



Predicting High-Risk Gambling Based on the First Seven Days of Gambling Activity After Registration Using Account-Based Tracking Data

Michael Auer¹ · Mark D. Griffiths²

Accepted: 13 April 2023
© The Author(s) 2023

Abstract

In recent years, several European regulators have introduced mandatory player tracking to identify potentially problematic online gambling. The present study's aim was to investigate the possibility of predicting future high-risk gambling based on a short time window (i.e., the first seven days after the registration for an online gambling site). The authors were given access to a secondary dataset comprising 37,986 gamblers who registered at a European online gambling operator between January 1 and April 30, 2022. The study examined the association between gambling behavior during the first week after registration and high-risk gambling during the first 90 days after registration. A logistic regression model with high-risk gamblers (based on the first three months of gambling data after initial registration) as the dependent variable and age, gender, and the first week's gambling behavior as independent variables explained 40% of the variance. Age, gender, and seven player tracking features from the first week after registration were significant. Machine learning models confirmed the high correlation between the first week of gambling and a high-risk classification during the first three months after registration. The most important features reported by a Random Forest and a Gradient Boost Machine model were the total amount of money deposited, the number of deposits, the amount of money lost, and the average number of deposits per session. The study showed that high-risk gambling during the first three months of a player's lifetime can be predicted very early after registration. These findings suggest that gambling operators should initiate preventive measures (such as limit setting, mandatory play-breaks, personalized messaging) and monitor gambling behavior at a very early stage after a gambler's initial registration.

Keywords Online gambling · Internet gambling · High-risk gamblers · Responsible gaming · Machine learning · Artificial intelligence

✉ Mark D. Griffiths
mark.griffiths@ntu.ac.uk
Michael Auer
m.auer@neccton.com

¹ Neccton GmbH, Vienna, Austria

² International Gaming Research Unit, Psychology Department, Nottingham Trent University, 50 Shakespeare Street, Nottingham NG1 4FQ, UK

Over the past 20 years, there has been a significant increase in internet use including various online activities such as online gambling. Online gambling has been described as more accessible, affordable, anonymous, and convenient than offline gambling (Griffiths, 2003). Furthermore, disinhibition, dissociation, and greater immersion have been described as risk factors that could link online gambling to a higher risk of problem gambling (Griffiths, 2003). King and Barak (1999) also argued that the global nature of the internet, combined with the limited (if not impossible) ability of local governments to effectively regulate or ban online gambling, would have profound psychological and social consequences.

Chóliz (2016) examined the effect of the legalization of online gambling in Spain with a sample of 1277 pathological gamblers in recovery at 26 gambling addiction treatment centers. The author claimed there had been a significant increase in young pathological gamblers since the legalization of online gambling. Based on a survey of 15,023 German individuals, Effertz et al. (2018) estimated that replacing 10% of offline gambling with online gambling would increase the likelihood of being a problematic gambler by 8.8–12.6%. They concluded that this increase was equivalent to 139,322 problem gamblers and €27.24 million per year of additional expenditures in the German health sector (based on the upper 12.6% figure). Since then, Germany has introduced legal online gambling in June 2021.

Many other European countries have legalized online gambling in recent years. Among these are Sweden, The Netherlands, and Spain. In Canada, the province of Ontario has now started to grant online gambling licenses to private operators (iGaming Ontario, 2022). Also, several states in the USA have now introduced legal online gambling (e.g., New Jersey, Kentucky, Nevada). Compared to land-based gambling, online gambling transactions are not anonymous which means that gambling operators know exactly how much gamblers are spending, what games they are gambling on, and when they are gambling. This means that researchers can use online gambling data to gain more insights into gambling behavior and the understanding of problematic gambling.

Two studies which compared self-reported gambling expenditure with actual data from online gambling operators have shown that players often wrongly assess their own gambling (Braverman et al., 2013; Auer & Griffiths, 2017). These two studies showed that regular gamblers often underestimate their losses and overestimate their winnings which gives reason to question the findings of self-report studies.

Several studies have used account-based player tracking data to understand potentially risky gambling behavior and identify problematic gambling. Finkenwirth et al. (2021) compared 2157 Canadian online gamblers with a record of voluntary self-exclusion (VSE) enrolment with 17,526 non-VSE controls, using 20 input variables of gambling behavior. The input variable with the greatest explanatory power was variance in money bet per session. Based on a sample of 25,720 online players from different European countries, Hopfgartner et al. (2022) predicted future VSE. The study found that the odds of future VSE across countries was associated with a (i) higher number of previous voluntary limit changes and self-exclusions, (ii) higher number of different payment methods for deposits, (iii) higher average number of deposits per session, and (iv) higher number of different types of games played. Adding monetary intensity variables such as the amount deposited or lost did not significantly increase the explanatory power of the statistical models.

