Capstone project - Car accident severity

Martynas Brazauskas

Introduction

A description of the problem and a discussion of the background

Serious situation with COVID-19 in the world should change travel habits. More people will start to avoid public transport, because of possibility to get sick. So, some of them will start to travel to work on feet or by bike.

Another reason, which will eventually change travel habits is the fact that the world is getting greener. People understand that they has to do something to reduce co2 emission. Of course, some of them is starting to use electric car, however, some people choose to travel on foot or by bike.

A description of the data and how it will be used to solve the problem.

Based on definition of my problem, factors that will influence our decision are:

- number of collisions, which includes pedestrians and bicycles.
- additional factors may have contributed to the collisions (weather, road condition, light condition and etc.)
- seriousity of collisions.

I decided to use various Machine Learning methods to analyze the data data and find the impact to collisions associated with pedestrians and bicycles.

Following data sources will be needed to extract/generate the required information:

- ArcGIS Metadata Form of all year collisions data set from Seattle city.

Data Set Review

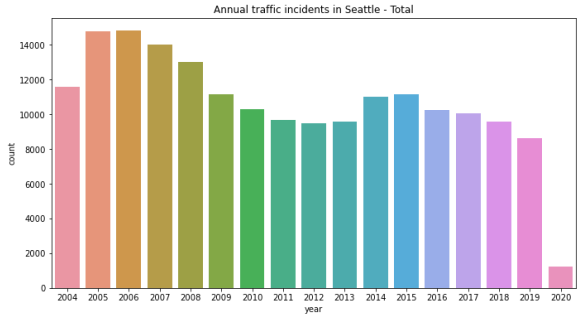
For most of columns missing data was removed. INATTENTIONIND, SPEEDING, PEDROWNOTGRNT data was updated for further research. The idea is to use this data as dummy variables. I will use following variables to classify the severity of the accidents, when bicycles and pedestrians is included:

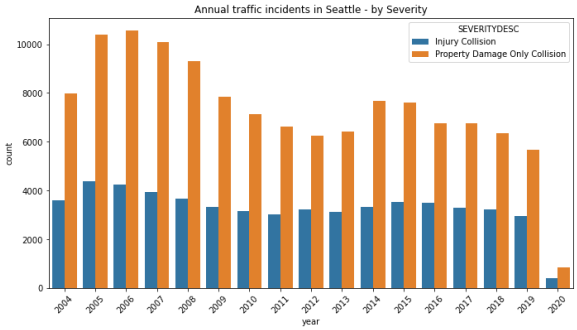
- PEDCOUNT - The number of pedestrians involved in the collision. This is entered by the state.
- PEDCYLCOUNT - The number of bicycles involved in the collision. This is entered by the state.
- INATTENTIONIND - Whether or not collision was due to inattention.
- WEATHER - A description of the weather 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.

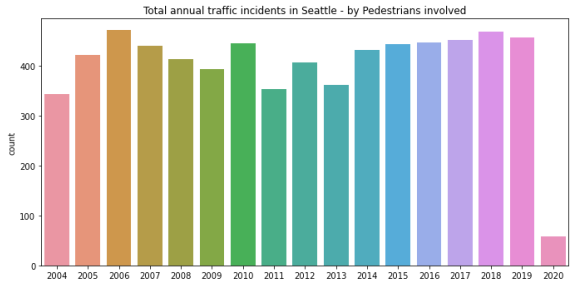
After data wrangling, data set has 182 660 samples and 38 columns. The target variable is SEVERITYCODE. This variable reflects the type of damage caused by the collision (property damage, injury).

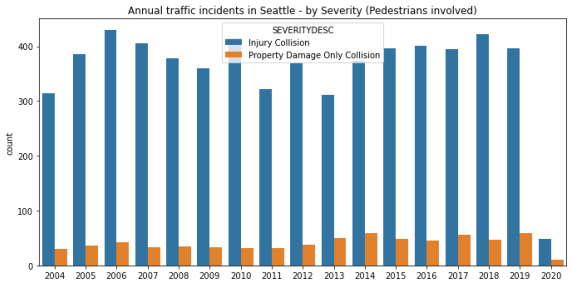
As we can see every year annual number of accidents is decreasing. Probably good decision was made to decrease accidents with car. But the number of pedestrians involved in collisions is stable over the year. It's a little disappointing number. Of course, the situation with collisions which includes bicycles is worse. According to the data, collisions which includes pedestrians and bicycles is more related with injuries than property damage.

