Capstone Project

IBM Coursera Data Science

PREDICTING THE SEVERITY OF ROAD VEHICLE ACCIDENTS

28.08.2020, Germany

I Table of content

1 Business Understanding

1.1 Background

In recent years, the number of vehicles in operation and the number of drivers holding a valid driving license has increased continuously worldwide. One of the busiest countries in terms of road traffic are the United States of America with almost 280 million vehicles in operation. Alongside the increase of traffic, the number of road vehicle accidents has increased as well. Looking at the statistics for the United States, in 2018 there have been around 6 million car accidents, with a total of 12 million vehicles involved**. [1]** The described situation of the road vehicle accidents results in different consequences. Surely, the financial influence is immense. In a worldwide average, road vehicle accidents result in a cost of approx. 3% of the gross domestic product. Furthermore, physical integrity, not only of the drivers but also of pedestrians is in danger. So, to speak, almost 3 million people are injured every year in car accidents just in the US alone. In 2018, 36.560 people died in car accidents, which results in a fatality rate of 11.18 deaths per 100.000 capita.**[2]** In order to reduce the number of fatalities in car accidents, it is of highest important, that paramedics arrive as soon as possible and with the right equipment, the right number of ambulances and an appropriate rescue team size. To do so, it is necessary to evaluate the severity of the occurred accident as soon and as accurate as possible. A potential situation, in which the paramedics arrive at the accident scene and realize that further support is required might result in critical delays regarding the treatment and transportation of the accident victims.

1.2 Goal and objectives

To prevent the described situation and allow a quick evaluation of the accident severity, it is desired to develop an algorithm which can predict the severity of an accident by external environmental input factors, that can be observed right away from whoever is calling the emergency number from the accident scene. This algorithm would help to reduce the fatality rate from accidents by ensuring a quick and appropriate arrival and treatment from paramedics and would be of a huge help for **rescue services** all over the country. In turn, this would also lead to financial benefits regarding several stakeholders as cost factors such as hospital stays, life insurance payments etc. would be reduced. Therefore, the implementation of the algorithm would be also advantageous for any **(health) insurance firm**.

In order to achieve the defined goal, following objectives have been stated:

- Obtain a data set which includes the severity of road vehicle accidents, external environmental factors, and further information
- Develop a supervised machine learning model based on the obtained and cleaned dataset which requires as few features as possible to predict the severity of an accident
- Ensure that all features of the model can be visually/directly obtained by whoever is calling the emergency number

2 Data Understanding

2.1 Data source

The dataset used in this project was created by SDOT Traffic Management Division, Traffic Records Group, and contains all recorded collisions in the city of Seattle from 2004 to present. The dataset is updated weekly. The dataset used in this project (as of 28.08.2020) contains 194673 datapoints with a total of 38 columns. As every accident (data point) is labelled with a severity code which indicates the severity of the accident it is suitable for the previously described goal of this project. The dataset distinguishes between “property damage”, labelled as 1, and “injury” labelled a 2. Included in the 38 columns are several attributes which fulfil the defined condition of being “external environmental”. Some of these are light condition, road condition, date, number of vehicles involved etc. The dataset can be downloaded here and the metadata can be downloaded here.

2.2 Feature selection

Taking a closer look into the dataset, it can be seen that many of the columns contain interorganizational codes which are not relevant for the case of this study. Therefore, these columns are dropped (e.g. Incident key, report number etc.). Furthermore, there are redundant and irrelevant columns such as “Severitycode1” which contains the same information as the column “Severitycode”. Additionally, there are several attributes which are interesting but cannot be used for the described use case as they cant be observed right away and visually. An example for this is the column “Underinf” which describes whether or not a driver involved in the accident was under the influence of drugs. Another example is the column “speeding”, which describes whether or not one of the drivers was speeding up. Even though these attributes are interesting to understand the reason of the accident, they cant be observed before a more thorough investigation has taken place. Therefore, these columns are also dropped. As it is desired to implement the model nationwide, the location of the observed accidents is not taking into consideration, as a bias due to local circumstances should be prevented. The location column (displayed as “X” and “Y”) is also dropped. The features which have been chosen for training the model are:

- PERSONCOUNT → Number of person involved in the accident
- VEHCOUNT → Number of vehicles involved in the accident
- DATE → Date of the accident
- JUNCTIONTYPE → Intersection, Mid-Block etc.
- WEATHER → Weather condition during accident
- ROADCOND → Road condition during the accident
- LIGHTCOND → Light condition during the accident

All of these attributes can be observed easily and by anyone, which makes them suitable for use in the to be trained model.

