PREDICTING CAR ACCIDENT SEVERITY IN SEATTLE, WASHINGTON

IBM APPLIED DATA SCIENCE CAPSTONE PROJECT

SARAH KRISTOLAITIS

THE PROBLEM

- Car accidents are a common occurrence in the United States, and can cause significant distress, disability, decreased quality of life, financial hardship, and even death
- According to the National Safety Council, 38,800 people died in car crashed in the US in 2019, and 4.4 million people had to seek medical attention for car accident related injuries

INTEREST IN THE PROBLEM

Government:

- Reducing car accidents would reduce healthcare costs and negative impacts on worker productivity, GDP, etc.
- The data from this project could drive more targeted improvement to infrastructure, policing strategies, or public education efforts

Private Businesses:

- May be able to use the data to create or improve products, such as GPS systems, that would assist their customers with reducing their accident risk

DATA ACQUISITION AND CLEANING

- Data was obtained from the Applied Data Science Capstone project site, with additional details on the data taken from the Seattle GeoData website
- The selected attributes were those deemed to be most likely to provide early warnings of accident risk (i.e. would be able to be implemented into a GPS type of system in order to provide useful warnings to drivers)
- The target variable was the accident severity, rated at 1 (property damage only) or 2 (injury to a person occurred)

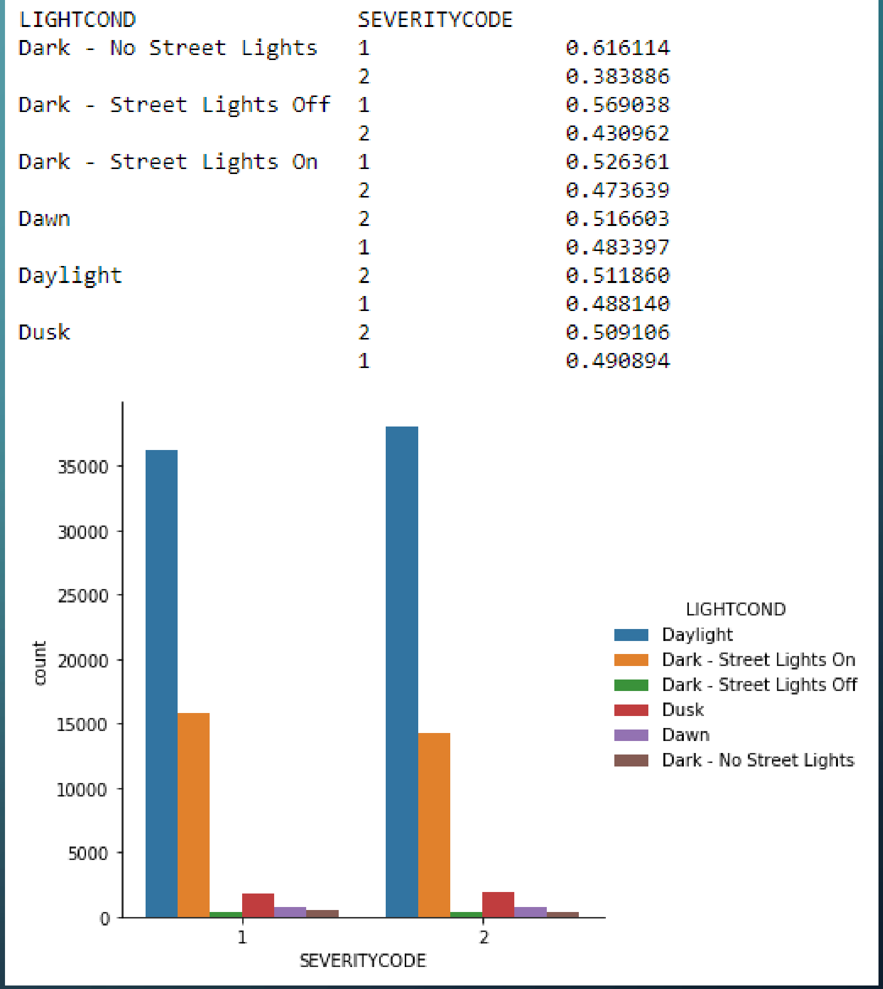
DATA ACQUISITION AND CLEANING

- The 7 chosen attributes for analysis were Road Conditions, Light Conditions, Inattentiveness, Being Under an Influence, Speeding, Weather, and Address Type
- Data was balanced so that type 1 accidents did not overpower type 2
- Null values for Inattentiveness, Speeding, and Being Under an Influence were determined to be “N”
- Null values for all other fields resulted in the observation being eliminated
- There were 111,308 observations remaining for analysis and model development

