3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11514

Regression scores to identify risky drivers from braking pulses

Shuai Sun[1,2], Jun Bi[1], Montserrat Guillen[2], Ana M. Pérez-Marín[2]
[1]Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Beijing Jiaotong University, China, [2]Riskcenter, Universitat de Barcelona, Spain.

Abstract

Driving data record information on style and patterns of vehicles that are in motion. These data are analysed to obtain risk scores that can later be implemented in insurance pricing schemes. Scores may also be used in on-board sensors to create risk alerts that help drivers to keep up with safety margins. Regression methods are proposed and a prototype real sample of 253 drivers is analysed. Conclusions are drawn on the mean number of brake pulses per day as measured within 30 seconds time-intervals. Linear and logistic regressions serve to construct a label that classifies drivers. A novel factor based on the driving range that is defined from geo-localization improves the results considerably. Driving range is expressed as measures the diagonal of a rectangle that contains the furthest North-South versus East-West weekly vehicle trajectory. This factor shows that frequent braking activity is negatively related to the square of driving range.

Keywords: *Telematics; logistic regression; insurance; risk measures; traffic safety.*

1. Introduction

The internet of vehicles is a new area providing lots of opportunities to develop big data applications before self-driving cars are fully available. In this paper, we analyze a set of 253 drivers that were monitored over one week. Data were collected every thirty seconds, including the position, speed, acceleration and the engine's revolutions per minute. No information on accident was available. We propose a new way to design driving risk alerts based on these data. Our predictive risk scores can serve as inputs to improve driving habits and to calculate insurance premiums by insurers in the Internet of Vehicle (IoV) environment. We also suggest that on-board vehicle telematics should encompass personalized risk-related alerts in their internal architecture.

Risk analysis in motor insurance or accident prevention usually studies traffic collisions (Handel et al, 2014). Some papers relate driving patterns or braking to accident risk (Jourbert et al., 2016; Guillen et al., 2019), as braking usually occurs before an accident happens, or before an accident is avoided.

Generalized linear models, have been used to predict the probability of a traffic accident or, alternatively, the expected frequency of insurance claims. As such, this is the main technique implemented by insurance companies to calculate the expected yearly number of claims and average cost, from which the basic premium prices are obtained (see, Jin et al., 2018; Verbelen et al., 2018; Ma et al., 2018). Paefgen et al. (2014) compared the performance of various machine learning methods, such as logistic regression, neural network, and decision tree classifiers, in driving risk prediction and insurance pricing. The interpretability of logistic regression models has made this method the outstanding technique to calculate risk scores. In many countries premium calculation is regulated and, as a result, the authorities prefer methods that are not black boxes (see, also Pesantez-Narvaez et al., 2019).

2. Data

Data used in our study were obtained from an IoV information service provider in China. Each vehicle in our database has a telematics box (T-box), including a GPS sensor, a vehicle condition sensor, and a wireless transmission unit. When the vehicle is turned on, data get recorded second-by-second and then they are aggregated on the device level to reduce costs of data transmission and storage. Every 30 seconds, the T-box transmits the latest piece of data to a central database. When the vehicle is turned off, the on-board device automatic restarts every 30 minutes and transfers a bunch of data to the base station.

The total number of effective vehicle data files is 253. Our statistical learning has to deal with unique vehicle identification, time-varying GPS trajectory data, and abundant vehicle

condition information. A summary of the per-vehicle averages can be found in Table 1. We also show in Figure 1 a Google Maps example of car travel trajectories in a Chinese region.

Table 1. The basic descriptive statistics of the variables in the driving data set 2019.

Variable	Definition	Mean	Standard Deviation	Minimum	Median	Maximum
Brakes	Brake times with speed>40km/h	1540.194	1266.109	26.000	1162.000	6633.000
Accelerator	Mean of acceleration pedal position (%)	19.640	7.313	0.185	20.124	39.480
Distance	Cumulative driving distance (Km)	2211.046	1578.700	17.140	1975.570	7163.830
Speed	Mean of speed (Km/h)	36.076	15.225	1.187	36.123	66.819
RPM	Mean of revolutions per minute	997.390	178.219	232.263	983.173	1622.257
Range	Range of driving (geographical units)	3.050	3.334	0.013	1.706	14.593

