Predicting Severity Code of Collisions in Seattle

CAROL YANG

Introduction

Predicting the severity of collisions using variables such as time, location, and number of people involved

Beneficial to police department and/or the development department

- Avoid collisions that cause most harm and damages by:
  - Implementing alternative city planning
  - Patrolling schedule

Data

Records collision data from 2015 to May 2020

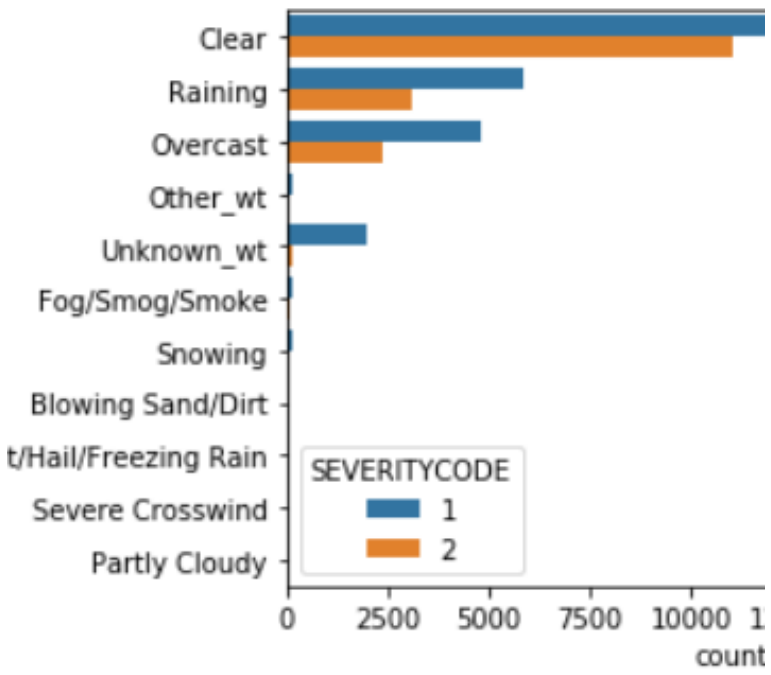
Target variable: *SEVERITYCODE*

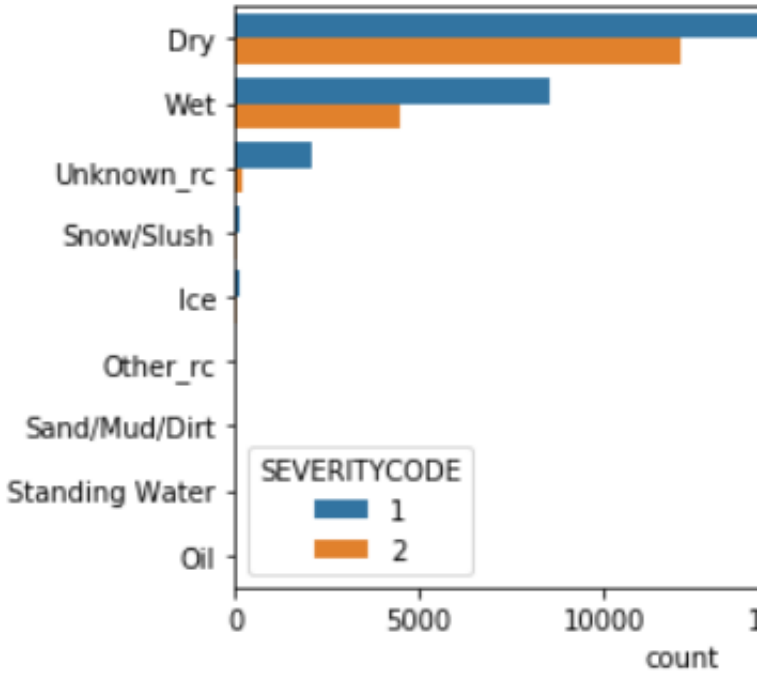
Predictor variables:

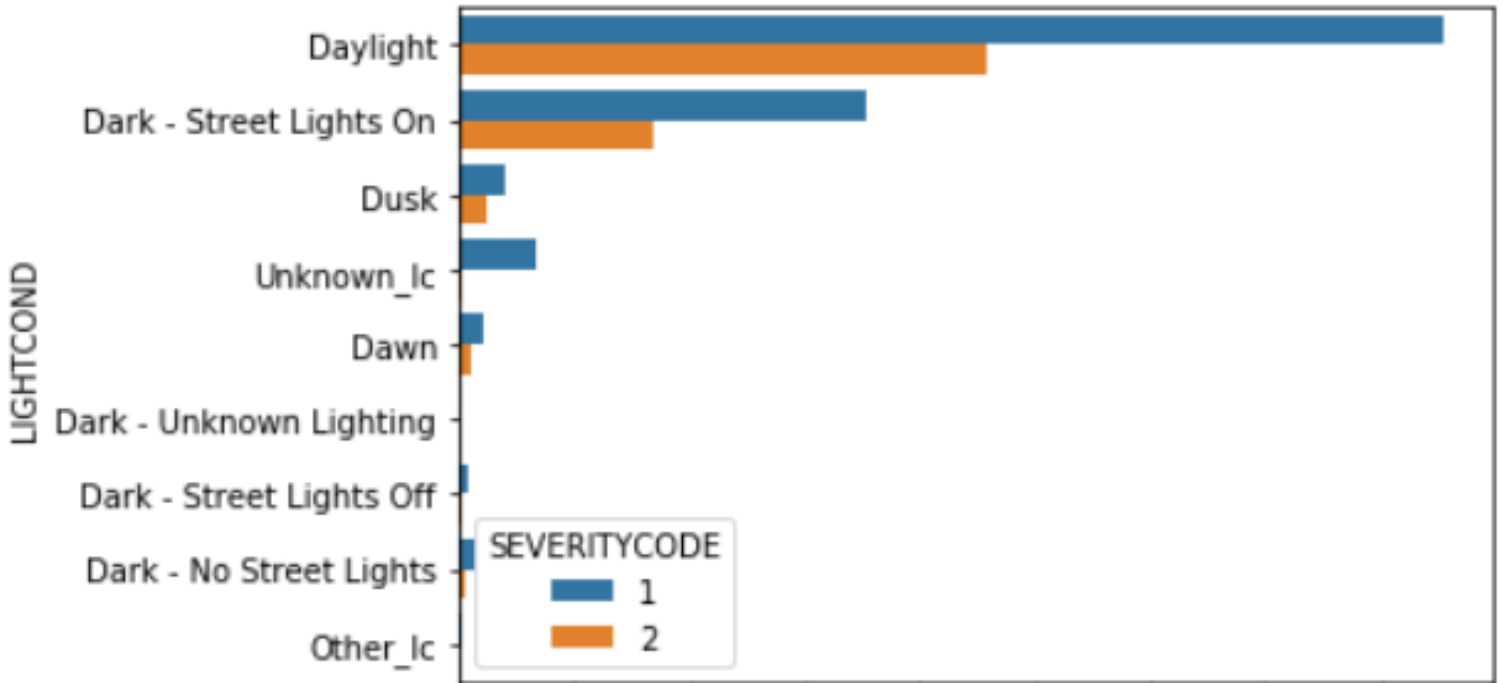
- Location factors: coordinates, street name, address type, junction type
- Timestamp factors: time of incident
- Environmental factors: weather, light condition, road condition
- Subjects involved: number of people involved, number of pedestrians involved, number of vehicles involved

Severity VS Environmental Factors

- Regardless of severity, most collisions occur on a clear day, on a dry road, during the day.
- People may be way more careful when driving in imperfect environment



