University of Zurich^UZH

Modelling Individuals' Car Accident Risk by Trajectory, Driving Events, and Geographical Context: A Comparison Study with Multiple Machine Learning Models on Two Countries

GEO 511 Master's Thesis

Author
Livio Brühwiler
15-711-385

Supervised by
Dr. Cheng Fu
Prof. Dr. Haosheng Huang (Haosheng.Huang@UGent.be)

Faculty representative
Prof. Dr. Robert Weibel

31.01.2021
Department of Geography, University of Zurich

Modelling Individuals’ Car Accident Risk by Trajectory, Driving Events, and Geographical Context:

A Comparison Study with Multiple Machine Learning Models on Two Countries

Abstract

With the advent of GPS-tracking technologies, car insurance companies have started to adopted usage- and behaviour-based insurance policies, where the insurance fee is calculated based on the safety of the customers' driving behaviour. These policies should provide a financial incentive for safer driving behaviour. Although many risk models for assessing an individual drivers’ accident risk based on driving behaviour and exposure exist, these models so far do not take the underlying geographical context of the driven trajectories and driving events into account. This study explores this research gap, by taking into account weather, land-use and points of interest (POI) as geographical context variables. GPS and driving events data from two study areas in the United Kingdom and Italy were available. Five different machine learning models, logistic regression, random forest, XGBoost, feed-forward neural network (FFNN), and a recurrent neural network with long short-term memory (LSTM) architecture were implemented and compared to perform a binary classification, which separates accident- from accident-free drivers. Several features derived from the trajectories, driving events, and geographical data were computed. The results show that the inclusion of geographical information can increase the relative predictive performance in terms of AUC by up to 10%, with XGBoost generally yielding the best performance and making the most use out of the spatial information in Italy. Random forest, logistic regression, and FFNN yield the best performance in the UK depending on the feature set and performance metric. Land-use contributes most to the performance improvement in Italy, while weather contributes most to the performance improvement in the UK, with higher levels of improvement in Italy. This study also confirms the results of previous studies that logistic regression is only very slightly outperformed by more expensive models if geographical information is not included. Therefore logistic regression can still be the preferred model for car accident risk prediction due to its simplicity and interpretability if the maximum performance is not the main aim. In terms of real-world application, the results outline the potential of including geographical information in the context of usage-based car insurance risk modelling, improving its accuracy, which should lead to fairer usage-based insurance policies.

Keywords: Pay-how-you-drive, Usage-based insurance, Machine-learning, Geographical information, Land Use, Points of interest (POI), Weather, Driving behaviour, Telematics

Acknowledgements

This thesis was written during a strange and difficult time for most of us. The more I am grateful for the excellent supervision, support and valuable inputs of my two supervisors Dr. Cheng Fu and Prof. Dr. Haosheng Huang. They were always able to help me with any difficulties and their ideas and advice substantially improved this thesis. I also want to thank the faculty representative Prof. Dr. Robert Weibel for being very helpful in sorting out the access to the data. Further, I want to acknowledge the European Community Horizon 2020 project no. 780754 “Track and Know” in the context of which this research took place and which provided me with access to the raw data. Lastly, I want to acknowledge everybody else who supported me during this year.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Motivation and Background

Every year around 1.35 million people die as the result of a road accident, with it being the leading cause of death among children and young adults between 5 and 29 years of age (WHO, 2020). A vast majority of these accidents are caused by human error. In both the United States and Switzerland, the percentage of car accidents that involve some form of human error lies around 94% (National Highway Traffic Safety Administration, 2018.; Swiss Federal Office of Statistics, 2018). Therefore, out of the three main factors which are mainly thought to influence road safety: driver, road, and vehicle (Eboli et al., 2017; Wang et al., 2015), the most important but also hardest to change is the driver. In order to turn unsafe drivers into safer drivers, they first need to be identified. However, according to Nees (2019), 80.3% of drivers self-describe their driving ability as above average. Therefore self-estimation does not seem to be an optimal method for assessing a driver's risk level, and further options should be explored.

The car insurance industry has been trying to classify drivers into different risk categories for years. However, traditionally car insurance companies only consider factors such as gender, age, or vehicle model for their rate-making (Lemaire et al., 2015), out of which only the vehicle model can be changed or influenced by the driver. In recent years with GPS tracking devices becoming more easily and cheaply available, insurers have started to adopt usage-based insurance policies (UBI), including Pay As You Drive (PAYD) and Pay How You Drive (PHYD): The former taking into account the mileage or exposure of a driver; The latter specific driving behaviour such as braking, acceleration or speeding (Tselentis et al., 2016). Such usage-based policies can provide a financial incentive for drivers to adopt a safer driving style. A similar approach has already partly been tested in the healthcare industry, where healthy behaviour is promoted by lower premiums and tracked through biometrics such as smartphone step counters (Ryder et al., 2019). The challenge in applying this methodology to the car insurance industry lies within converting the raw GPS-Data into meaningful and explainable variables that reflect a driver's risk profile and can assist the driver in adopting a safer driving style. Due to the increased amount of information available, these models should also provide a more accurate reflection of a driver's risk level, which in turn increases the fairness of the overall insurance process, since it stops cautious drivers from subsidizing the driving of more risky or dangerous drivers, which Tselentis et al. (2016) call the cross-subsidies phenomenon.

Additionally, risky driving behaviour such as frequent and hard acceleration or braking can be associated with higher fuel consumption and noise emissions, therefore providing feedback and incentives to the driver through PHYD policies with a resulting reduction in risky driving behaviour might also have other environmental and societal benefits (Bordoff & Noel, 2008). Furthermore, PHYD policies can render variables that are often seen as unfair or discriminatory such as gender redundant by replacing them with driving behaviour and exposure variables (Ayuso et al., 2014, 2016; Verbelen et al., 2017). For example, the higher accident rate of male drivers compared to female drivers can be largely explained by more frequent driving in general, more frequent speeding, and more frequent night driving (Ayuso et al., 2016). Through PAYD or PHYD policies, male drivers who exhibit a safe driving style, therefore, will not have to face unfair higher premiums anymore. The same is true for young drivers, who also often face higher premiums.

Although many insurance companies have already started to offer PAYD and PHYD policies, the underlying risk models usually don't take the geographical and environmental context of the driven trajectories such as weather conditions or land use into account, although these factors, especially weather, have shown to have a significant impact on accident risk (Kantor and Stárek 2014; Winlaw et al. 2019). Husnjak et al. (2015) describe the inclusion of environmental factors as the most critical step in the further development of UBI.

1.2 Problem Statement and Research Questions

Currently, the development of PAYD and PHYD risk models is still at an early stage. Furthermore, the geographical and environmental context of the driven trajectories is usually not being considered. Said context is also important for driving behaviour variables. For example, hard braking might be more dangerous on the highway than in the city and more dangerous in the rain than on sunny days (Husnjak et al., 2015). Including these geographical and environmental variables into a risk model might lead to a better assessment of individual driver risk, which, if communicated properly to the driver, can ultimately lead to drivers driving more carefully according to their circumstances and therefore higher road safety as well as lower and fairer insurance costs. This study aims to build on the existing research regarding PHYD risk modelling and improve it through the inclusion of geographical and environmental data, more specifically weather, land-use, and points of interest (POI) data. It is to be noted that this study does not aim to propose any specific rate-making methods for the car insurance industry, but rather focus on building an accurate and explainable classification model.

In order to achieve this task, several features will be derived from the driving behaviour and geographical data available to perform a binary classification to separate *accident* from *accident-free* drivers.

More specifically the following machine-learning techniques will be employed: Logistic regression, random forest, XGBoost and two types of neural networks: A simple feed-forward neural network (FFNN) and recurrent neural network (RNN). Furthermore, this study will also focus on the interpretability of these models, and compare the tradeoff between predictive performance and interpretability.

To achieve these goals, GPS positional data and driving events data for the year 2017 was available from two study areas: The greater London area in the United Kingdom, as well as the Tuscany and Rome area in Italy. In detail, the following research questions will be explored:

RQ1: Which driving behaviour variables are most suitable for predicting the accident risk of an individual driver, and to what extent can environmental and spatial information such as weather, points of interest, and land use improve this prediction?

Hypothesis 1: According to the literature, car accident risk is influenced by geographical factors, including this information should improve the predictive performance of risk models.

RQ2: Which machine learning techniques perform best for predicting car accident risk, and what is the trade-off between interpretability and predictive performance?

Hypothesis 2: More complicated models (e.g., random forest, XGBoost, neural networks) will perform better than the baseline model (logistic regression) but yield worse interpretability in turn.

RQ3: Are there geographical and cultural differences between London and Italy regarding the effects of driving behaviour on accident risk?

Hypothesis 3: Since the study areas are geographically and culturally different, it is to be expected that there will be some differences in model performance and variable importances.

The rest of this study will be structured as follows: Chapter 2 describes the state of the current research regarding PAYD and PHYD modelling, as well as the impact of several geographical factors on car accident risk. Chapter 3 provides an overview of the data, which was available and used in this study. Chapter 4 discusses the methodology of the data preprocessing, feature extraction and the model building. The results are presented in Chapter 5 and subsequently discussed in Chapter 6 where the limitations of this study are also shown. Finally, Chapter 7 serves as a conclusion and points out possible options for further research in this area.

Chapter 2: State of Current Research

2.1 Car Accident Risk Models

In general, car accident risk can either be predicted from a driver-centred or a location-centred perspective. In the first case, we want to know who will have an accident. In the second case, we want to know where accidents are most likely to happen. This thesis will mainly focus on the driver's perspective. Predicting accident risk from a driver's perspective is a key concept in the car insurance industry. Therefore, most existing researches lie in this context. Furthermore, there are several approaches to calculating a driver's accident risk (Huang & Meng, 2019). One is to model the expected number of claims (e.g., 0, 1, 2, 3), where generalized linear and additive models (GLMs and GAMs) such as Poisson regression are often used (Denuit et al. 2007). Another is a binary or, in some cases, multi-level classification approach, where the goal is to separate the drivers into different risk categories, e.g., those with and without an accident in the observed time period. In this case, GLMs such as Logistic regression, tree-based algorithms such as random forest or XGBoost, and several variations of neural networks are often applied. This thesis focuses on the latter approach, as in separating accident drivers from accident-free drivers.

2.1.1 Exposure Based (Pay As You Drive)

Pay as you drive (PAYD) policies are exposure-based insurance policies. Since each kilometre travelled by a driver significantly increases the risk of an accident (Litman 2005; Lemaire, et al., 2015; Boucher et al., 2013), logically, the premium should be higher for drivers who travel longer distances (Denuit et al. 2007). Early ideas of implementing mileage into insurance pricing include pay-at-the-pump (Sugarman, 1994), where a surcharge is applied for each litre of petrol, and self-reported mileage estimates, with occasional verification by the insurance company (Litman, 2011). However, with the advent of modern GPS technologies, much more detailed information about the mileage or exposure can be obtained. Further, this exposure can be divided temporally such as driving during peak hours or night driving, and geographically into urban driving versus rural driving, as demonstrated by Paefgen et al. (2013, 2014), Baecke & Bocca (2017), Guillen et al. (2019) and Ayuso et al. (2014, 2019). To summarize, exposure can be described as where, when, and how much someone drives (Baecke & Bocca, 2017).

One thing to note is that several researchers have stated that the relationship between mileage and accident risk is not always linear, as there seems to be a learning effect for high-mileage drivers. In other words, higher mileage results in a lower per-mile crash rate (Janke, 1991; Langford et al., 2008; Litmann, 2011; Paefgen et al., 2014; Guillen et al., 2019). Besides high-mileage drivers being more skilled, there are other explanations for this observation (Litman, 2005; Bordoff & Noel, 2008): First, high-mileage drivers tend to have a higher percentage of motorway driving, which is generally considered safer than city driving. Second, new vehicles that are technically safer tend to get driven more in comparison to older vehicles. Third, drivers who might be especially at risk, such as very old or very young drivers, on average drive smaller distances. This non-linear relationship should be considered when building a risk model based on exposure data.

Some examples of PAYD studies include Ayuso et al. (2019), who used Poisson regression to model the number of claims using a sample of young drivers in Spain. They used several traditional explanatory variables such as age, gender, and driving experience, as well as some vehicle-specific variables such as power and age. Furthermore, they included several exposure-based variables derived from telematics data. Besides the total distance travelled, these include the fraction of night driving, driving above the speed limit, and driving in urban areas. The effect of all these telematics variables proved to be significant, with improved performance over the traditional variables.

Baecke & Bocca (2017) used a classification approach, including logistic regression, random forest, and neural network. Driver, vehicle, and exposure specific variables were used. Compared to Ayuso et al. (2019), a larger number of explanatory variables was considered, specifically, more detailed information about the location (road type), time of day, as well as information about past claims. They also investigated the amount of driving data necessary to obtain an accurate model and found that three months is already sufficient.

Similarly to Baecke & Bocca (2017), Paefgen et al. (2014) used logistic regression to model the risk of accident involvement, although involving a smaller sample size and time window than Baecke & Bocca (2017). They found mileage to be the strongest predictor, with a non-linear relationship between mileage and accident risk. In an earlier study, Paefgen et al. (2013) also used the same dataset and similar explanatory variables, applying logistic regression, neural networks, and decision trees, with neural networks yielding the best predictive performance. However, due to the minor performance loss of logistic regression compared to neural networks and the vast improvement in interpretability, they recommended the usage of logistic regression, which lead to a follow-up study by Paefgen et al. (2014). One limitation of the studies by Paefgen et al. (2013, 2014), which they acknowledge, is that they did not have access to demographic data about the drivers, therefore making a comparison to traditional models impossible. This limitation also applies to this study. They further recommend the inclusion of speed limit violations, as well as route familiarity in future studies. Additionally, they point out the option of replacing mileage with time-driven as an exposure factor.

A different approach was proposed by Ayuso et al. (2014, 2016), who used survival analysis to model the distance and time to the first accident for different driver groups such as novice and more experienced drivers and men and women. Several exposure factors, such as night and urban driving, as well as speeding, were used, which correlated with lower distance and time driven to the first accident. Furthermore, they found that the difference in accident risk between men and women can largely be explained by factors such as more frequent urban driving, higher mileage, and more frequent speeding.

Verbelen et al. (2018) used generalized additive models (GAM), which help in modelling non-linear relationships between the explanatory variables and the target variable. Similarly to Ayuso et al. (2016), they found that the inclusion of exposure variables renders the gender variable redundant. In a similar fashion, Boucher & Turcotte (2020) used several exposure-based variables, such as mileage, time-driven and the number of trips to model claim frequency using GAM and generalized additive models for location, scale, and shape (GAMLSS).

Pesantez-Narvaez et al. (2019) compared logistic regression with XGBoost using several driver, vehicle, and exposure-based variables, as well as driving over the speed limit. Although XGBoost performed very well on the training set, it suffered from overfitting and an unbalanced dataset, which made the increased computational and explanatory efforts compared to logistic regression unviable. However, they state that XGBoost might perform better with a higher number of explanatory variables and further hyperparameter-tuning.

2.1.2 Behaviour Based (Pay How You Drive)

Pay how you drive (PHYD) is an extension of PAYD, where in addition to the exposure or how much someone drives, driving behaviour such as speeding, braking and acceleration is also considered (Tselentis et al., 2016). It should be noted that there is no clear definition separating PAYD and PHYD, with some authors such as Husnjak et al. (2015) classifying models that use certain exposure variables such as location or time of day as PHYD, while others classify them as PAYD (e.g., Tselentis et al., 2016; Verbelen et al., 2018).

Af Wåhlberg (2000, 2004, 2007, 2008a) has completed a lot of research on traffic safety in general and specifically regarding the impact of acceleration and braking behaviour on the accident involvement of bus drivers. He discovered a significant correlation between the frequency of acceleration events and involvement in accidents. Furthermore, he states that frequent speed changes might be indicators of other dangerous driving behaviour such as tailgating or harsh steering actions. Similarly, Stipancic et al. (2018) investigated the relationship between braking and acceleration events and crash frequency on a location level. They discovered a positive relationship between these driving events and locations with high crash frequency. While such driving events can also be used by skilled drivers as evasive manoeuvres in response to unexpected circumstances, those situations could still be avoided by having a safer and more cautious driving style in the first place (Musicant et al., 2010). From a statistical point of view, a high frequency of such events compared to other drivers can therefore be an indicator of dangerous driving behaviour (Musicant, et al., 2010).

Ma et al. (2018) found hard braking, acceleration, and speeding to have a high correlation with accident risk. Furthermore, they integrated observations with the traffic flow, as in comparing the drivers' speed with other vehicles on the same road segment. As opposed to other studies, Ma et al. (2018) implemented their models on a trip level instead of on a driver level, which lead to a largely imbalanced dataset.

In addition, Huang & Meng (2019) found several prediction models (logistic regression, Poisson regression, random forest, XGBoost, support vector machines (SVM), artificial neural network) which included braking, speeding, lane change and sudden turning as explanatory variables to have significantly better performance compared to the same models using only demographic data such as age and gender. As previously suggested by Paefgen et al. (2014), they also implemented variables about travel irregularity, calculated through dynamic time warping between each trajectory. Furthermore, they pointed out a lack of interpretability in previous studies and implemented a variable binning process based on regression trees in order to increase interpretability.

Bian et al. (2018) further proposed a bagging-based ensemble-learning approach for a multi-level risk-classification on a driver level, separating drivers into five distinct risk levels, which outperformed several benchmark models such as logistic regression and Naive Bayes.

In addition, Yan et al. (2020) recently used convolutional neural networks (CNN) in combination with SVM to perform a multilevel risk classification.

Table 1 presents a summary of the aforementioned PAYD and PHYD studies and embeds the study at hand within their context.

Table 1: Summary of previous PAYD and PHYD studies in reverse chronological order

Study	Data	Predictor Variables	Models	Research Scope
Boucher & Turcotte (2020)	26998 Vehicles	E	Generalized additive models	Claim Number Prediction
Yan et al. (2020)	2000 Vehicles	E, EV	Convolutional Neural Network, Support Vector Machine	Multilevel Risk Classification
Ayuso et al. (2019)	25014 Vehicles	D, V, E, S	Poisson Regression	Claim Number Prediction
Guillen et al. (2019)	25014 Vehicles	D, E, S	Zero Inflated Poisson Regression	Claim Number Prediction
Huang & Meng (2019)	2151 Vehicles	D, E, EV, S, V	Logistic Regression, Poisson Regression, Support Vector Machine, Random Forest, Neural Network, XGboost	Claim Number Prediction, Accident vs Accident-Free Classification
Pesantez-Narvaez et al. (2019)	2767 Vehicles	D, V, E, S	Logistic Regression, XGBoost	Accident vs Accident-Free Classification
Bian et al. (2018)	198 Vehicles	E, EV, S	Logistic Regression, Bagging, SMO, Naïve Bayes, Locally Weighted Learning	Multilevel Risk Classification
Ma et al. (2018)	503 Vehicles	D, E, EV, S	Logistic Regression, Poisson Regression	Accident Number Prediction, Accident vs Accident-Free Classification
Verbelen et al. (2018)	33259 Vehicles	D, V, E	Generalized additive model	Claim Number Prediction
Baecke & Bocca (2017)	6984 Vehicles	D, V, E	Logistic Regression, Random Forest, Neural Network	Accident vs Accident-Free Classification
Ayuso et al. (2016)	8198 Vehicles	D, V, E, S	Survival Analysis (Weibull Regression)	Distance Travelled to First accident
Ayuso et al. (2014)	15940 Vehicles	D, E, S	Survival Analysis (Weibull Regression)	Distance and Time Travelled to First Accident
Paefgen et al. (2014)	1567 Vehicles	E	Logistic Regression	Accident vs Accident-Free Classification
Guo & Fang (2013)	102 Vehicles	D, EV	Logistic Regression	Risk Classification
Paefgen et al. (2013)	1567 Vehicles	E	Logistic Regression, Neural Network, Decision Tree	Accident vs Accident-Free Classification
Af Wåhlberg (2000, 2004, 2007, 2008)	Various Bus Drivers	EV	Correlation Analysis	Effect of Acceleration and Braking on Bus Driver Accident Involvement
This Study	**14584 Vehicles**	**E, EV, G**	**Logistic Regression, Random Forest, XGBoost, Neural Networks**	**Accident vs Accident-Free Classification; Influence of Geographic Features**

D: Demographic; V: Vehicle Specific; E: Exposure; EV: Driving Events; S: Speeding, G: Geographic

2.2 Impact of geographical variables on car accident risk

2.2.1 Weather

There is a vast amount of literature regarding the effect of various weather conditions on car accident risk. Bad weather conditions have been shown to increase the frequency of car accidents (Peng et al., 2018), as well as their severity (Fountas et al., 2020). The effect of rainfall and its consequences such as slippery roads, bad visibility, and hydroplaning has been investigated extensively and linked to higher crash frequency numerous times (e.g., Andrey & Yagar, 1993; Caliendo et al., 2007; Chang & Cheng, 2005; Bergel-Hayat et al., 2013), with a stronger effect after long dry periods, indicating that people need some time to adapt their driving behaviour to the new weather conditions (Brijs et al., 2008; Eisenberg, 2004). However, in some cases, rainfall has also been linked to lower crash frequencies, perhaps due to adapted driving behaviour or different exposure levels (Yannis & Karlaftis, 2010; Bergel-Hayat et al., 2013). Winter precipitation, which includes snowfall, freezing rain, and ice pellets or sleet, has been associated with significantly higher crash risk compared to dry conditions, especially in the evenings and bad visibility conditions (Eisenberg & Warner, 2005; Black et al., 2014). Similarly, high temperatures have shown a negative effect on driving ability and result in increased crash risk (Wyon et al., 1996; Wahlberg, 2006; Maliyshkina et al., 2009; Yannis & Karlaftis; Hayat et al., 2013). Fog has also been linked to higher accident frequency as well as severity (Eisenberg & Warner, 2005; Black & Mote, 2015). Furthermore, the joint effect of bad weather and bad lighting conditions, e.g., night driving in the rain, can further amplify the probability of driving errors, hazardous driving, and the resulting accidents (Fountas et al., 2020). Table 2 presents an overview of weather conditions that have been linked to increased car accident risk.

Table 2: Summary of studies regarding the impact of certain weather conditions on car accident risk

Weather Condition	Studies	Findings
Rain	Caliendo et al. (2007) Chang & Chen (2005) Yannis & Karlaftis (2011) Brijs et al. (2008) Bergel-Hayat et al. (2013) Andrey & Yagar (1993) Eisenberg (2004) Fountas et al. (2020) Andrey et al. (2003)	Mostly increased risk, especially after long dry periods; in some cases reduced risk, possibly due to adaption of driving behaviour
High Temperatures	Wyon et al. (1996) Af Wåhlberg (2008b) Malyshkina et al. (2009) Yannis & Karlaftis (2011)	Negative effect on driving ability
Very Low Temperatures	Maliyshkina et al. (2008)	Increased risk
Winter Precipitation	Black & Mote (2015) Eisenberg & Warner (2005)	Increased risk, especially in the evenings
Fog	Abdel-Aty et al. (2011) Wu et al. (2018)	Increased risk and injury severity

2.2.2 Temporal Information

Car accident risk does not only vary spatially but also temporally. Several previous PAID and PHYD studies have observed higher crash frequencies on peak hours and weekdays probably attributed to higher traffic volumes. Furthermore, frequent night driving, especially on weekends and Fridays evenings, has been attributed to higher accident probability, due to bad visibility and other factors, such as intoxicated driving (Paefgen et al., 2014).

2.2.3 Points of Interest (POI)

Research linking points of interest to car accident risk is sparse, perhaps due to the connection not being very intuitive at first glance. However, POI data is easily accessible and has a high capability of reflecting human behaviour patterns (N. Wang et al. 2019; Siła-Nowicka et al. 2016). Jia et al. (2018) found areas with a high density of banks, hospitals, and residential areas to have a higher crash frequency. Kufera et al. (2020) found a newly built casino in Maryland contributes to an increased crash frequency in nearby areas, especially on the weekends, which could also be applied to other gambling venues. Furthermore, Ng et al. (2002) discovered a significant positive relationship between cinemas, hospitals, markets, railway stations, and the number of accidents in Hong Kong. Yao et al. (2018) found retail stores and restaurants are linked with a higher frequency of vehicle-pedestrian collisions in Shanghai. Similarly, Lee et al. (2015) investigated pedestrian-vehicle collisions in Florida and found touristic POIs such as hotels, motels, guest houses, rail and bus stations, ferry terminals, and schools to be linked with a higher amount of vehicle-pedestrian collisions.

2.2.4 Land-Use and Land-Cover

Similarly to POIs, research linking land-use with car accident risk is sparse. However, the data is also easily available, and different types of land-use can reflect different human behaviour, e.g., different land-use types might attract different trip purposes and therefore different driving styles with varying levels of risk (Kim & Yamashita 2002). Several studies observed a higher crash frequency in commercial and residential land-use areas (Kim & Yamashita, 2002; Loukaitou-Sideris et al., 2007; Lym & Chen, 2020; Yang & Loo, 2016). More specifically, Yang and Loo (2016) found commercial land-use mixed with residential land-use to be linked with the highest crash frequency, whereas Kim & Yamashita (2002) observed higher crash frequency in commercial than in residential areas. The mixture of commercial and residential land-use has also been shown to increase the frequency of vehicle-pedestrian collisions (Y. Wang & Kockelman 2013; Wier et al. 2009). Additionally, rural and agricultural areas have shown to have lower crash frequencies than urban areas. (Kim & Yamashita 2002; Alkahtani et al., 2019)

One common problem with classifying the land-use of a point-type location such as a crash location or a GPS waypoint lies in its heterogeneity, as its surroundings can have multiple types of land-use. Yang & Loo (2016) solved this by applying a buffer radius of 100m around each point and calculating the percentage of each land-use type within this buffer. This approach allows for a classification of mixed areas, which is a more accurate representation of reality.

Land-cover and land-use are often confused and used interchangeably. However, while they overlap each other spatially, land-use is a socio-economic interpretation of the way humans use the earth's surface, whereas land-cover is a direct observation of its physical properties, usually derived from satellite imagery (Fischer et al., 2005; Comber 2008). Less literature on the impact of land-cover and accident risk is available compared to land-use. Some previous studies explore the relationship between land-cover and traffic-wildlife collisions, e.g. (Neumann et al., 2012).

