1828
AKADÉMIAI KIADÓ

Insights into the temporal dynamics of identifying problem gambling on an online casino: A machine learning study on routinely collected individual account data

Journal of Behavioral Addictions

14 (2025) 1, 490–500

DOI:
10.1556/2006.2025.00013

SAM ANDERSSON[1*], PER CARLBRING[2,3], KEENAN LYON[4], MÅNS BERMELL[4] and PHILIP LINDNER[1]

[1] Centre for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet, & Stockholm Health Care Services, Region Stockholm, Stockholm, Sweden

[2] Department of Psychology, Stockholm University, Stockholm, Sweden

[3] School of Psychology, Korea University, Seoul, South Korea

[4] LeoVegas Group, Stockholm, Sweden

Received: October 7, 2024 • Revised manuscript received: January 29, 2025 • Accepted: February 1, 2025
Published online: February 27, 2025

FULL-LENGTH REPORT

ABSTRACT

Background and Aims: The digitalization of gambling provides unprecedented opportunities for early identification of problem gambling, a well-recognized public health issue. This study aimed to advance current practices by employing advanced machine learning techniques to predict problem gambling behaviors and assess the temporal stability of these predictions. *Methods:* We analyzed player account data from a major Swedish online gambling provider, covering a 4.5-year period. Feature engineering was applied to capture gambling behavior dynamics. We trained machine learning models, XGBoost, to classify players into low-risk and higher-risk categories. Temporal stability was evaluated by progressively truncating the training dataset at various time points (30, 60, and 90 days) and assessing model performance across truncations. *Results:* The models demonstrated considerable predictive accuracy and temporal stability. Key features such as loss-chasing behavior and net balance trend consistently contributed to accurate predictions across all truncation periods. The model's performance evaluated on a separate holdout set, measured by metrics like F1 score and ROC AUC, remained robust, with no significant decline observed even with reduced data, supporting the feasibility of early and reliable detection. *Discussion and Conclusions:* These findings indicate that machine learning can reliably predict problem gambling behaviors over time, offering a scalable alternative to traditional methods. Temporal stability highlights their potential for real-time application in gambling operators' Duty of Care. Consequently, advanced techniques could strengthen early identification and intervention strategies, potentially improving public health outcomes by preventing the escalation of harmful behaviors.

KEYWORDS

problem gambling, machine learning, temporal stability, predictive analytics, gambling behavior, public health

INTRODUCTION

Due to the high societal and individual costs associated with problem gambling, early identification is crucial (Eadington, 2003; Hofmarcher, Romild, Spångberg, Persson, & Håkansson, 2020; Jonsson, Abbott, Sjöberg, & Carlbring, 2017). The digitalization of gambling (Jonsson, Munck, Volberg, & Carlbring, 2017) offers unprecedented opportunities to do so, since every login, deposit, bet, and outcome is logged. Once identified, timely interventions can significantly increase the likelihood that individuals are helped before gambling-related harm accumulates (Clune et al., 2024). Existing methods for identifying problem gambling, such as

*Corresponding author.
E-mail: sam.andersson@ki.se

AKJournals

self-report questionnaires and behavioral tracking, have varying degrees of validity and reliability (Edgren et al., 2016; Jonsson, Munck, et al., 2017). Self-report methods depend on individuals accurately reporting their behaviors and experiences. However, these methods can be susceptible to underreporting and bias (Goldstein et al., 2017; Sato & Kawahara, 2011), whereas behavioral tracking requires sophisticated data analytics to interpret effectively (Bitar et al., 2017; Catania & Griffiths, 2021; Haeusler, 2016; Kuentzel, Henderson, & Melville, 2008). Thus, the multifaceted nature of gambling, which involves various psychological, social, and situational factors, makes it challenging to assess with a singular approach (Browne et al., 2017; Hahmann, Hamilton-Wright, Ziegler, & Matheson, 2021). For example, individuals may stop gambling for diverse reasons, including not only harm or financial loss but also personal or strategic considerations (Weatherly, Montes, Peters, & Wilson, 2012).

Much of the existing literature on identification focuses on cross-sectional data (Gainsbury, Sadeque, Mizerski, & Blaszczynski, 2013), which provides only a snapshot of gambling behavior at a single point in time and fails to capture the temporal dynamics by design. While understanding the temporal patterns of gambling behavior is crucial, as it allows for a more accurate identification of problem gambling at different timepoints (Braverman, LaPlante, Nelson, & Shaffer, 2013; Braverman & Shaffer, 2012; Deng, Lesch, & Clark, 2019), it is equally important to consider aggregated behavioral data that captures broader trends and patterns over time. Longitudinal studies, although more complex, offer the potential for deeper insights into the evolution of gambling behavior and the onset of problem gambling (Dowling et al., 2017) and could in theory extend the prediction window, allowing identification of not just current problem gamblers, but also future ones.

With access to player account data, predictive analytics can develop scalable, data-driven methods to identify problem gamblers (Auer & Griffiths, 2022; Perrot et al., 2022). Various machine learning models have shown promise in identifying problem gamblers (Kairouz et al., 2023; Murch et al., 2023; Perrot et al., 2022), revealing that they can leverage complex datasets and scientifically informed feature engineering to identify patterns of gambling behavior related to problem gambling. However, while these studies demonstrate significant potential, they also have limitations. For instance, many existing models often rely on self-reported data, which can be prone to biases and inaccuracies (Percy, França, Dragičević, & d'Avila Garcez, 2016). Moreover, in many predictive studies, researchers often not only utilize cross-sectional data but also frame the prediction problem itself as a cross-sectional analysis, rather than leveraging longitudinal or retrospective data windows (Paterson, Taylor, & Gray, 2020). This approach, particularly in how data is aggregated and features are engineered, often overlooks the temporal richness inherent in the raw data (Suzuki, Nakamura, Inagaki, Watanabe, & Takagi, 2019), potentially skewing the results towards recent data points while missing out on longer-term trends and broader progressions over time (Park, Eom, Seo, & Choi, 2020).

The Swedish Gambling Act, mandates counteracting excessive gambling through continuous monitoring of gambling behavior (Swedish Gambling Act, 2018). Whether the Duty of Care should extend to predictive analytics that foresee problematic patterns before they fully develop remains to be thoroughly examined and empirically validated. This study aims to enhance understanding of the temporal dynamics in identifying problem gambling by applying advanced machine learning methods focused on predicting manual assessments and evaluating the temporal stability of these predictions through truncating the training set at various time points. Our approach leverages aggregated data to capture broader behavioral indicators, ensuring comprehensive analysis and improved prediction accuracy. By transitioning from monitoring to proactive prediction, our research enables gambling operators to implement timely interventions to prevent the escalation of problem gambling behaviors. Such advancements align with legislative frameworks and could significantly improve public health outcomes by reducing gambling-related harms through early prediction.

