TU Dortmund

# Modelling ordinal data Case Studies

Lecturers:
Ms. Maria Iannario

Author: Swapnil Srivastava

Student ID: 230181

September 5, 2022

## Contents

# 1 Introduction

Given the intricate duties that many women continue to play in their houses, few other worldly roles are still disregarded by women. Opening an account and managing their finances at a financial institution is one such instance. In the current situation, when we are working to eradicate the term "gender-specific roles" from society, it is of utmost importance to comprehend whether there is a gender gap, the causes of it, and how to close the gaps.

In this report, we examine the Global Financial Inclusion (Global Findex) Database 2017, the most comprehensive data set available on how adults manage risk, save, borrow, and make payments. Over 150,000 persons in 144 economies—or more than 97 percent of the world's population—were surveyed for the indicators in the 2017 Global Findex database. The entire non-institutionalized civilian population aged 15 and over is the target population. We can further access the specifics of each variable in the data set at: "Overview of dataset".

We focused our research for this report on figuring out whether gender affects who is most likely to have an account with a financial institution or not. In order to do this, we first carry out our descriptive analysis. The gender gap is then examined in relation to several other factors like regions, age, education etc.

# 2 Statistical methods

In the following portion of the current report, the machine learning techniques that will be utilized to analyze and assess the data set, are introduced.

## 2.1 Logistic regression model

Logistic regression is a supervised machine learning technique used for binary classification. In logistic regression model, our main goal is to investigate the relationship and impact of a given set of explanatory variables (also known as independent variables, predictors, or regressors) $x_1, x_2, ..., x_k$ on the binary target variable $y \in (0, 1)$. The expected value of the target binary variable following a Bernoulli probability distribution is given by,

$E(y) = P(y = 0).0 + P(y = 1).1 = P(y = 1)$

The above equation can be rewritten as, $P(y = 1) = P(y = 1|x_1, x_2, ..., x_k) = \pi$ in presence of covariates and the probabilities are modeled as,

$\pi_i = P(y_i = 1) = F(\beta_0 + \beta_1 x_{i1} + .... + \beta_k x_{ik})$

where $\beta_0, \beta_1, ...., \beta_k$ are the unknown regression parameters or coefficients which needs to be determined, $K = (1, 2, ...k)$ are the number of covariates and the domain of the function $F$ is restricted between [0,1].

Choosing the logistic distribution function finally yields the logit model as given below,

$\pi_i = P(y_i = 1) = F(\eta_i) = \frac{exp(\eta_i)}{1+exp(\eta_i)}$

where $exp$ is the natural exponential function and $\eta_i$ is the linear predictor given as,

$\eta_i = \beta_0 + \beta_1 x_{i1} + .... + \beta_k x_{ik}$
(Fahrmeir et al., 2013, p. 33-34)

### 2.1.1 Dummy encoding for independent categorical variables

If the data set contains categorical independent variables $x_i \in 1, ..., c$ with $c$ categories, the dummy encoding is used to model the influence of these variables by introducing c-1 dummy variables in the regression model. These dummy variables can have one of the two values i.e., 0 or 1.

$$x_{i1} = \begin{cases} 1 & \text{if } x_{i1} = 1, \\ 0, & \text{otherwise,} \end{cases} \quad \cdots \quad x_{i,c-1} = \begin{cases} 1 & \text{if } x_{i,c-1} = c - 1, \\ 0, & \text{otherwise} \end{cases}$$

To make the model identifiable, one of the dummy variable is omitted. For the above equation, the dummy variable for category c is removed and this category is then known as the reference category. Direct comparison with the excluded reference category is then used to interpret the effects of other dummy variables.(Fahrmeir et al., 2013, p. 97)

### 2.1.2 Imbalanced classification

If the categorization categories are not approximately equally represented, a dataset is imbalanced, meaning that there are many more data points in the majority class than in the minority class. The majority of the predictions will match the majority class, while the minority class features will be treated as data noise and ignored. This will provide a significant bias in the model. One of the most frequently chosen methods for dealing with an imbalanced dataset is resampling the data. For this, there are primarily two types of methods: undersampling and oversampling. Oversampling is typically favored to undersampling methods. The reason is that when we undersample, we frequently leave out data points that might contain crucial information. The most widely used oversampling technique is called Synthetic Minority Oversampling Technique, or SMOTE which will be used in this project.

