CAR ACCIDENT SEVERITY STUDY BASED ON DATA FROM SEATTLE, WA

APPLIED DATA SCIENCE CAPSTONE PROJECT, COURSERA

SUBMITTED BY: MARCO SCHWAB

2020-11-06

INTRODUCTION

Every day, about 3,700 people die in road traffic worldwide. Road traffic accidents are an increasing problem even in times of modern vehicle technology and well-developed infrastructure. According to the World Health Organization (WHO), 1.35 million people die in traffic accidents every year. According to the Federal Statistical Office, the causes of traffic accidents involving personal injury in Germany are:

- Driver mistakes
- Road conditions, weather influences, obstacles
- Technical defects
- Incorrect pedestrian behavior

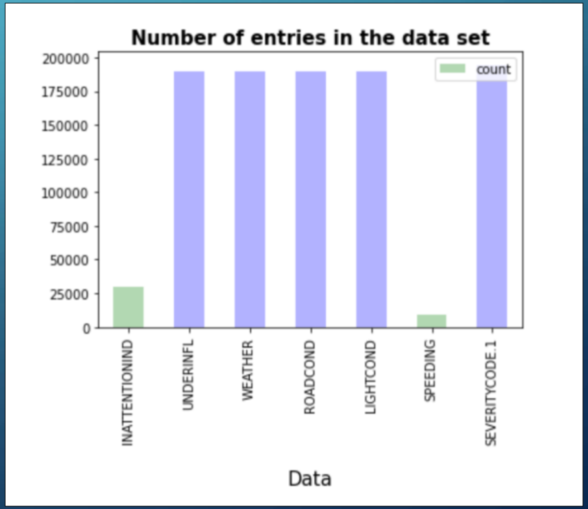
This information can probably be transferred to other countries.

DATA

The dataset used in this study is based on a document called 'Collisions—All Years' from the organisation 'SDOT Traffic Management Division, Traffic Records Group'. It includes all types of collisions counted by the Seatle Police Department and Traffic records in the city of Seatle, WA in the timeframe beginning 2004 until present. The document is weekly updated.

The project purpose is to analyze and predict the severity of an accident based on some particular features:

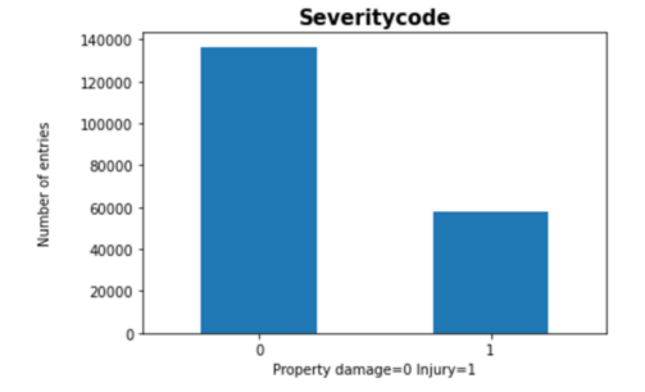
Feature	Description
INATTENTIONIND	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	A description of the weather conditions during the time of the collision.
ROADCOND	The condition of the road during the collision.
LIGHTCOND	The light conditions during the collision.
SPEEDING	Whether or not speeding was a factor in the collision. (Y/N)



DATA CLEANING

The number of usable entries in the columns of the datasets varies significantly. For this reason, a data cleaning was firstly done on the set of data. Missing or not usable entries were deleted, others were transformed or recoded. During the data cleaning the following steps were performed:

- Data were imported
- Questions marks were replaced by NaN
- Changing severity code from 1=Property Damage Only and 2=Physical Injury to 0=Property Damage Only and 1=Physical Injury



DATA PREPARATION

The data preparation section consists of:

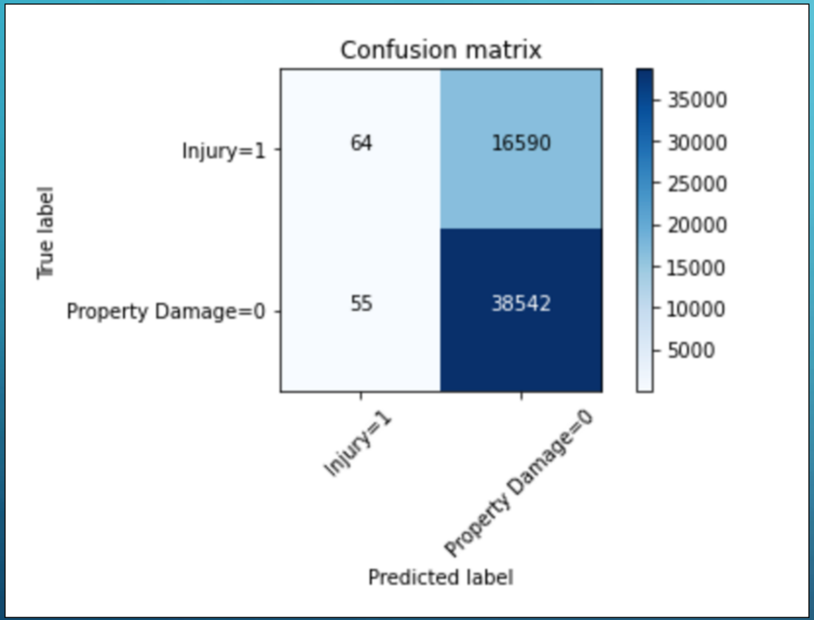
- Feature Selection
- Recoding to numeric values
- Replace Unknown or NaN to statistical representative values
- Convert data types to integer

METHODOLOGY

- Data understanding and Data cleaning
- Data preparation
- Machine Learning Section

RESULTS

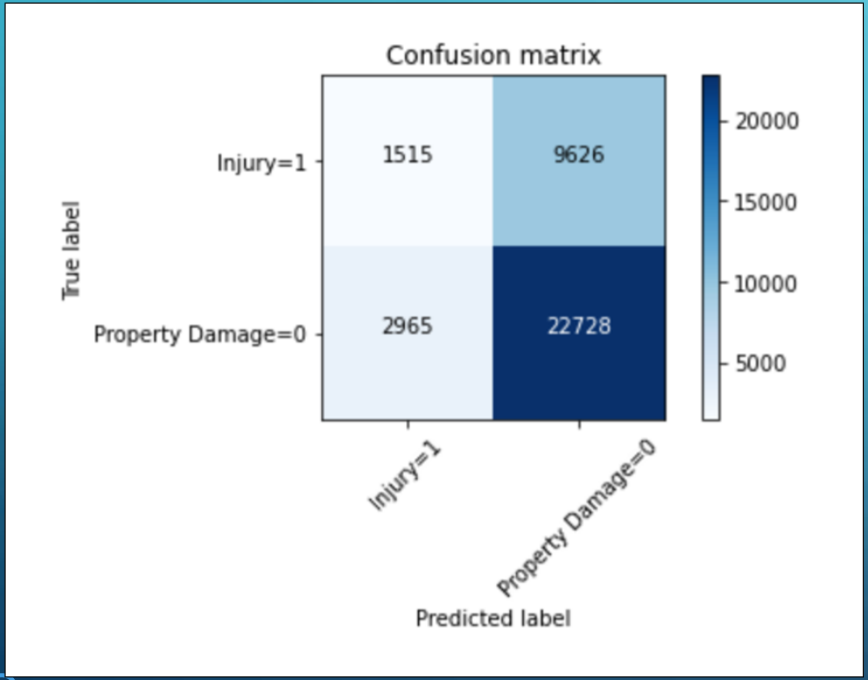
- Decision Tree Analysis



	Precision	Recall	f1-score
0	0.70	1.00	0.82
1	0.54	0.00	0.01
accuracy			0.70
macro avg	0.62	0.50	0.42
weighted avg	0.65	0.70	0.58

