ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

Master Thesis MSc Economics & Business

Predicting Traffic Congestion by Machine Learning:

A case study of the Netherlands

Abstract

This research shows that traffic congestion prediction systems can be highly accurate, even without the use of traffic flow data. Out of a comparison between multiple machine learning techniques, an XGBoost model is found to predict the occurrence of traffic congestion most accurately with an accuracy on unseen test data of 90.39%. The ratio of highway kilometres to the number of inhabitants of a region and the gross regional product per capita are found to be the most important predictors of traffic congestion. Other important predictors are road accidents and commuter obligations.

Keywords: traffic congestion, machine learning, logistic regression, XGBoost, neural network

Student: Rogier de Bruin
Student ID: 451943

Supervisor: Drs. M.A. Rosch
Second assessor: Dr. M.J.A. Gerritse

Date: 13-07-2020

Acknowledgements

I would like to thank my supervisor Drs. Rosch, whose efforts pushed this thesis to a higher level. This thesis was written during challenging and unprecedented times. The only contact we have had throughout the entire thesis process was via online meetings. It is almost eerie to realise how normal communicating with others from behind your desk has become. Disregarding the minor inconveniences of malfunctioning webcams and drilling neighbours, the restriction to online communication has never felt as a hinderance to the process. The quality of the feedback was from an exceptional quality and extremely helpful. I especially value the fact that he would always allow me to incorporate my own ideas and would just provide some minor hints about the direction that I should take. Without his guidance and persistent help, the writing of this thesis would not have been possible.

Furthermore, I would like to thank all family and friends that have supported me in one way or another during this thesis process. Without any obligation to do so, I could always reach out for any sort of unconditional support. This has been a great help while writing this thesis and has definitely enhanced its quality.

Table of Contents

1. Introduction

Road transport has been an omnipresent part of society for centuries. Over time, the number of passengers and goods being transported by road has increased exponentially (Barker & Gerhold, 1993). Currently, road transport forms an essential part of the contemporary economy: in 2018, more than 70,000 kilometres of highway extended over Europe and approximately 173,970 million vehicle-kilometres have been driven over these roads for freight transport alone (European Commission, 2020). Despite the imminent advantages of road transport, some negative externalities arise from this type of transport as well. An oversupply of vehicles on a certain road section inherently results in traffic congestion. The societal costs of traffic congestion are immense. The direct societal costs of traffic congestion mainly consist of longer travel times, travel time unreliability and deviation behaviour. In 2018, the costs related to these factors in the Netherlands were 1.2 billion, 0.7 billion and 1.4 billion Euros, respectively (Ministry of Infrastructure and Water Management, 2019). Additionally, traffic congestion also results in indirect costs. For example, people arrive late at their work, which lowers their overall productivity. Indirect traffic congestion costs were estimated to be between 0 and 1.0 billion Euros in 2018 in the Netherlands (Ministry of Infrastructure and Water Management, 2019). In total, societal costs related to traffic congestion are expected to be between 3.3 and 4.3 billion Euros in the Netherlands in 2018, which equals approximately 0.5% of the Dutch GDP. Throughout 2019, traffic congestion levels have reached record heights (Directorate-General for Public Works and Water Management (Rijkswaterstaat), 2020). Although exact numbers are not available yet, higher congestion levels are likely to result in even higher societal costs. The trend of growing congestion levels should be reversed to avoid continuous increases in societal costs related to traffic congestion.

Various initiatives have been proposed to mitigate the vast societal costs of traffic congestion. Simply constructing more roads is often not considered to be the proper solution; this policy is hard to execute for political, economic and environmental reasons and is said to even result in more congestion due to an increased demand for vehicle travel (Strickland & Berman, 1995). Therefore, congestion mitigation approaches have become more aimed at travel demand management (TDM) (Kitamura, Fujii, & Pas, 1997). A mixture of coercive and noncoercive TDM measures have proven to be the most efficient to reduce traffic congestion (Gärling & Schuitema, 2007). Coercive measures are, for example, creating car-free zones or road pricing. Noncoercive TDM policies are more aimed at nudging travellers to change their usual behaviour, such as providing up-to-date traffic information or rewarding public transport use. Regarding traffic

information, a distinction can be made between en-route and pre-route traffic information. En-route traffic information is provided while someone is already on its way, whereas pre-route information is provided before someone has started a trip. Multiple studies have shown that en-route traffic information can incentivise car users to deviate from their originally planned route or to use a different transport mode (e.g., Abdel-Aty, Kitamura and Jovanis (1997), and Jou, Lam, Liu and Chen (2005)). Pre-route traffic information can create the same incentive, but it also offers the choice to postpone or advance a journey (Polydoropoulou, Ben-Akiva, & Kaysi, 1994). Both traffic information types have in common that the quality of the information is an important determinant of the compliance behaviour of commuters (Chen, Srinivasan, & Mahmassani, 1999). Travellers are less probable to deviate from their original route when the provided traffic information is perceived as inaccurate or unreliable. The combination of traffic information being an effective TDM measure and the impact of the quality of the information systems demonstrates the need for a highly accurate and reliable traffic information system that can provide both pre-route and en-route traffic information.

Multiple companies or organisations already offer such information systems. In the Netherlands, the Royal Dutch Touring Club (ANWB) provides a daily forecast of traffic congestion levels during morning and evening rush hours. This forecast contains a country-level prediction of the traffic congestion severity and mentions a specific region in which traffic congestion is expected to be most severe. An example of such a forecast could be: traffic congestion severity is expected to be moderate during the morning rush hour with a high probability of traffic jams around Rotterdam[1]. The predictions are based upon historic traffic flow data on Dutch highways, weather forecast, planned road works and specific events. The data is not publicly available and performance levels of the predictive models are not reported. Although the traffic congestion forecast by the ANWB provides some insights into congestion levels on Dutch highways, the predictions are rather undetailed. The forecast only provides information about expected congestion levels on a country level and the forecast is only made for rush hour periods. A result of the lack of detail in the traffic forecast is that it does not allow road users to deviate from their original route, as it does not report traffic jam probabilities per road section. This research aims to improve the traffic forecast system in the Netherlands by providing a more detailed traffic flow forecast. A model is developed that predicts the probability of a traffic jam occurring for every hour of the day and every road section independently. The accuracy and reliability of this

[1] See https://www.anwb.nl/verkeer/nederland/verkeersverwachting for a real-life example of the ANWB traffic congestion forecast (in Dutch).

model are crucial factors to consider, as these are found to strongly affect the compliance behaviour of road users to deviate to a less congested route.

Existing literature on traffic information prediction systems mainly focuses on how congestion can be predicted by means of historic traffic flow data (e.g., Elfar, Talebpour and Mahmassani (2018), Min and Wynter (2011), and Zhang, Liu, Yang, Wei and Dong (2013)). The highest observed performance is found in research conducted by Elfar, Talebpour and Mahmassani (2018), with an accuracy of 93% and 96% of the congested observations being correctly labelled. This model predicts congestion by means of a logistic regression with detailed information about cars' trajectories (e.g., location, speed, and acceleration) on a highway stretch of 0.6 kilometres in California as predictors. However, in a recent overview of machine learning applications in traffic congestion prediction Akhtar and Maridpour (2021) identify some problems that arise when using traffic flow data. This data type can be gathered in two manners: by using sensors that collect spatiotemporal traffic data, or by using a GPS that collects traffic data continuously. A major disadvantage of the first method is that these sensors are costly and fail relatively often. As such, the number of road networks that are fully equipped with traffic sensors remains relatively low. If this data type is used in research, it is most commonly data about one specific highway stretch in California, of which the research by Elfar, Talebpour and Mahmassani (2018) is an example. Although the ANWB uses historic traffic flow data measured by sensors on Dutch highways in their traffic forecast, the coverage of the Dutch road network by these sensors is sparse. The second data type that is frequently used predominantly suffers from inaccuracies in the GPS software. This could result in significant fluctuations in the GPS data. Moreover, tracking the position of vehicle users by GPS comes with some serious privacy issues that disallow the usage of such tracking systems on a large scale. Lastly, it is impossible to use the model that is built with data for one city to predict traffic congestion in other cities, due to the unique characteristics of the GPS coordinates. This data collection method is currently only used on a large scale in the city of Beijing, where all taxis have been equipped with tracking software. The shortcomings of both data collection methods limit the geographical scope of the studies investigating the possibilities to predict traffic congestion.

Considering the disadvantages of using historic traffic flow data, this research tries to overcome these problems by avoiding the use of traffic flow data. Instead, the prediction of whether or not traffic congestion will arise on highways is solemnly based on external variables related to the supply and demand of traffic. Data about traffic congestion occurring on highways in the Netherlands between 2015 and early 2020 is gathered to construct a traffic congestion prediction

model. Sudden disruptions are incorporated in the model by determining so-called accident hotspots, which are road sections that have an increased probability of such disruptions. Models without historic traffic flow data can serve as complementary or even substitutionary models if their predictive performance is found to be comparable to the accuracy of models that do use historic traffic flow data. This results in the following research question:

To what extent can traffic jam occurrences be predicted without using traffic flow data?

The structure of the remainder of this research is as follows. Section 2 discusses the existing literature about traffic congestion prediction considered from a supply and demand perspective. The section is finalised by a conceptual framework covering the determinants of traffic congestion according to the existing literature. Then, section 3 presents the data that is used for this research, and section 4 explains the methods that are applied. Subsequently, section 5 shortly presents the results that are obtained. Section 6 provides an intensive discussion about the results that are found and their implications. Lastly, section 7 contains a conclusion of this research including an answer to the aforementioned research question.

2. Theoretical framework

The forming of traffic congestion can be considered as a typical non-equilibrium between supply and demand (Sugiyama, et al., 2008). The gap between the supply and demand for a piece of road is a classic example of market failure caused by road space being a public good (Rothenberg, 1970). As long as supply, the maximum capacity of a road section, is larger than demand, the number of vehicles willing to use a road section, traffic jams do not occur. The supply and demand of a road section can be visualised by means of a fundamental diagram of traffic flow, which is shown in Figure 1. Congestion levels on a road are generally measured in traffic flow and traffic density. Traffic flow is the number of vehicles that pass through a road section during a given time frame. Traffic density is the number of vehicles per kilometre on a road section. A low traffic flow in combination with a low traffic density indicates that traffic is flowing freely. However, as traffic flow starts to increase, traffic density will initially increase as well but traffic flow is still unimpeded. At the point that supply and demand are exactly equal, traffic flow is at its highest; the road section cannot handle more vehicles per hour. In this situation, road users already have to adjust their speed to the increased traffic density. If demand maintains increasing regardless, actual congestion starts to form. Traffic flow will decrease as vehicles are no longer able to achieve their original average speed, and traffic density keeps increasing. In

the unlikely case that the congestion has risen to such a severe level that not a single car is passing through the road section, traffic flow is equal to zero.

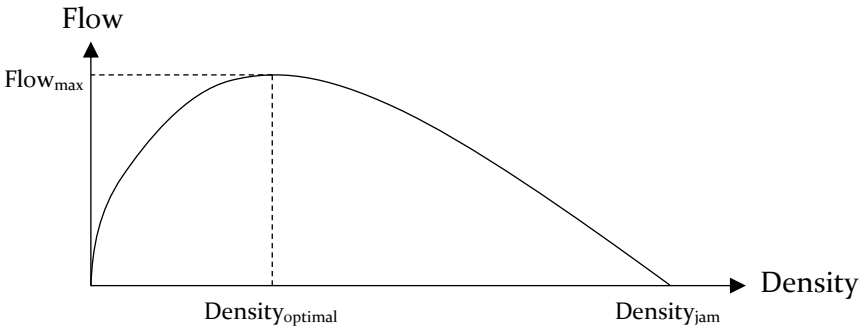


Figure 1: Fundamental diagram of traffic flow

Considering the imminent effect of supply and demand on congestion, these two factors are likely to play an important role in predicting the occurrence of a traffic jam. A problem that immediately arises, however, is that both supply and demand are hardly measurable on a large scale. Supply and demand data for every road section in a country on an hourly basis would be required to predict the occurrence of traffic congestion with enough detail. Therefore, the determinants of supply and demand will be reviewed and used to predict traffic congestion instead.

2.1. Road capacity

Supply can be defined as the maximum capacity of a road section. Changes in this supply can be divided into two types: long term and short term changes. Long term changes in road space supply are relatively seldom. Political, economic and environmental reasons have made the construction of new highways or new highway lanes less common (Strickland & Berman, 1995). Short term fluctuations in supply, however, are more frequently occurring. A short term change in supply means that the capacity of a road section is temporarily reduced, but that the road section's capacity will return to its original level. The main causes of such short term changes are road accidents, weather conditions and roadworks (Snelder, Van Zuylen, & Immers, 2012).

2.1.1. Constructing additional highway lanes

Expanding the number of lanes on a highway is undoubtedly the most straightforward way to increase the capacity of a road section. Creating more lanes indisputably means that more vehicles can pass through. One might expect that this would result in less traffic congestion accordingly. Research conducted on the relationship between travel demand and road capacity has shown that this is not the case though. Multiple efforts to estimate the long-run elasticity

between the length of highway lanes and vehicle kilometres travelled (VKT) have resulted in a range between 0.7 and 1.03 (Duranton & Turner, 2011; Hansen & Huang, 1997; Noland, 2001). Considering the high value of this elasticity, traffic density seems to be only marginally affected in the long term by adding more lanes to existing highways. Given that the traffic density is unchanged by an increase in the length of highway lanes, constructing more highway lanes seemingly does not influence the chance of a traffic jam occurring.

2.1.2. Road accidents

The occurrence of road accidents often results in a short term reduction in road capacity, for example due to one or multiple highway lanes being blocked, or having to be closed temporarily. With demand being unaffected, congestion can possibly emerge out of a road accident. Approximately 12% to 15% of the traffic jams in the Netherlands is caused by road accidents (Marchesini & Weijermars, 2010). It is not easy, though, to predict the happening of a road accident. The foremost predictors of a road accident taking place are age, gender, time of the day, weather, and a driver's physical condition (e.g. Åkerstedt and Kecklund (2001), and Park, Kim and Ha (2016)). How could one, for example, obtain future percentages of drivers' gender at a certain road section? It is found, however, that so-called road accident hotspots exist (e.g., Anderson (2009), Cheng and Washington (2005), Dong, Huang, Lee, Gao and Abdel-Aty (2016), and Montella (2010)). These are road sections where traffic accidents occur more frequently compared to other road sections. This indicates that the location of an accident happening is not random but dependent on the local characteristics of the road. The existence of accident hotspots also suggests that the likelihood of an accident occurring can be predicted by the number of accidents that have happened in the past. In addition to the accident hotspots, the sleep deprivation caused by the spring transition into daylight saving time results in an increased car crash risk (Smith, 2016). Accidents are found to be more likely to occur the day after daylight saving time has started.

2.1.3. Weather conditions

Local weather conditions can have a severe short term impact on the capacity of a road section. A clear negative relationship between road capacity and rainfall exists: the capacity of a road decreases by 4% to 10% during light rain conditions and up to 30% during severe rainfall (Chung, Ohtani, Warita, Kuwahara, & Morita, 2006; Smith, Byrne, Copperman, Hennessy, & Goodall, 2004). This is mainly a result of drivers adapting their speed to the weather conditions. Traffic flow is also negatively affected by snow, hail, poor visibility and high wind speeds (Kwon, Liping,

& Jiang, 2013). The effect of snow is particularly strong: road capacity decreases by 27% during snowfall (Agarwal, Maze, & Souleyrette, 2005). In the extreme scenario of a road blockage, capacity can even become zero. Eventually, the reduced road capacity that is caused by adverse weather conditions appears to result in a higher probability of traffic congestion occurring (Van Stralen, Calvert, & Molin, 2015). It is recognised, however, that an individual's travel behaviour is influenced by weather conditions as well, meaning that the probability of traffic congestion occurring could also be affected by changes in highway travel demand[2].

2.1.4. Road works

Although the general aim of road works is to enhance either safety or road capacity in the long term (Archondo-Callao, 2008), road works can temporarily lower the capacity of a road (Kerner, 2009). Possible explanations for this effect are lane closures or a reduction in the maximum speed limit to below-average free-flow speed required to carry out the maintenance (Yousif, 2002). The following reduction in road capacity makes traffic congestion more probable to arise. Approximately 4% of traffic jams in the Netherlands are caused by road works (Marchesini & Weijermars, 2010). A distinction can be made between planned and unplanned road works (Stopher & Stanley, 2014). Planned road works, such as road resurfacing, are announced beforehand and can be taken into account when predicting whether or not congestion will occur. Unplanned maintenance, on the other hand, is harder to incorporate in predictions. This is due to this type of road works resulting from unexpected events, such as the guard rail being damaged by a road accident or strong winds rupturing road information signs.

2.2. Travel demand

Travel demand can be defined as the number of people that want to travel from one location to another. A person's desire to travel is based on a multitude of individual choices, such as trip purpose, frequency, time of the day, destination, and mode of travel (McFadden, 1974). As these choices can change on a day-to-day basis, fluctuations in travel demand are inevitable. Managing the travel demand of individuals is considered an important instrument against traffic congestion (Kitamura, Fujii, & Pas, 1997). The most important determinants of highway travel demand are user costs (Levinson & Gillen, 1998), weather conditions (Maze, Agarwal, &

[2] The effect of weather conditions on the demand side of traffic congestion is discussed in detail in Section 2.2.2.

Burchett, 2006), commuter obligations (McKenzie & Rapino, 2011) and economic growth (Deming, 1975).

2.2.1. User costs

A general conception in economics is that the demand for a product is negatively related to its price. This holds true for travelling as well: car travel demand is negatively related to car user costs (Levinson & Gillen, 1998). People tend to lower their VKT by car as the price per kilometre rises. Examples of such costs are operating costs, maintenance costs, insurance costs, and licensing costs. It is found that especially fuel prices and toll costs have a profound impact on one's highway travel demand (De Jong & Gunn, 2001). These costs are more visible than other costs, making car users more aware of price levels. Although parking costs, for example, share the same characteristics, this is less relevant for highway travel demand. Considering that toll costs do not apply in the Netherlands, the main user cost aspect is limited to fuel costs.

Car user costs are not the only user costs that influence car travel demand, though. Changes in the user costs of substitutes of a car can affect car travel demand as well. Most notably, people tend to travel by car less frequently if public transport prices decrease, and vice versa (Redman, Friman, Gärling, & Hartig, 2013). Moreover, although not a typical monetary cost, the effort that is required to use a certain transport mode can be considered a user cost as well. It is found that people who live proximal to a public transport utility travel by public transport more often than people that live further away from public transport utilities (Badland, Garrett, & Schofield, 2010). In a similar manner, people tend to travel by car more frequently in areas with a high road density, as the road infrastructure provides more convenience than in regions with a lower road density (Tseng, et al., 2018).

2.2.2. Weather conditions

The effect of weather conditions on the capacity of highways has been discussed in an earlier paragraph. The impact of weather conditions is not limited to the supply-side of traffic congestion, though. For example, people tend to cancel leisure trips during rainy, snowy, or stormy conditions, resulting in less travel demand (Cools, Moons, Creemers, & Wets, 2010). Traffic volumes are found to be approximately 5% lower during rainfall, and 7% to 80% lower during snowstorms, dependent on its severity (Maze, Agarwal, & Burchett, 2006). The effect on highway travel demand is highly dependent on the type of adverse weather conditions. It is found that fewer people travel by public transport during extremely warm, extremely cold, or

rainy weather (Miao, Welch, & Sriraj, 2019). Under these circumstances, people prefer to travel by car, resulting in increased travel demand. Similarly, precipitation and fog result in people tending to travel by motorised vehicle instead of by bicycle or foot (Hranac, Sterzin, Krechmer, Rakha, & Farzaneh, 2006).

2.2.3. Commuter obligations

The obligatory journeys of people travelling to their work make up for a large part of travel demand, especially during weekdays (McKenzie & Rapino, 2011). Considering that most people start and stop working around the same time, a steep increase in travel demand can be seen during these moments. The peak rush hours in the Netherlands during weekdays are from 7 am to 9 am and from 4 pm to 6 pm (Oakil, Nijland, & Dijst, 2016). There are no peak rush hours during weekends and holidays since significantly fewer people have to travel to work (Ben-Elia & Ettema, 2011).

