INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA DE INTELIGÊNCIA ARTIFICIAL




Using Machine Learning to classify responsible from non-responsible online gamblers

DIOGO MOTA NEVES
Julho de 2023

P.PORTO

Using Machine Learning to classify responsible from non-responsible online gamblers

Diogo Mota Neves

Student No.: 1210057

A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science, Specialisation Area of Artificial Intelligence

Supervisor: Doctor Isabel Cecília Correia da Silva Praça Gomes Pereira, Coordinator Professor, Polytechnic of Porto - School of Engineering

Evaluation Committee:
President:
Doctor António Constantino Lopes Martins, Assistant Professor, Polytechnic of Porto - School of Engineering

Members:
Doctor Carlos Fernando da Silva Ramos, Principal Coordinator Professor, Polytechnic of Porto - School of Engineering
Doctor Isabel Cecília Correia da Silva Praça Gomes Pereira, Coordinator Professor, Polytechnic of Porto - School of Engineering

Porto, July 2, 2023

Abstract

In recent years, responsible gaming (RG) has become increasingly important to companies whose main activity relies on betting or casino activities and, as more and more people have begun to gamble online. With the proliferation of online gambling, it has become easier for individuals to access gambling sites and place bets from the comfort of their own homes, which can increase the risk of non-responsible gaming practices.

Betting companies, such as Betano, that prioritize responsible gaming are more likely to attract and retain customers, as individuals are more likely to trust and feel comfortable using a platform that takes steps to ensure that their gambling habits are safe and controlled.

In addition to the business benefits, responsible gaming is also important from a social and ethical perspective. Gambling addiction can have serious consequences for individuals and their families, including financial, relationship, and mental health problems. By taking steps to promote responsible gaming, Betano can help to minimize the negative impacts of gambling and contribute to the well-being of its customers.

This research paper will address the issue of responsible gaming for Betano by exploring the use of artificial intelligence (AI) and machine learning (ML) techniques to detect problematic gambling behaviors.

By utilizing AI and ML, Betano can better understand and predict gambling behaviors and take proactive steps to promote responsible gaming. This project will analyze the potential benefits and limitations of using these technologies in the context of responsible gaming.

Keywords: Non-Responsible Gaming, Responsible Gaming, Machine Learning, Artificial Intelligence, Gambling Problems, Gaming Problems

Resumo

O tópico de jogo responsável teve uma crescente importância nas empresas cuja atividade principal anda em torno do jogo de casino e apostas desportivas, principalmente pelo facto de cada vez mais apostadores utilizarem os meios eletrónicos para o fazer confortavelmente nas suas casas.

Este aumento de procura das apostas desportivas e casino online levou também ao aumento do número de casos de jogo não responsável.

No entanto, as empresas, como a Betano, dá uma grande prioridade e especial atenção ao cumprimento das normas estabelecidas, não só para a comunidade se sentir mais segura enquanto aposta, mas também para reforçar bons hábitos de jogo e não criar dependência que poderá ser fatal para muitas famílias.

Este projeto aqui apresentado irá analisar esta área de jogo responsável, quais os seus componentes e propor um algoritmo usando aprendizagem máquina para detetar padrões problemáticos em indivíduos.

Ao analisarmos todos os casos de forma automática, a Betano consegue de melhor forma prever e tomar ações até antes dos casos se tornarem graves e sobretudo promover o jogo responsável.

Palavras-chave: Non-Responsible Gaming, Responsible Gaming, Machine Learning, Artificial Intelligence, Gambling Problems, Gaming Problems

Acknowledgement

I'd like to take this opportunity to express my gratitude to Doctor Isabel Praça, who has provided guidance throughout this journey. Her feedback and constructive criticism helped me to see beyond and encouraged me to delve deeper into this project.

Furthermore, I'd like to thank Kaizen Gaming for the opportunity to work with the ML team, they provided me not only knowledge and guidance but their time to help me every day and had the patience to accompany me on this steep learning curve. Their passion for the field and their belief in my potential was a powerful source of motivation, a very special thanks to Maria and Stefanos.

Last but not least, I'd like to also thank to my family and my dear friends. They have been my biggest supporters, providing emotional support, motivation, and strength for this work. I truly appreciate their faith in me and their continuous encouragement during challenging times.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
RF	Random Forest
XGB	Extreme Gradient Boosting
GBM	Gadient-boosting machine
LSTM	Long short-term memory
SMOTE	Synthetic Minority Oversampling Technique
PR	Precision and Recall
ROC	Receiver operating characteristic
AUC	Area Under the ROC Curve
RG	Responsible Gaming
RTP	Return to Player
RNG	Random Number Generator
STDDEV	Standard Deviation

1. Introduction

Responsible gaming is the practice of ensuring that individuals have the ability to gamble in a way that is safe and controlled. This includes measures such as setting limits on the amount of money that can be bet, providing information and resources for individuals who may be at risk of developing a gambling addiction, and implementing self-exclusion programs to allow individuals to take a break from gambling if needed.

1.1 The Problem

Gambling habits refer to the patterns and behaviors that people exhibit when they engage in gambling activities.These habits can vary widely from person to person and may be influenced by a variety of factors, such as emotional state, cultural background, and financial situation.

Some people may engage in online gambling as a form of entertainment and may have a healthy attitude towards gambling, while others may develop unhealthy gambling habits that can lead to problems such as financial debt and addiction. It is important for individuals to be aware of their own gambling habits and to seek help if necessary to ensure that their gambling activities are safe and responsible.

Individuals who are in an addictive state may not be able to accurately report their behaviors or may downplay the extent of their addiction. This is because addiction can alter an individual's perception and judgment, leading them to underestimate the negative effects of their behaviors. As a result, self-reported behaviors may not provide an accurate picture of an individual's addiction.

Second, by the time an individual is reporting their behaviors, they may already be in the midst of an addiction and experiencing negative consequences as a result. These consequences may include health problems, financial difficulties, or relationship issues, among others. In some cases, these consequences may be irreversible, making it too late to intervene and address the addiction effectively.

For these reasons, relying on self-reported behaviors alone is not always the best solution for understanding and addressing addiction, therefore the solution will be to develop a ML model that can identify the increasing risk behavior and raise flags when needed.

1.2 Sports Betting

Sports betting is a type of product offered by many digital sports betting operator like Betano.

At its most basic, sports betting involves placing a wager on the outcome of a sports event. It is offered on virtually every sport imaginable - from mainstream games like football, baseball, basketball, soccer, and horse racing.

1.2.1 The Language of Sports Betting

Some common terms used in sports betting:

- Odds: These determine how much a bet will pay out. Decimal odds show how much you'll get back for every 1€ wagered (including the original bet).
- Bet/ Wager: This is the amount of money staked on the outcome of an event. The terms 'bet' and 'wager' are used interchangeably.
- Bookie/Bookmaker: The individual or organization that accepts bets from bettors.
- Over/Under (Totals): A bet on whether the total number of points scored by both teams will be over or under a set number.

There are more terms, but these ones are the basic ones.

1.3 Casino Slot Machines

Casino slot machines is the other product offered by some digital casino operators.

These digital operators offer slots that are virtual interpretations of traditional slot machines you would find in physical casinos.

The Return to Player (RTP) percentage is another important aspect. This statistic shows how much a customer could expect to get back over a long period of play. For instance, if an online slot has an RTP of 96%, you might expect, theoretically, to get back 96€ for every 100€ wagered.

Behind every spin in an online slot is software called a Random Number Generator (RNG). The RNG ensures that every spin is entirely random. All online casinos operate under strict regulations and their RNGs are tested for fairness.

1.4 Objectives

This project's objectives aim to detect non-responsible gaming behavior, including identifying patterns of behavior that indicate a player may be at risk for developing a gambling problem, such as excessive spending or prolonged periods of play. Once this is detected there will be a mechanism in place to flag such players with non-responsible gambling behavior and soon after can be analyzed by a human. Overall, is to ensure the players are enjoying the game in a safe and responsible way.

1.5 Contributions

Here's a list of contributions this project brings:

- Identifying at-risk players: By analyzing patterns of behavior, this project could identify players who may be at risk for developing a gambling problem and provide them with appropriate resources and support.
- Personalized feedback: This project will provide players with personalized feedback and recommendations based on their gaming habits, helping them to make more informed decisions about their play.
- Improve player experience: By identifying players that may be at risk for developing a gambling problem, this project will help to ensure that all players are having a positive and safe experience.

- Compliance: This project will help online casinos and betting companies to comply with regulations related to responsible gaming and gambling addiction prevention.
- Business benefits: By implementing a culture of responsible gaming, we will help to attract and retain more players.

1.6 Document Structure

Introduction, this section would provide an overview of the problem being addressed, including the importance of responsible gaming and the potential consequences of gambling addiction. It would also introduce the research question and explain the significance of the study.

State of the art and literature review, this section reviews relevant literature on responsible gaming, gambling addiction, and machine learning. It also discusses previous studies that have attempted to use machine learning to detect problem gambling behavior.

Methodology, this section describes the methods used to develop and evaluate the machine learning modal, including the data sources, algorithms, and evaluation metrics used.

Results, this section presents the results of the study, including the performance of the machine learning modal in detecting responsible gaming behavior, and also includes any findings or insights that emerged during the study.

Discussion, this section discusses the implications of the results and the limitations of the study. It suggests areas for future research.

In conclusion, the summary of the main findings of the study provides an overall assessment of the effectiveness of the machine learning modal for detecting responsible gaming behavior.

References, list of all sources cited.

Appendices, any additional materials, such as code, datasets, or detailed explanations of methods.

2. State of the Art

We will review the various techniques, technologies strategies that are being used by other betting companies and researchers that promote responsible gambling and make a contribution to identifying such cases or training a data set of self-reported people, including the use of artificial intelligence and machine learning to detect problematic gambling behaviors.

2.1 Responsible gaming behavior

Understanding the risks and warning signs of gambling problems, and knowing where to seek help if needed is what responsible gaming means. In the following sections of this document, we will explore the importance of responsible gaming behavior from various authors that have documented their projects.

2.1.1 Predicting voluntary self-exclusion among online gamblers with machine learning: A multi-country real-world study

Hopfgartner et al. 2022

This first paper wants to answer the same question this thesis is aiming to achieve, the prediction of people that will voluntarily self-exclude.

The authors want to identify behavioral variables that might cause gamblers to self-exclude in the future, which could then be used to target potential high-risk gamblers and introduce them to self-exclusion as a responsible gambling (RG) tool as early as possible.

This thesis not only aims to have a tool but to have a model that given the right customer behaviors will say how likely he is to self-exclude in the future.

The features used from the data set were grouped into three categories:

Control Variables: These were demographic features of the players.

Behavioral Features: These included: Number of different payment methods used by the player. Number of limit changes made by the player. Number of previous self-exclusions by the player. Number of different game types played by the player.

Monetary Intensity Features: These reflected the amount of money gambled, lost, or deposited by the player.

Study Conclusions

Frequent use of responsible gambling tools (e.g., limit-setting and self-exclusions), frequent deposits within a session, using multiple payment methods, and playing multiple types of games were significantly associated with future self-exclusions. And although the authors don't disclaim what models they've used they found that ML models were successful at generalizing the features.

This is important as we'll keep these features in mind when creating our own feature set.

2.1.2 How do gamblers start gambling: Identifying behavioral markers for high-risk internet gambling

The aim for this paper is to identify specific behavioural markers for high-risk internet gambling by analyzing the actual gambling behaviour of individuals and predict gambling-related problems based on the characteristics of a gambler's initial stage of gambling.

K-means cluster was used to identify groups of gamblers based on their betting behaviour during the first month of gambling. And the main features used were the intensity and frequency of gambling and the variability of wager sizes during their first month in the platform. Braverman and Shaffer 2010

The authors concluded that gamblers that presented high intensity and frequency of gambling and by high variability of wager sizes during their first month of gambling were at higher risk than other gamblers to report gambling-related problems upon closing their accounts. They suggested that methods like k-means cluster analysis could be beneficial for investigating other high-risk groups using different variables that describe gambling behaviours.

Study Conclusions

This research paper aimed to identify behavioral markers for high-risk internet gambling by examining the initial gambling behavior of individuals. Using a k-means cluster analysis to group gamblers based on betting behavior in their first month. The study found that gamblers displaying high intensity and frequency of betting, along with high variability in wager sizes during their initial month, were more likely to report gambling-related problems upon closing their accounts.

We can extract from this study is that some patterns might emerge right at the beginning of a customers journey, and some features can be extracted taking into consideration the time period.

2.1.3 Applying Data Science

This research has sought to identify high-risk gamblers using behavioral tracking data by looking at fine-grained analyses of bet-by-bet behavior within a play session. Loss chasing, the tendency to continue gambling or increase one's bet in an effort to recover debts, is often regarded as a defining feature of problem gambling and may be detectable behaviorally from player tracking data.

An analysis of 600,000 hands from the Full Tilt online poker site also examined how players responded following high-magnitude gains and losses. Other analyses of bet-by-bet behavior have focused on specific psychological phenomena such as winning or losing streaks. The text concludes that these studies show considerable promise in identifying behavioral markers of loss chasing. However, existing studies have yet to link these behavioral expressions with red-flag indicators of disordered gambling. Deng, Lesch, and Clark 2019

Study Conclusions

This research aimed to identify high-risk gamblers through an analysis of behavioral tracking data, focusing on loss chasing and response patterns to significant wins or losses in online poker. While the data shows potential in detecting loss chasing, a critical aspect of problem gambling, a connection between these behaviors and definitive signs of disordered gambling is yet to be established.

