STATS 101C FINAL PROJECT

Predicting Car Accident Severity in the U.S.

Lecture 1
Ruyi Lu (405592759)
Luoyao He (205709190)
Matthew Guo (805571825)
Ethan Vuong (905786142)
University of California, Los Angeles

TABLE OF CONTENTS

ABSTRACT……………………………………………………………………………………….. **3**

INTRODUCTION………………………………………………………………………………… **3**

DATA ANALYSIS ……………………………………………………………………………… **4**

MODELING………………………………………………………………………………………… **10**

MODEL ANALYSIS ……………………………………………………………………………... **12**

CONCLUSION………………………………………………………………………………………**13**

REFERENCES………………………………………………………………………………………**14**

ABSTRACT

By embarking on this Kaggle Project, our objective is to use the given data for U.S. car accidents to predict the severity of car accidents with the highest accuracy that we can manage. This report will outline our exact procedure, from data cleaning, feature selection, model building and analysis, as well as a detailed critique of our final model and conclusion.

Our final model utilizes the random forest classification algorithm and xGboost, with our chosen variables of description_severity, End_Lng, End_Lat, weather_timestamp_yr, state, distance.mi, time_length, timezone, pressure.in, wind.chill.F, weather_timestamp_mo, pop_density, weather_timestamp_hr, Humidity, Season, Nautical_Twilight, Wind_Speed.mph, weather_condition and full_coverage. Our public kaggle score is 0.94355, which places us in rank 10 in the public score rankings and rank 7 in the final score rankings.

INTRODUCTION

In 2021, the United States saw 42,915 deaths stemming from vehicular accidents, which has been the highest number of car crash fatalities since 2005. According to the Association for Safe International Road Travel (ASIRT), vehicular fatalities ranks as the number one cause of death for individuals aged 5 to 29. While most motor car accidents only yield mild injuries, if any, accidents at high speed occuring in dangerous areas can cause devastating consequences. Many campaigns have been launched focusing on alleviating the severity of car crashes, such as a push for sturdier cars, better-lit environments for driving and stricter seatbelt laws. Despite this, deaths by vehicular accident per capita in the United States seem to creep up every year. Thus, finding an efficient and accurate model for predicting car crash severity is an urgent issue which needs to be addressed quickly.

The provided dataset on U.S. car accidents covers 49 states from February 2016 to 2021. Accidents are classified as either ‘Mild’ or ‘Severe’, and we are given 45 predictors, varying from numerical, categorical and logical. These predictors offer us further insight into the conditions of the crash, such as the time when the crash happened, the weather, and the presence of a junction at the time.

DATA ANALYSIS

To further understand the hidden pattern of our data, we decided to perform EDA and other exploratory techniques on our original dataset to ensure the accuracy of our classification models.

(1). Replacement of NA values in the dataset

By taking a closer look at the out dataset, we found that there are many missing values in our dataset (13211 entries for the training data and 5842 entries for test data respectively). Due to this relatively large number of missing values, we decided to replace them with certain values instead of simply removing them. We also noticed that there are missing values of variables of different types, so we decided to handle NA values differently corresponding to the data types. For the numerical predictors, we used the mice() function with method "pmm" to handle the missing values; for the logical predictors, we used the mice() function with method "logreg" to handle the missing values; and for the categorical predictors, we replaced all NA values with the most frequent values in each column. Additionally, we noticed that two predictors, Country and Turning_Loop, had only one unique value for all entries, which meant it would not contribute to the prediction of accident severity. Therefore, we removed them from our dataset. Combining these data subsets by indices, we generated a new dataset without missing values.

(2). Creation of new predictors and transformation of current ones

Since our primary goal is to predict the severity of car accidents, and some of the predictors were not intuitively related to accident severity, we decided to transform some of the current predictors or create new ones out of the original dataset. And with the transformation of the original dataset, we were able to remove some collinearities among the predictors.

a. *Predictor transformed*: weather timestamp divided into new columns predictors
   We separated the weather timestamp predictor in new new predictors containing the exact year, month, day, hour, and minute that the accident happened

b. *Predictor transformed*: changed weather conditions to fewer categories.
   Since the original weather condition predictors contain too many levels than what could be imputed to the models, we decided to categorize the weather conditions into fewer groups.

