Investigating the Influence of Mobile Phone Use on Driving Behaviour with Machine Learning Analysis

Konstantinos-Eirinaios Kaselouris[1], Eva Michelaraki[1], Christos Katrakazas[1], Marianthi Kallidoni[1] and George Yannis[1]

[1]National Technical University of Athens, Iroon Politechniou 9, Athens, Greece, kkaselouris@mail.ntua.gr

This paper aims to investigate the impact of mobile phone use on driving behaviour and more specifically the sizes of the speed, acceleration and lead time through statistical analysis of imbalanced data using Machine Learning techniques. For classification and regression of mobile phone usage, telematics data from the OSeven, collected from naturalistic measurements, were used. Mobile phone use was defined as an indicator of risky behaviour and classification was performed on two levels of driving behaviour (risky and not risky). Variables related to travel speed were found to be the most significant independent variables, while according to the classification evaluation metrics, the most appropriate model was considered to be that of Linear Discriminant Analysis. Furthermore, a significant proportion of the drivers, about 25% of those recorded, used a mobile phone during driving without being aware of the activity for which they were using their mobile telephone.

Keywords: Machine Learning; road behaviour analysis; road safety; naturalistic driving; mobile phone usage; classification; regression

1 Introduction

According to the World Health Organization (WHO), every year about 1.3 million people lose their lives in road crashes and 20 to 50 million people are seriously injured as a result of the crash. The mobile phone usage is a main factor of driver distraction. Using a handheld cell phone while driving places drivers under the burden of three distinct tasks: first, the act of locating and glancing at the phone diverts their gaze away from the road; second, reaching for the phone and dialing impairs their control over the vehicle; and third, engaging in conversation via the phone distracts their attention from driving. Dialing a handheld cell phone, in particular, is an especially perilous activity, as it necessitates diverting one's eyes from the road, significantly elevating the risk of accidents and near-miss situations. The high prevalence of cell phone use during driving has led to its recognition as a primary contributing factor in automobile accidents, a topic extensively researched and discussed by the National Highway Traffic Safety Administration.

2 Data Collection

In this study, data were collected concerning 356.162 road trips in the year 2020 in Greece. The data was collected during the period of the COVID-19 pandemic outbreak. Indicators related to the drivers' rating (Stars) were extracted, as well as indicators related to the scores achieved by the drivers during the road trip (total_score, speeding_score, mu_score, hb_score, ha_score), in order to protect personal data. Finally, indicators were also removed, which have a direct dependence on each other and therefore might lead to unreliable subsequent analysis.

Table 1. Data Description

Variable Name	Measurement unit	Variable Description
Duration	sec	Total trip duration
total_distance	km	Total trip distance
risky_hours	km	Distance driven in risky hours (00:00-05:00)
ha/100km	-	Number of harsh accelerations per 100km
hb/100km	-	Number of harsh breakings per 100km
sum_speeding	sec	Total duration of speeding in a trip, i.e. driving over the “Speed Limit and Tolerance”
av_speeding_kmh	km/h	Average speed over the speed limit in a trip, i.e. driving over the “Speed Limit+ Tolerance”
avg speed	km/h	Average speed in a trip
avg_driving_speed	km/h	Average driving speed
time_mobile_usage	sec	Total duration of mobile usage in a trip
driving_duration	sec	Total duration of driving, i.e. duration of stops has been excluded
time_mobile_usage/ driving_duration	sec/sec	Total duration of mobile usage in a trip per total duration of driving
sum_speeding/ driving_duration	sec/sec	Total duration of speeding in a trip per total duration of driving

Using the available seaborn library in a Python programming environment, the following triangular heat map was drawn. Positive correlation between independent variables is denoted by shades of blue, while negative correlation is denoted by shades of brown. There is a high correlation between different descriptive statistics of the same variable. The above high correlation is logical since it concerns the relationship between different manifestations of the same item. For example, the independent variable driving_duration and the independent variable duration have a very strong correlation (r=0.98) because both are manifestations of the magnitude of driving duration.

Total_distance shows a strong correlation with the variable total driving duration either with or without stops (duration and driving_duration respectively), which follows logically from the equation of the speed definition, where driving duration and total distance travelled are proportional amounts. There is very little negative correlation between the independent variables, which demonstrates that increasing one independent variable does not decrease the magnitude of another.

Using a mobile phone while driving generally increases driving time, while it also slightly increases the average driving speed. This is due to the distraction of the driver when using a mobile phone, which leads to longer reaction times and larger speed variations. Mobile phone use has almost no effect on harsh accelerations and decelerations per 100km.

