

SAFARICOM COMMERCIAL BANK OF AFRICA



Mobile money: Your money, your way.



Understanding The Context

Safaricom's M-PESA mobile money platform has revolutionized the financial services sector in Kenya, providing millions of Kenyans with access to secure and convenient financial services, including money transfers, savings, and loans. M-PESA has played a major role in promoting financial inclusion in Kenya, and Safaricom is poised to play an even greater role in shaping the future of finance in the country.





Specifying The Data Analysis Question

-  **Identify the factors that contribute to loan default.**
-  **Build multiple classification models to predict loan default**

Defining The Metrics of Success

The project's objective is to estimate when a person will default on their loan. We construct a model that incorporates all the relevant data to reliably predict the outcome.

Loan defaulting occurs when a person does not fulfill their debt obligation in the specified period. A loan is considered to be in default for risk modelling purposes if it is more than 90 days old.

We will have achieved our objective when we get at least one model with an accuracy score of around 80%.



Recording The Experimental Design



Strategic Project

METHODOLOGY



1. Load the necessary libraries and datasets for our analysis.



2. Perform data cleaning and pre-processing where necessary.



3. Carry out our analysis



4. Build a predictive model



5. Interpret and summarize findings.



6. Provide recommendations.

Data Relevance

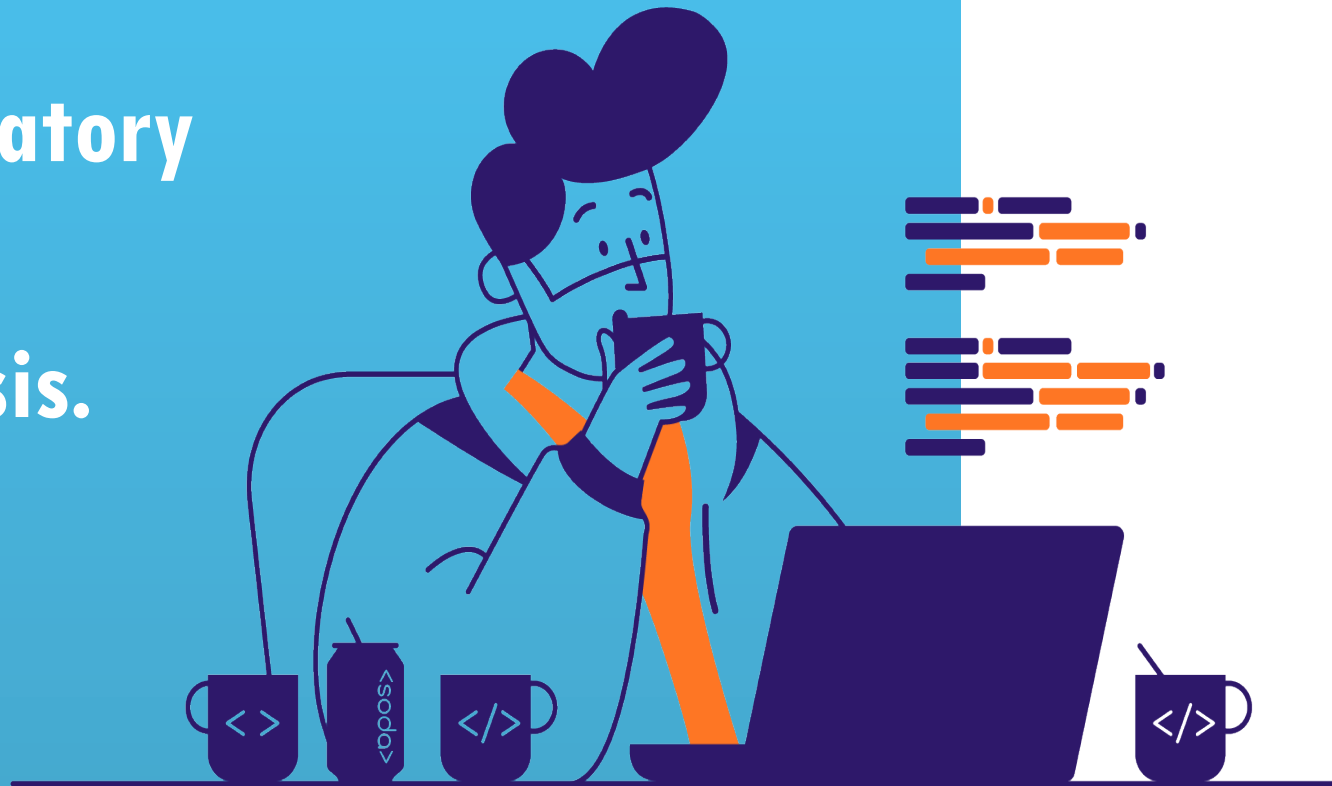
In December 2021, a household survey conducted by Central Bank of Kenya (CBK), FSD Kenya and the Kenya National Bureau of Statistics (KNBS) revealed that 50.9% of mobile loan borrowers had defaulted on their loans.**

<https://www.businessdailyafrica.com/bd/economy/half-mobile-phone-borrowers-default-3654550>



The data was relevant to answering our data analysis question.

Exploratory Data Analysis.



This process of examining and summarizing data sets using various techniques such as visualization and descriptive statistics to help to identify patterns and relationships in the data.



Univariate Data Analysis.



Bivariate Data Analysis.



