Machine Learning Based Real-Time Vehicle Data Analysis for Safe Driving Modeling

ABSTRACT

This paper identifies a necessity to evaluate the Meta features of vehicles which could be helpful in improving the vehicle driver's skill to prevent accidents and also evaluate the change in the quality of cars over passing time. This paper does an analysis of the vehicle data using supervised learning based linear regression model that is used as an estimator for Driver's Safety Metrics and Economic Driving Metrics. The data collected was obtained from fifteen different drivers over a span of one month which accumulated over 15000 data points. And the metrics that we have devised have potential application in automotive technology analysis for developing an advanced intelligent vehicles. Also, we have presented a system for performing the real-time experiment based on the On-Board-Diagnosis version II (OBD-II) scanner data. Finally, we have analyzed and presented the parameter accuracy over 80% for the driver's safety solution in real-world scenario.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *OBD-II, Dashboard Camera* • **Networks** → *Vehicular networks*, Intelligent Transportation System

KEYWORDS

Supervised Learning, Linear Regression, Statistical Analysis, Automotive Vehicle Data

ACM Reference format:

P. Yadav, S. Jung, D. Singh. 2019. In *Proceedings of ACM SAC Conference, Limassol, Cyprus, April 8-12, 2019 (SAC'19)*, 4 pages. DOI: 10.1145/3297280.3297584

1 INTRODUCTION

Automotive Technologies are providing improvised services to the driver's safety and vehicle security under the umbrella of Intelligent Transportation System (ITS). In the development of ITS, advanced Automotive Technologies shall play a crucial role in determining the overall experience of users by making it much at ease in terms of reducing the risk of road accidents, risk of cybercrime in the vehicle, buying a used car etc. It is often noted that judging the driver's driving skill is subjective and is difficult to set a standard for driver's skills [1]. The modern approach to transportation system is focusing on rapidly evolving with the intelligent vehicles. High rise in recorded traffic density, road accidents and crisis faced in regulating the effective management of traffic control in urban and rural areas have concerned us to develop a smart solution in context to ITS [2]. The automotive industry has great expectations from these futuristic solutions to improve the safety of people and security of vehicles. It is observed that the users are shifting from individualistic approach to the data-centric approach based on OBD-II scanner to avail the augmented driving experience. In spite of the modern command, control, communication, computers and intelligent systems, we are still facing numerous calamities in which thousands of precious human lives are lost in accidents. Therefore, it should be an immediate need to tackle the small scale yet serious issues using the state-of-the-art techniques. We are mainly focusing on analyzing the data which is collected from the vehicle using the OBD-II scanner and eventually providing the driver's safety solutions. We aim to obtain the solutions by observing the blind-spots accurately and efficiently using pattern recognition techniques from supervised learning.

In the paper, Section 1 describes the problem statement and brief analysis of the problem which we have attempted to work on in this paper. Section 2 discusses the system design used to solve the problem, which encapsulates the definition of the machine learning model, tuning parameters and the overall data analysis workflow that maintains the order of the whole computing process. Then, in Section 3 we made a detailed performance analysis in terms of accuracy of our solution. It covers the data generation, data preprocessing techniques and algorithms employed in obtaining the results. And finally, in Section 4 we concluded the paper with the overall insights and potential applications of the predictive modelling performed in the experiment.

2 MACHINE LEARNING BASED EXPERIMENTAL MODEL

2.1 System Design

The proposed system consists of some external hardware devices such as OBD-II scanner and Mini-dash camera, which are employed to gather the driving and vehicular data. OBD-II gathers various data related to vehicle performance such as speed, acceleration, idle time of engine, fuel consumption, distance travelled etc. Mini-dash camera gathers the real-time images and videos of the events happening in the surrounding of the vehicle. Then the gathered data is fed into our utility platform which transmits the data to the cloud server where the analysis is done and desired output is produced for our experiment. These results are shared with the driver, insurance companies, and other entities of interest.

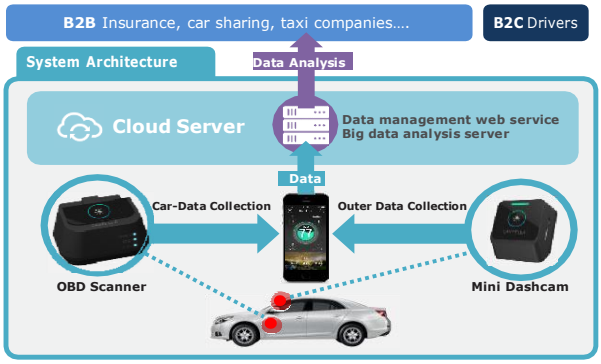

Figure 1: Overview of the OBD-II data collection and processing system.