c. *New predictor*: time duration of the accident. As introducing both the start time and end time into the model would produce problems with collinearity, we calculated the delta value of the time duration by subtracting the two values.

d. *New predictor*: season. We based our season predictor based on the meteorological seasons (i.e. March - May as Spring, June - August as Summer, September - November as Fall, and December - February as Winter). This way, we can reduce the levels to months while still maintaining the season in which the accident took place.

e. *New predictor*: description of severe accidents.
   The description of the accident contained in each data log gives valuable insight into the severity of the crash. We detected keywords in the description predictor, converted the descriptions with keywords to TRUE, and descriptions without keywords to FALSE.
   The keywords we chose are the following:
   i. closed due to accident
   ii. closed between
   iii. closed at
   iv. closed from
   v. Two lanes blocked
   vi. Secondary accident

f. *New predictor*: population density based on zip code.
   We used external population density data based on zip code, and used the left_join method to create a new predictor with population density.

g. *New predictor*: if the accident happened during night time.
   We have four predictors describing day and night time, so we define the accident to happen during day time if more than two of those predictors are DAY.

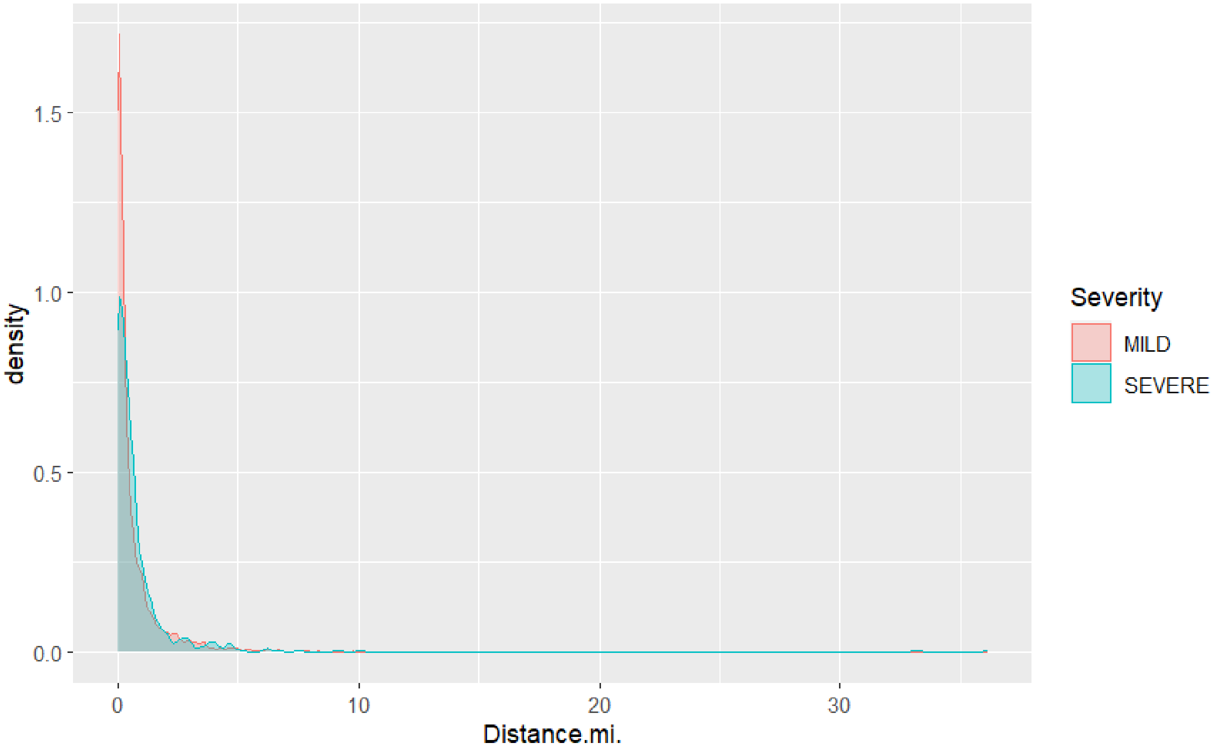
h. *New predictor*: change in displacement based on latitude and longitude
   We subtracted ending latitude and longitude from the starting one to find out the displacement, or the change in latitude and longitude of the car during the accident

i. *New predictor*: if the accident happened during rush hour. As rush hour traffic intuitively instigates more aggravated drivers, we added this new predictor to observe its effects on vehicular accident severity.