METHODS

Participants

We utilized player account data from one of Sweden's largest licensed online gambling providers, covering 4.5 years from January 1, 2019 (at which point Sweden switched to a licensed gambling market), to July 1, 2023. The dataset included extensive behavioral and transactional details for $n = 35{,}048$ unique, authenticated players, all of whom are based in Sweden, allowing for a comprehensive analysis of gambling behaviors within this specific context.

Measures

Data preprocessing and feature engineering. All data preprocessing and analyses were conducted using Python (3.11); the fully reproducible code is available online (https://github.com/SamAndersson-C/temporal-dynamics-problem-gambling). We performed extensive feature engineering on raw data consisting of 11 data frames, which included information on bets, transactions, sessions, demographics, payments, responsible gambling actions and predictions, manual risk assessments, and multiple accounts. Using SQL scripts, we combined the data shards into raw tables within a PostgreSQL database. As in past research (Hopfgartner, Auer, Griffiths, & Helic, 2022, 2024) features were derived to reflect various aspects of online gambling behavior, such as loss chasing, betting frequency, session lengths, and spending patterns. Accurate alignment of all tables was crucial due to the granularity of the timestamps, ensuring meaningful feature engineering (Wang et al., 2009) and to capture the evolution of gambling behaviors and detect significant changes or trends, we ensured temporal alignment of data tables and took specific care to avoid data leakage (using information from outside the training set in model training). We enhanced performance through indexing, partitioning, and query optimization. By precisely aligning and securely managing the data, we prevented inadvertent leakage that could produce overly optimistic results. All feature aggregations strictly used activity data up to each labeling date (Fig. 1).

Nominal variables were numerically coded to facilitate modeling. Features with more than 50% missing values were excluded, while others underwent median imputation to preserve data integrity. Heavily skewed features with an excess of zero values were log-transformed to achieve a more normalized distribution. To evaluate the temporal stability of our predictions, we implemented a temporal division for training and test sets, reserving the final year's data for testing. This ensured the model was evaluated on unseen data, simulating a real-world deployment scenario (Barros, Nascimento, Guedes, & Monsueto, 2023). We employed a data truncation strategy based on timestamps from the gambling operator's raw data to further assess temporal stability. Using June 1, 2022, as a general reference, we truncated each player's data by removing records 30, 60, and 90 days prior to their maximum timestamp in the training set. This resulted in three distinct datasets: 30-day, 60-day, and 90-day truncated data, each undergoing the same feature engineering for model training and evaluation. This allowed us to analyze how model performance varies with different amounts of historical data, providing insights into the temporal stability of the predictions over varying time horizons.

Labeling. The primary label indicating customer risk was derived from manual assessments conducted by the

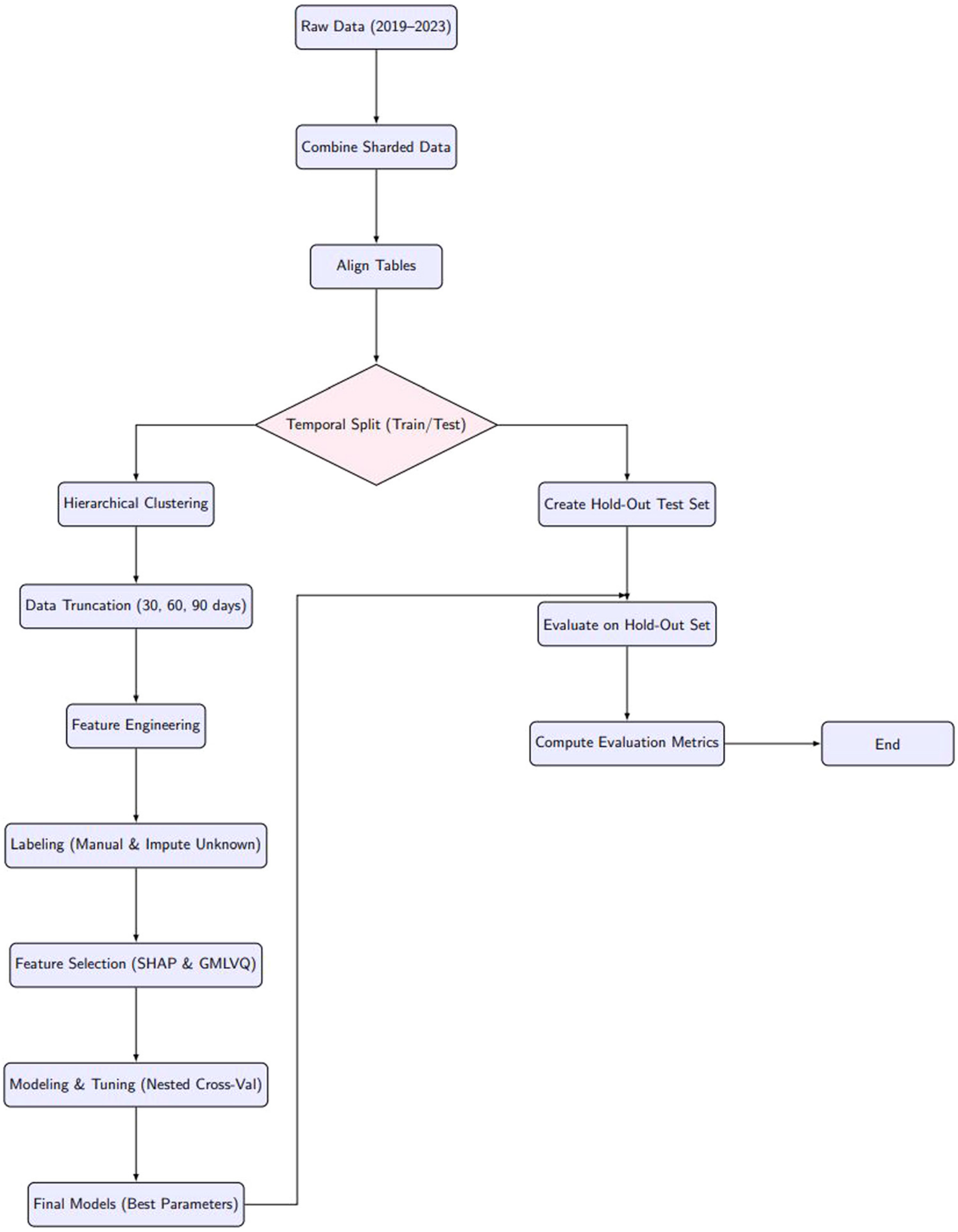
Raw Data (2019–2023)
Combine Sharded Data
Align Tables
Temporal Split (Train/Test)
Hierarchical Clustering
Data Truncation (30, 60, 90 days)
Feature Engineering
Labeling (Manual & Impute Unknown)
Feature Selection (SHAP & GMLVQ)
Modeling & Tuning (Nested Cross-Val)
Final Models (Best Parameters)
Create Hold-Out Test Set
Evaluate on Hold-Out Set
Compute Evaluation Metrics
End