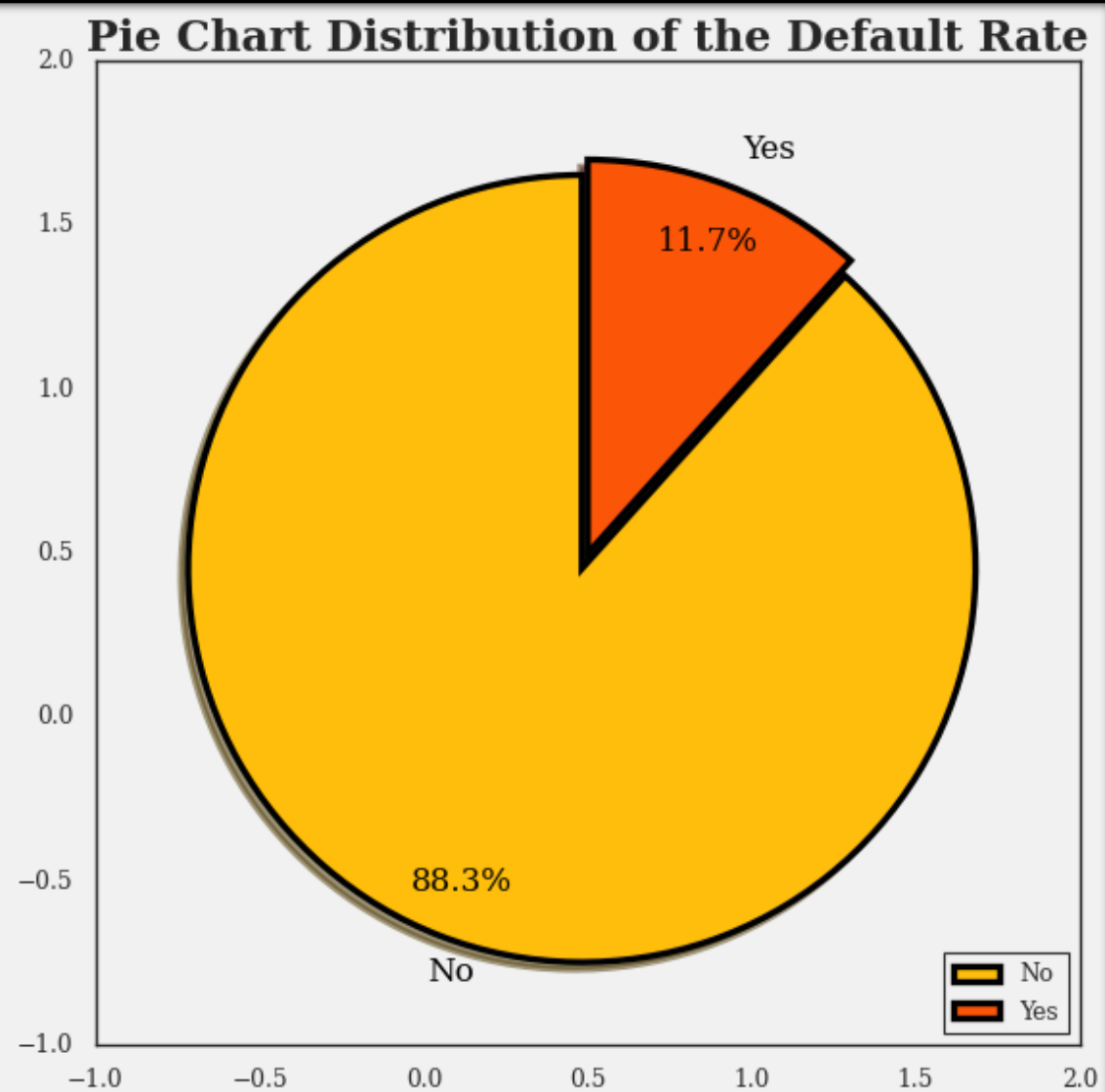
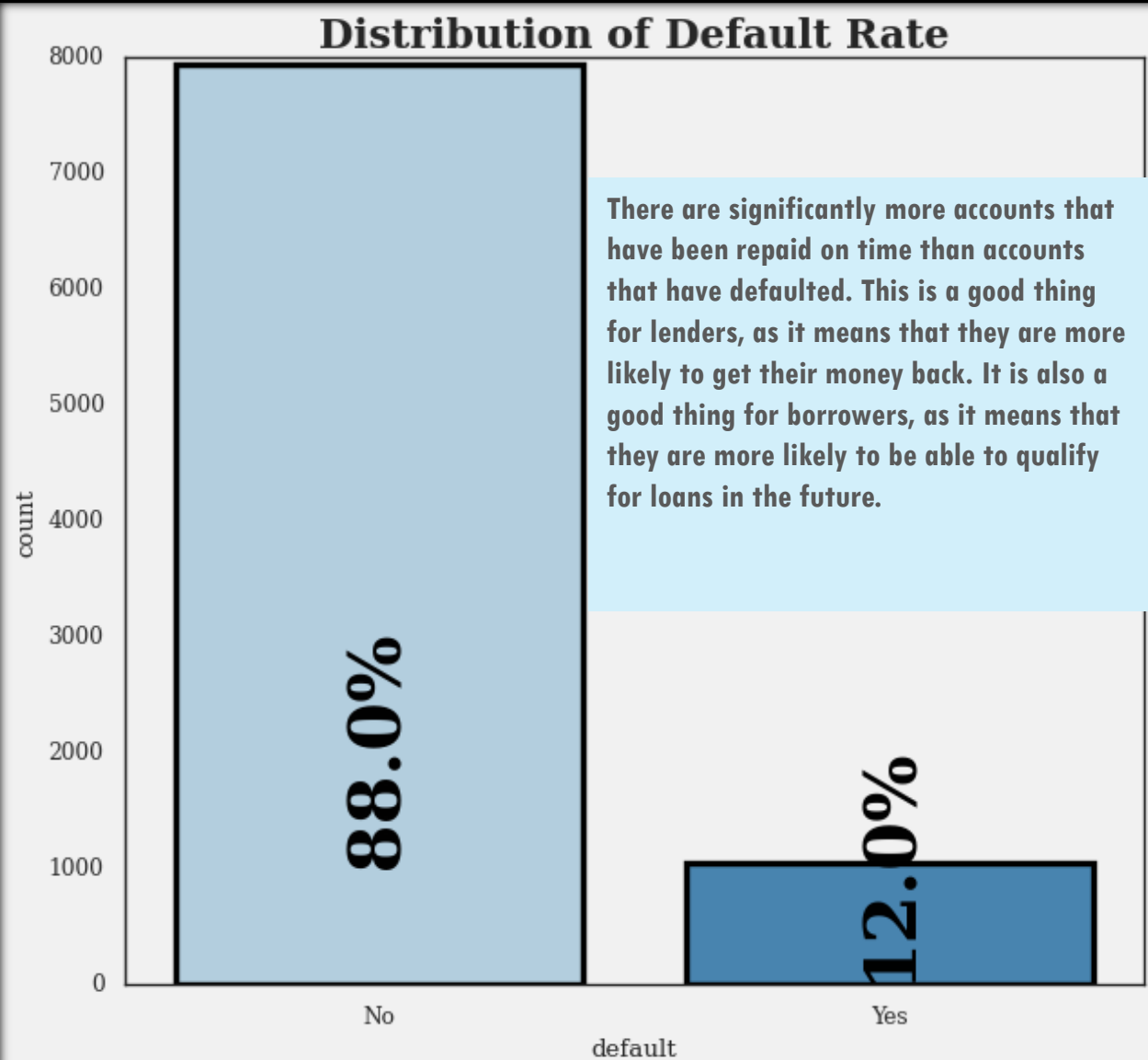
Multivariate Data Analysis.

Question:

What is the
distribution of the
default rate?