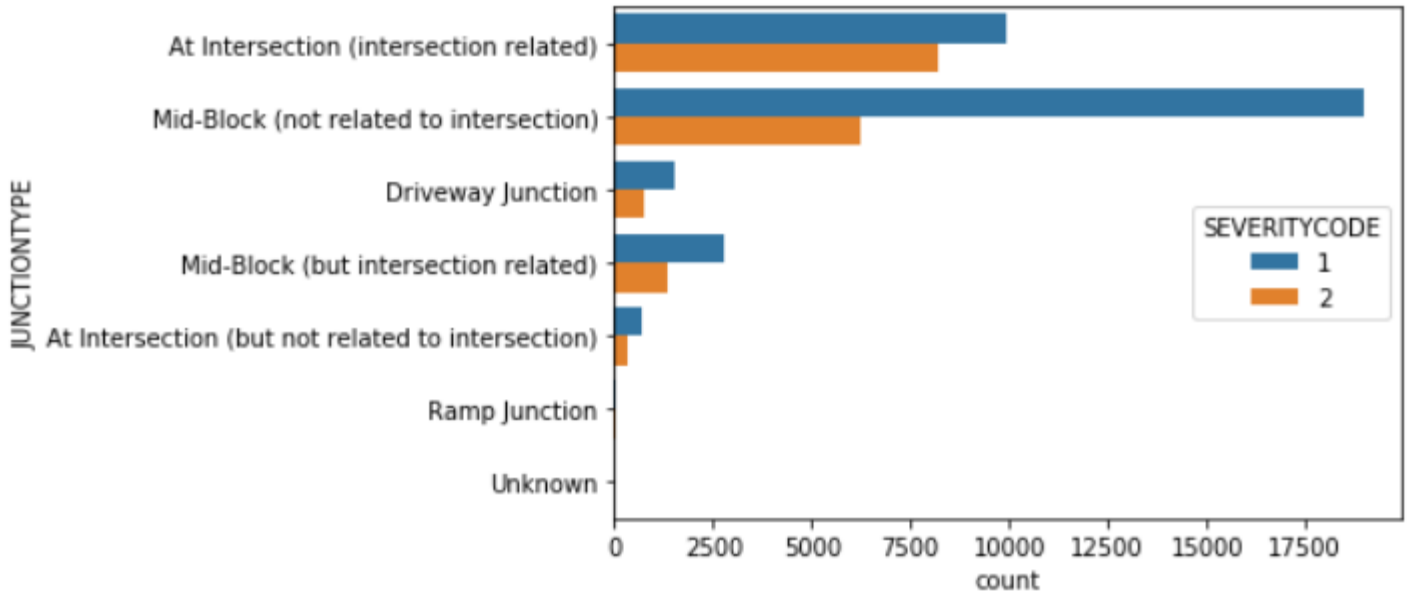




Severity VS Junction Type

Number of collision without fatalities is significantly higher at mid-blocks

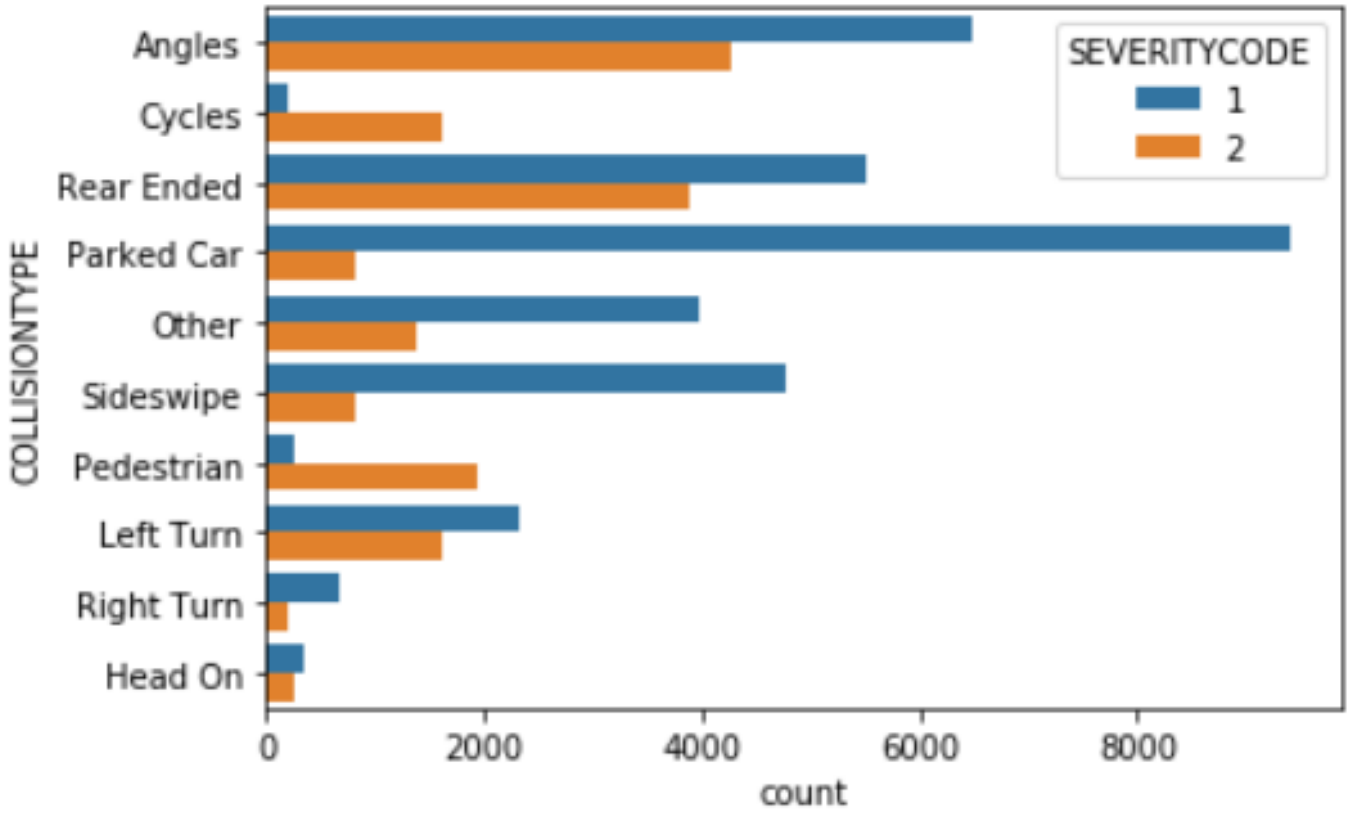
Most collisions with fatalities happen at intersections



Severity VS Collision Type

Number of collisions involved a parked car are significantly higher than other collision types