Ukhov et al. (2021) compared the gambling behavior of online casino players and online sports bettors using a sample of 10,000 online gamblers. They also knew which gamblers opted for a VSE during the study period. They found that the number of cash wagers per active day contributed the most to problem gambling-related exclusion in the case of sports betting, whereas the volume of money spent gambling contributed the most to problem gambling-related exclusion in the case of casino players. The contribution of the volume

of monetary losses per active day was noticeable in the case of both online casino players and online sports bettors. For online casino players, gambling via desktop computers contributed positively to problem-gambling-related exclusion. For online sports bettors, it was more concerning when the individual used mobile devices (e.g., smartphones, tablets). The number of approved deposits per active day contributed to problem-gambling-related exclusion to a larger extent for online sports bettors than online casino players.

Three studies have used artificial intelligence methods to predict self-reported problem gambling (Auer & Griffiths, 2022a, Louderback et al., 2021; Luquiens et al., 2016). Luquiens et al. (2016) collected responses to the Problem Gambling Severity Index (PGSI) from a sample of online poker players ($n = 14,261$). Their responses on the PGSI were correlated with transactional data from the respondents actual gambling. Almost one-fifth of the participants who completed the PGSI were classed as problem gamblers (18%). Gender, age, frequent wagering in a single session, high losses, frequent depositing within a 12-h period, and several other monetary variables were associated with self-reported problem gambling. Louderback et al. (2021) assessed self-reported problem gambling among a sample of online gamblers using the Brief Biosocial Gambling Screen (BBGS). Their goal was to assess thresholds for low-risk gambling. They identified thresholds with respect to wagering volume per month, the percentage of the annual income, monetary loss volume per month, and daily variability in the amount wagered. Auer and Griffiths (2022a) assessed self-reported problem gambling using the PGSI with a sample of 1287 European online gamblers. They applied AI methods to predict self-reported problem gambling based on a number of behavioral features derived from transactional data. The study found that frequent session depositing and frequently depleting the gambling account were most predictive of self-reported problem gambling.

Several other studies which investigated problem gambling have relied on the *PlayScan* problem gambling classification, a commercial player tracking tool (e.g., Challet-Bouju et al., 2020; Forsström et al., 2017, 2022). These studies did not explain in detail how *PlayScan* classifies high-risk gambling other than that is based on gambling behavior such as depositing, wagering, and playing duration. The present authors are not aware of a generally agreed approach to identify problem gambling based on player tracking data.

The Present Study

Several European countries (e.g., Spain, UK, Germany, Denmark, Sweden, The Netherlands) require online gambling operators to monitor player behavior and identify problematic gambling. Online gambling is a competitive market and several studies have found that gamblers continue to gamble with other operators when they have reached a mandatory limit or have self-excluded (Auer and Griffiths, 2022b; Håkansson and Widinghoff, 2020). For that reason, it is important that monitoring algorithms identify potentially problematic gambling as early as possible after gamblers have registered with a particular online gambling operator. Therefore, the present study investigated whether it is possible to identify risky behavioral patterns among online gamblers in the first week after registration that are predictive of future high-risk gambling. This could assist early prevention efforts and tailored responsible gaming measures by online gambling operators.

The authors examined a sample of European online gamblers to study the association between gambling behavior during the first week after registration and high-risk gambling during the first 90 days after registration. There was no specific hypothesis (as the study

was exploratory) other than the investigation of the correlation between the first week of gambling and high-risk gambling during the first three months after the registration. It was anticipated that the findings will be helpful for policymakers and regulators, as well as for online gambling operators.

Method

The authors were given access to an anonymized secondary dataset from a European online casino operator. The raw data contained every game played, every game won, every deposit, and every money transfer from the online gambling account to a player's bank account (i.e., a money withdrawal). Every transaction could be assigned to a single account. Furthermore, gamblers' age, gender, and registration details were known. The dataset comprised player data from January 1 to April 30, 2020 (inclusive).

Study Design

The dataset comprised all gamblers who registered during the aforementioned study period. For each gambler, the gambling behavior during the first seven days after the registration was carried out (see Appendix 1 for a list of all the variables). Apart from two demographic variables (i.e., age and gender), eight variables reflect session specific gambling behavior (e.g., average number of deposits per session, and number of games per session), and five variables report total monetary amounts for the first seven days after registration (e.g., total amount of money deposited, and lost). Only gamblers who had at least one playing session during the first seven days after registration were selected for further analysis.

