Algorithm Audit

Problem statement

Predicting irresponsible driving

Algoprudence
ALGO:AA:2025:01:P

January 13th 2025

Executive summary

This document describes several normative questions relating to algorithmic prediction of irresponsible driving in a database of a business-to-consumer car sharing platform. Through machine learning a risk model (balanced random forest) is trained to identify driving behaviour associated with users who have been previously banned from the platform for irresponsible driving. The model calculates a user's risk score following each new trip. If a risk score surpasses the warning threshold, a notification is sent to the user, advising them to improve their driving behavior and providing specific suggestions on how to do so. If a user's driving behavior does not improve in subsequent trips, the platform may block the user from its services following a human review. This document outlines specific normative questions that emerge from this use case. It provides technical, juridical and contextual background that is necessary to address these questions.

The car sharing platform has approached Algorithm Audit to provide independent advice on the responsible use of the risk prediction algorithm. To facilitate this, the platform allowed Algorithm Audit to carry out a due diligence assessment. All information presented in this document is based on interviews with the platform's data scientists and an exchange of relevant documents. While Algorithm Audit has carried out a due dilligence assesment on this information, the source code has not been independently reviewed by Algorithm Audit. To facilitate publication of this case study, the car sharing platform has been anonymized. Algorithm Audit has not received financial or any other compensation from the platform for conducting this independent review.

Algorithm Audit has brought together a focus group to gather user perspectives on data processing, algorithmic predictions and communication by shared mobility platforms. The insights from this focus group are shared in the document 'Focus group results'. On the basis of this problem statement and the focus group results, a normative advice commission convened by Algorithm Audit will issue advice on the questions outlined below.

Table of Contents

About Algorithm Audit

Algorithm Audit is a European knowledge platform for AI bias testing and normative AI standards. The goals of the NGO are four-fold:

Knowledge platform
Bringing together experts and knowledge to foster the collective learning process on the responsible use of algorithms, see for instance our white paper and public standards

Normative advice commissions
Forming diverse, independent normative advice commissions that advise on ethical issues emerging in real world use cases, resulting over time in *algoprudence*

Technical tools
Implementing and testing technical tools for bias detection and mitigation, e.g, bias detection tool and synthetic data generation

Project work
Support for specific questions from public and private sector organisations regarding responsible use of AI

1. Introduction

Car sharing platforms have become an increasingly popular option for using a car without having to own one. Operators purchase vehicles that are rented out to registered users on the platform. Not all users use the shared cars responsibly, however. Some users exhibit irresponsible driving behavior, such as speeding, excessive acceleration and heavy cornering. Such dangerous driving behavior is linked to traffic accidents and injuries, involving users of the car sharing platform and other road users. To promote road safety and to reduce damage costs, the platform monitors its users' driving behavior.

Costs for car sharing platforms due to vehicle damage range typically between €2k (small damage) to €20k (total loss damage). For the platform the annual damage costs in 2023 constituted approximately 7% (€2M-€3M) of total revenue (€25M-€45M).

The car sharing platform trains a prediction model to identify what driving behavior is associated with users previously banned from the platform due to irresponsible behavior. Driving behavior data is collected in a standardized format through a Driver Behavior Monitoring Systems (DBMS) installed in cars of the platform. Using these features, a balanced random forest (BRF) model is trained to predict a risk score indicating whether a user is likely to behave irresponsibly.

The focus of this document is to evaluate how the BRF model should be employed and calibrated in view of the potentially harmful impact of inaccurate predictions. Particular focus is placed on accurately identifying irresponsible behavior while avoiding unfair and excessive unjustified suspicion of users. This goal is pursued by evaluating the balance of false positive and false negative predictions and the selection of features and hyperparameters.

This problem statement provides the necessary technical, legal and contextual background for addressing these questions. It outlines the data collection, training and testing processes of the BRF model, and provides a brief analysis of the relevant legal frameworks governing this use case.

When contemplating the responsible use of algorithms, a key initial question to raise is why a data-driven approach is suitable for addressing the problem at hand. In this case, using a data-driven approach to collect and monitor driving behavior may be justified for various reasons. First, empirical research has demonstrated that collecting driving characteristics via a DBMS can enhance road safety.[1] Second, machine learning techniques have proven effective in identifying unsafe driving behavior in collected DBMS data.[2] Within the boundaries of the law, particularly on data collection and processing, it may therefore be worthwhile to explore algorithmic methods for identifying unsafe driving behavior to improve road safety. This case study examines how an algorithmic approach can be implemented in a responsible manner. Algorithm Audit conducted an independent examination of the case by interviewing the platform's data scientists and raising relevant questions. This problem statement was drafted on the basis of this examination. It distils key

[1] D. Lord, F. Mannering, The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives, Transportation Research Part A: Policy and Practice, Volume 44, Issue 5 (2010) https://doi.org/10.1016/j.tra.2010.02.001

[2] E. Lattanzi, V. Freschi, Machine Learning Techniques to Identify Unsafe Driving Behavior by Means of In-Vehicle Sensor Data, Expert Systems with Applications, Volume 176 (2021) https://doi.org/10.1016/j.eswa.2021.114818

normative questions that have emerged in the use case, and presents these questions alongside discussion of the relevant background. The problem statement serves as input for an independent normative advice commission convened by Algorithm Audit, which will issue advice on the questions outlined below.

In preparation for the advice commission, Algorithm Audit has convened a focus group, consisting of four users who regularly make use of the services provided by shared mobility platforms. In this case, the users are key stakeholders that are affected by implications of data collection and wrong risk prediction of the BRF model. The focus group was held in order to capture what users in particular think of different aspects of a shared mobility platform in the context of data processing, algorithmic predictions, and communication. Multiple questions were asked about their experiences with and ideas surrounding such a platform, which also lead to discussions within the group. The questions specifically focused on conceptions regarding data collection, driving and payment behavior as data points, use of the algorithm to predict irresponsible users, the procedure around blocking someone's account (including human interaction and redress), and the trade-off between false negatives and false positives. After conducting the focus group, the interview was transcribed and paraphrased into a report of questions and answers. The results can be found in the document 'Focus group results'. The insights from the focus group serve as input for the deliberation of the normative advice commission. Yet the commission is not obliged to include the views of the focus group in its final advice.

The structure of this problem statement is as follows: first, the data collection process is discussed in 2. Data collection, and question 1 is introduced. Thereafter in 3. Prediction model, details are shared about the training and testing process of the BRF model, where questions 2-5 are introduced. Then, a brief legal analysis of this case study is provided in 4. Legal aspects.