Also is important to note that the study wanted to identify high-risk customers, which although is a different subject the main points are there and these subjects can relate to one another.

2.1.4 A Proposed Set of Features on Implementing Responsible Gambling on Slot Games with G2S Technology

Although the paper does not explicitly mention the use of specific machine learning models, they have implemented a protocol for data exchange called Game-to-System (G2S) that intersect data between Electronic Gaming Machines and backend servers. The protocol is used to implement responsible gambling practices on slot games. Siu, Hoi, and Chan 2021

The features used in this paper are related to the implementation of responsible gambling practices such as Siu, Hoi, and Chan 2021:

- Time Interval: The minimum time interval between 2 consecutive game plays.
- Play Time Limit: The time period of a consecutive gameplay session; if reached, a warning message, an alert message, or a lock on the slot game or player card will be displayed respectively.
- Daily Time Limit: The maximum time to play in a day; if reached, the game and player card will be locked. Bet Limit: The cumulative wager limit for a gameplay session; if reached, a warning message, an alert message, or a lock on the game and player card will be displayed respectively.
- Daily Bet Limit: The maximum amount of wagers to play in a day; if reached, the game and player card will be locked.
- Game Played Limit: The maximum number of games played; if reached, a warning message, an alert message, or a lock on the game and player card will be displayed respectively.
- Daily Game Played Limit: The maximum number of games to play in a day; if reached, the game and player card will be locked.
- Credits Played Limit: The amount of credits wagered within a gameplay session; if reached, a warning message will be shown and a maximum amount of credits allowed to play is set.

Study Conclusions

They found that the proposed system can successfully limit the time, bets, and games played by a player, which can help to prevent problem gambling. The system also provides players with warning messages and alerts when they approach or reach their limits, which can help them to manage their gambling behavior more effectively.

2.1.5 Empowering responsible online gambling by real-time persuasive information systems

The paper discusses the use of persuasive information systems to help gamblers self-regulate their behavior, also, it explores the role of emotions in gambling and how they can be used in persuasive technology to facilitate responsible gambling. Drosatos et al. 2018

The authors propose a hybrid approach that includes an end-user intelligent agent and a research platform for responsible online gambling. The end-user intelligent agent is designed

to guide gamblers to self-regulate their problematic behavior, providing high privacy guarantees. The research platform is designed to support research about persuasive technologies that could be applied to users with addictive behavior.

The paper also discusses the implications of GDPR (General Data Protection Regulation) for data processing and transfer, suggesting that online gambling providers could offer an "opt-in" model for behavioral analysis aimed at improving public health.

The paper concludes by discussing the challenges and opportunities in this area, including the need for more research on emotions in gambling, the potential of smartphone technology in psychiatry and health behavior, and the balance between corporate social responsibility and commercially sensitive data.

Here's a list of features used in this paper:

- Online Gambling Experience Data: This includes data recorded by gambling operators, such as betting history, spending time, amount of money, and online status.
- Multimodal Data: This category includes data captured by the user's personal devices, such as geolocation, accelerometer data, heart rate, and eye tracking data.
- Emotional Analysis: The paper also discusses the development of emotion models that process and analyze signals captured from various sensors.
- User's Gambling Activity: The system aggregates users' gambling activity from online gambling operators' APIs.
- Sensor Data: The system has access to sensor data (geolocation and accelerometer) from the user's personal device and/or retrieves data from other third-party applications that manage wearable sensors (e.g., Fitbit).
- User's Profile Data and Goals: The system uses the user's profile data, gambling activity, and the goals that the user has set for themselves to provide personalized intervention.
- Account Statements: This includes a list of transactions in the user's virtual wallet that contains data such as date and time, description of the transaction, debit or credit amount, and balance.

Study Conclusions

The authors of the paper argue for the use of persuasive information systems to support responsible online gambling.

They propose that these systems should leverage gamblers' online behavioral, emotional, and profile data to help regulate their gambling habits. They believe that such technology can enhance the potential of behavior change techniques, but they also acknowledge that it may have unintended consequences.

The authors suggest that understanding the emotions involved in gambling is crucial for the success of these persuasive technologies.

The paper also discusses the challenges and risks associated with this approach. These include the potential for unsustainable change, loss of interest, failure to engage, and the risk of triggering alternative addictive behaviors.

The authors conclude that while persuasive information systems can support responsible online gambling, there are potential risks and unintended consequences that need to be carefully managed.

2.2 Datasets Research

In this section, we'll present the various public datasets found so we can have a starting point and also what researchers found useful in said data sets and the successful and non-successful approaches.

2.2.1 Public available datasets

In the field of ML we've some researchers that took into consideration a dataset from "The Transparency Project" *The transparency project* n.d. this dataset is a collection of self-reporting users who played in the "bwin" platform from November 2008 and November 2009 that did not comply with responsible gaming.

The Transparency Project provided three tables: Demographic information of the subscribers, the gambling history of each subscriber, and the information associated with the Responsible Gaming intervention on each flagged user. Rose n.d.

Here's the important field from each table in this data set Rose n.d. :

rg_case	Boolean	Whether the subscriber had an RG intervention
country_name	Text	The country the user registered.
language	Text	User's Primary Language
gender	Text	User's gender
registration_date	Date	User's registration date with Bwin.
first_deposit_date	Date	User's first cash deposit with Bwin.

Table 2.1: Demographic Information

date	Date	Date of associated data.
product_type	Integer	ID of the product type
turnover	Number	Total stakes on the given day and product (in Euros)
hold	Number	Total amount lost on the given day and product in (Euros).
num_bets	Integer	Number of "bets" placed on the given day and product.

Table 2.2: Gambling Behavior

events	Integer	Number of RG events the user had.
first_date	Date	Date of the user's first RG event.
last_date	Integer	Date of the user's last RG event.
event_type_first	Number	The ID of the type of the first RG event.
inter_type_first	Number	The ID of the type of the first intervention from Bwin.

Table 2.3: Responsible Gaming Intervention Information

2.2.2 Public available datasets problem

After researching most of the available studies published in this field it was clear that many datasets are kept private and only briefly mentioned in the studies, one of these examples comes from "Using machine learning to predict self-exclusion status in online gamblers on

the PlayNow.com platform in British Columbia" that although they do not share the dataset they give a general direction of what data they had.

Another very strong problem is that these data sets only contain very few features, and are fully made out of the positive cases, which would be "Non-Responsible Gaming" cases, and authors fail to mention how they overcome having only the data from positive cases, which is a very important aspect in this thesis.

2.2.3 State of the art conclusions

The current state of the art helped to understand a few aspects of this thesis scope, and where to aim:

A lot of research proved useful when authors are classifying data, we can aim to have some similar features that had good results in the past and add to it. An example of this would be the frequency of players in their early stage, and we can add to that by contrasting it with their late stage data.

Not every non-responsible customer has non-responsible behaviors since the beginning of their journey and since this model should have the most recent data at any given time, we should aim to understand if late stage data has more importance and can predict with more confidence these cases.

Another point was that while there is some overlap between responsible gaming and machine learning, there are relatively few studies that have combined them to explore the full potential of using machine learning to enhance responsible gaming practices. One example of this would be to add limitations on an account based on the predicted label.

Moreover, while there are some studies that use machine learning to detect patterns of problem gambling, these studies are often limited in their scope and the models used, and research on this area is still in its early stage. Therefore, there is a significant opportunity for future research to investigate the use of machine learning to enhance responsible gaming and detect problematic gambling behavior.

3. Data

Data plays an important role in the successful execution of any machine learning project. Relevant and diverse data enables the creation of robust, accurate models capable of making insightful predictions and understanding complex patterns.

This chapter will discuss how can we better understand the data at our hands, analyze the problems of imbalanced data, stratification, and labeling, and how that can lead to inaccuracy and bias, showcasing the importance of data in determining the overall success of the model.

3.1 Available data

This is the data provided by Betano for this project:

3.1.1 Customer Table

Name	Description
Customer Id	Unique identifier for each customer.
Age	Customer's age
Gender	Customer's gender
Registration Date	Customer's registration date with Betano.
Location	Customer's region or location.
Label	If the customer is Non-RG compliant or if the customer is RG compliant

Table 3.1: Customer data table

Transaction Table

Name	Description
Transaction Id	Unique identifier for each transaction.
Customer Id	Unique identifier for each customer.
Date	Transaction's date
Type	Type of transaction can be a deposit, withdrawal, sports book bet, or casino
Amount	Transaction Amount in Euro.

Table 3.2: Customer transaction Table

With this available data, we can extract some crucial information to support the model afterwards.

3.2 Understanding the data

One customer might have zero or more transactions, from different types, and usually, it starts with a deposit transaction with some amount of euros, then we have multiple

transactions for sports betting or casino slot machines, and we can have some withdrawals in between. This is the normal cycle for each customer.

Because each transaction is only connected to one customer we can have the count for every customer on how many transactions they made filtered by sports betting, casino, deposit, or withdrawal.

Each transaction has a date, so we can also count how many transactions a customer has made within 24h frames and how much money he spent during that period.

We can also see how many customers there are related to their country of origin in the data set. And which country has more non-RG cases.

3.3 Interview with the Responsible Gaming Team

The responsible gaming team is accountable for carrying out an evaluation for each customer presented as a possible non-RG-compliant case. They are the ones that must take action once this they are identified as such and dispatch the needed resources in order to help the customer at hand.

Because their every day is focused on analyzing player data they present themselves as a possible first step towards this project and understanding the data at our hands. They can give us the insights needed to better suit this data into new more valuable features that will help us and the machine learning model to better handle the data.

So we've elaborated a questionnaire that was promptly answered.

3.3.1 Question 1 - How does the RG team decide if a customer is playing normally or not?

Non-RG customers usually play on both products (sports betting and casino slots) and a big variety of games.

3.3.2 Question 2 - Do you see more RG cases based on personal details like age, gender, or location?

No. However statistical analysis based on label distribution in those groups to answer.

3.3.3 Question 3 - Do these customers often ask for more bonuses in the chat or talk more in general with customer support?

Yes, that is in fact the first reason for a customer to be checked by the RG team.

3.3.4 Question 4 - How far back in a player's history do you need to look to be sure it's an RG case? Can you tell from just a few initial bets?

Both all-time statistics and Profit and Loss (P&L) and more explicitly transactions from last week/ month. Mostly looking for increased tendencies and increased losses over time.

3.3.5 Question 5 - Do non-RG customers have the same pattern of play throughout the year?

Every non-RG player is examined based on himself and not the average of our customers. An RG player demonstrates standard losses through time, with no spikes in transactions.

3.3.6 Question 6 - What actions do you take following a detected RG Case?

In some cases, we may permanently or temporarily close the account. We may also apply restrictions of use.

3.3.7 Responsible Gaming Q&A Conclusions

The RG team provides a crucial role in analyzing player data and determining whether a player is complying with RG guidelines. Their daily focus on player data interpretation and behavioral patterns makes them an invaluable asset in understanding how to best utilize this data for creating better features.

The RG team uses a variety of metrics, from player betting behavior across games to communication patterns with customer support, to distinguish between responsible and non-responsible players.

It's clear from the interview that identifying non-responsible RG players is not a simple process. It requires a deep analysis of player behavior, both in their gaming habits and their interactions with customer support.

The fact that requests for more bonuses in the chat are a primary trigger for review by the RG team suggests that financial behaviors play a significant role in RG compliance. Additionally, the analysis involves both short-term and long-term assessment, examining trends in recent transactions as well as overall player history.

However, it's important to note that each player is evaluated individually, and demographic details such as age, gender, and location don't necessarily dictate the likelihood of non-responsibility. Once a non-responsible RG case is detected, the team can employ several measures, ranging from account restrictions to temporary or permanent closure, to manage the situation.

3.3.8 Data imbalance problems

It is important to note that we have a huge data imbalance between the labels, only 0.05% of the data is labeled as "non-responsible gaming customers", while the rest are labeled as "responsible gaming customers", this will affect how we proceed with the implementation.

