Gambling Addiction & Psychology - DSA210 Final Report

A. Bora YILDIZ – 00034546

Abstract

This project explores psychological and behavioral factors influencing gambling addiction using four datasets: ICPSR Gambling & Mental Health Study, Kaggle Gambling Behavior Data, Big Five Personality Traits, and Federal Reserve Consumer Debt Reports.

I have analyzed trends, performed hypothesis testing, and built classification models to predict problem gamblers. Results show personality traits like conscientiousness may relate to gambling risk, although predictive models struggled due to **data imbalance**.

1. Introduction

Gambling addiction is a behavioral disorder with increasing psychological and economic impact. This project examines how psychological traits, gambling behavior, and economic stress factors contribute to gambling addiction risk.

2. Datasets

- **ICPSR Gambling & Mental Health Study:** DSM-IV problem gambling symptoms and mental health indicators
- **Kaggle Gambling Behavior Data:** Bets, cashouts, and profit/loss tracking
- **Big Five Personality Traits:** OCEAN-based personality survey scores
- **Federal Reserve Consumer Debt Reports**: National-level economic stress metrics

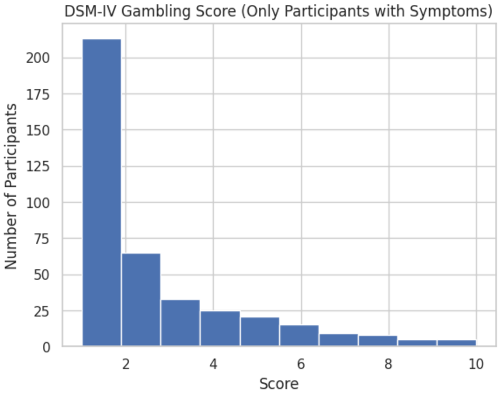
[All datasets are available on my github repo (github.com/BoraYldz)]

3. Methodology

For this project, I have performed exploratory data analysis, statistical hypothesis testing (independent t-tests), and built two classification models—Logistic Regression and Random Forest—to predict problem gambling based on personality traits and gambling behavior.

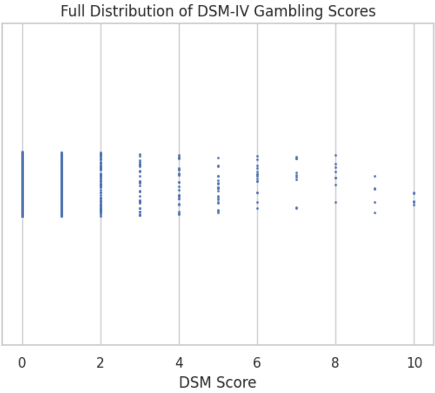
Finding a Problematic Gambler

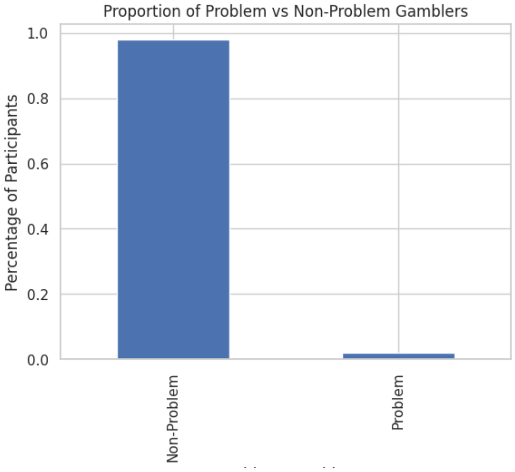
Participants were classified as **problem gamblers** if they met **five or more DSM-IV criteria**, based on the EVERPROB variable from the ICPSR dataset. This diagnostic threshold is standard in psychological studies and represents clinically significant gambling behavior.



This classification is visualized in the first graph, which shows that only a very small percentage of the sample (less than 5%) met this threshold — resulting in a stark class imbalance between problem and non-problem gamblers.

The second plot provides a **dot-based view of the full DSM-IV score distribution**, visually reinforcing the rarity of high-risk cases and highlighting the challenge this imbalance poses for predictive modeling.

