Capstone Project - Car accident severity

Jan Cinert

17th October 2020

Introduction / Business Problem

Every opportunity to reduce the number of traffic accidents is a good opportunity. 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). We will use X as a set of independent variables and Y as our dependent variable. We would like to identify the most influencing variables that cause the accident and the level of severity. Following standard data science steps we will:

- Review data further look for correlations, trends, ...
- Prepare data - transformation, filling missing data, cleaning the dataset and remove not relevant columns
- Model the prediction using different algorithms and compare results.
- Evaluate and conclude which factors may have more impact on the accidents

Our data set contains only 2 values of SEVERITYCODE:

- 1 = Property Damage
- 2 = 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)

Shape of the data
Data set has 194,673 samples and 38 columns

Missing data

Some features have over 40% of missing data and/or are not well described in the source data sheet. We will consider this in the model variable selection and remove such attributes away. We will use following variables to classify the severity of the accidents:

- WEATHER: Weather conditions during the time of the collision
- ROADCOND: The condition of the road during the collision
- LIGHTCOND: The light conditions during the collision
- SPEEDING: Whether or not speeding was a factor in the collision
- X: Longitude - Location of accident
- Y: Latitude - Location of accident

These features also contain missing values but are below 3% of the total amount of samples.

Target Variable

The target variable SEVERITYCODE corresponds to the severity of the collision:

- 1 - Property Damage
- 2 - Injury Collision

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

Methodology

We will use a limited number of features as attributes for our model to classify SEVERITYCODE. Selected features are:

- WEATHER
- ROADCOND
- LIGHTCOND
- SPEEDING
- location (X)
- location (Y)

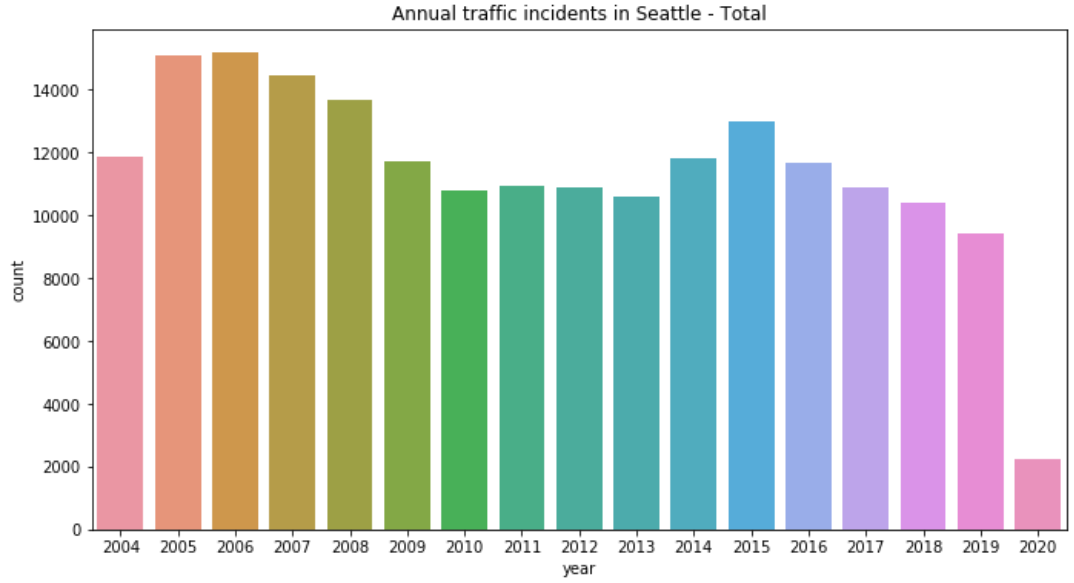
First we will need to prepare these features so it is suitable for a binary classification model. We will perform data cleaning, drop or fill in missing data and convert categorical variables to suitable format for machine learning algorithms.

