 traffic units (including vehicle, pedestrian, and bicycle) that involved in the crash records and (iii) 4,339,479 individual human beings involved, including drivers and passengers. In order to include all the related features provided by CTCDR, the three datasets were concatenated in our study. The observations with a unique and identical crash ID in all three data sources were merged. Table 1 shows the variables in each dataset. The resulting dataset contains a total of 4,339,220 records. Therefore, each record in the final dataset represents a unique person that was involved in each crash.

The combined dataset has 46 variables covering the crash, road, vehicle, driver, pedestrian, passenger and environmental characteristics (Table 1). Two variables, date and time of the crash, were transformed into season and time type of the accident. Some of the variables (measure distance, unit of measure, measure direction, average daily traffic, rural or urban, number of lanes, vehicle maneuver prefix, vehicle maneuver suffix, pedestrian maneuver, first object struck, second object struck) contain substantial missing data. Thus, they were eliminated from further analysis. After merging the three datasets and removing missing data, the final dataset contained a total of 50,034 records with 30 variables (Table 2).

Table 1. Variables in the three sources

Crash	Traffic Unit	Involved Person
date of crash	traffic unit type	seating position
time of crash	year of crash	involved person age
number of vehicles	commercial vehicle code	protection system use
number of pedestrians	vehicle type	airbag status
number of commercial vehicles	vehicle maneuver prefix	ejection status
town	vehicle maneuver suffix	
route class	pedestrian maneuver	
route or road number	driver or pedestrian sex	
route direction	driver or pedestrian age	
cumulative route mileage	traffic unit direction	
ramp or turning road number	alcohol or drug code	
at or between intersections	first object struck	
measure distance	second object struck	
unit of measure		
measure direction		
collision type		
weather condition		
road surface condition		
light condition		
crash occurred on		
other roadway feature		
median barrier penetration		
construction related		
at-fault traffic unit number		
contributing factor		
average daily traffic		
rural or urban		
number of lanes		

Table 2 Variables in the final dataset

Variable	Category
route class	interstate
	US route
	state route
	local road
route direction	north
	south
	east
	west
at or between intersections	at intersections
	between intersections
light condition	daylight
	dark-not lighted
	dark-lighted
	dawn
	dusk
crash occurred on	main roadway
	on ramp
	off ramp
	H.O.V. lane
	collector-distributor roadway
	service or rest area
	weigh station
	connector
median barrier penetration	full
	partial
	none
	not applicable
construction related	yes
	no
season of crash	winter or fall
	summer or spring
time type	daytime (6 am-6 pm)
	nighttime (6 pm-6 am)
driver or pedestrian sex	male
	female
alcohol or drug code	had been drinking (< 0.10)
	intoxicated (0.10 or more)

Table 3 Variables in the final dataset

Variable	Category
	had taken drugs
	had been drinking and taken drugs
	intoxicated and had taken drugs
airbag status	deployed
	not deployed
	not applicable
ejection status	not deployed
	totally ejected
	partially ejected
	trapped
protection system use	none used-vehicle occupant
	shoulder belt only
	lap belt only
	shoulder and lap belt
	child safety seat
	helmet/no high visibility clothing
	no helmet/high visibility clothing
	helmet/high visibility clothing
	restraint use unknown
collision type	turning
	sideswipe
	overturn/angle/head on
	rear-end/backing
	parking/jackknife
	pedestrian/object
	miscellaneous-non collision
contributing factor	driver related
	road related
	vehicle related
	else
vehicle type	automobile
	motorcycle/motor scooter
	Pedal-cycle
	special vehicle
	truck
	trailer
weather condition	no adverse condition

Table 4 Variables in the final dataset

Variable	Category
	rain
	sleet, hail
	snow
	fog
	blowing sand, soil, dirt or snow
	severe crosswinds
	other
other roadway feature	int.public road
	int.private road
	int.residential
	int.commercial Dr.
	on bridge
	at RR crossing
	at median crossover
	at on ramp
	at off ramp
road surface condition	dry
	wet
	snow/slush
	ice
	sand, mud, dirt or oil
	other
traffic unit direction	north
	south
	east
	west
seating position	front seat
	second seat
	third row
	else
commercial vehicle code	yes (it is a commercial vehicle)
	no (it is not a commercial vehicle)
number of vehicles	--
number of pedestrians	--
number of commercial vehicles	--
at-fault traffic unit number	--

cumulative route mileage	--
driver or pedestrian age	--
involved person age	--

In the original dataset the response variable has five levels of severity of the person (driver, pedestrian or passenger) involved in the crash. As shown in Table 3, the five levels are fatal, incapacitating injury, non-incapacitating injury, possible injury, and non-injury. Only 3.82% of the records pertain to fatal crashes while non-injury crashes account for the vast majority of the dataset with 64.08%. Incapacitating injury, non-incapacitating, and possible injury account for 4.66%, 16.55% and 10.89% of all crashes, respectively.

Table 5. Five-class injury classification

Class	**Amount**	**Proportion**
Incapacitating Injury	2331	4.66%
Non-incapacitating Injury	8279	16.55%
Possible Injury	5447	10.89%
Fatal Injury	1913	3.82%
Non-Injury	32064	64.08%
Total	50034	100%

Due to the imbalanced structure of the data, the original response variable was re-categorized into three classes, with the first class representing severe (fatal and incapacitating) injuries, the second class representing non-incapacitating and possible injuries, and the third class representing the observations with no injury. The allocation of observations to each category is shown in Table 4.

Table 6. Three-class injury classification

Severity	**Class Name**	**Amount**	**Proportion**
1	Fatal & Incapacitating	4244	8.48%
2	Injury	13726	27.43%
3	Non-Injury	32064	64.08%
Total	--	50034	100%

The following subsections explain the methods used in creating training and testing datasets, handling sampling error and treating categorical variables, respectively. Figure 1 shows the flow chart of our study.

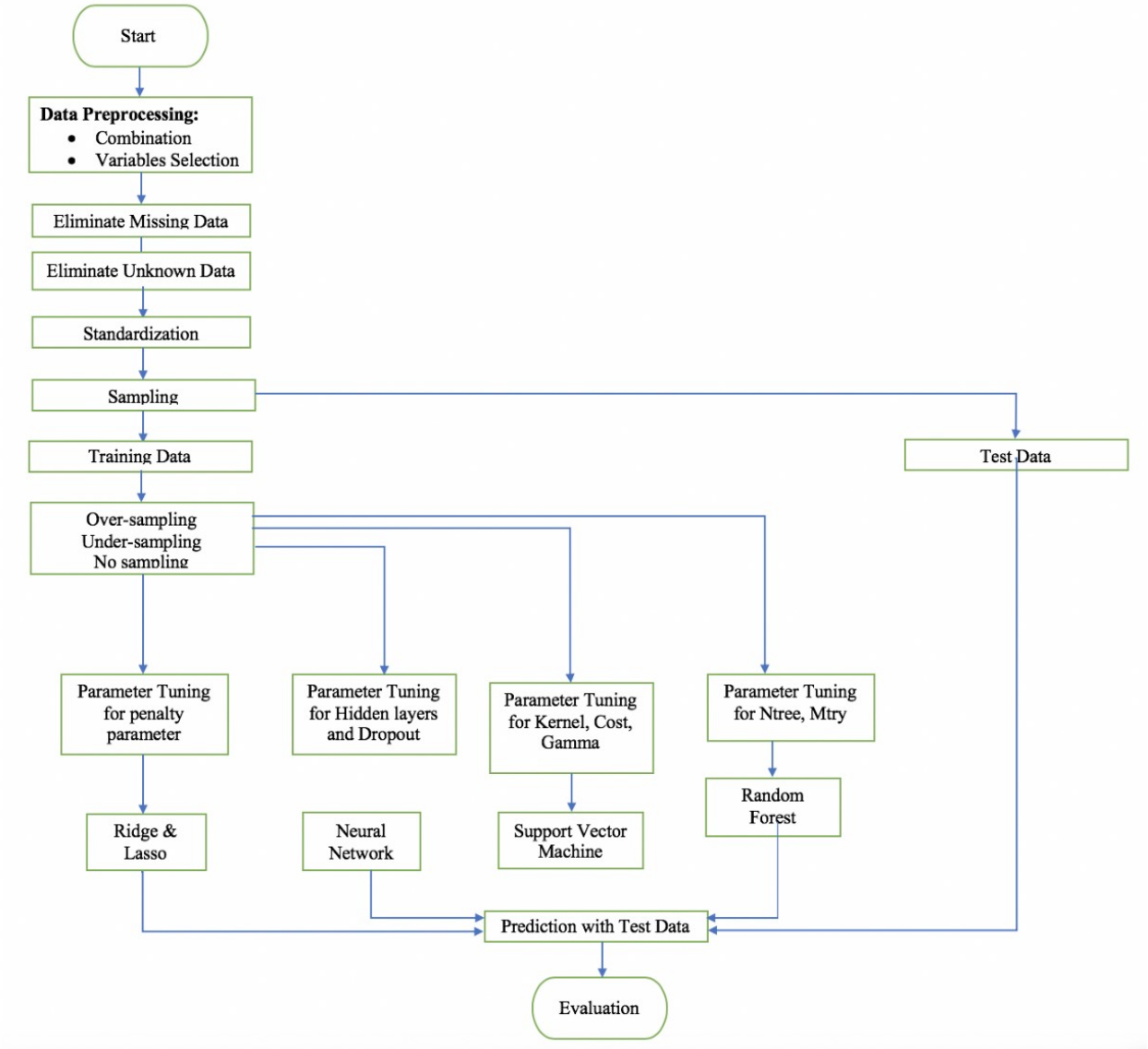


Figure 1. The flow chart of data processing approach

2.3.1 K-fold Cross-validation

K-fold cross-validation is a widely used approach that reduces the bias associated with random sampling of training and test data samples (Kohavi et al., 1995). In a k-fold cross-validation, the dataset is split into k mutually exclusive and similar sized subsets (i.e., folds). Each time, one of the k folds is taken as test data while the remaining k-1 folds are taken as training data. Each fold of the data is used once for testing and for training to eliminate the sampling bias (James et al., 2013). In this manner, the performance of the learning algorithm is evaluated based on the average of the k individual performance as shown in Equation 1 (James et al., 2013).

$$CV = \frac{1}{k}\sum_{i=1}^{k} PM_i \quad (1)$$

where CV stands for cross-validation, k is the number of folds and PM_i is the performance measure used for fold *i*.

2.3.2 Synthetic Minority Over-sampling Technique (SMOTE)

As the proportion of each response category in our dataset shows, the non-injury class in the training set contains disproportionally large amount of observations. In contrast, the fatal & incapacitating category only accounts for 8.48% of the total observations. When using a machine learning algorithm, the imbalance of the dataset may lead to misclassification, affecting the classification accuracy (Sun et al., 2009).

SMOTE deals with imbalanced data by synthesizing new minority instances based on the existing minority instances. The new synthetic instances are generated in the following process: (1) the k-nearest neighbors $\bar{x}$ of each minority instance are calculated based on Euclidean distance (Chawla et al., 2002). (2) depending on the oversampling rate, some neighbors are randomly selected and the difference between the feature vector of the instance under consideration and each selected

neighbor is calculated. (3) multiply this difference by a random number between 0 and 1 and add it to the feature vector of the instance under consideration (x) as follows (Xu et al., 2017).

$$x_{New} = x + rand(0,1) * (\bar{x} - x) \quad (2)$$

In this way, a random point along the line segment between two specific features are generated.

2.3.3 One Hot Encoding

In our dataset, most of the variables are categorical. The categorical independent variables are converted into integer variables via one hot encoding for Ridge and Lasso to improve the performances of these two algorithms (James et al., 2013). If a variable has *n* categories, each category is represented by a unique integer value. For each unique integer value, one hot encoding changes it into a binary variable. The n binary variables are taken as new variables in the dataset instead of the original categorical variables. One hot encoding transforms a variable with *n* categories into *n* binary variables. The binary variable is also called the dummy variable. In our study, one hot encoding was adopted for the Ridge and Lasso algorithms to get better performance in feature selection.

For example, in the "time type" variable, there are two categories. As shown in Table 5, "1" is "daytime", and "2" represents "nighttime". After one hot encoding, "time type" is removed from the dataset while "daytime" and "nighttime" are added as two variables:

Table 7. One hot encoding

Daytime	Nighttime
1	0
0	1

2.4 Methodology

This section explains the five machine learning models (Ridge, Lasso, NN, SVM, RF) used in our research and introduces the two evaluation approaches (confusion matrix, economic analysis) used to compare the performance of the five models. Section 2.4.8 lists the R packages used in our research.

2.4.1 Ridge Regression

When estimating coefficients of independent variables in a linear regression model, the least squares fitting procedure is used. This procedure minimizes the values of the residual sum of squares (RSS) as in Equation 3 (James et al., 2013).

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \quad (3)$$

here $i = 1, 2, 3, ..., n$, $j = 1, 2, 3, ..., p$. n is the number of data points, p is the number of independent variables, y_i is the value of the i th variable to be predicted; x_{ij} is the i th value of the j th independent variable, β_0 is the estimated value of the intercept term and β_j is the estimated value of the slope coefficient which could be interpret as the average effect on y_i of a one unit increase in x_{ij}.

Ridge regression adds a shrinkage penalty term ($\lambda \sum_{j=1}^{p} \beta_j^2$) to the RSS optimization to better estimate the coefficients (Hoerl and Kennard, 1970). λ ($\lambda \geq 0$) is the penalty parameter which needs to be tuned via a proper method, and it has the effect of control the impact of the penalty on the estimates. The shrinkage penalty term is also relative to the value of the coefficients $\beta_1, ..., \beta_p$. In this way, in order to minimize RSS, some coefficients can be shrunk to zero. Generally, the formula for the Ridge regression is given as follows (James et al., 2013).

$$\mathrm{RSS}_{Ridge} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \quad (4)$$

2.4.2 Lasso Regression

Lasso regression is another regularization technique to estimate the coefficients of the regression model. Both Ridge and Lasso regressions have the effect of limiting the size of the estimates. The only difference of the formula of Ridge and Lasso is the penalty parameter. The quantity that Lasso regression minimizes is given as follows (James et al., 2013):

$$\mathrm{RSS}_{Lasso} = \sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i} - \beta_0 - \sum_{\mathrm{j}=1}^{\mathrm{p}} \beta_{\mathrm{j}} \mathrm{x_{ij}})^2 + \lambda \sum_{j=1}^{p} |\beta_{\mathrm{j}}| = RSS + \lambda \sum_{j=1}^{p} |\beta_{\mathrm{j}}| \quad (5)$$

The penalty term in Ridge regression, $\lambda \sum_{j=1}^{p} \beta_{\mathrm{j}}{}^2$, is replaced by $\lambda \sum_{j=1}^{p} |\beta_{\mathrm{j}}|$. Compared to Ridge regression, Lasso uses an L1 penalty instead of an L2 penalty (James et al., 2013). Lasso makes feature selection through continuously shrinking feature coefficients to zero (Tibshirani, 1996).

2.4.3 Neural Networks

Neural network (NN) is a machine learning algorithm inspired by the processing mechanism of the biological neural system. In a biological neural system, groups of neurons interact with each other. These recurrent activities lead to the strengthening of connections between a certain set of neurons.

Neural network is comprised of three layers of neurons: the input layer, the hidden layer, and the output layer (Egilmez and McAvoy, 2017). For each layer, the number of units of neuron and the activation function need to be determined (Egilmez et al., 2019). The widely used activation functions include sigmoid, tanh, softmax and ReLU (Sagar, 2019). The tanh function is mainly used for binary classification (Sagar, 2019). However, evidence shows that when learning complex and high-dimensional data ReLU performs faster and more effectively than sigmoid and tanh (Farhadi, 2017; Groll et al., 2019). The softmax function is often used as the output layer for multiclass classification (Sagar, 2019). In our research, the response variable is multiclass, and our

dataset is high-dimensional, therefore ReLU was used for input and hidden layer and softmax was adopted for the output layer as the activation function.

The performance of NN largely depends on how the structure of the hidden layer is set, including the appropriate number of hidden layers and the number of neurons in each hidden layer. In order to reduce over-fitting, different sets of neurons in each layer can be dropped so that different neural networks can be trained. The dropout procedure significantly improves the performance of neural networks. In this study, the models with different structures were tuned to find the best settings.

2.4.4 Support Vector Machines

Support vector machine (SVM) has become one of the most widely used machine learning methods in recent years. The primary function of SVM is to construct optimal hyperplanes that separate the output classes from each other according to their class labels. The optimal separating hyperplane is also known as the maximal margin hyperplane. The margin is the perpendicular distance from each training observation to a given hyperplane. Therefore, the hyperplane for which the margin is the largest is the maximal margin hyperplane. In other words, it has the farthest perpendicular distance to the training observations. To accommodate a non-linear boundary between classes, SVM uses kernels to enlarge the space between features (James et al., 2013). Kernel functions include polynomial, Gaussian, Radial, and so on.

2.4.5 Random Forest

Random forest (RF) is a very popular variant of decision trees (Goddard, 2006). It performs well both when dealing with regression and classification problems (Liaw et al., 2002). The tuning parameters for a random forest include the number of trees to grow (Ntree) and the number of randomly sampled candidate variables for each split (Mtry). In building a random forest, at each

split in the tree, only a random subset of the predictors are considered by the algorithm, which results in a wide diversity that eliminate over-fitting (Liaw et al., 2002).

2.4.6 Performance Assessment

The performance criteria adopted in this study to compare the prediction models are accuracy, sensitivity, and precision. These three measurements provide a comprehensive picture of the models (Oztekin et al., 2018).

Since the response variable is multi-categorical in this study, the confusion matrix for each category (fatal & incapacitating, injury and non-injury crashes) was calculated. The final results of accuracy, sensitivity and precision are taken as the average of each category, given by the following formulas:

$$\text{Accuracy}_i = \frac{TP_i+TN_i}{TP_i+TN_i+FP_i+FN_i} \qquad i = 1,2,3 \tag{6}$$

$$\text{Sensitivity}_i = \frac{TP_i}{TP_i+FN_i} \qquad i = 1,2,3 \tag{7}$$

$$\text{Precision}_i = \frac{TP_i}{TP_i+FP_i} \qquad i = 1,2,3 \tag{8}$$

Where *i* represents three classes of severity. *TP, TN, FP* and, *FN* respectively denote true positive, true negative, false positive and false negative, defined as follows:

TP: number of samples predicted as true while their actual values were true.

FP: number of samples classified as true while their actual values were false.

TN: number of samples classified as false while their actual values were true.

FN: number of samples classified as false while their actual values were false.

The prediction results can be summarized in a confusion matrix (Table 6). In the confusion matrix, the injury severity level 1, 2, and 3 represent the fatality and incapacitation, injury, and non-injury, respectively. i is the index count of the actual severity and j represents the index count of the predicted severity level. p_{ij} denotes the number of involved human beings with predicted severity of j and the actual severity of i. N_i is defined as the actual number of involved human beings for level i. Therefore, the calculation formulas of the accuracy, precision, and sensitivity in our study are given as follows:

$$\text{Accuracy}_i = \frac{\sum_{i=1}^{3} p_{ii}}{\sum_{i=1}^{3} N_i} \qquad i = 1,2,3, j = 1,2,3 \tag{9}$$

$$\text{Sensitivity}_i = \frac{p_{ii}}{\sum_{j=1}^{3} p_{ij}} \qquad i = 1,2,3, j = 1,2,3 \tag{10}$$

$$\text{Precision}_i = \frac{p_{jj}}{\sum_{i=1}^{3} p_{ij}} \qquad i = 1,2,3, \text{ j=1,2,3} \tag{11}$$

Accuracy illustrates the probability of correct prediction. Sensitivity, also called as true positive rate, measures the proportion of actual positives that are correctly identified (Oztekin et al., 2018). Precision refers to the closeness of two or more measurements to each other (Oztekin et al., 2018).

Table 8. Confusion matrix of a crash severity prediction model

Injury Severity Level		**Predicted**			**Actual Number**
		1	**2**	**3**	
Actual	1	p11	p12	p13	N1
	2	p21	p22	p23	N2
	3	p31	p32	p33	N3

2.4.7 Economic Analysis

Using accuracy to evaluate the models assumes that all severity levels have the same economic losses. In reality, however, the costs of the three different severity levels are different. Iranitalab and Khattak (2017) utilized an alternative approach to compare prediction models. The method combined the prediction accuracy of each severity level with the economic losses caused by different levels of severity.

The Actual Overall Costs of Crashes (AOCC) ($) and Predicted Overall Costs of Crashes (POCC) ($) are defined as:

$$\mathrm{AOCC} = \sum_{\mathrm{i}=1}^{3} N_i C_i \tag{12}$$

$$\mathrm{POCC} = \sum_{\mathrm{i}=1}^{3} \sum_{j=1}^{3} p_{ij} C_j \tag{13}$$

here C_i is the economic costs of each crash with the severity level i. The overall prediction error (OPE) represents the ratio of prediction error in terms of dollar value to the overall actual costs, while Specific Prediction Error (SPE) is the average prediction error on each individual involved in each crash.

$$OPE = \frac{|POCC - AOCC|}{AOCC} \tag{14}$$

$$SPE = \frac{POCC - AOCC}{\sum_{i=1}^{3} N_i} \tag{15}$$

where N_i is defined as the actual number of crashes for severity level i (i = 1, 2, 3).

