A Data Analysis Framework to Rank HGV Drivers

Grazziela P. Figueredo1,2, Philip R. Quinlan1,2, Mohammad Mesgarpour3, Jonathan M. Garibaldi1,2, Robert I. John1
1. School of Computer Science, The University of Nottingham, NG8 1BB, UK
2. The Advanced Data Analysis Centre, The University of Nottingham, NG8 1BB, UK
3. Microlise, Farrington Way, Eastwood, Nottingham NG16 3AG, UK
4. The Automated Scheduling, Optimisation and Planning (ASAP) research group, NG8 1BB, UK
gzf, prq@cs.nott.ac.uk, Mohammad.Mesgarpour@microlise.com, jmg,rij@cs.nott.ac.uk

Abstract—We report on the details of the methodology applied to support shortlisting the nominees for the Microlise Driver of the Year awards. The aim was to recognize the United Kingdom’s most talented heavy goods vehicle (HGV) drivers, with the list of top 46 drivers across 16 different companies determined through the analysis of telematics data. Initial data for the awards was gathered from over 90,000 drivers engaging with Microlise’s telematics solutions. The data was analysed anonymously in order to identify the best criteria to establish top performing drivers. The initial selection was made based on a minimum number of miles driven across each of the four quarters in 2014. Outlier removal and a consensus clustering framework were subsequently employed to the dataset to identify subgroups of drivers. Three categories of drivers were identified: short, medium and long distance drivers. Each qualifying professional belonging to one of the three categories was then assessed using a range of criteria compared to other drivers from the same category. To determine the final winners, questionnaires for further evidence and indicators that might contribute to a driver being named as a winner was sent down to employers and their responses were evaluated.

I. INTRODUCTION

Microlise Driver of the Year Awards [1] has launched in 2015 to identify and honour the most talented heavy goods vehicle (HGV) drivers in 2014 across the UK. The competition is designed to encourage HGV drivers to continuously improve both safe and economic driving. It also aims to assist fleet managers to focus attention on reviewing and improving commercial fleet driver safety awareness and driving behaviour. The competition relies on Microlise’s telematics solutions which monitor drivers against set criteria in order to compare driver performance. All data generated by the telematics system are transmitted and collected from vehicles in real-time.

A study published by Frost & Sullivan [2] finds that installed-base commercial vehicles telematics was 14.7 million in 2014 and estimates that it will reach 37.9 million by 2020. Telematics solutions have traditionally been used to monitor the location of the vehicles and report some basic information, such as engine hours, odometer and speed [3]. With the recent development in cloud data storage, computing power, telecommunication, advances in automotive industry, increasing the capabilities of on-board units and advanced data analytics, various other services, such as fuel saving, fleet performance management, driving behaviour monitoring, dynamic routing, remote diagnostics and prognostics and telematics insurance are being offered by telematics providers [4].

Microlise telematics system links into the vehicle’s CAN databus (Controller Area Network) which allows Microlise to capture and process in real-time a whole range of driving style, engineering and fault code data. Driving style data includes measures like, engine revs (to identify greenband driving and over-revving), fuel consumption, accelerator position, idling, use of cruise control, use of PTO (power take off), use of primary and secondary (engine) braking. Moreover, Microlise’s Telematics device reports on harsh acceleration, harsh cornering and braking. All of this data is used to create "scorecards" which reports on each driver’s driving style against specific fuel economy and safe driving skills targets.

For the purpose of this competition, gathered telematics data between January 1st, 2014 and December 31st, 2014 from vehicle fitted with the Microlise telematics system in the UK have been aggregated. Digital tachograph driver IDs, which are unique identifiers of drivers have been used to indicate drivers who have worked in more than one contract (such as agency drivers) during the year. They also assisted to identify whether more than one driver has driven the same vehicle during the year. In addition, vehicle details such as make, model, gross weight, engine size provided by the Driver and Vehicle Licensing Agency (DVLA) are also used to classify the vehicles.

