Predicting the Severity of Missouri Traffic Accidents

Kelly McDowell

November 6, 2020

1. Introduction

1.1 Overview

The economic and societal impact of traffic accidents cost U.S. citizens hundreds of billions of dollars every year. Reducing traffic accidents, especially serious accidents, is an important challenge. A proactive approach would focus on education in order to prevent potentially unsafe driving and road conditions from occurring in the first place. For effective implementation of this approach, severity prediction based on road and weather conditions is critical. If we can identify the patterns of how these serious accidents happen and the key factors that contribute to severity, we may be able to implement well-informed driving education in Missouri.

1.2 Objectives

This project focuses on predicting the severity of an accident based on occurrence parameters, road, and weather conditions in the state of Missouri. The first objective will be to identify key factors affecting an accident's severity, with a focus on conditions that cause the most severe accidents. The second objective will be to develop multiple models that can accurately predict accident severity based on the key factors identified. Missouri is known for unpredictable and rapidly changing weather. This model will help Missourians make an informed decision regarding how and when they should drive based on the current weather and road conditions' likelihood to lead to a severe accident.

1.3 Data Overview

This project is starting with a countrywide traffic accident dataset, which covers 49 states of the United States. The data is continuously being collected from February 2016, using several data providers, including two APIs which provide streaming traffic event data. These APIs broadcast traffic events captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are approximately *3.5 million* accident records in this dataset. I will primarily utilize the accident data recorded for the state of Missouri for this project.

1.4 Acknowledgements

* Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. [“A Countrywide Traffic Accident Dataset.”](https://arxiv.org/abs/1906.05409), arXiv preprint arXiv:1906.05409 (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.”](https://arxiv.org/abs/1909.09638) In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

2. Data acquisition and cleaning

2.1 Data Sources

The dataset being used has been collected in real-time, using multiple Traffic APIs. Currently, it contains accident data collected from February 2016 to June 2020 for the Contiguous United States. Majority of the data came from MapQuest, with less than half coming from Bing. (Figure 1)

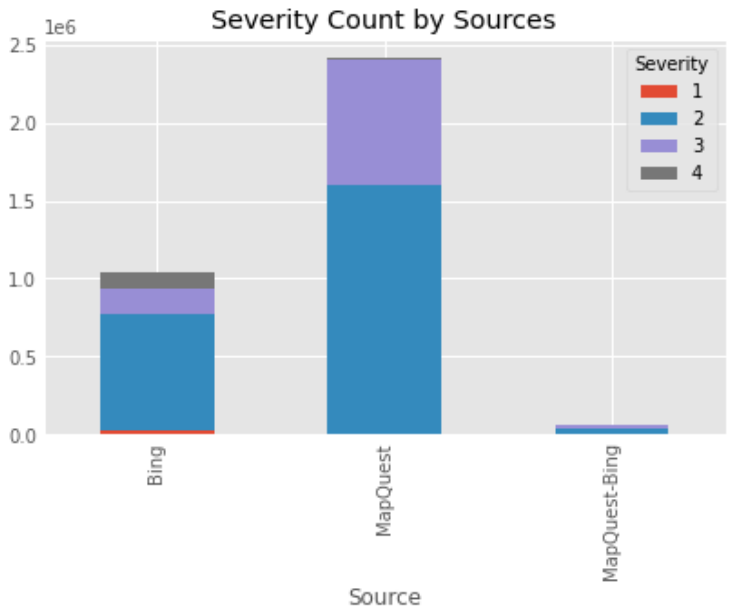

Figure 1. Data received from Bing and MapQuest with severity breakdown

2.2 Data Cleaning

Data cleaning began with fixing the dates and times associated with each accident. First, I converted the 'Start_Time' and 'End_Time' to date types. Then, extracted the year, month, day, hour, and weekday. I also extracted the amount of time in the unit of minutes for each accident. Any rows with a negative 'Time_Duration' were dropped. I removed the outliers for 'Time_Duration' and backfilled these with the 'Time_Duration' median.

