International Journal of Mental Health and Addiction
https://doi.org/10.1007/s11469-025-01602-2

ORIGINAL ARTICLE

Using Machine-Learning Algorithms to Predict Self-Reported Problem Gambling Among a Sample of Online Gamblers

Michael Auer[1] · **Mark D. Griffiths**[2]

Received: 19 June 2025 / Accepted: 19 November 2025

Abstract

Studies suggest that algorithms can effectively be used to predict self-reported problem gambling using player tracking data. The present study analyzed a sample of real-world online gamblers ($N=1{,}611$) who engaged in lottery playing, casino gambling, bingo playing, and sports betting. The data also comprised each player's actual gambling activity, as well as age and gender, in the 30 days prior to answering the Problem Gambling Severity Index (PGSI). Players who engaged in at least one lottery game 30 days prior to answering the PGSI were less likely to be problem gamblers compared to players who did not play lottery games. For all other game-categories the relationship was reversed. The results also indicated that specific behavioral tracking features—such as the average number of monetary deposits per session, total amount of money bet per day, session length, and casino gambling involvement—were among the most significant predictors of self-reported problem gambling. When evaluating different machine algorithms, logistic regression and random forest emerged as the most effective in predicting self-reported problem gambling. The present study is among the few which predicts self-reported problem gambling using a sample of online lottery players, casino gamblers, bingo players and sports bettors, and provides further empirical evidence supporting the use of machine learning models to identify self-reported problem gamblers based on player tracking data. These findings can inform responsible gambling strategies by enabling operators to identify and intervene before gambling-related problems escalate.

Keywords Online gambling · Behavioral tracking · Problem gambling · Machine learning · Algorithms

The proliferation of internet technologies has significantly transformed gambling, and raised concerns about the prevalence of gambling behaviors and the associated risks of problem gambling. In Europe, the online gambling sector has experienced substantial

✉ Mark D. Griffiths
mark.griffiths@ntu.ac.uk

Michael Auer
m.auer@neccton.com

[1] Neccton GmbH, Davidgasse 5, 7052 Müllendorf, Austria

[2] Psychology Department, Nottingham Trent University, 50 Shakespeare Street, Nottingham NG1 4FQ, UK

growth. According to the European Gaming and Betting Association ([EGBA] 2019), the gross gaming revenue (GGR) from online gambling was valued at €38.2 billion in 2022, accounting for 35% of the global online gambling market. The United Kingdom represented the largest proportion of this market, followed by Italy. Germany, France, and Spain. Sweden was the European country with the highest share (80%) of its gambling activity taking place online, followed by Latvia (75%), Lithuania (67%), Romania (65%) and the UK (65%). This upward trend is expected to continue, with projections indicating that the GGR will reach €54.3 billion by 2027 (EGBA, 2019).

The prevalence of online gambling in Europe, Canada, and the United States has seen a significant increase, driven by technological advancements and regulatory changes (e.g., American Gaming Association, 2025; EGBA 2022). While this growth has contributed to economic benefits, it has also raised concerns regarding problem gambling and associated harms (Connor, 2024; Håkansson, 2020; Wardle and Reith, 2021). Continuous monitoring, research, and the implementation of effective regulatory frameworks are essential to mitigate the negative impacts of online gambling and protect vulnerable populations.

European countries have not consistently assessed online gambling participation and online problem gambling (Carran, 2022). Although gambling disorder is included in both the eleventh revision of the *International Classification of Diseases* (ICD-11; World Health Organization, 2019) and the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5; American Psychiatric Association, 2013), European countries use different problem gambling screens, and national surveys are not common in each country. Carran (2022) also reported that gambling engagement ranged between 32.9% (Czech Republic, including lotteries) and 80% (Finland, including lotteries), and problem gambling ranged between 0.3% (Ireland in 2019) and 6.4% (Latvia in 2019).

Tran et al. (2024) conducted a systematic review and meta-analysis of gambling participation and problem gambling between January 2010 and March 2024. A total of 166 studies reported any gambling activity in the past 12 months with the majority of studies coming from western Europe, followed by North America and Australasia. Globally, 46.2% of adults were estimated to have engaged in a gambling activity in the past 12 months. This translates to 2.3 billion adults globally. Australasia had the highest estimated prevalence at 70% (63.5%), with similar levels in North America (61.3%). Gambling participation in Western Europe was 49% and 41.3% in Eastern Europe. Tran et al. (2024) also reported that globally, adult males (49.1%) had higher rates of gambling participation than adult females (37.4%).

Moreover, Tran et al. (2024) reported that among adults, the prevalence of problem gambling globally was 1.41%. This percentage is within the confidence interval of Gabellini et al.'s (2023) previous population estimate of 1.29%. The prevalence of problem gambling was higher among males (2.2%) than females (1%). Tran et al. (2024) reported that among those who had gambled in the previous 12 months, problem gambling for males was estimated to be 2.8% and problem gambling among females was estimated to be 1.2%

Factors Contributing to Online Gambling Prevalence

Costes et al. (2023) compared gambling practices among a sample of 24,412 adults from France, Italy, Germany, Switzerland, and Poland. They found that individuals with gambling-related problems were more likely to participate in online gambling activities other than lotteries. In all countries, individuals who reported high frequency of gambling, high spending, and participation in multiple gambling activities were more likely to experience problem gambling.

Mora-Salgueiro et al. (2021) conducted a systematic review examining the clinical and sociodemographic risk factors of online problem gambling. The review comprised 20 published studies between 2006 and 2019 and reported that that disordered online gambling among adults ranged from 2.7% to 20.3%. The two most common sociodemographic comorbidities of disordered online gambling were being single and being male, and that it was most likely to occur among those aged between 30 and 40 years.

In a more recent meta-analysis, Tran et al. (2024) reported the highest risk of problem gambling was for individuals who played online casino or slots (15.8%). Electronic gaming machines (EGMs), sports betting, any online gambling, and financial market gambling had a similar prevalence of problem gambling among individuals using those activities to gamble, ranging from 8.1% to 8.9%. Those engaged in buying lottery or raffle tickets (2%) and instant lottery/instant win games (2.6%) had the lowest prevalence of problem gambling.

In their meta-analysis regarding risk factors for problem gambling (based on 33 studies that met their inclusion criteria), Moreira et al. (2023) reported that gamblers who played more than one game and had longer gambling sessions were at greater risk of problem gambling. This is in line with Allami et al.'s (2021) meta-analysis of 104 studies. Allami et al. (2021) reported that the highest risk for problem gambling was associated with online gambling and continuous forms of gambling in general (e.g., EGMs).

Several factors have contributed to the increasing prevalence of online gambling. The convenience and accessibility of online platforms allow individuals to gamble at any time and place, removing traditional barriers associated with land-based venues (Gainsbury, 2015). The anonymity provided by online gambling can also reduce the social stigma associated with gambling behaviors, potentially encouraging higher participation rates (Gainsbury et al., 2015; Wardle & Griffiths, 2011).

Moreover, aggressive marketing strategies and promotions by online gambling operators have been identified as significant drivers of increased gambling participation (Hing et al., 2014). A report by the European Gaming and Betting Association (2019) highlighted that the regulatory trajectory in European countries has favored the legitimization of gambling industries, leading to increased exposure and normalization of gambling activities.

Regulatory Responses to Online Problem Gambling

The rapid expansion of online gambling has prompted regulatory bodies to implement measures aimed at mitigating associated harms. In Europe, regulatory approaches vary across countries, with some adopting stringent measures to control online gambling activities, while others have more liberal frameworks (Ukhova et al., 2024). The European Union has been moving toward the legalization and liberalization of gambling markets, necessitating continuous monitoring to assess the impact of these policies on gambling behaviors. In Canada, regulatory frameworks for online gambling are primarily managed at the provincial level, leading to variations in policies and enforcement across the country. The increasing prevalence of online gambling among adolescents has raised concerns, prompting calls for enhanced regulatory measures and public health interventions to address potential harms (Elton-Marshall et al., 2016).

The Application of Player Tracking in the Study of Online Gambling

A growing number of academic studies have utilized online gambling player tracking data to explore various responsible gambling measures, including limit setting (both time and money), self-exclusion, mandatory breaks/cool-off periods, personalized feedback, and nudging (e.g., Auer & Griffiths, 2013; Auer et al., 2014; Braverman & Shaffer, 2012; Håkansson & Henzel, 2021).

A number of studies have evaluated voluntary limits setting in online gambling settings (e.g., Auer & Griffiths, 2013; Auer et al., 2018, 2020a, 2020b). Attitudes towards limit setting are generally positive (Auer et al., 2020a, 2020b). Moreover, gambling expenditure is reduced after setting limits (Auer & Griffiths, 2013), players spend less when they are reminded of reaching their personal limits (Auer et al., 2018). and players who voluntarily set spending limits are more loyal to the operator (Auer et al., 2021). Personalized messages that inform players of increased time or money spent, chasing losses, gambling during the night time, and high losses have been shown to significantly reduce the amount of money spent (Auer & Griffiths, 2013).

Håkansson and Henzel (2021) analyzed the characteristics of online gamblers who opted for voluntary self-exclusion. They found that self-excluders often displayed higher levels of gambling intensity and were more likely to report gambling-related problems. This suggested that voluntary self-exclusion programs are utilized by individuals experiencing significant gambling issues, highlighting the importance of such programs in responsible gambling strategies. However, Catania and Griffiths (2021) analyzed a sample of 7,732 online gamblers who had voluntarily self-excluded. They found that almost one-fifth of the customers that used six-month VSE only had gambling activity for less than 24 h (19.15%). Moreover, half of the customers had less than seven days of account registration before using the six-month VSE facility (50.39%). Catania and Griffiths concluded that customers who use VSE were too different to be treated as a homogenous group and that VSE was not a reliable proxy measure for problem gambling.

Due to increased technical possibilities, online gambling companies can now provide players with personalized feedback about their gambling behavior. Studies have shown that players tend to underestimate the amount of money spent (e.g., Auer et al., 2024) and it has been found that personalized feedback can lead to an improved self-awareness and lower monetary spending (Auer & Griffiths, 2018, Auer and Griffiths, 2015). Personalized messages informing players about increased money spent gambling, increased time spent gambling, high losses, gambling during the night-time, and chasing losses have led to significantly reduced money and time spent gambling (Auer & Griffiths, 2020).