Source: Own calculations

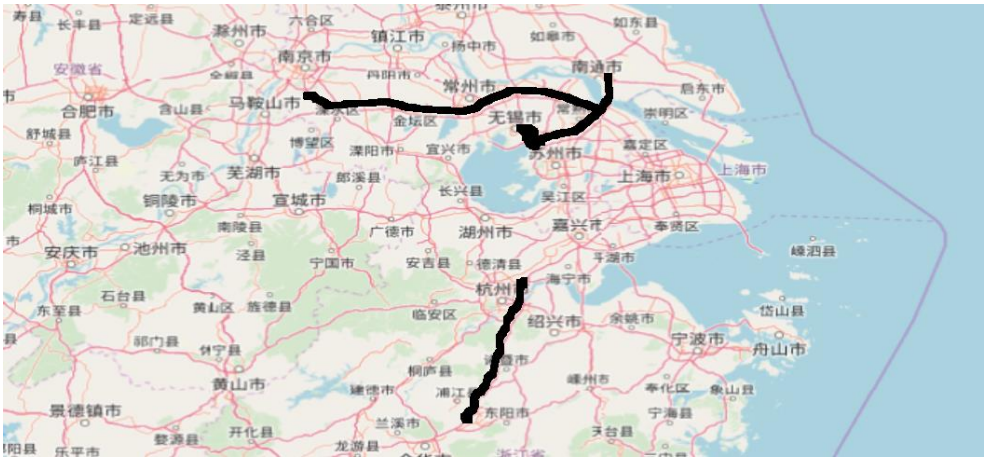

Figure 1. This map plots tow trajectories of monitored cars in July 2019.

Figure 2 presents the development of distance driven by one car in the upper plot. In the middle plot, there is a graph of the fuel consumption accumulated over time and finally, the speed and brake pulses are shown in the lower plot. For all cars in the sample, a daily average and variance was computed.

In addition, a new measure was defined to capture the driving pattern regarding the fact that a driver always stays in the same region. Driving range has a major influence on the analysis of driving risk. Whenever drivers brake or accelerate, there is either traffic congestion or moving requirements in limited area. However, this behavior must be relative to the driving radius. Indeed, a large number of brakes or accelerations for someone who drives longer distances, in a large radius from the starting home point may indicate that there may be safety hazards, compared to someone staying in a short circle distance. Driving range was calculated based on the available GPS trajectory as follows:

$$Range = \sqrt{(lon_{max} - lon_{min})^2 + (lat_{max} - lat_{min})^2}$$

where lon_{max} and lon_{min} represents the maximum and minimum observed Longitude value, lat_{max} and lat_{min} represents the maximum and minimum observed Latitude value.

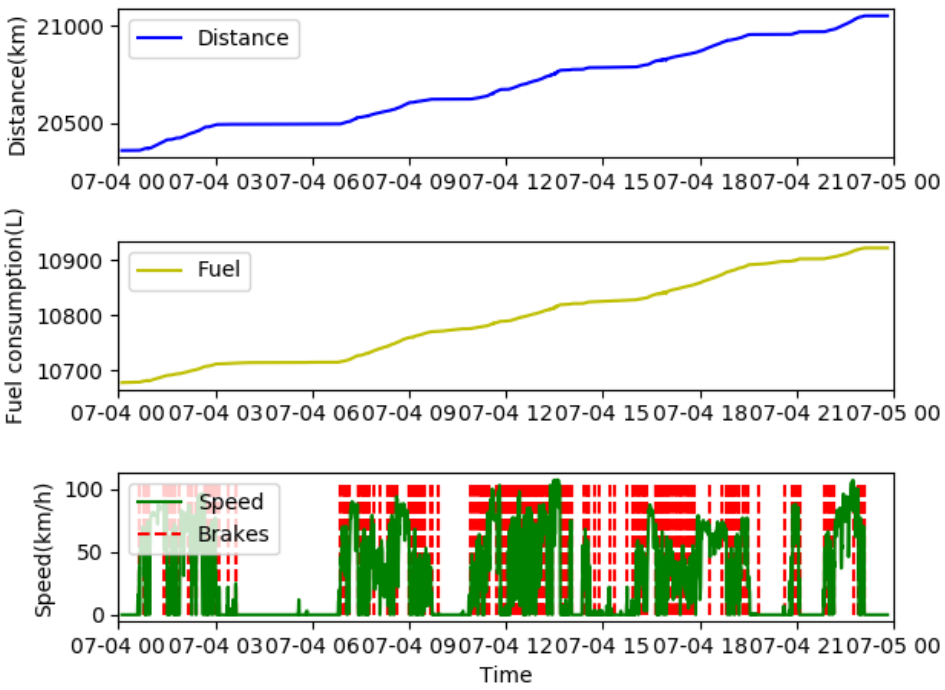

Figure 2. Example of accumulated distance driven (upper), fuel consumption (middle) and speed/breaking activity measured for a vehicle in July 2019.

3. Methods and results

OLS regression and logistic regression were respectively estimated by taking "Brakes" as the dependent variable. The variables derived from the multiplication of two pairs are taken as independent variables to reflect the interaction between variables.