In the appendices:

> Appendix A lists and describes all independent variables used in the BRF model
> Appendix B outlines how users are informed about warnings or exclusion of the car sharing platform
> Appendix C provides an overview of model predictions for specific BRF hyperparameters
> Appendix D includes technical details about sensitivity testing of the BRF model.

Box 1

Algoprudence: Building public knowledge for ethical algorithms

Algorithm Audit does not have a mandate to issue legal rulings or official judgements. In our case studies, we give non-binding ethical advice. Ethical advice often goes beyond advice on what is required for legal compliance. Yet in the absence of legal rulings or clear standards established by a supervisory body, our independent ethical advice also serves as a preliminary signpost for organizations. To our growing body of case-based normative advice is referred to as algoprudence. Algoprudence may also help elaborate official standards or support future decisions by legal bodies. In this sense, our ethical advice does have relevance for the legal domain.

2. Data collection

The process of data collection by the car sharing platform is described.

2.1 Driving characteristics

Risk predictions by the BRF model are based on data points related to driving characteristics. Personal data, such as age, ZIP code or years holding a driving license, are not included in the analysis. Data collection through a Driver Behavior Monitoring Systems (DBMS) is an industry standard for operators to measure standardized driving characteristics in real-time, specifically cornering, braking, acceleration and speed events. During a car rental, the DBMS collects data points on users' driving behavior as longitudinal data, representing pooled driving characteristics across the user's entire trip history. A detailed description of 32 features derived from the DBMS relevant for this case can be found in Appendix A.

User perspectives on data collection practices by shared mobility platforms can be found in the document Focus group results.

Previously, the platform considered user's payment data – specifically, the number of bank accounts linked to the platform and instances of late payments – as well as criminal offences, particularly the number of fines. However, concerns about potential violations of the General Data Protection Regulation (GDPR) and operational challenges related to the delay between a traffic offense and the arrival of the fine led the platform to discontinue the use of these variables.

2.2 Irresponsible driving behavior

The labels in the training and test datasets indicate users who have been previously blocked from the platform. Users are blocked when a human analyst concluded their behavior was irresponsible. The following behavior results in blocking the user account:

1. If a user doesn't respond to a damage report and there is evidence that they caused the damage
2. If a user doesn't return the car to the correct location for the second or third time, depending on the severity of the case
3. If a user doesn't pay their invoices for more than 6 months
4. If the user is reported for smoking in the car or leaving the car dirty, for the second or third time, depending on the severity of the case
5. If another driver drives the car who is not added as an extra driver, for the third time
6. If a user makes a fraudulent trip during a trip that was meant for operational purposes (e.g. discharging the car), for the third time.

Accidents that do not involve culpable negligence are not labelled as irresponsible. User perspectives on data collection practices by shared mobility platforms can be found in the document 'Focus group results'.

Q1

While only data about behavior on the platform and not about the person are collected and used for making risk predictions, a careful assessment should consider whether, or specifically which, features can be justifiably used for the purpose of predicting irresponsible driving behavior. For instance, the use of certain features may possibly introduce (indirect) discrimination, unwanted bias or other forms of unfairness in the risk predictions.

Question 1a: Should features related to aggressive driving behavior (aggressive acceleration, hard braking, heavy cornering) or speeding events be fed to the BRF model (see Appendix A)? Are there reasons why such driving characteristics should not be used as input variables for predicting irresponsible driving behavior?

Question 1b: Can it be justified to use other features related to trip events (specifically returning a car lately, leaving the car dirty) and payments details (specifically number of payment methods linked to a user account, late invoices) to make risk predictions for irresponsible drivers?

3. Prediction model

The training, evaluation and calibration of the Balanced Random Forest (BRF) model are described, followed by socio-technical questions regarding normative decisions that must be made during the data modelling process.

3.1 Risk predictions

Based on historical data, the BRF predicts which active users are likely to exhibit irresponsible behavior in the future. The trained BRF model assigns a risk score r_i to each user i, with $1 \leq i \leq n$ users in the database. Risk scores fall into three categories:

1. **No action:** the predicted risk score of user i falls below the warning threshold r_w, i.e., $r_i < r_w$;
2. **Warning:** the predicted risk score of user i falls between the warning threshold r_w and the blocking threshold r_b, i.e., $r_w < r_i < r_b$;
3. **Blocked:** the predicted risk score of user i exceeds the blocking threshold r_b, i.e., $r_i > r_b$.

A BRF model is fit on the longitudinal data of 75.919 users to link DBMS data (independent variables) with previously blocked accounts (dependent variable). Only 2.7% of the users in the training dataset have a blocked account status. The dataset is split into an 80:20 ratio for training and testing. For each user i in the test dataset the BRF predicts a risk score r_i between 0 and 1. Based on this model, existing users in the dataset are assigned a risk score. Users with a risk score between the warning and a blocking threshold are sent a warning, while those with a score above the blocking threshold are referred to a human analyst. After human review, the user either receives a warning message or is excluded from the car sharing service. The exact procedures for issuing warnings and blocking users are described in Appendix C.

Box 2

Other prediction models

Besides a Balanced Random Forest (BRF), also a logistic regression model has been trained on driving characteristics to predict irresponsible behavior. However, the BRF model consistently outperformed the logistic regression model across all confusion matrix-based evaluation metrics. As a result, the platform decided to deploy the BRF model and evaluate how it can be implemented in the most responsible way.

3.2 Evaluating the prediction model

Predictions of the BRF model on the train and test dataset are evaluated in terms of confusion matrix-based evaluation metrics. Let the positive class consist of all observed irresponsible drivers (previously blocked drivers due to irresponsible driving), then the confusion matrix consists of the following elements:

- **True Positive (TP):** irresponsible driver classified as irresponsible;
- **False Positive (FP)**: responsible driver classified as irresponsible;
- **True Negative (TN):** responsible driver classified as responsible;
- **False Negative (FN)**: irresponsible driver classified as responsible.

To assess the socio-technical implications of the BRF predictions, the impact of FN and FP predictions should be weighed, as both type of predictions come with certain risks.

For FNs, material risks for the company arise from not flagging users who are more likely to cause vehicle damage than responsible drivers. This drives up damage costs. Immaterial risks of FNs include allowing irresponsible behavior to go unnoticed, colloquially summed up by the saying: *"don't be gentle, it's a rental"*. This could, for instance, result in the platform's vehicles being associated with irresponsible driving behavior on the road. Minimizing the number of FNs contributes to road safety for all.

