12-2016

Safety Performance Prediction of Large-Truck Drivers in the Transportation Industry

Emily Moneka Francis Xavier
University of Arkansas, Fayetteville

Safety Performance Prediction of
Large-Truck Drivers in the Transportation Industry

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

by

Emily Moneka Francis Xavier
Sri Ramakrishna Engineering College
Bachelor of Engineering in Electronics and Instrumentation Engineering, 2013

December 2016
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

_____
Dr. Shengfan Zhang
Thesis Director

_____
Dr. Justin R. Chimka
Committee Member

_____
Dr. Haitao Liao
Committee Member

ABSTRACT

The trucking industry and truck drivers play a key role in the United States commercial transportation sector. Accidents involving large trucks is one such big event that can cause huge problems to the driver, company, customer and other road users causing property damage and loss of life. The objective of this research is to concentrate on an individual transportation company and use their historical data to build models based on statistical and machine learning methods to predict accidents. The focus is to build models that has high accuracy and correctly predicts an accident. Logistic regression and penalized logistic regression models were tested initially to obtain some interpretation between the predictor variables and the response variable. Random forest, gradient boosting machine (GBM) and deep learning methods are explored to deal with high non-linear and complex data.

The cost of fatal and non-fatal accidents is also discussed to weight the difference between training a driver and encountering an accident. Since accidents are very rare events, the model accuracy should be balanced between predicting non-accidents (specificity) and predicting accidents (sensitivity). This framework can be a base line for transportation companies to emphasis the benefits of prediction to have safer and more productive drivers.

TABLE OF CONTENTS

1. INTRODUCTION

1.1. Background and Motivation

The transportation system in the United States is the largest in the world and the commercial transportation industry is in an enviable position. One out of seven workers in the U.S. are in the transportation field (U.S. Department of Transportation 2016), serving a huge number of business establishments all over the country. According to the American Trucking Association (ATA 2016), trucks moved around 9.2 billion tons of commodities annually, which constitutes about 70% of total freight tonnage, requiring 3 million truck drivers. As a result, truck drivers play an important role in the safe and efficient delivery of freight. With an inevitable need for moving commodities, statistics show that accidents involving large trucks continue to take a toll on truck drivers, their passengers, and other road users. Driving a 53-foot truck, undoubtedly involves lot of concentration and focus. Developing and continuously improving preventive measures of such events (accidents), is the responsibility of any trucking company.

The National Highway and Traffic Safety Administration (NHTSA) reported that an estimate of 438,000 large-trucks was involved in traffic crashes in 2014 (NHISA 2016). Two federal agencies, the U.S. Department of Transportation (USDOT) and the Federal Motor Carrier Safety Administration (FMCSA), regulate all laws related to trucking companies and determine the cause and nature of an accident when occurred. Trucking companies and drivers must follow the laws on commercial driver licenses, hours of service, maximum weight permitted, quality control of trucks and hazardous waste, etc. Based on estimates by Blincoe et al. (2002), the average cost of highway crashes was $59,153 USD. This estimate includes costs from medical

and emergency services, property damage, lost productivity, and the monetized value of loss of quality of life that a family experiences due to death or injury (Blincoe et al. 2002).

In addition, the trucking industry with respect to truck drivers has some serious problems such as driver shortage, and aging drivers. The average age of drivers in the industry has been steadily increasing (Short 2014). Presently, a large percentage of drivers will be retiring, and too few younger drivers are entering the industry (Short 2014). According to the ATA data, the driver shortage could rise up to nearly 240,000 by 2022 with the forecasted demand. These potential issues reinstate the importance of safe driving habits of the existing and future drivers. As safety is one of the key concerns of any transportation company, the prediction of drivers at risk of an accident will help a company to target the right group of drivers for safety training in order to reduce accidents. Based on the size of the company and the number of drivers, the predictions can run weekly, monthly, quarterly or bi-annually. From the drivers' perspective, the act of predicting the possibility of an accident based on their history may not be well accepted, and so executing the training process based on the predictions needs to be done very carefully with the sole intention of helping the drivers. Accidents, by nature are rare events compared to non-accidents, and so the goal is to reduce the number of accidents or to decrease the intensity of non-preventable accidents with proper training in place.

1.2. Research Objectives

Technology has greatly transformed the trucking industry to have safer fleet and more productive drivers. Trucking-related safety metrics have been continuously enhanced over the past decade, lowering the truck-related fatality rate to a considerable extent. Achieving a high safety level is an increased need for the transportation industry. The high reliability of trucks for moving freight makes it more challenging to identify new methods that can further achieve the

desired safety improvements without lowering productivity. Being proactive by providing regular safety training and the willingness to learn from the previous mistakes would be an effective step towards accident prevention. Bob Joop Goos, Chairman of the International Organization of Road Accident Prevention stated that "More than 90 % of road accidents are caused by human error. We, therefore, have to focus on people in our traffic safety programs" (Global Driver Risk Management – Alert Driving 2016). The key is to focus on the human element with the "objective of stimulating good (driving) behavior" says Goos. Many industrial and academic researchers have examined statistical models (Al-Ghamdi 2002, Blower et al. 2008, Shankar et al.1997) and machine learning models (Abdelwahab & Abdel-Aty 2001, Mussone et al. 1999, Xie et al. 2007) to predict accidents and their severity using drivers' behaviors and various external factors (co-passengers on road, pedestrians, signals, intersections, etc.) associated with an accident.

The two main methodologies of finding the relationship between the response variable and predictor variables are statistical methods that are regression based and machine learning techniques that are algorithm based. Traditional regression methods are unarguably the baseline for prediction. But with the increasing amount of data and availability of high computational capability, machine learning techniques are gaining more popularity.

The objective of this research is to identify large-truck drivers who may meet with an accident in the next 30 days using prediction models including the generalized linear models, random forest, gradient boosting machine and deep learning. This research attempts to improve both driver and fleet safety by using predictive analytics to identify drivers who are prone to future accidents based on historical data. The safety managers can then act upon these predictions by training the drivers to improve their safety on road for mutual benefits.

1.3. Research Contribution

To the best of our knowledge, machine learning or statistical approaches have not been directly applied in the prediction of future accidents. Most of the previous research focused on prediction of severity of the injuries, number of fatalities, accident zone like intersections, highway, sideway sweep and other specific type of locations (Al-Ghamdi 2002, Jovanis & Chang 1986, Mussone et al. 1999). In addition, all of the accident-related predictions have been so far based upon the publicly available or government data, which may produce very generalized analysis, with missing or unreliable data. The main difference of this research from the others is that it concentrates on a very specific cause of accidents, "the drivers". Excluding the external factors associated with the accidents, this study focuses on the influence of the driver on an accident (e.g., age, tenure, number of previous accidents, number of citations, etc.) using data from a commercial transportation organization. Most importantly, this research proposes a method to find the root cause of majority of the accidents for any individual transportation firm where driving is "an occupation" or considered to be "an expertise of a person". Although commercial truck drivers are highly trained and are considered to be more cautious than most other road users, a deeper understanding of their concerns and an appropriate training program is mandatory. Once an accident occurs, the risk of life and cost involved is dramatic. While each transportation company operates differently, has a different size, and may require different training programs for their drivers, this research provides an example for an in-house system of prediction for accidents that would greatly improve the company's safety performance along with cost saving benefits. The technology and software used in this analysis are available as open source, making it possible for companies to have predictions at no cost except for the manpower involved. This research utilizes some of the commonly used algorithms in order to gain high prediction accuracy. The goal is to predict a higher number of accidents by having high

predictive accuracy and this research does not concentrate on comparing the results of the various algorithms used.

1.4. Predictive Analytics

Predictive analytics has become widely used in various industries as a powerful tool to analyze future expectations or outcomes of a specific targeted goal. It is an area of data mining that uses data, statistical algorithms, and machine-learning techniques to identify the trends and behavior patterns of historical data to predict the likelihood of future outcomes (SAS Institute Inc. 2016).

The need for machine learning algorithms relies on the fact that it can accommodate more predictor variables with fewer assumptions and the availability of tuning parameters that act as internal knobs. The success of machine learning algorithms depends on handling the tradeoff between the learning complexity and the ability to explain the inner workings of the models (Johansson 2007). Higher learning complexity makes the model inner workings less explanatory and falls under the category of the so-called "Black-Box Techniques" (Krishna 2012) which includes random forests, neural networks, deep learning, gradient boosting machine (GBM), etc.

Machine learning is a subfield of computer science while statistical modeling is a subfield of mathematics. In machine learning methods, there are only a few assumptions spared from statistical methods and less prior knowledge about the data is required. On the other hand, statistical methods require a good prior knowledge of the data and verification of assumptions. Machine learning consists of a huge variety of algorithms that suits different applications. Understanding the different algorithms is very important and no one single algorithm is just perfect. Choosing an appropriate machine learning algorithm depends on the data and the purpose of the study.

Machine learning techniques are generally applied to high dimensional data sets; the more data you have, the more accurate your prediction can be, which however may lead to a black box situation. On the contrast, statistical methods are used for low dimensional data and where delivery of a high-level explanation of the model is desired. Knowing the audience before starting the modeling would be the first step for any type of analysis.

The remainder of this thesis is organized as follows. Section 2 covers the literature review for statistical methods (Section 2.1) and machine learning methods (Section 2.2) applied to various accident related predictions. Section 3 introduces the data used for analysis, explains the data preparation steps, and the techniques to balance the data. Section 4 briefly explains the methodologies of various models used along with the parameters tuning to obtain the best model. Section 5 presents the results of all the models along with the performance measure to validate their estimates. Finally, Section 6 summarizes the overall results including discussion on the cost of an accident and ways to mitigate an accident.

2. LITERATURE REVIEW

Predicting truck drivers' future accidents in a transportation company is closely related to predictions for driver turnover, driver behavior models, transit bus driver distraction, and many more. This section summarizes the most relevant problems, compares and contrasts the modeling techniques used, and examines some methodologies that are used in this study such as logistic regression, penalized regression, random forest, gradient boosting machine, and deep learning.

2.1. Prediction Using Statistical Models

Regression is an integral part of data analysis when concerned about the relationship between the independent and the dependent variables. For the binomial classification problem

under study, logistic regression is appropriate. In logistic regression models, the response variable is binary or dichotomous (e.g., fatal or non-fatal). Jovanis & Chang (1986) studied the relationship of accidents to miles traveled using Poisson regression. The model was built using the accidents, travel mileage, and environmental data from the Indian Toll road. The model revealed that automobile and truck accidents are directly related to the automobile and truck travel mileage. As the truck Vehicle Miles Traveled (VMT) increases, the chance of collision also increases.

Murray et al. (2006) developed a model for predicting a truck crash involvement using logistic regression where the model uses driver's historical driving record and later used the significant factors identified to plan for effective enforcement actions to counteract the driving behaviors. The model suggested that drivers who had a past crash increase their likelihood of a future crash by 87%, where reckless driving and improper turn violation are the most important predictors.

Al-Ghamdi (2002) used logistic regression to estimate the factors influencing accidents as fatal or non-fatal and used statistical interpretation of the model estimates. The accident location (i.e., intersection and road section) and accident cause (i.e., speed too fast, run on red light, wrong way, not giving priority and others) were observed to be significant causing the fatal accident.

In another study using logistic regression, Blower et al. (2008) identified that driver errors (driver's contribution to the accident) are related to characteristics of the driver (i.e., age, sex, method of payment, and previous driving record) and bus operations (i.e., operation type and trip type). Driver characteristic i.e., violations and crashes within the previous three years

and bus operation types were the only statistically significant factors for the bus driver error crash.

