Candidate number: 43495

Word count: 1474

ST211 Individual Project

Contents

Introduction

The data in this research comes from the Next Steps Study, which follows over 16,000 UK residents born in 1989-90 who attended secondary school in England. The research began in 2004 with 14-year-old volunteers and continued annually for seven years. It seeks to investigate the experiences of young people in secondary school and beyond, such as further education, training, or employment. The analysis of Wave 8 statistics, which is the most recent survey completed when the participants were 25 years old, is the topic of this research. We focus on potential factors that may have impacted respondents' loans at the age of 25, such as their educational history, family circumstances, and physical and mental health.

Understanding the factors that influence a person's debt status at the age of 25 is crucial. By identifying these factors, we can suggest policy changes that address them, ultimately improving living standards for more individuals.

Exploratory Analysis

To identify any relevant outliers that could have affected the results, we analyzed residual plots, boxplots, skewness, and kurtosis of W8QDEB2. Our analysis revealed that some data points had high leverage values and standardized residuals, indicating a significant impact on model fit. Moreover, W8QDEB2 had a highly skewed and peaked distribution with several outliers (skewness = 11.90056, kurtosis = 227.5685), which could negatively impact model building and inference. To mitigate this, we transformed W8QDEB2 into binary numbers (hasdebt), using W8QDEB2 $\leqslant$ 0 as no debt (0) and all other values as has debt (1). We then analyzed the skewness and kurtosis of hasdebt, which indicated a distribution close to normal with no significant skewness or spikes (skewness = 0.0224, kurtosis = 1.0005).

After running the initial regression containing all the variables, we removed all predictors that were not significant at the 5% level, leaving us with 11 predictors for hasdebt.

When independent variables are highly correlated, the standard error of the regression coefficient can become very large, making it difficult to assess the significance of the coefficient. Therefore, we checked for multicollinearity among variables using the corMatrix and VIF, and found no significant covariance between predictors.

To ensure the study's accuracy and avoid bias, we decided to exclude missing values which less than 10%. This decision was made to avoid complications and maintain statistical power. Any missing data above 10% were grouped and transformed into plausible values.

We combined many categories for our categorical predictors to increase their significance and simplify interpretation, the way we merge is in the Appendix.

First to the final model

Table 1

Coefficients	Model 1	Model 2	Model 3	Model 4	Final Model
hasdebt				*	***
W1yschat1	*	**	**	***	***
W1wrk1aMP	*	**	**	**	***
W1truantYP	*				
W1alceverYP	***	**	**	**	***
W1disabYP	*				
W6acqno	*	**	**	**	**
W6DebtattYP	**	**	**	**	**
W8CMSEX		*			
W8DDEGP	***	***	***	***	***
W8DGHQSC	***	**			
W8DAGEYCH	***	***	***	***	***
W8DGHQSC:W8CMSEX			**	**	**
Null deviance	7802.7	7120.3	7120.3	7120.3	6965.1
Residual deviance	6732.1	6811.0	6802.7	6803.2	6550.6
AIC	7184.1	6849	6842.7	6835.2	6582.6

In this study, We used logistic regression to analyze all variables in Model 1 and selected the ones with significant associations. To investigate the relationship between gender and debt, we added another predictor called W8CMSEX, which showed significance with P-values less than 5%. After cleaning the data, we ran the logistic regression again to obtain Model 2. To assess the association between W8CMSEX, W8DGHQSC, and hasdebt, we used scatter plots and found a probable interaction.

We created a logistic regression model (Model 3) with an interaction term and analyzed its coefficients. After comparing the residual deviance of Model 2 (6811.0) and Model 3 (6802.7), we determined that Model 3 is a better fit because it has smaller residual deviance. Additionally, we compared the chisq value using Anova tables and found that Model 3 is significantly better than Model 2 (which is better than Model 1). Therefore, we included all significant variables from Model 3 into Model 4.

After using Cook's Distance, Residuals vs Leverage plot, and DFFITS to analyze outliers, we removed them and finalized the model based on 5054 observations.

When analysing the residual deviance and Anova data from Model 4 and the Final Model, it was clear that the Final Model performed better when outliers were removed. The AIC shifted from 6835.2 to 6582.6, indicating a substantial difference between the two models.The residual graphs demonstrated this as well. The final logistic regression model improved the fit to the data, as evidenced by the difference between the Null deviance (6965.1) and Residual deviance (6550.6), which was 414.5.

