PREDICTING CRASH SEVERITY

PROBLEM STATEMENT

Car crashes and road accidents could be considered an old topic. Yet, with the progress of the technology involved and the capabilities of these sophisticated machines, it is ever more important to have tools and means available to mitigate their occurrences, as well as their implications and consequences for the people involved. Thanks to the advancement of technological and analytical tools in the last two decades we are now able to better understand how crashes happen. This enables the transport, security and emergency agencies all around the world to have different (predictive) models for quickly analyzing crashes when they happen and dispatch an appropriate response swiftly. Many attempts have been taken by many professionals, scholars and government agencies to provide produce these models; each with different goals, ways of measuring success and precision of their results.

Predicting the severity of a car crash is no easy task. And even when possible, precision levels will vary significantly depending on the data available and how well the system or model has been defined. However, if the dataset's features are clearly defined and if there's a thorough description of how this data is collected we have much better chances of arriving at a usable model. In the dataset I'll use data associated with car crashes come in a hybrid mode; meaning we have both categorical and numerical features. This allows us to treat the problem from a mathematical approach and to use performance metrics such as R2 , ROC curves, precision, accuracy and F Scores.

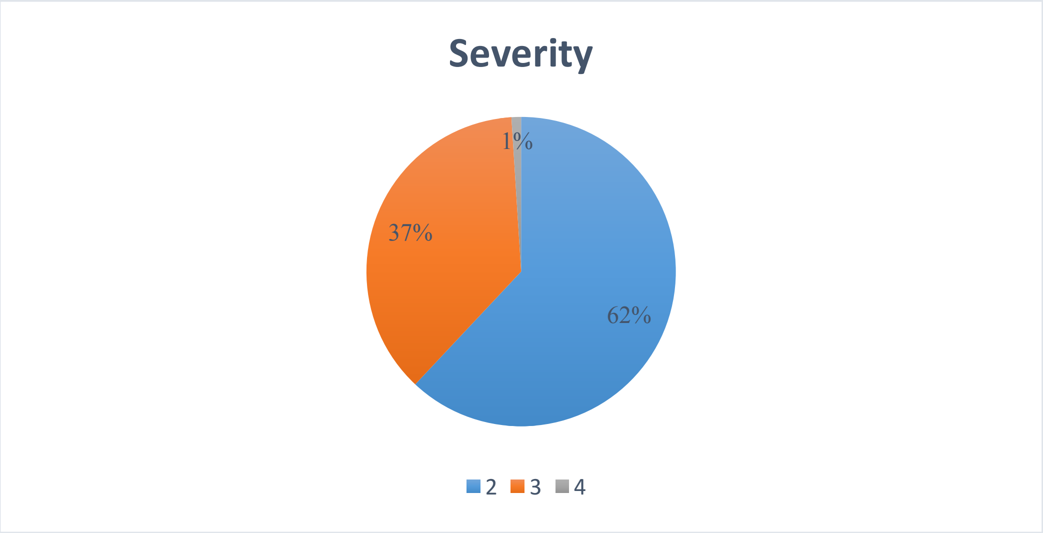
DATA

This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data are collect 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.