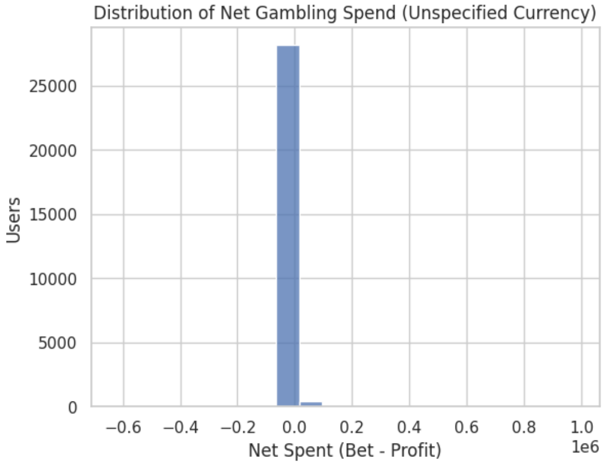


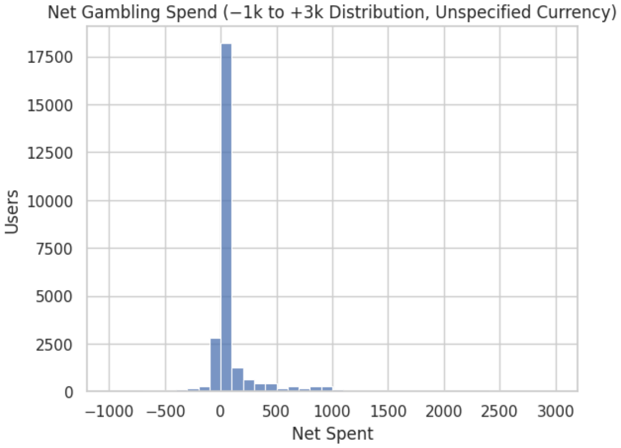


The third plot further illustrates this by showing the **distribution of DSM-IV scores** among participants with at least one symptom. The majority report mild issues (scores of 1–2), with very few reaching the problem gambler threshold.

Gambling Spending Habits

I have also checked participants spending habits on gambling to research impulsiveness. This data taken by Kaggle Gambling Behavior Data, does not specify a currency yet in the two plots below we can see These graphs main purpose is that a large margin of players quit playing while their losses and gains accumulation are close to zero.



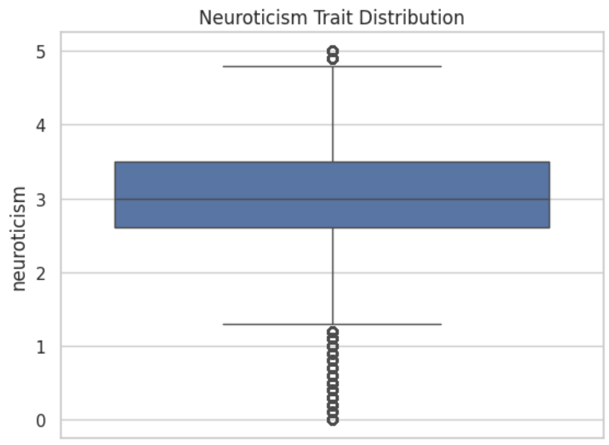


There are only outliers of big winners and big losers, and by above graphs these users are our main candidates for problem gamblers. Leftmost plot is the zoomed-out version, for clarity I have added a zoomed in version from -1000 currency debt to 3000 currency winnings.

Predicting Problem Gambling by Personality Traits

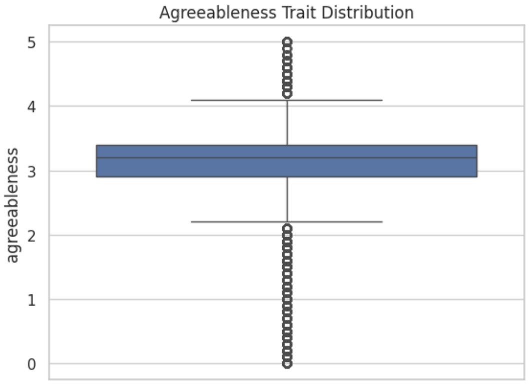
I tried to identify psychological patterns linked to problem gambling, using personality traits and behavioral data. In the ICSPR dataset, the variable EVERPROB counts how many DSM-IV criteria each participant met for gambling addiction A score of 5 or more (EVERPROB) is clinically recognized as problem gambling.

