Introduction

This report uses data from a survey conducted in England and Scotland, containing responses from over 13,000 individuals. The dataset includes detailed information on health, socio-economic, and lifestyle characteristics. The main outcome of interest is whether a respondent is a “problem gambler” or not. The analysis focuses primarily on socio-economic factors to answer the research question: what are the key socio-economic predictors which correspond to people having a gambling problem?

This report presents the steps taken to build the model, assess its performance, and interpret key findings. A lay summary has also been included to communicate the results in an accessible format for a non-technical audience.

EDA

A comprehensive exploratory data analysis was carried out on all variables in the dataset to investigate their relationship with the outcome variable, problem gambling (PROBGAM). This included all socio-economic, health, lifestyle, and demographic predictors.

For categorical variables, I created proportional bar plots to examine how individuals identified as problem gamblers varied across the different levels of each predictor. This made it easier to spot group-level differences in gambling behaviour.

For continuous variables, I plotted boxplots to show how the spread and central tendency varies between problem gamblers and non-problem gamblers.

Where necessary, variables were merged to simplify interpretation and ensure categories were meaningful. All missing categories were changed into a single “Missing” level for simplicity.

Model Building Process

Model development followed a forward selection approach, beginning with socio-economic factors. Decisions were made through the relevance of the variables and the results from the EDA. At each step, I evaluated the model improvements using likelihood ratio tests and checks for GVIF. The model-building process concluded after incorporating the key socio-economic predictors, followed by the inclusion of religion and country as it showed some significance from the EDA and in the final model.

Results

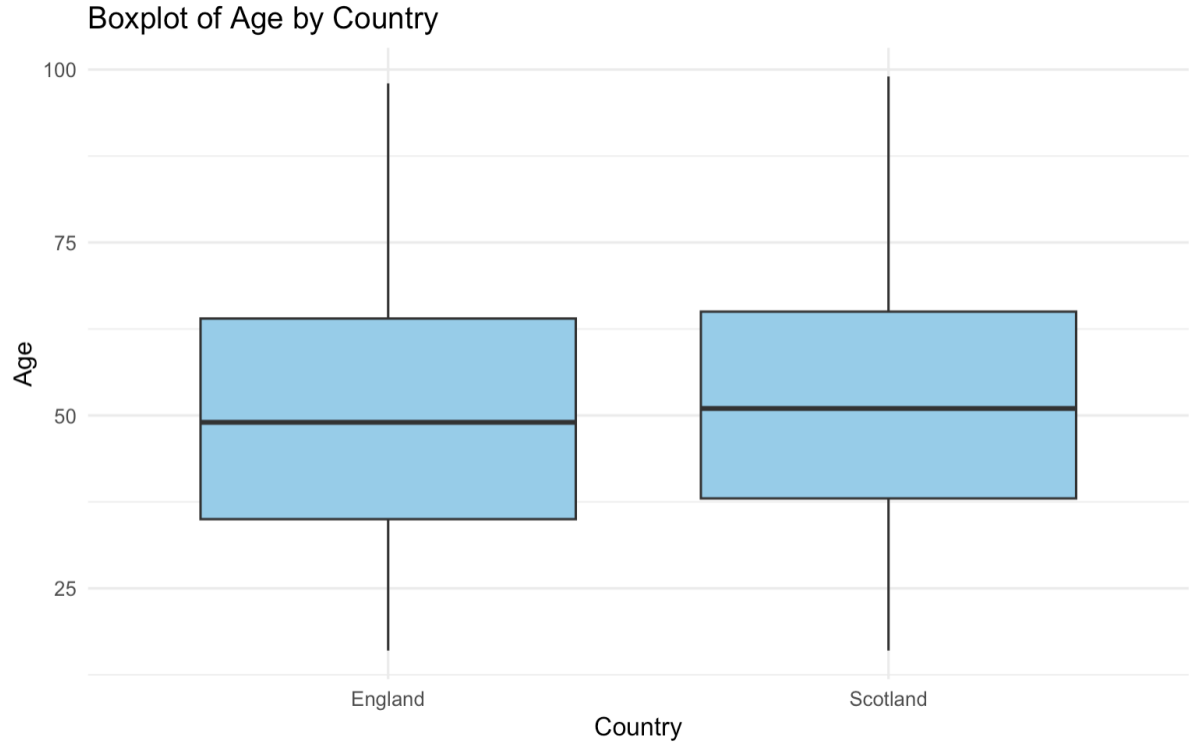
Predictor	Estimate	Std. Error	z value	Pr(|z |)	Significance
(Intercept)	3.52549	0.13872	25.415	<2e-16	***
HighQualMissing	-0.41490	0.62079	-0.668	0.50392	
HighQualDegree or Higher	0.15928	0.10428	1.527	0.12666	
HighQualGCSE or Equivalent	-0.09567	0.09772	-0.979	0.32759	
HighQualOther/None	-0.31644	0.09930	-3.187	0.00144	**
OwnRnt08Missing	0.53176	0.79817	0.666	0.50527	
OwnRnt08Mortgage	-0.00651	0.09463	-0.069	0.94516	
OwnRnt08Renting	0.12500	0.09408	1.329	0.18399	
numcarsMissing	-0.14811	0.08831	-1.677	0.09352	.
numcars3+ Cars	0.29316	0.14735	1.990	0.04664	*
totincMissing	-0.34810	0.22135	-1.573	0.11581	
totincLow Income	-0.20854	0.10120	-2.061	0.03933	*
totincHigh Income	0.01306	0.16170	0.081	0.93563	
totincDon't Know / Refused	-0.87368	0.10787	-8.099	5.53e-16	***
Econact_2Missing	-0.92047	0.76611	-1.201	0.22957	
Econact_2Student	-0.16377	0.17338	-0.945	0.34486	
Econact_2Retired	-0.22288	0.09672	-2.304	0.02120	*
Econact_2Unemployed	0.24029	0.18396	1.306	0.19149	
Econact_2Other Inactive	-0.24535	0.10916	-2.248	0.02461	*
countryScotland	-0.93997	0.07215	-13.027	<2e-16	***
ReligscMissing	-5.07732	0.13497	-37.617	<2e-16	***
ReligscChristian	-0.10410	0.08204	-1.269	0.20449	
ReligscOther Religion	-0.46951	0.20040	-2.343	0.01914	*
ReligscMuslim	-0.73161	0.20420	-3.583	0.00034	***

