Driver Behaviour Analysis

Course Project
DS203 : Programming for Data Science
Autumn 2021
IIT Bombay

Shiven Barbare
190100110
Department of Mechanical Engineering
IIT Bombay
Mumbai, India
190100110@iitb.ac.in

Akash Chodankar
1901000009
Department of Mechanical Engineering
IIT Bombay
Mumbai, India
190100009@iitb.ac.in

***Abstract*—This paper outlines a methodology that can be used to study and predict driving pattern and behaviours exhibited by different drivers, which can be used for fleet monitoring. The focus lies in analyzing the driving data obtained through a combination of GPS and accelerometer data by making use of Exploratory Data Analysis (EDA). Predictions are made using standard classification models such as Logistic Regression, Support Vector Classifier (SVC) and K-Nearest Neighbour Classification and their performance accuracy is compared to obtain the best possible results.**

***Index Terms*—Driver Behavior Analysis (DBA), Classification, Exploratory Data Analysis, Machine Learning (ML)**

I. Introduction

Driver Behaviour Analysis (DBA) is an emerging field worldwide, given the rising awareness about environment conservation and growing impetus on safety of passengers. According to a World Health Organization (WHO) report, in 2016 about 1.3 million people died worldwide in car crashes or related accidents. It is also worthwhile to analyze the environmental footprint of vehicles. As reported by the European Parliament in 2019, transport is responsible for nearly a third of the European Union's total carbon dioxide (CO_2) emissions.

DBA also forms the basis of Driving Score Prediction that is used extensively by companies like Frotcom for fleet analysis. The primary logic behind such scoring techniques is to analyze typical driving behaviours such as harsh acceleration/braking, sudden turns, etc. and use these metrics with machine learning models to quantitatively judge a driver's performance. Such an approach can lead to reduced costs for companies as well as improved passenger and driver safety; this is a win-win situation.

In this paper we have attempted to perform a classification of driver behaviour based on data recorded by inertial sensors. A dominating aspect of our approach is the Exploratory Data Analysis (EDA) that we have employed to visualize the data to better process it. This is followed by data pre-processing. The final stage involves predictions. For this, we have employed a set of five different classification models viz. Logistic Regression, Support vector Classifier (SVC), K-Nearest Neighbour Classification, and . Models are trained using the scikit-learn library. The performance of these models was analyzed by tuning hyper-parameters (if any). Comparison between performance of different models was done by evaluating the classification reports which included the metrics of precision, recall, f1-score and support.

II. Datasets

The UAH-DriveSet [1] is the primary dataset used in this project. It is a public collection of data captured using DriveSafe, a proprietary driving monitoring app by various testers in different environments. It provide more than 500 minutes of naturalistic driving data including raw data and additional semantic information, as well as the video recordings of the trips. The salient feature of this dataset are as follows:

- Data was obtained by six different drivers with six different vehicle models (D1-D6).
- Two road-types of different lengths were considered (motorway and secondary).
- Driver behaviour was classified into normal/drowsy/ aggressive behaviours.
- Raw data was obtained using GPS and accelerometer (with different sampling rates) in a smartphone that was used to capture videos of trips.
- The dataset also provides processed data about lane and vehicle detection that is logged in real-time using the app.

Additionally another dataset [4] was used to verify the prediction models that were used in this project. The data was obtained using the Carla Simulator platform that consisted of a 3-axis gyroscope and accelerometer

III. Exploratory Data Analysis (EDA)

Before performing any operations on the data, it is essential to visualize the data so that we get a *feel* of what the dataset actually looks like. Once we have a fair idea about the general

trends that the data follows, we can go ahead and process the data to convert it into a desired format.

The first task is to read the data, drop the inconsistent columns and rename the remaining ones for ease of analysis (data obtained from GPS and accelerometer is obtain in a .txt format without headers; corresponding column headers can be obtained from the dataset description [1]).

From the Raw GPS data, the variation of vehicle speed can be seen as a function of time for different behaviours 1. Clearly, aggressive behaviour is characterized by higher fluctuations in speed as well as higher average speed.

