Final Report

Introduction/Business Problem

In today's world many people can't imagine their lives without a car. Cars have become an integral part of modern humans, making it easier not only to perform everyday activities but also to travel long distance. Nonetheless, road vehicles are also the most unsafe of all available to humans. Traffic collisions are in first place in terms of the number of deaths and injuries. According to these parameters, cars significantly overtake railway, aviation and water transport.

Unfortunately, car accidents occur for many reasons, including both technological and human factors. An accident can happen due to the fault of a tired driver, due to icing of the road surface or a malfunction of the brake system. However, the risk of getting into an accident is often influenced by external factors, such as the day of the week, weather conditions and the quality of the road itself.

Many car companies are working on fully autonomous cars, which could potentially reduce the amount of car accidents. Until then regular cars will remain popular. Therefore, it would be great to have a mechanism that could warn you, given the weather and the road conditions, about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to.

Thus, **the goal of this project** is:

1. to identify and analyze the factors that cause traffic collisions, and
2. to create a model that will predict the severity of car accidents.

The findings may be useful for improving road safety or for insurance companies planning to introduce life and health insurance programs for drivers and passengers.

Data

For this project I will be using a shared dataframe about accident severity. The dataset includes all types of collisions in Seattle for the timeframe from 2004 to mid-2020. The dataset has an extensive amount of observations - it includes 194,673 samples. The total amount of attributes in the dataframe is 37. It is important identify relevant features in the dataset as well as to clean the data before building the model. The final size of the dataframe includes 187,524 entries and 9 attributes

Methodology

As mentioned above, not all attributes will be useful for this project. Thus, I will drop the columns that have no particular use for this project such as specialized codes and keys, GPS coordinates, etc.

Further data analysis will be structured as follows:

Firstly, I will visually illustrate the main characteristics of the dataset using the following attributes:

- 'SEVERITYDESC' - A detailed description of the severity of the collision;
- 'ADDRTYPE' - Collision address type (Alley, Block, Intersection);
- 'PERSONCOUNT' - The number of pedestrians involved in the collision;
- 'VEHCOUNT' - The number of vehicles involved in the collision;
- 'INCDATE' - The date of the incident;

- '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;

Secondly, I will create a model that will predict the severity of car accidents. In this project, 'SEVERITYCODE' is the target variable, which is used to measure the severity of an accident from 0 to 4 within the dataset.

In order to classify the target variable I will use 3 independent variables:

- '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.

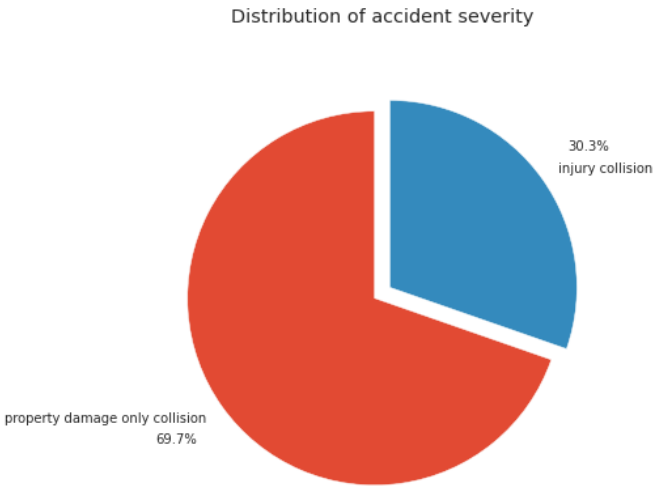
Note: initially the attribute “SPEEDING” was also included in the selection. However, it has more than 185,000 missing values, which can create a biased model. Therefore, this attribute was eventually dropped.

In this project I will use 3 algorithms, Logistic Regression, K-Nearest Neighbors and Decision Tree in order to verify accuracy of the model.