The odds ratio was utilized to analyze the final model's proportional impact on the dependent variable. The final model was also used to predict values, with the majority falling within the 0.2 to 0.8 range, supporting the assumption that response values would be binary, between 0 and 1. Also, we used classification table and got a accuracy of 0.61 of prediction.

Results

Table 2

	Estimate	**P-value**	**Odds ratio**
Intercept	0.119031	0.742591	1.1264053
W1yschat1	0.025763	1.36e-15	1.0260978
(W1wrk1aMP)full_time_employee	0.324657	0.000594	1.3835564
(W1wrk1aMP)part_time_employee	0.129458	0.178342	1.1382115
(W1wrk1aMP)look_after_family	-0.141922	0.180795	0.8676891
(W1alceverYP)alcoholic_drink	-1.405448	0.000331	0.2452572
(W1alceverYP)no_alcoholic_drink	-2.107775	1.12e-07	0.1215080
W6DebtattYP	-0.258899	0.004428	0.7719008
(W6acqno)Other	-0.020476	0.007099	0.9797321
(W6acqno)No_academic_study_aim	0.420708	0.779410	1.5230398
(W8DDEGP)no degree	0.138714	5.90e-08	1.1487950
(W8DDEGP)first or high	0.027846	0.061054	1.0282373
W8DGHQSC	0.017868	0.241844	1.0180287
(W8CMSEX)Female	-0.025618	0.729758	0.9747076
(W8DAGEYCH)with_child	0.643350	3.64e-13	1.9028455
W8DGHQSC:factor(W8CMSEX)Female	0.062861	0.001817	1.0648789

In Table 2, we can see how having debt or not at the age of 25 relates to other variables. Besides the results from the logistic regression, the way of how we use odds ratios in this context should be clarified first. These ratios indicate the relative likelihood of the target variable, not its probability. For example, if the odds of W1yschat1 are 1.023974, it means that a one unit increase in W1yschat1 will increase the odds of debt by 1.023974 times or about 2.4%.

Firstly, similar to the wage levels previously studied, the relationship between hasdebt and educational attainment (W8DDEGP) is that when observers do not have a degree they are 5 times more likely to take out a loan than those who have obtained a first or higher degree. Also observing the odds ratio shows that this variable has a greater impact on loans.Attitude towards school and highest grade obtained (represented by W1yschat1 and W6acqno) are also related to debt. Logistic regression analysis shows that attitude towards school at age 14 has little effect on loans, but better grades at age 17 increase the likelihood of being

debt-free.

Another significant factor to consider is whether or not someone has children at the age of 25 (W8DAGEYCH). On average, those with children are 0.643 points more likely to have a loan compared to those without children.

We can observe that gender (W8CMSEX) and mental disorders (W8DGHQSC) do not reach a significant level of 5%, but the interaction term has high significant level. Also, it has a positive effect on the outcome, meaning that if a person is female and has psychological issues, she is more likely to have debt.

Comments about the data analysis

We faced limitations when analyzing debt due to missing data, which we removed. This decision may have caused bias in the regression coefficients and potentially affected the impact of certain variables on loan status. Missing data in two variables(W8DDEGP and W8DAGEYCH) had a notable effect, but we lack information for thorough analysis. Additionally, we had to remove extreme outliers from the dataset as they can affect coefficient estimates and lead to inaccurate model predictions. Non-linear relationships must be considered when modeling variables, like the complex connection between female psychological problems and debt status.

After examining the specific variables, we discovered several issues. Firstly, our regression analysis indicated that students who did not report drinking at the age of 14 may not have been truthful due to the legal drinking age in the UK. This could potentially introduce bias. Secondly, we expected an interaction between academic qualifications, high school grades, and loan status. We assumed that good grades in high school would lead to better academic qualifications and influence loan decisions. However, our data did not support this assumption. Finally, attitudes towards loans at age 18 have a minor impact on the decision to take out a loan, contrary to our hypothesis. Without data from ages 18 to 25, we cannot draw a definitive conclusion.

Seen above of the results, accuracy and comments of the model, we don’t acknowledge this model as a good one. If we want to get a better model, we need more data and variables.

Interpretation & Conclusions for a lay audience

Our research examines how nine factors are associated with loan status, including education level, parental occupation, alcohol consumption, attitudes toward loans, mental health, gender, and parental status.

One of the most significant factors is whether or not the participant has children. Participants with children are over 60% more likely to have a loan compared to those without children. This factor has a significant impact on loans, increasing the probability of taking out a loan by approximately 90% when transitioning from having no children to having children. Furthermore, educational level is a key aspect. Individuals with a higher education are 70% less likely than those without a degree to have a debt.

