STANDARD BANK HACKATHON

Telematics Unleashed: The Truth Behind Taxi Drivers

Lise Prinsloo
Minette Farrell
Rachel Rawraway
Danel Adendorff

Contents

1 Introduction

A compelling and intellectually stimulating challenge presented itself in the context of this years hackathon hosted by Standard Bank and Mobalyz: modeling the qualitative assessment of taxi driver's behaviour as either *good* or *bad*. A driver's behaviour can be analysed using telematics data such as: idle time, acceleration, braking, location and speed to name a few.

A dataset consisting of 106 vehicle's telematics along with a claims dataset was given to extract features that are indicative of driver behaviour. From these datasets, a number of features were identified as important aspects that can identify driver behaviour including: ignition state, speed, acceleration, graviational forces and location. The features identified were then used to create a table that summarizes each driver's behaviour in terms of speeding, braking, acceleration, cornering, etc. From this summarized table, a k-nearest neighbour model was trained to identify driver behaviour.

Lise or Rachel talk more about the model, the results, etc.

A design choice was made to implement the solution as a cloud solution using Google Cloud Platform. Different Google Cloud products were used in the process including Cloud Storage, Data Prep, BigQuery, Google Colab and Looker studio. These products were chosen based off similar solutions, ease of implementation, good documentation and overall user friendliness both for the developer and user. Implementing the solution on a cloud platform, provides the future users with the flexibility of seamlessly scaling the solution, easily expanding on existing products and knowing that the solution is safe and secure from power outages and potential hackers.

2 Data Collection and Exploration

The raw data consists of 106 .csv files that contains thousands of rows of telematics data from each vehicle. Each vehicle is equiped with a telematics device that sends information that is used to model driver's behaviour. The features shown in Table 1 are telematics features that are considered important in determining the road safety of a vehicle.

Feature	**Description**
vehicleid	Unique Vehicle Identifier
time_stamp	Timestamp of data sample
ignitionState	Vehicle Ignition State
speed	Vehicle Speed at time of sample
event_description	Event Type
linear_g	Foward/Backward corrected G forces
lateral_g	Sideways corrected G forces
road_speed	Speed limit on road segment at time of sample
x_accel	Accelerometer uncorrected for direction - only gravity
y_accel	Accelerometer uncorrected for direction - only gravity
z_accel	Accelerometer uncorrected for direction - only gravity
SP_NAME	Suburb Name

Table 1: Driver Features

2.1 Defining behavior model

The main aspect to consider when defining the behavior of driver of a taxi is the safety of passengers. This is the objective for which the model will interpret good or bad driver behavior. In order to determine the behaviour, several features were considering including the total number of claims and the total cost of claims per vehicle, harshness of acceleration and braking, including the net gravitational force experienced by the passengers, harshness of corners taken, area that the vehicle drives in, time of day that the vehicle is most active in, the number of times the telematics device was off or not transmitting information and the number of times GPS signal was lost, the number of the times the vehicle was idle and the average number of stops and average distance travelled per day. These features were identified as possible indicators of driver behaviour. Each of these features were then implemented in a k-NN model to determine which of these features are contributing factors.

2.2 Exploratory Data Analysis

Google Cloud's Data Prep performs automatic exploratory data analysis on all the columns in the table created. The summary is given in Figures 1 and 2.

