COLLEGE PHYSICAL SCIENCES STATISTICS

STATS 101C FINAL REPORT - FALL 2022

Prediction of the U.S. Car Accidents' Severity

Gary Ramos
Marlyn Tanuwandi
Thi Nguyen
Juhyeon Seo

December 7, 2022

Abstract

The critical purpose of this project is to implement statistical learning methods to the given dataset thereby training a model to predict the severity of car accidents in the United States. This paper mainly focuses on the data analysis, modeling process we went through and our interpretation of the results. Our final model is a random forests model which predicted the severity with an accuracy of 94.59%.

Keywords

Car accidents; Random forest; Severity, Machine learning

1 Introduction

Traffic accidents result in associated fatalities and economic losses every year. According to the National Safety Council, car accidents cost the economy approximately $473.2 billion, and 4.8 million people were seriously injured in automobile collisions in 2020. Therefore, preventing potential unsafe road conditions will not only improve the safety of people and vehicles in our transportation system but also reduce the economic cost of vehicle accidents. In order to effectively deal with this, determining whether a certain condition is likely to lead to a severe accident is very important. Thus, predicting the severity of accidents is critical to allocating resources.

For this Kaggle project, we are given countrywide traffic accident data collected from February 2016 to December 2021. The training dataset contains 35000 observations, and the testing dataset has 15000 observations, where each observation contains 11 numerical variables and 32 categorical variables. Independent variables provided detailed information about the condition of traffic accidents such as the time of the accident, weather conditions, and geological information. Our role in the project is to build a binary classification model to predict the target variable "Severity" which is a categorical variable with two categories "Severe" and "Mild".

2 Data Analysis

2.1 Exploratory Data Analysis

Before performing any changes to the dataset, we performed some exploratory data analysis on variables first to understand the data in depth. By visualizing data, we were able to explore some important relationships between variables and response. Figure 1 is a good example in which looking at this map we should note that serious auto accidents are more common on the state's east coast. For the following steps , we will deeply focus on our strategy of picking potential predictors.