Generally, researchers use the goodness-of-fit statistics (Shankar et al. 1997, Miaou and Lord 2003) to determine which statistical model fits the data the best. On the other hand, a model that fits the data very well does not necessarily mean that it will be able to predict crashes successfully. Due to the problems where predictors like drivers' behaviors, drivers' characteristics, and factors related to accidents are nondeterministic and highly nonlinear, it is difficult for traditional methods to embody this kind of uncertain relationship to provide high accuracy in prediction.

2.2. Prediction Using Machine Learning Techniques

Mussone et al. (1999) used neural networks to analyze vehicle accident that occurred at intersections in Milan, Italy. They chose feed-forward neural networks with a back-propagation learning paradigm. The model has 10 input units, 4 hidden units and 1 output unit. The input nodes were day or night, traffic flow, road surface condition, number of conflict points, type of intersection, accident type, and weather condition. The output node was called an accident index and was calculated as the ratio between the number of accidents for a given intersection and the number of accidents at the most dangerous intersection. The model showed that the highest accident index for running over of pedestrian occurred at non-signalized intersections at nighttime.

Yang et al. (1999) studied the 1997 Alabama interstate alcohol related data using the neural network approach to detect safer driving patterns that have less chance of causing death and injury when a car crash occurs. The target variable in their study had two classes: injury and non-injury, in which the injury class included fatalities. They found that by controlling a single

variable (such as the driving speed or the light conditions), they could potentially reduce fatalities and injuries by up to 40%.

Abdelwahab et al. (2001) focused on two-vehicle accidents that occurred at signalized intersections. The accident data from the Central Florida area was used where the injury severity was divided into three classes: no injury, possible injury, and disabling injury. The performance of the multilayer perceptron (MLP) and fuzzy adaptive resonance theory (ART) neural networks was analyzed. MLP neural network gave better generalization performance than fuzzy ARTMAP and O-ARTMAP, where these two are types of ART. Fuzzy ARTMAP is a clustering algorithm that maps the set of input vectors to a set of clusters and O-ARTMAP is an ordered fuzzy ARTMAP algorithm. The authors also tested the result of the MPL model against the ordered logit model. The MPL model provided the best training and testing performance as opposed to the other two models.

Chong et al. (2005) used the National Automotive Sampling System (NASS) General Estimates System (GES) automobile accident data from 1995 to 2000 and investigated the performance of four machine learning paradigms to model the severity of injury that occurred during traffic accidents: 1.) neural networks trained using hybrid learning approaches, 2.) support vector machines, 3.) decision trees and 4.) a concurrent hybrid model involving decision trees and neural networks. Their research revealed that, the concurrent hybrid model involving decision trees and neural networks outperformed the other three approaches.

Moghaddam et al. (2011) used the artificial neural network (ANN) approach for crash severity prediction in urban highways and identification of significant crash-related factors. The model resulted in 25 independent variables as significant, having the highest value of crash severity as measured by fatality-injury crash percent. The model reflected the relationship

between crash severity on urban highways and the traffic variables including traffic volume, flow speed, human factors, road, vehicle and weather conditions. The finding of the study showed that the feed forward back propagation (FFBP) networks such as the MLP models yielded the best results.

Krishnaveni and Hemalatha (2011) investigated several classification techniques such as naive bayes, J48, adaboostm1, partial decision tree classifier, and random forest classifiers for predicting the severity of an injury that occurred during accidents. Data used in the analysis was traffic accident records of the year 2008 produced by the Transport Department of the Government of Hong Kong. The analysis revealed that random forest, instead of selecting all the attributes for classification, outperforms other classification algorithms. Genetic algorithm was used for feature selection to reduce the dimensionality of the data set.

Beshah et al. (2011) employed the classification and adaptive regression trees (CART) and random forest approaches in an effort to reduce road safety problems. The data was collected from three regional administrations in Ethiopia. The result showed that random forest modeling technique performs better by exhibiting lower error rate, higher ROC score and greater prediction accuracy than CART. The model performed well in determining non-injury risk of an accident based on the percentage of correct predictions of the non-injury case.

Guelman (2012) used the gradient boosting machine (GBM) method and tested against a conventional generalized linear model using an imbalanced data set for predicting an auto insurance loss cost modeling. The undersampling technique was used to initially balance the data. The results suggested that GBM presented a very good prediction compared to the generalized linear model. The author also discussed about the interpretability of the GBM model

using relative influence of the input variables and partial dependence plot that helps to understand the GBM output better, as opposed to other machine learning techniques.

Zhang & Haghani (2015) tested GBM against autoregressive integrated moving average (ARIMA) model and random forest for predicting the freeway travel time using the data provided by INRIX, a private sector company, where GBM was found to outperform the other methods. The data consists of two freeway sections in Maryland. GBM model captured the sharp discontinuities in traffic conditions (when traffic changes from uncongested to congested and vice versa) and handles the tree complexity (variable interaction).

Overall, most of the research that has been discussed in this section consists of only 10-15 independent variables in their models and have limitations on the reliability of the data. This research investigates generalized linear models (logistic regression and penalized logistic regression), deep learning networks, gradient boosting machine, and random forest to build models of high accuracy in predicting drivers at risk. Since the data is very specific to one company, the randomness involved in data is low and controllable. This research aims to reduce accidents of every individual trucking company, which will ultimately reduce the overall truck accident percentage in the country.

3. INTRODUCTION TO DATA

The data set used for this research is from a private transportation organization in the United States. The data set contains approximately 1.7 million records with 50 variables on drivers' weekly data starting from November 2012 until January 2015. Table 1 below lists all the predictor variables used for the analysis. Descriptive statistics of the data is not presented for confidentiality issues.

Table 1: List of Predictor Variables, their Category, Variable Type and Description

No.	Independent Variable	Category	Variable Type	Variable description
1	Gender	Demographic	Categorical	The gender of the driver categorized as male, female or undefined
2	Tenure	Demographic	Continuous	Tenure of the driver with the company
3	Previous Experience in the Same Company	Demographic	Categorical	Previous employment with the same company (in years)
4	Number of jobs previously held	Demographic	Continuous	Number of previous jobs held with different companies
5	ClassA Experience	Demographic	Continuous	Previous experience of driving ClassA trucks
6	ClassB Experience	Demographic	Continuous	Previous experience of driving ClassB trucks
7	Age	Demographic	Continuous	Age of the driver
8	Ethnicity	Demographic	Categorical	American Indian, Asian, Black, Hawaii/PAC, Hispanic, multiple, white, Not Specific
9	Number of Driver Inquiries	Demographic	Continuous	Number of inquiries on drivers updated on a weekly basis
10	Percentage Quit of Previous Jobs	Demographic	Continuous	Of the previous jobs held what is the percentage of quit
11	Weekly Pay	Financial	Continuous	Average weekly pay for the driver
12	Number of Cash Advances	Financial	Continuous	Number of cash advances received
13	Cash Advance Amount	Financial	Continuous	Amount given as cash advance for the driver
14	401k Participation	Financial	Categorical	Participation in the 401k Election (Y/N)
15	Million Miles Award Recipient	Financial	Categorical	Million miles award (Y/N)
16	401k Max Match	Financial	Categorical	401k match (Y/N)
17	Job Family	Operations	Categorical	Division of the trucking families - OTR, REG, LOC
18	Business Unit	Operations	Categorical	Business units within the organization (3 units)
19	Number of Miles driven	Operations	Continuous	Number of miles driven for the week

Table 1 (Cont.): List of Predictor Variables, their Category, Variable Type and Description

No.	Independent Variable	Category	Variable Type	Variable description
20	Number of Drivers on Board	Operations	Continuous	Number of drivers on board
21	Board Driver Turnover	Operations	Continuous	Board driver turnover rate
22	Number of Loads	Operations	Continuous	Number of loads per week
23	Number of Hazardous Loads	Operations	Continuous	Number of load requiring concerns per week
25	Number of Driver Failures	Operations	Continuous	Driver failure in the past 4 weeks
26	Number of Fuel Runouts	Operations	Continuous	Fuel runout in the past 4 weeks
27	Number of Hours-of-Violation	Operations	Continuous	Hours of service violation in the past 4 weeks
28	Number of Consecutive Days Off	Operations	Continuous	Number of consecutive days off of the driver per week
29	Truck Manufacturer	Operations	Categorical	Manufacturer of the truck – 5 different manufacturers
30	Tractor Manufacturer	Operations	Categorical	Manufacturer of the tractor – 11 different manufacturers
31	Number of Accident	Safety	Continuous	Number of accidents in the past 12 months
32	Number of Complaint	Safety	Continuous	Number of complaints in the past 12 months
33	Number of Incident	Safety	Continuous	Number of incidents in the past 12 months
34	Number of Observation	Safety	Continuous	Number of observations in the past 12 months
35	Number of Inspections	Safety	Continuous	Number of inspections in the past 12 months
36	Number of Citations	Safety	Continuous	Number of citations in the past 12 months
37	Number of Hard Breaking Events	Safety	Continuous	Number of hard breaking events captured by the device on truck in the past 12 months
38	Number of Roll Stability Events	Safety	Continuous	Number of roll stability events captured by the device on the truck in the past 12 months

Some predictor variables were interdependent. For example, three variables: citations in the past 3 months, in the past 6 months and in the past 12 months can be represented by one variable, the number of previous citations of the driver in the previous 12 months. After combining these predictor variables, the final data set has 38 predictor variables and 1 response variable. The binomial response variable is the accident flag (Yes/No). The predictor variables contain 28 continuous variables and 10 categorical variables.

The data set up is made to assure the predictions are monthly based and also provides the necessary time to act on the predictions. For example, data is set up in such a way that when considering a particular business date (usually Monday), if an accident had occurred on that business date or within 4 weeks following that date, then the driver is flagged as Y (having an accident).

3.1. Data Processing

Several preprocessing steps were undertaken to make sure that the data is ready to use for predictive modeling. Mismatch was noted between the historical data stored and the weekly new data that was collected. The other data issues include difference in the data type, extra space counted as characters, missing values, and different column names. Data preprocessing was done to combine all the collected data in one useable format.

Rather than using the entire data set for modeling, the original data set was divided into three categories namely the training set, validation set, and test set. The training set is the one on which the algorithms are trained and the models are built. Once training is complete, in order to estimate how well the model has been trained and to estimate the prediction error for model selection, the validation set is used (Friedman et al. 2001). Model assessment is done using the

test set only for the final chosen model to assess the generalization error (Friedman et al. 2001). Typically, the training set contains majority of the data in order to accommodate all possible information about the data set to provide a completely trained model. Remaining data is equally split between the validation and test set. This study follows the data split as 60% (training set), 20% (validation set), and 20% (test set).

