STATS 101C Final Project

Predicting the Severity of Car Crashes

Avani Kanungo, Kathy Nguyen-Ly, Jenna Schindele, Sarah Zhari

(Lecture 2, Group E)

ABSTRACT

The United States is home to the highest rates of vehicular accidents, fatal and non-fatal. As such, there is a high incentive to identify risk factors that are more conducive to accidents, especially severe ones. Using a given data set of 35,000 incidents from 2016 to 2021, we successfully predicted the severity levels of 15,000 testing incidents with a 94.43% accuracy rate using a final random forest model with 25 individual predictors. Ultimately, our results show that the text-mined variables of road closures (from our original description) and weather conditions were most significant in predicting traffic accident severity.

1. INTRODUCTION

According to Statista, the United States experiences the world's highest proportion of traffic incidents per million citizens (Carlier), and as a result, traffic incidents are the leading cause of death for Americans under 54 years of age (Center of Disease Control and Prevention). With so many deaths and injuries stemming from these incidents, understanding the conditions that are conducive to more severe accidents is critical in harm reduction. As such, for our project, we seek to predict the severity of accidents in the United States, reflected in the accident's impact on traffic. Our data includes both external conditions (weather and date) as well as descriptions of the accident (text and location) for 35,000 individual incidents between 2016 to 2021, from which we predicted accident severity.

2. METHODOLOGY

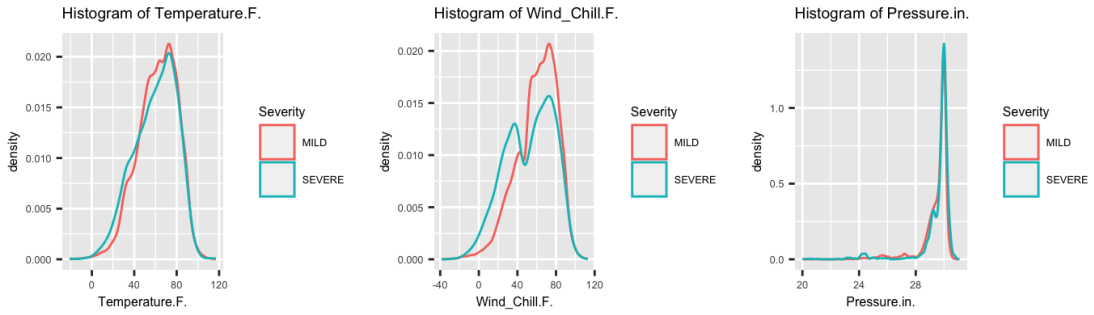
Data Analysis

A. Data Exploration

To begin our data analysis and modeling, we performed exploratory data analysis (EDA) which involves investigating the composition of data sets including the number of variables, the types of variables, and missing values (if applicable). From an initial exploration, we identified 44 variables, which includes 1 response variable and 43 predictors. Moreover, we flagged variables with missing, or NA, values, as shown below:

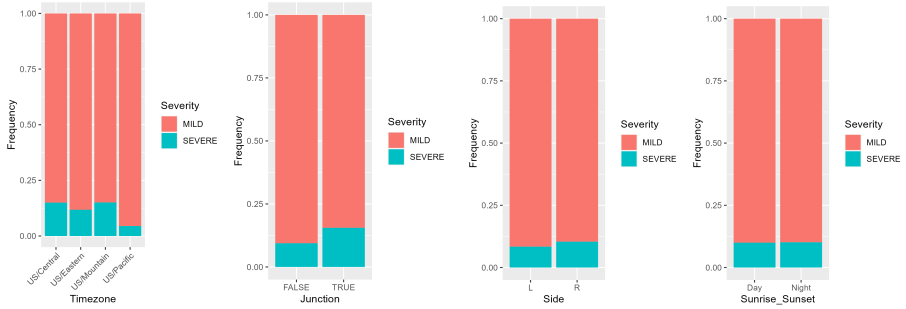
Variable.Names	Number.of.NAs
Wind_Chill.F.	5666
Wind_Speed.mph.	1870
Humidity...	847
Wind_Direction	845
Visibility.mi.	822
Weather_Condition	810
Temperature.F.	804
Pressure.in.	671
Weather_Timestamp	569
Airport_Code	117
Timezone	51
Sunrise_Sunset	30
Civil_Twilight	30
Nautical_Twilight	30
Astronomical_Twilight	30
Zipcode	18
City	1

To determine whether specific numerical variables would be more important and useful than others, we created density plots for each numerical predictor, separated by *Severity*.