**SMOTE resampling technique** SMOTE is a method of oversampling in which synthetic samples are produced for the minority class. The fundamental concept is to produce new synthetic points along the segments that connect a group of neighbors while taking into account the relationships that already exist between samples. Depending on how much oversampling is necessary, neighbors from the KNN (k-nearest neighbors) are randomly chosen. Consider $X_1$ is the sample for which k-nearest neighbors are being identified and let $X_2$ is one of its k-nearest neighbors. The new synthesized data will be given as,
$r_1 = X_1 + rand(0-1) * diff$
where $rand(0-1)$ is a random number between 0 and 1 and diff is the difference between the feature vector ($X_1$) and its neighbors in this case ($X_2$) calculated by any distance metric.(Chawla et al., 2002)

### 2.1.3 Model statistics

Most machine learning models provide a display of fundamental model statistics as well as a number of statistics crucial for evaluating model fit. The fundamental model statistics comprise the model intercept, one or more coefficients, and related standard errors, p-values, confidence intervals, and z statistics which will be discussed in this subsection.

**P-value and significance level** The P-value is a technique that helps in arriving at a statistical judgment while performing hypothesis testing. It is always checked and compared to the significance level to assess whether the null hypothesis should be rejected or not. The p-value ranges from 0 to 1. The minimal value of significance level for which the null hypothesis may be rejected is defined by the p-value. For example, if the p-value is 0.02, then the null hypothesis can be rejected for significance level 0.05 but cannot be rejected for significance level of 0.01 because 0.02 is the least value of significance level for which null hypothesis can be rejected. (Black, 2019, p. 302)

The probability of rejecting the null hypothesis even though it is true (Type-I error) is known as the significance level or level of significance. It is denoted as $\alpha$. The level of significance ranges take values between 0 and 1. The significance level is generally assigned to a value before the beginning of the experiment such as 0.05 which implies that there is a risk of a 5 percent chance of some difference in inferences when there is actually no difference.(Rasch et al., 2020, p. 39)

## 2.2 Collinearity analysis

Multicollinearity occurs when two or more independent or explanatory variables are correlated or linearly dependent on each other. Multicollinearity can affect the model fitting and the regression results as the model parameters become extremely sensitive to small changes in the model and as a result, it becomes hard to check explanatory variables that are statistically significant for the response variable. In order to avoid the multicollinearity issue, we use the variance influence factor (VIF) in this report.

## 2.3 Evaluation metrics

The capacity or the performance of the model to divide or classify two levels or categories of response variable is measured by the evaluation metrics of the model. AUC-ROC curve, cross-validation, confusion matrix and accuracy are just a few examples of classification tools which will be discussed in this report.

### 2.3.1 Classification report

It is an report which includes more parameters derived from the confusion matrix.

- Accuracy: It is a metric which predicts the model's overall accuracy. It is calculated by dividing the correct predictions (TP+TN) by the total number of predictions (TP+TN+FP+FN).
- Precision: Out of all positively predicted outcomes, precision measures how many are actually positive. It is calculated by dividing the true positive (TP) by the total number of predicted positive outcomes (TP+FP)
- Recall/Sensitivity/True positive rate (TPR): Out of all actual positive outcomes, recall measures how many are predicted positive. It is calculated by dividing the true positive (TP) by the total number of actual positive outcomes(TP+FN)
- Specificity/True negative rate (TNR): Out of all actual negative outcomes, specificity measures how many are predicted negative. It is calculated by dividing the true negative (TN) by the tota number of actual negative outcomes (TN+FP)
- F1-score: In reality, when we try to make our model more precise, the recall decreases and vice versa. Both trends are represented by a single value in the F1-score. It is the harmonic mean of Precision and Recall and is given by, F1-score $= 2 * \frac{Precision * Recall}{Precision + Recall}$

(Hilbe, 2015, p. 87-88)

