Exploring Problem Gambling Predictors with Logistic Regression

A personal investigation into socio-economic and behavioural factors affecting gambling behaviour

I wanted to explore logistic regression as a technique to understand factors associated with problem gambling. Using survey data from England and Scotland, I examined socio-economic and behavioural variables, built a logistic regression model, and evaluated predictions. The project emphasises interpretability through odds ratios and predictive comparisons, highlighting how statistical modelling can provide actionable insights.

Technologies Used:

R (tidyverse, dplyr, ggplot2, R Markdown reporting)

Statistical Modelling (various regression techniques)

Introduction

This report aims to present the findings of some statistical analyses performed on data collected by a survey in England and Scotland. I have been commissioned by the NHS to examine problem gamblers and what variables may affect this decision. I have chosen to focus on socio-economic factors as it has the greatest number of variables, as well as varied information I can utilise. The dataset is appropriate as it contains points on household, income, assets, education, and employment, as well as other non-socioeconomic factors that can be used as confounders to illustrate a broader result in the question of

"*What are the significant factors associated with a problem gambling status?"*

My main model for determining the effect of such predictors is a logistic regression model, which I chose because my outcome variable, PROBGAM, is binary, and it allows me to accurately determine classifications. My final tested model is checked for accuracy through a confusion matrix, and based on it, I create average predictive comparisons to get easy-to-interpret results.

These results are important for a range of people, particularly those who wish to understand factors that may possibly influence gambling problems and make policy decisions based on it. By focusing on socio-economic factors, I hope to enlighten the effect of a person's background and status on their likeliness to be a problem gambler.

Exploratory Data Analysis

First, I examined the variables and their definitions, choosing those I believe best represented the data and my investigation, removing some like total income (totinc), which seemed to be redundant to eqvinc (equivalised income), the latter of which I chose to go with as it contains more personalised information. Then, I coded boxplots of continuous variables and for categorical ones, created mosaic plots which show the proportions of each category to the outcome variable, PROBGAM. Using a combination of the graphs and the counts in each category, as well as testing for initial significance, I merged and relevelled some predictors, the details of which can be found in the appendix.

For my chosen variables, I ran a multicollinearity analysis using VIF diagnostics in which all predictors had acceptable GVIF^(1/(2*Df)) values, with all except 1 being less than 1.4.

Model Diagnostics

I employed a mostly forward selection strategy after starting from a baseline model with my chosen socio-economic factors. Initially, I had added age, sex, religion, and ethnicity as confounder variables, believing those to be significant; however, after running an Anova analysis on my initial logistic regression model, I removed ethnicity as it was the least significant. Despite non-significance for some, I kept all socio-economic variables to keep investigating, even if they had a limited effect.

Then, I tested out each additional variable individually, retaining it based on its significance threshold, $p < 0.05$. Through this, my model got the additions of drating, wemwbs, cigst1, and country.

The analysis of the deviance table shows that, compared to the null model, the final model has a decrease of deviance by 232.38, and the p value is less than 2.2e-16, so I can reject the null hypothesis that the predictors do not improve the model and so the final model is statistically better.

Next, the confusion matrix, which I assessed at a 0.15 probability threshold due to the proportion of class imbalance, reveals only a slight improvement in accurate predictions from my initial model to my final model. The final yielded a correct classification rate of approximately 71% for non-problem gamblers but a 53% for problem gamblers, indicating a moderate false negative rate. As the purpose of this model is to motivate underlying factors, a low correct testing rate of 53% does not provide much confidence in the model and is not optimal.

Results

The final model has a residual deviance of 4489.2 on 5957 degrees of freedom compared to the null deviance of 4721.6 on 5993 degrees of freedom; this decrease implies the model is an improvement. I can interpret the coefficients as well, but instead of looking at them in terms of log-odds, which can be difficult to conceptualise, I transform them to odds ratios by exponentiation. The highlighted variables are those which are significant based on their p-value.