Data Analysis

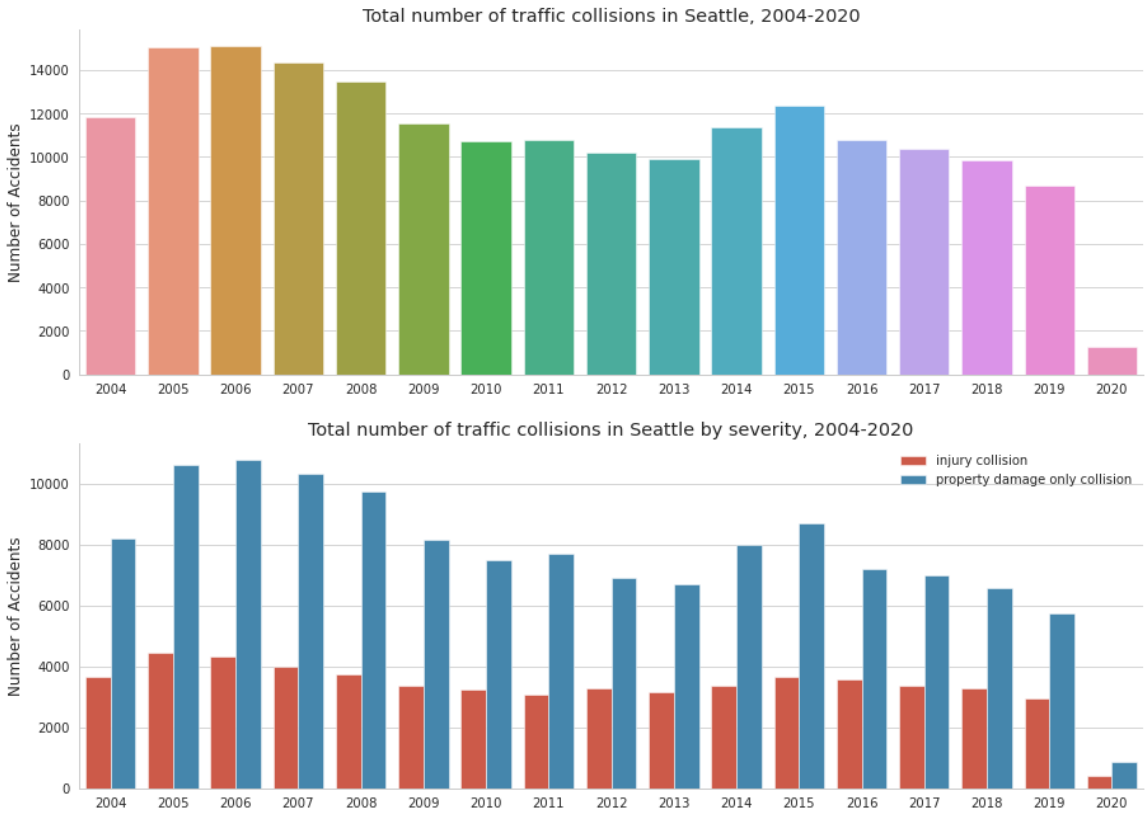
I began analyzing data by looking at several features:

- **Accident severity distribution**



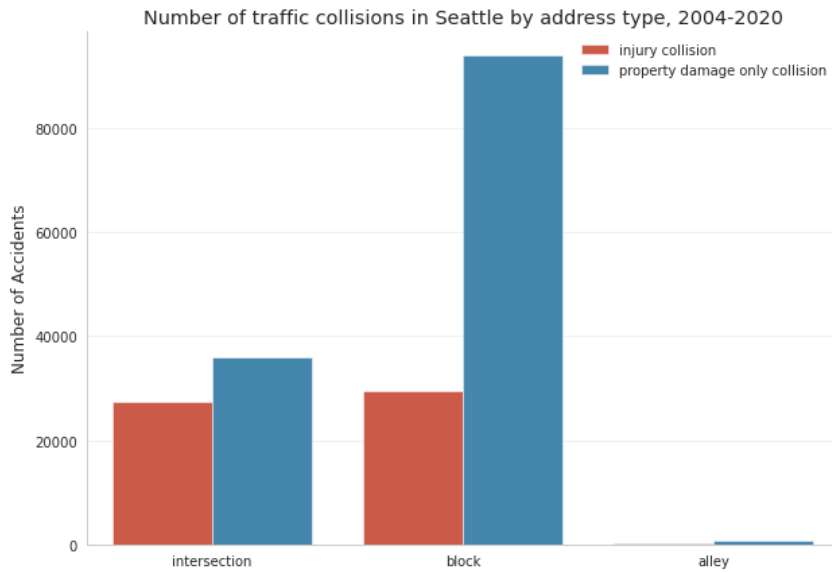
According to our dataset, almost 70% of accidents resulted in property damage (130,642 cases) while nearly 30% involved injuries (56,883 cases).