In this paper we present the methodology employed to shortlist the top 46 drivers in the competition, which is our main contribution. There were several challenges to overcome:

1) The method should prevent any sort of favouritism.
2) Given the diversity of vehicles and technologies, a common set of data per vehicle/driver across all customers had to be identified.
3) Instinctively, it was known that there were several types of drivers, with different driving profiles. It was part of the task therefore to identify and create the criteria of comparison within these groups.

The techniques were chosen to enable a fair, unbiased pipeline to compare and rank drivers based on the set of data defined to evaluate the driver’s performance. The resulting methods were produced and validated based on several interactions between data analysts, experts in transport and telematics and Microlise. We believe this methodology can be employed to other problems in the industry involving clustering analysis, profiling and ranking. In the next sections we introduce the details of the framework developed (Section II), how the profiles of drivers were determined (Section II-C) and present the conclusions and opportunities for improvements for the next competitions (Section III).
II. METHODOLOGY

The analysis methodology was developed to provide a fair, unbiased pipeline to enable the comparison and subsequent ranking of driver performance. There were three primary challenges: (i) How to determine which attributes should be included in the ranking? (ii) How to establish measurements of comparison, considering drivers from different vehicles, with distinct tasks and journeys? (iii) How to characterise the different profiles of drivers? To overcome these challenges, an empirical combination of established approaches from Knowledge Discover in Datasets (KDD) [5] was employed. As a result, a robust, replicable ranking framework was established that was independently verified and validated by Microlise and other domain experts. The pipeline developed consists of the following steps:

1) Data pre-processing, where the attributes, aggregates and data instances to be analysed were obtained from the raw data.
2) Outlier removal, in which drivers with attribute values with a larger distance from acceptable values predetermined by specialists were removed.
3) Application of a consensus clustering technique [6] to identify groups with specific characteristics (profiles of drivers) in the dataset.
4) Outlier removal within the clusters established.
5) Another application of clustering as an attempt to identify possible subgroups within the groups.
6) Removal of drivers whose values for the criteria adopted were above average compared to drivers from the same subgroups.
7) Determination of the final list of drivers, where qualitative questionnaires were sent to employers for further validation of the output obtained and final shortlisting of the winners.

A schematic representation of the framework employed is shown in Figure 1. Each step is explained in further details in the following sections.

A. Data Pre-Processing

The aim of the pre-processing step was to take the raw data, containing details of every journey and create an anonymised driver profile for the year. The data pre-processing steps are described in Figure 2 and consist of three key stages. The raw dataset contained information of over 5 million journeys. The first step therefore was to filter journeys that may not be valid. A journey was considered to be valid if either the amount of fuel used, the average engine torque or the distance travelled was greater than zero; therefore, if all of these fields were zero the journey would be excluded. Given the competition was open to HGV drivers only, journeys that involved a vehicle below 7.5 tonnes were also excluded. Once completed, the first stage of the pre-processing removed 169,294 journeys. The second stage was to move the dataset from each row representing a journey to each row representing a driver across the entire year. All journeys made by the same driver within the same class of vehicle were grouped together. As a result
motorists of multiple vehicle classes could be entered under each vehicle class i.e., they were considered for the competition in multiple categories. Given the aim of the competition was to find HGV drivers who consistently drove throughout the entire year, it was decided that drivers must have completed at least ten journeys every calendar quarter. In addition, a minimum of 5,000 miles of total driving across the year was required in order to qualify for inclusion in that vehicle class. The result of this pre-processing was the creation of a dataset, anonymised from the companies involved, driver names and contract numbers being passed into the next stage of the process. The final processed dataset for analysis contained 16,287 drivers.

The fields considered for evaluating driver performance were:

1. Driving time in seconds
2. Fuel used in litres
3. Distance in metres
4. Number of harsh braking events
5. Excessive idling duration
6. Excessive idling at depot
7. Excessive idling en route
8. Excessive idling at site
9. Over speed duration in seconds
10. Green band duration in seconds
11. Cruise control duration in seconds
12. Excessive throttle duration in seconds
13. Power take-off (PTO) duration
14. Number of over rev events
15. Number of journeys in the first quarter
16. Number of journeys in the second quarter
17. Number of journeys in the third quarter
18. Number of journeys in the fourth quarter
19. Median Engine Torque.