For FPs, material costs for the company stem from employing human analysts to manually review suspected users before sending warning and blocking messages. A risk is that users may experience being falsely accused of irresponsible driving behavior. This might lead them to switch to a competing platform, bringing material costs to the company. This risk is currently mitigated by the manual review conducted by human analysts, who filter out numerous false positives. It may turn out that certain groups are more subject to FPs than average, introducing a risk of bias and disparate impact. Another risk of false positives is that users may feel as if they are under constant surveillance while using the platform's services.

Q2

Question 2: Considering the actions following an assigned risk score as described in Appendix B, is it equally harmful to classify responsible drivers as irresponsible (false positive) as it is to classify irresponsible drivers as responsible (false negative)? Which outcome carries more weight, and why?

Based on the confusion matrix, the following evaluation metrics are considered for the BRF model:

- **False Positive Rate (FPR):** fraction of FPs with respect to all observed responsible drivers, i.e., FP/(FP+TN);
- **False Negative Rate (FNR):** fraction of FNs with respect to all observed irresponsible drivers, i.e., FN/(TP+FN).

Similarly, with more emphasis put on TPs, the following related evaluation metrics can be considered:

- Recall, also True Positive Rate (TPR): fraction of TPs with respect to all observed irresponsible drivers, i.e., TP/(TP+FN);
- Precision: fraction of TPs with respect to all predicted irresponsible drivers, i.e., TP/(TP+FP).

In a perfect world, both FP and FN predictions would be minimized. In practice, however, a balance should be struck. This trade-off is influenced by the choice of hyperparameters selected for the BRF model.

3.3 Hyperparameter tuning of the BRF model

The BRF model is trained using the Python scikit.learn package and is fitted on the 32 driving characteristics present in the training dataset. The model uses the following hyperparameters:

- Number of trees (n_est)
- Maximum depth per tree (max_depth)
- Minimum number of samples to split (min_samples_split)
- Minimum number of samples required to be in a leaf (min_samples_leaf)
- Number of features considered making a split (max_features)
- Whether the the minority class is balanced (sampling_strategy)
- Trees are trained on bootstrap samples, i.e., randomly selected samples from the training dataset (bootstrap);
- Datapoints for the bootstrap sample are randomly selected with or without replacement (replacement);
- Warning threshold (r_w).

Detailed explanations of these parameters are available in scikit.learn documentation.[3] A sensitivity analysis of the BRF model can be found in Appendix D.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Q3

Question 3: What sensitivity testing* should be performed to select hyperparameters n_est, max_depth, min_samples_split, min_samples_leaf and max_features?

* *sensitivity testing refers to evaluating how sensitive the BRF model's performance is to changes in its hyperparameters. The goal is to assess the stability and robustness of the model under varying conditions. This can help identify how changes in the model's design may impact its predictions. An example of sensitivity testing for a BRF model can be found in Appendix D.*

Assume that the following parameters are selected: n_est=100, max_depth=None, min_samples_split=2, min_samples_leaf=1, max_features=√32, sampling_strategy= 'not majority', bootstrap=False, replacement=True. The maximum depth per tree (max_depth) is set as none, which means nodes are expanded until all leaves are pure or until all leaves contain less than 2 (min_samples_split) samples. For this set of hyperparameters, evaluation metrics of the BRF model are reviewed.

3.4 ROC curve

Balancing FNs and FPs depends on the selection of the warning threshold r_w. For the hyperparameters specified above, the TPR and FPR are computed using 5-fold cross-validation for $0.01<r_w<0.9$ in intervals of 0.01 (see Box 3). Alternatively, precision and recall can be used as evaluation metrics. However, these metrics place less emphasis on the impact of FPs (see 3.2 Evaluating the prediction model). Since the platform aims to strike a balance between capturing as much risk as possible while avoiding excessive false suspicion of responsible drivers, the ROC curve is preferred for evaluating this model. In a ROC curve, the TPR is plotted against the FPR. See Figure 1.

The confusion matrix of the BRF model for thresholds r_w=0.25 and r_w=0.5 can be found in Appendix C.

Based on the results shown in Figure 1, three scenarios are considered to strike a balance between the TPR and FPR:

- **Scenario blue:** r_w=0.20 with TPR 0.4240 and FPR 0.0157. TPR is 27 times higher than FPR;
- **Scenario orange:** r_w=0.40 with TPR 0.2100 and FPR 0.0028. TPR is 75 times higher than FPR;
- **Scenario green:** r_w=0.60 with TPR 0.0740 and FPR 0.0003. TPR is 219 times higher than FPR.

A figure displaying the precision and recall rates for the same 5-fold cross-validations be found in Appendix C.

Q4

Question 4: Considering the socio-technical impact of FP and FN classifications, which of the three scenarios is most preferable?

Box 3

K-fold cross validation

K-fold cross validation refers to dividing the dataset into k equal, non-overlapping subsets or "folds". Each fold is used as the test set once, while the remaining k-1 folds are used for training. This method is widely used to prevent overfitting, ensuring that every part of the dataset is utilized for both training and testing, and providing a reliable estimate of model performance. After the model is trained using k-fold cross validation, the final model is applied on the test dataset, which was not included in the k-fold cross validation process.

More information on cross validation can be found in Chapter 5 of An Introduction to Statistical Learning, G. James, D. Witten, T. Hastie and R. Tibshirani