RELATIONSHIP BETWEEN ACCIDENT SEVERITY AND LIGHT CONDITIONS

Accidents were most likely to result in personal injury:

- At dawn
- At dusk
- During daylight hours

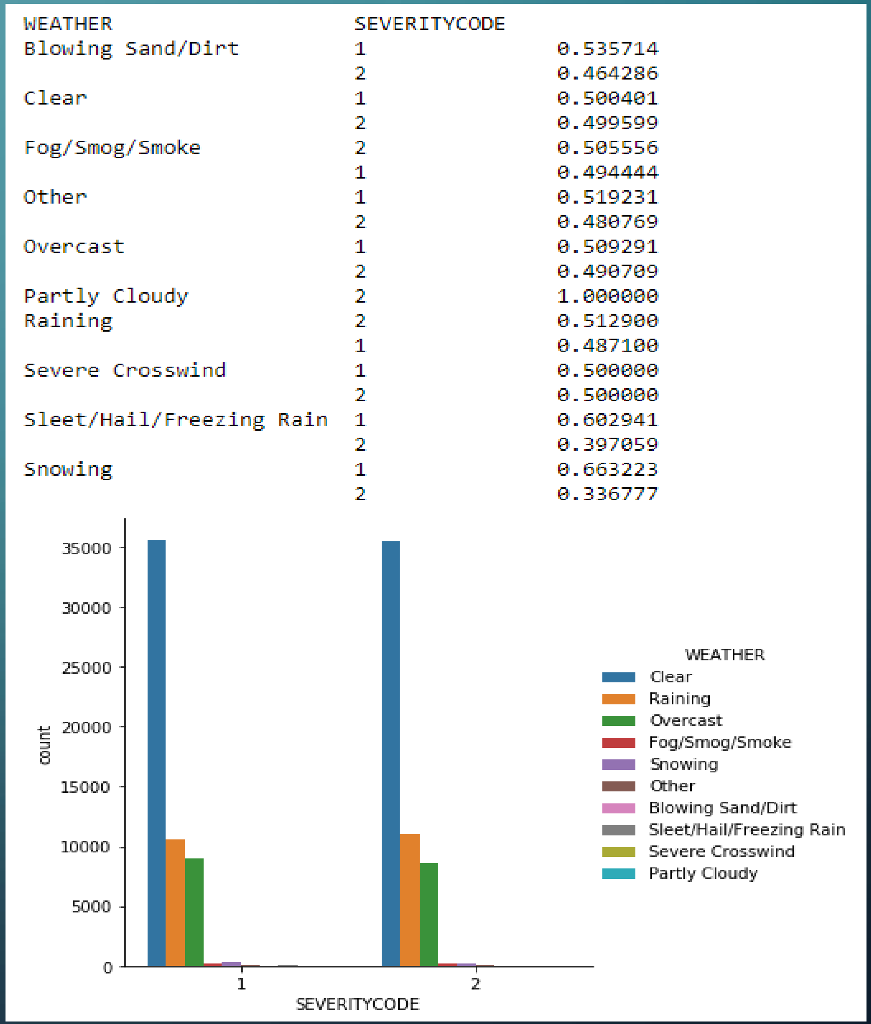


```
LIGHTCOND                 SEVERITYCODE
Dark - No Street Lights   1               0.616114
                          2               0.383886
Dark - Street Lights Off  1               0.569038
                          2               0.430962
Dark - Street Lights On   1               0.526361
                          2               0.473639
Dawn                      2               0.516603
                          1               0.483397
Daylight                  2               0.511860
                          1               0.488140
Dusk                      2               0.509106
                          1               0.490894
```

