Udacity Machine Learning Nanodegree

Capstone Proposal

Mohit Sharma

Domain Background

The performance of a vehicle control strategy, in terms of fuel economy improvement and emission reduction, is strongly influenced by driving conditions and drivers' driving styles. Also, Road safety rules and regulations are designed to prevent the citizens from fatal incidents. Although policies are in place, we observe negligent behaviour of the drivers which lead to serious injuries or death crashes. It is of utmost interest of the authorities to understand and analyse human behaviour to take necessary corrective and preventive actions.

Problem Statement

The major stakeholders are the citizens, road transport authorities, Insurers and Researchers/Data service providers. In order to design a driving assistance system there is a need to get an understanding of the data on the driving patterns and broadly distinguish bad drivers from good ones. This in turn will benefit Insurers in analysing underwriting risks, prevent frauds and designing No-claim-discount systems (NCD systems), etc. Additionally, the concerned authorities will need insights to design benchmarks for qualifications and driver licensing regulations, etc.

Datasets and Inputs

Every single vehicle is observed at various time stamps, to record the details of trips made, traffic conditions, vehicle details like length, weight, no of axles of the vehicle, road conditions, lanes switched, weather conditions etc. along with the driving styles are recorded. Driving styles are divided into three categories as mentioned below:

“1” indicates : “Aggressive”,

“2” indicates : ”Normal” and

“3” indicates : “Vague”

There are three datasets in total:

Training Data – The dataset contains 12994 observations and 5 variables which include Length of vehicle in cm, weight of vehicle in kg, Number of axles etc. Out of 12994 approximately 21.3% of the drivers are labelled as aggressive, 49.4% as Normal and rest 30%as vague.

Train_WeatherData – The dataset contains 162,566 observations and 9 variables.

Train_Vehicletravellingdata – The dataset contains 162,566 observations and 10 variables. The dataset provides information about the preceding vehicle and the road with respect to the weather.

Solution Statement

This is clearly a supervised problem. The goal here is to identify the driver type and utmost importance is on identifying aggressive drivers. As there are three levels which need to be classified, this implies the use of algorithms which can help us solve multinomial classification.

Evaluation Metric

The Recall is proposed as the evaluation metric for correctly identifying 1's. The 1 represents the aggressive drivers. As the total number of aggressive drives is just 21% and it becomes important to correctly classify these drivers IE out of total actual aggressive drivers how many our model is able to correctly classify. We want this to be as good as possible.

Project Design

First of all the data exploration needs to be done. It is useful to examine the percentage of missing values, to detect outliers and to figure out the type of each feature. In order to gain further inside the data will be visualized.

If the exploration step found outliers, then they will be removed or treated. Missing data will be either imputed or the whole row will be dropped. The new variables will be derived based on the information available. Then the three datasets needs to be merged together to create a one final dataset. This dataset will be split into a training and testing set. Many different models will be trained using standard parameters. The plan is to try multinomial logistic regression, Naive Bayes, decision trees, SVM, random forrest, AdaBoost, XGBoost and LightGBM as models. The models will be quantified based on the Recall metric and based on computation time. It will be checked whether the models are overfitting or underfitting. We can also use K-fold cross-validation to identify the best model. This ensures that the training set doesn't have to be reduced any further to create a separate validation set.

A principle component analysis might be used in order to make use of latent features and reduce data dimensions. This will also help in reducing the computation time. The performance of the updated models will be compared to the test dataset.

The best models will be further examined and tuned in order to improve the performance. Different sets of hyper-parameters will be tried out. It is planned to use grid search or random search to find the best hyperparameters. The tweaked models will be evaluated on the test set and compared based on the Recall performance metric.