1. Introduction

1.1 Background/Problem

Automobile accidents are the number one leading cause of deaths in the United States with over 35,000 people dying from accidents each year and medical/productivity costs exceeding 75 billion dollars per year (CDC). Approximately 22 percent of all accidents are weather-related (Federal Highway Administration). This project aims to address this problem by predicting what factors determine the severity of accidents to provide avenues for possible solutions.

1.2 Interest

This project is of particular importance not just to citizens (drivers) but to emergency services, the Federal Highway Administration, as well as new age driving related companies such as Uber, Lyft, DoorDash, etc. to minimize medical and productivity costs resulting from automobile accidents.

2. Data

2.1 Description

To solve this problem, we are using a set of traffic data from Seattle Washington, collected between the years of 2004-2020. This is a labelled dataset with "Accident Severity" as the outcome and contains 37 attributes including the weather conditions and timing (day of the week, time of day) at which the accident occurred. There is a total of 194673 unique reports. Accident Severity is encoded as 1 = Property Damage only, 2 = Injury Collision

2.2 Feature Selection

The attributes can be categorized into two main groups: predictor variables (i.e weather conditions, driver's state of mind), and resultant collision descriptors (type of collision, number of people/vehicles involved). To begin with, a selection of possible useful features in predicting the severity outcome was decided on. These included the following: INATTENTIONIND (collision due to inattention), UNDERINFL (collision due to influence of drugs/alcohol), WEATHER, ROADCOND (road condition), LIGHTCOND (light condition), PEDROWNOTGRNT (If pedestrian was given right of way), and SPEEDING.

From these possible features, INATTENTIONIND, PEDROWNOTGRNT, and SPEDING all had very few entries (29805, 4667, and 9333 respectively). We therefore focused any further analyses on the following final features:
UNDERINFL
WEATHER
ROADCOND
LIGHTCOND

2.3 Data Cleaning

The first task of data cleaning consisted of encoding the chosen features to numerical values.

For UNDERINFL : “N” was replaced with 0, “Y” was replaced with 1.
For WEATHER: 0 = Clear, 1 = mild weather (overcast/cloudy), 2 = severe weather (rain, snow, hail, wind, etc.)
For ROADCOND: 0 = Dry, 1 = Wet/Sandy/Oily, 2 = Icy, snowy, standing water
For LIGHTCOND: 0 = Light, 1 = Medium, 2 = Dark

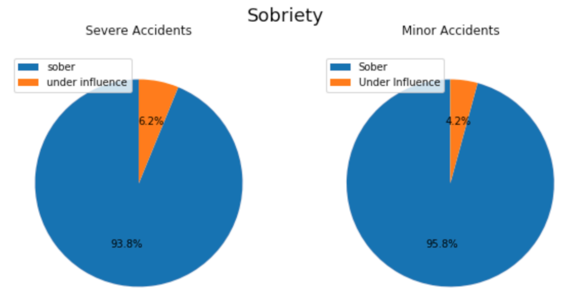
Any “Unknown” or “Other” values were encoded to NaN values.

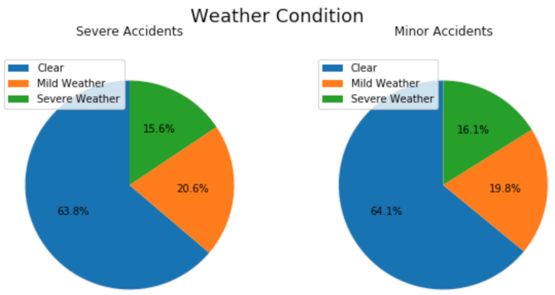
Following this, approximately 13% of the incidents are missing a value for at least one of these features. Since this is a fairly significant portion of the data, we instead interpolate values to fill the NaN values.