A second area of focus for cleaning was the weather features. Weather features such as 'Wind_Chill(F)' and 'Precipitation(in)' had more than 60% missing values. 'Wind_Chill(F)' was dropped because this feature is not highly related to severity, whereas 'Precipitation(in)' could be a useful predictor and was handled by separating the feature. I added a new feature for the missing values and replaced these missing values with median.

Lastly, since this project is focused on accidents in the state of Missouri, I copied the dataset to one with accidents occurring only in Missouri (Figure 2). I then dropped the 'State' feature.

Figure 2: Scatterplot of accidents in Missouri from 2/16 through 6/20

2.3 Feature Selection

After cleaning the data, my Missouri dataset contained 33,362 records and 53 features. It was immediately clear that some features would not be useful in the exploration or prediction of accident severity. I selected the pertinent features and left behind the features with less relevance to my analysis (Table 1).

Kept Features	Dropped Features	Reason for Dropping
Source, Severity	TMC, Description	Free-form text fields
Start_Time, Start_Lat, Start_Lng, Distance, Year, Month, Day, Hour, Weekday, Time_Duration(min)	End_Time, End_Lat, End_Lng,	Ending accident time or coordinates not needed.
Side, City, County	Number, Street, Zipcode, Country, Timezone, Airport_Code	Granular location details not needed. All accidents in analysis occur in Central Standard Timezone in the United States.
Tempurature(F), Humidity(%), Pressure(in), Visibility(mi), Wind_Direction, Wind_Speed(mph), Precipitation, Weather_Condition	Weather_Timestamp, Wind_Chill	Weather timestamp not as important as accident Start_Time. Wind chill is not as important as the temperature recorded.
Amenity, Bump, Crossing, Give_Way, Junction, No_Exit, Railway, Roundabout, Station, Stop, Traffic_Signal, Turning_Loop	Traffic_Calming	100% of dataset reported Traffic_Calming as False.
Sunrise_Sunset	Civil_Twilight, Nautical_Twilight, Astronomical_Twilight	Redundancy in recording whether accident occurred during night or day.

Table 1: Features selected or dropped for exploratory analysis

3. Methodology – Exploratory Data Analysis

3.1 Exploration of time features

I examined the impact of each time feature on accident severity, starting with the count of accidents by year (Figure 3). This dataset is up-to-date through June 2020, so the accidents showing for 2020 are potentially only 50% of the total for the year. I did not glean any important insights from the count and severity of accidents by year, beyond an increase in Severity 3 and 4 accidents from 2016 through 2019.

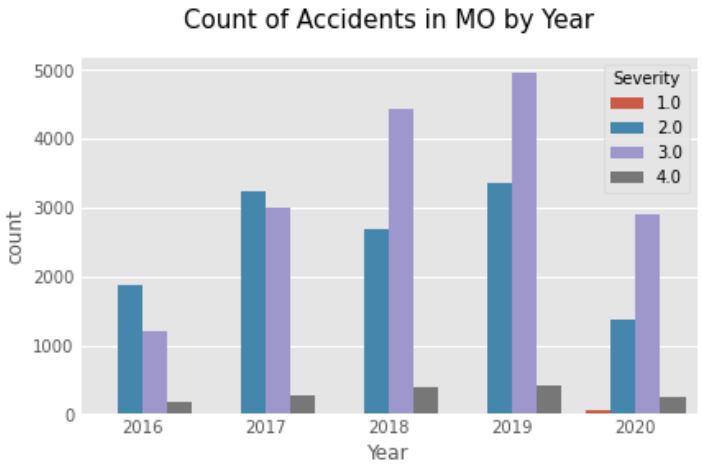

Figure 3: Count of accidents in Missouri by year with severity breakdown

The count of accidents by month (Figure 4) shows the months with the highest propensity for accidents are March, April, May, and June. Months with the lowest accident rates are July and September. The month with the highest propensity for both a Severity 3 and 4 (highest) accident is June.