Fig. 1. Data pre-processing and analysis pipeline

gambling provider as part of their Responsible Gambling operations as per their Duty of Care commitment (Cisneros Örnberg & Hettne, 2018). These assessments targeted players exhibiting concerning gambling behaviors, classifying them into five risk levels based on deposit patterns, session length, denied transactions, and responsible gambling tool use. Higher-risk cases involved persistent high deposits, prolonged play, or self-reported loss of control. Additionally, certain customers were flagged for manual review by the support team if communication raised concerns. Since assessments focused on flagged individuals rather than a random sample, the study population primarily reflects at-risk players rather than all gamblers, congruent with the aim of prediction algorithms in this context. A database logging error was discovered during the analysis, revealing that most of the manually assessed labels were concentrated at the beginning of the dataset's time frame, with some customers (n = 5,848) being flagged with "unknown risk" as their label. These were accounts that the operator's RG analysts began to review but could not complete due to unsuccessful attempts to contact the individual in question, resulting in the suspension of the review. These accounts were temporarily restricted from gambling until the company could establish contact with the individuals, allowing them to complete the review process in accordance with the Responsible Gambling (RG) procedures. These customers were included in the training data if they had a corresponding risk label from the RG prediction table on the same date as the "unknown risk" label (n = 1,844). If such a label was available, we replaced the "unknown risk" with the corresponding RG prediction label. Subsequent customers labeled as "unknown risk" without a corresponding RG prediction label were discarded from the training data (n = 4,004) while all "unknown risk" customers were discarded from the hold-out set (n = 1,902). This approach aimed to fill the gaps in the training data, thereby enhancing the dataset's coverage and the robustness of the predictive models and providing us with a more comprehensive and temporally distributed training dataset.

We acknowledge that this imputation method, while beneficial, does introduce some potential noise into the model. However, this noise can have a regularizing effect on the complex model. Since the imputation was applied only to the training data, we avoided any potential data leakage. Without imputation, large temporal gaps between manually assessed labels could have led to overfitting on sparse data patterns. By filling these gaps, we provided a more continuous and diverse set of training examples, helping the model generalize better to unseen data. This regularizing effect reduced overfitting, enhancing the robustness and reliability of the model's predictions.

Initially, customers were categorized into six risk levels, creating a multi-class classification problem. However, preliminary models performed poorly, as fine-grained classification can introduce unnecessary complexity and variance (Blanco, Perez-de-Viñaspre, Pérez, & Casillas, 2020; Elyan & Gaber, 2017). To gain deeper insight into risk labels and reduce the number of categories, we first generated average SHAP (SHapley Additive exPlanations) (Lundberg, Allen, & Lee, n.d.; Ukhov, Bjurgert, Auer, & Griffiths, 2021) value plots based on our preliminary models. From these plots, we selected the top 25 most influential variables and used them in a hierarchical clustering analysis with the complete linkage method, known for its robustness to noise and outliers (Laurikkala & Juhola, 2001). Euclidean distance was used; single linkage was also considered but found unsuitable due to data noise. This analysis revealed that the data naturally clustered into two groups: low-risk and higher-risk. This supported our decision to binarize the labels into low risk and all other risk levels. This binary framework allowed us to focus on distinguishing low-risk customers from those at elevated risk, aligning with responsible gambling objectives.

Procedure

Feature selection. After initial feature engineering in SQL, we conducted feature selection using SHAP values (Lundberg et al., n.d.; Ukhov et al., 2021) and Generalized Matrix Learning Vector Quantization (GMLVQ) (Lövdal & Biehl, 2024) to identify the most relevant features for the binary classification task. SHAP values decompose a model's prediction for an individual instance into contributions from each feature, providing local and consistent explanations. They ensure that the sum of SHAP values equals the difference between the model's prediction for that instance and the average prediction over the dataset, making them useful for interpreting complex models with clear, additive feature contributions.

In parallel, we applied GMLVQ, a supervised learning technique designed to enhance the discriminative power of features by optimizing a relevance matrix. GMLVQ adjusts the feature space to maximize the margin between classes, which is crucial for effectively distinguishing between classes. GMLVQ assigns different levels of relevance to each feature, thereby improving the model's ability to focus on the most discriminative features for accurate predictions. This approach not only aids in classification but also provides a way to interpret the contribution of each feature to the decision boundaries defined by the model.

To ensure equal contribution of all features during model training, we scaled them using a standard scaler. We reduced redundancy by calculating a correlation matrix and removing one feature from each pair of highly correlated features (Yu & Liu, 2003). Subsequently, we trained a model on the scaled training dataset: XGBoost (Chen & Guestrin, 2016). SHAP values were computed to evaluate the importance of each feature in the prediction process.

To finalize the feature selection, we combined the top 25 features identified by each method. Choosing 25 features was a heuristic decision to balance model complexity and interpretability. This subset size ensured the final models were both accurate and interpretable, maintaining manageable complexity. By merging the most informative and discriminative features from both methods, we created a comprehensive and optimized feature set for the classification task.

Statistical analysis

We used XGBoost to classify customers into low-risk and higher-risk categories. Comprehensive hyperparameter tuning was conducted using Optuna (Akiba, Sano, Yanase, Ohta, & Koyama, 2019), an automated optimization framework, to ensure the model's accuracy and generalizability across different datasets. We focused on optimizing the F1 score to balance precision and recall, exploring hyperparameters such as learning rate, number of estimators, maximum tree depth, subsampling ratio, column sampling ratio, and regularization parameters. We also optimized a probability threshold for converting predicted probabilities into binary classifications. To respect the chronological structure of the data during model selection and avoid any leakage from future observations, we employed a nested forward-chaining cross-validation procedure. Specifically, we sorted all training instances by date and split them into a 5-fold outer loop using a time-series split, ensuring that each validation fold came strictly after the training folds in time. Within each outer training fold, we performed a 3-fold time-series split in an inner loop to refine hyperparameters, again preserving the temporal order. This approach minimized overfitting and ensured that each step of hyperparameter tuning, and model evaluation respected the temporal sequence of events.