3.2. Handling the Imbalanced Data

With the advancement of efficient classification algorithms, high computational capabilities, and vast amount of data, data exploration has grown immensely with the goal to use the data productively. A data set is considered to be imbalanced when one class outperforms the other class severely. Extreme imbalance can be in the order of 100:1, 1000:1, or 10000:1 (He & Garcia 2009). The fundamental and standard algorithms currently in use were developed with the assumption of a balanced class distribution. As a consequence, imbalanced data leads to the questionability of the prediction results because the model may not obtain the necessary information from the minority class. There could be bias in the result leading to high misclassification cost. Moreover, the focus is usually on the minority class so attention should be given in evaluating the models with appropriate performance metrics. The data set that is considered for this research also suffers from imbalance issue having a ratio of 97.5%: 2.5% representing non-accidents to accidents, respectively. There are a lot of proposed methods in the literature to handle imbalanced data (He & Garcia 2009, Chawla 2005). For the purpose of this study, 5 different common methods, undersampling, oversampling, combination of undesampling and oversampling, Randomly Oversampling Examples (ROSE), and Synthetic Minority Oversampling Technique (SMOTE) are considered. The original data with no sampling method is also tested. In reality the imbalanced data can produce a good prediction if there is a

chance that the very small minority class had acquired all the information that would make the model to perform good classification. It is always good to test the model results of the original imbalanced data set against the results of the models built using some of the data balancing techniques (He & Garcia 2009). In this study, undersampling and oversampling are tested using the caret package (short for Classification And Regression Training) in R. Undersampling generally produces a random subset from the majority class to match the number of samples in the minority class. Oversampling on the other hand creates random duplicates of the minority class to match the number of samples in the majority class. The ROSE package (short for Random Oversampling Examples) in R provides a combination of undersampling and oversampling where the resulting data set is balanced by using both the techniques simultaneously. The same package also provides a fancier and more reliable technique (i.e., ROSE) which uses smoothed bootstrap technique (Lunardon et al. (2013 & 2014), Menardi & Torelli 2014) for balancing the data set. ROSE generated balanced data set contains new samples based on the distance of the neighborhood data point instead of just duplicating the original minority class. Similar to ROSE, another most popular method, SMOTE, is based on synthetic data generation and can be implement using the DMwR package in R. SMOTE utilizes bootstrapping and the k-nearest neighbor algorithm to produce artificial data points using an interpolation strategy (Chawla et al. 2002, Branco et al. 2015). Each of these methods has its own advantages and disadvantages, and so the fit of these methods for a data set can be found by trial and error. While the traditional methods (i.e., undersampling and oversampling) produce good results, the synthetic methods are gaining more focus due to the informed way of sampling other than mere randomness, and are considered to produce better classification results.

4. METHODOLOGY

In this section, we describe the predictive modeling techniques and algorithms used in this study. Our goal is to build a highly accurate model that can incorporate reliability and interpretability of the models to the possible extent. Two open source software namely, R and H2O was used to build the models. R is an extensively popular and adaptive software having thousands of built-in packages that makes it interesting for various applications. H2O is an in-memory prediction engine for big data analysis. It has a distributed, fast and scalable machine learning and predictive analytics platform. H2O is built with machine learning algorithms that can produce models at a much faster rate with additional easy-to-use features. The H2O R package contains the functions required to connect R into H2O environment and built models. More information on how to use H2O and its functionality can be found at H2O.ai with detailed documentations. While using H2O functions for model building, the actual models are built in the H2O environment and only the results are displayed on the R console.

4.1. Generalized Linear Models

Generalized linear models (GLM) are an extension of traditional regression models. They are similar to linear regression models that do not enforce the assumptions of linearity and constant variance structures in the data (Friedman et al. 2001). As opposed to the general linear relationship between the predictor variables and the response variable put forth by a linear model, GLM combines the linear predictors which are related to the mean of the response variable using a link function. GLM response variables can take any distribution among the exponential family (Guisan et al. 2002). Generalized linear model was formulated in 1972 by John Nelder and Robert Wedderburn in an effort to unify the typical regression models like linear regression, logistic regression, Poisson regression, etc. More specific information and

mathematical proof of GLM can be found in Nelder & Wedderburn (1972), McCullagh & Nelder (1989) and Friedman et al. (2001).

Regularization can be thought as a numerical re-formulating process by introducing additional terms in the loss function to solve modeling problems. GLM can utilize regularization for better prediction results. Regularization parameters (α and λ) can be introduced into models to serve any of the following purposes: large number of predictor variables, to reduce variance of the prediction error, to avoid overfitting, and collinearity (Nykodym et al. 2016). The models utilizing regularization methods are called penalized models where lasso, ridge, and elastic net are different regularization methods that can be used. Ridge regression is also considered to be a promising alternative to stepwise approaches using the shrinkage rule of L2 norm (Tibshirani 1996, Friedman et al. 2001, Harrell 2001, Guisan et al. 2002). The collinearity problem can be handled better using model selection and regularization in GLM as opposed to stepwise model approaches (Guisan et al. 2002). The regularization parameter α (ranges from 0 to 1) controls the influence of error relative to penalty distribution between L1 norm and L2 norm, while λ (ranges from 0 to infinity) controls the penalty strength (Nykodym et al. 2016).

L1 norm is the lasso penalty, which does both parameter shrinkage and variable selection by shrinking the sums of squares of the coefficients. L2 penalty is the ridge penalty that shrinks the sum of absolute values of the coefficients towards zero. Elastic net has $\alpha \in [0,1]$ where it is the same as lasso when $\alpha = 1$ and it becomes ridge when $\alpha = 0$. The GLM binomial optimization function (Nykodym et al. 2016) for an elastic net regularization can be represented as,

$$\max_{\beta,\beta_o} \sum_{i=1}^{N} \log f\,(y_i\,;\beta,\beta_o) - \lambda\left(\alpha\,|\beta|_1 + \frac{1}{2}\,(1-\alpha)|\beta|_2^2\right),$$

where $\alpha|\beta|_1$ accounts for lasso regularization and $(1-\alpha)|\beta|_2^2$ accounts for ridge regularization. The term y_i is the prediction value (accident/non-accident); β_o is the intercept; β corresponds to the coefficients of the predictor variables (e.g., age, tenure, etc.); and *N* is the number of samples (weekly data of the drivers) in the training data.

We implemented GLM model with a logit link function and binomial distribution function. Initially the model was built with all the predictor variables where significance of the variables was tested using the p-values. Following the backward elimination procedure, the insignificant variables were removed and then the model fit was tested. The process of backward stepwise elimination was repeated until the model is left only with the significant variables. The predictor variables were tested for collinearity by checking the variance inflation factor (VIF) values. The VIF values was in the range of (1.007, 1.5) indicating no confounding effect, except for gender with 4.01 as VIF. The high value (still acceptable) of gender may be due to its categorical nature and also gender was removed from the model being an insignificant variable. Since the GLM model built was not satisfactory in terms of prediction, an attempt was made to divide the data by job family and to run individual models for each of them. The data set consists of three types of drivers: over-the road drivers (who drive on long-distance loads, typically around 12 days), regional drivers (who drive on relatively long-loads, typically more than 2 days) and local drivers (who drive radially less than 150 miles and get to go home daily or every other day). The reason to build individual models is due to the curiosity to learn if they are any interesting findings or major improvements between the drivers belonging to different job families.

In order to further study the relationship between the predictor variables and the effect based on their combination in logistic regression, models with two level interaction terms were

tested. Due to the large number of predictor variables, building interaction terms directly in R was very cumbersome as it requires high memory capacity machine to run the model. In order to simply this requirement, H2O has a function for interaction that can deal with huge number of predictor variables. H2O interaction function has a requirement that all of the predictor variables should be categorical to run the model with interaction terms. So all the continuous variables were converted to categorical variables. It was challenging to decide on the number of levels that the categorical variables should take during conversion. Probably a histogram of each of the continuous variables could have helped in deciding the levels which was not possible due to the skewed distribution of the variables. So based on the knowledge of the data at hand, levels were assigned. The result of the model with interaction terms actually performed lower to models without interaction. The reason for poor performance may be attributed to the fact that adding extraneous interaction terms would result in loss of statistical power (Williams 2015). Detecting only the useful interaction terms between large number of predictor variable is crucial which can be a separate topic of concern and so not included in this research.

As mentioned earlier, since the data set under consideration has a large number of variables, introducing the regularization parameters would further improve the model. So penalized logistic regression models using the three regularization methods, i.e., the ridge, lasso and elastic net, were tested and compared. To aid in this process, grid search was very useful, which is a technique to build set of models that have different results based on the combination of parameter values used for each model. The grid search has hyper parameters that are complex to learn directly through normal training processes. Hyper parameters are defined when a grid search is initiated. There are two hyper parameters for penalized regression: α and λ, where α determines the type of regularization that should be used and λ adjusts the penalty

strength. For the purpose of this study, only the hyper parameter α is defined in the grid search, and the H2O software automatically selects the appropriate value of λ for each of the models in the grid search. Automatic selection of λ is considered to be more appropriate than manual entry (Nykodym et al. 2016). A grid search with α values ranging from 0 to 1 in 0.01 increments was carried on. As a result, this grid search output has 101 models with different values of α and λ.

4.2. Random Forest

Random forest is considered to be the user friendly and handy machine learning technique irrespective of the type of the data set and prior level of knowledge in predictive mechanism (Zhou and Hooker 2016, Biau and Scornet 2016). Random forest builds multitude of decision trees using the bagging strategy and then classify a sample by the mode or majority prediction of all the trees (Random Forest 2016). Bagging, also known as bootstrap aggregation, is a model averaging method by reducing the variance while retaining the bias (Friedman et al. 2001). In a nutshell, random forest is a fancier version of bagging where it averages approximately unbiased models with de-correlated trees to reduce variance. The idea of random decision tree was first proposed by Tin Ho in 1995 to overcome the problem of growing trees with traditional method. The goal was to increase the accuracy on both the training data and new/unseen data. The limitation on training the complex data is compensated by growing multiple trees, each having randomly selected feature space (Ho 1995). Later in 2001, Breiman developed random forest by combining two important aspects of machine learning such as bagging and feature selection (Breiman, 2001).

In bagging, each of the models developed pulls off a random training set that is bootstrapped from the training data. In contrast to the boosting method, where shallow trees are used to solve for the classification errors by learning from the previous trees, bagging mainly

concentrates on the diverse subset of the training data to grow relatively deep trees. Trees can capture the complex interaction structure (Friedman et al. 2001) in the data and can reduce bias if grown deeper. The sampling of the training data with replacement has an equal chance for all the samples to have multiple occurrence or no occurrence at all. The idea of perturbing the training data to achieve diverse model is very important in bagging. Since the data sets using in machine learning are usually large and multiple re-sampling is done, the bias is lower in a tree construction. The variance in the model is reduced by averaging the predictions of the number of trees built. There are three main factors to reduce variance as noted by Zhang & Haghani (2015): decrease correlation between any pair of trees, strengthening the individual performance of the tree and increasing the total number of trees in the forest.

Due to the fact that each tree is grown from the samples with replacement, the learning process tends to be on the same track introducing some bias. This problem is overcome by using the random feature selection process as introduced by Ho (1998, 1995) and Amit & German (1997). In feature selection, only subsets of the features are selected at each splitting node of the tree. Instead of the very few variables that dominate during the splitting process, feature selection allows most of the variables to take the role of the splitting node.

Introduction and mathematical details of the random forest algorithm for classification problems can be found in Friedman et al. (2001) and Zhang et al. (2015). We briefly summarize the algorithm below. Assume the training data set consists of a total number of r samples and total number of p predictor variables. Before starting the algorithm, the number of trees to be grown, M, and the number of selected predictor variables q is initialized. The number of selected predictor variables stays constant for all the trees that are grown and $q < p$. For classification, the

default value of q is $\sqrt{p}$ and the minimum node size is 1. Let $\hat{P}_m(x)$ be the prediction of the mth tree in the random forest, then for all the trees in the random forest the result is expressed as

$$\hat{P}_m^M(x) = \text{majority vote } \{\hat{P}_m(x)\}_1^M.$$

Majority voting is nothing but selecting the mode or majority decision of all the trees built in the random forest as the final prediction result. In order to build the best random forest model, various parameters values were tested based on trial and error with proper understanding of how each parameter would help in building a better model. The following are the parameter values used to tune the best model (for example, when growing more trees or less trees than the one mentioned below either did not improve the model or performed poorly),

1.) Number of tress (M), ntrees = 101

2.) Maximum depth of the tress, max_depth = 50

3.) Number of variables at each node (q), mtries = 15

4.) Number of rows to be selected at each tree (p), sample_rate = 0.75

4.3. Gradient Boosting Machine

Gradient Boosting Machine (GBM), one of the popular machine learning techniques in the current era of predictive analytics, is based on the concept of strategically combining the weak learning results to form a model of high accurate prediction rule (Gradient Boosting 2016). While GBM has the same advantages as other popular machine learning models (e.g., robustness to less clean data, less data preprocessing, handling missing values, feature selection and accounting for complex model interactions), it also has an added advantage of better model interpretability with less parameter tuning compared to other methods of machine learning (Guelman 2012, Zhang & Haghani 2015). The disadvantage of GBM is that it is a greedy algorithm that can overfit the training data easily and has scalability issues (Scikit Learn 2016).