Before continuing with cleaning and pre-processing the data, it has been checked whether the date has an influence on the severity of the accidents. To do so, first the date column has been changed to a “datetime” type. Afterwards, the interesting information has been extracted, which is the day of the week [df.dt.dayofweek]. Accordingly, a column was obtained containing numbers ranging between 0 and 5 which indicate the day of the week. After splitting the dataset by the severity code, two histograms have been plotted to observe whether the severity of accidents depend on the weekday (compare figure 1).

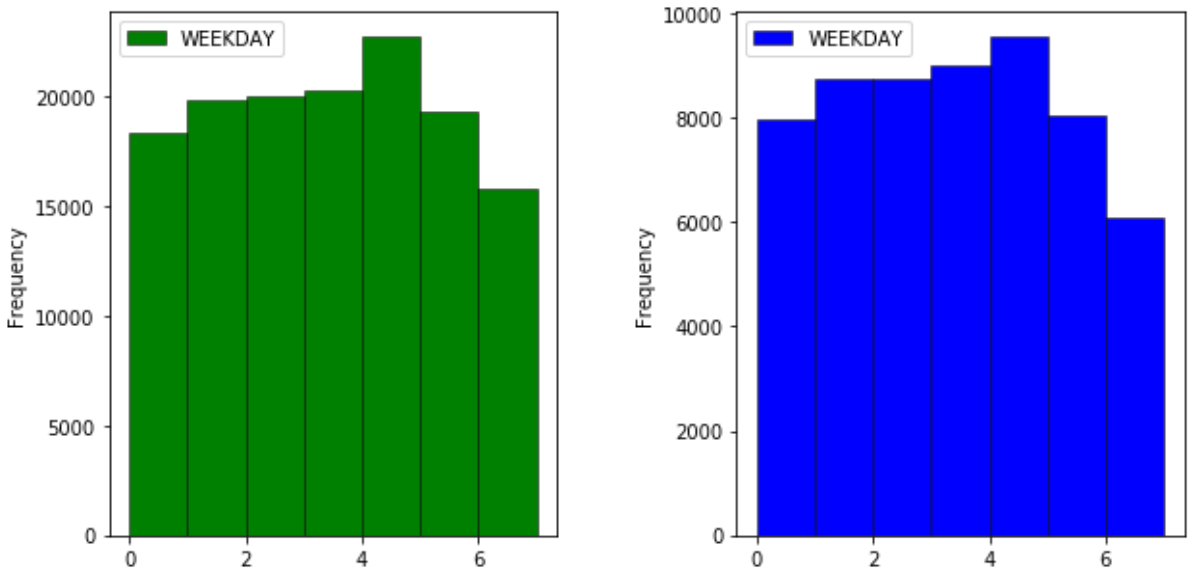

Figure 1: Frequency weekday for both severity codes

As can be seen in Figure 1, there is no relevant difference between the data set with severity=1 and severity=2. It appears that even though the day of the week seems to have an influence on the frequency of accidents, the day of the week doesn’t have a significant influence on the severity of the accident, therefore the date column has been dropped.

	SEVERITYCODE	PERSONCOUNT	VEHCOUNT	JUNCTIONTYPE	WEATHER	ROADCOND	LIGHTCOND
0	2	2	2	At Intersection (intersection related)	Overcast	Bad_Conditions	Daylight
1	1	2	2	Mid-Block (not related to intersection)	precipitation	Bad_Conditions	Dark - Street Lights On
2	1	4	3	Mid-Block (not related to intersection)	Overcast	Good_Conditions	Daylight
3	1	3	3	Mid-Block (not related to intersection)	Clear	Good_Conditions	Daylight
4	2	2	2	At Intersection (intersection related)	precipitation	Bad_Conditions	Daylight

Figure 2: Data Frame after dropping irrelevant columns

As a result of the data understanding section, the shown data frame of Figure 2 (shows first five entries of the data frame) has been obtained. In the following, the chosen features will be further observed, the data will be cleaned, visualized, and the data set will be balanced in order to prevent a bias.

3 Methodology

In a first step, we are going to apply data pre-processing methods and a explanatory data analysis. After this, the final data frame will be obtained, split into feature and target and several machine learning classification algorithms will be applied.