Distribution of Default Rate

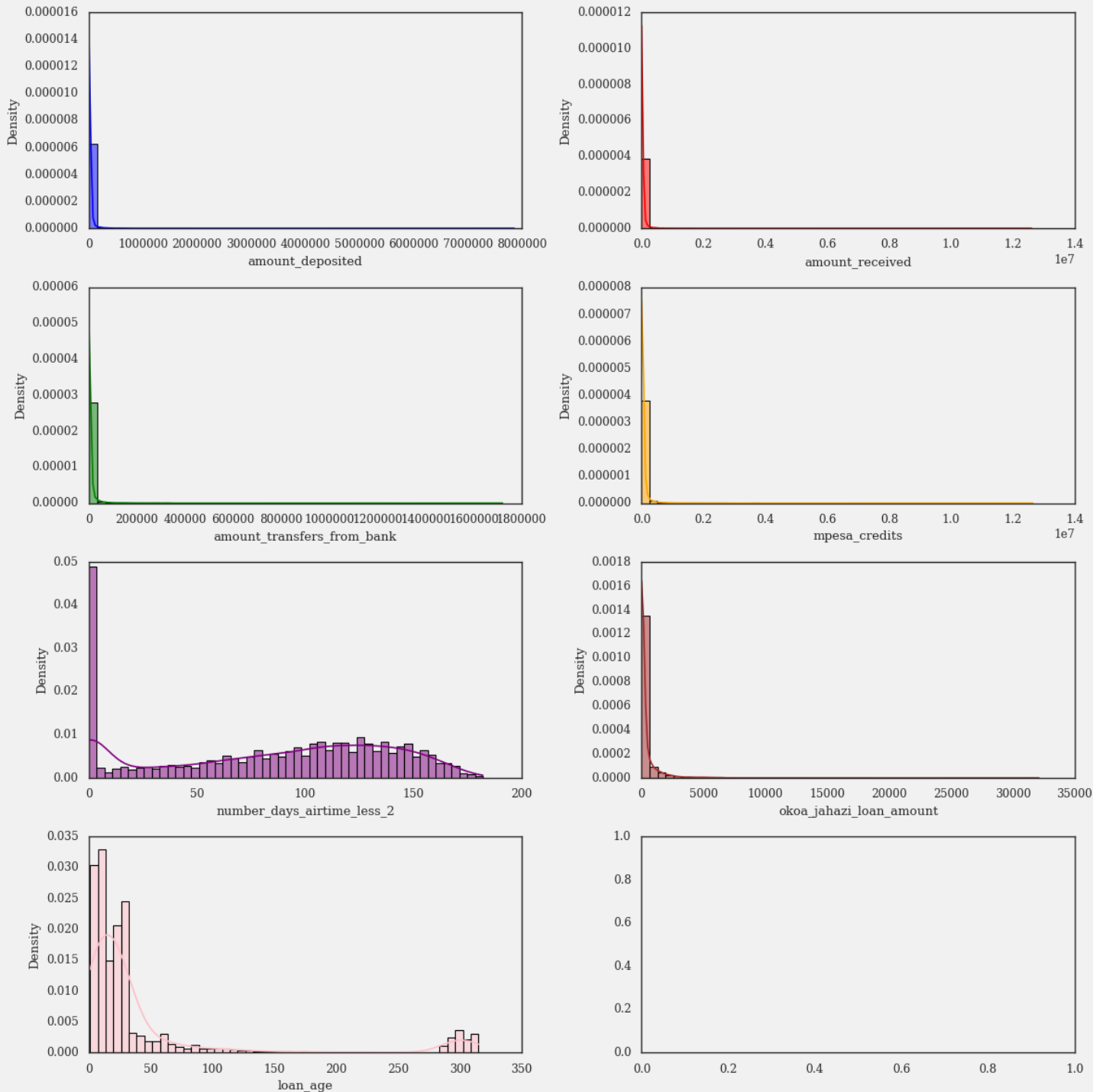


Distribution of Numerical Variables

Most customers have recently taken out loans, as the distribution of loan age is skewed right. This could be because the company is new or offers short-term loans.

The variable `number_days_airtime<2` is a mixture distribution, meaning it likely consists of multiple distributions. For example, a distribution of the number of days customers go without purchasing airtime could be a mixture of two distributions: one for regular customers and one for rare customers.

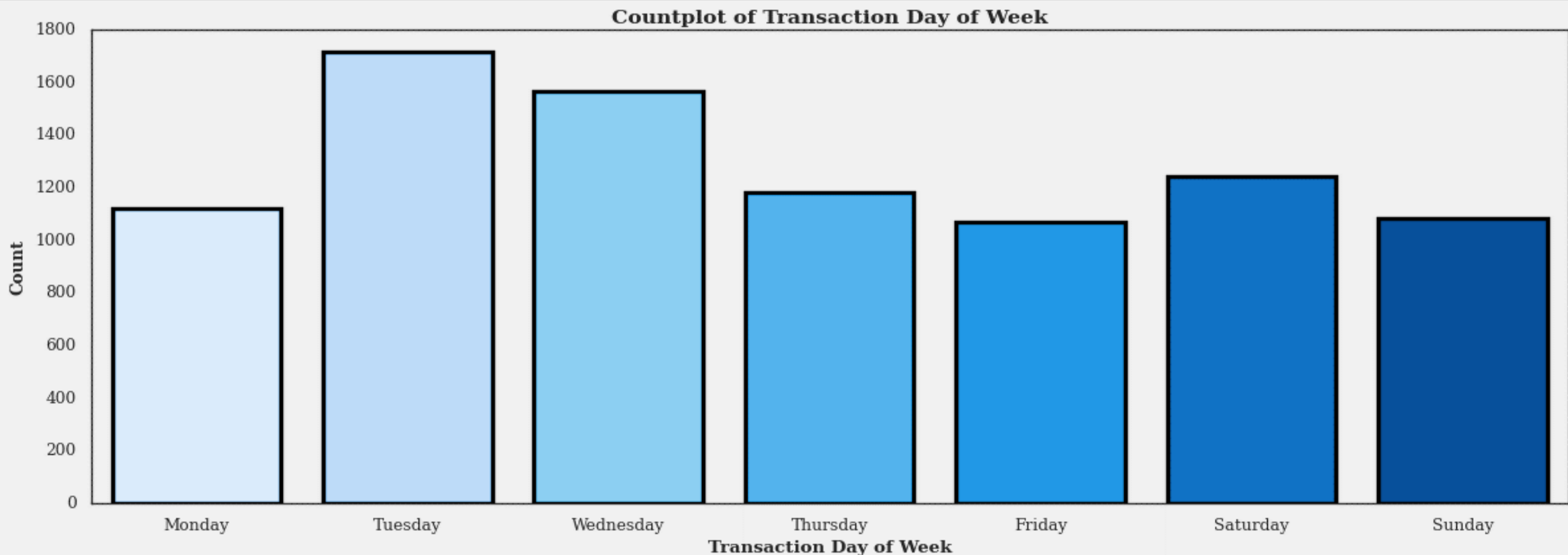
The variables `amount_deposited`, `amount_received`, `amount_transfers_from_bank`, `mpesa_credits`, and `okoa_jahazi_loan_amount` are skewed right, meaning most values are close to zero but there are more large positive values than large negative values. For example, a distribution of customer account balances would be skewed right if most customers have balances close to zero but a few customers have very large balances.



