Machine Learning Project

George Massey ID: 230973275

January 2024

1 Introduction and Problem Statement

I have chosen to do my project on the Bank dataset which means my task is to implement, describe and present (binary) regression/classification models of your choice to predict which customers will churn (i.e., leave the bank). The goal is to develop (binary) regression/classification models using any number of the variables provided, which describe customers' features, to predict their churn. The dataset includes 11 customer features which are: id, credit score, country of residence, gender, age, tenure (years at the bank), bank balance, if they own a credit card, if they are an active member and estimated salary. Firstly, the dataset has informed me that customer id, as expected, has no effect on churn and so I have ignored this feature in my project.

2 Analysis of the Dataset

To start my analysis of the Bank dataset I read the dataset csv file into a pandas data frame in order to assist my visualisation of the data. I then set up subplots of each features distribution as shown in Figure 1, which illustrated that:

- Credit score follows a normal like distribution with a mean around 650.
- There are only three countries of residence and there is twice as many customers in France than in Germany or Spain.
- There are 10,000 people in the dataset with there being slightly more males.
- Majority of customers are between the ages 25-50 with very few people over 65.
- Tenure is fairly uniform with around 1000 people at each year but only 400 people have joined in the last year and 1400 people have been there for 9+ years.
- Bank balance is skewed by the majority of the customers having around €0-25,000 with the rest of the customers being somewhat normally distributed with a mean of around €120,000.
- The credit card distribution is split 70:30 towards customers having a credit card.
- Very slight majority of customers being members of the bank.
- Customers estimated salary is almost uniform with around 1000 people being in each €20,000 interval from €0-200,000.
- 90% of customers have bought either 1 or 2 products from the bank.
- Around 20% of customers churn. This is the target variable!

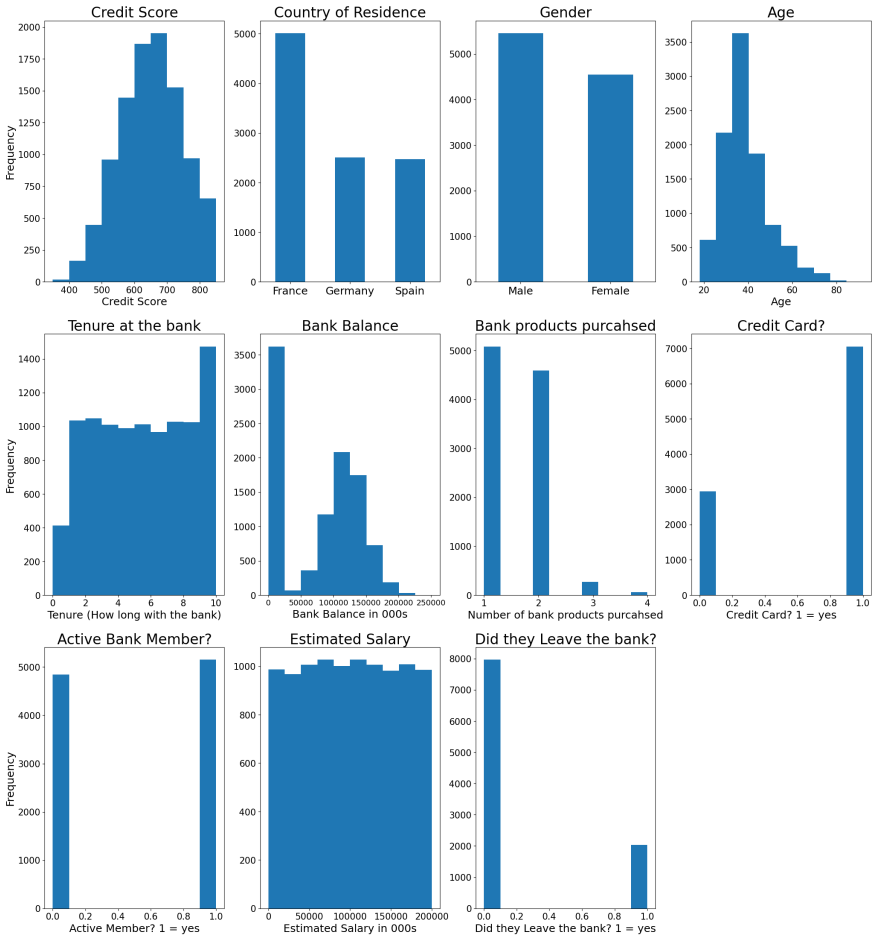

Figure 1: Distributions of Features and Target Variables.

