Applied Data Science Capstone

Predicting Traffic Accident Severity

Marius Stolz

October 2020

1 Introduction

1.1 Background

Nowadays, the automobile is one of the most frequently used method of transportation in many countries and car collisions occur everyday. Those collisions lead to different accident severities such as human fatality, injuries and property damage. To prevent car accidents and to improve people's life, the street network and its traffic has to be analyzed. A machine learning prediction model could predict the severity of a collision based on input data attributes such as weather conditions, junction types, time, etc. A trained, tested and evaluated model would help understanding the main causes of car accidents and support the decision-making of the architecture of new street networks.

1.2 Problem

Input data from reported car accidents in Seattle are to be used to develop a machine learning prediction model to predict accident *Severity*. Under severity a distinction is made between *Injuries* and *Property Damage*.

1.3 Interest

Obviously, big cities and places with many accidents are interested in predicting accident severity of reported crashes with unknown severity or of crashes that are eventually occur in the future. This may help to counteract against human and economic loss. An interest of navigation system developers is also conceivable. They could implement this prediction model to propose safer routes to the driver based on real-time data.

2 Data acquisition and cleaning

2.1 Data sources

The input data that will be used can be found and downloaded here or by going on the URL: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv. This data is the example data of the Coursera Course *Applied Data Science Capstone* and it is originally provided by the Seattle Police Department and recorded by Traffic Records. Its title is *Collisions - All Years*, as shown in the metadata file. This metadata file explains all attributes of the collisions recorded from 2004 to 2020. The file can be downloaded here or by visiting the URL: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

2.2 Data cleaning

The first reading in of the raw input data contains 194673 rows and 38 columns including the target label *SEVERITYCODE*. There are columns storing the X and Y coordinates, the category of junction, the weather conditions, time and various other information to describe the collisions. There is numerical data, categorical data and missing values. To analyze the data and develop a good machine learning prediction model, data cleaning is the first step to do. The cleaner the data, the fancier the algorithms. Without data cleaning predictions are most likely inaccurate, according to the saying ”*garbage in, garbage out*”. Therefore, missing values have to be handled by removing or replacing them, structural errors should be fixed, redundant and irrelevant information should be dropped. The following list shows the columns with missing values, their amount and the chosen cleaning type with a given reason:

Attribute	**Missing values**	**Missing values in %**	**Cleaning Type, Reason**
X	5334	0	drop column after visualisation purpose, irrelevant
Y	5334	0	drop column after visualisation purpose, irrelevant
ADDRTYPE	1926	1.0	replace with most frequent, few missing values
INTKEY	129603	66.6	drop column, too many missing values
LOCATION	2677	1.4	drop column, redundant
EXCEPTRSNCODE	109862	56.4	drop column, too many missing values
EXCEPTRSNDESC	189035	97.1	drop column, too many missing values
COLLISIONTYPE	4904	2.5	replace with most frequent, few missing values
JUNCTIONTYPE	6329	3.5	replace with most frequent, few missing values
INATTENTIONIND	164868	84.7	drop column, too many missing values
UNDERINFL	4884	2.5	replace with most frequent, few missing values
WEATHER	5081	2.6	replace with most frequent, few missing values
ROADCOND	5012	2.6	replace with most frequent, few missing values
LIGHTCOND	5170	2.6	replace with most frequent, few missing values
PEDROWNOTGRNT	190006	97.6	drop column, too many missing values

SPEEDING	185340	95.2	drop column, too many missing values
SDOTCOLNUM	79737	40.9	drop column, too many missing values
ST_COLCODE	18	0	drop column, redundant
ST_COLDESC	4904	0	drop column, irrelevant

The column *LOCATION* stores coordinates as in X and Y and is therefore redundant. The columns with more than 40% missing values were dropped also because there are too many missing values. The column *ST_COLCODE* is redundant to the column named *SDOTCOLCODE* (which is under the remaining columns with no missing values) and will be dropped together with its description column *ST_COLDESC*. The columns with few missing values contain categorical data and entries will be filled with most frequent values. All columns that are going to be dropped completely were colored in red.