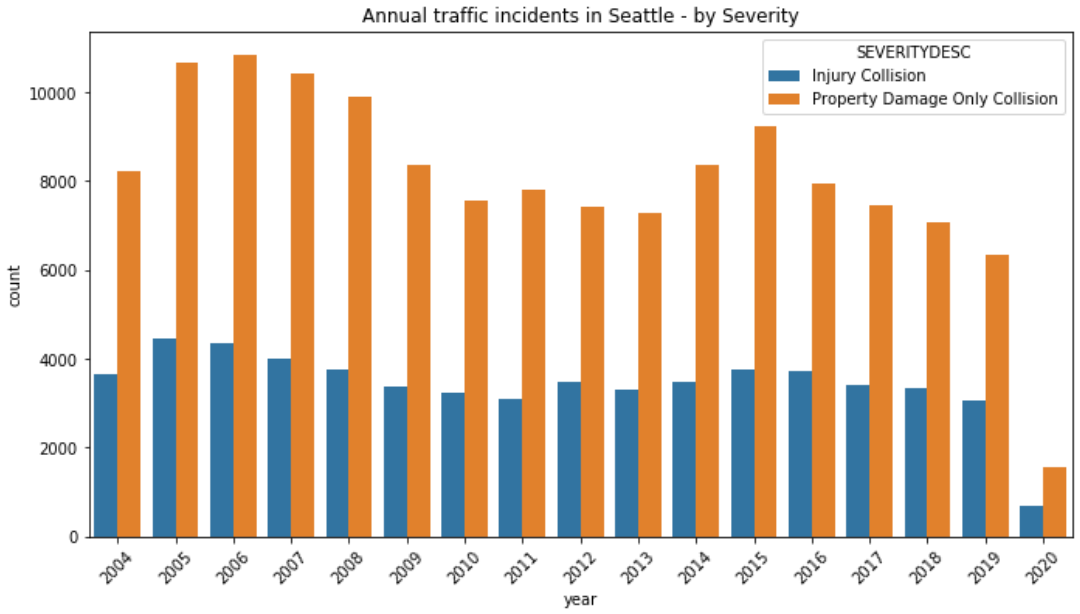
Next we will split the data for training and test groups with 80% of samples of training and 20% for testing.

Finally in our analysis will be calculation and exploration of different models to find out the main problem for severity. We will use 3 classification models which are Logistic Regression, Decision Tree and KNN. After obtaining each model's predictions we will evaluate their accuracy, precision, f1-score and compare and discuss the results.

Annual number of accidents

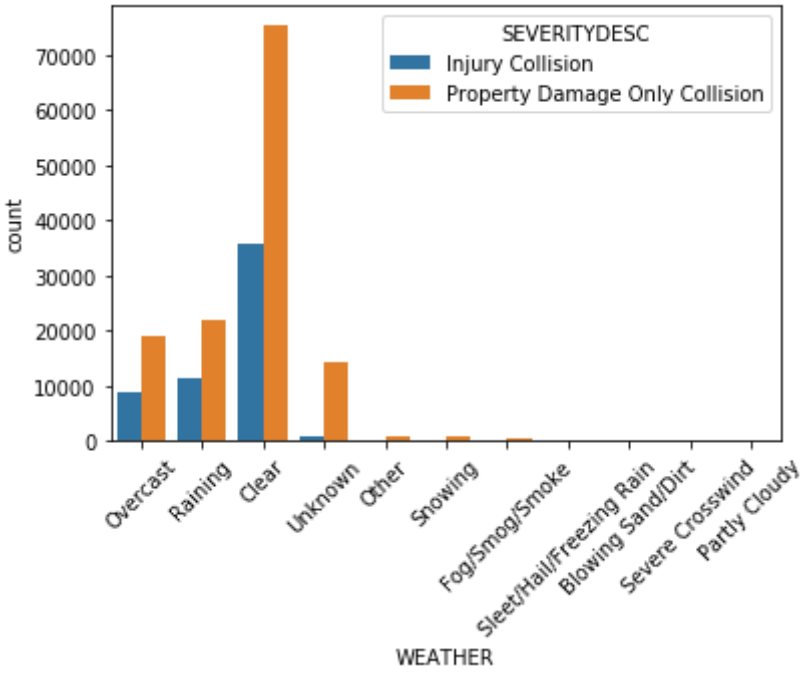
We can see that Injury collisions are less likely to happen compared to Property damage collisions. Below plot shows annual evolution of number of accidents. In the year 2020 we have data only till May, hence there is a decrease in the total.