We ran up to 1,000 Optuna trials, training XGBoost with different hyperparameters and selecting the best based on average F1 score across inner cross-validation folds. Using these optimal parameters, we trained final models on the full dataset and each truncation period (30-day, 60-day, 90-day) before evaluating them on a hold-out test set (the unused data) to assess real-world generalizability.

Predicted probabilities were converted into binary predictions using the optimized threshold, and performance metrics—including F1 score, ROC AUC, precision, recall, accuracy, and confusion matrices—were computed. To assess the stability of predictions across different amounts of historical data, we repeated this process for each truncation period (30-day, 60-day, 90-day, and full) and compared performance metrics. Finally, we applied linear regression to these metrics to identify trends as the amount of data decreased and used bootstrapping to compute confidence intervals for the slopes, determining whether changes in performance were statistically significant over time.

In addition to classification, we conducted a regression analysis to predict continuous risk scores, providing a more granular understanding of the model's predictive capabilities. We used XGBoost as a regressor to predict risk scores on a continuous scale, which were subsequently categorized into low, medium, and high-risk levels.

Ethics

The study procedures were carried out in accordance with the Declaration of Helsinki. The study was reviewed and approved by the Swedish Ethical Review Authority (Dnr 2023-07288-02). Informed consent was waived by the review board to permit research on pre-existing registry data.

RESULTS

Predictions of problem gambling exhibited considerable temporal stability, even with progressively truncated data. Across all truncation periods (30-day, 60-day, 90-day, and full data), "loss chasing behavior weekly log transformed," "net balance trend," "max deposit log transformed," "session sum p25," and "total bets daily log transformed" consistently had the highest SHAP values, indicating a strong influence on the model's predictions (Fig. 2). As shown in Fig. 3, hold-out set metrics improved slightly with more data, suggesting larger datasets enhance generalization and decision boundaries—particularly in identifying true positives. Overall, model performance was modest yet consistent (Table 1).

A bootstrap analysis of linear slopes across truncation periods (Full → 30-day → 60-day → 90-day) found no significant trend for most metrics, as their 95% confidence intervals included zero: Accuracy [−0.009, 0.031], Recall (Sensitivity) [−0.018, 0.095], F1 Score [−0.008, 0.035], and ROC AUC [−0.008, 0.008]. However, Precision (PPV) had a 95% CI entirely below zero [−0.005, −0.001], indicating a consistently negative slope with increasing truncation. Practically, Precision dropped slightly when moving from full to truncated data. Despite this, overall model performance remained stable across all truncation periods.

We used a regression model to predict continuous risk scores—grouped as low, medium, and high risk—to evaluate performance by risk level (Table 2 and Fig. 4). The model performed well for medium- and high-risk categories, with predicted means closely matching true means. For instance, in the 30-day dataset, the medium-risk group's actual mean was 0.618 vs. a predicted mean of 0.500, and the high-risk group's actual mean was 0.761 vs. 0.758. In the 60-day dataset, the high-risk group's actual mean was 0.765 vs. 0.755. However, the model consistently underestimated risk for the low-risk category in every dataset: for example, in the 60-day dataset, the low-risk group's true mean was 0.529 vs. a predicted mean of 0.248, and in the full dataset, 0.557 vs. 0.219. Thus, the model effectively identifies medium- and high-risk individuals but struggles to accurately capture low-risk cases.

Figure 4 shows better performance in medium and high-risk groups, with smaller gaps between true and predicted means, whereas the model underestimated risk in the low-risk group (the gap increased with longer truncation). This was most pronounced in the full dataset, where the difference for low-risk cases reached 0.337, compared to 0.188 for medium risk and 0.057 for high risk.

DISCUSSION

The results suggest that machine learning predictions of problem gambling, assessed manually or through proxy measures, show relative stability over time, with time being intrinsically linked to data amount. This indicates that early predictions are consistent and reliable, highlighting our