This economic analysis approach provides a managerial insight for transportation safety policy making. For example, SPE provides a prediction of the average economic loss of each person

involved in a crash for an insurance company or a hospital. OPE provides evidence for safety planners or government to predict annual crash costs (Iranitalab and Khattak, 2017).

2.4.8 R package

Table 7 lists the R packages used in this study.

Table 9. R package

Algorithm	R Package
Ridge	glmnet
Lasso	glmnet
NN	keras
SVM	e1701
RF	randomForest
SMOTE	DMwR

2.5 Parameter Tuning

Parameter tuning plays an essential role in improving prediction results. A 10-fold cross-validation is used for tuning the best lambda for Ridge and Lasso regression (James et al., 2013). Figure 2 and Figure 3 show the tuning results of Ridge and Lasso. Figure 4 and Figure 5 show the results of Ridge and Lasso obtained with over-sampling, and Figure 6 and Figure 7 depict the results obtained with under-sampling. In each plot, the red dotted line is the cross-validation curve, the error bars show the upper and lower standard deviation curves along the λ sequence and the vertical dotted lines indicate the two selected λ. The vertical dotted line on the left represents the value of *lambda.min* which is the value of λ that gives minimum mean-squared error. The line on the right side is *lambda.1se*, which gives the most regularized model where mean-squared error is within one standard error of the minimum (Hastie and Qian, 2016). In our research, *lambda.1se* was taken as the best lambda because more coefficients are shrunk toward zero (James et al., 2013). The best lambda values are shown in Table 8.

Table 10. Best lambda values

Model	Sampling Method	Lambda Value
Ridge	--	0.058
	over-sampling	0.059
	under-sampling	0.069
Lasso	--	0.006
	over-sampling	0.002
	under-sampling	0.005

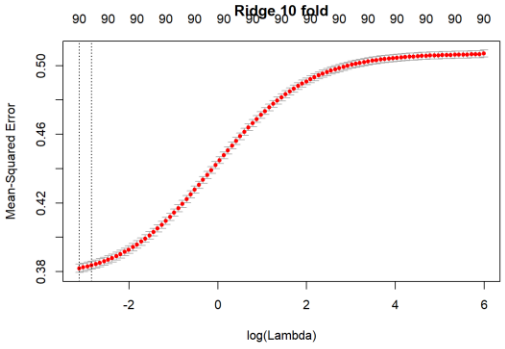


Figure 2. Lambda tuning for Ridge

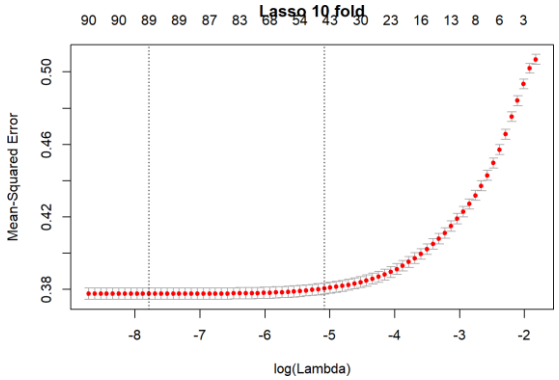


Figure 3. Lambda tuning for Lasso

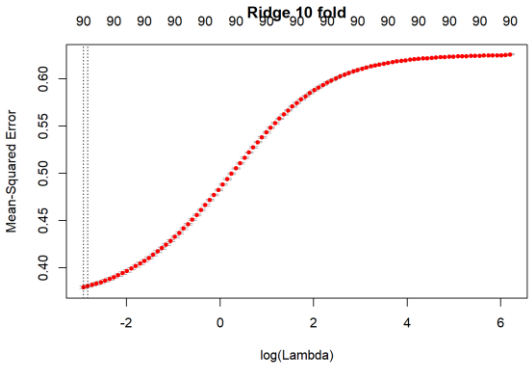


Figure 4. Lambda tuning for Ridge with over-sampling

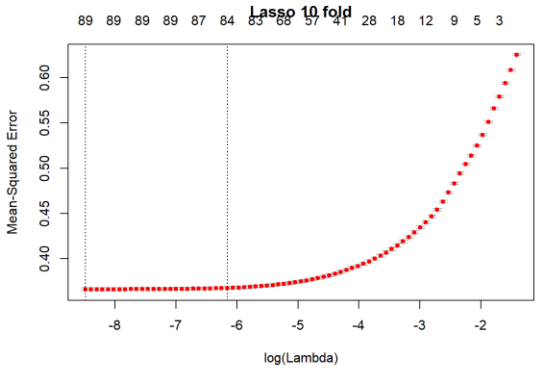
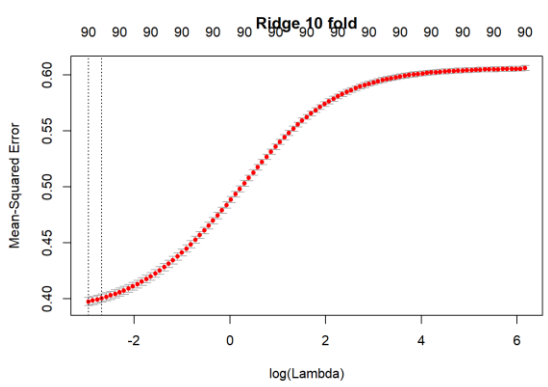


Figure 5. Lambda tuning for Lasso with over-sampling



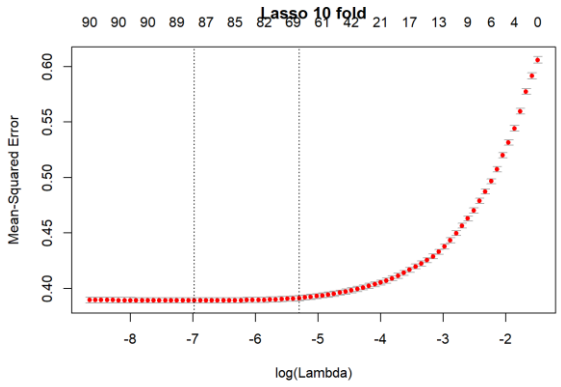


Figure 6. Lambda tuning for Ridge with under-sampling Figure 7. Lambda tuning for Lasso with under-sampling

For NN, the model structure was tuned by trying different activation functions and adjusting layer design features such as the number of hidden layers and the number of nodes in each layer. Five-fold cross-validation was used to detect overfitting and adjusted the dropout parameters. The best model has four layers, with the input layer containing 120 units, the first hidden layer 60 nodes, the second hidden layer 30 nodes, and the output layer 3 units. The activation function used for both the input and hidden layers are "ReLU". For the output layer, the activation function was chosen as "softmax". Forty percent of input units and 30% of each hidden layer's units were dropped out to reduce over-fitting. The categorical cross-entropy was used as the loss function. The metric function was chosen as "accuracy" to judge the performance of the model. Figure 8, Figure 9 show the cross-validation processes of the first two folds for NN. The plot of the first two folds was put here because other folds yielded similar results. The red line and the green line respectively show the changes in the loss and accuracy of training and validation data as the number of iterations increased.

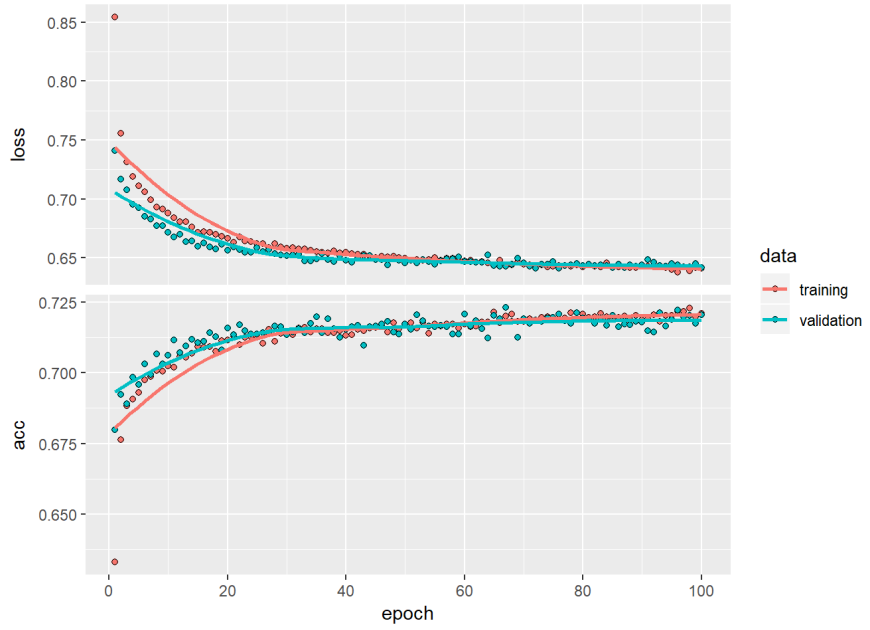


Figure 8. Cross-validation of fold one for NN

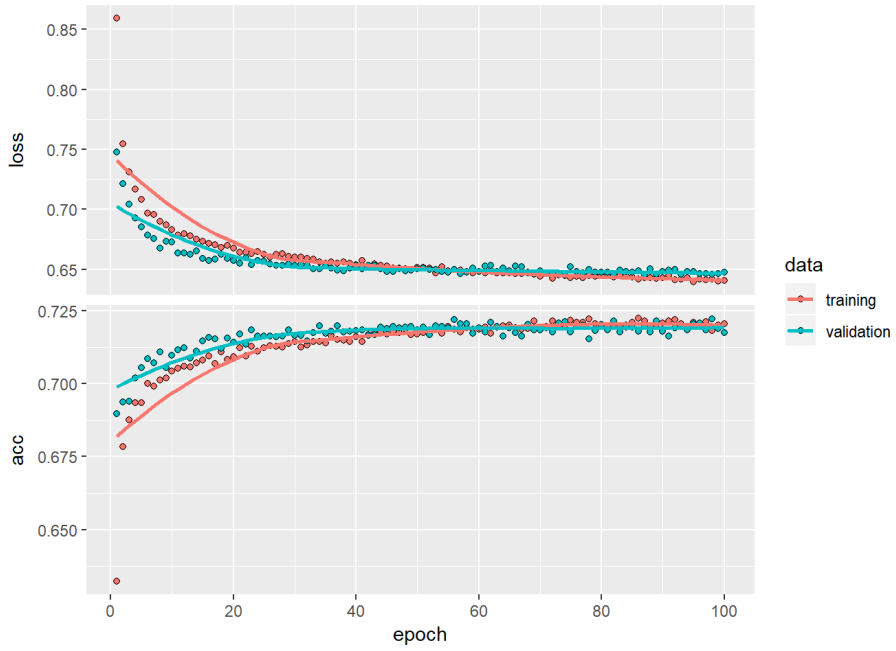


Figure 9. Cross-validation of fold two for NN

For NN with over-sampling, the best model was the same with NN without over or under-sampling. Although the result had over-fitting to some extent, 40% of input units and 30% of each hidden layer's units were dropped out to reduce it. For NN with under-sampling, the best model contained three layers, respectively, the input layer with 90 units, one hidden layer with 10 nodes, and the output layer with 3 units. 40% of input units and 30% of each hidden layer's units were dropped out to reduce over-fitting. The activation, loss and metric functions were adopted the same as normal sampling.

After trying different kernel functions for SVM, radial basis function (also known as RBF) gave the best performance. Cost and gamma are the two parameters of an SVM with an RBF kernel. Table 9 depicts the tuning parameters used in the experimentation.

Table 11. Tuning Parameters for SVM

Sampling method	Cost	Best Cost	Gamma	Best Gamma
--	180,190,200,209	190	0.001,0.01,0.1	0.01
over-sampling	120,170,200	170	0.001,0.01,0.1	0.01
under-sampling	135,150,180	150	0.001,0.01,0.1	0.01

RF has two parameters that need to be tuned: "Ntree" represents the number of trees in the forest and "Mtry" represents the number of variables randomly sampled as candidates at each split. As shown in Figure 10, when the number of trees reaches 200, the value of error no longer changes with Ntree. From Figure 11, the value of Mtry that gives the minimum Out-of-bag (OOB) error was picked. OOB error is a method of measuring the prediction error of random forests (Mitchell, 2011). As shown in Figures 12 - 15, the parameters for RF with over-sampling and RF with under-sampling were chosen in the same way.

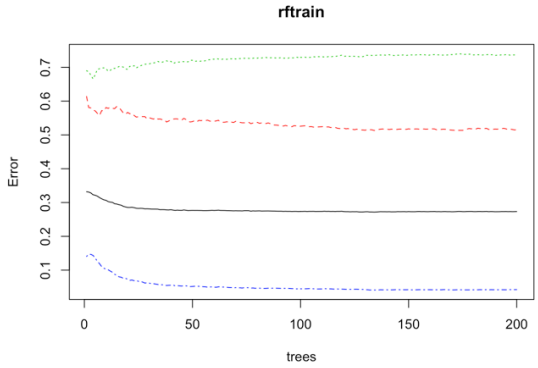


Figure 10. Tuning Ntree for RF

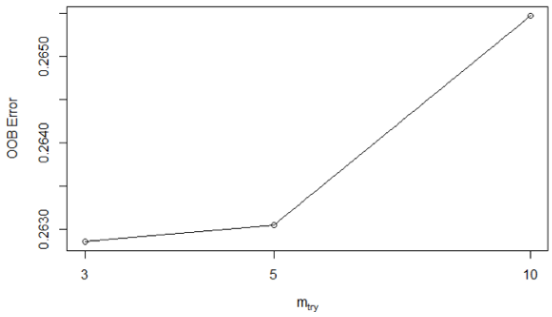


Figure 11. Tuning Mtry for RF

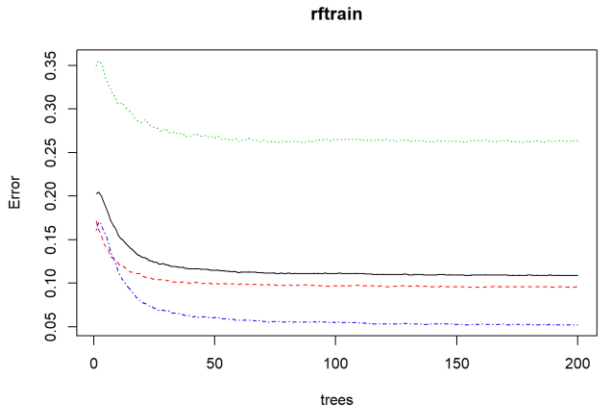


Figure 12. Tuning Ntree with over-sampling

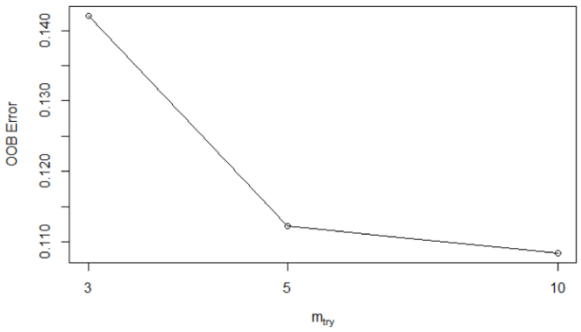


Figure 13. Tuning Mtry with over-sampling

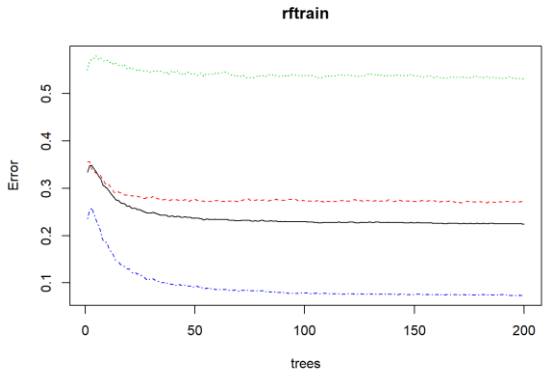


Figure 14. Tuning Ntree with under-sampling

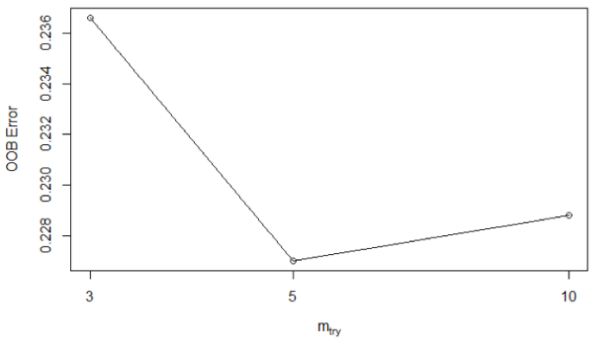


Figure 15. Tuning Mtry with under-sampling

The values of Mtry and Ntree are shown in Table 10.

Table 12. Tuning parameters for RF

Sampling method	Mtry	Ntree
--	3	200
over-sampling	10	200
under-sampling	5	200

Over-sampling and under-sampling were achieved with the SMOTE function in the R package "DMwR". The SMOTE function has two parameters to tune: perc.over and perc.under. perc.over/100 is the number of new examples of the minority class that will be created. The sample size of the majority class will become perc.under/100*(perc.over/100) of the original minority sample size. The values of perc.over and perc.under are shown in Table 11. Because our response variable contains three categories, it was difficult to adjust the three categories to be precisely the same. The distributions of the over and under-sampling datasets are shown in Table 12.

Table 13. Tuning parameters for SMOTE

Sampling method	perc.over	perc.under
over-sampling	400	250
under-sampling	10	3500

Table 14. Over and under-sampling datasets

Sampling method	Fatal & Incapacitating	Injury	Non-injury
over-sampling	15900	9577	22223
under-sampling	3498	3304	7826

2.6 Results

The performance results are shown in Table 13. For the overall accuracy without over-sampling or under-sampling, the five models all provided good prediction performance. RF provided the highest prediction accuracy among the five models of 73.85%, followed by SVM with 72.93%.

The accuracy of Ridge, Lasso, and NN were found to be 72.37%, 72.38%, and 72.22%, respectively.

Over-sampling and under-sampling did not lead to improvement in the overall accuracy but reduced the accuracy of the five models. With over or under-sampling, the accuracy of the five models approximately reduced to 70%. For easier comparison, the accuracy results are shown in Figure 16.

Without over or under-sampling, RF had the highest sensitivity of 51.32%, followed by NN with 50.16% for predicting fatality and incapacitation. The sensitivity results of the other three models were found to be as follows: Lasso (47.56%), SVM (47.74%), and Ridge (47.18%). RF not only had the best performance in predicting fatality and incapacitation, but also had the best sensitivity in predicting injury with a sensitivity of 36.38%. However, RF had the second lowest sensitivity in predicting non-injury with a sensitivity of 92.81%. In contrast, among the five models, NN had the lowest sensitivity in predicting injury, but NN's performance ranked at the second place in predicting the non-injury severity level with a sensitivity of 93.92%. For predicting fatality and incapacitation, Ridge had the lowest sensitivity of 47.18%. However, for non-injury, Ridge had the best sensitivity of 94.27%.

Sensitivity results describe the accuracy of each model in predicting crashes of each severity level. For the three severity levels, fatality & incapacitation obviously has the most significant societal impact. Both over-sampling and under-sampling improved the sensitivity of all the five models in predicting fatality and incapacitation. It means that over-sampling and under-sampling, while reducing overall accuracy, improve the models' ability to predict the type of severity at the most serious level. Under-sampling increased the sensitivity of RF in predicting fatality and incapacitation from 51.32% to 70.49%. With under-sampling, RF became the model with the

highest sensitivity at this severity level. However, both over-sampling and under-sampling decreased the sensitivity of predicting injury crashes. For non-injury crashes, over and under-sampling decreased the sensitivity except for under-sampling to improve the sensitivity of Ridge and SVM.

For the precision, the precision of injury crashes remained the lowest among the three severity levels. Comparing to other models, Lasso had the lowest precision of 54.70% in predicting injury severity level. NN not only had the lowest precision in predicting fatality and incapacitation, but also had the lowest precision in predicting non-injury severity level. For predicting injury crashes, RF had the highest precision of 59.57%. Meanwhile RF had the highest precision in predicting non-injury severity level among the five models with a precision of 77.06%. For fatality and incapacitation, SVM had the best precision of 72.47%.

Over-sampling and under-sampling did not significantly improve the precision for the five models. They also reduced the precision with all three levels of severity, especially for fatality & incapacitation.

.