Problem Gambling Risk Score

The authors wanted to evaluate whether the first week of gambling was predictive of becoming a high-risk gambler sometime during the first 90 days after registration. The online operator which provided the data uses the player tracking tool *mentor* which has been used in previous studies (Auer & Griffiths, 2015, 2020, 2022b). Based on gambling behavior, the system classifies gamblers daily into one of three categories: low-risk, medium-risk, high-risk. It uses a number of metrics such as monetary deposit volume, frequency of deposits, gambling session length, amount of money lost, frequency of gambling, and gambling during the night. The score takes into account up to six months of historical data. However, gamblers can sometimes be classified as a risky gambler the day after they register, given that they also gambled on the day of registration. Such gamblers usually deposit a lot of money, gamble most of the day, place large bets, do not withdraw any winnings, and chase their losses.

For each of the gamblers who registered during the study period and gambled during the first week after registration, a binary target variable was computed. The variable indicated if a gambler became high-risk on any day during the eight days after registration date up until 90 days into the future. Gamblers could have become high-risk at any day during the 90 days after the registration. Gamblers can remain high-risk for any number of days.

Statistical Analysis

A hierarchical logistic regression analysis was used to compute the correlation between demographics as well as gambling behavior and a future high-risk classification. The dependent variable was binary and indicated whether a gambler was classified high-risk at any time between the day after registration and 90 days after the registration. Variables were classified into three groups. Age and being female were the control variables, a second set of variables reflected behavioral features, and a final set of variables reflected monetary intensity features. First, a logistic regression which only included the control variables was carried out. Next, a logistic regression model which included the control variables and behavioral variables was carried out. In order to determine whether the explanatory power improved after including the behavioral variables, a likelihood ratio test (Feder, 1975) was carried out. The monetary intensity variables were added in a third logistic regression model and a likelihood ratio test was carried out between the third and the second model. The model quality was further investigated using Akaike's information criteria (AIC, Bozdogan, 1987). To reduce and prevent multicollinearity among the variables (James et al., 2013), the authors only included variables with a variance inflation factor (VIF) smaller than 10. This threshold was also used by Hopfgartner et al. (2022). The amount of money bet per session and the amount of money won per session were excluded from the analysis based on a VIF greater than 10. Finally, Nagelkerke's R^2 (Nagelkerke, 1991) was reported to assess the goodness of fit for each model. The Nagelkerke R^2 compares the log-likelihood of a model with explanatory variables to the null-model without any explanatory variables. Similar to an R^2 of a linear regression it is between 0 and 1. However, it does not report the percentage of explained variance, it reports the degree of the correlation between the independent variables and the binary dependent variable.

Additionally, two machine learning models, Random Forest (Rigatti, 2017) and Gradient Boost Machine (Doan and Kalita, 2015), were carried out. In contrast to classical statistical methods like logistic regression, machine learning methods use more parameters which can lead to overfitting. This means that models might explain data on which they were trained very well, but not be applicable to new datasets. For that reason, 80% of the data were used to train the models and 20% of the data were used to test the models. Model accuracy is reported based on the test data. The study was approved by the second author's university ethics committee.

Results

A total number of 37,986 gamblers registered between January 1 and April 30, 2022 with the online operator that provided the secondary dataset. Of these, 8657 were female (23%) and the average age was 31 years ($SD=11.62$). Out of the 37,986 gamblers, 2764 became high-risk for at least one day in the 90 days after registration (7.3%). Table 1 reports the mean average values for gamblers who became high-risk and gamblers who did not become high-risk during the 90 days after registration. Gamblers who did not become high-risk were on average 30 years old and gamblers who became high-risk were on average 38 years old. Moreover, 23% of gamblers who did not become high-risk were female and 25% of gamblers who became high-risk were female. Future high-risk gamblers also displayed higher values with respect to every metric carried out during the first seven

Table 1 First seven days' mean values for gamblers who became high-risk vs. gamblers who did not become high-risk

	Not high-risk	High-risk
<i>N</i>	35,222	2764
Age	30	38
Female	23%	25%
Average session lengths in minutes per session	22.35	49.81
Average number of monetary deposits per session	0.24	0.93
Average number of games played per session	108.32	357.8
Number of sessions	4.78	10.73
Number of playing days	2.44	4.35
Number of monetary deposits	0.91	7.45
Average amount of money deposited per session (€)	10.37	160.28
Average amount of money lost per session (€)	13.36	124.33
Average amount of money withdrawn per session (€)	4.07	33.97
Total amount of money deposited (€)	32.55	808.68
Total amount of money lost (€)	23.1	352.96

days after registration. Future high-risk gamblers' gambling sessions lasted an average of 49.81 min and non-high-risk gamblers' gambling sessions lasted on average 22.35 min. Future high-risk gamblers on average deposited €160.28 per session and non-high-risk gamblers on average deposited €10.37 per session.