2.2.5 Other

Several other environmental and anthropogenic factors that can have an impact on car accidents are described in the literature. This includes large events (Gutierrez-Osorio & Pedraza, 2020), air pollution (Wan et al., 2020), and even the stock market (Giulietti et al., 2020).

2.3 Prediction Models

The following section presents a brief overview of the most frequently used machine-learning models for car accident prediction on a driver level, which will later also be implemented in this study. The focus will be on (binary) classification models, which allow for the separation of accident- and accident-free drivers.

2.3.1 Logistic Regression

Logistic regression is a generalized linear model, which can be used for classification purposes. Unlike linear regression, the optimal parameters are found through maximum likelihood estimation rather than through ordinary least squares (OLS) estimation (DiGangi & Hefner, 2013). This, coupled with the binary response in logistic regression is the main difference compared to linear regression. The general equation can be written as follows:

$$P = \frac{1}{1+\mathrm{e}^{-(\beta_0+\sum \beta_i X_i)}}$$

Where in the context of this study P represents the probability of a driver having an accident, β_0 is the intercept and $\beta_i X_i$ represent the coefficients and the corresponding features belonging to the driver. Using a predefined threshold, usually 0.5, this probability can then be transformed into a classification. The advantage of logistic regression lies in its interpretability, where the effect-size and direction of each feature can be interpreted separately. In a real-world insurance application, this interpretability is in many places required by law (Baecke & Bocca, 2017). Many studies have used Logistic Regression to predict accident probability on a driver level, and it can be used as a benchmark for more sophisticated and black-box models, such as Huang & Meng (2019), Baecke & Bocca (2017) and Paefgen et al. (2013, 2014). In many of the aforementioned studies, logistic regression is only very slightly outperformed by more expensive and complicated algorithms such as XGBoost, random forest, and neural networks. Combined with its superior interpretability, logistic regression is the selected model in many classification scenarios.

2.3.2 Random Forest

Random forest (Breimann, 2001), is a machine-learning algorithm that can be used for classification and regression tasks. It consists of an ensemble of decision trees, where for each tree a random subset of the predictor variables and a bootstrapped sample of the data is chosen. In the end, the final classification consists of the majority classification of all trees. The main idea is to build a strong classifier consisting of many weak classifiers. Random forests can improve the tendency of decision trees to overfit (Baecke

& Van Den Poel, 2011). Another advantage is their ability to model nonlinear and highly-complex relationships, as well as the possibility to account for imbalanced datasets (Strobl et al., 2007). On top of that, they are relatively fast to train. In order to improve interpretability, an importance ranking of the explanatory variables can be computed in several ways. The most common way to calculate feature importance in a random forest is the so-called Gini impurity. In this way, the average decrease in Gini impurity is averaged for each feature over all trees. It should be noted that this way of measuring feature importance tends to favour continuous features and categorical features with many different values (Strobl et al., 2007). Random forests have been deployed in the context of car accident prediction in several PAYD and PHYD models (Huang & Meng, 2019; Baecke & Bocca, 2017). Furthermore, they have been employed in many other research fields, such as bioinformatics (Strobl et al., 2007), air pollution modelling (Kamińska, 2018), remote sensing (Liu et al., 2018), etc.

2.3.3 XGBoost

XGBoost is another tree-based method, which builds on gradient tree boosting (T. Chen and Guestrin, 2016). It is a state-of-the-art algorithm for both classification and regression problems. It has been used frequently in various machine-learning competitions and across many research fields and industries.

For a classification problem, XGBoost builds a series of classification trees, where each tree uses the residuals of the previous trees to correct their errors. In order to avoid overfitting, a regularization term is applied. For optimal performance, various hyperparameters need to be tuned, which will be described in more detail in the methodology section in Chapter 4. From an interpretation standpoint, similarly to a random forest, several feature importance measurements can be computed. One of them is so-called gain. Gain describes the accuracy improvement a feature yields to the branches it is on. (XGBoost, 2020). XGboost has been employed in many fields, such as price forecasting (Gumus & Kiran, 2017), engineering (Zhang et al. 2018), road accident prediction (Schlögl, 2020; Parsa et al., 2020) and more. In the context of car accident risk classification, XGBoost has been employed by Pesantez-Narvaez et al. (2019) and Huang & Meng (2019).

From an implementation point of view, XGBoost features high computational performance. It is available in several programming languages, such as Python, R, Julia, Java, C etc. (XGBoost, 2020).

2.3.4 Neural Networks

Neural networks are inspired by biology and try to simulate neurons in the human brain. Several different neural network architectures exist, which have been used for a plethora of tasks and across many research fields. They are able to produce very high-performance numbers. However, they generally require a lot of computing power but feature low interpretability and have a tendency of overfitting. Figure 1 provides an example of a simple neural network with one fully connected hidden layer.

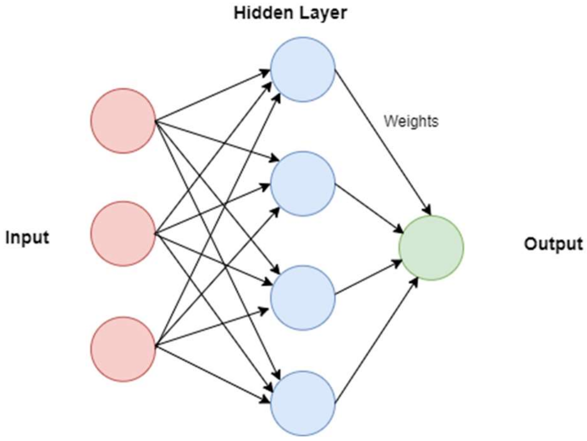

Figure 1: Example of a simple neural network consisting of one fully connected hidden layer.

For this study, besides a simple feed-forward neural network (FFNN), a variation of a long short-term memory (LSTM) architecture will be used, which belongs to the family of recurrent neural networks (RNN). Recurrent neural networks are a type of neural networks, which are able to process sequence-based data. LSTM models (Hochreiter and Schmidhuber, 1997) have been used in several applications recently, such as flood prediction (Fang et al., 2020), traffic forecasting (Cui et al. 2020; Y. Y. Chen et al., 2016), sentiment analysis (Behera et al., 2021; Zhao et al., 2020), marketing analytics (Sarkar & De Bruyn, 2021), power grid loss prediction (Tulensalo et al., 2020) and many more. LSTM models are able to capture long term dependencies in sequential data. Furthermore, in comparison to traditional models, instead of feature engineering and domain knowledge, they rely more on raw data (Sarkar & De Bruyn, 2021). In comparison to traditional RNN, LSTMs do not suffer from the vanishing gradient problem and are able to memorize information over many timesteps. (Oehmcke et al., 2018). This is done through the inclusion of different gating mechanisms (Figure 2), which allows the cell to decide, which information to keep and which information to forget.

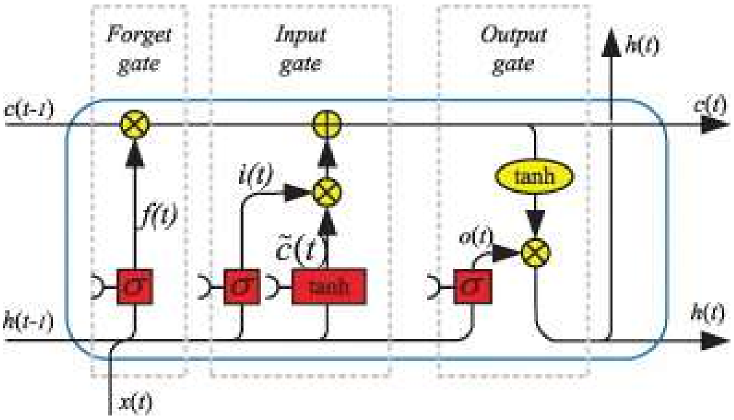

Figure 2: LSTM Cell with input, output and forget gate. Source: Yu et al. (2019).

2.4 Research Gaps

Many studies exist regarding various PHYD and PAYD models, using a large variety of exposure and behaviour-based driving variables as well as several different machine learning techniques. However, there seems to be no consensus as to which combinations of variables and techniques are optimal. Simultaneously, the impact of environmental conditions such as weather, POI, and land-use on car accident frequency is well documented, especially the relationship between weather and car accidents has been studied extensively. However, it has mainly been done from a location-centred or temporal-centred perspective while not from a driver's perspective. Therefore, based on the previously reviewed literature, the following research gaps can be identified:

- The first research gap exists in combining exposure and behaviour-based risk modelling on a driver level with information about the geographical and environmental context, possibly increasing prediction accuracy. This research gap was also identified by Husnjak et al. (2015), who pointed out that the inclusion of environmental factors is the most important step forward in the development of future PHYD models. A large amount of such environmental data, e.g., POI or land-use data, is easily and freely available, pointing out that a vast amount of potentially valuable information, which could be obtained with little effort, remains as of today unused by traffic accident researchers and the car insurance industry.

- Secondly, none of the previous studies uses multiple study areas in order to compare cultural and geographical differences in the impact of different driving behaviour features on individual car accident risk.

- Lastly, another research gap can be identified in the data aggregation format. The previous studies mainly rely on data aggregated over several months. There exists a potential for trip-based models with finer temporal granularity, where the sequence and order of the driven kilometres are taken into account as well. Recurrent neural networks have often been employed in sequence-based classification tasks and could potentially also be employed for the car accident risk classification problem at hand.

Chapter 3: Data

3.1 Telematics Data

The telematics data used in this study is provided by a telematics company in the context of the Track & Know research project[1]. The data was collected by boxes mounted to the vehicles, which collected GPS and accelerometer data. The data was stored in a MongoDB and contained the positional data, driving events and crashes of around 400000 drivers. All the data was collected during the year 2017 and in two main geographic areas, greater London and Italy, which will serve as the study areas for this thesis, shown in Figure 3. For both study areas, most of the drivers that had a crash (after filtering crashes, see section 3.1.3) were sampled as well as a random sample of the crash-free drivers. In total, data for 2322 drivers were retrieved from the UK sample, out of which 397 had at least one crash, and 12262 drivers from the Italy sample, out of which 3925 had at least one crash. Overall the UK study area is more homogeneous and mainly urban, focused on the city of London, while the Italy area includes both urban areas such as Rome and more rural areas such as Tuscany.

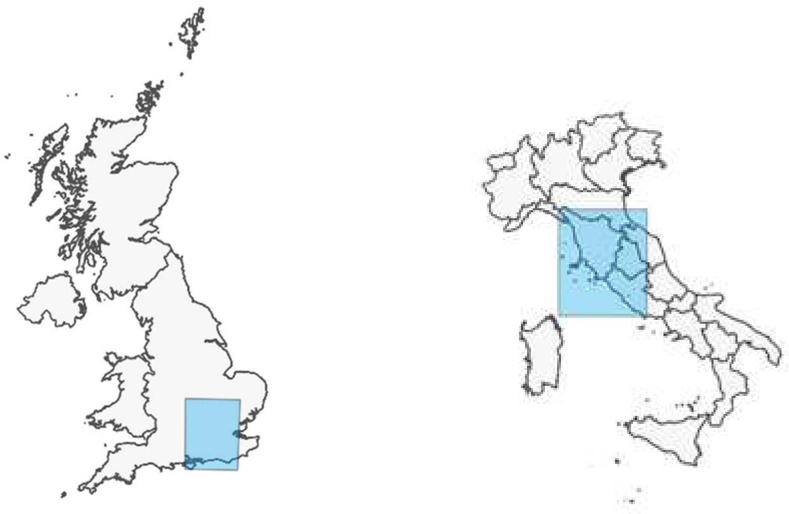

Figure 3: Bounding boxes of the study areas in the United Kingdom and Italy.

3.1.1 Positional Data

Positional data of the year 2017 for each driver was available in the following resolutions: One waypoint every 60 seconds for the London dataset and every 2000 meters for the Italy dataset. Table 3 provides an overview of the variables included in the positional data. Besides the coordinates, variables such as speed and heading are included. Overall the resolution of the London dataset is higher, while the Italy dataset contains a larger sample size. The impact of the different resolutions as well as different sample sizes will be discussed in the results section (Chapter 5). Furthermore, the location type variable had no value for most observations. There was no information on how it was derived. Therefore the location type was not used for the analysis.

[1] https://trackandknowproject.eu/

Table 3: Overview of positional data

Variable	Description	Example
Id	Unique waypoint ID	5c3ccf30ba39528a21458f51
T&K_VOUCHER_ID	Unique ID of the driver	3299
TIMESTAMP_LOCAL	Local time	2017-08-01 05:35:10
LATITUDE	Latitude in WGS 84 x 10^6	43942247
LONGITUDE	Longitude in WGS 84 x 10^6	10952413
SPEED	Instantaneous speed in km/h	44
HEADING	Heading from 0° to 360°	114
GPS_QUALITY	1 = No navigation 2 = partial navigation 2d 3 = full navigation 3d	3
STATUS	0 = Starting 1= Moving 2 = Stopping	1
DELTAPOS	Distance in meters to the previous position	2026
DELTATIME	Seconds since the previous position	353
PV	Province ID (Italy only)	69
LOCATION_TYPE	1 = City/Town 2 = Hamlet 3 = Urban 4 = extra urban 5 = Highway 6 = Others	2

3.1.2 Driving Events

In addition to the positional data, driving events, namely cornering, braking, acceleration, and quick lateral movement events were recorded if they exceeded a certain acceleration threshold. The exact threshold is not given by the data source. All drivers share the same threshold and labelling framework. Table 4 provides an overview of the variables included in the driving events data. Besides the coordinates and the variables included in the positional data, several other variables are present such as the maximum and average acceleration, as well as the duration of the event.

Table 4: Overview of driving events data

Variable	Description	Example
Id	Unique event ID	5c5004c9ba39528a213179e0
T&K_VOUCHER_ID	ID of the driver	396781
TIMESTAMP_LOCAL	Local time	2017-12-22 21:03:11
LATITUDE	Latitude in WGS 84 x 10^6	51439241
LONGITUDE	Longitude in WGS 84 x 10^6	275750
SPEED	Instantaneous speed in km/h	34
HEADING	Heading from 0° to 360°	234
GPS_QUALITY	1 = No navigation 2 = partial navigation 2d 3 = full navigation 3d	3
STATUS	0 = Starting 1= Moving 2 = Stopping	1
PV	Province ID (Italy only)	0
LOCATION_TYPE	1 = City/Town 2 = Hamlet 3 = Urban 4 = extra urban 5 = Highway 6 = Others	2

EVENT_TYPE	A = Acceleration
	B = Braking
	C = Cornering
	Q = Quick lateral movement
		B

AVG_ACCELERATION	Average acceleration of the event in 9.81 x 10^{-6} m/s^2	615
MAX_ACCELERATION	Maximum acceleration of the event in 9.81 x 10^{-6} m/s^2	1488
EVENT_ANGLE	Angle of the event from 0° to 360°	-4
DURATION	Duration of the event in milliseconds	850

Figure 5 shows the frequencies of the driving events. Figure 4 provides an overview of the hourly distribution of the driving events.

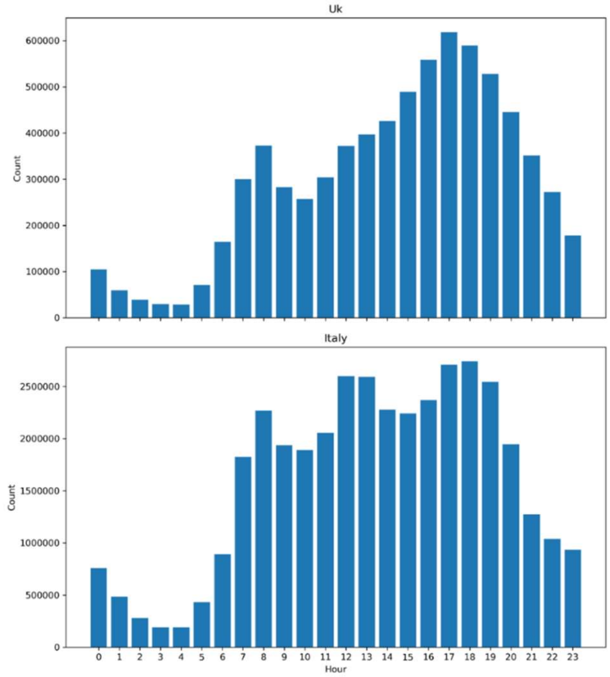

Figure 4: Hourly Count of Driving Events in the UK and Italy. In both study-areas, a clear peak is visible during 17:00 – 19:00. The peak is stronger in the UK area. Furthermore, there seems to be a second peak during 12:00 – 14:00 for the Italy data, which is not apparent in the UK data.

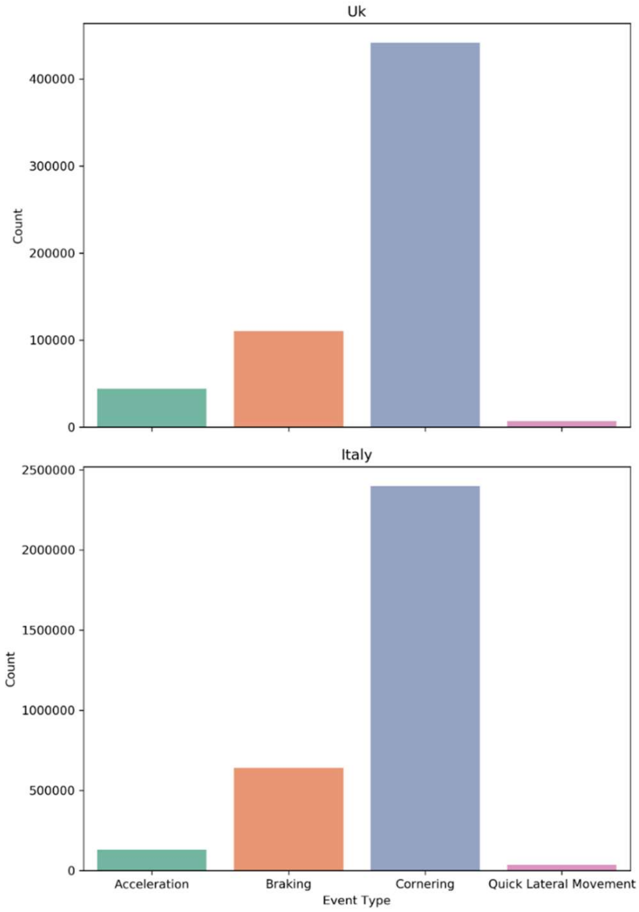

Figure 5 Events per Category in the UK and Italy. Cornering events are by far the most common driving events registered, followed by braking events. This distribution seems to be the same for both study-areas

3.1.3 Crashes

Similarly to driving events, crashes were recorded using the vehicle black box. Exceeding a certain acceleration threshold will trigger a crash alarm, which will then be validated by an automatic system or a crash assistance centre. In total, there were 145233 crash alarms, out of which 10118 were manually or automatically validated as real crashes. This number was further filtered down since it still included unrealistic sequences of many crashes in a row, crashes which had comments saying they are false alarms, and several duplicates. The final number of crashes included in this study is 4742, out of which 423 happened in the UK and 4319 happened in Italy. It is to be noted that there is still the possibility of false alarms being included in the final sample. However, due to the preprocessing, the vast majority should be real crashes or very close near-miss events. So even if some of the remaining crashes are still false alarms, they are most likely near-miss events due to the high acceleration. Figures 6 and 7 provide a general overview of the spatial distribution of the crashes in both study areas. Most crashes in the UK area are clustered in and around the city of London. In Italy, the crashes are mostly clustered in and around Rome as well as several bigger cities in the Tuscany area such as Florence, Pisa and Lucca.

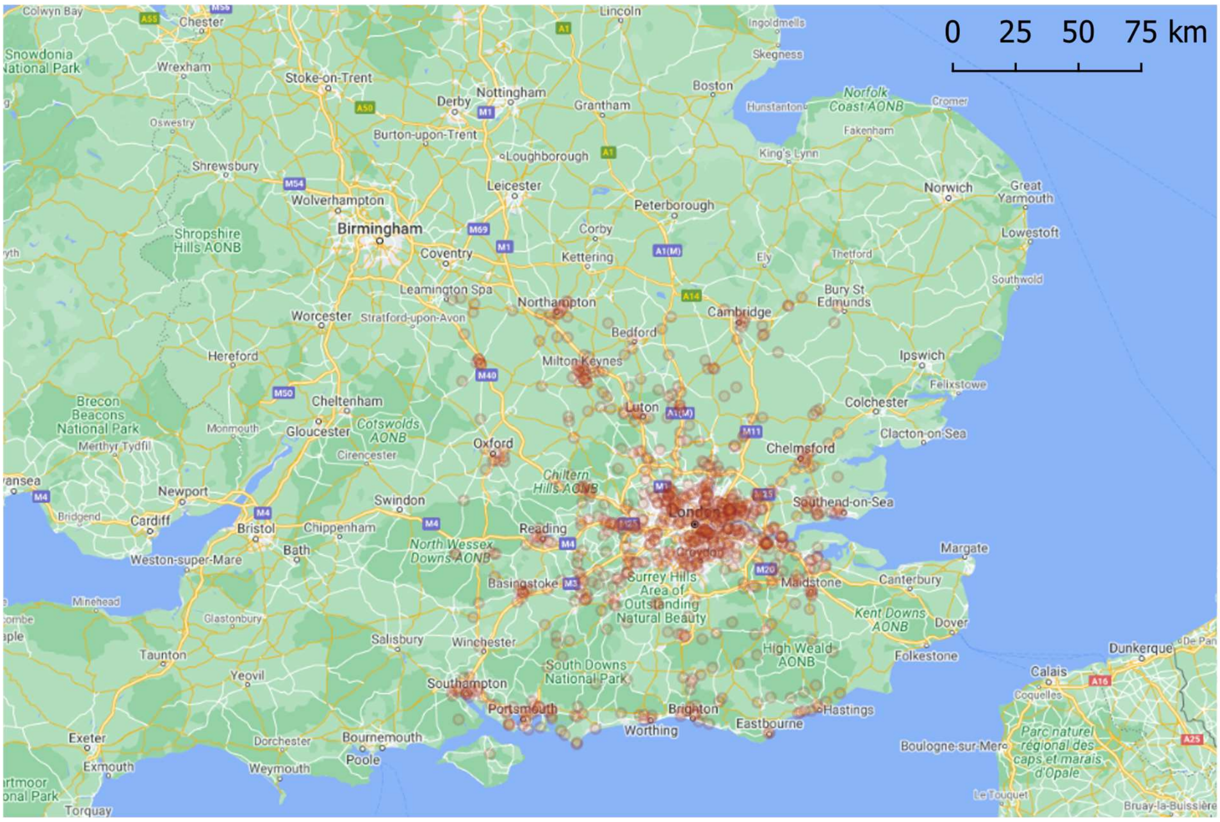

Figure 6: UK crash locations, background map: Google.

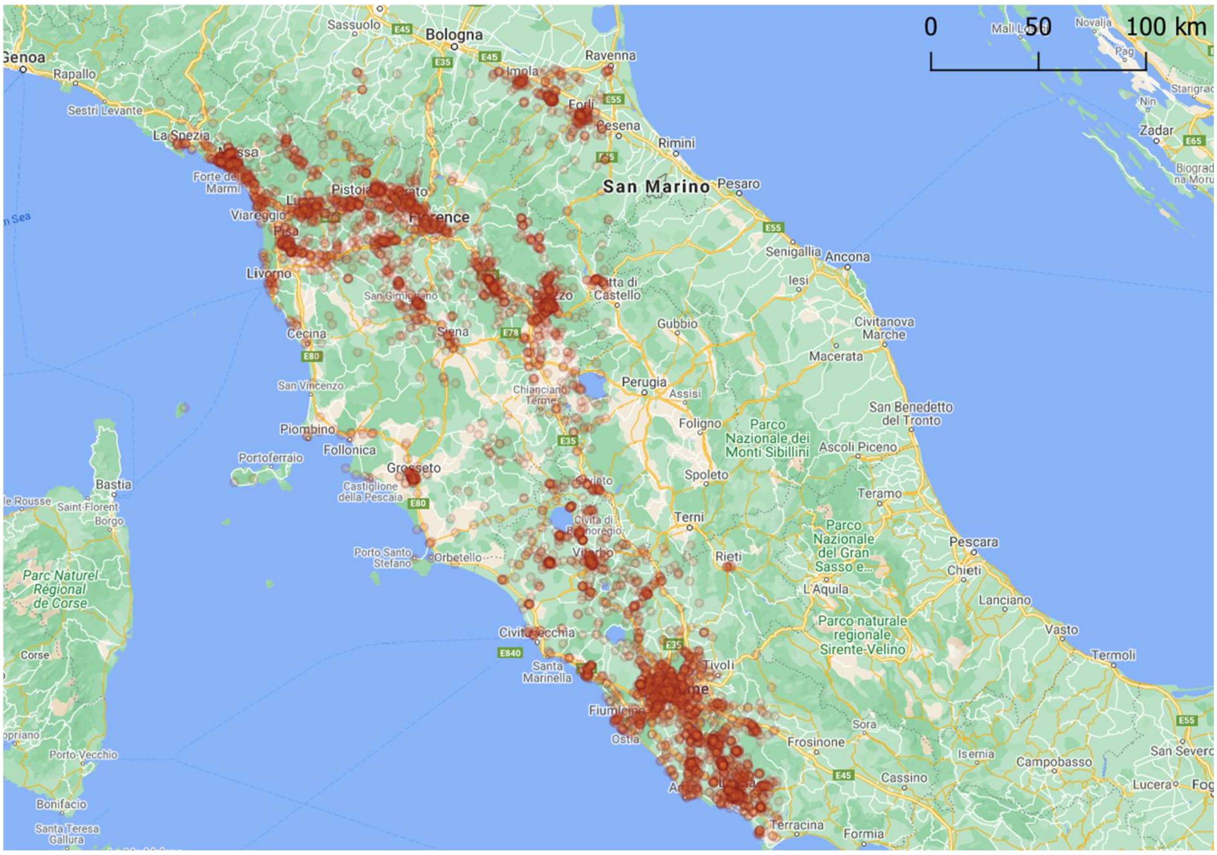

Figure 7: Italy crash locations, background map: Google.

