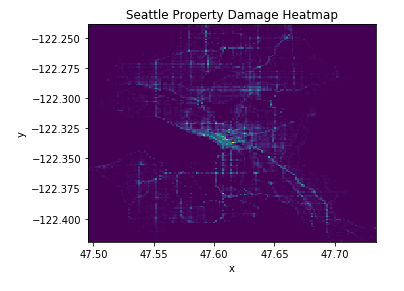
**Coursera IBM Data Science Capstone**

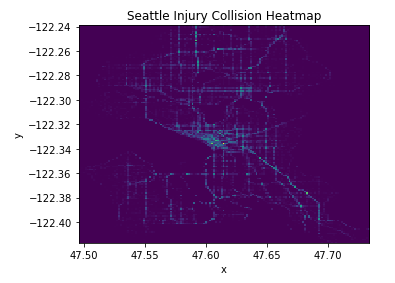
**Ryan Schmitt**

**Introduction**

Accidents happen. Responding to accidents requires resources. In this project, I am attempting to model the severity of an accident, given various factors. First Responders could use such a model to be able to respond on the scene with the correct resources given the accident conditions to provide the most effective assistance at the most critical time.

The stakeholders in this model project are any given city's first responders and city planners. With this model, city planners can better understand the environmental conditions that lead to more severe accidents and continue to work to mitigate those effects going forward. For first responders, this model could predict a mild accident where extra support vehicles would be unnecessary and should be preserved for responding to more severe accidents.





**Data**

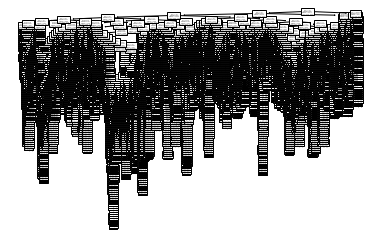
The dataset being used can be found at https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv. It includes data on auto accidents in the city of Seattle from 2004 to 2020. There is various information on driving conditions, locations, number of vehicles and people involved, etc.

There was a fair amount of preprocessing needed for this dataset, i.e. many column entries with no or missing data. I made my best assumptions on replacement values and whether to delete altogether. Any arbitrary codes or code descriptions were deleted from the dataset. I scanned the remaining features and determined default position values. For instance, if the driver was under the influence, that appears to be captured in the dataset with a "Y". However, all other cells in that feature were blank. I then assumed that positives were captured, and negatives were not, so I populated those remaining cells with "N".

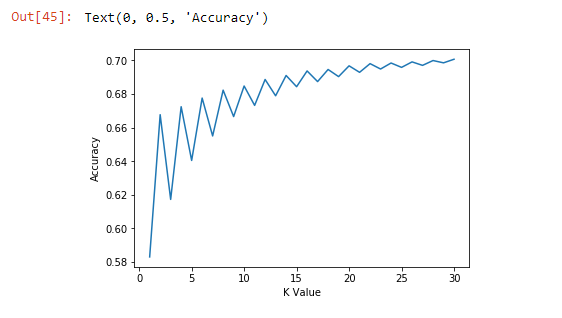
**Methodology**

After cleaning the data, replacing missing values, and deleting unnecessary features, I decided on Decision Trees as the first model to train the data set. Afterwards, I wanted to see how it would fare against a K-Nearest Neighbors and a Support Vector Machine (RBF kernel).

The resulting Decision Tree structure was gigantic, and the graphic is nearly impossible to utilize.



Next, I attempted to optimize the k-value for the K-Nearest Neighbor model. I started with a range of 1 to 10 and plotted it, but the accuracy score appeared to keep improving. I pushed the range up to 20, and still saw similar improvement. I finally pushed it up to 30, and can see the accuracy score starting to plateau around 70%, and I anticipate diminishing returns with higher k-values.



Lastly, the SVM model was ran for comparison, with a radial basis function kernel. I could have selected other kernels but chose to end the modeling here.

**Results**

The model with the highest accuracy score for predicting accident severity was the Support Vector Machine, yielding a model accuracy of 70.342%. K-Nearest Neighbor model, looking at the 30 nearest neighbors, modeled to an accuracy of 70.070%. Decision Trees came in last at 68.552% accurate.

**Discussion**

All three models were very similar in their accuracy scores, which lends confidence in what can be achieved given the input features. There were 18 features models, which could be on the high side. These models could be iterated through with a smaller number in a more granular feature selection process and see if an anomalous feature can be identified that has been negatively affecting the accuracy scores. I really did not drill down very hard on the time of day of the accidents, so that area could be explored more. Maybe grouping the events into hourly bands, or morning-noon-nighttime accidents would reveal more.

**Conclusion**

All 3 models produced similar accuracy score results. I feel confident that a ~70% accuracy score is a valid one, given the features used in these models.