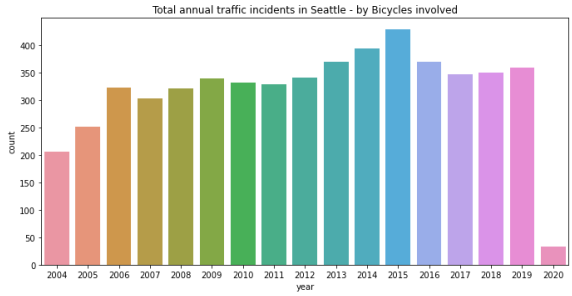
Based on these data, I can draw two conclusions. More and more people are walking and using bikes, leading to more collisions. Also, due to the growing number of road users, the safety of pedestrians and cyclists is not sufficiently ensured.

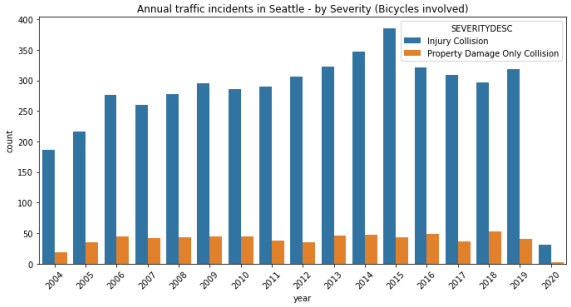










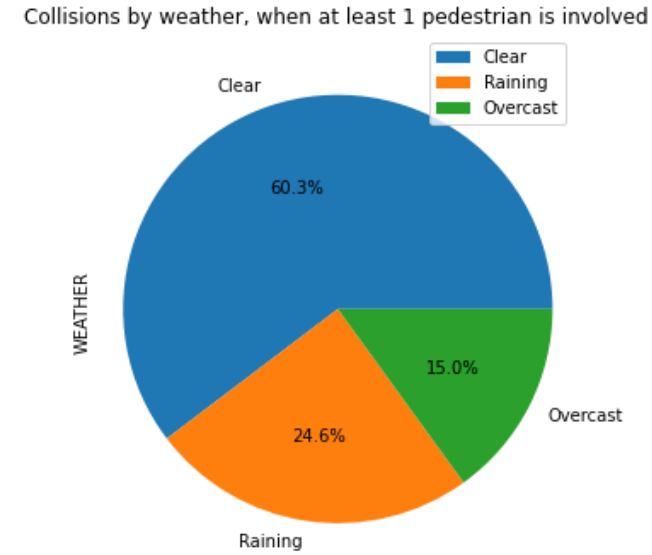


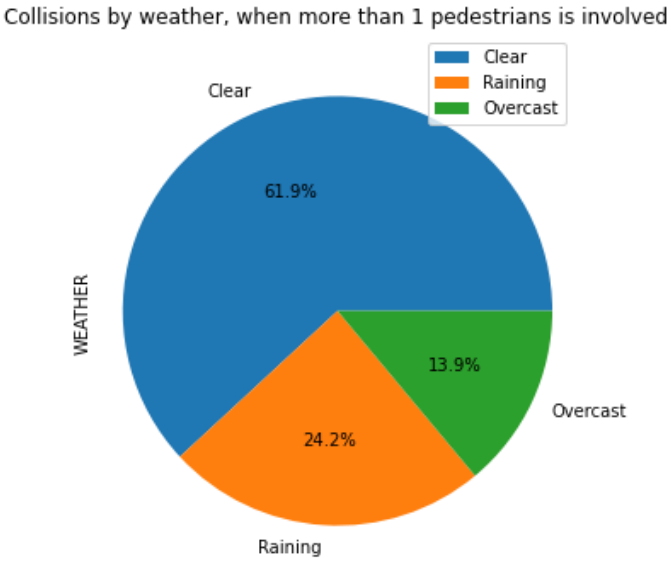
Almost 60 % of collisions occurs when the weather is clear. 17,86 % collisions occurs when raining. 14,87 % - when overcast.