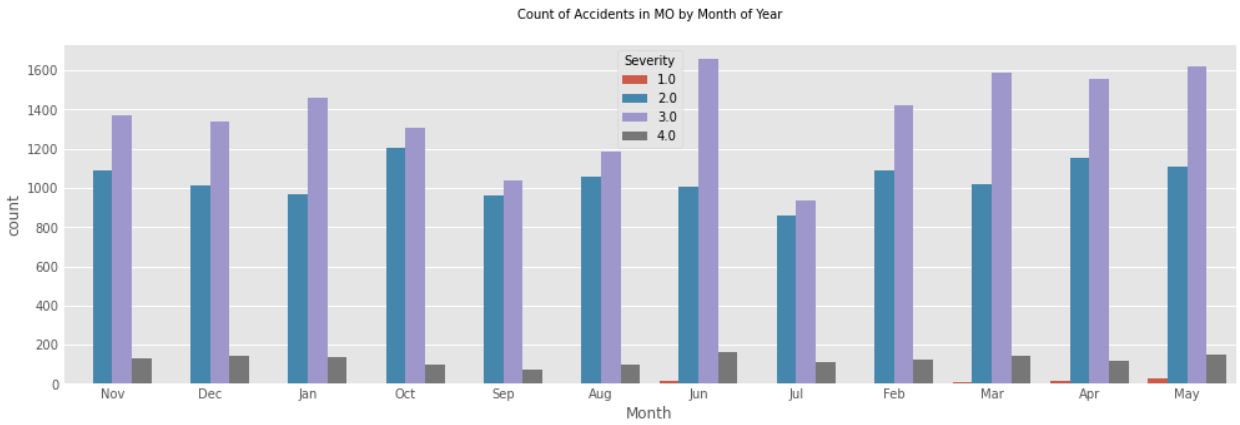

Figure 4: Count of accidents in Missouri by month with severity breakdown

The count of accidents by hour of day (Figure 5) shows most prominently that accidents occur during rush hour or commuting times of the day. Accidents are most likely to occur during the AM hours of 7 and 8 o'clock. They dip between the hours of 9 AM through 2 PM before ramping up at 3 PM. Accidents peak again and plateau from 4 to 5 PM before dropping off for the remainder of the day. Accidents are least likely from midnight to 3 AM when, presumably, the least amount of traffic is on the roads. Noon, 4, and 5 PM are the hours of the day that appear to have the highest propensity for a Severity 4 accident.

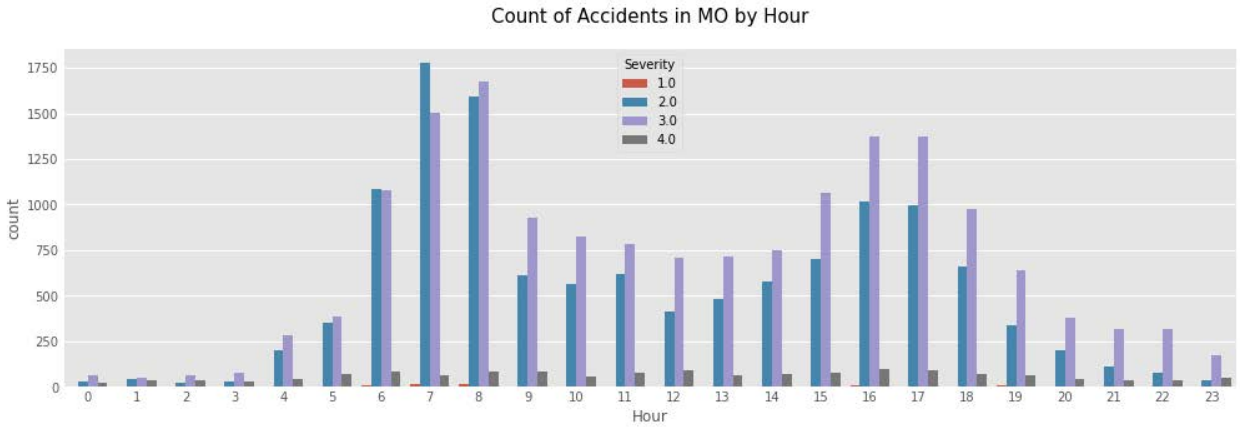

Figure 5: Count of accidents in Missouri by hour of the day with severity breakdown

Severity 1 (least severe) accidents are nearly non-existent. I believe the number Severity 1 accidents are insignificant in this dataset due to a lack of reporting and time needed for cleanup.