In the logistic regression model variable "Brakes" was dichotomized as follows. Brake levels above the median were identified as events, while those below the median were identified as non-events. The stepwise regression results after bidirectional elimination are shown in Table 2. The goodness-of-fit (pseudo-)R^2 statistic is 34% and 27% for the linear and logistic regression, respectively.

Table 2. Coefficient estimates and P-values for linear regression (left) and logistic regression (right) in the driving data set 2019.

	OLS		Logistic	
Variable	**Coefficient**	**P-value**	**Coefficient**	**P-value**
Intercept	-359.5976	0.173	-2.1345	0.000
Distance	1.4935	0.003	0.0016	0.000
Speed	-85.1223	0.045		
Accelerator	99.5726	0.014		
$Distance^2$	8.574e-05	0.012		
$Speed^2$			-0.0029	0.001
RPM^2			-5.309e-06	0.005
$Accelerator^2$			-0.0094	0.024
Distance*Speed	-0.0180	0.000		
Distance*RPM	-0.0018	0.006		
Distance*Accelerator	0.0555	0.000		
$Range^2$	-8.7141	0.007		
Range*RPM	0.2533	0.002		
Range*Accelerator	-7.2562	0.018		
Speed*RPM	0.1516	0.003	0.0001	0.045
Speed*Accelerator	-3.1860	0.002		
RPM*Accelerator	-0.0818	0.019	0.0004	0.017

Source: Own calculations.

A comparison between intuitive judgement and in-sample prediction shows that there is a mismatch between the conditional scores produced by the multivariate models, i.e. when taking into account the driver's information, and the intuitive judgment that is solely based on the univariate analysis of brake pulses.

This is easily seen in Figure 3, where there is a comparison between the predicted scores provided by the models (vertical axis) and the value of the observed brakes (horizontal axis).

A risky driver has a predicted score that is lower than his observed braking value. A non-risky driver has a predicted score that is higher than his real observed breaking value. Alternatively, the median of the scores and the median of the observed braking values are used as classifiers as shown in Figure 3. Drivers in the right bottom box are the ones that would be identified by our models as risky drivers. The two models produce slightly different results in terms of identified risky drivers.

Sensors should dynamically react to large values of brake pulses. The regression line or the logistic curve should act as the fundamental pieces of personalized alert systems. For example, a driver that has brake some of pulses equal to 300, but whose risk score predicts a total of 200, should be warned because his level of braking activity is above the risk limit predicted by the model.

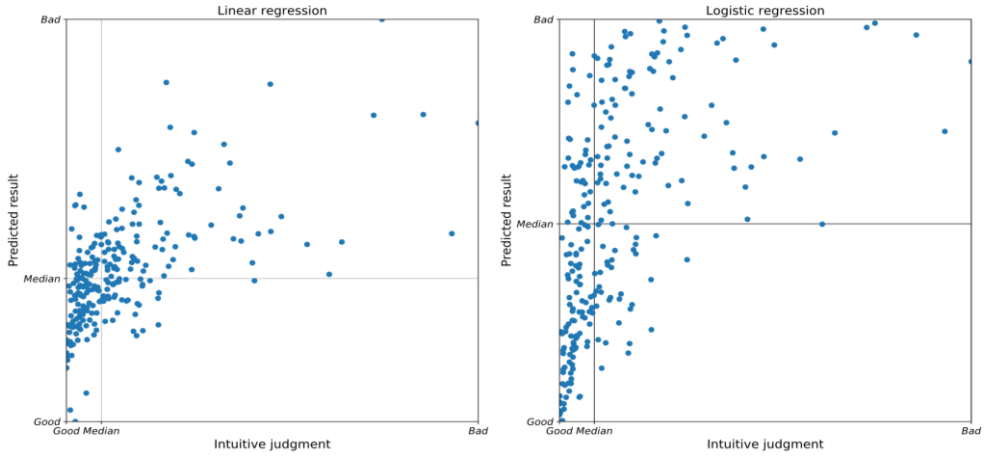

Figure 3. Comparison of predicted versus observed scores for "Brakes" in the telematics data set. Left plot corresponds to linear regression and right plot corresponds to logistic regression.

The points in Figure 3 indicate that the response variable has some outlying observations and that it is right skewed. The difference between the intuitive judgement, which corresponds to the purely observed values (horizontal axis) and the predicted scores (vertical axis) is that the later take into consideration the driving characteristics included in

the model. Those drivers exhibiting especially good habits and observed values below the predicted scores should be rewarded by the insurance company due to good driving habits. Insurance companies may deny access to insurance for drivers who cause extreme driving hazards in order to diminish claims.

Scores can also serve to reveal problems in the vehicle sensors. We argue that some on-board devices suffer quality deterioration over time. Predictive scores that are systematically above the levels of observed braking activity may indicate that the sensor fails to transmit the true driving activity correctly. Driving data providers should be aware that scoring drivers benefits traffic safety, insurance companies and their own business quality control.

