Improving road safety using on-board diagnostics

Jean Paul Dingemanse
STUDENT NUMBER: 2051504

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

Prof. dr. ir. P.H.M. Spronck
dr. S.H.P.Collin

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
June 2021

Improving road safety using on-board diagnostics

Jean Paul Dingemanse

As our society becomes more populated and denser, individual behavior more and more has an impact on humanity. This is specially the case in traffic, where every action has an influence on other people and their safety.

This thesis aims to contribute to road safety and cost reduction by predicting high-risk driving behavior. Although driving behavior has been researched for decades, a study which uses such detailed data as the current study is not common. This study uses a data set of more than 700 trips, with a data point for every 5 seconds during those trips. The data set contains road, vehicle, weather and behavioral data. These factors are used to predict high-risk behavior using a decision tree, logistic regression model and a neural network. Using the synthetic minority over-sampling technique (smote), the specificity scores were 0.84, 0.72 and 0.86 respectively for the decision tree, logistic regression and neural network. This study focused on cars and LCV's only, recommendations for further research therefore contain the suggestion to also look at other vehicle categories like vans and trucks.

1. Introduction

1.1 Motivation

People think of themselves as good drivers, and only see problems in other people's behavior on the road (Wohleber and Matthews 2016). Research done by the Dutch Institute for Road Safety (SWOV) shows that, despite the decrease in the number of fatal accidents in the recent years, the number of fatal injuries in the Netherlands has risen since 2014, as shown in Figure 1 (SWOV 2021). While not all accidents are the result of human error or driving behavior, fatalities may be prevented by improving driving behavior.

The official figures for the number of road deaths in 2020 are not yet known, but an estimate by SWOV shows that the target of a maximum of 500 road deaths in 2020 has been exceeded (Aarts et al. 2020). Besides high fatality numbers, the costs for accidents are also high, with the costs in 2018 being estimated at €18 billion in the Netherlands alone (SWOV 2020). Most of the costs are covered by the insurers, who are looking for solutions to bring down the costs. One of these solutions could be to improve driving behavior.

When the influence of weather, road, vehicle and behavioral data on high-risk driving behavior is known, personalized tips may be given to drivers. If drivers use this personalized advice to improve their behavior, high-risk driving behavior, and therefore accidents, may be prevented. This may possibly save the lives of people and bring down the costs of accidents with billions of euros.

1.2 Problem statement

A study (Dingus et al. 2016) shows that people who show aggressive driving behavior (e.g., illegal passing or following too closely), as well as drivers who drive above the speed limit, are eleven times more likely to be involved in traffic accidents. Individual behavior must be improved 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 contextual data, such as weather, road and vehicle information. The main purpose of this study is to determine if high-risk behavior can be predicted. The available features for the current study are presented in Section 3.1.3. Features.

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. 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:

RQ 1: *To what extent can we predict whether a driver will engage in high-risk driving behavior?*

The research question will be answered by examining the following sub-research questions:

Sub-Research question 1: *To what extent can Principal Component Analysis contribute to the performance of a prediction model?*

Sub-Research question 2: *How can a model without Principal Component Analysis contribute to the prediction whether a driver will engage in high-risk driving behavior?*

Sub-Research question 3: *How can the Synthetic Minority Oversampling Technique contribute to the performance of a prediction model?*

Sub-Research question 4: *How can a model without the Synthetic Minority Oversampling Technique contribute to the prediction whether a driver will engage in high-risk driving behavior?*

1.3 Methodology

To answer the research question, a Principal Components Analysis (PCA) model will be used to reduce the number of dimensions. This dimension reduction will also reduce the time needed to perform the analysis and the complexity of the model. A less complex model can handle more data using the same computing power, which will contribute to the robustness of the analysis. This analysis will also provide a deeper understanding of the predictor's influence on high-risk driving behavior. Subsequently, the high-risk driving behavior class will be balanced using Synthetic Minority Oversampling Technique (SMOTE). SMOTE uses oversampling to increase the minority class. After the dimension reduction and SMOTE is completed, a less complex model can be created to predict high-risk driving behavior. For the prediction of high-risk driving behavior, a decision tree, logistic regression model and neural network will be used. Research has shown that all these algorithms have their own strengths and weaknesses (Dreiseitl and

Ohno-Machado 2002). The performance of these algorithms will be compared in this study using the available data set.

1.4 Thesis outline

To cover these topics, this thesis is divided in different chapters. After the introduction, we will provide an overview of the theoretical framework in the second chapter, including an overview of the OBD-II system and previous research. In the third chapter we outline our experimental setup with the data collection, pre-processing and used algorithms. The fourth chapter focusses on all the results retrieved by the different algorithms. Finally, we discuss our findings in the fifth chapter and finalize with the conclusions in the sixth chapter.

2. Theoretical framework

This chapter will provide insight into on-board diagnostics and the previous research done in the related fields. The previous research first dives into traffic analysis in general, secondly into the definition of high-risk behavior, followed by the measure methods, the data and algorithms.

2.1 On-board diagnostics

On-board diagnostics (OBD) data are Controller Area Network (CAN) messages retrieved via the OBD-port in vehicles. The installation in all vehicles is required since 1994 in the United States and since 2004 in Europe. Since 1984, on-board diagnostics has evolved rapidly, starting as a small manufacturer-specific test and communication protocol evolving into a global standard (Baltusis 2004). Currently, almost all vehicle communication systems operate according to the OBD-II standard. With OBD-I - the precursor to OBD-II - the communication was mainly focused on repair and maintenance using Diagnostic Trouble Codes (DTCs) and freeze frame data (McCord 2011). Starting with the OBD-II, the possibility to retrieve real-time vehicle data was expanded and standardized. Real-time data regarding the engine status is available but also behavioral data like pedal position, speed and braking can be retrieved.

A vehicle tracking device is used to read the OBD-port on vehicles and send this data to the platform provided by the company Crossyn. This device, the FMB640, was manufactured by Teltonika in 2019 and is specially made for professional use and equipped to send real-time data (Teltonika 2019).

2.2 Previous research

In this section, the related literature regarding (1) traffic analysis, (2) the definition of high-risk behavior, (3) measure methods, (4) data and (5) an overview of the used algorithms in relevant research, will be discussed.

2.2.1 Traffic analysis. Traffic analysis has been around for decades, although most research focusses on crash prevention and crash investigation, real time and road condition analysis has also been done.

Research into crash prevention is practically always done in one of the three following ways: (1) focused on driver distraction with for example the use of electronic devices, (2) focused on driver alertness with for example studying sleeping behavior

(Kang 2013), work times (Akerstedt et al. 2005) or drugs and alcohol use (Liu and Ho 2010), and (3) focused on crash avoidance technologies with warning and prevention systems like the forward collision warning system (Jermakian 2011)

Crash investigation is mostly focused on vehicle specific characteristics, age groups (Durbin et al. 2015) or the behavior of the driver, for example seatbelt wearing behavior (Kidd and McCartt 2014).

