Predicting Severity Code of Collisions in Seattle

Applied Data Science Capstone

Carol Yang

Oct. 1, 2020

Introduction

This project predicts the severity of car accidents using variables such as time and location. It would be beneficial for the police department and/or the development department to acknowledge the attributes of accidents that cause the most harm and damages in order to avoid them by implementing alternative city planning from the development department and patrolling schedule from the police department.

Data

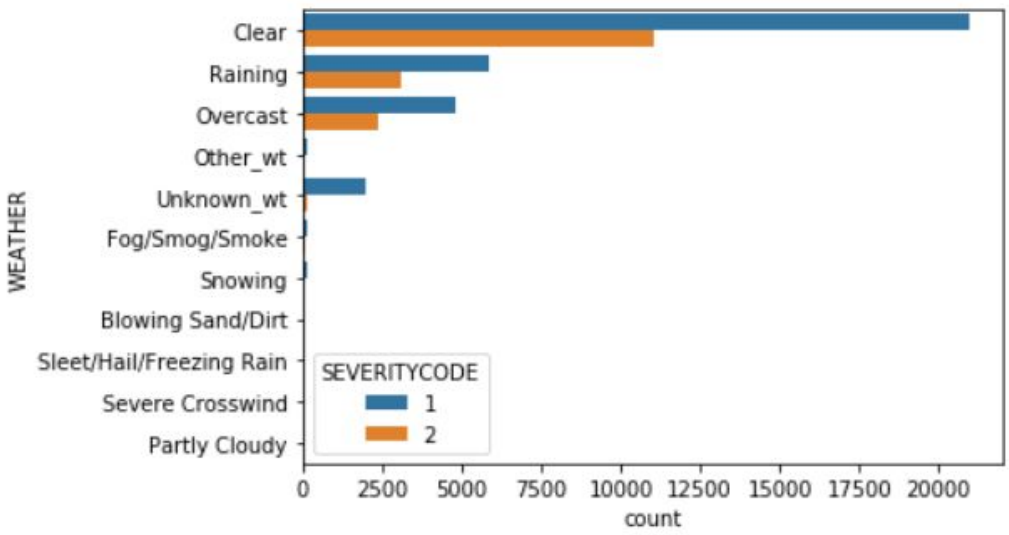
The dataset records the collision data in Seattle from 2004 to May 2020. The target variable we are looking at is the severity code. Predictor variables include location factors (i.e. coordinates, street name), timestamp factors, environmental factors (weather, light condition, road condition), and the subjects involved (number of people involved, number of pedestrians, number of cyclists involved, and number of vehicles involved).