3.1 Data pre-processing and exploratory data analysis

The data frame obtained after feature selection consists of 194673 data points with a total of 7 columns. First, all rows with na values are dropped. Afterwards, the unique values for each column/attribute are checked and listed:

- **Personcount**:
  [2 4 3 0 5 1 6 16 8 7 11 9 12 17 26 22 10 37 13 36 28 14 53 19 30 29 23 44 15 32 21 41 27 20 35 43 81 18 25 48 24 34 57 39 47 54 31]
- **Vehcount**:
  [2 3 1 4 0 7 5 6 8 11 9 10 12]
- **Junctiontype**:
  ['At Intersection (intersection related)', 'Mid-Block (not related to intersection)' 'Driveway Junction', 'Mid-Block (but intersection related)', 'At Intersection (but not related to intersection)' 'Unknown', 'Ramp Junction']
- **Weather**:
  ['Overcast' 'Raining' 'Clear' 'Unknown' 'Other' 'Snowing' 'Fog/Smog/Smoke', 'Sleet/Hail/Freezing Rain' 'Blowing Sand/Dirt' 'Severe Crosswind', 'Partly Cloudy']
- **Roadcond**:
  ['Wet' 'Dry' 'Unknown' 'Snow/Slush' 'Ice' 'Other' 'Sand/Mud/Dirt', 'Standing Water' 'Oil']
- **Lightcond**:
  ['Daylight' 'Dark - Street Lights On' 'Dark - No Street Lights' 'Unknown', 'Dusk' 'Dawn' 'Dark - Street Lights Off' 'Other', 'Dark - Unknown Lighting']

We can see that the Weather, Roadcondition and Lightcondition columns contain values marked as "Unknown" and "other". As those values are not adding any information, we are going to drop the corresponding rows using the following method:

```
a = ["Other","Unknown"]
df = df[~df['Column_name'].isin(a)]
```

It can be also seen, that each of the columns "Weather", "Roadcond" and "Lightcond" have various different entries, which only vary slightly. Furthermore it can be seen, that most of the entries are very few in comparison to the dominant ones (observed by grouping the dataframe by the according attribute such as weather, counting and normalizing the result)! Therefore, we are going to summarize several attributes and also drop others. The described data composition can also be seen in Figure 3, where the frequency for each entry of the regarding attributes has been displayed for both severity-codes.