Studies have reported that problem gamblers tend to regularly deplete their accounts (i.e., spending all of their funds by the end of a gambling session) (Auer & Griffiths, 2023a; Hopfgartner et al., 2024). This also means that they rarely withdraw any winnings. One way to promote responsible play could be to nudge players to withdraw money after they have won large amounts of money. Auer and Griffiths (2023b) studied the impact of personalized nudges on the withdrawal of gambling funds. After players won significant amounts of money, they received messages which attempted to nudge them to withdraw some of the winnings. Compared to matched controls, approximately 38% of gamblers withdrew money from their gambling account on the same day they read the message.

Predicting Self-Reported Problem Gambling

Hopfgartner et al. (2024) have emphasized the importance of using standardized problem gambling screens when studying the association between gambling behavior and problem gambling. Only a few studies have had access to self-reported problem gambling among samples of real-world online gamblers (e.g., Auer & Griffiths, 2023a; Hopfgartner et al., 2024; Louderback et al., 2021; Luquiens et al., 2016; Murch et al., 2023; Perrot et al., 2022).

Auer and Griffiths (2023a) conducted a study analyzing data from 1,287 European online casino players who completed the Problem Gambling Severity Index (Ferris & Wynne, 2001). The random forest model demonstrated superior predictive performance. Key behavioral indicators of problem gambling included higher losses per gambling day and session, more frequent deposits within sessions, and a greater tendency to deplete their monetary accounts.

Hopfgartner et al. (2024) expanded upon previous research by incorporating a cross-country approach to predict self-reported problem gambling using player-tracking data from 1,743 online casino gamblers in the UK, Canada, and Spain. The study tested five different machine learning models and found that behavioral variables, such as self-exclusions, frequent in-session deposits, and account depletion, were as predictive of problem gambling as monetary intensity variables. The models performed well across different countries, indicating the potential for generalized application in diverse cultural contexts.

Luquiens et al. (2016) correlated PGSI scores with player tracking data among a sample of 14,261 online poker players, and reported that 18% of their sample scored 5 or more on the PGSI (although scores of 8 or more indicate problem gambling). They reported that being male, having a mean loss of at least €170 per session, having a loss of at least €45 in 30 days, and engaging in at least 60 gambling sessions in 30 days were the main risk factors for scoring 5 or more on the PGSI.

Based on data from 1772 online casino players who completed the Brief Biosocial Gambling Screen (BBGS; Gebauer et al., 2010), Louderback et al. (2021) established criteria for low-risk gambling. The BBGS features three yes-or-no items, where a 'yes' response to any question signals a potential gambling problem. Their study found that a heightened risk of self-reported problem gambling was associated with wagering at least €167.97 per month, spending more than 6.71% of annual income on online bets, incurring losses of €26.11 or more each month, and having a daily wager variability (standard deviation) of at least €35.14 over a month.

Murch et al. (2023) examined PGSI responses from 9,145 adults (18 years and older) who gambled at lotoquebec.com—a platform offering lottery, casino, and sports betting services. They analyzed player-tracking data that detailed both time and money spent, as well as the use of responsible gambling tools, over the 12 months preceding the PGSI assessment (with scores of 5 or above indicating at-risk gambling and scores of 8 or above suggesting problem gambling). Their final predictive model incorporated 10 variables, including factors such as younger age and the pattern of making repeated deposits on a weekly basis after placing a bet. Other key predictors related to the amounts of money wagered and withdrawn. Notably, the study did not include any variables based on individual gambling sessions. The overall model achieved an area under the curve (AUC) of 84%, a strong performance that may partly reflect the inclusion of many inactive players in the sample—players only needed to place a single bet during the 12 months before completing the PGSI. Consequently, individuals with higher PGSI scores likely had recent gambling activity, while those with lower scores probably had minimal or no gambling activity in the weeks before the assessment. This pre-assessment activity likely contributed significantly to the model's high accuracy.

Perrot et al. (2022) analyzed two random samples of French online gamblers in skill-based games (i.e., poker, horse race betting, and sports betting, $n = 8{,}172$) and pure chance games (i.e., scratch-card games and lotteries, $n = 5{,}404$). They found that 7.4% of the skill-based players reported a PGSI score of 8 +, and 0.8% of the chance-based players reported a PGSI score of 8 +. Players were classified into four groups based on the PGSI score (i.e., non-problem gambling = 0, low-risk = 1–4, moderate-risk = 5–7, high-risk = 8 +). The predictive performances were good for the model for skill-based games (area under the receiving operating characteristic curves [AUROCs] from 0.72 to 0.82), but moderate for the model for pure chance games (AUROCs from 0.63 to 0.76), with wide confidence intervals due to the lower frequency of problem gambling in this sample. In the skill-based dataset, 84% of non-problem gambling, 24% of low-risk gambling, 12% of moderate-risk gambling, and 85% of problem gambling cases were correctly identified. The model in the chance-based dataset correctly identified 73% of non-problem gamblers, 27% of low-risk gamblers, 0% of moderate-risk gambling, and 67% of problem gambling cases. When predicting the four PGSI categories altogether, performances were good for identifying extreme categories (non-problem and problem gamblers) but poorer for intermediate categories (low-risk and moderate-risk gamblers), irrespective of game type.

The Present Study

The aforementioned studies suggest that artificial intelligence and machine learning algorithms can effectively be used to predict self-reported problem gambling using player tracking data. However, there are few published studies and additional evidence is needed as to which gambling behavior (e.g., amount of money bet, amount of money lost, gambling duration, depositing behavior, etc.) is most indicative of problem gambling. The present study aimed to contribute to the body of evidence by utilizing a new secondary dataset from a sample of real-world online gamblers. It is also one of the few studies which includes gamblers who engaged in gambling on lottery games, casino games, and bingo, as well as those engaging in sports betting. To the best of the authors' knowledge only Murch et al. (2023) have analyzed a similar sample of online gamblers. Apart from exploratory analysis, the present study aimed to answer the following research questions (RQs):

- Are problem gambling rates different between players who prefer lottery, casino, sports-betting or bingo games? (RQ1)
- Are behavioral variables (e.g., time spent, depositing behavior, demographics) sufficient compared to monetary variables (e.g., amount deposited, amount lost, amount bet) in explaining self-reported problem gambling? (RQ2)
- Which algorithms best predict self-reported problem gambling? (RQ3)

Numerous jurisdictions (e.g., UK, Germany, Spain, Sweden, Denmark, Ontario) require online gambling operators to assess problem gambling-related risk. In many jurisdictions, a high-risk assessment is connected to concrete actions such as the exclusion from marketing, bonuses or specific gambling products. For that reason, problem gambling risk assessments are of high importance to gambling businesses.

Method

The authors had access to a secondary dataset from a North American online gambling website which offers lottery games, casino games, and bingo games, as well as sports betting. Each of the participants answered the PGSI between April 2023 and February 2025 (Appendix I). The data also comprised each player's actual gambling activity (such as wins/losses, amounts of money deposited, amounts of money withdrawn after winning), as well as age and gender, in the 30 days prior to answering the PGSI.

Only players who wagered at least once during the 30 days prior to answering the PGSI or on the day of answering the PGSI were available to the authors. Using 30 days of behavior prior to answering the PGSI is in line with previous studies (e.g., Auer & Griffiths, 2023a; Hopfgartner et al., 2024). The authors hypothesized that if there was a correlation between self-reported problem gambling and prior gambling behavior, it should emerge in the most recent activity. Also, if a longer time-period was used (e.g., 12 months), some players might have been active for several months before answering the PGSI. Online gambling operators would see no reason to interact with a player who was inactive for an extended period of time. For that reason, the authors viewed players' most recent gambling activity as being more relevant for the prediction of problem gambling. The authors also computed the session length in line with previous studies (e.g., Auer & Griffiths, 2023a; Hopfgartner et al., 2024). More specifically, if two wagers were placed within 15 min of each other, these bets were operationalized as being in the same session. If there were more than 15 min between two wagers, the second wager was operationalized as the start of a new session.

Players were given four response options for each of the nine PGSI items: 'Never' (0), 'Sometimes' (1), 'Most of the time' (2), and 'Almost always' (3), resulting in scores ranging from 0 to 27. A score of 8 or greater indicates problem gambling. Additionally, the researchers had access to the time interval, measured in seconds, between a player's initial click on the PGSI site and the final submission after completing all nine PGSI questions. Appendix 2 provides an overview of the player tracking metrics that were calculated for each participant based on their gambling behavior in the 30 days leading up to their PGSI assessment. These metrics include the total number of monetary deposits and bets made during this period, as well as the mean average amount of money wagered per gambling day and per session. In line with Auer and Griffiths (2023c), the amount of money deposited per session is a key metric with respect to chasing losses. For each player, the authors computed the average number of deposits per session based on the number of deposits and sessions in the 30 days prior to answering the PGSI. The game-type specific bet ratios were computed based on the game-type specific amount of money bet, and total amount of money bet over the past 30 days before answering the PGSI. Every player answering the PGSI had to have at least one betting activity 30 days prior. Therefore, the total bet 30 days prior to answering the PGSI was greater than 0.

Further analysis identified a subgroup of those with gambling problems (greater harm problem gamblers [GHPGs]) based on specific PGSI items. Although all nine PGSI items are scored equally, specific items are more strongly associated with gambling harm than others. For example, selecting 'almost always' in response to statements such as *"Have you felt that you might have a problem with gambling?"*, *"Has gambling caused you any health problems, including stress or anxiety?"*, and *"Has your gambling caused any financial problems for you or your household?"* are a stronger indicator of gambling-related harm compared to items like borrowing money to gamble or receiving criticism for gambling.

Players who scored 8+on the PGSI and answered the three aforementioned harm-related items with 'almost always' were classified as 'greater harm problem gamblers' (GHPGs). This means that these players reported problem gambling and reported experiencing financial, social, and psychological gambling-related harm. The selection of these three items was in line with previous definitions of gambling harm which include social, psychological as well as financial aspects (Langham et al., 2015).

Statistical Analysis

First, a hierarchical logistic regression was used to compare a model which included all explanatory variables listed in Appendix 2 with a model that only computed the behavioral explanatory variables.

$$D = -2\left(\log L_0 - \log L_1\right)$$

where:

- L_0 = Log-likelihood of the null model (simpler model, typically without additional predictors).
- L_1 = Log-likelihood of the full model (more complex model with additional predictors).
- D = Deviance statistic, which follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the two models.

