PROJECT PROPOSAL

BLOCK 1D-2023-2024

Safer Roads in Breda

By Lars Kerkhofs, Natalia Mikes, Artjom Musaelans, Dominik Ptaszek and Luka Wieme

18-06-2024

Introduction

Throughout the years, the Netherlands have worked on increasing road safety for all road users with the intention of decreasing the number of road accidents. Many new programmes have been created with the focus on improved road safety. In 1998, the first phase of the Multi-year Traffic Safety Plan ('MPV' in Dutch) was implemented, the Start-up Programme Sustainable Safety. This programme resulted in a decrease of somewhere between 1600-1700 road deaths in the period 1988-2007 (SWOV, 2018, p. 7). To maximise safe road traffic, three principles were followed: eliminating, minimizing and mitigating. The project conforms well to the mitigating principle, which means that 'where people are exposed to risks, their consequences should as far as possible be mitigated by taking appropriate mitigating measures' (SWOV, 2018, p. 8). It also aligns well with one of the ANWB's missions that are stated on their website. One of their missions is 'Insurance', wherein their aim is preventing instead of insuring. Thus, the project aligns well with the client's goals as well as national ones.

The purpose of the project would be to focus on mitigating risks on urban roads in Breda. It has been noticed that there are a ton of applications being used in nowadays traffic, including ANWB's own application, the 'Onderweg en Wegenwacht' ('En route and roadside assistance) application. However, there is no information on a more local scale, streetwise application. Even if a street is not deemed busy concerning traffic, it could still be potentially dangerous to drive through due to several factors. With the application drivers and other road users will be informed about the risks of danger on a street level.

In the case of the project, Machine Learning techniques would be very beneficial, because an ML model would be able to identify previously unknown trends in the data. Additionally, it would be more accurate in predicting and forecasting high-risk areas, which would be the purpose of the project.

Problem Statement

As explained before, there are not enough applications that inform road users about streets that could be dangerous. An AI business model canvas was created to explain the project idea in detail, see figure 1. The business model canvas outlines what the project needs, the idea of the project and the costs of the project. The purpose of the project would be to mitigate risk or at least inform about potential risk depending on the street. A street could be considered dangerous to drive through due to several factors, including weather conditions, vegetation and side-streets from which children or cyclists could suddenly appear from. The aim of the project would be to combine the available datasets, which includes the ANWB safe driving dataset and the KNMI weather, to determine whether a street is high-risk or low-risk. Depending on the

project, it would also be possible to find additional datasets for more determining factors. With the help of the project application, drivers as well as other road users would be able to check whether a certain street is dangerous or not before they drive through it. A user of the application, the target user being drivers, could then decide to reroute if they do not wish to drive through high-risk streets or be made more aware of their surroundings when driving through them.

Additionally, the application would lower the costs of insurance for ANWB, since drivers will be more careful when driving through dangerous streets. The application could also be used to identify high-risk streets by ANWB, and they could confront the municipality of Breda about implementing measures to make them safer for all road users. ANWB has the right to do that since it is considered a road authority in the Netherlands and their objective is to aim for safe behaviour on the roads (SWOV, 2018).