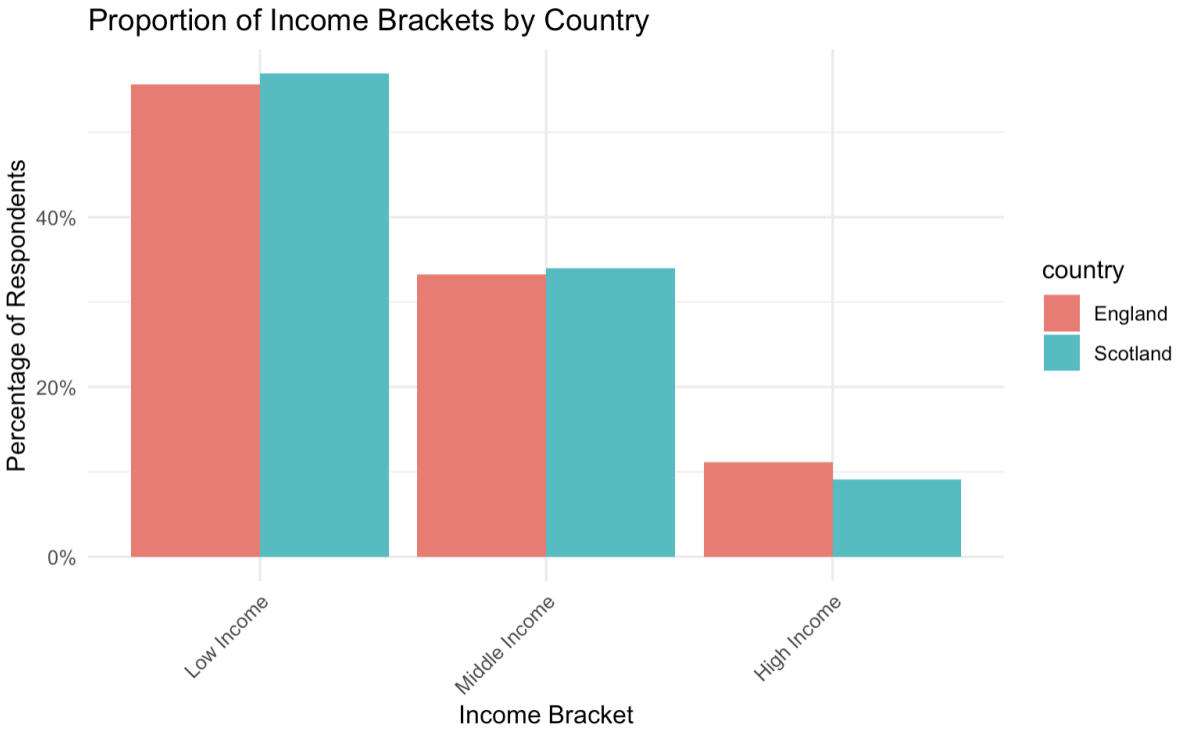
The final model identified several socio-economic factors that were significant. Individuals whose highest qualification was recorded as “Other/None” were significantly less likely to be problem gamblers compared to those with only high school-level education. Although not significant, those with missing qualifications also showed a lower likelihood of problem gambling. Respondents with GCSEs or equivalent had nearly identical probabilities to those with high school-level education, while those holding a degree or higher appeared slightly more likely to be problem gamblers. This pattern could suggest a counterintuitive relationship, where higher educational attainment may be associated with greater risk.