Since the missing values were being handled, the data could still contain redundant, irrelevant information and structural errors. The next column to be dropped is *SEVERITYCODE.1* because the column stores exactly the same data as the target label. By running the methods *unique()* it turned out that the columns *OBJECTID*, *INCKEY*, *COLDETKEY* and *REPORTNO* have as many unique entries as rows. Since those entries contain keys and codes, they all need to be decoded which needs too much time. Now the entries are seen as categories and in this case they dont contribute to any more understanding and will be dropped. Although *SDOT_COLCODE*, *SEGLANEKEY* and *CROSSWALKKEY* contain fewer unique codes, they will be dropped, because their investigation is too time consumptious and also the decoding description for *SEGLANEKEY* and *CROSSWALKKEY* is missing in the metadata file. Speaking of descriptions, they are also included in the dataframe and since many keys and codes have been removed they are of no use, so *SDOT_COLDESC* and *SEVERITYDESC* will be dropped. Also *STATUS* will be dropped, since there is no explanation for its meaning in the metadata file. Furthermore, by using the method *value_counts()* on all remaining columns with categorical entries, structural errors (like categories/bins named "Unknown") in the columns *JUNCTIONTYPE*, *UNDERINFL*, *WEATHER*, *ROADCOND* and *LIGHTCOND* have been found and then fixed by replacing the most frequent value. There is also numerical data such as *INCDATE* and *INCDTTM*. The column *INCDATE* stores the date of the collision like *INCDTTM*, therefore it is redundant information and has been dropped. Because *INCDTTM* includes the collision time too accurately, the new columns *YEAR* and *DAYTIME* storing the year and the daytime were being created and *INCDTTM* will be dropped. The column *YEAR* will only be used for visualisation purposes.

The following columns are now representing the clean data and the features to develop a machine learning prediction model:

- SEVERITYCODE
- ADDRTYPE
- COLLISIONTYPE
- PERSONCOUNT
- PEDCOUNT
- PEDCYLCOUNT
- VEHCOUNT
- JUNCTIONTYPE
- UNDERINFL
- WEATHER
- ROADCOND
- LIGHTCOND
- HITPARKEDCAR
- DAYTIME

3 Methodology - Exploratory Data Analysis

The goal is to develop a machine learning prediction model to predict accident severity. Data has been read in from a source containing information about collisions in Seattle. This data has been cleaned of missing values, redundant and irrelevant information and structural errors. Before developing the model, the data will be explored and analyzed more by visualization. The idea is to explore the features in terms of their influence on accident severity. Some conditions maybe tend to result in more injuries or property damage. This analysis leads to a better understanding of the data and to a easier assessment of the model's quality.

3.1 Number of collisions and their severity

Before we deal with the features, we want to gain a short overview. At first, the collisions will be plotted over the years and the proportions between their severity will be visualized. As shown in Fig. 1 the collisions took place between the years 2004 and 2020. In 2006 were the most collisions and in 2020 the lowest, probably because 2020 is still an ongoing year. It also can be seen that there are many more collisions reported with property damage than with injuries, which can be supported by the bar plot in Fig. 2 where *Severitycode* 1 stands for *Property Damage* and 2 for *Injuries*.