- **Annual amount of traffic accidents**



The plots above illustrate annual number of car accidents in Seattle between 2004 and mid-2020. According to the first histogram, the number of collisions yearly remains significant. The most amount accidents happened in 2006, while the least amount happened in 2013. Overall, we can see that there is a gradual decline in car accidents after years 2006 and 2015. The second graph shows that the amount of injury collisions is always lower than collisions with property damage.

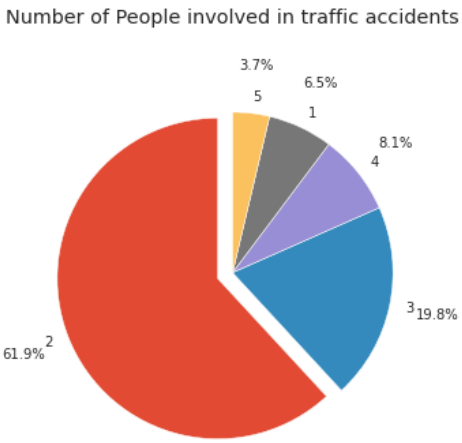
- **Address type of traffic accidents**

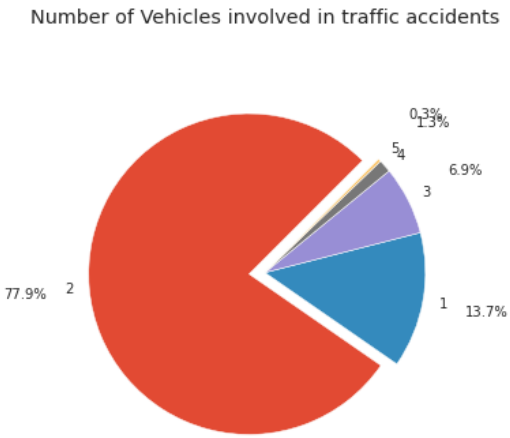


The plot above shows the number of traffic collisions in Seattle by address type. As we can see, most of collisions happened either at the block (123,321 cases) or at the intersection (63,462 cases). The least amount of accidents happened at the alley (742 cases). While the ratio of the two types of

collisions is nearly similar, the amount of property damage collisions at the block exceeds almost 3 times the amount of injury collision.

