**Introduction**

Accidents continue to block traffic and cause backups in the greater Seattle area. In order to help alleviate the congestion during peak hours of travel we need to know what causes them in the first place. I am exploring historical data provided to the general public from the Seattle Department of Transportation (SDOT).

**Data Acquisition, Cleaning, and Exploration**

The data, and metadata, was provided through the Coursera IBM Data Science Professional Certificate Applied Data Science Capstone course.

After looking at the provided excel sheet, I removed 5,639 of the 194,673 rows in the initial data set that were noted Not enough Information or Insufficient Location Information (NEI). After removing the ~5600, I also removed the following columns from the dataset because they were not of concern for the model; one of the SEVERITYCODE fields, X, Y, INCKEY, COLDETKEY, REPORTNO, STATUS, EXCEPTRSNCODE, EXCEPTRSNDESC, and INCDATE.

Focusing on the 5 levels of severity codes (0-4), 0 being a minor accident and 4 being the most severe with a high probability of a fatality involved. After checking the unique values and count of each for the Severity, Weather, Road Condition, Light Condition, and Speeding, I leveled data sets out so that there were equal amounts of data in severity codes 1 (Very Low Probability — Chance or Property Damage) and 2 (Low Probability — Chance of Injury).

Since there were only two options of Severity Code with 19,576 as the lowest, I set the data frames to 19,576 random records. This allows for the tests to be more accurate. After creating a new data frame, I converted the following fields to a coded value; Weather, Road Conditions, and Light Conditions.

**Results**

After running several models against the data, I found the best model to predict the severity of an accident is the Decision Tree model. The others that were ran were the Linear Regression, Support Vector Machine, and the K Nearest Neighbor models. The two values produced from all tests were a F1-Score and Jaccard Index. A F1-Score is score to test the accuracy of the model that is graded on a 0-1 scale, so the closer to 1 the better. A Jaccard Index is used for gauging the similarity of the two data frames that were created based on severity in this case. The scores are shown in Graph 1 below.

|  |  |  |
| --- | --- | --- |
|  | F1 – Score | Jaccard Index |
| K Nearest Neighbor (KNN) | 0.4974 | 0.5083 |
| Decision Tree | 0.5113 | 0.5484 |
| Support Vector Machine (SVM) | 0.5002 | 0.5143 |
| Logistic Regression | 0.4953 | 0.5106 |

**Conclusion**

After performing out tests we can conclude that the weather, light, and road conditions all are factors in the severity of an accident.