Capstone project: Car Accident Severity

Jan Cinert
Oct 2020

Introduction / Business Problem

- Accident prediction can be leveraged for optimizing infrastructure, providing police support in critical areas and/or during high accident frequency times.
- This project is focusing on predicting the accident severity based on different attributes like locations, time, visibility, rough driving by influence of drugs or alcohol and more.
- This could help to identify areas, times and conditions when accidents occur more and help prevention.

Data

Our data set contains accidents in the city of Seattle and was provided by SDOT Traffic Management Division, Traffic Records Group and contains data of all types of collisions that happened in Seattle city from 2004 till May 2020.

There are 195,673 accident records and 35 variables in our data set, excluding dependent variables.

In addition there is our dependent variable SEVERITYCODE (with SEVERITYDESC short text description).

Data set has 194,673 samples and 38 columns

Data - Dependents and Variables

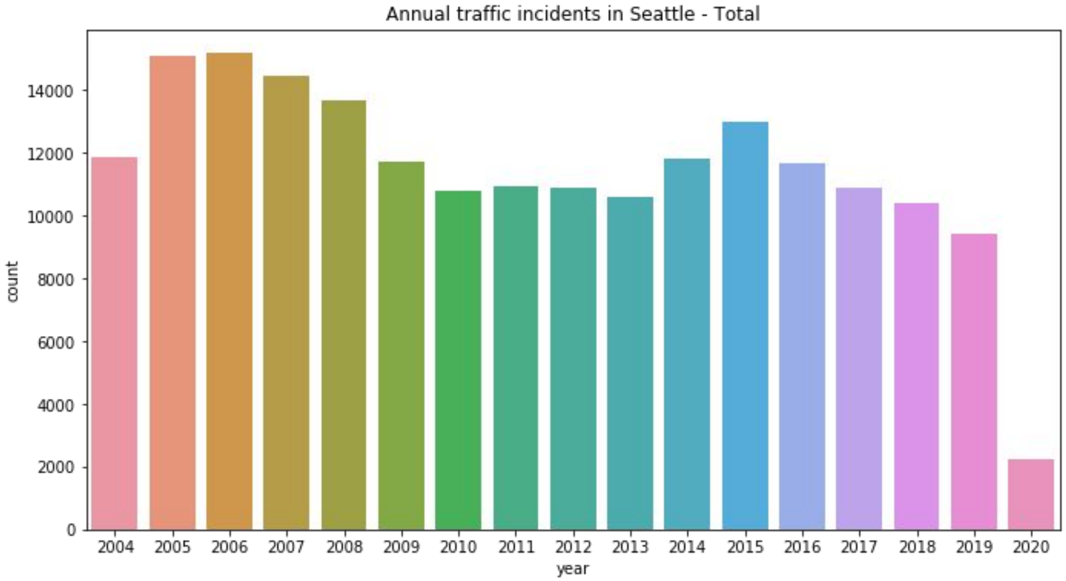
Our data set contains only 2 values of SEVERITYCODE:

	Count of accidents	**Accident Severity**
SEVERITYCODE		
1	136485	Property Damage
2	58188	Injury Collision

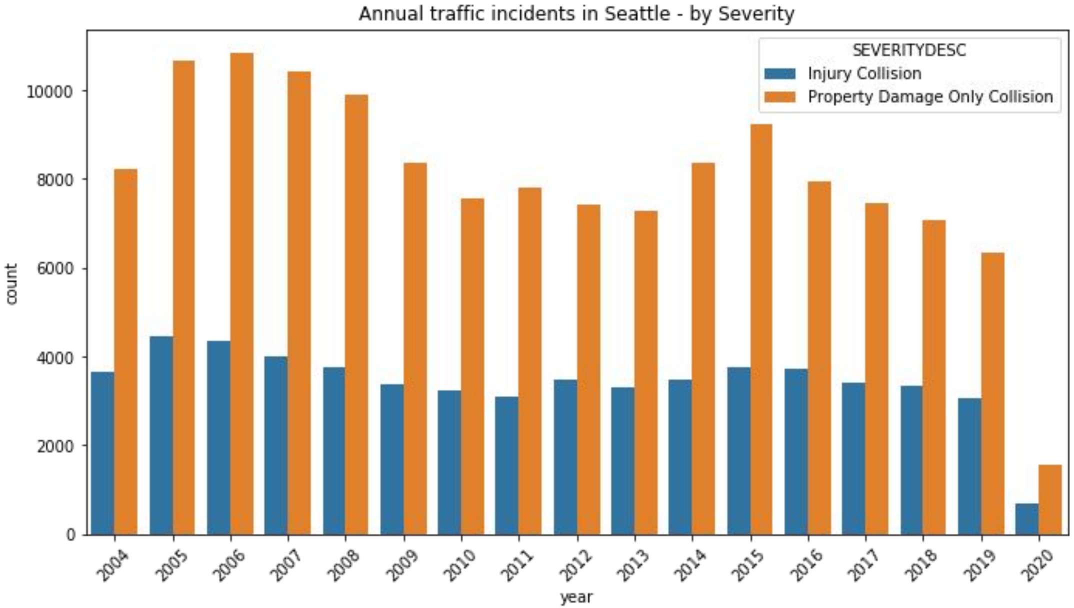
We will analyze the variables (columns) to understand if there is correlation with number of accidents and/or its severity. Key columns to be analyzed:

- WEATHER
- ROADCOND
- LIGHTCOND
- SPEEDING
- X (longitude)
- Y (latitude)

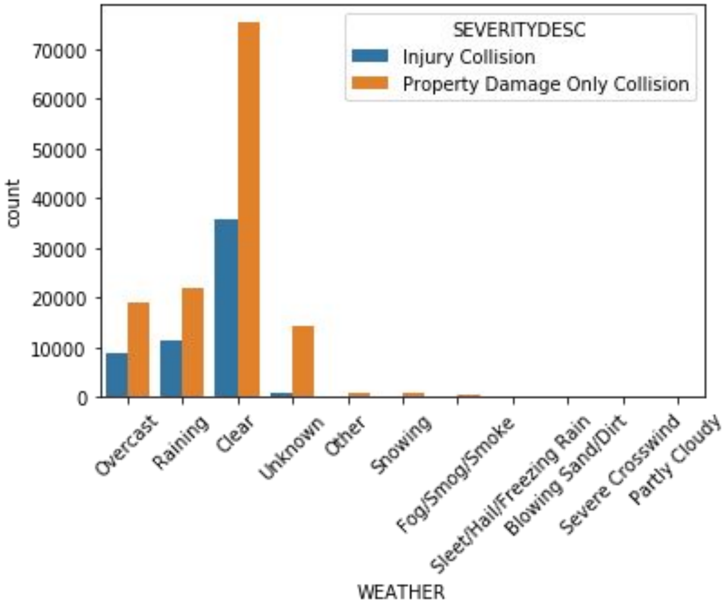
Analysis - Annual number of accidents



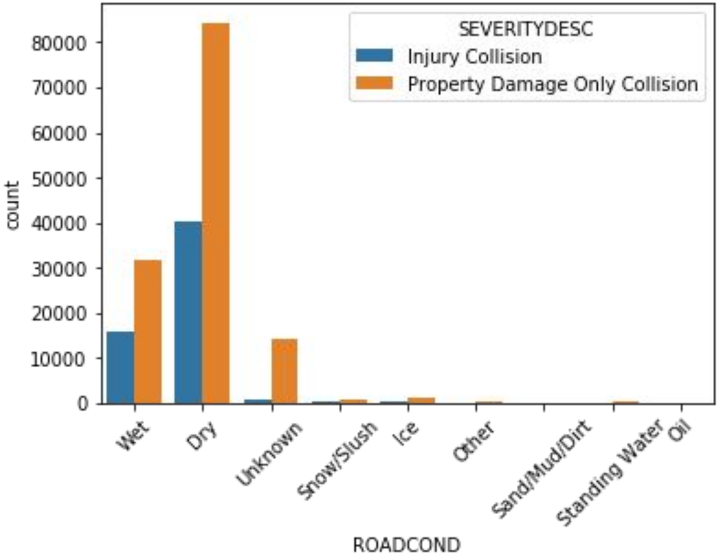
Analysis - Annual number of accidents



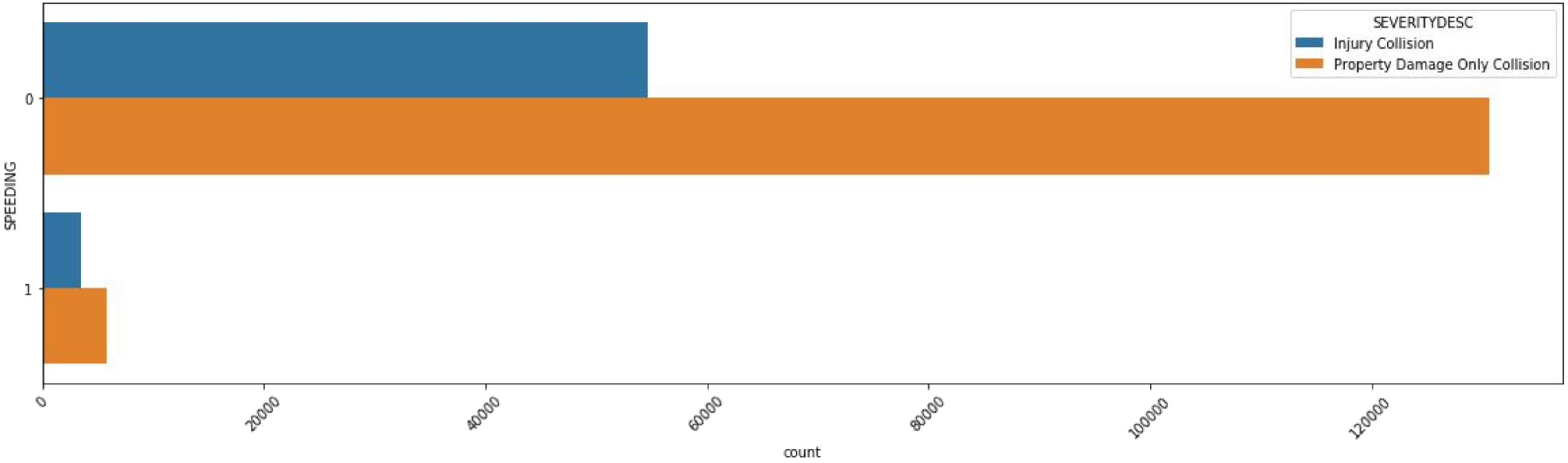
Analysis - Weather



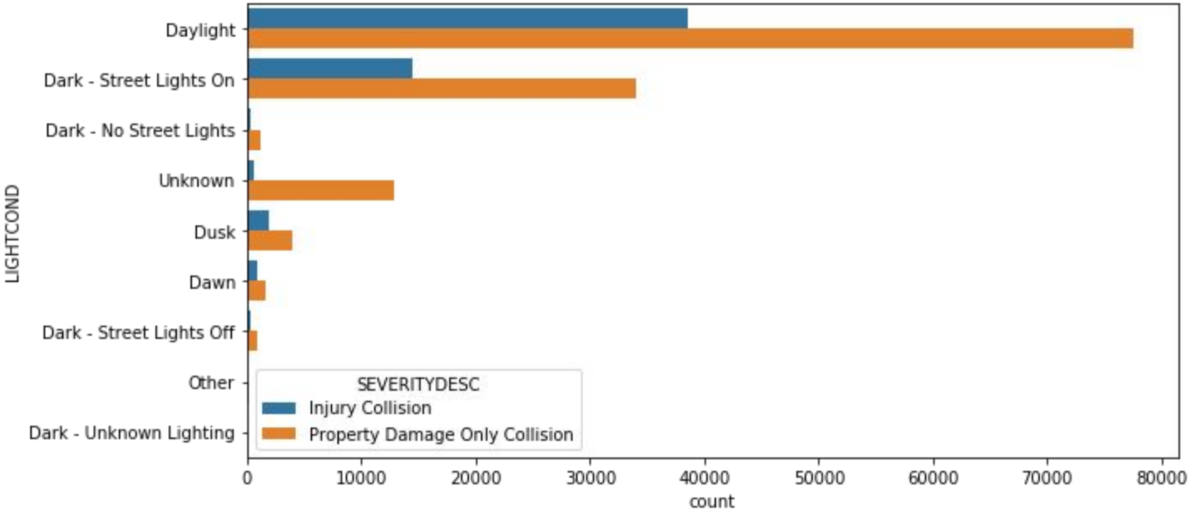
Analysis - Road Condition



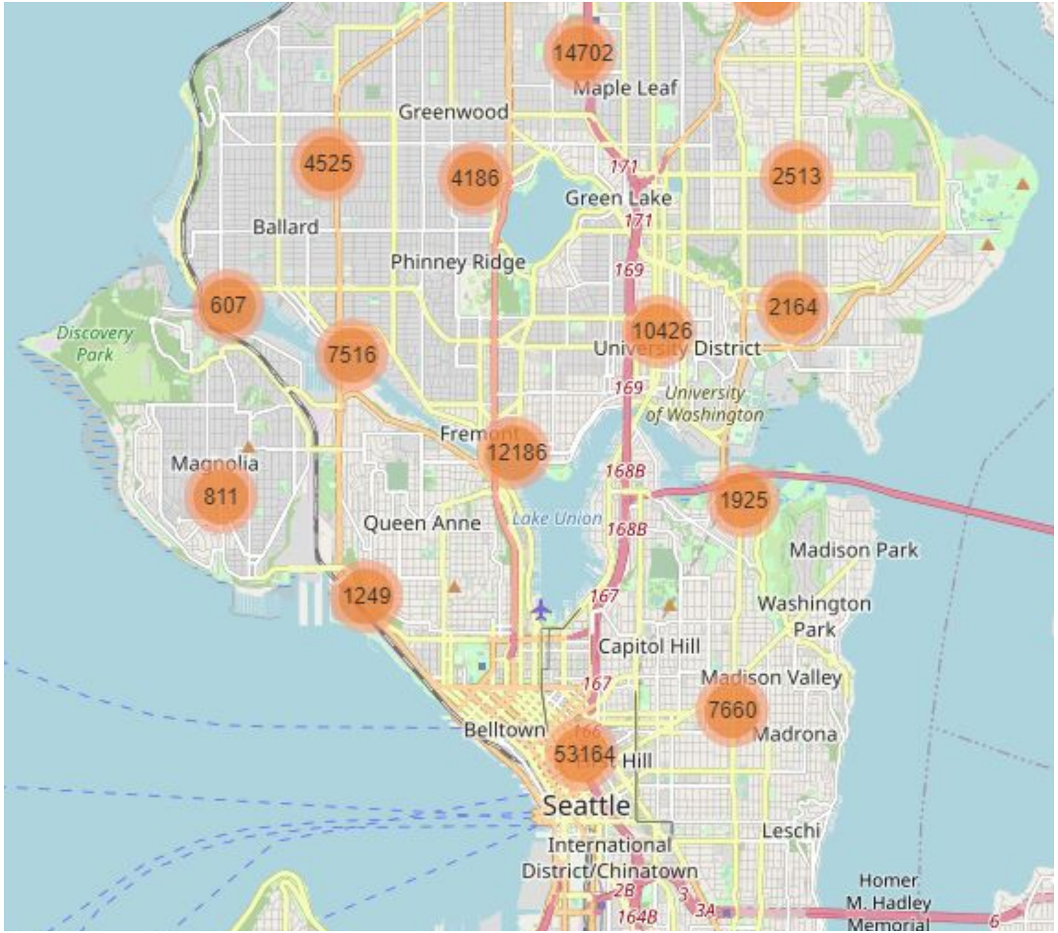
Analysis - Speeding



Analysis - Light Condition