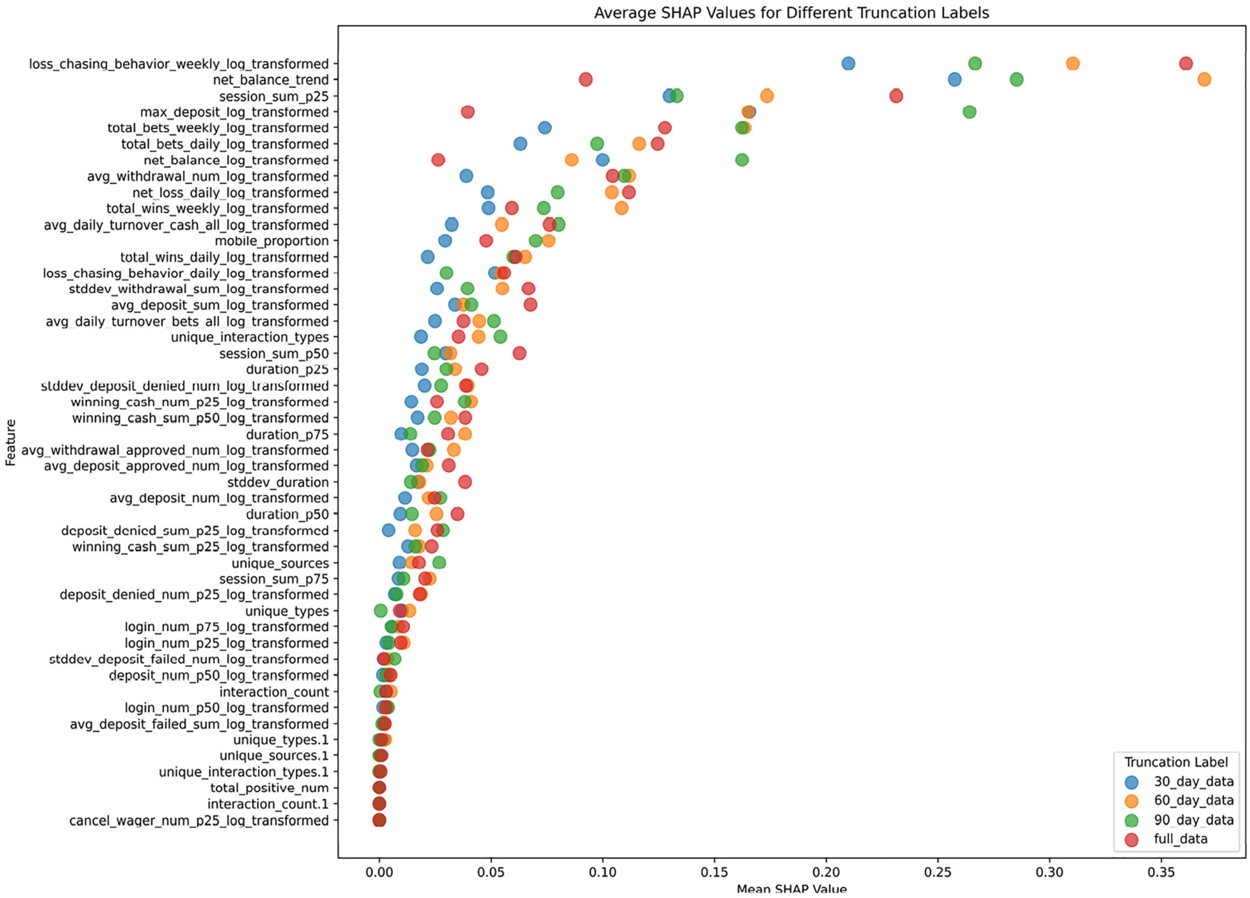

Fig. 2. Feature importance plot

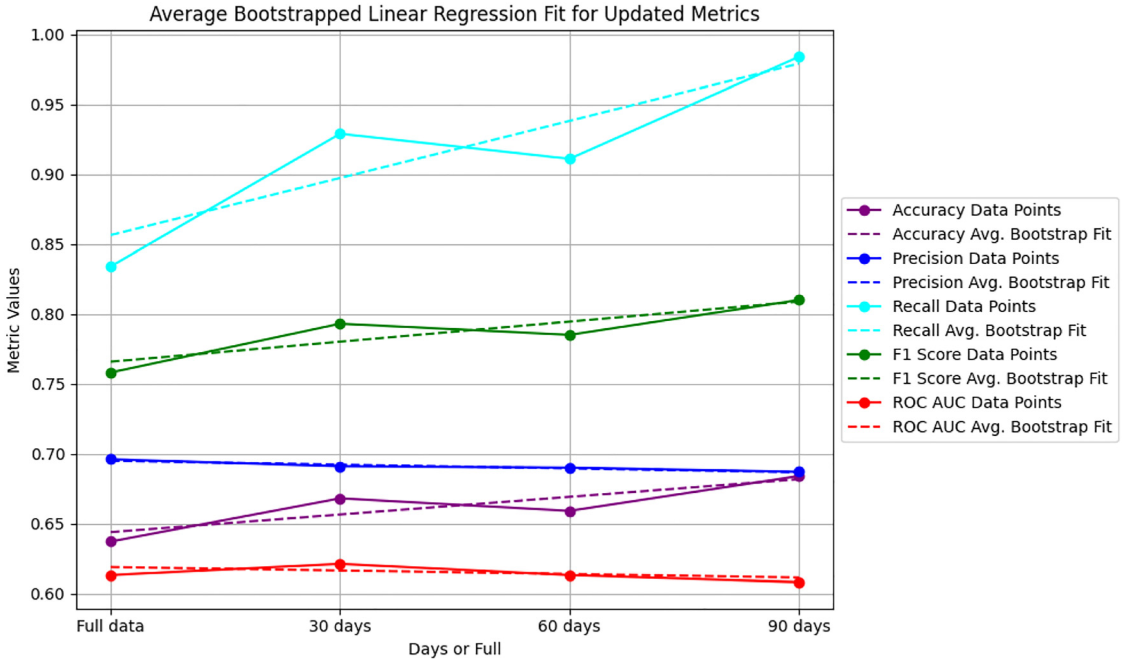

Fig. 3. Temporal evaluation and prediction stability

Table 1. Model performance metrics for different truncation labels

Truncation	Accuracy	Precision (PPV)	Recall (Sensitivity)	F1 Score	ROC AUC	Specificity	NPV
30-day-truncated-data	0.668	0.691	0.929	0.793	0.621	0.107	0.411
60-day-truncated- data	0.659	0.690	0.911	0.785	0.613	0.117	0.378
90-day-truncated- data	0.684	0.687	0.984	0.810	0.608	0.036	0.515
Full data	0.637	0.696	0.834	0.758	0.613	0.213	0.374

Table 2. Risk category prediction table with difference

Dataset	Risk Category	True Mean	Predicted Mean	Difference
30-day-truncated-data	Low Risk	0.600	0.242	0.358
30-day-truncated-data	Medium Risk	0.618	0.500	0.118
30-day-truncated-data	High Risk	0.761	0.758	0.003
60-day-truncated- data	Low Risk	0.529	0.248	0.281
60-day-truncated- data	Medium Risk	0.625	0.497	0.127
60-day-truncated- data	High Risk	0.765	0.755	0.010
90-day-truncated- data	Low Risk	0.588	0.255	0.333
90-day-truncated- data	Medium Risk	0.609	0.538	0.072
90-day-truncated- data	High Risk	0.771	0.724	0.047
Full data	Low Risk	0.557	0.219	0.337
Full data	Medium Risk	0.646	0.458	0.188
Full data	High Risk	0.779	0.722	0.057