Table 15. Model performance

Algorithm	Sampling	Accuracy	Precision			Sensitivity		
			fatal& Incapacitating	Injury	Non-injury	fatal& Incapacitating	Injury	Non-injury
Ridge	--	72.37%	72.44%	57.33%	74.92%	47.18%	28.80%	94.27%
	over	69.37%	41.03%	57.93%	75.71%	66.82%	18.12%	91.53%
	under	71.45%	55.82%	58.61%	74.78%	61.28%	20.58%	94.47%
Lasso	--	72.38%	71.57%	54.70%	76.26%	47.56%	33.50%	92.25%
	over	69.68%	40.94%	56.17%	77.07%	66.07%	23.78%	89.71%
	under	71.52%	52.32%	56.86%	76.49%	63.53%	26.16%	91.90%
Neural Network	--	72.22%	69.30%	57.90%	74.88%	50.16%	28.37%	93.92%
	over	71.21%	57.45%	58.14%	74.62%	54.19%	24.49%	93.47%
	under	70.41%	53%	56.33%	74.13%	60.72%	17.57%	94.31%
SVM	--	72.93%	72.47%	57.69%	75.91%	47.74%	32.27%	93.60%
	over	71.27%	53.49%	54.88%	76.77%	56.11%	31.39%	90.27%
	under	70.23%	47%	53.31%	76.57%	67%	23.40%	90.61%
Random Forest	--	73.85%	72.32%	59.57%	77.06%	51.32%	36.38%	92.81%
	over	71.21%	49.96%	54.29%	78.55%	66.07%	33.65%	87.90%
	under	70.44%	47.35%	52.92%	77.56%	70.49%	25.81%	89.45%

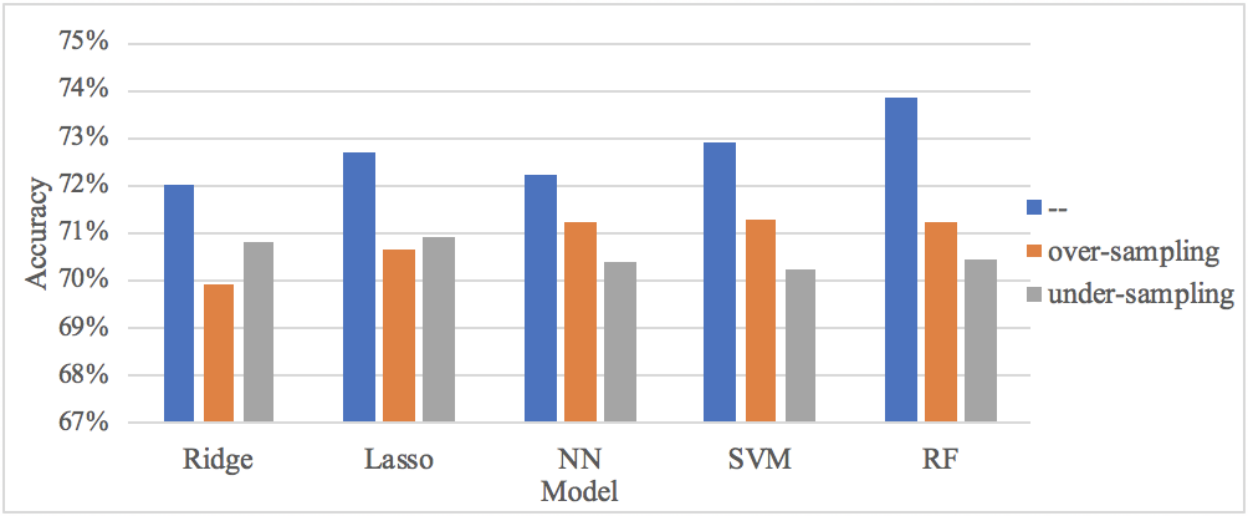


Figure 16. Accuracy comparison

2.6.1 Economic Analysis

As introduced in Section 2.4.7, an economic analysis was carried out to evaluate the performance of the five models to show the result in a more managerial way. The 2018 human capital cost of each severity level is shown in Table 14 which represents the C_i of each severity level. Although Iranitalab and Khattak (2017) used the comprehensive crash cost in their calculation, the human capital cost was adopted for further calculation in our study because each record in our final dataset represented each individual involved. Human capital crash cost estimates include the monetary losses associated with medical care, emergency services, property damage and lost productivity (Part, 2010). Comprehensive crash costs include the human capital costs in addition to non-monetary costs related to the reduction in quality of life in order to capture a more accurate level of the burden of injury (Part, 2010). The 2001 comprehensive crash costs were collected from the Highway Safety Manual (Part, 2010) and using the method that was introduced in it to convert the costs to 2018 costs with the Consumer Price Index (CPI) of 2018.

Table 16. 2018 Crash cost based on severity level

Crash Severity	2018 Human Capital Costs
Fatal & Incapacitating	$882,838
Injury	$51,813
Non-injury	$9,074

(*Rounded to the nearest hundred dollars)

The economic analysis results are shown in Table15. Because OPE represents the ratio of prediction error in terms of dollar value to the overall actual costs, 1-OPE illustrates the ratio of expenses that were predicted accurately. Without over-sampling or under-sampling, RF had the highest 1-OPE of 62.09% while Ridge had the lowest of 49.91%.

Table 17. Economic analysis

Accuracy	Sampling	Accuracy	POCC	OPE	1-OPE	SPE
Ridge	--	72.02%	$791,775,185.18	50.09%	49.91%	-$31,812.66
	over	69.93%	$1,672,879,767.96	28.96%	71.04%	$38,862.29
	under	70.84%	$1,184,802,280.49	0.30%	99.70%	-$287.27
Lasso	--	72.71%	$820,034,916.88	44.92%	55.08%	-$29,545.90
	over	70.64%	$1,675,012,078.01	29.05%	70.95%	$39,033.32
	under	70.90%	$1,309,005,020.85	9.21%	90.79%	$9,675.25
Neural Network	--	72.22%	$685,398,813.12	38.80%	61.20%	-$26,420.53
	over	71.21%	$842,680,891.51	13.12%	86.88%	-$10,702.98
	under	70.41%	$987,736,885.58	9.96%	90.04%	$3,793.07
SVM	--	72.93%	$807,056,616.46	47.25%	52.75%	-$30,586.91
	over	71.27%	$1,171,506,156.88	1.44%	98.56%	-$1,353.77
	under	70.23%	$1,495,534,465.83	20.54%	79.46%	$24,637.11
Random Forest	--	73.85%	$861,719,122.91	37.91%	62.09%	-$26,202.33
	over	71.21%	$1,432,737,701.72	17.06%	82.94%	$19,600.07
	under	70.44%	$1,568,161,352.65	24.22%	75.78%	$30,462.64

Although over-sampling or under-sampling did not help with accuracy, they decreased the OPE for all models. For RF and SVM, over-sampling reduced the OPE of RF from 37.91% to 17.06% and reduced the OPE of SVM from 47.25% to 1.44%. Therefore, the 1-OPE of SVM with over-sampling achieved 98.56%. Also, the SPE for SVM with over-sampling was found as $1,353.77. In contrary to RF and SVM, under-sampling was more effective for reducing OPE than over-sampling for Ridge, Lasso and NN. Over-sampling reduced the OPE of NN from 38.80% to 9.96% and reduced the OPE of Lasso from 44.92% to 9.21%. Ridge with under-sampling had the best performance among all models with an OPE of 0.30%. Therefore, the 1-OPE of Ridge with under-sampling achieved 99.70%. Also, the SPE for Ridge with under-sampling was $287.27. Therefore, the lowest averaged prediction error for each crash in terms of dollar value was found as $287.27. For comparison, the results of 1-OPE which is the prediction accuracy in terms of dollar value are shown as a bar chart in Figure 17.

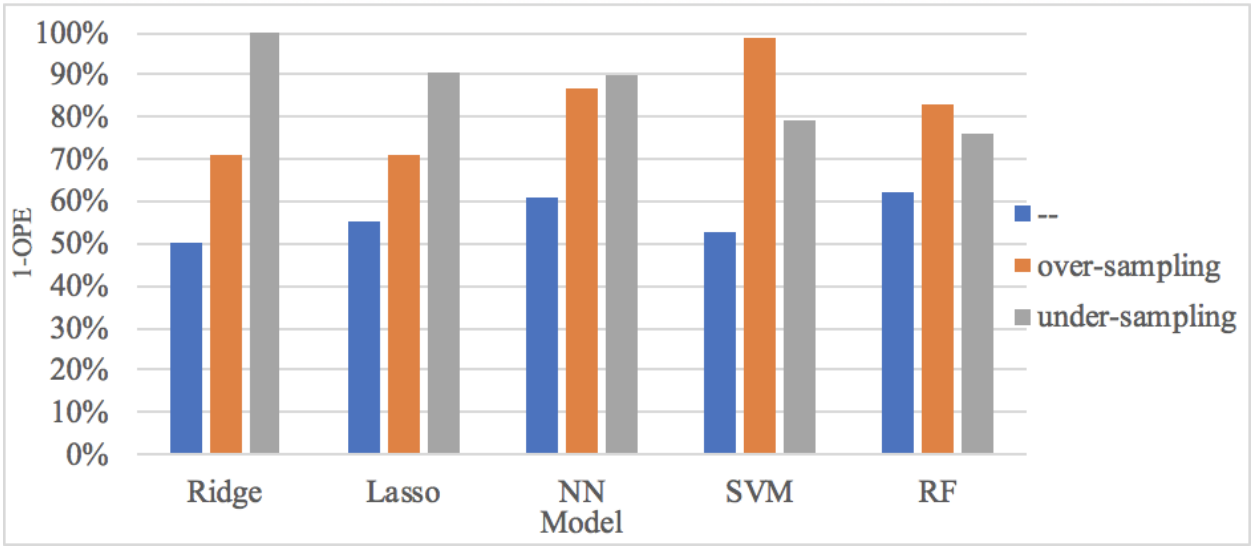


Figure 17. Prediction accuracy in terms of dollar value

2.6.2 Feature Importance

In summary, RF had an excellent and stable performance. For each severity level, RF ranked the influence degree of the 30 risk factors covered in the dataset of this study. The weight of the 30 variables for each severity level are listed in descending order in Table 22. The categories of each independent variable are shown in Table 1.

Table 16 shows that for fatality and incapacitation severity level, the actual weight of ejection status was 79.41. It means that ejection status was the most significant variables for fatality and incapacitation severity level. The actual weight of the ejection status far exceeded the actual weight of the usage of protection system. The usage of protection system was the second important predictor, with an actual weight of 32.37. Airbag status ranked at the third place with an actual weight of 26.43. Ejection status represented whether the passenger or driver was ejected or trapped in the vehicle after a crash happened. It had four categories: not applicable, totally ejected which represented that the passenger or driver was totally ejected from the vehicle, partially ejected, and trapped which means that the passenger or driver was trapped in the vehicle. Usage of protection system had eight different categories. It indicated the use of some protection systems such as seat belt, child safety belt, helmet, and high visibility clothing. Airbag status indicates whether the airbag was deployed when each crash happened. It had three categories: deployed, not deployed, and not applicable. For comparison, the actual weights of variables for fatality and incapacitation severity level are shown in Figure 18. The cumulative curve in Figure 18 is the cumulative weight of the variables. It can be seen that ejection status was far more influential than the other 29 variables. Moreover, the cumulative weight of the top half of the variables already reached 80%.

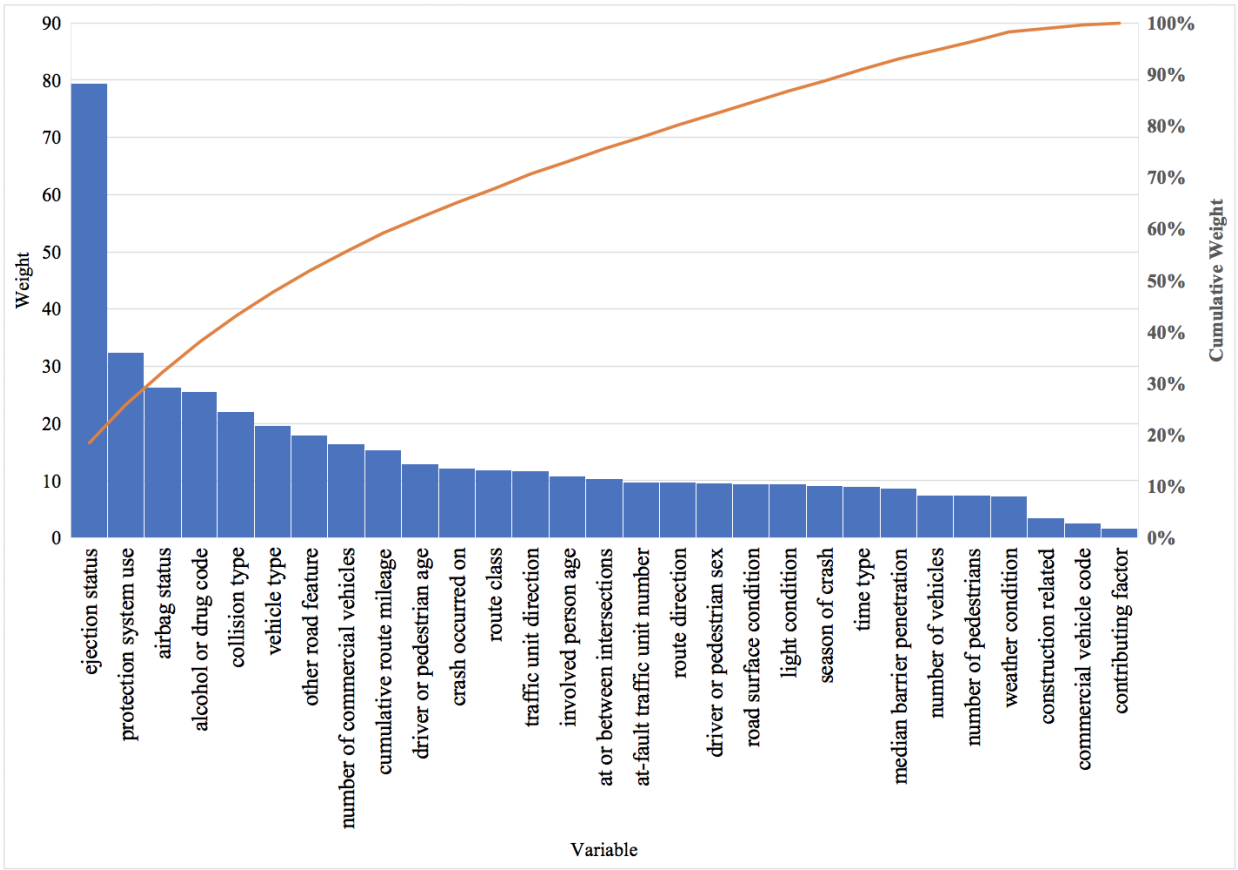


Figure 18. Significant variables for fatal and incapacitating crashes

Table 18. Features selected by random forest (importance in descending order)

Random Forest					
Fatal & incapacitating	Weight	Injury	Weight	No injury	Weight
ejection status	79.41	airbag status	41.42	ejection status	67.61
protection system use	32.37	protection system use	40.36	protection system use	62.08
airbag status	26.43	alcohol or drug code	40.16	alcohol or drug code	56.82
alcohol or drug code	25.56	route class	20.67	airbag status	52.77
collision type	22.09	number of pedestrians	17.49	involved person age	30.62
vehicle type	19.66	collision type	15.72	driver or pedestrian age	30.44
other road feature	17.95	other road feature	14.98	route class	29.35
number of commercial vehicles	16.41	at-fault traffic unit number	14.97	collision type	28.01
cumulative route mileage	15.37	crash occurred on	14.77	cumulative route mileage	27.11
driver or pedestrian age	12.93	at or between intersections	13.36	number of pedestrians	25.07
crash occurred on	12.24	number of vehicles	12.29	vehicle type	21.63
route class	11.90	driver or pedestrian sex	11.60	crash occurred on	21.14
traffic unit direction	11.80	traffic unit direction	10.96	route direction	20.36
involved person age	10.87	median barrier penetration	10.89	seating position	18.81
at or between intersections	10.44	season of crash	10.73	time type	18.61
at-fault traffic unit number	9.81	weather condition	10.13	other road feature	17.81
route direction	9.75	time type	9.09	traffic unit direction	16.47
driver or pedestrian sex	9.60	vehicle type	8.75	number of vehicles	16.08
road surface condition	9.48	road surface condition	8.67	at or between intersections	14.83
light condition	9.48	light condition	8.67	number of commercial vehicles	13.95
season of crash	9.12	cumulative route mileage	8.65	median barrier penetration	13.76
time type	9.05	driver or pedestrian age	7.32	contributing factor	12.91
median barrier penetration	8.70	number of commercial vehicles	5.60	road surface condition	12.62
number of vehicles	7.50	contributing factor	4.80	light condition	12.62
number of pedestrians	7.49	route direction	3.88	season of crash	12.32

weather condition	7.40	seating position	2.48	weather condition	11.58
construction related	3.53	involved person age	2.12	driver or pedestrian sex	8.21
commercial vehicle code	2.60	commercial vehicle code	1.19	at-fault traffic unit number	6.69
contributing factor	1.72	construction related	1.03	commercial vehicle code	3.25
seating position	-2.87	ejection status	-14.04	construction related	0.56
Average	14.26	Average	11.96	Average	22.80
Std.dev	14.43	Std.dev	11.68	Std.dev	16.70

For injury severity level, the weight of airbag status was 41.42. It was the most significant variables for injury crashes, followed by usage of protect system with a weight of 40.36. Alcohol and drug code ranked at third place with a weight of 40.16. Alcohol and drug code referred to the usage of alcohol and drugs by the passenger or driver. According to the different concentration of alcohol or drug used by the driver or passenger, alcohol drug code was classified into five categories. Route class ranked at the fourth place with a weight of 20.67. Route class were classified into four categories: interstate, US route, state route, local road. It indicated the class of the road on which the crash happens. The weights of variables for injury severity level and the cumulative weight curve are shown as a bar chart in Figure 19. The actual weight of ejection status was a negative value, which means that ejection status had negative impact on predicting injury crashes. However, ejection status was the most significant variable for fatality & incapacitation and non-injury severity level, so it was not eliminated from the model. On the other hand, airbag status, alcohol and drug code, usage of protection system were significant variables for these two severity levels.

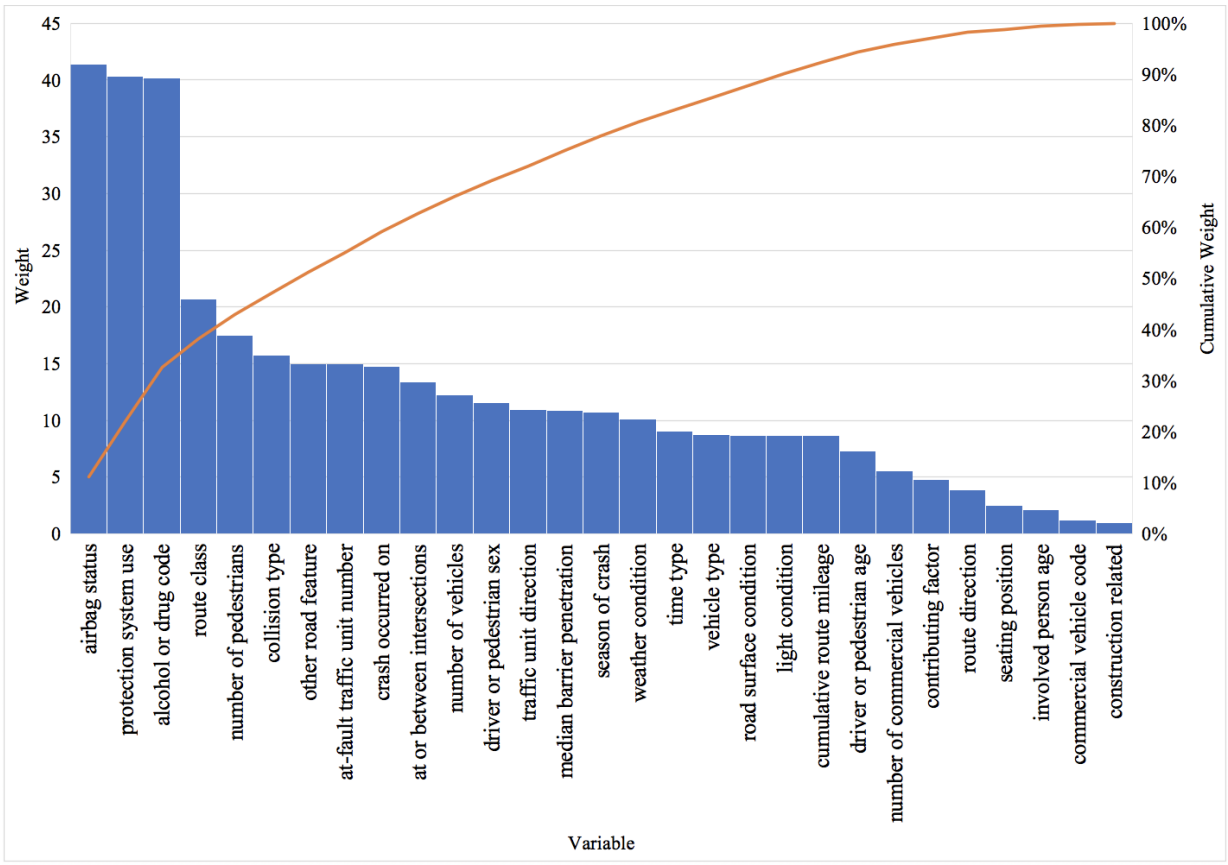


Figure 19. Significant variables for injury crashes

For non-injury severity level, the weight of ejection status was 67.61. It was the most significant variables for non-injury crashes, followed by usage of protect system with a weight of 62.08. Similar to the significant variables for fatality and incapacitation, ejection status, protection system, and airbag status were still the top variables among the 30 factors. The actual weights of variables for non-injury and the cumulative weight curve are shown as a bar chart in Figure 20. The cumulative relative weight of the top half of the variables also reached to 80%. Ejection status, alcohol and drug code, usage of protection system and airbag status were found as significant predictors for all severity levels.