For the elimination of drivers, the fields 2) Fuel used in litres; 3) Distance in metres; 4) Number of harsh braking events; 9) Over speed duration in seconds; 12) Excessive throttle duration in seconds and 14) Number of over rev events were considered. Drivers with values far distant from the median of their peers for these fields were eliminated, as further explained in sections II-B and II-D. These fields were chosen as they are common for all vehicles in the database. For further competitions, as the technology within the vehicles improve, the attributes considered need to be revised.

B. Outlier Removal

Graphs with the attributes’ values distributions were plotted and anomalies in the data were observed, such as those presented in Figure 3. Data specialists informed us about the cutting points for the outliers, that is, those values that were unacceptable for a driver. Individuals whose attributes had higher values for these cutting points were therefore eliminated. Table I shows the thresholds defined for each variable.

C. Consensus Cluster

Clustering techniques provide a powerful means of structuring data into groupings to be used either directly (e.g., for visualisation) or applied to data classification. Clustering also
assists determining central features within the data, which can be employed to interpret the data set or to inform further data analysis steps such as inference [5].

Intuitively, Microlise data managers knew there were different profiles of drivers based on their clients needs, types of vehicles, travel distances, fuel consumption, etc. This categorisation, however, has not been previously determined and therefore could not be employed for the selection of the driver of the year. Nevertheless, it did not seem appropriate directly compare, for instance, city drivers (with smaller distances traveled, higher chances of performing harsh brakes and less likelihood of activating the vehicles cruise control) with motorway drivers who were driving longer distances with less opportunity for harsh braking events.

To avoid comparing all drivers with the same criteria and therefore to establish a fairer ranking process, we employed data clustering [7]–[9]. The objective was to identify different groups and also to select those attributes important to characterise the profiles uncovered.

Data clustering techniques are part of a wider class of algorithms often referred to as “unsupervised algorithms”, i.e. algorithms that work without ground truth. There are no predefined labels or categories for the data instances and therefore the objective of performing clustering is to suggest possible groupings of elements based on certain criteria of similarity between data instances. Two challenges are common when choosing to apply clustering to a dataset: 1) There are a variety of clustering algorithms – which one should be used for the dataset? 2) How many clusters are present in the data?

Choosing which method to use is not an easy task, as different clustering techniques identify “groupings” in data in different ways. Consequently, it has been demonstrated that the use of multiple methods is often preferable in order to extract as much information as possible from the data [6]. When using more than one algorithm, a consensus across the results is established in order to integrate diverse sources of similarly clustered data and to provide insight on the stability of the results obtained from the different techniques. We adapted part of a framework proposed by Soria et al. [6] to elucidate core groups in our dataset (Figure 4). Their framework comprises the following steps, as tailored to our problem:

**Data pre-processing.** If necessary the data is cleaned, normalised and inconsistencies are fixed. For our dataset, this step has been performed previously in sections II-A and II-B. We did not normalise the data, as we wanted to observe whether any feature would have a higher impact on the group separation.

**Clustering.** We have employed the following clustering techniques: the Hierarchical Clustering Approach [10] (HCA), K-means (KM) [11], [12] and Partitioning Around Medoids (PAM) [13], [14]. These techniques differ from each other as they have different measurements to define proximity of the data instances to establish the clusters. This means that they group the data considering different characteristics present in the data. They were chosen as they are among the most widely used clustering methods in data mining.

**Determining the number of clusters.** Validity indices are applied to clustering results in this step. As for our problem the number of clusters is unknown, in an attempt to deter-
mine an ideal value, the indices Calinski and Harabasz [15], Hartigan [7] and Scott and Simons [8] were applied to KM, PAM and HCA, for which the number of clusters is an explicit parameter, over a range of number of clusters varying from 2 to 20. According to specific rules, these validity indices indicate the appropriate number of groups to consider in the analysis. For our dataset, the majority of the indices employed indicated three clusters present in the data.