Annual traffic incidents in Seattle - by Severity
SEVERITYDESC
Injury Collision
Property Damage Only Collision
count
year
0
2000
4000
6000
8000
10000
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020

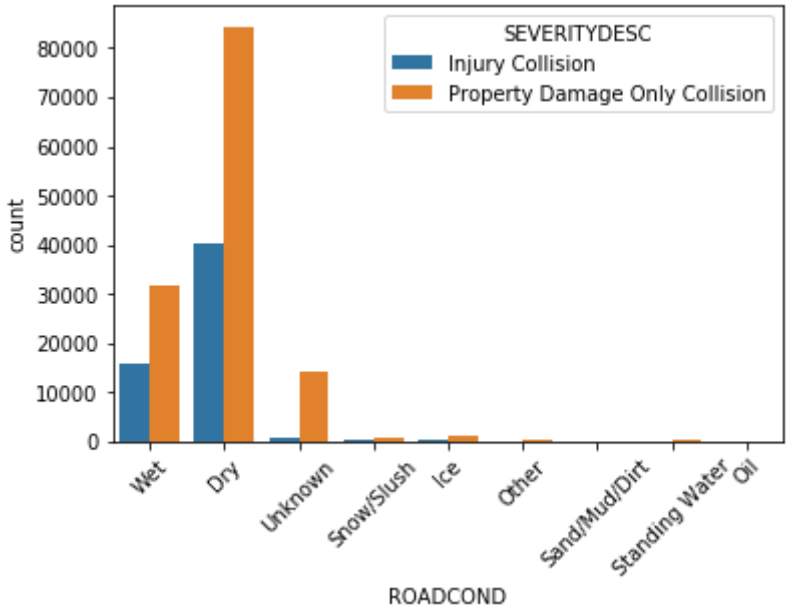
Weather Condition

Most incidents happened in a Clear weather conditions. That could be because most of the time area actually clear conditions and perhaps drivers are less careful compared to harsh conditions. Analysis of correlation between WEATHER and INATTENTIONIND (whether or not collision was due to inattention), did not confirm the fact that drivers would be less cautious during clear conditions. While we can see that more accidents with "inattention" occured during clear conditions, we can also see it is proportionally relevant to the number of accidents occuring during clear conditions.



Road Condition

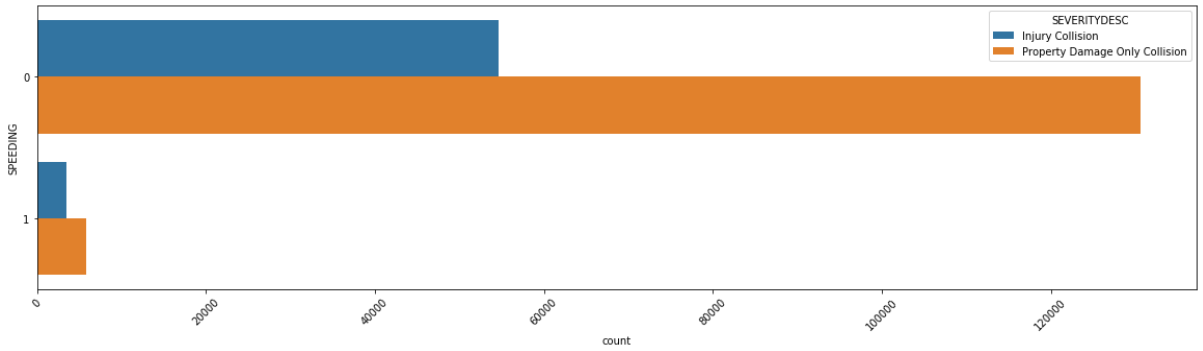
Majority of incidents happened during dry road conditions. This is because the majority of the time the condition is actually dry. We are also observing high numbers of incidents (approx 50%) with injury to happen in the dry road condition. Our hypothesis was that people would drive at higher speed in dry conditions. While this hypothesis did not prove to be true, we have noticed that the speeding variable has significant correlation with accident severity. Incidents with speeding would be more likely to be injury compared to incidents without speeding.