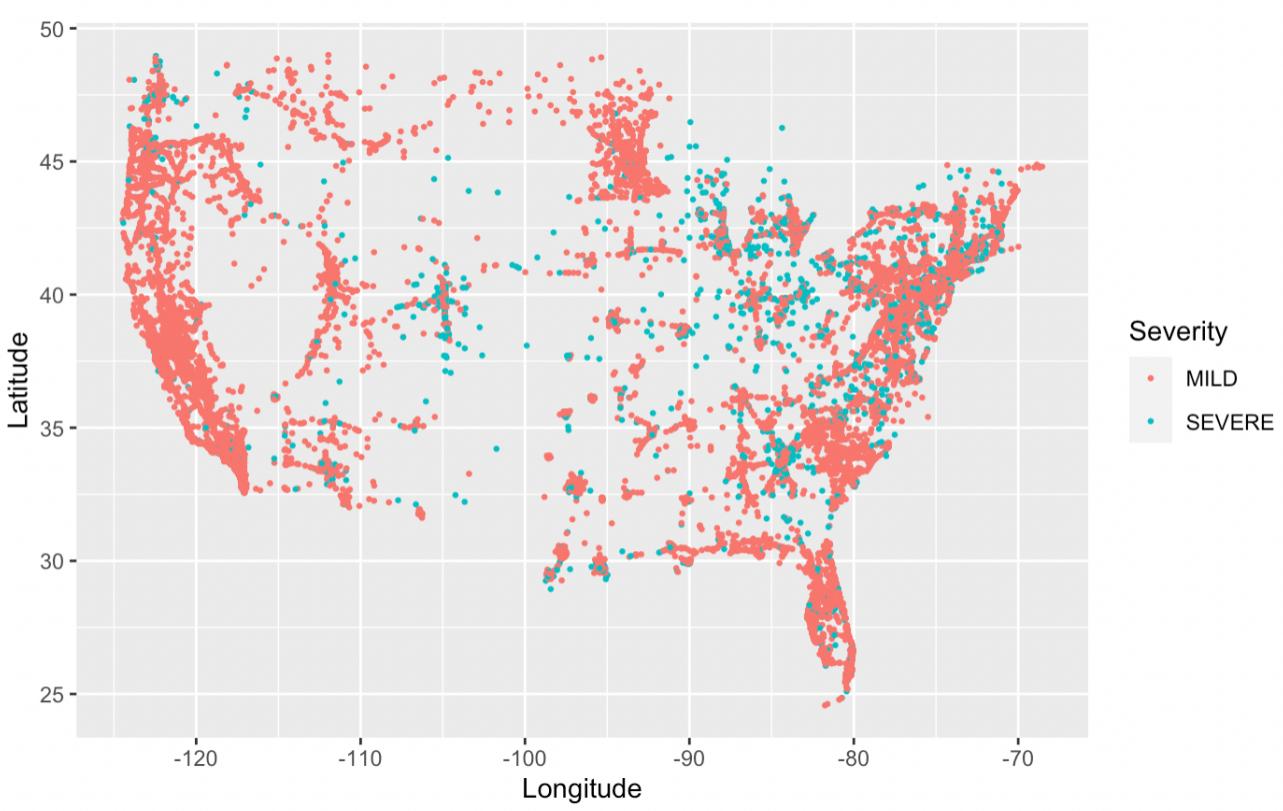

Figure 1: Geographical dispersion of accidents by severity

2.1.1 Distribution of the target variable

First, we started discovering how the response is distributed in the training dataset. We found that approximately 90% of accidents in the training dataset were mild accidents, which means simply predicting "mild" will have only a 10% misclassification error.This illustrates that a meaningful classification model should have higher accuracy than 90%.

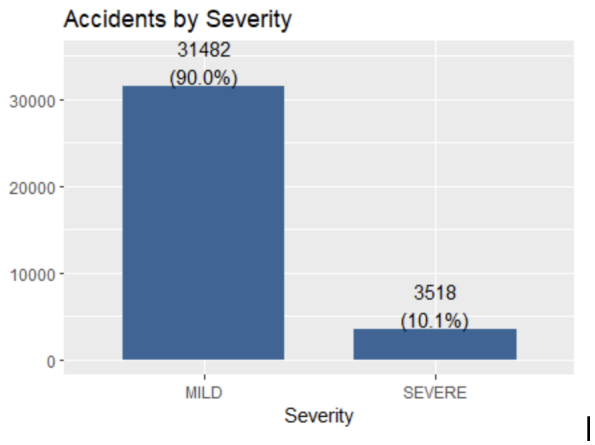

Figure 2: Numbers of accidents by severity

2.1.2 Missing values

Even though the data was relatively organized when we received it, we still need to take a few steps to prepare before making the actual model. As we can observe from the percentage of missing values plot below, some predictors such as wind_chill.F., Wind_Speed.mph, and Humidity contain large numbers of missing values. In general, we found that there are 13211 missing values in the training dataset and 5842 in the testing dataset. Since missing values can negatively affect the accuracy of the classification model, dealing with missing values was one crucial part in this project. We will give more detail on how we handled this problem in the 2.3 session.