Real time driving behavior analysis is used to give real time feedback to the driver and motivate drivers to improve behavior in a short term (Vaiana, Astarita, and Tassitani)).

A study from 2002 till 2005 requested by the United States National Highway Traffic Safety Administration used road structures and road surfaces to determine its influence on driving behavior (Neale et al. 2002).

2.2.2 Definition of high-risk behavior. Using the three factors found by Dingus et al. (2016) that increase the risk of a crash, namely: (1) driving above the speed limit, (2) strong deceleration and acceleration, and (3) traveling too fast for the weather conditions, could be used to determine high-risk driving behavior.

Speeding can be derived from the data directly, but strong deceleration, acceleration and driving too fast for weather conditions needs to be processed to a usable format. 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 and Maurya 2017). A study by (Hwang et al. 2018) uses the threshold of -2.74 and 2.74 m/s2 for strong deceleration and acceleration respectively, these thresholds are comparable to the mean of the study by Bokare and Maurya (2017) into different vehicle types. Based on previous research, -2.74 and 2.74 m/s2 will be used to generate the high-risk driving behavior labels for the thesis.

2.2.3 Measure methods. Numerous studies have been researching driving habits. A study by Eren et al. (2012) and a study by Li et al. (2016) focused on accelerometer, gyroscope and magnetometer in smartphones to identify patterns. These measure methods in this previous research had a frequency of less than one data point for every 10 seconds and will therefore not be able to take every detail into consideration when building models.

Studies using simulations (Kishimoto and Oguri 2008) and (Adler and McNally 1994) will not make it possible to draw solid conclusions applicable to real life situations.

2.2.4 Data. A study by Halim, Kalsoom, and Baig (2016) has shown that it is possible to predict high-risk driving behavior. The achieved accuracy on this receding study was more than 84%. However, this previous study focused on only fifty vehicles, and no distinguishment was made between vehicle types.

A study conducted in China shows that risk rating can be accurate (Shi et al. 2019). Nevertheless, the risk rating in this study was only focused on a 630-meter-long part of a road, which is not comparable to the environment of the current study. Research on German roads (which somewhat resemble the Dutch roads), only covers highways and is conducted using only 43 participants (Witt et al. 2019). To be able to draw solid conclusions, every road type should be included in the study.

Research using volunteers creates biased data sets, as high-risk drivers generally do not volunteer for driving behavior research. In the current study, people are not aware that the data is used for this specific research into driving behavior, which leads to an unbiased data set. 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, and Cendales 2017) These factors - although relevant for high-risk behavior - are not part of this study.

2.2.5 Classification methods. Multiply classification methods could be useful to predict high-risk driving behavior. For this research we will study the performance of decision trees, logistic regression and neural networks. Using decision trees, the importance of the features could be determined, this is done by counting the splits in the tree (Zhou and Hooker 2020). To find a linear coherence in the data set a Logistic regression model is trained, The Logistic regression is chosen because it can be conducted when the dependent variable is dichotomous. Logistic regression uses linear regression components on a logit scale to identify the strongest linear combination of variables (Stoltzfus 2011). A neural network is used to research the predictive power for neural networks using the used data set. According to a study into travel behavior neural networks are a solid supplementary method for data analysis, because of the relatively assumption-free reconnaissance on large data sets (Shmueli, Salomon, and Shefer 1996).

3. Experimental setup

This chapter provides an insight in the experimental setup. Section 3.1 presents an overview of the data including the collection, cleaning, pre-processing and a basic statistical overview. In section 3.2, the used algorithms will be provided.

3.1 Data

In this section the details about the data will be presented. This includes the data collection, pro-processing, an overview of the features and data oversampling.

3.1.1 Data collection. 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 system and published on the platform. The OBD system is installed in vehicles owned by logistic and leasing companies and sends data on vehicle ignition via the cellular network. Before the collected data reaches the database, the 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 recorded with a frequency between 1 and 5 seconds.

3.1.2 Data pro-processing. Using Python, all trips for each device are requested from the database. For each trip, it is checked if vehicle and weather information is available. For all points in a trip record, the following processing is done: (1) Acceleration and deacceleration are converted to a negative score if the value is below the threshold of -2.7 or above 2.7, otherwise a positive score is assigned. (2) The rush hour value is determined using the times 07:00 till 08:59 in the morning and 16:00 till 17:59 in the afternoon, these times are referred to as rush hour by the Dutch government (Rijkswaterstaat 2011). (3) For each record, the vehicle brand, model and type are removed because these records are hashed for privacy reasons in the database, therefore, this data cannot be used in this study. (4) The car category, composition, fuel type and energy label are converted to numeric values. (5) For cleaning purpose, the incomplete rows and rows with unknown variables are removed from the final data set. (6) Because the amount of

data is too much for reasonable processing of the models with the current computing power, two decisions are made. First, only the vehicle category 'Car' and 'LCV' will be used. The decision to only use these categories is based on the number of high-risk driving behavior datapoints in the data set. More than 90% of the high-risk driving behavior in the data set are done by one of those two categories. The second decision was to only use data for a specific month. Before excluding data, a comparison between months is made based on histograms of each feature. This comparison shows that the difference between distributions is negligible.

Finally, after all preprocessing is done, the records are split into a training and test set and written to a comma separated file (CSV) to get used for analyzing.

3.1.3 Features. The cleaned data set consists of more than 63700 data points, 700 trips and 130 vehicles. All trips include weather, vehicle, road and behavioral information. In Table 1, the weather features can be found. In Table 2, all available vehicle features can be found. All road features can be found in Table 3. Lastly, Table 4 contains all behavioral information.

weather_comfort	The numerical wind chill description
dewPoint	The degrees to which air must be cooled to become saturated with vapor
skyInfo	A sky descriptor value between 0 and 34
daylight	The point of the trip occurred during day or night
humidity	The amount of water vapor in the air
windSpeed	Speed of the wind
visibility	The visible distance in km
temperature	The temperature in Celsius degrees

Table 1: Weather information features

car_category	Category of the vehicle (e.g., Truck, LCV or Car)
power	Horsepower of the vehicle given by the manufacturer
acceleration	Acceleration speed of the vehicle
composition	Composition of the vehicle (e.g., Cabriolet, hatchback, coupe and station wagon).
fuel_type	The fuel type of the vehicle determined by a New European Driving Cycle (NEDC) or for vehicles manufactured after 2018 the Worldwide Harmonized Light Vehicles Test Procedure (WLTP) (Pavlovic et al. 2018)
model_year	The year the car was registered by the Dutch vehicle authority RijksDienst Wegverkeer (RDW)
unladen_mass	The mass of the vehicle determined by the manufacturer
euro_classification	The emission classification determined by the European union
fuel_consumption_combined	The fuel consumption determined by the manufacturer
energy_label	Score between A and G calculated using Co2-emissions, fuel consumption and vehicle size.
top_speed	The manufacturer determined top speed of the vehicle

Table 2: Vehicle information features

road_type	The type of the road
speed_limit	The speed limit on the current road

Table 3: Road information features

rush_hour	Current trip-point occurred during rush hour
traveled_distance	Traveled distance before this point in the trip during the same trip
speeding	Speeding happened during this point in the trip
aggressive_acceleration	Aggressive acceleration happened during this point in the trip
aggressive_behavior_combined	Speeding or aggressive acceleration has happened during this point in the trip

Table 4: Behavioral information features

To get a better idea of the data, an overview of the basic statistics can be found in Appendix A, including the mean, standard deviation, minimum value, 25th percentile, 50th percentile, 75th percentile and maximum value.