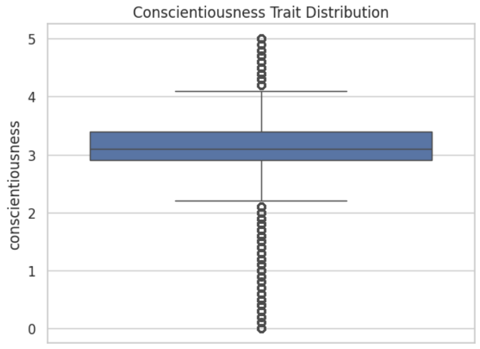
I have focused on **three** of the Big Five traits known to be most associated with addiction and impulsivity:



Neuroticism: Measures emotional instability and anxiety. Higher neuroticism is linked to gambling as an emotional escape.

Agreeableness: Measures empathy and trust. Lower agreeableness can indicate manipulativeness or antisocial behavior, traits correlated with risky gambling.





Conscientiousness: Measures self-discipline and organization. Lower scores suggest poor impulse control, a key predictor of gambling issues.

Each personality trait was calculated as the average score across 10 questions (e.g., EXT1 to EXT10 for Extraversion, from Kaggle Big Five Personalities Dataset). These values were added as model features alongside demographic and gambling behavior data. Then trained models to predict whether a person qualifies as a **problem_gambler.**

4. Hypothesis Testing

4.1 Do Problem Gamblers Have Higher DSM-IV Scores?

Hypotheses:

Null (H_0): There is no difference in DSM-IV scores between problem and non-problem gamblers.

Alternative (H_1): Problem gamblers have higher DSM-IV scores**.**

We used a two-sample independent t-test to compare DSM-IV scores (EVERPROB) between participants classified as, if problem_gambler = True; there are 5 or more symptoms.

Result:
$t = 78.10$, $p < 0.001$ **(very significant)**
The test confirms that problem gamblers have significantly higher DSM-IV scores, which supports our classification method (EVERPROB ≥ 5).

4.2 Is There a Difference in Conscientiousness Across Two Random Groups?

Hypotheses:

Null (H_0): No difference in conscientiousness between two random groups.

Alternative (H_1): The two groups differ in conscientiousness.

We took two random samples (n=500 each) from the Big Five dataset and compared their conscientiousness scores.

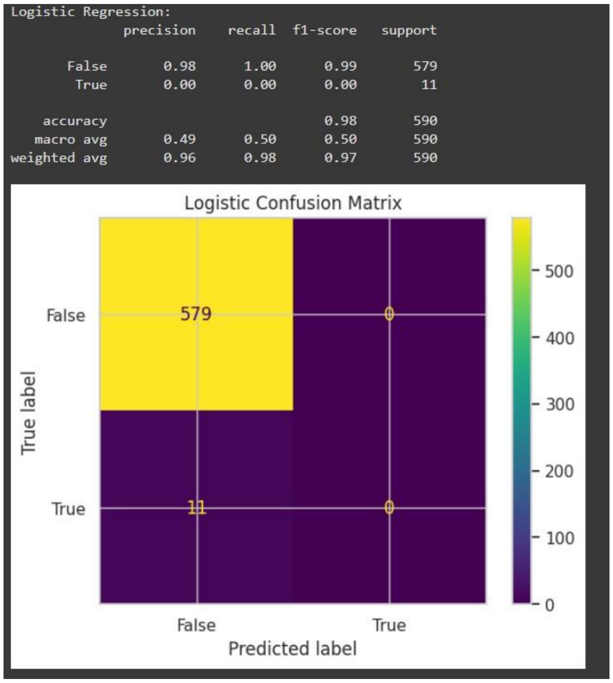
Result:
$t = -2.47$, $p = 0.014$
There is a small but statistically significant difference between the two random groups. This shows that personality traits can naturally vary across populations, important for understanding gambling risk patterns.