Data imbalance is not only a problem in this work many authors struggle to get a balanced dataset such as spam filtering, credit card fraud detection, and software defect prediction. 3.1

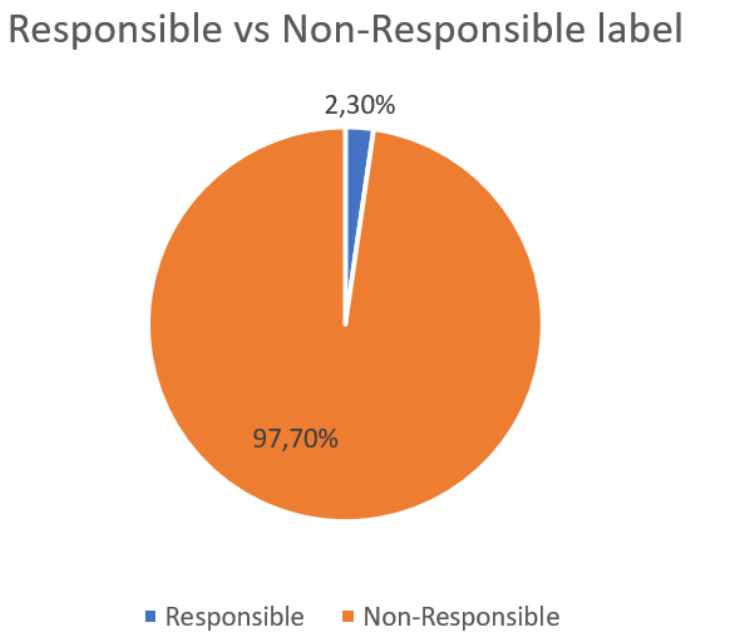

Figure 3.1: Data imbalance Responsible vs Non-Responsible label

Extreme learning machine autoencoders

Taking a look at the literature "Imbalanced Data Classification Based on Extreme Learning Machine Autoencoder" by Chu Shen, Su-Fang Zhang, Jun-Hai Zhai, Ding-Sheng Luo, and Jun-Fen Chen focuses on the problem of binary imbalanced data classification. Shen et al. 2018

The authors propose a method based on the extreme learning machine autoencoder (ELM-AE) to address this problem. The ELM-AE is a generative model, a type of two-layer feed-forward neural network with an encoder and a decoder. The weights and biases of the first layer (encoder) are randomly assigned, and the second layer (decoder) weights are analytically determined. Shen et al. 2018

The proposed method involves three steps:

- Positive instances are used as seeds, and new samples are generated using ELM-AE. These new samples are similar to the positive instances but not identical.
- Step 1 is repeated several times to obtain a balanced dataset.
- A classifier is trained with the balanced dataset and used to classify unseen samples.

The algorithm is iterative and terminates when the number of augmented positive instances equals the number of negative instances. The difference between the probability distributions of the input data and the output data is measured using Maximum Mean Discrepancy.

SMOTE (Synthetic Minority Oversampling Technique)

SMOTE is a method that addresses the imbalance by oversampling the minority class. Instead of duplicating existing examples from the minority class, SMOTE synthesizes new examples. It works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space, and drawing a new sample at a point along that line.Brownlee 2020

However, SMOTE isn't perfect, a downside is that synthetic examples are created without considering the majority class, possibly resulting in ambiguous examples if there is a strong overlap for the classes.

With this information in mind, and some authors agreeing that SMOTE might not be the best solution when dealing with a lot of features Seitz 2022 we'll need to take into consideration the data imbalance in the model itself.

3.3.9 Data labeling

Understanding how the data was labeled will help us understand better the results. This dataset like many other self-exclusion datasets is labeled by each and every individual customer, that have to flag themselves.

This is extremely important to understand because there are many reasons why customers opt to self-exclude. Some of these reasons include the amount of money spent, the amount of time spent, personal problems, and others.

This means that many of our "non-responsible" label is not always tied to money, there are other factors into play.

Not only is the positive label important but the negative label is also a critical point as many customers might be already in a state of "non-responsible" but because they've not identified themselves as such, we will have them as "responsible". This is why it is important to understand that although this dataset is good when we analyze some of these customers at large, we have some inconsistencies and this problem is not easily solved by manually labeling as it is needed a lot of time to label these cases taking into account each individual case.

3.3.10 Data stratification

Data stratification is a method used to divide a population into homogeneous subgroups, based on specific conditions. This technique is often used in data analysis and research settings, as it allows for more precise and robust conclusions to be drawn from the data. Huyen 2022

By segregating the data, we can ensure that the sample accurately represents the different subgroups within the population. One example of this could be to create a stratum evolving the amount of money spent, this means that we would ensure that there's a percentage group of customers that will have low amounts of deposits and an equal amount that would have high amounts of deposits. Huyen 2022

This approach is particularly valuable when the separation exhibits differences that could have a significant impact on the overall analysis.

However, for the purposes of this project, we've opted for simple random stratification. This method involves dividing the population randomly, without taking into account any specific characteristics or conditions. While this method may not deliver the depth of understanding that comes from stratification based on particular variables, it provides the advantage of simplicity and ease of execution. It ensures that our sample is representative of the population and that our findings can be generalized with fewer assumptions.

3.3.11 Business rules

Last, but not least we need to filter some of the data using some pre-defined business rules, this are the rules applied:

Rule Name	Description
Low Deposits	Customers must have at least some money in the platform since we might have customers that make the registration and then leave the platform.
Sports Betting Activity	Customers must have some significant bet behavior, they should have a couple of bets in the sports betting to be considered in the training data.
Casino, Slot Machines Activity	Customers must have some significant casino behavior, they should have a couple of games played in the casino to be considered in the training data.

Table 3.3: Business Rules

4. Feature Engineering

The principle of responsible gaming encourages practices that protect players from the potential risks associated with gambling, including excessive gaming and addiction. In this context, feature engineering involves creating representative features from raw data that can improve the performance of machine learning algorithms.

As such, this section will address all the features created taking into account the available data, and the state-of-the-art. As many authors agreed that their features produced good results we'd like to leverage some of the same features to this project, such as the amounts gambled in a certain period of time and the variance of the gambling values.

4.1 Sports betting Features

Taking the information about our knowledge in sports betting combined with the RG questionnaire we can elaborate on a few features that might help us with getting a better grasp of each customer activity.

These features should have the overall activity that a customer had until now, such as the total amount of bets he has made, to give the model a point if he bets a lot or not.

Also, we need features that represent his most recent activity, so we also gathered the last seven days of betting activity.

Here is a list of all the features related to sports betting 4.1:

4.2 Casino Slots Features

Just like we did for the sports betting features we'll take the same approach to the features for Casino Slot Machines, getting a full picture since the customer registration and the last 7 days of their casino activity.

Here is a list of all the features related to casino slots 4.2:

4.3 Deposit and withdrawal features

Like the other features were inspired by the RG team and the state of the art, these one are similar. Most of what the RG team work's is to go through and analyze deposits and withdrawal historical data, for that matter we've also engineered some features related to the deposits and withdrawals. 4.3

4.4 Missing Values

Like many others, our dataset has a lot of zero values and this is not a random, when we're joining features together, such as time and money spent of bets placed there will be many which didn't had such an activity during that period, and that's intended. If a customer didn't had any activity either by placing bets or playing in the slot machines we're expecting

Name	Description
Total_Number_of_Bets	Total number of bets a customer has made since his registration
Average_Bet_Amount	Average amount of euros in all bets a customer has made since his registration
Maximum_Bet_Amount	Maximum amount of euros in all bets a customer has made since his registration
Minimum_Bet_Amount	Minimum amount of euros in all bets a customer has made since his registration
Stddev_Bet_Amount	Standard deviation among all bets a customer has made since his registration
Betting_Frequency	Frequency of bets made per day a customer has made since his registration
Total_Number_of_Bets_last_7	Total number of bets a customer has made in his last 7 days of activity
Average_Bet_Amount_last_7	Average amount of euros in all bets a customer has made in his last 7 days of activity
Maximum_Bet_Amount_last_7	Maximum amount of euros in all bets a customer has made in his last 7 days of activity
Minimum_Bet_Amount_last_7	Minimum amount of euros in all bets a customer has made in his last 7 days of activity
Stddev_Bet_Amount_last_7	Standard deviation among all bets a customer has made in his last 7 days of activity
Betting_Frequency_last_7	Frequency of bets made per day a customer has made in his last 7 days of activity

Table 4.1: Sports betting Features

to miss these values and feed it as they are to the model, this is a precious features and it has a valuable meaning.

Name	Description
Total_Number_of_Casino	Total number of spins in a slot machine a customer has made since his registration
Average_Casino_Amount	Average amount of euros in all casino games a customer has made since his registration
Maximum_Casino_Amount	Maximum amount of euros in all casino games a customer has made since his registration
Minimum_Casino_Amount	Minimum amount of euros in all casino games a customer has made since his registration
Stddev_Casino_Amount	Standard deviation of the amount of euros in all casino games a customer has made since his registration
Casino_Frequency	Frequency of casino activity made per day a customer has made since his registration
Total_Number_of_Casino_last_7	Total number of casino spins a customer has made in his last 7 days of activity.
Average_Casino_Amount_last_7	Average amount of euros in all casino games a customer has made in his last 7 days of activity
Maximum_Casino_Amount_last_7	Maximum amount of euros in all casino games a customer has made in his last 7 days of activity
Minimum_Casino_Amount_last_7	Minimum amount of euros in all casino games a customer has made in his last 7 days of activity
Stddev_Casino_Amount_last_7	Standard deviation of the amount of euros in all casino games a customer has made in his last 7 days of activity
Casino_Frequency_last_7	Total number of spins in a slot machine a customer has made in his last 7 days of activity

Table 4.2: Casino Slots Features

Name	Description
Total_Deposit_Amount	Total amount of euros deposited since a customer's registration
Number_of_Deposits	Total number of deposits since a customer's registration
Average_Deposit_Amount	Average amount of euros deposited since a customer's registration
Standard_Deviation_of_Deposit_Amount	Standard deviation of amount of euros deposited since a customer's registration
Total_Withdrawal_Amount	Total amount of euros withdrawn since a customer's registration
Number_of_withdrawals	Total number of withdrawals since a customer's registration
Average_withdrawal_Amount	Average amount of euros withdrawn since a customer's registration

Table 4.3: Deposits and Withdrawal Features

5. Development

In this chapter, we'll dive deeper into the models we're going to be using alongside other techniques that will help us archive a good result. Since we know that we'd like to continue exploring this problem as a classification problem we're going to analyze a few alternatives, random forests, extreme gradient boosting (XGBoost), and LightGBM.

All the models here presented were made with the help of MLFlow which is a tool that is used to manage the machine learning lifecycle from initial model development. *MLFlow Documentation* n.d. This tool is great because it tries different approaches given a problem and it tries to get the best hyperparameters by giving a context.

5.1 Evaluating the model

Before we jump into the ML models it is important to note that we'll use the same metrics to draw conclusions.

5.1.1 Confusion Matrix

The Confusion Matrix, as an integral part of ML, is a potent tool for evaluating the performance of classification models, offering a detailed snapshot of the results obtained from such models. it presents a visual representation of the model's performance, it provides four primary metrics - True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), which help in understanding the intricate details of the classifier's accuracy, precision, recall, and F1-score Tharwat 2018.

5.1.2 ROC Curve

The ROC curve is a graphical representation used in statistics and machine learning to measure the performance of a binary classification system. It plots the true positive rate against the false positive rate for different threshold values. The area under the curve (AUC) is a single number summary of the ROC curve and provides an aggregate measure of performance across all possible classification thresholds Fawcett 2006.

The ROC curve and AUC are widely used because they are independent of the threshold set for classification. They provide a way to compare classifiers and measure accuracy, especially in cases where there is an imbalance in the observation between the two classes Metz 1978.

5.1.3 Precision and Recall

Precision and Recall are fundamental evaluation metrics in ML, specifically in the field of classification. These metrics offer insight into the performance of an algorithm in terms of its correctness and completeness Davis and Goadrich 2006.

Precision and Recall are often inversely related, an increase in Precision typically results in a decrease in Recall, and vice versa. This is known as the Precision-Recall trade-off. Fawcett 2006.

Precision and Recall formulas:

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

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

5.1.4 F1 Score

The F1 score is a measure of a test's accuracy and is used in many areas, including machine learning and statistics. It considers both the precision and the recall of the test to compute the score. The F1 score is the harmonic mean of precision and recall. Korstanje 2021

The F1 score is often used in situations where you want to balance precision and recall and there is an uneven class distribution. For example, in text classification where the number of negative samples greatly outweighs the number of positive samples. Korstanje 2021

F1 Score formula:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

5.2 Random Forests

Random Forests are among the most popular machine learning algorithms today. They're known for their simplicity, making them an attractive option for tackling a variety of tasks. For our purpose, this would be classification.

5.2.1 The Strengths of Random Forests

Random forests have been applied to a wide range of problems, including image classification, text classification, and financial forecasting. Fleiss 2023

They are considered:

- Robustness: Random Forests can handle noisy data and outliers. They're less likely to overfit the data, which means they can generalize well to new data.
- Accuracy: Random Forests are among the most accurate machine learning algorithms. They can handle both classification and regression problems and can work well with both categorical and continuous variables.
- Speed: Despite being a complex algorithm, Random Forests are fast and can handle large datasets. They can also be easily parallelized to speed up training.
- Feature Importance: Random Forests provide a measure of feature importance, which can help in feature selection and data understanding. Fleiss 2023

5.2.2 The Weaknesses of Random Forests

However, like any algorithm, Random Forests have their own drawbacks. Here are some potential pitfalls to be aware of:

- Overfitting: While Random Forests are generally less prone to overfitting than a single decision tree, they can still overfit the data if the number of trees in the forest is too high or if the trees are too deep.
- Interpretability: Random Forests can be less interpretable than a single decision tree because they involve multiple trees. It can be difficult to understand how the algorithm arrived at a particular prediction.
- Training Time: The training time of Random Forests can be longer than other algorithms, especially if the number of trees and the depth of the trees are high.
- Memory Usage: Random Forests require more memory than other algorithms because they store multiple trees. This can be a problem if the dataset is large. Fleiss 2023

5.2.3 Real-World Applications of Random Forest

Random Forest has found its way into many different sectors. Here are some applications:

- Banking Industry: Random Forest can classify transactions as fraudulent or legitimate by analyzing patterns in the data, which we can also learn from these patterns and apply in our context.
- It's also used for credit card fraud detection, as it can handle highly imbalanced datasets.
- Customer Segmentation: Random Forest can be used for customer segmentation by analyzing expenditure amounts of different customers.
- Healthcare and Medicine: It's used in predicting cardiovascular diseases, diabetes, and breast cancer, demonstrating its potential in the healthcare sector. Meena 2020

5.2.4 Random Forest Model key conclusions

Random Forest could prove highly effective in detecting non-responsible customers on an online platform. The process would start with the identification of relevant features that can distinguish non-responsible customers from responsible ones.