However, as evident in a few example density plots above, there was no prominent split between "MILD" and "SEVERE". Therefore, we concluded that using these numerical values as is would not be beneficial in distinguishing between Mild and Severe car accidents.

Similar to the numerical variables, we created 100% stacked bar charts for the categorical and logical variables. The three variables that presented possible differences in proportion of "MILD" and "SEVERE" were Timzone, Junction, Side, and Sunrise_Sunset, as shown below.



Finally, we identified two variables, *Description* and *Weather_Condition*, that acted as open text variables. As is, these variables do not act as useful predictors. However, as we will discuss later, we used text mining and were able to extract words and phrases from these variables to create new, more productive variables.

B. Data Cleansing and Variable Creation

To solve the issue of missing values, we used various data cleansing methods. The package *mice* allowed us to impute all numerical missing values. Below are the NA counts following the implementation of mice:

Variable.Names	Number.of.NAs
Wind_Direction	845
Weather_Condition	810
Weather_Timestamp	569
Airport_Code	117
Timezone	51
Sunrise_Sunset	30
Civil_Twilight	30
Nautical_Twilight	30
Astronomical_Twilight	30
Zipcode	18
City	1

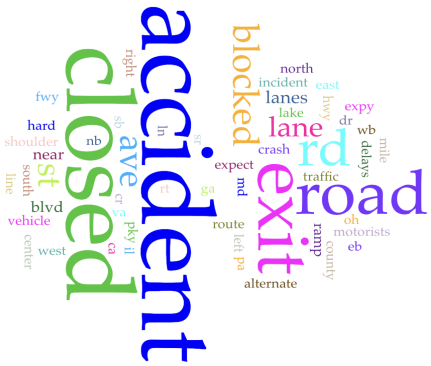
However, as identified through EDA, these numerical values were not of high importance in our data modeling. For the variable, *Timezone*, we imputed missing values with the mode of the timezone of other accidents in the same state.

After dealing with the missing values, we created new, useful numerical variables. From the variables involving time, latitude, and longitude, we created three new variables: change in time, change in latitude, and change in longitude. Moreover, we used the variable, *Start_Time*, to create three additional predictors (two categorical and one numerical): year, month, and hour. Specific years, such as 2020, had lower rates of driving and therefore witnessed fewer accidents. Similarly, specific months, such as the holiday months of November and December, experience higher rates of driving, resulting in an increase in accidents. Finally, the late, dark hours, ranging from dusk to dawn, seem to be correlated to more severe accidents.

We also created new, more useful variables from the text variables, *Description* and *Weather_Condition*. We created the following two word clouds from the variable *Description*, one for each type of Severity:

MILD

SEVERE



From the word clouds, we concluded that the following words would be useful in modeling: closed, accident, traffic, blocked, caution, and incident. We also identified that the phrase, “Road

closed due to accident" was useful in distinguishing between mild and severe accidents. Hence, we created new logical variables in which TRUE indicated that the *Description* for that accident record included the respective word/phrase.

Similar to *Description*, we used a similar methodology to create new variables from the variable, *Weather_Condition*. After tabling the responses for this column, we identified recurring weather conditions and the various phrasing used to identify said condition. Our final weather condition logical variables included: rain, cloudy, windy, thunderstorm, fog/haze, snow, and fair/clear.

C. Variable Selection

As discussed above, we concluded that the numeric variables and most categorical/logical variables would not be useful due to their minimal difference in densities of *Severity*. Therefore, we decided to select the new numeric, logical, and categorical variables we created, as well as *Timezone* and *Side* (from the 100% stacked charts above). Our final 25 predictors, compared to the original 43 predictors, are as follows: Year, Month, Hour, Side, State, Timezone, Change_Time, Change_Lat, Change_Lng, Closed, closed, Accident, accident, Traffic, Blocked, Caution, Incident, Rain, Cloudy, Windy, Thunderstorm, Haze, Snow, Clear, roadClosed.

Modeling

A. Logistic Regression

Logistic regression, compared to other classification algorithms, is rather simple. This algorithm uses the following probability formula to classify each record:

$$\theta = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}$$, where β_p is the coefficient for X_p

Since the number of predictors (25) is much smaller than the size of our training data set (35,000) and testing data set (15,000), logistic regression is very efficient to run. When we applied our logistic regression model, using all 25 predictors, we achieved a classification accuracy rate of 93.46%.