3.2 Geographical Data

3.2.1 Weather

To enrich the trajectory data with information about the weather conditions, historical weather data was downloaded through the API of worldweatheronline[2]. It should be noted that this data is historical hourly forecast data and not actual measurements. Actual measurements from local weather stations were available from meteostat[3]. Unfortunately, meteostat does not provide precipitation data prior to 2018. Therefore, worldweatheronline was chosen as the source for the weather data. On worldweatheronline, historical weather data for any given location can be queried, and the spatially nearest record will be returned. For this thesis, a grid consisting of 10 * 10 km squares was built for both study areas and the weather data API queried using the centroid of each square. For both study areas, the hourly weather data for each month was then retrieved from the API and stored in a JSON file. A list of important variables contained in the weather data can be found in table 5. As the number of API calls per day is limited, a finer grid than 10 *10 km was not used. In total, weather data was retrieved for 274 locations in the UK area and 651 locations in the Italy area, as shown in Figures 8 and 9, which provide good spatial coverage of both study areas.

[2] www.worldweatheronline.com

[3] www.meteostat.com

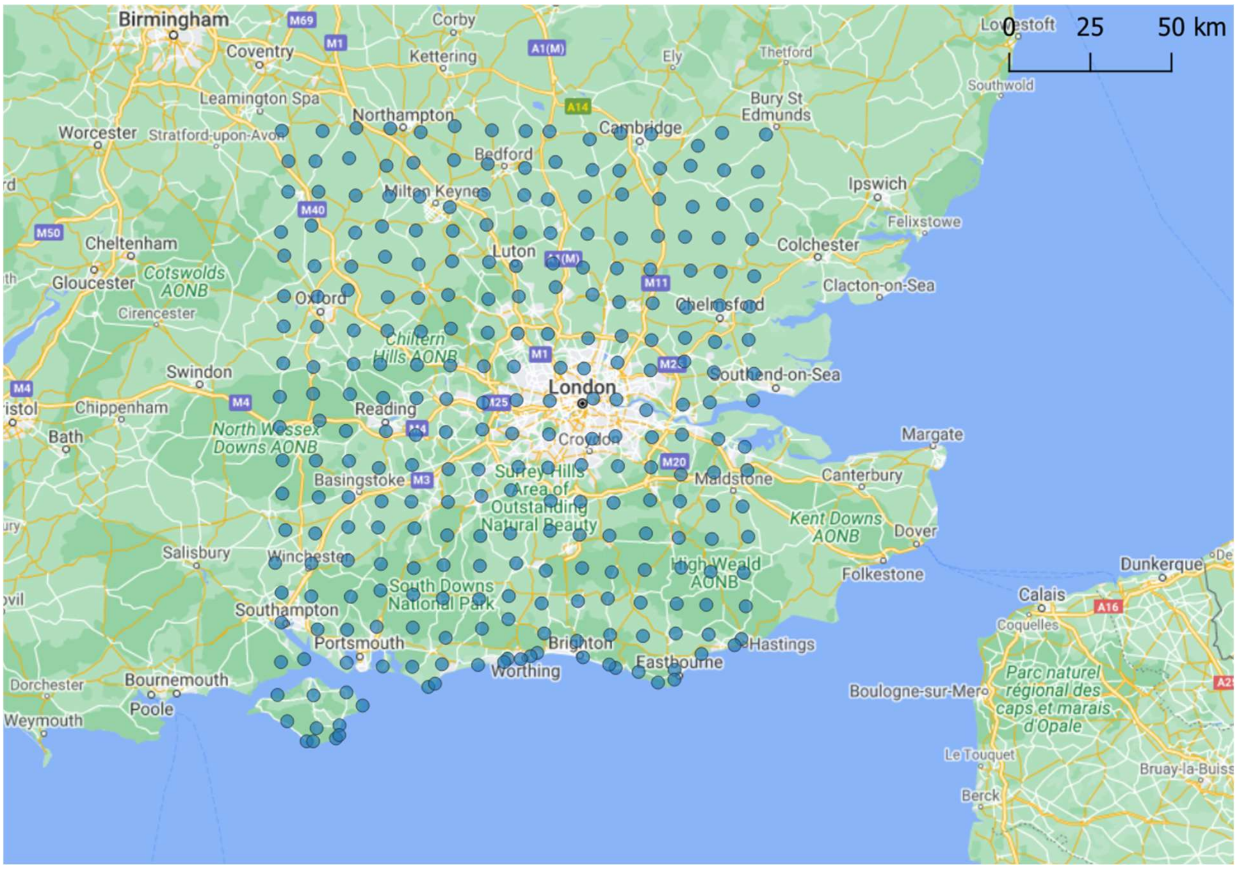

Figure 8: UK weather locations, background map: Google.

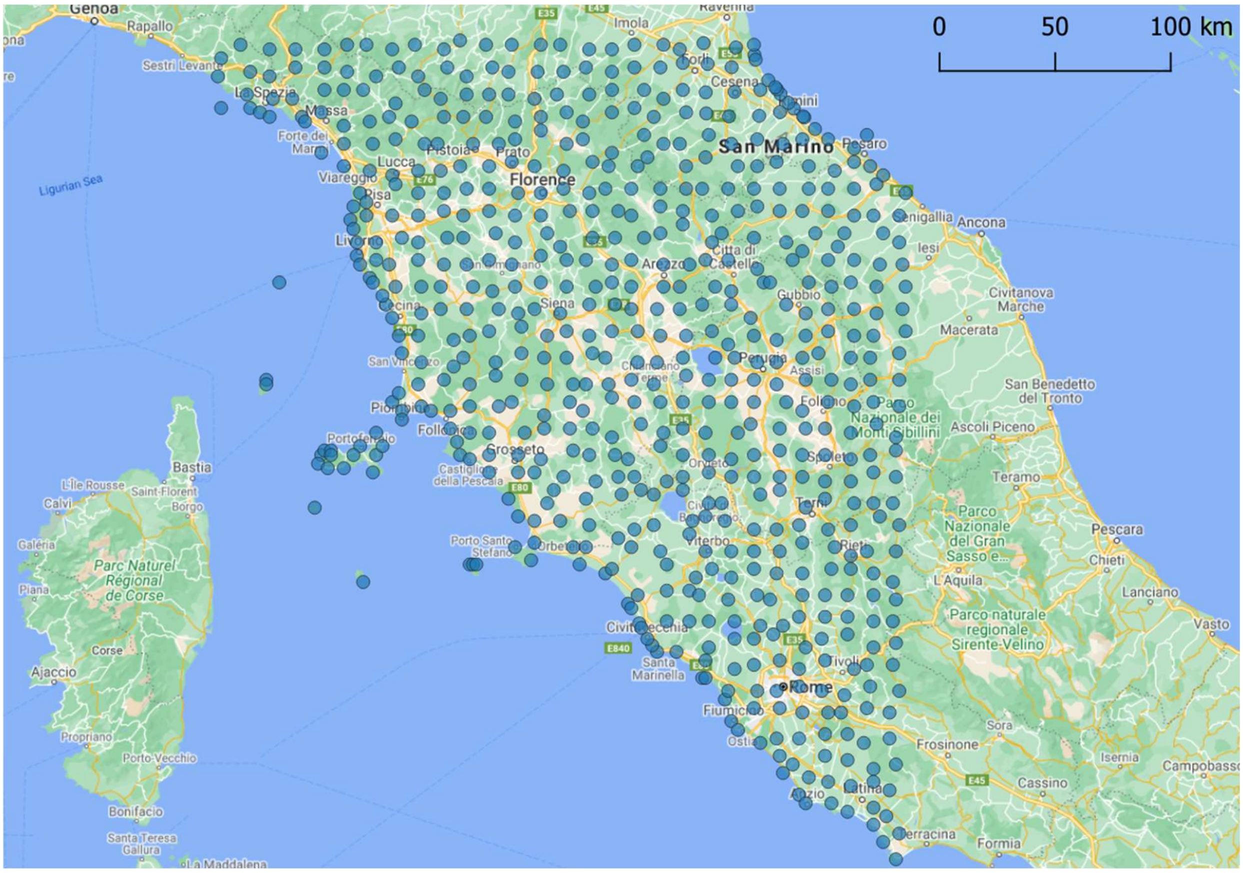

Figure 9: Italy weather locations, background map: Google.

Table 5: Overview of Weather Information Variables.

Variable	Unit
Temperature	°C
Humidity	%
Pressure	hpa
Precipitation	mm
Visibility	km
Weather Condition	Weather condition (e.g. Sunny, Rain)
Windspeed	km/h
Cloudcover	%

Although real weather data would have been preferred, this historical forecast data provides relatively high accuracy. The underlying model data used by worldweatheronline is provided by the World Meteorological Organization and the NCEP global forecast system (Worldweatheronline, 2020) and is also used by several big international companies, such as Qatar Airways, KLM, Coca-Cola and more. Furthermore, this forecast data follows the same terminology and classification methods for every region, which makes their output comparable, whereas local weather stations might have different standards of reporting their measurements and/or missing values. Lastly, since data is aggregated over a whole year, minor deviations from the real weather conditions do not matter that much and should even out over the course of the whole year.

3.2.2 POI

OpenStreetMap (OSM) POI data for both study areas were downloaded from geofabrik[4]. The following POI categories shown in Table 6 were considered, mostly according to section 2.2.5. This resulted in 458134 total POI for the Italy area and 924838 total POI for the UK area. Although the Italy area is bigger, the London area is more densely populated, which might be an explanation for the higher number of POI in the UK area, despite the overall study area being smaller. As OSM is volunteered geographic information (VGI), it is possible that the UK OSM community is more active than the Italian counterpart.

Table 6: Overview of POI data.

Category	Included POI
Commercial	Convenience Stores, Supermarkets, Pharmacies, Clothing Stores
Touristic	Attraction, Hostel, Hotel, Motel, Tourist Info
Nightlife	Restaurant, Pub, Cinema, Nightclub, Café, Bar, Fast Food
Public	Police station, School, Library, University, Kindergarten, Parks
Transportation	Bus stops, Railway stations, Taxi Stops

[4] www.Geofabrik.de

3.2.3 Land-Use

In a similar fashion, OSM land-use data was downloaded from geofabrik. The following land-use categories were considered:

- Industrial
- Commercial
- Farm
- Grassland
- Park
- Residential
- Forest
- Retail

For simplicity reasons and in order to avoid too many categories, commercial and retail were combined into *commercial*; Farm, Grassland and Park were combined into *rural*.

Chapter 4: Methodology

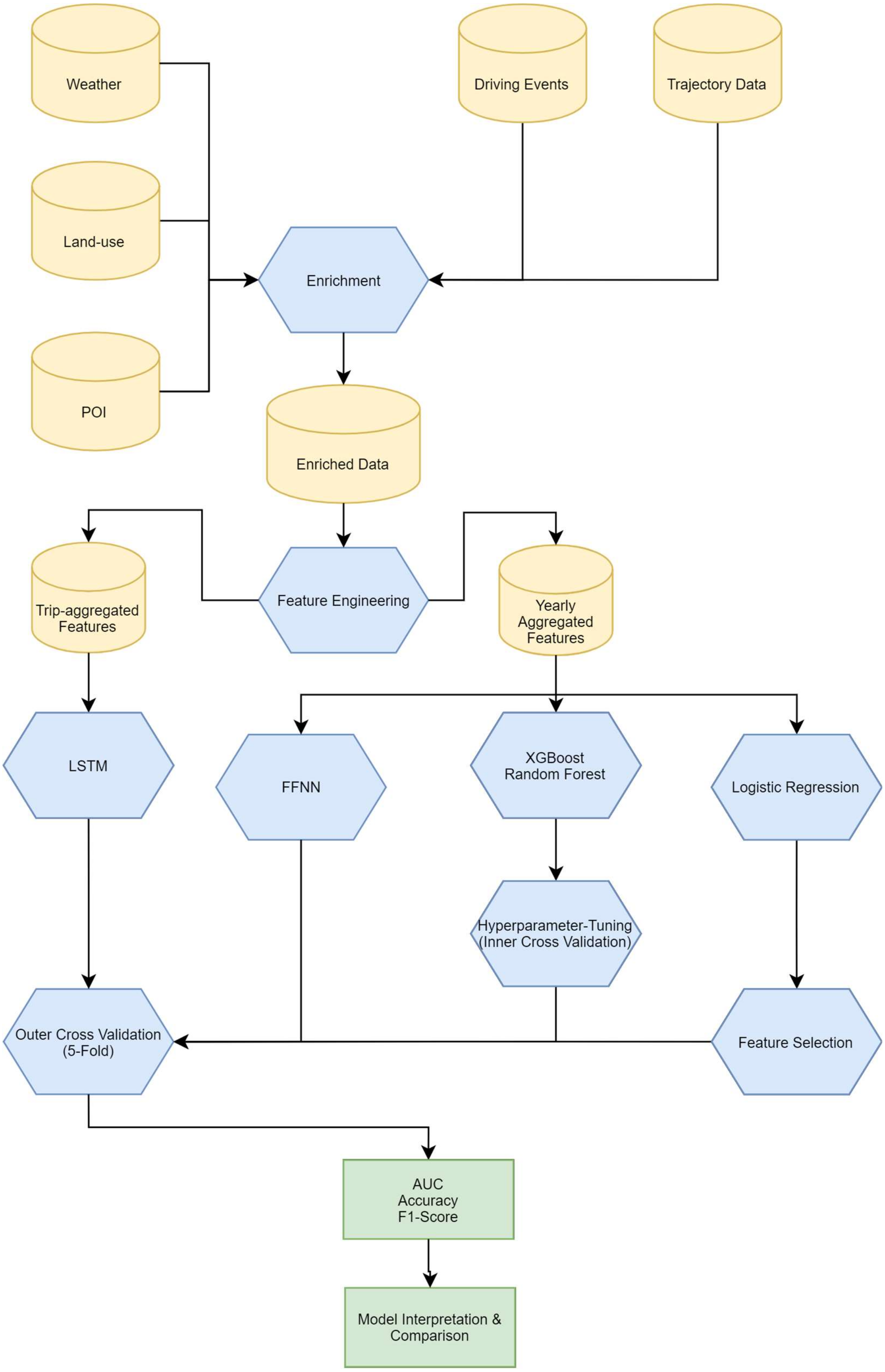

Figure 10: Simplified workflow, yellow signifies data, blue signifies computations, green signifies results

Figure 10 shows the (simplified) general workflow of this thesis. In the first step, data were cleaned and sampled. Then the raw events and positional data were enriched with geographical context data, e.g., weather, POI, and land-use data. In the following step, these data were aggregated and meaningful predictor variables (features) were computed. This was done on two levels: First on a yearly, aggregated level, and second on a per-trip basis. All models were evaluated through 5-fold cross-validation. If applicable hyperparameters were tuned using another 5-fold cross-validation on the training folds in each iteration, in order to prevent information leak. For logistic regression, a stepwise feature selection was performed. Several features and their combinations were tested, as well as different hyperparameter settings for the models.

The following sections will describe the tools used to perform the analysis, the methods for data enrichment, choice of feature sets, model building, performance assessment, and give an overview of the experimental design, which will be conducted. Furthermore, some summary statistics about the data at hand are included in this chapter.

4.1 Tools and Computation Setup

Python 3 has been used for most of the preprocessing and enrichment of the data as well as the implementation of the machine learning algorithms. Several libraries were used, as seen in Table 7. QGIS was used for parts of exploratory analysis, spatial visualizations, and preprocessing. All the calculations were computed on a machine with 4 cores at 3.6G Hz and 16 GB of RAM. The LSTM was sped up by the usage of a dedicated GPU.

Table 7: Most important python libraries.

Library	**Usage**
Pymongo	Connect to and extract data from MongoDB
Pandas	Data handling
Geopandas	Spatial data operations
Matplotlib	Plotting
Seaborn	Plotting
Numpy	Mathematical and matrix operations
Json	Handling of JSON files
Sklearn	Machine learning pipeline
Keras	Neural Network
Tensorflow	Neural Network

4.2 Data Preparation

Drivers with very low driving mileage were filtered out. Before the filtering, there was a large amount of mostly accident-free drivers that only drove a very small distance during the whole year. There is no information available if this is due to their recorder not working properly or if they just rarely use their car in general. If these low mileage drivers are included in the sample, the predictive performances of the final models increase significantly, since these drivers are very easy to classify as accident-free due to their low mileage. In order to get a better estimate about a drivers' driving behaviour and more reliable models, only drivers with more than 1500km driven for the Italy data set and more than 1000km for the UK data set during the given year were considered for the main results. The different cutoff values are due to the UK drivers driving shorter distances on average and the already small sample size. Furthermore, since the policy of some drivers started in the middle or ended before the end of 2017, only drivers with an observation period greater than one month were considered. The total exposure was normalized to 365 days in order to get an estimate of the yearly mileage. However, most of the drivers had data available over the whole year, as the median observation period was 364 days for Italy and 221 days for the UK. Furthermore, to account for outliers and the non-linear relationship between mileage and accident risk, the natural logarithm was taken of these weighted mileage measurements. However, a robustness test was also conducted that the models were run without this filtering strategy to see if the relationship between model and feature set performance changes. After filtering the low mileage drivers and transforming the mileage to the natural logarithm, the total weighted mileage roughly follows a log-normal distribution as depicted in Figure 11, which means that the natural logarithm of the data is normally distributed. This is consistent with previous findings (Paefgen et al., 2014). After filtering, there is still a surplus of accident-free drivers at the lower tail of the distribution, which is to be expected.

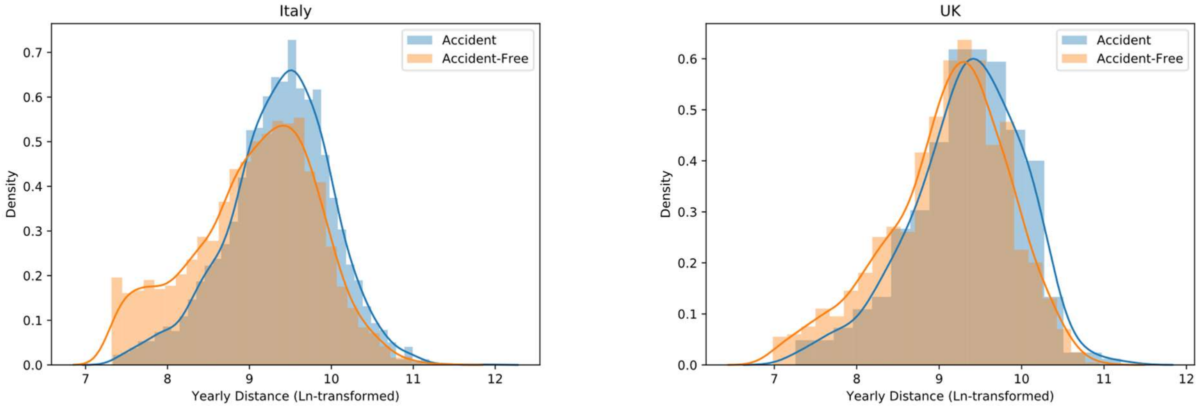

Figure 11: Distributions of yearly distance driven after outlier removal for both study areas (Ln-transformed).

4.3 Definition of the Classification Problem

There was a choice between two prediction strategies in regards to the classification problem: The first strategy was to classify the driver into *accident* and *accident-free* categories according to their whole year of driving behaviour. The other strategy was to take an early part of a year, e.g., January to June, for training to predict the possibility of having an accident in the future, e.g., July to December. Both approaches have been used by previous studies: Huang & Meng (2019) took the former strategy while Baecke & Bocca (2017) took the later strategy.

The first strategy has the disadvantage of taking data after an accident into account: After an accident, a driver's driving behaviour may change drastically, or a driver can be stopped from driving entirely by a severe crash (Mayou et al., 1993). The second strategy, however, does not take seasonal variability into account, which is not optimal if we want to use weather-related variables. Furthermore, the sample of drivers who had an accident would be further narrowed to only those that had an accident in the second half of the year. Concerning the UK sample is already quite small and this study focuses on the impact of environmental factors such as weather, the former prediction strategy was chosen. Ideally, more than one year of data would be available, so a prediction for the second year could be made based on the observations of the first year. To summarize, the classification problem can be described as follows:

Given a set of features derived from the driving behaviour data over the whole year, a binary classification is performed to separate *accident* from *accident-free* drivers. And in mathematical terms:

$$h : X \rightarrow y \in \{0, 1\}$$

Where:

h: Classification function.
X: Input vector consisting of n driving behaviour features $\{x_1 \ldots x_n\}$.
y: Binary label: 0 = accident-free, 1 = accident.

4.4 Data Enrichment

This section describes the technicalities of the enrichment of the trajectory and driving events data with geographical context data. Furthermore, the procedure to extract individual trips from the raw data will be outlined.

4.4.1 Weather Enrichment

In order to enrich the trajectory data with weather data, the closest virtual weather station (locations from where weather data was available, see Figures 5 and 6) was computed for each waypoint using a KDTree (Bentley, 1975). The python library cKDtree[5] was used for this task, which facilitates a very efficient nearest neighbour search, as this library is implemented in C. Once the nearest station for each waypoint was determined, the trajectory data was grouped by the closest weather station and then merged with the temporally closest hourly weather data using the timestamp. The same approach was used in order to enrich the driving events with the weather data.

4.4.2. POI Enrichment

First, it needs to be noted that due to the low spatial resolution of the trajectory data, only the driving events were enriched with POI and land-use data. This is due to the fact that it would be difficult to make meaningful aggregates about the whole trajectory when only one waypoint every 2000 meters is available and the exact route taken between those two points is unknown. The driving events however are not limited by this resolution, since they are recorded whenever they happen. In this way, it is still possible to include the POI and land-use information and provide a more detailed context about the driving events and get insight into the conditions under which driving events contribute to accident risk. Further, it can be argued that if no braking, cornering or acceleration events happen, the driving is stable and provides a low accident risk.

To enrich the whole trajectory, map matching would be necessary, which was not feasible in the scope of this study due to the aforementioned low temporal resolution of especially the Italy dataset. Map matching with a higher resolution dataset would however certainly be an option for follow-up studies.

For the enrichment of the driving events data with POI, two types of POI data had to be considered; point and polygon data. Due to their different structural natures, they require different approaches. For the point data, a KDTree was used again, and the number of POI of each category in a buffer of 200 meters around each driving event was counted. This buffer of 200 meters was selected according to

[5] https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.cKDTree.html

previous literature including POI data in transportation research who used values between 100 and 300 meters, such as Krause and Zhang (2019). However, this buffer can be changed according to the problem at hand. 100 meters and 300 meters buffers were tried as well but led to worse predictive performance in the final models.

Similarly, the polygon POI were buffered by 200 meters and a point in polygon join was performed and the number of POI of each category around each driving event counted using a Rtree. Rtree is a spatial indexing method that allows for very fast spatial searches, such as the aforementioned point-in-polygon operations (Guttman, 1984). Rtree is built-in in the spatial join method of the python library geopandas. A few POI with the same ID were present in both the point and the polygon POI data. In that case, the point data was used.

4.4.3 Land-Use Enrichment

In order to enrich the events with their corresponding land-use, in the first step, the land-use polygons were slightly buffered, using a buffer of 50 meters, which was done in order to deal with GPS uncertainty and edge cases. Then a point in polygon intersection was performed using a Rtree, in the same fashion as the polygon POI join. Due to the buffer points could sometimes intersect with different types of land-use polygons. This is intended and allows for a classification of mixed land-use areas.

4.4.4 Trip Extraction

In its raw form, the data does not contain information about individual trips. However, having access to certain measurements about trips such as distance and time allows for the extraction of more driving behaviour features, such as mean trip length, duration etc. Furthermore, for the LSTM, a sequence of trips is needed to train the model.

In order to assign a trip ID to each waypoint, the data first had to be grouped by each driver and sorted by the timestamp. Then for each waypoint, the time difference to the previous waypoint was calculated. If this time difference exceeds 15 minutes, the waypoint is assigned to a new trip. Previous studies mainly use ignition on/off events to determine trips, which were also available in the data at hand, however, this method tends to underestimate trip distance and duration, since drivers can turn off the ignition, e.g., at a traffic light or for a short toilet break. Therefore setting a waiting time of 15 minutes was deemed a more realistic approach. After a unique trip id was assigned to each waypoint, in a further step the data was grouped by the trip ID and for each trip, the following measures were derived:

- Trip Length (km)
- Trip Duration (minutes)
- Trip Start Time
- Trip End Time
- Average Speed
- Number of driving events per category
- Number of driving events per POI category
- Number of driving events per land-use category
- Binary dummy variables for each predominant weather condition (Sunny, rain, snow, overcast). Predominant weather condition refers to the weather condition in which the majority of kilometres in the trip were driven.
- Average Precipitation
- Average Temperature

- Binary variable for night driving (if either start or end time at night (23:00 – 7:00))
- Binary variable for rush hour driving (if either start or end time in rush hour (7:00 – 9:00 and 16:00 – 19:00) and weekday not Saturday or Sunday)
- Binary variable for weekend driving (if weekday is Saturday or Sunday)

In order to count the driving events per trip, both the events data and the trip data was grouped by the driver id. In the next step, each event was joined with the corresponding trip, which had a start time earlier than the event timestamp and an end time later than the event timestamp.

4.5 Computation of Features

Two types of feature sets had to be computed: One aggregates all the data in an N * M matrix, where N is the number of drivers and M the number of features. This type was used for the logistic regression, random forest, XGBoost and the FFNN. For the LSTM, the measures had to be aggregated on a trip level that is different from the first type. Since random forest, XGBoost and logistic regression can strongly benefit from feature engineering, several feature combinations were explored. The following section describes how the features, which were included in the results were derived. It is to be noted that a lot of feature engineering was done and several aggregations and feature combinations were tried and not all of them made it into the final results. In general, if a more complicated approach did not yield better predictive performance, for interpretability reasons the more simplistic one was chosen.

4.5.1 Exposure Features

As a general exposure variable, the log of the total mileage normalized by the observation period in 365 days was chosen. Furthermore, the total distance was divided into different timeslots and speed intervals, described in the following sections.

4.5.2 Speed Features

Several features were derived from the speed. Note that the average speed between two waypoints was used by dividing the distance travelled by the time delta, and not the instantaneous speed. The total distance driven was divided into 5 distinct speed intervals:

- 0-30km/h
- 30-60km/h
- 60-90km/h
- 90-130km/h
- > 130km/h

This was done similarly to Paefgen et al. (2013, 2014) and Huang & Meng (2019). The first two categories should capture city driving, 60-90km/h should capture extra-urban and rural driving and the last two categories should capture driving at motorway speeds. For each driver, the fraction of kilometres accumulated per category was calculated. The following boxplots show the distributions of driving in different speed intervals for both accident and accident-free drivers. According to Figure 12, accident drivers seem to drive a larger fraction of their total distance at a slower speed, whereas accident-free drivers drive more at highway speeds. This is in line with the expectations according to chapter two, where frequent city driving at lower speeds is associated with higher accident risk. The same is true for the UK dataset, as shown in the summary statistics later in section 4.5.8 and Figure 29 in the Appendix.