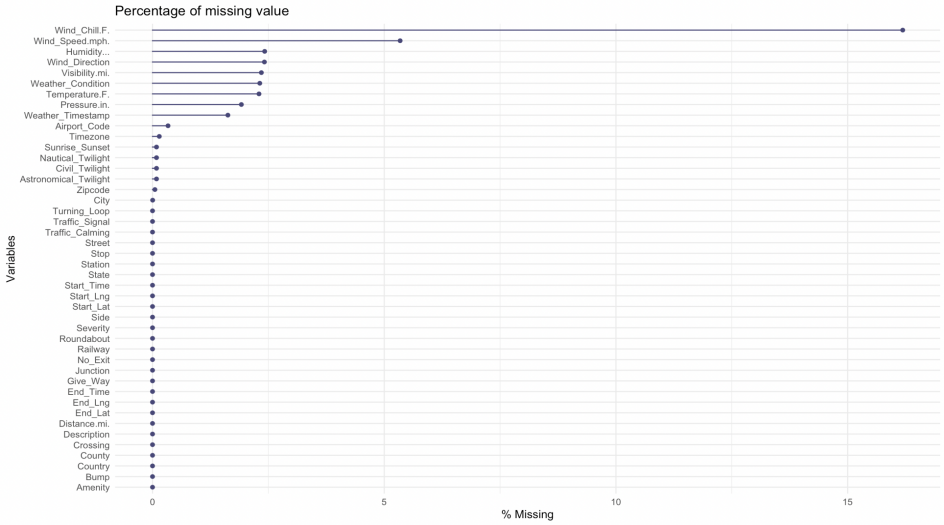

Figure 3: Percentage of missing values of training dataset

2.1.3 Numerical variables

To find variables that help separate severe accidents from mild accidents, we created density plots. Namely, if a predictor's probability density plots have a considerable overlap for both mild and severe, the predictor may not play a vital role in classification. Otherwise, if the plot shows a clear distinction from the other, then the predictor has the potential to be used as a predictor in models.

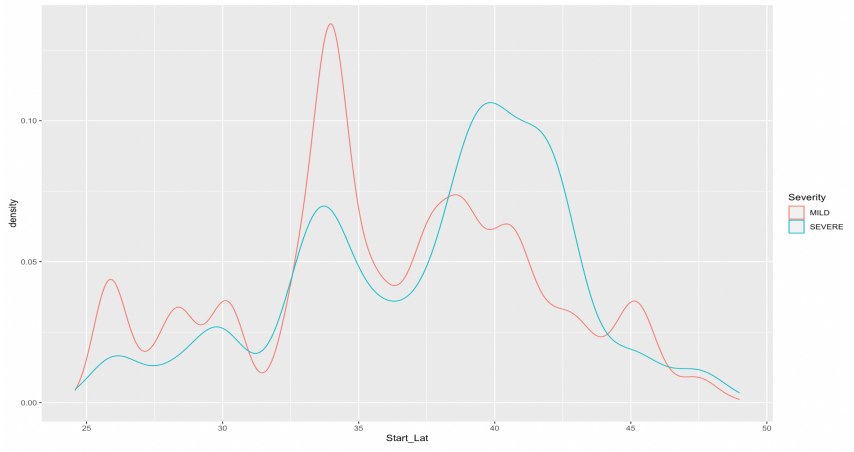

Figure 4: Density plot of Stat_lat

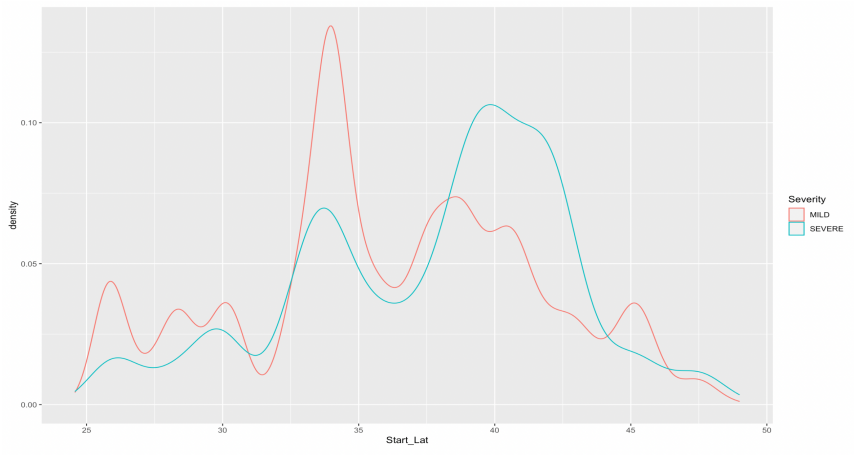

Figure 5: Density plot of wind_speed.mph.

To illustrate, the Density plot of "Start_Lat" is shown as an example of a predictor which successfully separates the target variable, whereas "Wind_Speed.mph." is an example of a predictor that failed to separate the data.

2.1.4 Categorical variables

Next, we created stacked bar charts for categorical variables to see the composition and comparison of each variable. If there is a significant difference in the proportion between each category of the variable, it is considered to be a useful predictor. On the contrary, it is not a good predictor if the category consists of a single value or if all categories in a predictor represent the same ratio.