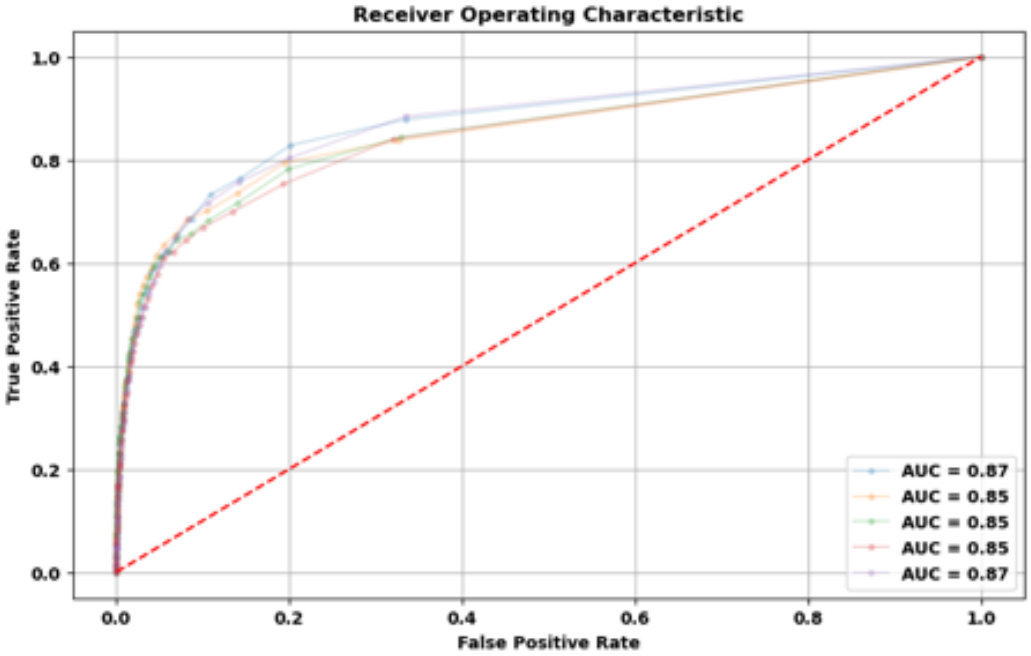

Figure 1 - ROC curve of BRF model for 5-fold cross-validation

3.5 Feature importance

Let r_w=0.25. This BRF model has a TPR of 0.3640, FPR of 0.0096, TNR of 0.9904 and FNR of 0.6360. The blocking threshold is set on r_b=0.4, which corresponds with TPR of 0.2250 and FPR of 0.0034. For this model, the BRF model is fit on the test data. The features holding the most predictive value in terms of mean decrease in impurity (based on the Gini-index) are displayed in Figure 2. The top three variables that best characterise irresponsible behavior are:

> **TOTAL_DRIVING_EVENTS_PERKM:** Total number of cornering and braking events divided by the number of driven kilometers;

> **TOTAL_DRIVING_EVENTS_PERMINUTE:** Total number of cornering and braking events divided by the total driving time in minutes per user;

> **BRAKING_PERKM:** The number of braking events recorded over all trips per user divided by the number of driven kilometers.

A full description of all features can be found in Appendix A.

Q5

Question 5: With the assumption that it is desirable to provide feedback to users when they receive a warning of irresponsible driving, what constitutes meaningful explanation here?

1. A general notification about the detection of the user's aggressive driving and speeding, without any further specification;
2. Specification of value 'total driving events' (cornering and braking events, per minute or km), acceleration and speeding events of the user;
3. Detailed specification of the user's cornering and braking events, and of different categories of speeding events.

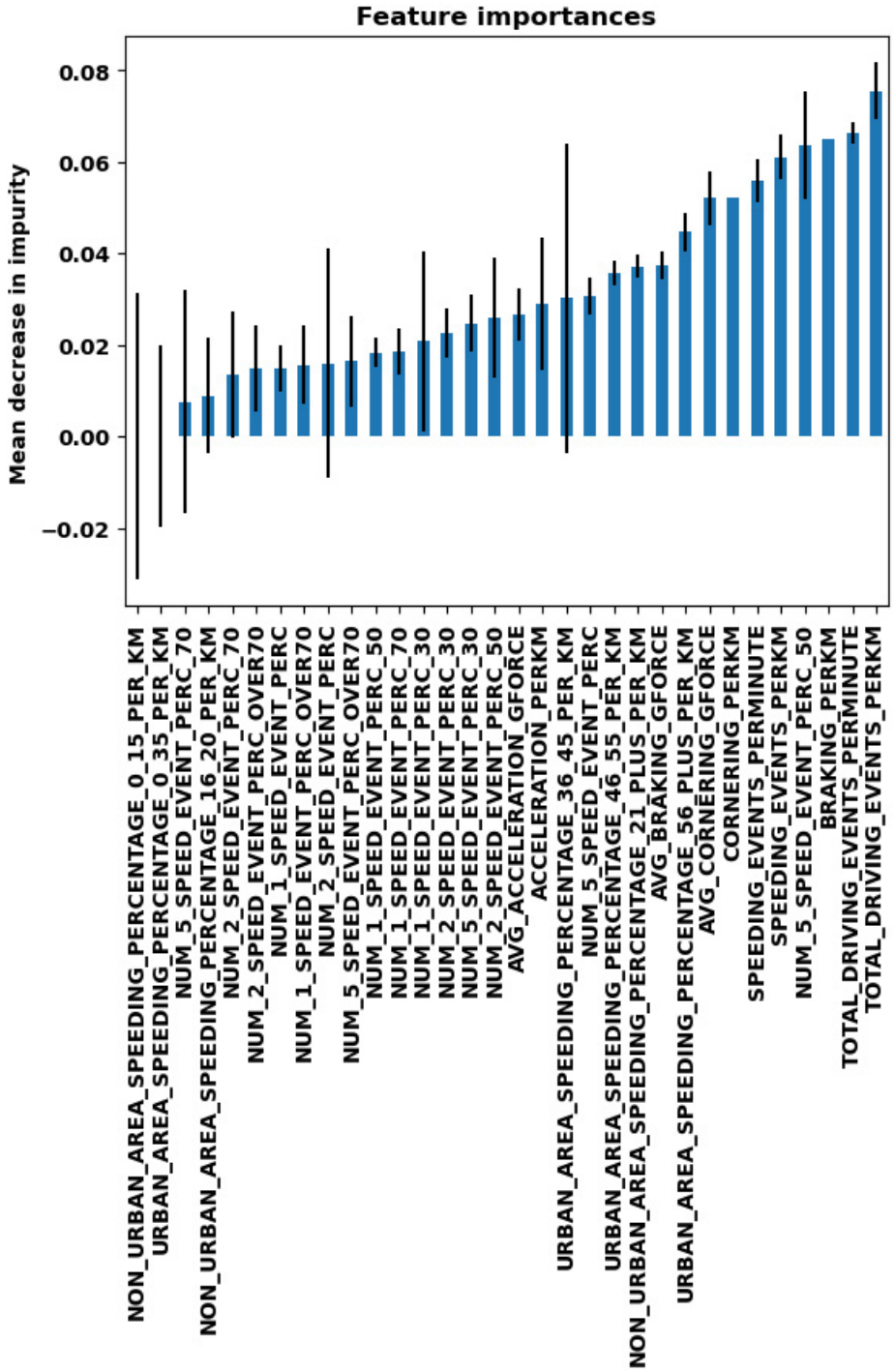

Figure 2 - Feature importance of BRF model predicting irresponsible behavior. Mean decrease in impurity is based on the Gini-index. A higher mean decrease in impurity indicates greater importance of the feature. The top three most predictive features are 1. TOTAL_DRIVING_EVENTS_PERKM, 2. TOTAL_DRIVING_EVENTS_PERMINUTE and 3. BRAKING_PERKM.

Box 4

Disclaimer legal analysis

This problem statement synthesizes insights from various domains to address normative questions related to the case under review, for which no clear legal guidance currently exists. tWhile the statistical and legal context is presented as accurately as possible, occasional limitations may remain. No rights may be derived from this document.