The hypotheses for the Likelihood Ratio Test (LRT) were:

- Null hypothesis H_0: The simpler model fits the data just as well as the more complex model.
- Alternative hypothesis H_1: The more complex model provides a significantly better fit.

Apart from logistic regression (Peng et al., 2002), a number of machine learning algorithms were used to predict self-reported problem gambling.

- *Decision tree:* A decision tree (Breiman et al., 1984) is a supervised machine learning algorithm used for classification and regression tasks. It follows a tree-like structure, where each internal node represents a decision based on a feature, each branch represents an outcome of the decision, and each leaf node represents a final prediction (class label or numerical value). Together with logistic regression, decision tree algorithms were the most basic approaches used in the present study.
- *Random forest:* Random forest is an ensemble learning method used for classification, regression, and feature selection tasks. It was developed by Breiman (2001) and is based on the bagging (bootstrap aggregating) technique. The model operates by constructing multiple decision trees during training and outputs the most common prediction (for classification) or the average prediction (for regression) across all trees.
- *Support vector machines:* Support vector machines (SVMs) (Cortes & Vapnik, 1995) are supervised machine learning algorithms used for classification, regression, and outlier detection. Developed by Vapnik and Chervonenkis (1995), SVMs are particularly effective in high-dimensional spaces and for datasets where the number of features is greater than the number of samples.

- *Gradient boost machine:* Gradient boost machine (GBM) is a powerful ensemble learning algorithm used for classification and regression tasks. It was developed by Friedman (2001) and has since become one of the most widely used techniques in machine learning due to its high predictive accuracy.

For the logistic regression, odds ratios (Bland and Altman, 2000) were computed, and for ML models, Shapley Additive exPlanations (SHAP) values (Kim & Kim, 2022). A supervised machine-learning approach was implemented to predict the target variable using multiple algorithms for comparison. The following models were trained and evaluated: linear regression, decision tree, random forest, support vector machine (SVM), and gradient boosting machine (GBM). The following hyperparameters were used:

- number of estimators (n_estimators): 100
- learning rate (learning_rate): 0.10
- maximum tree depth (max_depth): 3
- random state (random_state): 42 (to ensure reproducibility)

The decision tree and random forest models were initialized with default parameters, except for the random seed (random_state = 42). The SVM model used the default radial basis function (RBF) kernel. Model performance was evaluated using accuracy on a hold-out test set obtained via an 80/20 train–test split (train_test_split with random_state = 42). Given the small sample size, the number of univariate und multivariate statistical tests increases the chance of Type I errors. Therefore, the statistical tests should be seen exploratory and require validations using larger sample sizes.

The dataset was analyzed using the *Python 3.10* programming language (Van Rossum, 2007). To implement the machine learning algorithms, the scikit-learn library (Pedregosa et al., 2011) was utilized. The performance of the models was assessed visually through receiver operating characteristic (ROC) curves (Hanley & McNeil, 1982) and quantitatively by calculating the area under the curve (AUC) (Bradley, 1997). To evaluate the validity of the machine learning models, the dataset was divided into training and test sets, with 80% of the data allocated for training and 20% reserved for testing. Apart from AUC, precision, recall, overall classification accuracy, and F1 (Humphrey et al., 2022) score were computed.

- Precision refers to the percentage of predicted self-reported problem gamblers who were actual self-reported problem gamblers.
- Recall refers to the percentage of actual self-reported problem gamblers who were also predicted to be self-reported problem gamblers.
- Accuracy refers to the overall percentage of correctly classified records (across self-reported problem gamblers and non-problem gamblers).
- F1 is a score which measures recall and precision for both the classification of problem gamblers and non-problem gamblers. A value of 1 indicates a perfect model and 0 indicates a random model. In contrast to the accuracy, F1 takes into account how well non-problem gamblers as well as problem gamblers are classified by the model.

A high precision rate indicates that the majority of predicted problem gamblers are actually problem gamblers. A high recall rate indicates that the majority of actual problem gamblers are detected by the algorithm. If a sample is imbalanced as in the present study (i.e., the actual number of problem gamblers is far less than the number of non-problem gamblers) the values depend largely on the score threshold. If the chosen threshold from which records are classified into the minority class is low then there is a high recall value, but a low precision value. Every problem gambler is detected, but many non-problem gamblers are also falsely classified. If the chosen score threshold is high, very few records are classified in the minority class.

Data Cleaning

A total of 2,130 players answered the PGSI between April 2023 and February 2025 and placed at least one bet in the 30 days prior to answering the PGSI. The black bars in Fig. 1 represent the distribution of the scores. Lower scores are more frequent and higher scores are less frequent. In line with Auer and Griffiths (2023a) and Hopfgartner et al. (2024), players who answered the nine PGSI questions in less than one minute were removed from the analysis. The grey bars in Fig. 1 represent the distribution of the cleaned dataset. The cleaned dataset comprised 1,611 players. A slight drop in the percentage of players scoring 0 as well as 27 is visible between the black and grey bars. In the total sample of 2,130 players, 677 scored a PGSI score of 0 and 10 scored a PGSI score of 27. In the reduced sample of 1,611 players, 476 scored a PGSI score of 0 and two scored a PGSI score of 27. It is likely that players who rush through the nine PGSI questions answer each item with ‘never’ or ‘almost always’. These are the most extreme answers and are on the very left or very right side, respectively. The authors also examined alternate minimum response times for answering the PGSI.

- Minimum response time 30 s: 643 players with PGSI score zero and 5 with PGSI score of 27.
- Minimum response time 90 s: 75 players with PGSI score zero and 5 with PGSI score of 27.

Removing players who answered the nine PGSI questions in less than 30 s did not visibly change the sample size (from 2,310 to 2,069). Also, it did not visibly reduce the

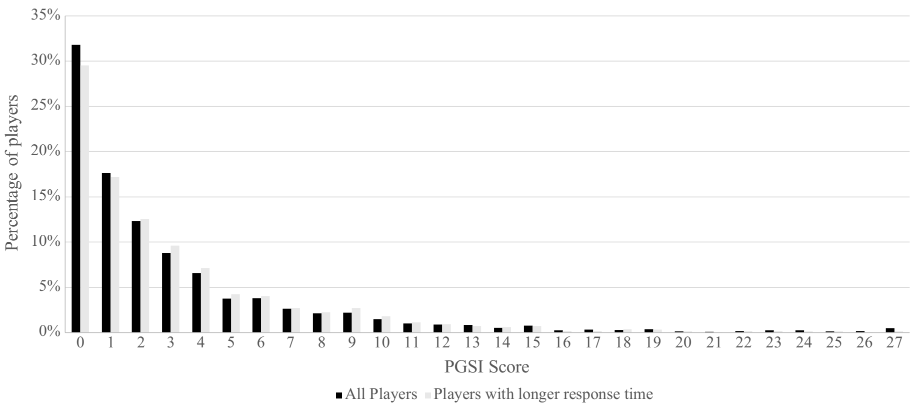

Fig. 1 Number of players who answered the Problem Gambling Severity Index before and after removing records with a short response time

number of players in the two extreme categories. In the entire sample, 677 players scored 0, and in the reduced sample, 643 players scored 0. Only retaining players who answered the PGSI questions in at least 90 s markedly reduced the sample size from 2,139 to 378. For that reason, the methodology of two previously published studies (Auer & Griffiths, 2023a; Hopfgartner et al., 2024) was followed, and a minimum response time of more than one minute was applied.

Results

The sample (N = 1611) comprised 690 females (42.8%), 693 males (43%), and 227 'other' (14%). Other could mean that players did not want to disclose their gender or did not identify as male or female. The average age among the 1,611 players was 55 years (SD = 16). Out of the 1,611 players, 209 had a PGSI score of 8 or more (13%) and were classified as problem gamblers. Figure 2 reports the percentage of players in each answer category for each of the nine PGSI items. Most players (91%) answered the question *"Have you borrowed money or sold anything to get money to gamble?"* with 'never'. The questions *"Have you gone back to try to win to back the money you'd lost?"* and *"Have you felt guilty about the way you gamble or what happens when you gamble*?" had the highest percentage of players who do not answer 'never'.

Out of the 1,611 players, 1,456 had at least one wager on lottery games (90%), 828 had at least once wager on casino games (51%), 121 had at least one wager on sports (7.5%), and 169 had at least wager once on bingo games (10%). These numbers are not exclusive. For example, the majority of the 1,456 lottery players also wagered on other games. Figure 3 reports the problem gambling rate in each game category. It was found that 12% of players who played lottery games reported problem gambling and 19% of players who did not play lottery games reported problem gambling; 18% of casino players reported problem gambling and 8% of players who did not play casino reported problem gambling. The respective z-tests comparing the percentage of problem gamblers are: (i) lottery: $z = -2.49$, $p = 0.013$; (ii) casino: $z = 6.28$, $p < 0.001$; (iii) bingo: $z = 2.94$, $p = 0.003$, and (iv) sport: $z = 1.73$, $p = 0.08$. The statistical tests compare the problem gambling rates for players who wagered in the respective game category and

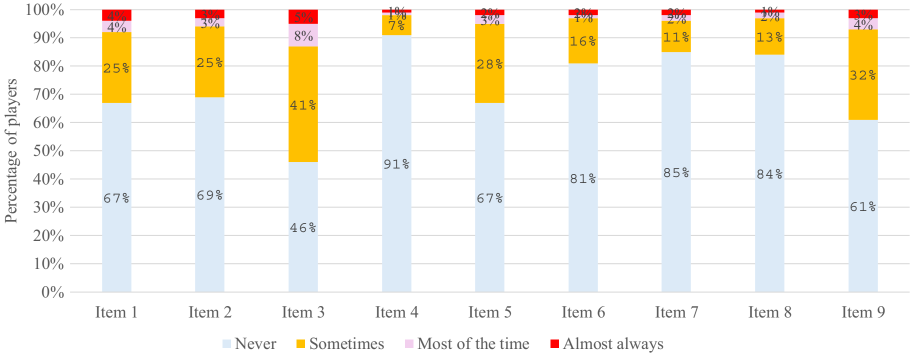

Fig. 2 Percentage of players answering each question on the Problem Gambling Severity Index 'never', 'sometimes', 'most of the time' or 'almost always'

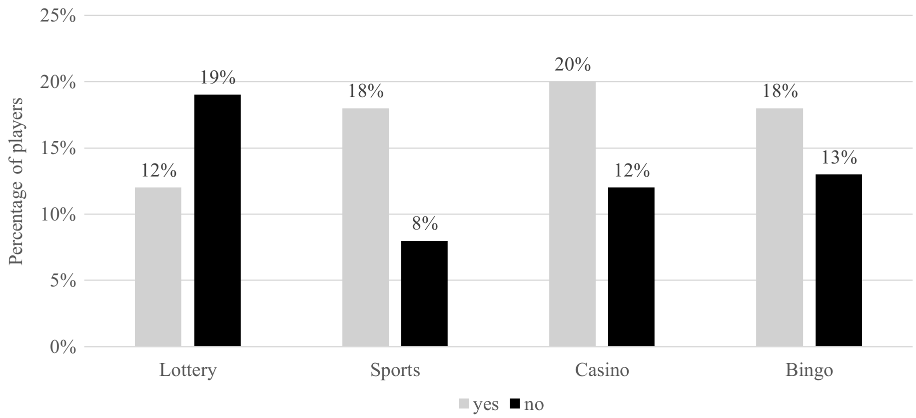

Fig. 3 Percentage of self-reported problem gamblers for each game-category**.** Note. 'Yes' refers to gamblers engaged in the respective game category and 'No' refers to gamblers who did not engage in the respective game category)

players who did not. Engaging in casino gambling, sports betting, and bingo playing increased the chance of problem gambling. Only playing lottery games was associated with a lower rate of problem gambling compared to players who did not play lottery games.