RELATIONSHIP BETWEEN ACCIDENT SEVERITY AND WEATHER

Accidents were most likely to result in personal injury:

- With Fog/Smog/Smoke
- On Partly Cloudy days
- When Raining

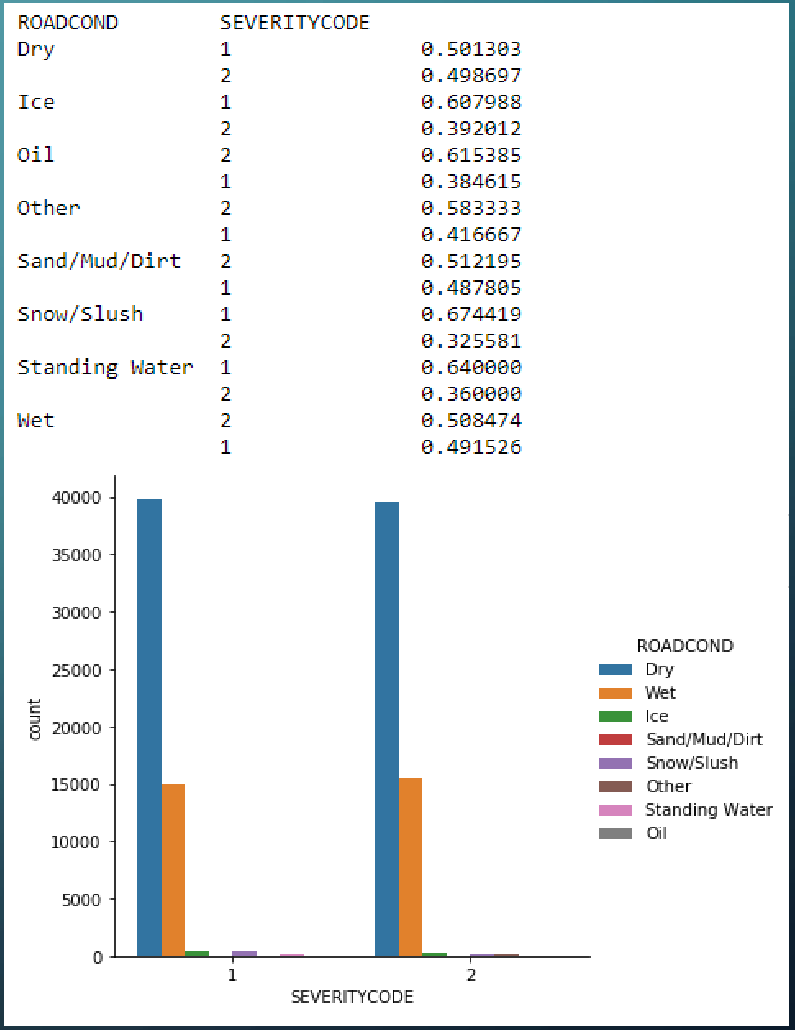


```
WEATHER                   SEVERITYCODE
Blowing Sand/Dirt         1               0.535714
                          2               0.464286
Clear                     1               0.500401
                          2               0.499599
Fog/Smog/Smoke            2               0.505556
                          1               0.494444
Other                     1               0.519231
                          2               0.480769
Overcast                  1               0.509291
                          2               0.490709
Partly Cloudy             2               1.000000
Raining                   2               0.512900
                          1               0.487100
Severe Crosswind          1               0.500000
                          2               0.500000
Sleet/Hail/Freezing Rain  1               0.602941
                          2               0.397059
Snowing                   1               0.663223
                          2               0.336777
```

RELATIONSHIP BETWEEN ACCIDENT SEVERITY AND ROAD CONDITIONS

Accidents were most likely to result in personal injury:

- In Oil
- In Sand/Mud/Dirt
- On Wet Roads
- In “other” conditions

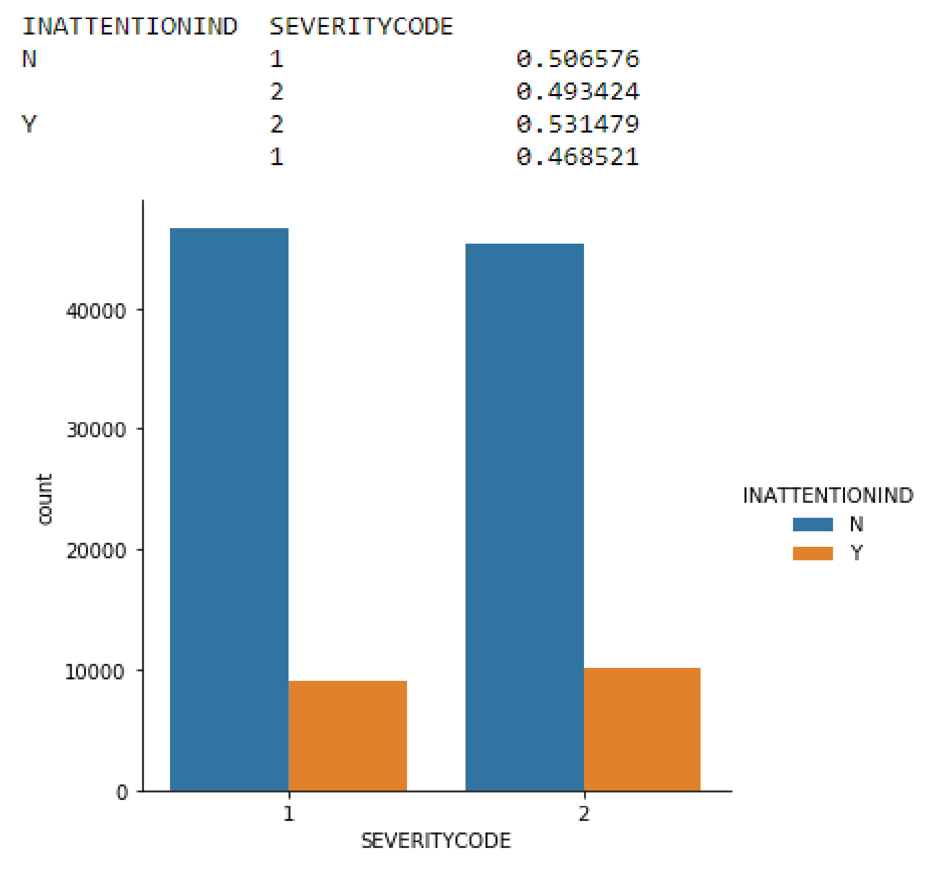


```
ROADCOND        SEVERITYCODE
Dry             1               0.501303
                2               0.498697
Ice             1               0.607988
                2               0.392012
Oil             2               0.615385
                1               0.384615
Other           2               0.583333
                1               0.416667
Sand/Mud/Dirt   2               0.512195
                1               0.487805
Snow/Slush      1               0.674419
                2               0.325581
Standing Water  1               0.640000
                2               0.360000
Wet             2               0.508474
                1               0.491526
```

RELATIONSHIP BETWEEN ACCIDENT SEVERITY AND DRIVER INATTENTION

Accidents were most likely to result in personal injury:

- When a driver is deemed to have been distracted

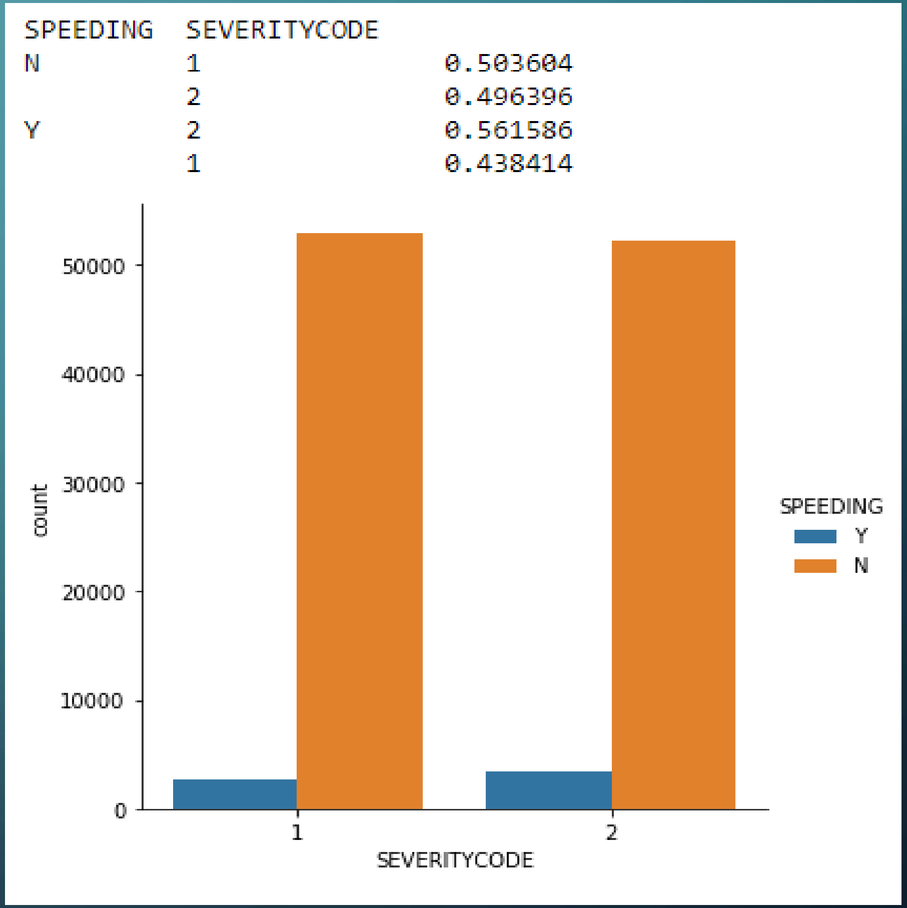


```
INATTENTIONIND  SEVERITYCODE
N               1               0.506576
                2               0.493424
Y               2               0.531479
                1               0.468521
```

RELATIONSHIP BETWEEN ACCIDENT SEVERITY AND SPEEDING

Accidents were most likely to result in personal injury:

- When a driver is speeding

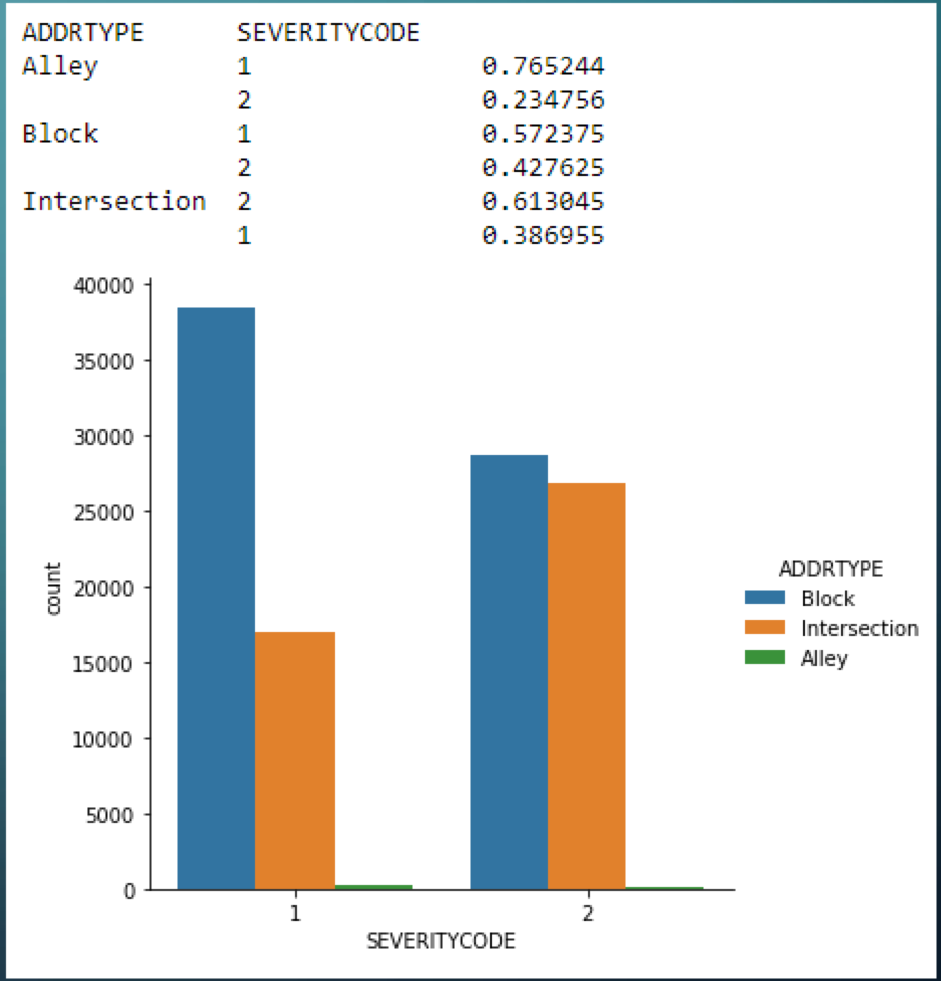


```
SPEEDING  SEVERITYCODE
N         1               0.503604
          2               0.496396
Y         2               0.561586
          1               0.438414
```