The COVID-19 pandemic has substantially changed the commuting behaviour of employees (e.g., Brynjolfsson et al. (2020), and Shibayama, Sandholzer, Laa and Brezina (2021)). There is a clear negative relationship between more stringent COVID-19 measures and workplace visits (Hale, et al., 2021). People are urged to practise social distancing and to stay at home, which decreases travel demand (De Vos, 2020). In April 2020, approximately 50% of the Dutch employees solemnly worked from home (Felstead & Reuschke, 2020). Therefore, months in which the traffic flow in the Netherlands was severely distorted by the COVID-19 pandemic are excluded from the analyses in this research.

2.2.4. Economic growth

The final predictor of highway travel demand that is discussed is economic growth. Economic growth is strongly correlated to economic production and, subsequently, results in more transport by truck (Deming, 1975). As a result of this, economic growth is positively related to the number of truck journeys. This effect is found to be particularly strong in the surroundings of cities that are highly economically active (Sweet, 2011). There is no evidence found that the number of car journeys also increases as a result of economic growth.

2.3. Conceptual framework

A conceptual framework in Figure 2 summarises the findings in the existing literature. Four factors affect the supply of road space: constructing more highway kilometres, road accidents, weather conditions, and road works. Out of these four, only the construction of more highway kilometres is found to be unrelated to traffic congestion occurring. This is due to more highway kilometres being almost perfectly correlated to more VKT. Road accidents and road works both diminish the capacity of a road section and, thus, positively relate to the probability of traffic jams arising. Factors influencing the happening of a road accident are a driver's age, gender, and physical condition, weather conditions, accident hotspots, and daylight saving time. Considering these predictors, only the latter two are suitable for predicting the probability of an accident taking place in the future. The effect of weather conditions on traffic congestion is dependent on the type of weather conditions. In general, it can be stated that traffic congestion is more likely to occur during cold and wet weather.

Four factors influence highway travel demand: user costs, weather conditions, commuter obligations, and economic growth. These four are all related to the probability of traffic congestion occurring as well. Higher car user costs per kilometre result in lower highway travel demand and, in this manner, a lower probability of traffic congestion arising. The main user costs in the Netherlands are fuel costs. Since user costs of substitutes of the car can also affect the highway travel demand, public transport user costs should also be taken into consideration. Additionally, the effort that has to be undertaken to use a transport mode can be seen as a non-monetary user cost. Therefore, the availability of both highways as public transport influences demand. The effect of weather conditions on highway travel demand is slightly ambiguous and highly dependent on the type of weather. For example, leisure trips are usually postponed or cancelled during wet weather conditions and people are more inclined to use motorised vehicles for their journeys during extreme temperatures or when it is raining. It appears, however, that the former effect is stronger, meaning that highway travel demand is lower during cold or wet weather conditions. Commuter obligations have a particularly strong effect on the probability of traffic congestion occurring. As people usually start and stop working at approximately the same time, sharp increases in highway travel demand are found during peak commuting hours. Factors influencing the number of commuters travelling to work are the time of the day, the day of the week, public holidays, and the COVID-19 pandemic. Lastly, economic growth is found to increase both highway travel demand and the probability of traffic congestion arising as a result

of an increased level of transport by truck. There is no evidence suggesting a relationship between economic growth and increased levels of transport by car.

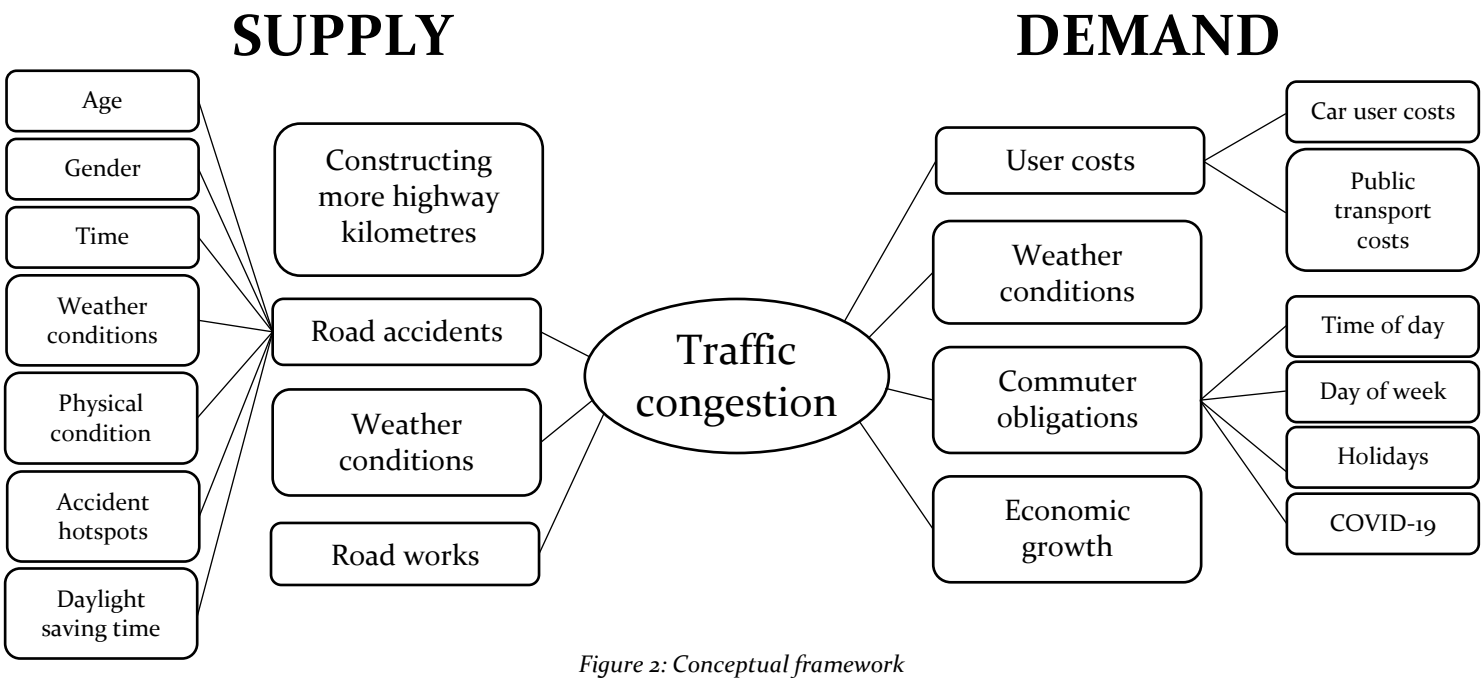


Figure 2: Conceptual framework

3. Data

The data that is used for this research is gathered from multiple sources. This section discusses which source the data originates from, and how the distinct data sources have been combined. In the end, some required data extrapolations are explained and descriptive statistics of the variables are provided.

3.1. Dependent variable

The primary data source is the Network Management Information System that is maintained by the Directorate-General for Public Works and Water Management (Rijkswaterstaat). A list containing all traffic jams on Dutch roads between January 1st 2015 and February 29th 2020 has been collected from this source. Traffic jams are added to this list as soon as they are reported by the Dutch Traffic Control Centre (VCNL). VCNL manually labels traffic jams arising on Dutch highways as part of their traffic information service. This specific period is chosen as traffic jam data was not administered before 2015 and various COVID-19 measures distorted the traffic behaviour in the Netherlands from March 2020 onwards. In total, 747,665 traffic jams were reported on Dutch highways within this period.

A data set is created dividing all days between January 1st 2015 and February 29th 2020 into 24 one-hour periods, and all highways in the Netherlands into sections of 20 kilometres. Only the last road section of a highway is shorter; the length of this section is the remaining road length when it is no longer possible to create a new road section of 20 kilometres. Lanes travelling in opposite directions are considered separate road sections. This results in a data set of 13,850,784 observations. A binary variable is created that states whether or not traffic congestion has occurred at any time or place within the specific time frame and road section of an observation. It is found that traffic congestion occurs for 3.164% of all observations.

3.2. Independent variables

Firstly, detailed daily weather data from the same period as the traffic jam data is assembled from the Royal Netherlands Meteorological Institute (KNMI), including data about wind speeds, temperatures, sunshine hours, visibility, overcast, precipitation and humidity. Considering that weather conditions can vary within a country, the data of weather stations in five different locations (i.e., Beek, De Bilt, De Kooy, Eelde, and Vlissingen) distributed over various regions in the Netherlands is gathered. To combine the weather conditions data with the existing traffic jam data set, it is assumed that the weather conditions for a road section are similar to the nearest weather station. The distance is measured from the middle of the section to the weather station as the crow flies.

Secondly, data about road works on Dutch highways is gathered from the National Road Traffic Data Portal (NDW). All road works in the Netherlands for the period between January 2015 and February 2020 are assembled from this source. In total, 381,907 instances of road works are reported on highways within this period. The road works data only provides geographic coordinates of the road works' locations instead of human-readable locations. Therefore, the coordinates of the road works' locations are transformed into an actual address or location by reverse geocoding. This is accomplished by means of web scraping from Bing Maps through an application programming interface (API). Ultimately, a binary variable is created labelling whether or not road works were in operation based upon the gathered locations.

Thirdly, data about user costs of both cars and public transport is collected from the Dutch Central Agency for Statistics (CBS). Three different car user costs measures are gathered: total user costs per kilometre, fuel costs, and road availability. When considering public transport user costs, only total user costs per kilometre and distance to public transport facilities are

collected. The total user costs per kilometre are measured monthly as a consumer price index with 2009 being the base year. Fuel costs are available daily for Euro 95 gasoline, diesel, and LPG and are measured in euro per litre. Road availability and distance to public transport facilities is measured on a provincial level. Road availability is calculated as the number of highway kilometres in a province per 100,000 inhabitants. The distance to public transport facilities is defined as the average distance in kilometres of all houses in a province to the nearest train station. Road sections are assigned the value of the province that the middle of the road section is in.

Lastly, yearly data about the gross regional product (GRP) per province is retrieved from the Dutch Central Agency for Statistics (CBS) and data about holidays is collected from the Government of the Netherlands. Public and school holidays are accounted for independently. Public holidays include New Year's Day, Easter, King's Day, Liberation Day, Ascension Day, Pentecost, and Christmas. From the same source, the start date of daylight saving time has been gathered.

3.3. Data extrapolations

Data about some variables is not yet available for the most recent years. Total user costs data is only available until November 2018, and data about the GRP per capita until the end of 2019. Firstly, the remaining values of both car and public transport user costs are estimated by means of an autoregressive integrated moving average (ARIMA) model. This is a model type that predicts future values based upon its own values and forecast errors in the past. Based upon the autocorrelation function (ACF) and the partial autocorrelation function (PACF), the values of both car and public transport user costs are forecasted by an ARIMA(1,1,2) model. This means that one lagged forecast error and two lags of the predictor in combination with a first-order difference of the variable are used to forecast future values. The ACF, PACF, and actual and predicted values of the ARIMA model are shown in Appendix A. Secondly, data about GRP per capita for 2020 have been extrapolated based upon time and province-level fixed effects. As ARIMA does not incorporate fixed effects into its forecasting, a different prediction method is used for this variable. A single linear regression model with time as independent variable and province-level fixed effects is applied instead. The coefficients of this model are displayed in Appendix B.

3.4. Descriptive statistics

Combining all variables results in a data set of 13,850,784 observations and 31 variables. Descriptive statistics of the continuous variables are shown in Table 1. Wind direction is the only categorical variable. The wind direction is west for 4,842,234 observations, south for 4,437,653 observations, east for 2,422,502 observations and north for 2,148,395 observations. More thorough descriptive statistics about differences in mean and standard deviation according to whether or not congestion arises for an observation are included in Appendix C.

Table 1: Descriptive statistics of continuous variables

	Mean	Std. dev.	Min.	Max.
Road conditions				
Congestion (1 = congestion, 0 = no congestion)	0.032	0.174	0.000	1.000
Road works (1 = road works, 0 = no road works)	0.285	0.442	0.000	1.000
Weather conditions				
Daily mean windspeed (in m/s)	4.521	2.349	0.700	18.000
Daily minimum hourly mean windspeed (in m/s)	2.332	1.889	0.000	16.000
Daily maximum hourly mean windspeed (in m/s)	6.763	3.060	1.000	24.000
Daily maximum wind gust (in m/s)	11.383	4.402	2.000	39.000
Daily mean temperature (in degrees Celsius)	10.910	6.053	-7.800	30.900
Daily minimum temperature (in degrees Celsius)	7.101	5.634	-9.700	23.200
Daily maximum temperature (in degrees Celsius)	14.582	7.026	-5.400	39.600
Daily sunshine duration (in hours)	5.152	4.238	0.000	15.700
Daily precipitation duration (in hours)	1.706	2.842	0.000	23.500
Daily precipitation (in mm)	2.161	4.371	0.000	49.700
Daily maximum hourly precipitation (in mm)	0.841	1.756	0.000	47.700
Daily minimum visibility (in km)	4.314	2.227	0.000	8.100
Daily maximum visibility (in km)	7.559	0.809	0.100	8.300
Daily average overcast (on a scale from 0 to 9)	5.920	2.184	0.000	8.000
Daily average humidity (in %)	79.799	10.031	33.000	100.000
Daily minimum humidity (in %)	63.850	15.025	16.000	99.000
Daily maximum humidity (in %)	93.467	6.113	43.000	100.000
User costs				
Public transport user costs (index with base year 2009)	121.240	1.471	118.800	122.800
Car user costs (index with base year 2009)	118.373	2.778	112.300	123.400
Gasoline price (in euro/litre)	1.574	0.080	1.368	1.743
Diesel price (in euro/litre)	1.260	0.096	1.022	1.444
LPG price (in euro/litre)	0.630	0.049	0.529	0.753
Road availability (in highway kms per 100,000 inhabitants)	33.857	9.206	72.840	21.990
Distance to public transport (in km)	5.392	2.443	3.500	17.300
Extraordinary days				
School holiday (1 = school holiday, 0 = no school holiday)	0.304	0.460	0.000	1.000
Public holiday (1 = public holiday, 0 = no public holiday)	0.023	0.151	0.000	1.000
Daylight saving time start (1 = start, 0 = no start)	0.003	0.056	0.000	1.000
Economic growth				
GRP per capita (in thousands of euros)	42.158	0.801	43.490	62.005

Figure 3 shows the distribution of traffic congestion encounters over multiple time-related variables. As various COVID-19 measures have distorted the regular flow from March 2020

onwards, it should be noted that the upper left and lower left graphs are corrected for the absence of data about March until December 2020. As a result of the absence of these months, fewer observations with congestion would be encountered in 2020 and in the months from March until December. Therefore, the distribution of congestion over years is divided by the number of months for which data was available per year, and the distribution of congestion over months is divided by the number of years for which data was available per month. Figure 3 illustrates that the occurrence of congestion is highly dependent on time-related factors. The monthly average of congestion encounters per year was increasing until 2020. In the first months of 2020, the relative number of traffic jams per month was slightly lower than in 2019. Also, a substantially higher number of traffic jams is reported during weekdays than during weekends. Throughout the week, it is noticeable that less congestion has been reported on Wednesdays and Fridays. When looking at the months, there are relatively fewer congestion encounters during July and August. The opposite effect is shown for October and November, however, where a peak in congestion occurrences is visible. The distribution of congestion over the time of the day illustrates two clear peaks during which congestion arises more often. These are the typical rush hour peaks between 7 to 9 am and 4 to 6 pm. Reports of congestion during the night are seldom.

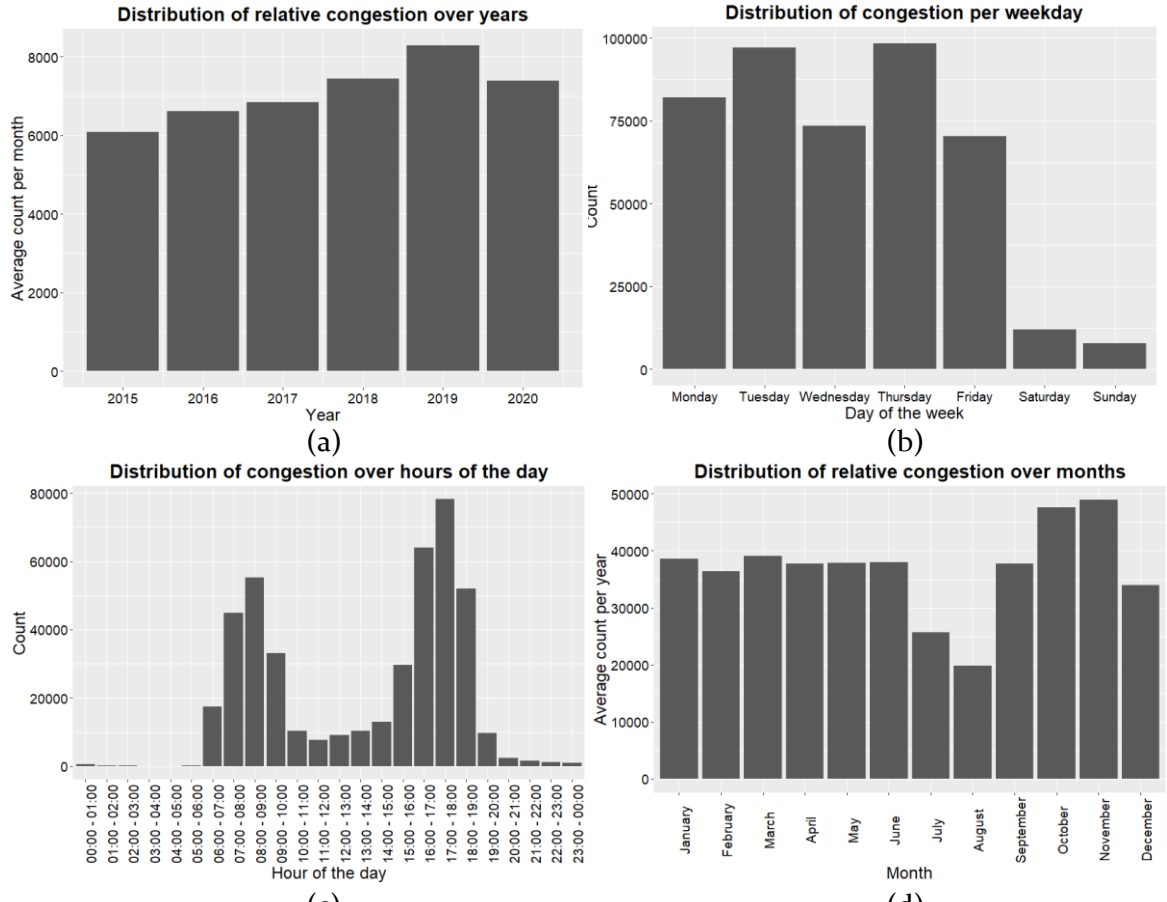


Figure 3: Distribution of traffic congestion encounters over time-related variables

3.5. Balancing

As previously mentioned, traffic congestion occurs for only 3.164% of the observations in the data set. This entails that the data set is highly unbalanced: the number of positive encounters is much lower than the number of negative encounters. A balanced data set would indicate that the proportion of congestion and free traffic flow is exactly equal. Balancing an unbalanced data set tends to result in improved predictive performance (Batista, Prati, & Monard, 2004). The original data set is balanced by randomly selecting a subset from the observations with negative encounters. The size of the subset is equal to the number of positive encounters in all observations. By combining all observations with congestion arising and the subset of the observations without congestion, a perfectly balanced data set consisting of 876,416 observations is created. A comparison of the descriptive statistics of the unbalanced and balanced data set is represented in Appendix C. It is found that the descriptive statistics of the balanced data set are similar to those of the unbalanced data set.

4. Methods

The aim of this research is to train a model that predicts whether or not traffic congestion will occur as accurately as possible. A common method to model a binary dependent variable is a logistic regression. This model type fits the coefficients of the independent variables to the data based upon the logistic function. Although this method is suitable for finding relationships between variables, it is less useful when the goal is to train a model with high predictive performance. This is due to the logistic regression being a high bias model: there is a relatively large difference between the average prediction of the model and the value that should be predicted. With the goal of training a high predictive performance model in mind, it is important to consider the bias-variance trade-off. This states that predictive models should balance bias and variance (e.g., Belkin, Hsu, Ma and Mandal (2019)). In this case, variance refers to the amount that the predicted values will change when provided with different training data. A combination of high bias and low variance results in underfitting: the model fails to capture all relationships in the data. On the other hand, a combination of low bias and high variance results in overfitting: the model represents the training data well but performs poorly on unseen or noisy data. The ideal predictive model is rich enough to express underlying structures in data sets but simple enough to avoid fitting spurious patterns. In this case, both bias and variance are low. The bias-variance trade-off is exemplified by Figure 4. Methods with higher levels of variance than the logistic regression are desired to achieve a balance between bias and variance.