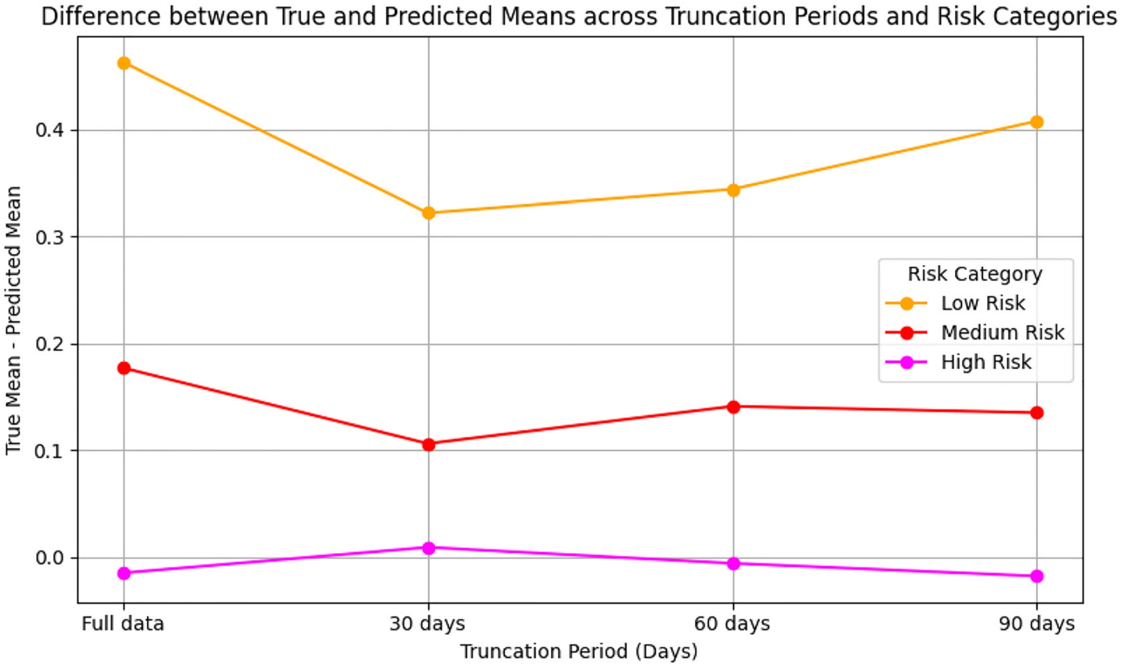

Fig. 4. Difference between true and predicted means

model's robustness. Our claims are based on the model's performance on a holdout validation set. By reserving a full year of data for validation, we evaluated the model on unseen data, mimicking real-world conditions for Duty of Care obligations. Our findings confirm that predictive analytics and machine learning are promising in identifying problem gamblers (Auer & Griffiths, 2022; Deng et al., 2019; Perrot et al., 2022), validating the effectiveness of these methods in a temporally robust manner. Metrics like ROC AUC and F1 score remained consistent across data truncation levels, indicating model reliability. Bootstrapping showed no significant slopes for Accuracy, Recall, F1, and ROC AUC, but Precision exhibited a slight, consistently negative slope from full data to 30-, 60-, and 90-day truncations. Despite this, overall performance stayed relatively stable. Unlike preliminary analysis based on training data and time series cross-validation, the holdout evaluation did not show a decline in performance metrics with the full dataset; instead, metrics such as recall and F1 score improved with increased dataset size, underscoring the importance of using a separate validation set for an accurate reflection of the model's true performance. Therefore, our methods avoid the limitations of traditional approaches like self-report questionnaires and simple behavioral tracking, which often suffer from validity and reliability issues (Edgren et al., 2016; Hodgins & Makarchuk, 2003; MacKillop, Anderson, Castelda, Mattson, & Donovick, 2006). Our machine learning approach offers a more reliable and scalable solution. The model consistently demonstrates reliable performance across different truncation periods, with SHAP values clarifying which features drive its predictions. This highlights the model's ability to effectively interpret complex behavioral data that traditional methods might not capture.

Finally, studies relying on cross-sectional data inherently struggle to capture the temporal dynamics of gambling behavior or (Castrén, Kontto, Alho, & Salonen, 2018; Gainsbury et al., 2013; Paterson et al., 2020). Our study addresses this gap by evaluating the temporal stability of predictions. The consistent importance of key features across different truncation periods, as shown by SHAP values and performance metrics, underscores this stability. This is critical for developing models that can accurately predict problem gambling over extended periods, enhancing our understanding of gambling behavior dynamics.

The findings have practical implications for early identification and intervention in problem gambling. The stability of predictions supports the timely implementation of preventive measures, which can mitigate the risks associated with problem gambling and aid stakeholders in developing effective public health monitoring and intervention programs (Jonsson, Munck, Hodgins, & Carlbring, 2023).

Limitations

This study has several limitations. First, inconsistent application of risk labels over time may cause the model to capture temporal biases rather than genuine risk patterns, especially in dynamic environments like gambling where user behavior and risk profiles can change rapidly. The presence of "unknown risk" labels lead to an imbalanced dataset, underrepresenting certain risk categories and potentially skewing the model's learning process toward more prevalent categories. Our imputation strategy—filling gaps with responsible gambling (RG) prediction labels —aimed to mitigate this by improving the quality and quantity of labeled training data. This approach increased the number of labeled data points and ensured a more uniform temporal distribution, allowing the models to learn from a broader and more representative sample. While this enhanced dataset reduced the risk of overfitting and increased generalizability, inherent imbalances may still pose challenges. Importantly, the hold-out validation data did not suffer from this limitation.

Second, potential bias introduced by manual assessments used for labeling must be acknowledged. Analysts' subjective judgments could have impacted the consistency and accuracy of the labels. Despite this potential bias, manual assessments are generally considered more reliable than self-assessments, which are often prone to inaccuracies (either deliberate or indeliberate) and inconsistencies.

Third, our truncation strategy intended to ensure temporal stability by focusing on consistent windows of activity. However, it may have inadvertently caused accounts with the most cumulative activity to contribute disproportionately to the predictions. Initially, we attempted to use accounts with 30, 60, or 90 days of total activity, but too few accounts met these criteria for meaningful model training. Consequently, we opted for an activity truncation strategy as a compromise, including enough data points for model training but possibly biasing the model toward accounts with more extensive histories.

Fourth, our dataset comes from a single gambling operator in a competitive market, and does not include any given gamblers' activity at other operators. Problem gamblers are typically more likely to gamble with multiple operators. Incomplete behavioral histories can lead to underestimation or misclassification of certain gambling behaviors and limit the broader applicability of our findings. Ideally, a "single customer view" mechanism—aggregating data from multiple operators—would yield more comprehensive insights and potentially more accurate predictive models. In lieu of a centralized system for sharing account-tracking data across operators, operator-specific predictions remain the pragmatic approach to minimizing gambling harms.

Lastly, although the analysis uses a robust setup—temporal holdout splits and nested cross-validation—the limited bootstrapping approach (four samples per metric) may reduce sensitivity to subtle trends. Even so, narrow confidence intervals suggest stable performance metrics over time, indicating temporal consistency. Future research with larger samples or alternative methods could further validate these findings.

Future research directions

Our findings suggest several avenues for future research. One key area is determining the optimal data window for reliable predictions, balancing data sufficiency with model performance. Exploring other machine learning techniques or refining labeling methods could further enhance accuracy. Validating the model with different datasets or in varied contexts will improve its generalizability and robustness.

To improve predictive capabilities, gambling operators should routinely collect relevent features reflecting various risk levels of problem gambling—beyond purely transactional data. This might include browsing patterns, time spent on different site areas, or engagement with specific features. Like how physical casinos observe customer behavior on the floor, incorporating such behavioral indicators online could enhance the model's ability to identify at-risk individuals more accurately.

CONCLUSIONS

This study demonstrates the value of advanced machine learning techniques and rigorous methodologies in gambling research. Our findings show stable long-term prediction performance, evidenced by consistent metrics across different truncation periods. This supports the feasibility of early detection and timely interventions, underscoring the importance of methodological rigor in developing reliable predictive models. These results have significant implications, providing a strong foundation for further research and development in the field.

Funding sources: This study was funded by the LeoVegas Group, a licensed gambling operator in Sweden.