#### 2.3.2 AUC-ROC curve

An indicator of performance for classification problems at different threshold levels is the AUC-ROC curve. The ROC curve plots the True positive rate (Sensitivity) on the y-axis against the False positive rate (1-Specificity) on the x-axis at different threshold values. The capacity of a classifier to differentiate between classes is measured by the Area Under the Curve (AUC), which is used as a summary of the ROC (Receiver Operator Characteristic ) curve. It shows how well the model can differentiate across classes. The higher the AUC the better at differentiating between the positive and negative classes. AUC = 1 indicates that the classifier can accurately distinguish between all Positive and Negative class points. AUC = 0 indicates a poor classifier, it would be predicting all Negatives as Positives and all Positives as Negatives. There is a good possibility that the classifier will be able to tell the difference between the positive class values and the negative class values when AUC lies between 0.5 and 1. This is the case because more True positives and True negatives can be detected by the classifier than False positives and False negatives.(Hilbe, 2015, p. 84-85)

#### 2.3.3 Data splitting and cross validation

**Data splitting** : A method for assessing a machine learning algorithm's performance is the train-test split. The method involves splitting the dataset into two distinct subsets: a training set and a test set to avoid over-fitting or under-fitting of the model and to properly evaluate the model.

- Training set: The dataset that we feed into our model during training to discover any potential underlying patterns and links. The training set should be as representative of the population we are trying to model.
- Test set: The test dataset is used to assess the fitted machine learning model based on the predictions made on the new unobserved data that were not used to train the model.

Overfitting and underfitting are two main factors contributing to poor performance of the model. A model that fits the training set of data too well and performs bad on the new unseen data is said to be over-fit whereas a model is said to be under-fit when it is unable to both model the training data and generalize to new data. Data splitting helps the model to find a sweet spot between over-fitting and under-fitting.Despite the fact that it is a simple technique to use and understand, the accuracy of the model decreases if the dataset is small and if the split is not random. (James et al., 2021, p. 21-22)

**Cross validation** Cross validation is helpful in evaluating machine learning model because it helps avoid the issues stated in the above train-test split evaluation. It is a resampling technique that contains a single parameter, k, that designates how many groups should be created from a given data sample. As a result, the process is frequently referred to as k-fold cross-validation. In this method, the collection of data is randomly divided into k folds or groups that are roughly similar in size. The first group is treated as the validation set and the model is fitted on the remaining k-1 groups. This process is repeated k times, with each iteration treating a different collection of observations as a validation set. Through this approach, k estimates of the test error are produced which is then averaged to get k-fold cross validation estimate.(James et al., 2021, p. 198,206)

- Validation set: The dataset we use to analyze the performance of our model across various model types and hyper-parameter selections is called the validation set. This data set uses the cross validation technique in order to find the best hyper-parameters.

**Hyper-parameter tuning** Model parameters are those that form the model's internal structure and are automatically computed by the model using the data whereas hyper-parameters are tweaked to provide the best fit after learning model parameters from the data. Finding the optimal combination of hyper-parameters to enhance the model's performance is known as hyper-parameter tuning. There are many ways to perform hyper-parameter tuning, one of the mostly used method which will be used in this report is Grid search CV(cross validation). Grid search CV selects a grid of hyper-parameter values and analyzes each one individually. Each iteration tries a set of hyperparameters in a certain sequence. It tracks the model performance when fitting the model with every possible set of hyper-parameters. The best model with the best hyper-parameters is then returned.(Pedregosa et al., 2011)

# 3 Statistical analysis

## 3.1 Understanding the Gender Gap

We can observe from our data set that there were 83593 female respondents out of a total of 154923 respondents, compared to 71330 male respondents. We also look into how many people have accounts with different financial intuitions. While 92925 had an account, 61998 did not, as we discovered. According to the figure 1, 45671 males and 47254 females have accounts. We also see that 25659 men and 36339 women do not have an account. Even though the figures show that more female have accounts, it is important to note that there is a gender imbalance in the original data set. Hence, it is interesting to investigate the gap in relation to the total number of respondents. Figure 1, shows that the gray line (depicting total with respect to gender in each category) rises while the yellow line falls (depicting difference with respect to gender in each category goes down). As a result, we can say that while more women are becoming financially included by opening accounts at financial institutions, the gender gap still remains.