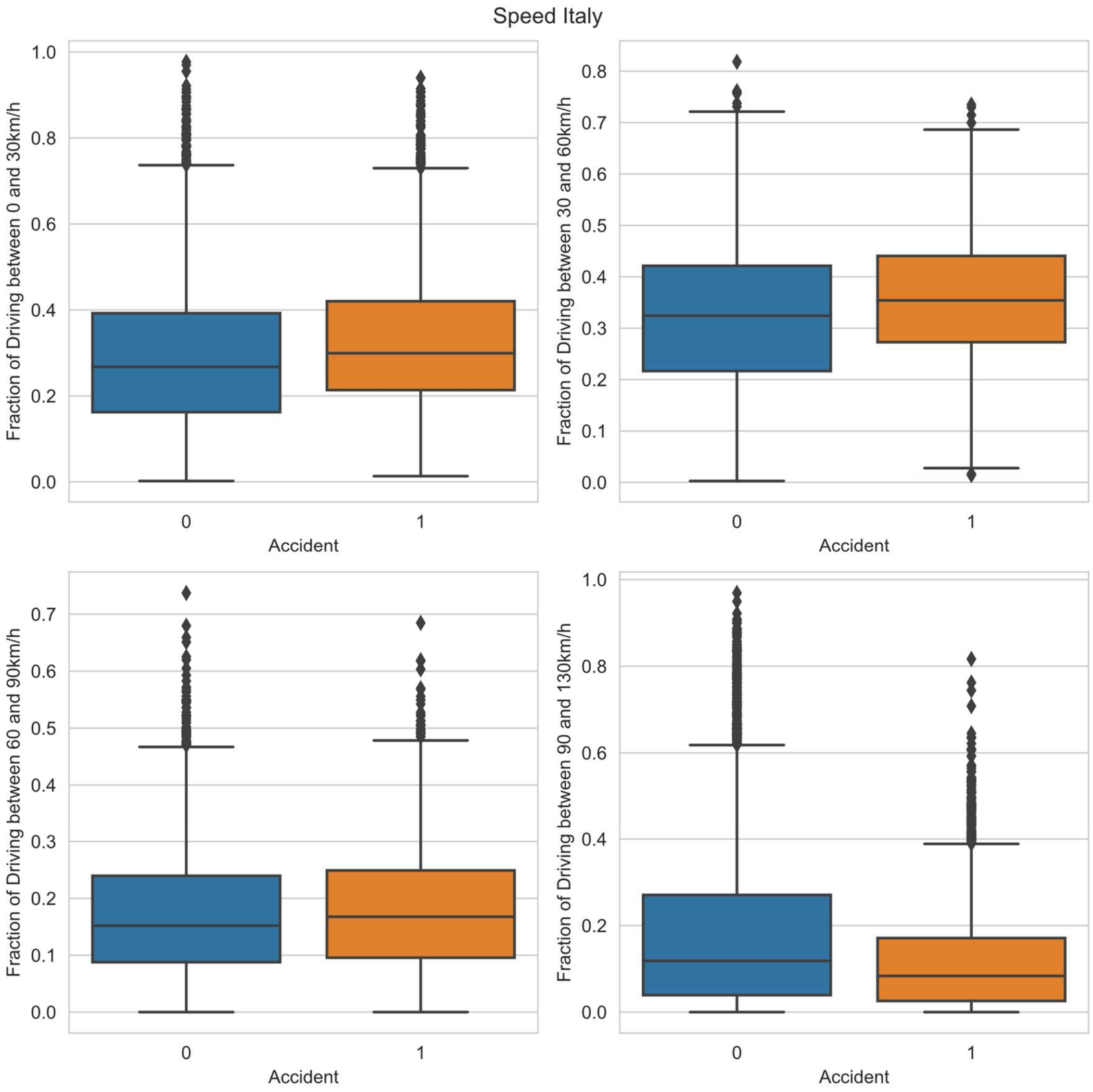

Figure 12: Boxplots of driving in different speed intervals in Italy.

4.5.3 Time-related Features

According to previous studies, the distance driven was divided into different time slots, which are based on the time of day and the day of the week. Especially of interest were driving at night, on the weekends and during peak hours, according to Chapter 2. Therefore the following timeslots were considered:

- Weekend driving (Saturday, Sunday)
- Night driving (23:00-7:00)
- Rush-hour driving during working day peak hours (Monday-Friday, 7:00 – 9:00 and 16:00 – 19:00)

For each of these timeslots, the fraction of the total distance accumulated in each slot was computed. From the boxplots in Figure 13, accident-free drivers seem to have a slightly higher fraction of weekend driving, whereas accident drivers have a higher fraction of night and rush-hour driving.

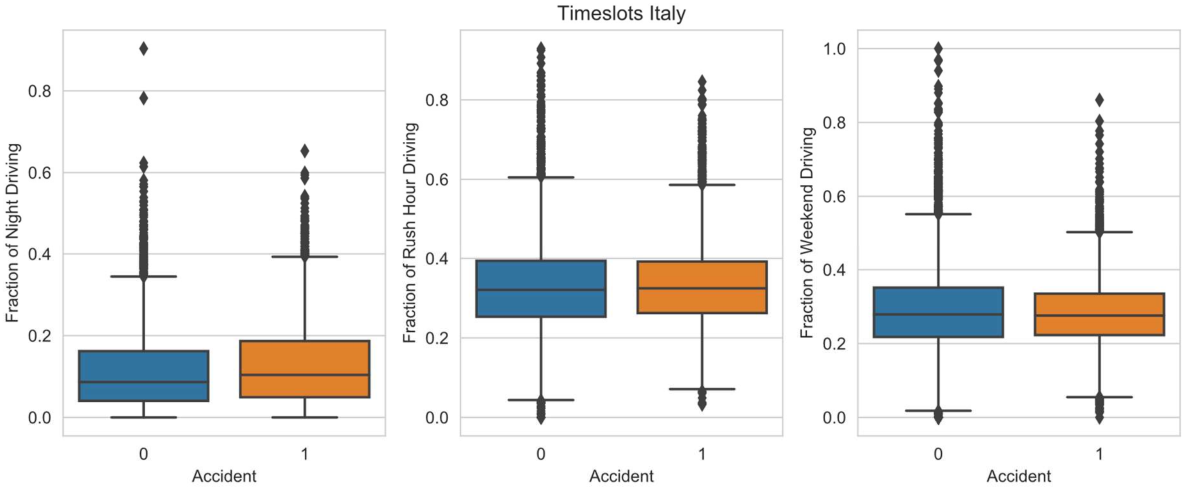

Figure 13: Boxplots of driving in different timeslots in Italy.

4.5.4 Weather-related Features

Weather-related features were computed based on the trajectory data. Originally, some weather-related features were also computed based on driving events, however, they did not add any performance improvement and were discarded in favour of simplicity.

In the first step, the total distance driven was aggregated by the weather condition given in the weather data. Since the weather conditions included over 25 different categories, many of which very rare, they were further condensed into five different categories: Good weather, overcast weather, rain, snow and fog. Furthermore, since according to Chapter 2 temperature can influence accident risk, the total distance driven was aggregated into different temperature levels: High (>25°C), moderate (0-25°C) and freezing (< 0°C). The following list summarizes the 9 features derived from the weather data:

- Fraction of km driven in good weather
- Fraction of km driven in overcast weather
- Fraction of km driven in rain
- Fraction of km driven in snow
- Fraction of km driven in fog
- Fraction of km driven above 25°C
- Fraction of km driven at night in the rain
- Fraction of km driven between 0 and 25°C
- Fraction of km driven below 0°C

According to Figure 14, the difference in driving behaviour according to the weather condition between accident and accident-free drivers is very small in the Italy dataset. The difference in the UK dataset is also small, which can be derived from the summary statistics in Section 4.5 and Figure 28 in the appendix.

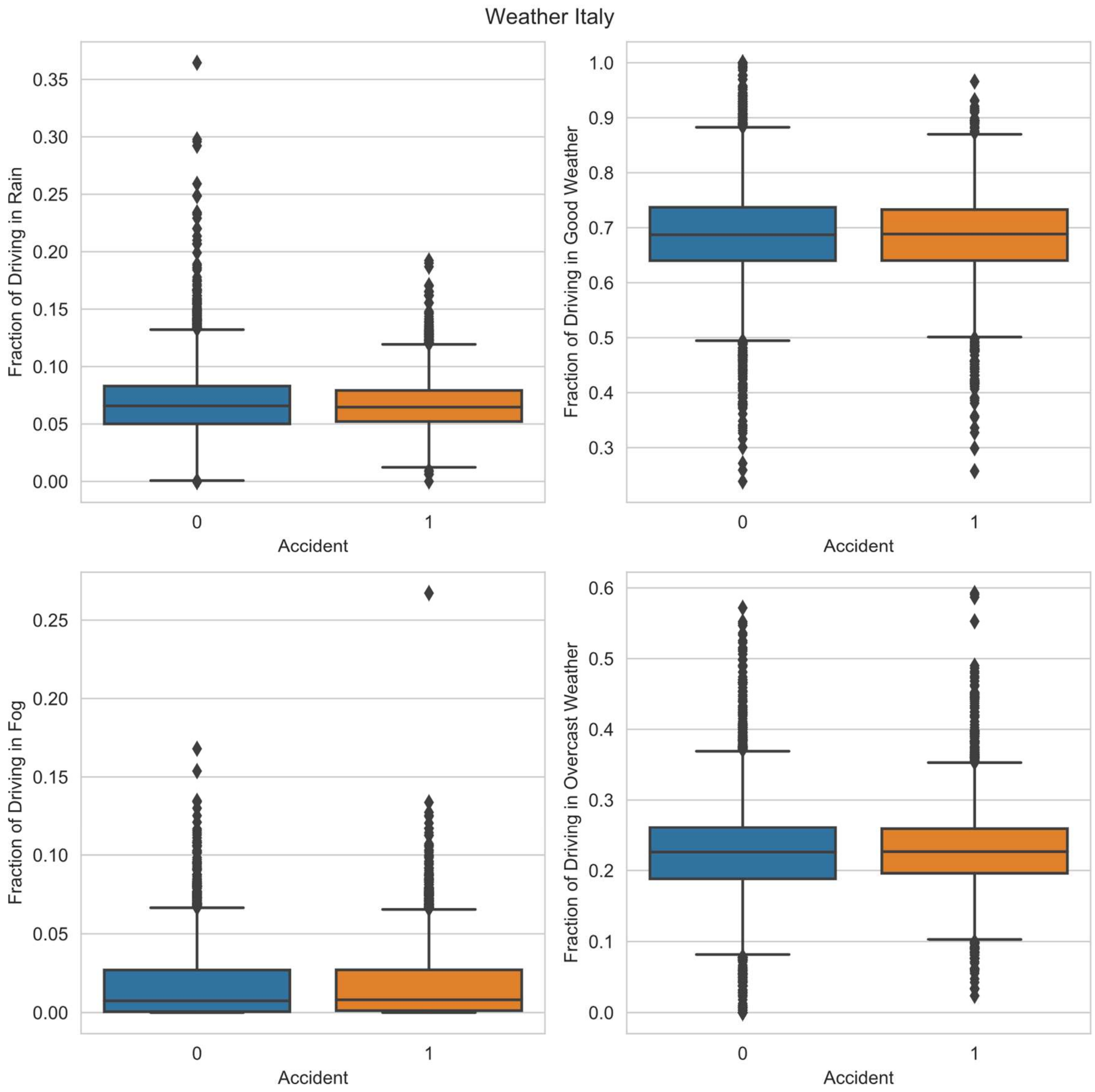

Figure 14: Boxplots of driving in different weather conditions in Italy.

4.5.6 Land-Use related Features:

The driving events were classified according to their type and their corresponding land-use category. This results in a large number of possible combinations, and for interpretability reasons, only the most common combinations were considered. Furthermore, events with more than 2 different types of land-use were classified as mixed. Note that the number and type of the combinations differ between the two data sets, due to different ratios of each land-use class in the two study areas, e.g., Italy had very few commercial land-use parcels. Part of it might be due to different taxonomy. In the end, the following example features were derived:

- Events in forest per 1000km
- Events in industrial land-use areas per 1000km
- Events in commercial land-use areas per 1000km
- Events in residential land-use areas per 1000km
- Events in rural land-use areas per 1000km

- Mixtures of above, e.g., events in rural-residential land-use areas per 1000km

This was done for braking, acceleration and cornering events. In total, 9 land-use related features were derived from the Italy set and 11 from the Uk set, as seen in Tables 8 and 9.

4.5.7 POI related Features

Different ways of aggregating the POI data were tried. Finally, the following 5 features were derived:

- Fraction of driving events near commercial POI
- Fraction of driving events near public POI
- Fraction of driving events near touristic POI
- Fraction of driving events near nightlife POI
- Fraction of driving events near transportation POI

More complicated aggregates such as several combinations of POI and event types were tried as well, as well as the average number of POI around each event. However, these aggregation methods did not yield better results. Therefore the more simplistic approach listed above was chosen.

4.5.8 Summary of Features

To provide a short overview of the features, Tables 8 and 9 show the sample medians of accident vs. accident-free drivers for both study areas for the yearly aggregated data. Note that for interpretation purposes the total weighted yearly distance is shown in the tables, while the natural log was used for the modelling.

Table 8: Summary Statistics Italy (after filtering low mileage drivers).

Feature	Sample Median	
	Accident (n = 3892)	Accident-Free (n = 4178)
Total yearly distance (km)	12318	9384
Fraction of night Driving	0.104	0.0865
Median of trip distance (km)	8.605	9.044
Standard deviation of trip distance (km)	13.971	15.479
Median of trip duration (minutes)	25.075	25.117
Mean of mean trip speed (km/h)	26.248	28.348
Fraction of rush hour driving	0.325	0.320
Fraction of weekend driving	0.280	0.288
Fraction of driving between 0 and 30 km/h	0.299	0.268
Fraction of driving between 30 and 60 km/h	0.354	0.324
Fraction of driving between 60 and 90 km/h	0.168	0.152
Fraction of driving between 90 and 130 km/h	0.084	0.118
Fraction of driving above 130km/h	0.001	0.003
Fraction of driving in rain	0.065	0.066
Fraction of driving in good weather	0.689	0.687
Fraction of driving in overcast weather	0.227	0.226
Fraction of driving in snow	0.0	0.0
Fraction of driving in fog	0.008	0.007
Fraction of driving above 25°C	0.122	0.123
Fraction of driving between 0 and 25°C	0.870	0.868
Fraction of driving below 0°C	0.003	0.002
Fraction of driving at night during rain	0.0006	0
Acceleration events per 1000km	6.702	3.945
Braking events per 1000km	72.547	50.661
Cornering events per 1000km	294.350	219.586
Quick lateral movement events per 1000km	3.081	1.833
Fraction of driving events near commercial POI	0.081	0.070
Fraction of driving events near nightlife POI	0.106	0.095
Fraction of driving events near public POI	0.195	0.171
Fraction of driving events near touristic POI	0.016	0.015
Fraction of driving events near transportation POI	0.111	0.103
Braking events in forest land-use per 1000km	0.473	0.650
Cornering events in forest land-use per 1000km	6.335	4.403
Cornering events in mixed forest-residential land-use areas per 1000km	0.721	0.430
Accelerations in residential land-use areas per 1000km	1.667	0.815
Braking events in residential land-use areas per 1000km	19.758	12.224
Cornering events in residential land-use areas per 1000km	56.881	35.088
Cornering events in mixed rural-residential land-use areas per 1000km	25.923	15.770
Cornering events in rural land-use areas per 1000km	26.865	19.163
Braking events in rural land-use areas per 1000km	1.451	1.156

Table 9: Summary statistics UK (after filtering low-mileage drivers).

Feature	**Sample Median**	
	Accident (n = 355)	**Accident-Free (n = 1157)**
Yearly Distance (km)	11738	9812
Fraction of night Driving	0.120	0.112
Median of trip distance (km)	10.123	10.894
Standard deviation of trip distance (km)	15.790	17.456
Median of trip duration (minutes)	32.483	30.642
Mean of mean trip speed (km/h)	23.263	25.787
Fraction of rush hour driving	0.311	0.297
Fraction of weekend driving	0.283	0.298
Fraction of driving between 0 and 30 km/h	0.225	0.200
Fraction of driving between 30 and 60 km/h	0.392	0.348
Fraction of driving between 60 and 90 km/h	0.162	0.172
Fraction of driving between 90 and 130 km/h	0.113	0.173
Fraction of driving above 130km/h	0	0
Fraction of driving in rain	0.074	0.070
Fraction of driving in good weather	0.265	0.256
Fraction of driving in overcast weather	0.610	0.622
Fraction of driving in snow	0.003	0.002
Fraction of driving in fog	0.034	0.036
Fraction of driving above 25°C	0.004	0.0007
Fraction of driving between 0 and 25°C	0.991	0.993
Fraction of driving below 0°C	0.001	0.0009
Acceleration events per 1000km	18.705	10.680
Braking events per 1000km	86.646	47.895
Cornering events per 1000km	417.295	259.649
Quick lateral movement events per 1000km	4.508	2.089
Fraction of driving events near commercial POI	0.085	0.090
Fraction of driving events near nightlife POI	0.096	0.094
Fraction of driving events near public POI	0.158	0.146
Fraction of driving events near touristic POI	0.0183	0.0186
Fraction of driving events near transportation POI	0.252	0.234
Cornering events in residential land-use areas per 1000km	115.592	62.170
Cornering events in mixed rural-residential land-use areas per 1000km	42.664	21.800
Braking events in residential land-use areas per 1000km	30.219	14.663
Cornering events in mixed commercial-residential land-use areas per 1000km	25.550	13.790
Cornering events in commercial land-use areas per 1000km	12.256	7.700
Cornering events in rural-forest land-use areas per 1000km	8.082	5.207
Cornering events in forest land-use areas per 1000km	7.824	5.398
Acceleration events in residential land-use areas per 1000km	4.712	2.436
Braking events in mixed rural-residential land-use areas per 1000km	7.846	3.952
Cornering events in industrial land-use areas per 1000km	2.493	2.0177
Driving events in mixed land-use per 1000km	116.724	79.266

4.6 Model Building

The data was rescaled before training the models. Since the data at hand contains quite a few outliers due to large differences in individual driving behaviour and due to the possibility of irregularities in the data recording due to malfunctioning recorders, a robust scaling approach was chosen, where the data was scaled between the 1st and 3rd quartile. This was implemented using the RobustScaler method from the Python library Scikit-Learn.[6] For the LSTM the MinMaxScaler method from the same library has been used to scale the data between 0 and 1. Since all the yearly aggregated variables were continuous, no other preprocessing steps had to be performed. Logistic regression, random forest and XGBoost were implemented using their respective implementations in Scikit-Learn. The neural networks were implemented using the Keras[7] library, which serves as a higher-level library for TensorFlow[8].

4.6.1 Logistic Regression

Since logistic regression is unable to handle co-correlated features, in the first step, the optimal number of features had to be computed. This was done using a step-wise cross-validation approach, where features were step-wise omitted and a 10-fold cross-validation performed to determine the accuracy after each step. The optimal number of features was chosen based on the combination, which resulted in the highest accuracy value. Note that there are some drawbacks to using this method, as depending on the order of the feature omission, the optimal combination might not be found. However, step-wise feature selection was used in other studies related to the topic at hand, such as (Paefgen et al., 2014; Baecke & Bocca, 2017; Huang & Meng, 2019) and was therefore deemed a suitable approach. This feature selection approach was independently performed on each of the 6 feature sets.

4.6.2 Random Forest

For all the different feature sets, a random forest was computed. In order to find the optimal hyperparameters, the Scikit-Learn method gridsearchCV was used, which also allows for a parallel computation of the grid search procedure. In this way, different combinations of hyperparameters were tested in a 5-fold cross-validation. Accuracy was chosen as the scoring parameter. The following Table 10 gives an overview of the hyperparameter grid that was searched. All other hyperparameters were set to their default value in the Scikit-Learn implementation of random forest. Since all features are continuous, Gini coefficient was chosen as a feature importance measurement.

Table 10: Random Forest Hyperparameters used during grid search.

Hyperparameter	**Searched values**
Number of estimators	100, 300, 400
Min samples leaf	1, 4, 5
Min sample split	2, 6, 10
Max depth	10, 30, 100, None

[6] https://scikit-learn.org/
[7] https://keras.io/
[8] https://www.tensorflow.org/

4.6.3 XGBoost

In a similar fashion, gridsearchCV was used to determine the optimal parameters for the XGboost algorithm. The following parameters in Table 11 were searched. All other parameters were set to their default value in the XGBoost implementation for Python. Information gain was chosen as a feature importance measurement.

Table 11: XGBoost Hyperparameters used during grid-search.

Hyperparameter	**Searched values**
Number of estimators	100, 300, 400
Learning rate	0.001, 0.01, 0.02
Colsample bytree	0.7, 0.8
Max depth	15, 35, None
Subsample	0.7, 0.9

4.6.4 Neural Networks

Two separate neural networks were employed: First, a simple FFNN, which uses the same aggregated data as the previous models was built. This FFNN consisted of 2 dense layers with rectified linear unit activation functions (RELU) and 256 neurons each. Furthermore, a dropout layer of size 0.4 was added after each hidden layer to prevent overfitting. The activation layer included a softmax activation function to return a probability between 0 and 1. The batch size was set to 32. Binary cross-entropy was chosen as the loss function. The model was trained for 50 epochs in Italy and 25 Epochs in the UK. Figure 15 provides a scheme of the FFNN which was used.

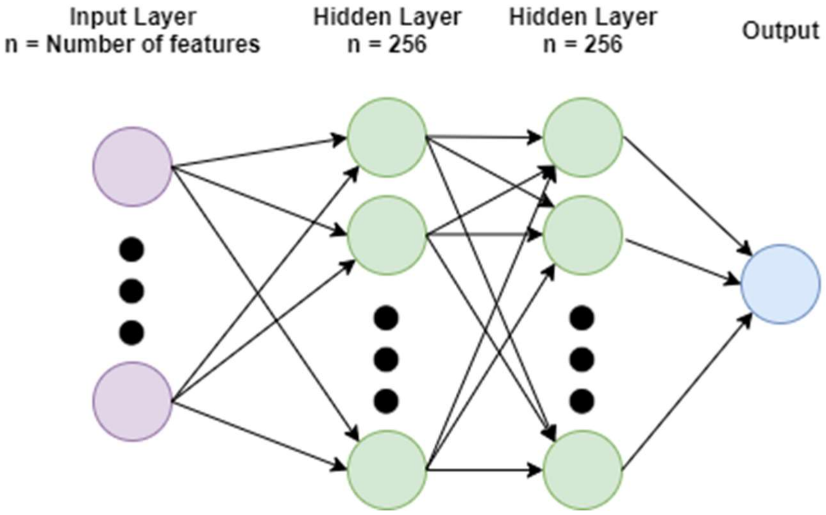

Figure 15: FFNN Architecture

Furthermore, in order to try a different aggregation approach, a recurrent neural network (RNN) was employed, which consisted of two LSTM layers, which had 64 LSTM cells each. In order to build the LSTM network, the trip data had to be converted into a nested array of the shape number of drivers * number of trips * features per trip. The features per trip are shown in section 4.4.4. For the LSTM, only trips longer than 3 km have been considered. Since the LSTM cell only accepts sequences of the same lengths, all the trip sequences of each driver had to be padded to the maximum amount of trips per driver in each dataset. This was done by filling the array with zeros, adding empty trips. Subsequently, a masking layer was added as the first layer of the neural network, which tells the LSTM cell to ignore these empty trips. Furthermore, since LSTM models tend to overfit three dropout layers of size 0.4 were added to combat this problem. An activation layer with a softmax function was used again, to return a probability between 0 and 1. Binary cross-entropy was chosen as the loss function. The model was trained for 50 epochs and the batch size set to 100 in Italy. For the UK, the batch size was set to 45 and the number of epochs to 50.

Since a systematic grid search cross-validation procedure would be too expensive to be used on a deep learning model considering the computational resources available, a few different values for neuron numbers, batch size, number of epochs and size of dropout layer were tried out manually for both neural networks.

4.7 Feature Combinations

Table 12: Overview of the different feature sets

Feature Combination	Exposure	Events	Weather	POI	Land Use
A	X				
B (Baseline)	X	X			
C	X	X	X		
D	X	X		X	
E	X	X			X
F	X	X	X	X	X

In order to assess the impact of different variable groups, six different feature combinations were considered for the yearly aggregated data, as seen in Table 12. Feature set B serves as a baseline, including all the information except the geographical information. C, D and E combine the baseline with weather, POI and land-use respectively, finally, feature combination F includes all the information available. This was only done for the yearly aggregated features and not for the trip-based features which were used in the LSTM model. The LSTM was assessed under two scenarios: With geographical information and without, which corresponds to Feature Set B and F in the yearly aggregated scenario.

4.8 Performance metrics and model evaluation

True Values

		Positive (Accident)	Negative (Accident-Free)
Predicted Values	Positive	TP	FP
	Negative	FN	TN

Figure 16: Example of a confusion matrix, green entries represent correct classifications.

Given the confusion matrix in Figure 16, the following measures can be derived in order to assess and compare model performances:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

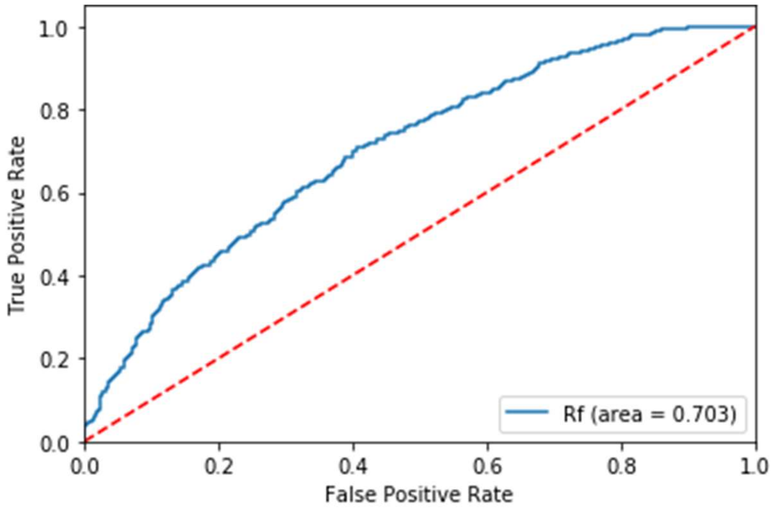

Figure 17: Example ROC curve, the red dotted line represents a random classifier with AUC 0.5, the blue line a Random Forest classifier with AUC 0.703.