In order to investigate whether there was a linear or non-linear relationship between age and being a high-risk gambler, the authors classified players into different age bands. There appeared to be a positive correlation between age and the percentage of high-risk gamblers with the largest value appearing among those aged 39–55 years. Those gamblers aged up to 21 years and those aged 22–28 years comprised the lowest percentage of high-risk gamblers. Gamblers older than 56 years had a lower percentage of high-risk gamblers compared to those aged between 39 and 55 years (Table 2).

Appendix 2 shows the correlations between each variable including the high-risk status. There is a correlation of 0.99 between the average amount of money bet per session and the average amount of money won per session. A combination of variance inflation factor analysis and examination of the bivariate correlations led to the exclusion of the average amount of money won per session and the average amount of money bet per session. This can be also explained by the fact that the difference between the amount of money won and amount of money bet is actually the amount of money lost. Variables which are derived

Table 2 Number of high-risk gamblers per age group

Age group	% High-risk	<i>N</i>
<21	1%	9573 (25%)
22–28	4%	10,291 (27%)
29–38	11%	9867 (26%)
39–55	14%	6604 (17%)
>56	11%	1651(4%)

from other variables do not add additional explanatory power, but increase collinearity and therefore add instability to regression models. The number of monetary deposits had the largest correlation with becoming high-risk (0.40), followed by the amount of money deposited (0.35), and the number of playing days (0.27).

A logistic regression model which included age and being female as independent variables and high-risk gambling as a binary dependent variable was carried out. The control model reported a Nagelkerke R^2 of 0.057 which indicates a low correlation between the two demographic control variables and the dependent variable. The AIC of the control model was 18,916.

In the next step, the behavioral variables were added to the logistic regression. The Nagelkerke R^2 was 0.32 which means that there was a substantial correlation between the independent behavioral variables and the dependent variable. The AIC was 14,531. The lower the AIC value, the better the model quality. The likelihood ratio test between the behavioral model and the control model was significant ($\chi^2 = 4397$, $p < 0.001$) which means that the behavioral variables contributed significantly to the model quality. Table 3 reports the coefficients for each independent variable. Except for the average number of games played per session, each independent variable's coefficient was statistically significant. Only being female was negatively correlated with becoming a high-risk gambler.

In the third step, the monetary intensity variables were added to the logistic regression. The Nagelkerke R^2 was 0.40 which means that the inclusion of the monetary intensity variables only slightly increased the accuracy of the logistic regression model. The AIC was 13,227 and therefore lower than for the model without the monetary intensity values. This means that the AIC also confirmed an improved model quality after adding the monetary intensity variables. The likelihood ratio test between the behavioral model and the monetary intensity model was significant ($\chi^2 = 1315$, $p < 0.001$) which means that the monetary intensity variables contributed significantly to the model quality.

Table 4 reports the coefficients for each independent variable in the third logistic regression model. Apart from the average session length in minutes per session, the average number of deposits per session, the average amount of money withdrawn per session, and the total amount of money lost, each independent variable's coefficient was statistically significant. In the multivariate logistic regression model, being female, average amount of money deposited per session, and average amount of money were negatively associated with becoming a high-risk gambler.

Table 3 Coefficients of the second logistic regression model which included the control and the behavioral variables

	Coefficient	Std. error	Z	$p > z $
Intercept	-5.58	0.08	-68.60	<0.001
Age	0.04	0.00	21.39	<0.001
Female	-0.40	0.06	-7.23	<0.001
Average session length in minutes per session	0.01	0.00	10.38	<0.001
Average number of monetary deposits per session	0.16	0.03	4.75	<0.001
Average number of games played per session	0.00	0.00	0.70	0.486
Number of sessions	0.02	0.00	4.73	<0.001
Number of playing days	0.30	0.02	18.23	<0.001
Number of monetary deposits	0.13	0.01	20.35	<0.001

Table 4 Coefficients of the third logistic regression model which included the control, the behavioral and the monetary intensity variables