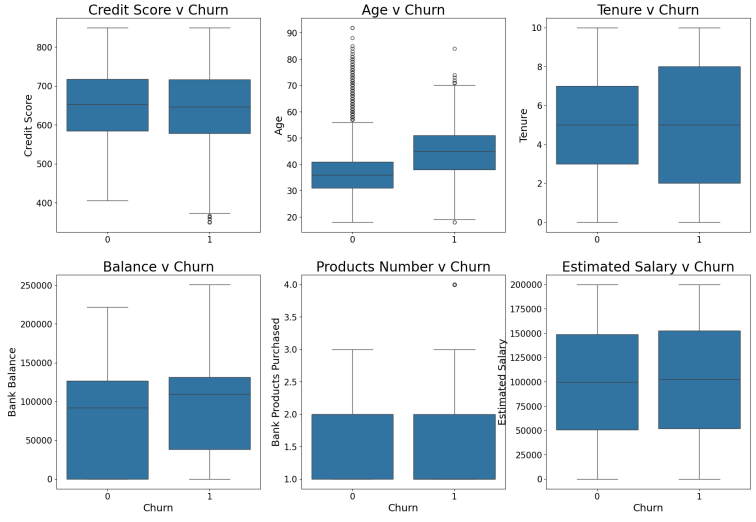

Figure 2: Effect of Numerical Features on Whether Customers Churn.

Next, I looked at the effect features had on the target variable. For the numerical features, shown in Figure 2 I created a Seaborn box-plot for customers that stayed at the bank and those that left. My main takeaways are as follows:

- Credit score box-plots are almost identical with a similar median as above at around 650, suggesting a minimal effect on the target variable. However the churn box-plot is slightly negatively skewed towards the lower credit score, with a few outliers below 400.
- Age has very different box-plots with the customers who stayed having a median age of around 36 compared to 43 of customers who leave. However, the non-churn box-plot has many more upper end outliers which suggests that young and old customers are more likely to stay where as middle aged customers are more likely to leave.
- The median for both Tenure box-plots are 5 but the churn box-plot has a much larger inter-quartile range suggesting that there's more variability to the length of stay at the bank for customers who churn.
- Large difference in the shape of the balance box-plots, with the churn median being around €110,000 compared to €90,000, suggesting the more money in your account the more likely you are to leave the bank. Furthermore, the non-churn boxplot is heavily skewed towards 0 suggesting that the majority of customers with a low bank balance stay at the bank.
- Lastly, both number of products purchased and estimated salary have identical boxplots suggesting they will not have much of an effect on whether customers churn.

I then looked at the categorical features of the dataset where box-plots would not be an effective visualisation so instead I created bar charts illustrating the percentage of customers leaving the bank in each category, with the highest category being highlighted pink. Figure 3 below shows that 32% of customers from Germany left the bank, which is double compared to the 16% of both France and Spain. Figure 4 displays that 25% of Female customers left the bank compared to 16% for Males. Figure 5 shows that 27% of customers who were not active members left the bank compared to 14% of members. These three plots all suggest that these features do have a reasonable effect on whether customers churn. I also, similarly plotted for whether customers own a credit card but the bars were identical and therefore will be unlikely to have an effect on churn.

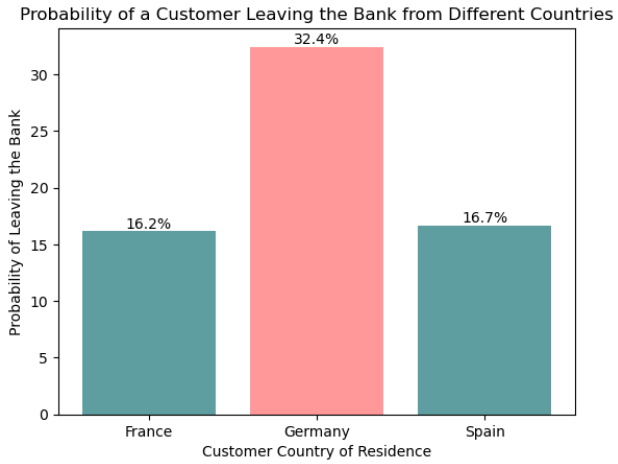

Figure 3: Effect of Country of Residence on Whether Customers Churn.