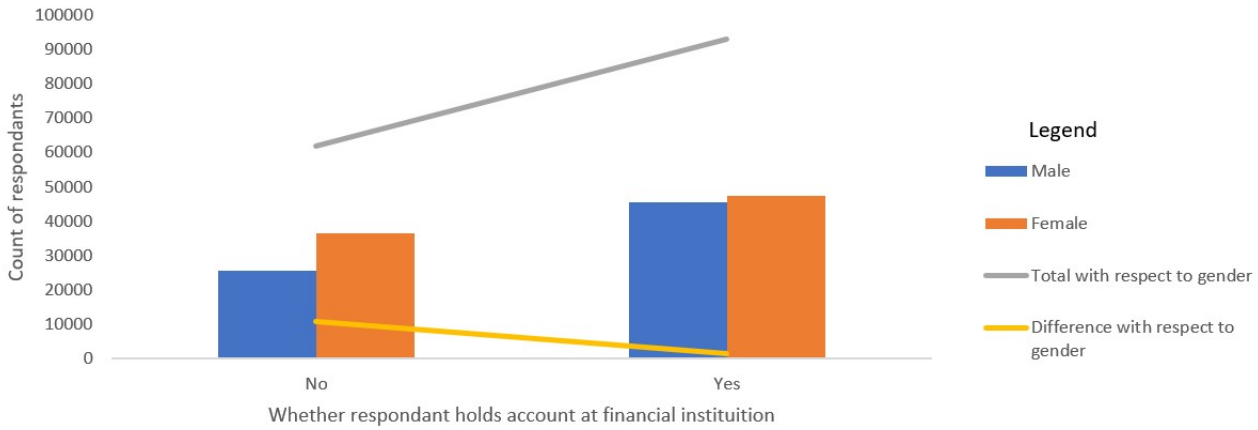


Figure 1: Comparison of Gender in holding an account at financial institution

## 3.2 Factors affecting the Gender Gap

To further investigate this gap, we explore various factors that may be involved. We concentrate on the 61998 respondents who do not have any accounts with financial institutions in order to study this gap and study the affect.

### 3.2.1 Region

One such interesting factor is, regions. From fig 2, we can observe, that the total respondents holding no account at financial institution were majorly from *Sub-Saharan Africa (excluding high income)* region and least from *High income: OECD* region. To understand the gender gap in these regions, we see the yellow line in fig 2. It is here we observe that while *Latin America and Caribbean (excluding high income)* has the most gender gap, *High income: nonOECD* has the least.

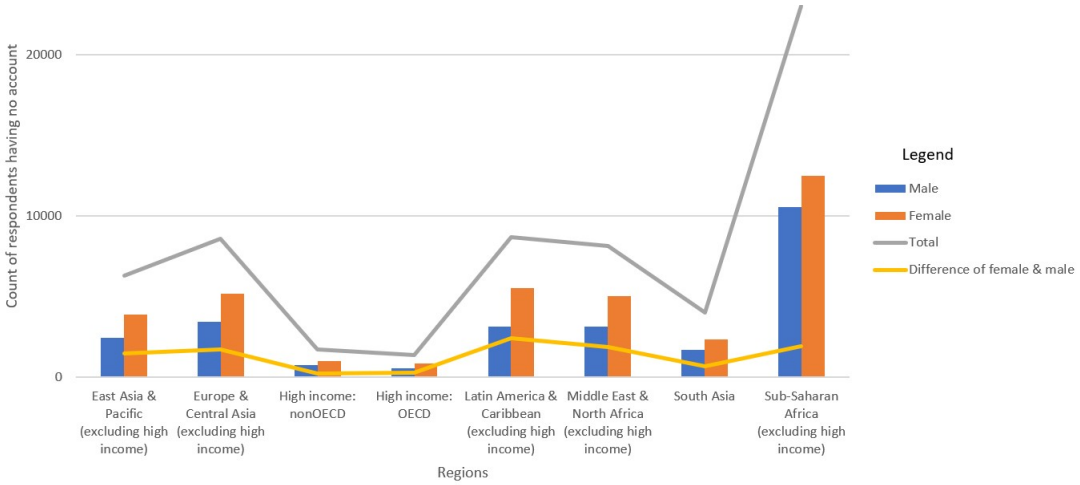


Figure 2: Comparison of Gender with respect to regions in holding an account at financial institution

### 3.2.2 Age

The other interesting factor is age. We introduced a column named, Age group depending on the values in the column "age". It has 4 categories as shown in figure 3. In this figure, we can observe that while the gray line (representing total number of respondents) is almost straight line sloping downwards, stating that age is inversely related to the count of people having account at financial institution i.e. older people are less likely to have an account at financial institution than younger people. To understand the gender gap, we focus on the yellow line which can be seen peeking for age group *Between 31 and 50 Years* while reaching its lowest value for age group *Greater than 71 Years* i.e., we can say, that for age group Between 31 and 50 years, more men are likely to have an account than women.