In the dataset, period of the day is also indicated (Figure 6). I focused on the Sunrise_Sunset feature which shows the period of day (i.e. night or day) based on sunrise/sunset. Clearly, far more accidents occur during the day. Severity 4 accidents occur at a higher number during the day, but the chance of a Severity 4 accident at night is greater.

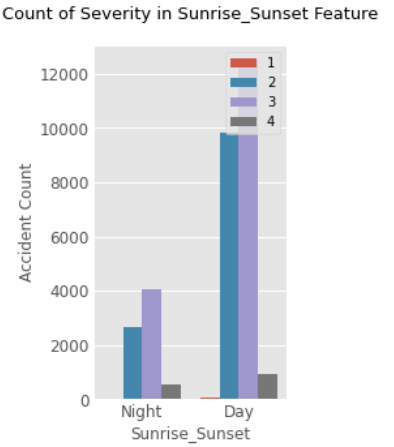

Figure 6: Count of accidents in Missouri by period of day with severity breakdown

3.1 Exploration of weather features

I explored the weather condition at the time of Missouri accidents for all accident severities combined (Figure 7) and for each severity level individually. Since this project focuses on the most severe accidents, I will only provide the visual for Severity 4 (Figure 8). I was surprised to find the most common weather conditions at the time of an accident included bland descriptors such as clear, fair, mostly cloudy, overcast, and so on. Light rain and rain combined were the weather condition for 9% of all severity levels accidents.

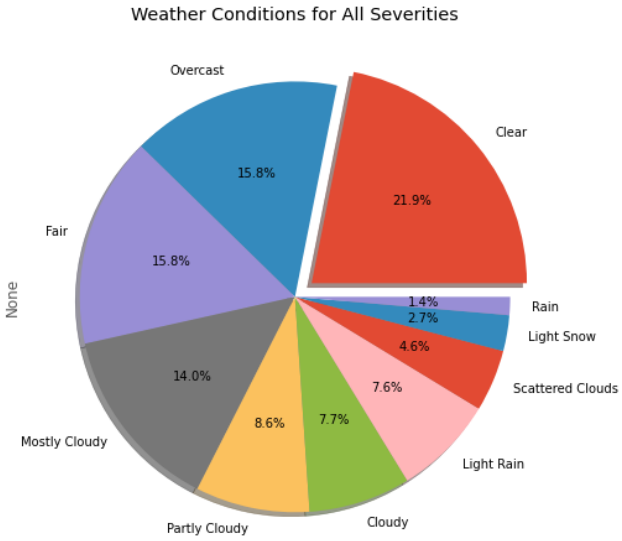

Figure 7: Top 10 weather conditions for all accident severity levels

Similar to the weather conditions for all severity levels, conditions for Severity 4 accidents most commonly do not include precipitation.