Authors' contributions: SA conceived the analysis pipeline, designed the methods, performed all statistical analyses, developed and executed the modeling and feature engineering, built the database, and drafted the manuscript. PL, as the main supervisor, was responsible for project organization and oversight, providing critical revision of the manuscript and engaging in discussions regarding study concept and methodological and analysis approaches particularly around temporal stability. PC, as co-supervisor, contributed to the study concept alongside PL and provided feedback during critical revision of the manuscript. KL, a data scientist at LeoVegas, supervised the data science aspects, provided technical input, facilitated data transfer, and participated in discussions on machine learning and statistical methods. MB, as head of data science at LeoVegas, contributed to data acquisition, addressed data-related queries, and provided expertise on study feasibility, technical aspects, and conceptual input. All authors were part of the steering group, which met monthly and was organized by PL. SA and PL had full access to all data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. All authors approved the final manuscript.

Conflict of interest: This study is part of an industry-academia collaboration on Responsible Gambling, financed by the LeoVegas Group, a licensed gambling operator in Sweden. The research was planned, performed and submitted under full academic freedom, guaranteed per a written agreement. The funders had no role in the design or execution of the study, nor the decision to publish.

SA's doctoral position is financed by the LeoVegas Group but is an employee of Karolinska Institutet and reports no other potential conflicts of interest. PL and PC reports past and ongoing industry-academia collaborations with several gambling providers, including project-specific funding, but have no personal ties to the gambling industry, financial or otherwise. MB is employed by the LeoVegas Group.

Generative AI use: During the preparation of this work, the corresponding author utilized ChatGPT to enhance writing style and correct grammar and spelling. Following the use of this tool, the authors thoroughly reviewed and edited the content as required, taking full responsibility for the final content of the publication. The AI's role was limited to enhancing clarity, coherence, and presentation of the manuscript.

Acknowledgements: We would like to acknowledge Prof. Ion Petre at the University of Turku, whose lecture series "Foundations of Machine Learning I-III" inspired much of the analysis pipeline used in this study.

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. http://arxiv.org/abs/1907.10902.

Auer, M., & Griffiths, M. D. (2022). Predicting limit-setting behavior of gamblers using machine learning algorithms: A real-world study of Norwegian gamblers using account data. *International Journal of Mental Health and Addiction*, *20*(2), 771–788. https://doi.org/10.1007/s11469-019-00166-2.

Barros, B. de M., Nascimento, H. A. D. do, Guedes, R., & Monsueto, S. E. (2023). Evaluating splitting approaches in the context of student dropout prediction. https://arxiv.org/abs/2305.08600.

Bitar, R., Nordt, C., Grosshans, M., Herdener, M., Seifritz, E., & Mutschler, J. (2017). Telecommunications network measurements of online gambling behavior in Switzerland: A feasibility study. *European Addiction Research*, *23*(2), 106–112. https://doi.org/10.1159/000471482.

Blanco, A., Perez-de-Viñaspre, O., Pérez, A., & Casillas, A. (2020). Boosting ICD multi-label classification of health records with contextual embeddings and label-granularity. *Computer Methods and Programs in Biomedicine*, *188*, 105264. https://doi.org/10.1016/j.cmpb.2019.105264.

Braverman, J., LaPlante, D. A., Nelson, S. E., & Shaffer, H. J. (2013). Using cross-game behavioral markers for early identification of high-risk internet gamblers. *Psychology of Addictive Behaviors*, *27*(3), 868–877. https://doi.org/10.1037/a0032818.

Braverman, J., & Shaffer, H. J. (2012). How do gamblers start gambling: Identifying behavioural markers for high-risk internet gambling. *The European Journal of Public Health*, *22*(2), 273–278. https://doi.org/10.1093/eurpub/ckp232.

Browne, M., Rawat, V., Greer, N., Langham, E., Rockloff, M., & Hanley, C. (2017). What is the harm? Applying a public health methodology to measure the impact of gambling problems and harm on quality of life. *Journal of Gambling Issues*, *36*. https://doi.org/10.4309/jgi.v0i36.3978.

Castrén, S., Kontto, J., Alho, H., & Salonen, A. H. (2018). The relationship between gambling expenditure, socio-demographics, health-related correlates and gambling behaviour—a cross-sectional population-based survey in Finland. *Addiction*, *113*(1), 91–106. https://doi.org/10.1111/add.13929.

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*(4), 2000. https://doi.org/10.3390/ijerph18042000.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. https://doi.org/10.1145/2939672.2939785.

Cisneros Örnberg, J., & Hettne, J. (2018). The future Swedish gambling market: Challenges in law and public policies. In *Gambling Policies in European welfare states* (pp. 197–216). Springer International Publishing. https://doi.org/10.1007/978-3-319-90620-1_11.

Clune, S., Ratnaike, D., White, V., Donaldson, A., Randle, E., O'Halloran, P., & Lewis, V. (2024). What is known about population level programs designed to address gambling-related harm: Rapid review of the evidence. *Harm Reduction Journal*, *21*(1). https://doi.org/10.1186/s12954-024-01032-8.

Deng, X., Lesch, T., & Clark, L. (2019). Applying data science to behavioral analysis of online gambling. *Current Addiction Reports*, *6*(3), 159–164. https://doi.org/10.1007/s40429-019-00269-9.

Dowling, N. A., Merkouris, S. S., Greenwood, C. J., Oldenhof, E., Toumbourou, J. W., & Youssef, G. J. (2017). Early risk and protective factors for problem gambling: A systematic review and meta-analysis of longitudinal studies. *Clinical Psychology Review*, *51*, 109–124. https://doi.org/10.1016/j.cpr.2016.10.008.

Eadington, W. R. (2003). Measuring costs from permitted gaming: Concepts and categories in evaluating gambling's consequences. *Journal of Gambling Studies*, *19*(2), 185–213. https://doi.org/10.1023/A:1023681315907.

Edgren, R., Castrén, S., Mäkelä, M., Pörtfors, P., Alho, H., & Salonen, A. H. (2016). Reliability of instruments measuring at-risk and problem gambling among young individuals: A systematic review covering years 2009–2015. *Journal of Adolescent Health*, *58*(6), 600–615. https://doi.org/10.1016/j.jadohealth.2016.03.007.

Elyan, E., & Gaber, M. M. (2017). A genetic algorithm approach to optimising random forests applied to class engineered data. *Information Sciences*, *384*, 220–234. https://doi.org/10.1016/j.ins.2016.08.007.

Gainsbury, S., Sadeque, S., Mizerski, D., & Blaszczynski, A. (2013). Wagering in Australia: A retrospective behavioural analysis of betting patterns based on player account data. *The Journal of Gambling Business and Economics*, *6*(2), 50–68. https://doi.org/10.5750/jgbe.v6i2.581.