Figure 1: Exploratory Data Analysis

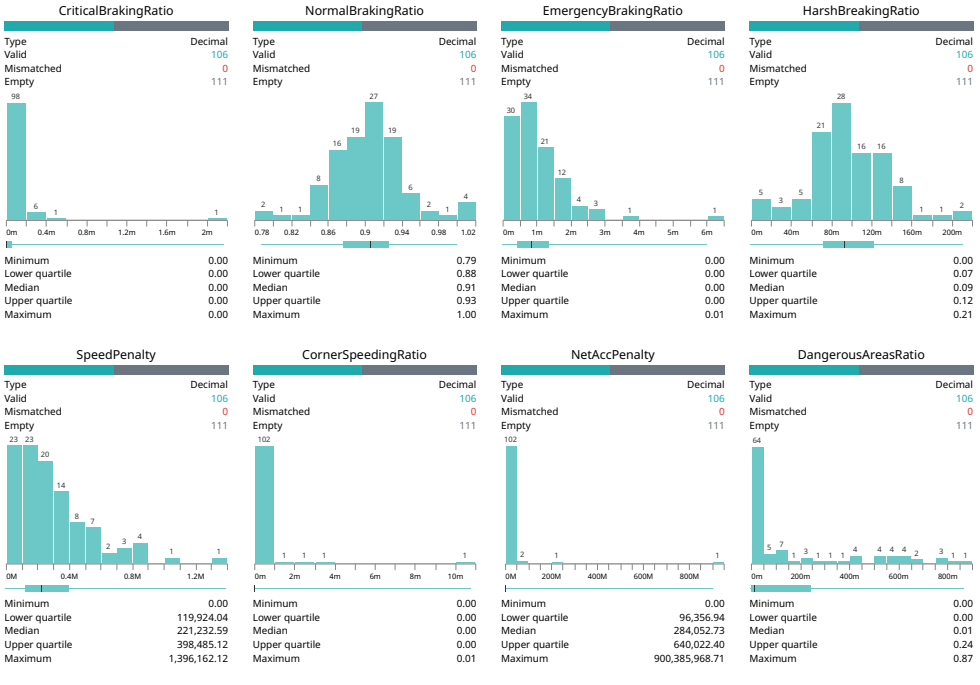

Figure 2: Exploratory Data Analysis

2.3 Data Pre-processing

Google Cloud's Data Prep automatically detects the data's schema when it is imported. One of the files, has issues with the automatic schema detection, but this was solved by opening the file using Google Sheets and saving it as a new .csv. After the new .csv file was imported, everything worked as expected. Following this logic, it might be easiest if the raw data was captured using Excel or Google Sheets to ensure that the schemas are correctly implemented and the .csv files are saved in an appropriate format.

To identify missing and null values, the raw data was examined using Python in Google Colab. The following features contained null values:

Column	**Replaced with**	**Number of instances**
x_accel, y_accel, z_accel	0.0	1886506
road_speed	60	50219
altitude	40.0	34841
odometer	0	372
SP_CODE	00000000.0	272
SP_NAME, MP_NAME, MN_NAME, DC_NAME, PR_NAME, country_name	Unknown area	272
coordinate_longitude	18.4	12
coordinate_latitude	-33.9	12

Table 2: Data Cleaning

There were many missing values for acceleration, thus it was assumed that null values indicated that there was no acceleration in the direction and was replaced with 0. Most of the cases where the speed was missing was residential areas such as estates and townships (residential areas), thus the missing values were replaced with 60kph. The data in the area's columns such as the municipality and country, were replaced with Unknown area and a dummy SP code of 00000000.0. Missing values for altitude, longitude and latitude were replaced with the rounded values for Cape Town, i.e. 40m above sea level, 18.4°East and 33.9°South. Missing values for the odometer was replaced by 0, since there were no direct way to estimate it, and the features did not depend on it.

Google's Data Prep was used to pre-process and clean the data by using a UI interface that uses SQL queries to implement the wanted replacement values. Data Prep combined all the csv files in one table by appending each driver's information to the bottom of the table. This allows Data Prep to apply the all the pre-processing steps on one table instead of many smaller tables. After pre-processing is applied to the raw data, Data Prep creates user defined features for each vehicle ID.

3 Feature Engineering

Variable	Explanation
TotalClaimsCost	Total cost of claims
NumberOfClaims	Number of claims
OverSpeedCount	Number of times the driver exceeded the speed limit
IdleRatio	Ratio of time the car is in idle
GpsLostCount	Number of times GPS signal was lost
TelematicsOffCount	Number of times tracking system was turned off
AverageNmrStopsPerDay	Average number of stops per day
AverageDistPerStop	Average distance per stop
EmergencyBraking	Instances of emergency braking
NormalBraking	Instances of normal braking
HarshBraking	Instances of harsh braking
DangerousTimes	Times of day considered dangerous
DangerousAreas	Number of times driving in dangerous areas
CriticalBraking	Instances of critical braking
CornerSpeeding	Number of times over speeding into corners
SpeedPenalty	Penalty score that takes into account the severity of over speeding
BrakingPenalty	Penalty score that takes into account the severity of braking as experienced by the passengers
AccelerationPenalty	Penalty score that takes into account the severity of acceleration experienced by the passengers
NetAccelerationPenalty	Penalty score that takes into account the severity of net acceleration experienced by the passengers

Table 3: Final features used in the modeling process

this part below is from chat gpt so maybe alter and format better

Total Claims Cost: High total claims cost could indicate that a driver has been involved in more accidents or incidents, suggesting potentially risky behavior or poor driving habits.

Number of Claims: A higher number of claims could indicate a driver's tendency to be involved in accidents or incidents, suggesting poor driving habits or risky behavior.

Over Speed Count: A high count of instances where the driver exceeds the speed limit might indicate aggressive driving behavior or lack of adherence to traffic rules.

Idle Ratio: A high idle ratio could suggest inefficient driving practices, such as unnecessary engine idling, which wastes fuel and contributes to environmental pollution.

GPS Lost Count: Frequent occurrences of lost GPS signal could indicate that the driver goes through areas with poor signal or tries to avoid being tracked intentionally, potentially indicating questionable behavior.

Telematics Off Count: Frequent instances of the tracking system being turned off might imply that the driver is trying to avoid monitoring or supervision, which could be considered suspicious behavior.

Average Number of Stops Per Day and Average Distance Per Stop: These metrics can provide insights into the efficiency of the driver's route planning and driving habits. A higher number of stops coupled with shorter distances might indicate efficient driving.

Emergency Braking, Normal Braking, Harsh Braking, Critical Braking: Frequent instances of emergency or harsh braking could suggest erratic driving behavior, po-

tentially indicating lack of caution or adherence to safe driving practices.