In addition, the area under the receiver operating characteristic (ROC) curve (AUC) as depicted in Figure 17 was used. The ROC curve shows the ratio of true positive rate versus false-positive rate under varying thresholds. It is a common robust metric used for classification problems. It can be interpreted as the probability that a random driver who had a crash during the observation period will have a higher crash probability compared to a random driver who did not have a crash during the observation period (Cheng et al., 2018). AUC is insensitive to class imbalance, which makes it suitable for imbalanced classification problems.

From the measures presented in this section, AUC, accuracy and F1-Score will be used to assess the results of this study. Since F1-Score includes the mean of precision and recall, precision and recall were not used for simplicity reasons. Because the final data is slightly imbalanced a stronger focus will be on AUC for the interpretation of the results.

All the models were assessed using 5-fold cross-validation. In order to prevent information leak from the validation set, hyperparameter-tuning and performance assessment for XGBoost and the random forest was implemented using a nested-cross validation approach. Information leak can cause biased and over-optimistic performance measurements if the hyperparameters of a given model are optimized using a dataset, which includes the validation data (Cawley & Talbot, 2010). In other words, some information about the validation data gets leaked indirectly through the process of hyperparameter-tuning. The following nested cross-validation approach was employed:

1. The data is split into 5 folds. This is the outer cross-validation
2. Each of the 5 folds is held out as a validation set once.
3. For each combination, a grid search using 5-fold cross-validation is performed on the 4 training folds. This is the inner cross-validation. The models then fit on the training folds using these optimal hyperparameters and validated on the validation fold.
4. Accuracy, AUC and F1-Score of all 5 folds are stored, as well as feature importance measurements

For logistic regression and neural networks, a simple 5-fold cross-validation was implemented as no hyperparameter tuning was necessary. Furthermore, since the UK sample was rather small, which resulted in large between-folds differences, in order to get more reliable results, the whole procedure was repeated 2 times.

It is acknowledged that the feature selection of logistic regression might also induce some information leak and cause the logistic regression results to be slightly over-optimistic, however since we are interested in the coefficients for interpretation reasons, performing a separate feature selection in each fold would make it impossible to take the average coefficients, since the selected features would be different each time.

As previously mentioned, all the yearly aggregated models were implemented across all feature sets, whereas the LSTM was computed for two scenarios: With geographical information and without. Furthermore, the UK dataset was largely imbalanced even after filtering out low mileage drivers, therefore in addition to the 355 drivers who had an accident, 380 accident-free drivers were randomly sampled. The Italy dataset was relatively balanced after filtering low-mileage drivers (3892 vs. 4178), and further balancing was not deemed necessary.

Chapter 5: Results

This Chapter presents the results to answer the research questions given at the end of chapter one. Mainly the impact of geographical features will be explored and the performance of the different machine-learning algorithms compared. Furthermore, the impact of the chosen minimum driving distance will be briefly illustrated. Note that a stronger focus will be on the Italy dataset results since the much bigger dataset returned more consistent and reliable results. Some differences between the UK and Italy datasets will be explored throughout the course of this chapter.

5.1 Comparison of Model Performance

Tables 13 and 14 provide a general overview of the model performances across all feature sets in Italy and the UK, respectively. As previously mentioned, since the LSTM uses a different aggregation method and due to computational reasons, it was only computed for two feature sets, B and F.

Table 13: Model and Feature Set comparison Italy, the best overall performance is indicated in bold, the best performance per feature set is underlined

Model	A	B (baseline)	C	D	E	F
			AUC			
Logistic Regression	0.642	0.647	0.654	0.652	0.665	0.666
FFNN	0.636	0.649	0.645	0.638	0.697	0.684
Random Forest	0.635	0.650	0.657	0.654	0.699	0.697
XGBoost	0.634	0.652	0.654	0.653	0.710	**0.714**
LSTM	-	0.591	-	-	-	0.618
			Accuracy			
Logistic Regression	0.591	0.597	0.601	0.600	0.610	0.612
FFNN	0.588	0.602	0.601	0.595	0.636	0.628
Random Forest	0.589	0.600	0.605	0.604	0.643	0.632
XGBoost	0.586	0.600	0.606	0.606	0.648	**0.648**
LSTM	-	0.553	-	-	-	0.563
			F1-Score			
Logistic Regression	0.603	0.605	0.604	0.607	0.612	0.614
FFNN	0.590	0.603	0.596	0.600	0.611	0.610
Random Forest	0.602	0.601	0.604	0.607	0.639	0.627
XGBoost	0.618	0.615	0.611	0.616	**0.643**	**0.643**
LSTM	-	0.584	-	-	-	0.641

Table 14: Model and Feature Set comparison UK, the best overall performance is indicated in bold, the best performance per feature set is underlined

Model	A	B (baseline)	C	D	E	F
AUC						
Logistic Regression	0.665	0.689	0.694	0.692	0.680	0.690
FFNN	0.663	**0.699**	0.693	0.696	0.664	0.689
Random Forest	0.634	0.670	0.681	0.675	0.676	0.685
XGBoost	0.640	0.664	0.673	0.667	0.660	0.675
LSTM	-	0.637	-	-	-	0.663
Accuracy						
Logistic Regression	0.618	**0.643**	0.638	0.627	0.628	0.632
FFNN	0.599	0.642	0.626	0.631	0.609	0.638
Random Forest	0.602	0.608	0.627	0.608	0.628	0.634
XGBoost	0.596	0.604	0.614	0.617	0.606	0.614
LSTM	-	0.569	-	-	-	0.582
F1-Score						
Logistic Regression	0.619	0.624	0.626	0.608	0.601	0.613
FFNN	0.612	**0.659**	0.636	0.646	0.623	0.639
Random Forest	0.606	0.607	0.620	0.606	0.629	0.639
XGBoost	0.592	0.603	0.612	0.618	0.604	0.610
LSTM	-	0.561	-	-	-	0.545

Since the scores are rather close, in order to get an overview of the variability, Figure 18 provides a boxplot of the AUC values over all folds for the logistic regression, random forest, XGBoost and FFNN. For the Italy data, Feature Set F outperforms the baseline across all performance metrics and models. Land-use seems to return higher performance improvements than weather and POI, which only provide minor or no improvement. In Italy, XGBoost returns the best performance across all metrics for feature set F, followed by random forest and the FFNN. In the baseline feature set B, XGBoost also performs best in terms of AUC and F1-Score, while the FFNN performs best in terms of accuracy, though the scores are very close. This small performance gap between models is consistent with other studies that also report between-model performance differences within 1-2 %, such as (Baecke & Bocca 2017; Huang & Meng 2019; Paefgen et al., 2013). The inclusion of driving events (Feature set B vs. A) returns higher performance across almost all metrics and models, with a more substantial effect in the UK dataset. Note that the AUC value is generally higher than the other metrics, attributed to the slight class-imbalance present in the dataset.

Interestingly, in the UK case, logistic regression slightly outperforms other algorithms in some feature sets, yielding the overall highest accuracy value for Feature Set B. A possible explanation for this observation is that random forest, XGBoost and FFNN are overfitting due to the small sample size. From the boxplots, it is apparent that the Italy results are more stable with fewer fluctuations and outliers over all folds. Especially the FFNN is prone to large between-fold differences. In general, as seen in Table 13 and 14, the LSTM performs worse than the yearly aggregated models across all performance metrics except for F1-Score in the case of Feature Set F and the Italy data, which might indicate that it is not the most optimal data aggregation or modelling approach for the data at hand.

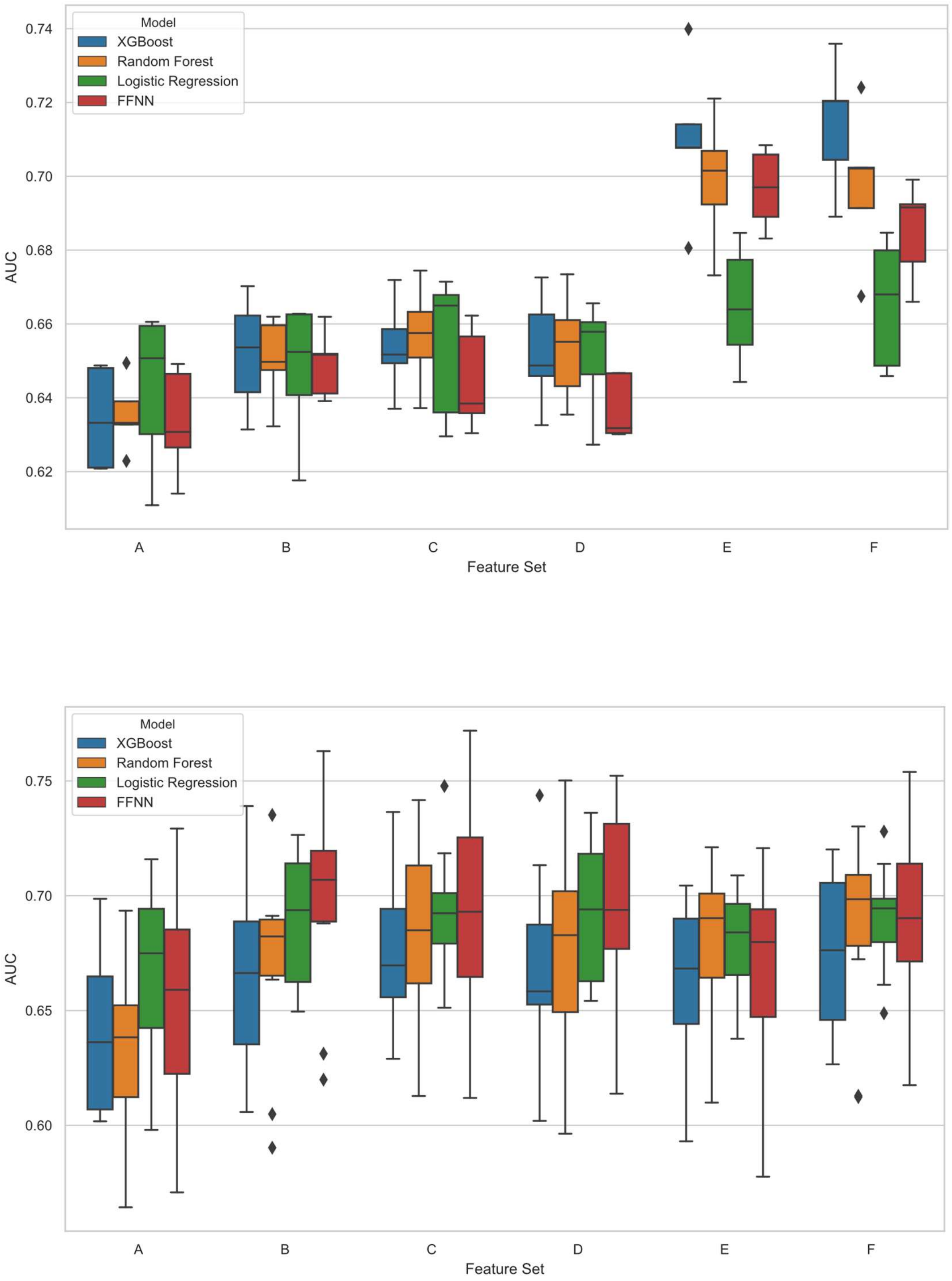

Figure 18: Boxplots of AUC values over all folds and feature sets, for Italy (top) and UK (bottom).

5.2 Model Interpretation

This section will provide an overview and interpretation of the logistic regression coefficients and the feature importances of random forest and XGBoost. Furthermore, findings will be compared to the literature to assess if the proposed models are robust and realistic.

5.2.1 Logistic Regression Coefficients

Table 15: Logistic Regression coefficients (Feature Set B, Italy)

Variable	**Exponentiated Coefficient**
Distance log	1.674
Fraction of driving between 0 and 30km/h	1.271
Cornering events per 1000km	1.184
Fraction of driving at night	1.163
Median trip duration	1.146
Fraction of driving between 60 and 90 km/h	1.144
Median trip distance	0.848
Fraction of driving between 90 and 130km/h	0.786

Table 16: Logistic Regression coefficients (Feature Set F, Italy) Gray shaded variables represent baseline-variables.

Variable	**Exponentiated Coefficient**
Distance log	1.660
Braking events per 1000km	1.264
Median trip duration	1.201
Cornering events per 1000km	1.143
Fraction of driving at night	1.121
Fraction of driving between 60 and 90km/h	1.079
Fraction of driving between 0 and 30km/h	1.070
Fraction of driving events near commercial POI	1.054
Quick lateral movement events per 1000km	1.048
Acceleration events per 1000km	1.033
Percentage of driving events near transportation poi	0.962
Fraction of driving above 25°C	0.955
Braking events in forest land-use areas per 1000km	0.952
Fraction of driving in fog	0.939
Cornering events in rural land-use areas per 1000km	0.930
Fraction of driving in rain	0.926
Mean trip speed	0.920
Standard deviation of trip distance	0.901
Braking events in residential land-use areas per 1000km	0.888
Median trip distance	0.868
Fraction of driving between 90 and 130km/h	0.835

Table 17 Logistic Regression coefficients of Feature Set B, UK

Variable	**Exponentiated Coefficient**
Median trip duration	1.735
Fraction of driving between 30 and 60km/h	1.36
Fraction of driving between 60 and 90km/h	1.284
Fraction of driving during rush-hour	1.176
Quick lateral movement events per 1000km	1.164
Braking events per 1000km	1.130
Acceleration events per 1000km	1.105
Distance log	1.07
Fraction of driving above 130km/h	1.06
Fraction of driving between 90 and 130km/h	0.929
Fraction of driving on the weekend	0.858
Standard deviation of trip distance	0.763
Mean of trip speed	0.747
Median trip distance	0.707
Fraction of driving between 0 and 30km/h	0.700

Table 18: Logistic Regression coefficients Feature Set F, UK, Gray shaded variables represent baseline-variables.

Variable	Exponentiated Coefficient
Median trip duration	1.726
Fraction of driving between 30 and 60km/h	1.320
Fraction of driving in rain	1.301
Acceleration events per 1000km	1.274
Percentage events near public POI	1.255
Fraction of driving between 30 and 60km/h	1.254
Braking events per 1000km	1.216
Cornering events in forest land-use areas per 1000km	1.173
Fraction of driving during rush hour	1.163
Quick lateral movement events per 1000km	1.162
Braking events in residential land-use areas per 1000km	1.155
Percentage of driving events near nightlife poi	1.138
Cornering events in mixed residential-commercial land-use areas per 1000km	1.122
Fraction of driving above 25°C	1.115
Fraction of driving at night	1.104
Cornering events in rural land-use areas per 1000km	1.080
Distance log	1.076
Fraction of driving above 130 km/h	1.068
Fraction of driving in snow	1.050
Cornering in commercial areas per 1000km	1.043
Percentage of driving events near transportation POI	0.956
Acceleration events in residential land-use areas per 1000km	0.927
Braking events in mixed rural-residential land-use areas per 1000km	0.890
Fraction of driving on the weekend	0.873
Fraction of driving in fog	0.864
Fraction of driving between 90 and 130km/h	0.856
Fraction of driving in overcast weather	0.849
Fraction of driving between 0 and 30km/h	0.798
Standard deviation of trip distance	0.793
Cornering events per 1000km	0.787
Median of trip distance	0.721
Mean of trip speed	0.689
Percentage of driving events near commercial POI	0.592

Tables 16, 17, 18 and 19 show the average coefficients of Feature Set B and F over all folds and for both study areas, derived from the logistic regression. For better interpretation, the coefficients were exponentiated. Since a feature selection was performed previously, not all original features of the corresponding feature sets were used. From an interpretation perspective, a coefficient greater than 1 indicates an increased risk with an increase in the feature, while a coefficient lower than 1 indicates that a decrease of the corresponding feature also decreases the accident risk. In detail, the coefficients describe the multiplicative change in odds of being an accident driver over being an accident-free driver for each unit increase in the corresponding feature, assuming that all other features are being kept constant.

The coefficients for the Italy dataset are mostly in line with the expectations. The natural log of the total mileage (distance log) seems to be the strongest coefficient in the baseline (Feature Set B). A few coefficients are close to 1, pointing out that they only have very little impact on average. Several coefficients related to city driving, such as *the fraction of driving between 0-30* and *60-90 km/h* seem to increase accident risk, as well as most of the driving events, which is also in line with the expectations. Furthermore, driving at higher speeds *between 90 and 130km/h* lowers the accident risk in both study areas. Interestingly, *the fraction of driving in rain and fog* seems to lower the accident risk in the Italian area. One possible explanation is that drivers who frequently drive in suboptimal conditions get used to driving in these conditions and adapt their driving style, which is also described in Chapter 2. Some studies found lower accident risk after prolonged periods of rain due to the aforementioned adapted driving behaviour.

For the UK data, however*, driving in the rain* seems to increase accident risk as expected. There are some more difficult to explain coefficients for the UK, such as *driving between 0 and 30km/h* seems to lower accident risk.

In both study areas, a higher *median of trip duration* results in increased accident risk, whereas a higher *median of trip distance* results in decreased accident risk. This points to the fact that driving long distances at higher speeds (highway trips) is safer than slower trips, e.g., city trips that take a longer time. Driver fatigue might also be a reason why trips with longer duration resulted in increased accident risk, as well as the increased risk resulting from *driving at night*. Also, the larger *standard deviation of trip distance* seems to lower accident risk, which is counterintuitive to the assumption that higher trip irregularity should result in higher risk. However, if a driver generally has longer distance trips that indicate safe highway driving, the standard deviation of the trip distance should also be higher. Therefore the *standard deviation of trip distance* might not be an optimal feature to measure travel irregularity.

For the temporal features, only *night driving* is included in the selected features of the Italy dataset, while *night, weekend,* and *rush-hour driving* are included for Feature Set F in the UK case. They all behave according to expectations, with *driving during rush-hour* resulting in increased risk, as well as *driving at night*, as previously stated. *Weekend driving* results in decreased risk in the UK case, possibly due to lower traffic volume.

In terms of temperature, driving in *high temperatures above 25°C* seems to lower accident risk in Italy, whereas it increases accident risk for the UK. The first interpretation of this observation might be that drivers in the UK are not used to high temperatures, which has a stronger effect on their ability to focus on driving.

Overall, no significant irregularities can be found in the logistic regression coefficients. Their directions are mostly in line with previous traffic accident research, which supports the quality of the data and the feature calculation process. Some coefficients that are not in line with previous research include *driving between 0-30km/h, cornering events per 1000km* and *driving in fog* for the UK, which all seem to decrease accident risk, despite the literature suggesting otherwise.

5.2.2 Feature Importances of Random Forest and XGBoost

In addition to the logistic regression coefficients, Figures 19 and 20 show the feature importances of the 20 most important features of feature set F and B for random forest and XGBoost, respectively, for the Italy dataset, coloured according to their category. Note that the categories Land-Use and POI are also based on the driving events. The driving events category (teal) only refers to the driving events without the inclusion of geographical information. All the importances represent average values over all folds. From the geographical features, *braking events in forest areas* seems to be an important predictor in both XGBoost and random forest, ranking 4 and 1, respectively. This might be because accident drivers had more driving events across almost all driving event categories, except for *braking in the forest* (see summary statistics in section 4.5.8). Further important features include the *fraction of driving between 0-30km/h*, which is the most important feature in the XGBoost model and ranked 3 and 2 in the random forest model. In addition, the *natural logarithm of the total distance (weighted distance)* is always among the most important features. *Weighted distance and 0-30km/h* being the most important features of set B is congruent with the results of logistic regression as seen in Table 15.

The *mean trip speed* is a further important feature, which probably helps to distinguish between city and highway driving. Some weather-related variables are also deemed important: *The percentage of driving in high temperatures, good weather*, and *at night in the rain* for XGBoost and the *percentage of km driven in rain for* the random forest model. No POI related features are found within the 20 most important features, which is also reflected in the fact that the inclusion of POI-related variables only yields minor performance improvements (see section 5.3).

In the UK case, driving events are more important than the *total distance driven* (Figure 21 and Figure 22), which can be attributed to a smaller difference in total mileage between accident and accident-free drivers in the UK dataset. Several land-use related features are among the most important features in the UK case, such as *braking events in residential* and *mixed rural-residential areas*. In addition, some weather-related features such as *driving in the rain*, *driving at night in the rain*, and *driving in temperatures above 25°C*. A single POI related feature can be found for both XGBoost and random forest in *the fraction of driving events near commercial POI.*

In general, logistic regression coefficients and feature importances of XGBoost and random forests tell a similar story. Furthermore, most of the selected features for logistic regression can be found in the most important features of random forest or XGBoost, further underlining their predictive power and overall importance.

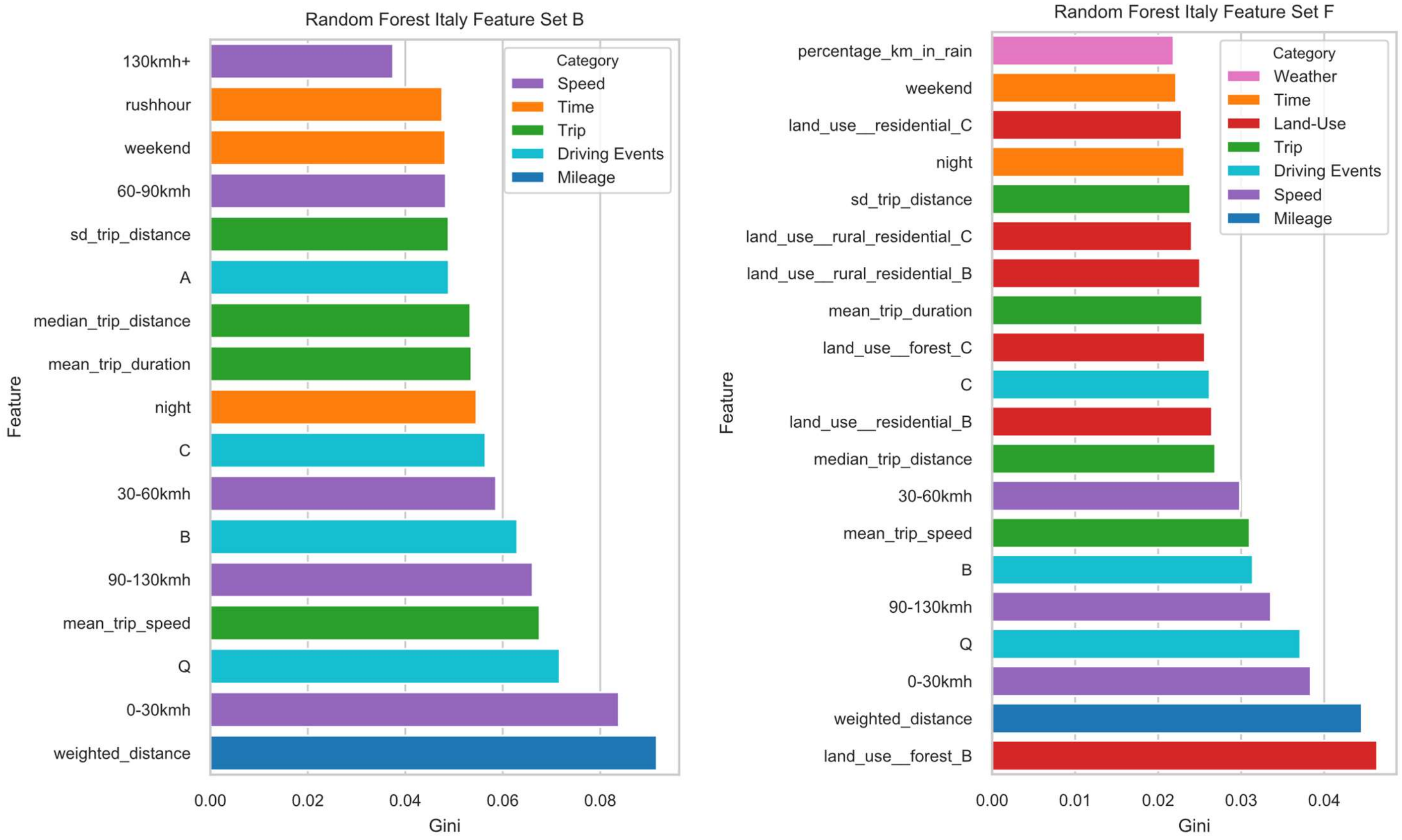

Figure 19: Italy Random Forest Feature Importance Feature Set B (left) and Feature Set F (right) (averaged over all folds).