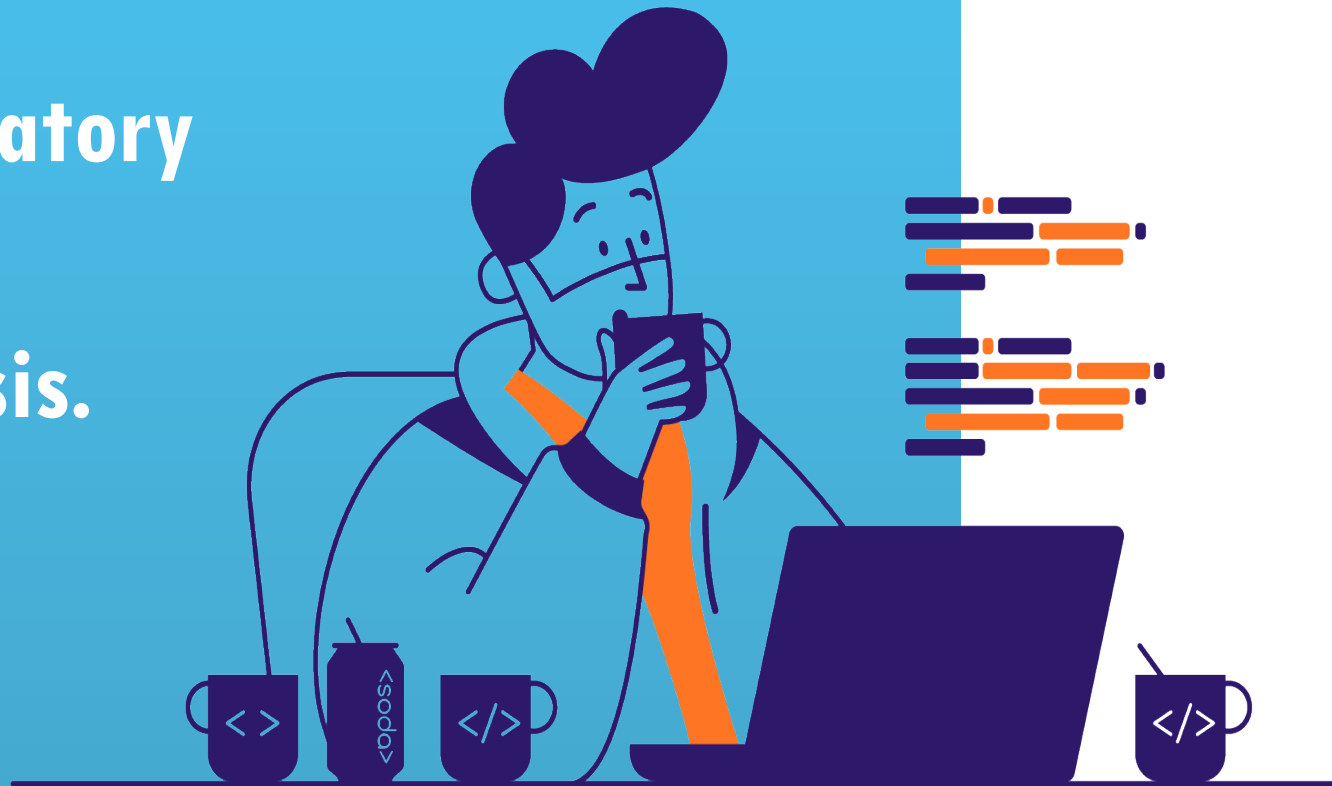
Distribution of the transaction_day_of_week

Mobile money transactions are more common on weekdays than on weekends. People are more likely to send and receive money using their mobile phones during the workweek, when they are more likely to be making purchases or paying bills. Possible explanations:

- People are more likely to be working or running errands on weekdays.
- Businesses are more likely to be open on weekdays.
- People may be more likely to use mobile money to pay for bills on weekdays.



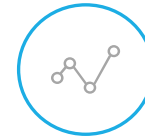
Exploratory Data Analysis.



This process of examining and summarizing data sets using various techniques such as visualization and descriptive statistics to help to identify patterns and relationships in the data.



Univariate Data Analysis.



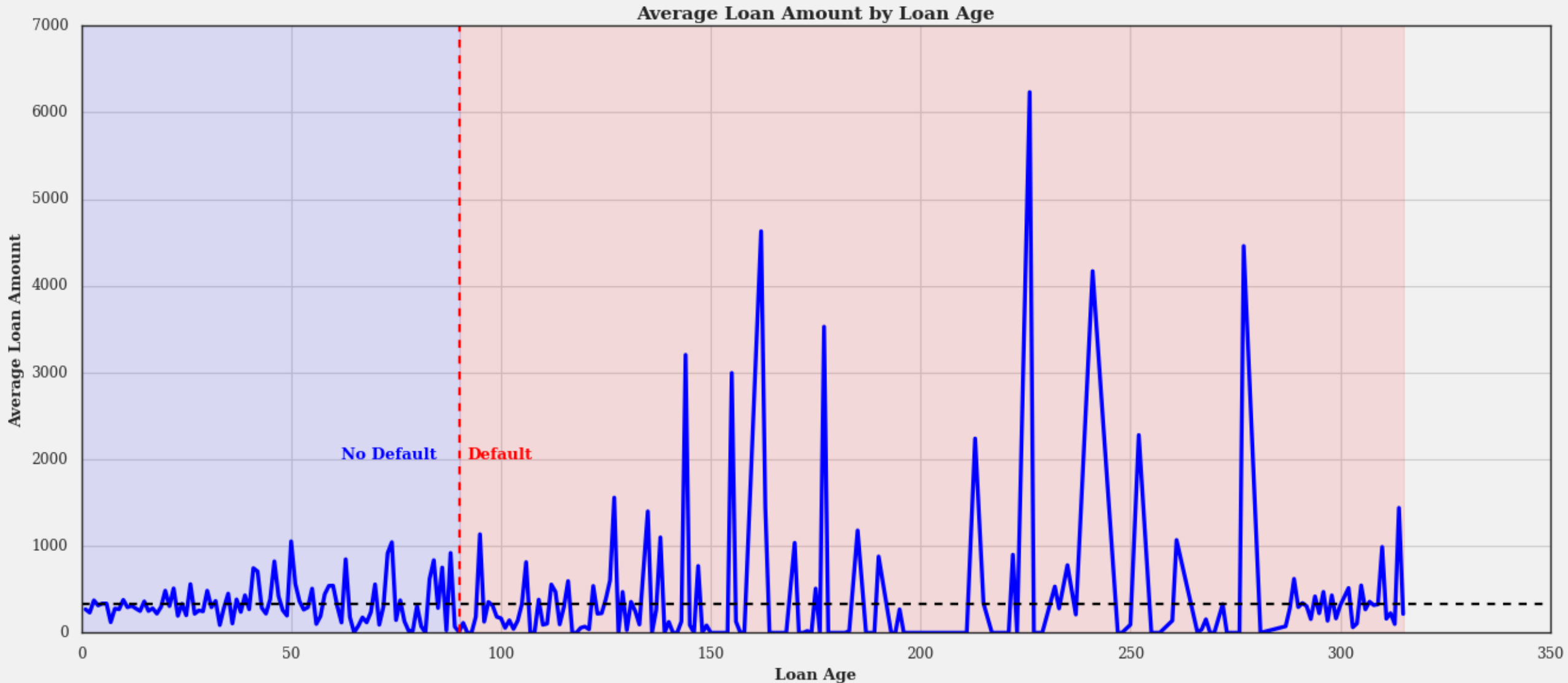
Bivariate Data Analysis.