Here we take a specific example in Table 3, varying any of gender, university qualifications,teenage

alcoholic status, and have children or not keeping the rest the same. We use our model to predict and conclude that:

Table 3

	Gender	Qualification	Alcohol	Children	Probability of has debt
Sample A	Female	First Degree	Yes	No	0.49
Sample B	Male	First Degree	Yes	No	0.46
Sample C	Female	First Degree	No	No	0.32
Sample D	Female	No degree	Yes	No	0.56
Sample E	Female	First Degree	Yes	Yes	0.65

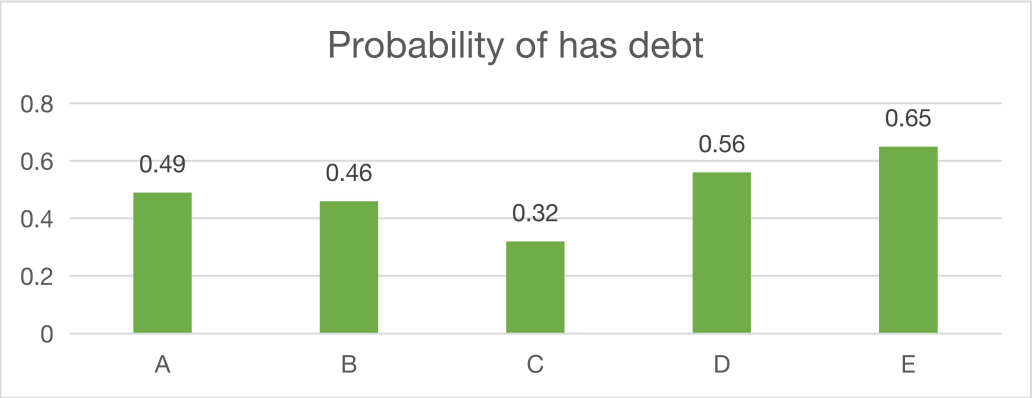
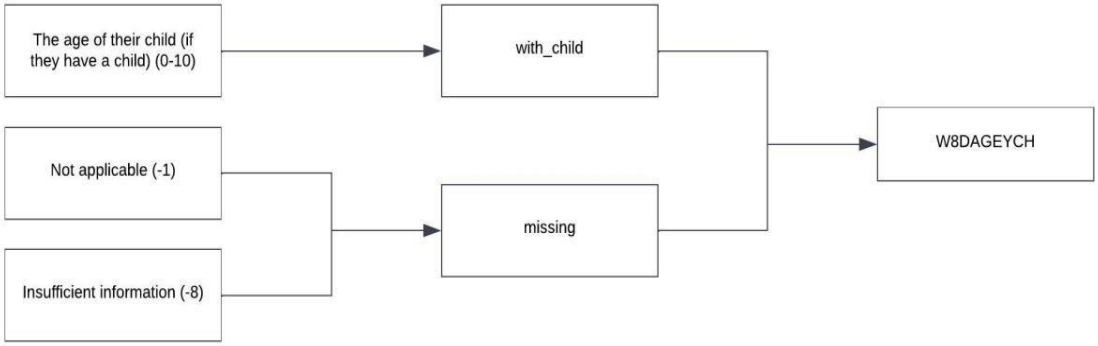