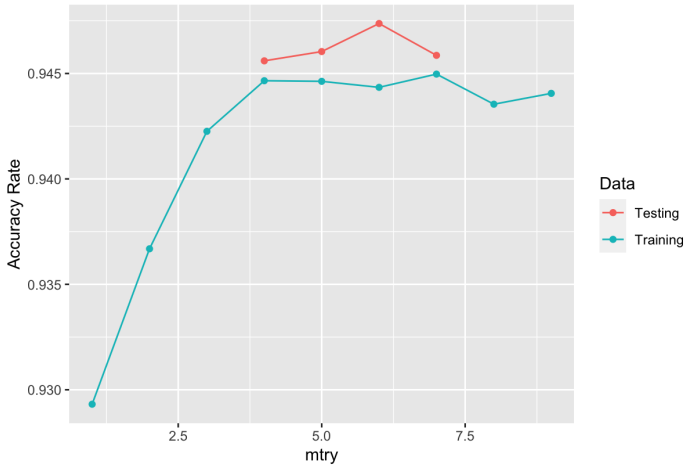
B. Tree Model

Similar to logistic regression, tree models are often fairly simple and efficient to run. Tree models are successful when the data can be split up into rectangular partitions. At each node of the tree, the algorithm tests and determines which branch the data will follow, ultimately ending at a terminal node for final classification. After running our tree model, we ended with a classification accuracy rate of 93.28%.

C. Random Forest

Random forest is one of the most powerful data modeling and classification algorithms. The algorithm randomly selects a subset of the training data to use in each tree model it generates. Moreover, when the parameter, *mtry*, is used, the algorithm also selects a subset of the predictors to use in the tree model. By doing so, random forest is able to reduce bias while also maintaining low variance. Furthermore, the algorithm runs rather efficiently on large datasets while still producing highly accurate results.

One of the keys to unlocking the power of random forest is to select the best value for the parameter, *mtry*. To do so, we ran several models of random forest using various values of *mtry*, ranging from 1 to 9. We compared the accuracy rates to determine which value(s) would be most effective for our data set.



We notice that *mtry* = 4, 5, 6, 7 achieve the highest training accuracy rates, with *mtry* = 7 being the highest at 94.5%. However, although *mtry* = 7 had the highest accuracy rate on our training data, *mtry* = 6 achieved the highest accuracy rate on our testing data at 94.74%. Therefore, we select our random forest model with *mtry* = 6 for comparison against other models.

D. Support Vector Machine (SVM)

Support vector machines are similar to logistic regression in terms of the outcome and accuracy. The kernel in each support vector machine determines the shape of the classifying boundary. By tuning this parameter, one can improve the effectiveness and predicting power of this algorithm. We ran 4 models of support vector machines using the most common kernel functions: linear, radial, sigmoid, and polynomial. Ultimately, the linear kernel produced the most accurate classification in both the training (left table) and testing (right table) data.

Kernel	Accuracy
Linear	0.9335
Radial	0.9334
Sigmoid	0.9285
Polynomial	0.9011

Kernel	Accuracy
Linear	0.9333
Radial	0.9329
Sigmoid	NA
Polynomial	NA

3. RESULTS & DISCUSSION

The final model used was a random forest model with *mtry* = 6. The final model used 25 predictors and utilized all 35,000 predictors in the testing data. The accuracy rate was 94.43%, with a final public Kaggle score of 94.74% and a private Kaggle score of 94.00%, averaging out to an overall score of 94.55%. Within the lecture, we placed third.

The most important predictors used in the model can be categorized into 4 categories: description, weather, time, and distance. We managed to create 9 new predictors simply from using the variable "Description" in the original dataset. The mention of "closed" or "blocked" roads, "accident(s)", and "traffic" were strong indicators of severe accidents. This aligns with our prior understanding of car accidents, as severe car accidents are more likely to result in road closures or major disruptions. One particularly interesting finding was the importance of separating the predictors "Closed", with an uppercase "C", and "closed" with a lowercase "c". Distinguishing the uppercase and lowercase letters as written in the description was helpful in categorizing the severe and mild accidents, where descriptions written with an uppercase "C" for "Closed" were more likely to be severe accidents.

Weather was also an important predictor in predicting the severity of car accidents. 7 new predictors were created from the "Weather" variable in the original dataset. The importance of weather in predicting car accidents was unsurprising – foggy, rainy, or hazy weather can impact the visibility of the drivers, leading to more severe accidents. Moreover, rain and snow can cause roads to be more slippery, which can also cause more severe accidents.