RESULTS

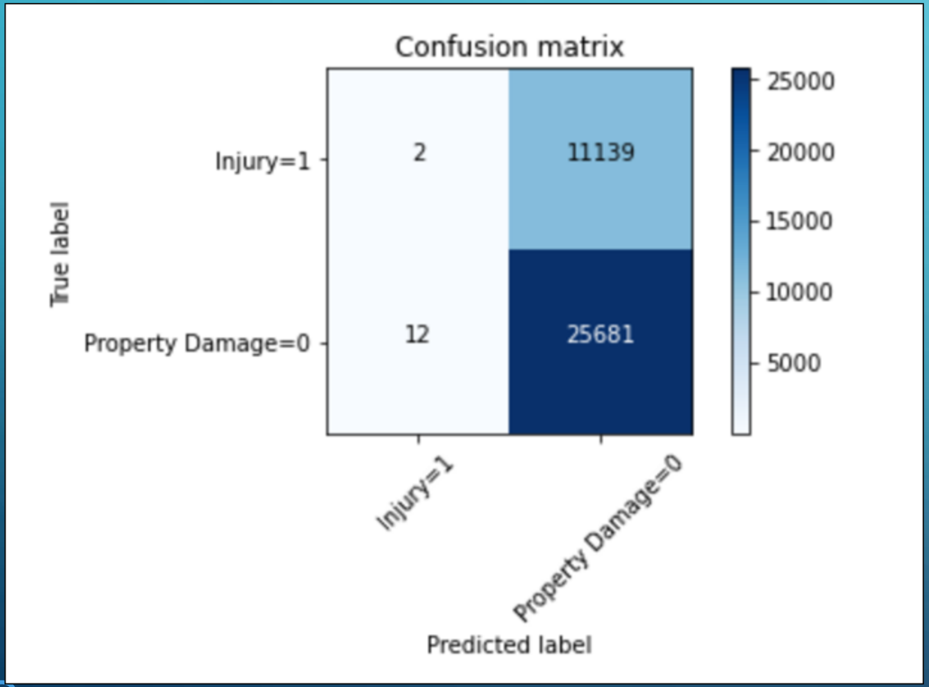
- K Nearest Neighbor (KNN) Analysis



	Precision	Recall	f1-score
0	0.70	0.88	0.78
1	0.34	0.14	0.19
accuracy			0.66
macro avg	0.52	0.51	0.49
weighted avg	0.59	0.66	0.60

RESULTS

- Logistic Regression Analysis



	Precision	Recall	f1-score
0	0.70	1.00	0.82
1	0.14	0.00	0.00
accuracy			0.70
macro avg	0.42	0.50	0.41
weighted avg	0.53	0.70	0.57

DISCUSSION

In the Decision Tree as well as in the Logistic Regression algorithm the Recall value for 1 (Injury) is calculated as zero or a very small value. That means that the injury cases could not very good predicted out of the real injury cases. Even for the KNN the Recall for 1 is with 0.14 very low. In contrast, the Recall value of 0 (Property damage) shows very good values. For the Decision Tree the Precision values are very good balanced between 0 and 1. Much better than the Precision value for Logistic Regression and also better than for the KNN.

With a weighted f1 score of 0.6, a good balanced Precision and the best Recall values for 1 (Injury) the KNN seems to be the best alternative ML algorithm for the analyzed data in the prepared condition.

Algorithm	f1-score weighted avg.	Property Damage (0) vs Injury (1)	Precision	Recall
Decision Tree	0.58	1	0.54	0.00
		0	0.70	1.00
KNN	0.60	1	0.34	0.14
		0	0.70	0.88
Logistic Regression	0.57	1	0.14	0.00
		0	0.7	1.00

CONCLUSION AND RECOMMENDATIONS

With the comparison made in table 5 an assessment on the chosen ML algorithm can be made. According to this comparison, the KNN seems to be the best alternative to use to predict car accidents on given conditions. With a closer look on the confusion matrix of the KNN, it can also be seen, that the predicted values are better balanced than for the other algorithm. According to this approximately 1500 injury cases can be predicted out the given 11100 injury cases and 22700 property damage cases out of 25600 property damage cases can be predicted.

Nevertheless there is potential for improvement. One improvement can be that the entry data should be better balanced. So, there are less than half the amount of injury cases included than cases with property damage. In addition the amount of entries for SPEEDING and INATTENTIONIND are much lower than the other features. The reason for this could be, that there are gaps in the data or they were deleted during the cleaning process. Artificial balancing processes can help here to get better data to feed the ML tool. In addition more features can be chosen to train the algorithm. There might be features which at first sight do not seem to have any influence, but in interaction with the others features can improve the output significantly.

Also, the given data could be analyzed regarding the locations where the accident happens. The amount of accidents took place on special areas in Seatle could be plotted on the map of the city. This can help the city planners to identify dangerous areas and optimize the traffic condition to avoid accidents.