Therefore, it is a challenge to obtain the correct and meaningful knowledge out of the data collected. With the growing community interest in Machine Learning, these techniques have been employed to obtain plausible results and further advance the knowledge domain in many industries to tackle some important challenges. These challenges may include and extend beyond driver's safety performance, estimation of car's life, fuel efficiency, and long-distance driving efficiency, all of them involve parametric learning of real-world vehicle related datasets. All the same the mini-dash cam has the additional function of recording the video; it can store the location of accident and condition of the vehicle in case of rear-end collision by adding the video information of head-on collision and GPS in the event of an accident. The system design has been categorized under the following steps.

2.2 Machine Learning Model

We have used supervised learning algorithm to the known target values (labels) for a problem. In order to train such a model which can be identified as the vehicle parameters – preferably with a variety of configurations – are required as input variables. Regression models includes determining continuous numerical values based on multiple input variables, for e.g., in a car, calculating its ideal speed to minimize the fuel consumption according to the road conditions, determining a financial indicator such as gross domestic product based on a changing number of input variables (use of arable land, population education levels, industrial production, etc.), and determining potential market shares with the introduction of new models. We have used a linear regression model for finding a future (unknown) values, where the model assumes a relationship between the dependent variable (in our case, economic driving index or driver safety index) and the independent variable (in our case, arithmetic combination of weights and features) [3].

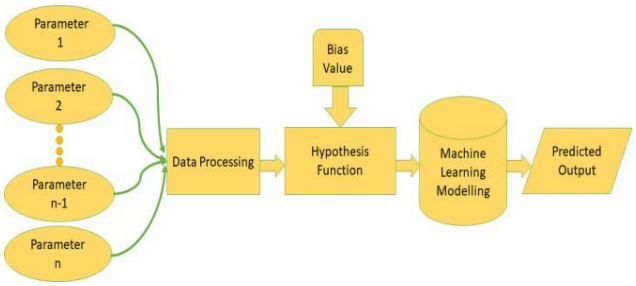

Figure 2: Machine learning model for predicting output values.

2.3 Hyperparameter Tuning

All of the data was processed as described above and then the hypothesis function was created to which a bias value was added. We used the Multivariate Linear Regression model as our base model to train on the data, since we found mostly a linear to quadratic relationship among the relevant features. To reduce the computation power and complexity of the training, linear regression model was found to be the best choice. Hyperparameter tuning was done using 'GridSearchCV' in ScikitLearn Library [8]. The grid parameter made an evaluation of all 3 × 5 = 15 combinations of n_estimators and max_features. The hyperparameters were fine-tuned automatically without involving too much human-labor tuning.

2.4 Process Structure

Data Collection involves OBD-II scanner for gathering data such as fuel efficiency, speed value parameters etc. We applied normalization and standardization methods to fit the data in the model. Then the processed data goes through the feature selection process, which was evaluated using correlation matrix. After obtaining the feature classes, the data was fed into the supervised learning based Linear Regression model. The model predicted the required output values which were then sent over the cloud to the Database servers where all the processing is done, whereafter the processed information is transmitted to the driver, insurance companies, nearby police stations. This overall system serves as a database, monitoring and alert system to prevent accidental risk and monitor the car's health [4].

3 PERFORMANCE ANALYSIS

3.1 Data Used

We analyzed the data collected over a span of one month from fifteen different drivers. Data was collected using the OBD-II scanner installed in the test vehicles, developed by MtoV Inc., Korea. In the

original dataset, we have recorded 51 features and after doing a correlation analysis we obtained some of the features such as fuel efficiency, average speed value, maximum speed value, fourth section speed value, interval driving distance, driving time value during green zone, travelling time value, emergency accelerated value, emergency decelerated value, fourth rpm time value and fifth rpm time value [5]. In our dataset, the total number of instances recorded were approximately 6000 which was divided in 80:20 ratio of training and test set. We derived the useful features for our scope using a correlation matrix.

TABLE 1: Correlation matrix of SFTY_DRVG_INDX

SFTY_DRVG_INDX	Correlation Matrix	
	Features	**Correlation Value**
ROF_XCH_SCR	Fuel Efficiency	1.000000
AVG_SPD_VAL	Average Speed Value	0.951148
MAX_SPD_VAL	Maximum Speed Value	0.850964
SPD_TH4_DRVG_TIM_VAL	Fourth Throttle Driving Time Value	0.676948
IDL_HCT	Idle Time Value	-0.296508

Table 1 shows the correlation values for our hypothesis #1, Safe driving Index. And **Table 2** shows the correlation values we used for our hypothesis #2, Economic driving Index (ECN_DRVG_INDX). We found that Fuel Efficiency in the following table showed a 100% correlation with our hypothesis #1, but considering it alone for the training would lead to over fitting of the data and eventually reduce the efficiency of the model. Therefore it was important to consider all the features.

