Jan Luca Neduchal
13 September 2020

Car accident severity

IBM Applied Data Science Capstone Project

Introduction/Business Problem:

Trying to reduce the frequency of car collisions, I'm going to use the given data for the Capstone. During this project I'll predict the severity of car accidents given the current weather, road and visibility conditions. The goal of my project is to alert drivers when the current conditions are bad, so that they can drive more careful.

Data:

Data requirements:

For our given problem we need a specific set of data. It should be a large amount of data, so that we can train our model as good as possible and predict severity of car accidents as precise as possible to prevent more accidents in the future.

Data description:

For my Capstone Project I have used the given data by Coursera called ‚Data-Collisions.csv‘.

The target variable for the Capstone is going to be 'SEVERITYCODE' - it is used measure the severity of a car accident from 0 to 4 within the dataset. To weigh the severity of a car accident the following attributes are used: 'WEATHER', 'ROADCOND' and 'LIGHTCOND'.

Severity codes are as follows:

0 : Little to no Probability (Clear Conditions)

1 : Very Low Probability - Chance or Property Damage

2 : Low Probability - Chance of Injury

3 : Mild Probability - Chance of Serious Injury

4 : High Probability - Chance of Fatality

Other important attributes:

OBJECTID: ESRI unique identifier

ADDRTYPE: Collision address type: Alley, Block, Intersection

LOCATION: Description of the general location of the collision

COLLISIONTYPE: Collision type

PERSONCOUNT: The total number of people involved in the collision

PEDCOUNT: The number of pedestrians involved in the collision

PEDCYLCOUNT: The number of bicycles involved in the collision

VEHCOUNT: The number of vehicles involved in the collision

INCDTTM: The date and time of the incident

PEDROWNOTGRNT: Whether or not the pedestrian right of way was not granted

SPEEDING: Whether or not speeding was a factor in the collision

In its given form, the data is not fit for analysis. There are many columns, that we don't need for our specific analysis and most of the features have to be converted to our desired data type.

We must use label encoding to covert the features to our desired data type.

Methodology:

After analysing and cleaning our dataset, the data is now ready to be fed through ML models. We will use the models we already know from the previous chapters:

- K-Nearest Neighbour -> predicting the severity code of an outcome by finding the most similar data point within k distance
- Decision Tree -> displaying all possible outcomes to analyse the consequences of a decision
- Logistic Regression -> predicting one of our two severity codes (1 or 2)

When we're finished with that process, we test the accuracy of our machine learning algorithms using the Jaccard Similarity Index, the F1-Score and LogLoss for Logistic Regression.

Results:

Machine Learning Model	KNN	Decision Tree	Logistic Regression
Jaccard Similarity Index	0.564001947698565	0.5664365709048206	0.5260218256809784
F1-Score	0.5401775308974308	0.5450597937389444	0.511602093963383
LogLoss	—	—	0.6849535383198887
Variable	**k**	**max_depth**	**C**
Most accurate amount of variable	**25**	**7**	**6**

Discussion:

In the beginning we changed the data type of some of our given data from 'object' to 'int8' - a numerical data type - to use it for our algorithm. After the first issue we had to take care of unbalanced data. To match the minority class we sampled down the majority class, which was class 1, until the values of both classes matched. That step was followed by analyzing and cleaning our data for the three machine learning models: K-Nearest Neighbor, Decision Tree and Logistic Regression. To see how accurate our models were, we lastly used the Jaccard Similarity Index, the F1-Score and LogLoss for Logistic Regression. To improve the accuracy of the models the different variables of each model (k, max_depth, C) had to be adjusted several times for the best possible result.

Conclusion:

Based on the given historical data about weather conditions in relation to car accidents, we can say that the weather does have an impact on whether or not driving could result in a car accident leading to property damages (class 1) or injuries (class 2).