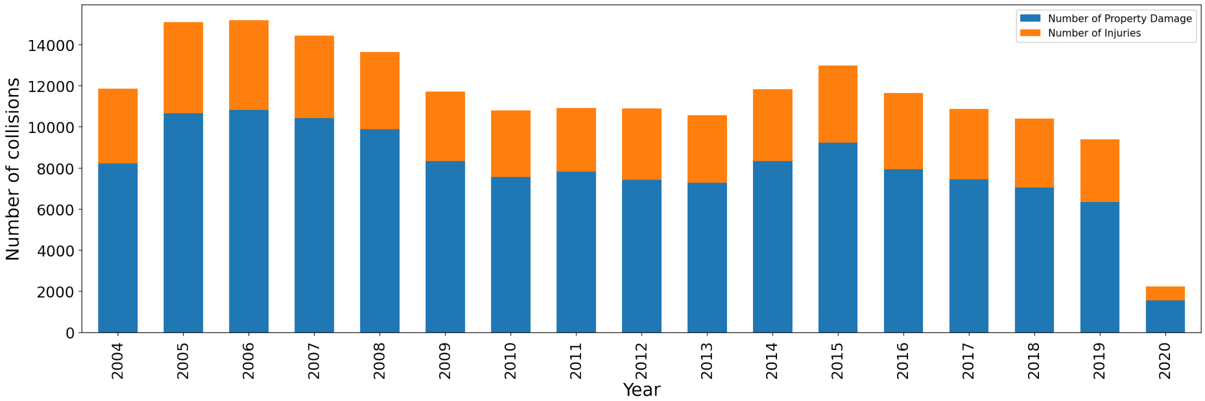


Figure 1: Distribution of Collisions from 2004 to 2020.

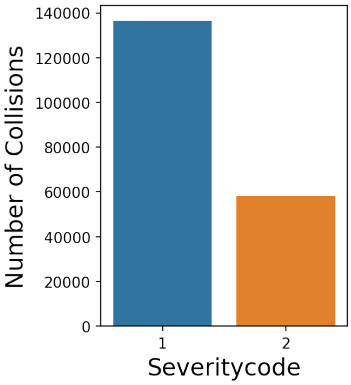
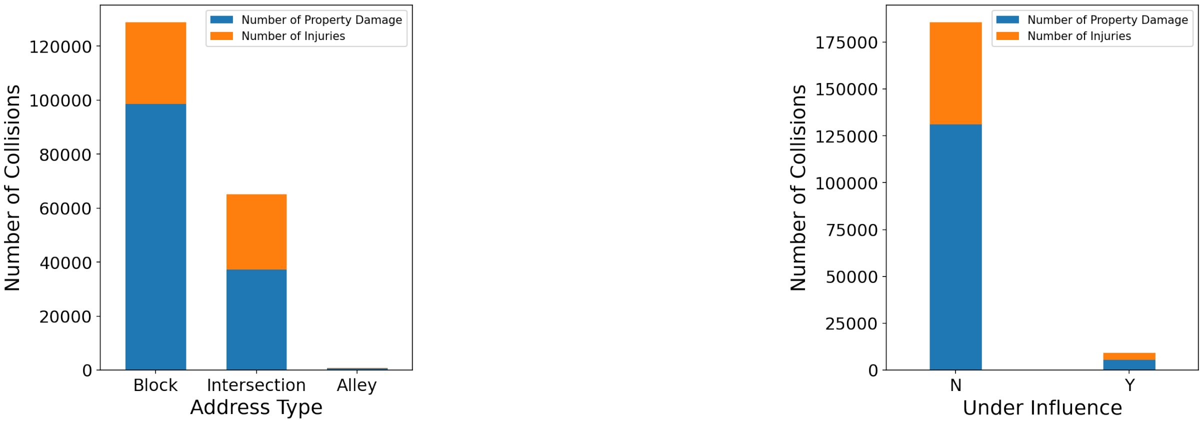


Figure 2: Distribution of severity of all recorded collisions.

3.2 Number of collision under other dependencies

In the following plots, the distribution of the value count of the remaining features are shown. In fig. 3a can be seen that most of the collisions occurred in a block and approximately half of that at an intersection. In terms of severity they have nearly the same amount of injuries. It appears, that intersections compared to blocks lead more likely to injuries and less to property damage.
Looking at the distribution in fig. 3b shows that most of the collisions are under no influence of drugs/alcohol and only very few are. In both cases, the number of property damage is dominant.