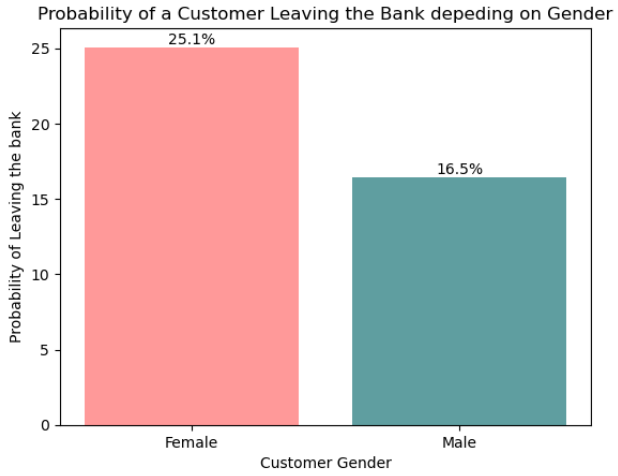

Figure 4: Effect of Gender on Whether Customers Churn.

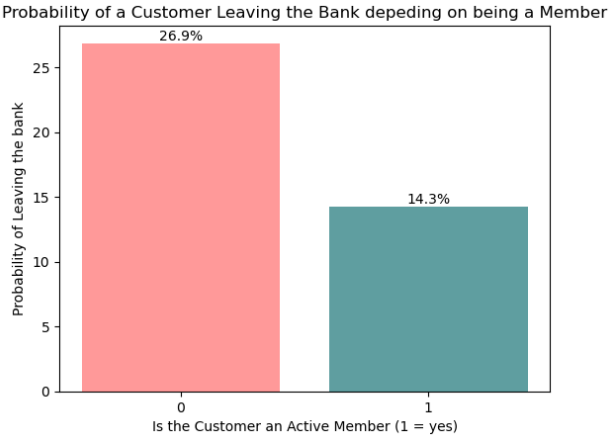

Figure 5: Effect of Membership on Whether Customers Churn.

Finally, as Figure 1 suggested that age has an interesting effect on customers leaving the bank, I wanted to visualise this in a scatter plot. To do this, as shown in Figure 6 I compared age to credit score and highlighted the dots of the customers that left the bank in red. This plot emphasises how the majority of red points (customers that leave) are located in the age region between 35 and 55. To conclude this initial analysis on the dataset I now have a good idea of which features may have the greatest effect on churn in our models. These are: age, balance, country, membership and gender.

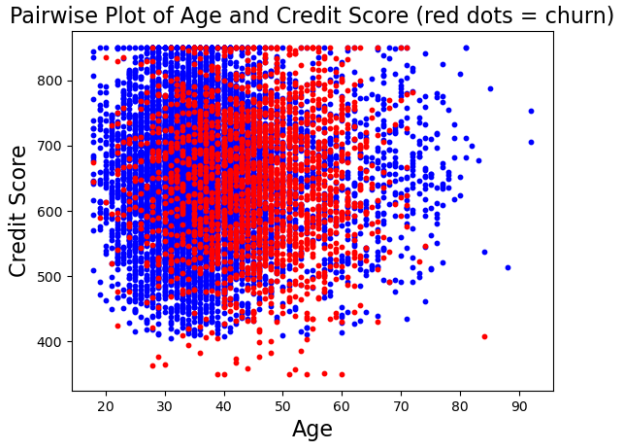

Figure 6: Pairwise Comparison of Age, Credit Score and Whether Customers Churn.