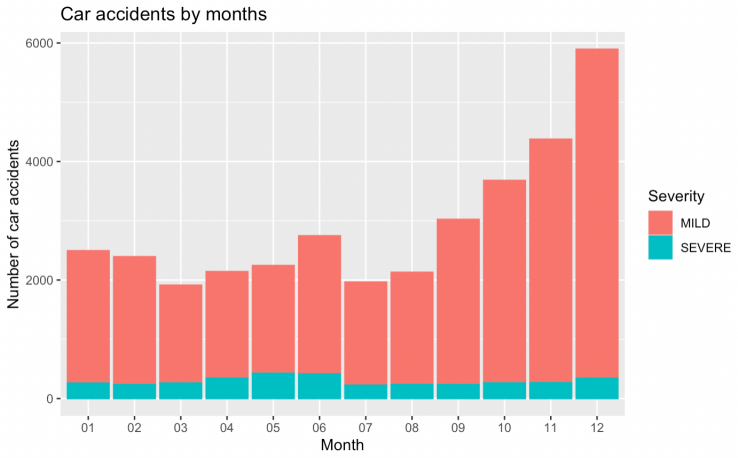

Figure 6: Numbers of car accidents vs month

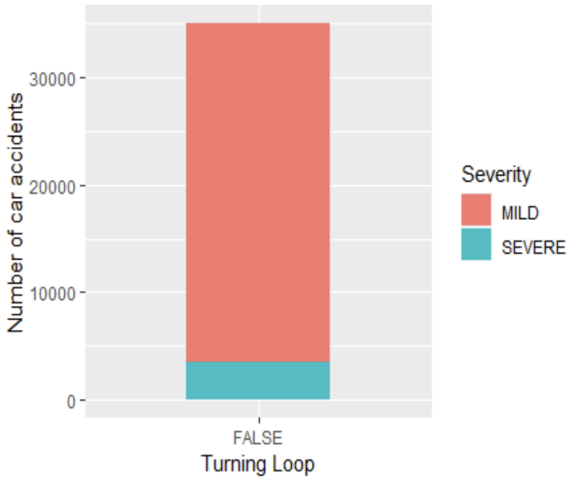

Figure 7: Numbers of car accidents vs Turning Loop

Figure 6 shows an example of a potentially helpful predictor. It represents the number of car accidents by month and the proportions vary between groups. By looking at the number of car accidents versus the monthly graph, we may notice that car accidents tend to occur more frequently at the end of the year. Intuitively, this means that the holiday season has a significant association with high rates of vehicle accidents. It is worth it to make a further investigation for this predictor. In contrast, the stacked bar chart of figure 7 can be an example of a less potentially helpful variable as it has only one value.

2.2 Feature engineering

We introduced some new variables such as "Population", "Fatal_total", "Duration", "is.closed", "is.road", and etc. Our thought process was that those variables would have significant associations with the target response so that they can help to classify the levels of severity.

2.2.1 Merge data using external sources

We wanted to include variables in our analysis that could capture some of the variability in severity levels due to factors associated with state population. We believe that the larger the population is, the more car accidents are likely to occur. Therefore, we decided to extract the population of each state from the Census Bureau website. We also included the factors that could be relevant to traffic accidents, such as the number of fatal crashes in the year where the accident occurred. After collecting the news data, we merge all the data with the left join function.

2.2.2 Creating features from existing variables

There were some predictors that were stored in terms of character data, including Start_Time and End_Time. We recognized that these variables would be very useful once we built the model, but we needed to convert them into the proper data type first. We began by splitting information from Start_Time and End_Timeare into five new variables such as new_date, month, year, new_time_start, and new_time_end. In fact, the longer the time of the car accident, the higher likelihood of the severity. Thus, we generated a new predictor named "duration" by subtracting the start time from the end time to get the duration of accidents.