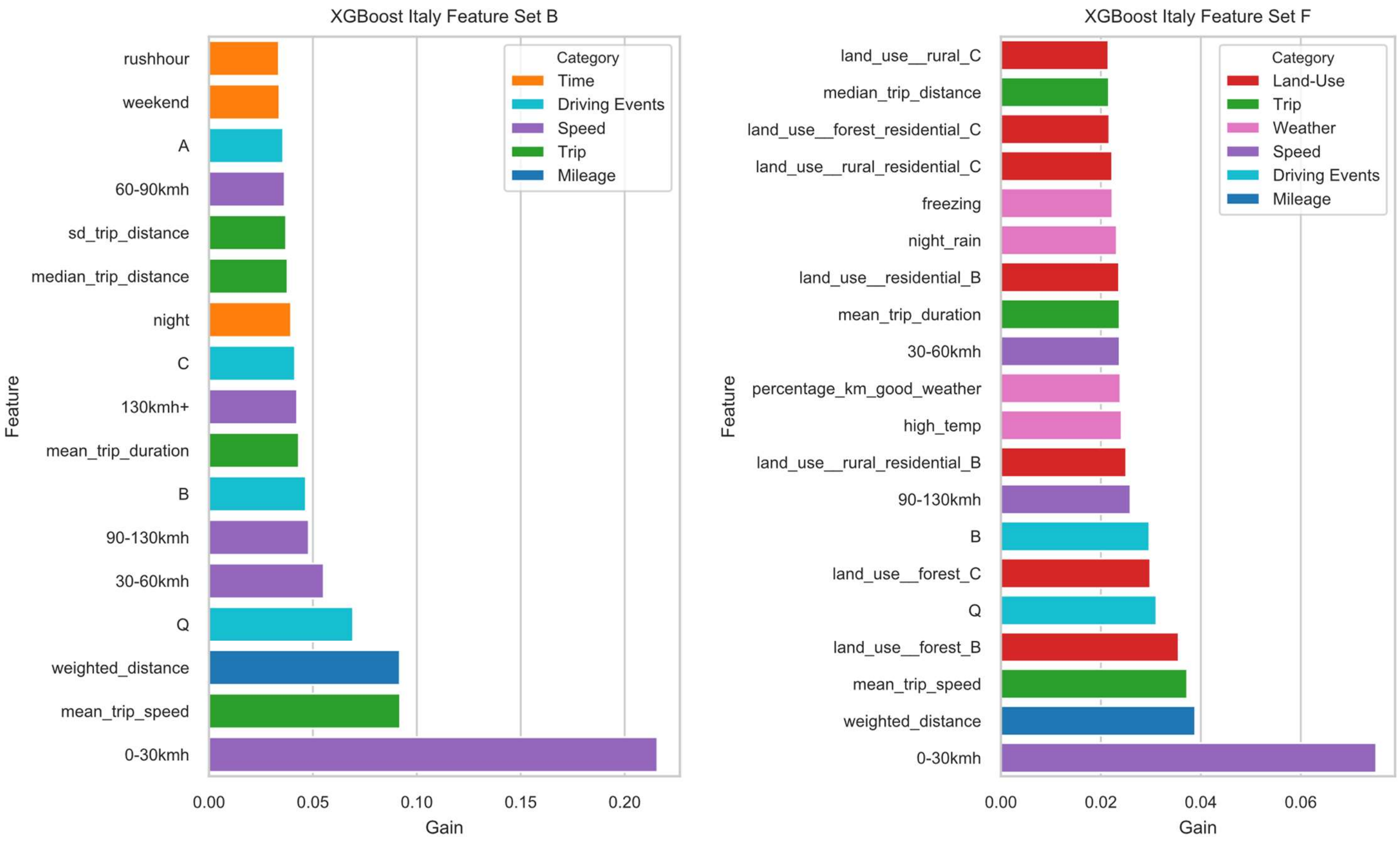

Figure 20: Italy XGBoost Feature Importance for Feature Set B (left) and Feature Set F (right) (averaged over all folds).

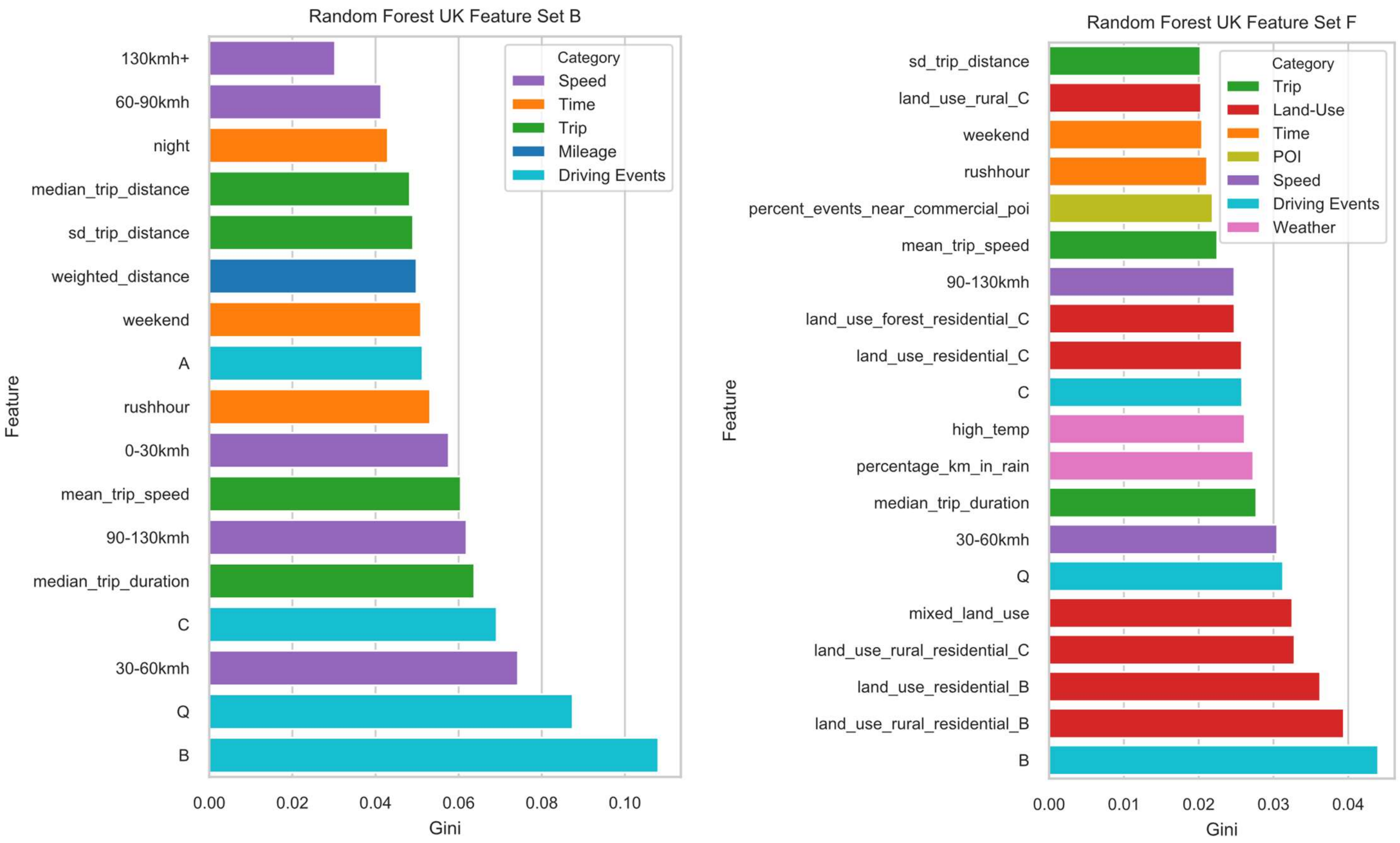

Figure 21: UK Random Forest Feature Importance for Feature Set B (left) and Feature set F (right) (averaged over all folds).

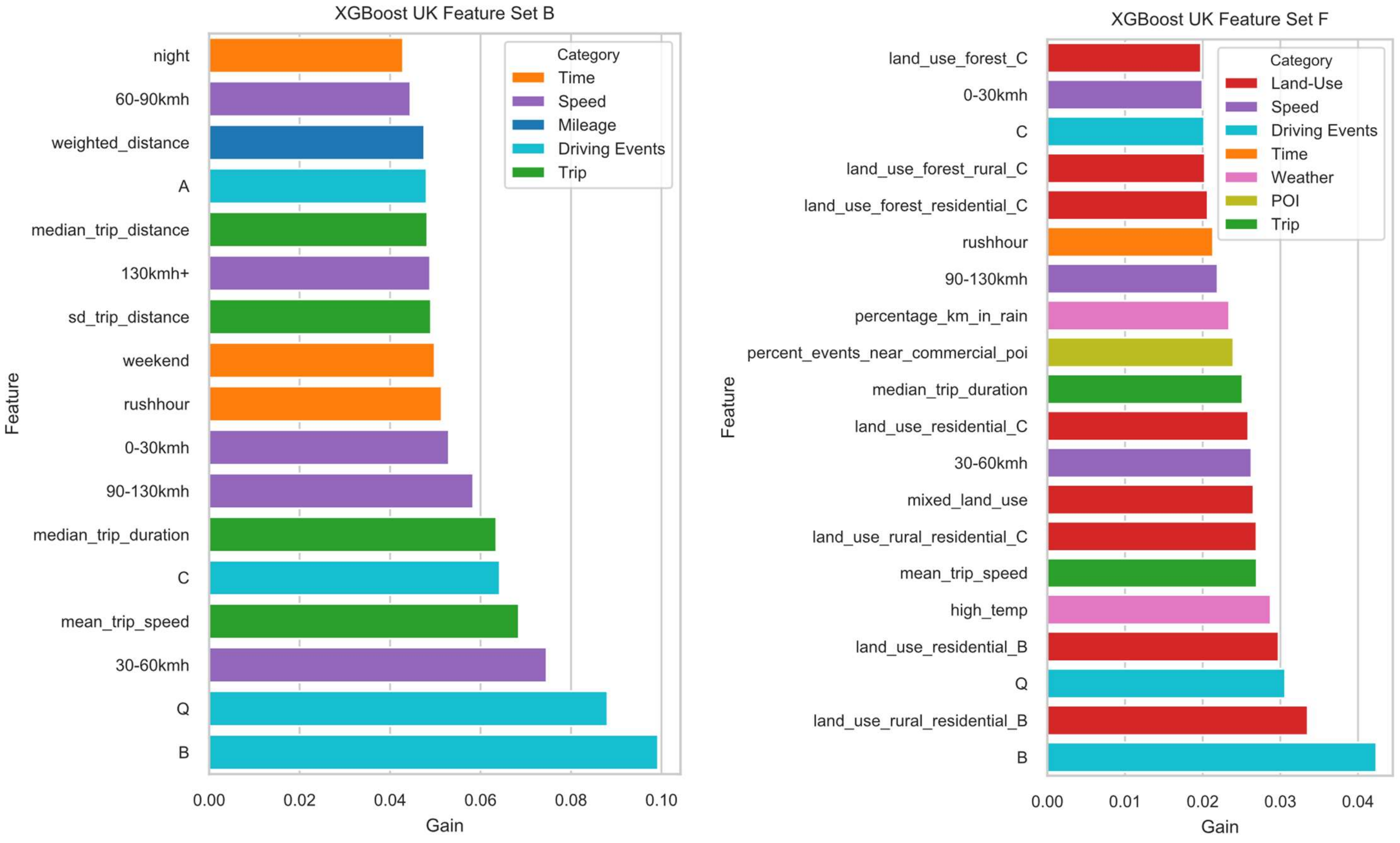

Figure 22: UK XGBoost Feature Importance for Feature Set B (left) and Feature Set F (right) (averaged over all folds).

5.3 Impact of Geographical Data

This section aims to quantify the performance improvement benefitting from geographical information. Specifically, the improvement from each type of geographical information will be compared for the two study areas.

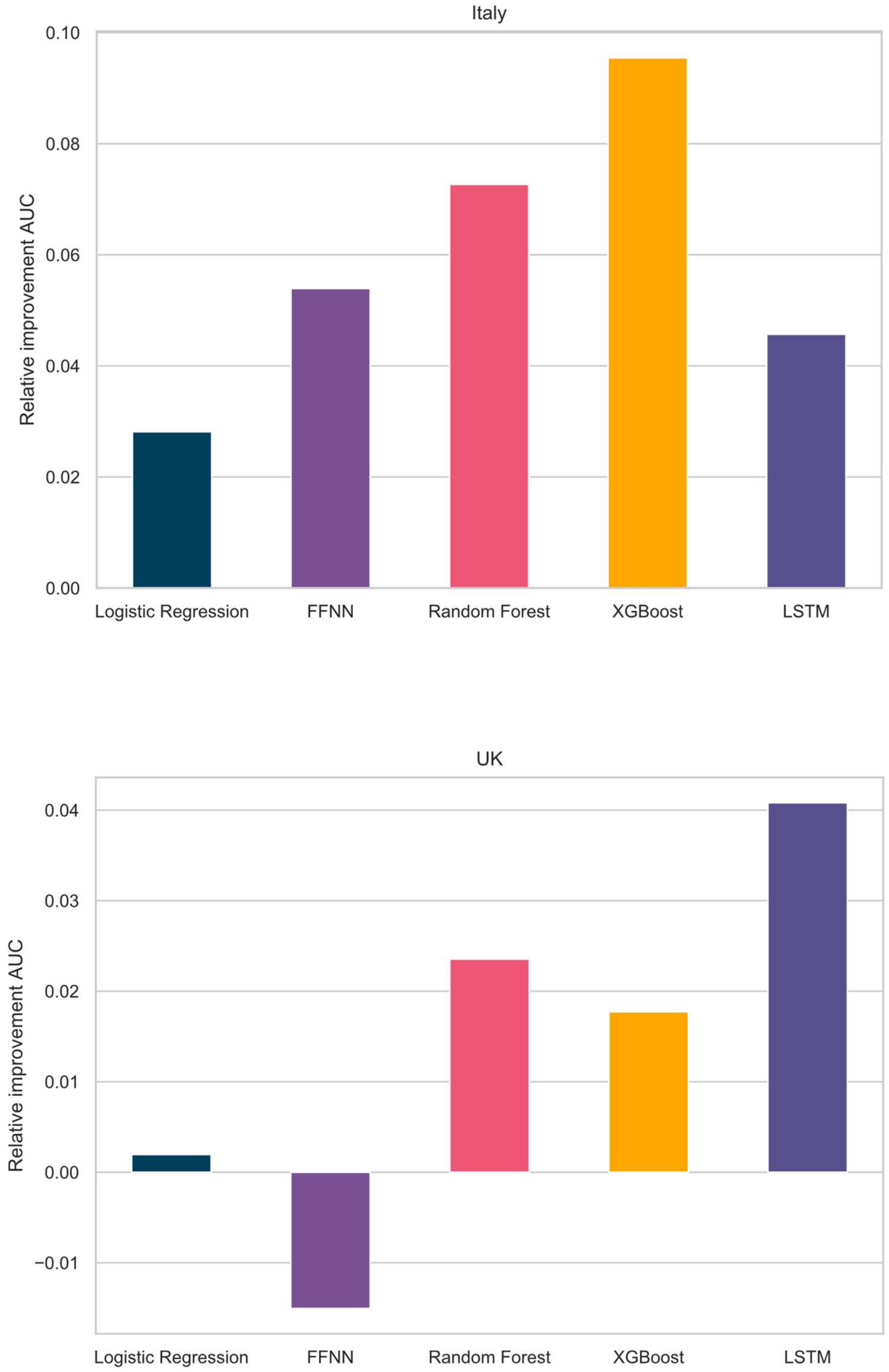

Figure 23: Relative improvement of feature set F over feature set B in terms of AUC for Italy (top) and UK (bottom). Note the different Y-axis scales.

As depicted in Figure 23, XGBoost yields the greatest relative performance improvement of Feature Set F (with all geographical information used) over Feature Set B (the baseline) in terms of AUC at 10% for the Italy case, followed by random forest. LSTM benefits the most for the UK data, but its overall performance is lower than the other algorithms, as previously shown in Section 5.1. In total, across the two study areas and five models, Feature Set F performs better than Feature Set B in 9 out of 10 times in terms of AUC.

In the UK, as seen in Figure 23, besides the LSTM that has a lower baseline accuracy, random forest and XGBoost again profit the most from geographical information, although on a smaller scale. While XGBoost yields higher performance improvements than the random forest in the Italy case, the relationship is inversed in the UK case. The performance of the FFNN even deteriorates. This can possibly be attributed to overfitting due to the small sample size and a larger number of features. However, since the UK results fluctuate a lot, as seen in Section 5.1, these results need to be considered with care.

The LSTM performance improves as well after the inclusion of geographical information for both study areas, although as previously stated, on a lower performance level in general. In general, logistic regression only slightly benefits from incorporating geographical information compared to the more complicated models. A possible interpretation for this observation might be that with the added amount of variables, the relationships within the dataset become too complicated for a relatively simple model like logistic regression to take advantage of. Furthermore, since a stepwise feature selection was performed for the logistic regression, there is a possibility that some feature combinations, which would yield a higher performance improvement were not included.

The effect of each type of geographical information is depicted in Figure 24. In Italy, land-use yields the highest performance improvement, whereas POI and weather only improve predictive performance by very small or even negative amounts in the case of the FFNN. In the UK, weather returns the highest performance improvement, slightly higher than POI, whereas land-use only improves the random forest model. The FFNN deteriorates with all geographical information in the UK case. Random forest is the only model that improves with each type of geographical information across both study areas, probably attributed to its robustness against overfitting.

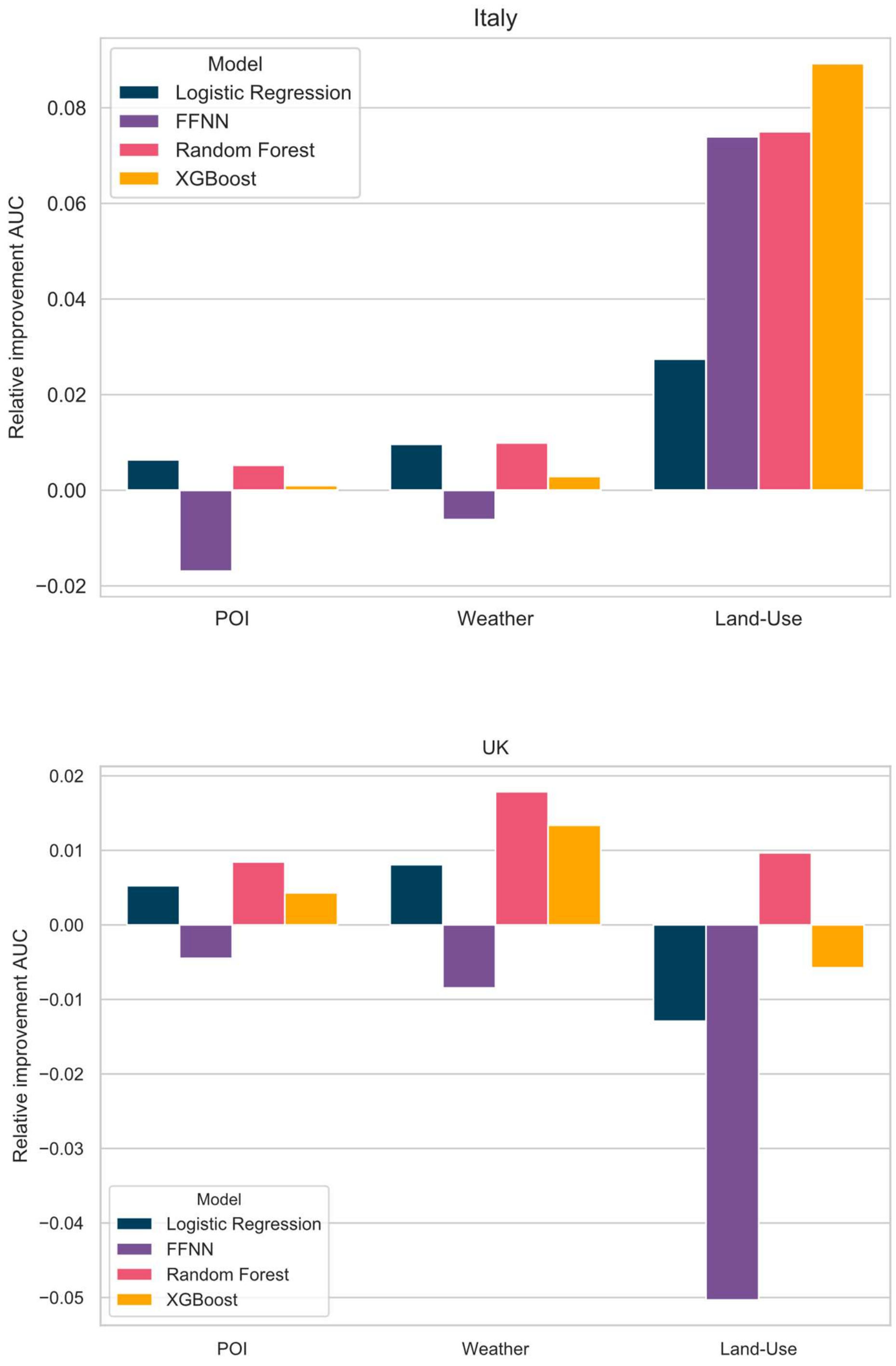

Figure 24: Relative improvement in terms of AUC from the inclusion of POI, weather and land-use for Italy (top) and UK (bottom). Note the different Y-axis scales.

5.4 Impact of Minimum Driving Distance

This section briefly illustrates the impact of the minimum chosen driving distance on the example of the yearly aggregated models and the Italy dataset. Table 19 shows the results in terms of AUC, accuracy, and F1-score of the Italy dataset if the minimum driving distance of 1500km is omitted. To avoid severe class imbalance and make the results comparable with the filtered scenario, 4000 drivers were randomly sampled from the accident-free drivers without filtering low mileage drivers in addition to the 3925 accident-drivers. The models were run in the same fashion as the filtered models, including hyperparameter-tuning for XGBoost and random forest and feature selection for logistic regression. The results show that the performance metrics compared to the unfiltered scenario are improved across all models and metrics (Figure 25 and Table 19). The relative model performance stays roughly the same as the filtered scenario, with XGBoost having the best performance, although the performance improvement due to geographical features is diminishing. Furthermore, the increase is mainly visible in terms of AUC. Accuracy only increases slightly after including geographic features, and the F1-Score is stagnating. The most likely explanation for this observation is that other features are less impactful and less room for improvement remains due to the increasing importance of the total distance.

Table 19: Performance metrics Italy without the minimum distance cutoff, the best overall performance is indicated in bold, the best performance per feature set is underlined

Model	A	B (baseline)	C	D	E	F
			AUC			
Logistic Regression	0.787	0.790	0.792	0.790	0.800	0.799
FFNN	0.787	0.791	0.795	0.789	0.819	0.816
Random Forest	0.787	0.799	0.801	0.801	0.820	0.818
XGBoost	0.787	0.800	0.800	0.800	0.822	**0.827**
			Accuracy			
Logistic Regression	0.747	0.747	0.745	0.747	0.749	0.748
FFNN	0.744	0.745	0.746	0.744	0.743	0.746
Random Forest	0.744	0.746	0.747	0.746	0.748	0.749
XGBoost	0.744	0.747	0.748	0.747	0.749	**0.753**
			F1-Score			
Logistic Regression	0.784	0.784	0.782	0.783	0.784	0.784
FFNN	0.785	0.785	0.780	0.780	0.775	0.771
Random Forest	0.784	0.785	0.785	0.786	0.786	0.787
XGBoost	0.786	0.789	**0.790**	0.789	0.784	0.783

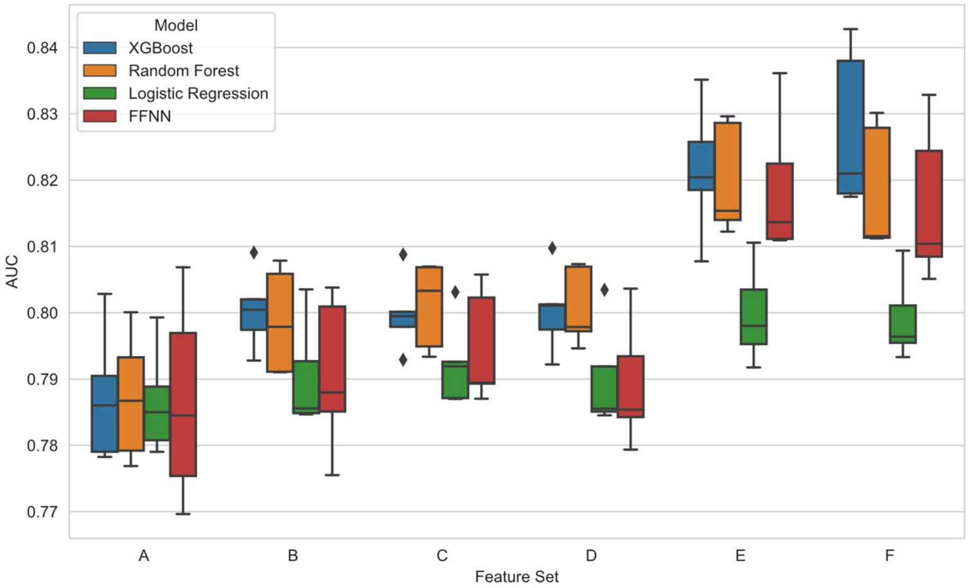

Figure 25: Boxplot of model performance (AUC) without applying the minimum distance cutoff over all folds, Italy

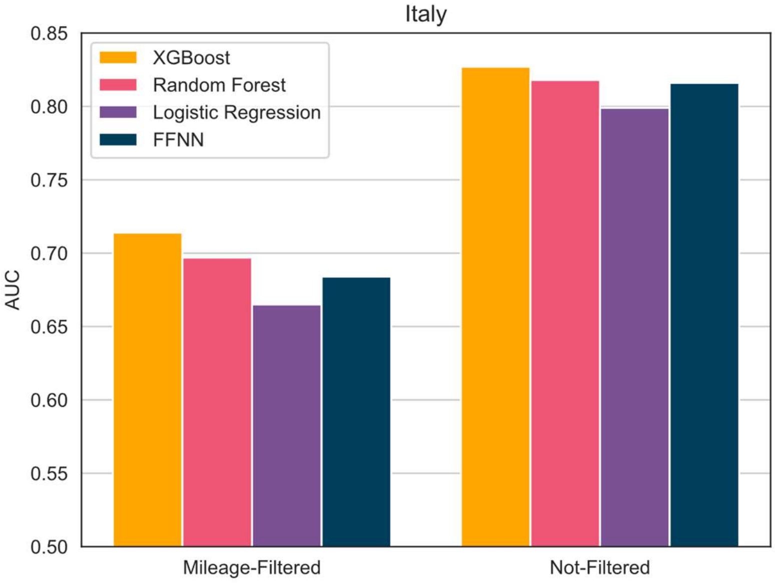

Figure 26: Italy model performance (AUC) of Feature Set F with and without applying the minimum distance cutoff.

The logistic regression coefficients in Table 20 are used to confirm the increased importance of the total distance. The coefficient value of *distance log* is almost doubled, with 3.270 against 1.660 of the unfiltered scenario (see Table 16). The rest of the coefficients are similar to the unfiltered scenario, with mileage, driving events, night driving, and driving at low speeds yielding the highest increase in accident risk, whereas driving at higher speeds and a higher median of the trip distance decreases accident risk. One irregularity can be found: all the weather conditions with the exception of *driving below 0°C* seem to lower the accident risk. This might be due to the feature selection procedure not being optimal and/or some co-correlated features still being present. Results for the UK dataset without the minimum distance cutoff can be found in the appendix (Figure 27 and Table 21). In general, they also yielded higher performance scores.