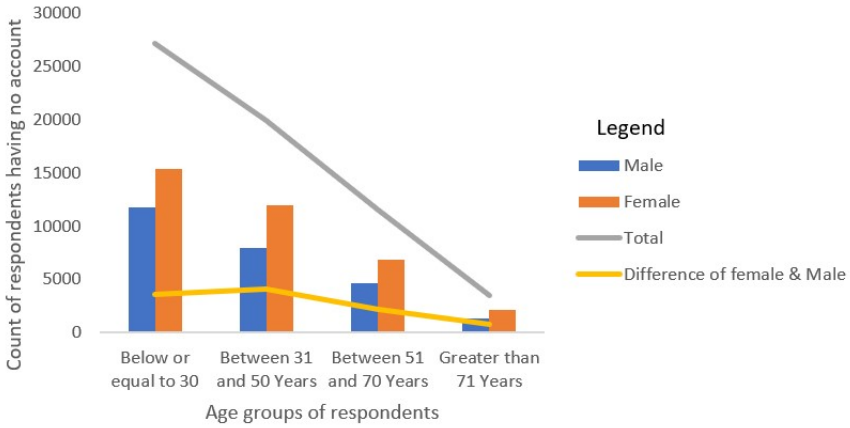


Figure 3: Comparison of Gender with respect to age group in holding an account at financial institution

### 3.2.3 Education

Another factor might be education. The data we received of the respondents also contained information of their education level. The levels were divided into following 4 categories: completed primary or less, secondary, completed tertiary or more and some default categories like dk and rf. We plotted a graph with respect to gender and if they had account at financial institution. It can be seen from figure 4, that the total number of people who do not hold an account are mostly who have completed primary level education or less. In other words, people who are less educated generally do not hold an account and it is also the case where we can see the gender gap as the highest. This means that there are most number of women who are less educated, do not hold an account as compared to men. While this number keeps on decreasing as the education level rises.

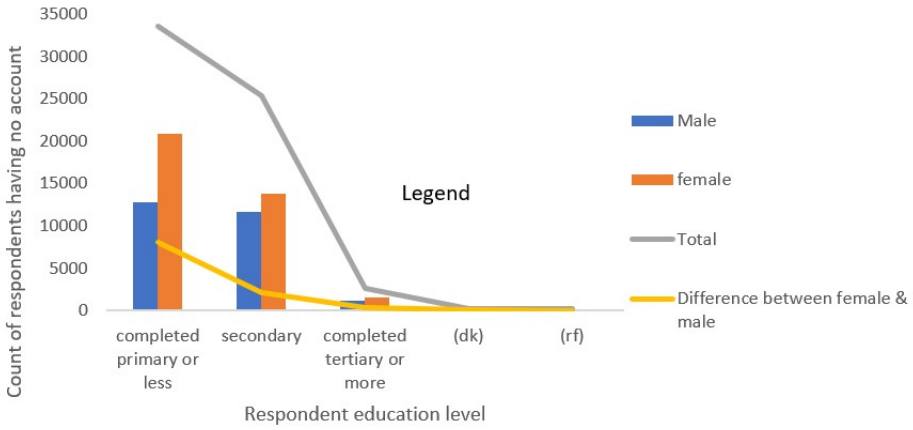


Figure 4: Comparison of Gender with respect to Education level in holding an account at financial institution

### 3.2.4 Employment status

Most interesting factor was the employment variable. We had two categories of respondent i.e., either they were in workforce or not. As we can see from fig 5, while the total number of people not having an account increased from being out of workforce to being in a workforce, the gender gap drops to **negative**. It implies that more women who are in workforce have accounts than the male working. Hence, it can be said that while the number of women working in a workforce was less than the male, the number of women working in a workforce and having an account at financial institution was more than such men, thus reducing the gender gap.