	β	Exp (β)	Std. error	Z	$p > z $
Intercept	-5.530	0.004	0.08700	-63.24	<0.001
Age	0.033	1.034	0.00200	17.22	<0.001
Female	-0.232	0.793	0.05800	-3.98	<0.001
Average session length in minutes per session	0.002	1.002	0.00100	1.73	0.083
Average number of monetary deposits per session	0.049	1.050	0.04400	1.11	0.266
Average number of games per session	0.000	1.000	0.00010	3.58	<0.001
Number of sessions	0.020	1.020	0.00400	5.09	<0.001
Number of playing days	0.297	1.345	0.01700	17.27	<0.001
Number of monetary deposits	0.049	1.051	0.00800	6.26	<0.001
Average amount of money deposited per session	-0.003	0.997	0.00001	-7.37	<0.001
Average amount of money lost per session	0.000	1.000	0.00003	11.12	<0.001
Average amount of money withdrawn per session	0.001	1.001	0.00001	1.34	0.182
Total amount of money deposited	0.003	1.003	0.00001	18.36	<0.001
Total amount of money lost	0.000	1.000	0.00004	-2.26	0.024

Table 4 also reports the odds ratios (exp (β)) for each independent variable in the logistic regression model. An odds ratio of 1 indicates that the chance of becoming high-risk is not related to an independent variable. An odds ratio of greater than 1 means that the chance of becoming high-risk increases with an increasing value of the independent variable. If a player gambles on one more day during the first seven days after registration, the chance of becoming high-risk increases by 34.5%. An odds ratio of smaller than 1 means that the chance of becoming high-risk decreases with an increasing value of the independent variable. Being female decreased the chance of becoming high-risk by 21.7%.

One of limitation of logistic regression is that the method can only identify linear relationships. It is also non-iterative and it has many assumptions regarding the distribution of the data. Therefore, the authors also carried out a Random Forest as well as a Gradient Boost Machine model. The independent variables and the dependent variable were the same as for the aforementioned third logistic regression model. Figure 1 displays the receiver operating curve (ROC) and the area under the curve (AUC) for the two models based on the test dataset. The Gradient Boost Machine's AUC (0.888) was slightly higher than the Random Forest's AUC (0.867). The most important variables with respect to explanatory power were the total amount of money deposited, the number of deposits, the amount of money lost, and the average number of deposits per session.

Discussion

The present study was carried out in an attempt to identify early patterns of gambling which are predictive of becoming high-risk during the first 90 days after registration. Player tracking data from a sample of 37,986 European online gamblers were used. The average age was 38 years which is in line with samples from other online gambling studies (e.g.,

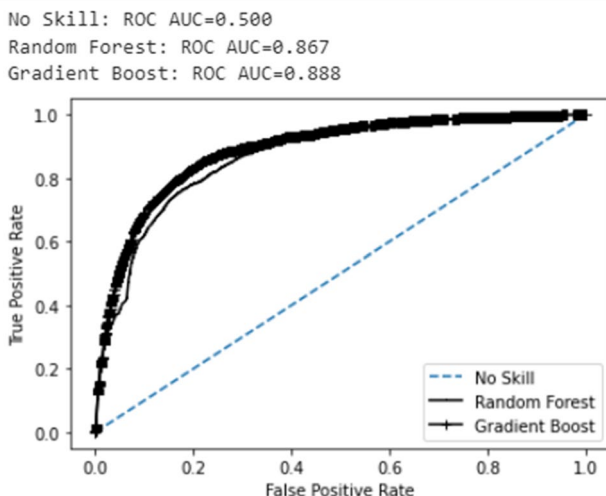


Fig. 1 Goodness of fit statistics for the two machine learning models

Auer & Griffiths, 2022a, 2022b). A univariate analysis found a lower percentage of high-risk gamblers among those aged up to 28 years compared to older gamblers. Various relationships between age and problem gambling have been identified previously. For example, Raisamo et al. (2015) found young age (<25 years) was associated with increased likelihood of reporting harms in nationwide telephone survey which was among 4484 Finns. In a study of 8165 16- to 86-year-old Australians, Abbott et al. (2018) found that the incidence rate for problem gambling among young adults was more than double than that of older adults.

The present study was based on player tracking data and did not assess gambling-related harm using a self-report gambling screen. Young adults were less likely to become high-risk gamblers (i.e., the percentage of high-risk gamblers was lower among younger players and in the regression analysis; and the age coefficient was positive which indicated that older players were more likely to be high-risk gamblers). However, this does not necessarily contradict the aforementioned findings because young adults might perceive lower losses as harmful due to a lower available income. Older gamblers are more likely to have more available income which can lead to higher losses and more frequent high-risk classifications. However, older adults might not perceive the losses as harmful because they do not negatively impact their financial situation.