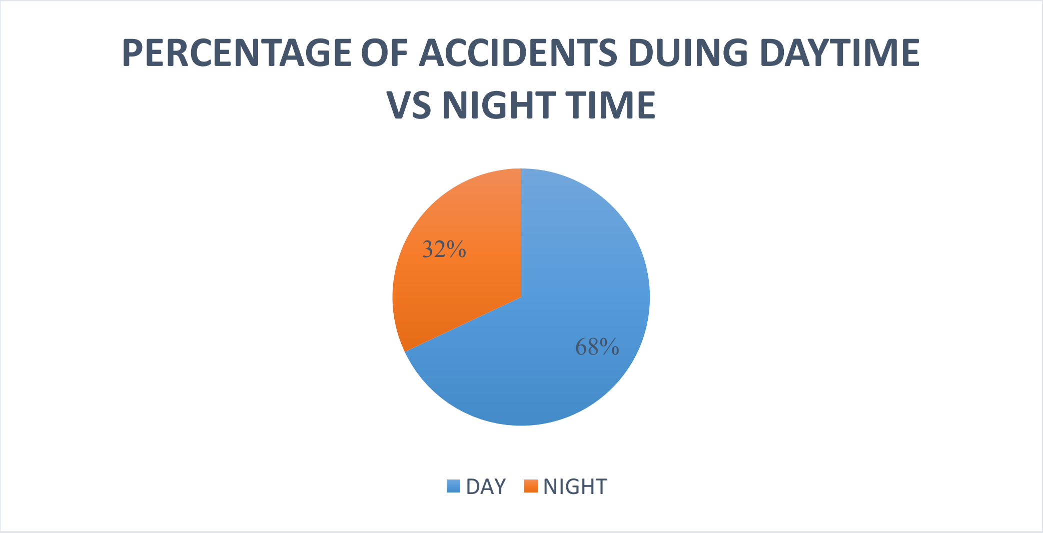
For the better understanding of the model, I'm focusing on the State of California. I will only select a few features I believe are more relavant to severity. Categorical data will be treated with Pandas get_dummies method. Rows with missing values will be dropped.

EXPLORATORY ANALYSIS

The good thing is that, although it varies state by state, most accidents are light accidents with severity level 2, with the most severe one (level 4) the least



According to the results, most accidents happened during the daytime. There are more accidents on weekdays than weekends. On weekdays, rush hours are most dangerous times while on weekends, early afternoon is more dangerous than other time. Based on this information, you may plan your travel better if possible, or pay extra attention while driving during highly risky time.



SEVERITY PREDICTION USING MACHINE LEARNING

Feature Engineering

Although there is more information available, it is desirable to select a few relevant features which have higher impact on accident and its severity. This can reduce the demand on computation and improve the accuracy of predictions. Mainly the time, location and weather conditions were selected.

```
feature_lst=['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)']
```

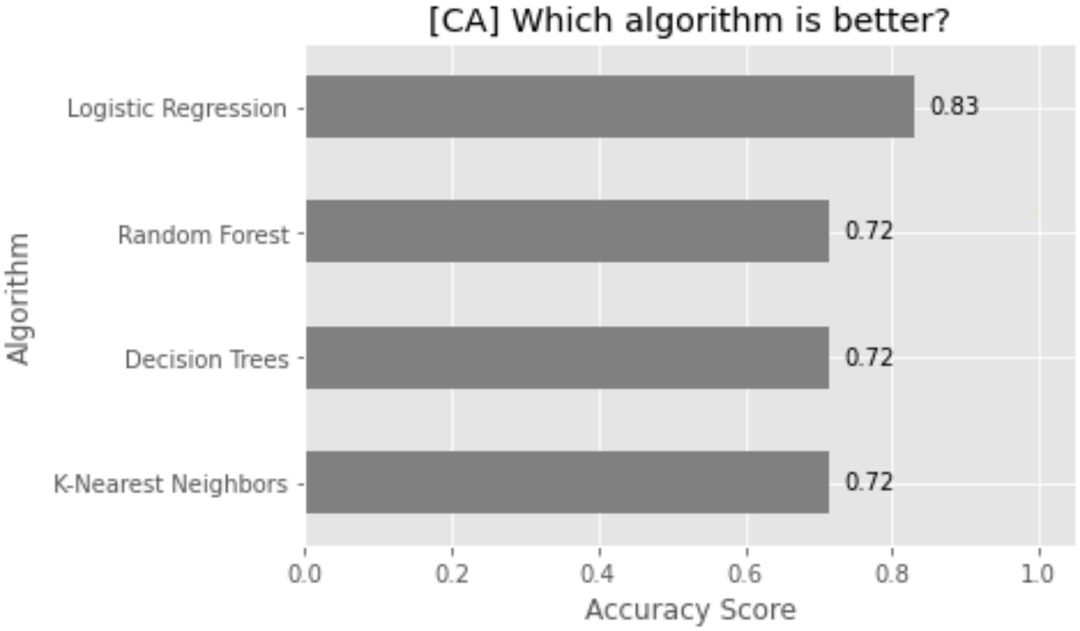
Due to the limit of personal computer, it is impossible to use the whole dataset. Predictions were made for individual state or county.