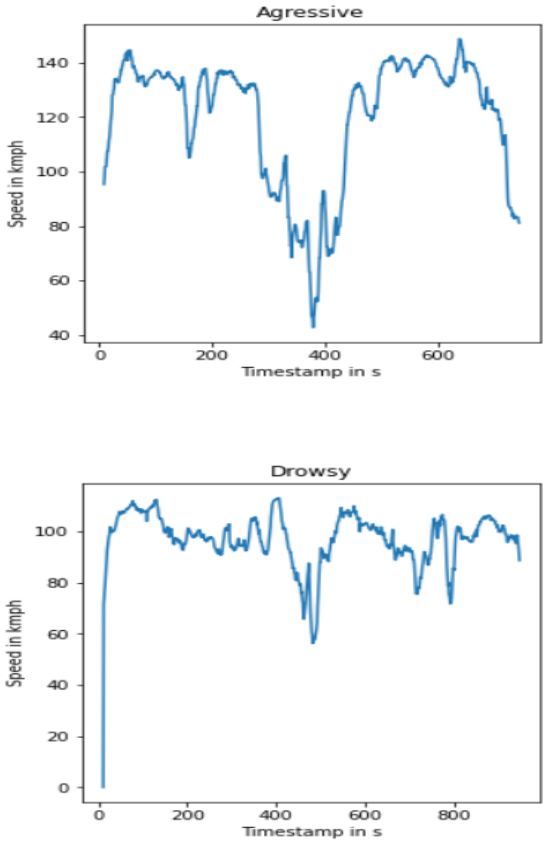

Fig. 1. Variation of speed with time

The accelerometer data gives more comprehensive information about driver behaviour and patterns. Scatter plots can be used to find variation of one component of acceleration with respect to another as shown in 2. the clustering of data points is around the origin and the variation is higher for aggressive and drowsy behaviour than normal behaviour, which is expected.

A Correlation Heat map is generated for different driving behaviours to see which variables are highly correlated and which ones are uncorrelated. The heat map for aggressive behavior is shown in 3. The acceleration values are highly correlated to their kalman filtered values. It is also seen that z-acceleration is uncorrelated with timestamp. The variation of acceleration is shown in 4. The hexagonal plot 5 shows that uncorrelated values of acceleration (normal behaviour).

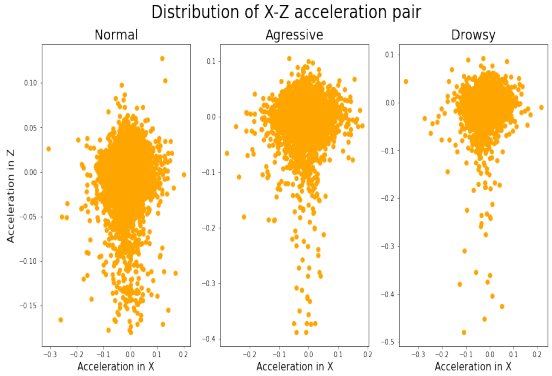

Fig. 2. Distribution of z-acceleration with respect to x-acceleration

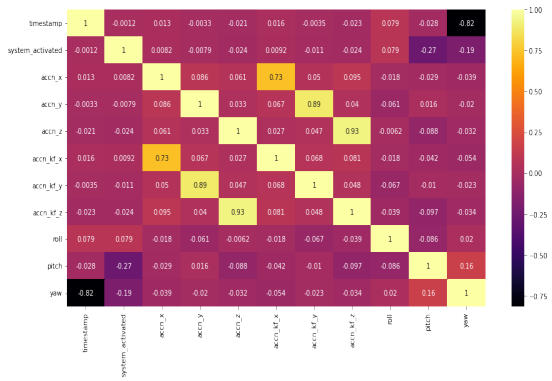

Fig. 3. Correlation Heat Map for aggressive behaviour

System activation is the only categorical variable, thus it's variation is plotted with another (slightly) correlated variable - roll using a boxplot as shown in 6.

IV. Models and Predictions

Driver Behaviour is considered to fall in one of the following three categories :

1) Normal
2) Drowsy
3) Aggressive

Predictions are done based on accelerometer data as GPS sampling rate is very low, abd thus the number of data points are limited. First, the data is pre-processed to eliminate correlated variables based on the correlation heat map. Thus only actual acceleration values are considered. Further, we have eliminated timestamp from our analysis, since it represents sequential data. So we have the three acceleration components as well as roll, pitch and yaw in as our input variables.

The dataset is now appended with a column for behaviour (0-Normal, 1-Aggressive, 2-Drowsy) and this column provides the actual labels. The Classification Models are trained using the scikit-learn library by adopting an appropriate train-test split. The models considered are :

- Logistic Regression
- Random Forests (RF)

