To Predict the Severity of an Accident

Suchithra Thirunthaiyan

September 06, 2020

1. Introduction

1.1 Background

As a nation on wheels, driving around is part of the life. Imagine You are driving to another city for work or to visit friends. It is rainy and windy and on the way, you come across the terrible traffic jam on the other side of the highway. Long lines of cars are barely moving. As you keep driving police car start appearing from afar shutting down the highway. Its an accident and there's a helicopter transporting the ones involved in the crash to the nearest hospital. They must be in critical condition for all of this to happening.

Now wouldn't be great if there is something in place that could warn you given the weather and road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even you change your travel if you are able to.

1.2 Problem

Data set collected for the past few years, insights on how to drive safely and keep accident free are highlighted here. This should provide a reference guide you may use to drive safely and avoid car accident.. This project aims to predict when and where the severity of an accident happens.

1.3 Interest

Obviously, the drivers would be very interested in predicting the severity of an accident to reach on time to the destination and to reach safely without delay .

2. Data acquisition and cleaning

2.1 Data Source

Most of the accidents features state, city, place, weather condition and POI data found in this Kaggle dataset. Please find the link -here

This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data are collected from February 2016 to June 2020, using two APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data 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 about 3.5 million accident records in this dataset.

This dataset has been collected in real-time, using multiple Traffic APIs. Currently, it contains accident data that are collected from February 2016 to June 2020 for the Contiguous United States.

2.2 Data cleaning

Data used from the source is maintaining in a single table and it has all the necessary attributes/features. It doesn't require any additional sources/datasets. It's well enough to predict the accident severity. Though it has many Nan and missing values with the data sets which need to handle.

The dataset millions of records, due to the limitation of my computation space unable to load all the data. So, instead focusing on to the entire country, I decided to focus in only one state and chose FLORIDA to predict the severity of an accident.

Second, the data came from two sources MapQuest and Bing. It is hard to choose one and we definitely can't use both. I decided to select MapQuest because serious accidents are we really care about and the sparse data of such accidents is the reality we have to confront. Finally, dropped data reported from Bing.

Found some chaos in categorical features. It is necessary to clean them up. And, there were 30 features which had missing values. The counts of missing values in some features are much smaller compared to the total sample. It is convenient to drop rows with missing values in these columns. Indeed, some missing values need to fill with median. Missing accuracies were imputed with 0 and 1.

2.3 Feature selection

After data cleaning, there were 2,08,000 samples and 49 features in the data. Upon examining the meaning of each feature, it was clear that there was some redundancy in the features.

Found some useless features that can be collected only after the accident has already happened and hence cannot be predictors for serious accident prediction, those were dropped from the dataset. After all, 42 features were selected (Table 1).

Table 1. Simple feature selection during data cleaning

Kept Features	Dropped Features	Reason for dropping features
Source, TMC, Severity, Start_Lng, Start_Lat, Distance(mi), Side, City, County, State, Timezone, Temperature(F), Humidity(%), Pressure(in), Visibility(mi), Wind_Direction, Weather_Condition, Amenity, Bump, Crossing, Give_Way, Junction, No_Exit, Railway, Roundabout, Station, Stop, Traffic_Calming, Traffic_Signal, Turning_Loop, Sunrise_Sunset, Hour, Weekday, Time_Duration(min)	ID, TMC, Description, Distance(mi), End_Time, End_Lat, End_Lng	Useless feature since the data collected only after the accident has already happened

3. Exploratory Data Analysis

3.1 Time Features

To avoid car accidents, it is meaningful to see when and where did most accidents happen.

I have chose to calculate the time features to predict when the severity of accidents happens. According to the results, It's quite interesting that the count of other levels accidents is mostly consistent from March to December, whereas the number of level 4 accidents rapidly increased from April to June and remained stable until July then increased again from August (Figure 1). The number of accidents was much less on weekends while the proportion of level 4 accidents was higher. Mostly on Tuesday records the highest accident (Figure 2).

Most accidents happened during the daytime, especially 7 and 8am peak. When it comes to night, accidents were far less but more likely to be serious (Figure 3). Period-of-Day-Accidents were less during the night but were more likely to be serious (Figure 4). Based on this information, you may plan your travel better if possible, or pay extra attention while driving during highly risky time.