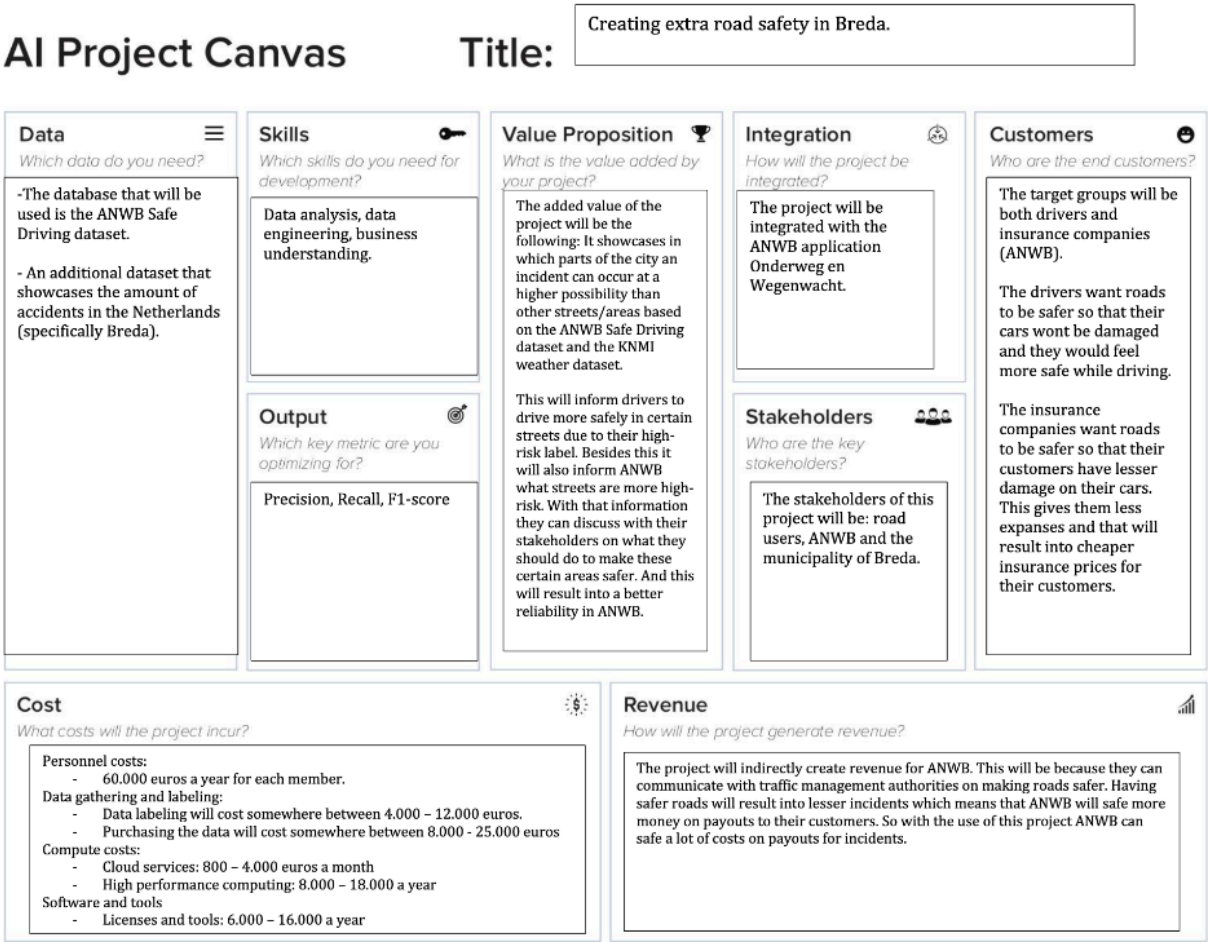

Figure 1 AI Project Business Canvas

Data Description

The main dataset that will be used in the project is the ANWB safe driving data source. The dataset exists out of incident-related data that ANWB have collected between the years 2017 and 2023. The project is focused on the municipality of Breda, so the

municipality_name variable will be used to ensure all data only pertains to incidents in Breda. From the available datasets, the most important ones for the project would be the categorical variables that pertain to incidents and the road where the incident took place. There are four categories of incidents: speed, harsh cornering, - accelerating and – braking. The maxwaarde variable is a numerical value that records the maximum speed (km/h) during the incident if the incident category is speed. If the incident is of a different category, it records the g-force value. This variable could be useful for the project. Furthermore, the incident_severity variable records the severity of the incident based on a given scale. Depending on how diverse the data is, it could be decided to focus only on the most reoccurring incident severity values. Lastly, the road_name variable is important for the placement of the incident, since the goal of the application is to provide the level of risk per street. To offer better location accuracy, the latitude and longitude variables could also be used.

Apart from the ANWB dataset, the additional Bron accidents dataset will also be incorporated in the project. This dataset offers information on accidents that happened in certain areas, which could be useful to predict accident-related data for the project. It also offers timestamp data, which could be useful for the accuracy on what time an accident could happen. Lastly, the latitude and longitude variables will be used alongside the ANWB ones for better accuracy in general. Interesting to note is that the dataset offers information on road conditions, which could be useful for the project.