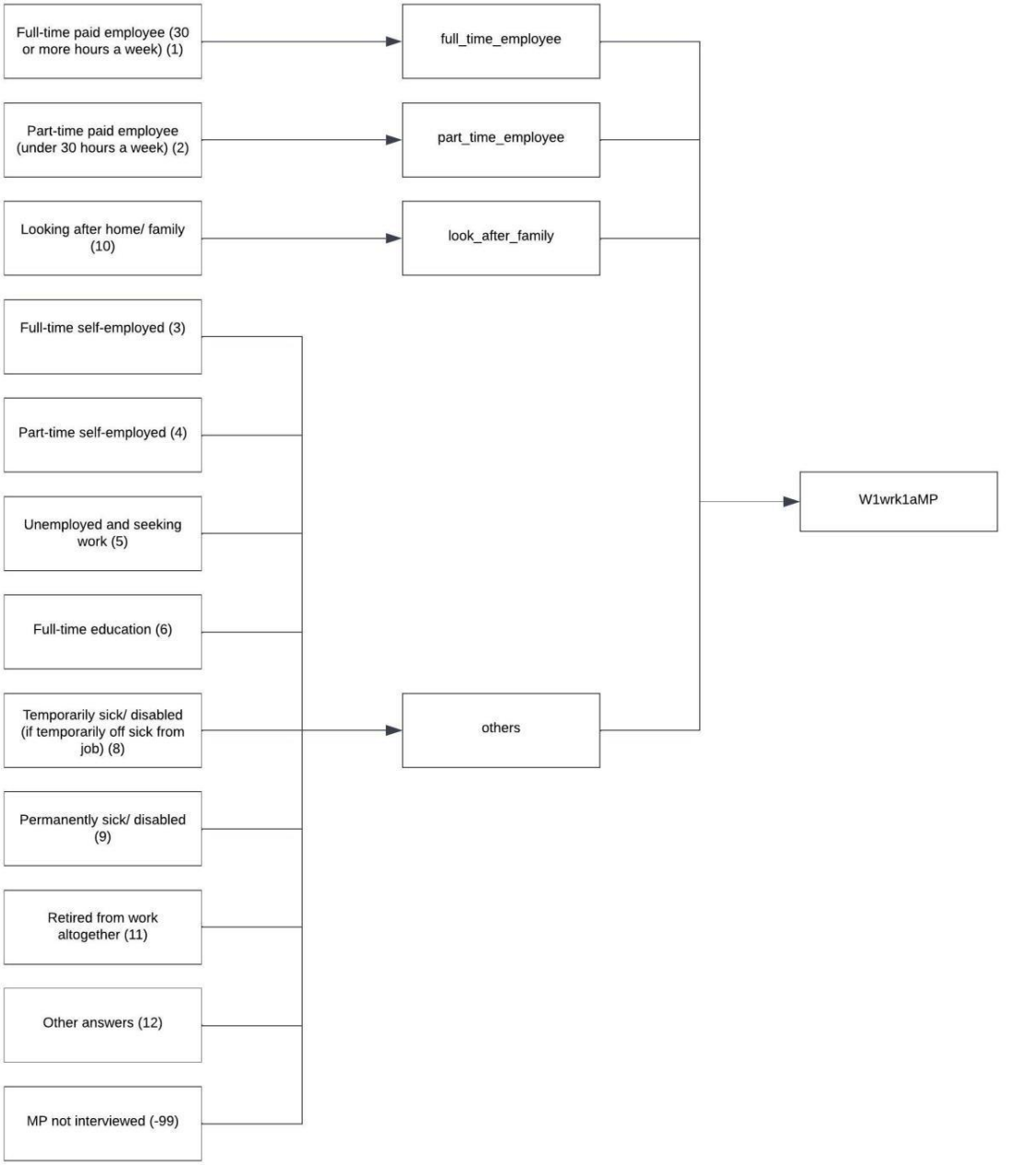
Chart 1: probability of has debt based on the table 3

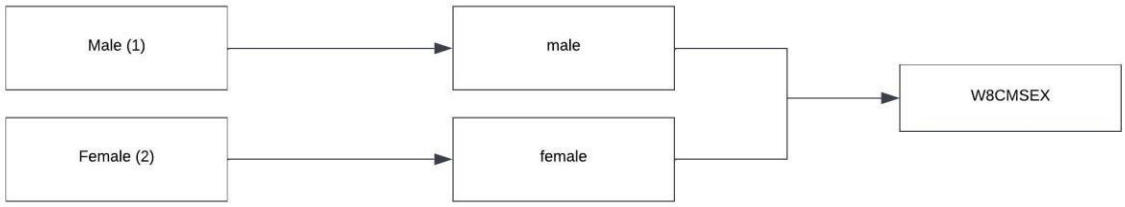
From the Table 3 and the Chart 1, It suggests that men have a slightly lower debt rate than women, individuals with a high level of education have lower debt rates than those with no degree, those who drink alcohol generally have higher debt rates than those who do not drink alcohol, and having children is associated with a higher debt rate. To conclude, Sample E had the highest debt ratio (0.65) and was twice as high as Sample C (0.32), which may be due to a combination of factors.

