Predicting problem gambling among online sports and race bettors:

Assessing the value of Machine Learning using behavioural and self-reported data

Robert M. Heirene, Eden Zhang, Darya Vanichkina, Charles T. de Leau, Eunice L. Y. Huynh, Sally M. Gainsbury

A RESEARCH SUMMARY

What this study set out to do:

The study aims to assess the performance of Machine Learning models in identifying online sports and race betting customers with self-reported gambling problems (Problem Gambling Severity Index; PGSI). We aimed to:

[1] Build machine learning models to predict self-reported problem gambling using behavioural account data from Australian sports and race betting consumers.
[2] Compare the performance of models built using past-30-day vs past-6-month behavioural account data.
[3] Compare the performance of models built with and without survey-based variables.

METHODOLOGY/ANALYSIS

- Customers from two Australian sports and race betting sites were invited to complete a gambling survey (n=1470).
- Survey responses were linked to their behavioural account data across two windows: **30 days (n=1470) and 6 months (n=1349) pre-survey.**

What Machine Learning means here?

Machine learning refers to a set of algorithmic approaches that learn statistical patterns from training data to make predictions on new cases.

We trained **five machine learning algorithms** (Logistic Regression, Random Forests, Decision Trees, XGBoost, and Neural Networks) and evaluated which approach best distinguished high-risk from lower-risk customers, as defined by PGSI.

Outcome

Each model was trained to predict customers' problem gambling status, measured by the Problem Gambling Severity Index (PGSI). We examined three classification approaches:
[1] Multi-Class PGSI (0, 1-2, 3-7, 8+)
[2] PGSI ≥ 5
[3] PGSI ≥ 8

Predictors

This study used both **account data** and **survey data** to build the AI learning models.

- Demographics
- 30-day variables
- Past 6-month variables
- Use of consumer protection tools (e.g., self exclusion)
- Income and related features (e.g, % of income spent on bets, % of income deposited into accounts)
- Employment status
- Number of other betting accounts
- Estimations of gambling expenditure
- Life and gambling satisfaction

Stages of Model Testing

PHASE 1: BUILDING AND EVALUATING THE MODELS

Five model types for each of the **three PGSI classification approaches** (Multi-class PGSI, PGSI ≥5. PGSI ≥8) using **predictors derived from 30-day account data**, including behavioural metrics such as deposit frequency as well as demographic information (i.e., age and gender).

PHASE 2: 30 DAYS VS 6-MONTH DATASET

The **best performing model from phase 1** was kept and expanded to test both **30-day and 6-month aggregate variables.**

PHASE 3: INTEGRATING SURVEY VARIABLES

Survey-based predictors (e.g., gambling satisfaction and # of betting accounts) were integrated into the machine learning models to assess whether customer self-report data improved classification beyond account data alone.

KEY FINDINGS

HOW WELL DID THE MODELS PERFORM?

- Models performed better when identifying customers with severe problem gambling (PGSI ≥8) compared to moderate problem gambling (PGSI ≥5).
- XGBoost consistently outperformed the other four machine learning algorithms.

SENSITIVITY	SPECIFICITY	PRECISION	ACCURACY
Correctly Identified **65–76%** customers who were actually high-risk.	Correctly Identified **71–84%** lower-risk customers as not high-risk.	**33–50%** Of the customers the model flagged as high-risk were actually at high risk	Correctly Identified **71–81%** of customers overall.

30 DAYS VS 6-MONTH DATASET

- Models built using **either 30 days or 6 months of customers' behavioural account data performed similarly.**
  - This suggests that extended periods of historical data may not be needed, and that these models could be implemented soon after customer registration.