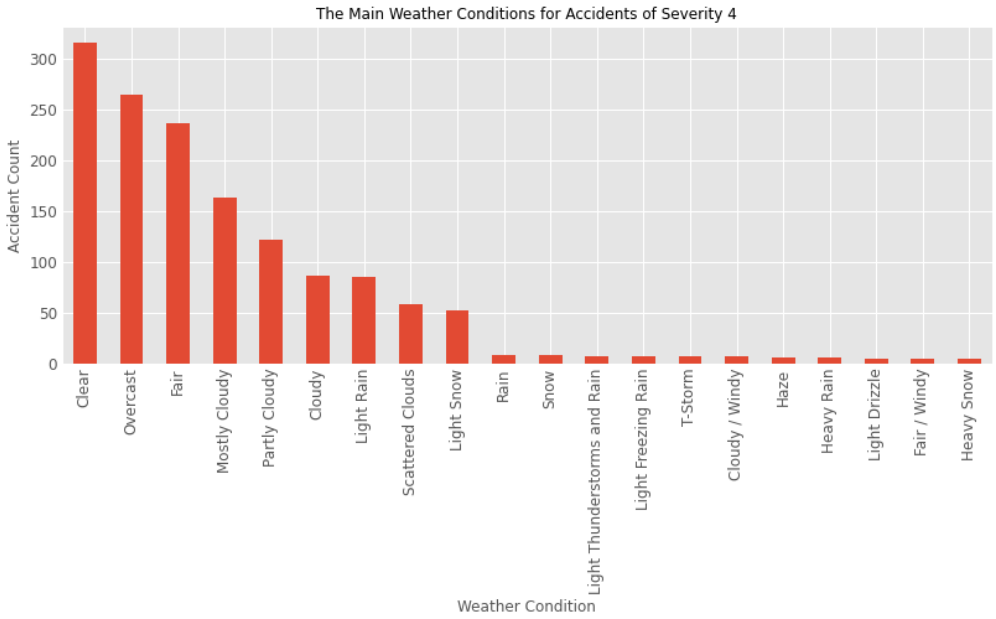

Figure 8: Top 20 weather conditions for Severity 4 accidents in Missouri

3.1 Exploration of where accidents are occurring

I wanted to find out if accidents were most likely to occur on a highway or other street types. Within the street type feature, I grouped freeway, expressway, and highway as 'Highway', leaving the remaining street types as 'Other'. This analysis (Figure 9) shows the probability of an accident occurring on a Highway is much greater than on other streets.

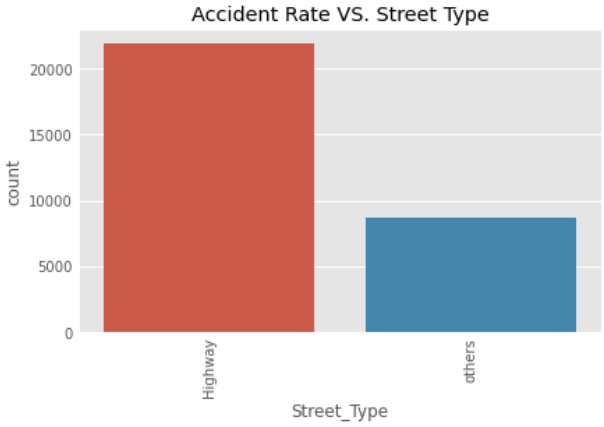

Figure 9: Accident rates in Missouri on Highway vs. other street type

Road features were also explored to identify which had an impact on the likelihood of an accident occurring (Figure 10). The two most common road features for an accident occurring were junctions and traffic signals. No accidents occurred at bump, roundabout, or turning loop road features.