3 Methods of Modelling

The ML algorithm I will use is Binary Logistic Regression. I am using this because I need to predict a binary categorical target variable, whether or not customers churn (denoted 0 for stay and 1 for leave). My main goal is to find which features have the greatest impact on customer churn, with finding the best model as a sub goal. Once I find my feature weight coefficients I will then bootstrap my data to create more instances of the sample and provide better estimations. I intend to repeat this process for multiple Logistic Regression models but each time I will remove the features which are having a minimal effect on the target variable.

Before I start my regression I turned any categorical string variables into binary variables. Therefore, gender became: 0 for Male and 1 for Female. Also I had to split the country variable into three separate variables: country_germany, country_spain and country_france, where 0 denoted a customer was not from that country and 1 denoted that they were. I now have all numerical variables, however they are not on the same scale so I standardised the data to allow all variables to have equal importance in my regression. For example, credit card has a scale from $[0, 1]$ where as estimated salary goes from $[0, 200, 000]$, which will result in them having wildly different means and standard deviations and will make the logistic regression coefficients difficult to interpret. Therefore the result of the standardisation is that all features will be re-scaled to have a mean of 0 and standard deviation of 1 (Perra, 2023a).

The first step of the logistic regression was to pretend as though I was doing linear regression and create my data matrix from the inputted dataset, where the rows represent each feature and the column each sample, with the first column being ones. Next I got the model function: $\langle \phi(\mathbf{x}_i), \mathbf{w} \rangle$, denoting the dot product of each data sample ($\mathbf{x}_i$) and the weights (w) which gives us our predicted values of our target variable (y) (Perra, 2023b). However, as classic MSE regression is not a good metric for this classification problem (Perra, 2023c), I applied the Sigmoid function: $\sigma(x) := \frac{1}{1+e^{-x}}$ to the linear model function above to convert

the predicted values of churn into probabilities based on the inputted feature weights. As the probability approaches 1, the more confident the model is of the customer leaving the bank (Piduguralla, 2023).

As I plan to use gradient descent to solve the logistic regression and find the optimal weights, I first need to find the cost function to minimise, and its gradient. The Binary Logistic Regression Cost Function is:

$$\mathrm{L}\left(\mathbf{w}\right) = \frac{1}{s}\left(\sum_{i=1}^{s} \log\left[1 + \exp\left(f\left(\mathbf{x}^{(i)}, \mathbf{w}\right)\right)\right] - y_i \cdot f\left(\mathbf{x}^{(i)}, \mathbf{w}\right)\right)$$

Where $\mathbf{x}^{(i)}$ is a vector representing the i-th data sample and f is the model function we introduced above. Furthermore, we will need to use the gradient which is as follows:

$$\nabla \mathrm{L}\left(\mathbf{w}\right) = \frac{1}{s}\left(\sum_{i=1}^{s} \phi\left(\mathbf{x}^{(i)}\right) \cdot \sigma\left(\left\langle \phi\left(\mathbf{x}^{(i)}\right), \mathbf{w}\right\rangle\right) - y_i \cdot \phi\left(\mathbf{x}^{(i)}\right)\right),$$

Where $\phi\left(\mathbf{x}^{(i)}\right)$ represents each sample from the new data matrix I have created with an additional column with value one.

As mentioned the method I have used to solve the classification problem is gradient descent. Gradient descent is an iterative procedure that, in this context, will give us the optimal weights to minimise the cost function above. The formula for gradient descent is $w_{k+1} = w_k - \tau \nabla L(w_k)$ where w_{k+1} represents w at the $k + 1^{th}$ iteration and τ is some step parameter which should follow $\tau \leq \frac{s}{\|X^\top X\|}$ (Perra, 2023d). As we know the function is convex, this formula computes the gradient (the point of max growth) at each iteration and subtracts from w which will move us towards the global minima of the function. Once I found my optimal weights from the gradient descent I then sampled my original data with replacement to create more instances (bootstrapping) to provide a better estimation of the weights (Perra, 2023e). I did this by sampling 80% of the original data (8000 samples) and repeating this to make 500 instances.