A Random Forest model would be trained on these new features. The robustness and accuracy of Random Forest would make it an ideal choice for this task, given its capacity to handle noisy data and outliers, which are typically common in real-world data. Dash 2022

Once trained, the model could predict the likelihood of a gamer being non-responsible based on their behavior patterns.

5.3 Applying Random Forest Models

5.3.1 Random Forest Hyper-parameters space

As we said previously we've used MlFlow to try to give us some hyperparameter adjustment taking into consideration our problem, here's the space generated in two of our tries 5.1 and 5.2:

```
space = {
    "bootstrap": True,
    "criterion": "entropy",
    "max_depth": 11,
    "max_features": 0.6275669638191145,
    "min_samples_leaf": 0.08948461919582683,
    "min_samples_split": 0.2321692357988701,
    "n_estimators": 58,
    "random_state": 553404597,
}
```

Figure 5.1: Hyper-parameters space for Random Forest No. 1

```
space = {
    "bootstrap": True,
    "criterion": "gini",
    "max_depth": 8,
    "max_features": 0.3004200651519776,
    "min_samples_leaf": 0.17324077219294692,
    "min_samples_split": 0.20372775377651348,
    "n_estimators": 97,
    "random_state": 654776397,
}
```

Figure 5.2: Hyper-parameters space for Random Forest No. 2

5.3.2 Results

Now that we've defined our hyper-parameters we just need to train the model and get the scores, here's the confusion matrix for the experiments, the ROC curve, and the precision.

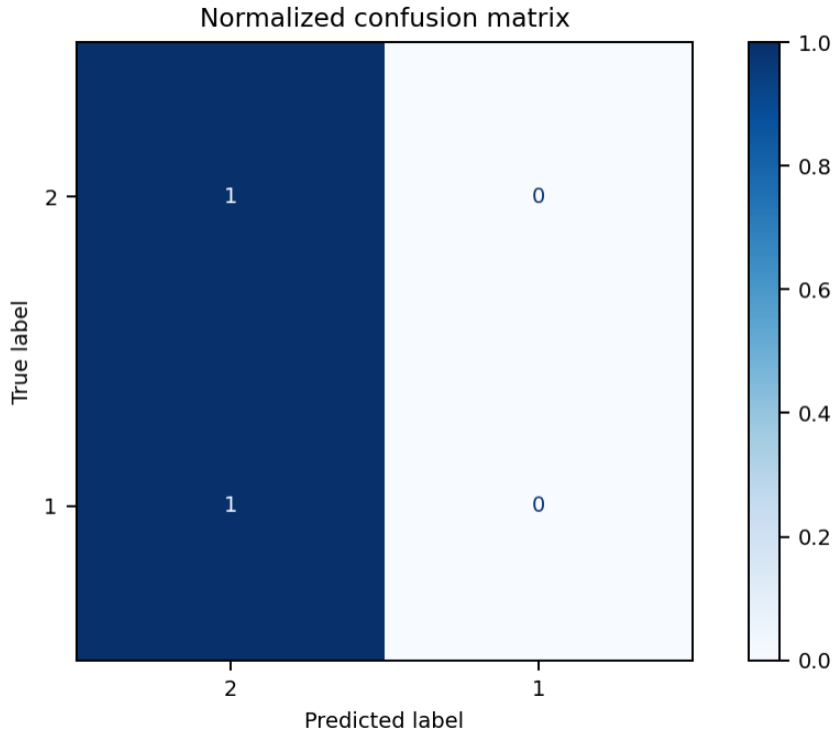

Figure 5.3: Confusion Matrix for Random Forest No. 1

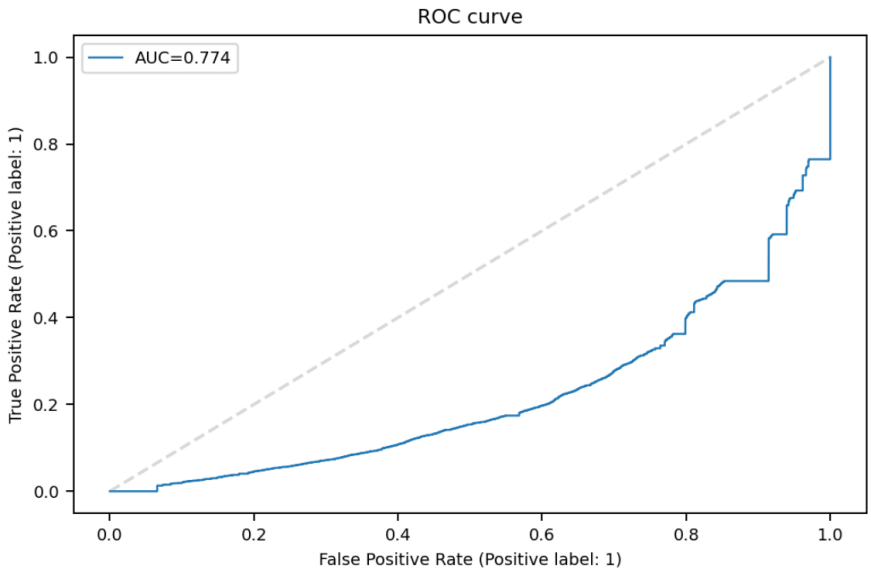

Figure 5.4: Roc Curve for Random Forest No. 1

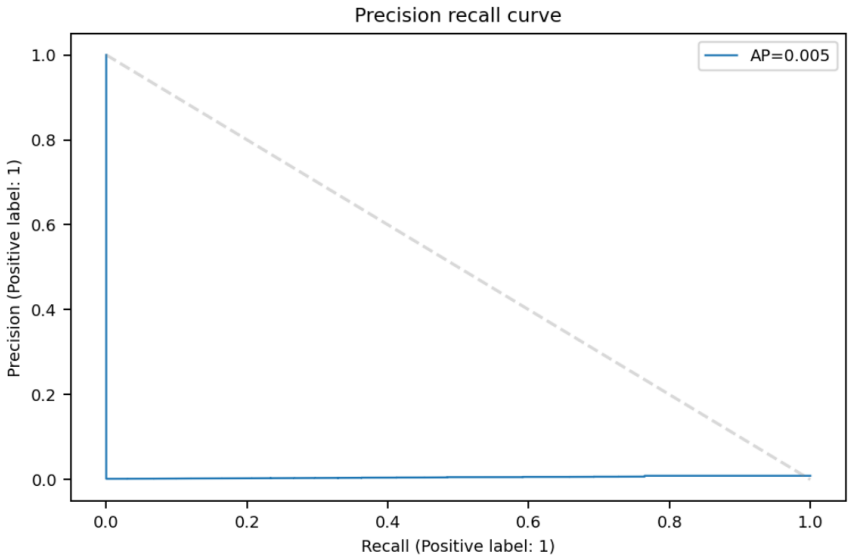

Figure 5.5: Precision Recall for Random Forest No. 1

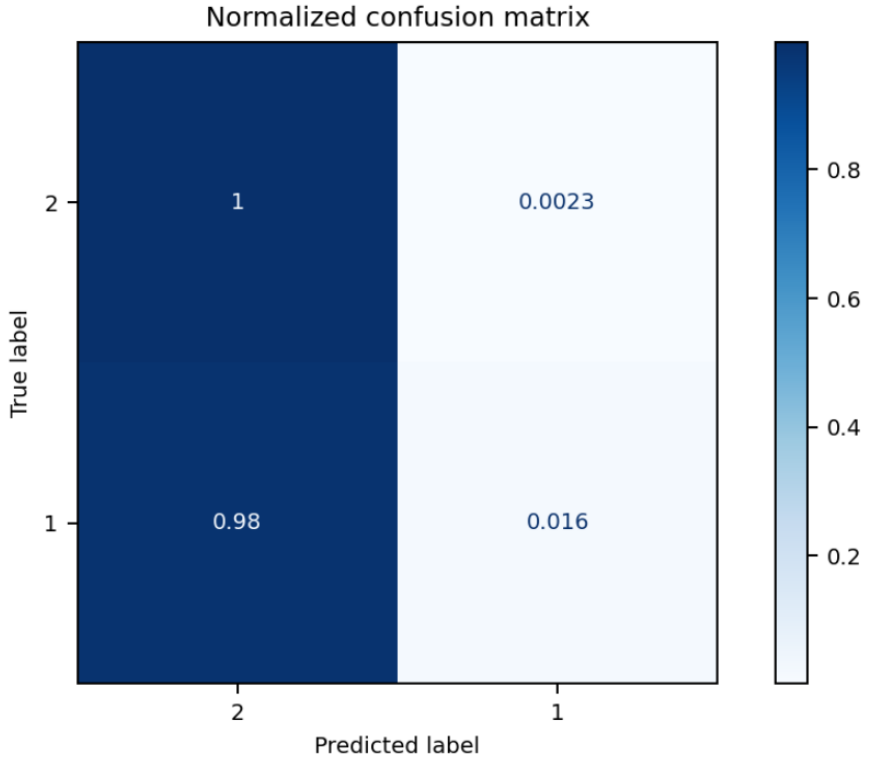

Figure 5.6: Confusion Matrix for Random Forest No. 2

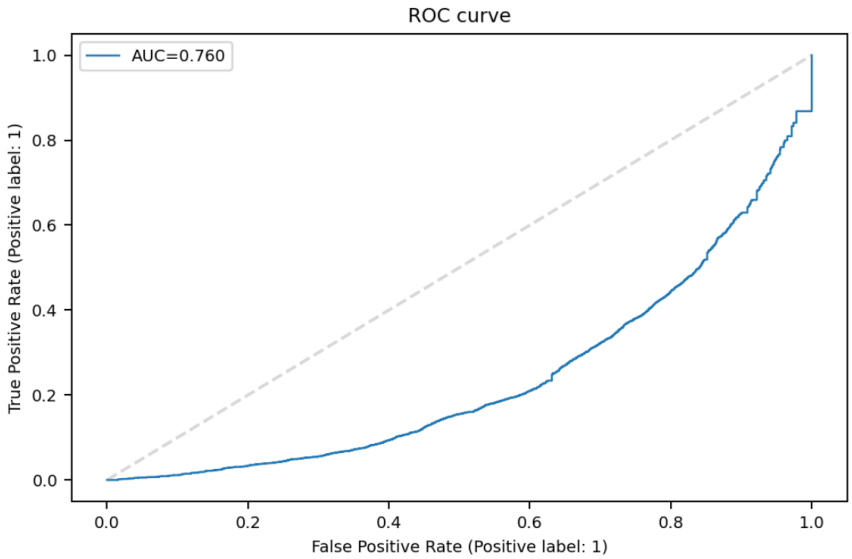

Figure 5.7: Roc Curve for Random Forest No. 2

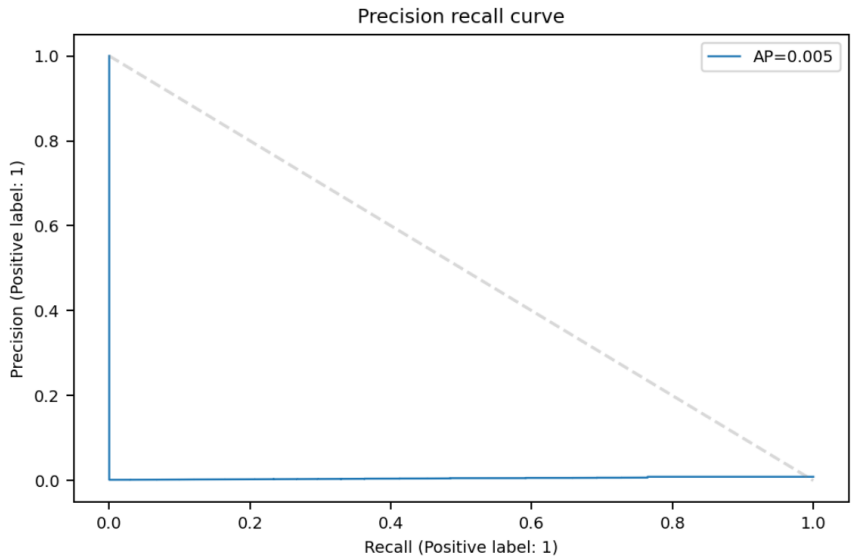

Figure 5.8: Precision Recall for Random Forest No. 2

5.3.3 XGBoost F1 scores

- Model 1 - 0.0
- Model 2 - 0.0

5.3.4 Random Forest Model Conclusions

We've seen that random forest is not very suitable for our project, the data imbalance makes the model always guess the "non-responsible" label which is the majority of our cases.

In the next sections, we'll analyze better models that are more capable of generalizing even with so much imbalance among our data.

5.4 Extreme Gradient Boosting (XGBoost)

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It provides a parallel tree-boosting solution that addresses many data science problems in a fast and accurate way. *XGBoost Documentation — xgboost 1.5.1 documentation* n.d.