Consensus. In this step, the clusters found by the different techniques are aligned. We established those samples assigned to the same group and further validated the outcomes with domain experts. For our data, there was a high agreement between the cluster techniques; therefore we considered the results obtained by KM.

Data Visualisation. Graphs such as box-plots are employed for a general characterisation of the clusters obtained. We identified distance as a variable that plays an important role in the separation of the classes, as shown in Figure 5. The Box-plots in the figure show a clear distinction between categories of drivers based on distance. Therefore, the data could be divided in: (1) Short Distance Drivers, with distances between 5,000Km and 45,000Km; (2) Medium Distance Drivers, with distances between 45,001Km and 80,000Km; and (3) Long Distance Drivers, with distances greater than 80,000Km.

D. Outlier removal per cluster

Once the data was separated in clusters, the same procedure from Section II-B was applied to each group. Guidance from Microlise experts assisted us to define new cutting points for each of the tree driver category based on the outliers and median values observed, as shown in tables II, III and IV, for short, medium and long distance drivers, respectively.

E. Subgroups within each Group and Final Shortlist

After the grouping and the elimination of outliers, we still had 2,928 drivers on the competition (1551 in the short distance, 986 in the medium distance and 391 in the long distance categories). For further elimination, maintaining a fair criteria, we repeated the clustering process described in Section II-C for each of the three groups determined, and obtained the subgroups shown in Figure 6. The results also indicate further heterogeneity between short distance driver profiles, as more subgroups appear.

![Fig. 5. Clusters (groups) found based on the attribute Distance.](image)

![Fig. 6. Diagram of the categories of drivers found.](image)
After the subgroups of the figure were found, the final shortlist of 46 drivers was obtained by iteratively removing those with values for harsh braking, over speed, over rev and excessive throttle above the average when compared to their peers in the subgroups. The final list was sent to Microlise, in which further evaluation and ranking of those drivers was conducted via questionnaires sent to the employers.

III. Conclusion

Microlise Driver of the Year Awards [1] aimed at honouring the most talented UK HGV drivers in 2014. The competition was launched with the purpose of establishing parameters for safe and economical driving. Another objective was to encourage companies to increase safety awareness and adopt technologies and practices to assist caring for drivers, prevent accidents and reduce vehicle and cargo damages. The ranking of drivers was performed based on data gathered from Microlise telematics solutions. In this paper we reported the framework developed to shortlist the candidates for the awards.

The methods employed were a result of numerous interactions with data analysts and experts in transport and HGV drivers. We focused on defining an unbiased, fair, replicable pipeline to determine the appropriate attributes for comparison and data filtering together with clustering methods to establish sub-categories of drivers.

The pipeline consisted of seven steps: (1) Data pre-processing; (2) Outlier removal, where those drivers with anomalous values (within the data fields selected for ranking) were excluded; (3) Clustering, which defined three categories of drivers (short, medium and long distance); (4) Outlier removal within the clusters established; (5) Another application of clustering as an attempt to identify possible subgroups within the groups; (6) Elimination of drivers whose values were above average compared to drivers from the same subgroups; and (7) Final listing of drivers, who were further evaluated qualitatively. As a result, we were able to shortlist 46 drivers out of 90,000 competitors.

During the development of the methodology we also determined several improvements in the area for future years in which the awards will take place. A normalisation of technologies within the vehicles is necessary to increase the standards of safe driving across the HGV community and to improve our framework for ranking drivers. We believe therefore the effort to put together the competition and the ranking of drivers constitute the first steps to identify the set of criteria necessary to determine the profile of outstanding drivers.

IV. Acknowledgements

We would like to thank Innovate UK (Technology Strategy Board) for funding this project. In addition, we would like to thank Dr Daniele Soria (The University of Nottingham) for providing us with the clustering method. We would also like to thank Bob Harbey, Matt Hague, Ian Dickinson, Bernie Warner, Lisa Keogh, Stephen Watson, Steve Redford, Jonathan Dolby and Andree Corden from Microlise for the support and for assisting us in the validation of the framework developed.

REFERENCES