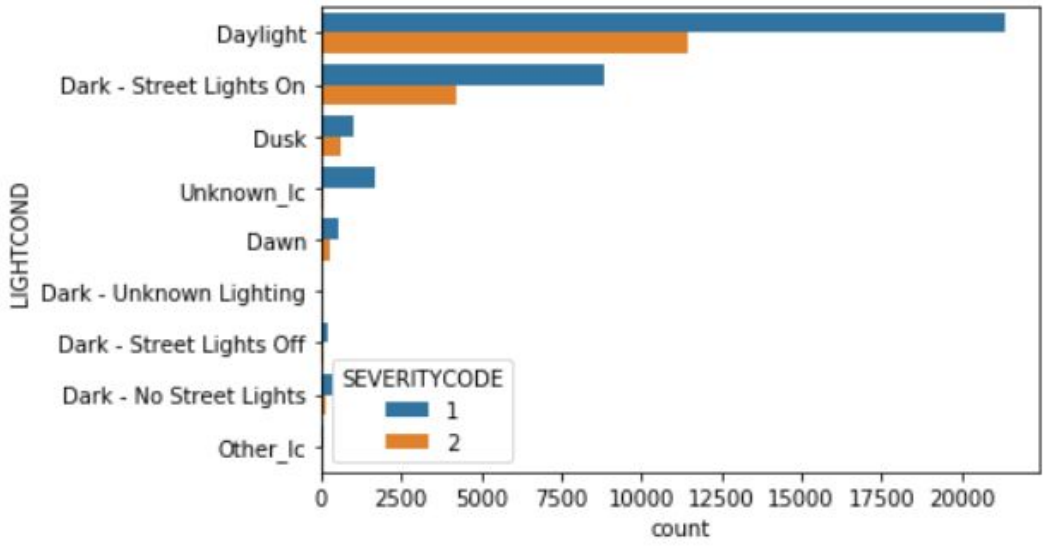
Pre-Processing

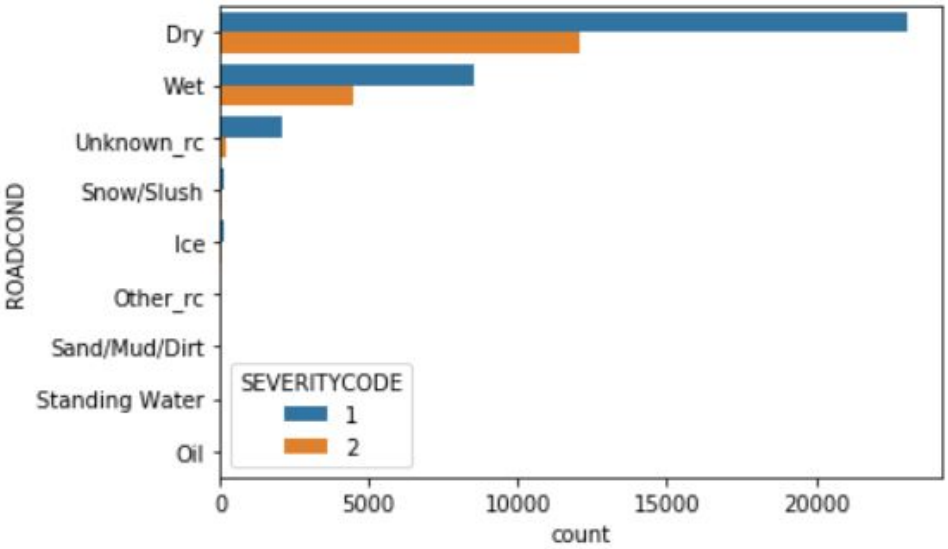
In order to provide relevant and up-to-date analysis and to speed up the process, I decided to use only the data starting 2015. I made a few graphs looking at severity versus the various groups of factors listed above as exploratory analysis.

First, I looked at severity versus the environmental factors: WEATHER, ROADCOND, and LIGHTCOND.

.

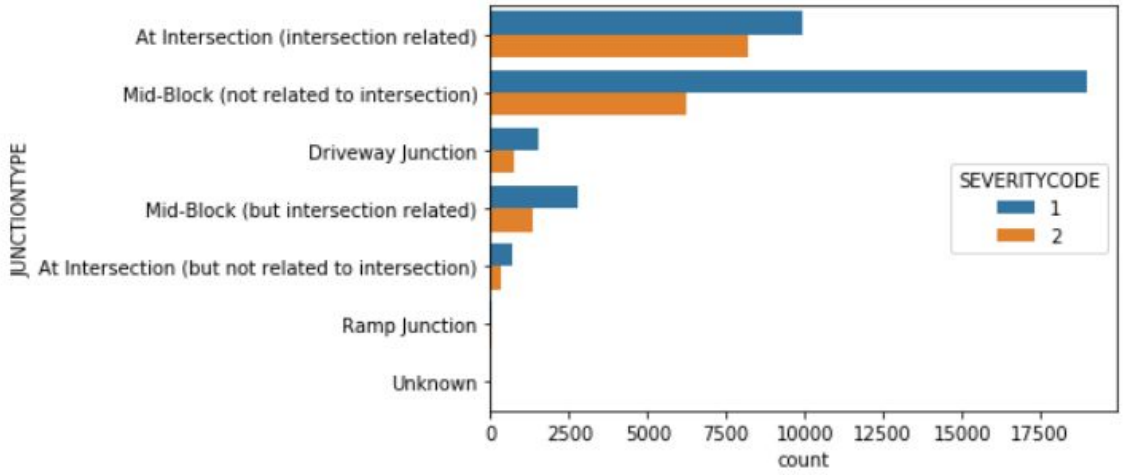




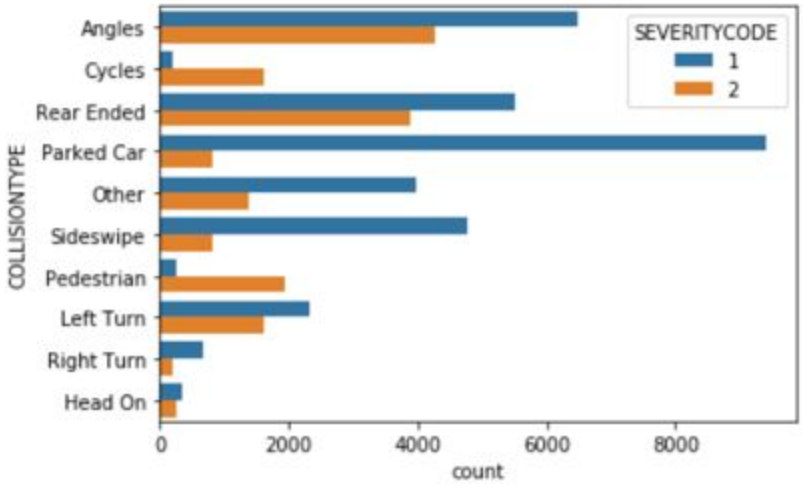


It is surprising to find that more collisions occurred in the most ideal environment possible (i.e. on a clear day, on a dry road, and during the day).