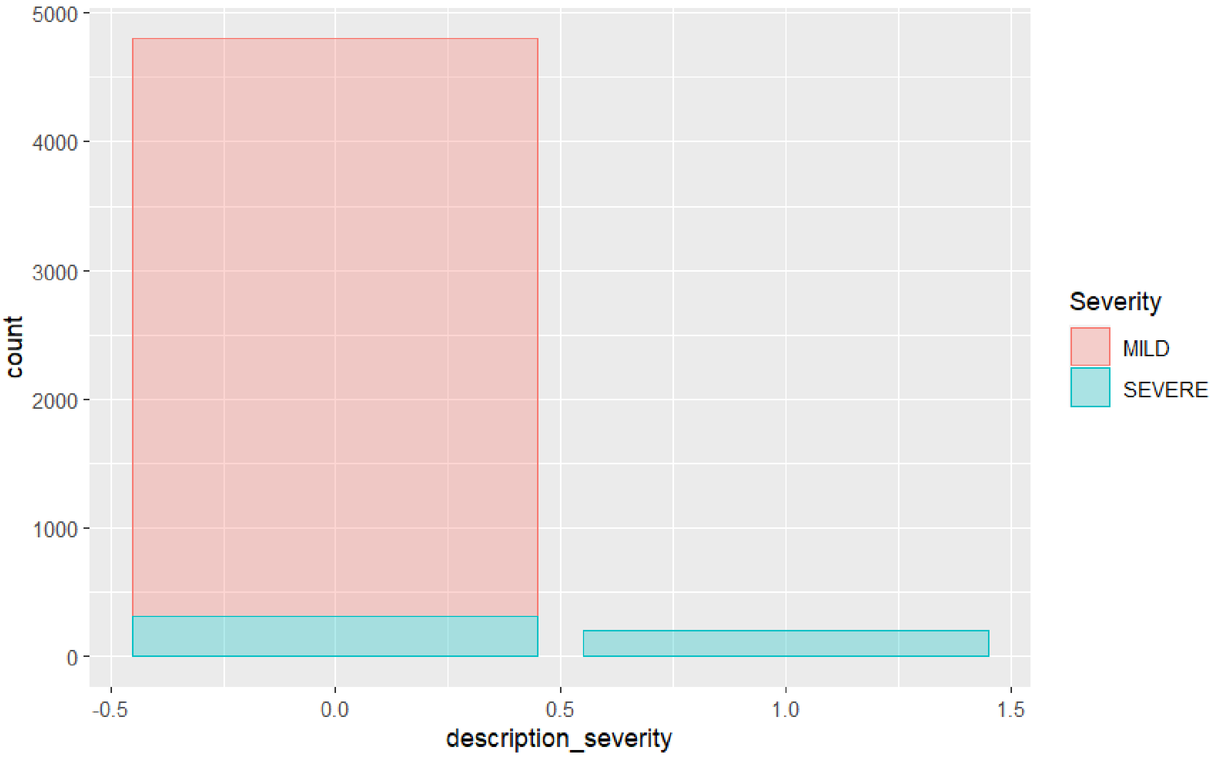
j. *New predicto*r: full car insurance coverage. Insurance fraud is common in the United States, and many individuals abuse the system to gain monetary benefits from crashing their cars.

(3). Predictor visualization: a closer look at selected predictor

To better understand the importance of certain predictors, we created density plots for several selected predictors. Here is an example of density plot using "Distance.mi."; and from the density plot created with part of the training data, we can see how different levels of accident severity is associated with different distance values: essentially, the peak values for "MILD" and "SEVERE" seems to be apart of each other, which indicates that "Distance.mi." might be a good predictor for distinguishing different severity levels.