3 Methodology

3.1 Oversampling technique

The most important problem with unbalanced data is the high misclassification rate for the minority class because the classifier favours the majority class. For this reason, balancing is required to continue the classification process. In particular, in this research project the data belonging to non-use of mobile phones is much more than those of use. Out of the three methods of resampling and balancing the data, the Oversampling method through the Synthetic Minority Oversampling (SMOTE) technique was finally chosen. In particular, the Undersampling method was not preferred in this research project for fear of losing important information, as the variation in variable values is strong.

3.2 Classification Analysis

In this subsection, data analysis is performed with mobile phone use as the dependent variable. More specifically, from the initial data provided by OSeven Telematics the requirement arose to create a binary variable, which will be the dependent variable for the classification. This continuous variable is the duration of mobile phone usage, which was transformed into the binary variable of mobile phone usage with the following values: i.e. value 0: For no mobile phone use during driving, value 1: For using a mobile phone while driving.

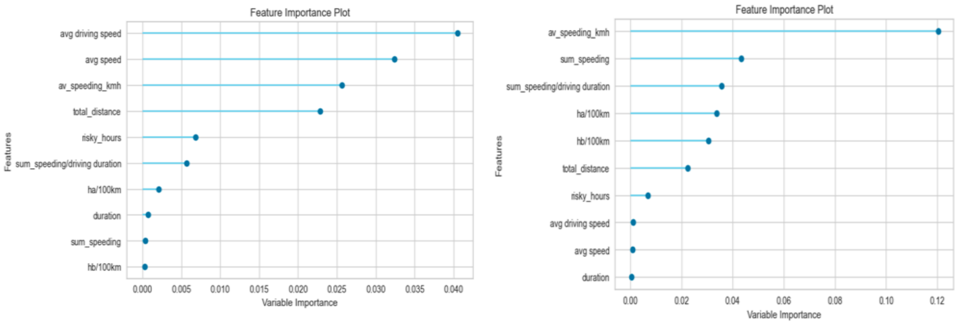

Fig. 1. Variable importance with dependent variable time_mobile_usage with Linear Discriminant Analysis and Logistic Regression

3.3 Regression Analysis

From the Figures below, the 5 most important independent variables were selected for further processing and analysis and the regression analysis was repeated only with these variables.

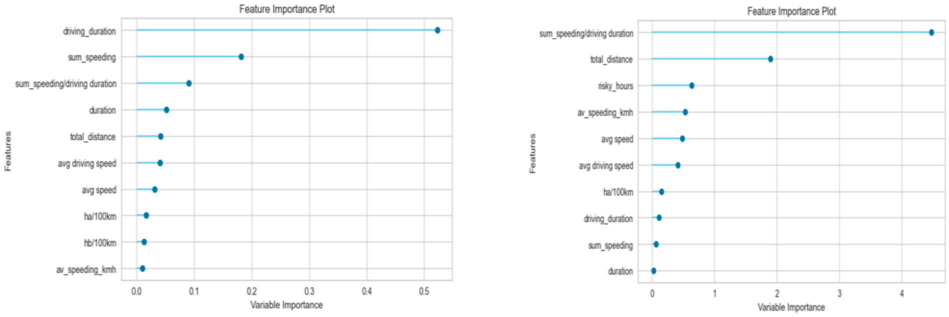

Fig. 2. Variable importance with dependent variable time_mobile_usage with AdaBoost Regressor and Linear Regression

4 Results

Tables 2 and 3 summarise the results of the classification and regression algorithms obtained through the statistical analysis of the data.

Table 2. Summary table of classification models.

Classification		Linear Discriminant Analysis	Logistic Regression
With all independent variables	Accuracy	84.4%	83.4%
	Precision	73.3%	67.4%
	Recall	65.1%	72.4%
	FNR	34.6%	26.4%
	F1-Score	68.8%	70.0%
	AUC Score	89.5%	89.1%
With most significant variables	Accuracy	83.5%	98.2%
	Precision	72.0%	99.8%
	Recall	62.4%	93.5%
	FNR	37.2%	6.4%
	F1-Score	66.9%	96.6%
	AUC Score	87.7%	99.9%

According to Table 2, the most suitable algorithms for the prediction of driving behaviour was the Linear Discriminant Analysis model with all the variables under consideration, which showed a high and realistic accuracy rate (realistic accuracy values range from 70% to 90% according to the international literature), as well as a moderate rate of False Negatives.

Table 3. Summary table of regression models.