Slightly more high-risk gamblers were female (25%) than non-high-risk gamblers (23%). In the multivariate logistic regression model, this correlation was reversed which means that being female meant a lower likelihood of being high-risk. The odds ratio for being female indicated that the chance of becoming high risk decreased by 21.7% compared to being male. It is not uncommon that the direction of a correlation is different in a multivariate analysis compared to a univariate analysis because independent variables are often correlated with each other.

Apart from the average session length in minutes per session and the average number of monetary deposits per session, each of the behavioral metrics in the logistic regression was significant. This was also indicated by a Nagelkerke R^2 value of 0.32. The addition

of monetary intensity variables only slightly increased the Nagelkerke R^2 value from 0.32 to 0.40. In the final model, only the average amount of money withdrawn per session and the total amount of money lost were not significant. Several previous studies have identified the number of deposits per session to be highly predictive of problematic gambling (Auer & Griffiths, 2022a, 2022b; Challet-Bouju et al., 2020). The non-significance of the average number of deposits per session in the present study could be related to high correlations between the independent variables. Although highly correlated variables were removed based on the variance inflation factor (VIF), it is still possible that another independent variable which is highly correlated with the average number of deposits per session was responsible for the non-significant correlation of the latter variable. The correlation matrix in Appendix 1 shows a correlation of 0.62 between the average number of monetary deposits per session and the number of monetary deposits. The latter remained in the model and the odds ratio indicated that the chance of becoming high-risk increased by 5.1% with one additional monetary deposit.

The logistic regression model with high-risk gamblers showed that the independent variables explained 40% of the variance with age, gender, and seven player tracking features from the first week after registration were significant. Other factors that were not assessed that may have contributed to high-risk gambling include factors specific to the gamblers themselves (e.g., individual characteristics such as personality factors or biological predispositions) or situational characteristics in the gamblers' environments (e.g., advertising and marketing). The logistic regression model's high explanatory power indicates that it is possible to identify future high-risk gambling in the first week after registration. This was further backed up by machine learning models which report an area under the curve (AUC) of 0.88. This is higher than any other AUC value reported by similar studies that the authors are aware of (e.g., Auer & Griffiths, 2022a; Louderback et al., 2021; Luquiens et al., 2016). This is a consequence of the fact that the variables used for the high-risk classification overlap with the variables measuring the first week's gambling. However, it indicates that first week's gambling can predict future high-risk gambling. Among all the variables, the number of days on which a player gambled during the first week increases the chance of becoming high-risk the most.

Limitations

The present study has a number of limitations. First, although the number of participants was large and representative of those who gambled on the website, the findings were based on a single anonymized secondary dataset from one European online casino operator. Data from different operators might lead to other results which limits the generalizability of the findings. Second, the results might have also been

influenced by the operator's responsible gaming interactions such as personalized messages, mandatory play-breaks, or the regulation in the specific countries (e.g., mandatory deposit limits and maximum bet size). For example, responsible gambling interactions could lower the number of high-risk players or prevent players from becoming high-risk at an early stage. Responsible gaming procedures can also lead to player suspensions which would also lead to a lower number of high-risk players over time. Third, there was no information available on the nationalities of the gamblers. Given there are often cultural differences between gamblers, it is not known if the participants predominantly came from one or two countries or whether the sample was more geographically diverse. Finally, there is a possibility that more than one person might have been gambling using the same account (e.g., a husband and wife). Future replication studies should be conducted with data from different operators with different types of gamblers.

Conclusions

The present study demonstrated that it is possible to predict a future classification of being a high-risk gambler based on the gambler's transactions during the first week after registering on an online gambling site. The metrics which were used to classify players' high-risk gambling overlap to a great extent with the metrics describing the first week's behavior. For that reason, the results of this study also mean that first week's high-risk gambling can foreshadow subsequent high-risk gambling. However, no causal conclusions are made because the study was based on secondary data. A controlled experiment was not possible because players' gambling behavior cannot be influenced in order to determine which variables might lead to future high-risk gambling. The findings of the study will be of interest to many different stakeholder groups of the gambling industry, gambling policymakers, and gambling regulators, as well as other researchers in the gambling studies field. The predictive power of the models in the present study's logistic regression model was high compared to previous studies which predicted VSEs or self-reported problem gambling (Auer & Griffiths, 2022a; Challet-Bouju et al., 2020; Finkenwirth et al., 2021). This means that online gambling operators could identify future high-risk players very early through monitoring metrics such as the amount of money deposited, number of monetary deposits, amount of money lost, and number of monetary deposits per session. The first week's behavior being predictive of future high-risk gambling is in line with findings by previous player tracking studies (e.g., Auer & Griffiths, 2022a; Finkenwirth et al., 2021; Hopfgartner et al., 2022). Such early detection suggests that gambling operators should engage in customer interactions such as e-mails, telephone calls, and/or on-site personal messaging which could prevent the development of high-risk gambling.