Out of the 1,611 players 793 engaged in one game-category (49%), 682 engaged in two game-categories (42%), 127 engaged in three game-categories (7.8%), and 9 engaged in all four game-categories (0.5%). Figure 4 reports the percentage of problem gamblers for each game-category frequency. There was a positive correlation between the number of game-categories engaged in and self-reported problem gambling. However, there was a very low number of players engaging in all four game categories (i.e., nine gamblers), therefore this should be treated with caution. The percentages of

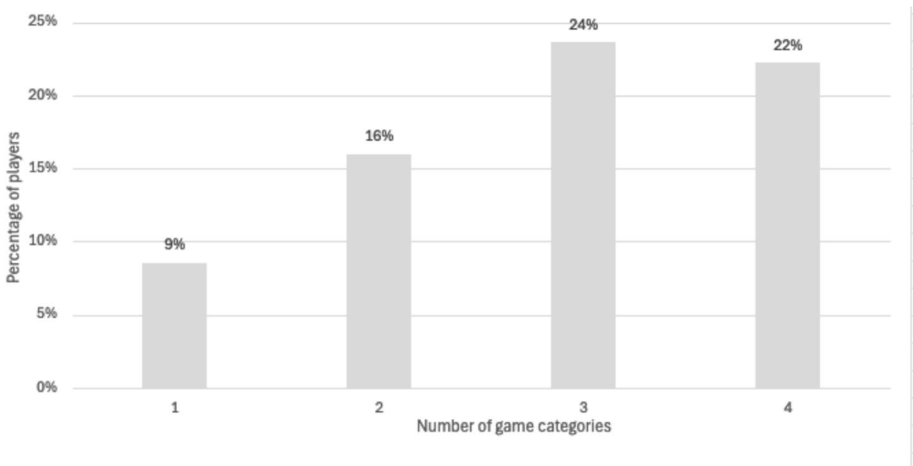

Fig. 4 Percentage of problem gamblers among players who wagered on one, two, three or four different types of gambling

problem gamblers among players who wagered in one, two, three or four different types of gambling were significantly different ($\chi^2 = 28.27$, $p < 0.001$).

In order to further understand the relationship between self-reported problem gambling and game-category, groups of players who engaged only in one specific game-category were computed:

- Out of the 1,611 participants, 648 exclusively played lottery games (40%). Out of these 41 reported problem gambling (6.3%).
- Out of the 1,611 participants, 125 exclusively played casino games (7.8%). Out of these 20 reported problem gambling (20%).
- Out of the 1,611 participants, four exclusively played bingo games (0.25%). Out of these none reported problem gambling (0%).
- Out of the 1,611 participants, 16 exclusively played sports games (1%). Out of these two reported problem gambling (12.5%).

The game-category specific percentage for lottery (6.3%) was significantly different from the overall percentage of self-reported problem gambling (13%) ($z = -4.85$, $p < 0.001$). The respective value for casino players was: $z = 2.21$, $p = 0.0274$. Due to the small sample size, no statistical tests were computed for bingo only and sports betting-only players. A total of 793 players engaged in only one game category (49%), and 648 out of these only played lottery games (82%). As reported above, these lottery-only gamblers reported a significantly lower percentage of problem gambling compared to the overall rate of problem gambling (6.3% vs. 13%). Lottery-only gamblers also represented a relatively large proportion of the overall sample (40%). Algorithms predicting problem gambling would heavily rely on lottery-only gamblers and this would conceal relevant behavioral patterns related to actual gambling behavior such as betting, playing time, depositing frequency, etc. The main outcome of any algorithm would be that players who only engage in lottery games have a significantly lower chance of reporting problem gambling. Therefore, in order to extract relevant patterns of play related to problem gambling, the 648 players who only played lottery games were excluded from the multivariate and machine learning algorithms. The analyses were performed on the 963 players who did not exclusively engage in lottery games. Every player in this sample engaged in either casino gambling, sports betting or bingo playing apart from lottery. Only players who exclusively played lottery games were excluded. Out of these 963 players, 168 scored 8 or more on the PGSI.

First, two logistic regression models were computed. In the first model, all variables listed in Appendix 1 were included. In the second model, only the behavioral variables in Appendix 2 were included. This hierarchical approach allowed for a significance test between the two models' reported log likelihoods. The log likelihood of the reduced model was −94.62 and the log likelihood of the full model was −99.55. The full model performed worse than the reduced model which was indicated by the more negative log likelihood. The respective chi-square test comparing the two log likelihoods was not significant ($\chi^2 = 9.86$, $p = 0.45$). the chi-square test statistic was computed as follows: $\chi^2 = -2 \times (-94.62 - [-99.55]) = -2 \times (-94.62 + 99.55) = -2 \times (4.93) = -9.86$. This means that the monetary player tracking features did not contribute significantly to the model explanation on top of the behavioral player tracking features. Self-reported problem gambling can be explained by behavioral aspects sufficiently. A model with the ten behavioral aspects explained self-reported problem gambling even better than a model with all 23 variables. The Akaike's information criteria (AIC) for the full model was 245.10 ($AIC = 2 \cdot 23 - 2[-99.55] = 46 + 199.10 = 245.10$), and for the reduced model was 215.24 ($AIC = 2 \cdot$

13 − 2[− 94.62] = 26 + 189.24 = 215.24). The AIC (which is computed as 2*df-2*LL) also indicated that the reduced model (lower AIC value) performed better than the full model.

Five algorithms were used to predict self-reported problem gambling using the player tracking features listed in Appendix 2. Figure 2 reports the area under the curve (AUC) for each of the five algorithms. The goodness of fit statistics were computed on a 20% test sample in order to avoid overfitting the training data. Logistic regression, which is the most basic algorithm, performed best and reached an AUC of 0.789, followed by random forest with an AUC of 0.776. Figure 5 also reports the 95% confidence intervals for the AUC values.

Figure 6 displays the trade-off between precision (proportion of actual problem gamblers correctly identified) and recall (proportion of predicted problem gamblers that are correct) for the five ML models. The random forest model achieved the highest precision-recall (PR) AUC (0.463), indicating the best overall trade-off between precision and recall. Gradient boosting followed closely (PR AUC = 0.422), while logistic regression, decision tree, and support vector machine models showed weaker discrimination (PR AUCs = 0.382, 0.356, and 0.335, respectively). The steep decline in precision as recall increases suggests a moderate class imbalance and a limited number of positive cases. Overall, ensemble-based methods (i.e., random forest and gradient boosting) demonstrated superior ability to correctly identify positive instances with fewer false alarms compared with single-model approaches.