Analysis - Location



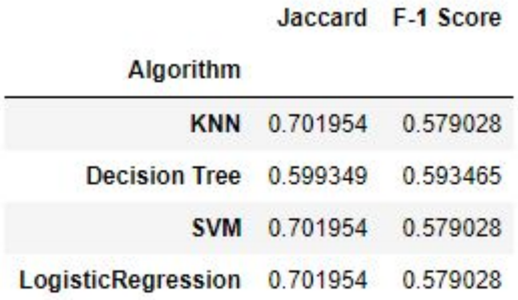
Methodology / Modelling

We will use the training set to build an accurate model. Then use the test set to report the accuracy of individual models. We will compare below popular ML algorithms:

- K Nearest Neighbor(KNN)
- Decision Tree
- Support Vector Machine
- Logistic Regression

Conclusion

In this analysis we evaluated the performance of 4 machine learning algorithms on the Seattle Collision dataset to predict the severity of an accident knowing the weather, road, light conditions and location. The three models performed very similarly, but Decision Tree stood out with a slightly higher F-1 score, but lower Jaccard index (lower accuracy). With KNN, SVM and Logistic regression we were able to meet 70% accuracy.



	Jaccard	F-1 Score
Algorithm		
KNN	0.701954	0.579028
Decision Tree	0.599349	0.593465
SVM	0.701954	0.579028
LogisticRegression	0.701954	0.579028