XGBoost utilizes the power of ensemble learning, which combines several learners (models) to improve overall performance, thereby increasing predictiveness and accuracy in machine learning and predictive modeling. The ensemble models can combine thousands of smaller learners trained on subsets of the original data, leading to a decrease in variance and bias due to bagging and boosting. Hachcham 2023

5.4.1 XGBoost Bagging

Bagging is a technique that decreases overall variance by averaging the performance of multiple estimates. It aggregates several sampling subsets of the dataset to train different learners chosen randomly. The core idea of bagging is to use a voting mechanism for classification. Hachcham 2023

5.4.2 XGBoost problems

Overfitting is especially acute in XGBoost because it is designed to be a high-performance model, which can lead to it fitting the training data too closely. This results in high training accuracy, but low test accuracy. *XGBoost Documentation — xgboost 1.5.1 documentation* n.d.

Overfitting can be controlled in XGBoost in two ways: by directly controlling model complexity through parameters like max_depth, min_child_weight and gamma, and by adding randomness to make training robust to noise with parameters like subsample and colsample_bytree. *XGBoost Documentation — xgboost 1.5.1 documentation* n.d.

Model interpretability is another limitation of XGBoost models. Machine learning models are often criticized for being "black boxes", and XGBoost is no exception. Despite the model's high performance, it can be difficult to understand why it makes the predictions it does.

5.4.3 XGBoost key conclusions

Like random forest XGBoost presents itself as a good model that would fit the needs of this work, the greatest takeaway is that it can handle large amounts of data, and can be tuned with hyperparameters to better fit our classification problem.

It lacks interpretability but this is nothing new, most models have this issue and we need to make sure that any prediction is later analyzed by the responsible gaming team.

5.5 Applying XGBoost Models

As in the previous model we'll first define our hyper-parameter space, and train the models, the following models we're the top three model results after some iterations.

5.5.1 XGBoost Hyper-parameters space

```
space = {
  "colsample_bytree": 0.6457121251890898,
  "learning_rate": 0.13266476149578058,
  "max_depth": 3,
  "min_child_weight": 3,
  "n_estimators": 588,
  "n_jobs": 100,
  "subsample": 0.4033767159578759,
  "verbosity": 0,
  "random_state": 654776397,
}
```

Figure 5.9: Hyper-parameters space for XGBoost No. 1

```
space = {
  "colsample_bytree": 0.44020068543380 28,
  "learning_rate": 0.937173760021371,
  "max_depth": 11,
  "min_child_weight": 1,
  "n_estimators": 11,
  "n_jobs": 100,
  "subsample": 0.2962888949853324,
  "verbosity": 0,
  "random_state": 654776397,
}
```

Figure 5.10: Hyper-parameters space for XGBoost No. 2

```
space = {
    "colsample_bytree": 0.6131017928832345,
    "learning_rate": 1.5288290440271002,
    "max_depth": 12,
    "min_child_weight": 4,
    "n_estimators": 32,
    "n_jobs": 100,
    "subsample": 0.6812029072903498,
    "verbosity": 0,
    "random_state": 654776397,
}
```

Figure 5.11: Hyper-parameters space for XGBoost No. 3

5.5.2 Results

These are the results obtained, from each iteration:

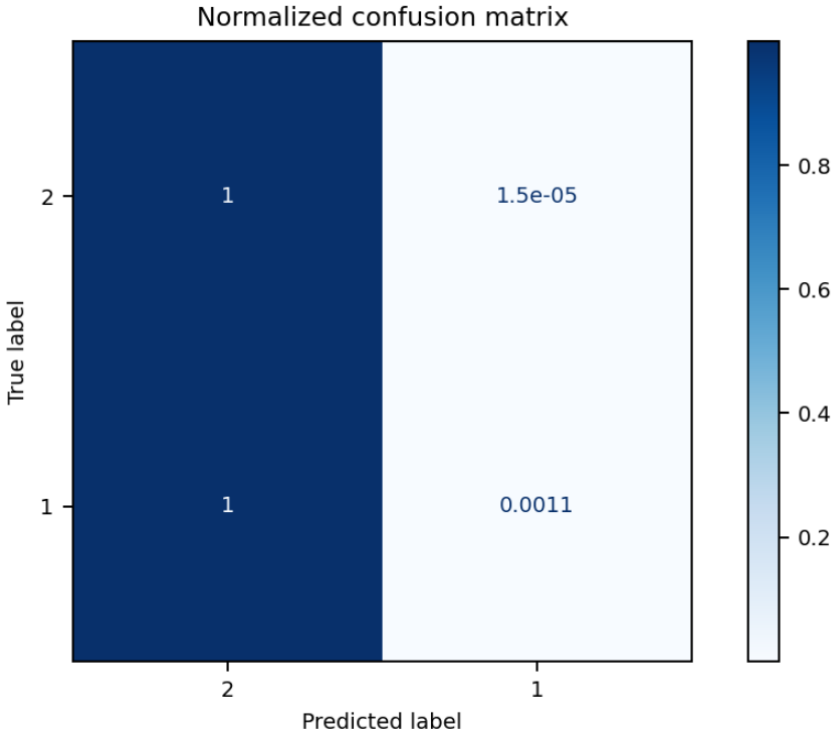

Figure 5.12: Confusion Matrix for XGBoost No. 1

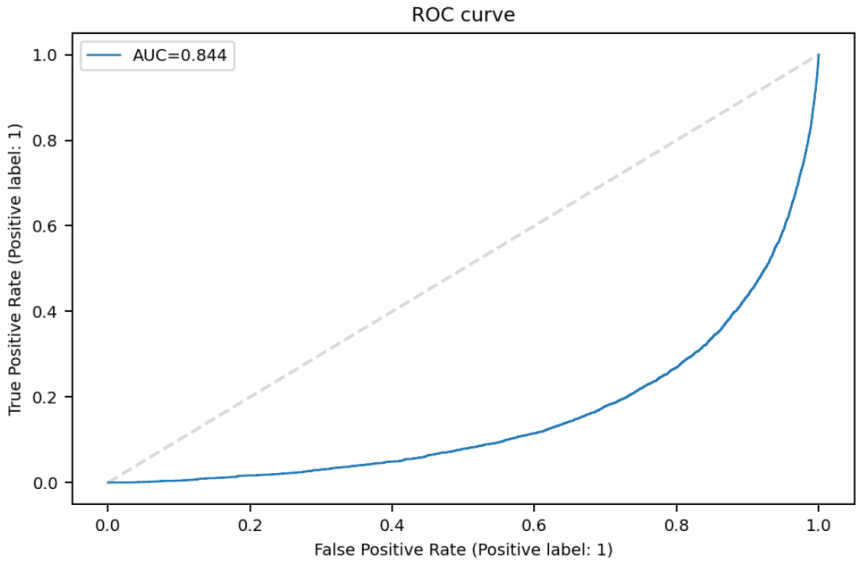

Figure 5.13: Roc Curve for XGBoost No. 1

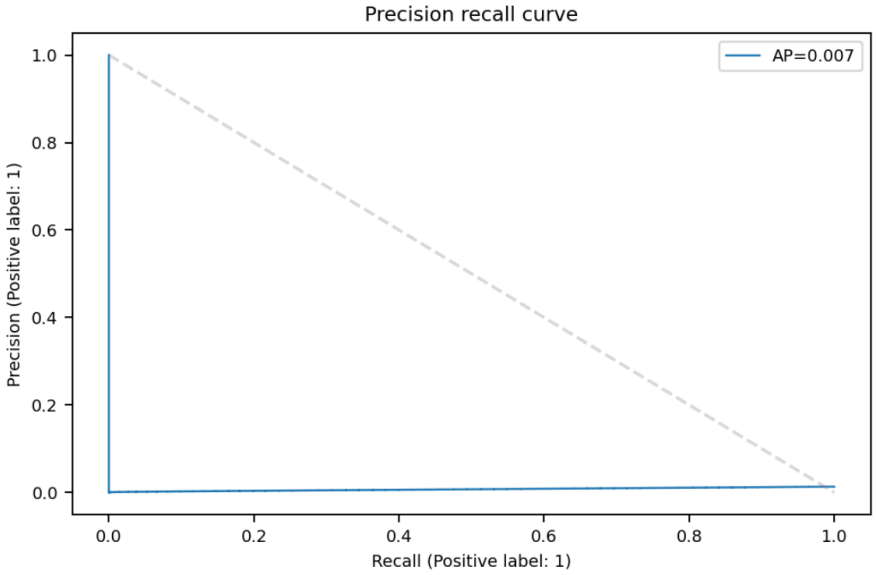

Figure 5.14: Precision Recall for XGBoost No. 1

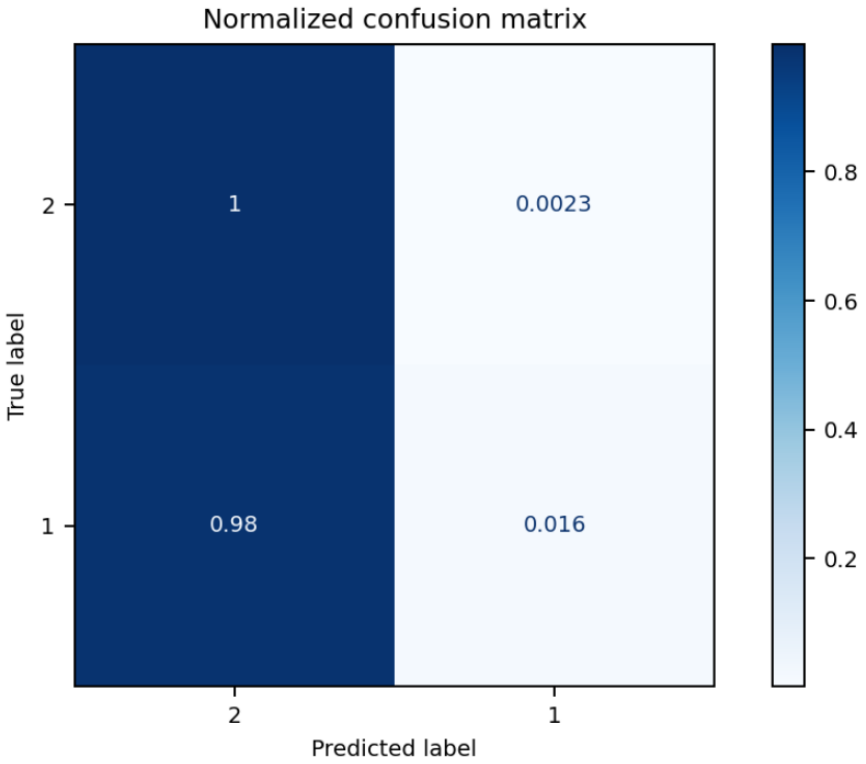

Figure 5.15: Confusion Matrix for XGBoost No. 2

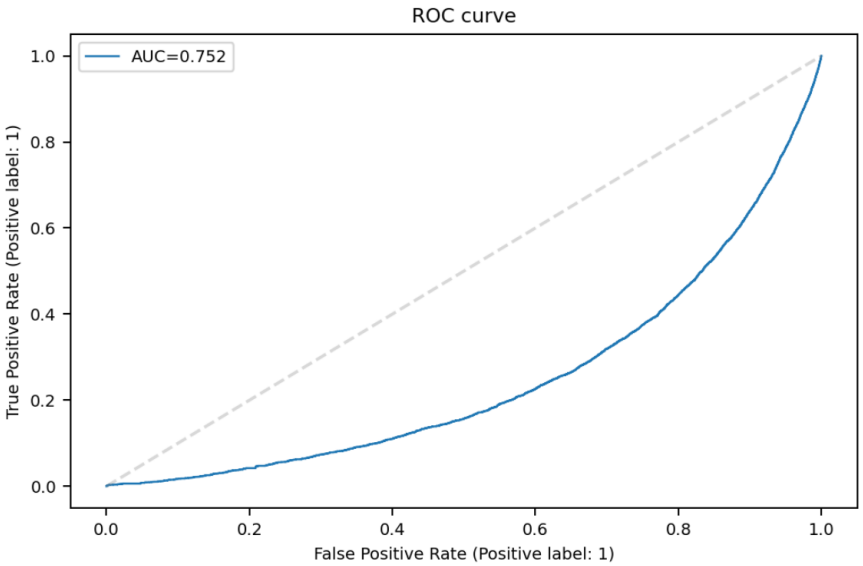

Figure 5.16: Roc Curve for XGBoost No. 2

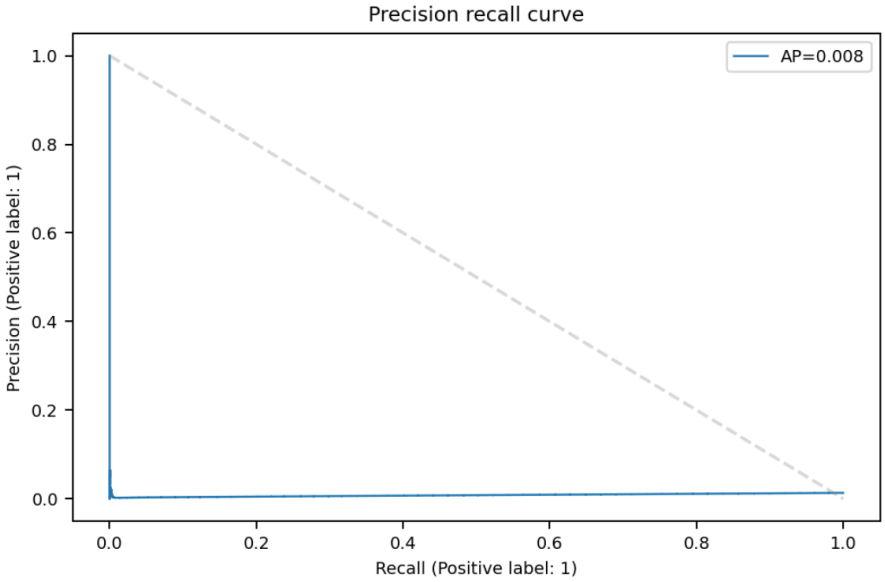

Figure 5.17: Precision Recall for XGBoost No. 2

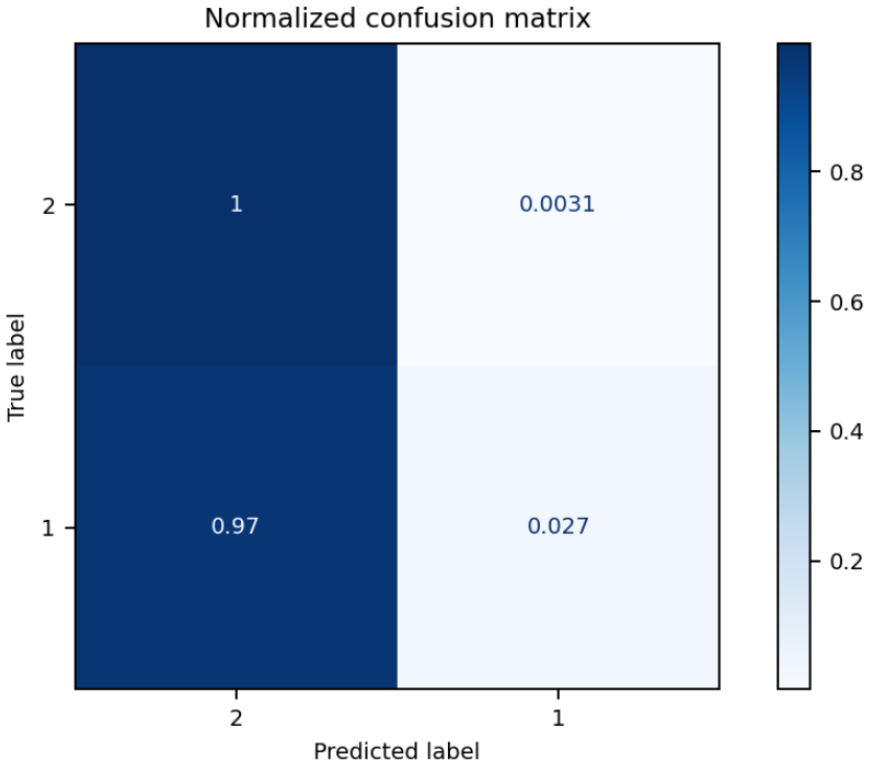

Figure 5.18: Confusion Matrix for XGBoost No. 3

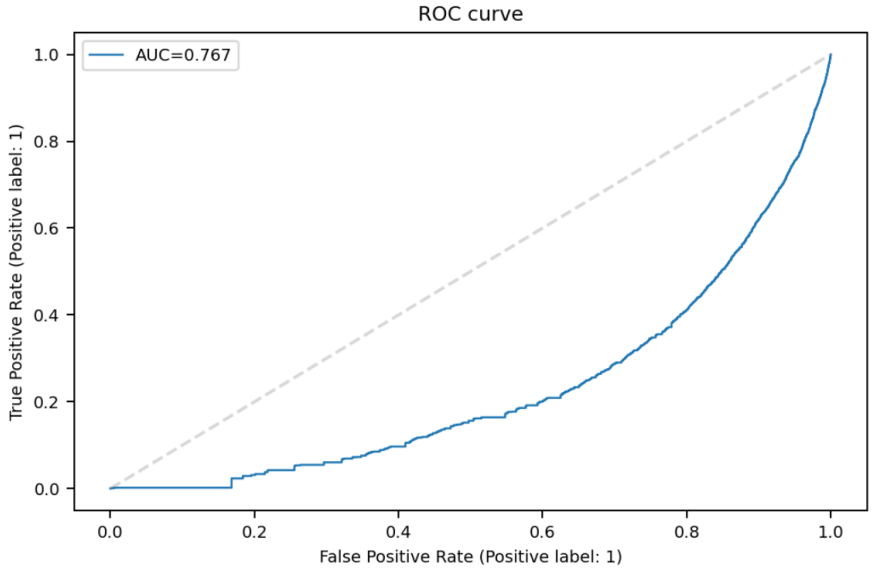

Figure 5.19: Roc Curve for XGBoost No. 3

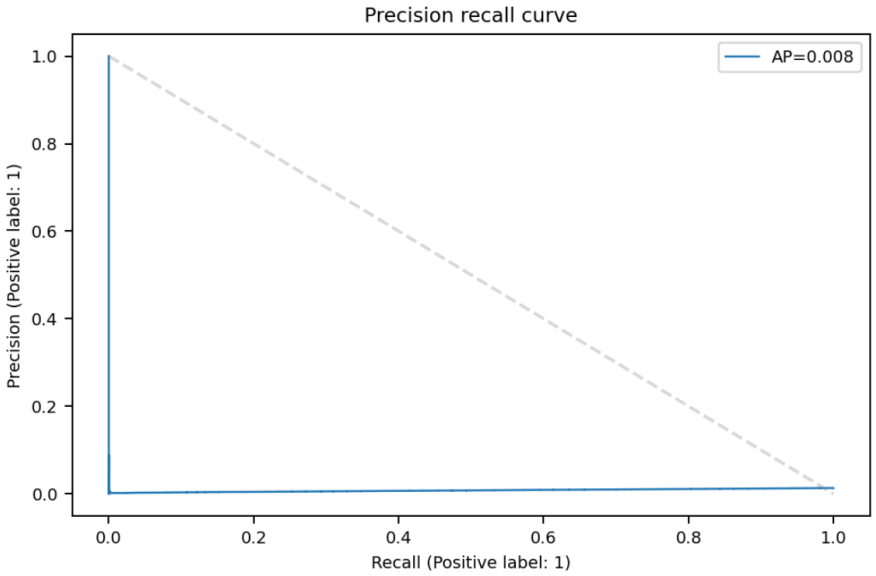

Figure 5.20: Precision Recall for XGBoost No. 3

5.5.3 XGBoost F1 scores

- Model 1 - 0.002
- Model 2 - 0.002
- Model 3 - 0.023

5.5.4 XGBoost Model Conclusions

Unfortunately, XGBoost did not perform very well to our needs, we might not have extracted all the potential from this algorithm and, like the random forest, it tried to just classify almost everything as a non-responsible gaming case which is not ideal.

This situation indicates the necessity for careful fine-tuning of the model's hyper-parameters, including a critical review of the cost function, to balance sensitivity and specificity. Additionally, it suggests that just like the random forest model the imbalance might have had a big impact, and the features might not have been tuned enough, like normalizing them. This is a great point for future work.

5.6 LightGBM Model

LightGBM is a gradient-boosting framework using tree-based learning algorithms. It's known for its high efficiency and speed, especially in handling large-scale data *LightGBM Documentation* n.d. , which we have a lot. The main question is to see how it will deal with the imbalanced data.

5.7 Applying LightGBM Models

5.7.1 LightGBM Hyper-parameters space

```
space = {
    "colsample_bytree": 0.5936896823710984,
    "lambda_l1": 1.3104439908020442,
    "lambda_l2": 0.11451907533058894,
    "learning_rate": 1.7512451306155203,
    "max_bin": 279,
    "max_depth": 12,
    "min_child_samples": 36,
    "n_estimators": 16,
    "num_leaves": 13,
    "path_smooth": 10.583627257176008,
    "subsample": 0.6380614316197261,
    "random_state": 654776397,
}
```

Figure 5.21: Hyper-parameters for LightGBM No. 1

```
space = {
  "colsample_bytree": 0.5014404359147553,
  "lambda_l1": 0.5423055438544158,
  "lambda_l2": 0.6711511606685296,
  "learning_rate": 2.317776809397732,
  "max_bin": 318,
  "max_depth": 12,
  "min_child_samples": 37,
  "n_estimators": 7,
  "num_leaves": 10,
  "path_smooth": 25.958972219902893,
  "subsample": 0.5116136713797462,
  "random_state": 654776397,
}
```

Figure 5.22: Hyper-parameters for LightGBM No. 2

```
space = {
  "colsample_bytree": 0.7997710321127646,
  "lambda_l1": 0.10909940107077358,
  "lambda_l2": 0.12127337252944859,
  "learning_rate": 4.8465409307602325,
  "max_bin": 489,
  "max_depth": 12,
  "min_child_samples": 24,
  "n_estimators": 4,
  "num_leaves": 688,
  "path_smooth": 0.6745942917983363,
  "subsample": 0.7964593777213169,
  "random_state": 654776397,
}
```

Figure 5.23: Hyper-parameters for LightGBM No. 3

```
space = {
    "colsample_bytree": 0.7532922489836188,
    "lambda_l1": 0.1073092070094222,
    "lambda_l2": 0.12460508007514762,
    "learning_rate": 4.2241755029276336,
    "max_bin": 494,
    "max_depth": 12,
    "min_child_samples": 20,
    "n_estimators": 5,
    "num_leaves": 790,
    "path_smooth": 1.8004043925921493,
    "subsample": 0.7996506739177236,
    "random_state": 654776397,
}
```

Figure 5.24: Hyper-parameters for LightGBM No. 4

5.7.2 Results

These are the results obtained, from each iteration:

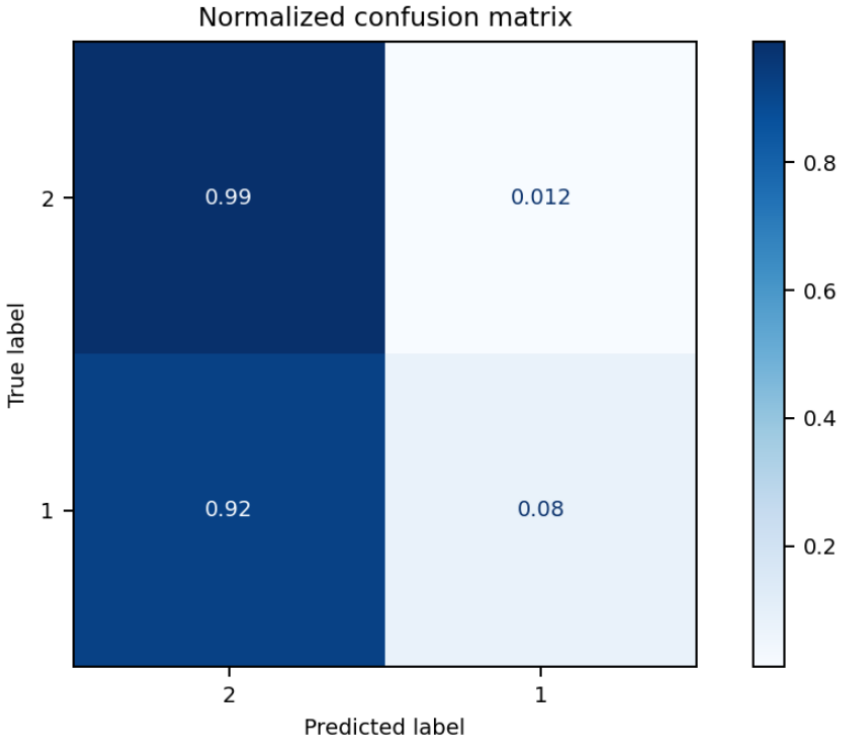

Figure 5.25: Confusion Matrix for LightGBM No. 1

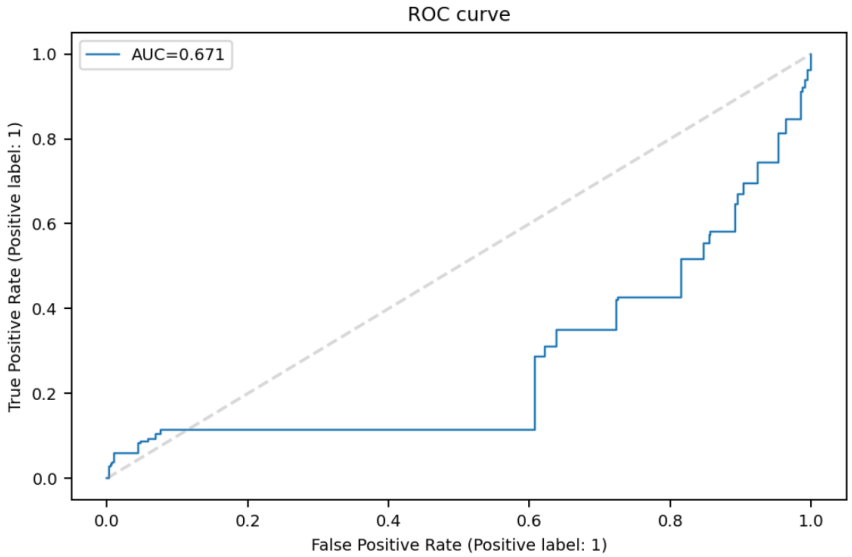

Figure 5.26: Roc Curve for LightGBM No. 1

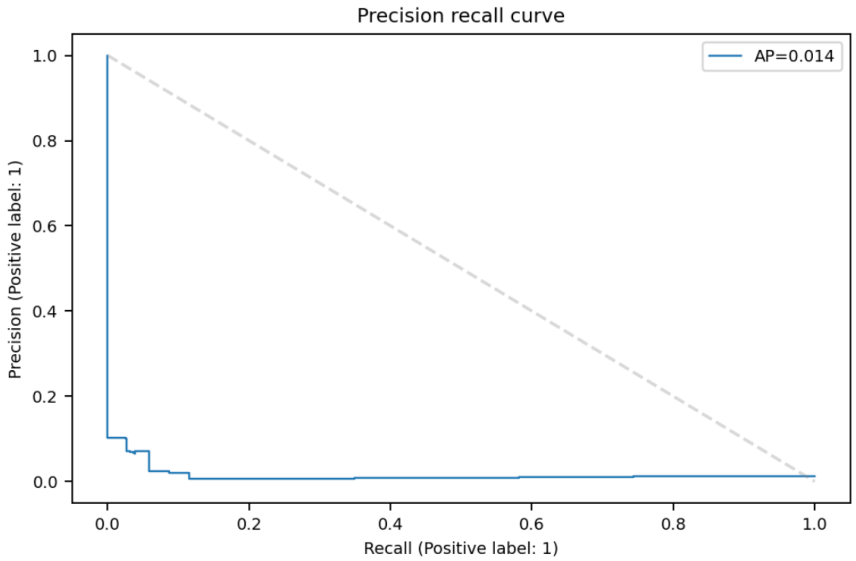

Figure 5.27: Precision Recall for LightGBM No. 1

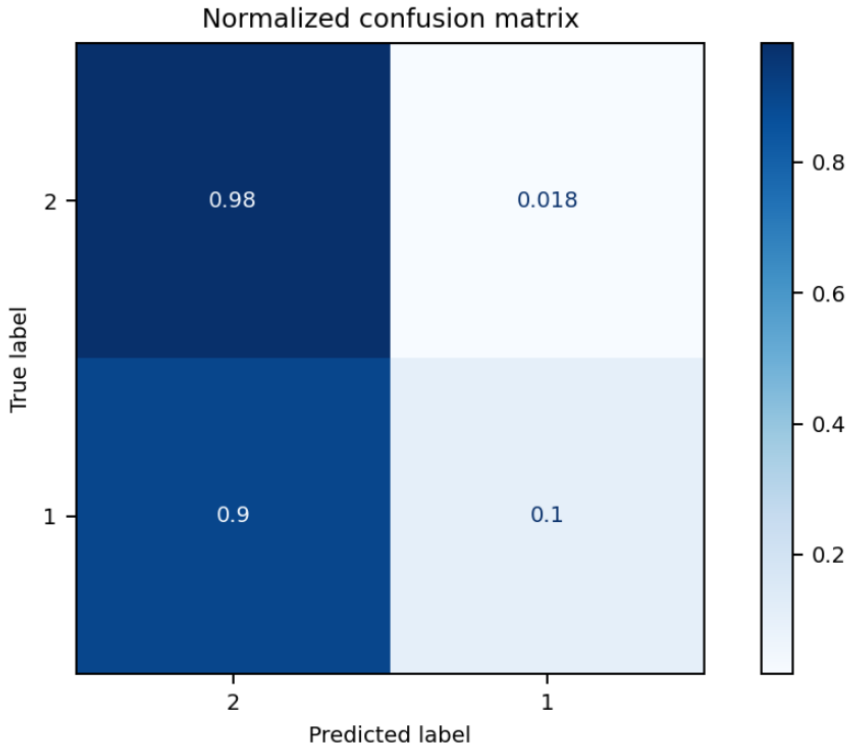

Figure 5.28: Confusion Matrix for LightGBM No. 2

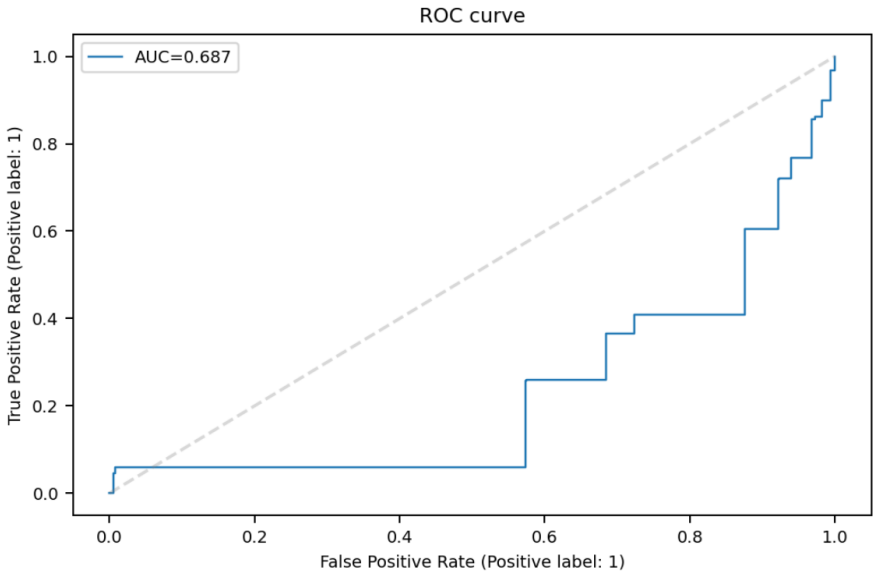

Figure 5.29: Roc Curve for LightGBM No. 2

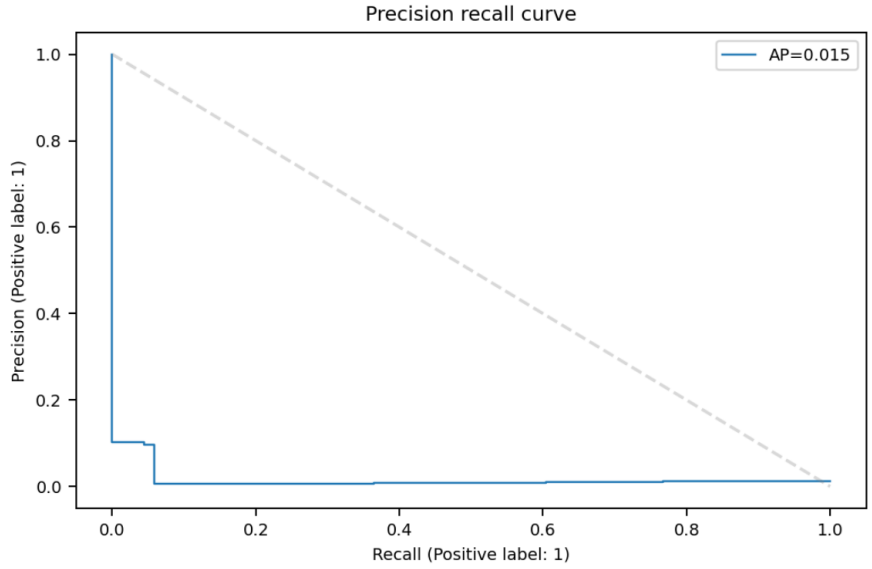

Figure 5.30: Precision Recall for LightGBM No. 2

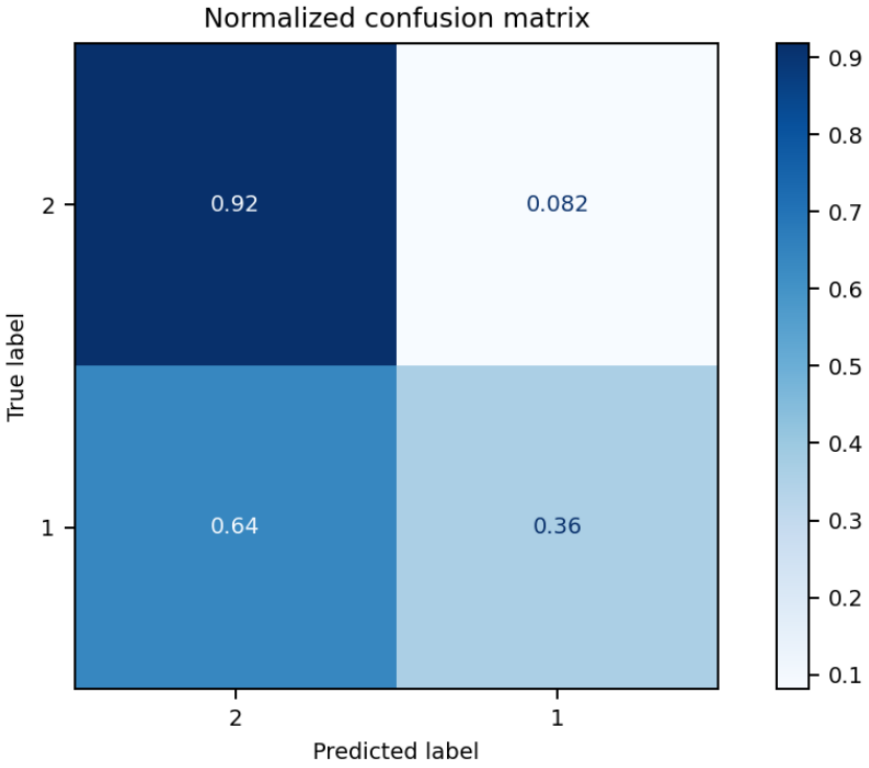

Figure 5.31: Confusion Matrix for LightGBM No. 3

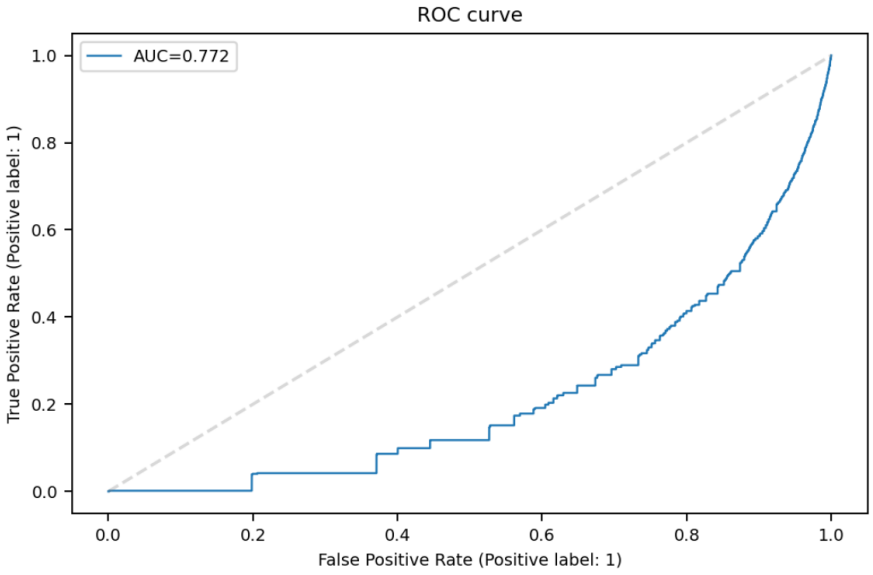

Figure 5.32: Roc Curve for LightGBM No. 3

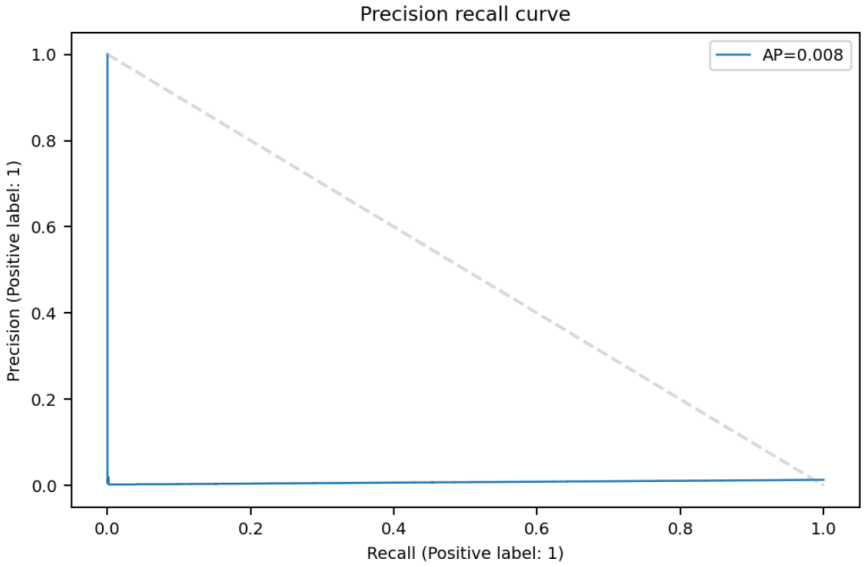

Figure 5.33: Precision Recall for LightGBM No. 3

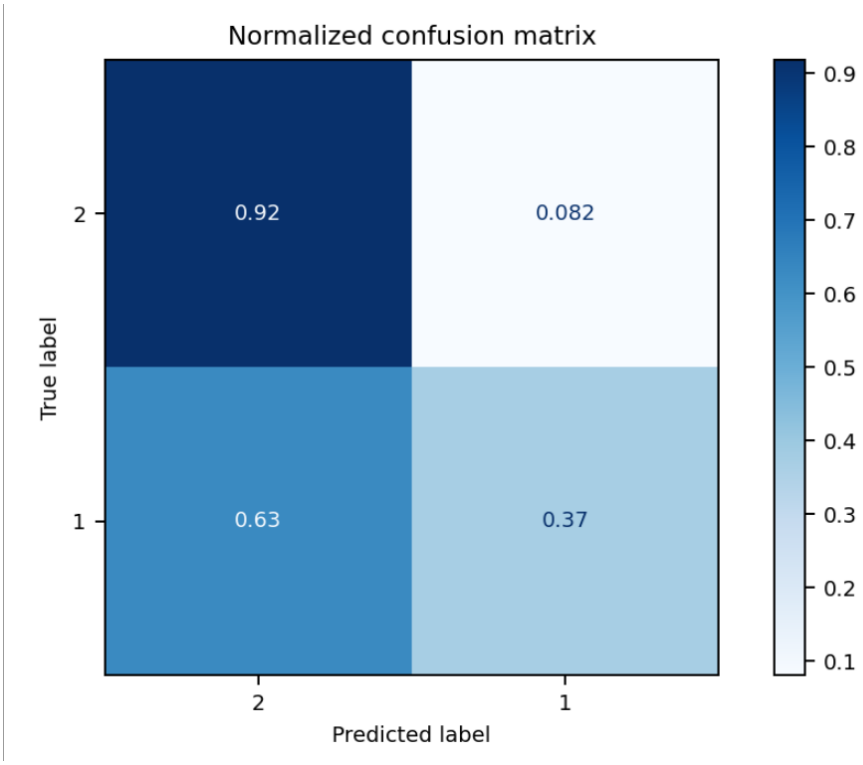

Figure 5.34: Confusion Matrix for LightGBM No. 4

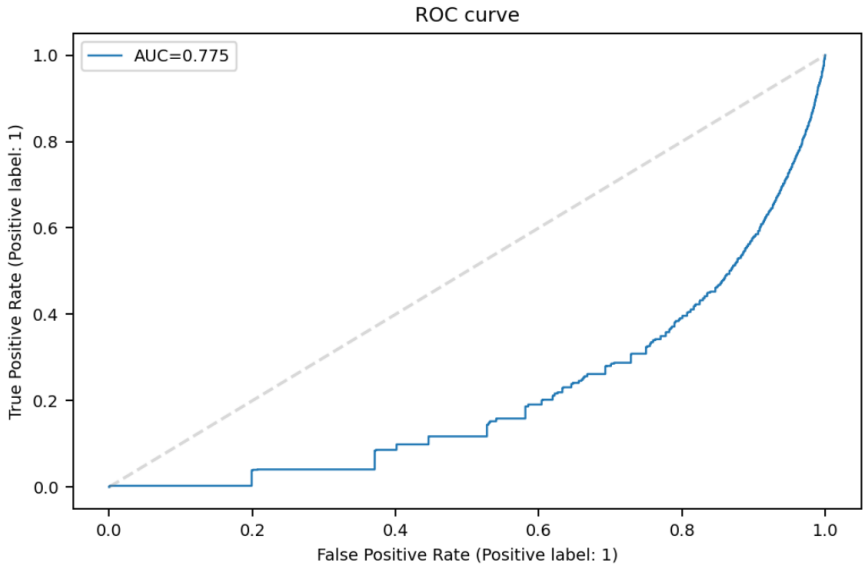

Figure 5.35: Roc Curve for LightGBM No. 4

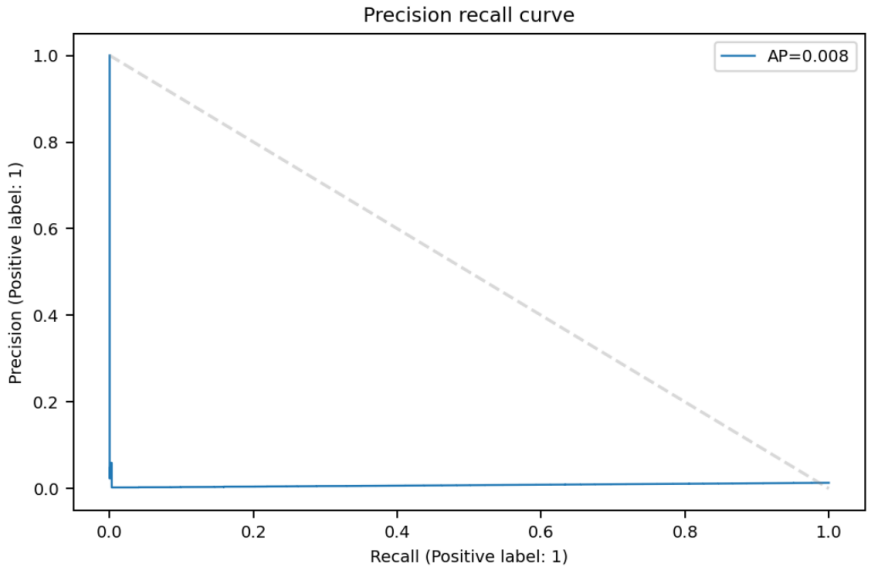

Figure 5.36: Precision Recall for LightGBM No. 4

5.7.3 LightGBM F1 scores

- Model 1 - 0.082
- Model 2 - 0.083
- Model 3 - 0.097
- Model 4 - 0.099

5.7.4 LighGBM Model Conclusions

LightGBM emerged as the most effective model overall, despite displaying certain limitations in predicting non-responsible customers. As depicted in the confusion matrix, it demonstrated a 37% accuracy rate when categorizing customers as non-responsible. Although less than ideal, still underscores the value of further investigation into these instances.

Also, this model improved a lot since random forests, and each new algorithm introduced produced better and better results.

5.8 Feature Importance

After obtaining the model's results we've run a feature importance on the latest model, that had the best results with SHAP.

"SHAP is a game-theoretic approach to explain machine learning models, providing a summary plot of the relationship between features and model output. Features are ranked in descending order of importance, and impact/color describes the correlation between the feature and the target variable." *SHAP Documentation* 2021

We can see draw some conclusions about the feature importance, starting from the top three features responsible for getting a better model were based on the last 7 days of activity, which 2 of them are from the casino, being the maximum amount spent on the casino and the total number of spins made. 5.37

The last 3 features might indicate that most players don't withdraw that often meaning that they might deposit the money and spend it all, the maximum amount of all time is actually not the best feature compared with his counterpart of the last 7 days. this might mean that people can have time that they want to put a higher stake but it doesn't mean they do it often or with the intent of being non-responsible.

Overall it is important to understand what features play the best role in making a good prediction so that for future work we can analyze better what features work and that should have our attention from the start.

Figure 5.37: SHAP graph

6. Conclusion

In this chapter will discuss the conclusions of this project and we'll end up with proposals for future work that we should take into consideration.

6.1 Main conclusion

We've managed to complete our main objectives with this work, we've explored a real world dataset, we've engaged in data analyses to try to understand the data at our hands, we've communicated with other teams to have their feedback while facing many challenges along the way.

One of this harder challenges was undertaking a ML project within a corporate environment which is notably complex and demanding. Unlike the neatly prepared datasets typically used in a more academic level, real-world data presents many challenges that complicate the ML process. This can range from missing or inconsistent data, the need for complex pre-processing, to the scale of the data that necessitates significant computational resources.

The amount of volume of the dataset was another big problem in this project, coupled with its nature, made feature computation made it the more challenging task.

Handling real-world data it's not an easy task and this project tried to be the best version of all the things that can go wrong will go wrong, but we have some results that proved that if we continued with more research we could have ended with a better, fine tuned model that would be a great help for the RG team.

6.2 Future work

This project has come a long way and it has still a lot to offer, as we see there's a lot of potential to get the model's predictions higher with more data analyses better data stratification, and better feature engineering. Here's what we can do in the future to have better results:

- Do data stratification on relevant features: We didn't split the data into groups for training, this could be done by choosing the most optimal features to stratify.