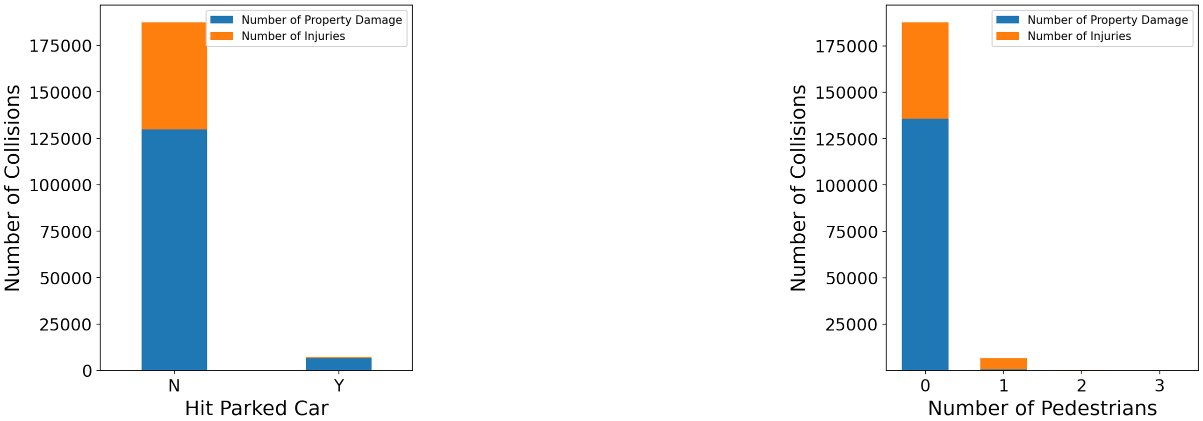


(a) Number of Collisions vs. Address Type.

(b) Number of Collisions under influence of drugs/alcohol.

Figure 3: Collisions with a) different address type and b) under drug/alcohol influence.

In fig. 4a the amount of collisions with parked cars involved is pictured. It shows, that only few collisions occurred by hitting a parked car which mostly ended in property damage. In fig. 4b you can see the distribution of the number of pedestrians. The majority of the collisions took place without pedestrians.



(a) Number of Collisions hitting a parked car.

(b) Number of Collisions and Number of Pedestrians involved.

Figure 4: Collisions with a) parked cars and b) pedestrians involved.

The number of bicycles involved in car accidents are shown in fig 5. Also here, the majority of the recorded collision were without bicyclist. But it appears, that in both cases, the severity is mostly injuries if there is one bicycle or pedestrian involved.

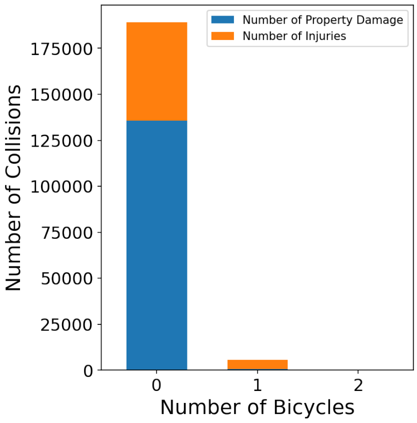
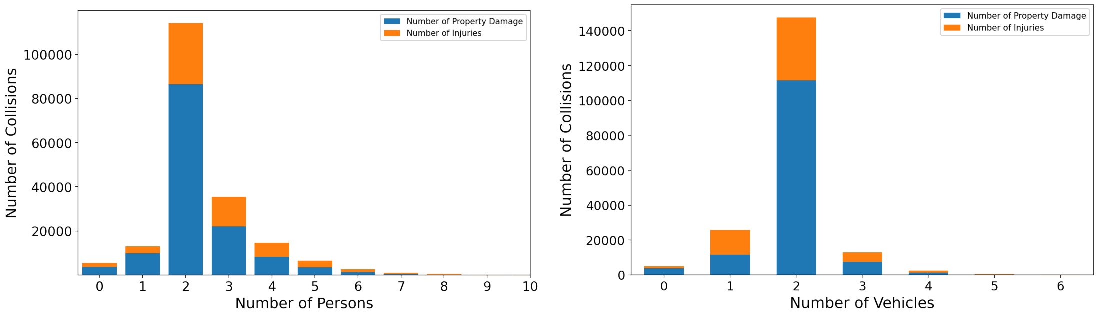


Figure 5: Number of collisions vs. number of bicycles involved.

In fig 6 the distribution of the number of persons and the number of vehicles involved can be seen. Most of the collisions happened with two persons and two cars and with property damage.



(a) Distribution of number of persons.

(b) Distribution of number of vehicles.

Figure 6: Distribution of a) number of persons and b) vehicles involved.

In the following, the distribution of collision types are presented in fig 7. Most collisions are recorded with parked cars, which mostly lead to property damage. Angles and rear ended collisions have significantly more injuries than any other collision types. The number of collisions with parked cars is much higher here than in figure 4a, because one car can hit many parked cars.