Appendix 1

List of variables carried out for the first seven days after registration

Number	Variable	Category
1	Age	Demographic
2	Female	Demographic
3	Average session lengths in minutes per session	Behavioral
4	Average number of deposits per session	Behavioral
5	Average number of games per session	Behavioral
6	Number of sessions	Behavioral
7	Number of playing days	Behavioral
8	Number of monetary deposits	Behavioral
9	Average amount of money deposited per session	Monetary
10	Average amount of money bet per session	Monetary
11	Average amount of money won per session	Monetary
12	Average amount of money lost per session	Monetary
13	Average amount of money withdrawn per session	Monetary
14	Total amount of money deposited	Monetary
15	Total amount of money lost	Monetary

Appendix 2

Correlation matrix of all variables

	Age	Female	Average session lengths in minutes per session	Average number of deposits per session	Average number of games per session	Number of sessions	Number of playing days	Average amount of money deposited per session	Average amount of money lost per session	Average amount of money won per session	Average amount of money drawn per session	Total amount of money deposited	Total amount of money lost	High-risk
Age	1	0.15	0.16	0.07	0.21	0.08	0.1	0.09	0.04	0.07	0.08	0.08	-0.03	0.17
Female		1	0.1	0.05	0.1	0.02	0.04	0.06	-0.01	0	0	0	0	0.01
Average session lengths in minutes per session			1	0.49	0.78	0.03	0.06	0.33	0.15	0.3	0.31	0.22	-0.1	0.14
Average number of deposits per session				1	0.44	-0.03	0.01	0.62	0.31	0.29	0.28	0.37	-0.21	0.12
Average number of games per session					1	0.01	0.07	0.33	0.12	0.23	0.24	0.19	-0.07	0.12
Number of sessions						1	0.79	0.28	-0.01	-0.01	0	0.11	0.04	0.03
Number of playing days							1	0.32	-0.01	0.01	0.02	0.13	0.04	0.07
High-risk														0.23

Table (continued)

Age	Female	Average session lengths in minutes per session	Average number of deposits per session	Average number of games per session	Number of sessions	Number of playing days	Number of monetary deposits	Average amount deposited per session	Average amount of money bet per session	Average amount of money won per session	Average amount of money lost per session	Average amount of money deposited	Total amount of money lost	Total amount of money deposited	High-risk
Number of monetary deposits					1		1	0.15	0.17	0.18	-0.06	0.14	-0.12	0.47	0.4
Average amount of money deposited per session							1	0.75	0.67	-0.87	0.07	0.67	-0.42	0.16	
Average amount of money bet per session								1	0.99	-0.62	0.16	0.61	-0.29	0.25	
Average amount of money won per session									1	-0.5	0.2	0.58	-0.21	0.26	

Table (continued)

Age	Female	Average session lengths in minutes per session	Average number of deposits per session	Average number of games per session	Number of sessions	Number of playing days	Number of monetary deposits	Average amount of money deposited per session	Average amount of money won per session	Average amount of money lost per session	Average amount of money withdrawn per session	Total amount of money deposited	Total amount of money lost	High-risk
								1	0.12	-0.49	0.64			-0.1
Average amount of money lost per session								1	0.19	0.21	0.15			
Average amount of money withdrawn per session												1	-0.37	0.35
Total amount of money deposited												1		-0.1

Data Availability The data for this study are not available due to commercial sensitivity.

Declarations

Ethical Approval Secondary data analysis. Ethical approval was provided by the ethics committee of Nottingham Trent University.

Informed Consent Not applicable. Secondary data analysis.