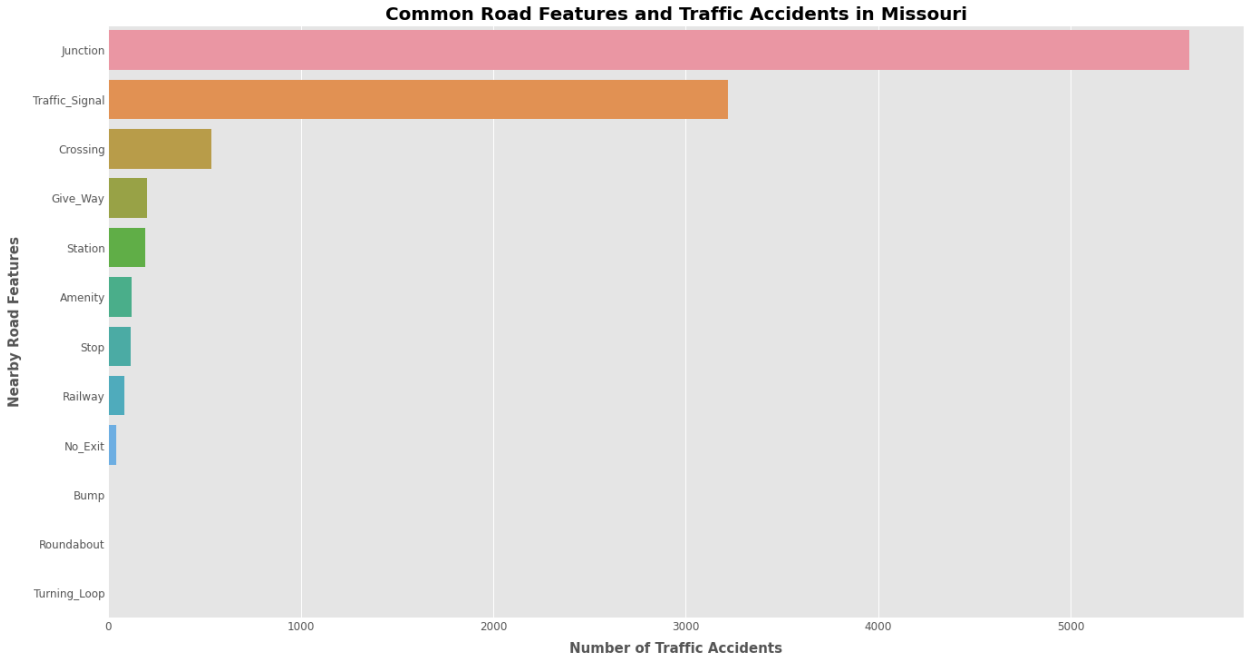

Figure 10: Accident rate for common road features in Missouri

After exploring the dataset features and producing a correlation heatmap of the features, I removed the features not correlated to severity from my feature set before moving on to machine learning. I removed 'Source' because the source of the accident data is not correlated to the likelihood of an accident occurring. I removed 'Street' and the road features 'Turning_Loop', 'Roundabout', and 'Bump' due to zero accidents being related to these features. I removed the 'Year', 'Month', and 'Day' features due to a lack of correlation. Lastly, the 'Distance' and 'Time_Duration' features needed to be removed because these features have a post-accident connection rather than an accident causation connection. 'Start_Time' also needed to be removed since this feature was previously broken down into smaller time units.

4. Methodology – Predictive Modeling

The feature list utilized for the machine learning portion of the project first needed object features reclassified as categorical features. I used One-hot encoding for the following features: 'Side', 'City', 'County', 'Wind_Direction', 'Weather_Condition', 'Sunrise_Sunset', 'Weekday', and 'Street_Type'.

I chose four different supervised machine learning algorithms to predict accident severity, including Logistic Regression, K-Nearest Neighbors, Decision Trees, and Random Forest. I standardized the features based on unit variance and split the data into X_train, X_test, y_train, and y_test. The number of points used for training was about 24,000 and the test size was about 6,000.

4.1 Logistic Regression

Logistic Regression is the appropriate regression analysis to utilize when the dependent variable is binary. The Logistic Regression predictive analysis is used to describe data and explain the relationship between variables. I increased the number of iterations to 1,000 to help the algorithm to converge. The Logistic Regression provided a very poor accuracy score of 0.654.

4.2 K-Nearest Neighbors (KNN)

The K-Nearest Neighbors algorithm is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. I began the KNN algorithm with 6 neighbors as a starting point which provided a KNN accuracy score of 0.601. I then optimized the number of neighbors and plotted the accuracy versus number of neighbors (Figure 11). I found k=7 provided the best KNN accuracy score of 0.617.