INTEGRATING SURVEY VARIABLES

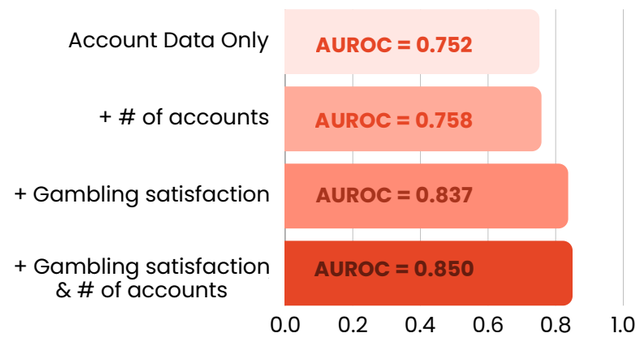

Figure 1. How well each model distinguished high-risk from lower-risk customers. Scores range from 0.5 (no better than chance) to 1.0 (perfect). Each row adds one more source of information to the model.

ADDING CUSTOMER SURVEY RESPONSES SUBSTANTIALLY IMPROVED THE MODELS' ABILITY TO ACCURATELY CLASSIFY CUSTOMERS.

- *The addition of* **gambling satisfaction improved the model the most**.
- The best model used **30 days of account data plus both survey variables** (No. of betting accounts and gambling satisfaction.

VARIABLE IMPORTANCE

Variable importance ranks which pieces of information the model relied on most when deciding whether a customer was high-risk.

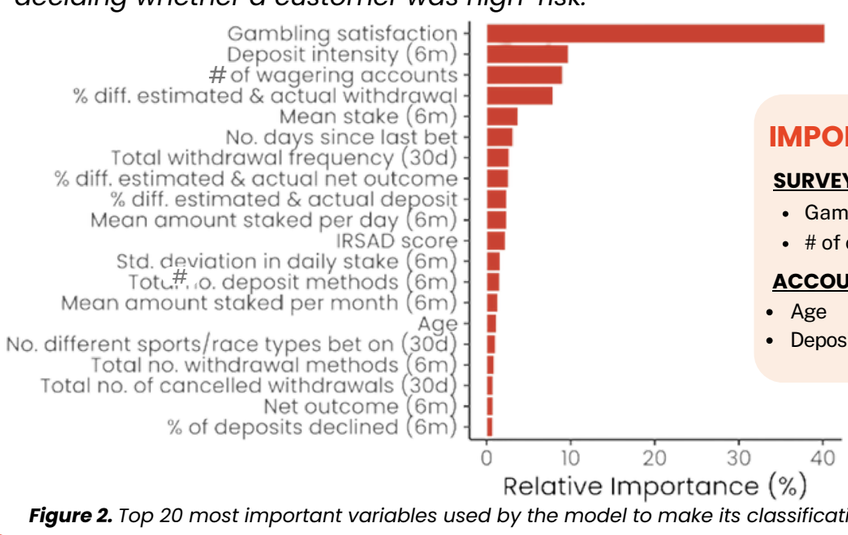

Figure 2. *Top 20 most important variables used by the model to make its classifications.*

IMPORTANT VARIABLES

SURVEY VARIABLES

- Gambling satisfaction
- # of online gambling accounts held

ACCOUNT BASED

- Age
- Deposit intensity
- Average stake size
- Days since last bet

RECOMMENDATIONS

Machine learning models can support early identification of at-risk customers:

- Operators do not need to wait for long account histories before applying risk-detection models, meaning interventions can occur earlier in a customer's gambling trajectory.
- Short, low-burden survey questions should be considered, especially around gambling satisfaction and number of active gambling accounts, as these substantially improved model performance.
- Gambling satisfaction should be further investigated as a potential early warning indicator, because it was the strongest self-reported predictor of high-risk gambling status.
- Model outputs should be used to guide proportionate support, not to definitively label customers as problem gamblers, because false positives remain likely.

Full paper can be accessed: https://doi.org/10.1556/2006.2025.00525

Please direct all Correspondence to: Dr. Robert Heirene (email: robert.heirene@sydney.edu.au)

This research was funded in part by a grant from the International Center for Responsible Gaming (ICRG), funded by Bally's Corporation. Its contents are solely the responsibility of the author(s) anddo not necessarily represent the official views of the ICRG or Bally's Corporation. RH was supported by a post-doctoral fellowship from the New South Wales (NSW) Responsible Gambling Fund as part of the Gambling Research Capacity Grant program, administered by the Office of Responsible Gambling (ORG)

THE UNIVERSITY OF SYDNEY