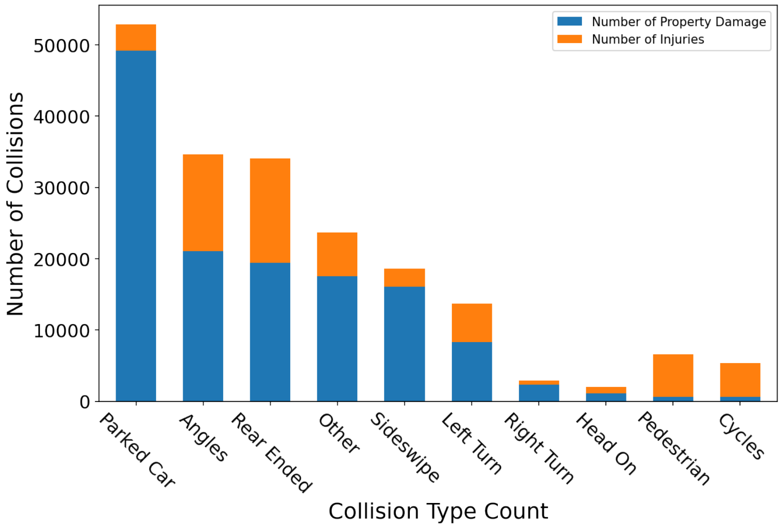


Figure 7: Distribution by type of collision.

Figure 8 shows the distribution of junction types in which the accidents occurred. At mid-blocks (not related to intersections) are the most recorded collisions in Seattle, followed by the number of collisions at intersections, which have a larger proportion of injuries than at mid-blocks, which supports the hypothesis that collisions at intersections are more dangerous.

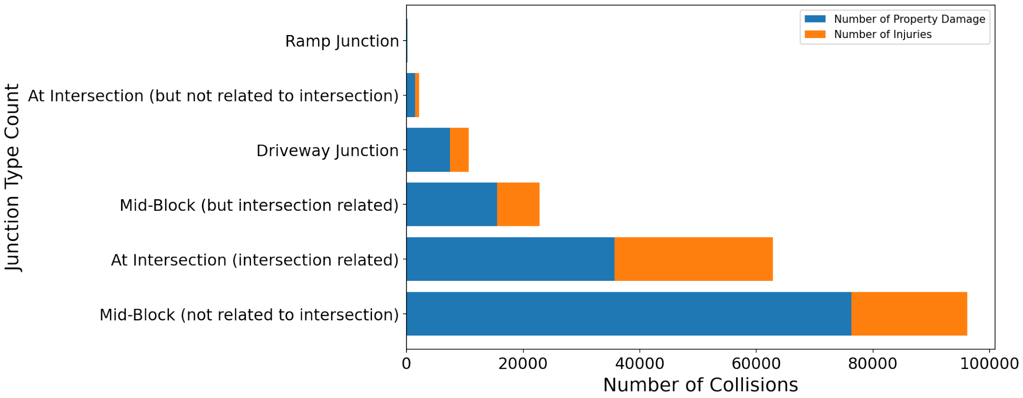


Figure 8: Distribution by type of junction.

Next up are the lighting conditions, which are displayed in fig. 9. With light by the day and with street lights on, most collisions took place.

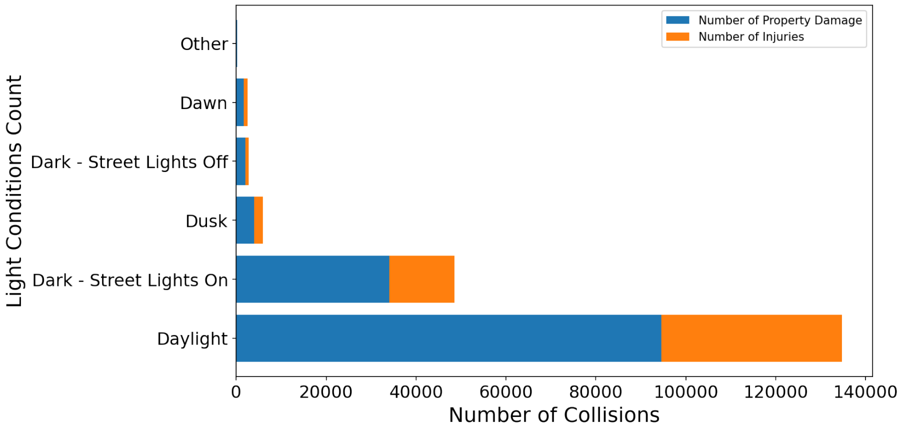


Figure 9: Distribution by light condition.

