Credit Risk Model Validation Task

Augustė Striogaitė, 2024

1.Data manipulation task

1.1. Select random subsample of data set;

A subsample of 100 entries was randomly selected from the bank dataset.

1.2. Filter desired rows using simple and more complex conditions;

Simple Filter: Rows where the age is greater than 43 were selected from the sample.

Complex Filter: An additional filter was applied to the sample to select rows where the age is over 43, either the balance is greater than 1000 or the response variable y equals "yes", and more than one marketing campaign was directed at the client.

1.3. drop unnecessary variables, rename some variables;

The month variable was removed from the filtered dataset;

The job variable was renamed to occupation.

1.4. Calculate summarizing statistics (for full sample and by categorical variables as well);

Summarizing statistics for full sample: General statistics such as the mean and median of age and balance were calculated for the entire bank dataset;

Summarizing statistics by categorical variables: Similar statistics were grouped by job and education categories to explore variations across different types of jobs and educational backgrounds.

1.5. Create new variables using simple transformation and custom functions;

Simple transformation: A new variable age_group was created by categorizing ages into predefined groups (e.g., Under 25, 25-35, etc.), which helps in demographic segmentation for analysis.

Custom functions: A custom function was applied to categorize the balance into groups such as Negative, Low, Medium, and High. This categorization simplifies the balance variable into more meaningful segments.

1.6. Order data set by several variables.

The dataset was sorted by age, balance (in descending order), and occupation.

2. Data visualization task

2.1. Correlation matrix

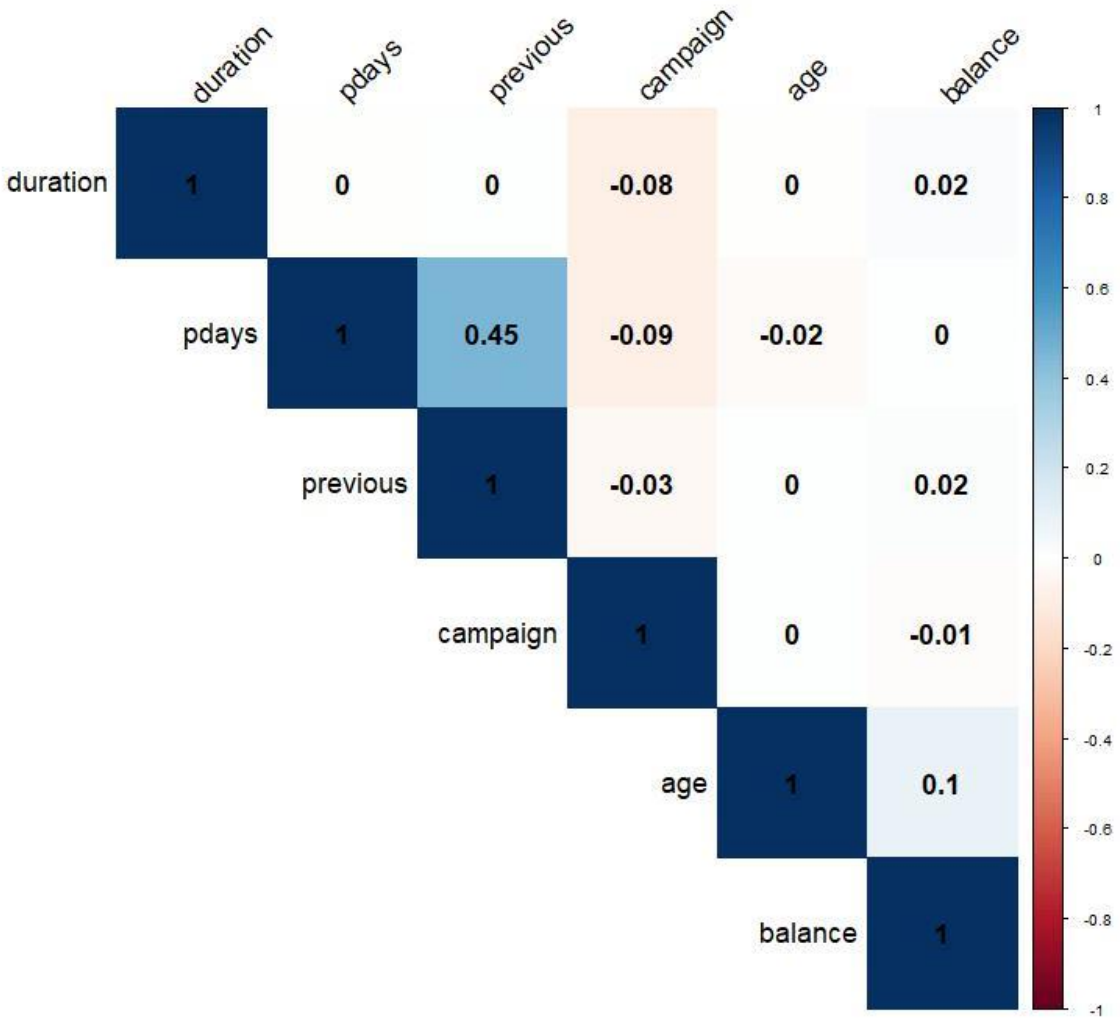

Figure 1 Correlation matrix

Each cell in the matrix *(figure 1)* shows the correlation coefficient between two numeric variables. The coefficient ranges from -1 to 1, where:

- 1 indicates a perfect positive correlation, where as one variable increases, the other also increases.
- -1 indicates a perfect negative correlation, where as one variable increases, the other decreases.

- 0 indicates no linear correlation between the variables.

Strongest positive correlation was between pdays and previous: 0.45. This suggests that clients who were contacted more frequently in previous campaigns also tended to have more days passed since they were last contacted.

Notably, there are slight negative correlations between campaign and previous (-0.03) and campaign and pdays (-0.09). This could imply that current campaigns are slightly less likely to target clients with extensive prior contact.