Therefore, multiple machine learning techniques are used to predict the occurrence of traffic congestion.

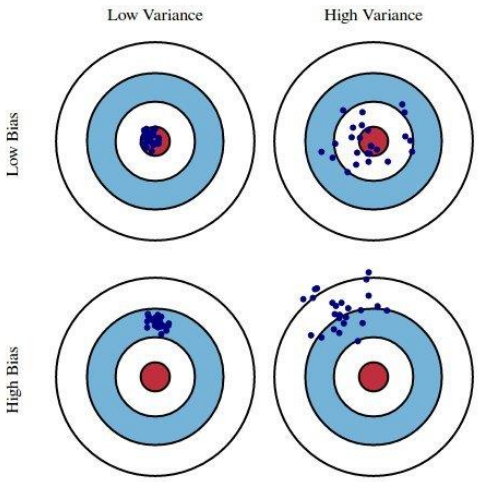


Figure 4: Examples of predicted values under possible combinations of bias and variance

The no free lunch theorem (Wolpert, 1996) is highly applicable to machine learning algorithms: it is impossible to know beforehand which model performs best on the task-specific data. Therefore, the predictive performance of multiple models is compared to determine which type of model performs best. For this purpose, 10% of the balanced data set is randomly selected and taken apart to serve as test set. This data is completely unseen by any of the trained models and is only used to assess the eventual predictive performance of the models. The traffic congestion data is highly structured. For this reason, machine learning techniques that are capable of handling this data type are used. Firstly, an unregularised logistic regression is constructed to serve as a benchmark for the other models. Then, three other models are trained to evaluate their performance in comparison to the benchmark model. The first model is a lasso-regularised logistic regression. The main difference with the benchmark model is that regularisation is applied, which reduces the variance of a model. The second model is an extreme gradient boosting (XGBoost) model. Various comparisons have shown that XGBoost models obtain the highest accuracies when models are trained on structured data (e.g., Mangal and Kumar (2016), and Nielsen (2016)). The third model is a feedforward neural network. Neural networks have been used regularly to predict traffic congestion (e.g. Alarcon-Aquino and Barria (2006), Chakraborty et al. (2018), and Fouladgar, Parchami, Elmasri and Ghaderi (2017)). However, unstructured data has been used in all of these cases. This means that traffic congestion is being predicted based on the input of visual maps on which traffic congestion levels have been marked. Although neural networks are generally used to model unstructured data, it has achieved high prediction accuracies on structured data before (e.g., Abdelwahab and Abdel-Aty (2001), and Yang, Yeo and Kim (2003)).

To mitigate the risk of overfitting, ten-fold cross-validation is applied to optimise the models' hyperparameters (Hastie, Tibshirani, & Friedman, 2001). Hyperparameters are parameters whose values are used to control the learning process (Goodfellow, Bengio, & Courville, 2016). Changing the hyperparameters of a model affects its predictive performance. Therefore, values of hyperparameters should be chosen in a manner that maximises a model's performance. It is essential to evaluate the performance of a hyperparameter set on unseen data to prevent overfitting. The requirement of unseen data is why ten-fold cross-validation is used. Ten-fold cross-validation randomly divides the training data into ten equally sized folds. Then, ten separate models are trained and every fold is withheld from the model once. Thus, the model is trained on the remaining nine folds. The performance of this model is evaluated on the withheld fold. The error of the model on the withheld fold is referred to as the out-of-sample error. The hyperparameter combination that results in the lowest out-of-sample error is used for the final model.

In the end, the accuracy, sensitivity, and specificity of the models on the test subset are used to assess the models' predictive performance. The accuracy is equal to the correctly predicted observations as a percentage of the total number of observations. The sensitivity and specificity refer to the percentage of correctly predicted positive and negative encounters, respectively. It is essential to note that maximising accuracy may not be the solemn goal of the model (Parikh, Mathai, Parikh, Chandra Sekhar, & Thomas, 2008). A trade-off between sensitivity and specificity exists, which comes down to the question: do we prefer wrong predictions in negative or positive encounters? This can be exemplified by COVID-19 tests (Kumleben, et al., 2020). You want to make sure that every person infected by the Coronavirus receives a positive test result. Providing a positive test result to someone that is not infected is inconvenient, but the consequences are less severe than providing a negative test result to someone that is infected. Therefore, COVID-19 tests should aim at achieving a high sensitivity rather than a high specificity. However, without neglecting the importance of high sensitivity, overall accuracy should still adhere to a high standard. Otherwise, people might deem the tests unreliable if the accuracy drops too low. In the same manner, it can be argued that road users consider the sensitivity more important than the specificity. People dislike encountering a traffic jam when free traffic flow was predicted but might be less displeased about facing free traffic flow when congestion was predicted. The remainder of this section discusses the theoretical background of the various techniques.

4.1. Lasso-regularised logistic regression

The lasso-regularised logistic regression is a relatively simple regression model to predict binary outcomes. The base of this model is an ordinary logistic regression. As the aim of the model is to predict traffic congestion as accurately as possible, all available variables and road, time, weekday, month, and year effects are included in the model. A common issue concerning unregularised regressions with large data sets is overfitting (Goodfellow, Bengio, & Courville, 2016). To avoid overfitting, Tibshirani (1996) has proposed to add a least absolute shrinkage and selection operator (lasso) penalty term to regressions, which makes the regression coefficients shrink. As a result of the shrinkage of the coefficients, the model's variance and the risk of overfitting reduce.

The coefficients in a logistic regression are calculated by minimising its loss function: the logistic loss. In a lasso-regularised logistic regression, a penalty term is added to the logistic loss function. Equation (1) depicts the resulting loss function $\mathcal{L}$ of a lasso-regularised logistic regression:

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} -y_i \log\Big(h_\beta(x_i)\Big) + (1-y_i)\log\Big(1-h_\beta(x_i)\Big) + \lambda \sum_{j=1}^{m} |\beta_j| \qquad (1)$$

In this equation, β is the set of coefficients, n is the number of observations, y_i is the observed value for observation i, $h_\beta(x_i)$ is the predicted value for observation i and coefficients β, λ is the lasso penalty parameter, and m is the number of variables. The last part of equation (1) contains the penalty term. It illustrates that the regression coefficients are shrunk by adding the sum of the absolute values of all beta coefficients to the loss function. To avoid prioritising one variable over another, all variables should be on the same scale. Therefore, the independent variables are standardised beforehand. Also, categorical variables have to be one-hot encoded. By one-hot encoding, a separate binary variable is created for all categories of a categorical variable. The severity of the lasso penalty is determined by the lasso penalty parameter λ. The magnitudes of the coefficients are negatively related to λ: the coefficients shrink more as λ becomes larger and vice versa. The coefficients tend to zero if λ tends to infinity, and the coefficients are similar to ordinary logistic regression coefficients if λ is exactly zero.

A disadvantage of applying lasso regularisation to regressions is that the coefficients lose their interpretability. This is due to two reasons. Firstly, the regression coefficients do not represent the strength of a relationship anymore. Due to the penalty term shrinking the coefficients, the regression coefficient simply signifies a numerical parameter rather than an actual relationship.

This makes it impossible to make exact statements about causal relationships between variables. Secondly, it is fiercely debated how standard errors and significance levels of lasso-regularised regressions should be calculated and to what extent these values are meaningful (e.g., Goeman, Meijer and Chaturvedi (2016), Kyung, Gill, Ghosh and Casella (2010), and Reid, Tibshirani and Friedman (2016)). The calculation method that is considered most appropriate for lasso-regularised regression still severely biases standard errors. Therefore, the original authors of the algorithm advise not to report standard errors and significance levels of predictors in lasso-regularised regressions, as it can result in an erroneous or biased representation of reality (Reid, Tibshirani, & Friedman, 2016).

An essential difference between an unregularised and lasso-regularised logistic regression is how the loss function as described in equation (1) is minimised. An unregularised logistic regression uses gradient descent to optimise its regression coefficients. The loss function is minimised by repeatedly altering the coefficients of the model in the opposite direction of the gradient of the loss function. This method cannot be applied to a lasso-regularised logistic regression, however, due to the absolute value of the regression coefficients in the penalty term. Absolute values are not differentiable, making it impossible to calculate the gradient of the loss function of a lasso-regularised regression. Therefore, the loss function of a lasso-regularised regression is optimised by the cyclic coordinate descent method (Friedman, Hastie, & Tibshirani, 2010; Tseng, 2001). This approach successively minimises along coordinate directions to find the minimum of the loss function. In other words, the coordinate descent method updates one parameter at a time, whereas the gradient descent method tries to update all parameters at once.

The eventual value of λ that will be used for this model is a hyperparameter. To find the optimal value of λ, a separate model is trained through ten-fold cross-validation with all values of λ of 10^{-s} with s between -10 and 10 by incremental steps of 0.25. The value that results in the highest out-of-sample cross-entropy is chosen as the optimal value of λ. Cross-entropy is calculated as:

$$H = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \qquad (2)$$

In this equation, y_i is the observed value and $\hat{y}_i$ is the probability predicted by the model for observation i.

4.2. Extreme gradient boosting

The second technique that is used to predict the probability of traffic congestion arising is extreme gradient boosting. The model is based on decision trees but incorporates various other statistical techniques. More specifically, the XGBoost algorithm developed by Chen and Guestrin (2016) is applied. XGBoost is a decision tree based ensemble algorithm that uses a gradient boosting framework. Like other boosting algorithms, XGBoost aims to come up with a high-quality prediction by training a sequence of weak models that compensate for the weaknesses of its predecessors. Again, the same variables including road, time, weekday, month, and year effects are used to predict traffic congestion. On the contrary to the lasso-regularised logistic regression and feedforward neural network, data standardisation is not required. As XGBoost demands numerical data, categorical variables are one-hot encoded.

Decision trees are non-parametric models that can be used for predictive purposes. A decision tree consists of nodes and branches. A schematical representation of a decision tree is shown in Figure 5. At every node, a certain variable is evaluated in order to split the observations to make a data point follow a certain path when making a prediction. The first node of a decision tree is the root node and the final nodes are leaf nodes. All data points are assigned the value of the leaf node that they end up in by following the decision tree. The value of a leaf node is equal to the class that is represented most often in that leaf node. Nodes are connected in a top-down manner through branches. At every split of a branch, the Gini index is minimised. A Gini index of zero means that all observations belong to the same class, while a score of one denotes that all observations are randomly distributed across classes. The Gini index can be calculated as:

$$G = 1 - \sum_{i=1}^{n} (p_i)^2 \tag{3}$$

In this equation, p_i denotes the probability of an observation being assigned to a particular class. To avoid that a tree continues to grow until every training observation has been assigned its own leaf node, multiple stopping conditions are set. Training a tree without stopping conditions would result in overfitting. The maximum number of nodes between the root node and a leaf node is referred to as the depth of a tree. This number cannot exceed a certain threshold that is set beforehand. Also, an additional split should result in a minimum loss. The height of this additional loss can be determined by means of the gamma parameter. Both the maximum depth of a tree and the gamma parameter are hyperparameters.

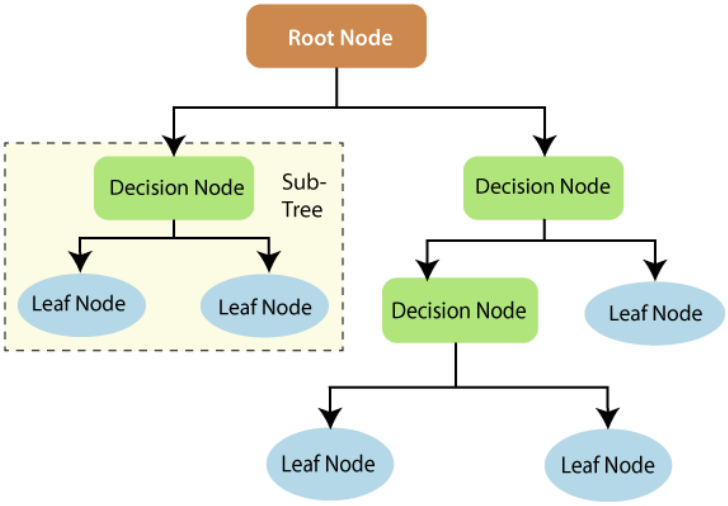


Figure 5: Schematic representation of a decision tree

The predictive performance of just one decision tree is generally limited. Therefore, multiple trees can be combined by the bootstrap aggregating method, abbreviated to bagging. This is a technique developed by Breiman (1996) that reduces variance and the risk of overfitting for predictive models. The bagging process is schematically shown in Figure 6. Firstly, l distinct bootstrap samples are taken from the original data set D of size $n \times m$, with n and m denoting the number of observations and variables, respectively. The bootstrap samples consist of pn observations and qm variables, in which p and q represent the proportion of the observations and variables that are used in every bootstrap sample. The samples are filled by collecting observations from the original data set randomly and with replacement. This means that observations can occur in a bootstrap sample multiple times or not at all. Eventually, l decision trees are constructed based upon the bootstrap samples. The actual prediction for an individual observation is determined by majority vote.

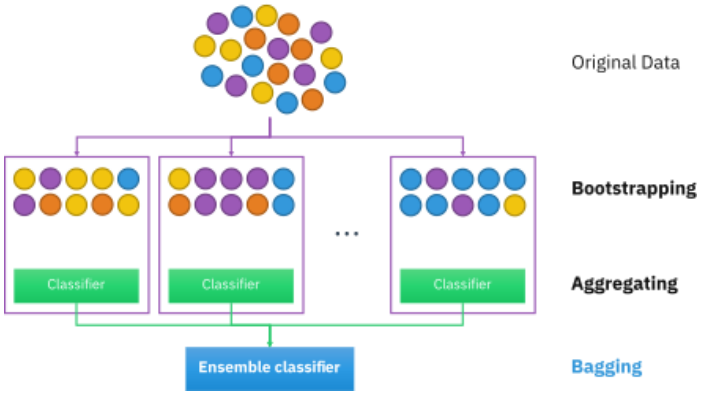


Figure 6: Schematic representation of the bagging process

The bagging method is a simultaneous process: the separate decision trees are constructed at the same time and do not exchange information between each other. The boosting technique

developed by Freund and Schapire (1996) is a variation on the bagging method that aims to teach decision trees the weaknesses of other trees by training them sequentially. Every time a decision tree has been trained, the model evaluates which observations are misclassified. The weight of these observations is increased making it more likely that they will be predicted correctly in the following decision tree. By repeating this process multiple times, the successive trees correct for the mistakes made by other trees. An extension of the boosting method is the gradient boosting technique that is developed by Friedman (2001). This algorithm starts by assuming an initial weights distribution across the original sample D_1 such that $D_{1,i} = 1/N$ for all $i \in N$. Then, some learning rate α_t is assumed for tree t and a new weak classifier c_t is created. The initial learning rate is a hyperparameter. The weight distribution is updated in the following manner:

$$D_{t+1,i} = \frac{D_{t,i} e^{-\alpha_t y_i c_t(x_i)}}{\sum_{i=1}^{N} D_{t,i} e^{-\alpha_t y_i c_t(x_i)}} \tag{4}$$

In this equation, y_i is the observed outcome of observation i, and $c_t(x_i)$ is the class predicted by tree t for independent variables x_i. With ε_t being the error of tree t, the learning rate of a tree is calculated as:

$$\alpha_t = 0.5 \log\left(\frac{1 - \varepsilon_t}{\varepsilon_t}\right) \tag{5}$$

This process is repeated until no significant improvements are made, or until the maximum number of iterations has been reached. The maximum number of iterations is a hyperparameter. The final prediction is computed by using the weighted average of the outputs of all trees, with α_t as the weight for each tree. The extreme gradient boosting method that is developed by Chen and Guestrin (2016) extends the gradient boosting algorithm. Although three different variants of the XGBoost algorithm exist, this research solemnly uses the basic exact greedy algorithm[3].

It is clear that training an XGBoost model comes with many hyperparameters that can be optimised: the number of iterations, tree depth, initial learning rate, training and variable proportions, and gamma parameter are all hyperparameters. The hyperparameters are optimised by the Bayesian Hyperparameter Optimisation (BHO) method (Snoek, Larochelle, & Adams, 2012). Multiple separate XGBoost models are sequentially trained through ten-fold cross-validation with different hyperparameter combinations. BHO calculates the probability that a certain hyperparameter set outperforms previous sets based upon the expected improvement as proposed by Jones, Schonlau and Welch (1998)[4]. This is opposed to most

[3] Discussing the entire functioning of this algorithm is beyond the scope of this research. See Chen and Guestrin (2016) for a complete description of the algorithm.

[4] The concept of expected improvement was originally proposed by Močkus (1975) but was not yet implemented in the optimisation process of black-box functions.

hyperparameter optimisation methods that do not use the information provided by previous combinations to pick the next set of hyperparameter values. The combination with the highest probability of performing better than the previous sets is trained and its out-of-sample performance is then evaluated by means of cross-entropy. Thus, every evaluated combination provides more information about the optimal hyperparameter values. In total, 30 different combinations are evaluated through the BHO method. The ranges of the hyperparameter values have to be pre-set. The number of iterations ranges between 10 and 400, and the depth per tree between 3 and 50. The initial learning rate varies between 0 and 1. Moreover, the optimal proportion of training observations and variables that is used to train the decision trees has to be determined. As it is a proportion, the possible values range from 0 to 1. Lastly, the gamma parameter has to be between 0 and 10.

An XGBoost model is a classic example of a black-box method: the user provides input data to the model and the model provides an output value, but what exactly happens in between is largely unknown. The number of parameters in black-box methods can grow very rapidly. This hinders the interpretation of such models and makes it hard to detect relationships between variables. It is still possible to gain some insights into the relationship between independent and dependent variables, though. Two methods are used in this research for this purpose. Firstly, a variable importance plot (VIP) is constructed that shows which variables have the most profound impact on the predictive performance of the model (Breiman, 2001). A VIP is constructed by randomly permuting the values of a variable and recording the drop in predictive performance. The variable that causes the largest drop in performance measured in terms of accuracy can be considered as the most influential variable for predicting the dependent variable. The second interpretation method that is applied is individual conditional expectation (ICE) curves, as developed by Goldstein, Kapelner, Bleich and Pitkin (2015). This technique considers the change in the predicted value for individual observations while the variable of interest x_s changes and all other variables $\boldsymbol{x}_c$ are held constant. A separate line is plotted in a graph for every observation. More formally, it can be stated that for each instance in $\left\{\left(x_s^{(i)}, \boldsymbol{x}_c^{(i)}\right)\right\}_{i=1}^{N}$ the curve $\hat{f}_s^{(i)}$ is plotted against the range of $x_s^{(i)}$, while $\boldsymbol{x}_c^{(i)}$ remains fixed. For the sake of clarity, only the ICE curves of 1,000 randomly selected observations are plotted.

4.3. Feedforward neural network

The third and last predictive model that is applied to the data is a feedforward neural network. Again, all available variables are included in the neural network, including road, time, weekday,

month, and year effects. The concept of a neural network was first conceptualised by McCulloch and Pitts (1943). A neural network consists of a set of links between the raw data in the input layer and the actual prediction in the output layer. A simplified schematic representation of the architecture of a neural network is shown in Figure 7. A neural network consists of at least one hidden layer that is responsible for transforming the data. Every hidden layer contains at least one neuron. These neurons carry a certain weight that alters the value that is fed to the neurons. Both the number of hidden layers and the number of neurons in a hidden layer are hyperparameters.