An interesting fact, most collision occured on clear weather conditions as can be seen in fig. 10.

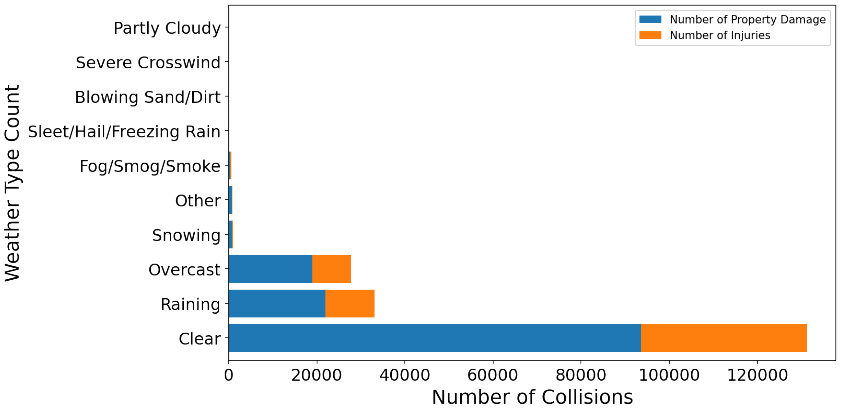


Figure 10: Distribution by weather condition.

The majority of collisions occured at dry road condition, but also wet roads show a significant amount of accidents as shown in fig 11.

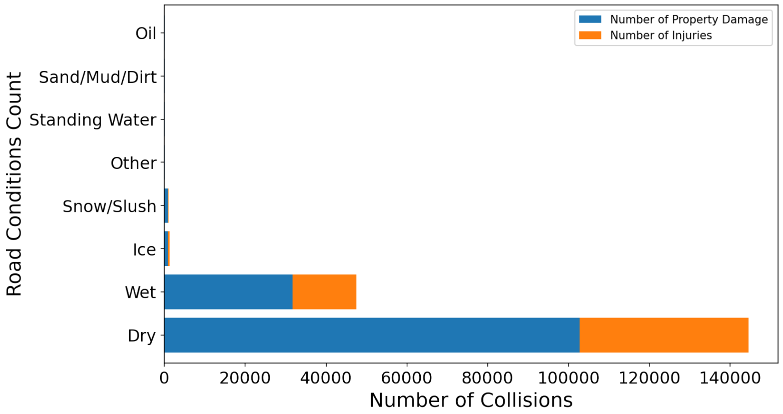


Figure 11: Distribution by road condition.

Analyzing the daytime in fig. 12, it seems that most of the collisions take place during noon and late night.

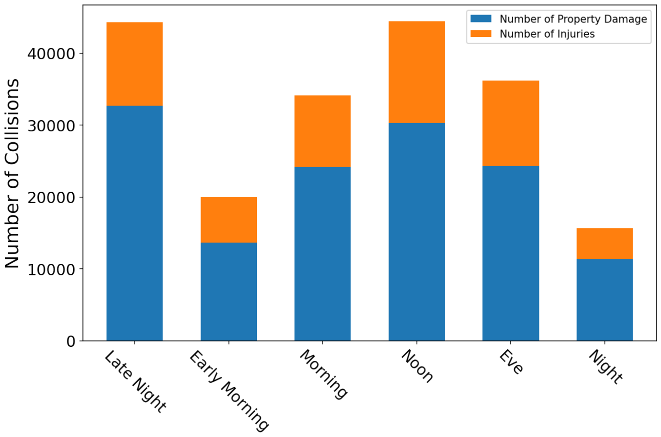


Figure 12: Distribution by daytime.