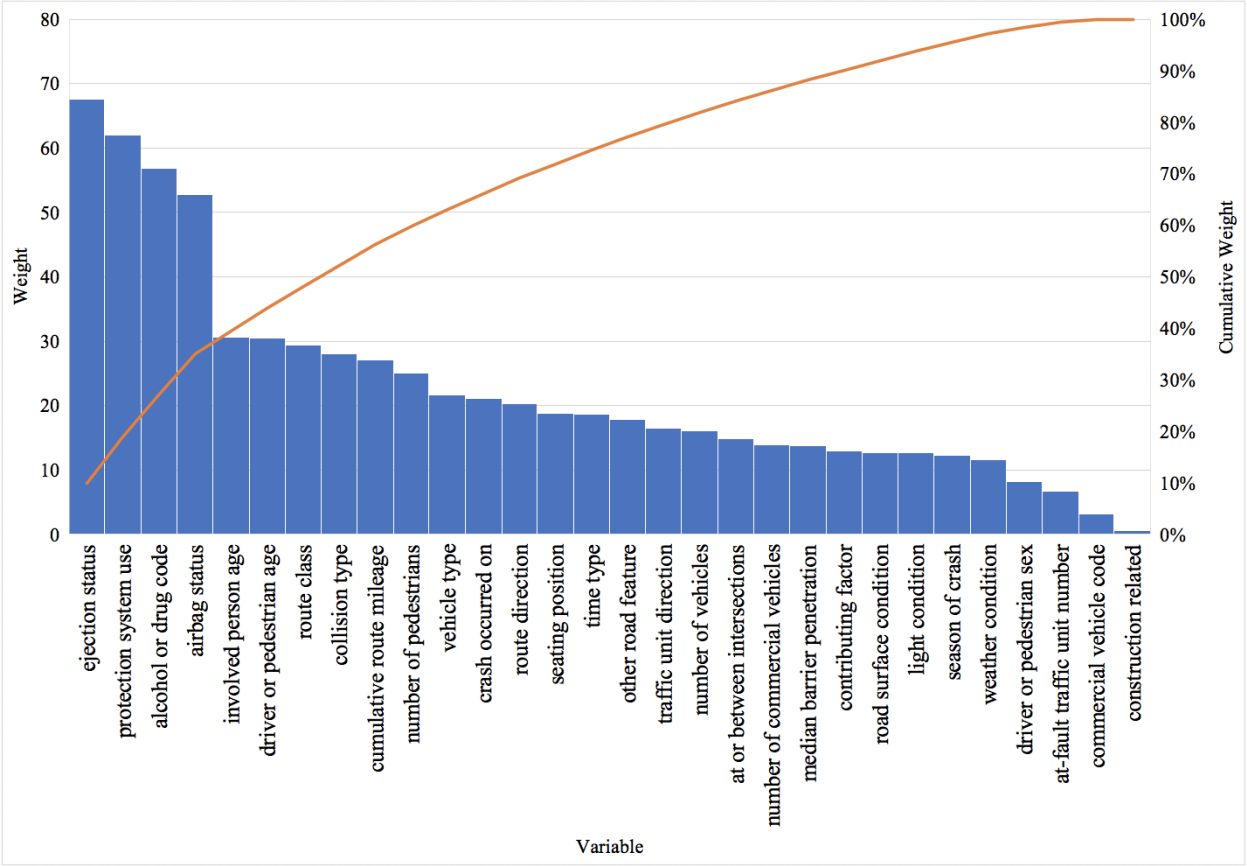


Figure 20. Significant variables for non-injury crashes

2.6.3 Grouped feature importance

The included features belonged to three different categories: crash-related features, traffic unit features, involved person features. Crash related factors included external environmental factors such as road class, weather condition, and so on. Traffic unit related factors included driver, pedestrian, or vehicle conditions. Information about the involved person-related factors included the pedestrian or the driver and all passengers in the vehicle. The grouped feature importance results for each severity level are shown in Table 17 - 19. For each group, the actual weights of the 30 factors were listed in descending order.

For fatality and incapacitation, the most significant crash-related feature was collision type. As is shown in Table 2, collision type had 7 categories. It indicated the type of collision. The most important factor for injury and non-injury was the same, which was route class. Route class had four categories: interstate, US route, state route, local road. Alcohol and drug code was the most important traffic unit related factor for all three severity levels. Ejection status was the most influential involved person related factor for both fatality and incapacitation and non-injury. For injury, airbag status ranked in the first place.

In order to improve road safety, it is difficult to control the crash related features such as light condition, weather condition, route class, etc. from the perspective of policy making. In contrast, factors related to traffic units and involved person could be more effectively controlled at policy level. Of the factors related to the driver, passenger and pedestrian, whether the driver or pedestrian drank or used drugs was the most critical factor for all severity levels of accidents. Therefore, from the perspective of policy making, strict inspection on "operating under the influence (OUI)" may have a significant effect on preventing traffic crashes and/or reducing its impact. The results also provided strong quantitative evidence for the policy of strictly prohibiting drunk driving and drug driving. The factors that are most relevant to the passenger were the passenger's ejection status and protection system usage. The use of seat belts reduced the risk of passengers being ejected out of the vehicle when the crash occurred. Protection systems, including seat belts and helmets, also had a significant impact on the severity of involved human beings. The results further supported the mandatory policies for airbags and seat belts.

Table 19. Grouped important features for fatality & incapacitation

Fatal & incapacitating					
Crash	**Weight**	**Traffic Unit**	**Weight**	**Involved Person**	**Weight**
collision type	22.09	alcohol or drug code	26.43	ejection status	79.41
other road feature	17.95	vehicle type	19.66	protection system use	32.37
number of commercial vehicles	16.41	driver or pedestrian age	12.93	airbag status	26.43
cumulative route mileage	15.37	traffic unit direction	11.80	involved person age	10.87
crash occurred on	12.24	driver or pedestrian sex	9.60	seating position	-2.87
route class	11.90	commercial vehicle code	2.60		
at or between intersections	10.44				
at-fault traffic unit number	9.81				
route direction	9.75				
road surface condition	9.48				
light condition	9.48				
season of crash	9.12				
time type	9.05				
median barrier penetration	8.70				
number of vehicles	7.50				
number of pedestrians	7.49				
weather condition	7.40				
construction related	3.53				
contributing factor	1.72				
Average	10.50	Average	13.83	Average	29.24
Std.dev	4.81	Std.dev	8.27	Std.dev	31.24

Table 20. Grouped important features for injury

Injury					
Crash	**Weight**	**Traffic Unit**	**Weight**	**Involved Person**	**Weight**
route class	20.67	alcohol or drug code	40.16	airbag status	41.42
number of pedestrians	17.49	driver or pedestrian sex	11.60	protection system use	40.36
collision type	15.72	traffic unit direction	10.96	seating position	2.48
other road feature	14.98	vehicle type	8.75	involved person age	2.12
at-fault traffic unit number	14.97	driver or pedestrian age	7.32	ejection status	-14.04
crash occurred on	14.77	commercial vehicle code	1.19		
at or between intersections	13.36				
number of vehicles	12.29				
median barrier penetration	10.89				
season of crash	10.73				
weather condition	10.13				
time type	9.09				
road surface condition	8.67				
light condition	8.67				
cumulative route mileage	8.65				
number of commercial vehicles	5.60				
contributing factor	4.80				
route direction	3.88				
construction related	1.03				
Average	10.86	Average	13.33	Average	14.47
Std.dev	5.00	Std.dev	13.66	Std.dev	25.03

Table 21. Grouped important features for non-injury

Non-injury					
Crash	**Weight**	**Traffic Unit**	**Weight**	**Involved Person**	**Weight**
route class	29.35	alcohol or drug code	56.82	ejection status	67.61
collision type	28.01	driver or pedestrian age	30.44	protection system use	62.08
cumulative route mileage	27.11	vehicle type	21.63	airbag status	52.77
number of pedestrians	25.07	traffic unit direction	16.47	involved person age	30.62
crash occurred on	21.14	driver or pedestrian sex	8.21	seating position	18.81
route direction	20.36	commercial vehicle code	3.25		
time type	18.61				
other road feature	17.81				
number of vehicles	16.08				
at or between intersections	14.83				
number of commercial vehicles	13.95				
median barrier penetration	13.76				
contributing factor	12.91				
road surface condition	12.62				
light condition	12.62				
season of crash	12.32				
weather condition	11.58				
at-fault traffic unit number	6.69				
construction related	0.56				
Average	16.60	Average	22.80	Average	46.38
Std.dev	7.37	Std.dev	19.25	Std.dev	20.90

2.6.4 Time-Series Analysis

Our dataset contained 20-year crashes data. The distribution of the number of crashes per year over the 20 years in the experimental dataset is shown in Figure 21.

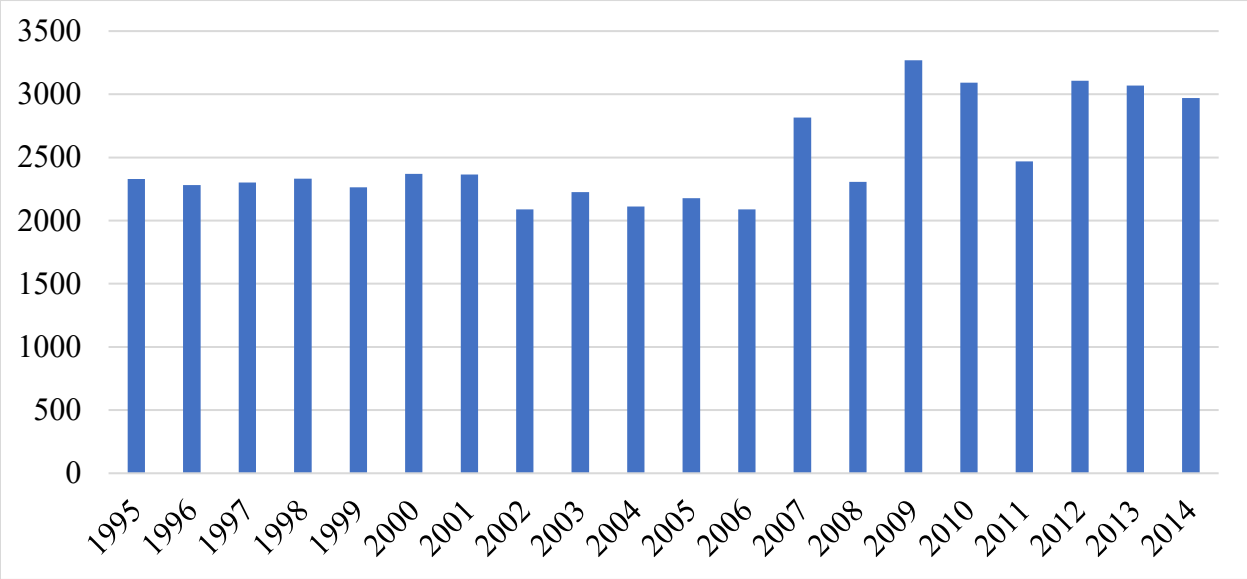


Figure 21. Distribution of the number of crashes over 20 years

To investigate how the behavior of data changes over time, the dataset was divided into four separate datasets to ensure that each period contained more than 10,000 records. Each dataset included 5-year crashes data. RF, which was the best model for the 20-year dataset, was applied to each of the 5-year datasets. The prediction accuracy for each period is shown in Table 20. Recall that the prediction accuracy for the 20-year data (1995-2014) was 73.85%. The prediction accuracies for the four periods increased over time as shown in Table 20, from 71.90% for 1995-1999 time period to 77.07% for 2010-2014.

Table 22. Prediction accuracy for each period

Period	Accuracy
1995-2014	73.85%
1995-1999	71.90%
2000-2004	74.07%
2005-2009	75.30%
2010-2014	77.07%

The top five critical factors for each period are shown in Table 21. For fatal and incapacitating severity level, the top two factors, namely ejection status and protect system usage, did not change over time. The importance of airbag status increased over time, and it became the third important factor in 2005-2009. The importance of alcohol drug code decreased over time, but it still ranked at the fifth place. Vehicle type became the fourth important factor in the 2005-2009.

For injury crash, route class was the most significant factor for 1995-1999 and 2000-2004 periods. However, its importance kept decreasing. It ranked 15th for the 2010-2014 period. Alcohol drug code and airbag status were the most important factors in 2005-2009.

For a non-injury severity level, ejection status was found as the most crucial factor. Airbag status became one of the top five significant variables from 2000-2004 period and its significance kept increasing over time. This result was identical to policy making that federal legislation made airbags mandatory since 1998. Alcohol drug code and protect system usage were essential factors for the period. The importance of route class kept decreasing. The age of the involved person became a vital factor since 2005.

Table 23. Top five important features for different time period

	Fatal&Incapacitating	Weight	Injury	Weight	Non-injury	Weight
1995-1999	ejection status	64.14	route class	27.89	ejection status	53.58
	protection system use	22.63	collision type	17.48	protection system use	49.85
	alcohol or drug code	17.02	number of vehicle	16.46	route class	28.73
	collision type	16.54	protection system use	14.53	alcohol or drug code	27.38
	other road feature	16.53	cumulative route mileage	14.41	number of pedestrians	25.76
2000-2004	Fatal&Incapacitating	Weight	Injury	Weight	Non-injury	Weight
	ejection status	39.93	route class	24.99	ejection status	37.50
	protection system use	20.47	alcohol or drug code	13.33	protection system use	33.94
	alcohol or drug code	15.60	at-fault traffic unit number	13.26	route class	25.82
	airbag status	13.84	other road feature	12.36	alcohol or drug code	25.11
	collision type	12.02	number of vehicle	11.45	airbag status	20.73
2005-2009	Fatal&Incapacitating	Weight	Injury	Weight	Non-injury	Weight
	ejection status	36.01	alcohol or drug code	27.76	ejection status	34.73
	protection system use	19.10	airbag status	23.84	protection system use	28.49
	airbag status	16.01	protection system use	18.38	alcohol or drug code	28.33
	vehicle type	12.70	collision type	10.42	airbag status	24.97
	alcohol or drug code	9.97	route class	8.61	driver or pedestrian age	18.12
2010-2014	Fatal&Incapacitating	Weight	Injury	Weight	Non-injury	Weight
	ejection status	40.00	airbag status	36.36	ejection status	36.52
	protection system use	18.74	alcohol or drug code	18.12	airbag status	32.21
	airbag status	14.67	protection system use	9.94	protection system use	31.22
	vehicle type	13.66	seating position	7.33	alcohol or drug code	22.19
	alcohol or drug code	9.13	other road feature	6.24	driver or pedestrian age	17.75

2.7 Conclusion and Discussion

All of the five machine learning models proposed in this study had good prediction performance. RF had the highest overall prediction accuracy of 73.85%. For each severity level, RF still had the best sensitivity in predicting both fatal and incapacitating and injury severity level. Ridge had the best sensitivity at predicting non-injury severity level. Although over-sampling and under-sampling did not improve the overall prediction accuracy, they did improve the models' ability to predict fatal & incapacitating severity level. The best accuracy in predicting fatal & incapacitating severity level increased from 51.32% to 70.49%, which was achieved by RF with under-sampling. The findings of the experimentation with machine learning models on crash data revealed implications about which variables should be focused on to most effectively reduce the negative outcomes of traffic accidents in policy making. In addition, this study employed economic analysis to evaluate the performance of the five models, which made the results more practical. Although over-sampling and under-sampling were not helpful in increasing prediction accuracy, they decreased the prediction error in dollar value. Ridge with under-sampling had the best 1-OPE and SPE. The best SPE was -$287.27, which means that on average the predicted economic loss for each crash was only $287.27 less than the actual economic loss. OPE and SPE can be adopted in a wide range of practical applications. For example, insurance companies and safety planners could use OPE and SPE to estimate the economic costs of crashes in a future year; hospitals and emergency management centers could use the model with the lowest SPE to evaluate the economic loss.

From the perspective of policy making, strict inspection on drunk driving and drug use could lead to substantial road safety improvement. Ejection status is the essential risk factors that affect fatal and incapacitating severity level. The use of seat belts significantly reduces the risk of passengers

being ejected out of the vehicle when the crash occurred. Usage of protection systems, including seat belts and helmets, has a significant impact on the severity of involved human beings. The findings have implications for which variables to be focused on to most effectively reduce the severity of involved human beings. The state transportation department, police officer and vehicle manufacturers could review the results to improve the safety of our transportation and road activities.

The time series analysis results showed that the prediction accuracy of RF increased over time, from 71.90% for 1995-1999 period to 77.07% for 2010-2014. The model had better accuracy when dealing with more recent data. Future work could compare the analysis results with the changes in policy over time to determine whether the improvement in accuracy and changes in important factors are related to policy.

With more powerful computation server, alternative methods can be adopted to handle the missing data more effectively so that one can fit the models with larger volume of crash records to improve the prediction performance and feature importance results. For the tuning process of NN, the parameters and structures are chosen based on a grid search, which is another major limitation of this study. Lam et al. (2001) presents a tuning of the structure and parameters of NN using an improved genetic algorithm (GA). Tsai et al.(2006) apply a hybrid Taguchi-genetic algorithm (HTGA) to solve the tuning problem of NN. These tuning methods may help us get better parameters to improve the performance of NN.

3. PREDICTING OUTCOME OF SOCCER GAMES

3.1 Introduction

Machine learning models have been profoundly used in sports analytics for a number of objectives such as predicting game results and extracting useful informations about important features that affect the performance of teams and organizations. Such applications have offered numerous managerial benefits to sports organizations, managers, athletes, and the media. Although the literature related to analytics of soccer games was not as sophisticated compared with other professional sports (Kerr, 2015), its application in the soccer field has gradually proliferated in recent years. Some studies have adopted pre-play features (match statistics from previous games) to predict future games and have demonstrated that the prediction model has an practically credible performance in predicting the game results (Hubacek et al., 2019; Lock and Nettleton, 2014). Numerous studies focused on predicting the game outcomes with features extracted from selected leagues' game results and statistics (Kerr, 2015) and investigating the features that significantly affect the game outcome.

For soccer analysis, statistical learning was frequently used to predict game results and to analyze whether certain factors affect the outcomes in the early years (Magel and Melnykov, 2014; Goddard, 2006). Descriptive statistics are used for drawing inferences about population from sample, while machine learning models are focused on improving the accuracy of prediction (Bzdok et al., 2018). In recent years, machine learning models with high prediction performance have been developed and introduced to sports analytics. Neural Network (NN) has been a popular machine learning model in sports prediction. Support Vector Machine (SVM) and Random Forest (RF) are relatively new supervised models but are proved to have excellent prediction performances in different problem domains (Baboota and Kaur, 2019; Lock and Nettleton, 2014;

Ulmer et al., 2013; Yezus, 2014). Kerr (2015) applied Ridge regression to soccer game prediction and demonstrated Ridge to be an effective model in soccer analysis.

In this study, five well-known and widely used machine learning approaches were applied: Ridge regression, Lasso regression, NN, SVM, RF. The abovementioned machine learning approaches have been adopted by many previous studies due to their statisfactory performance in predicting game results (Kahn, 2003; Kerr, 2015). The five machine learning models were applied to 5-season game results of three soccer leagues to compare the performance of the models in predicting the games of different leagues. The three leagues were English Premier League (EPL), Spainish La Liga and U.S. Major League Soccer (MLS). EPL and La Liga are two top Europe leagues, which have been studied in numerous studies (Baboota and Kaur, 2019; Zambom-Ferraresi et al., 2018). A U.S. league, MLS was also included in this study.

The prediction performance results could be used for the betting industry. The betting companies could use the results to more specifically select the optimal prediction model to calculate the odds. The results of feature importance could be used for the reference of coaches and head coaches of each league to improve the performance of soccer teams. According to the feature importance, coaches could focus on improving the most influential features to increase the probability of winning.

The purpose of this study was to understand the underlying statistical features that critically affect the soccer game results of the three major soccer leagues. The prediction performance of different machine learning models for game results of different leagues was examined and the feature importance for each league was investigated. This study also aimed to further improve the prediction accuracy on the basis of other studies.

The rest of the chapter is organized as follows. The literature review of the most relevant studies on soccer analytics is summarized in Section 3.2. Section 3.3 explains the data preparation. Section 3.4 introduces the methodology of the machine learning approaches and experimentation. Section 3.5 presents the best tuning parameters for each model. The prediction performance results, and feature importance results are shown in section 3.6. The last section provides the conclusion and discussion.

3.2 Literature Review

Soccer is one of the most popular sports in the world (Sawe, 2018). More than half of the global population identifies themselves as soccer fans (Sawe, 2018). Besides, soccer is leading the worldwide sports in terms of its market size with an annual revenue of $28 billion (A.T. Kearney, Inc., 2011). Thus, with the availability of more computational power, many researchers have focused on applications of data mining to a variety of problems related to soccer games, leagues, and players. Several machine learning algorithms have been applied to soccer datasets to predict the game outcome (Baboota and Kaur, 2019; Eggels et al., 2016; Hubacek et al., 2019; Mackay, 2017; Shin and Gasparyan, 2014; Ta and Joustra., 2015; Ulmer et al., 2013; Yezus, 2014). Some other researchers focused on feature selection to determine the impact of statistical features such as shots on target, number of fauls, etc. on the game outcome (Carmichael andThomas, 2005; Magel and Melnykov, 2014).