4. Legal aspects

Various legal aspects apply to the machine learning-driven decision-making process for blocking irresponsible drivers from the platform. This section provides a summary of relevant passages of a car sharing platform's privacy policy, alongside a selection of relevant legislation applicable to this case.

4.1 Privacy policy of car sharing platform

By using the platform's services, users agree with the privacy policy. In this policy, the car sharing platform informs users that DBMS data is collected and used as part of a risk prediction model. The platform also explains that data about the driving behavior of a user may qualify as personal data under the GDPR. The privacy policy indicates that users can request information about how their data was used. The data provided to the user will be limited to what was specifically used to make decisions about them. Additionally, the platform collaborates with private organizations and public authorities to address issues such as non-payments, fraud, and criminal offenses. User data may be retained even after an account is deleted, but only for as long as is strictly necessary. If data retention exceeds the timeframe deemed 'strictly necessary', the platform ensures that the data cannot be traced back to a former user by employing methods such as aggregation or anonymization.

4.2 Data Privacy Impact Assessment

At the platform's request, an external legal firm conducted a Data Privacy Impact Assessment (DPIA) in the fall of 2024. Based on the findings of this DPIA, the list of features used in the BRF model was revised, resulting in the 32 features listed in Appendix A. An impact assessment focusing on information security is scheduled for 2025, which may potentially lead to changes in the platform's privacy policy and BRF train-test pipeline.

4.3 General Data Protection Regulation (GDPR)

This section outlines some of the most relevant aspects of the GDPR in relation to this case and how these obligations are met by the platform. However, it provides only a brief legal analysis and is not exhaustive (see also Box 4).

Article 4(1) of the GDPR defines personal data as "any information related to an identified or identifiable natural person". The data used by the platform qualifies as personal data because, during the algorithmic process, driving behavior can be traced back to a specific user through their account. To create a user profile on the platform, users are required to provide personal data, such as their name, address, and date of birth. Additionally, if a user requests information on how their personal data has been processed for algorithmic-driven decision-making, the platform provides data linked to the individual in question.

Article 6 of the GDPR addresses the lawfulness of data processing. According to paragraph 1(a), data processing is considered lawful when the data subject has given consent. By accepting the platform's terms and conditions, a user agrees to its privacy policy, which outlines the types of personal data processed, the purposes for which the data is used, and the parties with whom it is shared. To increase awareness among users that DBMS data are collected, a pop-up screen at the beginning of every trip is shown stating that driving behaviour is monitored and containing a link to the privacy policy with more details. This encourages that the driver reads the privacy policy.

According to the Court of Justice of the European Union (CJEU), a decision based solely on automated processing also includes profiling. In the Schufa ruling, profiling – as defined in GDPR Article 4(4) – itself was considered an 'automated decision' within the scope of the GDPR Article 22, even if the final decision – such as granting credit to only suitable lenders – was made after human intervention. The definition of 'profiling' includes practices that involve automated generation of risk probability based on personal data, such as assessing an individual's creditworthiness and their ability to fulfill future obligations (e.g. repaying a loan).[4]

Article 22(3) GDPR, the controller must implement appropriate measures, such as the right of the subject to obtain human intervention for the purpose of expressing their viewpoints and potentially appeal.

Article 15 of the GDPR states that a data controller must inform the data subject about their rights in the context of processing personal data. This includes disclosing the existence of automated decision-making. In accordance with Article 15(1)(h), the platform must provide information about the functioning of the algorithm, including the logic underlying the automated processing of the personal data and the consequences of such processing.[5] The user can then make an informed decision to exercise their right to access or their right to object the processing of their personal data.

In the case under review, algorithmic-driven risk predictions are a form of profiling, as probabilities are generated whether a user is suitable for using the platform's services based on personal data. Following the Schufa ruling, this could be considered automated decision-making, even if human involvement is present in the decision-making process. Since the human analyst is granted discretionary power, which is exercised 40-50% of the time to deviate from the algorithm's prediction (see also Box 5 in Appendix B), it is likely that no fully automated decision-making, as defined in Article 22 of the GDPR, is involved. The privacy policy stresses that the decision to block an account, based on predictions made by an algorithm, is ultimately made by a human analyst. The privacy policy provides a high-level description of the algorithm, but there are no specific guidelines on the level of detail required in explaining its workings to fulfil the obligations of Article 15 of the GDPR. Therefore, the explanation provided is likely to be sufficient.

The platform is working on a full update of the privacy policy such that in much more detail is stated how the algorithm works, and what data are used to evaluate a user's driving behaviour.

[4] ECLI:EU:C:2023:957
[5] ECLI:EU:C:2023:957 §56

Users can appeal decisions regarding their removal from the platform. After being notified their account will be blocked, a user can appeal by emailing user support to request access to their data. The requested data will then be provided. If users identify any incorrect data, and can provide evidence supporting this, which contributed to the decision to block them, they can use this information to challenge the platform's decision.

So, based on the due diligence process conducted prior to drafting this document, including an analysis by Algorithm Audit of the platform's privacy policy, it appears that some safeguards are in place to ensure users have access to human intervention, can express their views, and can contest decisions made about them.

4.4 AI Act

The AI Act sets out requirements for the use of artificial intelligence (AI) within the European Union (EU). For high-risk applications of AI systems mandatory control measures are required to safeguard the safety, health and fundamental rights of EU citizens.

The algorithm system under review qualifies as an AI system because it uses machine learning to generate predictions based on input data, meeting the definition stated in Article 3 and elaborated in Recital 12 of the AI Act. However, according to Annex III of the Act, the system does not fall into the category of high-risk AI applications, as its use case is not included among the listed high-risk categories.

The closest relevant high-risk application listed in Annex III is under category 5: "*Access to and enjoyment of essential private services and essential public services and benefits*". However, using a car from a car sharing platform does not qualify as an essential private service as multiple car sharing options and public transportation alternatives (such as buses and trains) are generally available to users.

As a result, the mandatory control measures for high-risk applications are not applicable to the algorithmic system under review. Additionally, the control measures, as requested by the European Commission to be developed by standardization organization CEN-CENELEC, are expected to be finalized in the course of 2026 and are therefore not currently available for use as a benchmark.

Appendix A – Collected driving characteristics

Table 1 gives an overview of all driving characteristics considered by the BRF model to predict the risk score.

Table 1. Overview of Data Behavior Monitoring System (DBMS) features fed to the BRF-model.