Table 20: Logistic Regression coefficients Italy, feature set F, without applying the minimum distance cutoff

Variable	**Average Exponentiated Coefficient**
Distance log	3.270
Braking events per 1000km	1.257
Fraction of driving between 0 and 30km/h	1.242
Cornering events per 1000km	1.225
Fraction of driving at night	1.202
Mean trip duration	1.145
Fraction of driving between 60 and 90km/h	1.138
Quick lateral movement events per 1000km	1.074
Fraction of driving during rush hour	1.070
Fraction of driving events near commercial POI	1.053
Fraction of driving in moderate temperature	1.038
Braking events in mixed rural-residential land-use areas per 1000km	1.024
Cornering events in residential land-use areas per 1000km	1.024
Fraction of driving below 0°C	1.018
Acceleration events per 1000km	1.012
Acceleration events in residential land-use areas per 1000km	1.011
Braking events in rural land-use areas per 1000km	1.008
Fraction of driving above 130km/h	0.986
Fraction of driving above 25°C	0.971
Percentage of driving events near transportation POI	0.969
Standard devation of trip distance	0.963
Fraction of driving on weekends	0.963
Braking events in forest land-use areas per 1000km	0.963
Cornering events in rural land-use areas per 1000km	0.942
Cornering events in mixed rural-residential land-use areas per 1000km	0.942
Percentage of driving in Fog	0.885
Median of trip distance	0.866
Braking events in residential land-use areas per 1000km	0.856
Fraction of driving between 90 and 130km/h	0.768
Percentage of driving in rain	0.753
Mean trip speed	0.703
Percentage of driving in overcast weather	0.578
Percentage of driving in good weather	0.460

5.5 Summary of Results

The main findings can be summarized with the following points:

- XGBoost generally performs best in Italy, while FFNN, random forest and logistic regression perform best in the UK depending on the feature sets and performance metrics.

- Without the inclusion of geographical information, logistic regression performs almost as well as more complicated models.

- The LSTM, which uses data aggregated at a per-trip level, performs worse than the other models.

- The relative performance improvement resulting from the inclusion of geographical data is on a scale of up to 10% in terms of AUC in the case of Italy and the XGBoost model. The improvement mainly results from the usage of land-use-related features in Italy and weather-related features in the UK. In general, XGBoost and random forest benefit the most from the inclusion of geographical information, while logistic regression benefits the least amount.

- The chosen minimum driving distance matters: Without filtering low mileage drivers, model performance and importance of total distance increase significantly, although the relative performance stays roughly the same.

- The logistic regression coefficients and feature importances of random forest and XGBoost are mostly in line with what the literature suggests. Total distance and city-driving related features such as driving at lower speeds are the biggest contributors to higher accident risk. On the other hand, driving in rural areas and at higher speed is linked with lower accident risk. Temporal features behave as expected, with driving at night and during peak hours resulting in increased accident risk

- The impact of weather conditions differs between the study areas. For example, driving in rain increases accident risk in the UK while it lowers accident risk in Italy.

Chapter 6: Discussion

6.1 Model Comparison and Impact of Geographical Information

The current baseline method logistic regression employed in car PAYD and PHYD studies performs decently well and only gets slightly outperformed by more complicated algorithms. XGBoost generally yields the best predictive performance along with decent interpretability. However, due to the small performance difference from logistic regression, it makes sense to use a logistic regression rather than more complicated models, especially for the interpretation of size and direction of the effects of the individual features. Especially if the maximum possible performance is not the primary aim, logistic regression should be the preferred model. This confirms the findings of previous studies, namely Paefgen et al. (2013, 2014), Huang & Meng (2019), and Backe & Bocca (2017) that all report small benefits from the usage of more complicated models over logistic regression and Paefgen et al. (2014) specifically recommend the usage of the latter.

The LSTM model yields worse performance than the yearly aggregated models. This might be due to the data aggregation procedure, trip definition, model architecture, or the difference in number and types of features used. It can be argued that car accident risk prediction might rely heavily on feature engineering and domain knowledge. Therefore the LSTM model with its main strength in utilizing raw data might not be the optimal choice. However, the LSTM still has potential if the above-mentioned factors such as trip definition, feature selection, or parameter selection are changed. The usage of more complicated deep learning models in individual car accident risk prediction is worth further investigation. The LSTM showed significant improvement by including more features. Therefore adding more geographic or other information might increase its performance further.

The performance improvement by adding geographical data is moderate, especially for the UK area, yet still significant in both study areas. Land-use seems to be the most impactful geographical feature in Italy, with weather and POI only showing minor or no improvements over the baseline. It can be argued that since no map matching was performed, part of the performance improvement which results from the inclusion of geographical information could also be derived from, e.g., road types and network measures such as centrality. For the UK area, the inclusion of weather data yielded the most improvement. This could be explained by the different climate and higher variability of the weather conditions in the UK, whereas Italy usually has good weather all around. The higher impact of land-use in Italy compared to the UK area can possibly be explained due to the higher heterogeneity of the study area. The UK study area is mainly urban, focusing on London, while the Italian area includes cities of all sizes and rural areas.

Apart from geographical features, several features that were deemed important in previous studies were confirmed. The total mileage remains an important factor amongst almost all models. Simultaneously driving at low speed and a high frequency of driving events resulted in higher accident risk as expected. Furthermore, frequent trips that take a long time and night driving are linked to higher accident risk, possibly due to driver fatigue.

To sum up the answer to RQ 1: *Which driving behaviour features are most suitable for predicting individual car accident risk and to what extent can geographical improve this prediction*, it can be said that mileage, driving events, and driving in different speed- and time-intervals are all suitable features to predict individual car accident risk. It has been shown that the inclusion of geographical context features can further strengthen the prediction. However, more complicated models are required to make full use of these additional features, which results in lower interpretability. Furthermore, the importance of geographical features depends on the minimum distance chosen; if very low mileage drivers are included, total distance becomes more important compared to other features and diminishes their importance.

In regards to RQ2: *Which machine learning models are most suitable for predicting individual car accident risk and what is the trade-off between performance and interpretability*, it can be summed up that tree-based models such as XGBoost and random forest perform best for the prediction task at hand, possibly attributed to their robustness against overfitting and ability to model non-linear relationships. This is mainly true for the Italian dataset, as it is not entirely clear which model performs best in the UK due to the small sample sizes and large between-fold differences. From the results at hand, the trade-off between interpretability and predictive performance is relatively large when geographical information is not included since only small or even negative performance benefits result from the usage of models such as XGBoost and random forest over logistic regression. In contrast, their interpretability is significantly worse than that of logistic regression. However, the trade-off becomes smaller if geographical features are included, especially for the Italian dataset, as the performance advantage of the more complicated models gets larger.

It should be noted that car accidents are random events to a large degree, which are impossible to predict with very high accuracy. The margins for improvement are generally narrow. On the other hand, even a small improvement in accuracy can potentially translate into a large amount of money in an insurance context. Not to mention the value of human health, which could potentially be improved through the incentive of safer and more ecological driving, which PHYD car insurance provides. The results show that the inclusion of geographical information can improve model accuracy and therefore has the potential to improve PHYD car insurance and its benefits. Hence to improve model accuracy, further research about the inclusion of geographical information into car accident risk modelling is recommended. Ultimately though, the driver remains the biggest risk factor and not the environment.

6.2 Differences between UK and Italy

In general, the results for the Italian dataset are more reliable in the sense that they exhibited much smaller between-fold differences. This can mostly be attributed to the around 10 times larger sample. Overall, though the UK dataset helped in confirming the potential of geographical information and pointed out that depending on the area, different types of geographical information and machine-learning models might be optimal. Furthermore, more complicated models such as neural networks, random forest and XGBoost can suffer from overfitting if the sample size is too small, illustrated by the relatively high performance of logistic regression in the UK dataset. Especially in the context of car accident prediction where there can be large differences in each driver's driving behaviour, a sufficient sample size is needed to avoid this problem and produce reliable predictions.

In regards to RQ3: *Are there geographical and cultural differences between London and Italy regarding the effects of driving behaviour on accident risk*, it can be stated that there are indeed several differences which can be observed: Driving events are more important in the UK and mileage is less important than in the Italian dataset, although this could also be attributed to the data collection/sampling strategy. Furthermore, it can be stated that in a more homogenous area like the UK, spatial features such as land-use are less important, whereas weather becomes more important. Some weather conditions have reversed effects in the two study areas, e.g., rain lowers the accident risk in Italy while it increases the accident risk in the UK. However, the impact of speed- and time-related features such as driving at a lower speed or on the weekend remains roughly the same across both study areas.

6.3 Limitations

There are a few limitations to this study that need to be pointed out, mainly regarding data availability. One of the biggest limitations stems from the fact that there was no information available on whether the driver was actually at fault for an accident. This is different from previous studies, which usually had the information from the insurance company about whether an at-fault claim was made. Therefore some safe drivers, which do not exhibit any typical dangerous behaviour are labelled as accident drivers through circumstances outside of their control. Logically these drivers are very hard to classify as accident drivers since they do not exhibit any dangerous driving patterns. This goes hand in hand with the next limitation, which is that it is impossible to verify with 100% certainty that all the accidents included in the sample are real accidents and not a false alarm. Although a preprocessing of the crash alarms was performed to the best of the author's ability, most crashes do not contain any notes from the crash assistance centre. It would further be beneficial to have demographic data about the drivers, such as age, gender and driving experience to allow a comparison or combination with traditional car insurance risk models, as done in previous studies.

Another data-related limitation of this study lies in the relatively low spatial resolution of the Italy dataset, which only recorded one waypoint every 2000 meters. This makes potential map matching very difficult and inaccurate, especially for urban areas. Incorporating more information about the road network, such as various centrality measures or average traffic volumes are further variables that could be included if higher resolution data was available.

Furthermore, all the geographical data was simplified, it could be possible to achieve higher performance with more detailed variables. In regards to the weather data, real data might also yield better results than the historical forecast data which was used in this thesis. Also, the driving events could have been explored in more detail according to their acceleration values, in order to distinguish between events with different levels of severity. In addition, the total time driven could have been used instead of distance as the main exposure factor.

From a modelling perspective, it is possible that different neural network architectures or parameters might yield better results, as well as a more extensive grid search for random forest and XGBoost. A different feature selection process could have been used for the logistic regression, as stepwise-selection has its limitations. In addition, the model performance under different ratios of accident versus accident-free drivers and different classification thresholds could have been explored, since in a real-world application case this ratio can be highly imbalanced, although this has already been done by previous studies, e.g., Paefgen et al. (2013).

It also needs to be stated that the model performance numbers of this study are not really comparable to other studies. This is due to different recording types, different frameworks for registering a claim or accident, different class balances etc.

Lastly, it needs to be pointed out that as with all GPS-tracking applications, privacy concerns exist within the context of behaviour-based car insurance. However, the discussion of those lies outside the scope of this study.

Chapter 7: Conclusion and Future Work

7.1 Summary and Implications

Several machine learning models, logistic regression, random forest, XGBoost, FFNN, and LSTM were compared across two study areas in the UK and Italy and different feature subsets. These features were calculated based on a year of GPS and driving events data in order to predict car accident risk on a driver level and perform a binary classification to separate accident from accident-free drivers. Specifically, the research gap on the impact of the inclusion of geographical context data and the comparison of different machine learning algorithms and the two study areas was of interest. In order to fill these research gaps, the trajectory and driving behaviour data was enriched with weather, POI, and land-use data and several novel features derived.

It was found that the inclusion of geographical context can improve relative performance in terms of AUC by up to 10% using the XGBoost algorithm. This is a relatively high number, considering the inherent randomness in car accidents and a promising first result. From those geographical features, land-use has the biggest impact on predictive performance for the Italian dataset and weather for the UK dataset. In general, the Italian dataset yielded more reliable results due to the bigger sample size. The performance improvement resulting from the inclusion of geographical information was generally strongest if more complicated models were used, which in turn reduced interpretability.

Apart from the geographical information, findings of previous studies regarding model performance and the impact of several features such as mileage, speed, time of day and driving events on car accident risk were mostly confirmed, supporting the quality of the data and the feature calculation process. It was also pointed out that the chosen minimum mileage of drivers included in the modelling matters and can have a large effect on overall model performance.

A different data aggregation approach with finer temporal granularity on trip-level and the usage of an LSTM neural network returned worse predictive performance than the yearly aggregated approach. This might be due to the model parameters, aggregation levels, definition of a trip due to the different features which were used. However, it shows potential if more features are used and the usage of deep learning models for car accident risk is worth further attention.

In terms of real-world applicability, the results show that the car insurance industry and road traffic researchers could possibly benefit from including geographical context information into their risk models making them more accurate. Since most of this information is freely and easily accessible, the data enrichment could possibly be integrated directly into their risk modelling pipeline. Although further research about which features to include is recommended, as described in the next section regarding future work. Furthermore, as pointed out by Baecke & Bocca (2017) in many places models with high interpretability such as logistic regression are required by law, which might result in lower benefits from including aforementioned geographical information.

7.2 Future Work

In regards to future work, it is suggested to include real weather information instead of historical forecast data as well as map matching in order to possibly get more accurate results. Furthermore, other types of POI and land-use data and their combinations can be explored. Several other geographical features such as population density or elevation models could further be included as well to possibly derive more meaningful features with strong predictive performance. Furthermore, after map matching, information about the usual traffic condition on certain road segments could be included, as well as network measures such as road centrality.

In addition, to confirm the differences between the two study areas and the impact of geographical features, the proposed modelling and feature engineering strategy should be tested using a bigger dataset.

From a modelling perspective, a combination of previously suggested approaches through ensemble models could be tried. Furthermore, the usage of different data aggregation approaches and deep learning models is worth further investigation. In addition, for real-world applicability, the proposed strategy integrating geographical features should also be tested on claim-count data, using a regression instead of a classification strategy.

To simplify weather data collection, future GPS recorders could record factors such as temperature and precipitation directly from the cars built-in temperature and rain sensors, although there might be issues comparing different sensors from different car models.

Lastly, since the driver remains the most important risk factor, it is certainly beneficial to further the research about the traditional non-geographical driving behaviour features derived from telematics data and refine, e.g., the definition of driving events.

References

Abdel-Aty, Mohamed, Al Ahad Ekram, Helai Huang, and Keechoo Choi. 2011. "A Study on Crashes Related to Visibility Obstruction Due to Fog and Smoke." *Accident Analysis and Prevention* 43 (5): 1730–37. https://doi.org/10.1016/j.aap.2011.04.003.

Af Wåhlberg, A. E.. 2004. "The Stability of Driver Acceleration Behavior, and a Replication of Its Relation to Bus Accidents." *Accident Analysis and Prevention* 36 (1): 83–92. https://doi.org/10.1016/S0001-4575(02)00130-6.

Af Wåhlberg, A. E. 2000. "The Relation of Acceleration Force to Traffic Accident Frequency: A Pilot Study." *Transportation Research Part F: Traffic Psychology and Behaviour* 3 (1): 29–38. https://doi.org/10.1016/S1369-8478(00)00012-7.

Af Wåhlberg, A. E. 2007. "Aggregation of Driver Celeration Behavior Data: Effects on Stability and Accident Prediction." *Safety Science*. https://doi.org/10.1016/j.ssci.2006.07.008.

Af Wåhlberg, A. E. 2008a. "If You Can't Take the Heat: Influences of Temperature on Bus Accident Rates." *Safety Science* 46 (1): 66–71. https://doi.org/10.1016/j.ssci.2007.02.003.

Af Wåhlberg, A. E. 2008b. "Driver Celeration Behaviour and Accidents – an Analysis." *Theoretical Issues in Ergonomics Science* 9 (5): 383–403. https://doi.org/10.1080/14639220701596722.

Alkahtani, Khalid F., Mohamed Abdel-Aty, and Jaeyoung Lee. 2019. "A Zonal Level Safety Investigation of Pedestrian Crashes in Riyadh, Saudi Arabia." *International Journal of Sustainable Transportation* 13 (4): 255–67. https://doi.org/10.1080/15568318.2018.1463417.

Andrey, Jean, Brian Mills, Mike Leahy, and Jeff Suggett. 2003. "Weather as a Chronic Hazard for Road Transportation in Canadian Cities." *Natural Hazards* 28 (2–3): 319–43. https://doi.org/10.1023/A:1022934225431.

Andrey, Jean, and Sam Yagar. 1993. "A Temporal Analysis of Rain-Related Crash Risk." *Accident Analysis and Prevention* 25 (4): 465–72. https://doi.org/10.1016/0001-4575(93)90076-9.

Ayuso, Mercedes, Montserrat Guillen, and Jens Perch Nielsen. 2019. "Improving Automobile Insurance Ratemaking Using Telematics: Incorporating Mileage and Driver Behaviour Data." *Transportation* 46 (3): 735–52. https://doi.org/10.1007/s11116-018-9890-7.

Ayuso, Mercedes, Montserrat Guillén, and Ana María Pérez-Marín. 2014. "Time and Distance to First Accident and Driving Patterns of Young Drivers with Pay-as-You-Drive Insurance." *Accident Analysis and Prevention* 73 (December): 125–31. https://doi.org/10.1016/j.aap.2014.08.017.

Ayuso, Mercedes, Montserrat Guillén, and Ana María Pérez Marín. 2016. "Using GPS Data to Analyse the Distance Travelled to the First Accident at Fault in Pay-as-You-Drive Insurance." *Transportation Research Part C: Emerging Technologies* 68 (July): 160–67. https://doi.org/10.1016/j.trc.2016.04.004.

Baecke, Philippe, and Lorenzo Bocca. 2017. "The Value of Vehicle Telematics Data in Insurance Risk Selection Processes." *Decision Support Systems*. https://doi.org/10.1016/j.dss.2017.04.009.

Baecke, Philippe, and Dirk Van Den Poel. 2011. "Data Augmentation by Predicting Spending Pleasure Using Commercially Available External Data." *Journal of Intelligent Information Systems* 36 (3): 367–83. https://doi.org/10.1007/s10844-009-0111-x.

Behera, Ranjan Kumar, Monalisa Jena, Santanu Kumar Rath, and Sanjay Misra. 2021. "Co-LSTM: Convolutional LSTM Model for Sentiment Analysis in Social Big Data." *Information Processing and Management* 58 (1): 102435. https://doi.org/10.1016/j.ipm.2020.102435.

Bentley, Jon Louis. 1975. “Multidimensional Binary Search Trees Used for Associative Searching.” *Communications of the ACM* 18 (9): 509–17. https://doi.org/10.1145/361002.361007.

Bergel-Hayat, Ruth, Mohammed Debbarh, Constantinos Antoniou, and George Yannis. 2013. “Explaining the Road Accident Risk: Weather Effects.” *Accident Analysis & Prevention* 60 (November): 456–65. https://doi.org/10.1016/j.aap.2013.03.006.

Bian, Yiyang, Chen Yang, J. Leon Zhao, and Liang Liang. 2018. “Good Drivers Pay Less: A Study of Usage-Based Vehicle Insurance Models.” *Transportation Research Part A: Policy and Practice*. https://doi.org/10.1016/j.tra.2017.10.018.

Black, Alan W., and Thomas L. Mote. 2015. “Effects of Winter Precipitation on Automobile Collisions, Injuries, and Fatalities in the United States.” *Journal of Transport Geography* 48 (October): 165–75. https://doi.org/10.1016/j.jtrangeo.2015.09.007.

Bordoff, Jason E., and Pascal Noel. 2008. “Pay-As-You-Drive Auto Insurance: A Simple Way to Reduce Driving-Related Harms and Increase Equity.”

Boucher, J.-P, A M Peârez-Marõân, and Miguel Santolino. 2013. “Pay-As-You-Drive Insurance: The Effect of The Kilometers on the Risk of Accident.” *Anales Del Instituto De Actuarios Espanoles* 19 (January): 135–54.

Boucher, Jean Philippe, and Roxane Turcotte. 2020. “A Longitudinal Analysis of the Impact of Distance Driven on the Probability of Car Accidents.” *Risks* 8 (3): 1–19. https://doi.org/10.3390/risks8030091.

Brijs, Tom, Dimitris Karlis, and Geert Wets. 2008. “Studying the Effect of Weather Conditions on Daily Crash Counts Using a Discrete Time-Series Model.” *Accident Analysis and Prevention* 40 (3): 1180–90. https://doi.org/10.1016/j.aap.2008.01.001.

Caliendo, Ciro, Maurizio Guida, and Alessandra Parisi. 2007. “A Crash-Prediction Model for Multilane Roads.” *Accident Analysis and Prevention* 39 (4): 657–70. https://doi.org/10.1016/j.aap.2006.10.012.

Cawley, Gavin C., and Nicola L.C. Talbot. 2010. “On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation.” *Journal of Machine Learning Research*. Vol. 11. https://doi.org/10.5555/1756006.1859921.

Chang, Li Yen, and Wen Chieh Chen. 2005. “Data Mining of Tree-Based Models to Analyze Freeway Accident Frequency.” *Journal of Safety Research* 36 (4): 365–75. https://doi.org/10.1016/j.jsr.2005.06.013.

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System.” In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-August-2016:785–94. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785.

Chen, Yuan Yuan, Yisheng Lv, Zhenjiang Li, and Fei Yue Wang. 2016. “Long Short-Term Memory Model for Traffic Congestion Prediction with Online Open Data.” In *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 132–37. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ITSC.2016.7795543.

Cheng, Fan, Xia Zhang, Chuang Zhang, Jianfeng Qiu, and Lei Zhang. 2018. “An Adaptive Mini-Batch Stochastic Gradient Method for AUC Maximization.” *Neurocomputing* 318 (November): 137–50. https://doi.org/10.1016/j.neucom.2018.08.041.

Comber, A. J. 2008. “The Separation of Land Cover from Land Use Using Data Primitives.” *Journal of Land Use Science* 3 (4): 215–29. https://doi.org/10.1080/17474230802465173.

Cui, Zhiyong, Ruimin Ke, Ziyuan Pu, and Yinhai Wang. 2020. “Stacked Bidirectional and Unidirectional LSTM Recurrent Neural Network for Forecasting Network-Wide Traffic State with Missing Values.” *Transportation Research Part C: Emerging Technologies* 118 (September): 102674. https://doi.org/10.1016/j.trc.2020.102674.

Denuit, Michel, Xavier Marchal, Sandra Pitrebois, and Jean-Franois Walhin. 2007. *Actuarial Modelling of Claim Counts*. *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470517420.

DiGangi, Elizabeth A., and Joseph T. Hefner. 2013. “Ancestry Estimation.” In *Research Methods in Human Skeletal Biology*, 117–49. Elsevier Inc. https://doi.org/10.1016/B978-0-12-385189-5.00005-4.

Eboli, Laura, Gabriella Mazzulla, and Giuseppe Pungillo. 2017. “How to Define the Accident Risk Level of Car Drivers by Combining Objective and Subjective Measures of Driving Style.” *Transportation Research Part F: Traffic Psychology and Behaviour* 49 (August): 29–38. https://doi.org/10.1016/j.trf.2017.06.004.

Eisenberg, Daniel. 2004. “The Mixed Effects of Precipitation on Traffic Crashes.” *Accident Analysis and Prevention* 36 (4): 637–47. https://doi.org/10.1016/S0001-4575(03)00085-X.

Eisenberg, Daniel, and Kenneth E. Warner. 2005. “Effects of Snowfalls on Motor Vehicle Collisions, Injuries, and Fatalities.” *American Journal of Public Health* 95 (1): 120–24. https://doi.org/10.2105/AJPH.2004.048926.

Fang, Zhice, Yi Wang, Ling Peng, and Haoyuan Hong. 2020. “Predicting Flood Susceptibility Using Long Short-Term Memory (LSTM) Neural Network Model.” *Journal of Hydrology*, November, 125734. https://doi.org/10.1016/j.jhydrol.2020.125734.

Fischer, P, A Comber, and R Wadsworth. 2005. “Land Use and Land Cover: Contradiction or Complement.” In *Re-Presenting GIS*, edited by P Fisher and D Unwin, 85–98. Chichester, UK: Wiley.

Fountas, Grigorios, Achille Fonzone, Niaz Gharavi, and Tom Rye. 2020. “The Joint Effect of Weather and Lighting Conditions on Injury Severities of Single-Vehicle Accidents.” *Analytic Methods in Accident Research* 27 (September): 100124. https://doi.org/10.1016/j.amar.2020.100124.

Giulietti, Corrado, Mirco Tonin, and Michael Vlassopoulos. 2020. “When the Market Drives You Crazy: Stock Market Returns and Fatal Car Accidents.” *Journal of Health Economics* 70 (March): 102245. https://doi.org/10.1016/j.jhealeco.2019.102245.

Guillen, Montserrat, Jens Perch Nielsen, Mercedes Ayuso, and Ana M. Pérez-Marín. 2019. “The Use of Telematics Devices to Improve Automobile Insurance Rates.” *Risk Analysis* 39 (3): 662–72. https://doi.org/10.1111/risa.13172.

Gumus, Mesut, and Mustafa S. Kiran. 2017. “Crude Oil Price Forecasting Using XGBoost.” In *2nd International Conference on Computer Science and Engineering, UBMK 2017*, 1100–1103. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/UBMK.2017.8093500.

Guo, Feng, and Youjia Fang. 2013. “Individual Driver Risk Assessment Using Naturalistic Driving Data.” *Accident Analysis and Prevention* 61 (December): 3–9. https://doi.org/10.1016/j.aap.2012.06.014.

Gutierrez-Osorio, Camilo, and César Pedraza. 2020. “Modern Data Sources and Techniques for Analysis and Forecast of Road Accidents: A Review.” *Journal of Traffic and Transportation Engineering (English Edition)*, July. https://doi.org/10.1016/j.jtte.2020.05.002.

Guttman, Antonin. 1984. “R-Trees: A Dynamic Index Structure for Spatial Searching.” *ACM SIGMOD Record* 14 (2): 47–57. https://doi.org/10.1145/971697.602266.

Highway Traffic Safety Administration, National, and Us Department of Transportation. 2018. “TRAFFIC SAFETY FACTS Crash • Stats Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey.”

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. “Long Short-Term Memory.” *Neural Computation* 9 (8): 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.

Huang, Yifan, and Shengwang Meng. 2019. “Automobile Insurance Classification Ratemaking Based on Telematics Driving Data.” *Decision Support Systems* 127 (December). https://doi.org/10.1016/j.dss.2019.113156.

Husnjak, Siniša, Dragan Peraković, Ivan Forenbacher, and Marijan Mumdziev. 2015. “Telematics System in Usage Based Motor Insurance.” In *Procedia Engineering*, 100:816–25. Elsevier Ltd. https://doi.org/10.1016/j.proeng.2015.01.436.

