Predicting Accident Severity

Background/Problem

- 35,000 people die from automobile accidents each year (CDC)
- 75 billion dollars in medical/productivity costs each year (CDC)
- Approximately 22 percent of all accidents are weather-related (Federal Highway Administration)
- Of interest to: emergency services, the Federal Highway Administration, driving related companies such as Uber, Lyft, DoorDash, etc.

MAIN OBJECTIVE: address this problem by predicting what factors determine the severity of accidents to provide avenues for possible solutions.

Data - Description

- Seattle, Washington
- collected between the years of 2004-2020
- contains 37 attributes
- total of 194673 unique reports
- Accident Severity is encoded as 1 = Property Damage only, 2 = Injury Collision

Data – Feature Selection

- *predictor variables (i.e weather conditions, driver's state of mind)
- resultant collision descriptors (type of collision, number of people/vehicles involved)
- 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)
- SPEEDING

INATTENTIONIND, PEDROWNOTGRNT, and SPEDING all had very few entries (29805, 4667, and 9333 respectively).

Data – Final Features

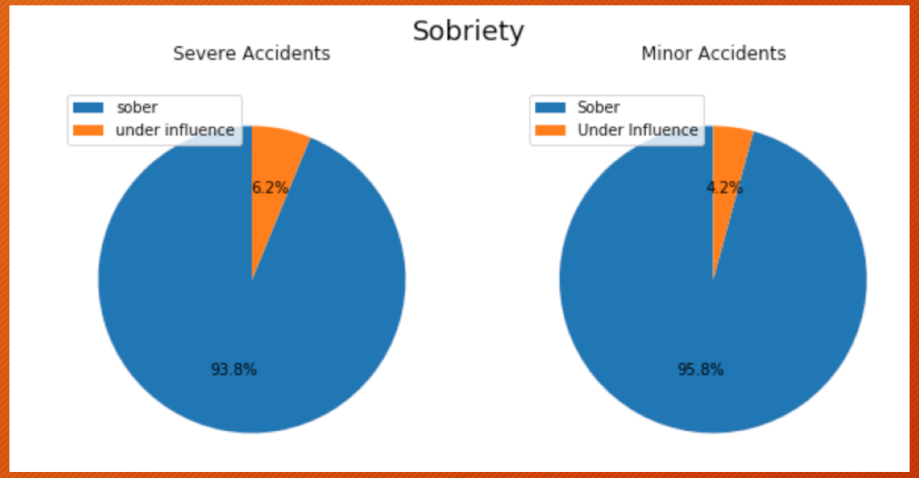
- UNDERINFL
- WEATHER
- ROADCOND
- LIGHTCOND

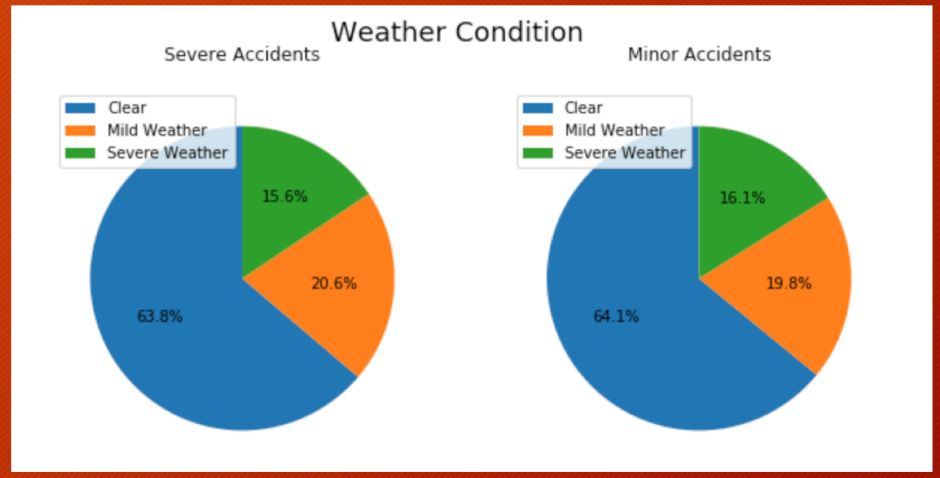
Data - Cleaning

- Encoding:
  - 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.
- Interpolate values to fill the NaN values.

Methodology – Data Exploration

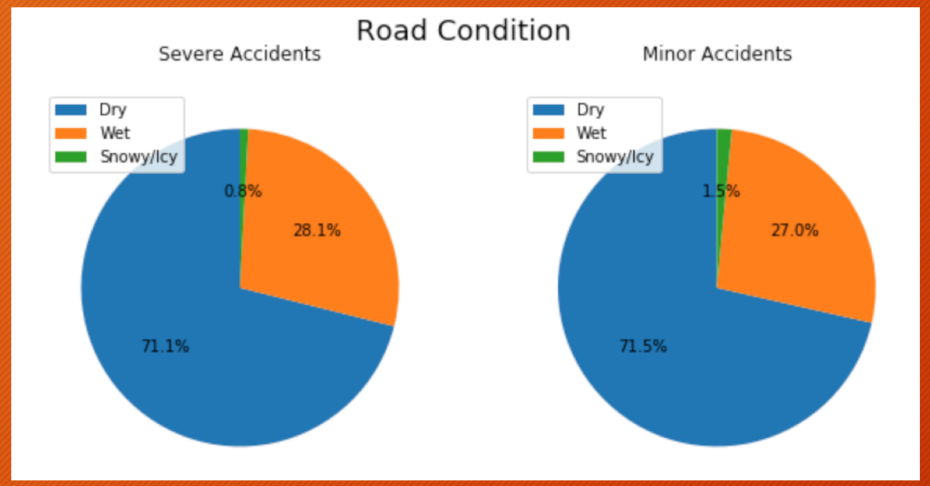
- Data split in minor vs. severe accidents