2.2.3 Text Mining Voyant Tools

Know that the description contains a lot of important information about a car accident. However, this variable is given as a disorganized data type. Instead of using the original description as a predictor, we extracted information from the unstructured texts in the description and created new variables based on that description. We used the Voyant Engine which provides the frequency of a word in the data to find commonly used terms. It is interesting to realize that "closed," "road," "accident," and "blocked" are the most common words that appear in descriptions of severe auto accidents, as shown in Figure 8.

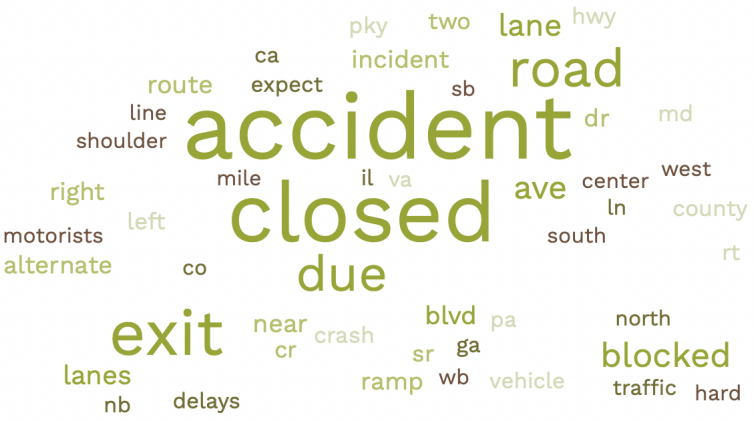

Figure 8: The common word used in Description

We also created the bar charts to investigate whether the description containing the common words can help predict the severity of an accident more accurately. As expected, the plots do prove that detecting the common words can be remarkably assisted in increasing the accuracy of our model . Therefore, we added five new predictors to our dataset, including "is.closed", "is.incident", "is.traffic", "is.blocked", and "is.road".

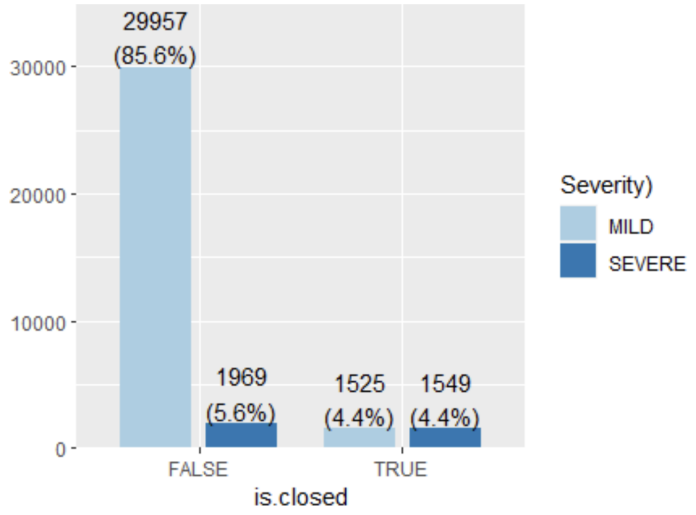

Figure 9: Severity based on is.closed

2.3 Data cleansing

Before handling the missing data, we looked at the data once again in order to carefully decide how to deal with missing values in this dataset. We generate another graph for testing data where this graph shows similar patterns of missing values to the one that we made for the training data.

2.3.1 Imputing missing variables using mice

The simplest way to deal with NA values in the data set is just dropping the rows with missing values. However, due to a large amount of missing data, removing the rows with missing values resulted in a significant loss of information. Therefore, we replaced missing values with imputed datasets by using the mice package.

2.3.2 Replacing missing values with median

Even after replacing missing values with imputed data, some values remained missing. Since the distribution of this dataset is heavily skewed, we decided to replace the remaining missing values with the medians rather than the mean. Technically, we calculated the median of the entire feature column and filled them all with median values. "Population_County", "Drive_County", "Humidity...", and "Pressure.in." are the columns we assigned the median to NAs.