However, most pairs of variables show very weak correlations close to 0. This suggests little to no linear relationship between these variables

2.2 Age histogram

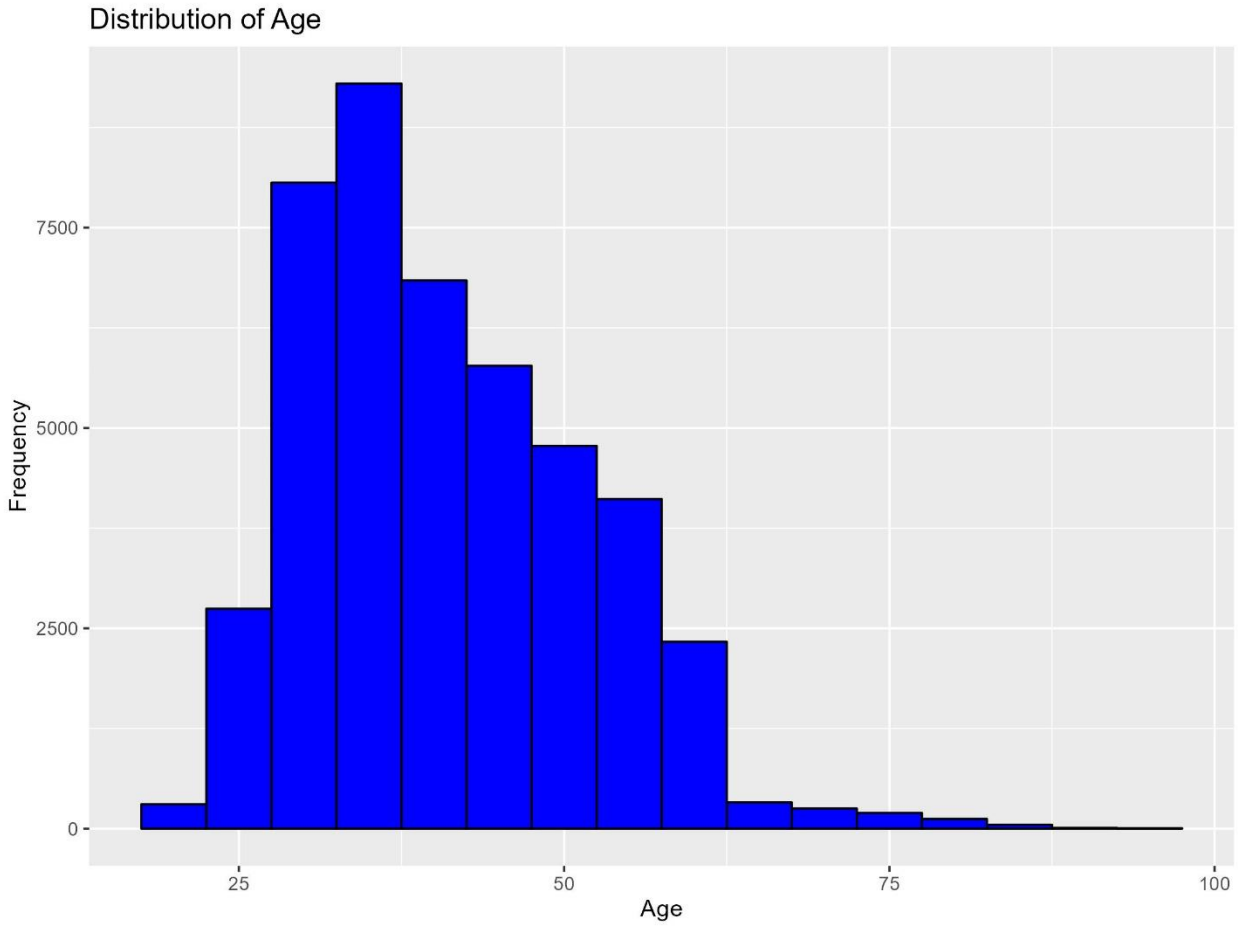

Figure 2 Age Histogram

The histogram *(figure 2)* is right skewed: meaning there are significantly more clients whose age is up to 50 than those whose age is 50 and more. Histogram peaks in the age groups from late 20s to mid-40s, which suggests that the majority of clients are middle aged. The frequency of clients is significantly lower below the age of 25 and past the age of 60. This suggests that the bank's client base predominantly consists of middle-aged individuals, with fewer young adults and senior citizens.