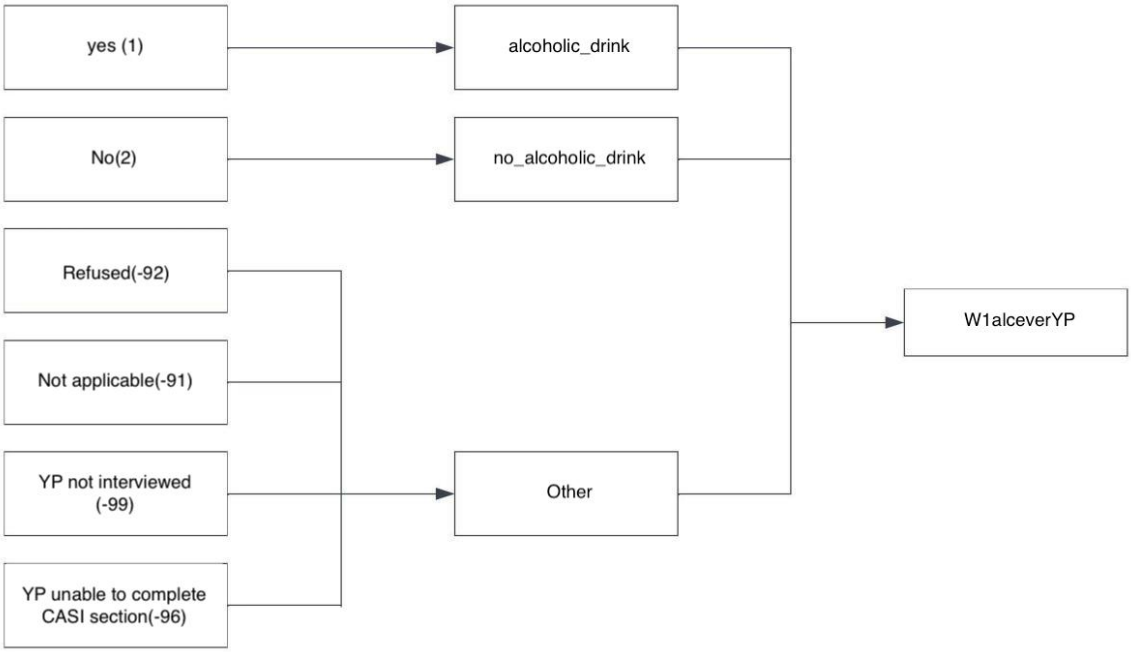
Assessing borrowing ability, income levels, and credit scores is crucial when evaluating loan status. Education level and potentially negative factors like income discrepancies should also be considered. Also,families with children often require loans to cover expenses like education and medical costs, but obtaining them can be challenging. Women's mental health can also affect their ability to secure loans, as they often face societal pressures and multiple responsibilities. This can result in stress and a greater need for loans to meet personal and family needs. However, these factors are not taken into account for the average British citizen when determining loan status. It's important to note that this study only focuses on a small portion of the population and has some limitations.

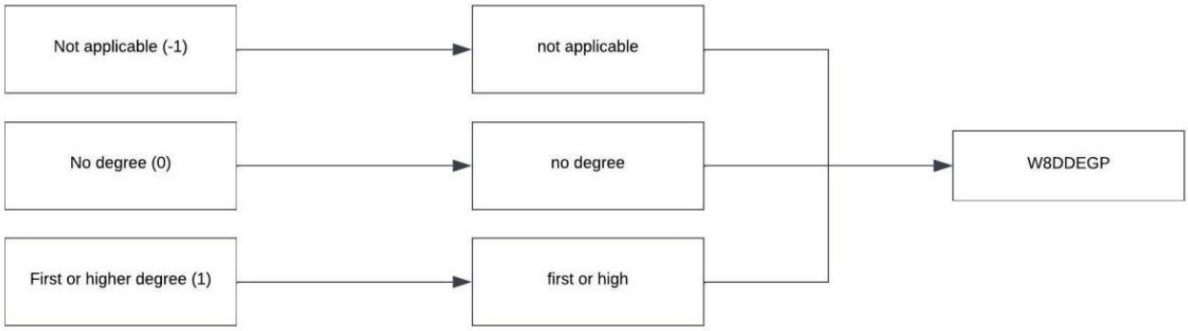
Appendix: The variable merging we did is displayed below