		Count of incidents	Count of speeding	Ration of speeding
ROADCOND	SEVERITYCODE			
Dry	1	84446	2425	0.028717
	2	40064	1797	0.044853
Ice	1	936	252	0.269231
	2	273	94	0.344322
Oil	1	40	1	0.025000
	2	24	2	0.083333
Other	1	89	17	0.191011
	2	43	8	0.186047
Sand/Mud/Dirt	1	52	6	0.115385
	2	23	6	0.260870
Snow/Slush	1	837	159	0.189964
	2	167	44	0.263473
Standing Water	1	85	30	0.352941
	2	30	10	0.333333
Unknown	1	14329	59	0.004118
	2	749	18	0.024032
Wet	1	31719	2851	0.089883
	2	15755	1551	0.098445

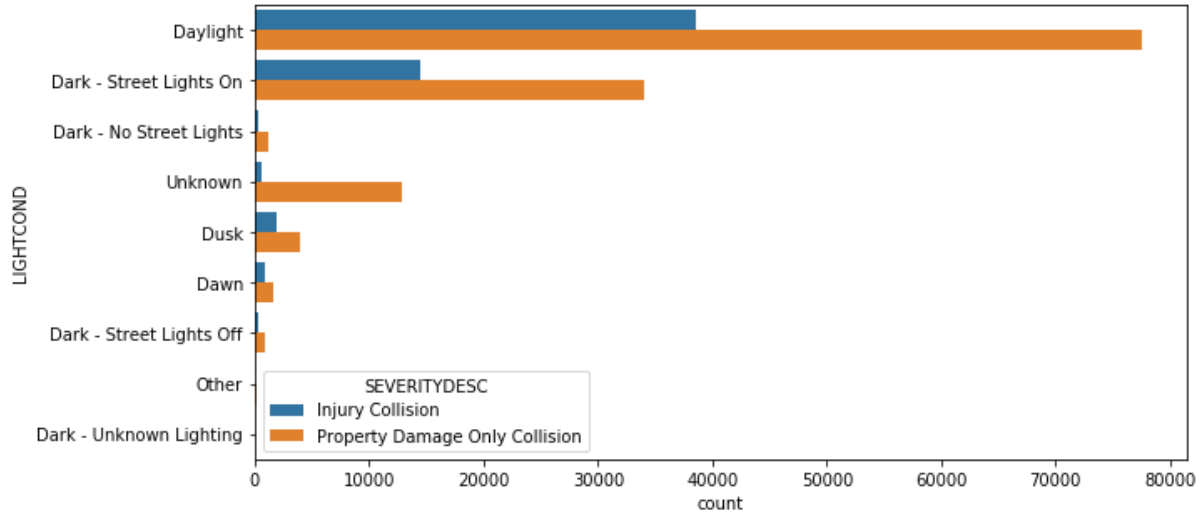
Speeding

We have observed that speeding has an influence on the accident severity. Incidents with speeding would be more likely to result in injury compared to incidents without speeding. Relatively more speeding is observed in incidents with injury.



Light Condition

Most of the accidents happened in Daylight whereas Dark-Street Lights On is also considerable for accidents.



Location

Some areas of the city tend to have traffic incidents more often. This is for example Seattle downtown area, major intersections etc. On the other hand suburbs, neighborhood with parks and other areas with less traffic will be less likely to observe accidents. Also incidents with injury seems to be more frequent in some areas.