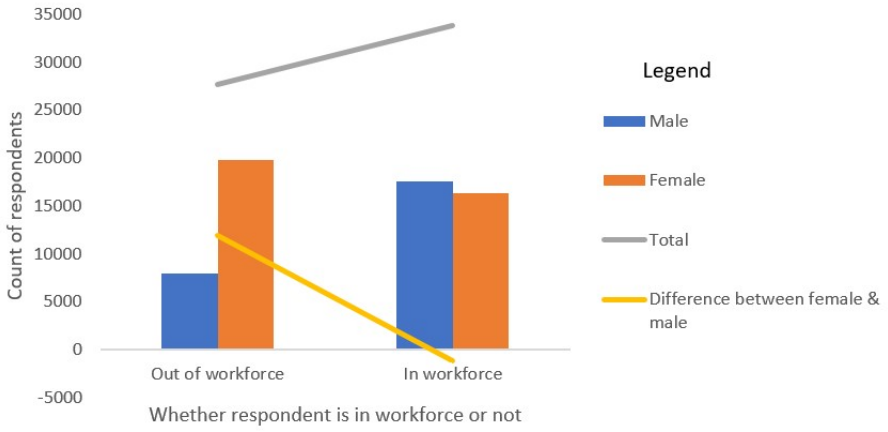


Figure 5: Comparison of Gender with respect to employment status in holding an account at financial institution

### 3.2.5 Other reasons

We further see from the data provided and try to understand the reasons why people do not hold an account. The graphical representation can be seen from figure 6. We can observe here, not having enough funds to open and maintain accounts is a major reason why people do not have accounts and the gender gap is the highest for it as well. To be sure, many financial institutions charge a fee for opening and maintaining a bank account, which can be intimidating to potential customers. Aside from that, transaction costs at many financial institutions remain high and strict personal identification requirements deter many women from opening accounts. We also observe that more women avoid opening an account because either one of the family member has an account or because of their mindset that they don't need to have one. Other subjective factors, such as religious factors, trust contribute to women's low rates of inclusion in traditional financial institutions.

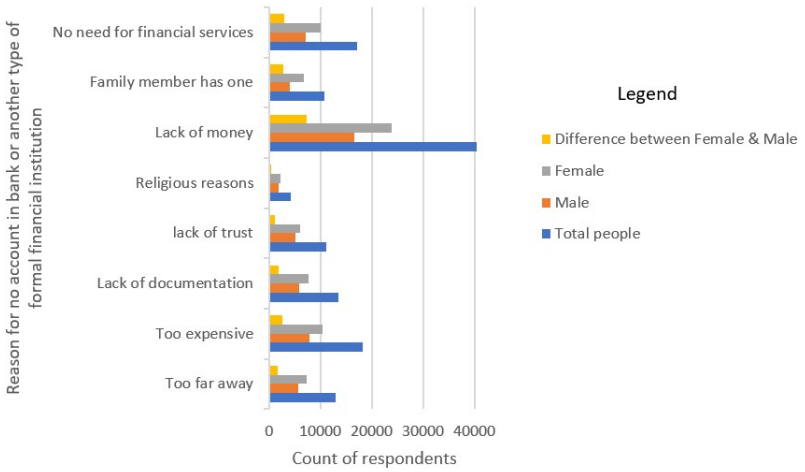


Figure 6: Reasons for not holding an account at financial instituition

## 3.3 Model Fitting

### 3.3.1 Collinearity analysis

After studying this, we now have a better understanding of the relationship between the various other variables that may influence a respondent's decision to open an account. For this, we generate the heat map for the multicollinearity between variables. Before, this we generate dummy variables for all our categorical variables like region, gender, income etc. For example, for gender: the gender_female becomes our reference variable and so on.