TABLE 2: Correlation matrix of ECN_DRVG_INDX

ECNM_DRVG_INDX	Correlation Matrix	
	Features	**Correlation Value**
TH5_RPM_TIM	Fifth Throttle RPM time	-0.567331
UGY_ACSD_OFT	Urgent Acceleration Number	-0.615989
UGY_RDSD_OFT	Urgent Deceleration Number	-0.621209
TH4_RPM_TIM	Fourth Throttle RPM time	-0.563859

3.2 Data Processing

We did the data processing by cleaning and eliminating the non-relevant features for our hypothesis. We performed normalization on the data where the values were shifted and modified on the scaling so that the range ends up within 0 to 5000 in case of hypothesis 1 and 0 to 200 in case of hypothesis 2 according to the different set of values. This was done by subtracting the minimum value and dividing by the maximum minus the minimum.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

3.3 Hypothesis Generation

After generating a correlation matrix, we found the appropriate features that were considered in our hypothesis generation. Following are our hypotheses:

$$h_1(x) = \sum_{i=1}^{i=5} x_i * \beta_i + bias \quad \text{(i)}$$

$$h_2(x) = \sum_{i=1}^{i=4} x_i * \beta_i + bias \quad \text{(ii)}$$

We hypothesize an outcome called Economic Driving Index (ECN_DRVG_INDX) represented using **h1** and another outcome called Safe Driving Index (SFTY_DRVG_INDX) represented using **h2**. Based on our correlation values, we frame a linear regression-based hypothesis where the feature values are represented using xi, the weights for each feature are represented using β.

$$y_1 = [\beta_1\beta_2\beta_3\beta_3\beta_4\beta_5]. \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} \quad \text{(iii)}$$

The weights are assumed in the initial stage but are refined with every iteration of the model while the initial noise in the system is assumed to be a bias value. However, before performing the product of features and weights, we need to perform an inverse operation as shown in **(iii)** and **(iv)** on the feature matrix due to the difference in the data storage formats.

$$y_2 = [\beta_1\beta_2\beta_3\beta_3\beta_4]. \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \quad \text{(iv)}$$

We added the bias value to represent an arbitrary but appropriate value to model a real-time scenario of roads. Based on the features that we have considered for deriving the hypothesis we found that five of the features are useful for EDI while four of the features are useful for SDI.

4. RESULTS AND DISCUSSION

The result analysis consisted of the collection of data from the OBD-II scanner through the app, which was then processed into the machine learning model and finally trained as shown in Fig. 3. Trained Output values were used as the benchmark for testing against the gathered data. To do so, we performed a k-fold cross-validation technique with k=10, to train the model. We performed several experiments on the parameters which are essential for the testing of vehicle's safety and economic efficiency. In our first experiment, a relationship between Maximum speed value and the travel time (red zone) is obtained. This relationship describes the total distances travelled while crossing the road given the signal was red. We clearly

\

observed, as shown in Fig. 4, that given the speed of the car was high, it was more likely for the driver to cross the road at red signal and in turn this implies the increasing likelihood of meeting with an accident.

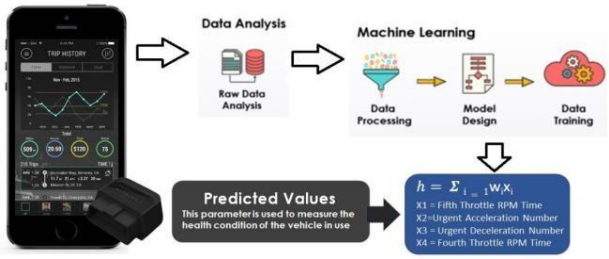

Figure 3: A schematic diagram of the analytics process.

We performed a k-fold cross-validation technique [6] with k=10, to train the model. In the Hypothesis-1, ECN_DRVG_INDX, we found that majority of the data was congested in the lower left part of the graph suggesting an inverse logarithmic growth of the trend based on the training data. This showed a positive growth of the ECN_DRVG_INDX based on the hypothesis value. However, the data scatters as the value on x-axis increases, hinting at a somewhat lesser correlation for predicted value based on hypothesis value. Therefore, the ECN_DRVG_INDX is found to be an inverse logarithmic function of the features.