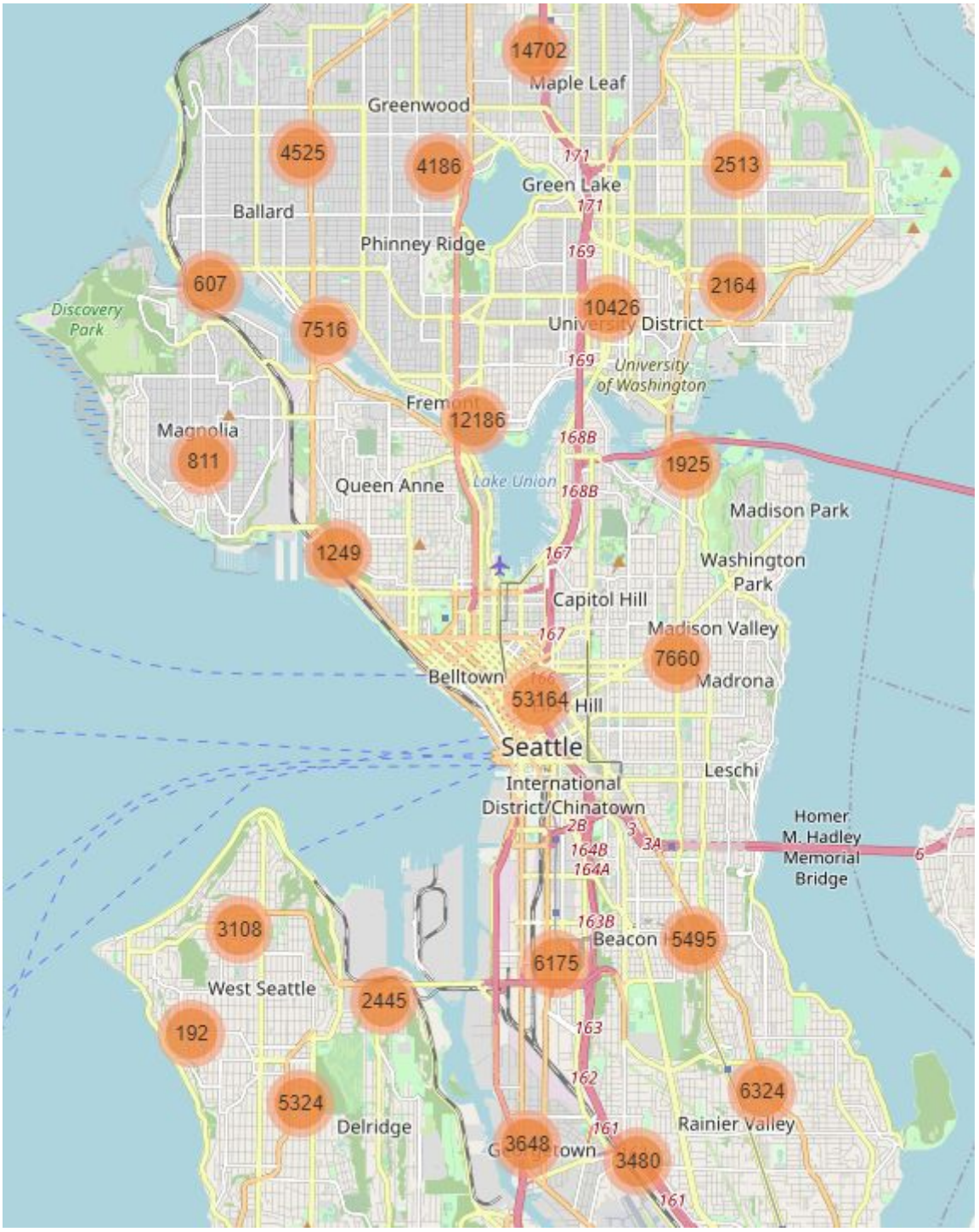


Image: Number of accidents in Seattle grouped by location

Variable selection

We will use WEATHER, ROADCOND, LIGHTCOND, SPEEDING and location (X,Y) as attributes to classify SEVERITYCODE. First we will need to prepare these features so it is suitable for a binary classification model. We will use popular machine learning algorithms like SVM, Logistic Regression, Decision Tree and KNN for build up models to analyze their performance and predict the collision severity.

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

To ensure best model settings we will try to select best parameters from for the model using Grid search, i.e. testing different parameters, like different number of K neighbors, and selecting one with highest estimate accuracy.

Results and Discussion

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.

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

Conclusion

Purpose of this project was to explore the relationship between traffic incident severity and characteristics describing the accident circumstances. We have selected 6 features from a set of 37 available that are being recorded. We have achieved over 70% accuracy. It is clear that selected features have impact on accident severity. Future possible improvement could be using additional variables to extend our model - both included in our data set and external. As an example to determine likelihood of injury, vehicle age, speed or vehicle condition might play a role.