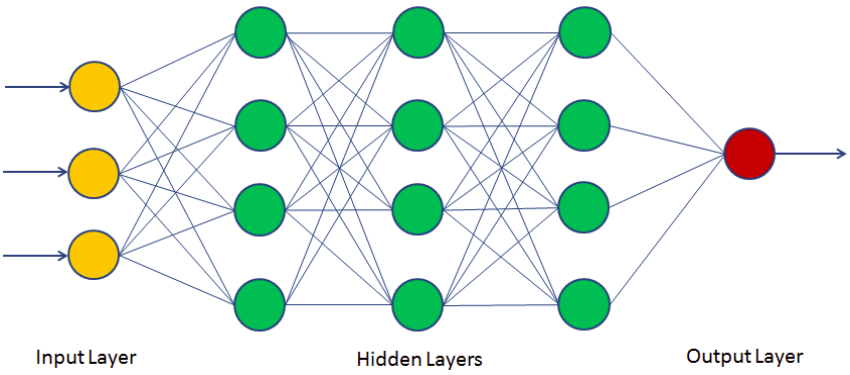


Figure 7: Schematic representation of a neural network

The neurons in a neural network behave similarly to the neurons in a human brain. Dendrites through which information is transmitted form connections between the neurons in successive hidden layers. Once a neuron receives a value from a preceding neuron, this value is altered by the weight of the neuron and then sent to the next neuron. The working of a neural network's neurons is schematically represented in Figure 8. The figure shows that three input values x_0, x_1 and x_2 are sent to the neuron. As for the lasso-regularised logistic regression, the data should be standardised beforehand and categorical variables should be one-hot encoded. The weights w_0, w_1 and w_2 belonging to values x_0, x_1 and x_2 are parameters in the network reflecting the importance of the connection between two neurons. The values of the neurons in the first hidden layer are calculated by:

$$\boldsymbol{h}^{(1)} = f^1\left(\boldsymbol{W}^{(1)T}\boldsymbol{x}\right) + \boldsymbol{b}^{(1)} \tag{6}$$

For all other layers, the values are calculated by:

$$\boldsymbol{h}^{(l)} = f^l\left(\boldsymbol{W}^{(l)T}\boldsymbol{h}^{(l-1)}\right) + \boldsymbol{b}^{(l)} \tag{7}$$

In these equations, $\boldsymbol{h}^{(l)}$ is the l-th hidden layer, with f^l its activation function. This activation function is a fixed nonlinear function that allows for nonlinear relationships between neurons. The neural network in this research uses the rectified linear unit (ReLu) activation function for

all layers except the last one[5]. The ReLu activation function states that $f(z) = \max(0, z)$. However, the desired output of the model is the probability of traffic congestion occurring at a certain stage. The ReLu function is not restricted to values between 0 and 1; z could theoretically grow towards infinity. Therefore, the sigmoid activation function is used in the final layer, which does result in a value between 0 and 1. The sigmoid function states that $f(z) = \frac{1}{1+e^{-z}}$. $\boldsymbol{W}^{(l)T}$ is the transpose of the weights of hidden layer l and $\boldsymbol{b}^{(l)}$ is the bias term of the hidden layer. The bias term can be seen as the constant in a linear regression model. $\boldsymbol{x}$ is the input data, which is only relevant for the first hidden layer. The manner in which the values in a hidden layer are calculated means that they are always dependent on the values in the previous layer.

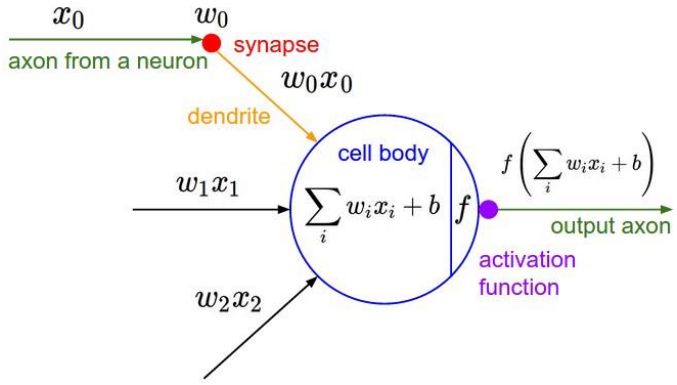


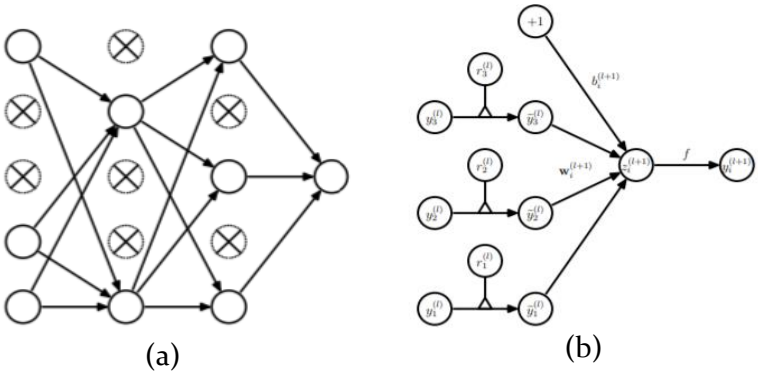
Figure 8: Schematic representation of the working of a neuron

Neural networks are extremely prone to overfitting due to their high complexity levels. To mitigate the risk of overfitting, dropout is applied to the network (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). Dropout is a technique that randomly changes the weights of neurons to zero during training. This is accomplished by multiplying the value of a neuron with a binary value. The probability that such a binary value is zero is referred to as the dropout rate, which is a hyperparameter. The architecture of a neural network with dropout applied to it is schematically shown in Figure 9. As a result of dropout being applied to the network, the calculation of the values of the neurons in hidden layers as described in equation (7) changes to:

$$\boldsymbol{h}^{(l)} = f^{l}\left(\boldsymbol{W}^{(l)T}\boldsymbol{h}^{(l-1)} + \boldsymbol{b}^{(l)}\right)\boldsymbol{r}^{l} \tag{8}$$

In this equation, $\boldsymbol{r}^{l}$ is the dropout rate of l-th hidden layer. The dropout rate follows the Bernoulli distribution.

[5] In an extremely detailed overview of multiple neural network techniques, Goodfellow, Bengio and Courville (2016) recommend to use the ReLu activation function as developed in multiple stages by Jarrett, Kavukcuoglu, Ranzato and LeCun (2009), Nair and Hinton (2010), and Glorot, Bordes and Bengio (2011) in all neural networks.



(a) (b)

Figure 9: Schematic representation of a neural network with dropout
Source: Srivastava et al. (2014)

Eventually, the neural network aims to find parameters that result in the best predictive performance. Neural networks use the gradient of the loss function, which is the cross-entropy, to train its parameters. The gradient provides the direction in which the loss function increases faster, meaning that the weight parameters should move in the opposite direction to minimise loss. Although an expression for the gradient is relatively straightforward, it would be too computationally expensive to numerically evaluate this expression. Therefore, the technique of backpropagation as developed by Rumelhart, Hinton and Williams (1986) is applied. Backpropagation uses multiple first-order derivatives to obtain the gradient of the loss function. The technique attempts to minimise loss by calculating the first-order derivative of the loss function of the model with respect to every weight in the network[6]. The most uncomplicated method to obtain the minimum loss is the gradient descent method. This method updates the weights after every iteration by altering them with the product of the learning rate α and the gradient of the loss function. Training the model takes several epochs and iterations. One epoch equals all data being propagated forward and backwards through the network once. As an entire data set is too large to feed to a neural network though, the data set is divided into multiple equally large subsets called batches. Both the number of epochs and the batch size are hyperparameters. Too few epochs or a too large batch size results in underfitting, while too many or a too small batch size results in overfitting. Every time a batch is passed into the neural network counts as an iteration. The change in the network parameters can be mathematically depicted as:

$$w_{ij,t+1}^{l} = w_{ij,t}^{l} - \alpha \frac{\partial \mathcal{L}(X, \theta)}{\partial w_{ij}^{l}} \tag{9}$$

[6] See Section 6.5. of *Deep Learning* by Goodfellow, Bengio and Courville (2016) for a detailed explanation and mathematical derivation of the backpropagation technique.

In this equation, $w_{ij,t}^{l}$ is the weight of neuron i in hidden layer l receiving input from neuron j during iteration t, α is the learning rate, and $\mathcal{L}(X, \theta)$ is the derivative of the loss function with using hyperparameter set θ. However, as the gradient descent method indicates that the gradient of the loss function with respect to the network weights has to be calculated for the entire data set, this method is highly computationally intensive. A solution for this problem is to perform a parameter update for every combination of input $\boldsymbol{x}_i$ and output y_i. This technique is referred to as the stochastic gradient descent method and is based upon work by Robbins and Monro (1951). The usage of this method changes equation (9) into:

$$w_{ij,t+1}^{l} = w_{ij,t}^{l} - \alpha \frac{\partial \mathcal{L}(\boldsymbol{x}_i, y_i\,, \theta)}{\partial w_{ij}^{l}} \tag{10}$$

A challenge that remains, though, is that it is hard to find the optimal learning rate. A learning rate that is too small results in slow convergence, whereas a too high learning rate hinders the convergence of the network. Also, by picking just one learning rate, the same rate applies to all parameter updates. To remedy this challenge, the adaptive moment estimation (Adam) method, as developed by Kingma and Ba (2015), is used in combination with the stochastic gradient descent method. The Adam method uses estimates of the exponentially decaying average of past squared and non-squared gradients to compute an adaptive learning rate. The weight parameters are then updated as follows:

$$w_{ij,t+1}^{l} = w_{ij,t}^{l} - \frac{\alpha}{\sqrt{v_t} + \epsilon} m_t \tag{11}$$

In this equation, v_t is the estimate of the average of the past squared gradients, m_t is the estimate of the average of the past non-squared gradients, and ϵ is a smoothing term that avoids division by zero. The learning rate α of the first iteration is a hyperparameter. The bias-corrected estimates of the decaying average of the squared and non-squared gradient are calculated as follows:

$$\begin{aligned} m_t &= \frac{\beta_1 m_{t-1} + (1 - \beta_1) g_t}{1 - \beta_1} \\ v_t &= \frac{\beta_2 v_{t-1} + (1 - \beta_2) g_t^2}{1 - \beta_2} \end{aligned} \tag{12}$$

In this equation, g_t is the gradient of the loss function, while β_1 and β_2 are hyperparameters.

As the XGBoost model, training a neural network comes with some hyperparameters that have to be optimised. To reduce the computational costs of the hyperparameter optimisation process, default values can be used for some hyperparameters. Kingma and Ba (2015) have shown that the effect of optimising the Adam parameters is neglectable. They propose values of 0.002 for the initial learning rate α, 0.9 for β_1, 0.999 for β_2 and 10^{-8} for ϵ. Considering the marginal effect

of optimising these hyperparameters, these values are used to train the model. The training of the model is halted as soon as the loss function has not improved by at least 0.1 over the last 20 epochs. In this manner, the number of epochs is dependent on the training behaviour of the neural network. The remaining hyperparameters are optimised through the BHO method. The number of hidden layers ranges from 1 to 10. The range of the number of neurons in a hidden layer stretches from 3 to 200. The dropout rate varies between 0.0001 to 0.9999 and the batch size from 500 to 100,000. For interpretational purposes, the same methods are used as for the XGBoost model.

5. Results

Firstly, the hyperparameters of all models are optimised. A grid search shows that a λ parameter of $10^{-5.25}$ results in the highest cross-validated cross-entropy for the regularised logistic regression. The BHO method indicates that the optimal hyperparameters of the XGBoost model are 342 iterations, a maximum tree depth of 35 nodes, a learning rate of 0.063, a gamma of six, and both a sample and variable proportion of one. In the same manner, three hidden layers, 113 neurons per layer, a batch size of 4,574 and a dropout rate of 0.0651 are found to be the optimal architecture of a neural network. It is interesting to note that regularisation levels are low for all three models. This suggests that models with relatively high variance levels perform best when predicting traffic congestion. The values of other hyperparameters are comparable to the values of other applications of these machine learning techniques.

A benchmark model in the form of an unregularised logistic regression is constructed. As common for an unregularised logistic regression, its loss function is optimised through the gradient descent method. The benchmark model achieves an out-of-sample accuracy of 84.69% and an accuracy on the unseen test data of 84.85%. A comparison between the predictive performance of the benchmark model and the other three models is shown in Table 2. The predictive performance of a regularised logistic regression is found to be only marginally better than the performance of the benchmark model. This is expected as the loss functions of the benchmark model and the regularised logistic regression differ only slightly due to the small penalty parameter. The coefficients of the both regressions are depicted in Appendix D. There are substantial differences between the coefficients of the benchmark model and the regularised logistic regression though. These differences are a result of the loss function of the regularised logistic regression being optimised through the cyclic coordinate descent method instead of the gradient descent method. To highlight the marginal effect of such a small penalty parameter,

the coefficients of an unregularised logistic regression optimised through the cyclic coordinate descent method are also included in Appendix D. In contrast to the regularised logistic regression, the XGBoost model performs substantially better than the benchmark model. The XGBoost model achieves an out-of-sample accuracy of 90.13% and an accuracy on the test set of 90.39%. Although the performance of the neural network is better than the logistic regressions' performance, it does not perform as well as the XGBoost model. An accuracy of almost 89% is achieved both on out-of-sample observations and the test data set. The mean ICE curves of the independent variables visualising their relationship with the probability of congestion arising are depicted in Appendix E for both the XGBoost model and neural network.

Table 2: Comparison between the predictive performance of the benchmark model and other predictive models

	Benchmark model	Regularised logistic regression	XGBoost model	Neural network
Training accuracy	84.70%	84.69%	94.39%	89.88%
Out-of-sample accuracy	84.69%	84.69%	90.13%	88.96%
Test accuracy	84.85%	84.86%	90.39%	88.96%
Test sensitivity	86.70%	86.70%	91.25%	86.95%
Test specificity	83.00%	83.02%	89.55%	90.97%

6. Discussion

The findings in the previous sections indicate that traffic congestion is well-predictable by machine learning techniques, even without the use of traffic flow data. The XGBoost model achieves an accuracy on unseen test data of almost 91%. Its accuracy is comparable to the best-performing model based upon historic traffic flow data that is found in the existing literature, which achieves an accuracy of 93%. Despite the comparable predictive performance, the manner in which the predictions have been obtained are completely different. The best-performing model in the existing literature uses highly detailed data about vehicles' trajectories on a highway section of less than a kilometre. The model that has been trained in this research uses a broader variety of data resources to predict congestion states on a nationwide level.

Notwithstanding the high predictive performance that has been obtained by the XGBoost model, it should be noted that, as mentioned beforehand, it is not only about maximising accuracy, but also about the sensitivity-specificity trade-off. The best observed historic traffic flow model achieves an accuracy of 93% in combination with a sensitivity of 96%. The accuracy, sensitivity, and specificity of models can be influenced by changing the threshold at which a positive encounter is predicted to different values than the default of 0.5. For example, if a traffic congestion occurrence is predicted as soon as the probability of traffic congestion exceeds 0.6, fewer congestion encounters will be predicted relative to a threshold of 0.5. This increases

sensitivity but decreases accuracy and specificity. The dependency of the accuracy, sensitivity and specificity on the prediction threshold is visualised in Figure 10. The graph shows that the accuracy is relatively stable between thresholds of 0.15 and 0.85, but deteriorating rapidly outside this range. To achieve a sensitivity of 96%, the prediction threshold has to be set at 0.875. This would result in an overall accuracy of 84.62% and a specificity of 71.34%. Improving the sensitivity of the XGBoost model clearly comes at the cost of lower overall accuracy. The relatively low accuracy that results from a high sensitivity illustrates that the XGBoost model is still lacking some predictive performance compared to the best-observed performance of a historic traffic flow model.

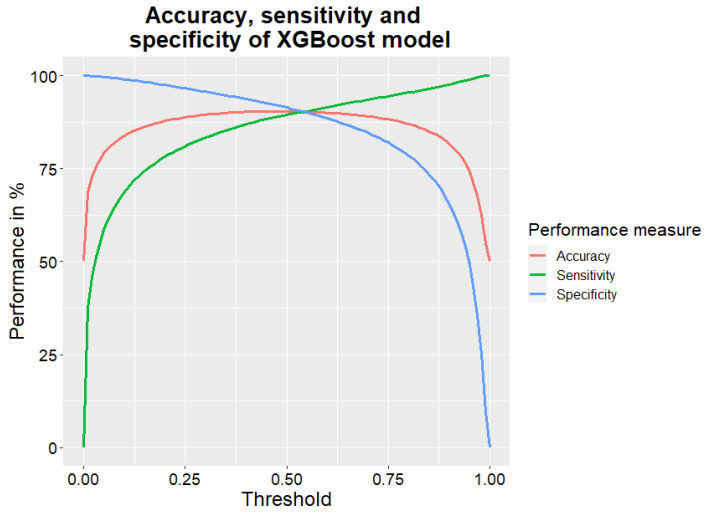


Figure 10: Dependency of accuracy, sensitivity and specificity on the prediction threshold

It is desired that models not only predict accurate results but also confident results. A predicted probability of 0.501 comes with more uncertainty than a predicted probability of 0.999. Despite these probabilities both resulting in a traffic jam being predicted, the latter value is much more confident about its prediction. The predicted probabilities on the test set per model in Figure 11 visualise that both logistic regressions predict values that are further away from zero or one than the other two models. This effect is particularly strong when it comes to predicted probabilities that are close to one. Whereas the neural network and XGBoost model predict relatively many probabilities that are extremely close to one, the number of predicted probabilities by the logistic regressions eases as the predicted probabilities approach one. When comparing the two best performing models, the XGBoost model is found to predict slightly more confidently than the neural network. However, the difference between the neural network and XGBoost model is less strong than between those two models and the logistic regressions. In the end, Figure 11 confirms that the XGBoost model performs best when predicting traffic congestion not only when it comes to actual accuracy, but also when looking at the confidence of the predictions.

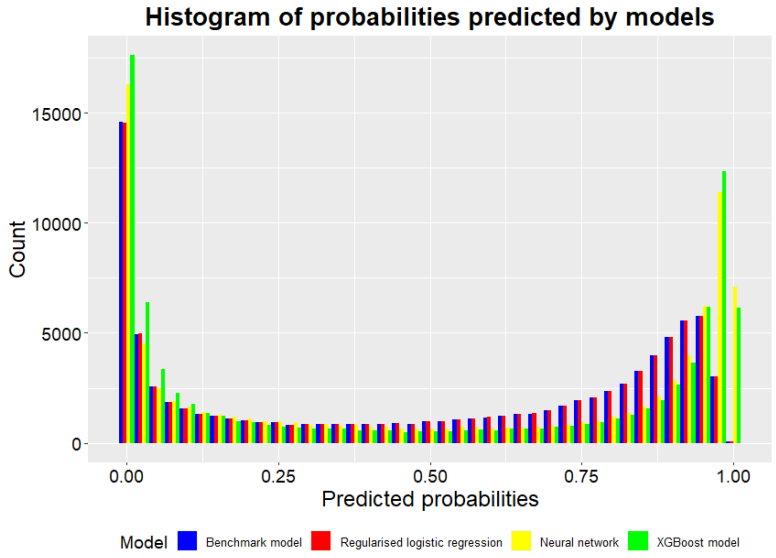


Figure 11: Histogram of predicted probabilities per model

The difference in prediction confidence between the models is also visible in their respective logistic loss. The logistic loss function does not only account for whether or not a prediction is correct, but also for the actual difference between the observed outcome and the predicted probability. A low logistic loss indicates that the predictions are closer to the observed values in comparison to a high logistic loss, and vice versa. To determine whether or not the models' prediction confidence is dependent on certain variables, the logistic loss per variable quartile is represented in Appendix F. As expected, differences between logistic losses per quartile of the benchmark and regularised logistic regression are largely similar, whereas the neural network and XGBoost model achieve substantially lower logistic losses. Despite this, major differences in logistic loss per variable quartile dependent on which model the predictions are based upon are not found.