Kerr (2015) developed and applied machine learning models to derive insights from a dataset of soccer ball events collected from OptaPro (https://www.optasportspro.com/). In the first experiment, an L2-regularized logistic regression model was used to predict the soccer game results and investigate the critical features related to the outcome of the games. The researchers used two of the outcomes, win and loss, and did not consider any tie game in their analysis. The

dataset contained 19 ball-events features, each of which was calculated for every away-team, home-team, and the difference between the home-team and away-team. The dataset also contained a variable indicating whether the studied team was a home-team or an away-team. The final dataset consisted of 58 variables. Further classification was made to classify 10 variables as obvious (home-team/away-team factor, number of shots on target, number of shots, number of crosses) and 48 of them as non-obvious variables (number of passes, number of tackles, number of cards, etc.). The first and second L2-regularized logistic regression models contained only the obvious variables and non-obvious variables, respectively, and their third model contained all the variables. The highest accuracy rate in predicting the outcome (84%) was achieved by the model that contained both obvious and non-obvious variables. Based on the results obtained from this model, the difference in the number of shots on target between the home-team and away-team was the most important feature in determining the winning team. The most influential variables for the models with only obvious and non-obvious variables were found to be the difference in the number of shots on target and the difference in the number of crosses between the home-team and away-team, respectively.

Regression models have been used in sports results prediction and feature selection analysis. Mackay (2017) applied Ridge regression with a sliding window approach to compute the probability of a possession becoming a goal in English Premiere League (EPL). Magel and Melnykov (2014) developed least square regression models to predict the point spread of a soccer game and used logistic regression to predict the winner of games during 2011-2012 season from the three top European soccer leagues: EPL (England), La Liga (Spain), and Serie A (Italy). Their model successfully predicted the winner of the games with 73% to 80% accuracy, and t-test

analysis showed that there was a significant difference between the number of cards received by home-teams and the number of cards received by away-teams.

The home-field effect is a popular feature that favors the home-team. Carmichael and Thomas (2005) employed regression analysis and showed that the home advantage was an essential factor in predicting the outcome by analyzing EPL games. Goddard (2006) used statistical analysis on the dataset of 35 seasons (from1970/1971 to 2004/2005) of EPL and Football League, and they concluded that the magnitude of the home-field effect was dependent on the geographical distance the away-team had to travel.

To improve prediction performance, some studies focused on developing effective methods for selecting predictor variables. Hucaljuk and Rakipovic (2011) developed a software that assigns a quantitative value to the features and later selects based on these values the necessary features that must be taken into account in predicting the outcome of games. After determining the optimal combination of features, they applied six different machine learning algorithms (naive Bayes, Bayesian network, logit boost, the k-nearest neighbor algorithm, RF, and artificial neural network (ANN)) to predict the game outcomes of Europe Champions League. ANN model achieved 68.8% accuracy for three outcomes prediction (win, loss, draw) and surpassed the other 5 models in terms of accuracy. Berrar et al. (2019) introduced two novel approaches for modeling process and compared their role in improving the performance of gradient boosted trees (XGBoost) and a k-nearest neighbor (k-NN) model. Zambom-Ferraresi et al. (2018) used Bayesian model averaging (BMA) to analyze the relative importance of possible determinants of football performance during 2012/13–2014/15 seasons in the Europe 'Big Five' leagues (EPL, Bundesliga, La Liga, Serie A, and Ligue One) and found that the number of saves made by goal keeper was a determinant that had been ignored before. Some of the studies used pre-play data (statistics from previous games)

to predict future matches. Hubacek et al. (2019) used Gradient Boosted Trees to predict future matches within a selected timeframe from leagues around the world. (Lock and Nettleton, 2014) estimated the win probability before each play in a game of a National Football League (NFL).

Some studies predicted game results from other perspectives. Shin and Gasparyan (2014) predicted the game results with virtual data collected from a videogame (FIFA 2015) and compared the performance with a model that used real data model. They found that it was effective to use virtual data in predicting games result. Eggels et al. (2016) predicted the game results by estimating the probability of a goal scored. Mackay (2017) focused on computing the goal probability of a possession in EPL games for the season 2016/2017.

ANN is a popular machine learning algorithm in predicting the outcome of sport games. Kahn (2003) trained the structure of ANN with the first 192 matches in the 2003 season of NFL and used the last two rounds (weeks 14 and 15) as the test data. The optimal structure they got was 10-3-2 (10 nodes for input layer, 3 nodes for hidden layer, 2 nodes for output layer). The best accuracy for the two outcomes prediction was found to be 75 %. Bunker and Thabtah (2017) proposed a sport result prediction framework using an ANN. McCabe and Trevathan (2008) used multilayer perceptron (MLP), which was a class of feed forward ANN to predict game results of four rugby and football teams. The highest average accuracy of 67.5% was achieved when predicting the outcomes (win, loss, draw) of three-season Super Rugby games. The lowest average accuracy of 54.6% was achieved when predicting three-season EPL games.

RF is another popular machine learning model implemented to predict the outcome of games. By using a RF model, Lock and Nettleton (2014) combined pre-play variables to estimate win probability (WP) of leagues in NFL. Ulmer et al. (2013) applied five machine learning models: Linear from stochastic gradient descent, naive Bayes, hidden Markov model, SVM, RF to predict

the game outcome (win, loss, and draw) of the soccer matches of EPL. They collected the historical data about the soccer matches from a website called Football-Data (https://www.football-data.org/). The classification error rates were 0.48 (linear classifier), 0.5 (RF) and 0.5 (SVM) for the three outcomes prediction. Tax and Joustra (2015) compared the performance of nine machine learning algorithms: Continuous High-resolution Image Reconstruction using Patch priors (CHIRP), Decision Table Naive Bayes Hybrid Classifier (DTNB), Fuzzy Un-ordered Rule Induction Algorithm (FURIA), Hyper-Pipes, J48, Naive Bayes, MLP, RF and Logit Boost in predicting the game results of Dutch Eredivisie. The highest accuracy was achieved by FURIA with an accuracy of 55.297%. Baboota and Kaur (2019) applied Gaussian Naive Bayes, SVM, RF, Gradient Boosting to build a generalized predictive model for EPL game results. The best model was Gradient Boosting, followed by RF with an accuracy of 0.57. Yezus (2014) tested the ability of machine learning models in predicting the games with good precision and found RF had the best precision of 0.634.

Different from other studies, in this study the machine learning models were not only applied to the combined dataset of the three leagues but also applied separately to each league's data. In this way, the prediction performance and critical factors affecting the game results were investigated for different leagues and then compared. Carmichael and Thomas (2005) demonstrated the existence of home-field advantage in EPL, and Goddard (2006) further showed that the magnitude of the home-field advantage was related to the geographic distance the away team had to travel. The United States is a geographically huge country with different time zones from the east to the west coasts. This study aimed to provide further evidence that the home-field advantage of MLS is more evident than the European teams. In addition to the home-team and away-team factor, all features from the team match statistics and situation report of each game were collected. The

features included not only the noticeable features like the number of shots on target and possession rate, which directly affect the game results but also indirect features like the number of tackles, passes, crosses, cards.

Table 22 summarizes the contributions of soccer related studies in terms of focus and methods employed.

Table 24. Literature review

ID	Citation	Focus	Method
1	Hubacek et al. [2019]	Predict outcomes of future matches within a selected time-frame from different leagues over the world	Gradient boosted trees
2	Berrar et al. [2019]	Present two novel ideas for integrating soccer domain knowledge into the modeling process	Extreme Gradient Boosting (XGBoot) K nearest neighbors
3	Baboota and Kaur [2019]	Build a generalized predictive model for predicting the results of the English Premier League	Gaussian naive Bayes Support Vector Machine Random Forest Gradient Boosting
4	Zambom-Ferraresi et al. [2018]	Analyze the relative importance of possible determinants of football performance during the period 2012/13–2014/15 for the Europe 'Big Five' leagues (the English Premier League, the German Bundesliga, the Spanish Liga, the Serie A Italian Calcio, and the French Ligue One)	Bayesian model averaging (BMA) Statistical Analysis
5	Bunker and Thabtah [2017]	Analyze some recent research on sport prediction that have used ANN and propose a sport result prediction framework for the complex problem of sport result prediction	Neural Network
6	Mackay [2017]	Compute the probability a possession becomes a goal for the English Premier League for the season 2016/2017	Ridge Regression with a sliding window approach
7	Eggels [2016]	Predict the soccer game results by estimating the probability of scoring for the individual goal	Logistic Regression Decision Tree Random Forest Adaptive Boosting

8	Kerr [2015]	1.Predict soccer game result by ball-event features 2. Investigate the feature weights to learn relationships between game events and a team's chances of success. 3. Recognizing a team based on their style of play 4. Recognizing a team base on a random sample of passes made by a single team	L2-regularized logistic regression (Ridge Regression)
9	Tax and Joustra [2015]	Evaluate the performance of different machine learning algorithms in predicting the match results of Dutch Eredivisie	Continuous High-resolution Image Reconstruction using Patch priors (CHIRP) Decision Table Naïve Bayes Hybrid Classifier (DTNB) Fuzzy Unordered Rule Induction Algorithm HyperPipes J48 NaiveBayes Neural Network (Multilayer Perceptron) RandomForest LogitBoost
10	Magel and Melnykov [2014]	1. Use statistical models to estimate game results of three top European soccer leagues: England, Spain, and Italy games 2. Investigate the influence of cards given in games on the game results with statistic analysis	Least squares regression Logistic regression
11	Lock and Nettleton [2014]	Estimate win probability before each play of an National Football League game with preplay variables	Random Forest

12	Yezus [2014]	Test the ability of machine learning models in predicting the outcome of soccer games with good precision.	K nearest neighbors Random Forest
13	Shin and Gasparyan [2014]	Compare the performance of soccer game result prediction with virtual data collected from a video game(FIFA 2015) and with real data	Linear Support Vector Machine Logistic Regression
14	Ulmer et al. [2013]	Predict soccer match results in the English Premier League with different machine learning algorithms.	Linear from stochastic gradient descent, Naive Bayes, Hidden Markov Model, Support Vector Machine, Random Forest
15	Hucaljuk and Rakipovic [2011]	Predicting football scores using machine learning techniques with features selected by a software solusion	LogitBoost Neural Network Random Forest Bayesian Network k Nearest Neighbor Naive Bayes
16	McCabe and Trevathan [2008]	Predict game results of four rugby and football teams with artificial intelligence	Neural Network (Multilayer Perceptron)
17	Goddard [2006]	Investigate the home-field advantage and prove that magnitude of the home-field effect is dependent on the geographical distance the away-team has to travel (EPL and football)	Statistical Analysis
18	Carmichael and Thomas [2005]	Provides further evidence from English Premier League games regarding the existence of home-field advantage	Regression Analysis
19	Kahn [2003]	Prediction of NFL Football Games with Neural Network	Neural Network

3.3. Data Preparation

The data was scraped from whoscored.com (https://www.whoscored.com/) and consisted of game statistics from five seasons (2014/2015 – 2018/2019) in three soccer leagues (La Liga, EPL, and MLS). Four datasets were created (Table 23): the first 3 datasets corresponded to the 3 leagues respectively and the fourth dataset contained all the three leagues' game statistics data, which was labeled as ALL.

The response variable of a game was binary (win:1, or lose:0) as used in (Kerr, 2015). All tie-games were removed from the dataset to increase the prediction accuracy and reliability of machine learning models. For each game, one of the team was randomly designated as team A and the other team was designated as team B (Kerr, 2015). If the team A was the winner, the outcome was labeled as 1. Otherwise, the outcome was labeled as 0. After eliminating the tie games, the total number of games for each dataset is shown in Table 23.

Table 25. Number of games for each dataset

Dataset	League	Number of games
1	La Liga	1424
2	EPL	1442
3	MLS	1474
4	ALL	4340

Table 24 describes the predictor variables. The home-team/away-team was coded as 1 if team A was the home-team, and 0 otherwise. Except for the home-team/away-team, other features were collected from the match report of each game. The predictor variables were categorized into four classes on the website: 1) team match statistics, 2) attempt types, 3) card situations, and 4) pass types, which are depicted in Table 24. Except for the home-team/away-team variable, the other 26 variables were collected for both the home-team and away-team. As performed by (Kerr, 2015),

the difference between the home-team and away-team was calculated for each variable (feature) in our dataset. The difference was calculated as the value of the home-team minus the value of the away-team. Therefore, including the home-team/away-team variable, our final dataset contained 26×3+1=79 variables. Among the 79 variables, 22 variables belonged to the team match statistics, 15 variables were related to attempt types, 24 variables were related card situations type and the remaining 18 variables belonged to pass types. As shown in Figure 22, a total of 79 variables were standardized. In the experiments, 20% of the dataset was sampled as test data and the rest 80% was taken as training data. Only the training data was used to tune the machine learning models.

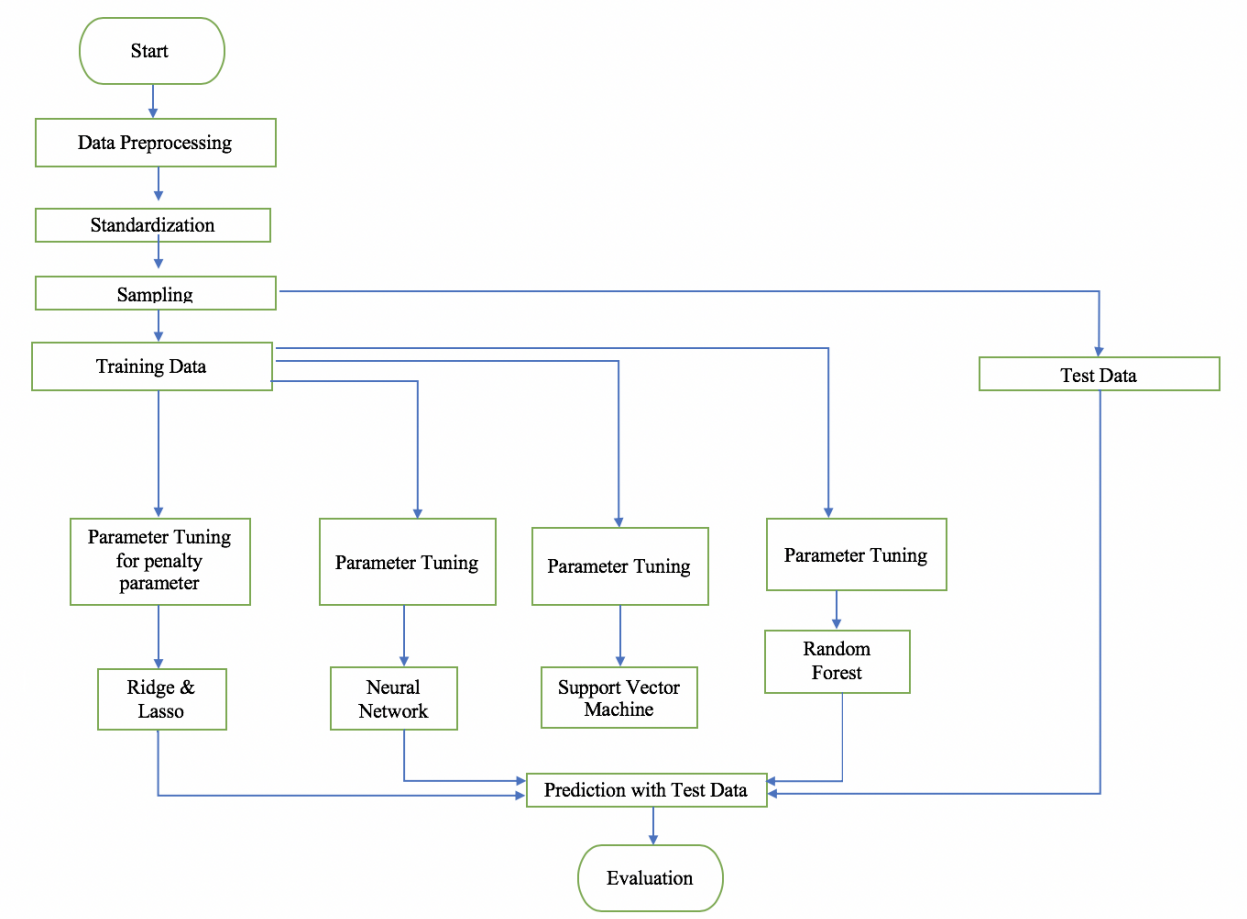


Figure 22. The flow chart of data processing approach

Table 26. Predictor variables of the soccer game data

Feature types	Feature	Description
Team match statistics	home-team /away-team	1 if team A plays as home team, 0 otherwise
	shots	number of shots
	shots on target	number of shots on target
	pass success rate	percentage of passes that succeed
	aerial duel success rate	percentage of aerial duel that succeed
	dribbles won	number of dribbles won
	tackles	number of tackles
	possession rate	percentage of time a team controlled the ball
Attempt types	open play	number of open plays
	set-piece	number of set-piece
	counter attack	number of counter attacks
	penalty	number of penalty
	own goal	number of own goal
Card situations	red cards	number of red cards a team got
	yellow cards	number of yellow cards a team got
	cards for fouls	number of cards received for foul
	cards for unprofessional	number of cards received for unprofessional
	cards for dive	number of cards received for dive
	cards for other reason	number of cards received for other reason
	cards per foul rate	cards per foul
	fouls per game	number of fouls
Pass types	passes	number of passes
	crosses	number of crosses
	through balls	number of through balls
	long balls	number of long balls
	short passes	number of short passes
	average pass streak	average number of pass streak

3.4 Methodology

This section introduces the proposed five machine learning models (Ridge, Lasso, NN, SVM, RF) and explains how the performance of the models was measured.

3.4.1 Machine Learning Models

The machine learning models (Ridge, Lasso, NN, SVM, RF) used in this study were the same as the crash severity prediction study. See the explanations of them in detail in Section 2.4.1 to Section 2.4.5.

3.4.2 Confusion Matrix

The confusion matrix, shown in Table 25, is a visualization tool that displays the performance of a classification model. The definitions of TP, FP, TN, FN are as follow:

T P: number of samples predicted as true while their actual values were true

F P: number of samples classified as true while their actual values were false

T N: number of samples classified as false while their actual values were true

F N: number of samples classified as false while their actual values were false

Three criteria to compare the performance of models: accuracy, sensitivity, and specificity are calculated from the confusion matrix by using the equations 4, 5, and 6, respectively:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{5}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{6}$$

Accuracy represents the probability of correct classification; that is, the predicted class is the same as the actual class. Sensitivity, which is defined as the true positive rate, illustrates the probability of the correct classification of an actual positive (Delen et al., 2017). It measures the model's ability when classifying the positive class. Relatively, specificity measures the ability to predict

the negative class, which illustrates the probability of the correct classification of an actual negative.

Table 27. Confusion matrix

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Negative (TN)	False Positive (FP)
	Negative	False Negative (FN)	True Positive (TP)

3.5 Parameter Tuning

This section explains the parameter tuning process used to heighten the prediction performance. In fact, parameter tuning is the most essential step in improving the prediction results. For each dataset, 80% of the data was sampled as training data to train the machine learning models. Because python was used in this study, the tuning process was slightly different from the crash study.

As mentioned above, the parameters to be tuned for RF are 1) the number of trees in the forest (n_estimators), 2) the maximum depth of each tree in the forest (max depth), 3) the minimum number of samples for each node that can be split (min samples split), 4) the number of features to consider when splitting each node (max features), 5) the minimum number of samples for each leaf node (min samples leaf). The grid search method was used to tune the parameters with a scoring metric of accuracy. The tuning range of each parameter is shown in Table 26. The parameters of the optimum RF models for the four datasets are shown in Table 27.

Table 28. Tuning range of parameters for RF

Parameter	Range
n_estimators	50,60,70,80,…,200
max_depth	3,5,7,9,11,13
min_samples_split	2,3,4,…,20
max_features	2,4,6,8,10,12,14
min_samples_leaf	1,2,3,…,15

Table 29. Optimum RF model for each dataset

Parameter	La Liga	EPL	MLS	ALL
n_estimators	190	190	130	170
max_depth	13	11	13	13
min_sample_split	4	12	3	6
max_features	14	12	14	14
min_samples_leaf	4	1	2	7

For NN, the tuning parameters are 1) the structure of the layers, which includes the number of layers and the number of units in each layer, 2) the dropout parameter, 3) the number of epochs, which represents the number of times the entire dataset pass forward and back through the NN model. Different sets of neurons in input layer and hidden layer were dropped out when training the model to reduce overfitting. The dropout parameter indicates the proportion of neurons to be dropped in the layer. The activation used for the input and hidden layers was ”ReLU”. For the output layer, the activation was ”sigmoid”. The criterion used for evaluation was ”binary cross-entropy”.

In order to obtain more stable results, stratified 5-fold cross-validation was used to train each NN model. The prediction criteria of NN were the average of five folds, and a standard deviation of the five folds was calculated to show the stability of the results.