Data Preprocessing

Besides no missing value is allowed, most machine learning algorithms work only with numerical data. For that, records with missing values were dropped from the calculation. Some outliers, especially with extremely short or long time to clear the accident, were processed and replaced with median values. As it doesn’t make sense to take 0 minute or many years to clear an accident.

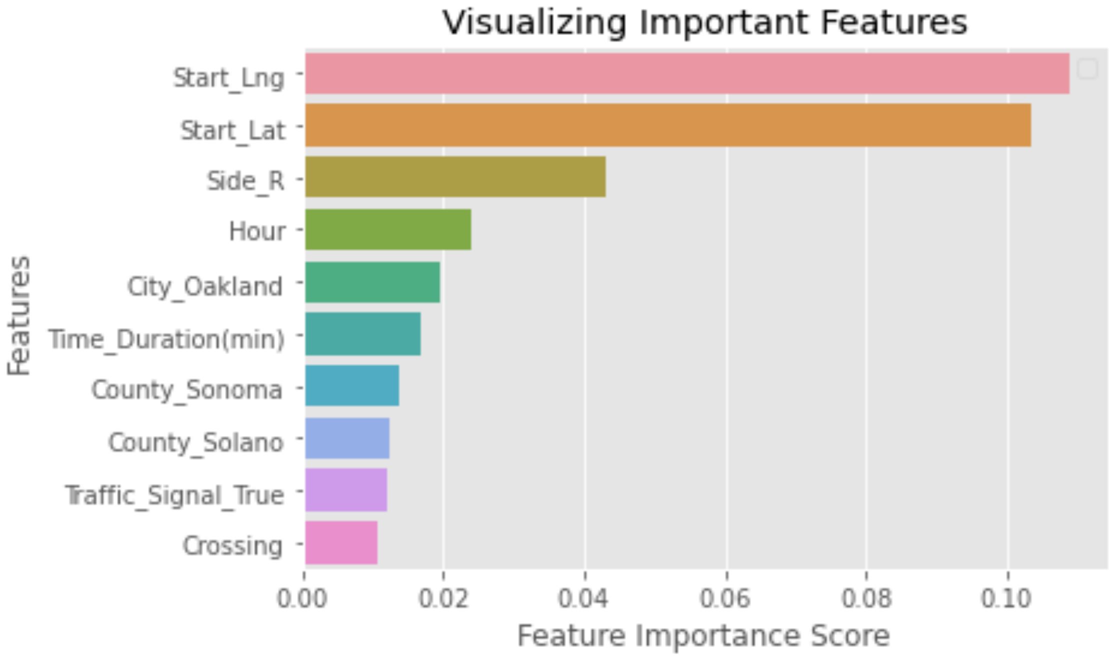
Predicting Accident Severity with various Machine Learning Algorithms

In this study, four classification machine learning algorithms were evaluated. These are Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees and Random Forest. As you can see, Random Forest algorithm is the winner and KNN is the last on the list in terms of accuracy.



Most important features

Shown below are the top 10 features affecting the accuracy of predicting accident severity for the data from the State of California. Prediction accuracy can be further improved by removing less important or irrelevant features.


Visualizing Important Features
Features
Start_Lng
Start_Lat
Side_R
Hour
City_Oakland
Time_Duration(min)
County_Sonoma
County_Solano
Traffic_Signal_True
Crossing
Feature Importance Score