---

5. Implementing Machine Learning Models

In the final phase of this project, two machine learning models are implemented to predict whether a participant qualifies as a problem gambler, based on combined features from demographic (ICPSR), psychological (Big Five traits), and behavioral (gambling activity) datasets.

5.1 Logistic Regression



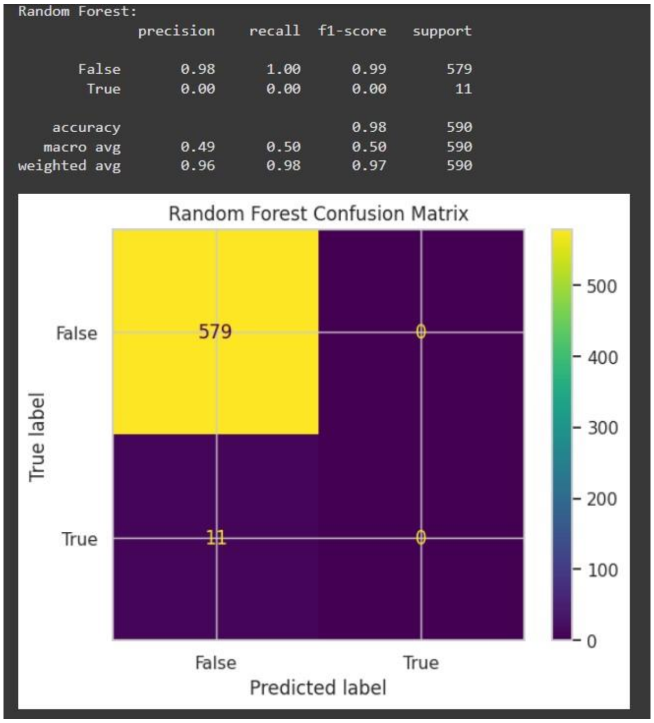
```
Logistic Regression:
              precision    recall  f1-score   support

       False       0.98      1.00      0.99       579
        True       0.00      0.00      0.00        11

    accuracy                           0.98       590
   macro avg       0.49      0.50      0.50       590
weighted avg       0.96      0.98      0.97       590
```

Logistic Regression is a commonly used classification algorithm designed to model binary outcomes. It estimates the probability that a participant belongs to one of two classes (problem gambler vs non-problem gambler), based on a linear combination of input features.

I have selected logistic regression as a baseline model as its well to linearly separable data, and also outputs probability scores for further analysis.

However, due to significant class imbalance **(only 11 out of 590 participants labeled as problem gamblers)**, the model defaulted to always predicting the majority class (non-problem gamblers). This led to poor recall and precision for the minority class, even though overall accuracy remained high. The confusion matrix confirmed that the model failed to identify any problem gamblers correctly.

5.2 Random Forest Classifier



```
Random Forest:
              precision    recall  f1-score   support

       False       0.98      1.00      0.99       579
        True       0.00      0.00      0.00        11

    accuracy                           0.98       590
   macro avg       0.49      0.50      0.50       590
weighted avg       0.96      0.98      0.97       590
```

Random Forest is a non-linear ensemble learning method that builds multiple decision trees and averages their predictions. It typically performs well on datasets with complex interactions between features.

Used Random Forest model to see if it could outperform logistic regression in detecting rare cases of gambling addiction. Unfortunately, due to the same class imbalance, Random Forest also classified all participants as non-problem gamblers, mirroring the logistic regression outcome.