4 Results of Prediction

My first logistic regression model included all features of the dataset. After bootstrapping the average classification accuracy of the model across the 500 samples was 81.20% and this gave the following weights, represented as a boxplot in Figure 7. As you can see here the features age and active_member have the largest effect on the target variable with median coefficients of 0.78 and -0.57 respectively. For my next logistic regression I filtered out any features that had an absolute median coefficient of less than 0.1 and therefore a very minimal effect on churn. These features were: credit_score, country_spain, tenure, products_number, credit_card and estimated_salary. This was all to be expected from my plots above in part 2. However country_france looked to have a slightly less effect than country_spain in Figure 3 with 16.2% of people from France leaving the bank compared to 16.7% of Spain, but the opposite is suggested in the regression with country_france's absolute coefficient being the only one out of the two to be greater than 0.1 and therefore stay in the model. This could be explained by France having double the amount of people in the dataset and therefore the model can be more sure of its effect.

The second model therefore included the six remaining features and actually had a slightly worse average classification accuracy of 81.19%. The main takeaway from this model is that the feature country_france now has a negligible effect on churn whilst the other features remain similar. However, for my next model I was stricter and removed any features with an absolute median value less than 0.2, which was indeed country_france and balance. Balance being removed was quite interesting as earlier in my visualisations I bookmarked it as a feature which seems to have a potential effect on churn.

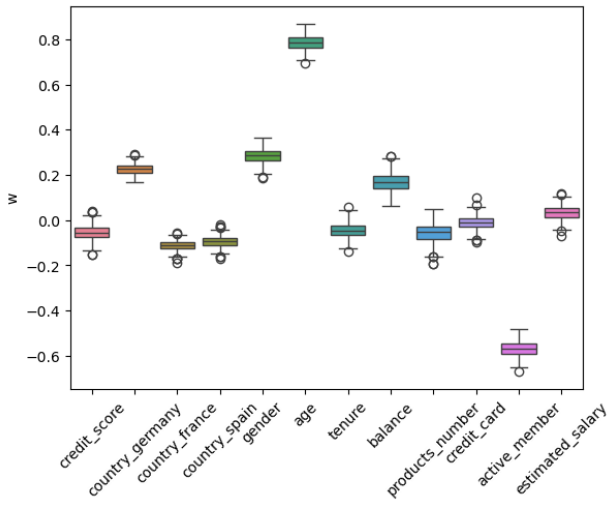

Figure 7: Box plots of Bootstrapped Weights for every Feature.

My third model included only 4 of the features and had the lowest average classification accuracy so far of 81.18%. The weights here are shown by the box-plot in Figure 8 and the main difference here to the last model is that the feature country_germany seems to have become more important to predict customers leaving the bank. For my final model I removed any features from the previous model that were below 0.3, which was only gender. This model had the highest average classification accuracy of 81.27% and the bootstrapped weights of the three variables are shown in Table 1.

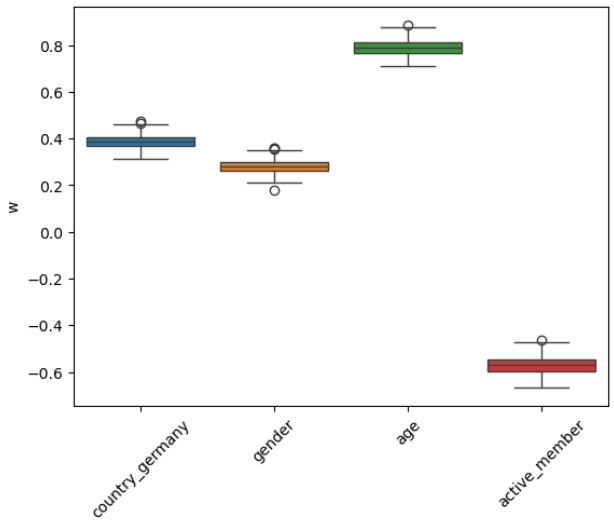

Figure 8: Box plots of Bootstrapped Weights for selected Features.

