Road safety improvement using on-board diagnostics

Jean Paul Dingemanse

j.p.dingemanse@tilburguniversity.edu

u561450

Supervisor: Prof. dr. ir. P.H.M. Spronck

Company supervisor: Dhr. B. van Weerdenburg

bas.van.weerdenburg@crossyn.nl

Master thesis proposal | Data Science & Society

Tilburg University | School of Humanities and Digital Sciences

Tilburg, The Netherlands March 2021

Project Definition

Research by Dingus et al. (2016) shows that people who show aggressive driver behavior (e.g., illegal passing or following too closely), as well as drivers who drive above the speed limit are 11 times more likely to be involved in traffic accidents. Individual behavior must be improved in order to make roads safer. Individual driving behavior can be determined based on the on-board diagnostics (OBD) data. This OBD data can be enriched using context data, such as weather, road and vehicle information. The main purpose of this project is to find factors which influence high-risk driving behavior. The factors presented in Table 1 and Table 2 are available in the dataset for the current research.

Outcomes of Dingus et al. (2016) show that three factors increase the risk of a crash, (1) driving above the speed limit, (2) Strong deceleration and acceleration, and (3) traveling too fast for the weather conditions. Research into deceleration and acceleration show that strong deceleration and strong acceleration is different for each vehicle type and depends on the current speed (Bokare & Maurya, 2017). Research by Hwang et al. (2018) uses the threshold of -2.74 and 2.74 m/s^2 for strong deceleration and acceleration respectively, these thresholds are comparable to the mean of the research by Bokare & Maurya (2017) into different vehicle types. For the current research -2.74 and 2.74 m/s^2 will be used for the baseline predictions, however this threshold is subject to change if more insights appear. Driving above the speed limit, strong deceleration, strong acceleration and traveling too fast for the weather conditions are referred to as *high-risk driving* from this point onward. This thesis aims to answer the following research question:

Which factors influence high-risk driving behavior?

To answer this question, principal component analysis (PCA) models will be used to determine the factors which influence high-risk driving behavior most. The current research will provide a deeper understanding of influence on high-risk driving behavior. After the factors with the most influence have been determined, a less complex model can be created to predict high-risk driving behavior. For the prediction of high-risk driving behavior, a logistic regression and Dynamic Bayesian Network (DBN) model are trained. The logistic regression model serves as a baseline for comparing the DBN model.

Motivation

People think of themselves as good drivers, and only see problems in other people's behavior on the road (Wohleber & Matthews, 2016). Research shows that, despite the improvement of the number of fatal accidents and of road safety, the number of fatal accidents in the Netherlands has risen again (SWOV, 2021). While not all accidents are the result of human error or driving behavior, most fatalities could have been prevented by improving driving behavior. In traffic behavior, various components, such as traffic aggression, unaware behavior and unintended unsafe traffic actions, can be detected.

Moreover, the latter two can evoke the former in other drivers. The official figures for the number of road deaths in 2020 are not yet known, but an estimate by SWOV shows that the target (maximum of 500 road deaths by 2020) has not been achieved (Aarts et al. 2020). Besides high fatality numbers, the costs for accidents are high, with the costs in 2018 being estimated at €18 billion (SWOV, March 2020). Most of the costs are covered by the insurers, who are looking for solutions to bring down the costs. One of these solutions is to improve driving behavior.

When the influence of certain factors on high-risk driving behavior is known, personalized tips could be given to drivers. If drivers use this personalized advice to improve their behavior, high-risk driving behavior, and therefore accidents, could be prevented.

Background

Research into the field of driving behavior is mostly done using smartphone data with a low frequency instead of OBD data. Research has shown that it is possible to predict high-risk driving behavior (Halim, Kalsoom & Baig, 2015). However, this previous research focused on only 50 vehicles with an interval of 30 seconds between the data measurements. The available dataset for the current research contains 1000 vehicles with an interval of 1-5 seconds. This detail in the data is necessary to ensure that every important event is taken into account.

