IBM capstone project report

Vicente Zehnder

Problem description

The amount of road traffic incidents is immense, it causes great damage to all the US families and is estimated a loss of $810bn every year. Therefore, being able to identify the factors that lead to a greater number and severity of accidents is really important and has a big incentive.

One way that this problem can be addressed is with data science tools, using machine learning to create models that can forecast concentration of road accidents, for the people to prepare. It seems pretty obvious that the more important factors will be about the climate, day of the week, the influence of substances, and some other few. But the important thing is to be able to determine how these variables interact with each other.

This could lead to a far better understanding of the traffic accidents, and being able to predict them with much higher accuracy. So in this project, we will be using a dataset of the severity and conditions of different accidents to create and test a model that is able to predict the severity of possible car accidents.

Data description

The data is about collisions in Seattle, provided by the SDTO. It contains around 200,000 samples, with 37 different characteristics from the crashes. The idea of this project is to train a machine learning model to be able to predict the severity of the crashes, that have value 1(property damage) or 2(injury damage). So it will be a binary classification model. There are a lot of missing values, around 40% of the samples have at least one NaN value. But many characteristics probably don't affect the outcome of the model, so before doing something about the missing values is better to remove the columns that are not relevant. The variables that will be included in the model will be

- If the driver was under alcohol or drugs

- Collision type

- Weather conditions

- The light conditions

- The condition of the road

After removing all the other variables there are only 3% of the samples with missing values, so those can be removed with no problem. So then the data needs to have further processing, for the training and testing of the model. Like for example balancing the dataset and transforming the categorical values into numerical ones.

Results and conclusions

	Algorithm	F1-score	Precision
0	KNN	0.67	0.71
1	Decision Tree	0.69	0.78
2	Logistic Regression	0.58	0.68

After apliying all the content learned on the course specialization, we where able to do a proper the data preparation, model training, model testing and report evaluation.

On this report we can see that the best option to predict the outcomes of the severity of the crashes(with the data that was used for training and testing) is the Decision tree.