Next, I looked at severity versus junction type to find out that the number of collisions without fatalities is significantly higher at mid block and most collisions with fatalities happen at intersections.



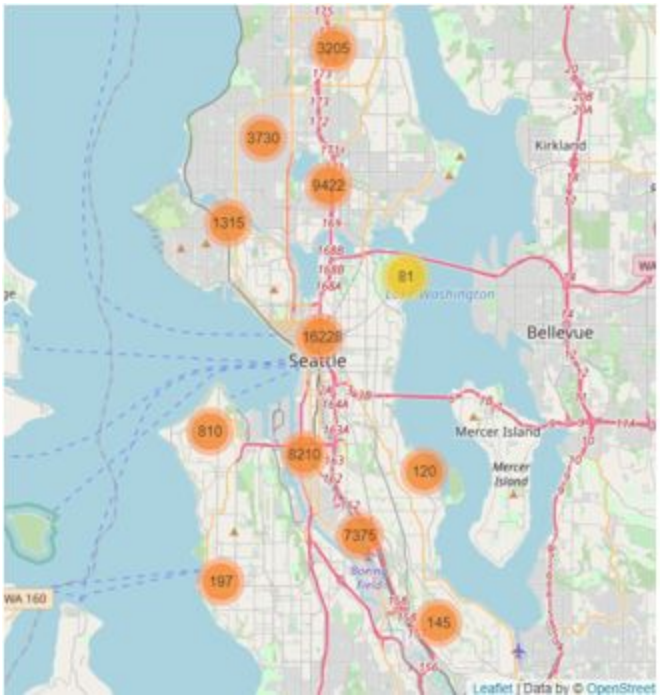
Collision types such as left turns, right turns and pedestrians may be a good indicator of severity code as well since in some circumstances, more damages are done compared to others, for example, left turns usually cause more damage than right turns in an intuitive sense.



One remark from this graph is that any collisions involving cycles and pedestrians almost always result in fatalities. A lot of collisions without fatalities involve a parked car, which makes sense as one of the subjects involved is not moving at all.

I then looked at the locations of where the collision occurred by generating a map. Most of the clusters are centred along the I5 highway and the clusters get larger as they approach the

downtown area. This is not particularly surprising as there is heavier traffic surrounding or on the highway and in the downtown area compared to the suburb and residential area.



Data Cleansing

The original dataset with almost 200,000 lines of data. After extracting the data starting from 2015, only less than 58,000 lines of data are left. The following steps were used to clean the data:

1. Empty columns or columns with only one unique value like EXCEPTRSNCODE, EXCEPTRSNDESC, and PEDROWNOTGRNT are dropped.
2. Categorical variables such as WEATHER, ROADCOND, LIGHTCOND are converted to numerical values.
3. Binary variables (i.e. UNDERINFL, HITPARKEDCAR) are standardized to 1 and 0 to represent Yes's and No's

Methodology

The predictor variables used in this analysis include: JUNCTIONTYPE, UID (UNDERINFL), WEATHER, ROADCOND, LIGHTCOND, HPC (HITPARKEDCAR), PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, COLLISIONTYPE. I then split the data into train and test sets for the sake of model evaluation later. Using the train set, I built a model for each of the K-nearest neighbours, random forests, and support vector machines methods. The test

set is then used to evaluate the models. Jaccard score, F1 score, and the accuracy score were used as measures.

Algorithm	Jaccard	F1_score	Accuracy_score
KNN	0.673368	0.692379	0.714311
Random Forest	0.709938	0.675957	0.732431
SVM	0.709885	0.679607	0.733385

Results & Discussion

Using the variables in the predictors set, I modeled using KNN, random forest, and SVM. All three models performed fairly well, however, SVM resulted in the highest accuracy score of 73.4% in predicting the severity code.

Based on the graphs shown above, environmental factors (weather, road condition, light condition) play a big role in the collisions. Surprisingly, the number of collisions is significantly higher in the ideal environment (ie on a clear day, dry roads, and in the daylight). It may be because people are way more careful when they are driving in a worse environment, therefore, resulting in less collisions. Collisions with fatalities are half the number of collisions without fatalities in the ideal environment. On the other hand, there is a significantly higher number of collisions happening at an intersection and at mid-block, where the highest number of collisions with fatalities happens at an intersection, and the highest number of collisions without fatalities happens at mid-block. The collision map also revealed that clusters of collisions centre along the I5 highway, which may infer that heavier traffic leads to higher chance of collisions.

Conclusion & Recommendations

Through this project, a lot of unexpected insights are found. For example, there are drastically higher numbers of collisions in the best environment possible, and any collisions involving pedestrians and cyclists almost always result in fatalities. Therefore, drivers must be cautious at all times regardless of the environment. The police department, on the other hand, should schedule more patrols surrounding the highway (i.e. on the highway and the highway exits) and the downtown area and impose penalties on jaywalking or crossing a red light on major intersections or major streets to keep pedestrians and/or cyclists safe.