The concept of combining the weak classifiers through a system of boosting was put forth by Schapire and he proved the equivalence of the weak and strong learnability. The accuracy of a strong classifier using a probably approximately correct (PAC) learning is similar to the weak learning method, which is better than a random selection executed through a boosting mechanism (Schapire 1990). Practically, it is also computationally easy to develop a shallow tree like a stump that has a single split and two leaves, which forms a weak learner. Intuitively, the idea of boosting works better because there is a higher probability for a hard sample (i.e., with classification error) to occur multiple times in the model (Zhang & Haghani 2015). These misclassified samples reoccur multiple times gaining higher weights. Boosted trees are not identically distributed due to the adaptive nature and hence reduce bias greatly (Friedman et al. 2001). GBM is based on the constructive strategy that each consecutive tree built is fitted solving for the net error of the prior trees. This can be explained in simple steps such as choosing the loss function based on the output (regression or classification), creating a base model for learning (like a stump model which is a tree with a single split), and using an additive model that can add the trees at each successive steps using procedures like the gradient descent to reduce the loss function. The statistical formulation of GBM and the algorithm used to reduce the loss function is explained in detail by Friedman (2001) and Friedman et al. (2001). A simple overview of the steps in GBM is explained below.

Consider the data set in the form $(x, y)_{i=1}^{N}$ where $x = (x_1, x_2, \dots \dots, x_m)$ and the binomial output y which is to be used for supervised learning. Here x refers to the set of inputs or the explanatory variables (e.g., age, gender, tenure), y refers to the corresponding output or the response variable representing an accident or non-accident and N is the number of samples in the data set. Functional dependence is mapped from x to y to obtain an approximation such that the

objective is to minimize the loss function over the joint distribution of all values of (y, x). Being a binomial function the response variable is coded as $y = \{0,1\}$ where the classifier can take only one of these two values.

From a boosting tree perspective, the predictor variables partition the total space into disjoint regions $R_j, j = 1,2, \dots, J$ representing the terminal nodes. Friedman et al. (2001) assigns a constant γ_j to each such region based on the joint values of the predictor variables such that, $x = R_j \; and \; f(x) = \gamma_j$. So the predictive rule is between the joint values of the predictor variables and the resulting prediction of the response variable (Friedman et al. 2001). The formulation of a tree can be expressed as $T(x; \Theta)$ where the parameter $\Theta = \{R_j, \gamma_j\}_1^J$.

Optimization of the parameters can be divided into two parts as γ_j and R_j (Friedman et al. 2001). Generally, γ_j is the mean of all y_i falling in the corresponding region R_j and also, finding R_j entails γ_j. The additive or the sum of the boosted tree is represented as $f^M(x) = \sum_{m=1}^M T(x_i\,; \Theta)$. The procedure is followed in steps in a forward stagewise manner solving for all the iteration m, of the set (R_{jm}, γ_{jm}) having the current model as $f_{m-1}(x)$ and y_i is the actual classification label (Friedman et al. 2001).

$$\widehat{\Theta} = arg \min_{\Theta_m} \sum_{i=1}^{N} L(y_i\,, f_{m-1}(x_i) + T(x_i; \Theta_m)\; \gamma$$

As a sum of all the tree at each step, the estimate f^M is given as $f^M = \sum_{m=0}^M f_m$ where f_0 is the initial guess which boosts up to M (Friedman et al. 2001). Steepest gradient descent is based on consecutive improvements to reduce the loss function such that $f_m = -\rho_m g_m$, where the parameter ρ_m is a scalar representing the step length and g_m is the gradient of the loss function $L(f)$ at $f = f_{m-1}$ and also $g_m \in IR^N$(Friedman et al. 2001). Conceptually, gradient boosting tree is dependent on the previous trees. One of the main differences is that each

consecutive tree is fitted solving for the net error of the prior trees. The tree function is $T(x;\ \Theta_m)$ with m iteration, where the predictions close to the negative gradient, and it follows the least squares minimization, $\widetilde{\Theta_m} = arg\min_{\Theta} \sum_{i=1}^{N}(-g_{im} - T(x_i;\ \Theta))^2$

In this study, the main parameters used to build the best GBM models and its corresponding values are as follows:

1.) Number of trees, ntrees = 1001

2.) Maximum depth of the tree, max_depth = 50

3.) Learning rate used, learn_rate = 0.2

4.) Row sample rate, sample_rate = 0.75

5.) Column sample rate, col_sample_rate = 0.75.

4.4. Deep Learning

Deep learning is an improved version of neural network with multiple hidden layers consisting of both linear and non-linear transformations to solve complex problems for which high-level data abstraction is required. Deep learning application can be found in areas like image recognition, automatic speech recognition, robotics, etc. Shallow networks are more expensive compared to deep networks because the neuron function computation in deep networks follows a subroutine concept (Le 2015), which can be re-used multiple times.

The birth of neural nets dates back to 1943 based on computational models (Pitts & McCulloch 1943) and later it was developed based on an algorithm having a threshold logic (Piccinini 2004) where each neuron has an excitatory or an inhibitory level which determines whether they are active or not. Various improvements and findings to this initial neural net (Anderson & Rosenfeld 1988, Hebb 1949, Johnson & Brown 1988) were explored which led to the flourishing growth of neural nets in various application areas. The two main drawbacks of

neural net that led to the diminishing use of the initial neural nets are the lack of machines with high processing capability and the inability of processing an exclusive-OR circuit with single layer perceptron (Monsky & Papert 1969). Two key algorithms called the perceptron (Rosenblatt 1958) and backpropagation (Werbos 1975) played the key role in advancing the neural net to the next level.

Basically a neural network consists of an input layer, hidden layer and output layer where all layers are fully connected. Each layer has neurons or cells. The neurons of the hidden layer consist of an activation unit which is a function of the input neurons. The connection between layers are given by weights. Each node within hidden layer has a sigmoidal activation function which is bounded between 0 and 1. The weights are determined by assigning various learning rates. Since the weights can be adjusted until the learning is complete, neural net is also termed as adaptive system. Deep learning is applicable to two types of learning, the supervised and unsupervised learning. This research is concentrated on the supervised learning approach where a set of data is given to the algorithm, based on which the learning happens in sequential steps. In this study, the input layer consists of neurons that represent the predictor variables (i.e., age, gender, incidents, complaints, tenure, etc.) and the output neuron is either 0 (non-accident) or 1 (accident). Each row from the training set passes through this network in-order to be classified.

Considering one row of the data (weekly record of the driver) at a time, the decision function $h(x)$, which can be considered as a weighted linear combination of the predictors (Le 2015), can be represented as

$$h(x;\, \theta, b) = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_{38} x_{38} + b,$$

where $\theta_1, \theta_2, \ldots, \theta_{38}$ are the weights associated with the 38 corresponding predictor variables as listed in Table 1, and b is the bias. The goal is to reduce the classification error in each step by

finding the values of two parameters, θ and b such that these parameters minimizes the following objective function,

$$J(\theta, b) = \sum_{i=1}^{n} (h(x^{(i)}; \theta, b) - y^{(i)})^2,$$

where $x^{(i)}$ corresponds to weekly record of the driver and $y^{(i)}$ is the accident label for that driver. In order to minimize the objective function, the parameters (θ and b) are iteratively updated using a non-negative scalar quantity α in the direction of global minima such that $\theta = \theta - \alpha\Delta\theta$ and $b = b - \alpha\Delta b$. This process is called stochastic gradient descent (SGD), and α is known as the learning rate and satisfies the following relationship,

$$\alpha = \sum_{i=1}^{m} \theta x + b$$

SGD initializes the parameters and assigns it to each pair $(x^{(i)}, y^{(i)})$ where it follows chain rule and partial derivatives to update the decision function $h(x; \theta, b)$ (Le 2015). In order to reach the global minimum with least error, the backpropagation algorithm is used so that each step is directed to the steepest value of the vector surface. This nonlinear multi-layer feedforward backpropagation network is referred to as the deep learning architecture. More detailed formulation and mathematical proofs can be found in Friedman et al. (2001) and Le (2015). While modeling with the deep learning algorithm, two important parameters need to be chosen are: 1) the number of hidden layers along with number of neurons in each layer, and 2) the learning rate. In neural nets, although having many hidden layers leads to additional cost, it is commonly recommended as it better captures the nonlinearities (Le 2015), and additionally, not having enough hidden layers may result in incomplete learning. A common practice to find the optimal hidden layer and the number of neurons in each layer is by trial and error method and

then estimating the model. Similarly, having a large learning rate may miss an actual global minimum while having a small learning rate can be too conservative leading to very slow tuning. To select a good learning rate, the key is to monitor the training where α of 0.01 is a good start (Le 2015).

Instead of using the sigmoidal activation function, another improvement can be made by using a rectified linear activation function (Nair & Hinton 2010), which will lead to a better approximation than the sigmoidal function. The rectified linear activation function is given by $f(x) = \max(0, x)$ and sigmoidal function is given by $f(x) = \frac{1}{1 - e^{-x}}$ where x is an input to the neuron. Considering the sigmoidal function, the range of f(x) is between [0,1] so the gradient of this function vanishes as the value of x increases or decreases, whereas the rectilinear activation function has a range between [0,∞] leading to a gradient function that vanishes only if x decreases. Due to this property the rectified linear function increases sparsity and dispersion of the hidden layer that helps to improve the performance with better approximation quality (He et al. 2015, Mass et al. 2013, Le 2015). Generally deep learning works very well for really huge data set. Although the data set used in this research is big, it is not considered huge compared to the capability of deep learning algorithms. Since the data has non-linear distributions and a non-linear output function (binomial), deep learning provides a different perspective from random forest and gradient boosting algorithms which are tree based.

In this study, the main tuning parameters used in building the best deep learning model are,

1.) The number and size of each hidden layer, or the hidden_layer_size is set at (2048, 2048) after trial and error representing two hidden layers with 2048 units each.
2.) Activation function used is the rectifier activation function

3.) Number of iterations (i.e., epochs) are set to be 100

Some of the additional tuning parameters used in both GBM and deep learning in order to save time and to obtain better flexibility are stopping rounds, stopping metrics and stopping tolerance. The stopping metric can be logloss, MSE, AUC, etc. where stopping round handles the early stopping concept based on stopping metric. Early stopping is a form of regularization technique to avoid overfitting. Metric-based stopping criterion is defined by a relative tolerance criterion called the stopping tolerance. The model stops if the relative improvement is not equal to the defined criterion. More information on these parameters can be found in Click et al. (2016) and Candel et al. (2015).