RELATIONSHIP BETWEEN ACCIDENT SEVERITY AND ADDRESS TYPE

Accidents were most likely to result in personal injury:

- In an intersection

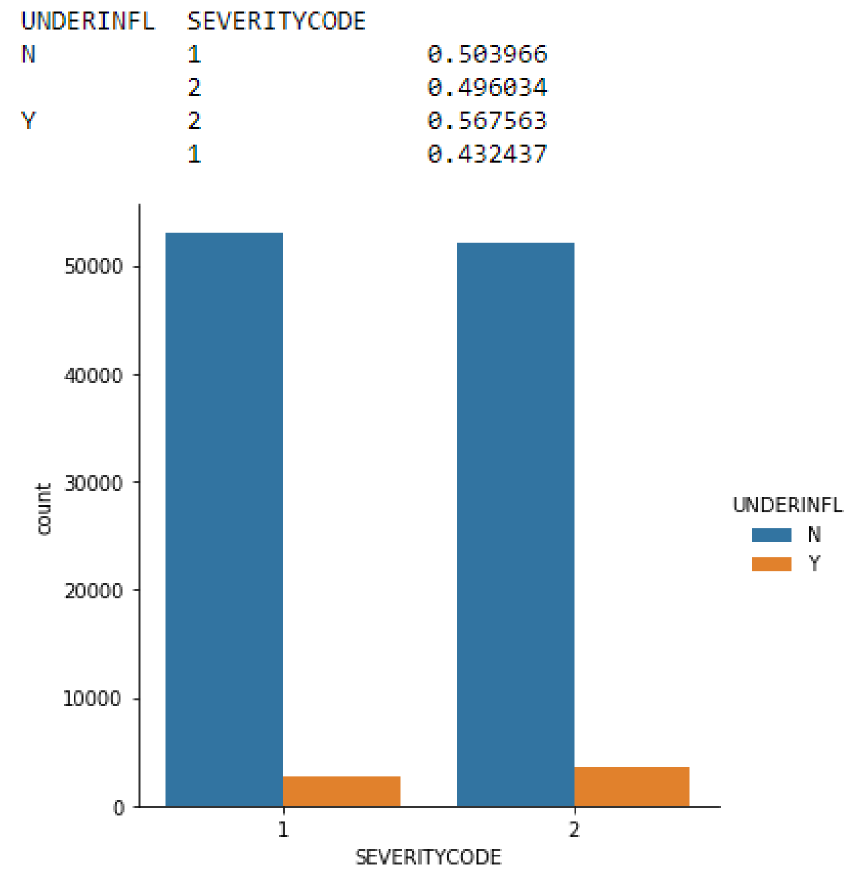


```
ADDRTYPE      SEVERITYCODE
Alley         1               0.765244
              2               0.234756
Block         1               0.572375
              2               0.427625
Intersection  2               0.613045
              1               0.386955
```