Furthermore, the KNMI weather data source will be used for the project. It is divided into three datasets: wind, precipitation and temperature. All of the data only pertains to Breda but does not offer data localised data. This means that the data could only be used on a broad scale for all of Breda. It could be used to predict what type of weather could be present during specific timestamps in Breda alongside incident data. It also contains the variable dtg, which records timestamp data. This could be used alongside the timestamp data in the Bron accidents dataset.

Methodology

Machine Learning Process

The machine learning process will be outlined. Firstly a look will be taken at the available datasets, which include the ANWB, KNMI and Bron datasets. The data cleaning process will take place as one of the first things, which includes dealing with duplicates and irrelevant observations. Since the data has been initially looked at, there is some understanding of what steps are needed to be taken during the data cleaning. There is very little duplicate data, so duplicated rows will be dropped. Unwanted outliers will be identified and filtered out using the IQR technique with quantiles lower than 0.25 and greater than 0.75. Structural errors will be fixed, which includes fixing

strange naming conventions, typos and incorrect capitalization. These could cause mislabelling of categories and classes if they are not dealt with. Columns that do not provide a lot of value will be dropped. Unknown and empty values will be converted to nan values. Data will also be scaled using Standard Scaler and transformed to appropriate formats. To make sure that the distribution of data is equal, certain columns will be converted to a binary column type.

After data cleaning follows the pre-processing step. Data was previously transformed to make it suitable for analysis. Additional datasets that were cleaned will be merged and data conflicts will be removed. The EDA (Exploratory Data Analysis) follows the last step along with feature engineering. During the EDA, the data will be explored to find new patterns and relationships in the data. Techniques like plotting value distributions will be done to better understand the data. A correlation heatmap will be plotted to see the correlation coefficients between certain variables, which will be used to understand what variables would be beneficial. Depending on whether it is needed, feature engineering will be done to create new features.

For the model selection the decision has been made to use an XGBClassifier. It prioritises computational speed and model performance, which is preferred for the project. It is most used for classification and regression predictive modelling problems (J. Brownlee, 2021). Gradient boosting will be useful, because new models are added which correct the errors of the previous models. The models will be trained on the aforementioned datasets as well as additional datasets if necessary. The performance of the model will be validated on a separate validation set.

Several models will be trained, some with coordinates included and some without, to see which is more accurate and to check for overfitting problems. The models will be evaluated using the RMSE (Root Mean Square Error), recall, and MAE (Mean Absolute Error). The best model will be chosen to iterate on. GridSearchCV will be used to tune the hyperparameters and pinpoint the best possible ones. This method systemically explores different combinations of hyperparameters to improve the model. The process is very time-consuming and resource-intensive, but it will provide the best results (R. Shah, 2024). Lastly, the idea is to deploy the final model with an API application which will be made using Streamlit.

Level of Risk

According to the recently released EU AI Act, wherein levels of AI risk are explained, it has been discussed that the AI system in the project would be classified as limited risk. This is in accordance with Chapter 3, Section 1, Article 6, it states that if the AI system is not the safety component of a product and is not a product itself (European Commission, 2024).

The nature of the AI system is also to merely predict and provide suggestions to the user based on historical data. The final decision still lies with the user, the driver, and they are in control of their own driving behaviour, since the application merely provides a suggestion.

The purpose of the model is to inform drivers of high-risk streets, thus potentially leading to safer driving behaviour, but it is not able to intervene with a driver's decisions. Since the AI system itself does not put the lives and health of citizens directly at risk, it could not be deemed high-risk according to the EU AI Act regulations. It does not infringe on any rights, apart from the privacy of certain groups during the data collection process.