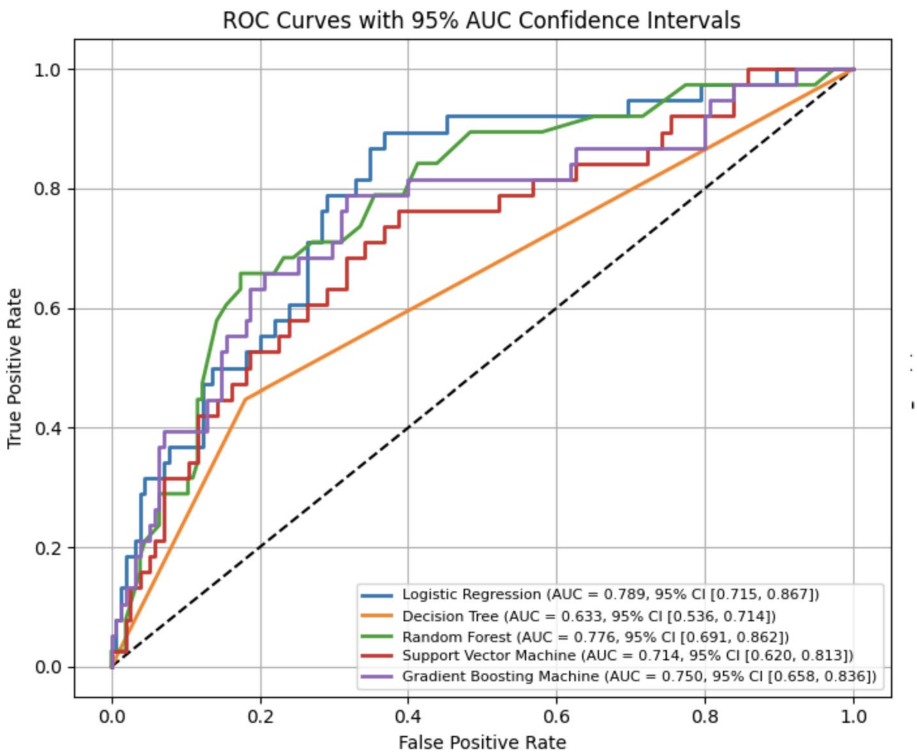

Fig. 5 ROC chart for the five algorithms

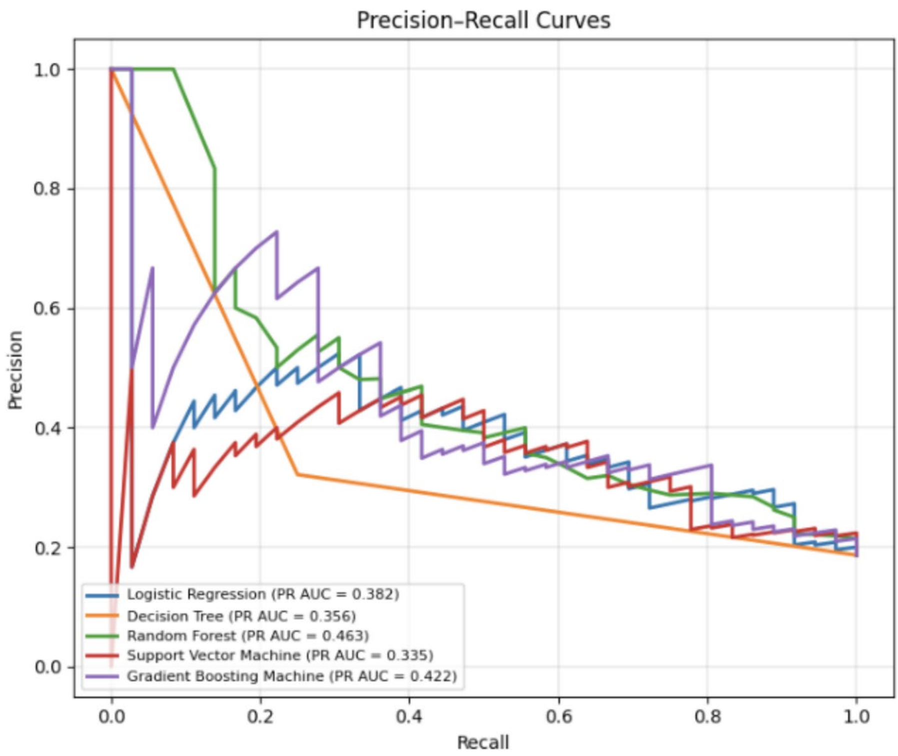

Fig. 6 Precision-recall curves for the five algorithms

Figure 7 displays the probability output by the model (e.g., 0.2 means "20% chance of positive") on the *x*-axis and the proportion of true positives among samples with that predicted probability on the *y*-axis. The dashed diagonal line represents perfect calibration — predictions exactly match observed outcomes. The calibration analysis showed that none of the evaluated models produced perfectly reliable probability estimates. Logistic regression and random forest exhibited relatively stable (though biased) calibration curves. Logistic regression tended to slightly underpredict and random forest consistently underestimated event likelihoods. Decision tree and support vector machine showed substantial deviation from the ideal diagonal, indicating overconfidence and irregular probability scaling. Gradient boosting showed moderate alignment at low probabilities but diverged for higher predicted risks.

Table 1 reports goodness of fit statistics for the five machine learning models. These statistics were heavily influenced by the imbalanced dataset. Only 17.4% of the total dataset were self-reported problem gamblers. In the logistic regression, 23% of predicted self-reported problem gamblers were actual problem gamblers, and 77% of predicted self-reported problem gamblers were non-problem gamblers. This high percentage of false-positives is mostly due to the imbalanced dataset. In the logistic regression, 52% of actual self-reported problem gamblers were also classified as such, and 48% of actual self-reported problem gamblers were wrongly classified. The logistic regression displays the highest F1 score which means that it balanced the correct prediction of problem gamblers and non-problem gamblers better than the other algorithms.

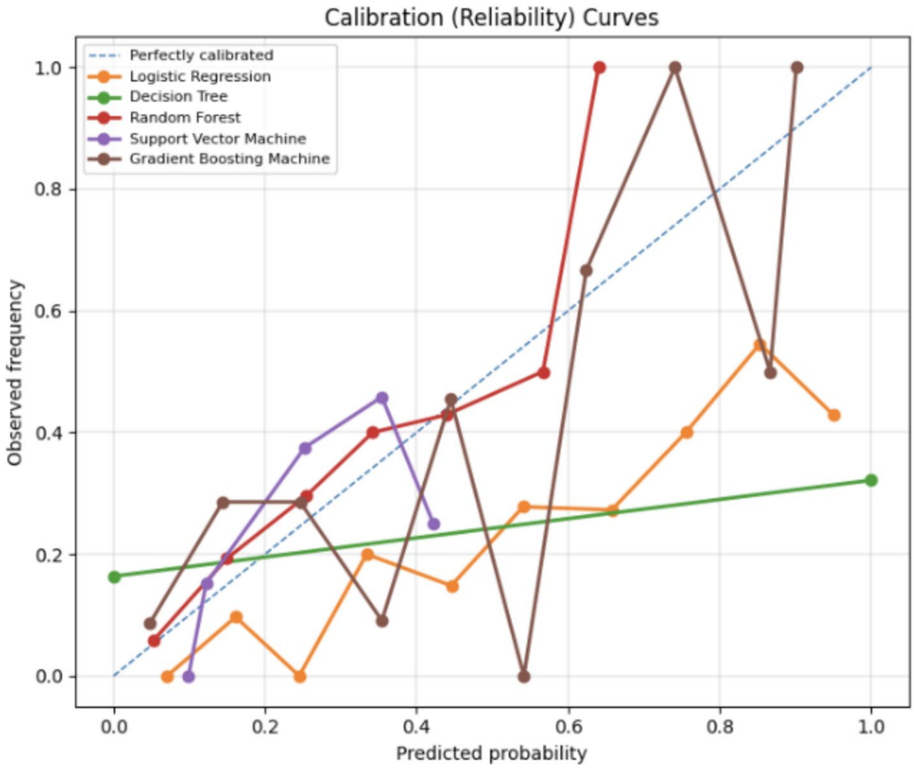

Fig. 7 Calibration (reliability) curves for the five algorithms

Table 1 Goodness of fit statistics for the five machine learning models

	Precision (minority class)	Recall (minority class)	Accuracy	F1
Logistic regression	0.23	0.52	0.67	0.32
Decision tree	0.24	0.31	0.75	0.27
Random forest	0.27	0.10	0.82	0.15
Support vector machines	0.17	0.03	0.83	0.06
Gradient boost machine	0.36	0.17	0.83	0.23

A more detailed analysis of the player tracking features impact for the logistic regression and the random forest reported the following features:

- *Logistic regression:* gender, percentage of amount bet on lottery and casino games, number of game types played, age, average number of deposits per day, average number of monetary deposits per session, and total number of monetary deposits.

- *Random forest:* average number of monetary deposits per day, average amount of money bet per day, average amount of monetary deposits per day, average number of monetary deposits per session, percentage of amount of money bet in lottery games, age, average session length, and number of monetary deposits.

For the random forest model, the overall accuracy was also computed across six age groups (younger than 25 years, 25–34 years, 35–44 years, 45–54 years, 55–64 years, and older than 64 years). The accuracy ranged between 85% and 89% which reflects a stable pattern across all age categories. The random forest's accuracy was 82%, both for female and male players.

Table 2 reports the odds ratios for the logistic regression and the SHAP values for the random forest model including the respective probabilities. The odds ratio showed that the chance of reporting problem gambling was 2.18 times higher for males compared to females. Similarly, one deposit increased the chance of problem gambling by 1.8%. The SHAP value's *p*-values also reflected the significant variables. The odds ratios of the logistic regression model were computed for all the independent variables. Being male (OR $=$ 2.18; $p = 0.037$), amount of money deposited (OR $=$ 0.99, $p = 0.043$), and total session length (OR $=$ 1,0006, $p = 0{,}02$) were significant.

Among the 168 players who scored eight or higher on the PGSI, 44 individuals responded with 'almost always' to at least one of the three items indicating greater gambling harm. Additionally, this subgroup (GHPGs) accounted for 4.6% of the 963 players included in the further analysis.

Table 3 reports average values for non-problem gamblers, problem gamblers, and GHPGs. Problem gamblers (i) were younger, (ii) were more likely to be male, (iii) bet more money, (iv) bet more money on casino games and less money on lottery games, (v) deposited more money frequently in total, on a day and in session, (vi) had more failed deposits (GPHPs value was lower than NPGs), and (vii) had longer session lengths.

Figure 8 displays the ROC and AUC values for the five algorithms predicting greater-harm problem gambling. Because the target variable (i.e., self-reported problem gambling) was extremely unbalanced, the ROC chart is only of descriptive value. Like the predictions for all PGs, logistic and random forest performed best with respect to AUC. The goodness of fit statistics were computed using a 20% test sample of the GHPGs. Table 3 reports the average values across relevant player tracking features for the GHPGs. Compared to those without gambling problems, the average values of the GHPGs were in the same direction as for the PGs. Similar to PGs, GHPGs were younger, more likely to be male, wagered more money, spent more money on casino games and less money on lottery games, deposited money more frequently, played longer, and played more frequently.

Precision, recall and F1 for the prediction of the GHPGs are reported in Table 4. Only 4.6% of the sample were self-reported problem gamblers reporting greater harm and the goodness of fit statistics were expected to perform poorly. Moreover, the goal of the study was not to find a perfect score threshold for prediction accuracy, but to identify the most significant player tracking features indicative of self-reported problem gambling.