Using the coordinates given by X and Y, the first 1000 collision of the dataset are visualized in fig. 13. It can be seen that most of the collisions took place in the city center of Seattle.

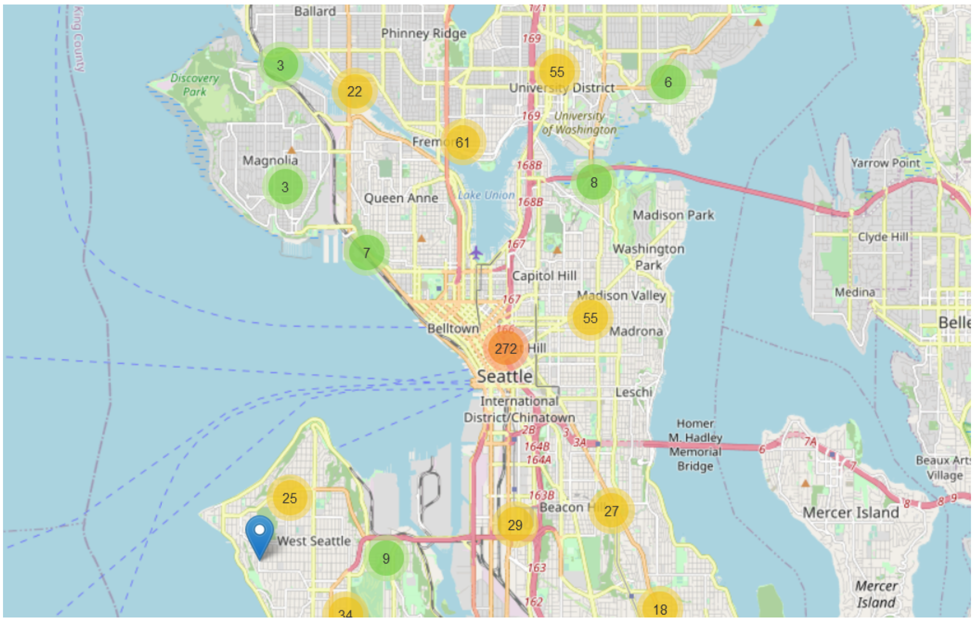


Figure 13: Map showing places of 1000 collisions of dataset.

In summary it can be said that most collisions occurred under the following circumstances: at a block or intersection, under no influence of drugs/alcohol, without any pedestrians and bicycles, with two persons and two vehicles, colliding at angles or rear ended, at daylight, clear weather, dry road, at late night or noon, in the city centre, in a mid block.

4 Methodology - Model Development

4.1 Data preparation

The target label *SEVERITYCODE* is imbalanced, because there are more collisions with the severity of property damage than with injuries, as shown in fig 2. An imbalanced dataset would skew the prediction accuracy of the model. There are two options to deal with this: either up-sampling the minority-class by randomly duplicating observations or down-sampling the majority-class by removing observations. In this study, the dataset has been down-sampled with the *resample* module of *Scikit-Learn*. As next step, the feature columns with categorical data have to be encoded by using *LabelEncoder* to convert it to numerical data. After scaling the independent variables/features (predictors) with *StandardScaler*, the dataset has been splittet into training and testing data to begin with the prediction model development.

4.2 Machine Learning Models

In this study, three different machine learning models were chosen: Logistic Regression, Decision Tree and Random Forest. Logistic Regression is a typical approach to predict the probability of binary classification problems. Decision Trees are popular in their intelligibility and simplicity, because they can visualize the decision-making and their probability. However, after hypertuning parameters the decision tree becomes very complex which could cause over-fitting. That's when Random Forest are a good choice. Because Decision Trees tend to overfitting, Random Forests are used to correct that.
After training and testing the models, metrics are used for evaluation. The Classification Accuracy (accuracy), Area Under Curve (AUROC) and F1-Score, the harmonic mean between Precision and Recall, will be calculated. Furthermore, the feature importance has been visualized using integrated functions of the machine learning models.

5 Results

In the following, a table has been created to list the scores that are evaluating the performances of the different models. The greater the scores, the better the model. It turns out that Decision Tree has got the best scores towards the other models.