Even though the AI system would be classified as limited risk, which means there are less regulations to follow, there are still certain legal requirements and obligations that need to be addressed during the development of the AI system. Since the AI system is not high risk, it would not need to undergo a conformity assessment procedure as per the EU AI Act regulations. However, there are other obligations that would need to be taken into consideration. As providers of the AI system, a documentation would be needed that provides the assessment and reasoning of the AI system as limited risk. Before that, as providers, or an authorised representative would have to register in the EU database, including the AI system, even if it is not high risk (European Commission, 2024, Article 49). It is important to mention Article 7 of the EU AI Act, which states that the Commission could decide to change the list of high-risk AI systems. This could happen if the state of the current AI system grows and starts to be used in other areas which do correspond to high risk AI systems. It is also important to note Article 8, since it is the duty of the providers to ensure that the AI system complies with all the requirements of the Union harmonisation legislation. The necessary testing and reporting processes, information and documentation would have to be provided in line with Article 8. In accordance with Article 52 as well as Article 51, as a generic AI system, it will have to be assessed if it poses systemic risk. If so, the Commission would have to be notified of this within two weeks.

Additionally, transparency about the use of AI is important. Users would have to be notified that the predictions are generated by an AI system, so users and all involved parties are sufficiently informed. This can be achieved by writing a code of conduct wherein the data collection process is explained and to what purpose.

Project Timeline

The project timeline was created with the help of a visual roadmap, see figure 2. The project timeline starts at week 3, because the prior weeks were individual learning weeks. In the first project week, the legal and ethical framework will be studied

pertaining to the project and client. The responsibilities and roles will be divided between team members. The idea for week 3 is to brainstorm potential project ideas and present an initial project proposal at the end of the week. In the second week the plan is to scrape the internet for additional datasets and identify if any of these datasets could be used. The project will also be refined during this week. At the end of week 4, the data cleaning process will be started as well as the pre-processing. In the fifth week both the EDA and feature engineering will be conducted. The pre-processing will also be finalised. Ideally, some initial models will also be created by the end of the week. In the sixth week, model training and model iteration starts. The models will also be evaluated in the same week and once that is done the hyperparameters will be fine-tuned. The last week will be for model deployment. The documentation of the project will also be finalised and sent in the last week.

Figure 2 Project Timeline

Concerning role distribution, the idea is as follows:

Dominik and Lars will be the data engineers during the project. Their job will be to clean, pre-process and prepare the data for later use. Luka will be the data analyst. He will perform the EDA process and analyse the data for relationships, patterns and other

interesting features found in the data. Artjom will be the data scientist. He will further explore the data and create appropriate machine learning models. He will be supported by the other team members during this process. Natalia will be the business analyst. She will focus on the legal framework concerning the AI system and the client.

References

1. *AI Act*. (2024, April 30). Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
2. *OpenAI.* ChatGPT. Generated an answer to a question pertaining to the legalities. Prompt: 'What are the legal obligations and requirements that have to be addressed for the development of the AI system'. (05-06-2024)
3. SWOV. (2018). *Sustainable Safety 3rd Edition – The Advanced Vision for 2018-2030.* In swov.nl. SWOV Institute for Road Safety Research. https://swov.nl/system/files/publication-downloads/dv3_en_kort_rapport.pdf
4. Schermers, G. (1999). *Sustainable Safety A preventative road safety strategy for the future*.
5. *EU AI Act: first regulation on artificial intelligence | Topics | European Parliament*. (2023, August 6). Topics | European Parliament. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
6. *General Data Protection Regulation (GDPR) – legal text*. (2024, April 22). General Data Protection Regulation (GDPR). https://gdpr-info.eu/
7. *EU Artificial Intelligence Act | Up-to-date developments and analyses of the EU AI Act*. (n.d.). https://artificialintelligenceact.eu/
8. Kraay, J. H., Ministry of Transport, Public Works and Water Management, Directorate-General of Public Works and Water Management (RWS), & Transport Research Centre (AVV). (2001). *ROAD SAFETY AT THE START OF THE THIRD*

MILLENNIUM. Ministerie van Verkeer en Waterstaat, Rijkswaterstaat, Adviesdiensst Verkeer en Vervoer (RWS, AVV). https://open.rijkswaterstaat.nl/publish/pages/134153/road_safety_at_the_start_of_the_third_millennium-paper_presented_to_the.pdf

9. Brownlee, J. (2021, February 17). *A Gentle Introduction to XGBoost for Applied Machine Learning.* Machine Learning Mastery. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/
10. Shah, R. (2024, June 4). *Tune Hyperparameters with GridSearchCV.* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/