In the logistic regression, 5% of predicted self-reported problem gamblers were actual problem gamblers, and 95% of predicted self-reported problem gamblers were non-problem gamblers. This high percentage of false-positives is mostly due to the imbalanced dataset. In the logistic regression, 60% of actual self-reported problem gamblers were also classified as such, and 40% of actual self-reported problem gamblers were wrongly classified. The random forest displayed the highest F1 score which means that it balanced the

Table 2 Odds ratios for the logistic regression and SHAP values for random forest

Feature Number	Feature	Odds ratio	*p*-value	SHAP	*p* value
1	Age (in years)	0.96	< 0.001*	−0,00127	< 0.001*
2	Male	2.18	0.037 *	−0.0001	0.002*
3	Number of bets	1.005	0.047*	−0.00035	0.023*
4	Amount of money bet	0.999	0.99	0.0015	0.231
5	Amount of money deposited	1.003	0.043*	0.0012	0.213
6	Amount of money lost	0.999	0.99	0.0012	0.2234
7	Amount of money won	1	0.99	0.0015	0.83
8	Amount of failed deposits	1	0.97	0.00015	0.65
9	Number of failed deposits	0.97	0.83	0.00015	0.31
10	Amount of cancelled withdrawals	1	0.88	0.000013	0.45
11	Number of cancelled withdrawals	1	0.87	0.000013	0.46
12	Total session length (minutes)	1.0006	0.02*	0.000062	0.017*
13	Number of different gambling days	0	0.32	0.000242	0.03
14	Number of monetary deposits	1.018	0.012*	0.000576	0.15
15	Average number of monetary deposits per day/session	1.31/0.80	0.021/0.018*	0.000063/0.000683	0.023/0.014*
16	Average amount of money bet per day/session	1.001/0.99	0.38/0.25	0.00079/−0.0004	0.04/0.02*
17	Average amount of money deposited per day/session	1.006/1.009	0.57/0.46	0.00087/0.00135	0.45
18	Average amount of money lost per day/session	0.99/1.003	0.59/0.73	0.0015/0.0011	0.12
20	Percentage of amount of money wagered on lottery games	1.24	0.02*	−0.000076	0.02*
21	Percentage of amount of money wagered on casino games	1.88	0.018*	0.00013	0.66
22	Percentage of amount of money wagered on bingo games	0.23	1	0.000062	0.76
23	Percentage of amount of money wagered on sports betting	0.32	1	0.0001	0.23

* Significant finding

Table 3 Average values for non-problem gamblers and problem gamblers

	Non- problem gambler	Problem Gamblers	Greater-harm problem gamblers
N	795	168	44
Age	56	48	46
Female	46%	38%	32%
Male	38%	49%	48%
Amount of money bet (€)	4.01	20.63	34.79
Percentage of money spent on lottery	33%	25%	18%
Percentage of money spent on casino	59%	70%	78%
Average number of monetary deposits per session	0.62	1	0.92
Average number of monetary deposits per day	0.83	1.76	1.63
Number of monetary deposits	14	34	29
Number of failed monetary deposits	0.26	0.45	0.25
Average session length (in minutes)	23	33	37
Number of gambling days	15	16	16

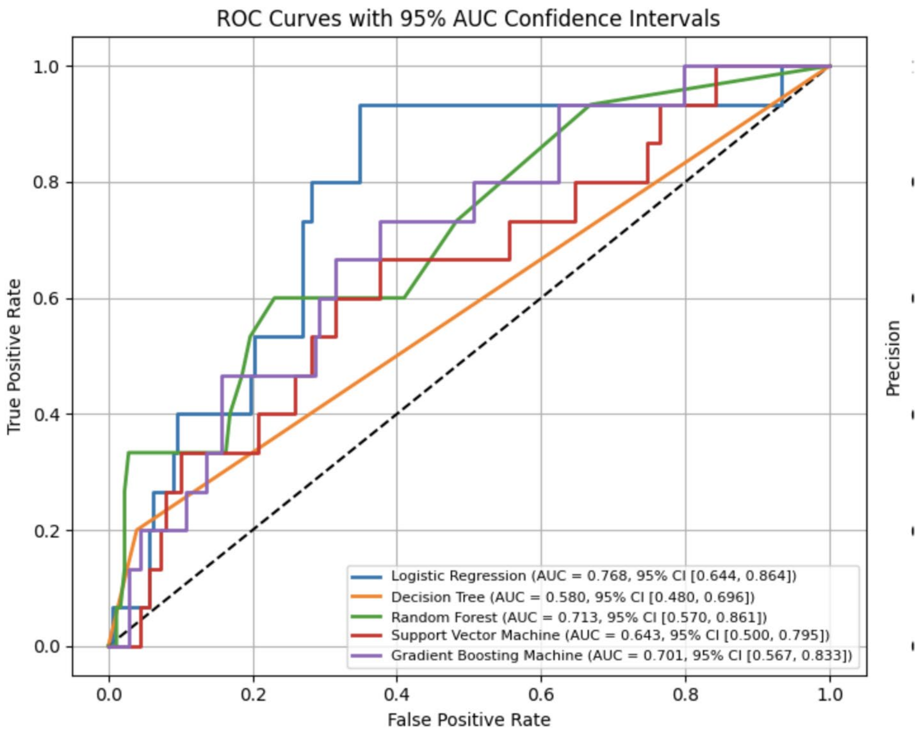

Fig. 8 ROC chart for the five algorithms based on the greater-harm problem gamblers

Table 4 Goodness of fit statistics for the five machine learning models and the prediction of the greater-harm problem gamblers

	Precision (minority class)	Recall (minority class)	Accuracy	F1
Logistic regression	0.05	0.60	0.67	0.09
Decision tree	0.06	0.60	0.73	0.10
Random forest	0.10	0.60	0.84	0.17
Support vector machines	0.02	0.20	0.69	0.03
Gradient boost machine	0.07	0.40	0.84	0.11

correct prediction of problem gamblers and non-problem gamblers better than the other algorithms.

Discussion

In the present study, a sample of 2,130 real-world online gamblers answered the nine items on the PGSI. The sample was from a gambling website that offered lottery games, casino games, bingo games, and sports betting. After removing gamblers who completed the PGSI items in a very short response time, 1611 gamblers remained. This data cleaning procedure was in line with previous studies (e.g., Auer & Griffiths, 2023a; Hopfgartner et al., 2024). All players who exclusively played online lottery games were removed. This means that the results refer only to players who did not exclusively play lottery games. The average age among the 1,611 players was 55 years (SD = 16), 42.8% players were female, 43% were male, and 14% were categorized as 'other'. It should be noted that 'other' could also mean that players were not willing to disclose their gender. This has to be taken into consideration when interpreting the findings regarding gender. The average age was higher and the percentage of females was higher compared to other player tracking studies (e.g., Auer & Griffiths, 2023a; Dragicevic et al., 2015; Hayer & Meyer, 2011; Hopfgartner et al., 2023a, 2023b; 2024; Percy et al., 2016). In fact, the average age and gender distribution were more in line with studies that have incorporated lottery players (e.g., Murch et al., 2023; Perrot et al., 2022). These findings indicate that lottery players tend to be older and have a more balanced gender distribution.

Are Problem Gambling Rates Different Between Players Who Prefer Lottery or Casino Games? (RQ1)

Players who engaged in at least one lottery game 30 days prior to answering the PGSI were less likely to be problem gamblers (12%) compared to players who did not play lottery games (19%). For all other game-categories the relationship was reversed. For example, 18% of players who engaged in casino games reported problem gambling, whereas 8% of players who did not engage in casino games reported problem gambling. The respective rates for bingo games and sports betting were 20%/12% and 18%/13%, respectively. Several studies have reported higher problem gambling rates for casino games, sport betting, and bingo games, respectively (Allami et al., 2021; Lombardi et al., 2024; Mazar et al., 2020; Tran et al., 2024).

Based on a sample of 5046 gamblers, Mazar et al. (2020) reported that monthly gamblers participating in casino gambling, bingo playing, and sports betting contained a higher proportion of problem gamblers. They also found that high gambling involvement was also positively associated with problem gambling, and gambling involvement was also positively associated with intensity of gambling. Therefore, intensity of gambling may be partly driving the relationship between involvement and problem gambling. Specific gambling formats (i.e., casino gambling, bingo playing, and sports betting) mediated the relationship between involvement and problem gambling.

In a meta-analysis of 273,946 adults, Allami et al. (2021), reported an association between problem gambling and continuous forms of gambling such as casino gambling, bingo playing, and sports betting. In another meta-analysis of 366 studies, Tran et al. (2024) concluded that the prevalence of problem gambling was greatest among online casino gamblers or slots gamblers (15.8%). Using a sample of adolescents from 33 countries, Lombardi et al. (2024) concluded that playing slot machines demonstrated the highest predicted probability of risky gambling behavior when combined with online gaming.

In the present study, there was also a positive association between the number of game-categories played and the percentage of self-reported problem gambling. However, out of the 793 players who only engaged in one game-category, 648 only played lottery games. This means that most players who played at least two game-categories played casino, bingo or sports games. Lombardi et al. (2024) also reported that engaging in more gambling activities was associated with a higher prevalence of problem gambling.

Are Behavioral Variables Sufficient in Explaining Self-Reported Problem Gambling? (RQ2)

In the present study, there was no significant difference between the logistic regression model which included only the behavioral player tracking features and the full set of player tracking features. This means that self-reported problem gambling can be sufficiently explained without taking into account how much money was wagered, deposited, lost or won. This finding concurs with a similar study by Hopfgartner et al. (2024), and supports the notion that machine learning models which predict problem gambling can be applied across countries and jurisdictions. Moreover, it is important to note that the present study is first to predict self-reported problem gambling using a sample of online lottery players, casino gamblers, bingo players, and sports bettors.

Which Algorithms Best Predict Self-Reported Problem Gambling? (RG3)

Out of the 1,611 players which also included the group playing lottery games exclusively, 209 reported a PGSI score of 8 or more (13%). Out of the 963 players who did not exclusively play lottery, 168 reported a PGSI score of 8 or more (17%). Comparable behavioral tracking studies have reported similar percentages of problem gambling (e.g., Auer & Griffiths, 2023a; Hopfgartner et al., 2024; Murch et al., 2023; Perrot et al., 2022). The lower percentage of problem gambling among lottery-only players is also in line with previous findings that playing the lottery is a less risky form of gambling than other types of gambling (e.g., Allami et al., 2021; Hing et al., 2022; Tran et al., 2024). The PGSI items *“Have you gone back to try to win to back the money you’d lost?”* and *“Have you felt guilty about the way you gamble or what happens when you gamble?”* had the highest percentage of players who did not answer ‘never’. The PGSI item *“Have* *you felt that you might have a problem with gambling"* was answered by 33% of participants at least 'sometimes'.

In evaluating different machine algorithms, logistic regression and random forest emerged as the most effective in predicting self-reported problem gambling. Random forest was superior to the other four algorithms when taking into account ROC, precision-recall (PR) and calibration plots. Logistic regression was only superior in the PR plot. This result is consistent with Auer and Griffiths (2023a), where random forest outperformed other models in identifying gambling-related harm among a sample of online casino players. Similarly, Murch et al. (2023) found that random forest models achieved the highest classification performance in predicting PGSI scores among Canadian online gamblers.

The superior performance of logistic regression in the present study may be due to its interpretability and robustness in handling structured data with well-defined behavioral markers. While gradient boosting machines (GBMs) and support vector machines (SVMs) have demonstrated high predictive accuracy in other gambling studies (e.g., Louderback et al., 2021; Perrot et al., 2022), they did not outperform random forest or logistic regression in the present study.

