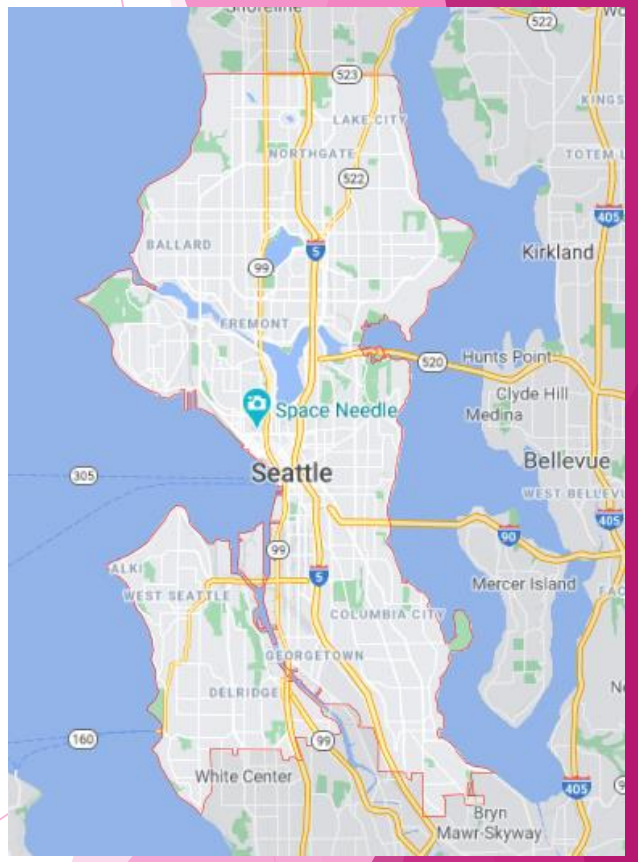
Seattle Traffic Analysis

Capstone Project Presentation

Introduction / Business Understanding

- Everyone who commutes daily to work would know that traveling can be a stressful and time wasting activity if not planned properly
- The aim of this project was to see if we can build a model to be able to predict the severity of an accident taking place taking into account different environmental/traffic/geographical factors
- This will be able to help both, law enforcing agencies as well as daily commuters.
- In the end, we all stand to gain collectively as a society as we will have less accidents, safer roads, less pollution (noise and air) due to less traffic jams and an over improvement in daily commute both in terms of time and stress.



Data Sources

- Collision data from Government of Seattle Website (https://data.seattle.gov/Land-Base/Collisions/9kas-rb8d)
- Description is available at: https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf
- There are a total of **40 Variables** and **221267 Data points/Observations**
- The aim is to be able to predict **SEVERITYCODE** from all the independent variable available

Exploratory data analysis

- Some Columns has too many empty observations
- These were removed from the data set

PEDROWNOTGRNT	97.654807
SPEEDING	95.515586
EXCEPTRSNDESC	94.679501
INATTENTIONIND	86.364273
INTKEY	67.530455
EXCEPTRSNCODE	54.385268
SDOTCOLNUM	42.542312
LIGHTCOND	11.973946
WEATHER	11.933746
ROADCOND	11.897158
COLLISIONTYPE	11.847924
ST_COLDESC	11.847924
UNDERINFL	11.838890
JUNCTIONTYPE	5.407676
ST_COLCODE	4.251792
X	3.374603
Y	3.374603
LOCATION	2.072370
ADDRTYPE	1.676687
SDOT_COLCODE	0.000452
SEVERITYCODE	0.000452
SDOT_COLDESC	0.000452
OBJECTID	0.000000
INCKEY	0.000000
COLDETKEY	0.000000
REPORTNO	0.000000
STATUS	0.000000
HITPARKEDCAR	0.000000
SEVERITYDESC	0.000000
PERSONCOUNT	0.000000
PEDCOUNT	0.000000
PEDCYLCOUNT	0.000000
VEHCOUNT	0.000000
CROSSWALKKEY	0.000000
SERIOUSINJURIES	0.000000
FATALITIES	0.000000
INCDATE	0.000000
INCDTTM	0.000000
SEGLANEKEY	0.000000
INJURIES	0.000000

- Other columns with empty observation were dropped owing to the criticality of the data

X	7475
Y	7475
ADDRTYPE	3712
SEVERITYCODE	21616
JUNCTIONTYPE	11995
UNDERINFL	26293
WEATHER	41634
ROADCOND	41561
LIGHTCOND	40124

Model Development

- Since the problem at hand needs a Supervised Machine Learning Algorithm, we looked at the options we have. These were
  - Classification Models (Used for Categorical Values)
  - Regression Models (Used for continuous values)

- Since we were dealing with categorical variable we used classification models for prediction, namely
  - K-Nearest Neighbour (KNN)
  - Decision Tree

Prediction Accuracy

- To assess the accuracy of our models, we will be using the following metrics
  - F1 Score
  - Jaccard Score

- Following is the table that we get as a result

Algorithm	Jaccard	F1-score
KNN	0.6416	0.5867
Decision Tree	0.6563	0.5310

Conclusion

- Both our models were able to predict the outcome with considerable accuracy
- The decision tree was able to outperform KNN by a slight margin when it comes to Jaccard score but F1 was better for KNN
- The prediction accuracy can be further improved by using other complex models available
- The Seattle department can now take it account these findings to be able to improve safety standards on the roads of Seattle.