Coursera Capstone
Final Assignment

Seattle Traffic Accident Study

By Geoffrey Longtin

Predicting accident is valuable for Seattle Drivers

- In present day, applications allow driver to find best driving routes, avoid traffic and so.
- However, there is a need to have an application to allow Seattle drivers to evaluate the risks of accident.
  - If the drivers are warned of the potential risk levels of accidents, then they might avoid driving in certain combination of road and environment conditions.
  - Overall, this will help increase road safety for all.

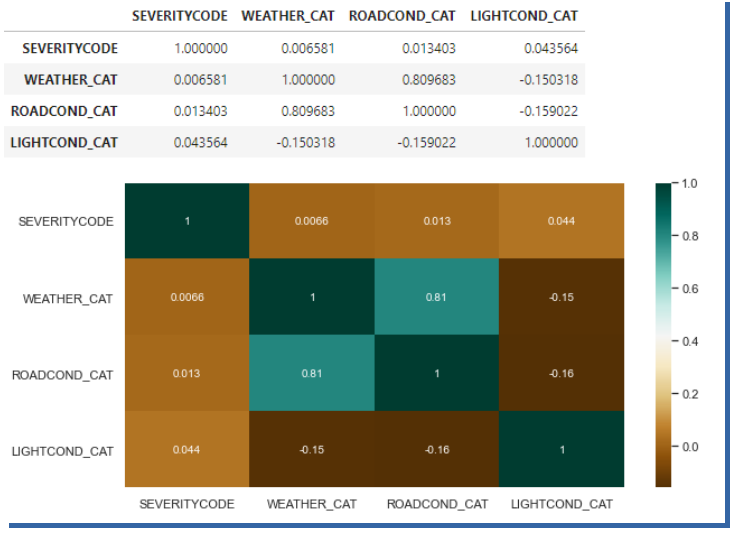
Data Acquisition and Cleaning

- Data used for this project is the collection of all collisions provided by the Seattle Police Department and recorded by Traffic Records since 2004 until the present and was downloaded from https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv
- There are 194 673 rows and 38 features in the raw dataset.
- Accident related to speed, inattention or alcohol intoxication were dropped.
- Also, all other non relevent features were dropped.
- Data was balanced to have the same number of accident cases.
- Cleaned and balanced data contains 81 714 rows and 4 features.

Data Correlation

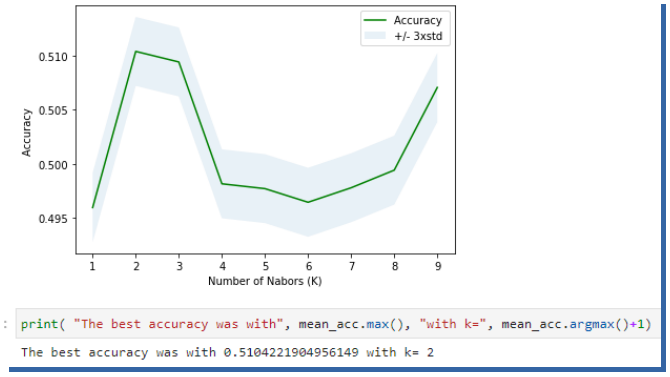
- There seems to be a mild correlation with the 3 variables and the target variable
- The Weather and Road Conditions are strongly correlated this is due to the fact the weather directly affects road conditions.
- From these observations, we should expect the model accuracy to be negatively impacted.

	SEVERITYCODE	WEATHER_CAT	ROADCOND_CAT	LIGHTCOND_CAT
SEVERITYCODE	1.000000	0.006581	0.013403	0.043564
WEATHER_CAT	0.006581	1.000000	0.809683	-0.150318
ROADCOND_CAT	0.013403	0.809683	1.000000	-0.159022
LIGHTCOND_CAT	0.043564	-0.150318	-0.159022	1.000000



K-Nearest Neighbor (KNN) Model

- The best KNN Model that we could build using the data used a K value of K = 2 as shown in figure below



- The model evaluation show that
  - Jaccard score is 0.3496
  - F1 score is 0.5103

Decision Tree Model

- The Decision Tree Model was built using a entropy criterion and a maximum depth of 7.

- The model evaluation show that
  - Jaccard score is 0.1912
  - F1 score is 0.4748

Logistic Regression Model

- The Logistic Regression Model was built using a C value of 0.1, and liblinear solver.

- The model evaluation show that
  - Jaccard score is 0.2759
  - F1 score is 0.5062
  - Logloss score is 0.6923

Conclusion and Recommendations

- Built models that would prove useful of Seattle drivers in evaluating the risk of accidents.
  - The best model uses the Logistic Regression methodology.
- The models should be improved as it has room to gain accuracy
  - Improvements could come by gathering driver data such as age, sex, physical disabilities which could impair driving.