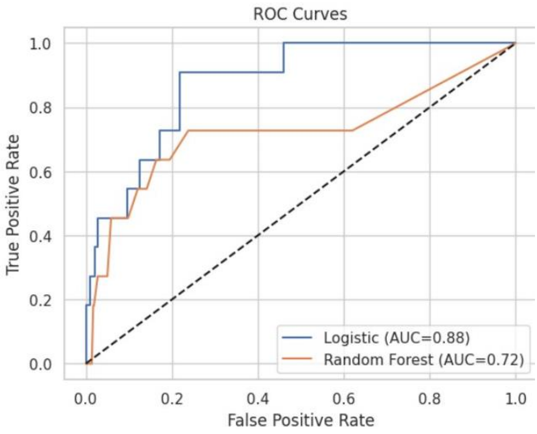
While the model demonstrated high accuracy and low total error, the confusion matrix showed that all 11 actual problem gamblers were misclassified, indicating zero sensitivity to the minority class.

Although both models achieved high accuracy, their inability to identify the minority class limits their practical value for this prediction task. **This highlights the importance of addressing class imbalance when dealing with real-world health or addiction datasets.**

5.3 ROC Curve Interpretation

Despite the poor classification outcomes, the ROC curve still offers meaningful insights, as it is based on the underlying probability scores the models assign rather than their final class predictions.

- **Logistic Regression (AUC = 0.88):**
  This model gave relatively higher probability scores to actual problem gamblers. While it ultimately labeled all participants as non-problem gamblers, still showed *some ability to distinguish* between the two groups. Which lead to stronger ROC curve and high AUC.
- **Random Forest (AUC = 0.72):**
  In contrast, this model failed to adequately separate problem gamblers from others. Lower AUC suggests weaker discriminatory power.



The higher AUC for logistic regression demonstrates that it internally recognized risk patterns even if it did not act on them due to class imbalance. This reinforces the importance of using AUC, not just accuracy, to evaluate model performance — especially in imbalanced datasets like this one.

6. Conclusion

In this project, I set out to investigate the psychological and behavioral dimensions of gambling addiction through a data-driven lens. By combining four datasets — including clinical self-report data (ICSPR), personality surveys (Big Five), online gambling behavior logs (Kaggle), and national debt trends (Federal Reserve) I aimed to identify patterns and predictive indicators of problem gambling.

I defined problem gambling using DSM-IV clinical criteria, labeling individuals with 5 or more reported symptoms as “problem gamblers.” This classification became the target variable for my machine learning models.

In Phase 3, I implemented two models — Logistic Regression and Random Forest — using demographic, psychological, and behavioral features. While both models achieved high overall accuracy (~98%), neither was able to detect the minority class of problem gamblers. As shown in the confusion matrices, all positive cases were misclassified. This was primarily due to severe class imbalance, with less than 2% of the dataset labeled as problem gamblers.

Despite this limitation, the ROC analysis offered additional insight. The Logistic Regression model achieved an AUC of 0.88, compared to 0.72 for the Random Forest. This suggested that, while Logistic Regression did not make correct predictions for problem gamblers, it still assigned higher probability scores to them — indicating a deeper underlying awareness of risk patterns.

Why Did I Do This Project?

This project confirmed that even with well-prepared features and multiple data sources, machine learning struggles with extremely imbalanced, real-world mental health data. The classifiers defaulted to the majority class, missing critical minority predictions — a common but important limitation in applied ML.

Still, the project achieved several goals:

- It successfully merged four diverse datasets; psychological, behavioral, and socioeconomic (finance and gambling) data.
- It demonstrated that certain traits (e.g., low conscientiousness) and loss patterns are associated with risk.
- It showed how ROC and AUC offer better insight than raw accuracy in imbalanced problems.

Moving forward, I would like to revisit this project with techniques better suited to minority detection, such as SMOTE (Synthetic Minority Over-sampling Technique), ensemble balancing, or cost-sensitive learning. I also hope to experiment with scraping social media data, from twitter or reddit.

Mental health data mostly points out the minority, and our ML solutions cant predicting rare outcomes from messy, imbalanced data. Even with decent features and real-world data, standard ML models can’t detect rare but critical cases unless you **specifically address the imbalance**.

I want to return to this project on a future date to learn how can I represent minority-classes (problem gamblers in this case) better.