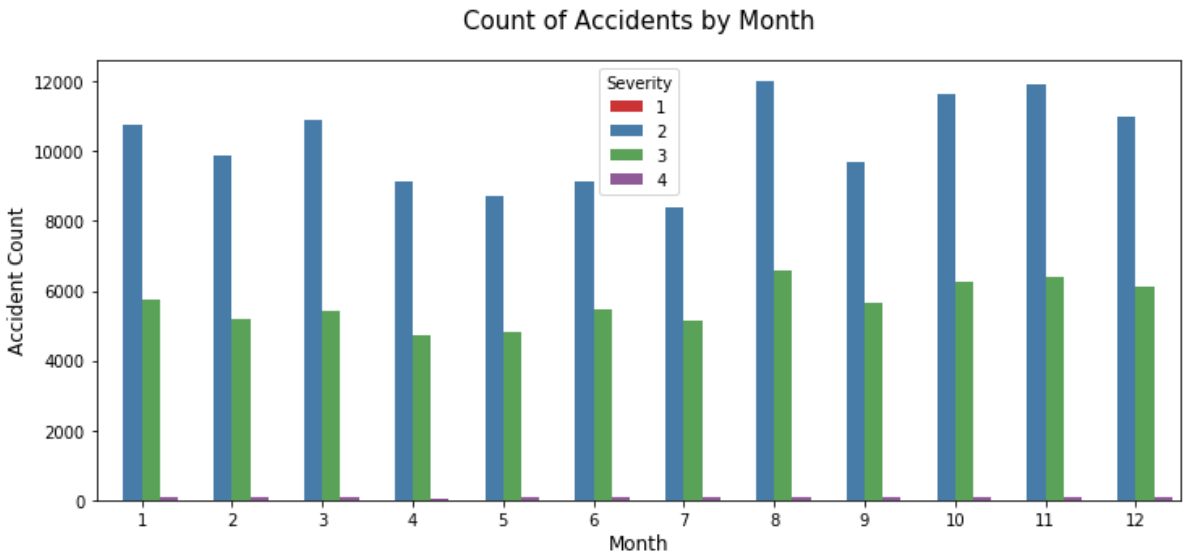

Figure 1. To predict the Count of Accident severity by Month

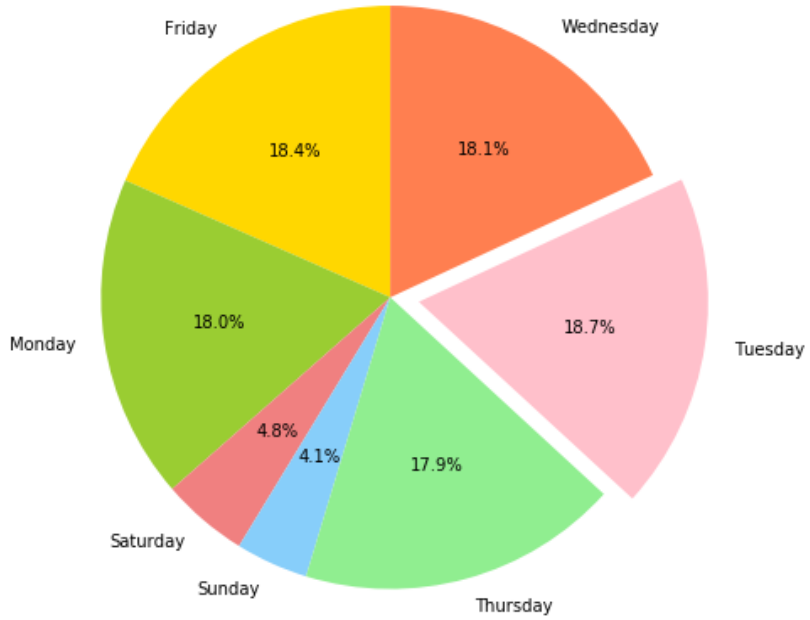

Figure 2. To predict the Count of Accident severity by Weekdays

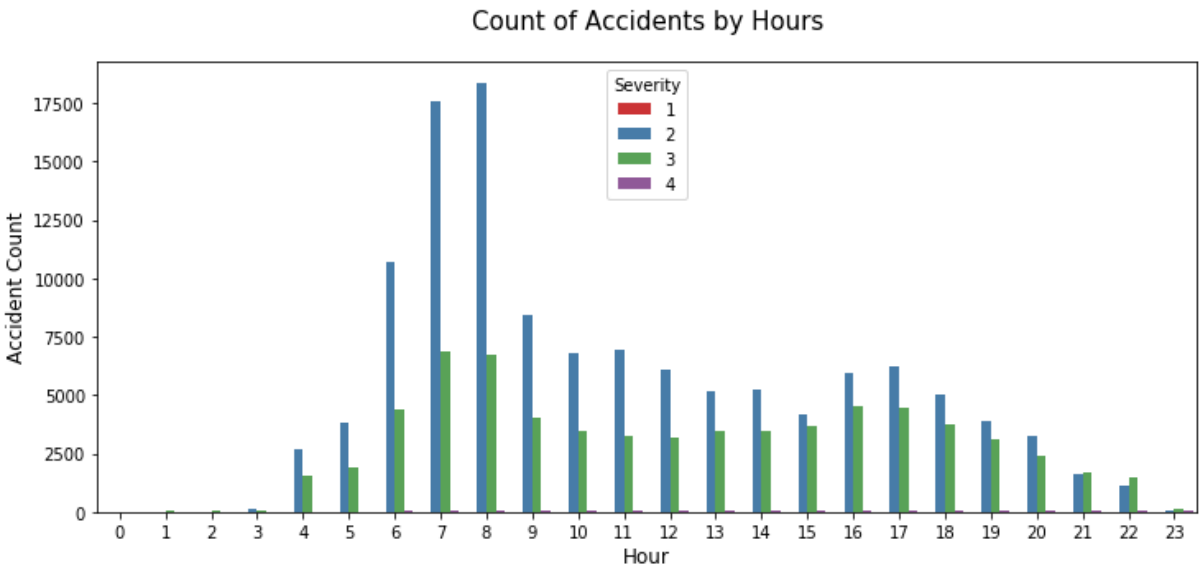

Figure 3. To predict the Count of Accident severity by Hours

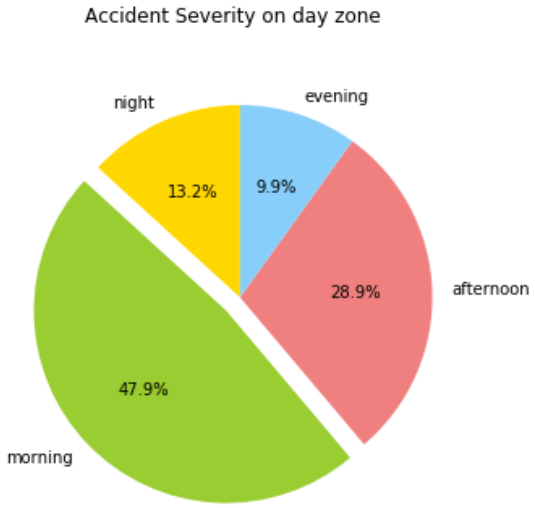

Figure 4. To predict the Count of Accident severity by Period-of-days

3.2 Address Features

To predict the location when did the accident happens I chose some attributes to calculate it. Most of the accidents happens on the right side of the line is much more dangerous than left side (Figure 5). And plotted the location of accidents using latitude and longitude (Figure 6)