Different NN models with different parameters were fitted for each dataset to find the best structure. The dropout parameter was 0.5 to reduce overfitting. The tuning range of each parameter is shown

in Table 28. The epoch vs. accuracy plot for the last fold validation for each dataset is shown in Figure 23, Figure 24, Figure 25, Figure 26. The values of epoch and dropout that minimized the overfitting in the plots were selected. The best NN model for each dataset is shown in Table 29.

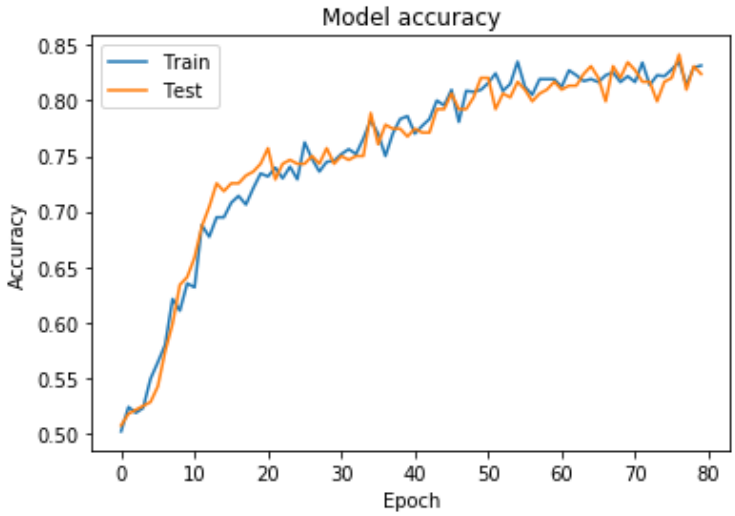


Figure 23. Epoch vs accuracy plot for the last fold validation of La Liga

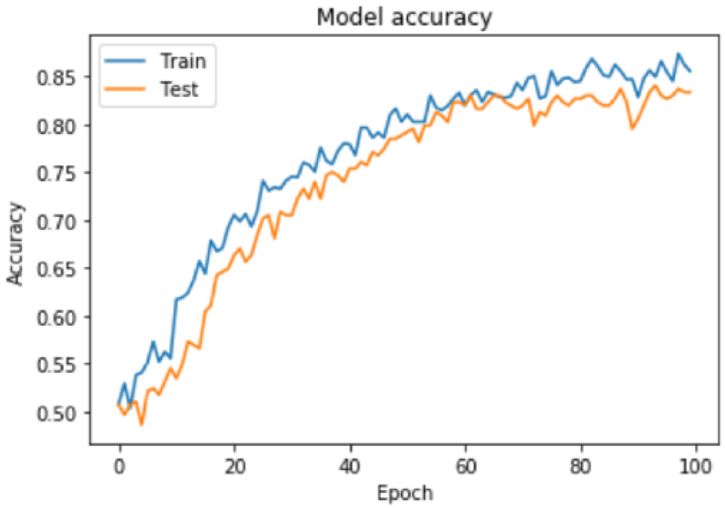


Figure 24. Epoch vs accuracy plot for the last fold validation of EPL

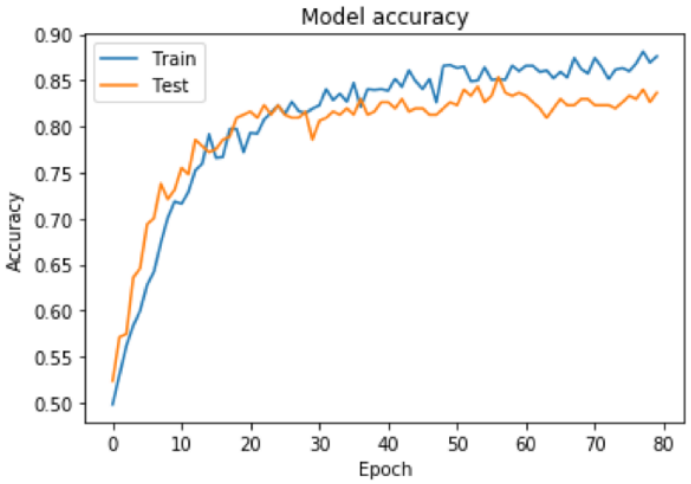


Figure 25. Epoch vs accuracy plot for the last fold validation of MLS

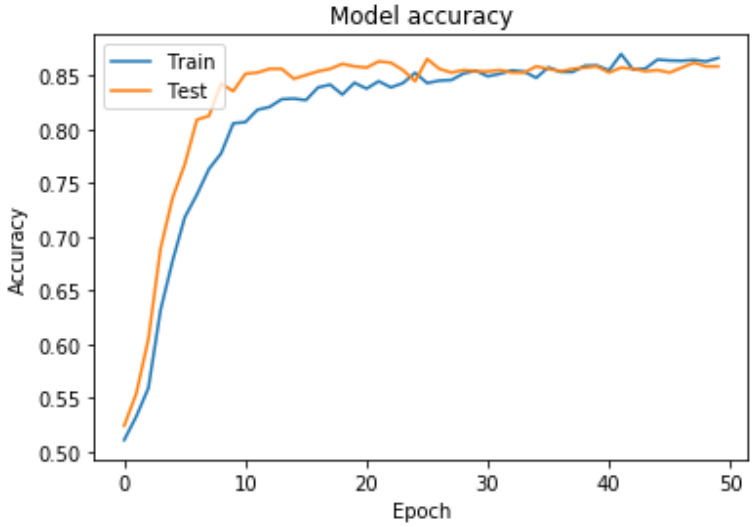


Figure 26. Epoch vs accuracy plot for the last fold validation of ALL

Table 30. Parameters for NN

Parameter	**Model**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Number of hidden layer	2	1	1	1	1	1	1	1	1	1
Nodes for input layer	160	50	40	40	30	20	20	18	18	12
Nodes for hidden layer	50	20	30	30	10	15	10	9	9	12
Nodes for hidden layer	10	--	--	--	--	--	--	--	--	--
Nodes for output layer	1	1	1	1	1	1	1	1	1	1
Dropout	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Epoch	100	100	100	50	100	100	100	100	80	100

Table 31. Optimum NN model for each dataset

Parameter	La Liga	EPL	MLS	ALL
Nodes for input layer	18	30	18	40
Nodes for hidden layer	9	10	9	30
Nodes for output layer	1	1	1	1
Dropout	0.5	0.5	0.5	0.5
Epoch	80	100	80	50

For SVM, two kernel functions, linear and RBF, were tested. The one that had the best performance was chosen (James et al., 2013). The grid search method was applied to tune the parameters with a scoring metric of accuracy. The tuning range is shown in Table 30. The parameters of the optimum SVM models for the four datasets are shown in Table 31.

Table 32. Parameters for SVM

Kernel	Parameter	Range
linear	C	1,10,100,1000
RBF	C	1,10,100,1000,10000
	gamma	0.5,0.1,0.01,0.001,0.0001

Table 33. Optimum SVM model for each dataset

Parameter	La Liga	EPL	MLS	ALL
kernel	RBF	RBF	RBF	RBF
C	100	10000	100	1000
gamma	0.001	0.001	0.001	0.001

For both Ridge and Lasso regression, the tuning parameter is the penalty parameter (alpha). Five-fold cross-validation was employed to tune the parameter for the two models with a scoring metric of “negative mean squared error”. The best penalty parameters for the four datasets are shown in Table 32.

Table 34. Best penalty parameters of Ridge and Lasso for each dataset

Model	La Liga	EPL	MLS	ALL
Ridge	1	0.76	7.05	0.25
Lasso	0.001	0.001	0.0007	0.0008

Figure 27 shows the classification mean square error with the penalty parameter of Ridge for La Liga. The values of penalty parameter that minimized the mean square error were selected. The plots for the other datasets are shown in Figure 28, Figure 29, Figure 30.

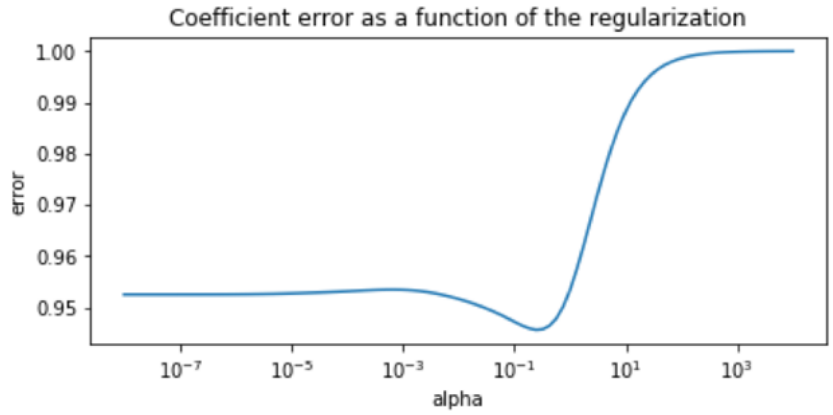


Figure 27. Penalty parameter vs MSE plot for Ridge of La Liga

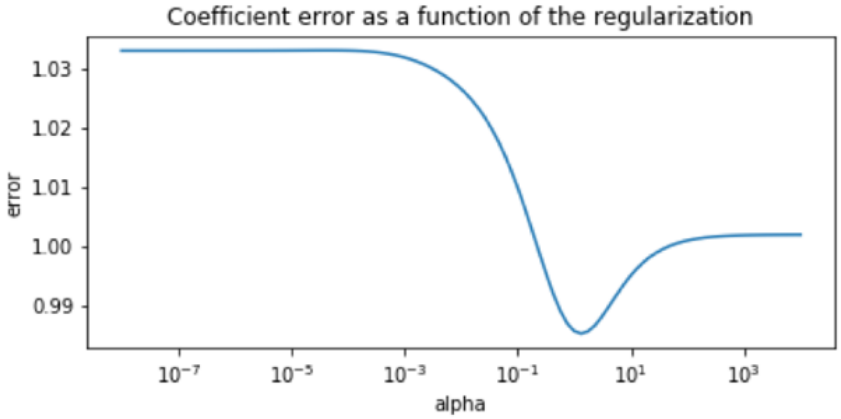


Figure 28. Penalty parameter vs MSE plot for Ridge of EPL

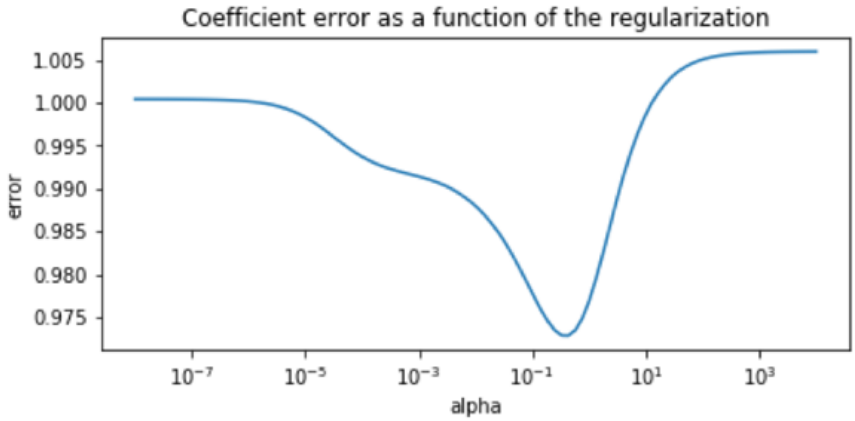


Figure 29. Penalty parameter vs MSE plot for Ridge of MLS

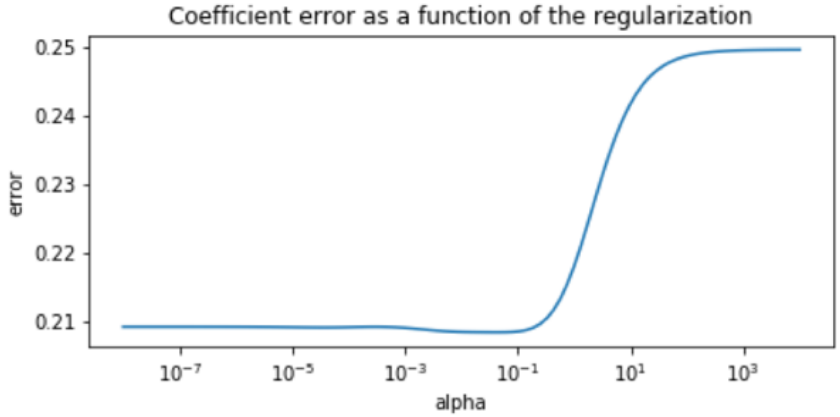


Figure 30. Penalty parameter vs MSE plot for Ridge of ALL

In a similar vein, the plots for Lasso are shown in Figure 31, Figure 32, Figure 33, Figure 34.

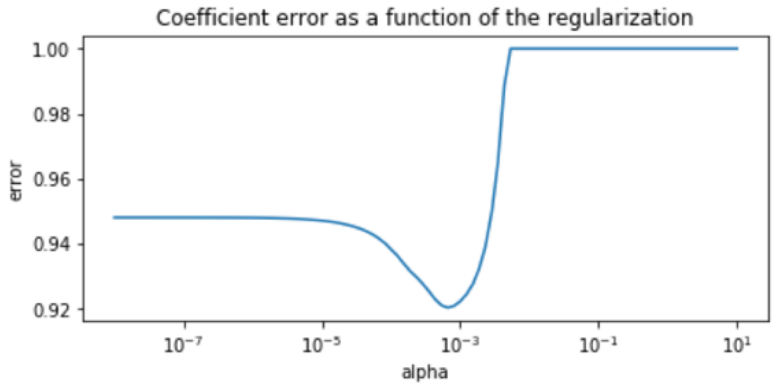


Figure 31. Penalty parameter vs MSE plot for Lasso of La Liga

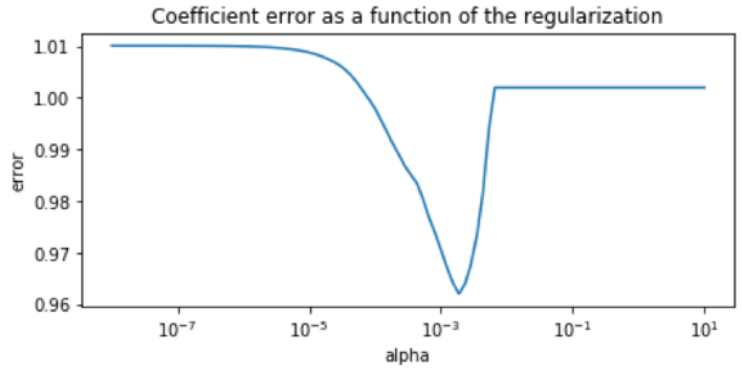


Figure 32. Penalty parameter vs MSE plot for Lasso of EPL

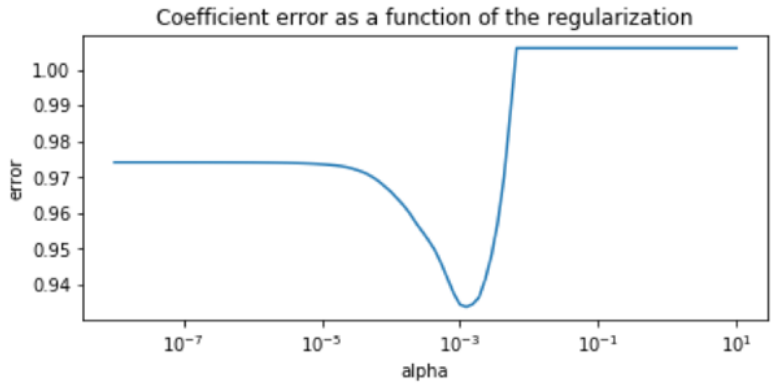


Figure 33. Penalty parameter vs MSE plot for Lasso of MLS

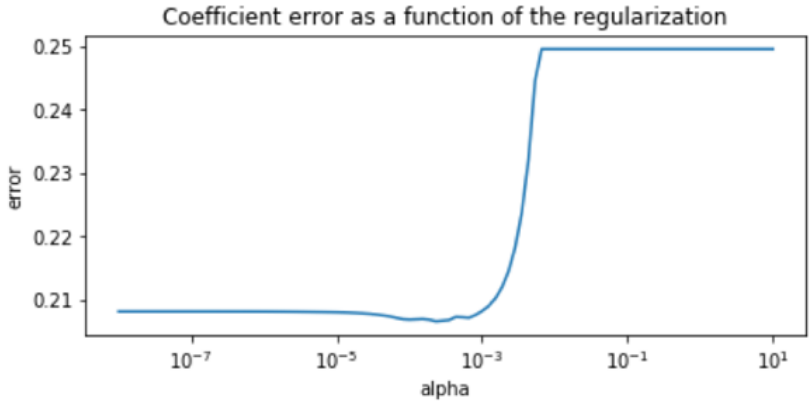


Figure 34. Penalty parameter vs MSE plot for Lasso of ALL

3.6 Results

Results are explained in twofold: 1) prediction performance, 2) feature importance of predicted variable.

3.6.1 Prediction Performance

Table 33 shows the prediction accuracy results of the five models for each dataset. For the dataset containing the data of all three leagues, NN had better performance than other models with an accuracy of 85.71% $\pm$ 0.73%, here 85.71% was the average prediction accuracy and 0.73% was the standard deviation of the five-folds. As mentioned in section 3.5, the prediction criteria of NN were the average of five folds, and the standard deviation was calculated to show the stability of the results. The standard deviation was 0.73%, which means that the accuracy result was reliable and stable. Although with an accuracy of 84.68% RF ranked in the second place, it was the best model for the other three datasets. The best performance of La Liga, EPL and MLS datasets was 83.86%, 83.39%, 84.07%, respectively.

Table 35. Accuracy results

	RF	NN	SVM	Ridge	Lasso
Laliga	83.86%	82.59% (+/- 1.84%)	65.96%	58.25%	65.26%
EPL	83.39%	82.39% (+/- 1.50%)	70.24%	58.13%	60.55%
MLS	84.07%	82.63% (+/- 3.42%)	74.58%	71.19%	70.85%
ALL	84.68%	85.71% (+/- 0.73%)	78.46%	64.40%	64.40%

Figure 35 depicts the results as a column chart. As shown in the column chart, the accuracy results of NN and RF were more stable and better than the other three algorithms in predicting the game outcomes in all datasets. Notably, the accuracy results of Ridge and Lasso regression varied by more than 10% depending on the dataset. Ridge predicted MLS with an accuracy of 72.89%, 17.52% higher than the La Liga result of 55.37%.

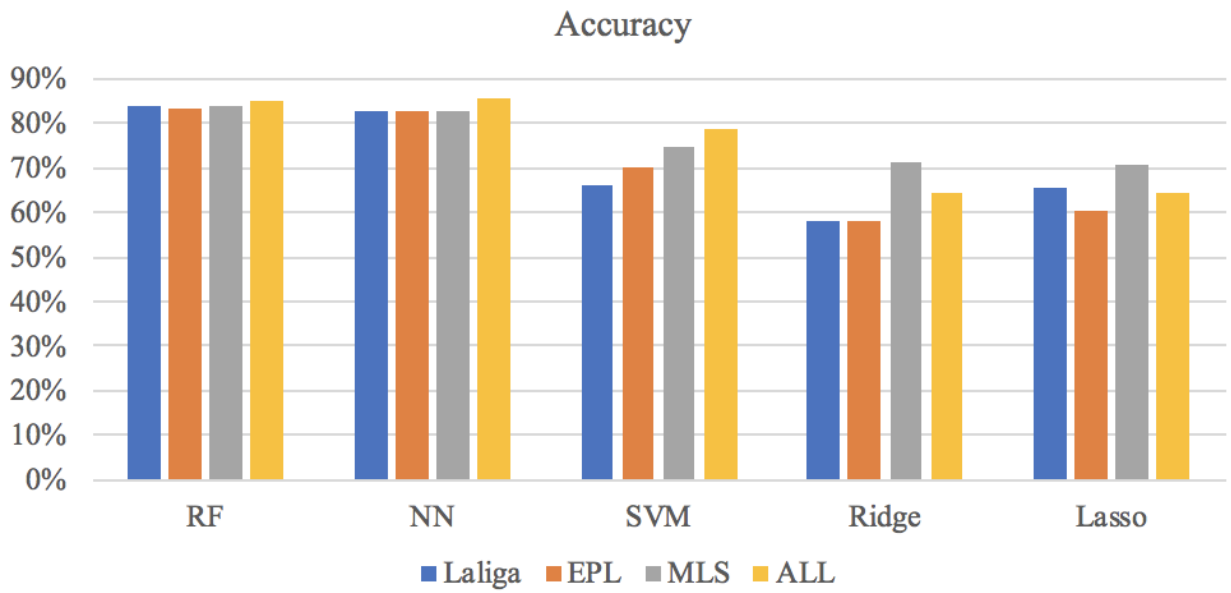


Figure 35. Comparison of the accuracy results

Table 34 and Table 35 show the sensitivity and specificity results of the five models for each dataset. The sensitivity measures a model's ability when predicting the positive class of the

response variable, while specificity measures the ability to predict the negative class. Positive class of the response variable indicates team A wins, and the negative class indciates team A loses. RF was the best model for the EPL and MLS datasets when predicting positive class. Although the sensitivity result of NN were slightly higher than RF when predicting La Liga and ALL datasets, the results were observed to possess significant standard deviations. Thus, RF was still the best performing machine learning approach.