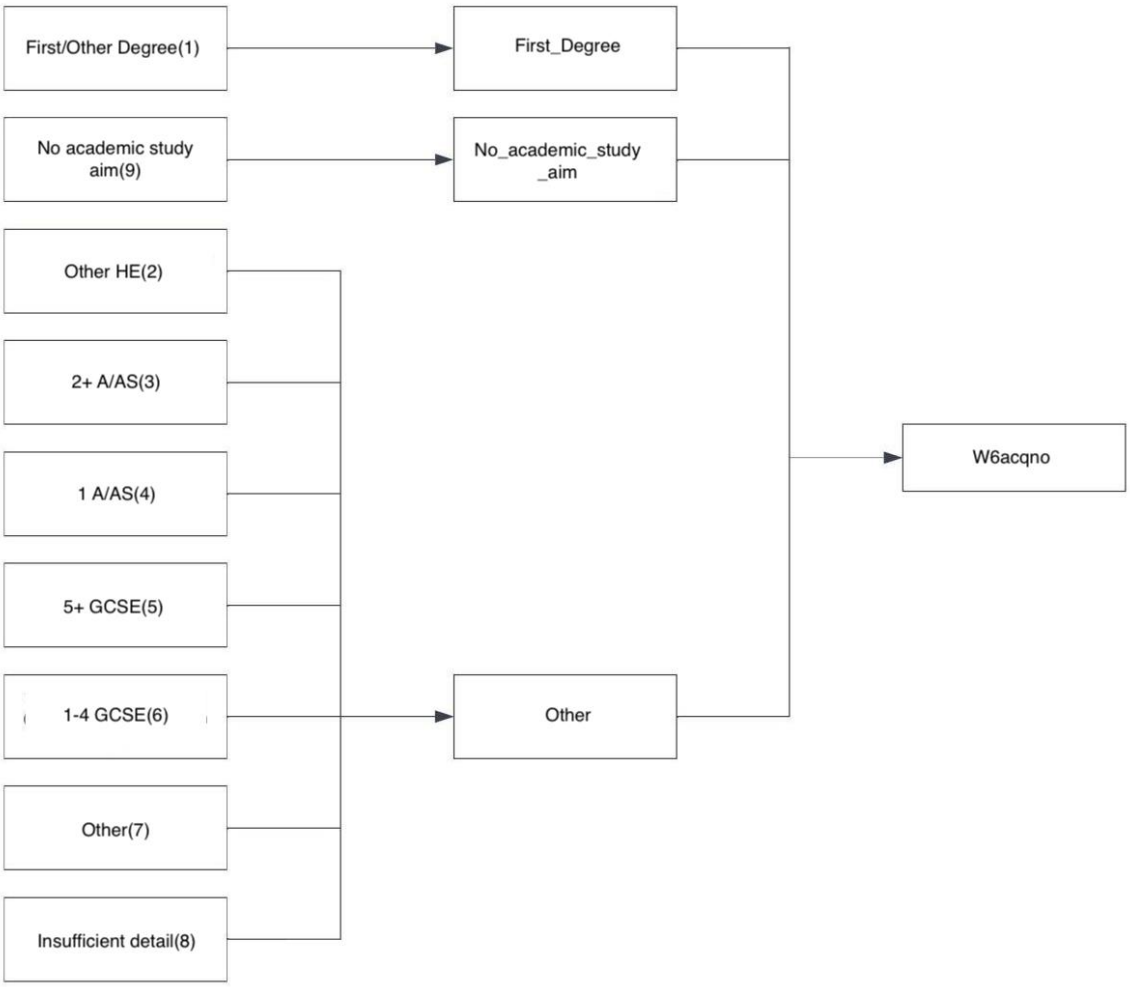



Full-time paid employee (30 or more hours a week) (1)
Part-time paid employee (under 30 hours a week) (2)
Looking after home/ family (10)
Full-time self-employed (3)
Part-time self-employed (4)
Unemployed and seeking work (5)
Full-time education (6)
Temporarily sick/ disabled (if temporarily off sick from job) (8)
Permanently sick/ disabled (9)
Retired from work altogether (11)
Other answers (12)
MP not interviewed (-99)
full_time_employee
part_time_employee
look_after_family
others
W1wrk1aMP


Male (1)
Female (2)
male
female
W8CMSEX


yes (1)
alcoholic_drink
No(2)
no_alcoholic_drink
Refused(-92)
Not applicable(-91)
YP not interviewed (-99)
YP unable to complete CASI section(-96)
Other
W1alceverYP


Not applicable (-1)
not applicable
No degree (0)
no degree
First or higher degree (1)
first or high
W8DDEGP


First/Other Degree(1)
No academic study aim(9)
Other HE(2)
2+ A/AS(3)
1 A/AS(4)
5+ GCSE(5)
1-4 GCSE(6)
Other(7)
Insufficient detail(8)
First_Degree
No_academic_study _aim
Other
W6acqno