4.5. Model Validation

Once building the model after training, appropriate model evaluation is necessary. For model evaluation, performance metrics of the resulting model should be studied. Although models built from imbalanced data can produce high overall accuracy, the sensitivity may be low due to the low presence of accidents compared to non-accidents in the data. Especially while evaluating a model using the validation set, per-class-accuracy may be more informative compared to the overall accuracy. For GLM, some of the metrics that can be used are deviance, Akaike Information Criterion (AIC), Hosmer–Lemeshow test, etc. Deviance is the difference between the maximized log-likelihoods of the fitted model and the saturated model, where too large value explains that model is not a good fit (Nykodym et al. 2016). The AIC score depends on the number of parameters in the model and so it is not a good indication of model fitness but helps in model comparison (Nykodym et al. 2016). Hosmer–Lemeshow test is a goodness-of-fit test for logistic regression models. The Hosmer–Lemeshow output has a p-value between 0 and 1 with higher values indicating a better fit (Allison 2014).

Classification based machine learning model evaluation techniques such as the ROC chart and area under the curve (AUC), confusion matrix, and F1 score are widely used (Chawla 2005). AUC is computed using all possible values of the classification threshold (i.e., the cut-off value to decide whether the sample should be classified as an accident or non-accident). AUC produces a summarized curve showing the worst and the best classification of a binomial class as opposed to other metrics which use a particular threshold. So AUC is a reliable measure for choosing the best model. The best possible classification is obtained based on the optimal cutoff point which is the value that corresponds to the minimum distance to the upper left corner (0,1) on the ROC chart (Hajian-Tilaki 2013). The minimum distance is calculated as

$$\text{Minimum distance to } (0,1) = \sqrt{(1 - \text{sensitivity})^2 + (1 - \text{specificity})^2}.$$

The confusion matrix represents the false positive, true positive, false negative and true negative values directly. Although sensitivity (recall or true positive rate) and specificity (true negative rate) can be directly read from the confusion matrix table, indirect measure that can be calculated using the confusion matrix like precision, F1 score, dominance, etc. can be more useful (Braince et al. 2015). Mean squared error (MSE) is also a very good measure which shows the difference between the mean squared error of the predicted value and the actual value.

Given different models and different performance metrics, the ultimate goal is to select an appropriate model to place in production. Focusing on the data at hand and the presence of imbalance data predicting accidents is more important than non-accidents. Potentially the accuracy of the model may be still high with low sensitivity due to the proportion of the actual accidents compared to non-accidents. It is also important to remember that only the training data is balanced and not the entire data set. Multiple models are built using the grid function available in H2O and R to reduce manual efforts. All the models are estimated using the validation set and

the one that has the best performance indicated by a good AUC value, balanced sensitivity and specificity along with high accuracy is selected. As preventing as many accidents as possible is the primary purpose, it is worth sacrificing the overall accuracy to improve sensitivity. To be more specific, in order to capture more accidents that would cause huge cost and risk of a life, lowering the specificity is acceptable as training more drivers is less costly compared to dealing with one accident.

5. RESULTS

The results from all the models developed using the algorithms and the various data balancing techniques discussed above are presented in this section. Overall, the oversampling method for balancing the data worked the best for this data set.

5.1. Results of Generalized Linear Models

Using the traditional logistic regression, 16 predictor variables was found insignificant (at significance level of 0.05) and removed. The AUC value of the validation set was less than 0.66. The p-value from the Hosmer–Lemeshow goodness-of-fit test was close to zero (< 2e-16) indicating a bad fit. Similar results were observed when separate regression models were run for each job family. Table 2 summarizes the coefficient estimates, standard errors, and the corresponding p-values for the significant variables from the final logistic regression model built using stepwise backward elimination process. It can be seen that ethnicity has a surprisingly huge impact on prediction as show by the coefficient estimates. Other important variables in terms of their estimates are board turnover, number of accidents, and number of failures all with a positive sign.

Table 2: Results for the Logistic Regression Model

Variable	Estimate	Standard Error	p-value
(Intercept)	-3.7700	0.0788	< 0.0001
Age	0.0110	0.0007	< 0.0001
Tenure	-0.1100	0.0023	< 0.0001
factor(Ethnicity) Asian	0.3200	0.0847	0.00016
factor(Ethnicity) Black	0.3460	0.0681	< 0.0001
factor(Ethnicity) Multiple	0.4210	0.1210	0.0004
factor(Ethnicity) Not Specific	0.3970	0.0701	< 0.0001
factor (Ethnicity) White	0.2270	0.0677	0.0008
factor (401K max match)Y	0.0487	0.0142	0.0005
ClassA Experience	-0.0002	0.0000	< 0.0001
ClassB Experience	-0.0001	0.0000	< 0.0001
factor (Job Family) OTR	-0.2330	0.0443	< 0.0001
Number of Accidents	0.0705	0.0045	< 0.0001
Number of Complaints	0.0464	0.0080	< 0.0001
Number of Incidents	0.0186	0.0025	< 0.0001
Number of Observe	-0.0146	0.0031	< 0.0001
Number of Inspections	-0.0488	0.0051	< 0.0001
Number of Roll Stability event	0.0263	0.0056	< 0.0001
Number of Fuel runouts	0.2630	0.0702	0.0001
Number of Failure	0.0602	0.0058	< 0.0001
Weekly pay	-0.0002	0.0000	< 0.0001
factor(Business_unit) JBI	-0.0676	0.0191	0.0004
factor(Business_unit) VAN	0.1080	0.0243	< 0.0001
Number of drivers per board	0.0037	0.0004	< 0.0001
Board turnover	0.0823	0.0049	< 0.0001
Number of Miles per stop	-0.0003	0.0000	< 0.0001
Number of miles driven	0.0003	0.0000	< 0.0001
Number of Loads	-0.0133	0.0018	< 0.0001

These estimates are reasonable because a driver having many failures shows his/her lack of responsibility; having many past accidents indicates the requirement of training on precautions and defensive driving; and a driver in a board with high turnover rate (each broad represents a group of drivers, typically 12 drivers or more, where turnover rate is the percentage of drivers

leaving the company during a time period and low turnover rates are expected to maintain consistency of drivers) might relate to the lack of responsibility of the fleet manager or any related complaints not being addressed leading to dissatisfied drivers. But, the remedy to decrease accidents cannot be relied solely on the variables that have the higher absolute values of the coefficient estimates.

Penalized logistic regression was carried out for the entire training set and by job family. Unfortunately, the performance metrics for none of these models indicated a good fit. Tables 3 and 4 show the AUC and MSE values of the validation set respectively produced by the penalized regression models for the entire training set. We keep more decimal places in the table to capture the precise performance between the models. Table 4 indicates that ridge regularization with oversampling is the best model having an AUC of 0.65 although all models have poor performance.

Table 3: AUC Values of the Validation Set for Penalized Regression Models

Validation Set - AUC	Lasso	Ridge	Elastic Net
Oversampling	0.6521903	0.6522428	0.6521956
Undersampling	0.6518728	0.6518055	0.6518885
Both Sampling	0.6445556	0.647471	0.64466
ROSE	0.6385426	0.6376074	0.6377037
SMOTE	0.628624	0.6288381	0.628627
No Sampling	0.6385426	0.6376074	0.6377037

Table 4: MSE Values of the Validation Set for Penalized Regression Models

Validation Set - MSE	Lasso	Ridge	Elastic Net
Oversampling	0.2342191	0.2342158	0.234219
Undersampling	0.2342928	0.2342906	0.2342927
Both Sampling	0.2402303	0.226971	0.240164
ROSE	0.2339009	0.2249245	0.2362407
SMOTE	0.1903037	0.1902917	0.1903319
No Sampling	0.2339009	0.2249245	0.2362407

The models were also validated based on MSE values as shown in Table 4. Although MSE for the models using the SMOTE sampling method had the lowest errors, their corresponding AUC values were the lowest, around 0.62.

Figure 1 shows the standardized coefficients for the best penalized logistic regression model, which is the ridge regularization using the oversampling data set. The blue bars correspond to the positive coefficients and the red bars correspond to the negative coefficients. Standard coefficients are useful in comparing the relative importance of each predictor in the model. It can be seen from the graph that variables like tenure, classA experience, number of miles per stop, etc. has negative coefficients similar to the logistic regression model indicating less chance for an accident. As the drivers' tenure with the company increases and their experience of driving a classA truck increases, the driver would be less prone to an accident as a result of good training programs. On the other hand, it is intuitive that drivers who have been involved in accidents should participate in more trainings. It also suggests that the aging drivers may lack physical strength and concentration, which may lead to a higher probability of an accident.

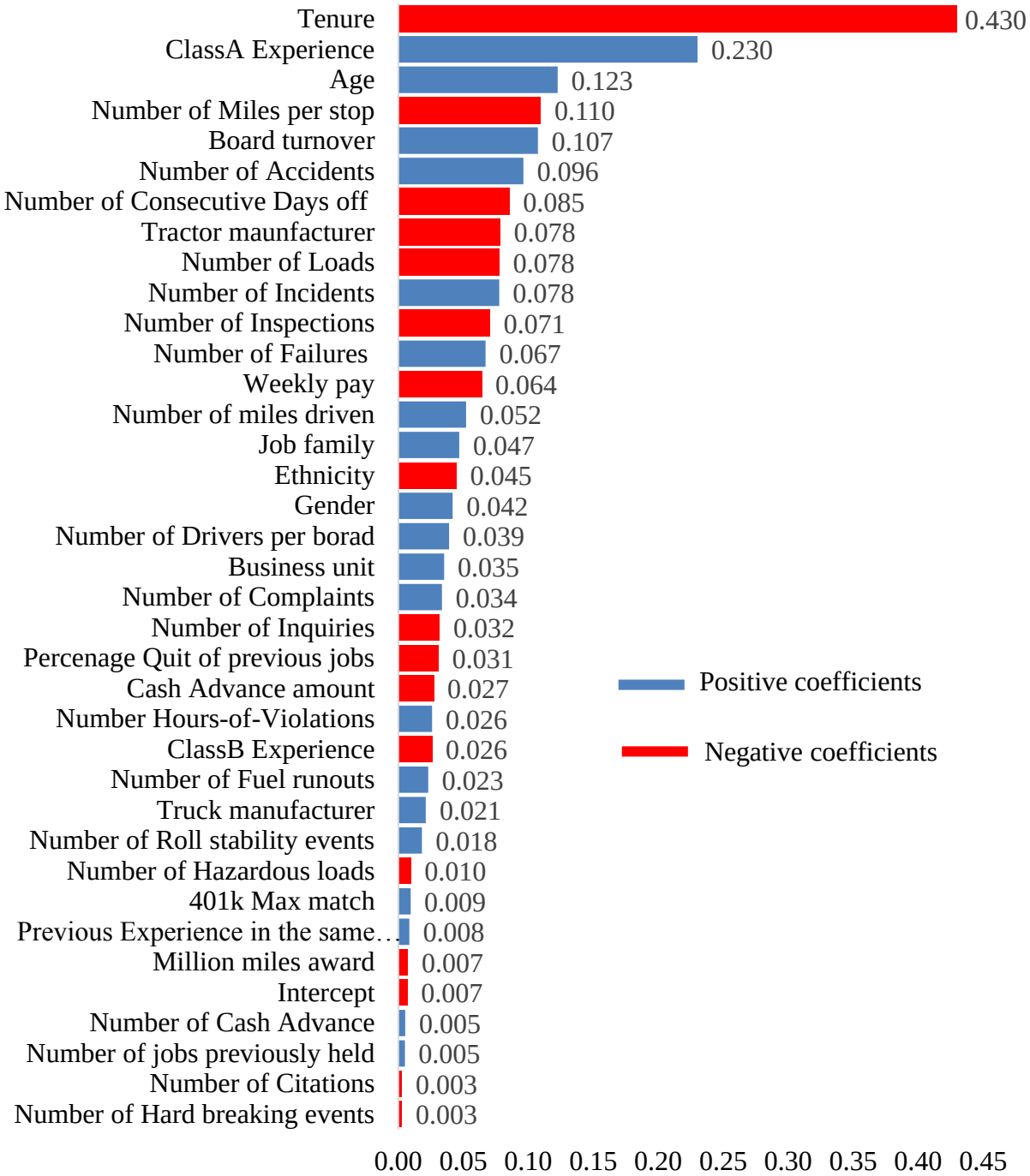

Figure 1: Standardized Coefficients of Penalized Logistic Regression (Ridge Regularization)

Figure 2: ROC Curve for Penalized Logistic Regression (Ridge Regularization)

Table 5: Confusion Matrix for Penalized Logistic Regression (Ridge Regularization)

		Predicted		
		N	Y	
Actual	N	201482	141331	Specificity= 58.77%
	Y	3232	5847	Sensitivity= 64.40%
		NPV[a] = 98.42%	PPV[b] = 3.97%	Accuracy= 58.92%

[*] NPV = negative predictive value; [b] PPV = positive predictive value. Same abbreviations are used for future tables.