Three main weather conditions had biggest influence to collisions. I will skip 'Unknown' variable, because weather conditions was unknown. Further I will analyse the influence of these three weather conditions. According to the data 60,3 % of collisions, when at least 1 pedestrian was involved, occurred when weather was clear. However, 61,3 % of collisions, when more than 1 pedestrian was involved, occurred when weather was clear.

72,5 % of collisions, when at least 1 bicycle was involved, occurred when weather was clear. Also, 75,6 % of collisions, when more than 1 bicycle was involved, occurred when weather was clear.

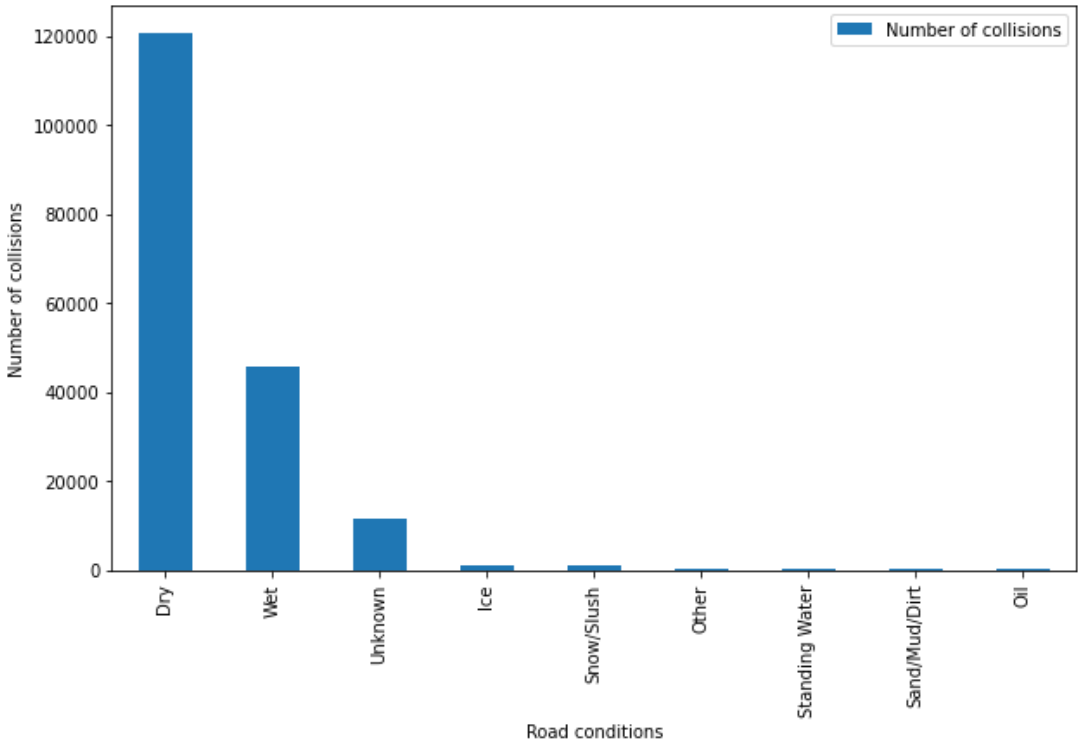
It can be said, that weather wasn't main reason of seriousness of collisions, which involves pedestrians and cyclist.



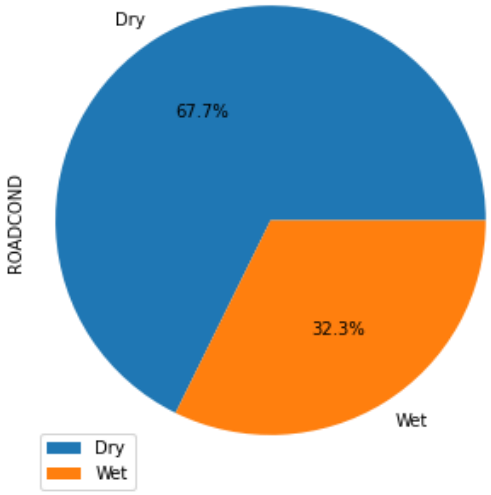


Almost 67 % of collisions occurs when the road was dry. 25,55 % collisions occurs when road was wet.