Table 36. Sensitivity results

	RF	NN	SVM	Ridge	Lasso
Laliga	81.63%	82.69% (+/- 4.46%)	61.90%	46.26%	67.35%
EPL	86.23%	82.62% (+/- 1.26%)	68.12%	64.49%	63.77%
MLS	79.10%	75.99% (+/- 8.81%)	71.64%	66.42%	70.90%
ALL	85.08%	85.73% (+/- 2.51%)	78.55%	66.43%	65.50%

Table 37. Specificity results

	RF	NN	SVM	Ridge	Lasso
Laliga	86.23%	83.98% (+/- 4.42%)	70.29%	71.01%	63.04%
EPL	80.79%	81.59% (+/- 2.24%)	72.19%	52.32%	57.62%
MLS	88.20%	88.52% (+/- 3.99%)	77.02%	75.16%	70.81%
ALL	84.28%	84.74% (+/- 1.58%)	78.36%	62.41%	63.33%

Similarly, although the specificity results of NN were slightly higher than RF when predicting EPL and ALL datasets, the results had large standard deviations. Thus, RF was still the best model considering the specificity results. Figure 36 shows the sensitivity results, and Figure 37 displays the specificity results. The patterns of the plots were the same as the accuracy results. For both sensitivity and specificity results, NN and RF had more stable and better results than the other three algorithms when dealing with different datasets.

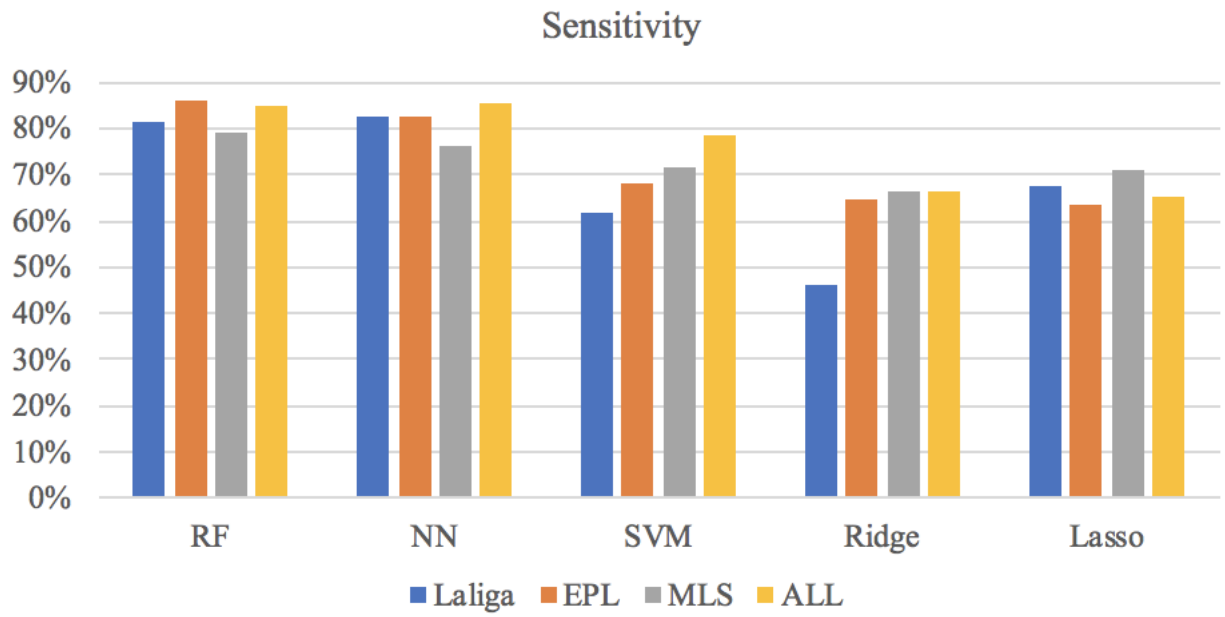


Figure 36. Comparison of the Sensitivity results

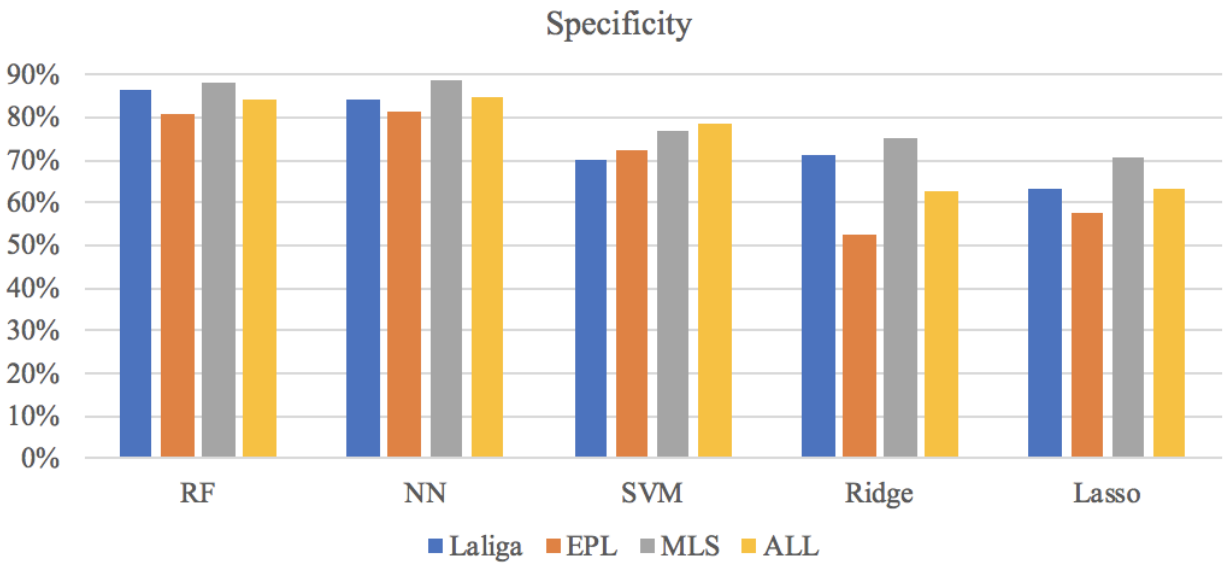


Figure 37. Comparison of the Specificity results

3.6.2 Feature Importance

As discussed above, RF had the best prediction performance among the proposed machine learning approaches. It also provided the importance of each predictor in predicting the outcome of soccer

games. All predictors had positive importance. The relative weight of the *i* th (*i=1,2,…n*) predictor was calculated as follows: Relative weight$_i$ = $\frac{weight_i}{\sum_{j=1}^{79} weight_j}$, wehre 79 was the number of predictors. If each predictor had the same contribution to the prediction, then the average weight of each factor would be 1÷79≈1.27%. Therefore, a predictor was considered important when its relative weight was higher than 1.27%.

Table 36, Table 37, Table 38, and Table 39 present the important variables in predicting game outcomes for all four datasets. The variables were grouped into two categories: obvious variable and non-obvious variable as practiced in Kerr (2015). A total of 10 variables included the home-team/away-team factor, and the variables related to the number of shots on target, shots, cross were considered as the obvious variables in parallel with the previous literature, and the remaining 69 variables were grouped as the non-obvious variables.

For the obvious variables, the difference in the number of shots on target was the most important predictor variable across the datasets except for the MLS dataset. The home-team/away-team factor was the most influential variable for the MLS dataset with a relative weight of 11.30%, 3.75% more than the difference in the number of shots on target. In addition to these two most important variables, the difference in the number of crosses was found to be as the third important variable for all three leagues. The variables that related to the number of shots had the lowest relative weight relative to other obvious variables. While all the other obvious variables were important variables for all leagues, the variables related to the number of shots were not always important. The relative weights of the three variables related to the number of shots were lower than the most important non-obvious variables for the four datasets.

Table 38. Relative weights of the predictor variables in La Liga

Obvious Variable	Relative Weight	Non-obvious Variable	Relative Weight
difference in the number of shots on target	10.91%	difference in the rate of cards per foul	1.93%
home-team /away-team	5.92%	away team tackles	1.84%
difference in the number of crosses	5.18%	difference in the rate of aerial duel success	1.71%
away team shots on target	4.57%	away long balls	1.68%
home crosses	4.25%	away short passes	1.64%
home team shots on target	3.43%	difference in the number of tackles	1.59%
away crosses	2.58%	home short passes	1.57%
difference in the number of shots	1.48%	difference in the number of through balls	1.56%
away team shots	1.47%	difference in the number of counter attacks	1.55%
		difference in the number of passes	1.53%
		away passes	1.53%
		difference in the number of open plays	1.47%
		away team aerial duel success rate	1.46%
		home cards per foul rate	1.46%
		difference in the number of short passes	1.40%
		home team aerial duel success rate	1.38%
		difference in the number of long balls	1.33%
		away cards per foul rate	1.33%
		home passes	1.30%
		home team tackles	1.28%
		away team pass success rate	1.28%

Table 39. Relative weights of the predictor variables in EPL

Obvious Variable	Relative Weight	Non-obvious Variable	Relative Weight
difference in the number of shots on target	7.30%	away short passes	2.11%
home-team /away-team	6.02%	home team pass success rate	1.99%
difference in the number of crosses	4.46%	difference in the number of shots	1.94%
home crosses	3.80%	away passes	1.92%
home team shots on target	3.71%	home short passes	1.91%
away crosses	3.36%	difference in the number of tackles	1.88%
away team shots on target	3.11%	difference in the number of through balls	1.79%
away team shots	1.59%	away team open plays	1.75%
home team shots	1.36%	home team tackles	1.72%
		difference in the number of open plays	1.72%
		difference in the number of long balls	1.68%
		home long balls	1.66%
		away team pass success rate	1.62%
		difference in the number of counter attacks	1.62%
		away team tackles	1.60%
		home passes	1.58%
		difference in the rate of cards per foul	1.50%
		difference in the number of dribbles won	1.44%
		difference in the number of passes	1.34%
		away long balls	1.33%
		difference in the number of short passes	1.32%
		difference in the rate of pass success	1.31%

Table 40. Relative weights of the predictor variables in MLS

Obvious Variable	Relative Weight	Non-obvious Variable	Relative Weight
home-team /away-team	11.30%	difference in the number of long balls	1.93%
difference in the number of shots on target	7.55%	away long balls	1.84%
difference in the number of crosses	6.68%	away passes	1.68%
home crosses	4.75%	away short passes	1.67%
away team shots on target	3.33%	difference in the number of tackles	1.62%
away crosses	3.13%	home short passes	1.61%
home team shots on target	2.67%	difference in the number of through balls	1.60%
difference in the number of shots	1.41%	home team tackles	1.56%
home team shots	1.34%	difference in the rate of cards per foul	1.41%
		home passes	1.41%
		home long balls	1.39%
		home team pass success rate	1.38%
		difference in the number of set-piece	1.35%
		home cards per foul rate	1.35%
		away team tackles	1.30%
		home team open plays	1.27%

Table 41. Relative weights of the predictor variables in all leagues together

Obvious Variable	Relative Weight	Non-obvious Variable	Relative Weight
difference in the number of shots on target	13.46%	difference in the number of counter attacks	1.93%
home-team /away-team	10.69%	difference in the number of through balls	1.73%
difference in the number of crosses	8.42%	away short passes	1.56%
home crosses	5.24%	away passes	1.44%
away team shots on target	4.71%	home short passes	1.41%
home team shots on target	4.31%	difference in the number of open plays	1.32%
away crosses	3.21%		
difference in the number of shots	1.32%		

In terms of the non-obvious variables, the most important variables varied across the four datasets. The difference in the rate of cards per foul had the highest relative weight of 1.93% in the La Liga dataset. The number of shots of away-team was the most important variable affecting the outcome of EPL matches of the relative weight of 2.11%. The difference in the number of long balls and the number of long balls of away-team were the top two most important variables affecting the MLS matches with relative weights of 1.93%, 1.84%. For the ALL dataset, which was the combined data of three leagues, the difference in the number of counter attacks was the most influential variable with a relative weight of 1.93%.

Therefore, whether the team was the home-team or away-team had the most significant influence on the outcome of the MLS matches. Although the home-team and away-team factor also influenced the outcome of La Liga and EPL, the degree of influence was not as high as that of MLS. La Liga and EPL games were mainly affected by the difference in the number of shots on target between home-team and away-team. In general, the home-team/away-team factor, the difference in the number of shots on target, the difference in the number of crosses were the three most important factors affecting the outcome of soccer matches. The number of shots was not one of the most important predictors of the outcome of soccer games. The most important non-obvious variables that affected the outcome varied from league to league.

As described in Section 3.3, the features included have four types. Table 40, Table 41, Table 42 and Table 43 list the important features of each dataset by feature types. A total of 22 variables belonged to the team match statistics type, 15 variables were attempt types, card situations type had 24 variables, and the remaining 18 variables were pass types.

Table 42. Different types of important variables in the La Liga dataset

Team match statistics	Relative Weight	Pass types	Relative Weight
difference in the number of shots on target	10.91%	difference in the number of crosses	5.18%
home-team /away-team	5.92%	home crosses	4.25%
away team shots on target	4.57%	away crosses	2.58%
home team shots on target	3.43%	away long balls	1.68%
away team tackles	1.84%	away short passes	1.64%
difference in the rate of aerial duel success	1.71%	home short passes	1.57%
difference in the number of tackles	1.59%	difference in the number of through balls	1.56%
difference in the number of shots	1.48%	difference in the number of passes	1.53%
away team shots	1.47%	away passes	1.53%
away team aerial duel success rate	1.46%	difference in the number of short passes	1.40%
home team aerial duel success rate	1.38%	difference in the number of long balls	1.33%
home team tackles	1.28%	home passes	1.30%
away team pass success rate	1.28%	**Card situation**	**Relative Weight**
Attempt types	**Relative Weight**	difference in the rate of cards per foul	1.93%
difference in the number of counter attacks	1.55%	home cards per foul rate	1.46%
difference in the number of open plays	1.47%	away cards per foul rate	1.33%

Table 43. Different types of important variables in the EPL dataset

Team match statistics	Relative Weight	Pass types	Relative Weight
difference in the number of shots on target	7.30%	difference in the number of crosses	4.46%
home team /away team	6.02%	home crosses	3.80%
home team shots on target	3.71%	away crosses	3.36%
away team shots on target	3.11%	away short passes	2.11%
home team pass success rate	1.99%	away passes	1.92%
difference in the number of shots	1.94%	home short passes	1.91%
difference in the number of tackles	1.88%	difference in the number of through balls	1.79%
home team tackles	1.72%	difference in the number of long balls	1.68%
away team pass success rate	1.62%	home long balls	1.66%
away team tackles	1.60%	home passes	1.58%
away team shots	1.59%	difference in the number of passes	1.34%
difference in the number of dribbles won	1.44%	away long balls	1.33%
home team shots	1.36%	difference in the number of short passes	1.32%
difference in the rate of pass success	1.31%	**Attempt types**	**Relative Weight**
Card situation	**Relative Weight**	away team open plays	1.75%
difference in the rate of cards per foul	1.50%	difference in the number of open plays	1.72%
		difference in the number of counter attacks	1.62%

Table 44. Different types of important variables in the MLS dataset

Team match statistics	Relative Weight	Pass types	Relative Weight
home-team /away-team	11.30%	difference in the number of crosses	6.68%
difference in the number of shots on target	7.55%	home crosses	4.75%
away team shots on target	3.33%	away crosses	3.13%
home team shots on target	2.67%	difference in the number of long balls	1.93%
difference in the number of tackles	1.62%	away long balls	1.84%
home team tackles	1.56%	away passes	1.68%
difference in the number of shots	1.41%	away short passes	1.67%
home team pass success rate	1.38%	home short passes	1.61%
home team shots	1.34%	difference in the number of through balls	1.60%
away team tackles	1.30%	home passes	1.41%
Card situation	**Relative Weight**	home long balls	1.39%
difference in the rate of cards per foul	1.41%	**Attempt types**	**Relative Weight**
home cards per foul rate	1.35%	difference in the number of set-piece	1.35%
		home team open plays	1.27%

Table 45. Different types of important variables in the ALL dataset

Team match statistics	Relative Weight	Pass types	Relative Weight
difference in the number of shots on target	13.46%	difference in the number of crosses	8.42%
home-team /away-team	10.69%	home crosses	5.24%
away team shots on target	4.71%	away crosses	3.21%
home team shots on target	4.31%	difference in the number of through balls	1.73%
difference in the number of shots	1.32%	away short passes	1.56%
Attempt types	**Relative Weight**	away passes	1.44%
difference in the number of counter attacks	1.93%	home short passes	1.41%
difference in the number of open plays	1.32%		

The results of experiments with La Liga dataset indicated that 13 of the 22 team match statistics variables, 12 of the 18 pass types variables, 2 of the 15 attempt types variables, and 3 of the 24 card situation variables were considered as essential variables that affected the match outcome. The most crucial team match statistics variable was the difference in the number of shots on target. Variables related to the number of crosses were the top three influential pass types variables. The difference in the number of counter attacks and the difference in the number of open plays were significant attempt types variables. Variables related to the rate of cards per foul were the most critical card situation variables.

The results of experiments with EPL dataset indicate that 14 of the 22 team match statistics variables, 13 of the 18 pass types variables, 3 of the 15 attempt types variables, and 1 of the 24 card situation variables were considered as essential variables that affected the match outcome. The most important team match statistics variable was the difference in the number of shots on target. Variables related to the number of crosses were the top three influential pass types variables. The away-team open plays number and the difference in the number of open plays were important attempt types variables. The difference in the number of open plays was also an essential attempt type variable. The most critical card situation variable was the difference in the rate of cards per foul.

The results of experiments with MLS dataset indicate that 10 of the 22 team match statistics variables, 11 of the 18 pass types variables, 2 of the 15 attempt types variables, and 2 of the 24 card situation variables were considered as important variables that affected the match outcome. Different from other datasets, the home-team/away-team factor was the most important team match statistics variable. Variables related to the number of crosses were the top three influential pass types variables. The difference in the number of set-pieces and the difference in the number

of open plays were important attempt types variables. The most critical card situation variable was the difference in the rate of cards per foul.

The results of experiments with ALL dataset indicated that 5 of the 22 team match statistics variables, 7 of the 18 pass types variables, 2 of the 15 attempt types variables, and none of the 24 card situation variables were considered as important variables that affected the match outcome. The most important team match statistics variable was the difference in the number of shots on target. Variables related to the number of crosses were the top three influential pass types variables. The difference in the number of counter attacks and the difference in the number of open plays were important attempt types variables.

To sum up, the number of crosses variable had the most significant impact on the outcome of the game for all leagues in the pass type category. The difference in the rate of cards per foul was found to be the most influential card situation for all leagues. The difference in the number of counter attacks and open plays were the most critical attempt types for La Liga and EPL. The difference in the number of set-piece was the most critical attempt type for MLS.

3.7 Conclusion

In this study, the predictive performance of five machine learning models was compared, both separately on the dataset of each of 3 major soccer leagues and on a combined dataset that included all the 3 leagues. The three leagues were the two European leagues, La Liga, EPL, and one U.S league, MLS. The results showed that RF and NN had the best and most stable performance in predicting the outcomes of different leagues. Due to the high standard deviation of the results obtained by NN models, RF was selected as the best and most stable modeling approach.

Whether a team was a home-team or away-team was the most crucial predictor variable in MLS. For other leagues, the home-field advantage was also a significant factor. Still, the most critical factor determining the outcomes of the La Liga and EPL games was the difference in the number of shots on target. This result provides further evidence that the magnitude of the home-field advantage was related to the geographic distance the away team traveled (Goddard, 2006). Although MLS is divided into eastern and western coference according to geographical location, most away-teams of MLS league still have to travel great distances, so the effect of the home-field advantage is more significant than other leagues.

For the three leagues, the number of crosses was the most significant pass type, and the difference in the rate of card per foul was the most crucial card situation. The referee primarily determines the difference in the rate of card per foul. For the Europe leagues, the difference in the number of counter attacks and open plays were consequential attempt types affecting a game result in La Liga and EPL, while in the MLS, the difference in the number of set-piece was the most crucial predictor variable. By comparing the important factors affecting the three leagues, the important factors affecting La Liga and EPL matches were the same, while the factors affecting MLS were slightly different. Coaches of different leagues could refer to these results to more specifically improve the probability of winning.

For future work, other machine learning models need to be applied and more predictors need to be included to improve prediction accuracy further. For the current study, the predictors included were result statistics collected after the predicted games. Predictors obtained from pre-play statistics (statistics collected before the predicted games) will be added, for example, the number of winnings for the last season, the referee, the number of injured players, team value, the number of years the coach has been in charge and so on. The analysis of these factors could be more helpful

in improving the team's winning probability from the perspective of team management. Also, more leagues need to be included to analyze the significant factors affecting the games of different leagues to further evident our results. If better data sources could be obtained to get data from more seasons, time-serie analysis could be conducted to analyze how the influencial predictors change over time .