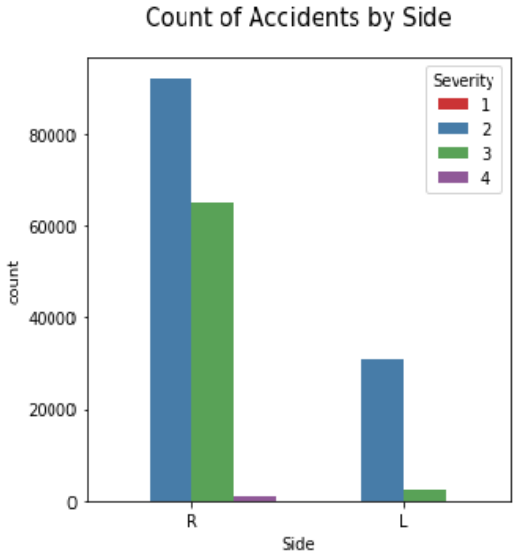

Figure 5. To predict the Count of Accident severity by side

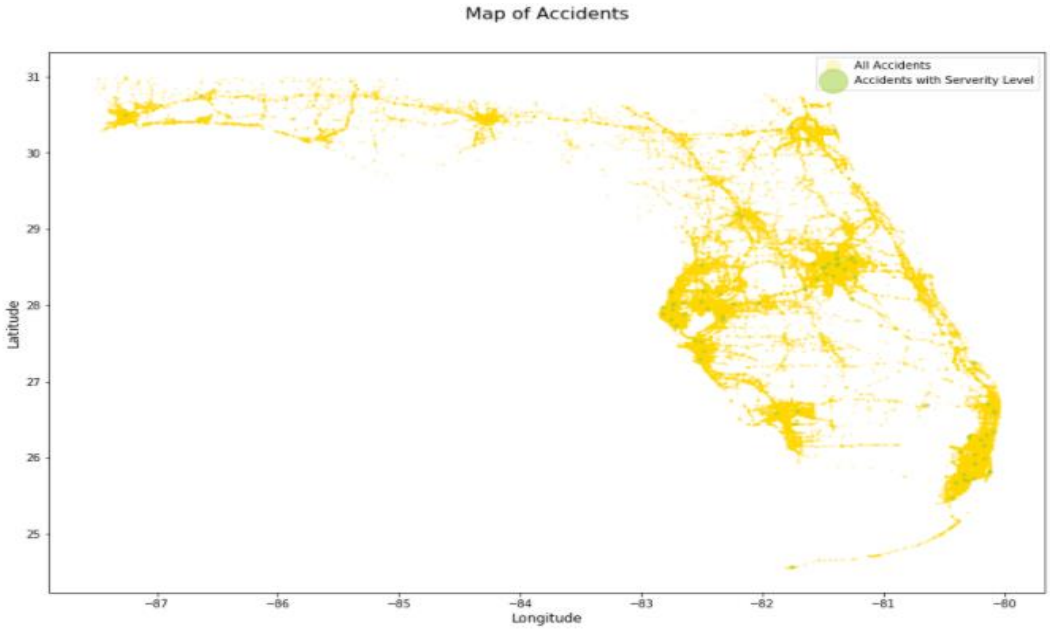

Figure 6. To predict the Count of Accident severity by Latitude & Longitude

3.3 Weather Features

Weather-related vehicle accidents kill more people annually than large-scale weather disasters(source: weather.com). According to Road Weather Management Program, most weather-related crashes happen on wet-pavement and during rainfall.

But, I found the result as maximum number of accident records when the weather condition is in CLEAR state (Figure 7). During wind in the East direction most of the accidents happens (Figure 8)