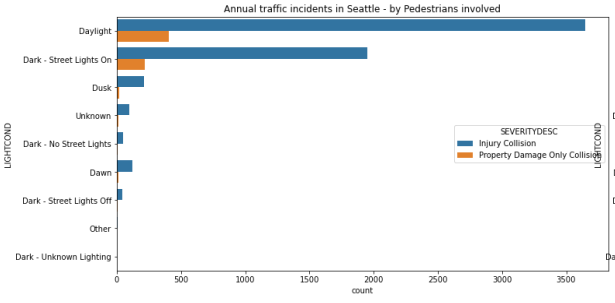
Two main road conditions had biggest influence to collisions. I will skip 'Unknown' variable, because road conditions was unknown. Further I will analyse the influence of these three weather conditions. According to the data 67,7 % of collisions, when at least 1 pedestrian was involved, and 82,5 % of collisions, when at least 1 cyclist was involved, occurred when road was dry. It can be said, that road conditions wasn't main reason of seriousness of collisions, which involves pedestrians and cyclist.

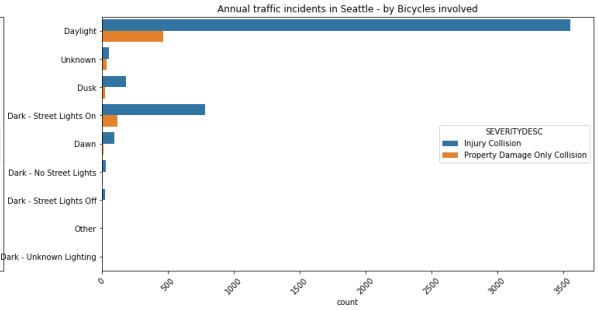


Collisions by road condition, when at least 1 pedestrian is involved



Most of accidents happens in daylight and when dark-street lights on.





Inattention is rather common factor of collisions. Inattention was one of the factors that led to the collisions. This factor related to 14,99 % collisions, when property was damaged, and to 18,21 % collisions, when someone was injured.

Speeding is common factor of collisions. Speeding was one of the factors that led to the collisions. This factor related to 4,3 % collisions, when property was damaged, and to 5,92 % collisions, when someone was injured.

Methodology

I will use WEATHER, ROADCOND, LIGHTCOND, INATTENTIONIND, SPEEDING and location (X,Y) as attributes to classify SEVERITYCODE. First, I will need to prepare these features, so it is suitable for a binary classification model. I will use all data and few machine learning algorithms like KNN, SVM, Logistic Regression and Decision Tree for build up models to analyze their performance and predict the collision severity. Later, I will use best performed model to analyse data of collisions which includes pedestrians and cyclist.

Also, the data was normalized and splitted in following ration:

- 80% to train model
- 20% to test model

Modeling

I will use the training set to build an accurate model. Then use the test set to report the accuracy of individual models. Most popular ML algorithms for data classification will be compared:

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

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

Algorithm	Jaccard	F-1 score
KNN	0.652913	0.616497
DT	0.634698	0.619784
SVM	0.688232	0.561136
LR	0.688232	0.561136

In this analysis was 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, speeding, inattention and location. Two models (Logistic regression, Support Vector Machine) performed very similarly. K-Nearest Neighbours stood out with a slightly higher F-1 score, but lower Jaccard index. SVM and Logistic regression we were able to meet 68,8% accuracy (measured by Jaccard index). But these two algorithms is characterized by lowest accuracy in estimating it by F-1 score. According to average of accuracy measures, best

performance was reached with KNN algorithm. I will test this algorithm with pedestrians collision data and cyclist collision data.

KNN	Jaccard	F-1 score
Pedestrians	0.895007	0.845420
Cyclist	0.871058	0.816186

Results and discussion

According to this research, created model was even better for classification of collisions severity, when pedestrians and cyclist is included. I have reached almost 90 % accuracy for pedestrians involved collisions and 87 % accuracy for bicycles involved collisions. These results are important for the safety of pedestrians and cyclists in big cities.

Conclusion

As a result, more often people are using public transport, bicycles and travel on foot in big cities. For this reason, we need better understanding of collisions, which includes pedestrians and cyclist. Historically, data shows that total situation of collisions is getting better. But situation of collisions when pedestrians or bicycles is involved doesn't change over the year. Research results shows that weather, road condition, light condition, inattention, speeding and location factors is very important to classify collisions severity. Understandably, more detailed analysis and more detailed data are needed to achieve better results.