As the geographical landscape of the Netherlands varies strongly, it is possible that differences in predictive performance between regions exist. The Netherlands is divided into twelve provinces. The three provinces Zuid-Holland, Noord-Holland and Utrecht, commonly referred to as the Randstad, are the most densely populated and form the economic heart of the country. On the other hand, the provinces Zeeland, Drenthe, Groningen and Friesland are the least densely populated. Two maps in Figure 12 illustrate the accuracy and logistic loss per province of the XGBoost model. It is noticeable that the highest predictive performance both in terms of logistic loss and accuracy is achieved in the sparsely populated provinces. The model achieves a slightly lower predictive performance when predicting traffic congestion in the Randstad provinces. The maps clearly show that the performance of the XGBoost model is dependent on

the geographical location of the prediction. Despite the geographical dependency, the model is still useful for predicting traffic congestion. The lowest achieved accuracy in any province is 88.30% in Noord-Limburg, which is still a relatively high accuracy. A possible explanation of the better predictive performance in the sparsely populated regions could be that it is somewhat easier to predict traffic congestion in these regions. Traffic jams in the four sparsely populated provinces are scarce. This means that a relatively high accuracy can already be achieved simply by predicting no traffic congestion for all observations in these provinces.

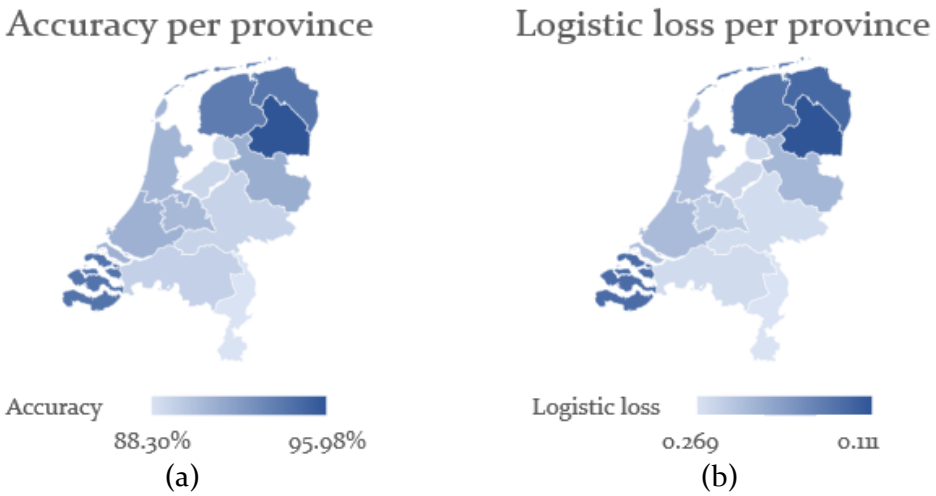


Figure 12: Accuracy and logistic loss per province

VIPs containing the fifteen variables with the highest variable importance scores of the models have been created to find which variables are the most important when it comes to predicting traffic jams. This is measured by assessing the change in predictive performance when the observations of a certain variable are randomly permuted. The VIPs in Figure 13 show that road availability plays an important role in predicting traffic congestion. Road availability belongs to the two most important variables in every model. GRP per capita is also found to have a strong influence on the predictive performance of the neural network and the XGBoost model, albeit the effect is weaker than road availability's effect. However, GRP per capita seemingly does not play a major role in predicting traffic congestion in both logistic regressions. Public and school holidays are important determinants of traffic jams in these regressions, though. Holidays are of much lesser importance in the neural network and XGBoost model. The two aforementioned differences can potentially explain the gap in predictive performance between the logistic regressions and the other models. It should be noted, though, that this difference is also likely to be at least partly caused by the fact that interaction effects are not included in the logistic regressions, whereas the other two models automatically incorporate those. When comparing the most important variables of the neural network and XGBoost model, the extreme importance of road availability in the XGBoost model stands out. It appears that the neural network uses

multiple variables to acquire accurate predictions, whereas the XGBoost model mainly relies on road availability to predict traffic congestion.

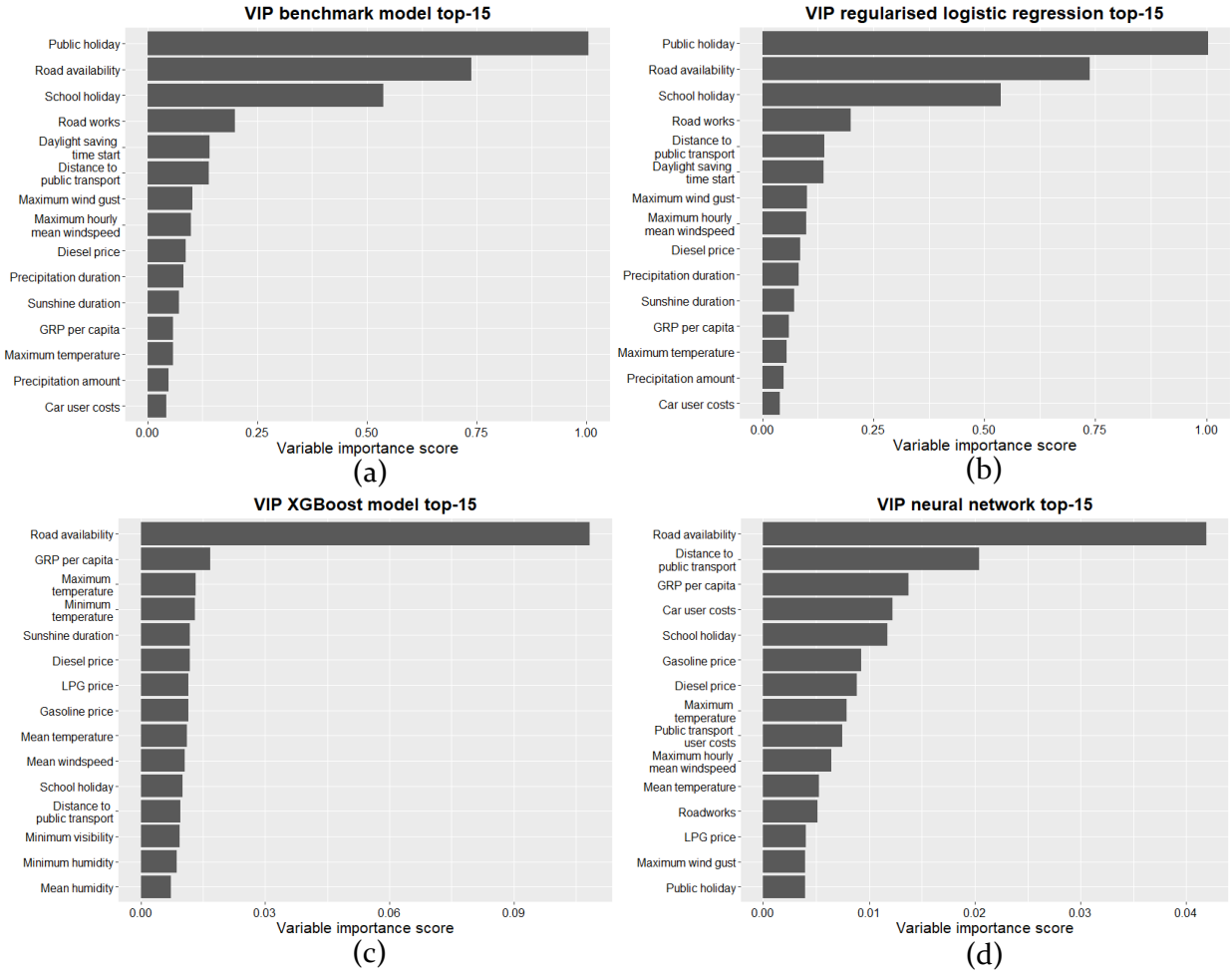


Figure 13: VIPs containing the fifteen most influential variables per model

Figure 2 in section 2.3 contained a conceptual framework summarising the factors that influence either the supply or demand of road space according to the existing literature. Four factors are found to affect the supply-side of road space: constructing more highway kilometres, road accidents, weather conditions, and road works. Literature states that only the first factor of these four is not directly related to the occurrence of traffic congestion. Moreover, the probability of congestion arising is influenced by four factors from the demand-side: user costs, weather conditions, commuter obligations, and economic growth. To assess to what extent the results of this research are in accordance with the findings in the existing literature, the effect of the aforementioned factors on the arising of traffic jams in the models is evaluated.

The first factor that is said to affect the supply-side of road capacity is the construction of more highway kilometres. According to the existing literature, constructing more highway kilometres is unrelated to traffic congestion, as an increase in road capacity is nullified by a potentially even

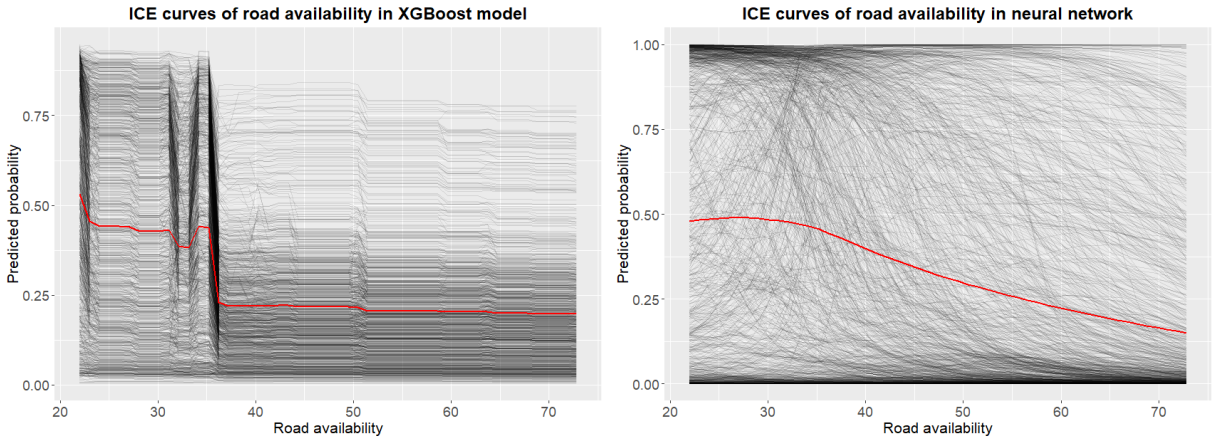
steeper increase in demand. The findings in this research are in sharp contrast with this conception, though. The VIPs in Figure 13 show that road availability does have a severe impact on the probability of traffic congestion arising. To understand the effect of road availability in the black-box methods, ICE curves of road availability in both the XGBoost model and neural network are visualised in Figure 14. The grey lines are individual curves depicting the change for one observation, while the red line is the average of all individual curves. The ICE curves indicate that a higher road availability actually does lower the probability of congestion occurring. More specifically, a steep descent in the probability of congestion occurring is visible around 35 kilometres of highway per 100,000 inhabitants. This is supported by the regression results in Table 3. The unregularised regression indicates a negative relationship between road availability and the probability of traffic congestion, which is significant at the 1% level. Although causal statements based upon the regularised regression should be made cautiously, the negative coefficient of this regression also suggests that a higher road availability lowers the probability of a traffic jam. A potential reason why this result seemingly does not correspond to findings in the existing literature could be how road availability is measured. In this research, road availability has been calculated as the single length of all highways in a province divided by its population. In contrast, articles stating that constructing additional highway kilometres does not reduce the probability of traffic jams arising all multiply the number of highway kilometres by the respective lanes on a section. For example, a one-kilometre high section consisting of three lanes counts for three highway kilometres, whereas it would only account for one highway kilometre in this research. Combining the findings of this research with those in the existing literature, it appears that adding more lanes to an existing highway does not reduce the probability of traffic congestion arising, but that constructing a completely new highway does.

Table 3: Summary of both logistic regressions' coefficients containing all variables related to road capacity

	Dependent variable Congestion	
	Benchmark model	Lasso-regularised regression
Road availability	-0.088***	-0.736
	(0.001)	

Standard errors are reported in parentheses[7].
, **, * indicate significance at the 10%, 5%, and 1% level, respectively.*

[7] As thoroughly discussed in Section 4.1., it is preferred not to report standard errors of the regularised logistic regression.



(a) (b)

Figure 14: ICE curves of road availability in XGBoost model and neural network

A second factor that is found to influence the forming of traffic congestion through the supply-side is road accidents. Although road accidents are not directly included in the models, they are incorporated as so-called accident hotspots. Research has found that the likelihood of accidents occurring on a road section is mainly dependent on the road section's individual characteristics. Therefore, the addition of road fixed effects to the model accounts for differences in the likelihood of accidents occurring on a certain road. Moreover, existing literature found that the start of daylight saving time in the spring can increase the likelihood of road accidents happening. The effect of incorporating road accidents into the models is evaluated by training an XGBoost model excluding the daylight saving time variable and road fixed effects. All other variables and hyperparameters are unchanged. This model obtained an accuracy on the unseen test data of 86.29%, which is approximately four percentage points lower than the original XGBoost model. Although the obtained accuracy might be slightly higher after reoptimizing the hyperparameter, this result suggests that incorporating accident hotspots in a traffic congestion prediction model improves its performance. It must be noted, however, that road fixed effects also account for other differences between road sections, such as variations in the supply and demand ratio between roads. Hence, it is hard to make strong statements about the effect of adding accident hotspots to the models.

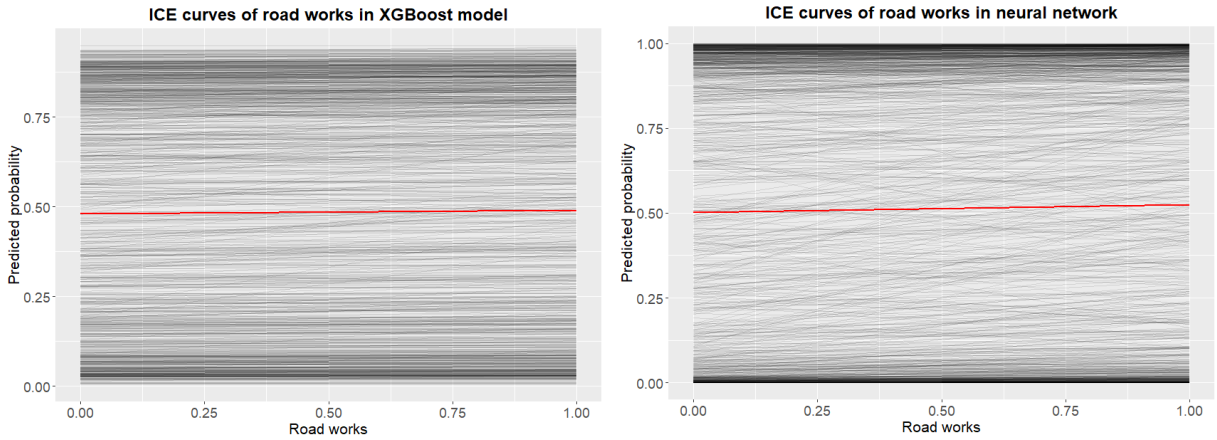
Another factor that affects the capacity of a road section and, subsequently, the probability of congestion occurring is road works. Figure 13 shows that road works are of major importance for both logistic regressions. The unregularised regression coefficients in Table 4 indicate a positive relationship between road works and the probability of traffic congestion, which is significant at the 1% level. Although this positive relationship is supported by the coefficient of the regularised regression, it is disputed by both the XGBoost model and the neural network. The VIPs in Figure

13 illustrate that road works are not an important variable in neither the XGBoost model nor neural network. Moreover, the ICE curves in Figure 15 show that the effect of road works being carried out to a road section has a neglectable effect on the predicted probability. This finding contradicts the general notion that a limitation to the capacity of a road section results in an increased chance of traffic congestion occurring. A possible explanation could be that the information provision in the Netherlands about road works reaches enough people to sufficiently reduce the number of vehicles travelling via the road section. If this is the case, the reduction in capacity is offset by a reduction in traffic flow.

Table 4: Summary of both logistic regressions' coefficients containing all variables related to road works

	Dependent variable Congestion	
	Benchmark model	Lasso-regularised regression
Road works	0.290***	0.195
	(0.022)	

Standard errors are reported in parentheses.
, **, * indicate significance at the 10%, 5%, and 1% level, respectively.*



(a) (b)

Figure 15: ICE curves of road works in XGBoost model and neural network

When looking at the demand-side of traffic congestion, user costs are one of the factors that is found to affect the probability of traffic congestion occurring. For example, higher fuel costs result in fewer people travelling by car and, thus, less traffic congestion arising. The VIPs in Figure 13 confirm that user costs are important to predict the occurrence of traffic congestion. Most variables related to transport user costs are frequently mentioned as important variables in the models. Especially fuel costs are highly ranked when it comes to variable importance. Despite the clear importance of user costs to predict traffic congestion, it is harder to make firm statements about possible causal relationships. This is mainly due to the models disagreeing with each other on the strength and even the sign of the relationship. For example, the unregularised logistic regression coefficients in Table 5 indicate that car user costs, public

transport user costs and the gasoline price are all negatively related to traffic congestion arising. The relationship of all three variables is found to be significant at the 1% level. However, mean ICE curves in Appendix E indicate that changes in these variables do not substantially affect the average predicted probability in the XGBoost model or neural network. Although conclusions about causal relationships cannot be made, it does appear that user costs interact with other variables. This is based upon the notion that user costs are important variables to predict traffic congestion when looking at the VIPs, but the ICE curves showing only marginal changes in the average predicted probability, especially in the XGBoost model.

Table 5: Summary of both logistic regressions' coefficients containing all variables related to user costs

	Dependent variable Congestion	
	Benchmark model	Lasso-regularised regression
Public transport user costs	-2.063*** (0.181)	0.007
Car user costs	-0.017*** (0.006)	-0.045
Gasoline price	-1.128*** (0.206)	0.003
Diesel price	1.279*** (0.202)	0.089
LPG price	0.750*** (0.228)	-0.023
Distance to public transport	0.070 (0.004)	0.140

Standard errors are reported in parentheses.
, **, * indicate significance at the 10%, 5%, and 1% level, respectively.*

A second factor that influences the demand for road space, according to the existing literature, is commuter obligations. Figure 3 already partially confirmed this conception by pointing out that traffic jams usually occur during weekdays and peak rush hours. Moreover, the figure illustrated that the number of traffic jams drops substantially during holiday months. However, just as road accidents, commuter obligations are not directly incorporated into the models, but are accounted for by fixed effects. Only public and school holidays are directly included in the models. Consequently, it is only possible to evaluate the effect of these two holiday types on the probability of congestion arising. All models undoubtedly point at a significantly lower probability of traffic congestion occurring during both public and school holidays. The unregularised regression coefficients in Table 6 indicate a negative relationship significant at the 1% level. The ICE curves of public and school holidays are depicted in Figure 16. The figure shows that the predicted probability of the XGBoost model and neural network decreases by approximately twenty and twenty-five percentage points on average, respectively, during public

holidays, and by approximately seven and five percentage points on average, respectively, during school holidays. To assess the overall effect on the predictive performance of variables related to commuter obligations, an XGBoost model is trained with the same variables and hyperparameters, but excluding the holiday variables and fixed effects related to commuter obligations. This model achieved an accuracy on the unseen test data of 80.65%, confirming the importance of accounting for commuter obligations when predicting traffic congestion.