Density Plot of Distance.mi. by Levels of Severity

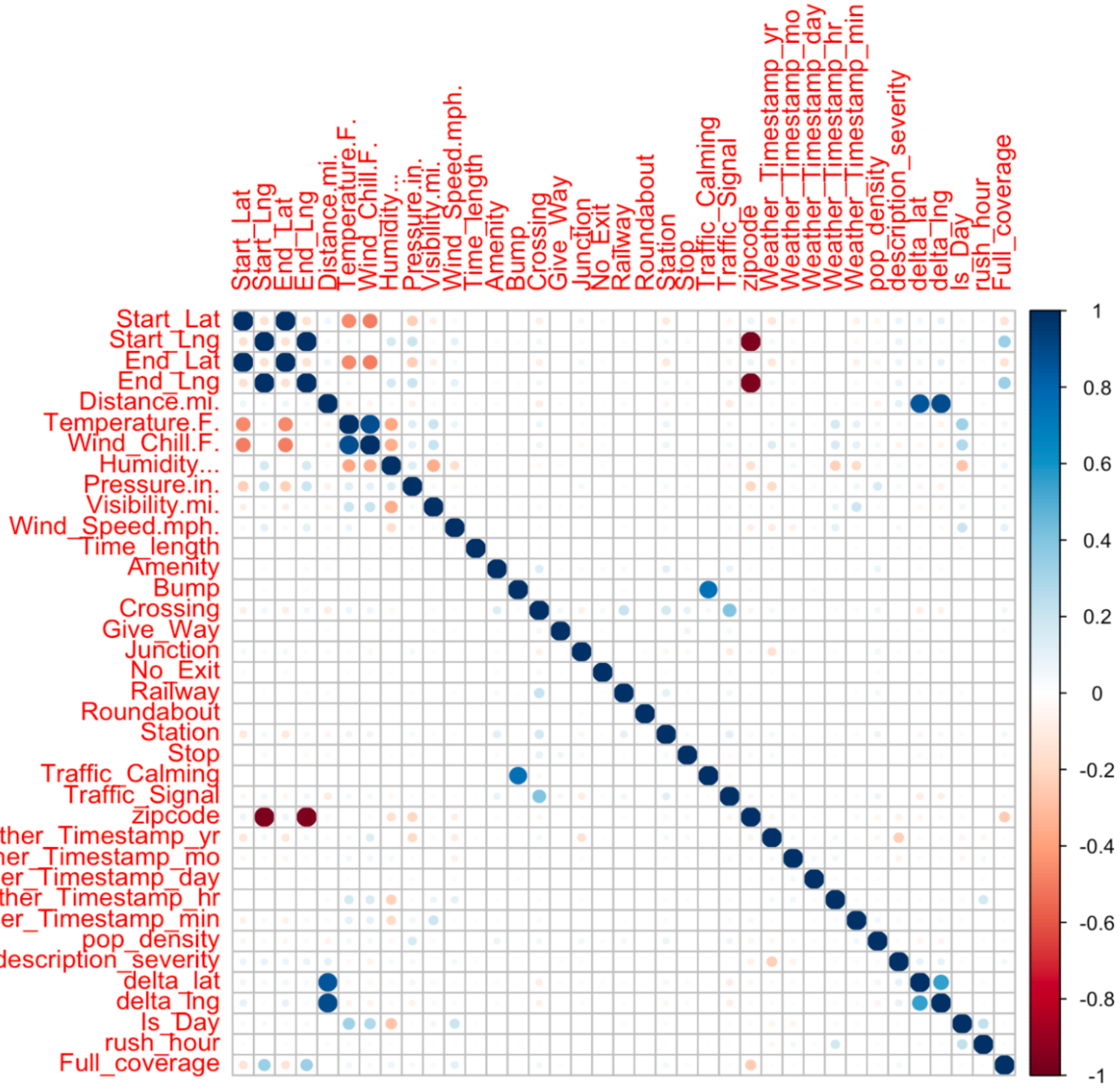


Bar Plot of descriotion_severity by Levels of Severity

Here is another example of a bar plot using “descriotion_severity”. From the bar plot above, which is also created with part of the training data, it is obvious that none of the MILD accident cases have had a description containing keywords that indicates the accident is severe, creating a clear distinction between MILD and SEVERE cases based on “description_severity”. Therefore, we might need to include this predictor when we construct our final model.