Conflict of Interest The second author's university has received funding from *Norsk Tipping* (the gambling operator owned by the Norwegian Government). The second author 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 social responsibility in gambling.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abbott, M., Romild, U., & Volberg, R. (2018). The prevalence, incidence, and gender and age-specific incidence of problem gambling: Results of the Swedish longitudinal gambling study (Swelogs). *Addiction, 113*(4), 699–707.
- Auer, M., & Griffiths, M. D. (2015). The use of personalized behavioral feedback for online gamblers: An empirical study. *Frontiers in Psychology, 6*, 1406.
- Auer, M., & Griffiths, M. D. (2017). Self-reported losses versus actual losses in online gambling: An empirical study. *Journal of Gambling Studies, 33*(3), 795–806.
- 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*, 106402.
- Auer, M., Reiestad, S. H., & Griffiths, M. D. (2020). Global limit setting as a responsible gambling tool: What do players think? *International Journal of Mental Health and Addiction, 18*(1), 14–26.
- Auer, M., & Griffiths, M. D. (2022a) Using artificial intelligence algorithms to predict self-reported problem gambling with account-based player data in an online casino setting *Journal of Gambling Studies*. Advance online publication <https://doi.org/10.1007/s10899-022-10139-1>
- Auer, M., & Griffiths, M. D. (2022b). An empirical attempt to operationalize chasing losses in gambling utilizing account-based player tracking data. *Journal of Gambling Studies*. Advance online publication. <https://doi.org/10.1007/s10899-022-10144-4>
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345–370.
- Braverman, J., Tom, M. A., & Shaffer, H. J. (2014). Accuracy of self-reported versus actual online gambling wins and losses. *Psychological Assessment, 26*(3), 865–877.
- Challet-Bouju, G., Hardouin, J. B., Thiabaud, E., Saillard, A., Donnio, Y., Grall-Bronnec, M., & Perrot, B. (2020). Modeling early gambling behavior using indicators from online lottery gambling tracking data: Longitudinal analysis. *Journal of Medical Internet Research, 22*(8), e17675.
- Chóliz, M. (2016). The challenge of online gambling: The effect of legalization on the increase in online gambling addiction. *Journal of Gambling Studies, 32*(2), 749–756.
- Doan, T., & Kalita, J. (2015, November). Selecting machine learning algorithms using regression models. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)* (pp. 1498–1505). IEEE.

- Effertz, T., Bischof, A., Rumpf, H. J., Meyer, C., & John, U. (2018). The effect of online gambling on gambling problems and resulting economic health costs in Germany. *European Journal of Health Economics*, *19*(7), 967–978.
- Feder, P. I. (1975). The log likelihood ratio in segmented regression. *The Annals of Statistics*, *3*(1), 84–97.
- Finkenwirth, S., MacDonald, K., Deng, X., Lesch, T., & Clark, L. (2021). Using machine learning to predict self-exclusion status in online gamblers on the PlayNow.com platform in British Columbia. *International Gambling Studies*, *21*(2), 220–237.
- Forsström, D., Jansson-Fröjmark, M., Hesser, H., & Carlbring, P. (2017). Experiences of PlayScan: Interviews with users of a responsible gambling tool. *Internet Interventions*, *8*, 53–62.
- Forsström, D., Rozental, A., Wiklund, E., Carlbring, P., & Lindner, P. (2022). Gamblers' perception of the PlayScan risk assessment: A mixed-methods study. *Journal of Gambling Studies*, *38*(2), 591–606.
- Griffiths, M. (2003). Internet gambling: Issues, concerns, and recommendations. *CyberPsychology & Behavior*, *6*(6), 557–568.
- Håkansson, A., & Widinghoff, C. (2020). Gambling despite nationwide self-exclusion—A survey in online gamblers in Sweden. *Frontiers in Psychiatry*, *11*, 599967.
- Hopfgartner, N., Auer, M., Griffiths, M. D., & Helic, D. (2022). Predicting self-exclusion among online gamblers: An empirical real-world study. *Journal of Gambling Studies*. Advance online publication. <https://doi.org/10.1007/s10899-022-10149-z>
- iGaming Ontario (2022). Ontario's legal, regulated iGaming market has launched! November 18. Retrieved November 20, 2022, from: <https://igamingontario.ca/en/operator/operators>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.
- King, S. A., & Barak, A. (1999). Compulsive internet gambling: A new form of an old clinical pathology. *CyberPsychology & Behavior*, *2*(5), 441–456.
- 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.
- Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692.
- Raisamo, S. U., Mäkelä, P., Salonen, A. H., & Lintonen, T. P. (2015). The extent and distribution of gambling harm in Finland as assessed by the Problem Gambling Severity Index. *The European Journal of Public Health*, *25*(4), 716–722.
- Rigatti, S. J. (2017). *Random Forest*. *Journal of Insurance Medicine*, *47*(1), 31–39.
- Ukhov, I., Bjurgert, J., Auer, M., & Griffiths, M. D. (2021). Online problem gambling: A comparison of casino players and sports bettors via predictive modeling using behavioral tracking data. *Journal of Gambling Studies*, *37*, 877–897.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.