3.1.4 Data oversampling. The data consists of more than 60000 data points labeled as non-high-risk behavior and over 3700 data points labeled as high-risk behavior. A study with unbalanced data shows the importance of balanced data when fitting logistic regression models (Salas-Eljatib et al. 2018). To balance the data set, oversampling or undersampling may be used. In this case, oversampling is preferred to prevent important information being removed from the data set while undersampling.

The technique Synthetic Minority Over-sampling Technique (SMOTE) has proven to improve the accuracy and specificity on different data sets (Chawla et al. 2002). SMOTE creates larger and less specific decision regions, rather than smaller and more specific regions. These less specific decision regions cause the used classifier to generalize better on unseen data.

To prove the importance of oversampling on this data set, both an analysis with and without SMOTE will be performed and the results will be shown in Chapter 4 Results.

3.2 Algorithms

In this paragraph, the used models will be outlined and an overview of used the libraries, tuned parameters, evaluation metrics and baselines will be provided. First, we will outline the dimension reduction techniques followed by the prediction algorithms.

3.2.1 Dimension reduction. To reduce computation time and complexity, less relevant information is removed from the data set. The dimensionality reduction technique Principal components analysis (PCA) is used to reduce the number of dimensions. Using this technique, redundant features are found and eliminated from the study with a minimum of information loss.

PCA uses a linear transformation of the data to create a data set with less dimensions. The dimensions are referred to as principal components (PCs). PCs are ordered

based on how much variation they explain in the original data set, beginning with the highest variation and ending with the PC with the lowest variation.

The following steps are involved in a PCA: (1) Standardize the data to get all the features along the same scale, (2) Compute the covariance matrix, (3) Perform 'eigendecomposition' on the covariance matrix to create eigenvalues and eigenvectors, (4) Order eigenvectors based on the eigenvalues, (5), Determine the number of principal components, (6) Construct the new matrix based on the number of principal components chosen and (7) Compute the new feature space (O'Sullivan 2020).

The determination of the number of principal components is done using the GridSearchCv function. All evaluated parameters by this function can be found in Table 5.

Parameter name	**Parameters**
n_components	[2, 3, 4, 5, 6, 7, 8, 10, 15, 20, 22]

Table 5: GridSearchCV parameters

PCA reduces the dimensionality by replacing the original variables with derived variables. Often, this is possible while retaining most of the variability from the original data. The number of components used is determined by the explained variance. To get all variables on the same scale, the data is standardized to a normal distribution (Scikit-learn (n.d.). Scaling is done using the Sklearn.preprocessing StandardScaler function in Python. All used libraries can be found in Table 6.

Used libraries
Sklearn.preprocessing StandardScaler
Sklearn.decomposition PCA
sklearn.model_selection.GridSearchCV

Table 6: Used libraries for dimension reduction

3.2.2 Prediction. For the prediction, three different algorithms are used: (1) decisions trees to determine the feature importance, (2) logistic regression for the linear coherence and (3) neural network to find the non-linear coherence. In this section, the used algorithms will be outlined, starting with the decision tree, followed by logistic regression and neural networks. The Python libraries used for the models can be found in Table 7.

Used libraries
Sklearn.preprocessing.StandardScaler
Sklearn.model_selection.train_test_split
Matplotlib.pyplot
Imblearn.over_sampling.SMOTE
Sklearn.tree.DecisionTreeClassifier
Sklearn.linear_model.LogisticRegression
Sklearn.metrics.roc_curve
Tensorflow.Keras
Sklearn.metrics.classification_report
Sklearn.metrics.confusion_matrix

Table 7: Used libraries for the prediction models

3.2.2.1 Decision tree.
In a similar study by Osman et al. (2019), a decision tree was found to be a solid classification method. Another reason to use decision trees is because of their simple analysis and precision on multiple data forms (Tijo and Abdulazeez 2021). The precision on multiple data forms is accomplished because of the robustness to noise and tolerance against missing values of the decision tree model.

3.2.2.2 Logistic regression.
When the dimension reduction is completed, the reduced factors were used to build a logistic regression model to predict high-risk driving behavior. Logistic regression is commonly used with traffic safety data (Ghasemzadeh and Ahmed 2018) and with behavioral data (Agresti 2007). The logistic regression model was created using the Sklearn package. To determine the optimal parameters for tuning the logistic regression model, a GridSearchCv function with different solvers and C parameters is used. For the solver parameter, the most common solvers are tested (e.g., newton-cg, lbfgs and liblinear). The C parameter can be explained as it is the inverse of regulation strength, this is an applied penalty to reduce overfitting. To test the best performing C parameter, five different parameters are tested with the GridSearchCv reaching from 0.01 till 100. For the penalty parameter, 'l2' is used because all used solvers are compatible with this penalty method. An overview of all the considered parameters van be found in Table 8.

Parameter	**Value**
Solvers	Newton-cg, lbfgs, liblinear
C parameter	0.01, 0.1 1.0 10, 100

Table 8: Logistic regression tuning parameters

3.2.2.3 Neural network.
To test the ability to predict high-risk driving behavior, a neural network is used. For the neural network model, different combinations of layers and activation functions will be used. Starting with zero hidden layers to test the linearity in the data, followed by more hidden layers till the best performing number is known. The input layer is a flatten layer with 24 dimensions, and an output layer with 1 dimension will be used. All most common activation functions like relu, softmax, sigmoid and tanh are used to determine the best performing ones for the hidden layers. Khan suggests that for binary classification problems a sigmoid or logistic function would be most appropriate (Khan 2019).

