Capstone Project - Car accident severity

1. Introduction of business problem and interested stakeholders

People die in traffic accidents because of risks that cannot be eliminated. But fatal accidents also happen due to avoidable risks. And these avoidable risks are to be identified below for the city of Seattle (WA) in order to reduce the risk for road users to save human lives, and lower insurance costs due to unnecessary damages.

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

The Seattle (WA) collisions data set is used, which contains traffic accidents in Seattle from 2004 to the present day. The data set is used because it is maintained and made available by a state body, the Seattle Department of Transportation, and thus contains highly up-to-date (weekly updated) and credible data. The data set can be found here: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

The following features and labels are used as data. SEVERITYCODE (severity of collision) is used as the label and corresponds to the severity of the collision. The features include ADDRTYPE (collision address type), ROADCOND(ition), and LIGHTCOND(ition).

	SEVERITYCODE	ADDRTYPE	ROADCOND	LIGHTCOND
count	170823.000000	170823	170823	170823
unique	NaN	3	8	8
top	NaN	Block	Dry	Daylight
freq	NaN	108552	121839	112825

3. Methodology: Discussion and description of exploratory data analysis and machine learning algorithms

First a qualitative, exploratory data analysis was carried out with the help of the metadata (descriptions of features and label) in form of thinking about which features can have an influence and which features cannot have an influence on the label SEVERITYCODE. Features without believed influence, e.g., OBJECTID (ESRI unique identifier) were not included in the analysis. Then all double features, features that describe exactly the same, for example INCDATE (the date of the incident) and INCDTTM (the date and time of the

incident), were identified, and one of the double features was not considered further. Then it was checked whether all variables were assigned the correct data type, which was the case; the number of nan values was determined: the selected variables contained a maximum of 3% nan per variable and therefore the nan data points were removed from the data set; the categorical variables were encoded with dummy variables (1 or 0); and finally the data set was split into 80% training and 20% testing data sets.

Three machine learning algorithms were used for classification: GradientBoostingClassifier, KNeighborsClassifier, and SVC (support vector machine). GradientBoostingClassifier was applied because the algorithm works well with categorical data and, compared to KNeighborsClassifier and SVC, often provides a higher predictive accuracy. KNeighborsClassifier was used because the algorithm can be understood more easily than the other two algorithms and still offers a certain flexibility by setting hyperparameters. And the SVC was chosen because the SVC, like GradientBoostingClassifier, provides a good level of prediction accuracy with categorical data, and while it is the most computationally demanding of the algorithms used, linear and polynomial kernels can be tried, which led to the choice of the SVC algorithm.

4. Results of GradientBoostingClassifier, KNeighborsClassifier, and SVC

The GradientBoostingClassifier achieved an accuracy of 0.6723 ~ 67% with the testing data set. KNeighborsClassifier achieved an accuracy of 0.6599 ~ 66%. And SVC achieved an accuracy of 0.6723 ~ 67%. GradientBoostingClassifier has predicted from 34,165 data points in the training data set 22,968 true positive, 2 true negative, 6 false positive, and 11,189 false negative. KNeighborsClassifier predicted 21,373 true positives, 1,172 true negative, 1,601 false positive, and 10,019 false negative. And SVC predicted 22,966 true positives, 4 true negatives, 8 false positives, and 11,187 false negatives. Thus, the GradientBoostingClassifier and SVC achieved the highest prediction accuracy of the three models, with KNeighborsClassifier only slightly behind. Consequently, ADDRTYPE (alley, block, intersection), ROADCOND(ition), and LIGHTCOND(ition) together can explain to some extend the label SEVERITYCODE.

5. Discussion of observations noted and recommendations based on results

The predictability of GradientBoostingClassifier, KNeighborsClassifier, and SVC strongly depend on the choice of the values for the hyperparameters. With the selected values (k-nearest neighbors '3', SVC kernel 'poly' and SVC degree '8'), the best forecast accuracy could be achieved. In addition to using these three algorithms, it would be interesting to compare different neural network models with the predictive accuracy of GradientBoostingClassifier, KNeighborsClassifier, and SVC. Furthermore, the obvious is confirmed that poor road conditions and poor visibility (poor lighting) conditions result in a higher probability that if an accident occurs, it is more severe in degree.

It is recommended to make the maximum permissible speed dependent on the current light and road conditions. If the light and road conditions are poor, a lower, adjusted maximum speed limit must be set. Street areas with poor lighting should also get better lighting and the street lighting should be connected to sensors that switch on the lighting when the lighting conditions fall below a certain threshold, even if it is not yet dark outside.

6. Conclusion

Maximum speed limits must be dynamically adapted to the road conditions (weather) and lighting conditions, and poorly lit areas of the street need to be better lit. Neural networks should also be used to see whether they can achieve a higher level of prediction accuracy.