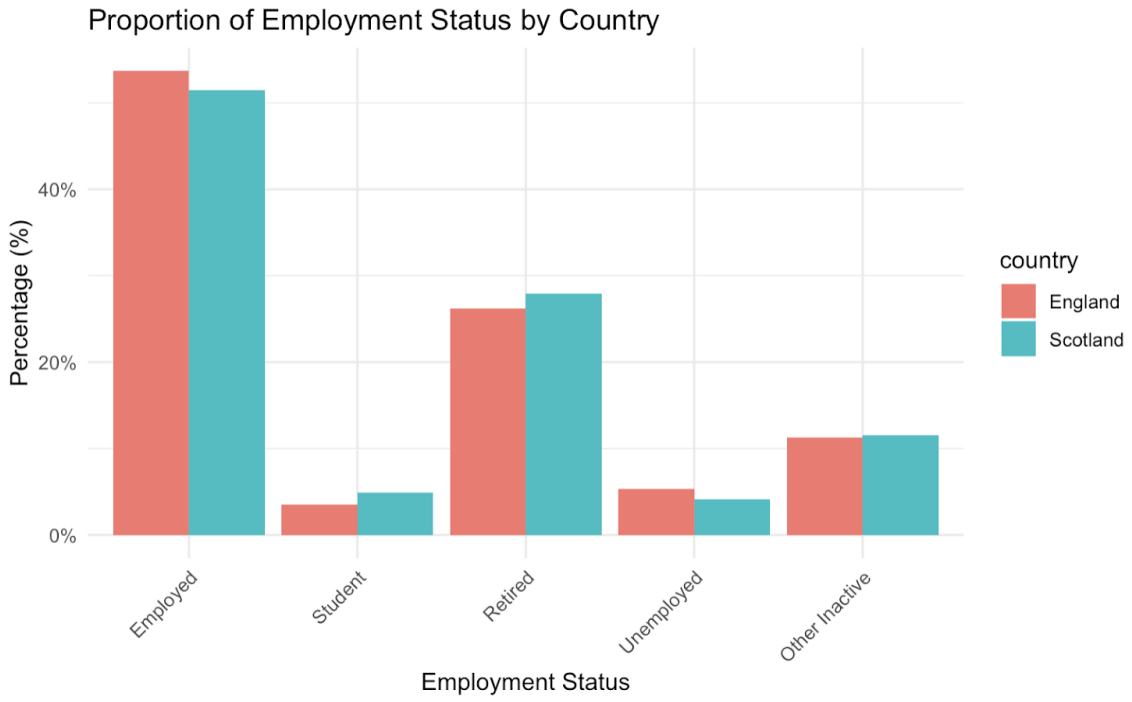
In terms of income, individuals in the lowest household income group (TotIncLow) were significantly less likely to be problem gamblers than those in middle or higher income brackets. Supporting this, individuals in households with three or more cars (NumCars3+) were more likely to be problem gamblers. Econtact_2 followed the same trend. Those who were “Retired” “Other Inactive” were significantly less likely to be problem gamblers compared to those in employment. Unemployed individuals, on the other hand, showed a higher likelihood of problem gambling than those employed, although this result was not statistically significant. These results overall suggest that actively earning or having a higher income may be associated with an increased risk of problem gambling.