Table 1: Variable Results

Variable Name	Description	p-value	Estimate	Odds Ratio
age	Age	0.0000534	-0.017	0.983
SexFemale	Sex: Female vs Male	0.000000000000002	-0.677	0.508
Religsc_groupNo Religion	Religion: No Religion vs Catholic-Christian	0.317	-0.119	0.888
Religsc_groupNon-Catholic Christian	Religion: Non-Catholic Christian vs Catholic-Christian	0.736	-0.039	0.962
Religsc_groupOther Religion	Religion: Other Religion vs Catholic-Christian	0.857	0.04	1.041
HHSize_group3–5	Household Size: 3–5 vs (1-2)	0.342	-0.121	0.886
HHSize_group6+	Household Size: 6+ vs (1-2)	0.199	0.367	1.443
numcars_group2 cars	Number of Cars: 2 cars vs 1 car	0.021	-0.217	0.805
numcars_group3 or more	Number of Cars: 3 or more cars vs 1 car	0.654	-0.068	0.934
eqvinc	Equivalised Income	0.133	-0.000003	1

hhdtypb_groupMultiple Adults	Household Type: Multiple Adults vs Family Household	0.717	0.045	1.046
hhdtypb_groupSenior Household	Household Type: Senior Household vs Family Household	0.18	-0.292	0.747
hhdtypb_groupSingle Adult	Household Type: Single Adult vs Family Household	0.931	0.017	1.017
eqv5_groupHigh Income	Income Quintile: High Income vs Medium Income	0.875	0.019	1.019
eqv5_groupLow Income	Income Quintile: Low Income vs Medium Income	0.169	-0.157	0.855
HighQual_groupHigher Education	Highest Qualification: Higher Education vs Secondary	0.036	-0.194	0.823
HighQual_groupLow/No Quals	Highest Qualification: Low or No Qualifications vs Secondary	0.857	-0.022	0.978
Econact_2Education	Economic Activity: Education vs Employed	0.855	-0.043	0.958
Econact_2Retired	Economic Activity: Retired vs Employed	0.006	0.439	1.551

Econact_2Unemployed	Economic Activity: Unemployed vs Employed	0.993	-0.002	0.998
Econact_2Other	Economic Activity: Other vs Employed	0.764	-0.049	0.952
OwnRnt08Shared Ownership	Household Tenure: Shared Ownership vs Owner	0.138	-0.172	0.842
OwnRnt08Social Rent	Household Tenure: Social Rent vs Owner	0.501	0.271	1.311
OwnRnt08Private Rent	Household Tenure: Private Rent vs Owner	0.243	0.125	1.133
OwnRnt08Other	Household Tenure: Other vs Owner	0.152	0.512	1.668
hpnssec5Intermediate	NS-SEC 5 Classification: Intermediate vs Managerial	0.083	0.237	1.267
hpnssec5Self-Employed	NS-SEC 5 Classification: Self-Employed vs Managerial	0.764	0.043	1.043
hpnssec5Technical/Supervisory	NS-SEC 5 Classification: Technical/Supervisory vs Managerial	0.163	0.202	1.224

hpnssec5Semi-Routine	NS-SEC 5 Classification: Semi-Routine vs Managerial	0.096	0.193	1.213
SrcInc7Yes	Income Support: Yes vs No	0.187	-0.419	0.658
drating	Alcohol Units per Week	0.001	0.006	1.006
wemwbs	WEMWBS *Mental* Wellbeing Score	0.003	-0.014	0.986
cigst1Current smoker	Smoking Status: Current Smoker vs Never Smoked	0.005	0.304	1.355
cigst1Occasional smoker	Smoking Status: Occasional Smoker vs Never Smoked	0.28	0.178	1.194
cigst1Regular smoker	Smoking Status: Regular Smoker vs Never Smoked	0.76	-0.03	0.97
countryScotland	Country: Scotland vs England	0.015	-0.195	0.823

Key Findings:

- The odds ratio of 0.983 of age indicates that each additional year of age reduces the odds of problem gambling by about 1.7%.
- Living in a household with 6+ people is associated with higher odds by 44% as opposed to living alone or with one other person.

- Having two cars implies a 20% decrease in the odds of problem gambling than having one car.
- Higher education reduces odds by 18% compared to lower qualifications.
- Being retired is associated with 55% higher odds of gambling.
- Current smokers have a 36% higher chance of being a problem gambler as opposed to a person who's never smoked.
- A person living in Scotland has lower odds (17.7%) of being a problem gambler.

Similarly, the Average Predictive Comparison findings in the table below highlight changes across factors.

Table 2: Average Predictive Comparisons

Predictor	Change Compared	APC (Mean Δ Probability)
age	20 years to 50 years	-0.066
Sex	Male to Female	-0.076
Household Size	1–2 persons to 3–5 persons	-0.008
Household Size	1–2 persons to 6+ persons	0.053
Equivalised Income	£20,000 to £80,000	-0.019
Education Level	(Low/No Quals to Higher Education)	-0.019
Alcohol Consumption	0 to 50 units	0.036
Mental Wellbeing Score	30 to 70	-0.065

Smoking Status	Never smoked to Current smoker	0.036
Country	England to Scotland	-0.022

Comments on Data

This analysis has a few limitations. First, there's a substantial amount of missing data which I treated randomly rather than investigating whether there's a deeper reason. It could be that by generalising all the missing data, I lost some key insights. The final model also results in 5,994 rows being deleted due to missingness.