3.2.2.4 Evaluation method.
The unbalanced data set causes that evaluation using the accuracy score will not give any useful insights in the performance of the models. Therefore, the Specificity also known as True Negative Rate is used to compare the models, calculated using the formula as shown in Equation 1. The specificity measures the proportion of the actual negatives that are correctly identified as such. For this experiment, non-high risk driving behavior will be referred to as positive and high-risk driving behavior as negative.

$$\text{Specificity} = \frac{\text{True negative}}{\text{True negative + false positive}}$$

Equation 1: True negative rate

4. Results

In this chapter, the results will be presented. First, we will present the dimension reduction results, followed by the prediction results. The results are used to answer the sub-research questions in the discussion. These answers will then be used to answer the main research question. The sub-research questions were:

Sub-Research question 1: *To what extent can Principal Component Analysis contribute to the performance of a prediction model?*

Sub-Research question 2: *How can a model without Principal Component Analysis contribute to the prediction whether a driver will engage in high-risk driving behavior?*

Sub-Research question 3: *How can the Synthetic Minority Oversampling Technique contribute to the performance of a prediction model?*

Sub-Research question 4: *How can a model without the Synthetic Minority Oversampling Technique contribute to the prediction whether a driver will engage in high-risk driving behavior?*

4.1 Dimension reduction result

To answer the first sub-research question, a PCA was done. To determine the number of components, a GridSearchCV pipe was setup with the parameters as shown in Table 7. The results for the GridSearchCV pipe can be found in Figure 2 and Figure 3. Figure 2 shows the explained variance for each number of components. Figure 3 shows a classification accuracy based on a non-tuned logistic regression model. This accuracy shows that that the most ideal number of components is 7.

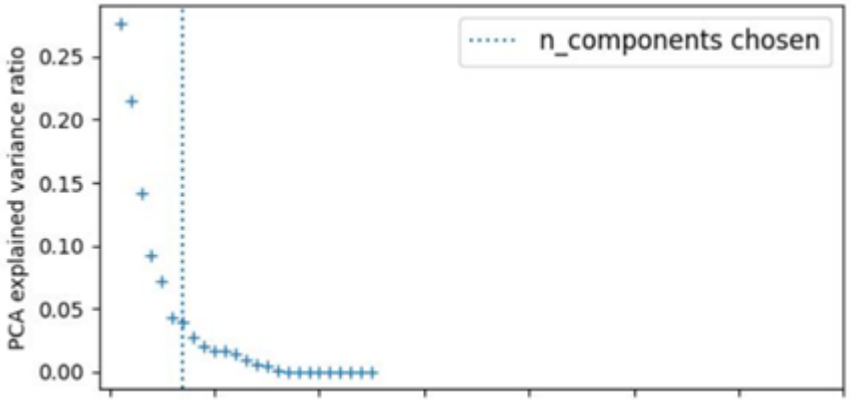

Figure 2: PCA explained variance and n_components chosen

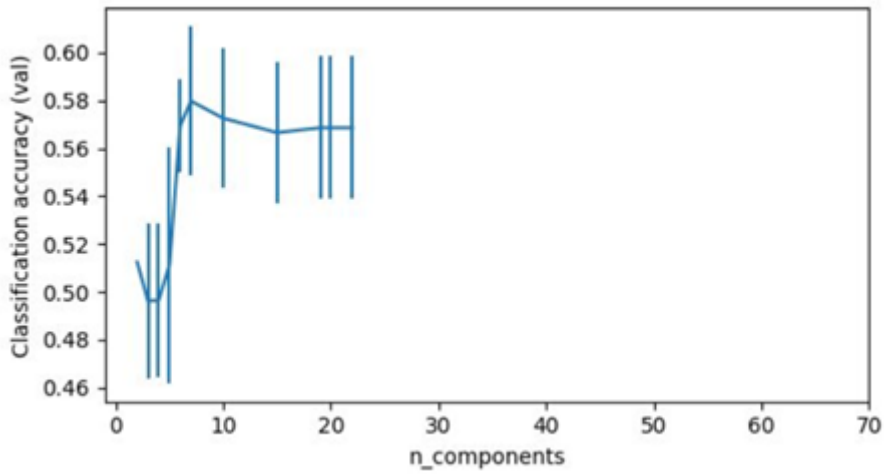

Figure 3: PCA Classification accuracy with n_components

The number of seconds it took to fit the data using a logistic regression model with or without PCA and SMOTE can be found in Table 9.

	Time in seconds	**Difference from base (%)**
Non-reduced or balanced data (base)	589	100.00%
Reduced dimensions with PCA	18	3.05%
Oversampled with SMOTE	958	162.65%
SMOTE and PCA	36	6.11%

Table 9: Duration of fitting the data on a logistic regression model

4.2 Prediction result

To answer all sub-research questions a decision tree, logistic regression and a neural network were used to create the predictions. First, the tuning results for each algorithm are shown. This is followed by the results - with and without using Principal Component Analysis and Synthetic Minority Oversampling Technique - of these algorithms.

4.2.1 Logistic regression tuning. First, we needed to find the best performing logistic regression parameters. With a high score of 0.942 on the training data, the results from the GridSearchCv showed that 'newton-cg' was the best solver parameter for our data in combination with a value of 1.0 for the C parameter. The 'newton-cg' solver is a method which uses the Hessian matrix to optimize the model. A detailed comparison between the solvers is outlined in "ADMM-Softmax: An ADMM Approach for Multinomial Logistic Regression" (Fung et al. 2019). The 'newton-cg' solver and a value of '1.0' for the C parameter were used with all logistic regression models in the current study.

To get a visual view into the performance of the logistic regression model an ROC curve is added see Figure 4. The curve was plotted using the specificity and sensitivity pair, the area under the curves is a measure of the usefulness of a test in general (Simundic 2009). A greater area under the curve means a more useful test.

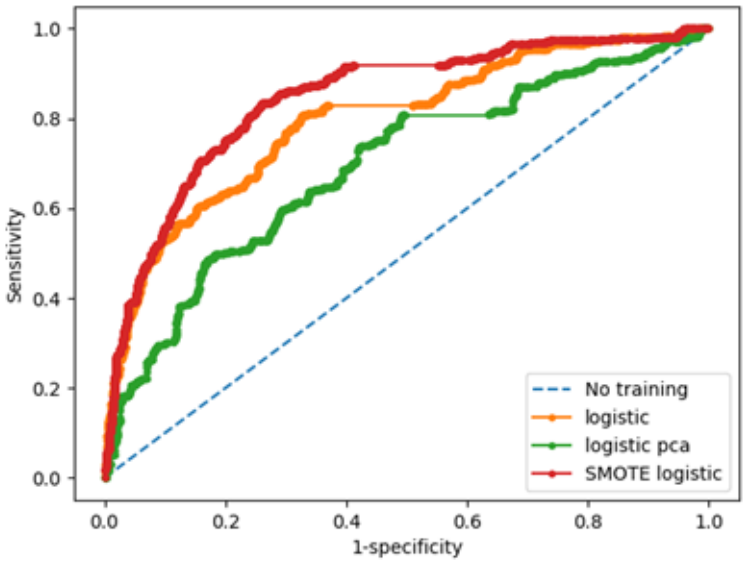

Figure 4: ROC curves logistic regression

4.2.2 Decision tree results. To identify if the decision tree model is overfitted the training resulsts are compared to the test results. The results of the analysis can be found in Table 10 for the training set and in Table 11 for the test set. The calculated specificity scores were 0.9948 and 0.8997 for the training and test set respectively.