Table 1 here highlights quite clearly the rank of importance of the top three features on the effect of the outcome. To better interpret what these coefficients are actually telling us I have used the Odds Ratio (OR). The OR is the odds of the target variable occurring given an increase of 1 standard deviation in a selected predictor variable. The OR is calculated by $OR_p = \exp_p^w$ where p is the variable p. If the result

Feature	Weight ($\hat{w}$)
Age	0.79
Active_Member	-0.58
Country_Germany	0.39

Table 1: Table of Selected Features' bootstrapped weight coefficients in Final Model.

is > 1, exposure to an increase in this variable increases the odds of the target variable and the reverse for results < 1 (Szumilas, 2010). So in this context, using Age as an example, I calculate it's OR to be $\exp^0 .79 = 2.20$ which means this model suggests that a customer is 2.20 times more likely to leave the bank given an increase of one standard deviation in their age. Alternatively, it can be seen that a customer being an active member will decrease the odds of the customer leaving the bank (OR = 0.56).

5 Conclusion

To conclude, the problem I was faced with was to implement a model to predict what customers will leave the bank. I did this through binary logistic regression and making use of bootstrapping to give my model more support. I initially modelled all 12 predictor variables and with each new model I removed some of the least effective features at predicting churn. This left me with, as I mentioned before, my final model (3 predictor variables), which was on average the best performing model at predicting the churn of customers. This could be explained by the earlier models potentially over-fitting due to including redundant features that have a minimal effect on churn.

However, my main goal was to discover which features were the most effective in predicting churn. From my models it seemed to suggest that, in order of importance, customers age, whether they were a bank member and whether they resided in Germany were the three most effective features at predicting whether a customer would leave the bank. This was unsurprising to me as my earlier visualisations in Figures 2,3 and 5 had already suggested that these features must have some impact. I was surprised that the best model didn't include the customers gender and their bank balance as these were features that looked to have a significance difference in my initial plots and also had somewhat significant coefficients in my first model (0.28 and 0.17 respectively). What further intrigued me was that my initial model placed gender as having more of an impact than the customer residing from Germany with coefficients 0.28 and 0.22 respectively, but as we removed other features, customers residing from Germany became more important, ending with a weight coefficient of 0.39.

Although my model was somewhat successful in fitting the data this doesn't necessarily mean it will be useful for predicting whether future customers will leave the bank. One of the reasons is that there could be other features that effect churn that I have not used in the classification. Furthermore, I only used one type of solution to this problem and the bank could use other ML algorithms like for example, random forest to compare and evaluate the performance against my logistic regression.

6 References

Bank Dataset: `https://qmplus.qmul.ac.uk/pluginfile.php/4077403/mod_folder/content/0/BANK.zip?forcedownload=1`

Perra, N. (2023a), 'Coursework 6 solutions' [Jupyter Notebook], *MTH786P: Machine Learning with Python*. Queen Mary University of London. 19 November. `https://qmplus.qmul.ac.uk/mod/resource/view.php?id=2452298`

Perra, N. (2023b), 'Coursework 9 solutions' [Jupyter Notebook], *MTH786P: Machine Learning with Python*. Queen Mary University of London. 12 December. https://qmplus.qmul.ac.uk/mod/resource/view.php?id=2467607

Perra, N. (2023c), 'Classification tasks' [Pdf], *MTH786P: Machine Learning with Python*. Queen Mary University of London. 20 November. https://qmplus.qmul.ac.uk/mod/resource/view.php?id=2451016

Perra, N. (2023d), 'From ridge regression to the LASSO' [Pdf], *MTH786P: Machine Learning with Python*. Queen Mary University of London. 17 November. https://qmplus.qmul.ac.uk/mod/resource/view.php?id=2437648

Perra, N. (2023e), 'Interpreting regression models and logistic regression' [Pdf], *MTH786P: Machine Learning with Python*. Queen Mary University of London. 24 November. https://qmplus.qmul.ac.uk/mod/resource/view.php?id=2456565

Piduguralla, S. (2023), *Understanding the Sigmoid Function in Logistic Regression: Mapping Inputs to Probabilities*, Linkedin, accessed 23 January 2024. https://www.linkedin.com/pulse/understanding-sigmoid-function-~:text=The%20sigmoid%20function%20is%20essential,model%20fits%20the%20training%20data.

Szumilas, M. (2010), *Explaining Odds Ratio*, National Library of Medicine, accessed 23 January 2024. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/#:~:text=An%20odds%20ratio%20(OR)%20is,the%20absence%20of%20that%20exposure.