#	Variable name	Description
Cornering events		
1	AVG_CORNERING_GFORCE	The average G-force of cornering events based on all trips per user
2	CORNERING_PERKM	The number of cornering events recorded over all trips per user divided by the number of driven kilometers
Breaking events		
3	AVG_BRAKING_GFORCE	The average G-force of braking events based on all trips per user
4	BRAKING_PERKM	The number of braking events recorded over all trips per user divided by the number of driven kilometers
Acceleration events		
5	AVG_ACCELERATION_GFORCE	The average G-force of acceleration events based on all trips per user
6	ACCELERATION_PERKM	The number of acceleration events recorded over all trips per user divided by the number of driven kilometers
Driving events		
7	TOTAL_DRIVING_EVENTS_PERKM	Total number of driving events (cornering and braking events) per user divided by the number of driven kilometers
8	TOTAL_DRIVING_EVENTS_PERMINUTE	Total number of driving events (cornering and braking events) per user divided by the total driving time in minutes per user
Speed events		
9	SPEEDING_EVENTS_PERKM	Total number of speeding events (>15 km/h above the limit) per user divided by the number of driven kilometers
10	SPEEDING_EVENTS_PERMINUTE	Total number of speeding events (>15 km/h above the limit) per user divided by the total driving time in minutes per user
11	NUM_1_SPEED_EVENT_PERC	Percentage of trips of a user with at least 1 speeding event

12	NUM_2_SPEED_EVENT_PERC	Percentage of trips of a user with at least 2 speeding events
13	NUM_5_SPEED_EVENT_PERC	Percentage of trips of a user with at least 5 speeding events
14	NUM_1_SPEED_EVENT_PERC_30	Percentage of trips of a user with at least 1 speeding event in a limit up to 30 km/h
15	NUM_2_SPEED_EVENT_PERC_30	Percentage of trips of a user with at least 2 speeding events in a limit up to 30 km/h
16	NUM_5_SPEED_EVENT_PERC_30	Percentage of trips of a user with at least 5 speeding events in a limit up to 30 km/h
17	NUM_1_SPEED_EVENT_PERC_50	Percentage of trips of a user with at least 1 speeding event in a limit between 30 and 50 km/h
18	NUM_2_SPEED_EVENT_PERC_50	Percentage of trips of a user with at least 2 speeding events in a limit between 30 and 50 km/h
19	NUM_5_SPEED_EVENT_PERC_50	Percentage of trips of a user with at least 5 speeding events in a limit between 30 and 50 km/h
20	NUM_1_SPEED_EVENT_PERC_70	Percentage of trips of a user with at least 1 speeding event in a limit between 50 and 70 km/h
21	NUM_2_SPEED_EVENT_PERC_70	Percentage of trips of a user with at least 2 speeding events in a limit between 50 and 70 km/h
22	NUM_5_SPEED_EVENT_PERC_70	Percentage of trips of a user with at least 5 speeding events in a limit between 50 and 70 km/h
23	NUM_1_SPEED_EVENT_PERC_OVER70	Percentage of trips of a user with at least 1 speeding event in a limit at least 70 km/h
24	NUM_2_SPEED_EVENT_PERC_OVER70	Percentage of trips of a user with at least 2 speeding events in a limit at least 70 km/h
25	NUM_5_SPEED_EVENT_PERC_OVER70	Percentage of trips of a user with at least 5 speeding events in a limit at least 70 km/h
26	URBAN_AREA_SPEEDING_ PERCENTAGE_0_20_PER_KM	Number of trips per total km driven in an urban area, where the speed limit was up to and including 50 km/h and the speeding percentage above the limit was between 0 and 20%
27	URBAN_AREA_SPEEDING_ PERCENTAGE_21_40_PER_KM	Number of trips per total km driven in an urban area, where the speed limit was up to and including 50 km/h and the speeding percentage above the limit was between 21 and 40%
28	URBAN_AREA_SPEEDING_ PERCENTAGE_41_55_PER_KM	Number of trips per total km driven in an urban area, where the speed limit was up to and including 50 km/h and the speeding percentage above the limit was between 41 and 55%

29	URBAN_AREA_SPEEDING_ PERCENTAGE_56_PLUS_PER_KM	Number of trips per total km driven in an urban area, where the speed limit was up to and including 50 km/h and the speeding percentage above the limit was at least 56%
30	NON_URBAN_AREA_SPEEDING_ PERCENTAGE_0_10_PER_KM	Number of trips per total km driven in an non-urban area, where the speed limit was greater than 50 km/h and the speeding percentage above the limit was between 0 and 10%
31	NON_URBAN_AREA_SPEEDING_ PERCENTAGE_11_20_PER_KM	Number of trips per total km driven in an non-urban area, where the speed limit was greater than 50 km/h and the speeding percentage above the limit was between 11 and 20%
32	NON_URBAN_AREA_SPEEDING_ PERCENTAGE_21_PLUS_PER_KM	Number of trips per total km driven in an non-urban area, where the speed limit was greater than 50 km/h and the speeding percentage above the limit was at least 21%

Appendix B – Communication to irresponsible drivers

The procedures are described for when a user's predicted risk score exceeds the warning or blocking threshold. Each time a user completes a new trip, a risk score is calculated using the BRF model.

B1. Warning users

When a user's risk score exceeds the warning threshold, the following procedure is followed.

The platform sends an initial mild warning mail to the user to notify that their driving behavior shows signs of irresponsibility. A human analyst will review the user's DBMS data, identify deviations and communicate these observations to the user to help improve their driving behavior.

If a new trip results in a risk score that still exceeds the warning threshold, the user's driving profile is forwarded to a human analyst, who manually checks the user's profile (see Box 5). The analyst will then decide whether a second, more stringent warning message, is sent. If a subsequent trip results in a lower risk score, no additional warning message is sent.

B2. Blocking users

When a user's risk score exceeds the blocking threshold, the following procedure is followed.

The user's profile is sent to a human analyst for a manual review (see Box 5). If a user exceeds the blocking

threshold without having previously received a warning, the analyst may choose to issue a warning message first. However, in cases of serious violation of the terms and conditions of the platform, such as excessive speeding, the analyst may decide to block the user immediately.

The user is provided transparency about the variables contributing to their blocking. A human analyst reviews the user's DBMS data to identify the deviating features that led to a high risk score, which are shared with the user. The notification email informing users about their blocking, also offers users the opportunity to appeal the platform's decision. If users have concerns about the accuracy of the DBMS data collected about them, they can contact the platform for clarification.

Appendix C – Confusion matrices and PR curve

The confusion matrices of the BRF model for thresholds r_w=0.25 and r_w=0.5 on the test dataset with 12.147 datapoints, with hyperparameters as specified in 3.3 Hyperparameter tuning of the BRF model, are shown in Figure 3 and Figure 4.