Table 6: Summary of both logistic regressions' coefficients containing all variables related to commuter obligations

	Dependent variable Congestion	
	Benchmark model	Lasso-regularised regression
School holiday	-0.541***	-0.534
	(0.011)	
Public holiday	-1.002***	-0.981
	(0.029)	

Standard errors are reported in parentheses.
, **, * indicate significance at the 10%, 5%, and 1% level, respectively.*

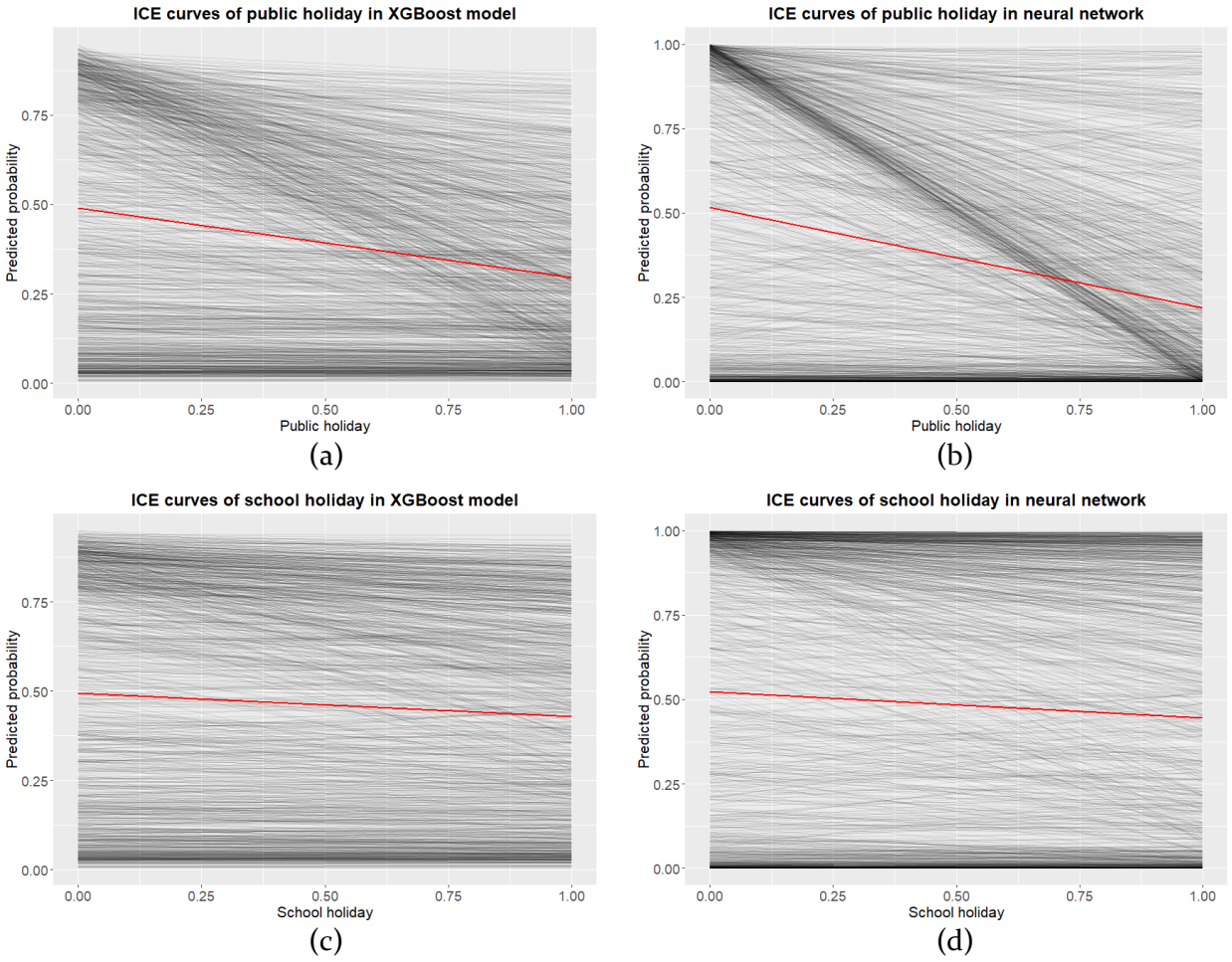


Figure 16: ICE curves of public and school holiday in XGBoost model and neural network

The third factor that is affecting the probability of traffic congestion arising through the demand-side is economic growth. The existing literature states that an increase in economic activity results in more congestion. The VIPs in Figure 13 show that the impact of GRP per capita is limited in the logistic regressions, but that GRP per capita substantially influences the predictions in the neural network and XGBoost model. The limited impact of GRP per capita in the logistic regressions is confirmed by the regression coefficients in Table 7. ICE curves of GRP per capita in **Error! Reference source not found.** show the effect of this variable on the predicted probability of traffic congestion arising in these models. Despite the high importance of GRP per capita in predicting traffic jams, the average predicted probability is relatively stable. On average, slightly higher probabilities are predicted for high GRP levels than for low GRP levels, but the difference is less than ten percentage points. The grey individual curves do show substantial changes between low and high GRP levels, though. As the changes are both positively and negatively related to the predicted probability of traffic congestion arising, it appears that other variables strongly interact with the relationship between GRP per capita and traffic congestion. For example, the combination of a high GRP per capita and low road availability in a region is likely to result in a high probability of traffic congestion occurring. However, high GRP levels combined with high road availability might cause fewer problems from a traffic flow perspective. This exemplifies how GRP per capita can be a major determinant of traffic jams arising, despite a clear direct relationship between the two is absent.

Table 7: Summary of both logistic regressions' coefficients containing all variables related to economic growth

	Dependent variable Congestion	
	Benchmark model	Lasso-regularised regression
GRP per capita	0.000***	0.058
	(0.000)	

Standard errors are reported in parentheses.
, **, * indicate significance at the 10%, 5%, and 1% level, respectively.*

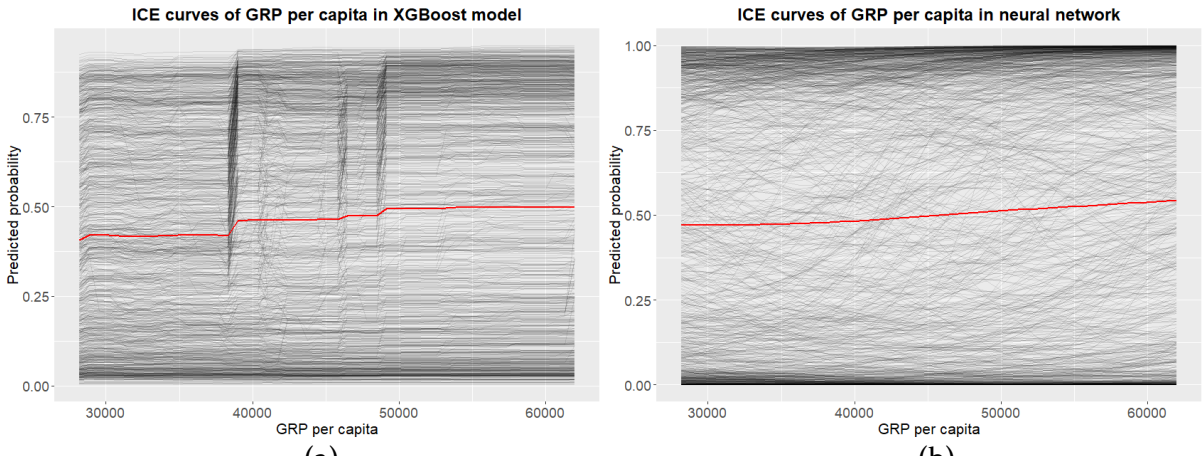


Figure 17: ICE curves of GRP per capita in XGBoost model and neural network

Lastly, weather conditions affect the probability of traffic congestion occurring both through supply and demand, according to the existing literature. Adverse weather conditions are found to decrease a road section's capacity but also changes people's individual travel behaviour. It is impossible to assess the effect of weather conditions on supply or demand based upon the models, as neither of them is directly incorporated into the models. Nevertheless, it is possible to evaluate how certain weather conditions affect the probability of traffic congestion arising. Based upon the sign of the effect, it can be argued whether the demand-side or supply-side effect is stronger. Variables about the temperature, wind speed, precipitation, or sunshine are frequently mentioned among the most important variables in Figure 13. The regression coefficients of the variables related to weather are depicted in Table 8. To evaluate the combined importance of all weather-related variables, an XGBoost model is trained with the same variables, fixed effects and hyperparameters, but excluding all weather-related variables. The accuracy on the unseen test data of this model is still 89.94%, which is only marginally lower than the accuracy of the original model. This suggests that weather variables are not a critical component of congestion prediction models.

Table 8: Summary of both logistic regressions' coefficients containing all variables related to weather conditions

	Dependent variable Congestion	
	Benchmark model	Lasso-regularised regression
Mean windspeed	-0.012	-0.024
	(0.007)	
Minimum hourly mean windspeed	-0.003	-0.008
	(0.005)	
Maximum hourly mean windspeed	-0.032***	-0.096
	(0.005)	
Maximum wind gust	0.023***	0.095
	(0.002)	
Wind direction north	-0.016	-0.015
	(0.013)	
Wind direction south	0.007	0.013
	(0.012)	
Wind direction west	-0.033***	-0.031
	(0.012)	
Mean temperature	-0.002	0.009
	(0.007)	
Minimum temperature	0.000	-0.010
	(0.003)	
Maximum temperature	0.009**	0.046
	(0.004)	
Sunshine duration	0.016***	0.070
	(0.002)	
Precipitation duration	0.029***	0.081
	(0.002)	

	Dependent variable Congestion	
	Benchmark model	Lasso-regularised regression
Precipitation amount	0.011***	0.052
	(0.002)	
Maximum hourly precipitation	-0.011***	-0.020
	(0.004)	
Minimum visibility	0.000***	-0.029
	(0.000)	
Maximum visibility	0.000***	0.023
	(0.000)	
Mean overcast	-0.005*	-0.013
	(0.003)	
Mean humidity	0.001	0.014
	(0.001)	
Minimum humidity	0.000	.
	(0.001)	

Standard errors are reported in parentheses.
, **, * indicate significance at the 10%, 5%, and 1% level, respectively.*
. indicates that the respective coefficient is shrunk to zero.

7. Conclusion

In recent years, some concerns have been voiced about the use of traffic flow data to predict traffic congestion. Although accurate and reliable traffic information has proven to be an efficient method to mitigate traffic congestion, gathering traffic flow data either is highly expensive or comes with serious privacy and data management issues. The results of this research indicate that the performance gap between predictive models trained on traffic flow data, and a model trained on more general determinants of traffic congestion is small. An XGBoost model is found to predict traffic congestion most accurately, out of a comparison between three machine learning techniques differing in complexity. The accuracy of this model (i.e., 90.39%) is only slightly lower than the accuracy of the best-performing model found in the existing literature (i.e., 93%). However, the XGBoost model appears to lack some sensitivity compared to the best-performing model based on traffic flow data. Despite scientific research conducted into this subject is missing, it seems logical that people prefer encountering a situation of free-traffic flow while congestion was predicted, instead of the opposite situation. According to this assumption, it would be beneficial to improve the model's sensitivity by changing the prediction threshold. However, doing so lowers the overall accuracy of the model, which might lower the users' trust in the system. For this reason, further research should be conducted into the optimal balance between accuracy, sensitivity and specificity when it comes to traffic congestion prediction models. Although the data used in this research is limited to the

extent that it only contains traffic jams in the Netherlands between January 2015 and February 2020, there is no reason to assume that the same results cannot be obtained for different countries or time frames. The high adaptability of the method proposed in this research is a major advantage over methods using traffic flow data. The methods used in this research can be applied to construct a model, with newly-optimised hyperparameters, for any other region or time frame.

When looking at the most important determinants of traffic congestion occurring, one predictor stands out in all compared models: road availability. It is clear that an increase in the number of highway kilometres in a region relative to the number of inhabitants decreases the probability of traffic congestion arising. By combining the findings of this research with the findings in the existing literature, it seems that building new highways is more efficient than adding more lanes to an already existing highway. Other factors that are found to be influential predictors of traffic congestion are GRP per capita in a region, changes in commuting behaviour, for example due to holidays, and road accidents.

Some policy recommendations can be made based upon the findings of this research. Firstly, although further research into this subject is definitely required, the results provide some preliminary evidence that simply constructing more highways mitigates the traffic congestion problem. Nevertheless, executing such a policy will presumably result in resistance from various other parties. In an era of increasing sustainable and environmental awareness, it seems undesirable to change the road infrastructure of a country into a frenzy of countless separate highways. Therefore, one of the main questions to be answered by further research into this subject is how a balance between sustainability and connectivity can be maintained from a road capacity perspective. This problem is generally approached from a demand perspective: coercive and non-coercive TDM measures are combined to lower the traffic flow on a road section. However, the effectiveness of such measures is not confirmed by this research. Secondly, the expected future GRP per capita of regions in the Netherlands should be monitored and infrastructural projects should be planned accordingly. GRP per capita is an important predictor of traffic congestion and a positive relationship appears to exist. Therefore, infrastructural investments have to be made in regions with strong economic growth. Preferably, these investments are made well in advance of the actual growth as existing literature warns that congestion slows economic growth. Lastly, employees should be nudged to spread their in-office working days over the week. The effect of commuting behaviour on the occurrence of traffic congestion is enormous. The COVID-19 pandemic provides an excellent opportunity to make

long-lasting changes to people's commuting behaviour. It is likely that people will continue working from home on a regular basis even after the pandemic. This could lower the number of traffic jams on highways, but it is essential to avoid that everyone decides to work in-office on the same days. Nudging employees, either directly or through their employers, appears to be the easiest manner to achieve such a change in commuting behaviour.

Finally, the findings in this research open up possibilities for companies and (semi-)governmental organisations to use models trained without traffic flow data to predict traffic congestion. It is recognised that this research does not offer a fully implementable traffic congestion prediction algorithm as of yet. However, the results do show that such an algorithm is capable of achieving an excellent predictive performance. Organisations that are currently predicting traffic congestion at least partly by traffic flow data, for example the ANWB in the Netherlands, could start experimenting with predictions based upon general supply and demand characteristics. The method proposed in this research solves many of the problems that arise when using traffic flow data, such as the high data collection costs and the relatively undetailed end result. An additional feature that could be incorporated in the model when applied on a large scale is information about the severity of traffic congestion. Currently, the model only predicts whether or not a traffic jam will occur, but does not provide information about the duration of the delay. An example of how such a prediction can be achieved is by changing the dependent variable into a multinomial variable. Potential classes of this variable could be 'no delay', 'between 0 and 19 minutes delay', 'between 20 and 39 minutes delay' and 'more than 40 minutes delay'. Almost the exact same method as in this research can be applied. If doing so, the model does not only indicate whether or not a traffic jam will arise, but also offers an approximation of the length of the delay.

References

Abdel-Aty, M. A., Kitamura, R., & Jovanis, P. P. (1997). Using stated preference data for studying the effect of advanced traffic information on drivers' route choice. *Transportation Research Part C: Emerging Technologies, 5*(1), 39-50. doi:10.1016/S0968-090X(96)00023-X

Abdelwahab, H. T., & Abdel-Aty, M. A. (2001). Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections. *Transportation Research Record: Journal of the Transportation Research Board, 1746*(1), 6-13. doi:10.3141/1746-02

Agarwal, M., Maze, T. H., & Souleyrette, R. (2005). *Impacts of weather on urban freeway traffic flow characteristics and facility capacity.* Retrieved from https://www.researchgate.net/profile/Reginald-Souleyrette/publication/228720996_Impact_of_Weather_on_Urban_Freeway_Traffic_Low_Characteristics_and_Facility_Capacity/links/0f31752f8fd061ea14000000/Impact-of-Weather-on-Urban-Freeway-Traffic-Low-Characterist

Åkerstedt, T., & Kecklund, G. (2001). Age, gender and early morning highway accidents. *Journal of Sleep Research, 10*(2), 105-110. doi:10.1046/j.1365-2869.2001.00248.x

Akhtar, M., & Maridpour, S. (2021). A Review of Traffic Congestion Prediction Using Artificial Intelligence. *Journal of Advanced Transportation, 2021*. doi:10.1155/2021/8878011

Alarcon-Aquino, V., & Barria, J. A. (2006). Multiresolution FIR neural-network-based learning algorithm applied to network traffic prediction. *IEEE Transactions on Systems, Man, and Cybernetics, Part C, 36*(2), 208-220. doi:10.1109/TSMCC.2004.843217

Anderson, T. K. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention, 41*(3), 359-364. doi:10.1016/j.aap.2008.12.014

Archondo-Callao, R. (2008). *Applying the HDM-4 Model to Strategic Planning of Road Works.* Retrieved from http://hdl.handle.net/10986/17419

Badland, H. M., Garrett, N., & Schofield, G. M. (2010). How Does Car Parking Availability and Public Transport Accessibility Influence Work-Related Travel Behaviors? *Sustainability, 2*(2), 576-590. doi:10.3390/su2020576

Barker, T., & Gerhold, D. (1993). *The rise and rise of road transport, 1700-1990.* Cambridge, UK: Press Syndicate of the University of Cambridge.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter, 6*(1), 20-29. doi:10.1145/1007730.1007735

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences, 116*(32), 15849-15854. doi:10.1073/pnas.1903070116

Ben-Elia, E., & Ettema, D. (2011). Rewarding rush-hour avoidance: A study of commuters' travel behavior. *Transportation Research Part A: Policy and Practice, 45*(7), 567-582. doi:10.1016/j.tra.2011.03.003

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123-140. doi:10.1007/BF00058655

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5-32. doi:10.1023/A:1010933404324

Brynjolfsson, E., Horton, J. J., Ozimek, A., Rock, D., Sharma, G., & TuYe, H.-Y. (2020). COVID-19 and Remote Work: An Early Look at US Data. *National Bureau of Economic Research*. doi:10.3386/w27344

Chakraborty, P., Adu-Gyamfi, Y. O., Poddar, S., Ahsani, V., Sharma, A., & Sarkar, S. (2018). Traffic Congestion Detection from Camera Images using Deep Convolution Neural

Networks. *Transportation Research Record: Journal of the Transportation Research Board, 2672*(45), 222-231. doi:10.1177/0361198118777631

Chen, P. S.-T., Srinivasan, K. K., & Mahmassani, H. S. (1999). Effect of Information Quality on Compliance Behavior of Commuters Under Real-Time Traffic Information. *Transportation Research Record: Journal of the Transportation Research Board, 1676*, 53-60. doi:10.3141/1676-07

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785-794. doi:10.1145/2939672.2939785

Cheng, W., & Washington, S. P. (2005). Experimental evaluation of hotspot identification methods. *Accident Analysis & Prevention, 37*(5), 870-881. doi:10.1016/j.aap.2005.04.015

Chung, E., Ohtani, O., Warita, H., Kuwahara, M., & Morita, H. (2006). Does Weather Affect Highway Capacity? *Proceedings of the 5th International Symposium on Highway Capacity and Quality of Service, 1*, 139-146. Retrieved from http://www.plan.civil.tohoku.ac.jp/kuwahara/publications/2006-026.pdf

Cools, M., Moons, E., Creemers, L., & Wets, G. (2010). Changes in Travel Behavior in Response to Weather Conditions: Do Type of Weather and Trip Purpose Matter? *Transportation Research Record: Journal of the Transportation Research Board, 2157*(1), 22-28. doi:10.3141/2157-03

De Jong, G., & Gunn, H. (2001). Recent Evidence on Car Cost and Time Elasticities of Travel Demand in Europe. *Journal of Transport Economics and Policy, 35*(2), 137-160. Retrieved from https://www.ingentaconnect.com/content/lse/jtep/2001/00000035/00000002/art00001#expand/collapse

De Vos, J. (2020). The effect of COVID-19 and subsequent social distancing on travel behavior. *Transportation Research Interdisciplinary Perspectives, 5*, 100-121. doi:10.1016/j.trip.2020.100121

Deming, W. E. (1975). On Some Statistical Aids Toward Economic Production. *Journal on Applied Analytics, 5*(4), 1-15. doi:10.1287/inte.5.4.1

Directorate-General for Public Works and Water Management (Rijkswaterstaat). (2020). *Ontwikkeling verkeersdrukte 2019*. Retrieved from https://www.rijkswaterstaat.nl/nieuws/2020/03/ontwikkeling-verkeersdrukte-2019.aspx

Dong, N., Huang, H., Lee, J., Gao, M., & Abdel-Aty, M. (2016). Macroscopic hotspots identification: A Bayesian spatio-temporal interaction approach. *Accident Analysis & Prevention, 92*, 256-264. doi:10.1016/j.aap.2016.04.001

Duranton, G., & Turner, M. A. (2011). The Fundamental Law of Road Congestion: Evidence from US Cities. *American Economic Review, 101*(6), 2616-2652. doi:10.1257/aer.101.6.2616

Elfar, A., Talebpour, A., & Mahmassani, H. S. (2018). Machine Learning Approach to Short-Term Traffic Congestion Prediction in a Connected Environment. *Transportation Research Record: Journal of the Transportation Research Board, 2672*(45), 185-195. doi:10.1177/0361198118795010

European Commission. (2020). *Eurostat*. Retrieved from https://ec.europa.eu/eurostat/data/database

Felstead, A., & Reuschke, D. (2020). *Homeworking in the UK: before and during the 2020 lockdown.* Retrieved from http://orca.cf.ac.uk/id/eprint/134545

Fouladgar, M., Parchami, M., Elmasri, R., & Ghaderi, A. (2017). Scalable deep traffic flow neural networks for urban traffic congestion prediction. *2017 International Joint Conference on Neural Networks (IJCNN)*, 2251-2258. doi:10.1109/IJCNN.2017.7966128

Freund, Y., & Schapire, R. E. (1996). *Experiments with a New Boosting Algorithm.* Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.51.6252&rep=rep1&type=pdf

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics, 29*(5), 1189-1232. Retrieved from https://www.jstor.org/stable/2699986

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software, 33*(1), 1-22. doi:10.18637/jss.v033.i01

Gärling, T., & Schuitema, G. (2007). Travel Demand Management Targeting Reduced Private Car Use: Effectiveness, Public Acceptability and Political Feasibility. *Journal of Social Issues, 63*(1), 139-153. doi:10.1111/j.1540-4560.2007.00500.x

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 15*, 315-323. Retrieved from http://proceedings.mlr.press/v15/glorot11a

Goeman, J. J., Meijer, R., & Chaturvedi, N. (2016). L1 and L2 Penalized Regression Models. *Biometrical Journal, 52*(1), 70-84. Retrieved from https://mran.microsoft.com/snapshot/2016-05-19/web/packages/penalized/vignettes/penalized.pdf

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics, 24*(1), 44-65. doi:10.1080/10618600.2014.907095

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* Cambridge, MA: MIT Press.