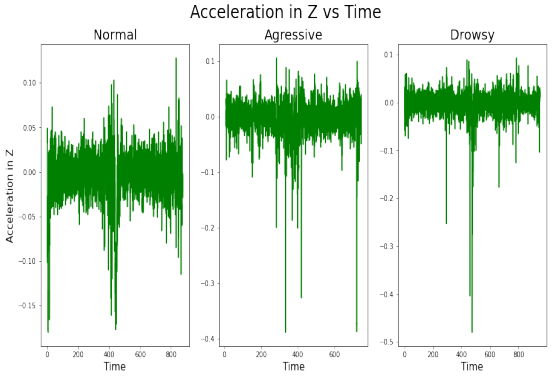

Fig. 4. Acceleration vs time

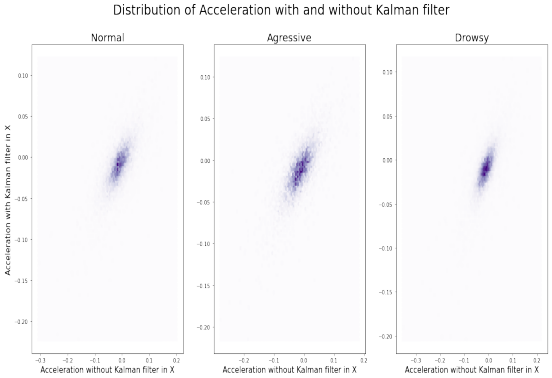

Fig. 5. Hexagonal Plot

- K-Nearest Neighbours (KNN)
- Support Vector Classifier (SVC)
- Multi-layer Perceptron Classifier (MLPC)

Out of the above classifiers, MLPC is a feedforward artificial neural network (ANN).

The models were trained using a two-step processes:

1) Training of all five models on a single driver's (D1) data
2) Training of all five models using complete dataset (D1-D6)

The performance of the models was compared by generating their corresponding classification reports. The metrics considered were precision, recall, f1-score and support .

The classification report for Logistic Regression is shown in 7. Clearly, this model is not precise. Thus, further models (mentioned earlier) were explored and their performance vis-a-vis driver behaviour classification were considered.

V. Libraries

The following set of libraries were used in the project:

- matplotlib
- numpy
- pandas
- seaborn
- scikit-learn

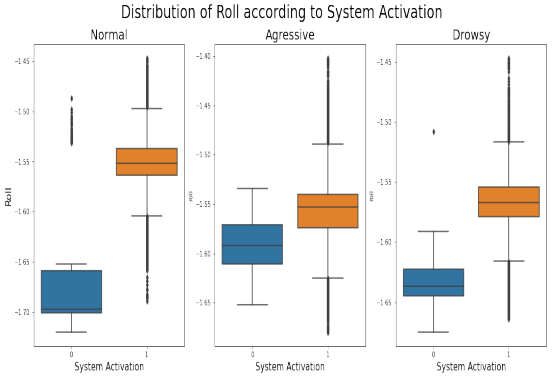

Fig. 6. Box Plot for Roll vs System Activation

	precision	recall	f1-score	support
0	0.39	0.49	0.44	11093
1	0.39	0.18	0.25	9951
2	0.39	0.47	0.42	12120
accuracy			0.39	33164
macro avg	0.39	0.38	0.37	33164
weighted avg	0.39	0.39	0.38	33164

Fig. 7. Classification Report (Logistic Regression)

VI. Conclusions

The observations drawn from the EDA plots have been mentioned in III.

The performance (accuracy) of different classifiers has been summarized in I

TABLE I
Classifier Performance

Classifier	Accuracy
Logistic	0.3899
RF	0.8859
KNN	0.6683
SVC	0.3209*
MLPC	0.7085

* The low accuracy of SVC is because an upper limit was set on the number of iterations, since the running time for SVC was significantly high.

Thus, MLPC was found to be the best classifier. The performance can be further improved by considering parameters other than those obtained using accelerometer.

VII. Acknowledgment

We would like to thank our Course instructors Prof. Amit Sethi, Prof. Manjesh Hanawal, Prof. Sunita Sarawagi and Prof. S. Sudarshan for giving us this wonderful opportunity to apply the learnings obtained through this course. We would also like to thank our TAs who have put in incredible efforts to ensure that we are able to make the best use of the content covered in this course.

VIII. Code Demonstration Video

The demonstration video for the code can be found here.

References

[1] E. Romera, L.M. Bergasa and R. Arroyo, "Need Data for Driving Behavior Analysis? Presenting the Public UAH-DriveSet", IEEE International Conference on Intelligent Transportation Systems (ITSC), pp. 387-392, Rio de Janeiro (Brazil), November 2016.

[2] Peppes, N., Alexakis, T., Adamopoulou, E., & Demestichas, K. (2021). Driving Behaviour Analysis Using Machine and Deep Learning Methods for Continuous Streams of Vehicular Data. Sensors, 21(14), 4704.

[3] FROTCOM Intelligent Fleets : https://www.frotcom.com/features/driving-behavior-analysis.

[4] Carla Simulator Platform : https://www.kaggle.com/dasmehdixtr/carla-driver-behaviour-dataset