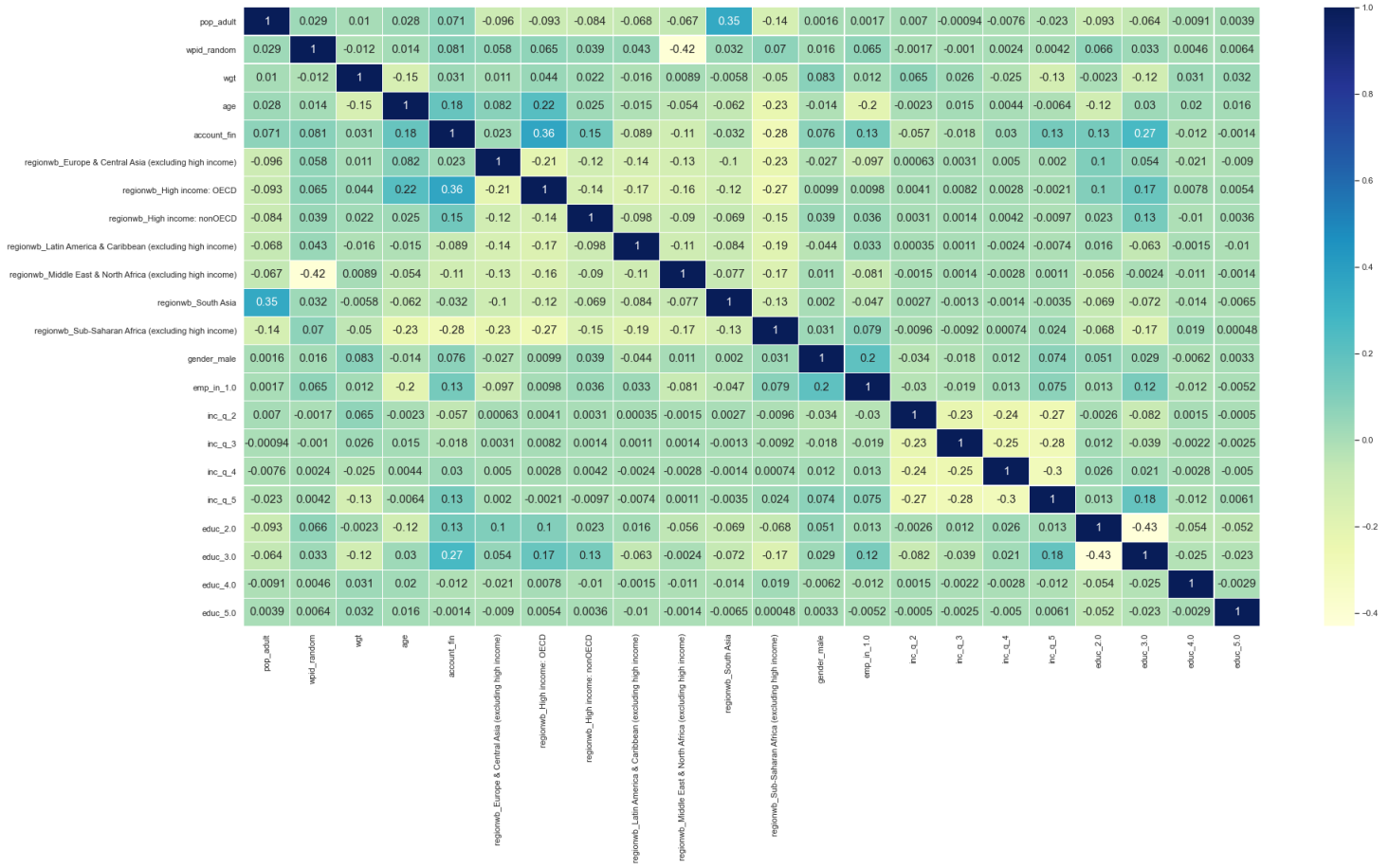


Figure 7: Heat map for multicollinearity between variables

From fig 7, we can observe that none of the variables highly correlated as none of them are close to -1 or 1. It is here, that we also observe that *gender_male* has slightly positive correlation with *account_fin* i.e. 0.076. Hence,we take all the variables for our model building.

### 3.3.2 Evaluation metrics

We now divide our data set using 70% as our training data set and remaining 30% as test data set. When we fit our logistic regression model here, we find that the accuracy of this model is 51.37% which is quite low.

Table 1: Classification report

| Account Class | precision | recall | f1-score |
|---|---|---|---|
| 0 | 0.35 | 0.25 | 0.29 |
| 1 | 0.58 | 0.69 | 0.63 |

From table 1, we can see that the precision & recall both are low. Also, there is an imbalance in the prediction of both the classes. This may be due to the fact that the original data that we had was not perfectly balanced. So,in order to achieve a better model, we use the **class_weight** system as now instead of using weight as 1 for each class we adjust weights inversely proportional to the class frequencies. We also perform **hyper parameter tuning** by introducing L2 regularization with primal formulation and newton-cg algorithm to use in the optimization problem. We also use 5 fold cross validation for our new model and observe the result. We now see that the accuracy has risen upto 73.89% and the classification report reads as in table 2

Table 2: Classification report after hyper parameter tuning

| Account Class | precision | recall | f1-score |
|---|---|---|---|
| 0 | 0.65 | 0.75 | 0.70 |
| 1 | 0.82 | 0.73 | 0.77 |

We see that now the precision and recall are balanced and we also see that the f1 score is now nearer to 1 than our previous model, making it a better one. The ROC curve for our model can be seen in figure 8. We can observe that the Area under the curve (AUC) is 0.74 which is acceptable.