	Predicted positive	**Predicted negative**
Positive	35627	12
Negative	78	2305

Table 10: Decision tree result train set

	Predicted positive	**Predicted negative**
Positive	11826	74
Negative	111	664

Table 11: Decision tree result test set

The results of the analysis using a decision tree with the reduced dimensions can be found in Table 12. The calculated specificity for this analysis had a score of 0.8981.

	Predicted positive	**Predicted negative**
Positive	11755	81
Negative	125	714

Table 12: Decision tree result using PCA

To identify if the model is overfitted using SMOTE, the training set results as shown in table 13 are compared to the test results in Table 14. The specificity scores are

0.9985 and 0.9865 for the training and test set respectively.

	Predicted positive	**Predicted negative**
Positive	35594	51
Negative	117	35528

Table 13: Decision tree result using SMOTE train set

	Predicted positive	**Predicted negative**
Positive	11724	152
Negative	753	11123

Table 14: Decision tree result using SMOTE test set

The results of the analysis using decision tree with oversampling through SMOTE and with the through PCA reduced can be found in Table 15. The calculated specificity for this analysis had a score of 0.9867.

	Predicted positive	**Predicted negative**
Positive	11727	149
Negative	800	11076

Table 15: Decision tree result using SMOTE and PCA

To determine which factors contribute to the scores, a feature importance matrix was made. The sorted results for this matrix can be found in Table 16.

Feature	Importance
traveled_distance	0.42746364808037374
power	0.13395390456015568
speed_limit	0.11857819844958438
road_type	0.059481750547607896
weather_comfort	0.05882480185301585
windSpeed	0.043493942250678416
rush_hour	0.0333786780818325
distance	0.027955540086833125
temperature	0.021986245495815713
model_year	0.015160485286628667
dewPoint	0.013467375974703549
fuel_consumption_combined	0.013214193441412371
visibility	0.007796448586518279
top_speed	0.007599191167551518
humidity	0.006067655684921963
unladen_mass	0.004140824780721146
skyInfo	0.002483386283475392
composition	0.001827686855375893
acceleration	0.0012585099585756424
daylight	0.0009306220203068175
energy_label	0.0006072400492423296
fuel_type	0.00032967050466918535
car_category	0.0
euro_classification	0.0

Table 16: Feature importance matrix

To determine the impact of the traveled_distance feature on the decision tree prediction, an analysis without this feature was done. The results of this analysis using oversampling with SMOTE can be found in Table 17. The calculated specificity for this analysis had a score of 0.8380.

	Predicted positive	Predicted negative
Positive	9831	2074
Negative	1176	10729

Table 17: Decision tree result using SMOTE without traveled_distance

To determine the impact of the traveled_distance and power feature on the decision tree prediction, an analysis without these features was done. The results of this analysis using oversampling with SMOTE can be found in Table 18. The calculated specificity for this analysis had a score of 0.8426.

	Predicted positive	Predicted negative
Positive	9857	2013
Negative	1089	10781

Table 18: Decision tree result using SMOTE without traveled_distance and power

4.2.3 Logistic regression results. The results of the analysis using logistic regression for both the training and test set can be found in Table 19 and Table 20. The calculated specificity for these two analyses had a score of 0.00.

	Predicted positive	**Predicted negative**
Positive	11876	0
Negative	799	0

Table 19: Logistic regression train result

	Predicted positive	**Predicted negative**
Positive	11876	0
Negative	799	0

Table 20: Logistic regression test result

The results of the analysis using logistic regression with the through PCA reduced dimensions can be found in Table 21. The calculated specificity for this analysis had a score of 0.00.

	Predicted positive	**Predicted negative**
Positive	11871	0
Negative	804	0

Table 21: Logistic regression result using PCA

The results of the logistic regression analysis with oversampling using SMOTE can be found in Table 22. The calculated specificity for this analysis had a score of 0.7196. Although the focus in this study was mainly on the specificity of the model, the F1 score was used to show the overall performance. The F1 score of 0.7236 shows that the predictive power of the model for true positives did not drop after using SMOTE.

	Predicted positive	**Predicted negative**
Positive	8348	3521
Negative	2857	9034

Table 22: Logistic regression result using oversampling with SMOTE

The results of the logistic regression analysis with the through PCA reduced dimensions and with the oversampled data using SMOTE can be found in Table 23. The calculated specificity for this analysis had a score of 0.6908.

	Predicted positive	**Predicted negative**
Positive	8108	3867
Negative	3146	8639

Table 23: Logistic regression result using PCA and oversampling with SMOTE

4.2.4 Neural network results. The tuning of the neural network is done by testing different configurations. First the model is trained with 0 hidden layers followed by a model with more hidden layers. Also, combinations of several hidden layers and number of neurons with different combinations of activation functions are evaluated. The combination with one layer and the parameter configuration as shown in Table 24 showed the highest performance.

Parameter	Value
Number of neurons	512
Activation function	Sigmoid
Epochs	200
Bach size	2000

Table 24: Neural network configuration

The results of the trained Neural network using the previous mentioned configuration and without oversampling can be found in Table 25. The calculated specificity for this analysis has a score of 0.00.

	Predicted positive	Predicted negative
Positive	11909	0
Negative	766	0

Table 25: Neural network without oversampling

The results of the trained Neural network with the oversampling only on the training set can be found in Table 26. The calculated specificity for this analysis has a score of 0.2602.

	Predicted positive	Predicted negative
Positive	10153	1751
Negative	155	616

Table 26: Neural network with oversampling on the training set

The results of the trained neural network with the oversampling on both the training and test set can be found in Table 27. The calculated specificity for this analysis had a score of 0.8559.

	Predicted positive	Predicted negative
Positive	10132	1724
Negative	1618	10238

Table 27: Neural network results with oversampling on the training and test set

5. Discussion

In this thesis different classifiers were used to predict if a driver will engage in high-risk driving behavior. These predictions could be used to give tips and create awareness among high-risk drivers. With these classifiers it is also determined if the data mining techniques PCA and SMOTE contribute to the prediction of high-risk driving behavior. This section provides a discussion on the results as presented in Chapter 4. First, the results will be evaluated and an answer for the sub-research questions will be provided. Subsequently, an answer to the main research question will be stated based on the sub-

research questions. This will be followed by an outline of the limitations for the current study and a recommendation for future research.