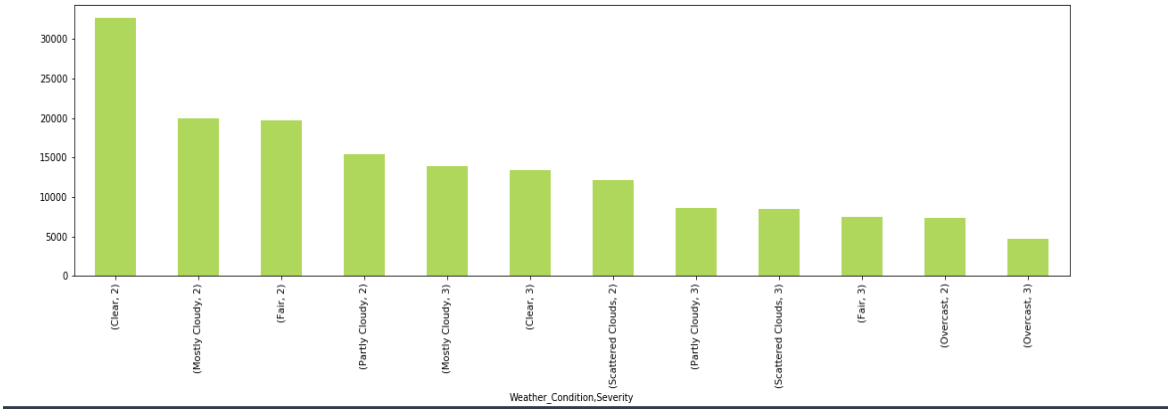

Figure 7. To predict the Count of Accident severity by Weather condition

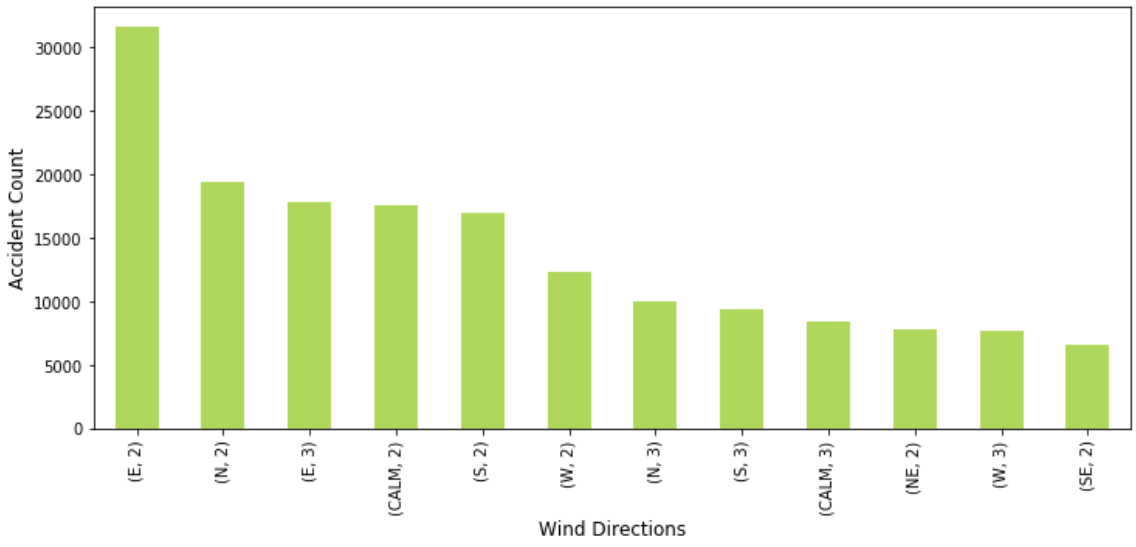

Figure 8. To predict the Count of Accident severity by Wind Direction

3.4 POI Features

Accidents near traffic signal and crossing are much less likely to be serious accidents while little more likely to be serious if they are near the junction (Figure 9). Maybe it is because people usually slow down in front of crossing and traffic signal but junction and severity are highly related to speed. Other POI features are so unbalanced that it is hard to tell their relation with severity from plots. So, I dropped those features.