Multivariate Data Analysis.

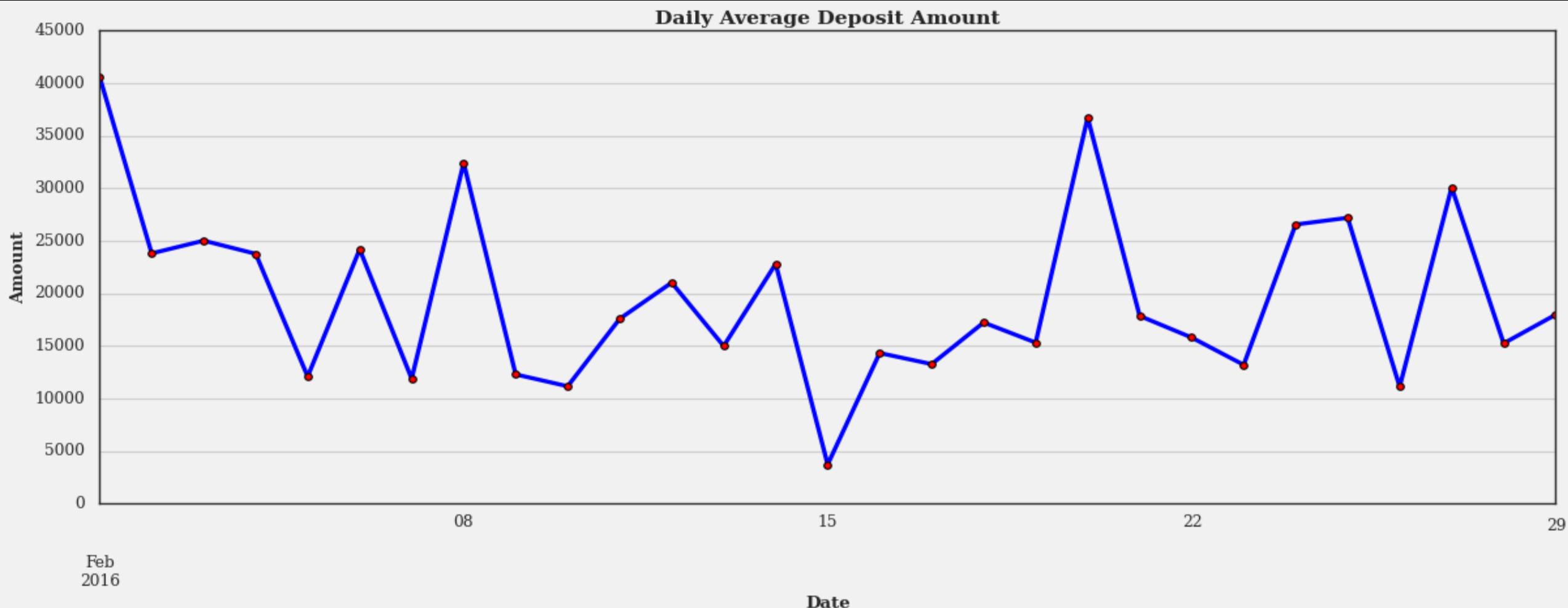
Average Loan Amount by Loan Age

The graph shows that the average loan amount for both types of loans increases with loan age. This is likely because borrowers are able to qualify for larger loans as they build up their credit history and become more established financially.



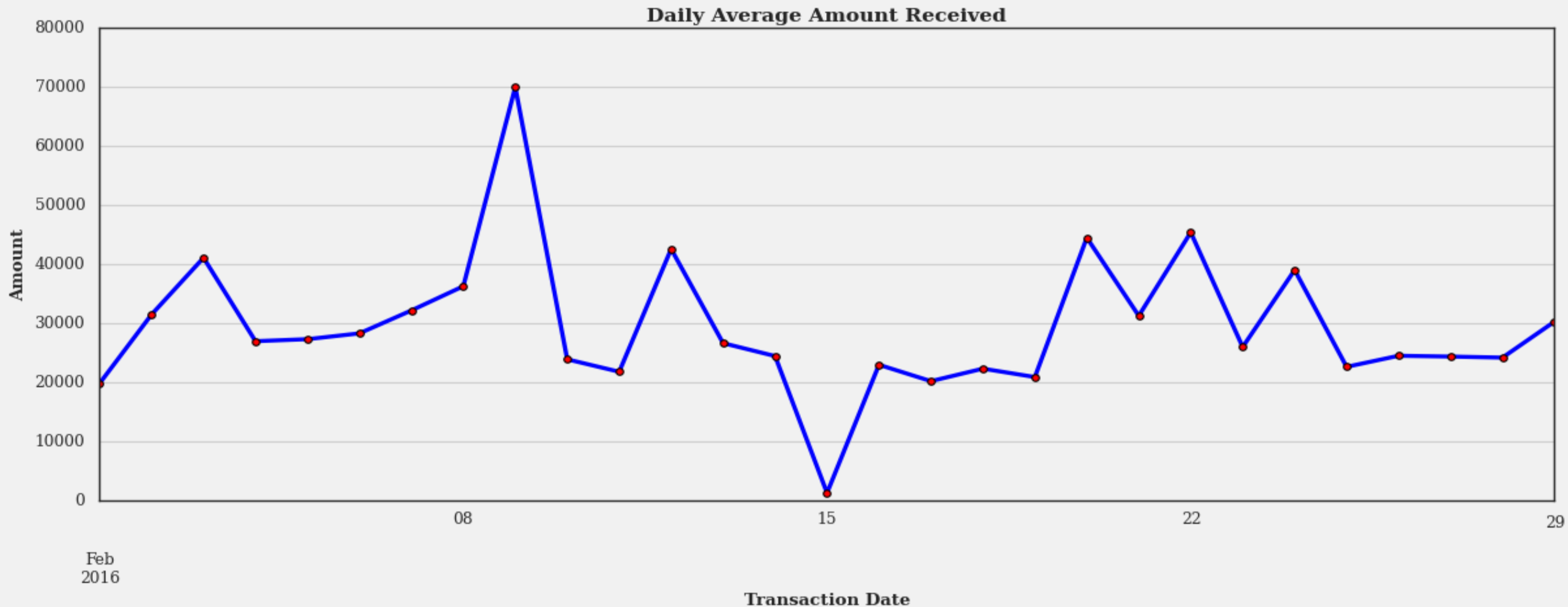
Daily average deposit amount

The highest daily average deposit amount was on February 1st, with about 40000. The lowest daily average deposit amount was on February 15th, with about 5000.