Algorithmus	**Jaccard**	**Accuracy**	**F1-Score**	**Precision**	**Recall**	**AUROC**
Logistic Regression	0.51	0.66	0.68	0.65	0.71	0.66
Decision Tree	0.51	0.71	0.68	0.76	0.61	0.71
Random Forest	0.51	0.69	0.67	0.73	0.62	0.69

The models used in this study also have techniques implemented that are able to score the input features regarding to their influence of the severity of a collision. For future studies, this can help to reduce the dimensions of the dataset and to improve the prediction algorithm. In fig. 14 the feature importance scored by each model is shown. In the Logistic Regression model scoring, the coefficients are both negative and positive, where negative scores indicate a feature that predicts class 0 (”property damage”) and positive scores indicate a feature with class 1 ("injuries"). The other models just score the feature importance without distinguishing between the two classes. It turns out that many attributes have very little impact on the dependent variable. The results dont show any clear influence and also dont agree with one another.

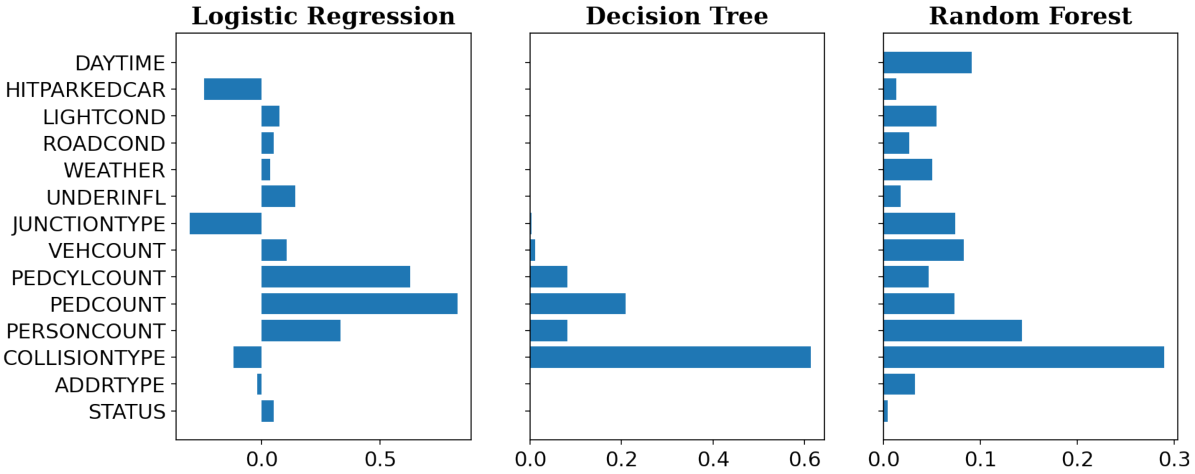


Figure 14: Feature importances by using different machine-learning models.

6 Discussion

A dataset of car collisions in Seattle has been downloaded, read in and cleaned of missing values, structural errors and redundant information. The target label *SEVERITYCODE* has been identified and 14 under 37 features were selected to conduct an exploratory data analysis where the features were visualized. It turned out, that the dataset was imbalanced due to approximately 140.000 collisions with the outcome of property damage and circa 60.000 collisions with severity of injury. The features were visualized to get an overview of the distribution of collisions and severity on their indicators. During visualizing no real trend or dependencies between the features could be found, only the peak numbers of the indicators. After balancing the data, creating a training and test data set, three different machine-learning models were developed and evaluated. The Decision Tree model achieved the highest scores. In the end, a comparison of the feature importance of the input features was made by using implemented techniques of the prediction models, but no clear statement could be made, because the models did not agree one another or the scores were not very high.

7 Conclusion

The provided data set has been used to train and test prediction models. The Decision Tree model reached the highest accuracy with 71% which is not satisfying. This means, the models must be trained on more data to develop better algorithms. Furthermore, the data collection should be checked and improved, because a lot of missing data was noticed, that lead to removing columns while data cleaning. Finally, with a better trained prediction model, selecting the important features by using the implemented functions of the machine-learning models, the severity of a car accident can be better predicted and people's lives can be improved.