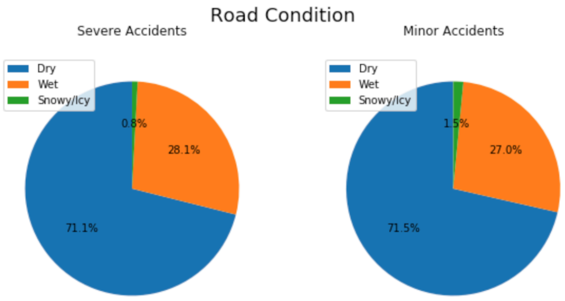
3. Methodology

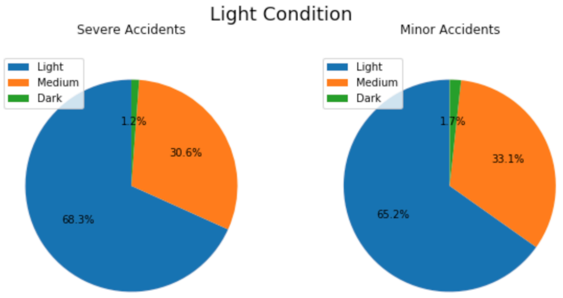
3.1 Exploratory Data Analysis

For initial data exploration, the data was split into minor and severe accidents. Then for each feature a pie chart was generated to compare how the proportion of different weather, road, light, and sobriety conditions changes in minor vs. severe accident situations.









From this exploration we can see that severe accidents contain a higher ratio of drivers under the influence and worse road and weather conditions. Interestingly, severe accidents seem more likely to occur during light conditions.

3.2 Model Selection

The machine learning algorithms we chose to use include K nearest neighbor (KNN), Decision/Classificiation Tree, and Logistic Regression. Support Vector Machine (SVM) does not work well for large datasets

KNN: clusters data into groups by matching a point with the closest neighbor in some multi-dimensional space. It can be used for categorical data

Decision/Classification Tree: go from features/variables as inputs to come to a conclusion/target. It can be used for categorical data.

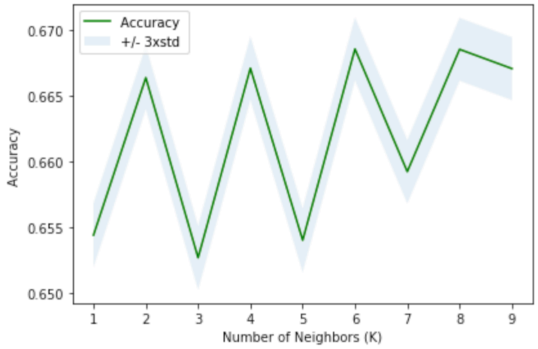
Logistic Regression: uses logistic function to estimate a binary dependent variable.

4. Results

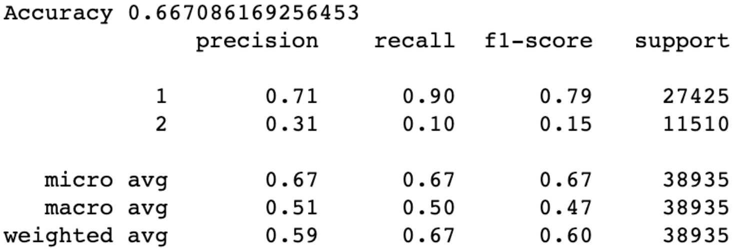
The data was split into 80% train, 20% test size with a random state value of 4.

4.1 KNN

The accuracy for different values of k were plotted (shown below). The best k value was 6 which had an accuracy of 0.67. The classification report is also shown below



KNN Classification Report



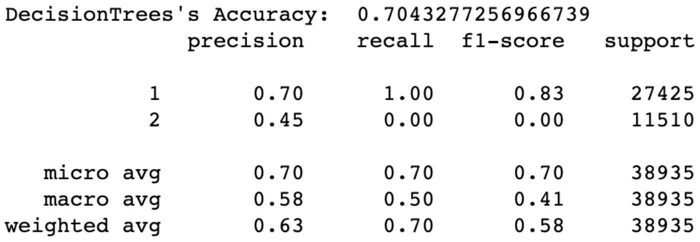
```
Accuracy 0.667086169256453
              precision    recall  f1-score   support

           1       0.71      0.90      0.79     27425
           2       0.31      0.10      0.15     11510

   micro avg       0.67      0.67      0.67     38935
   macro avg       0.51      0.50      0.47     38935
weighted avg       0.59      0.67      0.60     38935
```

4.2 Decision/Classification Tree

The classification tree method produced an accuracy of .70. The classification report for this method is below.