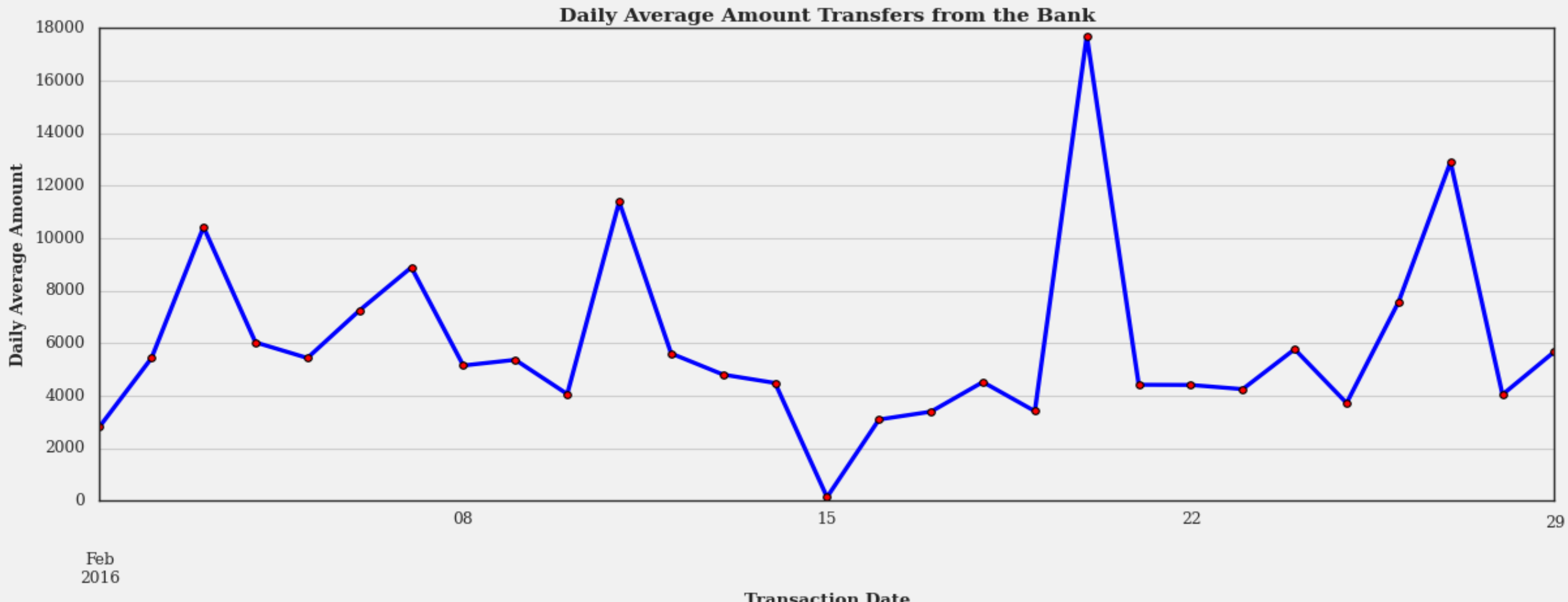
Daily average amount received

The line graph shows the daily average amount received for the month of February 2016. The graph has a peak around February 8th and a trough around February 22nd



Daily average amount transfers from bank

The graph shows that there were a few fluctuations in the daily average amount of transfers during this period, but the overall trend was upward.