However, this result could show that the observed relationship between education and gambling risk may not be directly due to education itself. Instead, it may stem from the fact that individuals with lower/no qualifications typically earn lower incomes, which is what keeps them at a lower risk of having a gambling problem.

Now consider the non-socioeconomic predictors. Individuals living in Scotland were significantly less likely to be problem gamblers compared to those in England. To investigate why this is the case, I've investigated the data further through the plots below to investigate the difference between the Scotland and England respondents.




Proportion of Income Brackets by Country
Percentage of Respondents
40%
20%
0%
Low Income
Middle Income
High Income
Income Bracket
country
England
Scotland


Proportion of Employment Status by Country
Percentage (%)
40%
20%
0%
Employed
Student
Retired
Unemployed
Other Inactive
Employment Status
country
England
Scotland

Scottish respondents had a higher proportion in the “Low Income” category, which was associated with a lower risk of problem gambling. In contrast, a greater proportion of English respondents were in the “High Income” group, which was linked to a higher gambling risk. Scotland had a higher proportion of “Retired”, “Student” and “Other Inactive” respondents, which again are all linked to a lower problem gambler risk.

Therefore, the result that Scottish people are significantly less likely to be problem gamblers is most likely due to the overall demographic and socio-economic profile of the sample. These proportions may accurately reflect the real population characteristics, suggesting the sample is representative of the Scottish population. This implies that geographic differences in problem gambling rates may be driven more by the makeup of local populations rather than regional effects themselves.

Among religious beliefs, Muslims followed by ‘Other Religions' were significantly associated with a lower probability of problem gambling compared to those having no religion. Christians also followed this trend but the result was not significant. Upon further research, Islam is the only religion which very strongly prohibits gambling. Other religions discourage gambling but do not categorically forbid it. Therefore, it's likely that these religious beliefs keep you away from gambling.

While many socio-economic factors showed significant associations with problem gambling, some predictors did not demonstrate a statistically meaningful relationship in the final model. No level across housing tenure came to give significant results. This is somewhat surprising, as housing situations are often closely tied to income.

Finally, some of the more significant results were observed among those who did not respond to questions regarding household income (TotInc) and religion. These patterns suggest the presence of informative missingness, where not responding may itself indicate important underlying differences. Exploring these patterns further will likely provide further insights but it lies outside the scope of this report.

To quantify the effects of key predictors, average predictive comparisons (APCs) were calculated to estimate the change in probability of being a problem gambler associated with the corresponding socio-economic predictors.

Those with "Other or None" qualifications were about 9 percentage points less likely to be problem gamblers compared to those with high school-level education. Individuals with a GCSE or equivalent showed a 1.3 percentage point reduction in risk, while those with degrees were more likely problem gamblers by 1.2 percentage points. Those who did not report their highest qualification (“Missing”) were dramatically less likely to be problem gamblers, with a 50 percentage point lower predicted probability compared to those with only high school-level education.

Income levels were also associated with different likelihoods of problem gambling. Individuals in the low-income group were 3.7 percentage points less likely to be problem gamblers than those

in the middle-income bracket. Those in the high-income group were no different to the middle-income group. In terms of employment, “retired”, “inactive” and “student” were 5.2, 6.3 and 1.5 percentage points respectively less likely to be problem gamblers. “Unemployed” had no difference compared to those in employment.

Classification table

	obs=0	obs=1
pred=0	79.00	20.00
pred=1	9631.00	1679.00
%corr	0.01	0.99

The classification table shows the results of my final model and reveals a severely imbalanced classification pattern. The model demonstrates excellent sensitivity, correctly identifying 99% of true problem gamblers. However, its specificity is extremely low, with only 1% of non-problem gamblers correctly classified. This skewed performance suggests that the model is overwhelmingly biased toward predicting individuals as problem gamblers. While it rarely misses true cases, its near-total inability to rule out false positives means it performs only marginally better than a naive model that assumes everyone is a problem gambler.

Lay Report

One of the strongest findings was that people with lower qualifications were less likely to be problem gamblers. For example, individuals with “Other/None” qualifications were around 9 percent less likely to have gambling problems compared to those with high school-level education. In contrast, those with degree-level qualifications showed a slightly higher likelihood of gambling problems. Income may help explain this pattern. Individuals in the lowest income categories are the least likely to experience problem gambling (compared to those of higher income). It is possible that those with lower educational qualifications tend to earn less over their lifetimes than those with degrees. This connection to income may be a key factor influencing the risk of problem gambling.