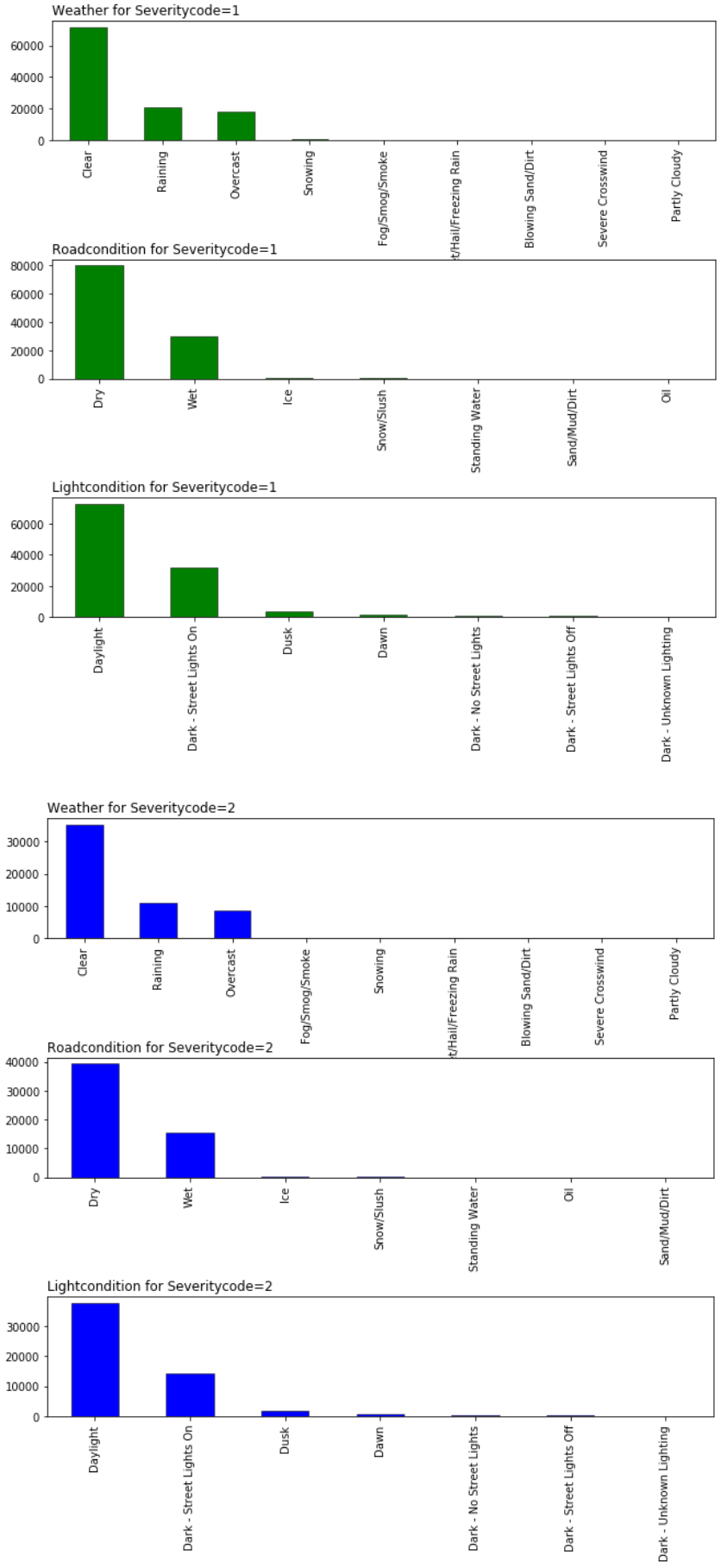

Figure 3: Frequency of entries for columns “weather”,”roadcond” and “lightcond”

After cleaning the three described entries, we are left with following unique values:

- **Weather**:
  ['Overcast' 'precipitation' 'Clear' 'Fog/Smog/Smoke']
- **Roadcond**:
  ['Bad_Conditions' 'Good_Conditions']
- **Lightcond**:'
  ['Daylight' 'Dark - Street Lights On' 'Dark' 'Dusk' 'Dawn']

In the next step, we have checked the data points for the columns “Personcount” and “Vehcount” by plotting further hisograms, boxplot and scatter plots. These plots can be seen in Figure 4.

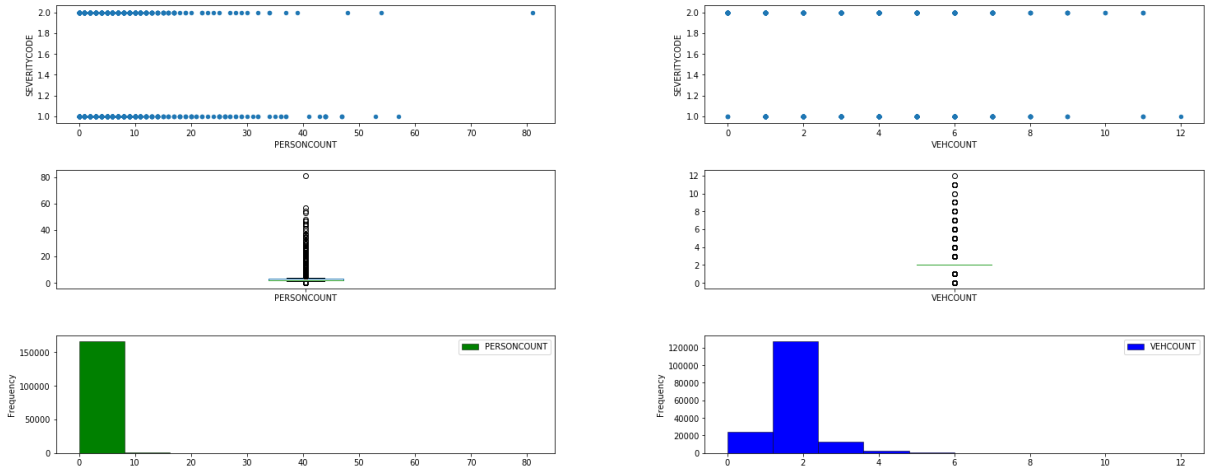

Figure 4: Explanatory Data Analysis for Personcount and Vehcount

It can be seen that there are a lot of outliers we should get rid of (compare boxplot and hist). We are going to drop data points where personcount > 6 an/or vehcount >4. The resulting boxplot and hisograms of the cleaned data can be seen in Figure 5.