Regression R2	AdaBoost	Linear Regression
With all independent variables	0.842	0.497
With the most significant variables	0.840	0.422

From Table 3 it can be noted that the Adaptive Empowerment model is the optimal regression model as it has a very satisfactory R2 coefficient of determination value both with all the variables under consideration as well as with the most significant ones. In general, satisfactory R2 values are considered to be those in the range from 0.8 to 0.9, because higher values may indicate overfitting and for values less than 0.8 the model becomes less reliable. It is worth mentioning that the coefficient of determination R2 becomes the most informative simple metric in the evaluation of regression analyses (Chicco et al., 2021).

5 Conclusions

The main influencing parameter of mobile phone use according to the classification models turned out to be speed (km/h). This makes sense, as the duration of mobile phone use transformed into the binary variable of mobile phone use leads to driver distraction and indirectly influences driver speed either by increasing it by exceeding its limit or by decreasing it, confirming the international literature.

The main parameters influencing the duration of mobile phone use according to the regression models were both the total driving time with exceeding the speed limit and tolerance per unit of non-stop road trip duration (sec/sec) and the non-stop road trip duration (sec). The duration of mobile phone use increases as a road trip increases. Furthermore, there is a dependence between the duration of mobile phone use and the duration of driving above the speed limit, as it has been observed that drivers whose attention is distracted by their mobile phone increase their reaction time and lead to more abrupt driving behaviour and thus to speeding.

In this study, two classification algorithms and two regression algorithms were trained with all the variables under consideration as well as with the most important ones. The best algorithm for the classification models was the Linear Discriminant Analysis model, while for the regression the AdaBoost Regressor model gave more reliable results. The total distance driven affects the duration of mobile phone use, as evidenced by the Pearson's triangular matrix combined with the significance of the variables, and this is explained by the fact that the greater the road distance driven, the more time the driver has to use his/her mobile phone. In addition, by distracting the driver with the mobile phone, he can take a longer route to reach his destination. It was observed that mobile phone use had almost no effect on sudden incidents and more specifically harsh accelerations and decelerations in drivers using a mobile phone. This may be due to the fact that being preoccupied with the mobile phone sometimes causes a reduction in speed, as confirmed by the international literature (Gazder & Assi, 2022).

Dangerous driving hours, i.e. between 00:00 and 05:00, do not seem to have much effect on mobile phone use. Despite the general occurrence of symptoms of dangerous driving behaviour during this period, drivers do not seem to be particularly concerned with their mobile phone, so it is not practically responsible for causing serious crashes during this period. Lastly, it was observed that the regression algorithms with the most appropriate variables showed a lower coefficient of determination R^2 compared to the same algorithms considering all the independent variables under consideration. This conclusion lies in the fact that the R^2 coefficient indicates the predictive ability of the model with the more independent variables.

References

1. Chicco,D. ,Warrens, M. J. & Jurman, G. (2021), The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. Available at: https://peerj.com/articles/cs-623/
2. Gazder U. & Assi J. K. (2022), Determining driver perceptions about distractions and modeling their effects on driving behavior at different age groups, Available at: https://doi.org/10.1016/j.jtte.2020.12.005
3. Khan, I., Rizvi, S. S., Khusro, S., Ali, S., & Chung, T. S. (2021). Analyzing drivers' distractions due to smartphone usage: evidence from AutoLog dataset. Mobile Information Systems.Available at: https://www.hindawi.com/journals/misy/2021/5802658/
4. Kontaxi ,A., Ziakopoulos, A., Katrakazas, C. & Yannis, G. (2022). Measuring the impact of driver behavior telematics in road safety. Available at: https://fersi.org/wp-content/uploads/2022/10/Armira-Kontaxi-et-al.pdf
5. Michelaraki, E., Sekadakis, M., Katrakazas, C., Ziakopoulos, A. & Yannis, G. (2023). One year of COVID-19: Impacts on safe driving behavior and policy recommendations. Available at: https://doi.org/10.1016/j.jsr.2022.10.007
6. Scikit-learn: Machine Learning in Python[WWW Document] (2022). Available at: https://scikit-learn.org/
7. Seaborn: statistical data visualization [WWW Document], 2022. Available at: https://seaborn.pydata.org/
8. Stevens, M., Sunseri I. & Alexanderian, A. (2022). Hyper-differential sensitivity analysis for inverse problems governed by ODEs with application to COVID-19 modeling. Available at: https://doi.org/10.1016/j.mbs.2022.108887
9. Yannis G., Laiou A., Papantoniou P., & Christoforou C. (2014). Impact of texting on young drivers' behavior and safety on urban and rural roads through a simulation experiment. Available at: https://doi.org/10.1016/j.jsr.2014.02.008
10. Ziakopoulos, A., Kontaxi, A. & Yannis, G. (2023), Analysis of mobile phone use engagement during naturalistic driving through explainable imbalanced machine learning, Available at: https://doi.org/10.1016/j.aap.2022.106936