Note that this can be misleading that young drivers are not prone to accidents, which is not true according to the statistics (NHTSA 2008, Curry et al. 2014). Therefore, finding a direct relationship with the signs of the regression is not very useful and might be confusing. In the future research, results can be compared between age groups (i.e., young drivers versus old drivers), but since the performance of the best model among (ridge regularization with

oversampling) penalized regularization was not satisfactory on the test set (as indicated by the ROC curve in Figure 2 with AUC = 0.65, we did not perform additional analysis using regression. Additionally, as shown in the confusion matrix (Table 5), the model has low sensitivity (64.40%) and overall accuracy (58.92%). The negative predictive value (NPV) and positive predictive value (PPV) are usually highly affected by the imbalanced data (Vihinen 2012, Gagliano et al. 2015). PPV or precision is very low (3.97%) as opposed to very high NPV (98.42%) as the result of a very small size of the minority class (positive / accident class) as compared to the majority class (negative / non-accident class).

5.2. Results of Machine Learning Methods

Each of the machine learning algorithms used in this study were tested independently using the different data balancing techniques discussed in Section 3.2. Table 6 represents the results of different models in terms of AUC and MSE values for the validation set. Similar to performance of the generalized linear models, the oversampling method had the best result with AUC at 0.95 for random forest, 0.94 for gradient boosting machine and 0.89 for deep learning. The MSE for these respective models were also the lowest as desired.

Table 6: AUC and MSE Values of the Validation Set for Machine Learning Methods Using Data Balancing Techniques

Validation test - AUC	**No Sampling**	**Under-sampling**	**Over-sampling**	**Both**	**SMOTE**	**ROSE**
Random Forest	0.9548	0.9058	0.9568	0.7626	0.9239	0.6233
GBM	0.9402	0.9133	0.95.43	0.7239	0.9120	0.6456
Deep Learning	0.5988	0.8209	0.9004	0.6883	0.8525	0.6544
Validation test - MSE	**No Sampling**	**Under-sampling**	**Over-sampling**	**Both**	**SMOTE**	**ROSE**
Random Forest	0.01593	0.16457	0.01563	0.02318	0.090070	0.066380
GBM	0.01888	0.17497	0.01648	0.02294	0.067114	0.042200
Deep Learning	0.02713	0.22866	0.03341	0.05870	0.162490	0.056360

Random forest among all machine learning algorithms has many advantages such as efficient runs on large databases, high predictive power, fast speed, and the ability to produce good results without data preprocessing (Krishnaveni & Hemalatha 2011, Li et al. 2008, Xie et al. 2007). This is reflected in the results as random forest performed better compared to gradient boosting and deep learning. Figure 3 shows the ROC curve of the random forest test set with an AUC of 0.95.

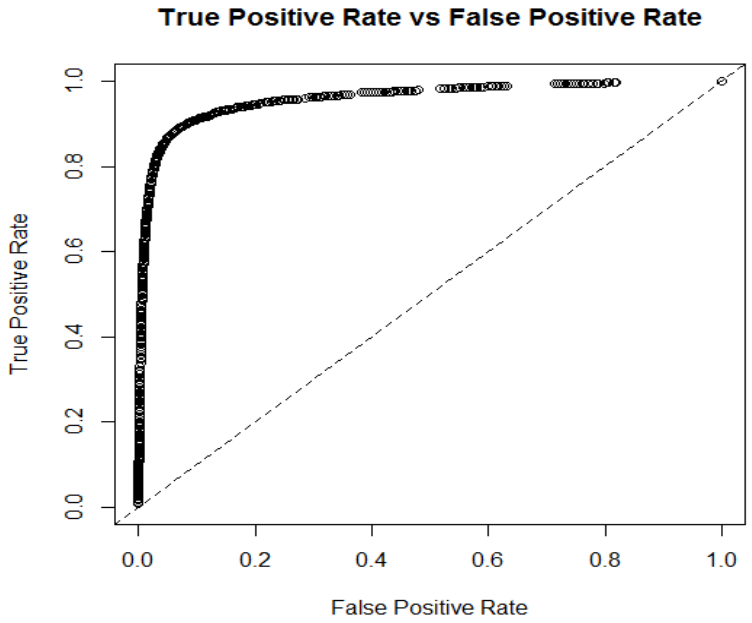

Figure 3: ROC of the Test Set Using Random Forest

The overall accuracy of the model where it correctly predicts both the accidents and non-accident events is 91.12%. The model specificity is 90.16% and the sensitivity is 89.78% as shown in Table 7. The same discussion on NPV and PPV as for the penalized logistic regression holds for all the machine learning methods due to the data imbalance issue.

Table 7: Confusion Matrix for Random Forest

		Predicted		
		N	Y	
Actual	N	312503	30310	Specificity= 90.16%
	Y	928	8151	Sensitivity= 89.78%
		NPV = 99.70%	PPV = 21.19%	Accuracy= 91.12%

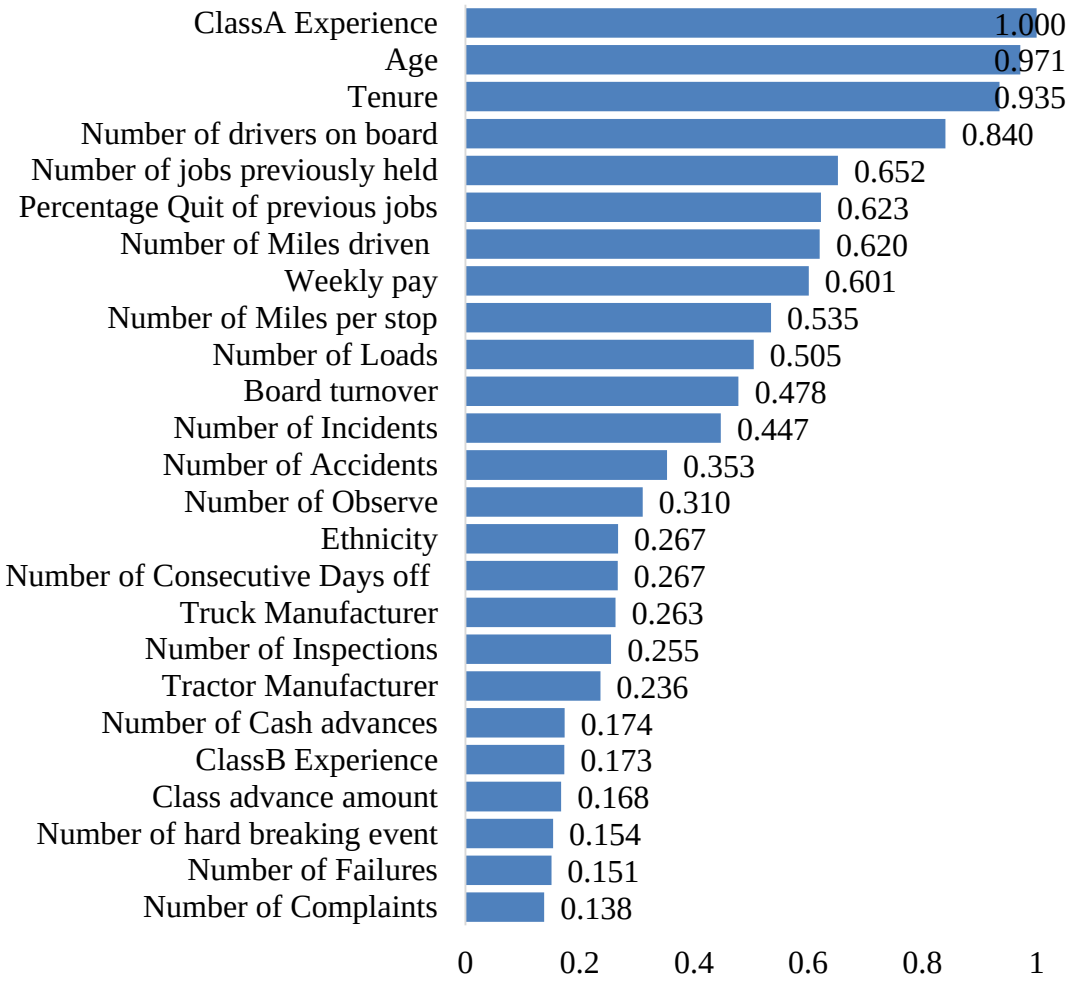

Figure 4: Variable Importance of the Test Set Using Random Forest

The Figure 4 shows the top 25 scaled variable importance of the predictor variables produced by random forest where the top 5 variables are classA experience, age, tenure, drivers per board and average miles driven per week. This aligns well with the results of the penalized regression standard coefficients (Figure 1). Although the absolute values are different from the regression coefficients, the most important variables seem to be the same, just in a slightly different order. Inference from random forest variable importance can be summarized as to enhance the experience of classA truck drivers, focusing on age groups of drivers, improving drivers' tenure with the company, having a balanced number of drivers per board as per the demand, etc. This proves why machine learning techniques are called black box because more detailed information is difficult to capture as the signs and the magnitude are not defined.

Gradient boosting machine has similar results as random forest. As discussion on GBM and its advantage has been provided in Section 4.3, the results of the model are shown below. The overall accuracy of the GBM model on test set is 91.56%. As shown in the confusion matrix (Table 8), the specificity is 91.64% and the sensitivity is 88.48%. The ROC curve is shown in Figure 5 having an AUC of 0.95.

Table 8: Confusion Matrix for GBM

		Predicted		
		N	Y	
Actual	N	314170	28643	Specificity= 91.64%
	Y	1046	8033	Sensitivity= 88.48%
		NPV = 99.67%	PPV = 21.90%	Accuracy= 91.56%

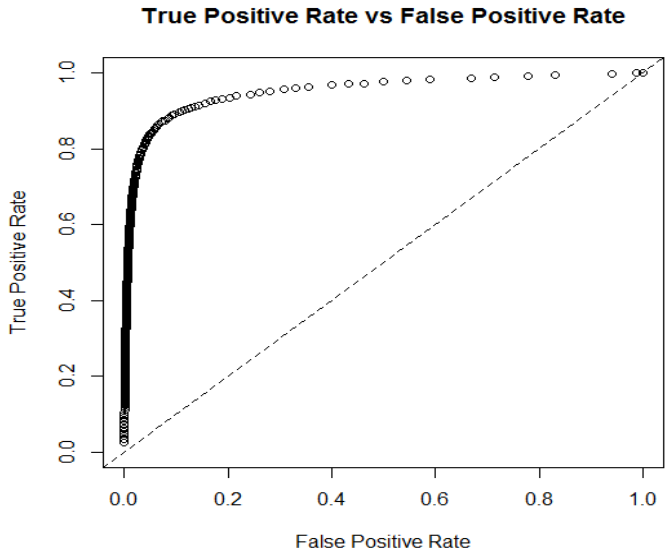

Figure 5: ROC of the Test Set Using GBM

Similar to random forest, variable importance for GBM is shown in Figure 6. It is interesting that random forest and GBM has exactly the same top 6 variables with a little difference in ranking. The interpretation of variable importance is also similar to random forest. Since the relative variable importance for the regression and machine learning methods are mostly similar, it is an evidence that statistical methods and machine learning methods lead to similar findings. In reality, the goal is to take necessary actions based on predictions, and thus, the order of variable importance is less critical. Accurate flagging of drivers who are at more risk of an accident is the ultimate purpose of running these models. As most transportation companies have safety reports for drivers, training and safety programs do not just focus on one important factor identified from the models.