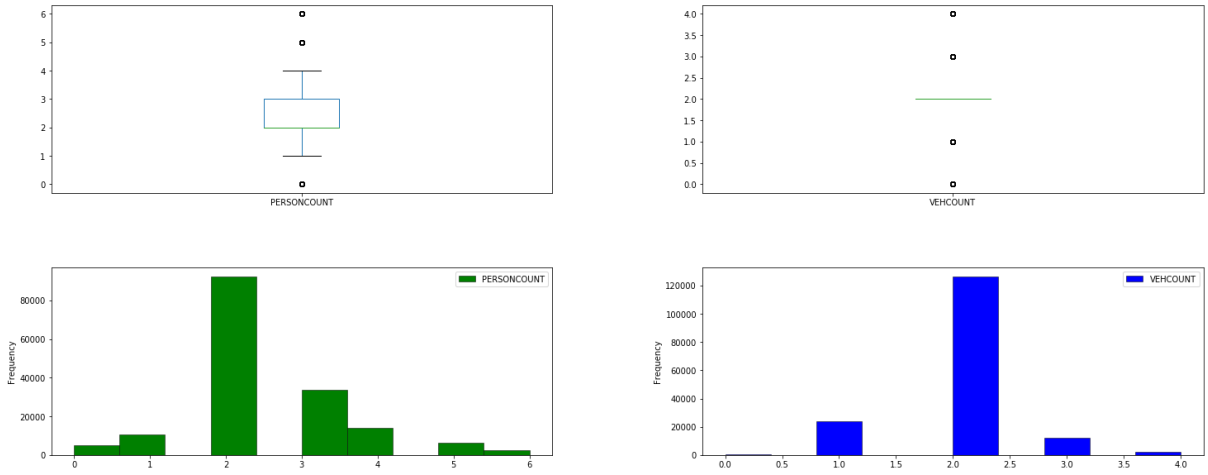

Figure 5: Explanatory Data Analysis for Personcount and Vehcount after cleaning

Finally, we have checked the data for the column “Junctiontype” and again we have discovered several outliers and attributes which are similar to each other, so that we have dropped some

values and also summarized similar attributes to create a leaner data set. As a result, the “Junctiontype” entries have been limited to:

- Intersection
- Mid-Block
- Driveway Junction

After counting the entries for severitycode=1 and severitycode=2 it can be observed, that the data set is imbalanced, as there are 118.401 entries for severitycode=1 and only 56815 entries for severitycode = 2. As the data in imbalanced, we need to balance the data. There are potentially several ways to do so. The two most popular ways are oversampling the minoriy class and undersampling the majority class. As it is also desired to reduce the needed computing power (as we want to apply severel algorithm and the computing power is unfortunately limited) we are choosing the undersampling method. In this particular case, it has been chosen to limit the dataset to 20.000 data points, equally split into severitycode=1 an severitycode=2 data points. As a result, a balanced data set has been reached.

As a last step in the data cleaning process, we change the categorical features into numerical features by using one hot encoding. The resulting, final data frame consists of 20.000 data points and 17 features.

3.2 Predictive Modeling

The machine learning algorithm should be able to predict the severity of an accident (1 or 2, meaning “property damage” or “injury”). The described problem is a typical case for a classification model, where a probability is calculated to which class a certain data point should be classified. Therefore, in the present project four different classification algorithms could be used and have been modelled:

- k nearest neighbors
- decision tree
- random forest
- logistic regression

Each algorithm will be trained several times iteratively with different parameters and will be evaluated using the metrics accuracy-score, f1-score and log_loss. For training the models, sklearn library has been used. Furthermore, the data has been standardized using StandardScaler and splitted using sklearn.model_selection.train_test_split with a test size of 20%. For each algorithm, the same train and test data set has been used.

3.2.1 **K-nearest neighbor**

“k” is the number of neighbours we consider in our algorithm to determine the class of a certain new data point. As k is unknow, an iterative approach has been used evaluating values of k in the range [1,10]. Each, in that way trained model, has been evaluated and a data frame with

the evaluation scores has been created and sorted by the accuracy_score in order to find out the k with the highest accuracy and best f1_score.

3.2.2 Decision Tree

For the decision tree algorithm, again an iterative approach has been used, to create several decision trees with varying maximal depth of the tree. Here a total of 50 trees have been modelled with a maximum depth ranging from 1 to 50. Similar to the knn algorithm, the decision trees have been evaluated using accuracy_score and f1_score.