2.3. Density histogram of balance

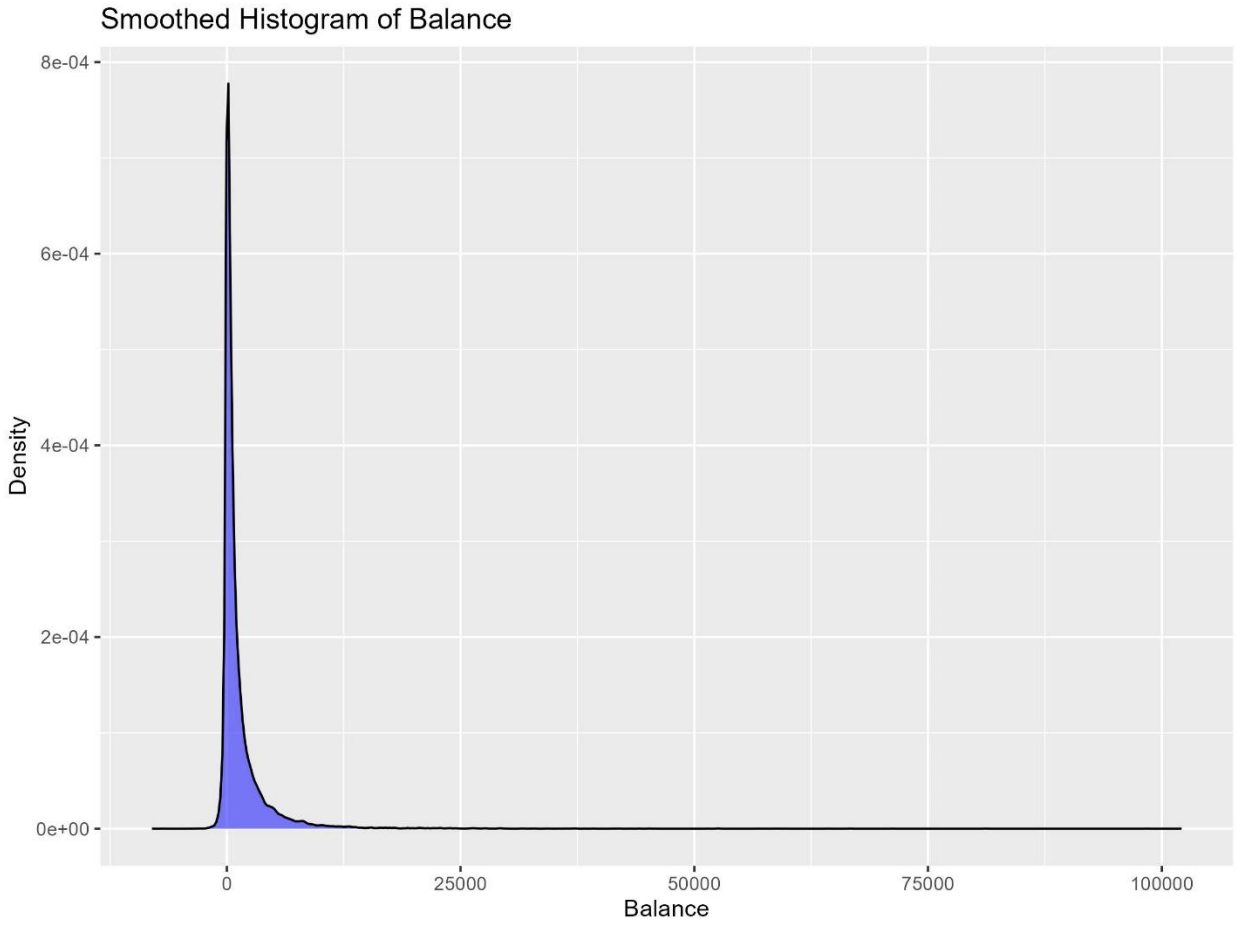

Figure 3 Density Histogram

The graph *(figure 3)* is also right skewed: it displays a peak at or near zero, indicating that a large number of clients have a balance close to zero. The long tail to the right suggests that a smaller number of clients have much higher balances, although these are the exceptions.

2.4. Box plot of balance by education and marital factors

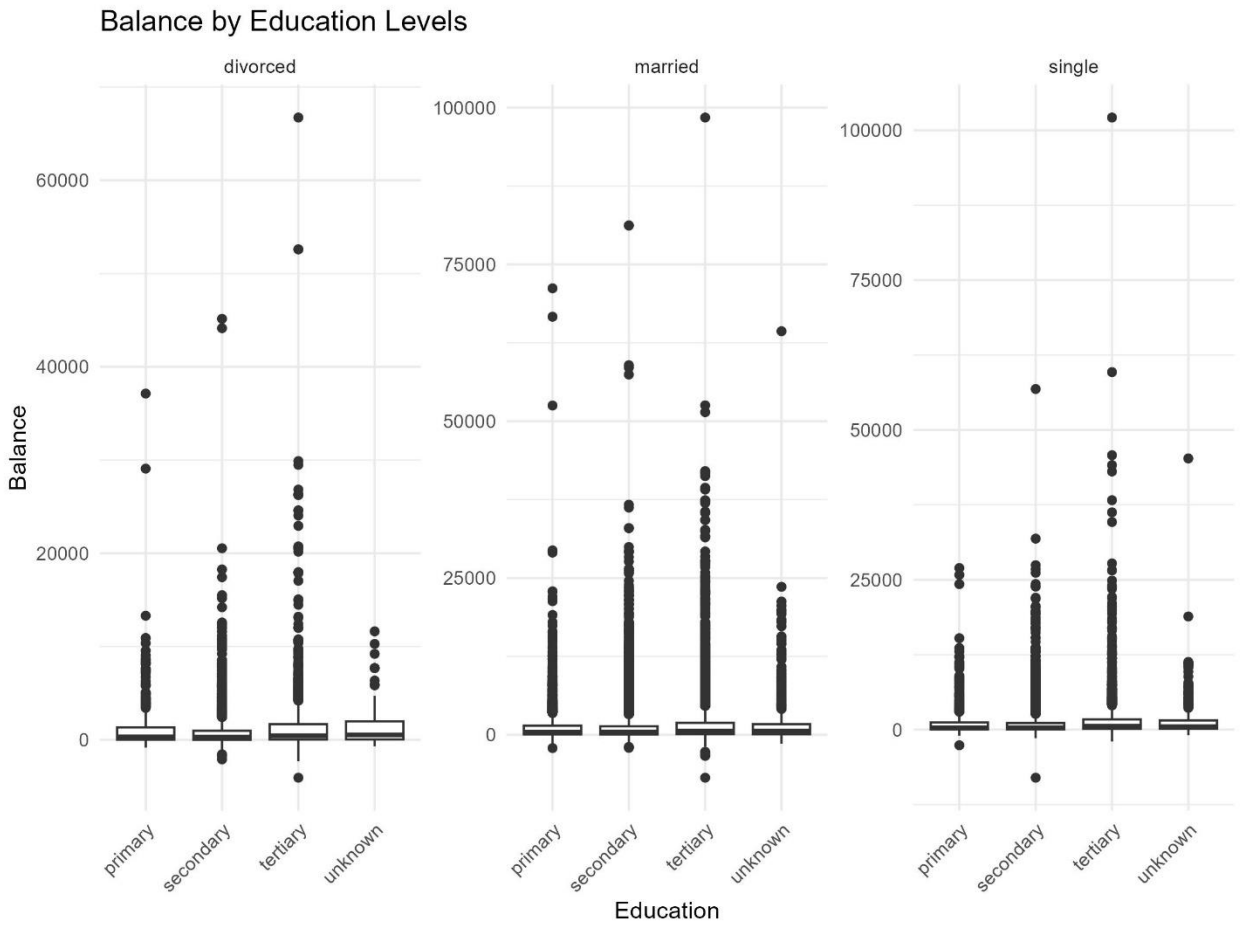

Figure 4 Box Plot

The box plot *(figure 4)* visualizes the distribution of balances among bank clients, categorized by their education levels and marital status.

Median, which is indicated by the line within each box, is close to 0 in all cases, which correlates with the results seen in the graph *(figure 3)* seen above.

Divorced: In the divorced group, the spread of balances across different education levels is relatively compact, except for a few outliers, especially in the tertiary education category. The significant number of outliers are indicative of a subset of clients with substantially higher wealth.

Married: This group exhibits a broader range of balances, particularly in the secondary and tertiary categories, where higher outliers are more prominent. However, it also has a few outliers

below 0 in the tertiary education section, suggesting that married couples with education section could suffer from negative balance more.

Single: Similar to the married group, the single clients with secondary and tertiary education also show a wide range of balances with numerous high outliers.

Married and single clients show more variability and higher balances compared to divorced clients, suggesting possible differences in financial stability or savings behavior sometimes associated with marital status.

Higher education levels (tertiary) tend to show higher median balances and greater variability in balances, indicating that higher education might correlate with higher earnings or savings.