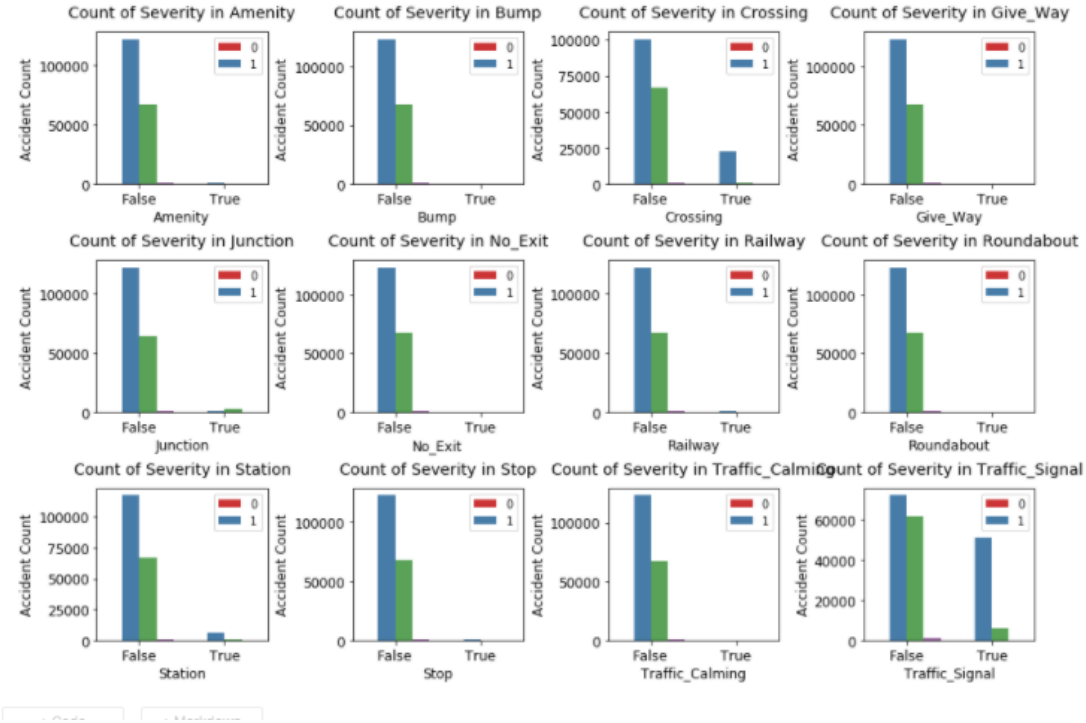

Figure 9. To predict the Count of Accident severity by POI

4. Predictive Modeling

There are two types of models, regression and classification, that can be used to predict player improvement. Regression models can provide additional information on the amount of improvement, while classification models focus on the probabilities a player might improve. The underlying algorithms are similar between regression and classification models, but different audience might prefer one over the other.

For example, to predict the severity of an accident might be more interested in the amount of improvement (regression models) but a general NBA fan might find the results of classification models more interpretable Therefore, in this study, I carried out classification modeling.

4.1 Classification Models

4.1.1 Applying standard algorithms and their problems

I applied linear models Logistic regression, Decision tree, random forest models to the dataset. The results of logistic regression is very poor. Decision tree and Random forest accuracy of result is more similar. Even with the best parameter setting, logistic regression yielded very poor results for both training data and test data.

4.1.2 Solution to the problems

I have performed the grid search over choices of min_samples_split and max_features for decision tree and for random forest algorithm n_estimators and max_depth to get the more accuracy for train and test data. As a result, the top 15 important features of random forest model are almost as same as decision tree model (Figure 10 & 11)