RELATIONSHIP BETWEEN ACCIDENT SEVERITY AND BEING UNDER AN INFLUENCE

Accidents were most likely to result in personal injury:

- When a driver is under the influence of drugs or alcohol



```
UNDERINFL  SEVERITYCODE
N          1               0.503966
           2               0.496034
Y          2               0.567563
           1               0.432437
```

PREDICTIVE MODELING

Four different machine learning models were developed and tested:

- K Nearest Neighbour
- Decision Tree
- Support Vector Machine
- Logistic Regression

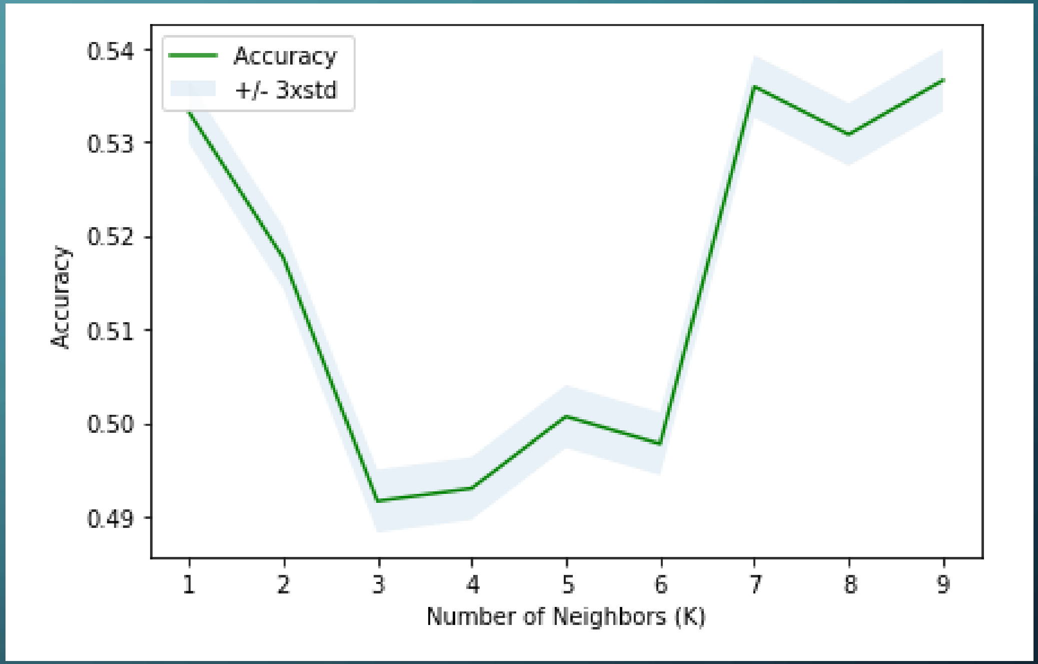
DATA SPLITTING

The 111, 308 observations were split into a training and a test set

- 80% of the data became the training set
- 20% of the data became the test set

K NEAREST NEIGHBOUR

- An initial value of k = 4 was used
- Further assessment revealed an optimum k value of 9
- The model was retrained with k = 9
- Train set accuracy = 0.542
- Test set accuracy = 0.536



DECISION TREE

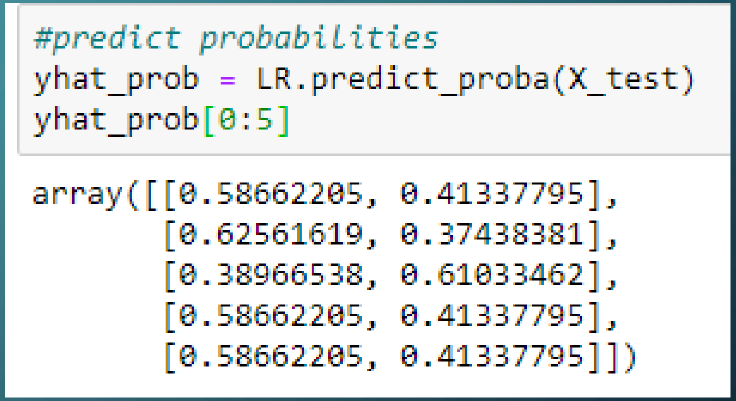
- A decision tree model was fit using the Scikit-learn library
- Prediction accuracy was determined to be 0.596