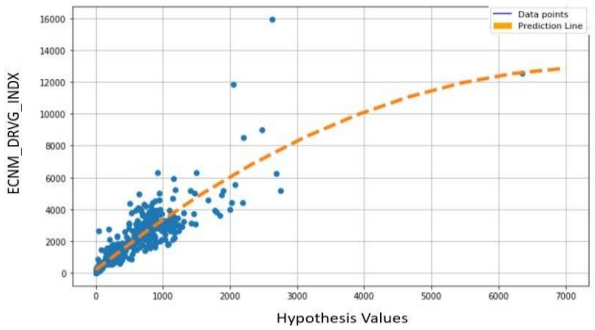

Figure 4: Linear regression based prediction trend of ECN_DRVG_INDX against hypothesis value.

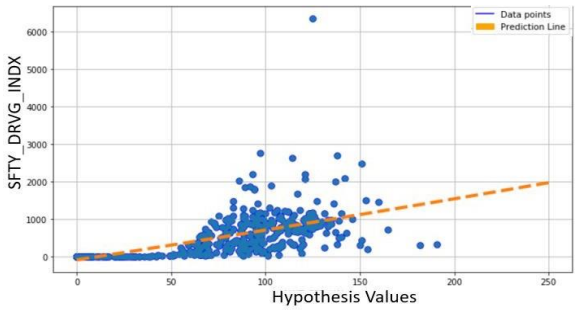

Figure 5: Linear regression based prediction trend of SFTY_DRVG_INDX against hypothesis value.

In the Hypothesis-2, safety driving index (SFTY_DRVG_INDX), we observed a slower initial growth in the trend but rapidly picks up the momentum after crossing a certain threshold value. After this value the trend moves in a logarithmic fashion up towards the positive y-axis. We do observe few outliers that were safely ignored for consideration. Therefore, the safety driving index (SFTY_DRVG_INDX) is found to be a logarithmic function of the features considered in our hypothesis. However, a good example for future work would be to obtain different datasets on vehicle metrics and train our model on that data to further improve the accuracy from 80% to more than 90%. However, an accuracy of 80% in the economic driving index, which represents a car's health condition, is good enough to include a human-in-the-loop for the prediction work. We aim to implement our model in an IoT system [7] for driving school scenario to judge the driver driving ability and frame a standard whether the driver should be eligible to obtain a driver's license or not.

5. CONCLUSIONS

In this paper we have obtained some newer insights about the car data analysis such as economic driving index (ECN_DRVG_INDX) and safety driving index (SFTY_DRVG_INDX.) The results have proven to be approximately 80% fitting the given features and are very helpful to be used in different use cases such as a parameter in finding the driver's driving performance in a driving school, as a good estimate for finding an optimal price for a used car that can be based on several factors which we have analyzed in this paper etc. We also found that the model used to train the data can be improved further by finding better hyper parameter values for the features. It is also possible that different features can be considered for improving the hypothesis.

ACKNOWLEDGMENT

This work was partially supported by the MtoV Inc., VESTELLA Inc., and Hankuk University of Foreign Studies research fund.

REFERENCES

[1] Singh D, Singh M., "Internet of Vehicles for Smart and Safe Driving", *International Conference on Connected Vehicles and Expo (ICCVE)*, Shenzhen, 19-23 Oct., 2015.

[2] Zhang, Y., Lin, W., and Chin, Y., "Data-Driven Driving Skill Characterization: Algorithm Comparison and Decision Fusion," SAE Technical Paper 2009-01-1286, 2009, https://doi.org/10.4271/2009-01-1286.Azevedo, C. L Cardoso.

[3] J. E. Meseguer, C. T. Calafate, J. C. Cano and P. Manzoni, "DrivingStyles: A smartphone application to assess driver behavior," *2013 IEEE Symposium on Computers and Communications (ISCC)*, Split, 2013, pp.000535-000540. doi: 10.1109/ISCC.2013.6755001.

[3] Schneider, A., Hommel, G., & Blettner, M. (2010). Linear Regression Analysis: Part 14 of a Series on Evaluation of Scientific Publications. *Deutsches Ärzteblatt International*, 107(44), pp. 776–782.

[5] Kenneth L. Clarkson. 1985. *Algorithms for Closest-Point Problems (Computational Geometry)*. Ph.D. Dissertation. Stanford University, Palo Alto, CA. UMI Order Number: AAT 8506171.

[6] Schneider, A., Hommel, G., & Blettner, M. (2010). Linear Regression Analysis: Part 14 of a Series on Evaluation of Scientific Publications. *Deutsches Ärzteblatt International*, 107(44), pp. 776–782.

[7] Goszczynska H., Kowalczyk L., Kuraszkiewicz B. (2014) Correlation Matrices as a Tool to Analyze the Variability of EEG Maps. In: Piętka E., Kawa J., Wieclawek W. (eds) *Information Technologies in Biomedicine*, Volume 4. Advances in Intelligent Systems and Computing, vol 284. Springer.

\