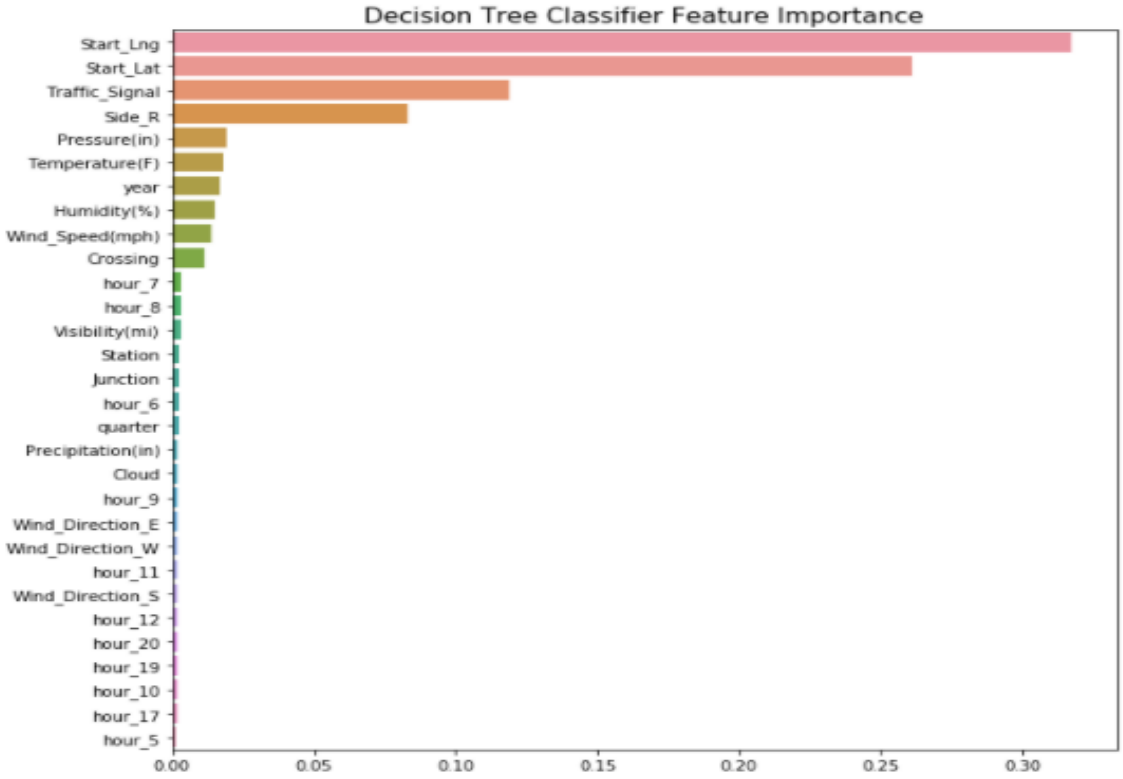

Figure 10. Decision Tree classifier importance features

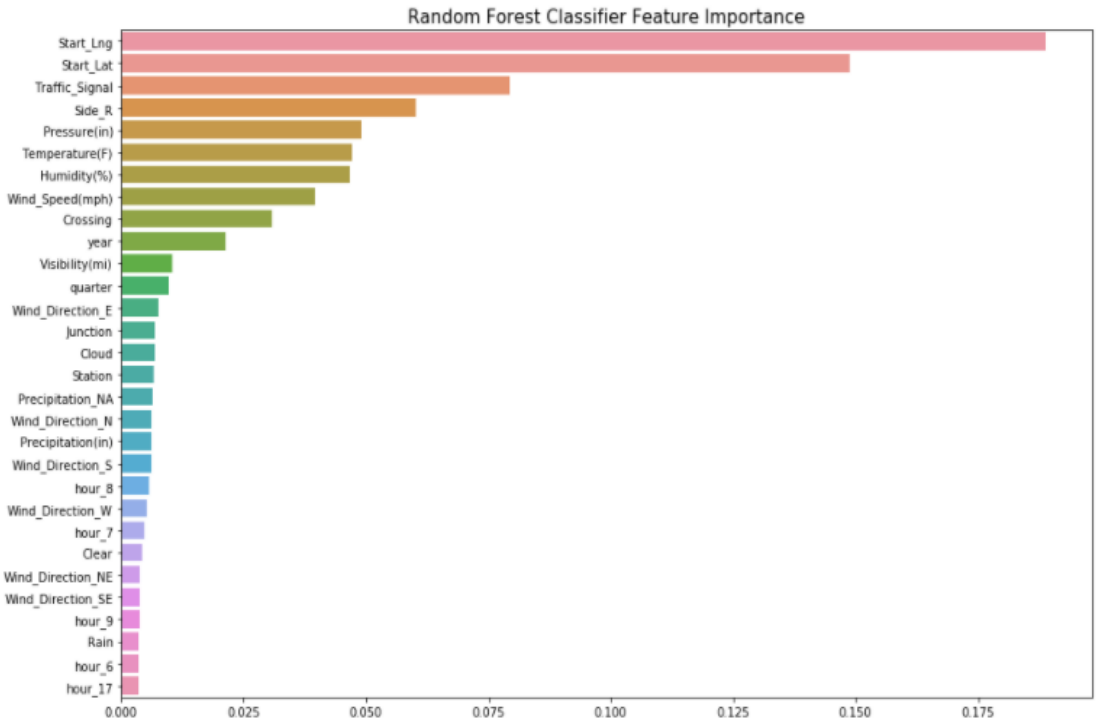

Figure 11. Random Forest classifier importance features

4.1.3 Performances of different models

Using the new approach of different sample weights, I built Logistic regression, Decision Tree and Random Forest algorithms. For each algorithm, I have calculated the F-1 score and Jaccard to predict the accuracy.

Table 2. Performance/Accuracy of classification algorithms

Algorithm	**Train Data**	**Test Data**	**Jaccard**	**F1-Score**
Logistic Regression	73.580524	73.664023	0.736640	0.734916
Decision Tree	98.170859	88.239745	0.882397	0.881742
Random Forest	99.947161	85.619977	0.882397	0.881742

5. Conclusion

In this study, I predicted the severity of an accident to drive safely and avoid car accident. I found patterns and study indicates that when, where and under what weather conditions did most accidents occurred. I built a various machine learning algorithms to predict severity of each accident and it predicted quite accurately. These models can be very useful in helping drivers for safe drive, with this prediction they can avoid distracted rush hours, signals.

6. Future Directions

I was able to predict the accuracy score not 100%. To incorporate this model in a real-time accident risk prediction model or develop a new real-time severe accident risk prediction on grid cells. Needs to find detailed relations between some key factors and accident severity can be further studied by predicting speed of vehicle, people who involved. Although many accidents are just merely out of luck. For instance, I was once waiting for the red signal to turn green and the car in front of me started to back up and bumped into mine. Still it is desirable to understand what are possible to control and reduce in terms of chance of accident or its severity.