Goldstein, A. L., Vilhena-Churchill, N., Munroe, M., Stewart, S. H., Flett, G. L., & Hoaken, P. N. S. (2017). Understanding the effects of social desirability on gambling self-reports. *International Journal of Mental Health and Addiction*, *15*(6), 1342–1359. https://doi.org/10.1007/s11469-016-9668-0.

Haeusler, J. (2016). Follow the money: Using payment behaviour as predictor for future self-exclusion. *International Gambling Studies*, *16*(2), 246–262. https://doi.org/10.1080/14459795.2016.1158306.

Hahmann, T., Hamilton-Wright, S., Ziegler, C., & Matheson, F. I. (2021). Problem gambling within the context of poverty: A scoping review. *International Gambling Studies*, *21*(2), 183–219. https://doi.org/10.1080/14459795.2020.1819365.

Hodgins, D. C., & Makarchuk, K. (2003). Trusting problem gamblers: Reliability and validity of self-reported gambling behavior. *Psychology of Addictive Behaviors*, *17*(3), 244–248. https://doi.org/10.1037/0893-164X.17.3.244.

Hofmarcher, T., Romild, U., Spångberg, J., Persson, U., & Håkansson, A. (2020). The societal costs of problem gambling in Sweden. *BMC Public Health*, *20*(1), 1921. https://doi.org/10.1186/s12889-020-10008-9.

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*, *39*(1), 447–465. https://doi.org/10.1007/s10899-022-10149-z.

Hopfgartner, N., Auer, M., Helic, D., & Griffiths, M. D. (2024). Using artificial intelligence algorithms to predict self-reported problem gambling among online casino gamblers from different countries using account-based player data. *International Journal of Mental Health and Addiction*. Advance online publication. https://doi.org/10.1007/s11469-024-01312-1.

Jonsson, J., Abbott, M. W., Sjöberg, A., & Carlbring, P. (2017). Measuring gambling reinforcers, over consumption and fallacies: The psychometric properties and predictive validity of the Jonsson-Abbott scale. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01807.

Jonsson, J., Munck, I., Hodgins, D. C., & Carlbring, P. (2023). Reaching out to big losers: Exploring intervention effects using individualized follow-up. *Psychology of Addictive Behaviors*, *37*(7), 886–893. https://doi.org/10.1037/adb0000906.

Jonsson, J., Munck, I., Volberg, R., & Carlbring, P. (2017). GamTest: Psychometric evaluation and the role of emotions in an online self-test for gambling behavior. *Journal of Gambling Studies*, *33*(2), 505–523. https://doi.org/10.1007/s10899-017-9676-4.

Kairouz, S., Costes, J.-M., Murch, W. S., Doray-Demers, P., Carrier, C., & Eroukmanoff, V. (2023). Enabling new strategies to prevent problematic online gambling: A machine learning approach for identifying at-risk online gamblers in France. *International Gambling Studies*, *23*(3), 471–490. https://doi.org/10.1080/14459795.2022.2164042.

Kuentzel, J. G., Henderson, M. J., & Melville, C. L. (2008). The impact of social desirability biases on self-report among college student and problem gamblers. *Journal of Gambling Studies*, *24*(3), 307–319. https://doi.org/10.1007/s10899-008-9094-8.

Laurikkala, J., & Juhola, M. (2001). *Hierarchical clustering of female urinary incontinence data having noise and outliers* (pp. 161–167). https://doi.org/10.1007/3-540-45497-7_24.

Lövdal, S., & Biehl, M. (2024). Iterated relevance matrix analysis (IRMA) for the identification of class-discriminative subspaces. *Neurocomputing*, *577*, 127367. https://doi.org/10.1016/j.neucom.2024.127367.

Lundberg, S. M., Allen, P. G., & Lee, S.-I. (n.d.). *A unified approach to interpreting model predictions*. https://github.com/slundberg/shap.

MacKillop, J., Anderson, E. J., Castelda, B. A., Mattson, R. E., & Donovick, P. J. (2006). Divergent validity of measures of cognitive distortions, impulsivity, and time perspective in pathological gambling. *Journal of Gambling Studies*, *22*(3), 339–354. https://doi.org/10.1007/s10899-006-9021-9.

Murch, W. S., Kairouz, S., Dauphinais, S., Picard, E., Costes, J., & French, M. (2023). Using machine learning to retrospectively predict self-reported gambling problems in Quebec. *Addiction*, *118*(8), 1569–1578. https://doi.org/10.1111/add.16179.

Park, Y., Eom, D., Seo, B., & Choi, J. (2020). Improved predictive deep temporal neural networks with trend filtering. In *Proceedings of the first ACM international conference on AI in finance* (pp. 1–8). https://doi.org/10.1145/3383455.3422565.

Paterson, M., Taylor, M., & Gray, M. (2020). Trajectories of social and economic outcomes and problem gambling risk in Australia. *Social Indicators Research*, *148*(1), 297–321. https://doi.org/10.1007/s11205-019-02194-w.

Percy, C., França, M., Dragičević, 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. https://doi.org/10.1080/14459795.2016.1151913.

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. https://doi.org/10.1556/2006.2022.00063.

Sato, H., & Kawahara, J. (2011). Selective bias in retrospective self-reports of negative mood states. *Anxiety, Stress & Coping*, *24*(4), 359–367. https://doi.org/10.1080/10615806.2010.543132.

Suzuki, H., Nakamura, R., Inagaki, A., Watanabe, I., & Takagi, T. (2019). Early detection of problem gambling based on behavioral changes using shapelets. In *Proceedings - 2019 IEEE/WIC/ACM international Conference on web intelligence, WI 2019* (pp. 367–372). https://doi.org/10.1145/3350546.3352549.

Swedish Gambling Act, Pub. L. No. 2018:1138, Swedish code of statutes (2018).

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*(3), 877–897. https://doi.org/10.1007/s10899-020-09964-z.

Wang, T. D., Plaisant, C., Shneiderman, B., Spring, N., Roseman, D., Marchand, G., … Smith, M. (2009). Temporal summaries: Supporting temporal categorical searching, aggregation and comparison. *IEEE Transactions on Visualization and Computer Graphics*, *15*(6), 1049–1056. https://doi.org/10.1109/TVCG.2009.187.

Weatherly, J. N., Montes, K. S., Peters, D., & Wilson, A. N. (2012). Gambling behind the walls: A behavior-analytic perspective. *The Behavior Analyst Today*, *13*(3–4), 2–8. https://doi.org/10.1037/h0100725.

Yu, L., & Liu, H. (2003). Efficiently handling feature redundancy in high-dimensional data. In *Proceedings of the Ninth ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 685–690. https://doi.org/10.1145/956750.956840.