To interpret the results, I had to group categories, which I tried to make as accurate as possible without losing data. However, there were some categories, such as some religions, I had to forego as they had too few datapoints to properly analyse.

Finally, while most of the data isn't subjective, it could be that some people were reluctant to give accurate information and instead are misrepresented by the dataset. It is unlikely that many people would want to categorise themselves as problem gamblers.

Interpretation for a Lay Audience

I developed a model to be able to predict the probability of a person being a problem gambler based on other information I have about them. While the final model is marginally

better than no model at all, it still has its limitations and has only a moderate predictive classification accuracy.

The most significant factors I found were age, sex, household size, number of cars, education, employment, alcohol consumption, mental wellbeing, smoking status, and country. To consider the extent of the effect, I can interpret the odds ratios shown in Table 1, which consider the impact of a factor by holding all other variables constant.

The analysis shows that older individuals are less likely to be problem gamblers, while women are about half as likely as men. Living with six or more people increases the chances of problem gambling, and having two cars is associated with a smaller probability of having one car. Responders with higher education qualifications are also less likely to be problem gamblers, and retired people are more likely. Alcohol consumption as well as smoking are both directly related to the probability of being a problem gambler. Lastly, people in Scotland are less likely to be problem gamblers than those in England.

Other variables have little effect on the likelihood of being a problem gambler and it is interesting to note that among these are income group and job position.

Interpreting the average predictive comparisons, I can compare two people with all except one factor held constant. This means that a person earning £20,000 is 1.9 percentage points more likely to be a problem gambler as compared to their counterpart earning £80,000. Or

a person who currently smokes is 3.6 percentage points more likely to be a problem gambler than someone who has never smoked.

If policymakers wish to reduce the probability of someone becoming a problem gambler, they should consider targeted interventions that address the significant factors reported in this analysis and also look into further research for more accurate data assessments.

Appendix (Merging and Relevelling):

EducEnd

Merging: For the variable at what age education ended, as there were 8 values, I looked at the table and plot and based on those I merged 1 “Not yet finished”, 2 “Never went to school” and 3 “14 or under” into one “Under 14” category. I put values 4, 5, 6, and 7 into “Under 18” and the rest as “19 or over”.

Relevelling: I set the baseline as under 18 as that is the typical age education ends also supported by this category having the highest count.

numcars

Relevelling: 1 car as the reference category.

HHSize

Merging: I recoded the household size into three categories. 1-2 people, 3-5 people, and 6+. These categories were chosen both on count as well as similar proportionalities when looking at the contingency table when compared to the target variable.

Relevelling: The reference level is 1-2 people.

eqv5

Merging: By looking at the data set, I grouped the five categories into 3. A high income category (from quintiles 1-2), a middle income category (quintile 3), and a low income category (quintiles 4-5). The quintiles that were merged together had similar effects in the graphs.

Relevelling: Middle income is used as the baseline.

HighQual

Merging: Merged into three categories, Higher Education (which includes degrees and sub-degrees), Secondary (including GCSE and A-level), and Low/No Qualification

Relevelling: Secondary is used as the baseline.

Econact_2

Relevelling: Employed individual (1) is used as the reference group.

hhdtypb

Merging: I grouped the household types into Single Adult, Multi-Adult, Family Household, and Senior Household. These categories are based on both logical interpretations of the data as well as the sample size distribution and the observed problem gambling rates.

Relevelling: The reference level is set to Family Household.

OwnRnt08

Relevelling: Baseline is Owner (mortgage/freehold).

hpnssec

Relevelling: The reference level is managerial.

Sex

Relevelling: Baseline is set to Male.

ethincC

Merging: Combined into 5 broader categories, where I put White British and White Other in the same category, and the other categories are Black, Asian, Mixed, and Other.

Relevelling: The reference level is White as that's the most common ethnicity.

Religsc

Merging: Even though the graph and data shows variety between religions, most of the religions don't have enough data points that can be used to come to an accurate conclusions, which is why I merged to four categories, focusing on the ones with a good amount of data points; Christian Catholic, Christian other, No Religion, and Other Religion.

Relevelling: The most prominent group is Chritian Other which is why its used as the reference category.

cigst1

Relevelling: Changed the baseline to never smoked.

Country

Relevelling: Relevelled to England, as it has more datapoints.