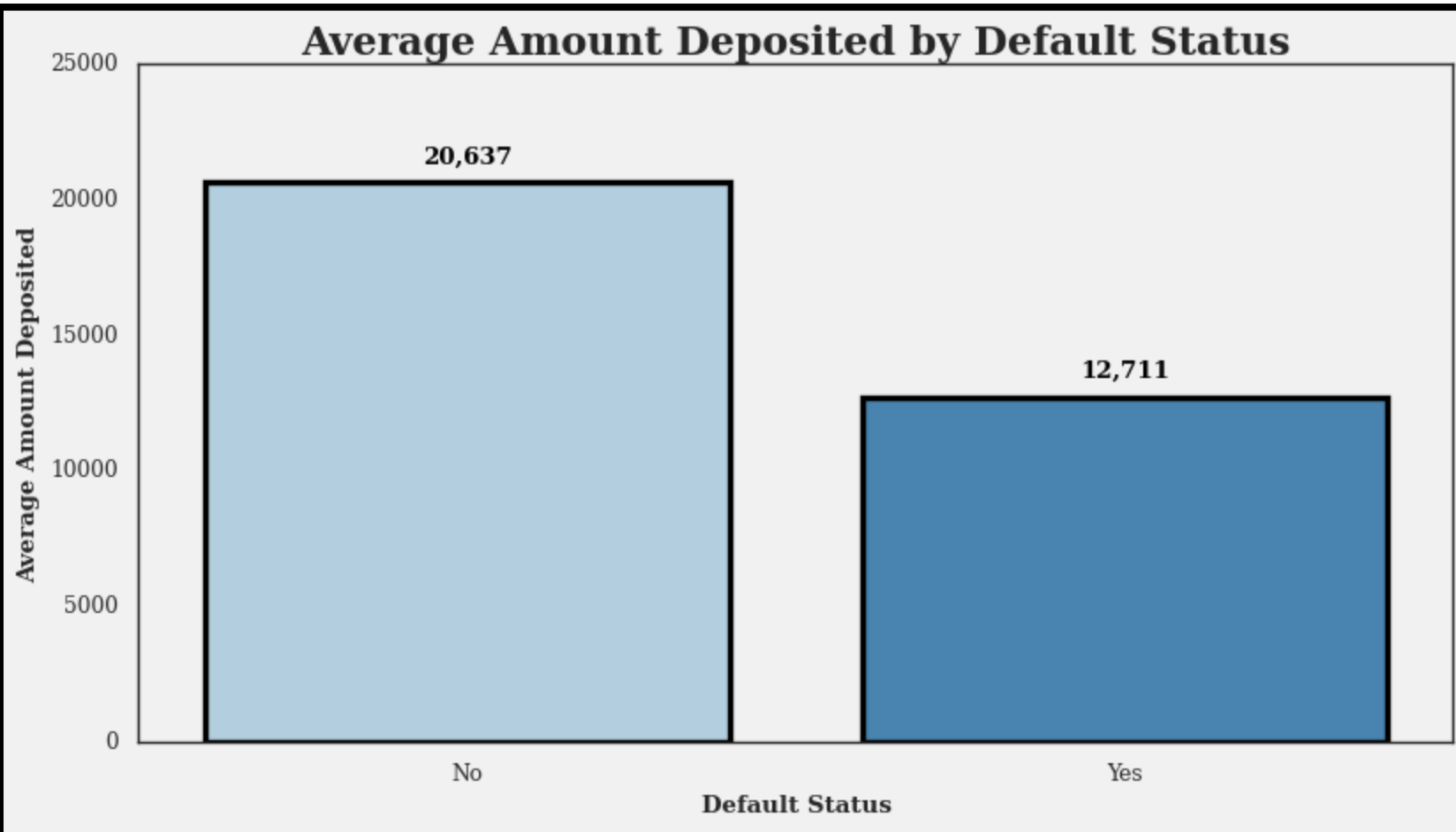


Question:

What is the
distribution of the
different average
amounts by default
status?