5.1 Evaluating results

The research question for the current study was formulated as: *To what extent can we predict whether a driver will engage in high-risk driving behavior?* Before structuring an answer to this question, we focus on answering the sub-research questions.

5.1.1 Sub-Research Question 1: To what extent can Principal Component Analysis contribute to the performance of a prediction model? . The motive to use PCA was to reduce the complexity and duration of the prediction algorithm to fit on the data. Before applying the PCA, the number of components was determined first by using a GridSearchCv function. As shown in Figure 2, the output of the GridSearchCv function shows that the number of components with a maximum explained variation loss of 0.05 is 7 (n_components = 7).

The results show that the complexity and duration of the prediction model were reduced. The results of the decision tree with the through PCA reduced dimensions (Table 12) show that some information was lost during the reduction compared to the results without PCA (Table 11). Using logistic regression and SMOTE (Table 22 and 23), this loss of information was also visible by comparing the predicted true negatives. Although these differences between the specificity were not significant - a difference in specificity of 0.0002 for the decision tree model and 0.0288 for the logistic regression model – but still 395 extra high-risk drivers were misclassified as non-high-risk drivers with PCA (Table 22), compared to the model without PCA (Table 23). While performing PCA, some information was lost. Although it was only an explained variation loss of maximum 0.05 per dimension, this lost explained variance caused the model to perform less. This loss of information is not uncommon to influence the performance of a model (Janecek and Gansterer 2008)

The advantage of PCA becomes clear when looking at the performance of the models in Table 9. Fitting the models while using PCA took a fraction of the time compared to the models without PCA. With reduced dimensions, the performance of the logistic regression took 3.05% (non-SMOTE) and 6.11% (SMOTE) of the time needed when performing a logistic regression without PCA. The reduced complexity ensured that the fitting time dropped significantly, this creates possibilities for using more data in future research. When more data is used during analysis and computing power is limited, PCA can play an important role.

5.1.2 Sub-Research Question 2: How can a model without Principal Component Analysis contribute to the prediction whether a driver will engage in high-risk driving behavior?. To answer this sub-research question, the results of the decision tree results (Table 10 and Table 11) and the logistic regression results using SMOTE (Table 22 and Table 23) can be used. As previously mentioned, the specificity score for the results without PCA was slightly higher than with reduced dimensions. With the aim for this study being to detect as much high-risk behavior drivers as possible, we must conclude that PCA does not contribute to that goal. Therefore a model without PCA can contribute to the prediction whether a driver will engage in high-risk driving behavior.

5.1.3 Sub-Research Question 3: How can the Synthetic Minority Oversampling Technique contribute to the performance of a prediction model?. To answer this sub-

research question, the results of the logistic regression model without SMOTE (Table 20 and Table 21) are compared with the results of the logistic regression model with SMOTE applied (Table 22 and Table 23). Table 20 and Table 21 show that the logistic regression model may not predict the high-risk behavior of drivers at all. The data was too unbalanced for the model to generalize the predictions. Due to this result, SMOTE was applied and an increase in the predictive power of the specificity was achieved. With a specificity score of 0.7196 for the model using SMOTE compared to a specificity score of 0 without smote, the model became useful in predicting high-risk driving behavior.

Because no significant difference was found in the specificity results for SMOTE used only on the training set (Table 25) and SMOTE used on both the training and test set (Table 26), it can be concluded that an oversampled test set does not contribute to higher specificity.

5.1.4 Sub-Research Question 4: How can a model without the Synthetic Minority Oversampling Technique contribute to the prediction whether a driver will engage in high-risk driving behavior?. The specificity of 0.8997 using a decision tree model (Table 10) suggests that a model without SMOTE could contribute to the prediction of high-risk driving behavior. However, the overall high performance and the accuracy of 0.9948 on the train set (Table 10) suggests overfitting, therefore, an overview of the feature importance was made and presented in Table 16. This table shows that the traveled distance had the highest impact on the performance. By removing this feature, it could be determined if the high performance of the model was caused by overfitting. The specificity without the 'traveled distance' feature becomes 0.8380 (Table 17). Considering these results it suggests that traveled distance is specific for each trip point and therefore not contributing to the prediction of unseen data. The feature with the second highest impact on the decision tree prediction was 'power'. Table 18 shows that the specificity score of 0.8426 without the 'traveled_distance' and 'power' is higher compared to the results were only the 'traveled_distance' feature is disregarded. This suggests that 'power' does not overfit the decision tree model. The feature 'power' or Horsepower of a car is a less trip specific feature and would be a logical predictor for high-risk driving behavior and should therefore not be removed from the data set.

The results for the logistic regression model (Table 20 and Table 21) and neural network (Table 25) without SMOTE shows that the models could not generalize properly. With no predictions for the high-risk driving behavior class (true negatives), logistic regression and neural network without undersampling or oversampling cannot be used to predict high-risk behavior accurately. The data used for the current study was unbalanced, with only 5.5% of the data belonging to the high-risk behavior class. The decision tree model outperformed the logistic regression and neural network without SMOTE.

5.1.5 Main Research Question: To what extent can we predict whether a driver will engage in high-risk driving behavior?. The F1 score of 0.9628, 0.7236 and 0.8584 for decision trees, logistic regression and neural networks respectively while using SMOTE, suggests that high-risk driving behavior can be predicted. Because of the importance of finding the high-risk driving behaviors inside the data set as previously mentioned, we focus on the specificity evaluation metric (true negatives rate). If we focus on the specificity, the score for these models become 0.9865, 0.7196 and 0.8559, this shows that we can find between 72% and 98% of the high-risk drivers.

The tree-based method decision tree outperforms logistic regression could be due the fact that no strict applicability terms are needed, decision trees are less influenced by the multicollinearity of the features, handle missing values well and have a low prevalence compared to logistic regression models (Nagy et al. 2010). The reason neural network outperforms logistic regression could be due the non-linearity in the data and complexity of the neural network model. The complexity of the neural network model is caused by the weight parameters who are adjusted for each iteration based on the used hyper parameters (Sanderson et al. 2019).

Because both the decision tree and the neural network are outperforming the logistic regression model it suggests that the data is not linear.

5.2 Limitations

Despite these results, a couple of limitations should be pointed out. Firstly, the amount of data the system could handle. Although, the system used an Intel Xeon gold Central Processing unit with 20 cores and 32 gigabytes of RAM, the available computing power had to be shared with other core processes. To keep the analysis within a reasonable duration for the available time, a subset of the data was selected.

A second limitation is the unbalanced data. To solve the unbalance in the data, SMOTE had to be applied to oversample the data. Oversampling in general consumes time and computing resources and may possibly lead to overfitting (Elrahman and Abraham 2013). This limitation may be solved by removing non-high-risk driving behavior data, however, removing useful information from the data set could lead to a lower performance.