2.5. Mosaic Plot of Education vs Marital Status

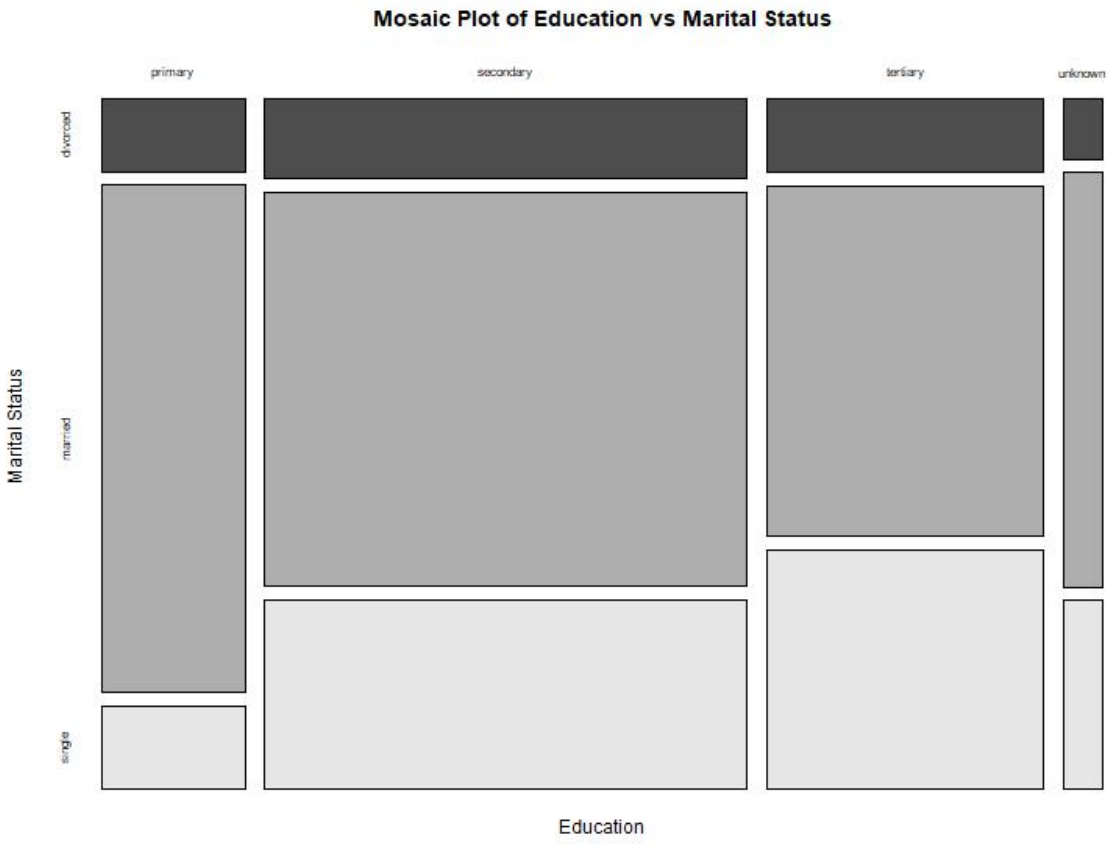

Figure 5 Mosaic Plot

The mosaic plot *(figure 5)* illustrates the relationship between education levels and marital status.

Each rectangle in the mosaic plot corresponds to a combination of education level and marital status. The area of each rectangle is proportional to the frequency count of the combination it represents.

Column width corresponds to the education level. The width of each column is proportional to the overall number of individuals in each education category.

Row height is proportional to the number of individuals with that marital status within the specific education category.

The width of the columns indicates that secondary and tertiary education categories contain more individuals compared to primary and unknown, suggesting a higher number of clients with at least a secondary level of education. The married category is consistently the tallest across all education levels, indicating that married individuals form the largest group within each educational category.

Single ad divorced marital status show a bit more variability in their proportions relative to each education category. They appear to be more prominent in tertiary and secondary education compared to primary.