4. Conclusions

Accident risk analysis is difficult because collisions and crashes seldom occur. The inspection of IoV data even if there is no information on motor accidents can be done by comparing drivers' ratings and observed patterns. Basic machine learning models were used to classify observations and to identify risky clusters of drivers in the sample. The mean of the braking pulses when the vehicle exceed 40 Km/h was used as a response variable and it was also dichotomized to reveal an association with other driving factors. This solution is promising for insurers and even car manufacturers that design new safety procedures and gadgets. Data analysis of a big source of information when accident data for vehicles were not available is still valuable to produce relevant scores. The relationship between accident risk has been found in previous studies. So, the level of braking pulse intensity can be related positively with proportionally higher insurance prices (Bian et al., 2018; Carfora et al., 2019; Tselentis, 2016 and 2017).

The linear relationship between the response variable and the covariates is a limitation of the linear regression model. Further analysis of more flexible specifications is recommended. Pérez-Marín and Guillen (2019) showed that excess speed is one of the main factor influencing driving risk, however the results obtained for this dataset indicate that mean speed is inversely related to braking, but positively related to braking when interacting with engine RPM. The higher the speed, the more acceleration action is required. Moreover, the higher the RPM, the more braking action is need to reduce speed, and so the greater the driving risk.

Driving range was not informative by itself, but it did have a substantial influence when combines with other factors.

Some of the limitations of this case study are related to the model approaches. Other machine learning algorithms and classifying techniques should also be examined. Some

additional efforts remain in the research agenda. For example, geolocation information could be used to define driving zones (urban versus nonurban) and time stamps could be transformed into day and night driving percent. The volume of information contained in each vehicle daily file opens an opportunity to explore patterns that may improve driving habits and produce recommendations for safety on the road.

Funds

The paper was granted by the Fundamental Research Funds for the Central Universities 2019YJS091.

Acknowledgements

MG thanks ICREA Academia. The authors thank Fundación BBVA and the China scholarship council.

References

Bian, Y., Yang, C., Zhao, J. L., Liang, L. (2018). Good drivers pay less: A study of usage-based vehicle insurance models. *Transportation Research Part A: Policy and Practice*, 107, 20-34.

Carfora, M. F., Martinelli, F., Mercaldo, F., Nardone, V., Orlando, A., Santone, A., & Vaglini, G. (2019). A "pay-how-you-drive" car insurance approach through cluster analysis. *Soft Computing*, *23*(9), 2863-2875.

Guillen, M., Nielsen, J. P., Ayuso, M., & Pérez-Marín, A. M. (2019). The use of telematics devices to improve automobile insurance rates. *Risk Analysis*, 39(3), 662-672.

Handel, P., Skog, I., Wahlstrom, J., Bonawiede, F., Welch, R., Ohlsson, J., et al. (2014). Insurance telematics: Opportunities and challenges with the smartphone solution. *IEEE Intelligent Transportation Systems Magazine*, 6(4), 57-70.

Jin, W., Deng, Y., Jiang, H., Xie, Q., Shen, W., & Han, W. (2018). Latent class analysis of accident risks in usage-based insurance: Evidence from Beijing. *Accident Analysis & Prevention*, 115, 79-88.

Joubert, J. W., de Beer, D., de Koker, N. (2016). Combining accelerometer data and contextual variables to evaluate the risk of driver behaviour. *Transportation Research Part F: Traffic Psychology and Behaviour*, 41, 80-96.

Ma, Y., Zhu, X., Hu, X., Chiu, Y. (2018). The use of context-sensitive insurance telematics data in auto insurance rate making. *Transportation Research Part A: Policy and Practice*, 113, 243-258.

Paefgen, J., Staake, T., Fleisch, E. (2014). Multivariate exposure modeling of accident risk: Insights from Pay-as-you-drive insurance data. *Transportation Research Part A: Policy and Practice*, 61, 27-40.

Pérez-Marín, A. M., Guillen, M. (2019). Semi-autonomous vehicles: Usage-based data evidences of what could be expected from eliminating speed limit violations. *Accident Analysis and Prevention*, 123, 99-106.

Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression. *Risks*, *7*(2), 70.

Tselentis, D. I., Yannis, G., Vlahogianni, E. I. (2016). Innovative insurance schemes: pay as/how you drive. *Transportation Research Procedia*, 14, 362-371.

Tselentis, D. I., Yannis, G., Vlahogianni, E. I. (2017). Innovative motor insurance schemes: A review of current practices and emerging challenges. *Accident Analysis and Prevention*, 98, 139-148.

Verbelen, R., Antonio, K., Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5), 1275-1304.