- Have more time-related features: During the investigation, some features were created in regard to how much time a user spends on the platform, unfortunately, due to constraints these features could not be computed in time. An example of this would be the session time a user spends in the casino, where we had short sessions < 2 hours of play time, and long sessions, $>= 2$ hours of play time.
- Having more loss chasing features: Many papers talked about how much a player tries to chase the loss money in a single session, this could be a major factor but we didn't manage to get this feature in place.

- Using different models: In the beginning we've defined that this project scope would only try to present classification algorithms, this is because there's was state of the art researchers that found this approach successful and it was an obvious choice giving the situation, that said, some recommendations for this case would be to try a LSTM model.

Bibliography

Braverman, Julia and Howard J. Shaffer (Jan. 2010). "How do gamblers start gambling: identifying behavioural markers for high-risk internet gambling". In: *European Journal of Public Health* 22.2, pp. 273–278. issn: 1101-1262. doi: 10.1093/eurpub/ckp232. eprint: https://academic.oup.com/eurpub/article-pdf/22/2/273/1366659/ckp232.pdf. url: https://doi.org/10.1093/eurpub/ckp232.

Brownlee, Jason (Jan. 2020). *SMOTE for Imbalanced Classification with Python*. url: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.

Dash, Sunil Kumar (May 2022). *Handling Missing Values with Random Forest*. url: https://www.analyticsvidhya.com/blog/2022/05/handling-missing-values-with-random-forest/.

Davis, Jesse and Mark Goadrich (2006). "The relationship between Precision-Recall and ROC curves". In: *Proceedings of the 23rd international conference on Machine learning - ICML '06*. doi: https://doi.org/10.1145/1143844.1143874. url: http://pages.cs.wisc.edu/~jdavis/davisgoadrichcamera2.pdf.

Deng, Xiaolei, Tilman Lesch, and Luke Clark (2019). "Applying data science to behavioral analysis of online gambling". In: *Current Addiction Reports* 6.3, pp. 159–164. doi: 10.1007/s40429-019-00269-9.

Drosatos, George et al. (May 2018). *Empowering responsible online gambling by real-time persuasive information systems*. doi: https://doi.org/10.1109/RCIS.2018.8406671. url: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8406671.

Fawcett, Tom (June 2006). "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27.8, pp. 861–874. doi: https://doi.org/10.1016/j.patrec.2005.10.010.

Fleiss, Alex (Feb. 2023). *What are the advantages and disadvantages of random forest?* url: https://www.rebellionresearch.com/what-are-the-advantages-and-disadvantages-of-random-forest.

Hachcham, Aymane (2023). *XGBoost: Everything You Need to Know*. url: https://neptune.ai/blog/xgboost-everything-you-need-to-know.

Hopfgartner, Niklas et al. (2022). "Predicting self-exclusion among online gamblers: An empirical real-world study". In: *Journal of Gambling Studies* 39.1, pp. 447–465. doi: 10.1007/s10899-022-10149-z.

Huyen, Chip (2022). *Designing Machine Learning Systems [Book]*. url: https://www.oreilly.com/library/view/designing-machine-learning/9781098107956/.

Korstanje, Joos (Aug. 2021). *The F1 score*. url: https://towardsdatascience.com/the-f1-score-bec2bbc38aa6.

LightGBM Documentation (n.d.). url: https://lightgbm.readthedocs.io/en/stable/.

Meena, Mansi (Dec. 2020). *Applications of Random Forest*. url: https://iq.opengenus.org/applications-of-random-forest/.

Metz, Charles E. (Oct. 1978). "Basic principles of ROC analysis". In: *Seminars in Nuclear Medicine* 8.4, pp. 283–298. doi: https://doi.org/10.1016/s0001-2998(78)80014-2.

MLFlow Documentation (n.d.). url: https://mlflow.org/docs/latest/what-is-mlflow.html#:~:text=MLflow%5C%20is%5C%20used%5C%20to%5C%20manage.

Rose, Alex (n.d.). *ALRO5921/problem-gambling-study*. url: https://github.com/alro5921/problem-gambling-study.

Seitz, Sarem (June 2022). *Why SMOTE is not necessarily the answer to your imbalanced dataset*. url: https://towardsdatascience.com/why-smote-is-not-necessarily-the-answer-to-your-imbalanced-dataset-ef19881da57a.

SHAP Documentation (2021). url: https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%5C%20introduction%5C%20to%5C%20explainable%5C%20AI%5C%20with%5C%20Shapley%5C%20values.html.

Shen, Chu et al. (July 2018). *Imbalanced Data Classification Based on Extreme Learning Machine Autoencoder*. doi: https://doi.org/10.1109/ICMLC.2018.8526934. url: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8526934.

Siu, Ka-Meng, Lap-Man Hoi, and Ka-Hou Chan (Nov. 2021). *A Proposed Set of Features on Implementing Responsible Gambling on Slot Games with G2S Technology*. doi: https://doi.org/10.1109/IC2SE52832.2021.9792030. url: https://ieeexplore.ieee.org/document/9792030.

Tharwat, Alaa (Aug. 2018). "Classification assessment methods". In: *Applied Computing and Informatics*. doi: https://doi.org/10.1016/j.aci.2018.08.003.

The transparency project (n.d.). url: http://www.thetransparencyproject.org/.

XGBoost Documentation — xgboost 1.5.1 documentation (n.d.). url: https://xgboost.readthedocs.io/en/stable/.