Time was also crucial in our model. Specifically, the duration of the accident was an important indicator of severity. Accidents that take longer to be cleared out were more likely to be severe.

Furthermore, we also utilized the reported year, month, and hour of the accidents. The dataset included years affected by the COVID-19 pandemic, which caused fewer cars to be on roads, leading to less severe accidents. The increased driving rates during holiday months and decreased visibility during the hours from dusk to dawn also led to more severe accidents.

Finally, the distance covered by the accident, as measured by a difference in the latitudes and the longitudes, were also important in our model. A larger distance covered by the accident could indicate more severe accidents. Overall, description and weather were the most important predictors used to create this model.

4. LIMITATIONS & CONCLUSION

We encountered a few limitations regarding the variables available to us as well as handling missing values. Out of the 11 numerical variables provided, none of them were significant in our models as-is. Calculating the change in time, latitude, and longitude over the course of the accident proved to be beneficial to our model, which used 6 of 11 numerical variables to create 3 new variables. The 32 categorical and logical variables mostly weren't helpful as is, but extracting phrases from categorical variables became the basis of our model. The variables Side, Timezone, and State were included, as well as 19 variables created from Description, Weather_Condition, and Start_Time.

None of the variables used could use MICE to impute NAs, as they either had no NAs or were categorical variables. In the variable Timezone, NAs were handled by using the most common Timezone for that state, but this does not account for states that have more than one timezone. For Weather_Condition, NAs were set to the value "None" which was not included in the final model. There were relatively few categories in Weather_Condition, so we could condense them

into the 7 weather variables easily. Our new variables were logicals based on the presence of a phrase, so if there was an NA converted to “None,” the phrase wasn’t present, so the value was set to FALSE.

For the variables extracted from Description, each value was unique, so we used str_detect() to determine if a phrase was present or not. str_detect() is very specific and case sensitive, which we used to our advantage with “Accident” vs “accident,” but we may have missed out on variations of words (ex. Blocked vs blocking). This could have been prevented by using the root of the word and variable characters in our data mining process.

Additional variables that are often found in police reports could have been helpful in creating a more accurate and maybe simpler model. Details of the car’s make, model, size, and other factors could be beneficial in assessing the severity of the accident, as well as classifying the area as urban or rural. It would be extremely useful to have information on the condition of the driver or other people involved in the accident, as a drunk or otherwise impaired driver can cause a lot of damage. Knowing the emergency responders who reported to the scene also can indicate the severity based on who was needed.

In our efforts to make the best model, we attempted logistic regression, tree models, random forests, as well as support vector machines. We did not try splines, PCA, boosting, or bagging, which may have given us a better fit model. Despite these limitations, our 25 variables created a model with a final R^2 of 94.55% and performed very well in the Kaggle competition. We believe that our detection of capital and lowercase key words like accident and closed gave us the edge against other teams, and we were able to place 3rd in lecture 2 with a random forest model containing only 25 predictors.

5. ACKNOWLEDGEMENT

We would like to acknowledge Professor Almohalwas for the creation and guidance throughout this project. We would not have been able to achieve such a successful model without his support, insight, and encouragement to continue to strive for greatness.

REFERENCES

Almohalwas, Akram. “Bagging and Random Forests” (Lecture, STATS 101C, University of California, Los Angeles, November 14, 2022).

Almohalwas, Akram. “Classification Methods” (Lecture, STATS 101C, University of California, Los Angeles, October 12, 2022).

Almohalwas, Akram. “Chapter 8 CART” (Lecture, STATS 101C, University of California, Los Angeles, November 14, 2022).

Almohalwas, Akram. “Merging Data and Imputation of Missing Values” (Lecture, STATS 101C, University of California, Los Angeles, November 16, 2022).

Almohalwas, Akram. “Support Vector Machines” (Lecture, STATS 101C, University of California, Los Angeles, November 21, 2022).

Carlier, Mathilde. “Road accidents in the United States - Statistics & Facts.” *Statista*, Ströer Media, 25 January 2022. https://www.statista.com/topics/3708/road-accidents-in-the-us/#dossier-chapter4

Center for Disease Control and Prevention. “Global Road Safety.” *Center for Disease Control and Prevention*, United States Department of Health and Human Services, 14 December 2020. https://www.cdc.gov/injury/features/global-road-safety/index.html#:~:text=Road%20Traffic%20Injuries%20and%20Deaths%E2%80%94A%20Global%20Problem&text=Road%20traffic%20crashes%20are%20a,citizens%20residing%20or%20traveling%20abroad.