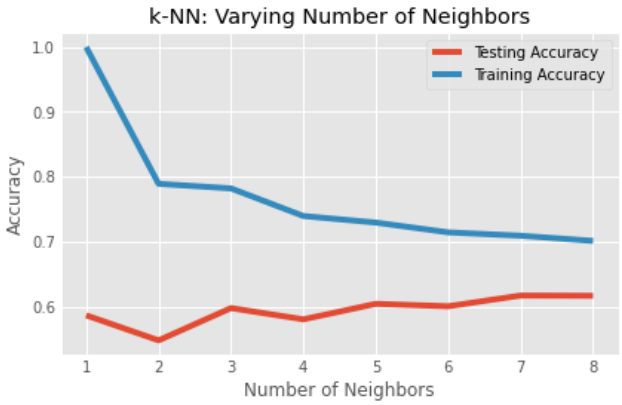

Figure 11: Testing and training accuracy for k = 1 through 8

4.3 Decision Tree

Decision Tree Classifier was used to improve accuracy. In this method, a set of training examples is broken down into smaller and smaller subsets while at the same time an associated decision tree gets incrementally developed. I used two different splitting measures: Entropy and Gini Index. Entropy is the measurement of the impurity or randomness in the data points. The Decision Tree – Entropy accuracy score was 0.657. This indicates there is a lot of variance and uncertainty in the feature dataset.

Next up was Gini Index which calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly. A value of 0 expresses the purity of classification, while 1 indicates the random distribution of elements across various classes. The resulting Decision Tree – Gini Index accuracy score was 0.665 which is close to an equal distribution of elements. Accuracy was not improved as much as I expected through Decision Tree machine learning.

4.4 Random Forest

To improve accuracy further, the Random Forest algorithm was used. Random Forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes the model's prediction. By using Random Forest Classifier, the test accuracy improved to 0.709.
A visualization of the important k features (Figure 12) indicates the top 30 features in predicting accident severity. There are only 7 features with an importance threshold of more than 0.03, including: 'Start_Lng', 'Start_Lat', 'Temperature(F)', 'Humidity(%)', 'Hour', 'Pressure_bc', and 'Wind_Speed_bc'.

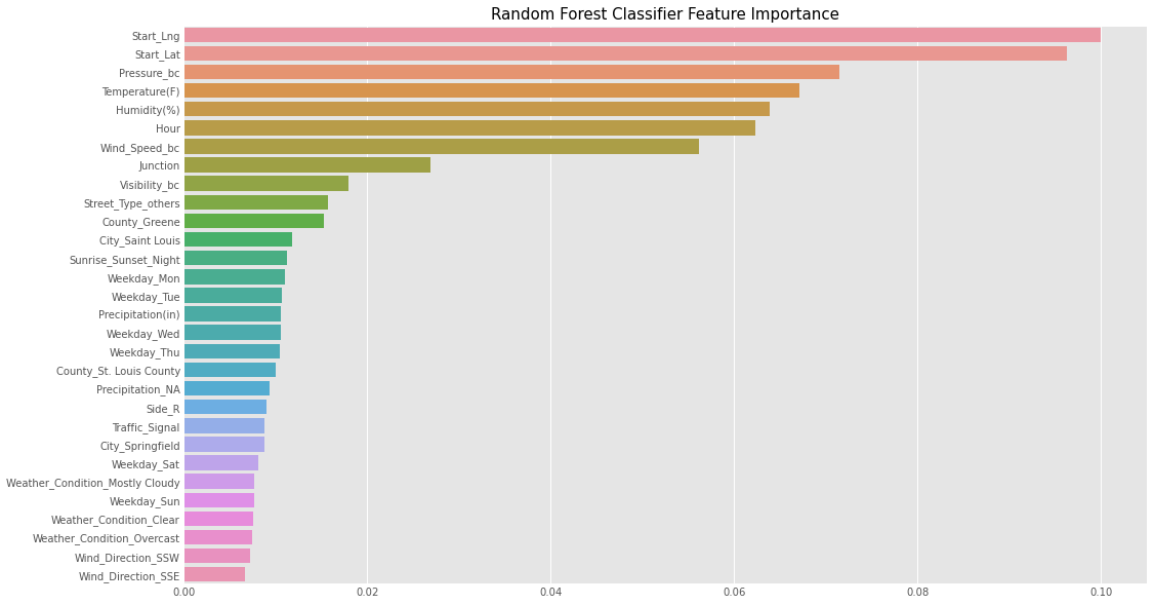

Figure 12: Visualization of top 30 important feature to predicting accident severity