Although the goal of the study was not to identify the best score threshold to classify self-reported problem gamblers from non-problem gamblers, the authors computed a number of classification accuracy statistics. These statistics also supported the superiority of the logistic regression because 52% of actual self-reported problem gamblers were predicted accurately. This was the highest recall value across the five machine learning algorithms. The logistic regression also displayed the highest F1 (0.32) score among the five machine learning algorithms. The lower accuracy of the logistic regression (0.67%) is simply due to the fact that it tended to classify more players as problem gamblers. Consequently, the overall accuracy cannot be interpreted independently. If the algorithm classified the entire sample as non-problem gamblers it would be correct in 83% of the cases. This is because 83% of players were actual non-problem gamblers and 17% were actual problem gamblers. The overall low precision and recall was in line with Murch et al. (2023) and was most likely due to the imbalanced sample.

Out of the 168 players who scored 8 + on the PGSI, 44 answered the three greater harm items (i.e., *"Have you felt that you might have a problem with gambling?", "Has gambling caused you any health problems, including stress or anxiety?",* and *"Has your gambling caused any financial problems for you or your household?")* with 'almost always'. In line with Auer and Griffiths (2023a), this subgroup was regarded as experiencing significant harm caused by gambling. Based on the AUC metric, logistic regression performed best in predicting self-reported problem gambling. Due to the highly imbalanced dataset the ROC curves were much more discrete compared to the prediction of the overall group of problem gamblers. The GHPGs wagered larger amounts, had longer sessions, played more casino games and less lottery games than the overall group of PGs. This is another indication for the validity of the answers to the PGSI because players who experience greater harm from problem gambling would also be expected to gamble more intensely.

The present study contributes to the growing body of research on the application of machine learning techniques to predict self-reported problem gambling using player tracking data. The findings align with previous research, particularly Auer and Griffiths (2023a) and Hopfgartner et al. (2024), in demonstrating that behavioral indicators such as deposit frequency, bet size, and session length serve as strong predictors of gambling-related harm. By incorporating a diverse dataset that included lottery players, casino gamblers, sports bettors, and bingo players, the present study expands upon prior work, which typically focused on single gambling verticals.

The results also indicated that specific behavioral tracking features—such as the average number of monetary deposits per session, total amount of money bet per day, session length, and casino gambling involvement—were among the most significant predictors of self-reported problem gambling. These findings are in line with Auer and Griffiths (2023a), who identified frequent deposits within a session as key indicators of problem gambling in their study of 1,287 European online casino players.

The findings have important policy and regulatory implications, particularly in jurisdictions where gambling operators are required to monitor gambling-related risk. Many regulatory frameworks, such as those in the UK and Sweden, emphasize consumer protection through responsible gambling tools such as voluntary self-exclusion (VSE), personalized feedback, and deposit limits (Gainsbury, 2014; Håkansson & Henzel, 2021).

Moreover, the present study suggests that behavioral indicators alone may be sufficient to identify at-risk players, even without explicit self-reported problem gambling scores. This aligns with Hopfgartner et al. (2024), who found that monetary intensity variables were not necessarily the strongest predictors of problem gambling, reinforcing the idea that engagement patterns and depositing behavior may be more indicative of gambling risk.

Limitations and Future Directions

Despite the strengths of this study, several limitations must be acknowledged. First, the dataset was limited to a single online gambling operator which may not fully generalize to other platforms with different game offerings, marketing strategies, or responsible gambling interventions. Second, while the PGSI is a widely used instrument for assessing problem gambling, self-reported problem gambling scores remain subject to bias, and some players may underreport or overreport their gambling-related harms. Third, compared to most previous studies using account-based tracking data, the sample size was relatively small. Future research should (i) aim to expand datasets to include a broader range of gambling operators and jurisdictions, (ii) analyze player behavior over a longer period of time, and (iii) improve the interpretability of ML models, ensuring that predictive algorithms align with regulatory expectations and consumer protection policies.

Conclusion

The present study provides further empirical evidence supporting the use of machine learning models to identify self-reported problem gamblers based on player tracking data. The findings reinforce prior research (e.g., Auer & Griffiths, 2023a; Hopfgartner et al., 2024; Louderback et al., 2021; Murch et al., 2023; Perrot et al., 2022) by confirming that behavioral indicators—such as monetary deposit frequency, session length, and game-type engagement—are significant predictors of gambling-related harm. Given the increasing regulatory focus on data-driven gambling harm prevention, these insights can inform responsible gambling strategies by enabling operators to identify and intervene before gambling-related problems escalate. However, continued research is necessary to refine these predictive models and ensure their effectiveness in diverse gambling environments.

Gambling operators could build upon the findings of this and previous studies and develop prediction models based on self-reported problem gambling. An important aspect is the balance between specificity and sensitivity. Given that real-world samples will always be unbalanced (only a small proportion of players are typically problem gamblers) prediction algorithms will either produce a high percentage of false positives or a high percentage of false negatives. Gambling operators would need to decide whether they prefer a larger number of falsely classified players as PGs if most of the actual PGs are detected or if they want a low number of false positives which means that fewer PGs will be detected.

Appendix 1

Items in the Problem Gambling Severity Index (Ferris & Wynne, 2001)

Item number and question
(1) Have you bet more than you could really afford to lose?
(2) Have you needed to gamble with larger amounts of money to get the same excitement?
(3) Have you gone back to try to win to back the money you'd lost?
(4) Have you borrowed money or sold anything to get money to gamble?
(5) Have you felt that you might have a problem with gambling?
(6) Have you felt that gambling has caused you any health problems, including stress or anxiety
(7) Have people criticized your betting, or told you that you have a gambling problem, whether or not you thought it is true?
(8) Have you felt your gambling has caused financial problems for you or your household?
(9) Have you felt guilty about the way you gamble or what happens when you gamble?

Items can be answered: Never (0), Sometimes (1), Most of the Time (2), Almost Always (3)

Problem Gambling: Score ≥ 8

Appendix 2

Player tracking features based on the 30 days prior to answering items on the Problem Gambling Severity Index

Feature Number	Feature	Category
1	Age (in years)	Demographic
2	Gender	Demographic
3	Number of bets	Behavioral
4	Amount of money bet	Monetary
5	Amount of money deposited	Monetary
6	Amount of money lost	Monetary
7	Amount of money won	Monetary
8	Amount of failed deposits	Monetary
9	Number of failed deposits	Behavioral
10	Amount of cancelled withdrawals	Monetary
11	Number of cancelled withdrawals	Behavioral
12	Total session length (minutes)	Behavioral
13	Number of different gambling days	Behavioral
14	Number of monetary deposits	Behavioral
15	Average number of monetary deposits per day/session	Behavioral
16	Average amount of money bet per day/session	Monetary
17	Average amount of money deposited per day/session	Monetary
18	Average amount of money lost per day/session	Monetary
19	Average amount of money won per day/session	Monetary
20	Percentage of amount of money wagered on lottery games	Monetary
21	Percentage of amount of money wagered on casino games	Monetary
22	Percentage of amount of money wagered on bingo games	Monetary
23	Percentage of amount of money wagered on sports betting	Monetary

Authors' Contributions Both authors contributed to the preparation and writing of this paper.

Funding Not applicable.

Data Availability The data are not publicly available due to commercial sensitivity.

Declarations

Ethics This study was approved by the second author's university ethics committee (Nottingham Trent University).

Conflict of interest Both authors have received research funding from *Norsk Tipping* (the gambling operator owned by the Norwegian government). The second author (MDG) has received funding for a number of research projects in the area of gambling education for young people, social responsibility in gambling and gambling treatment from *Gamble Aware* (formerly the *Responsibility in Gambling Trust*), a charitable body which funds its research program based on donations from the gambling industry. Both authors undertake consultancy for various gaming companies in the area of player protection, harm minimization, and social responsibility in gambling.

References

Allami, Y., Hodgins, D. C., Young, M., Brunelle, N., Currie, S., Dufour, M., Flores-Pajot, M., & Nadeau, L. (2021). A meta-analysis of problem gambling risk factors in the general adult population. *Addiction, 116*, 2968–2977.

American Gaming Association (2025). AGA commercial gaming revenue tracker. Retrieved February 17, 2025, from: https://www.americangaming.org/resources/aga-commercial-gaming-revenue-tracker/

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders (*5th ed.). American Psychiatric Publishing.

Auer, M., & Griffiths, M. D. (2013). Voluntary limit setting and player choice in most intense online gamblers: An empirical study of gambling behaviour. *Journal of Gambling Studies, 29*(4), 647–660.

Auer, M. M., & Griffiths, M. D. (2015). The use of personalized behavioral feedback for online gamblers: an empirical study. *Frontiersin psychology, 6*, 1406.

Auer, M., & Griffiths, M. D. (2018). Cognitive dissonance, personalized feedback, and online gambling behavior: An exploratory study using objective tracking data and subjective self-report. *International Journal of Mental Health and Addiction, 16*, 631–641.

Auer, M., & Griffiths, M. D. (2020). The use of personalized messages on wagering behavior of Swedish online gamblers: An empirical study. *Computers in Human Behavior, 110*, Article 106402.

Auer, M., & Griffiths, M. D. (2023a). Predicting self-reported problem gambling using machine learning and player tracking data: An empirical study. *Journal of Gambling Studies, 39*(4), 1121–1140.

Auer, M., & Griffiths, M. D. (2023b). Nudging online gamblers to withdraw money: The impact of personalized messages on money withdrawal among a sample of real-world online casino players. *Journal of Gambling Studies, 40*(3), 1227–1244.

Auer, M., & Griffiths, M. D. (2023c). An empirical attempt to operationalize chasing losses in gambling utilizing account-based player tracking data. *Journal of Gambling Studies, 39*(4), 1547–1561.

Auer, M., Malischnig, D., & Griffiths, M. D. (2014). Is “pop-up” messaging in online slot machine gambling effective as a responsible gambling strategy? An empirical research note. *Journal of Gambling Issues, 29*, 1–10.

Auer, M., Hopfgartner, N., & Griffiths, M. D. (2018). The effect of loss-limit reminders on gambling behavior: A real-world study of Norwegian gamblers. *Journal of Behavioral Addictions, 7*(4), 1056–1067.

Auer, M., Hopfgartner, N., & Griffiths, M. D. (2020a). The effects of voluntary deposit limit-setting on long-term online gambling expenditure. *Cyberpsychology, Behavior, and Social Networking, 23*(2), 113–118.