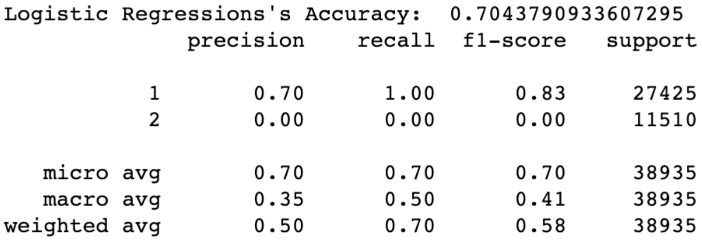
```
DecisionTrees's Accuracy:  0.7043277256966739
              precision    recall  f1-score   support

           1       0.70      1.00      0.83     27425
           2       0.45      0.00      0.00     11510

   micro avg       0.70      0.70      0.70     38935
   macro avg       0.58      0.50      0.41     38935
weighted avg       0.63      0.70      0.58     38935
```

4.3 Logistic Regression

The logistic regression method produced an accuracy of .70. The classification report for this method is below.



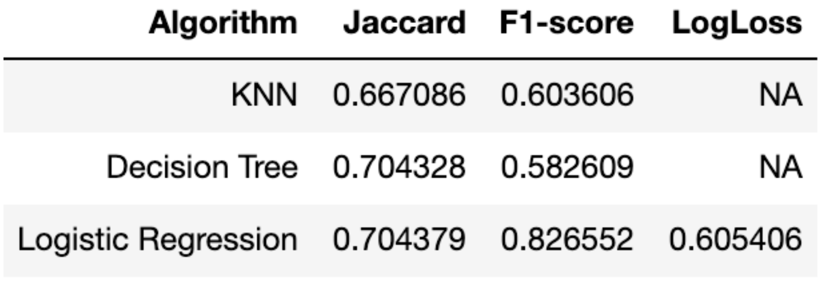
```
Logistic Regressions's Accuracy:  0.7043790933607295
              precision    recall  f1-score   support

           1       0.70      1.00      0.83     27425
           2       0.00      0.00      0.00     11510

   micro avg       0.70      0.70      0.70     38935
   macro avg       0.35      0.50      0.41     38935
weighted avg       0.50      0.70      0.58     38935
```

4.4 Model Evaluation

A summary table of each of the models and their jaccard index, f1-score, and logloss score (for logistic regression only) is shown below.



Algorithm	Jaccard	F1-score	LogLoss
KNN	0.667086	0.603606	NA
Decision Tree	0.704328	0.582609	NA
Logistic Regression	0.704379	0.826552	0.605406

5. Discussion

5.2 Jaccard Index

The Jaccard Index is the ratio of correct predicted values to wrongly classified values. A higher jaccard score means a better model (more accurate). The logistic regression and decision tree model both have the highest jaccard index as compared to KNN.

5.3 F1-score

The F1-score is a measure of the balance between the precision and recall of a model. The higher (closer to 1) value of the F1-score, the better the model. Using this metric, the logistic regression model is best.

We can conclude from this analysis that the KNN method is the least accurate of the models. The best model is the logistic regression model, although the decision tree model is close behind.

The models would have performed better had the dataset been more balanced (an equal number of low and high severity accidents). Also, it would have been useful to have a complete set of data for the features inattention and speeding, as these likely have a major impact on severity of accidents.

Based on the data suggesting that severe accidents contain a higher ratio of drivers under the influence and worse road and weather conditions, I would recommend addressing these issue first.

Under the influence: I would recommend further delving into the data to see what days of the week and time of the day intoxication accidents occurred. Then during these day/timepoints increase the number of patrol cars out on the roads.

Road and weather conditions: I would recommend developing an app, or potentially an add on to google/apple maps, that takes into consideration the current weather/road conditions to make suggestions/recommendations about when and when not to go out driving (using the logistic regression model that was developed).

6. Conclusion

To conclude, I would recommend using the logistic regression model for future predictions. These predictions should include accident severity predictions based on the current weather and road conditions. Additionally, a new model should be generated using date and time data to come up with a prediction of likelihood of an intoxication related accident that can be used to determine times to increase highway/city patrol cars.