Applied Data Science Capstone – Car accident severity

A report from S. Greven

Business Problem

Vision Zero – to reduce road deaths to almost zero by 2050 - is one goal of the European Union. There are many aspects und factors which can lead to an accident. One big factor is the weather and the related conditions.
With available data like weather forecasts, road conditions and the light conditions, it could be possible to warn drivers in order to avoid potential accidents. Since data is real-time available, it would be helpful f there would be a ML-model which could use these parameters in order to predict an accident probability and severity.

Data

In this Notebook the suggested data from Coursera was used. This data set includes a huge variety of data related to accident incidents.
Data like geographical location, index, type and number of involved and injured parties are included.
Since the explained objective of this Business Problem is to detect the relation between Weather conditions and severity of accidents, most of the data is not relevant for this use case. It will be evaluated, how big is the impact between the attributes Road conditions, Weather condition and light conditions and the accident severity. Target variable is severity Code, which differs 1 or 2.
Interest
Car manufacturer, public health organizations, insurance companies could be interested in such a model. It can be used for public health or as a business case for an extra service, for which car drivers might pay.

Data cleaning

All of the relevant data (Severity, Weather, Road, Light) are categorical and therefore outliers are irrelevant. First of all, a general overlook will be done.
According to the first methods for an initial overview, the data set is 194673 rows and 38 attributes / columns big. The target variable severity code is imbalanced, since severity level 1 is 2.5 times more represented within the data than severity level 2. Out of the approx. 195000 rows there are around 5000 values for each input variable, where the data is not available. Since most of the time, if one value so the relevant attributes is missing, also the other values are NaN, those rows will be deleted. In order to have a balanced data, the author decided to set the sample data at 57104.
The categorial data will be set from string to a numerical value in order to prepare it for modelling.

Methodology

Since it is not a numerical task, Regression models are not the right tool. Classification models like KNN or Decision Tree Model will be used in order to predict the severity of the accident.

kNN

Since the accuracy score for a rather low k (e.g. 7) was quite low with approx, the range for possible ks was increased. The best k was found at --- with an Accuracy score of ---. Because of the high ressources neede for this operations, it needs to be evaluated if this is the correct model type.

Running KNN-model, the best k-value would be 91 with an accuracy with approx. 55%. This score is rather poor and is not usable for a proper model. Therefore, a second classification model will be evaluated.

Support Vector Machines

Results

Both SVM and kNN offer no ideal solution for the discussed question. kNN offers aaccuracy with aprox 55%. SVM offers

Discussion

After using two types of classification models with a rather bad the question is whether

Conclusion

This project had the objective to find out, whether weather, road or light conditions have an impact on the severity of accidents. After finishing the modeling with the respective results, following can be stated:

Features like road condition, light condition, weather do not influence the accident rate as we expected. It appears that the accident occurred the most during daylight, when the weather was clear, and road condition was dry.

This leads to the point, that further investigations on the data need to follow. Attributes like day, time or accident location could lead to better prediction of model and could help to reach the goals of vision zero.