Employment status also followed a similar link with gambling risk. Retired/inactive individuals and students were less likely to be problem gamblers compared to those in employment. Meanwhile, unemployed people showed a similar level of risk to employed individuals. These findings suggest that being actively employed and having more disposable income comes with a higher chance of gambling problems.

Religious beliefs also played a role. Individuals who identified as Muslim were significantly less likely to be problem gamblers compared to those with no religion. Other religious groups also showed reduced risk, but to a smaller degree.

Lastly, there were also notable differences between England and Scotland. People living in Scotland were significantly less likely to be problem gamblers. A closer look at the data shows that Scotland generally has more, retired, and lower-income individuals who are all less likely to have a problem gambling risk.

Overall, the findings suggest that a person's background, particularly their level of education, income and employment status can influence their risk of gambling problems. In particular, individuals who are less economically active, have lower qualifications, or identify with a religion were generally less likely to be problem gamblers. While gambling problems can affect anyone, these patterns highlight the importance of targeting prevention efforts toward groups that may be at higher risk.

Appendix

Predictor Name	Original Levels	New Levels	Baseline Level
HHSize	1–11	'Small (1–3)', 'Medium (4–5)', 'Large (6+)'	Small (1–3)
Sex	['1', '2', '-1', '-2', '-8', '-9']	'Male', 'Female', 'Missing'	Male
maritalg	['1', '2', '3', '4', '5', '6', '-8', '-9']	'In a Relationship', 'Not in a Relationship', 'Missing'	In a Relationship

hhdtypb	['1', '2', '3', '4', '5', '6', '7', '-8', '-9']	'Young Adults – No kids', 'Family', 'Older Adults – No kids', 'Missing'	Family
OwnRnt08	['1', '2', '3', '4', '5', '-8', '-9']	'Mortgage', 'Ownership/No Rent', 'Renting', 'Missing'	Ownership/No Rent
numcars	['1', '2', '3', '-1']	'1–2 Cars', '3+ Cars', 'Missing'	1–2 Cars
SXORIEN	['1', '2', '3', '4', '-1', '-6', '-9']	'Heterosexual', 'LGBTQ+', 'Missing'	Heterosexual
Religsc	['1', '2', '3', '4', '5', '6', '7', '8', '9', '-1', '-8', '-9']	'No Religion', 'Christian', 'Muslim', 'Other Religion', 'Missing'	No Religion
ethnicC	['1', '2', '3', '4', '5', '6', '-1', '-8', '-9']	'White', 'Black', 'Asian', 'Other', 'Missing'	White
Econact_2	['1', '2', '3', '4', '5', '-1', '-8', '-9']	'Employed', 'Student', 'Retired', 'Unemployed', 'Other Inactive', 'Missing'	Employed
cigst1	['1', '2', '3', '4', '-1', '-2', '-6', '-8', '-9']	("2", "3") Ex-smoker	Never Smoked
EducEnd	['1', '2', '3', '4', '5', '6', '7', '8', '-1', '-2', '-8', '-9']	("2", "3", "4") Left Early or No Education ("5", "6", "7", "8") Left at 16+	Left at 16+
HighQual	['1', '2', '3', '4', '5', '6', '-9']	("2", "3") High School ("5", "6") Other/None ("2", "3") High school	High School
eqv5	['1', '2', '3', '4', '5', '-1']	['High Income', 'High Income', 'Medium Income', 'Medium Income', 'Low Income', 'Missing']	Medium Income
ghq12scr	['0', '1', '2', '3', '4', '5', '6', '7', '8', '9', '10', '11', '12', '-1', '-2', '-6', '-8', '-9']	(0:2) Low distress (3:5) Moderate distress (6:12) High Distress	Low Distress
totinc	['1', '2', '3', '4', '5', '6', '7', '8', '9', '10', '11', '12', '13', '14', '15', '16', '17', '18', '19', '20', '21', '22', '23', '24', '25', '26', '27', '28', '29', '30', '31', '96', '97', '-1', '-2', '-8', '-9']	(1:15) Low income (16:23) Middle income (24:31) High income ("96", "97") don’t know	Middle Income