The values in the confusion matrices have the following meaning:

> **True Positive (TP):** True label blocked, predicted label blocked;
> **False Positive (FP):** True label active, predicted label blocked;
> **True Negative (TN):** True label active, predicted label active;
> **False Negative (FN):** True label blocked, predicted label active.

The Precision-Recall (PR) curve for 5-fold cross validation of the BRF model is displayed in Figure 5.

Appendix D – Sensitivity testing

Box 5

Review of a user profile by a human analyst

Manual checks involve human interference, i.e., human analyst checking if the data attached to a user appear to be sound and complete. This review does not include an analysis of personal data, i.e., only a database ID is shown in combination with driving characteristics. No names, dates of births, registered addresses and other personal data are shown when driving characteristics are analysed. Human analysts follow a training before they conduct reviews. On average, one human analyst spends 30-60 minutes per day to inspect approximately 20 warning cases and 5 blocking cases. 50-60% of the users for which a risk score is predicted that surpasses the blocking threshold are blocked after human inspection.

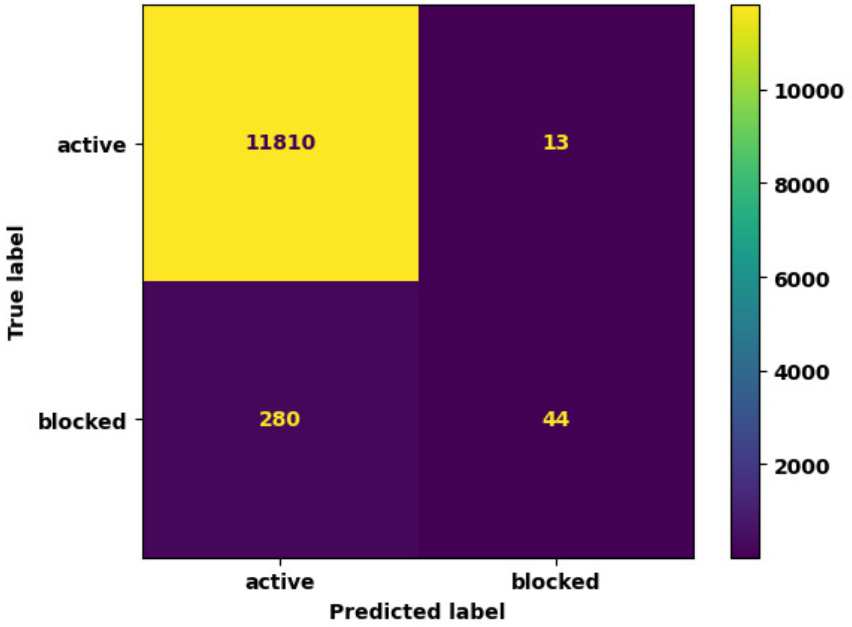

Figure 3 - Confusion matrix voor drempelwaarde 0.5

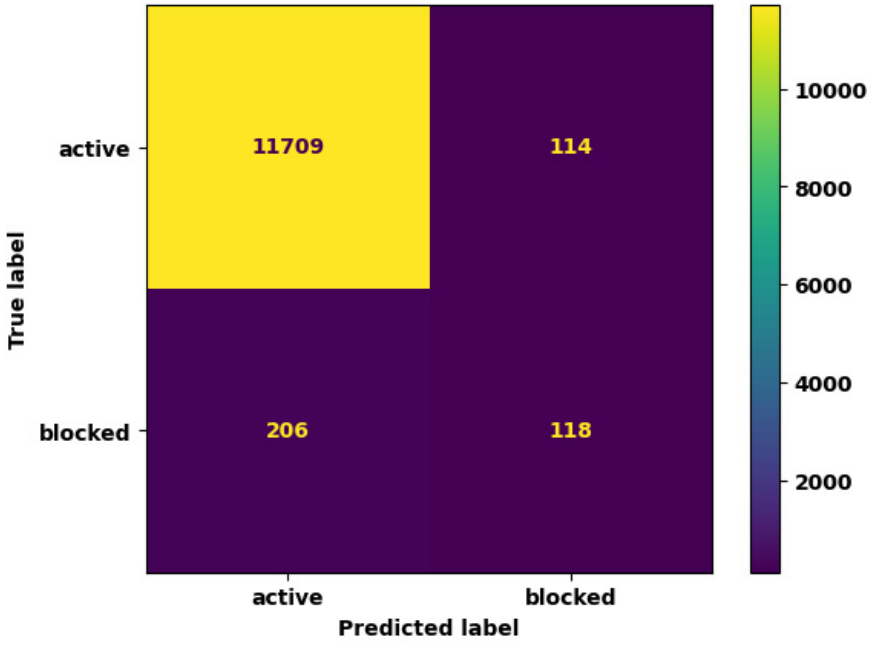

Figure 4 - Correlation matrix voor drempelwaarde 0.25

The sensitivity of the BRF model is tested by training the model using 3*3*4*3*1=108 different hyperparameter specifications:

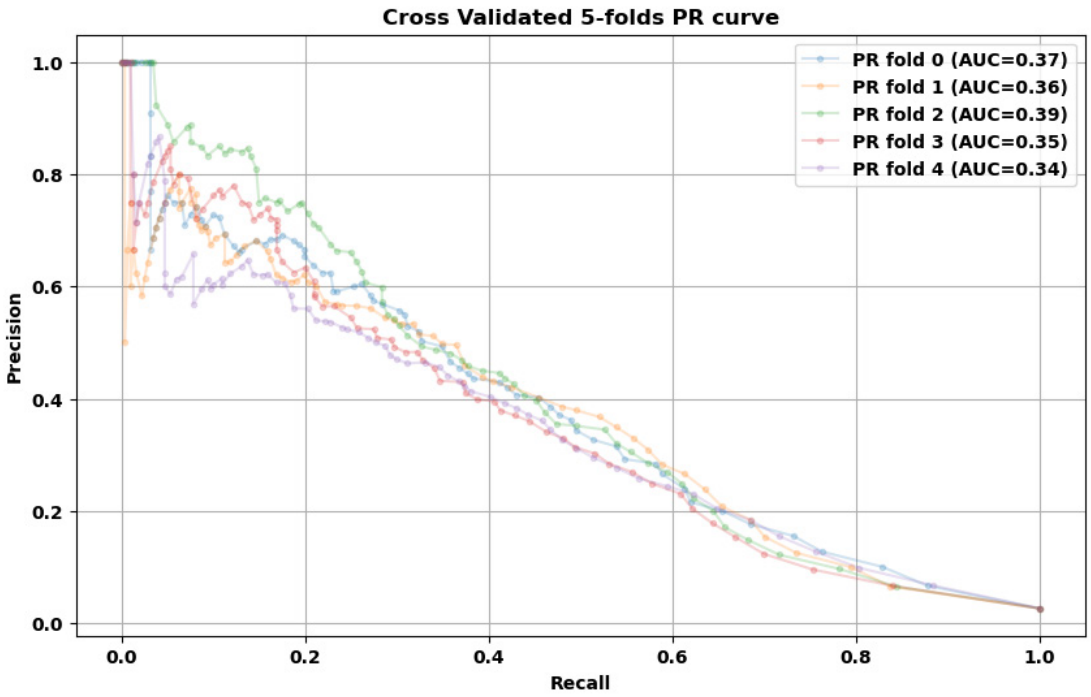

Figure 5 - PR curves for 5-fold cross validation of the BRF model

> Number of trees (n_est): 100, 200, 300
> Maximum depth per tree (max_depth): None, 1, 5
> Minimum number of samples to split (min_samples_split): 2, 3, 5, 10
> Minimum number of samples required to be in a leaf (min_samples_leaf): 1, 3, 5
> Number of features considered making a split (max_features): 32.

For each set of hyperparameters, the BRF model generates predictions using 5-fold cross-validation.[6] The averages of the 5-fold cross-validation predictions for all 108 combinations are shown in Figure 6.

From this analysis it can be concluded that the BRF model is not highly sensitive to changes in hyperparameters, as the PR curve remains largely unchanged.

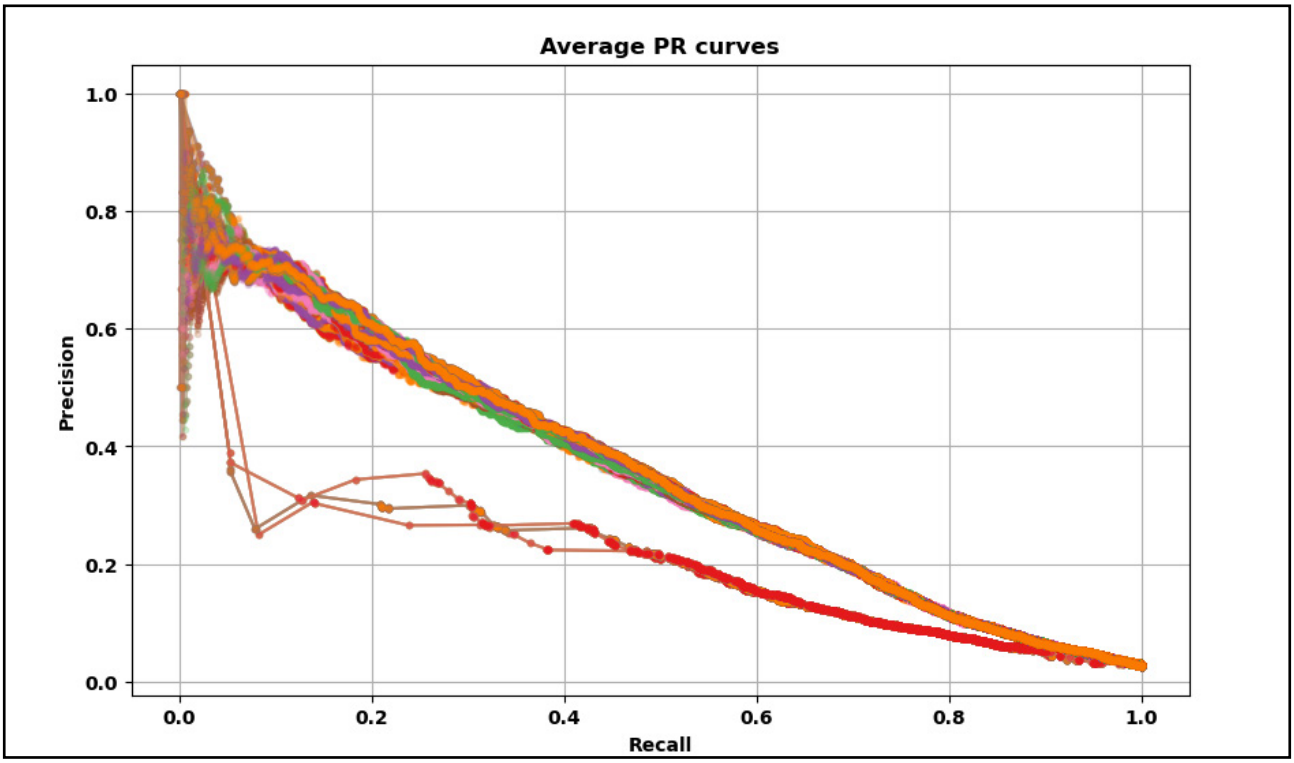

Figure 6 - The averages of 5-fold cross-validation predictions for 108 combinations of a BRF model.

[6] For an open-source example on a similar dataset see section 5 of this notebook: https://github.com/NGO-Algorithm-Audit/ML-pipeline/blob/main/BRF_pipeline.ipynb

About Algorithm Audit

Algorithm Audit is a European knowledge platform for AI bias testing and normative AI standards. The goals of the NGO are three-fold:

Knowledge platform	Bringing together experts and knowledge to foster the collective learning process on the responsible use of algorithms, see for instance our AI Policy Observatory and position papers
Normative advice commissions	Forming diverse, independent normative advice commissions that advise on ethical issues emerging in real world use cases, resulting over time in algoprudence
Technical tools	Implementing and testing technical tools for bias detection and mitigation, e.g, bias detection tool, synthetic data generation
Project work	Support for specific questions from public and private sector organisations regarding responsible use of AI

Structural partners of Algorithm Audit

SIDN Fund

The SIDN Fund stands for a strong internet for all. The Fund invests in bold projects with added societal value that contribute to a strong internet, strong internet users, or that focus on the internet's significance for public values and society.

European AI&Society Fund

The European AI&Society Fund supports organisations from entire Europe that shape human and society centered AI policy. The Fund is a collaboration of 14 European and American philantropic organisations.

Building *AI auditing* capacity from a *not-for-profit* perspective

www.algorithmaudit.eu

www.github.com/NGO-Algorithm-Audit

info@algorithmaudit.eu

Parkstraat 22, 2514 JK The Hague

Algorithm Audit

Stichting Algorithm Audit is registered as a non-profit organisation at the Dutch Chambre of Commerce under license number 83979212