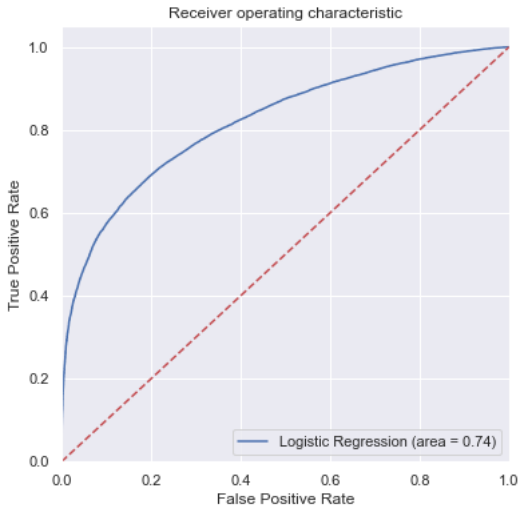


Figure 8: ROC Curve

Hence, from all the above parameters, we can say that our model is able to perform a decent classification i.e. it has the ability to classify the respondent as whether they have an account at a financial institution or not. So, we now we look at the snippet of coefficient summary to understand how gender is playing the role which can be seen from 3.

We can see that the coefficient of *gender_male* variable is 0.07 i.e. it means that on an average for a respondent being a male, one would expect e(0.07) i.e. 1.07 times more probability of having an account at financial institution than female.

Table 3: Model summary

| | coef | std err | z | P > \|z\| |
|---|---|---|---|---|
| const | 2.964 | 0.054 | -54.667 | 0.000 |
| gender_male | 0.070 | 0.016 | 4.436 | 0.000 |

## 4 Summary

This project report focuses on figuring out whether gender affects who is most likely to have an account at a financial institution or not. The data set which we used for our analysis is the Global Financial Inclusion (Global Findex) Database 2017. The target audience was people aged 15 years or more.

It is here that we found that while 58.6% of the respondents who didn't have an account at financial institution were female while only 41.4% were male. To understand this gap more, we tried to dig this gap more. We judged this gap on the factors like region, age, education level, employment status and others. We found out that Latin America and Caribbean (excluding high income) region has the highest gender gap while High income: nonOECD region has the least. We further found that even though as the people gre older, they were less likely to have an account than younger people, it was slightly different while comparing genders as well. While the gender gap peaked for respondents of age between 31 to 50 it lowered for people greater than 71 age. The most interesting factor was the employment status of respondent who said no to having financial account. We found out that there exists a gender gap between people who were in workforce, but it was reverse. In other words, women who were in workforce are more likely to have an account at financial institution than male working at financial institution. Apart from this, we also tried to study the various reasons which may have lead to no financial accounts.

Furthermore we tried to fit a logistic regression model to find out how these factors may exactly affect the respondent to have an account or not. We performed hyperparameter tuning to tune our model as our original model was not that accurate. After that, we studied the most important coeffecient for this report i.e. gender and found out that on an average for a respondent being a male, one would expect 1.07 times more probability of having an account at financial institution than female.

Hence, to conclude, there still exists a gender gap where percentage of men who own an account at financial institution is almost 30% more than women. In order to bridge this gap, we can further drill down the various reasons and pin point the pain areas. Government may take necessary actions to remove them and try to bridge the gap.

## Bibliography


Ken Black. *Business statistics: for contemporary decision making.* John Wiley, 2019.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–330, 2002.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. *Regression: Models, Methods and Applications.* Springer, 2013.

J.M. Hilbe. *Practical Guide to Logistic Regression.* Chapman and Hall/CRC, 2015.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R.* Springer, 2021. URL `https://faculty.marshall.usc.edu/gareth-james/ISL/`.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Rasch et al. *Applied statistics : theory and problem solutions with R.* John Wiley and Sons, USA, 2020.