SUPPORT VECTOR MACHINE

- An SVM modeled was constructed and fit using the Scikit-learn library
- The model accuracy was determined to be 0.597

LOGISTIC REGRESSION

- A logistic regression model was developed and fit using the Scikit-learn library
- Model accuracy was determined to be 0.596



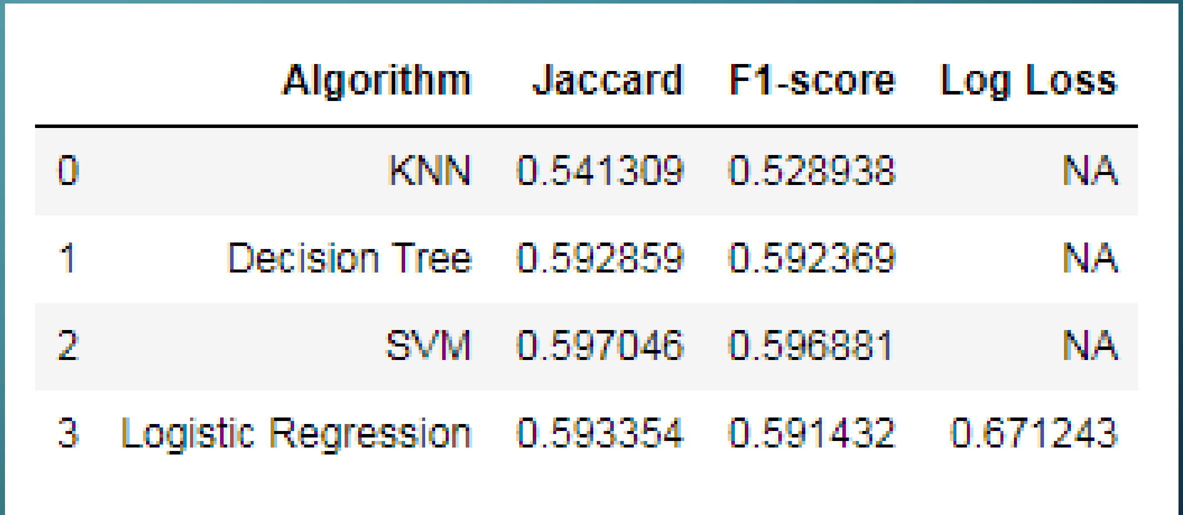
```
#predict probabilities
yhat_prob = LR.predict_proba(X_test)
yhat_prob[0:5]
```

```
array([[0.58662205, 0.41337795],
       [0.62561619, 0.37438381],
       [0.38966538, 0.61033462],
       [0.58662205, 0.41337795],
       [0.58662205, 0.41337795]])
```

MODEL METRICS

The metrics performed rather poorly, with Jaccard Index and F1-scores being between 0.54 and 0.6 for all models

Log loss for the Logistic Regression model was higher than desired, at 0.671



	Algorithm	Jaccard	F1-score	Log Loss
0	KNN	0.541309	0.528938	NA
1	Decision Tree	0.592859	0.592369	NA
2	SVM	0.597046	0.596881	NA
3	Logistic Regression	0.593354	0.591432	0.671243

CONCLUSIONS

- The models definitely need improvement prior to being deployed for any practical purpose
- The models may benefit from greater or fewer attributes, a reduction in the number of potential values associated with each attribute, or from a different set of attributes altogether
- Visualization of purely categorical data is surprisingly challenging!

FUTURE DIRECTIONS

- The project may be revisited once the inexperienced author further develops her data science and Python skills, as the project has highlighted important areas where further knowledge is needed
- Repeating the project with a dataset for a region with different weather and other conditions may provide some interesting information regarding potential regional differences in accident patterns
- New questions emerged from the data analysis section, such as: “Why are accidents more likely to be severe when they occur in intersections?” This question could potentially be answered with the same data set

THANK YOU :)