Jia, Ruo, Anish Khadka, and Inhi Kim. 2018. “Traffic Crash Analysis with Point-of-Interest Spatial Clustering.” *Accident Analysis and Prevention* 121 (December): 223–30. https://doi.org/10.1016/j.aap.2018.09.018.

Kamińska, Joanna A. 2018. “The Use of Random Forests in Modelling Short-Term Air Pollution Effects Based on Traffic and Meteorological Conditions: A Case Study in Wrocław.” *Journal of Environmental Management* 217 (July): 164–74. https://doi.org/10.1016/j.jenvman.2018.03.094.

Kantor, S., and T. Stárek. 2014. “Design of Algorithms for Payment Telematics Systems Evaluating Driver’s Driving Style.” *Transactions on Transport Sciences* 7 (1): 9–16. https://doi.org/10.2478/v10158-012-0049-5.

Kim, Karl, and Eric Yamashita. 2002. “Motor Vehicle Crashes and Land Use: Empirical Analysis from Hawaii.” *Transportation Research Record: Journal of the Transportation Research Board* 1784 (1): 73–79. https://doi.org/10.3141/1784-10.

Krause, Cory M., and Lei Zhang. 2019. “Short-Term Travel Behavior Prediction with GPS, Land Use, and Point of Interest Data.” *Transportation Research Part B: Methodological* 123 (May): 349–61. https://doi.org/10.1016/j.trb.2018.06.012.

Kufera, Joseph A., Ahmad Al-Hadidi, Daniel G. Knopp, Zachary D.W. Dezman, Timothy J. Kerns, Olasunmbo Eva Okedele, Geoffrey L. Rosenthal, and J. Kathleen Tracy. 2020. “The Impact of a New Casino on the Motor Vehicle Crash Patterns in Suburban Maryland.” *Accident Analysis & Prevention* 142 (July): 105554. https://doi.org/10.1016/j.aap.2020.105554.

Langford, Jim, Sjaanie Koppel, Dennis McCarthy, and Sivaramakrishnan Srinivasan. 2008. “In Defence of the ‘Low-Mileage Bias.’” *Accident Analysis and Prevention* 40 (6): 1996–99. https://doi.org/10.1016/j.aap.2008.08.027.

Lee, Jaeyoung, Mohamed Abdel-Aty, Keechoo Choi, and Helai Huang. 2015. “Multi-Level Hot Zone Identification for Pedestrian Safety.” *Accident Analysis and Prevention* 76 (March): 64–73. https://doi.org/10.1016/j.aap.2015.01.006.

Lemaire, Jean, Sojung Carol Park, and Kili C. Wang. 2015. “THE USE OF ANNUAL MILEAGE AS A RATING VARIABLE.” *ASTIN Bulletin* 46 (1): 39–69. https://doi.org/10.1017/asb.2015.25.

Litman, Todd. 2005. “Pay-as-You-Drive Pricing and Insurance Regulatory Objectives.” *Journal of Insurance Regulation* 23 (3): 35.

Litman, Todd. 2011. “Distance-Based Vehicle Insurance Feasibility, Costs and Benefits: Comprehensive

Technical Report," June. https://trid.trb.org/view/1549618.

Liu, Dan, Elizabeth Toman, Zane Fuller, Gang Chen, Alexis Londo, Xuesong Zhang, and Kaiguang Zhao. 2018. "Integration of Historical Map and Aerial Imagery to Characterize Long-Term Land-Use Change and Landscape Dynamics: An Object-Based Analysis via Random Forests." *Ecological Indicators* 95 (December): 595–605. https://doi.org/10.1016/j.ecolind.2018.08.004.

Lym, Youngbin, and Zhenhua Chen. 2020. "Does Space Influence on the Frequency and Severity of the Distraction-Affected Vehicle Crashes? An Empirical Evidence from the Central Ohio." *Accident Analysis and Prevention* 144 (September): 105606. https://doi.org/10.1016/j.aap.2020.105606.

Ma, Yu Luen, Xiaoyu Zhu, Xianbiao Hu, and Yi Chang Chiu. 2018. "The Use of Context-Sensitive Insurance Telematics Data in Auto Insurance Rate Making." *Transportation Research Part A: Policy and Practice* 113 (July): 243–58. https://doi.org/10.1016/j.tra.2018.04.013.

Malyshkina, Nataliya V., Fred L. Mannering, and Andrew P. Tarko. 2009. "Markov Switching Negative Binomial Models: An Application to Vehicle Accident Frequencies." *Accident Analysis and Prevention* 41 (2): 217–26. https://doi.org/10.1016/j.aap.2008.11.001.

Mayou, R., B. Bryant, and R. Duthie. 1993. "Psychiatric Consequences of Road Traffic Accidents." *British Medical Journal* 307 (6905): 647–51. https://doi.org/10.1136/bmj.307.6905.647.

Musicant, Oren, Hillel Bar-Gera, and Edna Schechtman. 2010. "Electronic Records of Undesirable Driving Events." *Transportation Research Part F: Traffic Psychology and Behaviour* 13 (2): 71–79. https://doi.org/10.1016/j.trf.2009.11.001.

Neumann, Wiebke, Göran Ericsson, Holger Dettki, Nils Bunnefeld, Nicholas S. Keuler, David P. Helmers, and Volker C. Radeloff. 2012. "Difference in Spatiotemporal Patterns of Wildlife Road-Crossings and Wildlife-Vehicle Collisions." *Biological Conservation* 145 (1): 70–78. https://doi.org/10.1016/j.biocon.2011.10.011.

Ng, Kwok Suen, Wing Tat Hung, and Wing Gun Wong. 2002. "An Algorithm for Assessing the Risk of Traffic Accident." *Journal of Safety Research* 33 (3): 387–410. https://doi.org/10.1016/S0022-4375(02)00033-6.

Oehmcke, Stefan, Oliver Zielinski, and Oliver Kramer. 2018. "Input Quality Aware Convolutional LSTM Networks for Virtual Marine Sensors." *Neurocomputing* 275 (January): 2603–15. https://doi.org/10.1016/j.neucom.2017.11.027.

Paefgen, Johannes, Thorsten Staake, and Elgar Fleisch. 2014. "Multivariate Exposure Modeling of Accident Risk: Insights from Pay-as-You-Drive Insurance Data." *Transportation Research Part A: Policy and Practice*. https://doi.org/10.1016/j.tra.2013.11.010.

Paefgen, Johannes, Thorsten Staake, and Frédéric Thiesse. 2013. "Evaluation and Aggregation of Pay-as-You-Drive Insurance Rate Factors: A Classification Analysis Approach." *Decision Support Systems* 56 (1): 192–201. https://doi.org/10.1016/j.dss.2013.06.001.

Parsa, Amir Bahador, Ali Movahedi, Homa Taghipour, Sybil Derrible, and Abolfazl (Kouros) Mohammadian. 2020. "Toward Safer Highways, Application of XGBoost and SHAP for Real-Time Accident Detection and Feature Analysis." *Accident Analysis and Prevention* 136 (March): 105405. https://doi.org/10.1016/j.aap.2019.105405.

Peng, Yichuan, Yuming Jiang, Jian Lu, and Yajie Zou. 2018. "Examining the Effect of Adverse Weather on Road Transportation Using Weather and Traffic Sensors." Edited by Sergio A. Useche. *PLOS ONE* 13 (10): e0205409. https://doi.org/10.1371/journal.pone.0205409.

Pesantez-Narvaez, Jessica, Montserrat Guillen, and Manuela Alcañiz. 2019. "Predicting Motor Insurance Claims Using Telematics Data—XGboost versus Logistic Regression." *Risks* 7 (2).

https://doi.org/10.3390/risks7020070.

Sarkar, Mainak, and Arnaud De Bruyn. 2021. "LSTM Response Models for Direct Marketing Analytics: Replacing Feature Engineering with Deep Learning." *Journal of Interactive Marketing* 53 (February): 80–95. https://doi.org/10.1016/j.intmar.2020.07.002.

Schlögl, Matthias. 2020. "A Multivariate Analysis of Environmental Effects on Road Accident Occurrence Using a Balanced Bagging Approach." *Accident Analysis and Prevention*. https://doi.org/10.1016/j.aap.2019.105398.

Siła-Nowicka, Katarzyna, Jan Vandrol, Taylor Oshan, Jed A. Long, Urška Demšar, and A. Stewart Fotheringham. 2016. "Analysis of Human Mobility Patterns from GPS Trajectories and Contextual Information." *International Journal of Geographical Information Science* 30 (5): 881–906. https://doi.org/10.1080/13658816.2015.1100731.

Stipancic, Joshua, Luis Miranda-Moreno, and Nicolas Saunier. 2018. "Vehicle Manoeuvers as Surrogate Safety Measures: Extracting Data from the Gps-Enabled Smartphones of Regular Drivers." *Accident Analysis and Prevention* 115 (June): 160–69. https://doi.org/10.1016/j.aap.2018.03.005.

Strobl, Carolin, Anne Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution." *BMC Bioinformatics* 8 (1): 25. https://doi.org/10.1186/1471-2105-8-25.

Swiss Federal Office of Statistics. 2018. "Strassenverkehrsunfälle | Bundesamt Für Statistik." Accessed January 5, 2021. https://www.bfs.admin.ch/bfs/de/home/statistiken/mobilitaet-verkehr/unfaelle-umweltauswirkungen/verkehrsunfaelle/strassenverkehr.html.

Tselentis, Dimitrios I., George Yannis, and Eleni I. Vlahogianni. 2016. "Innovative Insurance Schemes: Pay as/How You Drive." In *Transportation Research Procedia*. https://doi.org/10.1016/j.trpro.2016.05.088.

Tulensalo, Jarkko, Janne Seppänen, and Alexander Ilin. 2020. "An LSTM Model for Power Grid Loss Prediction." *Electric Power Systems Research* 189 (December): 106823. https://doi.org/10.1016/j.epsr.2020.106823.

Verbelen, Roel, Katrien Antonio, and Gerda Claeskens. 2018. "Unravelling the Predictive Power of Telematics Data in Car Insurance Pricing." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67 (5): 1275–1304. https://doi.org/10.1111/rssc.12283.

Wan, Yue, Yuhang Li, Chunhong Liu, and Zhongqiu Li. 2020. "Is Traffic Accident Related to Air Pollution? A Case Report from an Island of Taihu Lake, China." *Atmospheric Pollution Research* 11 (5): 1028–33. https://doi.org/10.1016/j.apr.2020.02.018.

Wang, Jianqiang, Jian Wu, and Yang Li. 2015. "The Driving Safety Field Based on Driver-Vehicle-Road Interactions." *IEEE Transactions on Intelligent Transportation Systems* 16 (4): 2303–14. https://doi.org/10.1109/TITS.2015.2401837.

Wang, Ningcheng, Yufan Liu, Jinzi Wang, Xingjian Qian, Xizhi Zhao, Jianping Wu, Bin Wu, Shenjun Yao, and Lei Fang. 2019. "Investigating the Potential of Using POI and Nighttime Light Data to Map Urban Road Safety at the Micro-Level: A Case in Shanghai, China." *Sustainability* 11 (17): 4739. https://doi.org/10.3390/su11174739.

Wang, Yiyi, and Kara M. Kockelman. 2013. "A Poisson-Lognormal Conditional-Autoregressive Model for Multivariate Spatial Analysis of Pedestrian Crash Counts across Neighborhoods." *Accident Analysis and Prevention* 60 (November): 71–84. https://doi.org/10.1016/j.aap.2013.07.030.

Wier, Megan, June Weintraub, Elizabeth H. Humphreys, Edmund Seto, and Rajiv Bhatia. 2009. "An

Area-Level Model of Vehicle-Pedestrian Injury Collisions with Implications for Land Use and Transportation Planning." *Accident Analysis and Prevention* 41 (1): 137–45. https://doi.org/10.1016/j.aap.2008.10.001.

Winlaw, Manda, Stefan H. Steiner, R. Jock MacKay, and Allaa R. Hilal. 2019. "Using Telematics Data to Find Risky Driver Behaviour." *Accident Analysis and Prevention* 131 (October): 131–36. https://doi.org/10.1016/j.aap.2019.06.003.

Worldweatheronline. 2020. "Weather API | JSON | World Weather Online." Accessed January 5, 2021. https://www.worldweatheronline.com/developer/.

Wu, Yina, Mohamed Abdel-Aty, and Jaeyoung Lee. 2018. "Crash Risk Analysis during Fog Conditions Using Real-Time Traffic Data." *Accident Analysis and Prevention* 114 (May): 4–11. https://doi.org/10.1016/j.aap.2017.05.004.

Wyon, David P., Inger Wyon, and Fredrik Norin. 1996. "Effects of Moderate Heat Stress on Driver Vigilance in a Moving Vehicle." *Ergonomics* 39 (1): 61–75. https://doi.org/10.1080/00140139608964434.

XGBoost. 2020. "XGBoost Documentation." XGBoost Documentation. 2020. https://xgboost.readthedocs.io/en/latest/.

Yan, Chun, Xindong Wang, Xinhong Liu, Wei Liu, and Jiahui Liu. 2020. "Research on the UBI Car Insurance Rate Determination Model Based on the CNN-HVSVM Algorithm." *IEEE Access* 8: 160762–73. https://doi.org/10.1109/ACCESS.2020.3021062.

Yang, Bruce Zi, and Becky P.Y. Loo. 2016. "Land Use and Traffic Collisions: A Link-Attribute Analysis Using Empirical Bayes Method." *Accident Analysis and Prevention* 95 (October): 236–49. https://doi.org/10.1016/j.aap.2016.07.002.

Yannis, George, and Matthew G Karlaftis. 2011. "Weather Effects on Daily Traffic Accidents and Fatalities: A Time Series Count Data Approach."

Yao, Shenjun, Jinzi Wang, Lei Fang, and Jianping Wu. 2018. "Identification of Vehicle-Pedestrian Collision Hotspots at the Micro-Level Using Network Kernel Density Estimation and Random Forests: A Case Study in Shanghai, China." *Sustainability* 10 (12): 4762. https://doi.org/10.3390/su10124762.

Yu, Yong, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. "A Review of Recurrent Neural Networks: Lstm Cells and Network Architectures." *Neural Computation*. MIT Press Journals. https://doi.org/10.1162/neco_a_01199.

Zhang, Dahai, Liyang Qian, Baijin Mao, Can Huang, Bin Huang, and Yulin Si. 2018. "A Data-Driven Design for Fault Detection of Wind Turbines Using Random Forests and XGboost." *IEEE Access* 6 (March): 21020–31. https://doi.org/10.1109/ACCESS.2018.2818678.

Zhao, Jinghua, Dalin Zeng, Yujie Xiao, Liping Che, and Mengjiao Wang. 2020. "User Personality Prediction Based on Topic Preference and Sentiment Analysis Using LSTM Model." *Pattern Recognition Letters* 138 (October): 397–402. https://doi.org/10.1016/j.patrec.2020.07.035.

Appendix A: Figures

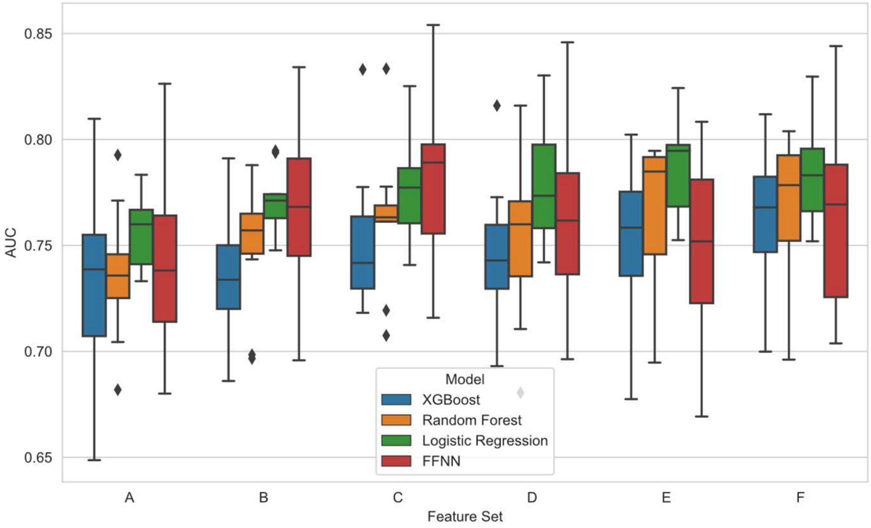

Figure 27: AUC values UK over all folds without the minimum mileage filter

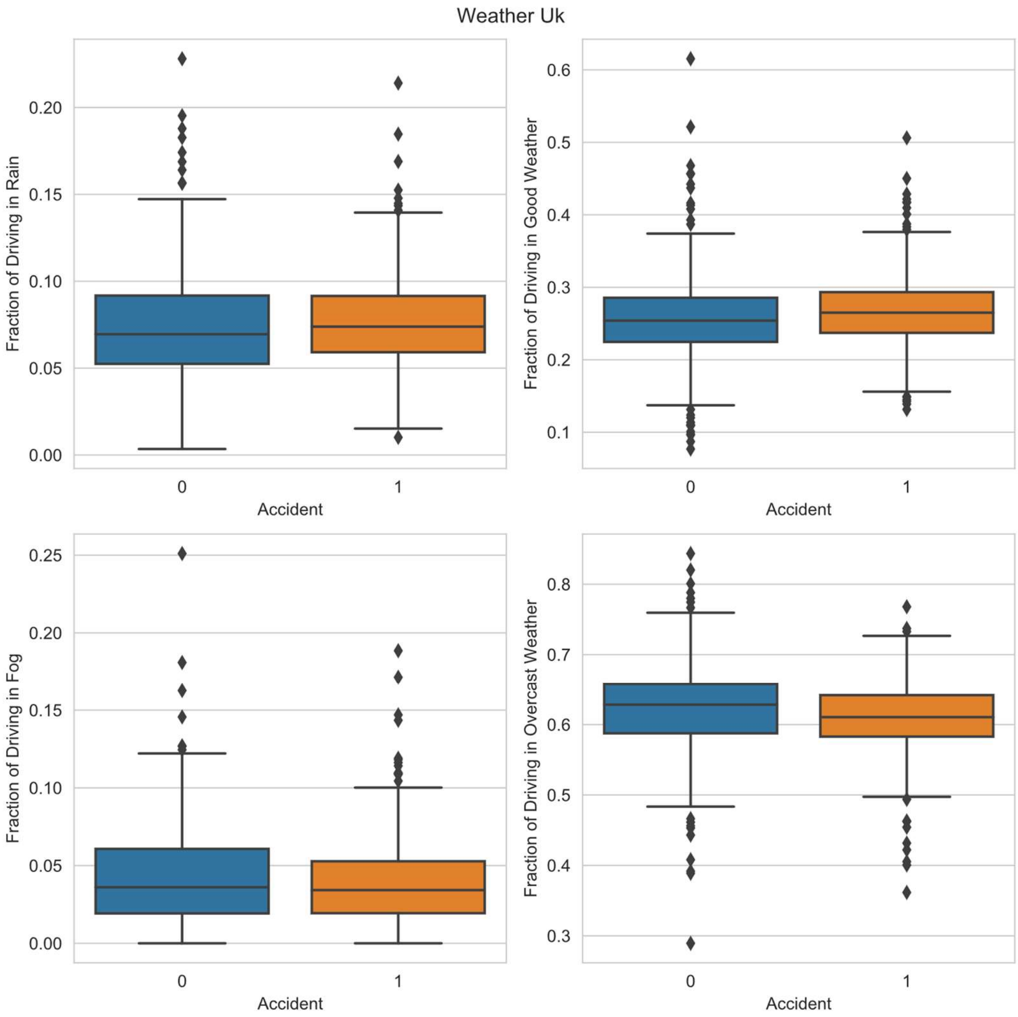

Figure 28: Boxplots of driving in different weather conditions, UK

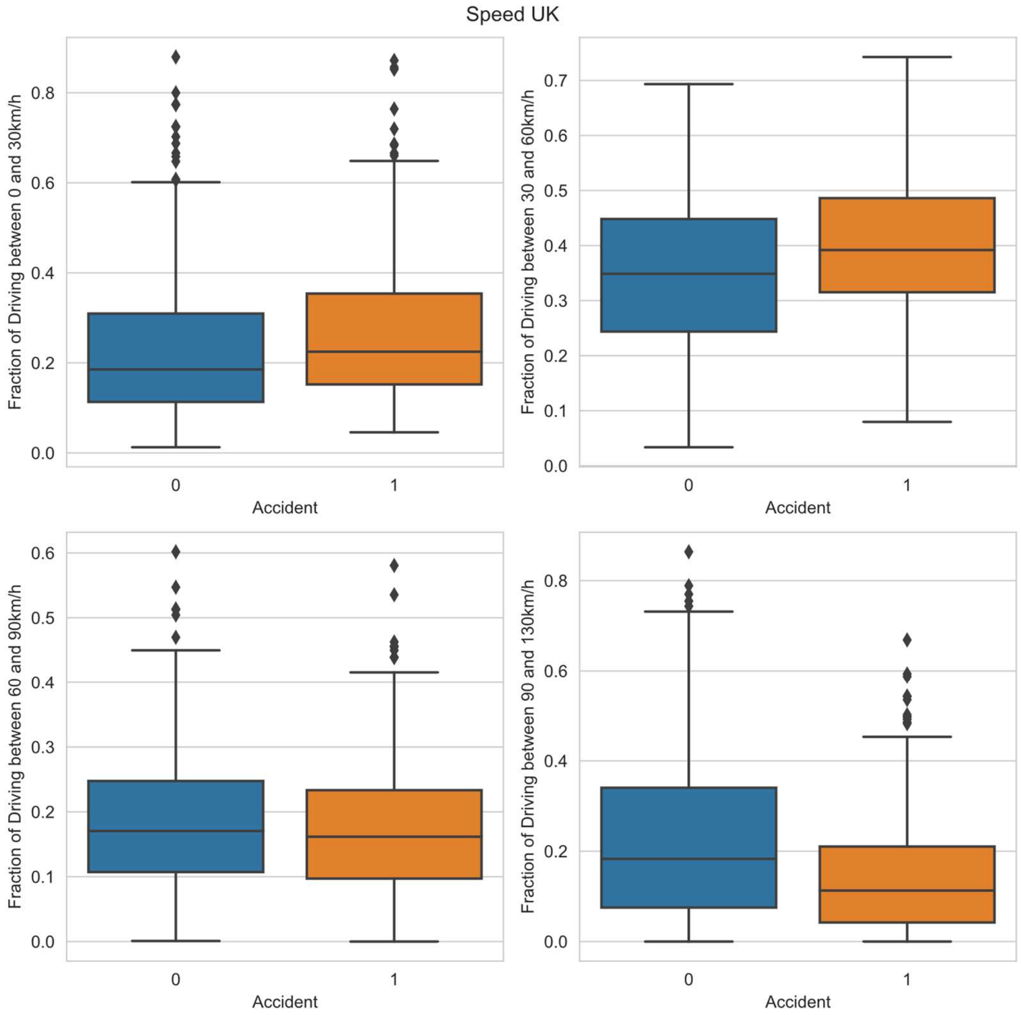

Figure 29: Boxplots of driving in different speed intervals, UK

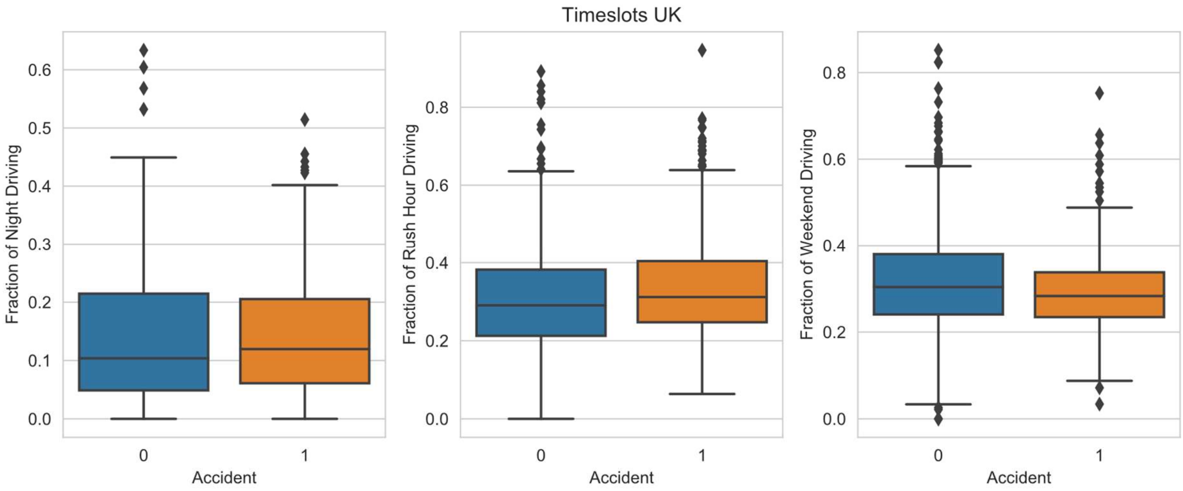

Figure 30: Boxplot of driving in different Timeslots, UK

Appendix B: Tables

Table 21: Performance metrics UK without the minimum distance cutoff

Model	A	B (baseline)	C	D	E	F
			AUC			
Logistic Regression	0.757	0.770	0.776	0.777	0.786	0.785
FFNN	0.740	0.768	0.781	0.762	0.747	0.762
Random Forest	0.737	0.750	0.762	0.752	0.764	0.764
XGBoost	0.731	0.735	0.752	0.745	0.751	0.760
			Accuracy			
Logistic Regression	0.670	0.686	0.695	0.696	0.704	0.719
FFNN	0.664	0.686	0.694	0.684	0.678	0.703
Random Forest	0.660	0.670	0.678	0.672	0.676	0.684
XGBoost	0.642	0.640	0.672	0.670	0.670	0.688
			F1-Score			
Logistic Regression	0.702	0.713	0.715	0.715	0.719	0.736
FFNN	0.706	0.726	0.731	0.724	0.713	0.739
Random Forest	0.682	0.688	0.698	0.693	0.695	0.705
XGBoost	0.666	0.658	0.685	0.691	0.684	0.702

Personal Declaration

I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in this thesis.

Horgen, 31.01.2021

Signature: L. Brühwiler

Livio Brühwiler