2.6. Staced bar chart

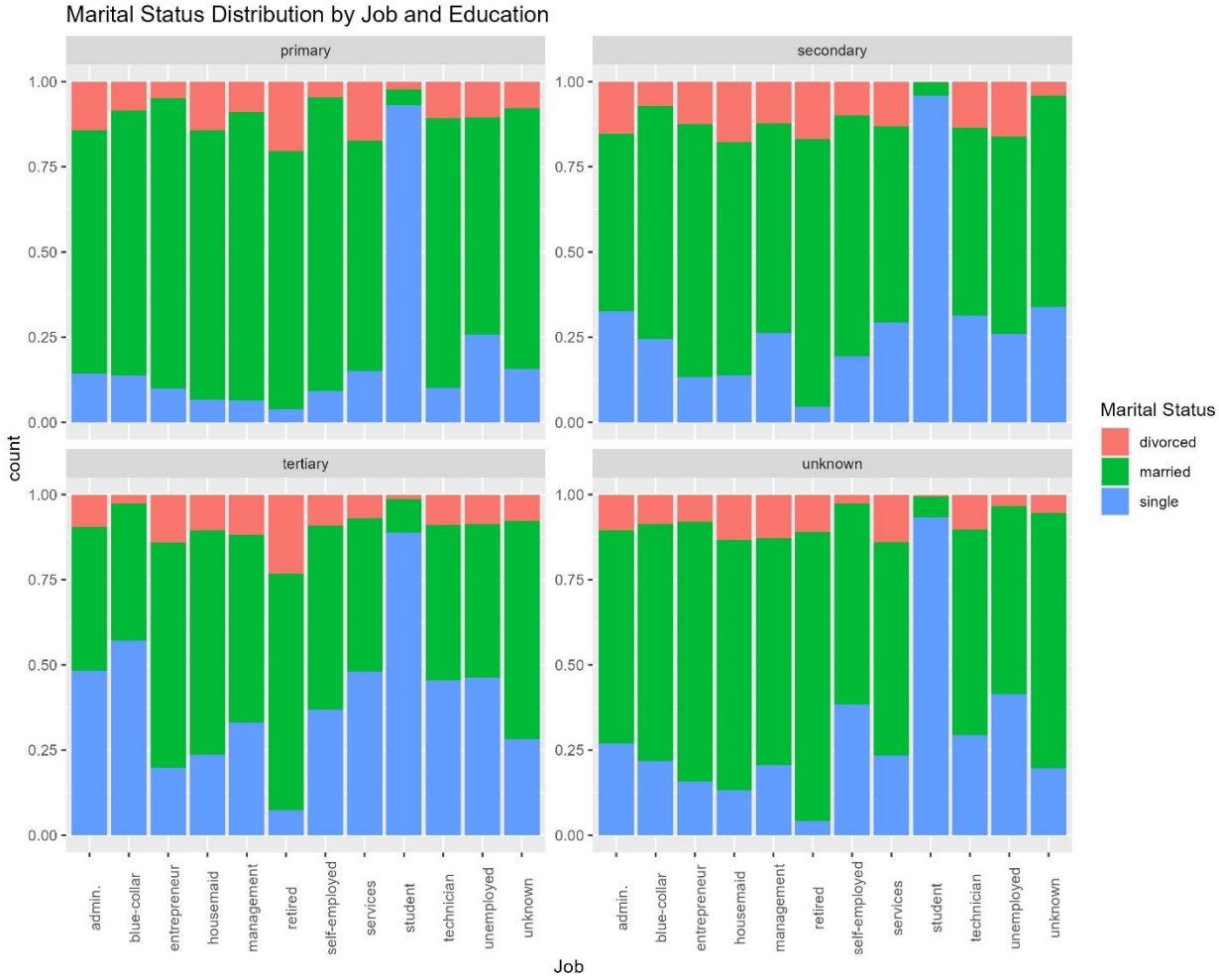

Figure 6 Stacked Bar Chart

The stacked bar chart *(figure 6)* on marital status, job and education distribution shows that students from any education group are mostly single. Single people proportions were bigger between people with tertiary education then with other education levels, seen especially with admin and blue-collar jobs. The majority of people having primary or secondary education with any job title excluding 'student' were married.

2.7. Bar chart for 'job' types

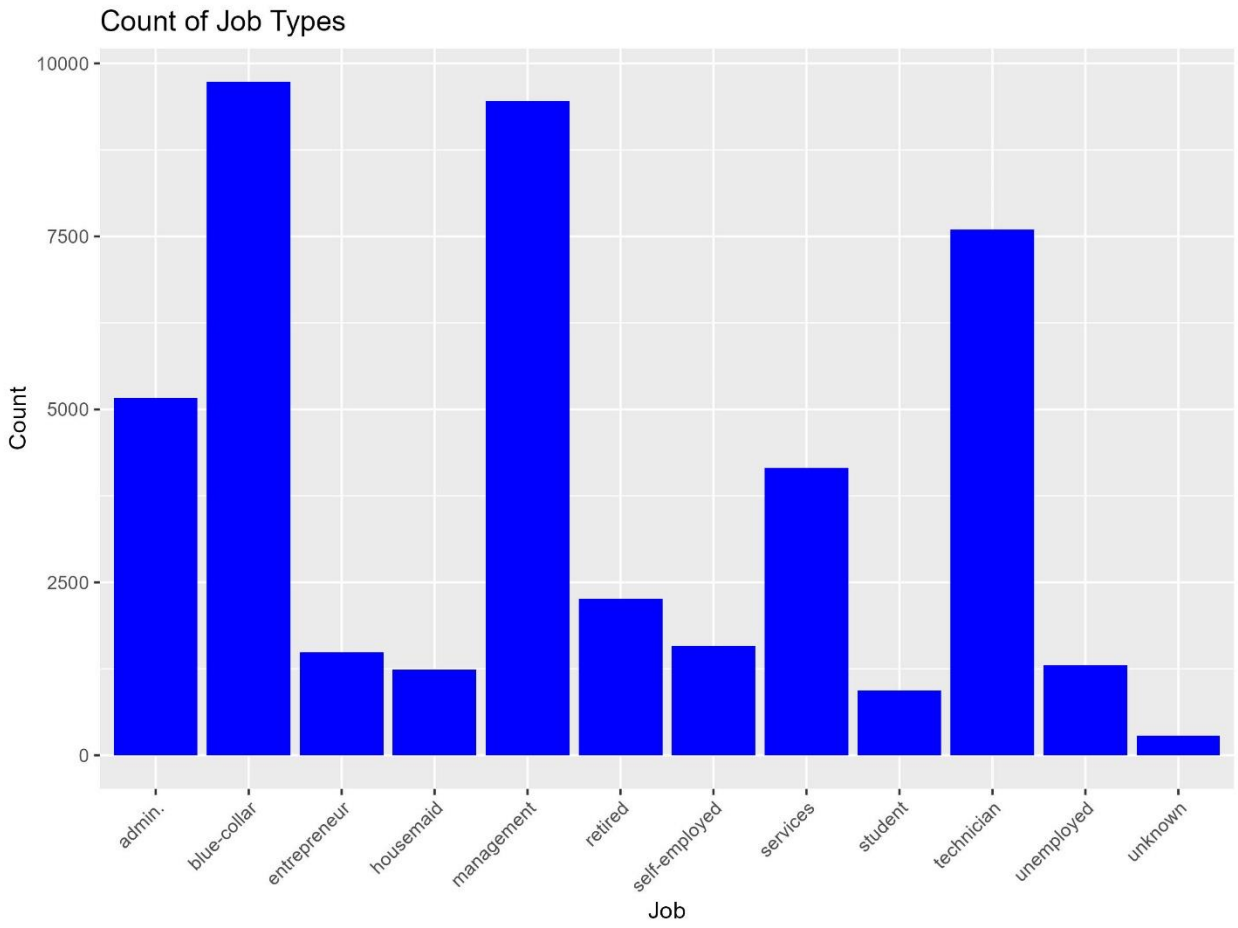

Figure 7 Bar Chart

The bar chart above *(figure 7)* visualizes the distribution of job types among clients in the dataset.

- Most common Job Types: The highest bars appear in categories such as blue-collar, management, and technician, indicating these are the most common job types among the clients.

- Middle Range Job Types: Categories like admin., services, and self-employed show moderate counts**.**

- Less Common Job Types: Job types like student, unemployed, and unknown have significantly lower counts, suggesting these categories have fewer individuals.