(4). Correlation plot for numerical variables, investigating the collinearity issues among variables

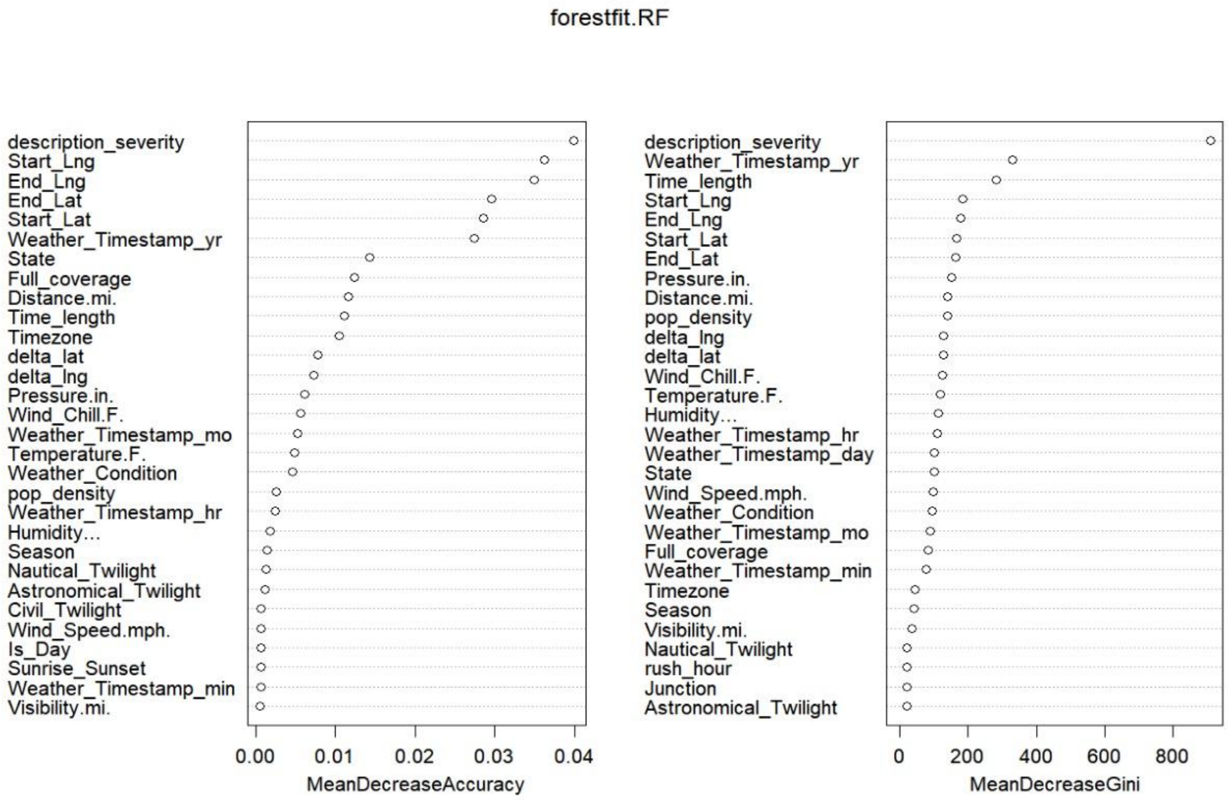
Based on this Correlation Plot, we easily see that Start_Lat, Start_Lng, End_Lat and End_Lng display heavy collinearity, which is intuitively logical as the end position is always based on the start position of the vehicle, with an added displacement. Thus, even if both the Start and End points are shown to be significant in the importance plot, we only need to select either the Start points or the end points.



Correlation Plot for Numerical Predictors

(5). Feature selection for model setup

We generated an importance plot with random forest to look for the most important predictors in our dataset.



Based on the importance plot above, we chose the following predictors that performed well on both plots:

- description_severity
- End_Lng
- End_Lat
- Weather_Timestamp_yr
- State
- Distance.mi.
- Time_length
- Timezone
- Pressure.in.
- Wind_Chill.F.
- Weather_Timestamp_mo
- pop_density
- Weather_Timestamp_hr
- Humidity...
- Season
- Nautical_Twilight
- Wind_Speed.mph.
- Weather_Condition
- Full_coverage

MODELING

After exploring viable models, we applied our processed training data onto 5 models: XGBoost, Random Forest, kNN (k-Nearest Neighbors algorithm), Logistic Regression, and LDA/QDA (Linear and Quadratic Discriminant Analysis). To compare the performance of different models, we divided our original training data into two subsets, one with 70% of the data representing our new training data and the other with 30% of the data representing the test set.

