Contents

Predicting the “Driving Style”

Problem Description

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.

The 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.

About Data:

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.

Objective:

You are expected to create an analytical and modelling framework to predict the driving style of each id categorizing into “Aggressive”, ”Normal” and “Vague” and also obtain the actionable top 10 data insights (patterns) for “Aggressive” class using the tree based algorithms.

Details of the dataset

The datasets are provided as cited below for the analysis:

1. **Vehicle Data:**
   - **"Train.csv" & "Test.csv"**
   - These files consist of the vehicle details of each ID, like ID, length, weight of the vehcile, etc .
   - Train.csv has the target attribute "DrivingStyle" also, where as Tes.csv doesn't have as it has to be predicted .

2. **Trip data :**
   - **"Train_Vehicletravellingdata.csv" & "Test_Vehicletravellingdata.csv"**
   - These files consist of the triip information like ID, vehicle speed, road condition based on weather, timegap with the preceding vehicle, etc

3. **Weather information:**
   - **"Train_WeatherData.csv" & "Test_WeatherData.csv"**
   - These files consist of the details about weather during the trip like ID, date and time, weather details like temperature, humidity, etc

4. **Attributes Details:** "AttributeInformation.docx"
   - This has the details of attributes for the datasets cited above (1 to 3)

Note: For analysis, consolidate/aggregate all the datasets cited above (1) to (3)

Note: Missing values are denoted as "","NA" in general in the datasets. Please go through the document "AttributeInformation.docx" thoroughly to address different aspects in the data.

Main Tasks:

1. Exploratory Data Analysis using visualizations in R Notebook or Jupiter notebook format (Use only Train data for this task)
2. You are expected to build a framework that predicts the driving sytele ("Aggressive" or "Normal" or "Vague" in target attribute "DrivingStyle").
3. *You are expected to build a framework that generates the patterns for "Aggressive" on target attribute "DrivingStyle", using any decision tree algorithms only.* ***(Use only train data for this)***

4. *Viva*

Note: Use the aggregated/consolidated train dataset for model building and tuning the model and then apply the model on test data for obtaining the predictions.

Submission 1 (3rd June 2018 by 6 p.m.) :

i. Preprocessing of Data

1. Data preparation for model building should be done

ii. Visualization

1. Train data should be used for data analysis and visualizations.

iii. Documents to be submitted into Grader tool are cited below:

Visualization :

1. Commented Code developed for Preprocessing, visualisation with the name “submission.R” or “submission.ipynb” as the case may be.
2. Report on your understanding about the problem, data analysis based on the visualizations you made and pre- processing steps required etc, with the name “VisualizationReport.doc”.

Submission 2 (4th June 2018 by 6 p.m.):

i. Model Building and predictions

i. Train data should be used not only to build the model but also to tune and conclude the model for submission
ii. Test data should be used for evaluation of model
iii. Test data is totally unseen data and hence it does not have the target attribute, "DrivingStyle".
iv. The predictions obtained for test dataset should be uploaded to Grader tool with the name “predictions.csv”.

ii. Generating the patterns using any decision tree algorithm only:

i. Traindata should be used for generating the patterns

ii. Generate the patterns for fraud cases (ie., “Aggressive” level in the target attribute) in data and also compute evaluation metrics for patterns

iii. Documents to be submitted into Grader tool are cited below:

Visualization & Modelling & Patterns:

1. Commented Code developed for Preprocessing, visualisation and modelling & Patterns with the name “submission.R” or “submission.ipynb” as the case may be.
2. Upload the predictions obtained on the Test dataset to Grader tool with the name “predictions.csv”.
3. Upload the top 10 patterns for fraud (ie., “Aggressive” level in the target attribute) to the Grader tool with the name patterns.csv”

Submission 3 (8th June 2018 by 6 p.m.):

1. Documents to be submitted into Grader tool are cited below:

a. Improved versions of Submission 1 & 2

i. Final Commented R code / python code, predictions, patterns with additional efforts and improvements made during the week

b. Viva Presentation

i. Final presentation for viva with the name “viva.ppt”

Note: Please follow the naming convention for the submission files as cited above only.

Error Metrics:

- The score for evaluation on Grader would be the mean f1 scores for the three classes
- Consider “Recall” for “Aggressive” level of Target attribute as error metric for deciding the top 10 patterns for fraud on target attribute.