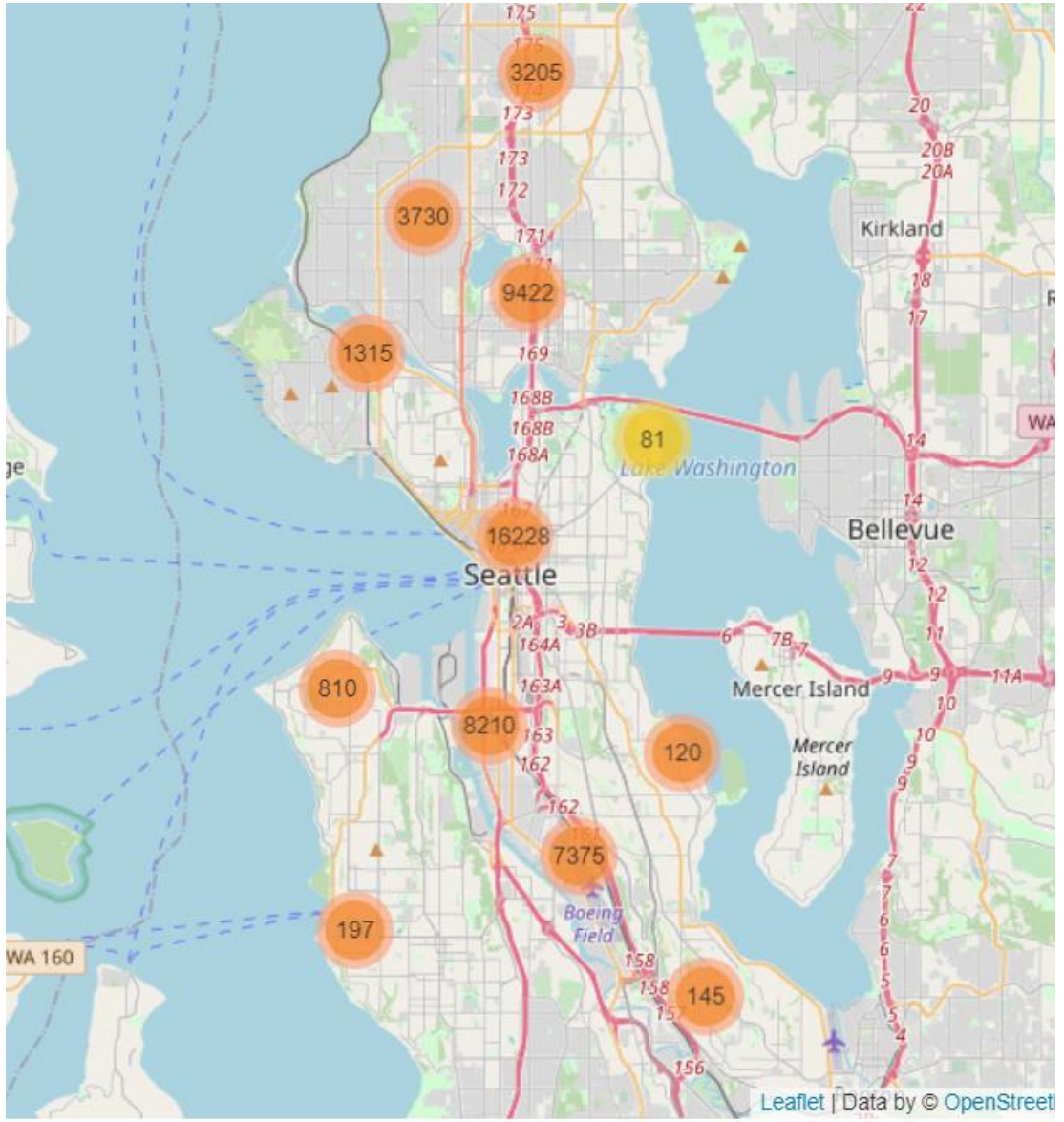
Collisions involved cycles and pedestrians almost always result in fatalities



Collision Clusters

Most collision clusters center along I5 highway

The clusters get bigger as they approach the downtown area



Data Cleansing

1. Drop empty columns (i.e. EXCEPTRSNCODE, EXCEPTRSNDESC)
2. Convert categorical variables such as WEATHER, ROADCOND, and LIGHTCOND to numerical values
3. Standardize measures of UNDERINFL and HITPARKEDCAR to 1 and 0
4. Predictor Variables:
   - JUNCTIONTYPE
   - UID
   - WEATHER
   - ROADCOND
   - LIGHTCOND
   - HPC
   - PERSONCOUNT
   - PEDCOUNT
   - PEDCYLCOUNT
   - VEHCOUNT
   - COLLISIONTYPE

Methodology

1. Split the dataset to train and test sets
2. Use ML methods to do the classification using the predictors:
   - K-Nearest Neighbours
   - Random Forests
   - Support Vector Machines

Model Evaluation

Jaccard score, F1 score, and accuracy score are computed against the test set on the three models mentioned above

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

All three models perform well, while SVM performs slightly better than the other two

Environmental factors (weather, road condition, light condition) play a big role to collisions:

- Number of collisions is significantly higher in the ideal environment

Number of collisions is much higher at an intersection and at mid-block:

- the highest number of collision with fatalities happens at an intersection
- highest number of collision without fatalities happens at mid-block

Collisions involving pedestrians and cyclists almost always result in fatalities

Clusters of collisions center along the I5 highway:

- Size of clusters increase as they approach the downtown area
- Heavier traffic infers higher chance of collisions

Recommendations & Conclusion

To the drivers: Drive cautiously regardless of the environment

To the police department:

- Schedule more patrols surrounding the highway (i.e. on the highway and the highway exits) and the downtown area
- impose penalties on jaywalking or crossing a red light on major intersections or major streets to keep pedestrians and/or cyclists safe