Candidate Model 1: XGBoost

XGBoost is a supervised machine learning technique that stands for 'Extreme Gradient Boosting', which is applied to decision trees and predicts the response variable based on estimates of weaker models. Upon running the XGBoost model, we arrive at an error rate of 0.056.

Candidate Model 2: Random Forest

Random Forest is a supervised machine learning algorithm commonly used as a classification method. By implementing multi-layered decision trees, Random Forest classifies data points using significant and relevant predictors. After running a Random Forest model with the optimized tree size onto our data, we arrived at an error rate of 0.0557.

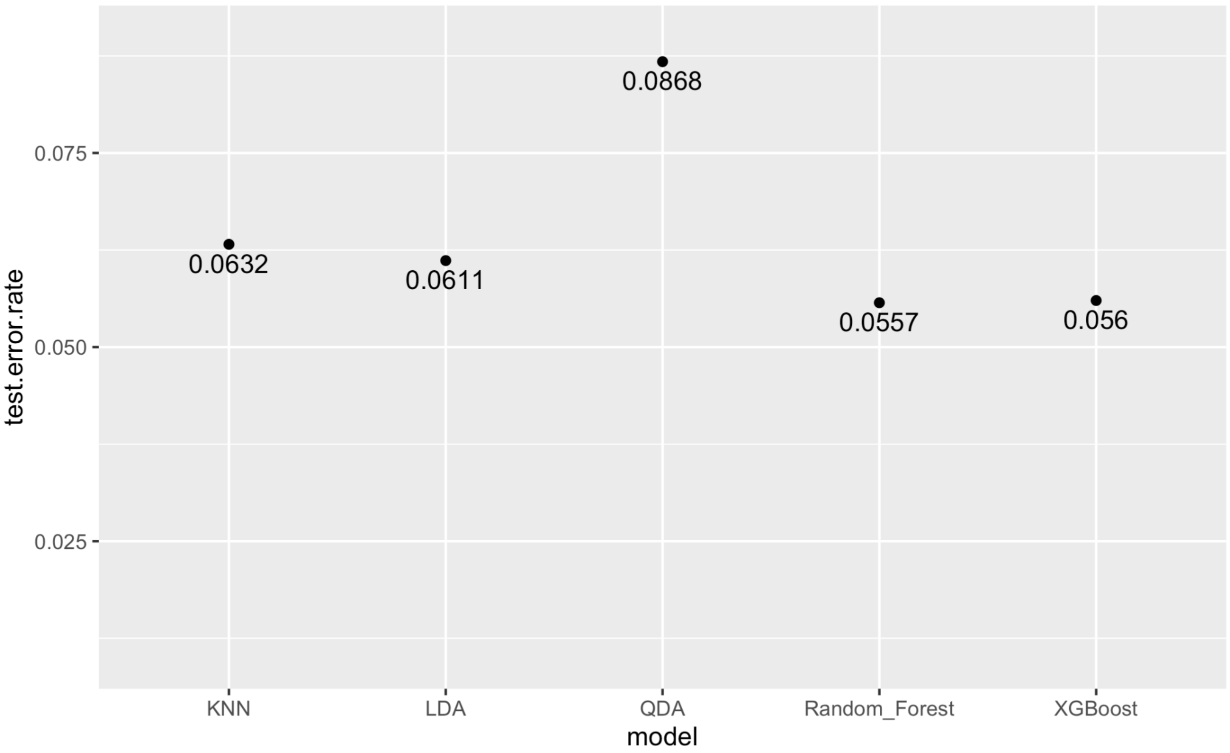
Candidate Model 3: KNN

The kNN (k-Nearest Neighbors algorithm) is a supervised, non-parametric machine learning algorithm that groups data observations based on proximity with other observations. This is often measured with the euclidean distance of the data point, in relation to another point, and then evaluating the classification of the data point based on the classification of others. Thus, kNN is commonly used for classification analysis. Upon running kNN, we arrived at an error rate of 0.0632.