2.4 Methodology

2.4.1 Random Forests

Random forests is a classification algorithm that combines the output of many decision trees to reach a single result. It is used in many fields of application because of its excellent predictive performance. It reduces the risk of overfitting as a robust number of decision trees in a random forest lowers overall variance and prediction error.
One of the most important advantages of random forests is that it does not require feature selection as it automatically calculates variable importance and provides its own interpretation. Because of this, we did not use additional methods to select less predictive variables even though the data had a large number of variables. Instead, we removed predictors gradually based on their effect on the accuracy of the model.
With random forests, we achieved the highest accuracy with 18 predictors(Start_Lat, Start_Lng, Distance.mi., State, Temperature.F., Humidity..., Pressure.in., Population_County, new_start, new_end, duration, month, year, is.closed, is.incident, is.traffic, is.road, Fatal_total). In the important variable plot below, we illustrate the 18 important features we previously selected through correlations and sort them by their Gini Importance from the Random Forest Models. Gini importance, also known as mean decrease impurity, measures how each feature decreases the impurity of the resulting split. The feature that decreases the impurity of the split the most is selected for an internal node, and the average impurity is calculated for each feature. Gini Importance provides us insight into what is happening in a Random Forest model and allows us to determine which features contributed the most to the creation of the decision trees.