Auer, M., Reiestad, S. H., & Griffiths, M. D. (2020b). Global limit setting as a responsible gambling tool: What do players think? *International Journal of Mental Health and Addiction, 18*, 14–26.

Auer, M., Hopfgartner, N., & Griffiths, M. D. (2021). An empirical study of the effect of voluntary limit-setting on gamblers’ loyalty using behavioural tracking data. *International Journal of Mental Health and Addiction, 19*(6), 1939–1950.

Auer, M., Hopfgartner, N., Helic, D., & Griffiths, M. D. (2024). Self-reported deposits versus actual deposits in online gambling: An empirical study. *Journal of Gambling Studies, 40*(2), 619–637.

Bland, J. M., & Altman, D. G. (2000). The odds ratio. *BMJ (Clinical Research Ed.), 320*(7247), 1468.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159.

Braverman, J., & Shaffer, H. J. (2012). How do gamblers start gambling: Identifying behavioural markers for high-risk internet gambling. *European Journal of Public Health, 22*(2), 273–278.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Wadsworth & Brooks/Cole.

Carran, M. (2022). Consumer protection in European online gambling. Regulation. Monitoring gambling engagement and problem gambling prevalence within selected European jurisdictions. City Law School City, University London. London. Retrieved February 17, 2025, from: https://www.egba.eu/uploads/2022/05/Monitoring-Gambling-Prevalence-Study-01May2022.pdf

Catania, M., & Griffiths, M. D. (2021). Understanding online voluntary self-exclusion in gambling: An empirical study using account-based behavioral tracking data. *International Journal of Environmental Research and Public Health, 18*, 2000.

Connor, B. (2024). The consequences of America’s big bet on sports gambling. *The Regulatory Review*. Retrieved February 17, 2025, from: https://www.theregreview.org/2024/12/10/connor-the-consequences-of-americas-big-bet-on-sports-gambling/

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

Costes, J. M., Kairouz, S., Fiedler, I., Bartczuk, R. P., Lelonkek-Kuleta, B., Minutillo, A., & Notari, L. (2023). Online gambling practices and related problems in five European countries: Findings from the Electronic Gam(bl)ing Multinational Empirical Survey (E-GAMES) Project. *Journal of Gambling Studies*. Advance online publication. https://doi.org/10.1007/s10899-023-10229-8

Dragicevic, S., Percy, C., Kudic, A., & Parke, J. (2015). A descriptive analysis of demographic and behavioral data from internet gamblers and those who self-exclude from online gambling platforms. *Journal of Gambling Studies, 31*(1), 105–132.

Elton-Marshall, T., Leatherdale, S. T., & Turner, N. E. (2016). An examination of internet and land-based gambling among adolescents in three Canadian provinces: Results from the Youth Gambling Survey (YGS). *BMC Public Health, 16*, 477.

European Gaming and Betting Association (2019). Market reality report 2018–2019. Retrieved February 17, 2025, from: https://www.egba.eu/uploads/2019/06/EGBA-Market-Reality-Report-2019.pdf

European Gaming and Betting Association (2022). Annual report 2022: The state of online gambling in Europe. Retrieved February 17, 2025, from: https://www.egba.eu/uploads/2023/02/230203-European-Online-Gambling-Key-Figures-2022.pdf

Ferris, J., & Wynne, H. (2001). The Canadian problem gambling index: Final report. Ottawa: Canadian Centre on Substance Abuse

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gabellini, E., Lucchini, F., & Gattoni, M. E. (2023). Prevalence of problem gambling: A meta-analysis of recent empirical research (2016–2022). *Journal of Gambling Studies, 39*(3), 1027–1057.

Gainsbury, S. M. (2014). Review of self-exclusion from gambling venues as an intervention for problem gambling. *Journal of Gambling Studies, 30*(2), 229–251.

Gainsbury, S. M. (2015). Online gambling addiction: The relationship between internet gambling and disordered gambling. *Current Addiction Reports, 2*(2), 185–193.

Gainsbury, S. M., Russell, A., Hing, N., Wood, R., Lubman, D., & Blaszczynski, A. (2015). How the internet is changing gambling: Findings from an Australian prevalence survey. *Journal of Gambling Studies, 31*(1), 1–15.

Håkansson, A. (2020). Changes in gambling behavior during the COVID-19 pandemic—a web survey study in Sweden. *International Journal of Environmental Research and Public Health, 17*(11), 4013.

Håkansson, A., & Henzel, V. (2021). Who chooses to enroll in a new national gambling self-exclusion system? A general population survey in Sweden. *Harm Reduction Journal, 18*(1), Article 82.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36.

Hayer, T., & Meyer, G. (2011). Internet self-exclusion: Characteristics of self-excluded gamblers and preliminary evidence for its effectiveness. *International Journal of Mental Health and Addiction, 9*(3), 296–307.

Hing, N., Cherney, L., Blaszczynski, A., Gainsbury, S. M., & Lubman, D. I. (2014). Do advertising and promotions for online gambling increase gambling consumption? An exploratory study. *International Gambling Studies, 14*(3), 394–409.

Hing, N., Russell, A. M., Black, A., Rockloff, M., Browne, M., Rawat, V., Greer, N., Stevens, M., Dowling, N. A., Merkouris, S., King, D. L., Salonen, A. H., Breen, H., & Woo, L. (2022). Gambling prevalence and gambling problems amongst land-based-only, online-only and mixed-mode gamblers in Australia: A national study. *Computers in Human Behavior, 132*, Article 107269. https://doi.org/10.1016/j.chb.2022.107269

Hopfgartner, N., Auer, M., Griffiths, M. D., & Helic, D. (2023a). Predicting self-exclusion among online gamblers: An empirical real-world study. *Journal of Gambling Studies, 39*(1), 447-465.5.

Hopfgartner, N., Auer, M., Helic, D., & Griffiths, M. D. (2023b). The efficacy of voluntary self-exclusions in reducing gambling among a real-world sample of British online casino players. *Journal of Gambling Studies, 39*(4), 1833–1848.

Hopfgartner, N., Auer, M., & Griffiths, M. D. (2024). Cross-country prediction of problem gambling using player tracking data: A machine learning approach. *International Journal of Mental Health and Addiction, 22*(1), 237–255.

Humphrey, A., Kuberski, W., Bialek, J., Perrakis, N., Cools, W., Nuyttens, N., Elakhrass, H., & Cunha, P. A. C. (2022). Machine-learning classification of astronomical sources: Estimating F1-score in the absence of ground truth. *Monthly Notices of the Royal Astronomical Society, Letters, 517*(1), L116–L120.

Kim, Y., & Kim, Y. (2022). Explainable heat-related mortality with random forest and SHapley additive explanations (SHAP) models. *Sustainable Cities and Society, 79*, Article 103677.

Langham, E., Thorne, H., Browne, M., Donaldson, P., Rose, J., & Rockloff, M. (2015). Understanding gambling related harm: A proposed definition, conceptual framework, and taxonomy of harms. *BMC Public Health, 16*(1), 80.

Lombardi, G., Molinaro, S., Cotichini, R., Cerrai, S., Scalese, M., & Benedetti, E. (2024). The cards they're dealt: Types of gambling activity, online gambling, and risk of problem gambling in European adolescents. *Social Science & Medicine, 363*, Article 117482.

Louderback, E. R., LaPlante, D. A., Currie, S. R., & Nelson, S. E. (2021). Developing and validating lower risk online gambling thresholds with actual bettor data from a major internet gambling operator. *Psychology of Addictive Behaviors, 35*(8), 921–938.

Luquiens, A., Tanguy, M. L., Benyamina, A., Lagadec, M., Aubin, H. J., & Reynaud, M. (2016). Tracking online poker problem gamblers with player account-based gambling data only. *International Journal of Methods in Psychiatric Research, 25*(4), 333–342.

Mazar, A., Zorn, M., Becker, N., & Volberg, R. A. (2020). Gambling formats, involvement, and problem gambling: Which types of gambling are more risky? *BMC Public Health, 20*, 711.

Mora-Salgueiro, J., García-Estela, A., Hogg, B., Angarita-Osorio, N., Amann, B. L., Carlbring, P., Jiménez-Murcia, S., Pérez-Sola, V., & Colom, F. (2021). The prevalence and clinical and sociodemographic factors of problem online gambling: A systematic review. *Journal of Gambling Studies, 37*, 899–926.

Moreira, D., Azeredo, A., & Dias, P. (2023). Risk factors for gambling disorder: A systematic review. *Journal of Gambling Studies, 39*(2), 483–511.

Murch, W. S., Kairouz, S., Dauphinais, S., Picard, E., Costes, J. M., & French, M. (2023). Using machine learning to retrospectively predict self-reported gambling problems in Quebec. *Addiction, 118*(8), 1569–1578.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96*(1), 3–14.

Percy, C., França, M., Dragicevic, S., & d'Avila Garcez, A. (2016). Predicting online gambling self-exclusion: An analysis of the performance of supervised machine learning models. *International Gambling Studies, 16*(2), 193–210.

Perrot, B., Hardouin, J. B., Thiabaud, E., Saillard, A., Grall-Bronnec, M., & Challet-Bouju, G. (2022). Development and validation of a prediction model for online gambling problems based on players' account data. *Journal of Behavioral Addictions, 11*(3), 874–889.

Van Rossum, G. (2007). Python programming language. Retrieved February 17, 2025, from: https://www.python.org

Tran, L. T., Wardle, H., Colledge-Frisby, S., Taylor, S., Lynch, M., Rehm, J., Volberg, R., Marionneau, V., Saxena, S., Bunn, C., Farrell, M., & Degenhardt, L. (2024). The prevalence of gambling and problematic gambling: A systematic review and meta-analysis. *The Lancet Public Health, 9*(8), e594–e613.

Ukhova, D., Marionneau, V., Nikkinen, J., & Wardle, H. (2024). Public health approaches to gambling: A global review of legislative trends. *The Lancet Public Health, 9*(1), e57–e67.

Wardle, H., & Griffiths, M. D. (2011). Defining the "online gambler": The British perspective. *World Online Gambling Law Report, 10*(2), 12–13.

Wardle, H., & Reith, G. (2021). Gambling-related harms to others: A framework for research and policy. *International Gambling Studies, 21*(2), 155–171.

World Health Organization (2019). International statistical classification of diseases and related health problems (11th revision). World Health Organization.