Candidate Model 4&5: LDA/QDA

Linear and Quadratic Discriminant Analysis is a supervised learning algorithm, commonly used for data classification, and dimension reduction. Through the use of Bayes classification, LDA and QDA create boundaries that separate the data by linear or quadratic decision boundaries. After running LDA and QDA on our dataset, we arrive at an error rate of 0.0632 and 0.0868, respectively.

The graph below shows a detailed comparison of the test error rate for the 5 models we constructed.



Error Rate Comparison of the 5 Models

MODEL ANALYSIS

By analyzing the models mentioned above, we can clearly compare the advantages and disadvantages of different models.

KNN model:

Pros: Simplest Model and flexible
Cons: Suffers from high dimensionality, time consuming

LDA model:

Pros: Simple and Fast
Cons: Not as good for categorical variables

QDA model:

Pros: More flexibility for the covariance matrix
Cons: Cannot be used for dimensionality reduction

Random Forest model:

Pros: Accurate and Flexible
Cons: Computationally Intensive

XGBoost model:

Pros: Accurate and Flexible
Cons: Cannot perform on unstructured data

In summary, Random Forest and XGBoost are more accurate and flexible, whereas kNN, LDA, and QDA models are more simple. However, kNN, LDA, QDA are not good for high dimensionality categorical variables.

Through the graph of error rates above, we observe that the QDA model yields the highest error rate. Furthermore, considering its limitations, which do not outweigh its poor misclassification rate score, we deleted it from our choices. The same applied to kNN and LDA, which did better comparatively but still did not perform as well as Random Forest or XGBoost, which were our top two choices. Considering that Random Forest slightly edges out XGBoost, and the fact that XGBoost has slightly more steps and is more computationally intensive than Random Forest, we finalized Random Forest as our chosen model.

Meanwhile, when submitting our prediction to Kaggle, we decided to combine the results of multiple models in order to incorporate the performance of different models in one single prediction. To achieve this, we selected the prediction of three models (kNN, Random Forest, and XGBoost) and did a majority vote on all the prediction results. With the combined result, we hope to generate a final result that is adaptable to multiple models and have a lower variance for our prediction.

CONCLUSION

In summary, we selected our 19 predictors as a subset of our transformed and data-cleaned variables, and then decided to use the Random Forest algorithm to predict the Severity of Car Accidents on the testing dataset. We selected the 20 variables based on our generated Importance Plot, which provided us with a detailed look on how significant our variables were. In addition, we selected the Random Forest Algorithm as it yielded the lowest misclassification rate and the best results on the public kaggle scoreboard, with an error rate of 0.94355.

One strength of our modeling process was the holistic nature of our investigation. To ensure that we implemented a well-performing model for our dataset, we used five separate modeling methods, including kNN, LDA, QDA, XGBoost Random Forest. From this, we then calculated the misclassification rate of each model to determine the model that was best suitable in our case from a large pool of candidates.

We also attribute the high score we achieved in our model to the use of advanced functions to impute the NA values, namely mice. Instead of deleting all the NA values, which would negatively impact our prediction score due to the large number of NA values in the dataset, or simply replacing all NA values with the median or mode, which may also have been a problem as the median without the missing values may not represent the true median, we worked around this by using the mice function.

Of course, there are also shortcomings to our model that we were not able to overcome, especially in the given time constraints. Firstly, our selected data provides a very narrow time range, only considering accidents from 2015 to 2021. While this may better represent the severity of accidents that are happening nowadays, the lack of range in year means our data is narrow, and older car crashes are either underrepresented or not represented at all.

Secondly, we heavily relied on model accuracy to determine our best model. Whilst Random Forest produced the lowest misclassification rate and the highest kaggle score for us, this ignores many other factors that could prove to be useful when evaluating models. For example, we did not consider efficiency or runtime, which could be a major factor if we were working with a significantly larger dataset. Our best model happened to be Random Forest, but in different circumstances (such as if efficiency is another key factor) this may have been different.

REFERENCES

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

Simon, Shelby. “How Many People Die from Car Accidents Each Year?” *Forbes*, Forbes Magazine,7 Nov. 2022, https://www.forbes.com/advisor/legal/auto-accident/car-accident-deaths/.