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Philips, T., . . . Tatlow, H. (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour*. doi:10.1038/s41562-021-01079-8

Hansen, M., & Huang, Y. (1997). Road supply and traffic in California urban areas. *Transportation Research Part A: Policy and Practice, 31*(3), 205-218. doi:10.1016/S0965-8564(96)00019-5

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The Elements of Statistical Learning.* New York, NY: Springer.

Hranac, R., Sterzin, E., Krechmer, D., Rakha, H., & Farzaneh, M. (2006). *Empirical Studies on Traffic Flow in Inclement Weather.* Retrieved from https://rosap.ntl.bts.gov/view/dot/42251

Jarrett, K., Kavukcuoglu, K., Ranzato, M., & LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? *2009 IEEE 12th International Conference on Computer Vision*, 2146-2153. Retrieved from https://ieeexplore.ieee.org/abstract/document/5459469

Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization, 13*, 455-492. doi:10.1023/A:1008306431147

Jou, R.-C., Lam, S.-H., Liu, Y.-H., & Chen, K.-H. (2005). Route switching behavior on freeways with the provision of different types of real-time traffic information. *Transportation Research Part A: Policy and Practice, 39*(5), 445-461. doi:10.1016/j.tra.2005.02.004

Kerner, B. S. (2009). *Introduction to Modern Traffic Flow Theory and Control: The Long Road to Three-Phase Traffic Theory.* Heidelberg: Springer.

Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, 1-13. Retrieved from https://arxiv.org/abs/1412.6980

Kitamura, R., Fujii, S., & Pas, E. I. (1997). Time-use data, analysis and modeling: toward the next generation of transportation planning methodologies. *Transport Policy, 4*(4), 225-235. doi:10.1016/S0967-070X(97)00018-8

Kumleben, N., Bhopal, R., Czypionka, T., Gruer, L., Kock, R., Stebbing, J., & Stigler, F. L. (2020). Test, test, test for COVID-19 antibodies: the importance of sensitivity, specificity and predictive powers. *Public Health, 185*, 88-90. doi:10.1016/j.puhe.2020.06.006

Kwon, T. J., Liping, F., & Jiang, C. (2013). Effect of Winter Weather and Road Surface Conditions on Macroscopic Traffic Parameters. *Transportation Research Record: Journal of the Transportation Research Board, 2329*(1), 54-62. doi:10.3141/2329-07

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized Regression, Standard Errors, and Bayesian Lassos. *Bayesian Analysis, 5*(2), 369-412. doi:10.1214/10-BA607

Levinson, D. M., & Gillen, D. (1998). The full cost of intercity highway transportation. *Transportation Research Part D: Transport and Environment, 3*(4), 207-223. doi:10.1016/S1361-9209(97)00037-0

Mangal, A., & Kumar, N. (2016). *Using big data to enhance the bosch production line performance: A Kaggle challenge.* doi:10.1109/BigData.2016.7840826

Marchesini, P., & Weijermars, W. (2010). *The relationship between road safety and congestion on motorways.* Retrieved from https://www.swov.nl/sites/default/files/publicaties/rapport/r-2010-12.pdf

Maze, T. H., Agarwal, M., & Burchett, G. (2006). Whether Weather Matters to Traffic Demand, Traffic Safety, and Traffic Operations and Flow. *Transportation Research Record: Journal of the Transportation Research Board, 1948*(1), 170-176. doi:10.1177/0361198106194800119

McCulloch, W. S., & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics, 5*, 115-133. doi:10.1007/BF02478259

McFadden, D. (1974). The measurement of urban travel demand. *Journal of Public Economics, 3*(4), 303-328. doi:10.1016/0047-2727(74)90003-6

McKenzie, B., & Rapino, M. (2011). *Commuting in the United States: 2009.* Retrieved from https://www.infrastructureusa.org/wp-content/uploads/2011/09/census-commuting.pdf

Miao, Q., Welch, E. W., & Sriraj, P. S. (2019). Extreme weather, public transport ridership and moderating effect of bus stop shelters. *Journal of Transport Geography, 74*, 125-133. doi:10.1016/j.jtrangeo.2018.11.007

Min, W., & Wynter, L. (2011). Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C: Emerging Technologies, 19*(4), 606-616. doi:10.1016/j.trc.2010.10.002

Ministry of Infrastructure and Water Management. (2019). *Mobiliteitsbeeld 2019*. Retrieved from https://www.kimnet.nl/publicaties/rapporten/2019/11/12/mobiliteitsbeeld-2019-vooral-het-gebruik-van-de-trein-neemt-toe

Močkus, J. (1975). On Bayesian Methods for Seeking the Extremum. In G. I. Marchuk, *Optimization Techniques IFIP Technical Conference. Lecture Notes in Computer Science* (pp. 400-404). Berlin: Springer.

Montella, A. (2010). A comparative analysis of hotspot identification methods. *Accident Analysis & Prevention, 42*(2), 571-581. doi:10.1016/j.aap.2009.09.025

Nair, V., & Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. *Proceedings of the 27 th International Conference on Machine Learning*. Retrieved from https://www.cs.toronto.edu/~hinton/absps/reluICML.pdf

Nielsen, D. (2016). *Tree Boosting With XGBoost: Why Does XGBoost Win "Every" Machine Learning Competition?* Retrieved from https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2433761/16128_FULLTEXT.pdf

Noland, R. B. (2001). Relationships between highway capacity and induced vehicle travel. *Transportation Research Part A: Policy and Practice, 35*(1), 47-72. doi:10.1016/S0965-8564(99)00047-6

Oakil, A. T., Nijland, L., & Dijst, M. (2016). Rush hour commuting in the Netherlands: Gender-specific household activities and personal attitudes towards responsibility sharing. *Travel Behaviour and Society, 4*, 79-87. doi:10.1016/j.tbs.2015.10.003

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology, 56*(1), 45-50. doi:10.4103/0301-4738.37595

Park, S.-H., Kim, S.-M., & Ha, Y.-G. (2016). Highway traffic accident prediction using VDS big data analysis. *The Journal of Supercomputing, 72*, 2815-2831. doi:10.1007/s11227-016-1624-z

Polydoropoulou, A., Ben-Akiva, M., & Kaysi, I. (1994). Influence of Traffic Information on Drivers' Route Choice Behavior. *Transportation Research Record: Journal of the Transportation Research Board*(1453), 56-65. Retrieved from http://onlinepubs.trb.org/Onlinepubs/trr/1994/1453/1453-006.pdf

Redman, L., Friman, M., Gärling, T., & Hartig, T. (2013). Quality attributes of public transport that attract car users: A research review. *Transport Policy, 25*, 119-127. doi:10.1016/j.tranpol.2012.11.005

Reid, S., Tibshirani, R., & Friedman, J. H. (2016). A Study of Error Variance Estimation in Lasso Regression. *Statistica Sinica, 26*(1), 35-67. doi:10.5705/ss.2014.042

Robbins, H., & Monro, S. (1951). A Stochastic Approximation Method. *The Annal of Mathematical Statistics, 22*(3), 400-407. doi:10.1214/aoms/1177729586

Rothenberg, J. (1970). The Economics of Congestion and Pollution: An Integrated View. *The American Economic Review, 60*(2), 114-121. Retrieved from https://www.jstor.org/stable/1815795

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533-536. doi:10.1038/323533a0

Shibayama, T., Sandholzer, F., Laa, B., & Brezina, T. (2021). Impact of COVID-19 lockdown on commuting: A multi-country perspective. *European Journal of Transport and Infrastructure Research, 21*(1), 70-93. doi:10.18757/ejtir.2021.21.1.5135

Smith, A. C. (2016). Spring Forward at Your Own Risk: Daylight Saving Time and Fatal Vehicle Crashes. *American Economic Journal: Applied Economics, 8*(2), 65-91. doi:10.1257/app.20140100

Smith, B. L., Byrne, K. G., Copperman, R. B., Hennessy, S. M., & Goodall, N. J. (2004). *An Investigation into the Impact of Rainfall on Freeway Traffic Flow.* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.721.983&rep=rep1&type=pdf

Snelder, M., Van Zuylen, H. J., & Immers, L. H. (2012). A framework for robustness analysis of road networks for short term variations in supply. *Transportation Research Part A: Policy and Practice, 46*(5), 828-842. doi:10.1016/j.tra.2012.02.007

Snoek, J., Larochelle, H., & Adams, R. P. (2012). *Practical Bayesian Optimization of Machine Learning Algorithms.* Retrieved from https://arxiv.org/pdf/1206.2944.pdf

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research, 15*(1), 1929-2958. Retrieved from https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf?utm_campaign=buffer&utm_content=buffer79b43&utm_medium=social&utm_source=twitter.com

Stopher, P., & Stanley, J. (2014). *Introduction to Transport Policy.* Northampton, MA: Edward Elgar Publishing.

Strickland, S. G., & Berman, W. (1995). Congestion Control and Demand Management. *Public Roads, 58*(3), 1-7. Retrieved from https://trid.trb.org/view/514442

Sugiyama, Y., Fukui, M., Kikuchi, M., Hasebe, K., Nakayama, A., Nishinari, K., . . . Yukawa, S. (2008). Traffic jams without bottlenecks—experimental evidence for the physical mechanism of the formation of a jam. *New Journal of Physics, 10*. doi:10.1088/1367-2630/10/3/033001

Sweet, M. (2011). Does Traffic Congestion Slow the Economy? *Journal of Planning Literature, 26*(4), 391-404. doi:10.1177/0885412211409754

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288. doi:10.1111/j.2517-6161.1996.tb02080.x

Tseng, F.-H., Hsueh, J.-H., Tseng, C.-W., Yang, Y.-T., Chao, C.-H., & Chou, L.-D. (2018). Congestion Prediction With Big Data for Real-Time Highway Traffic. *IEEE Access, 6*, 57311-57323. doi:10.1109/ACCESS.2018.2873569

Tseng, P. (2001). Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. *Journal of Optimization Theory and Applications, 109*, 475-494. doi:10.1023/A:1017501703105

Van Stralen, W. J., Calvert, S. C., & Molin, E. J. (2015). The influence of adverse weather conditions on probability of congestion on Dutch motorways. *European Journal of Transport and Infrastructure Research, 15*(4), 482-500. Retrieved from https://ojs-libaccp.tudelft.nl/index.php/ejtir/article/view/3093/3279

Wolpert, D. H. (1996). The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation, 8*(7), 1341-1390. doi:10.1162/neco.1996.8.7.1341

Yang, I.-H., Yeo, M.-S., & Kim, K.-W. (2003). Application of artificial neural network to predict the optimal start time for heating system in building. *Energy Conversion and Management, 44*(17), 2791-2809. doi:10.1016/S0196-8904(03)00044-X

Yousif, S. (2002). Motorway roadworks: effects on traffic operations. *Highways and Transportation*, 20-22. Retrieved from http://usir.salford.ac.uk/id/eprint/9704

Zhang, L., Liu, Q., Yang, W., Wei, N., & Dong, D. (2013). An Improved K-nearest Neighbor Model for Short-term Traffic Flow Prediction. *Procedia - Social and Behavioral Sciences, 96*(6), 653-662. doi:10.1016/j.sbspro.2013.08.076

Appendix A: Results of ARIMA model to extrapolate car and public transport user costs data

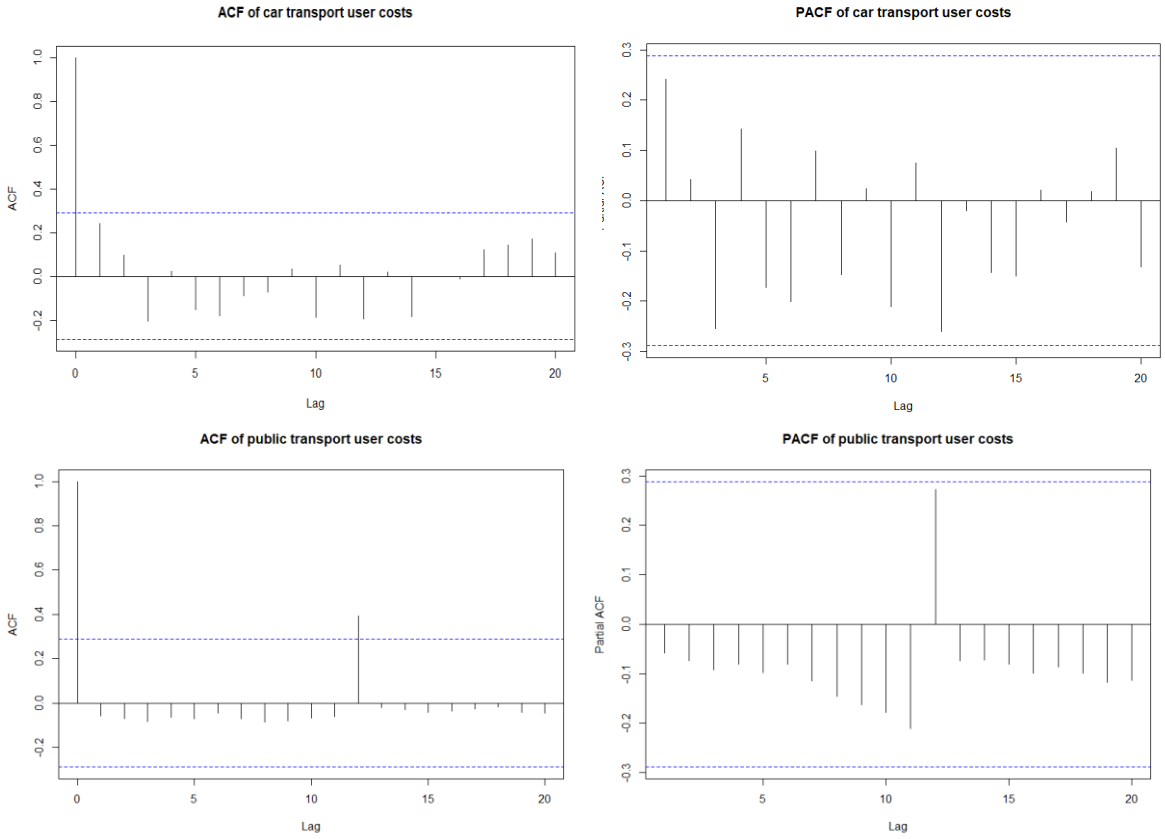


Figure 18: ACF and PACF of car and public transport user costs

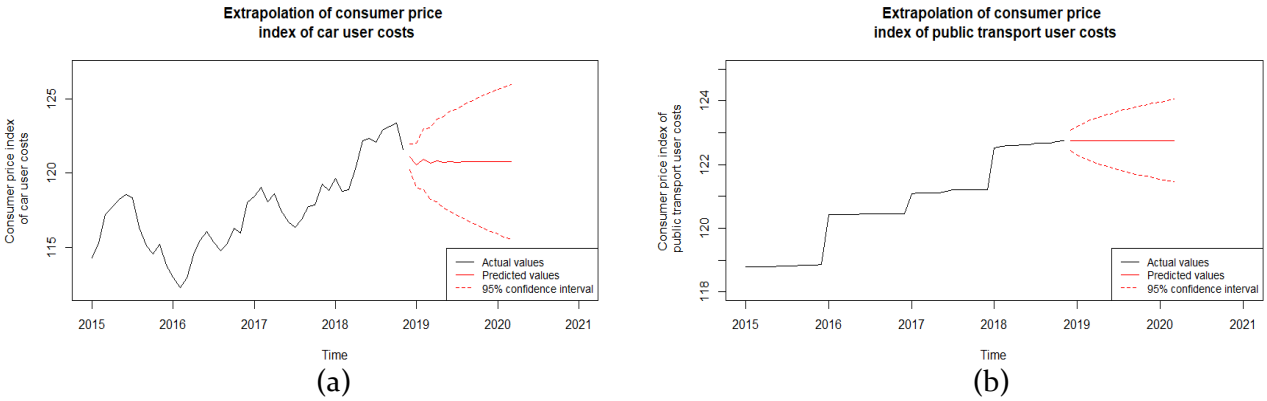


Figure 19: Extrapolations of the consumer price index of public transport and car user costs per kilometre

Appendix B: Coefficients of the fixed effects single regression model of GRP per capita

Table 9: Regression results of the province-fixed effects single regression model of GRP per capita

	Dependent variable GRP per capita
Constant	25,689.07***
	(507.05)
t (2015 is t = 1, 2016 is t = 2, etc.)	1,351.51***
	(88.27)
Province: Flevoland	2,668.20***
	(611.52)
Province: Friesland	61.60
	(611.52)
Province: Gelderland	6,689.00***
	(611.52)
Province: Groningen	11,939.60***
	(611.52)
Province: Limburg	8,338.00***
	(611.52)
Province: Noord-Brabant	14,104.00***
	(611.52)
Province: Noord-Holland	27,336.80***
	(611.52)
Province: Overijssel	6,343.20***
	(611.52)
Province: Utrecht	23,418.40***
	(611.52)
Province: Zeeland	4,866.60***
	(611.52)
Province: Zuid-Holland	13,233.60***
	(611.52)

Standard errors are reported in parentheses.
, **, * indicate significance at the 10%, 5%, and 1% level, respectively.*

Appendix C: Descriptive statistics of balanced and unbalanced data set

Table 10: Mean and standard deviation of continuous variables in unbalanced and balanced data set