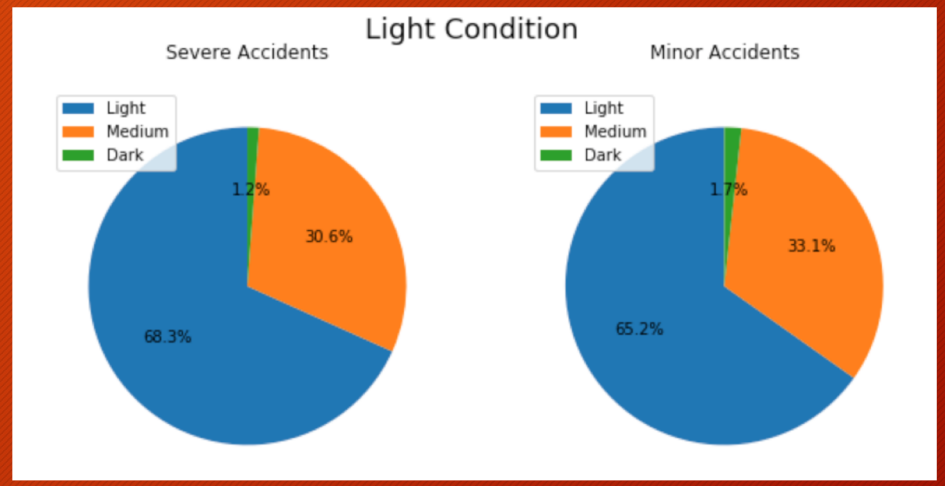




Methodology – Data Exploration

- Data split in minor vs. severe accidents





Methodology – Data Exploration

- 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.

Methodology – Model Selection

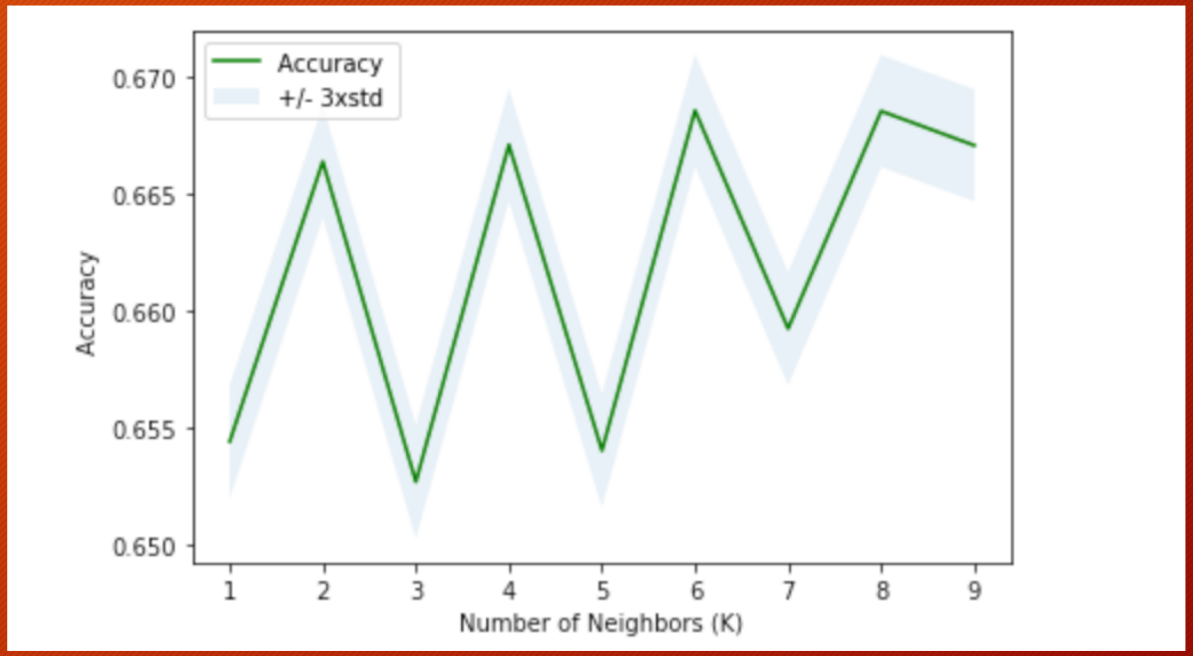
- KNN: clusters data into groups by matching a point with the closest neighbor in some multi-dimensional space. It can be used for categorical data
- Decision/Classification Tree: go from features/variables as inputs to come to a conclusion/target. It can be used for categorical data.
- Logistic Regression: uses logistic function to estimate a binary dependent variable.
- ** Did not choose SVM since doesn't work well for large datasets

Results - KNN

- The data was split into 80% train, 20% test size with a random state value of 4.
- Best k value = 6
- Accuracy = 0.67

Accuracy 0.667086169256453

	precision	recall	f1-score
1	0.71	0.90	0.79
2	0.31	0.10	0.15
micro avg	0.67	0.67	0.67
macro avg	0.51	0.50	0.47
weighted avg	0.59	0.67	0.60



Results – Classification Tree

- Accuracy = 0.70

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

Results – Logistic Regression

- Accuracy = 0.70

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

Results – Model Evaluation Summary

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

Discussion – 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.

Discussion – 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.

Discussion – Best Model

- 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 if
  - dataset been more balanced (an equal number of low and high severity accidents).
  - Complete set of data for the features inattention and speeding, as these likely have a major impact on severity of accidents.

Discussion/Conclusion - Recommendations

- **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).