5.3 Recommendations

Four recommendations can be pointed out. First, as previously stated, the available computing power was limited. It is recommended to use more computing power in future research, this way, more data could be analyzed and prediction scores might become more accurate.

The second recommendation is to combine SMOTE with undersampling techniques. This has been proven effective (Chawla et al. 2002). The study by Chawla shows that SMOTE is a useful tool and should not be replaced with other techniques, however, combining it with undersampling may potentially be highly effective.

A third recommendation would be regarding the available features. Features with high potential, for example: the physical state of the driver, gear change behavior, revolutions per minute and lane change behavior – which also may be considered high-risk behavior (Li, Lu, and Xu 2017) – could be added to the available data set. When these features are used in future research, the accuracy of the drawn conclusions may potentially be improved.

A final recommendation is regarding the vehicle categories. The current study only focused on the 'car' and 'LCV' categories. As shown in research, medium and heavy trucks are also prone to perform high-risk driving behavior (Castillo-Manzano J. 2016). To improve road safety and reduce insurance costs, other vehicle categories should be part of future research.

6. Conclusion

The purpose of this study was to determine to what extent engaging in high-risk behavior may be predicted. To summarize our findings, we will conclude what the answer will be to the main research question.

To what extent can we predict whether a driver will engage in high-risk driving behavior?

By comparing the prediction specificity between using and not using the two data mining techniques SMOTE and PCA, a conclusion can be made about their contribution to the prediction. The contribution of PCA to the specificity score is negative and therefore not contributing to the predictions whether a driver will engage in high-risk driving behavior. SMOTE on the other hand contributes to the specificity (true negatives) and balances the data, which is beneficial for the model to generalize on the high-risk driving behavior class.

For the three evaluated models in this study, decision tree, logistic regression and neural network, the best performing models were the decision tree and neural network. With a score above 0.85 these models performed well above the baseline of 0.5. These results suggest that high-risk driving behavior can be predicted and personalized tips based on driving behavior could be prescribed.

Acknowledgement

I would like to thank Dr. P.H.M. Spronck for his continue support and insights during this thesis process. Furthermore, I would like to thank colleagues, friends and family who contributed to this project in all the different ways.

References

Aarts, L. T., J. P. Schepers, Ch. Goldenbeld, R. J. Decae, N. M. Bos, F. D. Bijleveld, M. J. A. Doumen, A. Dijkstra, C. Mons, J. J. F. Commandeur, and F. Hermens. 2020. *De staat van de verkeersveiligheid 2020 (The state of road safety 2020) SWOV 2020.*

Adler, Jeffrey L. and Michael G. McNally. 1994. In-laboratory experiments to investigate driver behavior under advanced traveler information systems. *Transportation Research Part C: Emerging Technologies*, 2(3):149–164.

Agresti, A. 2007. An introduction to categorical data analysis (second edition). John Wiley Sons Inc.

Akerstedt, T., B. Peters, A. Anund, and G. Kecklund. 2005. Impaired alertness and performance driving home from the night shift: a driving simulator study j. *Sleep Research*, 14(1):17–20.

Baltusis, P. 2004. *On board vehicle diagnostics*, convergence international congress exposition on transportation electronics edition.

Bokare, P. S. and A. K. Maurya. 2017. Acceration-deceleration behaviour of various vehicle types. *Transportation Research Procedia*, 25:4733–4749.

Castillo-Manzano J., Fageda X., Castro-Nuño M. 2016. Exploring the relationship between truck load capacity and traffic accidents in the european union, transportation research part e: Logistics and transportation review. *Volume*, 88:94–109.

Chawla, N., K. Bowyer, L. Hall, and W. Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(2002):321–357.

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

Dreiseitl, S. and L. Ohno-Machado. 2002. Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, 35(5-6):352–359.

Durbin, D., J. Jermakian, M. Kallan, A. McCartt, K. Arbogast, and MyersR. Zonfrillo M. 2015. Rear seat safety: Variation in protection by occupant, crash and vehicle characteristics. *Accident Analysis Prevention*, 80(2015):185–192.

Elrahman, S. and A. Abraham. 2013. A review of class imbalance problem. *Journal of Network and Innovative Computing*, 1:332–340.

Eren, H., S. Makinist, E. Akin, and A. Yilmaz. 2012. Estimating driving behavior by a smartphone. *IEEE Intelligent Vehicles Symposium*, pages 234–239.

Fung, S., S. Tyrvainen, L. Ruthotto, and E. Haber. 2019. Admm-softmax: An admm approach for multinomial logistic regression,. *Cornell University*, 1901.

Ghasemzadeh, A. and M. Ahmed. 2018. Utilizing naturalistic driving data for in-depth analysis of driver lane-keeping behavior in rain: Non-parametric mars and parametric logistic regression modeling approaches. *Transportation Research Part C: Emerging Technologies*, 90:379–392.

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

Hwang, C., M. Chen, C. Shih, H. H. Chen, and W. K. Liu. 2018. Apply scikit-learn in python to analyze driver behavior based on obd data. pages 636–639, 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA).

Janecek, A. and W. Gansterer. 2008. A comparison of classification accuracy achieved with wrappers, filters and pca. *University of Vienna, Research Lab Computational Technologies and Applications*.

Jermakian, J. 2011. Crash avoidance potential of four passenger vehicle technologies,. *Accident Analysis Prevention*, 43(2011):732–740.

Kang, H. 2013. Various approaches for driver and driving behavior monitoring: A review. pages 616–623, Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops.

Khan, R. 2019. *Nothing but NumPy: Understanding Creating Binary Classification Neural Networks with Computational Graphs from Scratch*. Towards data science.

Kidd, D. and A. McCartt. 2014. Drivers attitudes toward front or rear child passenger belt use and seat belt reminders at these seating positions. *Traffic Injury Prevention*, 15(3):278–286.

Kishimoto, Y. and K. Oguri. 2008. A modeling method for predicting driving behavior concerning with driver's past movements. pages 132–136, CIEEE International Conference on Vehicular Electronics and Safety.

Li, F., H. Zhang, H. Che, and X. Qiu. 2016. Dangerous driving behavior detection using smartphone sensors. pages 1902–1907, IEEE 19th International Conference on Intelligent Transportation Systems (ITSC).

Li, Y., J. Lu, and K. Xu. 2017. Crash risk prediction model of lane-change behavior on approaching intersections. *Discrete Dynamics in Nature and Society*, pages 332–340. Article ID 7328562.

Liu, Y. and C. Ho. 2010. Effects of different blood alcohol concentrations and post-alcohol impairment on driving behavior and task performance. *Traffic Injury Prevention*, 11(4):334–341.

McCord, K. 2011. Automotive diagnostic systems: understanding obd i and obd ii. *CarTech*, 1.