	Unbalanced			Balanced		
	All	Congestion		All	Congestion	
		0	1		0	1
Road conditions						
Congestion (1 = congestion, 0 = no congestion)	0.032 (0.174)	0.000 (0.000)	1.000 (0.000)	0.500 (0.500)	0.000 (0.000)	1.000 (0.000)
Road works (1 = road works, 0 = no road works)	0.285 (0.442)	0.248 (0.431)	0.309 (0.462)	0.282 (0.450)	0.249 (0.433)	0.314 (0.464)
Weather conditions						
Daily mean windspeed (in m/s)	4.521 (2.349)	4.516 (2.346)	4.686 (2.444)	4.585 (2.388)	4.494 (2.333)	4.677 (2.438)
Daily minimum hourly mean windspeed (in m/s)	2.332 (1.889)	6.757 (3.055)	6.947 (3.196)	2.377 (1.916)	6.731 (3.045)	6.936 (3.192)
Daily maximum hourly mean windspeed (in m/s)	6.763 (3.060)	2.328 (1.886)	2.448 (1.963)	6.834 (3.124)	2.314 (1.874)	2.440 (1.955)
Daily maximum wind gust (in m/s)	11.383 (4.402)	11.376 (4.400)	11.568 (4.459)	11.449 (4.427)	11.343 (4.393)	11.554 (4.458)
Daily mean temperature (in degrees Celsius)	10.910 (6.053)	10.913 (6.065)	10.820 (5.683)	10.926 (5.888)	11.002 (6.078)	10.849 (5.691)
Daily minimum temperature (in degrees Celsius)	7.101 (5.634)	7.096 (5.643)	7.256 (5.363)	7.230 (5.509)	7.184 (5.641)	7.277 (5.374)
Daily maximum temperature (in degrees Celsius)	14.582 (7.026)	14.592 (7.039)	14.289 (6.595)	14.504 (6.840)	14.683 (7.064)	14.325 (6.604)
Daily sunshine duration (in hours)	5.152 (4.238)	5.154 (4.237)	5.091 (4.269)	5.117 (4.252)	5.134 (4.232)	5.101 (4.272)
Daily precipitation duration (in hours)	1.706 (2.842)	1.701 (2.836)	1.850 (3.009)	1.762 (2.898)	1.700 (2.820)	1.824 (2.971)
Daily precipitation (in mm)	2.161 (4.371)	2.154 (4.359)	2.397 (4.711)	2.267 (4.537)	2.161 (4.372)	2.373 (4.694)
Daily maximum hourly precipitation (in mm)	0.841 (1.756)	0.839 (1.751)	0.898 (1.896)	0.870 (1.836)	0.847 (1.768)	0.894 (1.901)
Daily minimum visibility (in km)	4.314 (2.227)	4.314 (2.228)	4.334 (2.190)	4.315 (2.211)	4.298 (2.226)	4.332 (2.195)
Daily maximum visibility (in km)	7.559 (0.809)	7.560 (0.810)	7.529 (0.793)	7.542 (0.803)	7.554 (0.813)	7.530 (0.794)
Daily average overcast (on a scale from 0 to 9)	5.920 (2.184)	5.918 (2.185)	5.989 (2.163)	5.963 (2.164)	5.940 (2.164)	5.986 (2.164)
Daily average humidity (in %)	79.799 (10.031)	79.785 (10.040)	80.246 (9.755)	80.064 (9.828)	79.909 (9.898)	80.220 (9.756)
Daily minimum humidity (in %)	63.850 (15.025)	63.819 (15.033)	64.796 (14.751)	64.351 (14.861)	63.954 (14.957)	64.748 (14.753)
Daily maximum humidity (in %)	93.467 (6.113)	93.471 (6.118)	93.370 (5.964)	93.471 (5.947)	93.574 (5.925)	93.369 (5.968)
User costs						
Public transport user costs (index with base year 2009)	121.240 (1.471)	121.236 (1.471)	121.370 (1.442)	121.286 (1.458)	121.213 (1.470)	121.360 (1.443)
Car user costs (index with base year 2009)	118.373 (2.778)	118.366 (2.778)	118.603 (2.751)	118.459 (2.772)	118.331 (2.783)	118.588 (2.755)
Gasoline price (in euro/litre)	1.574 (0.080)	1.574 (0.080)	1.580 (0.079)	1.577 (0.080)	1.574 (0.081)	1.580 (0.079)

	Unbalanced			Balanced		
	All	Congestion		All	Congestion	
		0	1		0	1
Diesel price (in euro/litre)	1.260 (0.096)	1.259 (0.096)	1.271 (0.096)	1.265 (0.096)	1.259 (0.096)	1.270 (0.096)
LPG price (in euro/litre)	0.630 (0.049)	0.629 (0.049)	0.634 (0.048)	0.632 (0.049)	0.630 (0.049)	0.634 (0.048)
Road availability (highway kms per 100,000 inhabitants)	33.857 (9.206)	34.013 (9.242)	29.138 (6.413)	31.595 (8.330)	34.049 (9.253)	29.141 (6.414)
Distance to public transport (in km)	5.392 (2.443)	5.412 (2.470)	4.783 (1.234)	5.101 (1.984)	5.419 (2.479)	4.783 (1.235)
Special days						
School holiday (1 = school holiday, 0 = no school holiday)	0.304 (0.460)	0.308 (0.462)	0.182 (0.385)	0.248 (0.432)	0.313 (0.464)	0.183 (0.387)
Public holiday (1 = public holiday, 0 = no public holiday)	0.023 (0.151)	0.024 (0.153)	0.007 (0.083)	0.015 (0.123)	0.024 (0.152)	0.007 (0.083)
Daylight saving time start (1 = start, 0 = no start)	0.003 (0.056)	0.003 (0.057)	<0.001 (0.017)	0.001 (0.038)	0.003 (0.052)	<0.001 (0.017)
Economic growth						
GRP per capita (in thousands of euros)	42.158 (8.012)	41.625 (8.120)	44.958 (7.396)	43.295 (7.947)	41.629 (8.127)	44.960 (7.397)

Standard deviations are reported in parentheses.

Table 11: Proportion of the wind directions in unbalanced and balanced data set

		North	East	South	West
Unbalanced	All observations	15.207%	17.147%	31.411%	34.274%
	Observations with congestion	15.668%	17.354%	32.618%	34.507%
	Observations without congestion	15.193%	17.142%	31.374%	34.268%
Balanced	All observations	15.541%	17.354%	32.345%	34.760%
	Observations with congestion	15.692%	17.319%	32.599%	34.389%
	Observations without congestion	15.390%	17.388%	32.091%	35.131%

Appendix D: Coefficients of logistic regressions

Table 12: Results of the various logistic regressions

	Dependent variable		
	Congestion		
	Benchmark model	Lasso-regularised regression	Comparative unregularised regression
Constant	245.695***	-2.801	-2.875***
	(21.603)		(0.224)
Road works	0.290***	0.195	0.199***
	(0.022)		(0.022)
Mean windspeed	-0.012	-0.024	0.026
	(0.007)		(0.020)
Minimum hourly mean windspeed	-0.003	-0.008	-0.008
	(0.005)		(0.017)
Maximum hourly mean windspeed	-0.032***	-0.096	-0.100***
	(0.005)		(0.010)
Maximum wind gust	0.023***	0.095	0.102***
	(0.002)		(0.011)
Wind direction north	-0.016	-0.015	-0.021
	(0.013)		(0.015)
Wind direction south	0.007	0.013	0.005
	(0.012)		(0.013)
Wind direction west	-0.033***	-0.031	-0.036**
	(0.012)		(0.013)
Mean temperature	-0.002	0.009	-0.006
	(0.007)		(0.046)
Minimum temperature	0.000	-0.010	-0.007
	(0.003)		(0.021)
Maximum temperature	0.009**	0.046	0.058*
	(0.004)		(0.033)
Sunshine duration	0.016***	0.070	0.072***
	(0.002)		(0.009)
Precipitation duration	0.029***	0.081	0.083***
	(0.002)		(0.008)
Precipitation amount	0.011***	0.052	0.048**
	(0.002)		(0.012)
Maximum hourly precipitation	-0.011***	-0.020	-0.020**
	(0.004)		(0.009)
Minimum visibility	0.000***	-0.029	-0.028***
	(0.000)		(0.006)
Maximum visibility	0.000***	0.023	0.027**
	(0.000)		(0.006)
Mean overcast	-0.005*	-0.013	-0.010
	(0.003)		(0.007)
Mean humidity	0.001	0.014	0.015
	(0.001)		(0.017)
Minimum humidity	0.000	.	0.005
	(0.001)		(0.016)

	Dependent variable		
	Congestion		
	Benchmark model	Lasso-regularised regression	Comparative unregularised regression
Public transport user costs	-2.063***	0.007	-0.011
	(0.181)		(0.030)
Car user costs	-0.017***	-0.045	-0.043**
	(0.006)		(0.020)
Gasoline price	-1.128***	0.003	0.005
	(0.206)		(0.016)
Diesel price	1.279***	0.089	0.088**
	(0.202)		(0.022)
LPG price	0.750***	-0.023	-0.021
	(0.228)		(0.011)
Road availability	-0.088***	-0.736	-0.737***
	(0.001)		(0.010)
Distance to public transport	0.070	0.140	0.140***
	(0.004)		(0.008)
School holiday	-0.541***	-0.534	-0.537***
	(0.011)		(0.012)
Public holiday	-1.002***	-0.981	-1.004***
	(0.029)		(0.033)
Daylight saving time start	-0.128	-0.169	-0.141
	(0.119)		(0.103)
GRP per capita	0.000***	0.058	0.059***
	(0.000)		(0.007)
Observations	788,774	788,774	788,774
Optimisation method	Gradient descent	Cyclic coordinate descent	Cyclic coordinate descent
Regularisation	No	Yes	No
Road fixed effects	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes
Weekday fixed effects	Yes	Yes	Yes
Month fixed effects	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes

Standard errors are reported in parentheses[8].
, **, * indicate significance at the 10%, 5%, and 1% level, respectively.*
. indicates that the respective coefficient is shrunk to zero.

[8] As thoroughly discussed in Section 4.1., it is preferred not to report standard errors of the regularised logistic regression.

Appendix E: Mean ICE per variable of the neural network and XGBoost model

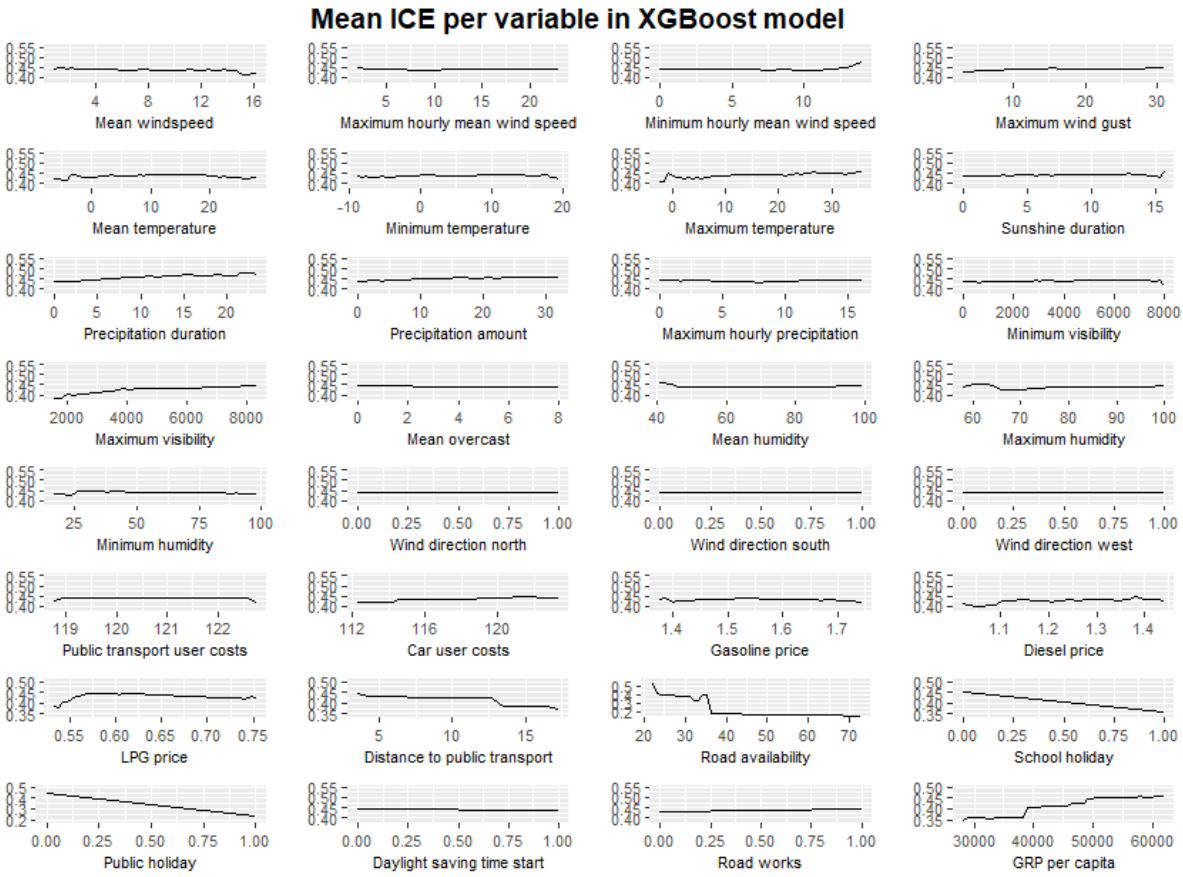


Figure 20: Mean ICE per variable in XGBoost model. The vertical axis represents the predicted probability. Note that the scale of the vertical axis of LPG price, distance to public transport, road availability, public holiday, and GRP per capita differs from the scale of the other variables.

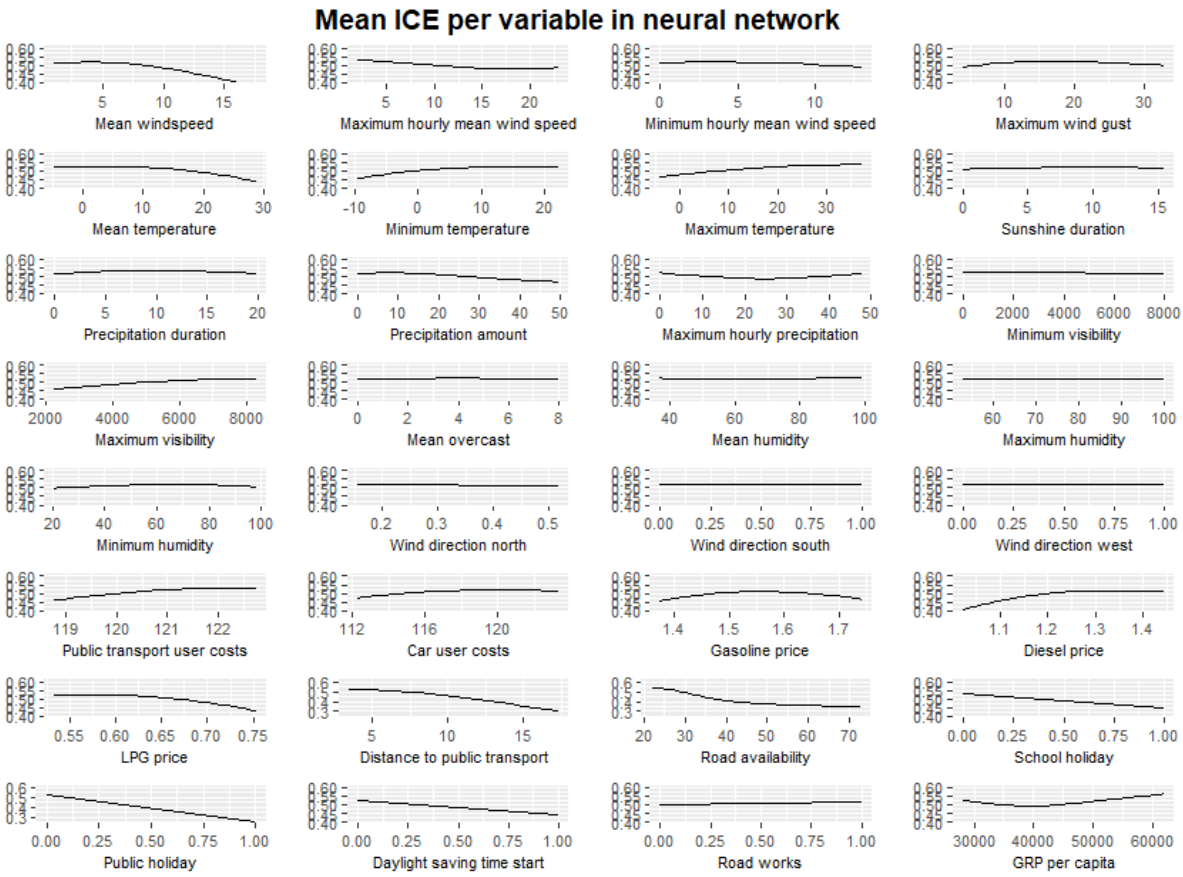


Figure 21: Mean ICE per variable in neural network. The vertical axis represents the predicted probability. Note that the scale of the vertical axis of distance to public transport, road availability and public holiday differs from the scale of the other variables.

Appendix F: Logistic loss per variable quartile per model

Table 13: Logistic loss per variable quartile per model

	Benchmark model				Regularised logistic regression				XGBoost model				Neural network			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Daily mean windspeed	0.349	0.358	0.348	0.332	0.349	0.358	0.348	0.333	0.236	0.245	0.234	0.228	0.267	0.271	0.262	0.257
Daily minimum hourly mean windspeed	0.352	0.352	0.347	0.337	0.352	0.352	0.347	0.334	0.238	0.245	0.232	0.229	0.266	0.275	0.260	0.259
Daily maximum hourly mean windspeed	0.352	0.353	0.340	0.337	0.352	0.353	0.340	0.334	0.240	0.237	0.228	0.229	0.270	0.266	0.256	0.257
Daily maximum wind gust	0.352	0.355	0.344	0.332	0.352	0.355	0.344	0.330	0.240	0.241	0.235	0.223	0.269	0.268	0.265	0.255
Daily mean temperature	0.321	0.331	0.342	0.394	0.321	0.331	0.342	0.394	0.220	0.226	0.238	0.258	0.249	0.250	0.266	0.295
Daily minimum temperature	0.327	0.328	0.345	0.386	0.327	0.328	0.345	0.387	0.224	0.222	0.237	0.259	0.254	0.247	0.265	0.292
Daily maximum temperature	0.322	0.327	0.348	0.391	0.322	0.327	0.348	0.390	0.221	0.224	0.241	0.255	0.250	0.248	0.270	0.292
Daily sunshine duration	0.319	0.349	0.343	0.377	0.319	0.349	0.343	0.377	0.217	0.240	0.232	0.254	0.243	0.270	0.263	0.285
Daily precipitation duration	0.354	0.328	0.349	0.333	0.354	0.328	0.349	0.333	0.239	0.213	0.239	0.227	0.269	0.241	0.265	0.258
Daily precipitation	0.354	0.336	0.347	0.334	0.354	0.336	0.347	0.334	0.239	0.224	0.238	0.228	0.269	0.251	0.265	0.257
Daily maximum hourly precipitation	0.354	0.344	0.339	0.339	0.354	0.344	0.339	0.341	0.239	0.229	0.231	0.234	0.269	0.258	0.256	0.263
Daily minimum visibility	0.335	0.338	0.351	0.363	0.335	0.338	0.351	0.364	0.230	0.233	0.240	0.240	0.258	0.264	0.267	0.270
Daily maximum visibility	0.320	0.339	0.364	0.367	0.320	0.339	0.364	0.365	0.221	0.231	0.246	0.246	0.248	0.261	0.273	0.278
Daily average overcast	0.362	0.348	0.330	0.330	0.362	0.348	0.330	0.330	0.244	0.234	0.228	0.228	0.273	0.266	0.254	0.254
Daily average humidity	0.374	0.350	0.338	0.323	0.374	0.350	0.338	0.319	0.252	0.238	0.228	0.222	0.284	0.267	0.255	0.252
Daily minimum humidity	0.372	0.355	0.338	0.322	0.372	0.355	0.338	0.320	0.250	0.242	0.225	0.223	0.283	0.270	0.255	0.250
Daily maximum humidity	0.356	0.349	0.343	0.336	0.356	0.349	0.343	0.332	0.241	0.236	0.233	0.229	0.269	0.266	0.262	0.258
Public transport user costs	0.344	0.358	0.338	0.341	0.345	0.358	0.338	0.342	0.229	0.244	0.234	0.233	0.261	0.272	0.254	0.260
Car user costs	0.344	0.365	0.326	0.351	0.344	0.365	0.326	0.354	0.234	0.243	0.228	0.237	0.263	0.275	0.254	0.266
Gasoline price	0.342	0.346	0.348	0.351	0.342	0.346	0.348	0.351	0.231	0.239	0.236	0.236	0.261	0.266	0.264	0.268
Diesel price	0.351	0.347	0.355	0.334	0.351	0.347	0.355	0.334	0.238	0.236	0.239	0.228	0.269	0.264	0.268	0.257
LPG price	0.366	0.353	0.328	0.339	0.366	0.353	0.328	0.339	0.246	0.238	0.226	0.231	0.276	0.272	0.250	0.260
Road availability	0.312	0.359	0.380	0.343	0.366	0.353	0.328	0.339	0.226	0.242	0.264	0.210	0.250	0.272	0.291	0.251
Distance to public transport	0.365	0.325	0.343	0.351	0.365	0.325	0.343	0.305	0.250	0.233	0.239	0.259	0.278	0.258	0.268	0.262
GRP per capita	0.354	0.356	0.336	0.347	0.354	0.356	0.336	0.341	0.222	0.247	0.238	0.235	0.255	0.280	0.265	0.264

The logistic loss is calculated based upon all observations of the test set that belong to a certain quartile. For example, the logistic loss of the first quartile of daily mean windspeed is the logistic loss of all observations in the test set for which the values of daily mean windspeed belong to the first quartile.