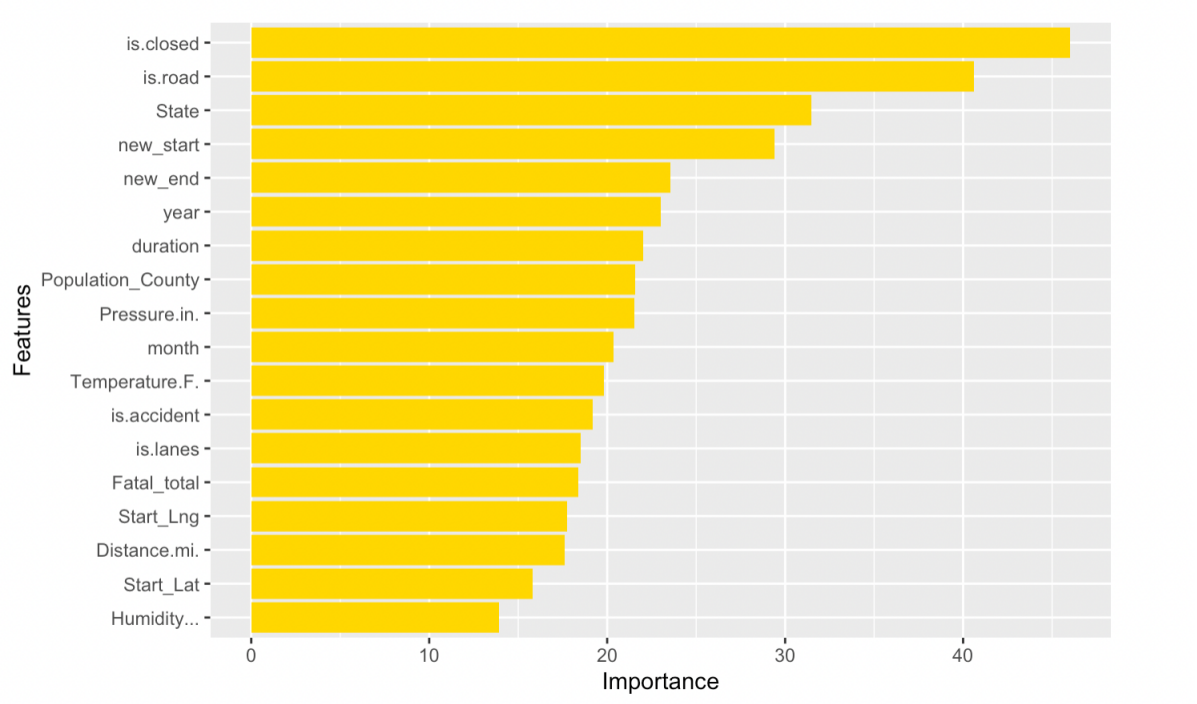

Figure 10: Mean Decrease Gini

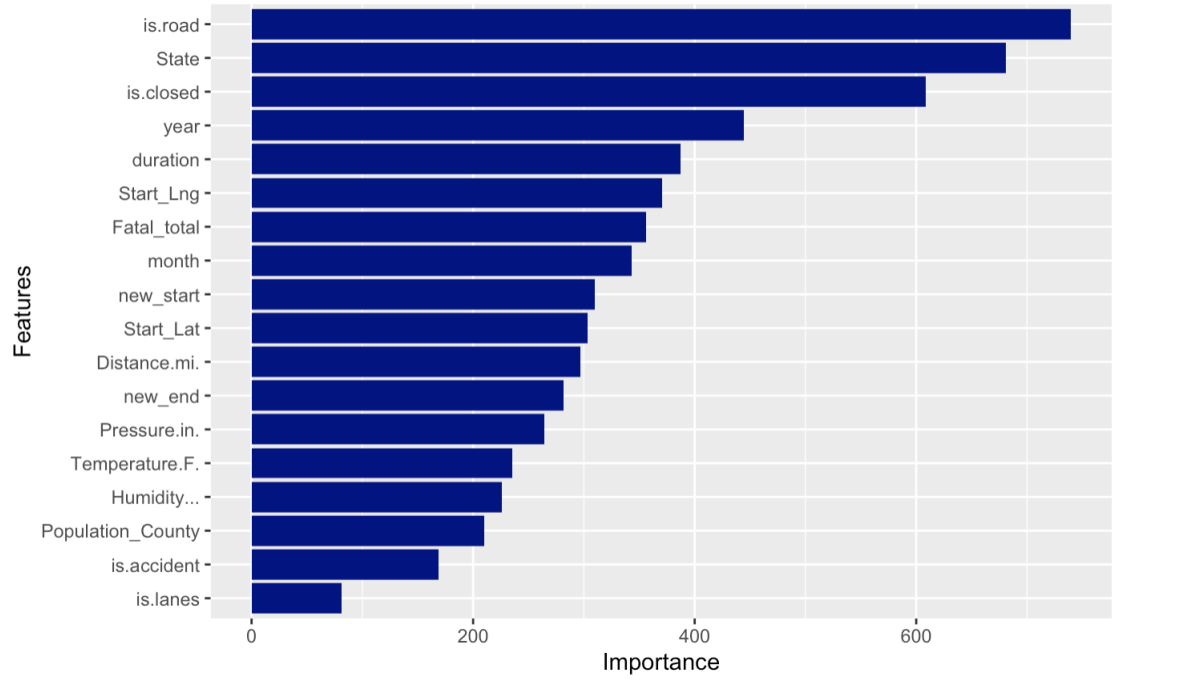

Figure 11: Mean Decrease Accuracy

It can be seen above that the top 5 most important features for classifying current level of car accidents are is.closed, is.road, States, new_start, and new_end. Note that the ordering of the 18 important features using Random Forest's Gini Importance differs from the ordering using correlation with the response variable.
We also built different random forest models using different numbers of important predictors, truing multiple mtry, and compared their error rates. We found that for our dataset, reducing the number of predictors actively led to an increase in the error rate , as shown in Figure 12.

2.4.2 Generalized Logistic regression

In addition to the random forests model, we applied the generalized linear model (GLM) to the same data. Logistic regression is a regression analysis widely used in statistics for predicting binary target variables as it predicts the probability that an observation belongs to one of two categories. The testing error rate of the GLM model was 0.05834 on Kaggle.