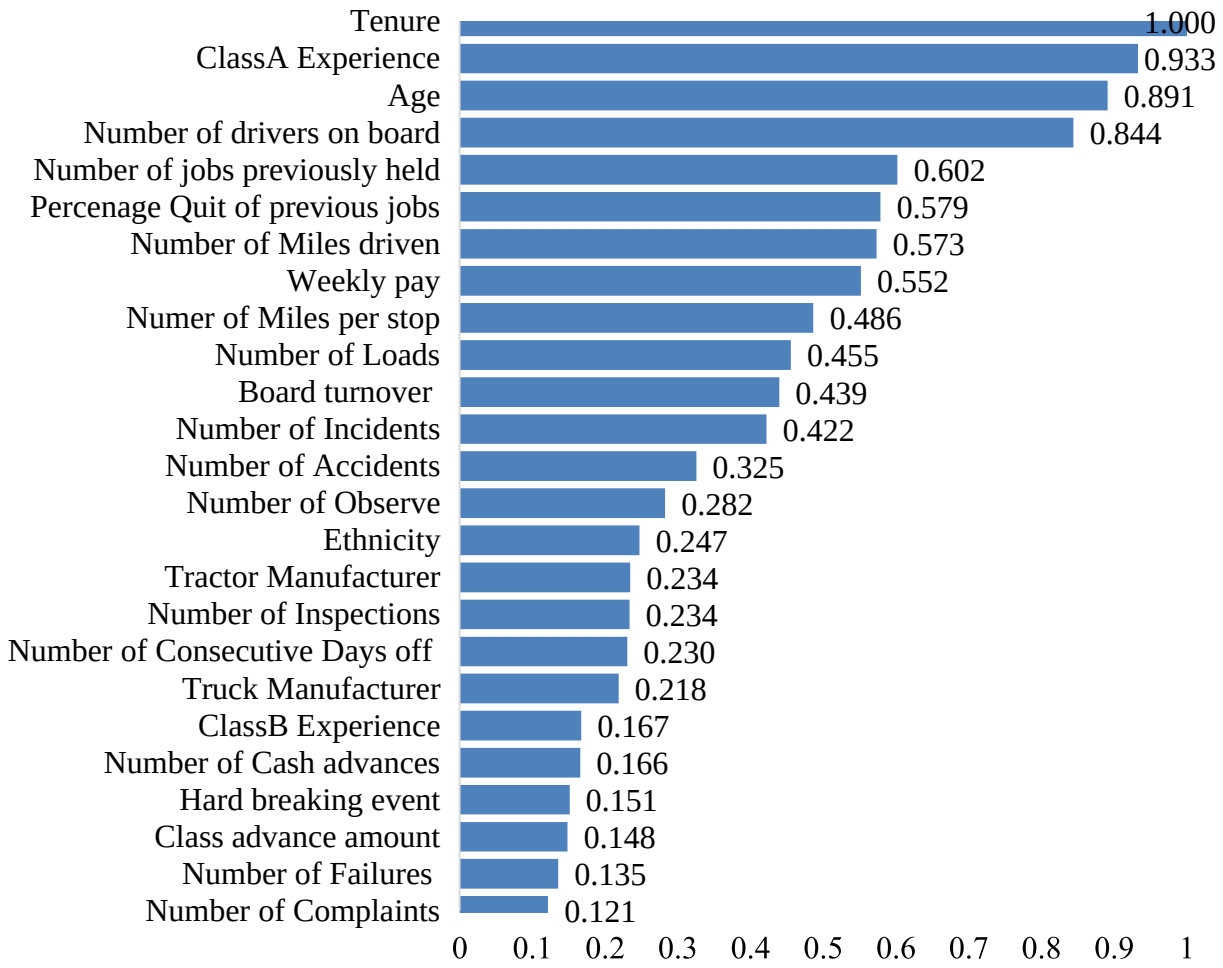

Figure 6: Variable Importance Chart of the Test Set Using GBM

Next, the results for the deep learning method are presented similar to the other methods. Deep learning method does not generate variable importance like the tree based models due to the inherent nature of the algorithm. The interpretability is still more confined in deep learning as these models concentrate more on accurate classification rather than finding the relationships.

The AUC was found to be 0.90 as depicted in Figure 7 with overall accuracy of 85.28% assuring to be the best model. The sensitivity and specificity of the model were found to be 81.91% and 85.37%, respectively as shown in Table 9.

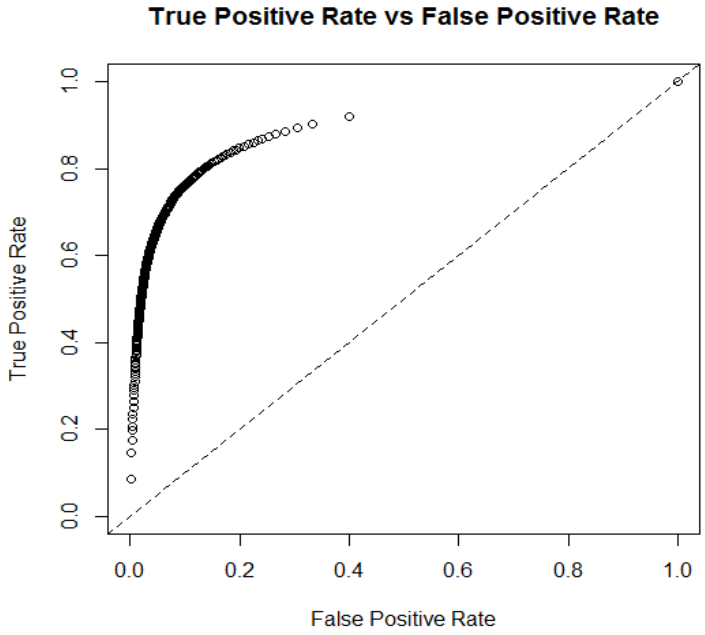

Figure 7: ROC of the Test Set using Deep Learning

Table 9: Confusion Matrix for Deep Learning Model

		Predicted		
		N	Y	
Actual	N	292658	50155	Specificity= 85.37%
	Y	1642	7437	Sensitivity= 81.91%
		NPV = 99.44%	PPV = 12.91%	Accuracy= 85.28%

Based on the results shown above, random forest had the highest sensitivity and accuracy as desired. Random forest and GBM are tree-based models which produces similar predictions compared to deep learning. Considering accidents predicted by random forest, 65.5% of the time GBM also agrees, while deep learning only predicts 54.7% of these. Random forest covers 70.1% of GBM predicted accidents, while deep learning agrees only 58.3%. Random forest and GBM do not agree with deep learning's accident classification 62.9% and 63.6% of the time,

respectively. Thus, the predictions produced from the three methods vary significantly, but they all predict significantly more accidents than actuals as depicted by the low PPV values. Among these methods random forest can be implemented since it has the best sensitivity. However, in order to gain higher accuracy and take advantage of other algorithms, these individual models can be combined to obtain better predictions. Using an ensemble combination rule based on confidence estimation (Polikar 2009), if majority of classifiers agree with a decision (Y/N), such an outcome can be interpreted as high confident ensemble. On the other hand, if half of the classifier predicts Y and other half of the classifier predicts N, it is termed as low confident ensemble. According to Polikar (2009), when the independent classifiers outputs are combined for majority voting, the result of majority ensemble always lead to performance improvement. As an example, Table 11 shows the confusion matrix of majority voting ensemble combining the result of random forest, GBM and deep learning. It proves that sensitivity, specificity, and accuracy are better than any of the individual models. However, the improvement is less than 1% using ensemble compared to the best individual model, it can be interpreted that since ensemble uses combination of different algorithms the output can be relied better than the output from one individual model. Using a different cutoff for the individual models the majority ensemble voting can be improvement .

Table 11: Confusion Matrix for Majority Voting Ensemble

		Predicted		
		N	Y	
Actual	N	314115	28698	Specificity= 91.63%
	Y	918	8161	Sensitivity= 89.89%
		NPV = 99.80%	PPV = 10.55%	Accuracy= 91.58%

The above metrics in the confusion matrix are obtained based on the cutoff value that produces the minimum distance to the (0,1) point in the ROC chart. This may be due to the assumption of equal cost assigned to accidents and non-accidents. However, in reality the cost of false negative or type II error (i.e., not predicting an actual accident) should be significantly higher to the cost of false positive or type I error (i.e., wrongly predicting a non-accident as accident). This is due to the huge difference between the cost of training a driver and the cost of bearing an accident. A small modification that can be applied to have higher sensitivity (i.e., to correctly predict the true positives) is by sacrificing the specificity and overall accuracy. We select the cutoff point that attains high sensitivity level, which corresponds to a false negative rate between 5% and 7% for all the 3 models. . The result of this assumption is shown in Table 12.a to Table 12.c which represents the confusion matrix for all the three individual models. It can be seen from Table 12.a, 12.b and 12.c that sensitivity of random forest, GBM and deep learning are 95.08%, 94.49%, and 93.09% respectively with a corresponding drop in specificity and accuracy. So depending on the desired level of sensitivity the cutoff point could be changed. The management team can determine the level of sensitivity they would like to achieve, and find the best model for prediction; or set the specificity at a pre-defined level considering training capacity, and then identify the best prediction model with highest sensitivity.