The prevalence of blue-collar, management, and technician jobs may suggests that the bank's clientele is diverse but has significant representations from traditionally stable and middle-income professions**.**

2.8. Age vs Balance Scatter Plot

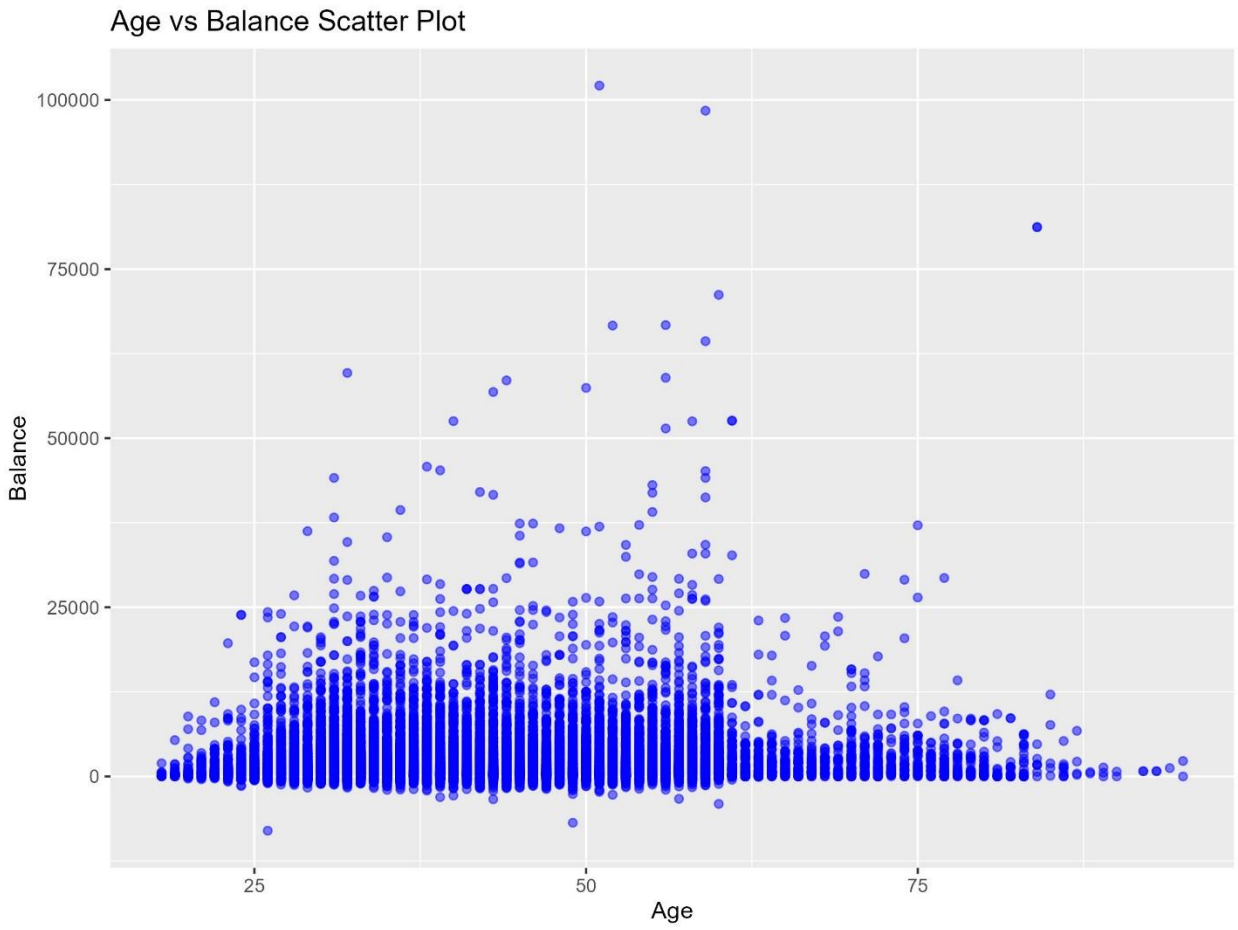

Figure 8 Scatter Plot

The scatter plot *(figure 8)* shows the relationship between age and balance for a set of clients. Each point on the plot represents an individual client, with their position determined by their age and corresponding bank balance.

Most of the data points are concentrated near the bottom of the plot, indicating that a large portion of clients have low balances, regardless of age. There is a visible density of points from about age 30 to 60, which is typically considered prime working age.

There does not appear to be a strong relationship between age and balance, as the data does not form a clear trend or pattern. Balances are dispersed across all ages without a discernible increase or decrease associated with age. However, people over 60 do have fewer negative balances and people over 50 have slightly higher balances. The presence of higher balances in older age brackets could suggest that savings accumulate as clients grow older, though the effect is not strong enough to form a clear trend across the dataset.