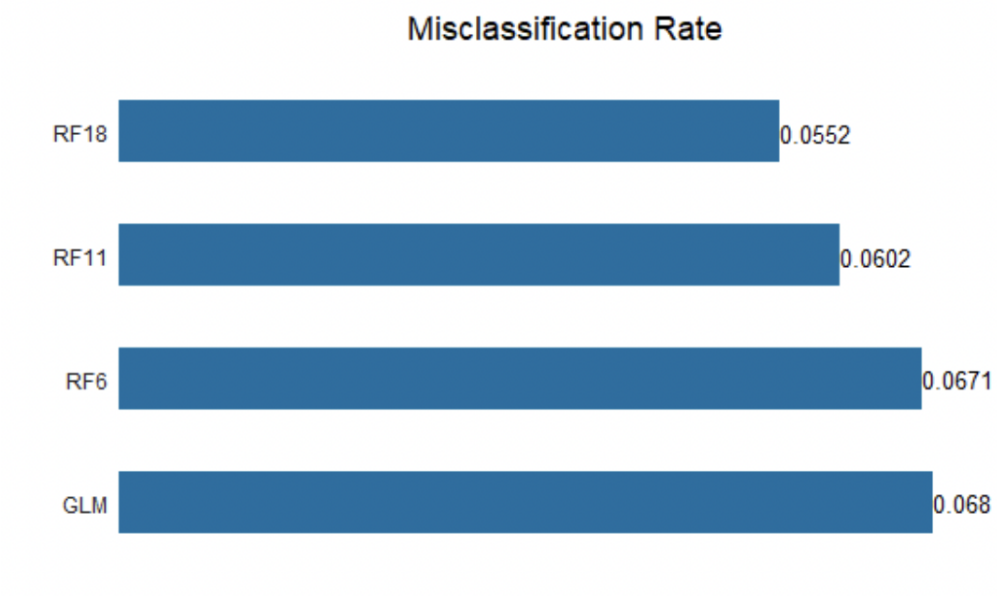

Figure 12: Error Rates of different models

As a result, we determined that our best model considerably includes 18 important predictors with 500 trees and 4 mtry since it yielded the smallest error rate.

3 Discussion and Limitation

The performance of our model was satisfactory for us. We won second place in the competition and the final score was 0.9459. However, one main limitation of random forests made us reluctant to use it as a final model at first despite its high accuracy. Even though it provides an estimate of important variables, it does not tell us its internal workings and thus we cannot see how it arrived at a certain decision. We decided to use random forests because interpretability was not the main concern in this project but it was easy to notice that this lack of interpretability limits its use in some fields such as health and economics. Another drawback of the random forests classifier is that it requires more trees to improve the accuracy. The more trees it has, the more

expensive it is to construct random forests and the more time it takes to process. In other words, random forests can become time-consuming and computationally extensive.

4 Conclusion and Recommendation

From this project, we learned that data cleansing and feature engineering for predictive modeling is as important as understanding the characteristics of models and choosing the right method. Based on the variable importance plot, new features created by extracting utilizable parts of the existing features were very meaningful to the model. Even though random forests is a powerful algorithm, it does not outperform other models with better datasets when applied without data handling.

If we had more time, we would have spent time studying the detailed relationship between key factors and accident severity. As mentioned earlier, random forest does not provide information about the complex interactions inside it so we could not gain a full understanding of the decision process. Further research would have allowed us to find the connection between them and improve the model.

Lastly, although we have a decent accuracy rate, we have to admit the fact that our model is not really efficient enough given the disadvantage of the random forest method. Technically, random forests favor categorical variables with a high number of levels. This proves that the predictor State, consisting of 51 levels, is one of the most important predictors for our model, demonstrating that our model is biased when choosing the important features. For further work, we will need to reduce the level of the State by using its original information to create a new variable related to it in order to solve the issue.

5 Acknowledgement

Thanks to this project, we learned a lot of things like how to deal with missing values, how to improve the classification model by applying multiple methods, etc. This project also gives us the opportunity to apply what we have learned in the lecture to solve real-world problems, thereby consolidating our knowledge in a deep and thorough way. Most importantly, we would like to especially thank Professor Akram Almohalwas for his help and support during the course of this thesis.

References

[1] IBM Cloud Education. (n.d.). *What is Random Forest?* IBM. (2020, December 7). Retrieved December 7, 2022, from https://www.ibm.com/cloud/learn/random-forest

[2] Aria, M., Cuccurullo, C., Gnasso, A. (2021).*A comparison among interpretative proposals for random forests*, Machine Learning with Applications, 6, 100094. https://doi.org/10.1016/j.mlwa.2021.100094

[3] Almohalwas, Akram. "Predicting Car Accidents' Severity." https://www.kaggle.com/competitions/predicting-car-accidents-severiy

[4] Insurance Institute for Highway Safety (IIHS) – https://www.iihs.org/about-us