Table 12.a: Confusion Matrix for Random Forest with Higher Sensitivity

		Predicted		
		N	Y	
Actual	N	257622	85191	Specificity= 75.15%
	Y	447	8632	Sensitivity= 95.08%
		NPV = 99.83%	PPV = 9.20%	Accuracy= 75.66%

Table 12.b: Confusion Matrix for GBM with Higher Sensitivity

		Predicted		
		N	Y	
Actual	N	261622	81191	Specificity= 76.32%
	Y	500	8579	Sensitivity= 94.49%
		NPV = 99.81%	PPV = 9.56%	Accuracy= 76.79%

Table 12.c: Confusion Matrix for Deep Learning with Higher Sensitivity

		Predicted		
		N	Y	
Actual	N	200214	142599	Specificity= 58.40%
	Y	627	8452	Sensitivity= 93.09%
		NPV = 99.69%	PPV = 5.60%	Accuracy= 59.30%

6. SUMMARY AND DISCUSSION

This study provides a framework for a transportation company to build their own predictive models to save the life and the cost involved, by avoiding an accident. Regression can be helpful to interpret but unfortunately achieving high accuracy is difficult. Similarly, machine learning methods have proved their purpose by producing better accuracy with high specificity and sensitivity. This result also suggests that that instead of completely relying on one model or a specific algorithm, ensemble techniques like voting, weighted average, etc. might produce better results. This study may not be an example just for accident prediction but also applicable for driver turnover, fuel consumption, tractor and trailer maintenance, etc. with their own related data. According to the data from Department of Transportation (USDOT), the National Center for Statistics and Analysis (NCSA), and the National Highway Traffic Safety Administration

(NHTSA) cited by TruckDrivingJobs.com, the average cost of a truck accident with no fatality is $62,000 and the average cost of truck accidents with fatality is $3 million. Some of the major causes for these costly and deadly accidents include the longer stopping distance required by a truck (typically nearly thrice the distance required by other vehicles), requirement of more space to make wide turns, the height and weight of the truck contributing to easy rollover events, blind spots while making a turn, passing and lane changing, etc. Based on the prediction results, it can be argued that training the drivers in the false positive cell of the confusion matrix is an extra cost but those are nothing but investments to avoid unexpected accidents that are beyond predictions. Training the drivers based on prediction would cost only a few thousands of dollars while bearing an accident might cost in millions along with the risk of a life. So the possibility of training the entire driver work force can be questioned which will turn to be a very boring practice and the drivers would not take it seriously. For this study, the predictions are done for every month, each driver flagged by the model are taken very seriously and made sure all the concerns are addressed with rigorous training. Once the driver is trained, she/he is not trained again for a defined period of time (e.g., 5 months) even if the model again flags the same driver. By this way, the effect of training can also be analyzed and the comfort zone of the drivers is also not disturbed, as very frequent and repetitive training can be annoying. Awareness programs, interactive sections, counseling groups, regular feedbacks are some of the steps that can be taken to act towards the prediction aiming at reducing accidents. Following these procedures and emphasizing the importance of drivers, safety can easily become a habit.

7. REFERENCES

Abdelwahab, H., & Abdel-Aty, M. (2001). Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. *Transportation Research Record: Journal of the Transportation Research Board*, (1746), 6-13.

Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis & Prevention*, *34*(6), 729-741.

Allison, P. D. (2012). *Logistic regression using SAS: Theory and application*. SAS Institute.

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural computation*, *9*(7), 1545-1588.

Anderson, J. A., & Rosenfeld, E. Neurocomputing: Foundations of research, 1988. *Cambridge, MA*, 729.

ATA. (2016). *American Trucking Association*. Retrieved 11 20, 2016, from Reports, Trends & Statistics: http://www.trucking.org/News_and_Information_Reports_Industry_Data.aspx

Beshah, T., Ejigu, D., Abraham, A., Snasel, V., & Kromer, P. (2011, December). Pattern recognition and knowledge discovery from road traffic accident data in ethiopia: Implications for improving road safety. In *Information and Communication Technologies (WICT), 2011 World Congress on* (pp. 1241-1246). IEEE.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197-227

Blincoe, L., Seay, A., Zaloshnja, E., Miller, T., Romano, E., Luchter, S., & Spicer, R. (2002). *The economic impact of motor vehicle crashes, 2000* (No. HS-809 446,). Washington, DC, National Highway Traffic Safety Administration.

Blower, D., Green, P. E., & Matteson, A. (2008). Bus operator types and driver factors in fatal bus crashes: results from the buses involved in fatal accidents survey. *University of Michigan Transportation Research Institute.*

Branco, P., Torgo, L., & Ribeiro, R. (2015). A survey of predictive modelling under imbalanced distributions. *arXiv preprint arXiv:1505.01658*.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199-231.

Candel, A., Parmar, V., LeDell, E., & Arora, A. (2015). Deep Learning with H2O

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook* (pp. 853-867). Springer US

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.

Chong, M., Abraham, A., & Paprzycki, M. (2005). Traffic accident analysis using machine learning paradigms. *Informatica, 29*(1).

Click, C., Malohlava, M., Candel, A., Roark, H., & Parmar, V. (2016). Gradient Boosted Models with H2O.

Curry, A. E., Peek-Asa, C., Hamann, C. J., & Mirman, J. H. (2015). Effectiveness of parent-focused interventions to increase teen driver safety: A critical review. *Journal of Adolescent Health*, *57*(1), S6-S14.

Dasarathy, B. V., & Sheela, B. V. (1979). A composite classifier system design: concepts and methodology. *Proceedings of the IEEE, 67*(5), 708-713.

Dimitriadou, E., Weingessel, A., & Hornik, K. (2003). A cluster ensembles framework, Design and application of hybrid intelligent systems.

Federal Motor Carrier Safety Administration Analysis Division. (2016, March). *Large Truck and Bus Crash Facts 2014.* Report No. FMCSA-RRA-16-001. Washington, DC: U.S. Department of Transportation.

Federal Motor Carrier Safety Administration Analysis Division. (2016, March). *Large Truck and Bus Crash Facts 2014.* Report No. FMCSA-RRA-16-001. Washington, DC: U.S. Department of Transportation.

Flom, P. L., & Cassell, D. L. (2007, November). Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. In *NorthEast SAS Users Group Inc 20th Annual Conference: 11-14th November 2007; Baltimore, Maryland*.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics.

Gagliano, S. A., Paterson, A. D., Weale, M. E., & Knight, J. (2015). Assessing models for genetic prediction of complex traits: a comparison of visualization and quantitative methods. *BMC genomics*, *16*(1), 1.

Global Driver Risk Management - Alert Driving. (2016). *Human Error Accounts for 90% of road accidentst - Fleet Alert Magazine - International News - April 2011*. Retrieved 11 20, 2016, from Global Driver Risk Management: http://channel.alertdriving.com/home/fleet-alert-magazine/international/human-error-accounts-90-road-accidents

Gradient boosting. (2016, October 21). In *Wikipedia, The Free Encyclopedia*. Retrieved, October 21, 2016, from https://en.wikipedia.org/w/index.php?title=Gradient_boosting&oldid=745575269

Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, *39*(3), 3659-3667.

Guisan, A., Edwards, T. C., & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological modelling*, *157*(2), 89-100.

Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian Journal of Internal Medicine*, *4*(2), 627–635.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263-1284.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1026-1034).

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological approach*. John Wiley & Sons.

Ho, T. K. (1995, August). Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on* (Vol. 1, pp. 278-282). IEEE.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, *20*(8), 832-844.

Johansson, U. (2007). *Obtaining accurate and comprehensible data mining models: An evolutionary approach*. Linköping University, Department of Computer and Information Science.

Johnson, C., & Johnson, R. C. (1988). Cognizers: Neural networks and machines that think.

Jovanis, P. P., & Chang, H. L. (1986). Modeling the relationship of accidents to miles traveled. *Transportation Research Record*, *1068*, 42-51.

Krishna, T. G. (2012). Abstract methods used in data mining. Reg.No. PU14PHD0462. *International society of thesis publication - A Scoiety of Reseach Publication*.

Krishnaveni, S., & Hemalatha, M. (2011). A perspective analysis of traffic accident using data mining techniques. *International Journal of Computer Applications*, *23*(7), 40-48.

Le, Q. V. (2015). A Tutorial on Deep Learning Part 1: Nonlinear Classifiers and The Backpropagation Algorithm.

Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: A Package for Binary Imbalanced Learning. *A peer-reviewed, open-access publication of the R Foundation for Statistical Computing*, 79.

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013, June). Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML* (Vol. 30, No. 1)

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (Vol. 37). CRC press.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (Vol. 37). CRC press.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133.

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, *28*(1), 92-122.

Miaou, S. P., & Lord, D. (2003). Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. *Transportation Research Record: Journal of the Transportation Research Board*, (1840), 31-40.

Minsky, M., & Papert, S. (1969). Perceptrons.

Moghaddam, F. R., Afandizadeh, S., & Ziyadi, M. (2011). Prediction of accident severity using artificial neural networks. *International Journal of Civil Engineering*, *9*(1), 41.

Murray, D., Lantz, B., & Keppler, S. (2006, March). Predicting truck crash involvement: Developing a commercial driver behavior model and requisite enforcement countermeasures. In *Transportation Research Board 85th Annual Meeting* (No. 06-2850).

Mussone, L., Ferrari, A., & Oneta, M. (1999). An analysis of urban collisions using an artificial intelligence model. *Accident Analysis & Prevention*, *31*(6), 705-718.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (pp. 807-814).

National Center for Statistics and Analysis. (2016, May). Large trucks: 2014 data. (Traffic Safety Facts. Report No. DOT HS 812 279). Washington, DC: National Highway Traffic Safety Administration.

Nelder, J. A., & Baker, R. J. (1972). Generalized linear models. *Encyclopedia of statistical sciences*.

Nykodym, T., Kraljevic, T., Hussami, N., Rao, A., & Wang, A. (2016). Generalized Linear Modeling with H2O.

Piccinini, G. (2004). The First computational theory of mind and brain: a close look at mcculloch and pitts's "logical calculus of ideas immanent in nervous activity". *Synthese*, *141*(2), 175-215.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, *6*(3), 21-45.

Polikar. R. (2009). Ensemble learning. Scholarpedia, 4(1):2776.

Random forest. (2016, November 2). In *Wikipedia, The Free Encyclopedia*. Retrieved, November 2, 2016, from https://en.wikipedia.org/w/index.php?title=Random_forest&oldid=747497674

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1-2), 1-39.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

SAS Institute Inc. (2016). Predictive Analytics. Retrieved from http://www.sas.com/en_sg/insights/analytics/predictive-analytics.html

Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, *5*(2), 197-227.

Scikit Learn. (2016, 11 9). Ensemble Methods. Retrieved from Scikit Learn: http://scikit-learn.org/stable/modules/ensemble.html

Shahzad, R. K., & Lavesson, N. (2013). Comparative analysis of voting schemes for ensemble-based malware detection. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, *4*(1), 98-117.

Shankar, V., Milton, J., & Mannering, F. (1997). Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. *Accident Analysis & Prevention*, *29*(6), 829-837.

Short, J. (2014). *Analysis of Truck Driver Age Demographics Across Two Decades – 2014* . Atlanta, GA: American Transportation Research Institute

Sohn, S. Y., & Lee, S. H. (2003). Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. *Safety Science*, *41*(1), 1-14.

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

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

Truck DrivingJobs.com. (2016). *Truck Driving Accidents – Causes, Fatalities, Statistics and Costs*. Retrieved from https://www.truckdrivingjobs.com/faq/truck-driving-accidents.html

U.S. Department of Transportation (USDOT), Bureau of Transportation Statistics (BTS). (2016). *Growth in the nations's freight shipments - Highlights*. Retrieved 11 20, 2016, from http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/freight_shipments_in_america/html/entire.html

Vihinen, M. (2012). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC genomics*, *13*(4), 1.

Werbos, P. (1974). Beyond regression: New tools for prediction and analysis in the behavioral sciences.

Williams, R. (2015, February 20). Interaction effects and group comparisons.

Xie, Y., Lord, D., & Zhang, Y. (2007). Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis. *Accident Analysis & Prevention*, *39*(5), 922-933.

Yang, W. T., Chen, H. C., & Brown, D. B. (1999, November). Detecting Safer Driving Patterns by A Neural Network Approach. In *ANNIE'99 for the Proceedings of Smart Engineering System Design Neural Network, Evolutionary Programming, Complex Systems and Data Mining* (Vol. 9, pp. 839-844).

Zhang, C. X., & Zhang, J. S. (2008). A local boosting algorithm for solving classification problems. *Computational Statistics & Data Analysis*, *52*(4), 1928-1941.

Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, *58*, 308-324.

Zhang, Y., Burer, S., & Street, W. N. (2006). Ensemble pruning via semi-definite programming. *Journal of Machine Learning Research*, *7*(Jul), 1315-1338.

Zhou, Y., & Hooker, G. (2016). Interpreting Models via Single Tree Approximation. *arXiv preprint arXiv:1610.09036*