Accuracy comparison of each machine learning method (Figure 13) shows the K-Nearest Neighbors performed worst with a 0.617 accuracy, Logistic Regression performed with a 0.654 accuracy rate, Decision Tree Classification came in third at 0.665, and Random Forest had the best predictability at 0.709.

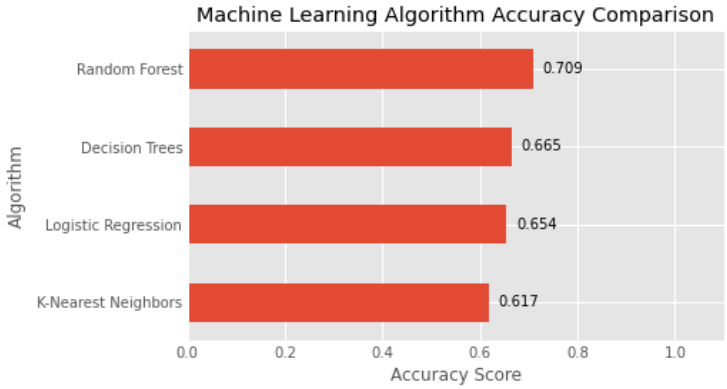

Figure 13: Accuracy comparison of Algorithms used to predict accident severity

5. Results & Discussion

Four machine learning algorithms were utilized to predict accident severity. The accuracy of all four was below my expectation and what I believe to be an acceptable accuracy rate within the data science field. The Random Forest machine learning algorithm proved to be the best predictor of accident severity with an accuracy rate of 0.71.

I believe one detractor to accurate machine learning was unbalanced data in my feature set. For example, the road feature 'Stop' has 117 Trues versus 30,468 Falses. There are many other features with this imbalance. In the future, I would balance these binary features. Likewise, the target class 'Severity' was also imbalanced with severity 1 and 4 having very little data in comparison to severity levels 2 and 3.

Though the machine learning did not perform well, there were features identified through the Random Forest machine learning as having the greatest impact on predicting accident severity (Figure 11). Weather features with the greatest importance were air pressure ('Pressure_bc'), temperature ('Temperature(F)') and humidity ('Humidity(%)'). A deeper dive into the pressure amount, temperature, and humidity percentage would be necessary in order to provide actionable information to Missouri drivers. One road feature bubbled to the top of the list of important features:

'Junction'. Missouri drivers should proceed with caution when approaching and driving through a junction, as the possibility of a high severity accident is greater.

The 'Hour' of the accident occurrence is an important feature when predicting accident severity. Hours of the day in which Missouri drivers should be most cautious are between 7 and 8 AM and 4 and 5 PM (Figure 5). Monday is the most likely day of the week to involve a high severity accident in Missouri and Greene County, Missouri is the most likely county in which a severe accident can occur. Accidents in Missouri are most likely to impact the right side ('Side_R') of the vehicle. Lastly, contrary to my thoughts going into this project, accidents are most likely to occur without the involvement of precipitation ('Precipitation_NA').

6. Conclusion

The purpose of this project was to predict the severity of an accident based on occurrence parameters, road, and weather conditions in the state of Missouri. I explored the various weather, road, and accident occurrence features associated with accidents in Missouri between February 2016 and June 2020. I built both regression models and classification models to predict the severity of accidents in Missouri and through these models, identified the top important factors in determining severity. Though there may be no actionable steps Missourians can take to lower their risk of being involved in a severe accident as a result of this project, I believe an awareness of important factors involved is a good first step.