3.2.3 Random Forest

In addition to the decision tree algorithm, a random forest algorithm has been used and evaluated. In this case, the estimator range has been iteratively changed in order to evaluate the best random forest model. The number of estimators has been varied between 1 and 100.

3.2.4 Logistic Regression

As the last algorithm, a logistic regression has been modeled with varying c values. The c value indicate the inverse of the regularization strength. A total of 3 different values have been tested [0.001,0.01,0.1]. In addition to the accuracy and the f1_score, the logistic regression has also been evaluated with the log_loss metric.

4 Results

As described above, all models have been evaluated regarding the accuracy_score, f1_score and log_loss for the logistic regression and for several different parameters. All of the results have been added to data frames, which then have been sorted by the accuracy in order to find the overall best model. The results for the best 5 models for each algorithm can be seen in Figure 6 and 7.

	k	Accuracy_Score	F1_Score
4	5.0	0.61225	0.611384
9	10.0	0.60975	0.606648
6	7.0	0.60875	0.605882
8	9.0	0.60500	0.604183
3	4.0	0.60350	0.594858

	Max_Depth	Accuracy_Score	F1_Score
6	7.0	0.64850	0.647881
5	6.0	0.64750	0.646833
7	8.0	0.64425	0.644000
4	5.0	0.64375	0.643603
3	4.0	0.64275	0.641532

Figure 6: Results for knn (left) and decision tree(right)

	N_estimators	Accuracy_Score	F1_Score
54	55.0	0.64125	0.641044
4	5.0	0.64075	0.640637
97	98.0	0.64000	0.639960
30	31.0	0.64000	0.639937
66	67.0	0.63950	0.639420

	C	Accuracy_Score	F1_Score	Log_loss
2	0.001	0.62300	0.622618	0.656204
0	0.100	0.62125	0.620931	0.654979
1	0.010	0.61950	0.619010	0.654798

Figure 7: Results for random forest (left) and logistic regression (right)

Even though the results are quite similar for all evaluated models, the best scores have been reached for a decision tree with a maximum depth of 7 with an accuracy_score of 0.6485 and a f1_score of .647881.

	precision	recall	f1-score	support
0	0.66	0.61	0.63	1979
1	0.64	0.69	0.66	2021
micro avg	0.65	0.65	0.65	4000
macro avg	0.65	0.65	0.65	4000
weighted avg	0.65	0.65	0.65	4000

Figure 8: Classification report for the decision tree with max_depth = 7

A further investigation of that particular model is seen in Figure 8, which shows the classification report. It can be seen, that similar results have been reached for both severities and for both, precision and recall. Therefore, it can be said, that the model is not biased.

5 Discussion

As described in the result section, a maximum accuracy of 0.6485 has been reached by the developed machine learning classification model. It appears, that the selected features are not enough to predict the severity of an accident with a high precision. This can have several reason. The first reason might be, that the data set is not big enough, as we have limited the number of entries to 20.000 in order to reduce the required computing power. Another possible explanation is, that the correlation between the selected features and the target variable is not high enough in order to create a model with a high accuracy. There might be the need to include further data points and/or further features to the data set in order to increase the accuracy of the classifier. It also has to be discussed, whether or not the target variable is sufficient, as it only differentiates between no injury and injury. There is no way to predict the severity of a potential injury, which might be of high interest for the rescue services, at which this model is aimed. Therefore, a further acquisition of data is recommended. The so acquired greater distinction of the target variable might also lead to a higher accuracy of the potential ML model. Also, at this point it has to be stated, that it would be also of great interest to train a second model which would be able to predict the likelihood of an accident to occur. This would not only be interesting for the rescue services but also for drivers and potentially for insurers.

(conclusion on next page)

6 Conclusion

In this study a machine learning classifier has been developed based on a data set of the severity of car accidents from the city of Seattle. The model takes several external environmental features and predicts the severity of an accident. As best algorithm for the described problem a decision tree with a maximum depth of 7 has been found. The developed model can be of great help for rescue services and might lead to a reduction of fatality rate in road vehicle accidents. Nevertheless, room for further improvement has been pointed out and potential further research has ben discussed.

7 List of sources

[1] https://www.driverknowledge.com/car-accident-statistics/

[2] https://en.wikipedia.org/wiki/Motor_vehicle_fatality_rate_in_U.S._by_year