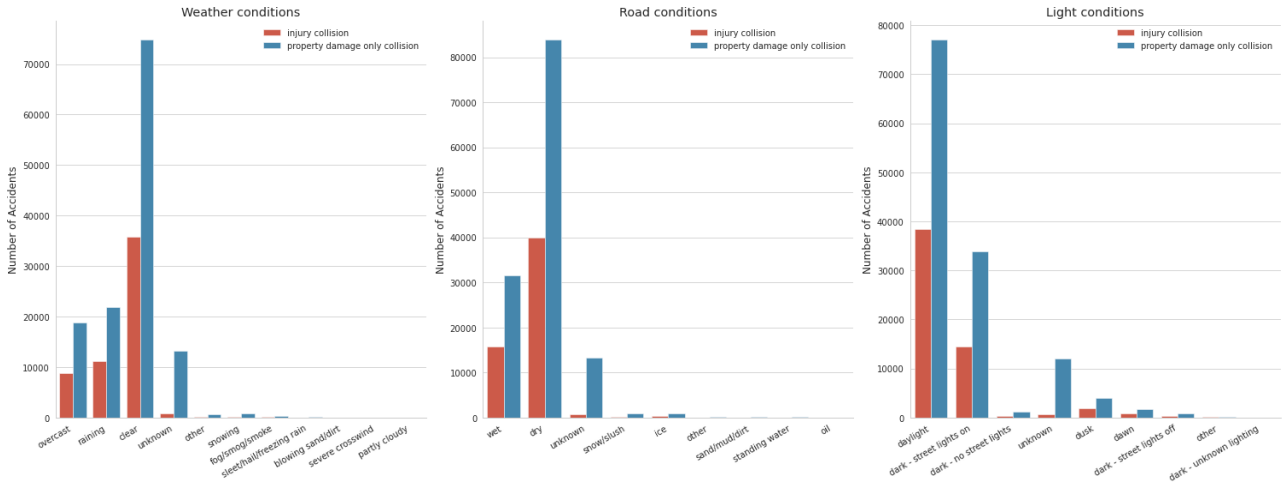
- **Number of People and Vehicles involved in traffic accidents**





The pie charts above show the number of People and Vehicles involved in traffic accidents in Seattle. According to the chart on the left, top 5 most common numbers of people involved in the collision is in a range between 1 and 5. In almost 62% of cases 2 people get in a car accident, followed by 3 people in ~20% cases. One person gets in a collision in 6.5% of cases. The chart on the right illustrates that top 5 most common numbers of vehicles involved in the collision is in a range between 1 and 5. In nearly 78% 2 cars are affected and in approximately 14% 1 car is affected.

- **Weather, Road and Light conditions and Accident Severity**



The three graphs above describe the accident severity based on weather, road and light conditions. In the 1st histogram we can see that the most accidents happened during clear weather, followed by rainy and overcast weather. In the 2nd histogram we notice that when collisions happened the road condition was dry in most cases or wet. The 3rd histogram shows that on average traffic accidents happen during the daylight followed by darker time of the day with streetlights on.

Results and Discussion

In order to create a model for predicting car accident severity I chose “SEVERITYCODE” as a dependent variable and “WEATHER”, “ROADCOND” and “LIGHTCOND” as independent

variables. I have created a separate dataset with the variables and converted categorical variables to binary variables. After that I used the Train/Test Split to split the dataset into training and testing sets respectively, which are mutually exclusive. The ratio for Train set 70% and for Test set is 30%.

To verify accuracy of the model I used 3 algorithms: Logistic Regression, K-Nearest Neighbors and Decision Tree. Summary of the results in presented below.

	Algorithm	Jaccard	F1-score	Precision
0	Logistic Regression	0.6979	0.5737	0.4871
1	KNN	0.6673	0.6051	0.5958
2	Decision Tree	0.6979	0.5737	0.4871

Among the three algorithms, Jaccard's score varies between 66.7% and 69.7%. F1-score is between 57.3% and 60.5%. Precision is between 48.7% and 59.5%.

Conclusion

This project was dedicated to the analysis of accidents severity in Seattle from 2004 until mid-2020. After data selection and cleaning, I have identified that almost 70% of accidents resulted in property damage while other 30% involved injuries. Over the years the number of collisions remains significant. Most of collisions happened either at the block or at the intersection. In almost 62% of cases 2 people get in a car accident, followed by 3 people in ~20% cases. One person gets in a collision in 6.5% of cases. In nearly 78% 2 cars are affected by the collision. Traffic accidents usually happen during daytime with clear weather and dry road condition.

In order to identify the relationship between accident severity and accident features, I have created a model and tested it using Logistic Regression, K-Nearest Neighbors and Decision Tree algorithms. All three algorithms showed similar results, while Logistic regression and Decision Tree showed better results in evaluating model accuracy. While I have chosen only 3 features for the model, it was able to achieve ~70% accuracy.