4. CONCLUSION

In this thesis, two problem domains have been focused on: 1) traffic crash severity prediction, 2) soccer game result prediction. The objective was to explore and compare the prediction performance of five machine learning approaches to inform policy making in road safety and contribute to the state of art in soccer analytics. In the first part, machine learning approaches were applied to a time series crash data of Connecticut to predict the injury severity of involved human beings. The crash data covers the traffic crashes occurred in Connecticut over 20-year period (1995-2014). In the second study, five machine learning approaches were applied to three major soccer leagues' (MLS, EPL, and La Liga) data to predict the outcome of soccer games. The soccer data includes 5-season game results and statistics of the soccer leagues. Consistent across the two studies, RF and NN had better prediction performance, especially in the soccer datasets. By investigating the importance of predictors estimated from RF, this study also provided valuable knowledge that could be applied to real-life situations. Both the prediction accuracy and the predictor importance results were compared and analyzed from different perspectives to extract more information from the data. In the first study, for the crash severity prediction, economic analysis and time-series analysis were conducted. The results showed that over-sampling and under-sampling methods helped improve the prediction accuracy in terms of financial cost and the prediction accuracy of fatal & incapacitating injuries. In the second study, for the soccer games outcome prediction, the performance of the models in predicting game outcomes of different league were compared. The comparison results revealed that the home-field advantage of MLS was more considerable than EPL and La Liga. Overall, it was found that the selected machine learning approaches provided stable, reliable and highly-accurate prediction performance to both soccer and crash datasets.

REFERENCES

A.T. Kearney, Inc. (2011). The sports market. Retrieved from https://www.atkearney.com/documents/10192/6f46b880-f8d1-4909-9960-cc605bb1ff34

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting, 35(2), 741-755.

Berrar, D., Lopes, P., & Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. Machine Learning, 108(1), 97-126.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Bunker, R. P., & Thabtah, F. (2017). A machine learning framework for sport result prediction. Applied computing and informatics.

Bzdok, D., Altman, N., Krzywinski, M. (2018). Statistics versus machine learning. Nature Methods, pp. 233-234.

Carmichael, F., & Thomas, D. (2005). Home-field effect and team performance: evidence from English premiership football. Journal of sports economics, 6(3), 264-281.

Chang, L. Y., & Wang, H. W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis & Prevention, 38(5), 1019-1027.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Chen, C., Zhang, G., Qian, Z., Tarefder, R. A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. Accident Analysis & Prevention, 90, 128-139.

Delen, D., Tomak, L., Topuz, K., & Eryarsoy, E. (2017). Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. Journal of Transport & Health, 4, 118-131.

Eggels, H., van Elk, R., & Pechenizkiy, M. (2016). Explaining Soccer Match Outcomes with Goal Scoring Opportunities Predictive Analytics. In MLSA@ PKDD/ECML.

Egilmez, G., & McAvoy, D. (2017). Predicting nationwide road fatalities in the US: a neural network approach. International Journal of Metaheuristics, 6(4), 257-278.

Egilmez, G., Erdil, N. Ö., Arani, O. M., & Vahid, M. (2019). Application of artificial neural networks to assess student happiness. International Journal of Applied Decision Sciences, 12(2), 115-140.

Farhadi, F. (2017). Learning Activation Functions in Deep Neural Networks (Doctoral dissertation, École Polytechnique de Montréal).

Fox‐Wasylyshyn, S. M., & El‐Masri, M. M. (2005). Handling missing data in self‐report measures. Research in nursing & health, 28(6), 488-495.

Glorot, X., Bordes, A., & Bengio, Y. (2011, June). Deep sparse rectifier neural networks. In Proceedings of the fourteenth international conference on artificial intelligence and statistics (pp. 315-323).

Goddard, J. (2006). Who wins the football?. Significance, 3(1), 16-19.

Groll, A., Ley, C., Schauberger, G., & Van Eetvelde, H. (2019). A hybrid random forest to predict soccer matches in international tournaments. Journal of Quantitative Analysis in Sports, 15(4), 271-287.

Hastie, T. and Qian, J. (2016, 9 20). Glmnet vignette. Retrieved from http://www. web. stanford. edu/~ hastie/Papers/Glmnet_Vignette. pdf

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

Hubáček, O., Šourek, G., & Železný, F. (2019). Learning to predict soccer results from relational data with gradient boosted trees. Machine Learning, 108(1), 29-47.

Hucaljuk, J., & Rakipović, A. (2011, May). Predicting football scores using machine learning techniques. In 2011 Proceedings of the 34th International Convention MIPRO (pp. 1623-1627). IEEE.

Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. Accident Analysis & Prevention, 108, 27-36.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Jeong, H., Jang, Y., Bowman, P. J., & Masoud, N. (2018). Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data. Accident Analysis & Prevention, 120, 250-261.

Kahn, J. (2003). Neural network prediction of NFL football games. World Wide Web electronic publication, 9-15.

Kerr, M. G. S. (2015). Applying machine learning to event data in soccer (Doctoral dissertation, Massachusetts Institute of Technology).

Khattak, A. J., Pawlovich, M. D., Souleyrette, R. R., & Hallmark, S. L. (2002). Factors related to more severe older driver traffic crash injuries. Journal of Transportation Engineering, 128(3), 243-249.

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).

Leung, F. H. F., Lam, H. K., Ling, S. H., & Tam, P. K. S. (2003). Tuning of the structure and parameters of a neural network using an improved genetic algorithm. IEEE Transactions on Neural networks, 14(1), 79-88.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Lock, D., & Nettleton, D. (2014). Using random forests to estimate win probability before each play of an NFL game. Journal of Quantitative Analysis in Sports, 10(2), 197-205.

Mackay, N. (2017). Predicting goal probabilities for possessions in football. Master of Science. Vrije Universiteit Amsterdam. URl: https://beta.% 20vu.% 20nl/nl/Images/werkstuk-mackay% 5C_tcm235-849981.% 20pdf.

Magel, R., & Melnykov, Y. (2014). Examining influential factors and predicting outcomes in European soccer games. International Journal of Sports Science, 4(3), 91-96.

McCabe, A., & Trevathan, J. (2008, April). Artificial intelligence in sports prediction. In Fifth International Conference on Information Technology: New Generations (itng 2008) (pp. 1194-1197). IEEE.

Mitchell, M. W. (2011). Bias of the Random Forest out-of-bag (OOB) error for certain input parameters. Open Journal of Statistics, 1(03), 205.

Oztekin, A., Al-Ebbini, L., Sevkli, Z., & Delen, D. (2018). A decision analytic approach to predicting quality of life for lung transplant recipients: A hybrid genetic algorithms-based methodology. European Journal of Operational Research, 266(2), 639-651.

Prato, C. G., Gitelman, V., & Bekhor, S. (2012). Mapping patterns of pedestrian fatal accidents in Israel. Accident Analysis & Prevention, 44(1), 56-62.

Part, D. (2010). Highway Safety Manual. Washington, DC: American association of state highway and transportation officials (aashto).

Sagar Sharma (2019, 9). Activation functions in neural networks. Retrieved from towardsdatascience: https:// towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6.

Sawe, B. E. (2018, April 5). The Most Popular Sports In The World. Retrieved from https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html

Shin, J., & Gasparyan, R. (2014). A novel way to soccer match prediction. Stanford University: Department of Computer Science.

Singh, G., Sachdeva, S. N., & Pal, M. (2018). Comparison of three parametric and machine learning approaches for modeling accident severity on non-urban sections of Indian highways. Advances in transportation studies, 45.

Sun, A., Lim, E. P., & Liu, Y. (2009). On strategies for imbalanced text classification using SVM: A comparative study. Decision Support Systems, 48(1), 191-201.

Tax, N., & Joustra, Y. (2015). Predicting the Dutch football competition using public data: A machine learning approach. Transactions on knowledge and data engineering, 10(10), 1-13.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

Tsai, J. T., Chou, J. H., & Liu, T. K. (2006). Tuning the structure and parameters of a neural network by using hybrid Taguchi-genetic algorithm. IEEE Transactions on Neural Networks, 17(1), 69-80.

Ulmer, B., Fernandez, M., & Peterson, M. (2013). Predicting soccer match results in the English Premier League.

Xu, Y., Wu, C., Zheng, K., Niu, X., & Yang, Y. (2017). Fuzzy–synthetic minority oversampling technique: Oversampling based on fuzzy set theory for Android malware detection in imbalanced datasets. International Journal of Distributed Sensor Networks, 13(4), 1550147717703116.

Ye, F., & Lord, D. (2014). Comparing three commonly used crash severity models on sample size requirements: multinomial logit, ordered probit and mixed logit models. Analytic methods in accident research, 1, 72-85.

Yezus, A. (2014). Predicting outcome of soccer matches using machine learning. Saint-Petersburg University.

Zambom-Ferraresi, F., Rios, V., & Lera-López, F. (2018). Determinants of sport performance in European football: What can we learn from the data?. Decision Support Systems, 114, 18-28.

Zhang, J., Li, Z., Pu, Z., & Xu, C. (2018). Comparing Prediction Performance for Crash Injury Severity Among Various Machine Learning and Statistical Methods. IEEE Access, 6, 60079-6008

APPENDIX

A. Variables included in each dataset (soccer)

Home team factor	Away team factor	Difference
home-team/away-team		
home team shots	away team shots	difference in the number of shots
home team shots on target	away team shots on target	difference in the number of shots on target
home team pass success rate	away team pass success rate	difference in the rate of pass success
home team aerial duel success rate	away team aerial duel success rate	difference in the rate of aerial duel success
home team dribbles won	away team dribbles won	difference in the number of dribbles won
home team tackles	away team tackles	difference in the number of tackles
home team possession rate	away team possession rate	difference in the rate of possession
home team open plays	away team open plays	difference in the number of open plays
home team set-piece	away team set-piece	difference in the number of set-piece
home team counter attacks	away team counter attacks	difference in the number of counter attacks
home team penalty	away team penalty	difference in the number of penalty
home team own goal	away team own goal	difference in the number of own goal
home team red cards	away team red cards	difference in the number of red cards
home team yellow cards	away team yellow cards	difference in the number of yellow cards
home team cards for fouls	away team cards for fouls	difference in the number of cards for fouls
home team cards for unprofessional	away team cards for unprofessional	difference in the number of cards for unprofessional
home team cards for dive	away team cards for dive	difference in the number of cards for dive
home team cards for other reason	away team cards for other reason	difference in the number of cards for other reason
home cards per foul rate	away cards per foul rate	difference in the rate of cards per foul
home fouls per game	away fouls per game	difference in the number of fouls per game
home passes	away passes	difference in the number of passes

home crosses	away crosses	difference in the number of crosses
home through balls	away through balls	difference in the number of through balls
home long balls	away long balls	difference in the number of long balls
home short passes	away short passes	difference in the number of short passes
home average pass streak	away average pass streak	difference in the number of average pass streak

B. Relative weight of variables in La Liga dataset (soccer)

Variable	Relative Weight	Variable	Relative Weight
difference in the number of shots on target	10.91%	home team shots	0.99%
home-team /away-team	5.92%	away fouls per game	0.93%
difference in the number of crosses	5.18%	home team open plays	0.87%
away team shots on target	4.57%	difference in the rate of possession	0.86%
home crosses	4.25%	home team possession rate	0.83%
home team shots on target	3.43%	home through balls	0.83%
away crosses	2.58%	home team set-piece	0.78%
difference in the rate of cards per foul	1.93%	home team counter attacks	0.78%
away team tackles	1.84%	away team set-piece	0.76%
difference in the rate of aerial duel success	1.71%	difference in the number of yellow cards	0.72%
away long balls	1.68%	away through balls	0.71%
away short passes	1.64%	away team possession rate	0.71%
difference in the number of tackles	1.59%	home team yellow cards	0.63%
home short passes	1.57%	away team counter attacks	0.62%
difference in the number of through balls	1.56%	away team yellow cards	0.62%
difference in the number of counter attacks	1.55%	difference in the number of cards for fouls	0.61%
difference in the number of passes	1.53%	home team cards for fouls	0.58%
away passes	1.53%	difference in the number of cards for other reason	0.52%
difference in the number of shots	1.48%	home team cards for other reason	0.51%
difference in the number of open plays	1.47%	home average pass streak	0.49%
away team shots	1.47%	difference in the number of average pass streak	0.43%
away team aerial duel success rate	1.46%	away team cards for fouls	0.40%

home cards per foul rate	1.46%	away average pass streak	0.33%
difference in the number of short passes	1.40%	difference in the number of penalty	0.32%
home team aerial duel success rate	1.38%	home team penalty	0.25%
difference in the number of long balls	1.33%	difference in the number of red cards	0.25%
away cards per foul rate	1.33%	away team cards for other reason	0.22%
home passes	1.30%	away team penalty	0.15%
home team tackles	1.28%	away team cards for unprofessional	0.11%
away team pass success rate	1.28%	home team cards for unprofessional	0.11%
away team dribbles won	1.22%	away team red cards	0.08%
difference in the number of dribbles won	1.22%	home team red cards	0.08%
difference in the number of set-piece	1.22%	difference in the number of cards for unprofessional	0.07%
home long balls	1.21%	difference in the number of own goal	0.07%
away team open plays	1.09%	away team own goal	0.05%
home team dribbles won	1.06%	home team own goal	0.01%
home team pass success rate	1.04%	home team cards for dive	0.00%
difference in the rate of pass success	1.03%	away team cards for dive	0.00%
difference in the number of fouls per game	1.01%	difference in the number of cards for dive	0.00%
home fouls per game	1.00%		

C. Relative weight of variables in EPL dataset (soccer)

Variable	Relative Weight	Variable	Relative Weight
difference in the number of shots on target	7.30%	difference in the number of set-piece	1.00%
home-team /away-team	6.02%	home team open plays	0.98%
difference in the number of crosses	4.46%	away fouls per game	0.97%
home crosses	3.80%	away team possession rate	0.88%
home team shots on target	3.71%	home team set-piece	0.85%
away crosses	3.36%	home through balls	0.84%
away team shots on target	3.11%	difference in the rate of possession	0.76%
away short passes	2.11%	home team possession rate	0.74%
home team pass success rate	1.99%	away team cards for fouls	0.74%
difference in the number of shots	1.94%	difference in the number of cards for fouls	0.73%
away passes	1.92%	away through balls	0.71%
home short passes	1.91%	home team counter attacks	0.70%
difference in the number of tackles	1.88%	away team set-piece	0.65%
difference in the number of through balls	1.79%	away team counter attacks	0.63%
away team open plays	1.75%	difference in the number of cards for other reason	0.60%
home team tackles	1.72%	away team yellow cards	0.56%
difference in the number of open plays	1.72%	difference in the number of yellow cards	0.54%
difference in the number of long balls	1.68%	difference in the number of average pass streak	0.46%
home long balls	1.66%	away average pass streak	0.46%
away team pass success rate	1.62%	home team yellow cards	0.45%
difference in the number of counter attacks	1.62%	away team cards for other reason	0.42%
away team tackles	1.60%	home average pass streak	0.40%
away team shots	1.59%	home team cards for fouls	0.40%

home passes	1.58%	home team cards for other reason	0.38%
difference in the rate of cards per foul	1.50%	difference in the number of penalty	0.31%
difference in the number of dribbles won	1.44%	difference in the number of own goal	0.31%
home team shots	1.36%	difference in the number of red cards	0.28%
difference in the number of passes	1.34%	home team penalty	0.21%
away long balls	1.33%	home team own goal	0.18%
difference in the number of short passes	1.32%	difference in the number of cards for unprofessional	0.17%
difference in the rate of pass success	1.31%	away team penalty	0.14%
home team dribbles won	1.24%	home team cards for unprofessional	0.12%
difference in the number of fouls per game	1.23%	home team red cards	0.11%
difference in the rate of aerial duel success	1.22%	away team cards for unprofessional	0.10%
home team aerial duel success rate	1.22%	difference in the number of cards for dive	0.07%
away team dribbles won	1.21%	away team own goal	0.06%
away team aerial duel success rate	1.16%	away team cards for dive	0.05%
home fouls per game	1.14%	away team red cards	0.05%
home cards per foul rate	1.10%	home team cards for dive	0.01%
away cards per foul rate	1.03%		

D. Relative weight of variables in MLS dataset (soccer)

Variable	Relative Weight	Variable	Relative Weight
home-team /away-team	11.30%	away team dribbles won	0.97%
difference in the number of shots on target	7.55%	difference in the rate of pass success	0.95%
difference in the number of crosses	6.68%	difference in the number of counter attacks	0.89%
home crosses	4.75%	home fouls per game	0.86%
away team shots on target	3.33%	away through balls	0.83%
away crosses	3.13%	away team set-piece	0.80%
home team shots on target	2.67%	home through balls	0.77%
difference in the number of long balls	1.93%	away team counter attacks	0.74%
away long balls	1.84%	away team possession rate	0.70%
away passes	1.68%	home team possession rate	0.69%
away short passes	1.67%	difference in the number of cards for fouls	0.65%
difference in the number of tackles	1.62%	difference in the rate of possession	0.59%
home short passes	1.61%	home team cards for fouls	0.55%
difference in the number of through balls	1.60%	away team yellow cards	0.52%
home team tackles	1.56%	difference in the number of cards for other reason	0.51%
difference in the rate of cards per foul	1.41%	difference in the number of yellow cards	0.50%
home passes	1.41%	difference in the number of red cards	0.49%
difference in the number of shots	1.41%	away team cards for fouls	0.46%
home long balls	1.39%	home team yellow cards	0.43%
home team pass success rate	1.38%	away team cards for other reason	0.42%
difference in the number of set-piece	1.35%	away average pass streak	0.41%
home cards per foul rate	1.35%	difference in the number of average pass streak	0.39%
home team shots	1.34%	home team cards for other reason	0.31%
away team tackles	1.30%	home average pass streak	0.28%

home team open plays	1.27%	difference in the number of penalty	0.27%
difference in the number of short passes	1.26%	home team counter attacks	0.25%
away team pass success rate	1.25%	home team red cards	0.24%
difference in the number of open plays	1.21%	away team red cards	0.21%
home team dribbles won	1.20%	home team penalty	0.20%
away team shots	1.16%	away team penalty	0.17%
home team set-piece	1.15%	difference in the number of own goal	0.09%
difference in the number of passes	1.15%	difference in the number of cards for unprofessional	0.08%
away cards per foul rate	1.14%	away team own goal	0.07%
difference in the number of dribbles won	1.12%	home team cards for unprofessional	0.06%
home team aerial duel success rate	1.12%	away team cards for unprofessional	0.05%
difference in the rate of aerial duel success	1.10%	home team own goal	0.03%
away team open plays	1.08%	home team cards for dive	0.00%
away team aerial duel success rate	1.06%	away team cards for dive	0.00%
difference in the number of fouls per game	1.04%	difference in the number of cards for dive	0.00%
away fouls per game	0.99%		

E. Relative weight of variables in ALL dataset (soccer)

Variable	Relative Weight	Variable	Relative Weight
difference in the number of shots on target	13.46%	away team dribbles won	0.78%
home-team /away-team	10.69%	home team dribbles won	0.76%
difference in the number of crosses	8.42%	difference in the number of fouls per game	0.73%
home crosses	5.24%	home team open plays	0.72%
away team shots on target	4.71%	home fouls per game	0.71%
home team shots on target	4.31%	home team counter attacks	0.68%
away crosses	3.21%	away through balls	0.68%
difference in the number of counter attacks	1.93%	away team possession rate	0.62%
difference in the number of through balls	1.73%	home through balls	0.60%
away short passes	1.56%	difference in the rate of possession	0.60%
away passes	1.44%	home team possession rate	0.49%
home short passes	1.41%	away team set-piece	0.47%
difference in the number of shots	1.32%	home average pass streak	0.47%
difference in the number of open plays	1.32%	difference in the number of fouls	0.45%
away team pass success rate	1.25%	home team fouls	0.41%
home passes	1.24%	away team fouls	0.38%
home cards per foul rate	1.23%	difference in the number of yellow cards	0.34%
difference in the rate of cards per foul	1.20%	away team yellow cards	0.33%
away long balls	1.12%	away average pass streak	0.31%
home team pass success rate	1.10%	home team yellow cards	0.31%
away team shots	1.10%	difference in the number of average pass streak	0.29%
home team shots	1.08%	difference in the number of penalty	0.29%
difference in the number of passes	1.08%	away team cards for other reason	0.27%

difference in the number of long balls	1.08%	difference in the number of cards for other reason	0.26%
away team tackles	1.05%	home team cards for other reason	0.22%
difference in the number of tackles	1.05%	difference in the number of red cards	0.21%
difference in the number of short passes	1.04%	away team penalty	0.16%
home long balls	1.04%	home team penalty	0.15%
away team open plays	1.00%	difference in the number of own goal	0.12%
home team aerial duel success rate	0.97%	away team red cards	0.08%
difference in the number of dribbles won	0.89%	home team red cards	0.07%
difference in the rate of aerial duel success	0.88%	away team own goal	0.06%
away cards per foul rate	0.88%	home team own goal	0.06%
away team counter attacks	0.86%	difference in the number of unprofessional	0.04%
home team set-piece	0.86%	home team unprofessional	0.03%
difference in the number of set-piece	0.84%	away team unprofessional	0.02%
home team tackles	0.83%	home team dive	0.00%
away team aerial duel success rate	0.80%	away team dive	0.00%
difference in the rate of pass success	0.79%	difference in the number of dive	0.00%
away fouls per game	0.78%		