Nagy, Krisztina, Jenő Reiczigel, Andrea Harnos, Anikó Schrott, and Péter Kabai. 2010. Tree-based methods as an alternative to logistic regression in revealing risk factors of crib-biting in horses. *Journal of Equine Veterinary Science*, 30(1):21–26.

Neale, V. L., S. G. Klauer, R. R. Knipling, T. A. Dingus, G. T. Holbrook, and A. Petersen. 2002. The 100 car naturalistic driving study, phase i-experimental design. HS-809:536.

Osman, O., M. Hajij, S. Karbalaieali, and S. Ishak. 2019. A hierarchical machine learning classification approach for secondary task identification from observed driving behavior data. *Accident Analysis Prevention*, 123(2019):274–281.

O'Sullivan, C. 2020. *A Step-By-Step introduction to PCA*. Towards datascience.

Pavlovic, J., B. Ciuffo, G. Fontaras, V. Valverde, and A. Marotta. 2018. How much difference in type-approval co2 emissions from passenger cars in europe can be expected from changing to the new test procedure (nedc vs wltp)? *Transportation Research Part A: Policy and Practice*, 111(2018):136–147.

Rijkswaterstaat. 2011. Quick scan kruispunt nieuwe postbaan – mauritslaan te stein. Retrieved from: https://www.rijkswaterstaat.nl/rws/docmgt/Stein_01_2-def.pdf.

Salas-Eljatib, C., A. Fuentes-Ramirez, T. Gregoire, A. Altamirano, and V. Yaitul. 2018. A study on the effects of unbalanced data when fitting logistic regression models in ecology. *Ecological Indicators*, 85:502–508.

Sanderson, Michael, Andrew G.M. Bulloch, JianLi Wang, Tyler Williamson, and Scott B Patten. 2019. Predicting death by suicide using administrative health care system data: Can feedforward neural network models improve upon logistic regression models? *Journal of Affective Disorders*, 257:741–747.

Scikit-learn. (n.d.). Importance of feature scaling. Retrieved from: https://scikit-learn.org/stable/auto$_examples/preprocessing/plot_scaling_importance.html$.

Shi, X., Y. D. Wong, M. Z. F. Li, C. Palanisamy, and C. Chai. 2019. A feature learning approach based on xgboost for driving assessment and risk prediction. *Accident Analysis Prevention*, 129:170–179.

Shmueli, Deborah, Ilan Salomon, and Daniel Shefer. 1996. Neural network analysis of travel behavior: Evaluating tools for prediction. *Transportation Research Part C: Emerging Technologies*, 4(3):151–166.

Simundic, A. 2009. Diagnostic accuracy - part 1 basic concepts: sensitivity and specificity, roc analysis, stard statement. *Point of Care: The Journal of Near-Patient Testing Technology*, 11:6–8.

Stoltzfus, J. 2011. Logistic regression: A brief primer. *Academic Emergency Medicine*, 18(10):1099–1104.

SWOV. 2020. Kosten van verkeersongevallen. (costs of traffic accidents). Retrieved from: https://www.swov.nl/feiten-cijfers/factsheet/kosten-van-verkeersongevallen. SWOV-factsheet.

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

Teltonika. 2019. Fmb640. *Professional trackers*, Retrieved from: https://teltonika-gps.com/product/fmb640/.

Tijo, B. and A. Abdulazeez. 2021. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2:20–28.

Useche, S., V. G. Ortiz, and B. E. Cendales. 2017. Stress-related psychosocial factors at work, fatigue, and risky driving behavior in bus rapid transport (brt) drivers. *Accident Analysis Prevention*, 104(2017):106–114.

Vaiana, Iuele, Caruso Astarita, and Giofré Tassitani, Zaffino. Driving behavior and traffic safety: An acceleration-based safety evaluation procedure for smartphones. *Modern Applied Science*, 8:88–96.

Witt, M., K. Kompaß, L. Wang, R. Kates, M. Mai, and G. Prokop. 2019. 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*, 64(2019):361–376.

Wohleber, R. W. and G. Matthews. 2016. Multiple facets of overconfidence: Implications for driving safety. *Transport Research, Part F: Traffic Psychology*, 43:265–278.

Zhou, Z. and G. Hooker. 2020. Unbiased measurement of feature importance in tree-based methods. *ACM Transactions on Knowledge Discovery from Data*, 25(26):1–21.

Appendix A: Basic statistical information

	Mean	Std.	Min.	25%	50%	75%	Max.
aggressive_acceleration	0.0246	0.154	0.0	0.0	0.0	0.0	1.0
speeding	50.381	74.636	0.0	0.0	0.0	69.0	300.0
speeding_binary	0.476	0.499	0.0	0.0	0.0	1.0	1.0
aggressive_combined	0.500	0.500	0.0	0.0	1.0	1.0	1.0
road_type	2.546	1.415	1.0	1.0	2.0	4.0	5.0
traveled_distance	27.252.076	31.804.004	0.0	5680.0	15965.5	35425.0	270931.0
speed_limit	74.191	27.212	0.0	50.0	80.0	100.0	130.0
rush_hour	0.242	0.428	0.0	0.0	0.0	0.0	1.0
car_category	3.652	0.478	2.0	3.0	4.0	4.0	4.0
power	6.411	3.598	2.0	2.0	8.0	10.0	10.0
acceleration	81.540	26.373	49.0	56.0	75.0	95.0	184.0
composition	6.297	5.873	0.0	0.0	8.4	12.1	16.4
fuel_type	1.606	0.488	1.0	1.0	2.0	2.0	2.0
model_year	2.016.793	13.320	1901.0	2018.0	2020.0	2020.0	2021.0
unladen_mass	1.424.630	529.472	775.0	953.0	1295.0	2073.0	3320.0
euro_classification	5.869	0.395	4.0	6.0	6.0	6.0	6.0
fuel_consumption	6.381	1.670	3.7	4.8	6.0	7.8	12.7
energy_label	0.673	10.119	0.0	0.0	0.0	1.0	5.0
top_speed	170.871	24.004	90.0	160.0	160.0	187.0	250.0
weather_comfort	11.655	3.705	-0.61	9.17	12.06	13.79	283.0
dewPoint	7.826	1.864	-2.0	7.0	8.22	9.0	13.22
skyInfo	14.676	5.851	1.0	9.0	18.0	18.0	25.0
daylight	0.904	0.294	0.0	1.0	1.0	1.0	1.0
distance	15.328	7.942	0.28	9.5	14.15	20.64	43.24
humidity	77.420	16.082	25.0	65.0	79.0	93.0	100.0
windSpeed	8.453	5.539	0.0	3.71	7.41	11.12	33.36
visibility	17.020	16.863	0.0	0.1	11.1	30.25	75.48
temperature	12.082	33.610	-0.61	10.0	12.22	14.11	24.0