Dangerous Times and Dangerous Areas: Dangerous times and areas could indicate that the driver operates in risky conditions, potentially resulting in unsafe behavior or higher chances of accidents.

Corner Speeding: Frequent corner speeding could suggest aggressive driving behavior or lack of adherence to safe speed limits, which can increase the risk of accidents.

Speed Penalty: A penalty is added that reflects by how much a driver is going over the speed limit, by taking the ratio of the vehicle speed and the road speed.

Braking Penalty: A penalty is added that reflects how harsh the braking is, by taking the ratio of the negative linear g-forces and a value that viewed as an unacceptable g-force for braking (0.4g).

Acceleration Penalty: A penalty is added that reflects how harsh the acceleration is, by taking the ratio of the positive linear g-forces and a value that viewed as an unacceptable g-force for acceleration (0.4).

Net Acceleration Penalty: A penalty is added that reflects how harsh the net acceleration is, by taking the net g-force as $net\_g = \frac{\sqrt{x\_accel^2 + y\_accel^2 + z\_accel^2}}{9.81}$ and summing over the total force experienced.

4 Model

lol so chatgpt gave me this, but maybe this is actually something we can use??

Explain how Google AutoML was utilized for creating a predictive model. Describe the steps involved in setting up the AutoML experiment, including selecting the target variable and features. Highlight the advantages of using AutoML, such as its ability to automate model selection and hyperparameter tuning

A simple unsupervised machine learning model was used to cluster the vehicles in different classes and observe different behaviours of drivers. The most optimal way of clustering the given data was to use kmeans clustering. Different values for k were investigated as can be seen in figure .

insert figure of diff kmeans clusters

The each line and line colour in a parallel coordinate plot represents a vehicle and its class respectively. So it can be observed that one vehicle in figure has this and that features.

insert parallel coordinate plot

4.1 Training and Evaluation

Discuss how the training of the predictive model was conducted using the prepared data. Explain the evaluation metrics used to assess the model's performance, such as accuracy, precision, recall, and F1-score.

4.2 Deployment and Prediction

Describe how the trained model was deployed to make predictions on new data. Outline the process of inputting new taxi trips data and receiving predictions on driver performance.

4.3 Results

Present the results of the model's performance on the evaluation metrics. Discuss the implications of the predictive model for improving taxi service and driver management. Address any limitations or areas for improvement in the current solution.

5 Google Cloud Platform Integration

Google Cloud Platform (GCP) was chosen to provide the client with a cloud based solution. GCP was chosen over other cloud platforms based on two main factors: cost of the solution and user friendliness of the platform. Google Cloud allows a developer to create buckets and instances that aren't as expensive as other platforms for a project of this size. For instance, in total project was billed: 0.58 USD. In terms of development, GCP has many free APIs that makes implementing the solution much easier.

Figure 3, shows the infrastructure used to implement this solution. The data is collected and stored as .csv files. These files are then uploaded to Google Cloud Storage in Buckets. Data Prep is used to clean pre-process and clean the data using a step by step recipe as defined by the user. Once new data is imported, Data Prep runs the recipe and loads the pre-processed and cleaned data to BigQuery. BigQuery allows the user to perform normal SQL queries on the data tables. Google Colab is used to create models and plot graphs from the data using Python. Finally, Looker is used to create a dashboard that gets updated as new data is imported.

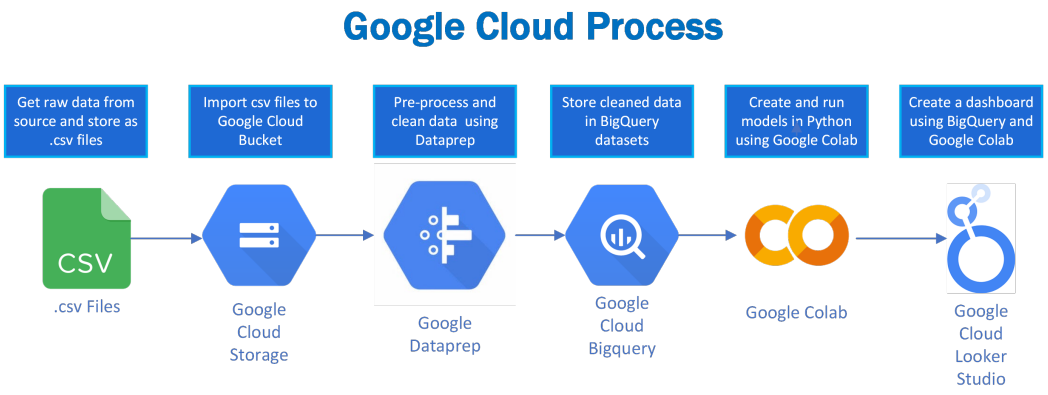

Figure 3: Google Cloud Platform

6 Conclusion

Summarize the key points of the document. Emphasize the success of utilizing Google Cloud Platform for data cleaning, modeling, and prediction in the Hackathon competition.

7 Building onto this

Suggest potential enhancements or extensions to the current solution. Discuss possibilities for further refining the model and incorporating additional features.