2.9. Balance and age group by y factor

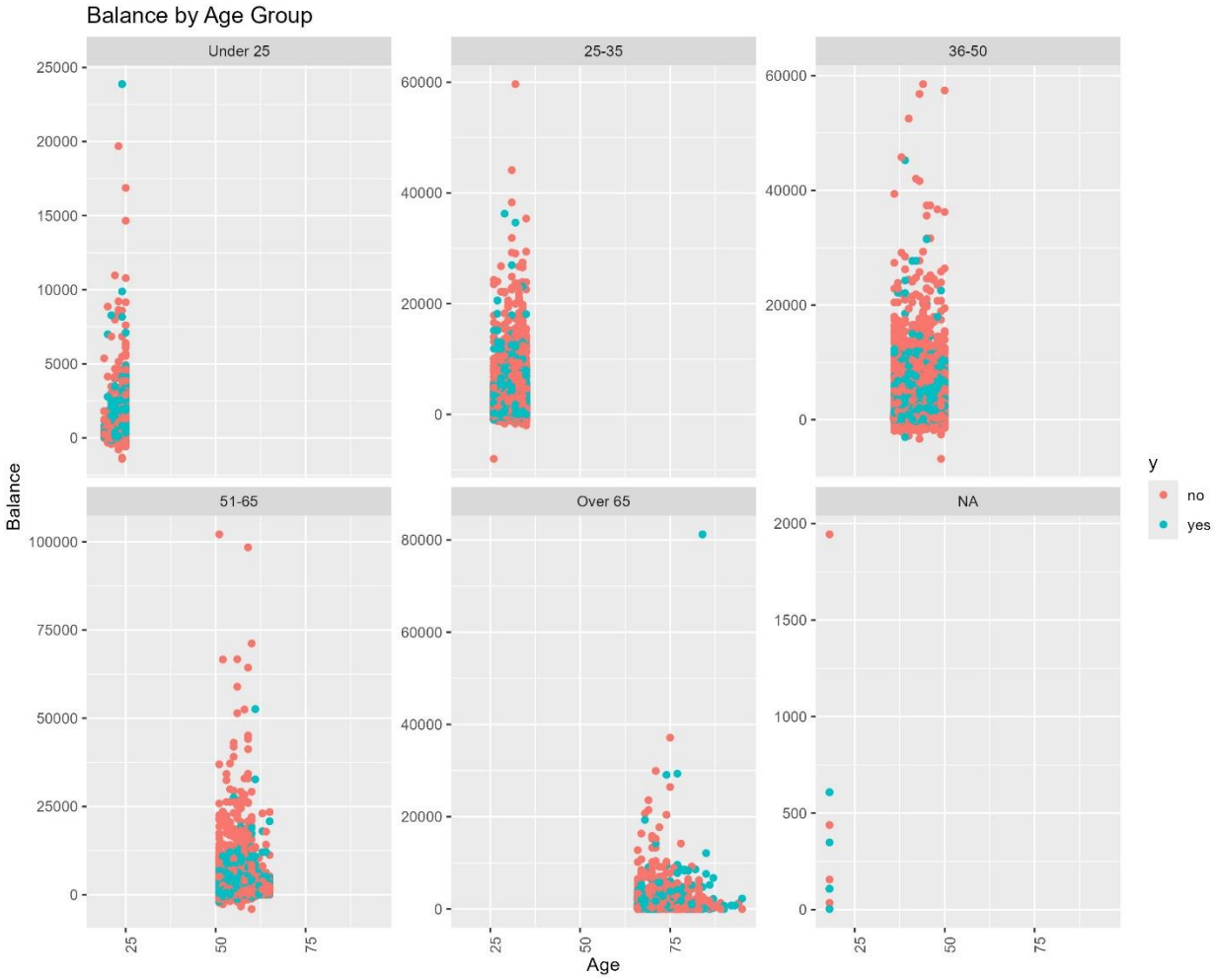

Figure 9 Faceted Scatter Plot

Faceted Scatter Plot *(figure 9)* displays the relationship between client age and bank balance, further categorized by whether clients have subscribed to a term deposit

- Younger age groups (Under 25, 25-35) show lower balance amounts overall, with balances generally not exceeding 20,000 euros. The distribution in these groups is relatively tight, centered at lower balance values.
- Middle age groups (36-50, 51-65) display a wider range of balances, with the '51-65' group showing some particularly high balances, peaking around 100,000 euros.
- The 'Over 65' group generally shows moderate balance amounts, although there is a noticeable cluster of balances around 5 thousand euros and few higher balances.
- The 'NA' group, lacking specific age data, also shows a broad range of balances but predominantly on the lower end.

The proportion of clients subscribing to a term deposit appears relatively consistent across age groups.

2.10. Distribution of Term Deposit Subscription Pie Chart

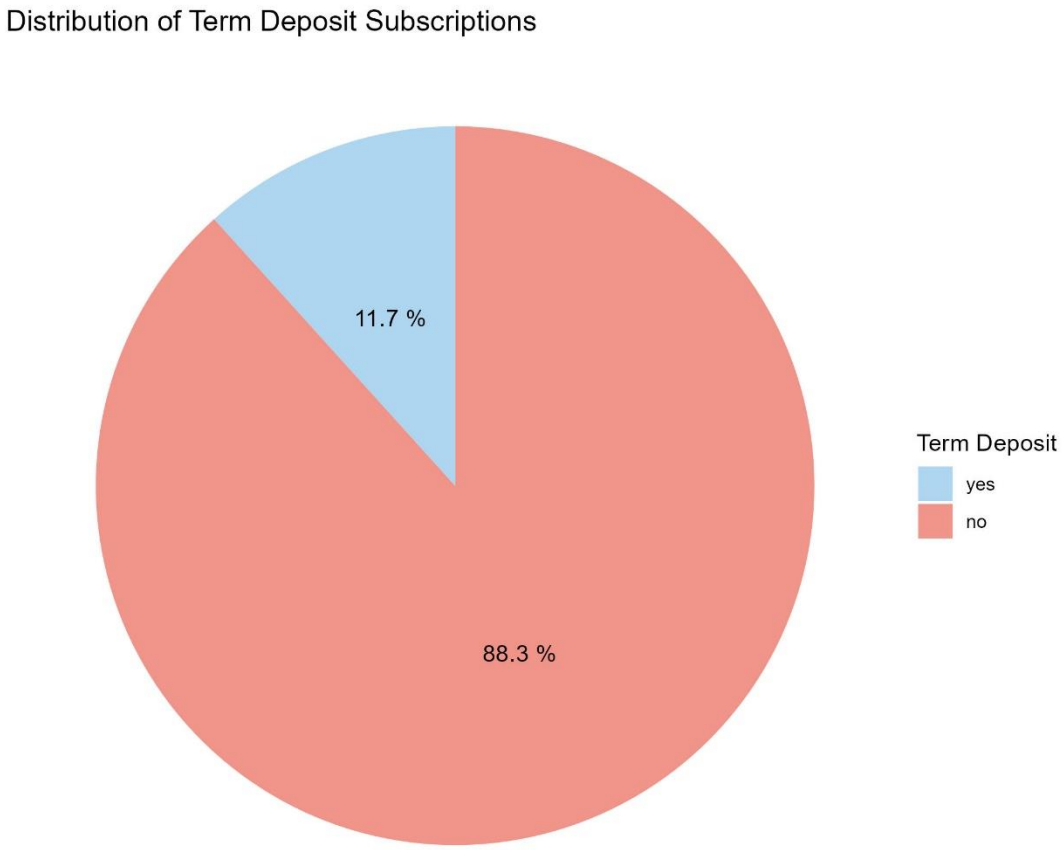

Figure 10 Pie Chart

Pie chart *(Figure 10)* visualizes the percentage of bank clients who have and have not subscribed to a term deposit. A significant majority (88.3%) of the clients have not subscribed to a term deposit, as indicated by the large teal segment of the pie. A smaller fraction (11.7%) of the clients have chosen to subscribe, shown by the smaller coral segment. Similar results are seen in the faceted scatter plot *(Figure 9)*, where you could see that significantly more responses were „no“.

3. Modeling task

Perform a logistic regression to obtain the predicted probability that a customer has subscribed for a term deposit.
Use continuous variables and dummy variables created for categorical columns. Not necessarily all variables provided in data sample should be used.
Evaluate model goodness of fit and predictive ability. If needed, data set could be split into training and test sets.
Data: attached (response variable y).

3.1. Logistic Regression Model Summary

Deviance Residuals: measure how well the model fits each observation. Results show a range from -5.6884 to 3.4356, indicating variability in how well the model predicts individual outcomes. The residuals suggest that while many predictions are close to their actual values (residuals near zero), there are some outliers or instances where the model predictions deviate significantly from the observed values.