A study conducted in China shows that risk rating can be accurate (Shi, Wong, Li, Palanisamy & Chai, 2019). Nevertheless, this risk rating in the previous research was only focused on a 630-meter-long part of a road, which is not comparable to the environment of the current research. Research on German roads (which somewhat resemble the Dutch roads), only covers highways and is conducted using only 43 participants (Witt, Kompaß, Wang, Kates, Mai & Prokop, 2019). To be able to draw solid conclusions, every road type should be included in the research. Studies using simulations (Kishimoto & Oguri, 2008) or demarcated tracks (Hamada et al., 2016) will not make it possible to draw solid conclusions applicable to real life situations. The study by Hamada et al. (2016) shows that Bayesian methods may be used for risk prediction based on driving behavior.

Research using volunteers creates biased datasets, as high-risk drivers generally do not volunteer for driving behavior research. In the current research, people are not aware that the data is used for this specific research into driving behavior, which leads to an unbiased dataset.

Other studies are focusing on the psychological and consciousness factors on high-risk driving behavior, namely drinking, cannabis use and risk-taking attitude (Useche, Ortiz, Cendales, 2017). These factors - although relevant for high-risk behavior - are not part of this research.

Dataset

The data that will be used is provided by the Crossyn data platform. This platform consists of vehicle Controller Area Network (CAN) messages collected by the OBD and published on the platform. The collected data is enriched with external factors and context data, namely weather, location, and vehicle data from companies like Here and RDC. The collected data currently consists of circa 750 cars and 400 trucks which have created over 1.900.000 trip records. This data is available in a ClickHouse and PostgreSQL database, and can be accessed using Structured Query Language (SQL).

Table 1 shows the available data with their unit/type. A few different features could be determined based on the available data. An overview of this determined data may be found in Table 2.

Available data	Data unit/type
- Temperature - Humidity - Wind direction - Wind speed	- Degrees of °C - Percentage - E.g., SW, NE, N - Km/h
Trip start	datetime
Trip end	datetime
Acceleration	m/s
Fuel consumption	L/100km
Weight	Kg
Location coordinates	Latitude and longitude
Deceleration	m/s
Vehicle information - Brand - Model - Year - Horsepower - Energy label	- Text - Text - Integer - Integer - Euro energy classification A - Z
X, Y and Z accelerometer measures	Units in mg
Speed	Km/h
Speed limit	Km/h
Euro classification	Classification euro 1 - 6
Expected fuel consumption	L/100km

Table 1: Available data with their unit.

Determined value:	Data unit/type	Determined using:
Road Type	urban, motorway, highway	Speed limit
Traveled distance	Km	Location coordinates
Rush hour	Boolean	Datetime

Table 2 Data that could be determined.

Algorithms and software

To reduce computation time and complexity, less relevant factors need to be removed. Principal components analysis (PCA) is a dimensionality reduction technique, which will be used to reduce the number of features. Using PCA, redundant features can be found and eliminated from the research with a minimum of information loss.

To gain a better understanding of the data, a baseline prediction is made using a logistic regression model. The result of this model is a binary prediction of the features to decide whether it belongs to high-risk driving behavior. When the dimension reduction and baseline prediction are completed, the reduced factors will be used to build a model to predict high-risk driving behavior. Previous research has demonstrated that Dynamic Bayesian Networks (DBNs) retrieve a high accuracy using similar data (Kishimoto & Oguri, 2008). DBNs will be used, and the results will be compared to the baseline.

The software used in this project is DBeaver. With DBeaver, the Clickhouse and PostgreSQL databases can both be reached through SQL queries. The access to DBeaver and those databases is arranged through a system owned by Crossyn and made available after signing an NDA. The programming language Python will be used to create the algorithms.

Evaluation method

The trained models will be evaluated using precision, recall and accuracy on the prediction results. Precision is used to evaluate the ratio between the number of true positives and all positive predictions. The recall will be used to determine how many participants the models correctly identified as high-risk drivers. The accuracy will be used as a general score on the performance of the models.

In this case, the most important evaluation method is the recall since it is important to find all high-risk drivers and improve their behavior. If a driver is falsely seen as a high-risk driver and thus receives tips on their driving behavior, the impact is reasonable. When this research is used for a different purpose, the importance of these evaluation methods needs to be reconsidered.

Milestones and plan

To achieve a solid model, an investigation of the features with an influence on high-risk driving behaviour must be made first. For the next milestone, it must be determined which combinations of features are most useful for predicting high-risk driving behavior.

Version 0.1 – March 19th, 2021 Final proposal deadline

Version 0.2 – March 31st, 2021 Data exploration finished.

Version 0.4 – April 9th, 2021 Data reduction finished.

Version 0.5 – April 16th, 2021 Research possible features which influence high-risk behavior.

Version 0.6 – April 23rd, 2021 Baseline prediction finished.

Version 0.7 – April 30th, 2021 Predicting High-risk behavior research finished.

Version 0.9 – May 5th, 2021, First version ready for feedback.

Version 1.0 – May 21st, 2021, Submission.

References:

Aarts, L.T.; Schepers, J.P.; Goldenbeld, Ch.; Decae, R.J.; Bos, N.M.; Bijleveld, F.D.; Doumen, M.J.A.; Dijkstra, A.; Mons, C.; Commandeur, J.J.F.; Hermens, F. De staat van de verkeersveiligheid 2020 (The state of road safety 2020) SWOV 2020, https://www.swov.nl/publicatie/de-staat-van-de-verkeersveiligheid-2020

Bokare P. S., Maurya A. K. Acceration-Deceleration Behaviour of Various Vehicle Types, Transportation Research Procedia, Volume 25, 2017, Pages 4733-4749, ISSN 2352-1465,

Dingus T. A., Guo F., Lee S., Antin J.F., Perez M., Buchanan-King M., Hankey J. (2016), Driver crash risk factors and prevalence evaluation using naturalistic driving data. Proceedings of the National Academy of Sciences March 2016, 113 (10) 2636-2641; DOI: 10.1073/pnas.1513271113

Hwang C., Chen M., Shih C., H. Chen H., and Liu W. K., Apply Scikit-Learn in Python to Analyze Driver Behavior Based on OBD Data, 2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA), Krakow, Poland, 2018, pp. 636-639, doi: 10.1109/WAINA.2018.00159.

Kishimoto Y. and Oguri K., "A modeling method for predicting driving behavior concerning with driver’s past movements," 2008 IEEE International Conference on Vehicular Electronics and Safety, Columbus, OH, USA, 2008, pp. 132-136, doi: 10.1109/ICVES.2008.4640888.

Witt M., Kompaß K., Wang L., Kates R., Mai M., Prokop G., Driver profiling – Data-based identification of driver behavior dimensions and affecting driver characteristics for multi-agent traffic simulation, Transportation Research Part F: Traffic Psychology and Behaviour, Volume 64, 2019, Pages 361-376, ISSN 1369-8478

Shi X., Wong Y. D., Li M. Z. F., Palanisamy C., Chai C., A feature learning approach based on XGBoost for driving assessment and risk prediction, Accident Analysis & Prevention, Volume 129, 2019, Pages 170-179, ISSN 0001-4575

SWOV. Kosten van verkeersongevallen. (Costs of traffic accidents) SWOV-factsheet, March 2020, SWOV Den Haag. https://www.swov.nl/feiten-cijfers/factsheet/kosten-van-verkeersongevallen

SWOV. Riskant verkeersgedrag, verkeersagressie en veelplegers (Risky traffic behavior, traffic aggression and frequent offenders) SWOV-factsheet, January 2021. https://www.swov.nl/feiten-cijfers/factsheet/riskant-verkeersgedrag-verkeersagressie-en-veelplegers.

Useche S., Ortiz V. G., Cendales B. E., Stress-related psychosocial factors at work, fatigue, and risky driving behavior in bus rapid transport (BRT) drivers, Accident Analysis & Prevention, Volume 104, 2017, Pages 106-114, ISSN 0001-4575

Wohleber, R.W., Matthews, G., 2016. Multiple facets of overconfidence: Implications for driving safety. Transp. Res. Part F: Traff. Psychol. Behav. 43, 265–278.

Halim Z., Kalsoom K., Baig K.R., (2016), Profiling drivers based on driver dependent vehicle driving features. Applied Intelligence March 2016, Vol. 44 Issue. 3