Coefficients:

Estimate: Represents the effect size of each predictor. Positive values increase the log-odds of the outcome (subscribing to a term deposit), and negative values decrease the log-odds.

Std. Error: Measures the standard error of the estimate, which indicates the level of uncertainty around the coefficient estimate.

z-value: The coefficient divided by its standard error, used to test the null hypothesis that the coefficient is zero (no effect).

Pr(>|z|): P-value associated with the z-test. Values less than 0.05 typically suggest that the predictor is statistically significant.

jobblue-collar: Coefficient of -0.3323 with a p-value of 4.63e-05 suggests that being in a blue-collar job is significantly associated with a lower likelihood of subscribing compared to the baseline job category.

housingno: Coefficient of 0.7121 with a p-value less than 2e-16indicates that not having a housing loan is strongly associated with a higher likelihood of subscribing.

duration: Each additional second of contact duration increases the log-odds of subscribing by 0.004145, highly significant with a p-value less than 2e-16.

Null Deviance shows the deviance of a model with only an intercept, which is 26111.

Residual Deviance is the deviance of the model with predictors, at 17285, indicating a substantial improvement when including predictors.

AIC: A measure of the relative quality of the statistical model for a given set of data. Lower AIC values indicate a model is more parsimonious**.**

Fisher Scoring iterations: Indicates the number of iterations taken to converge on a solution, with 6 iterations suggesting the model required several steps to find optimal coefficients.

Implications and Observations

Predictor Importance: Variables like duration, housingno, and poutcomesuccess are particularly influential, as indicated by their coefficients and significance levels.

Model Performance: The difference between null and residual deviance shows that the model has good explanatory power and fits the data significantly better than a model without predictors.

Potential Model Issues: Large ranges in deviance residuals might suggest the presence of outliers or influential observations that could affect the robustness of the model.

3.2. ROC Curve

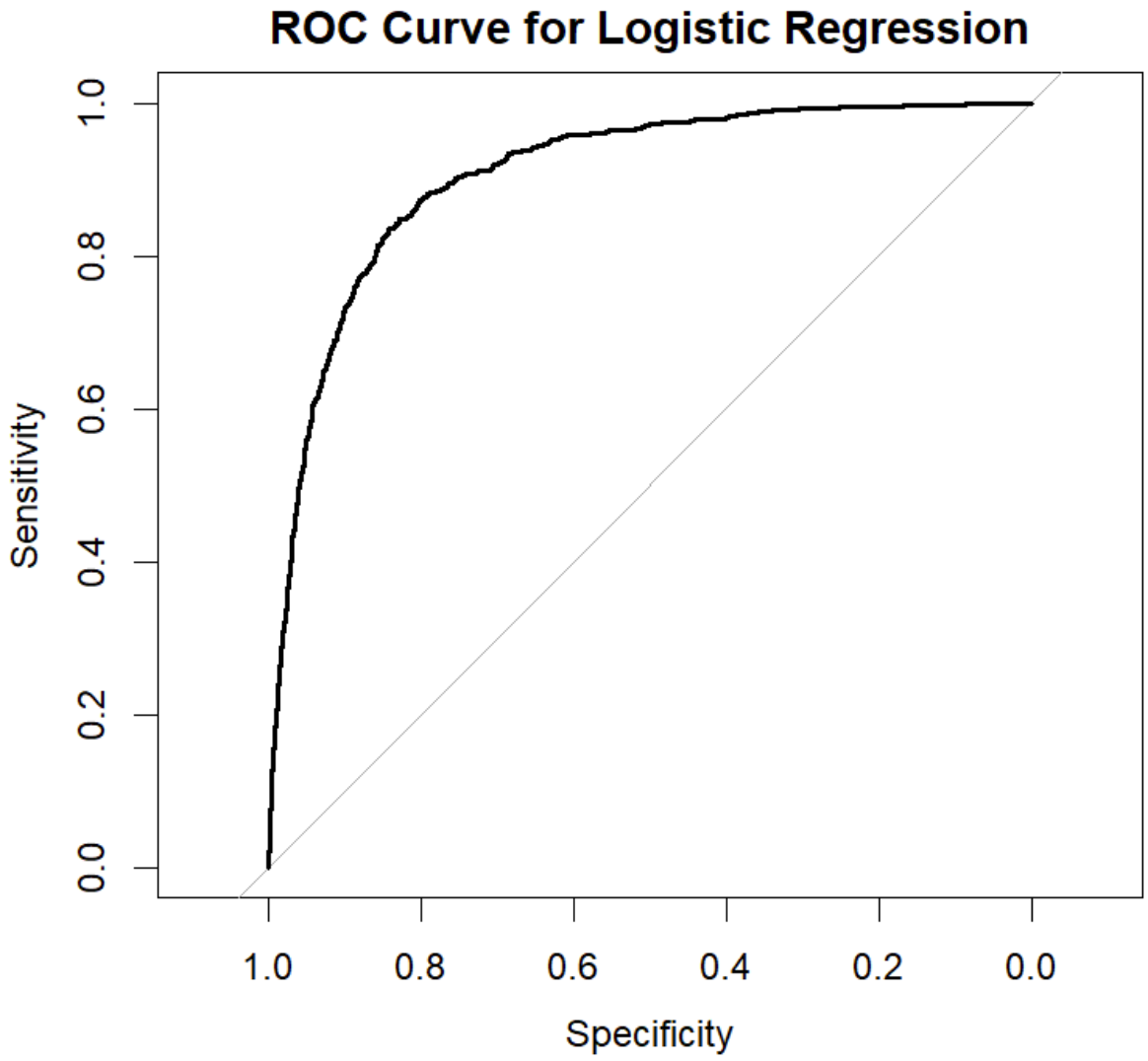

Figure 11 ROC Curve

ROC is a graphical representation used to evaluate the diagnostic ability of a binary classifier system as its discrimination threshold is varied. This curve helps in selecting an optimal threshold for maximizing the true positive rate while minimizing the false positive rate, based on the specific costs of false negatives versus false positives for the business context.

The graph indicates that the logistic regression model performs well in distinguishing between the customers who will and will not subscribe to a term deposit. The steep initial rise suggests that the model can achieve high sensitivity without sacrificing much specificity at lower thresholds.

3.3. Goodness-of-fit

A McFadden's R-squared value of 0.3380347 indicates a moderate level of explanatory power of the logistic regression model. This value suggests that the model is substantially better at predicting the outcome than a model without any predictors, but it's still leaving a considerable amount of variation in the response variable unexplained.

Sources:

1. [Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology.
2. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011.