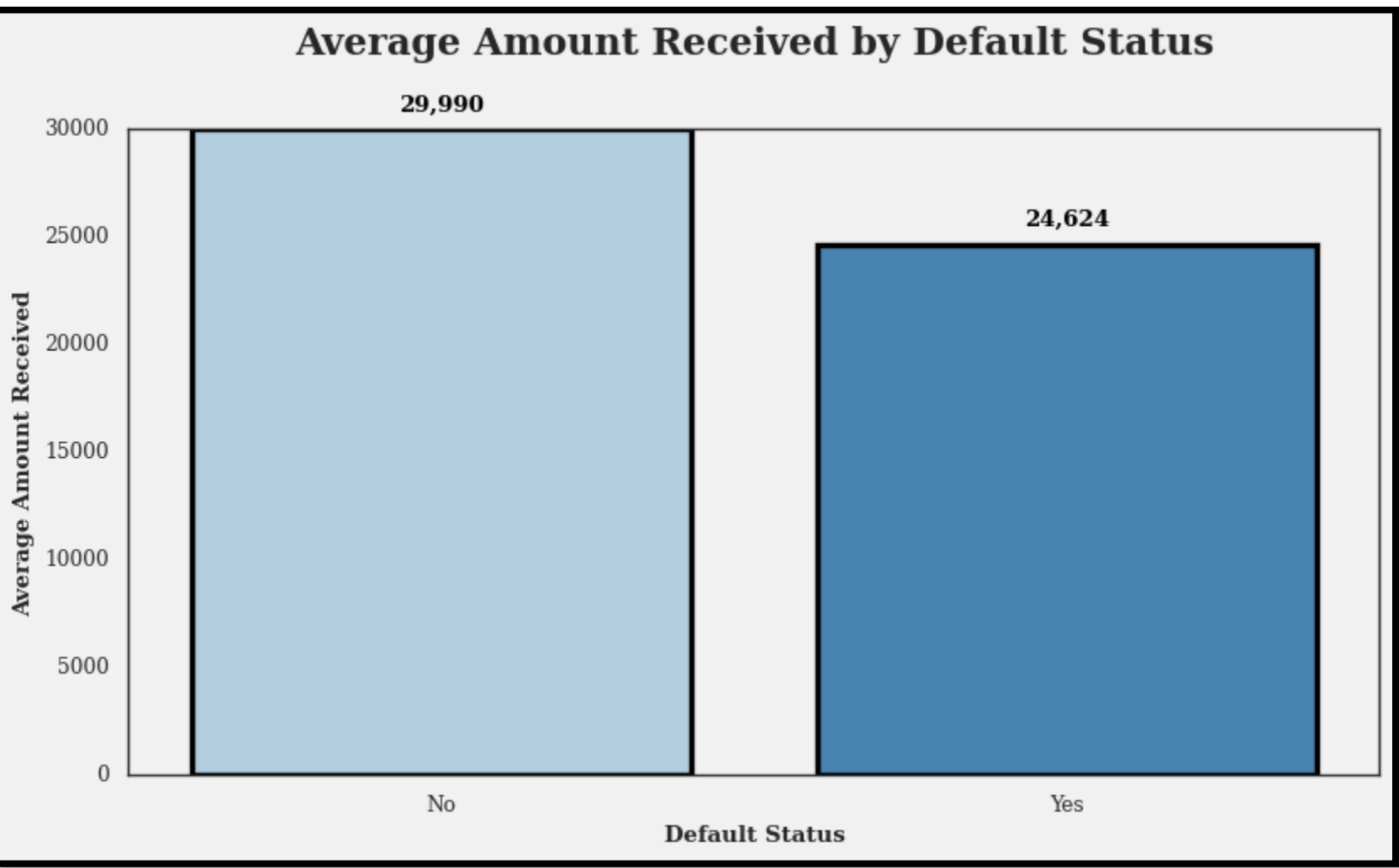


Average Amount Deposited by Default Status



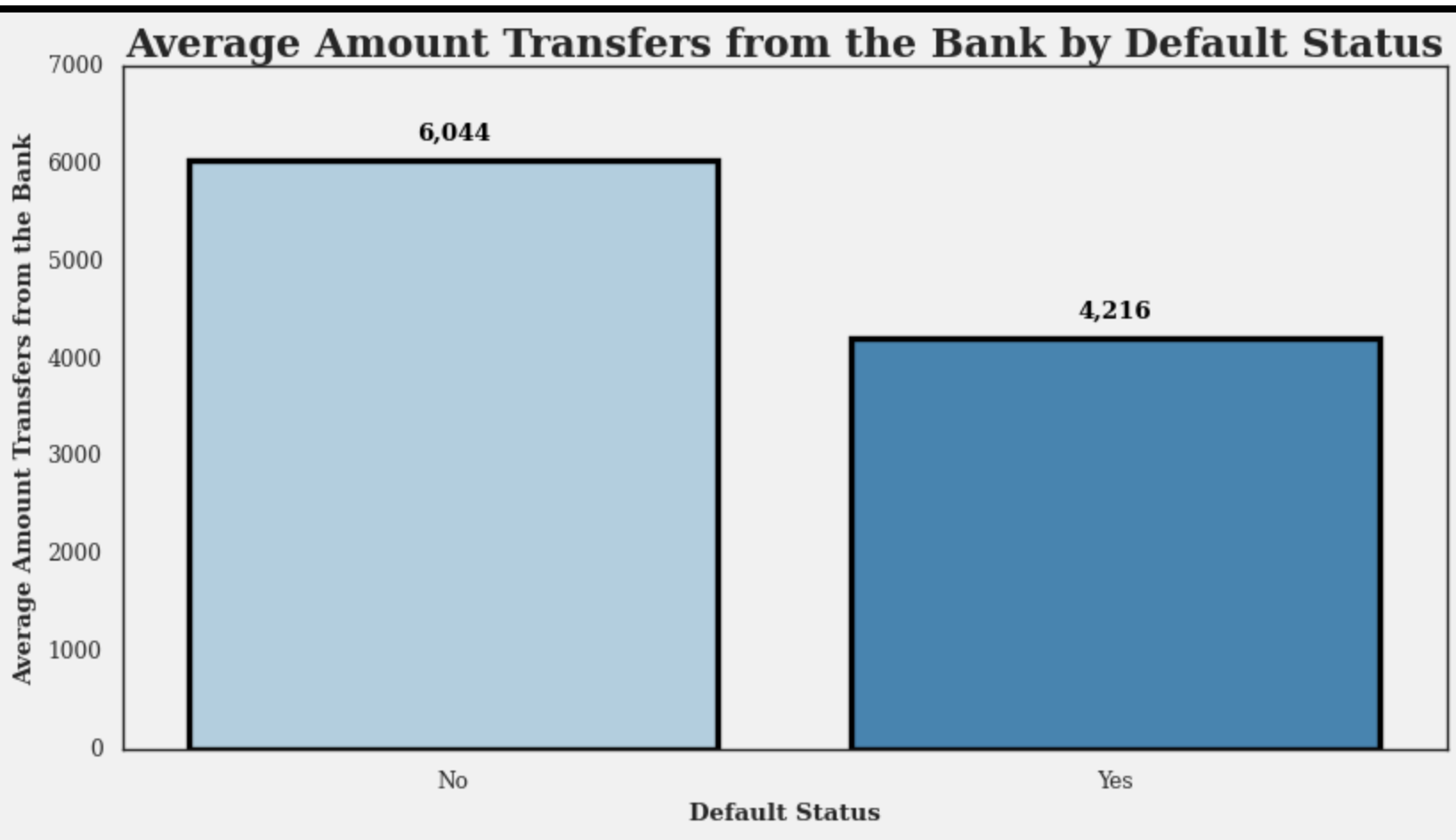
The graph indicates that customers who have not defaulted have deposited more money on average than those who have defaulted. The difference between the two groups is about KES 8,000. This is a large difference, so this variable could be a good predictor of whether or not a customer will default on their loan.

Average Amount Received by Default Status



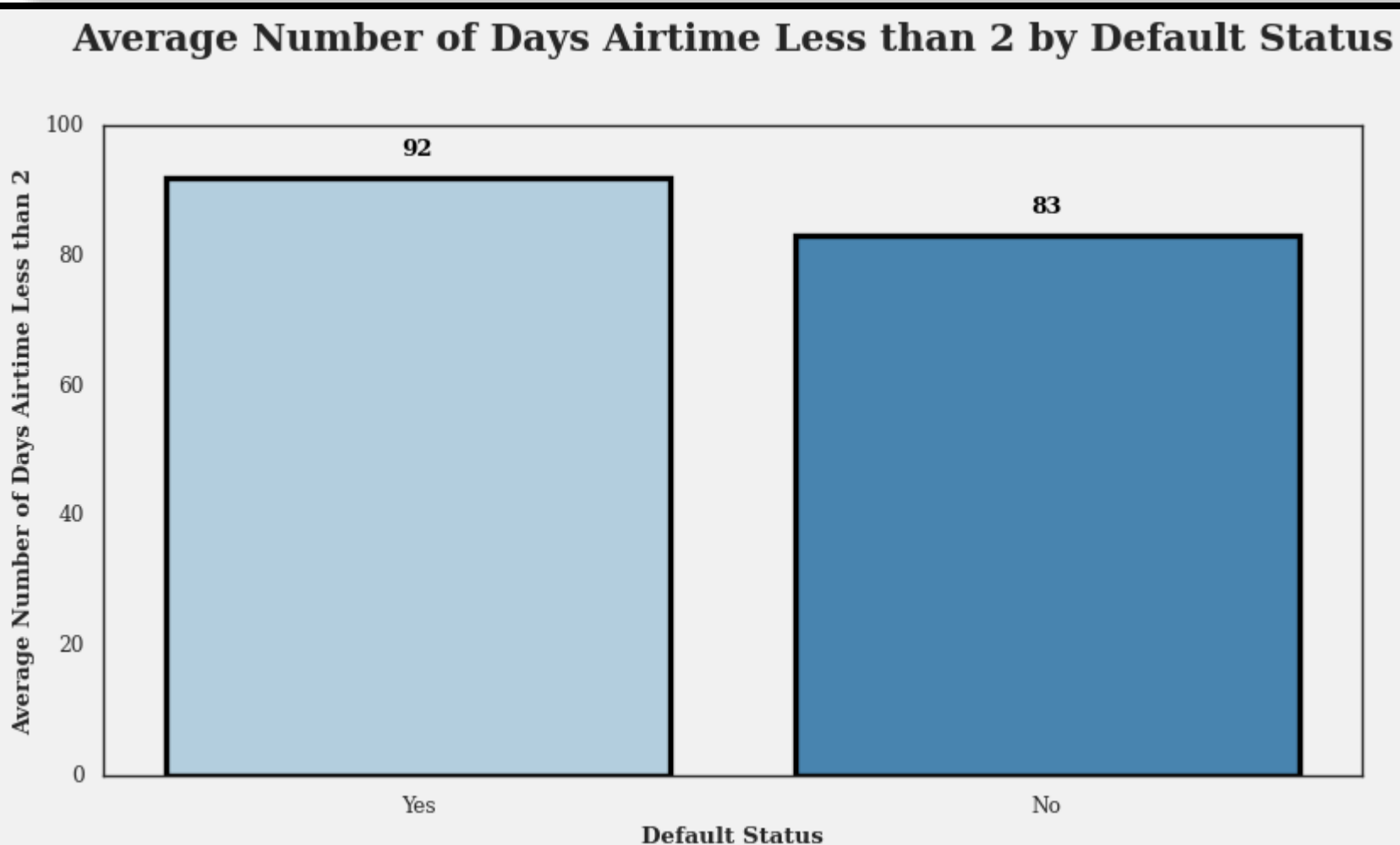
The default status category that has a higher average amount received is No. This means that customers who have not defaulted on their loans tend to receive more money than customers who have defaulted on their loans. This could be because customers who receive more money are more likely to be able to pay back their loans. It could also be because customers who receive more money are more likely to be able to qualify for larger loans.

Average Amount Transfers from the Bank by Default Status



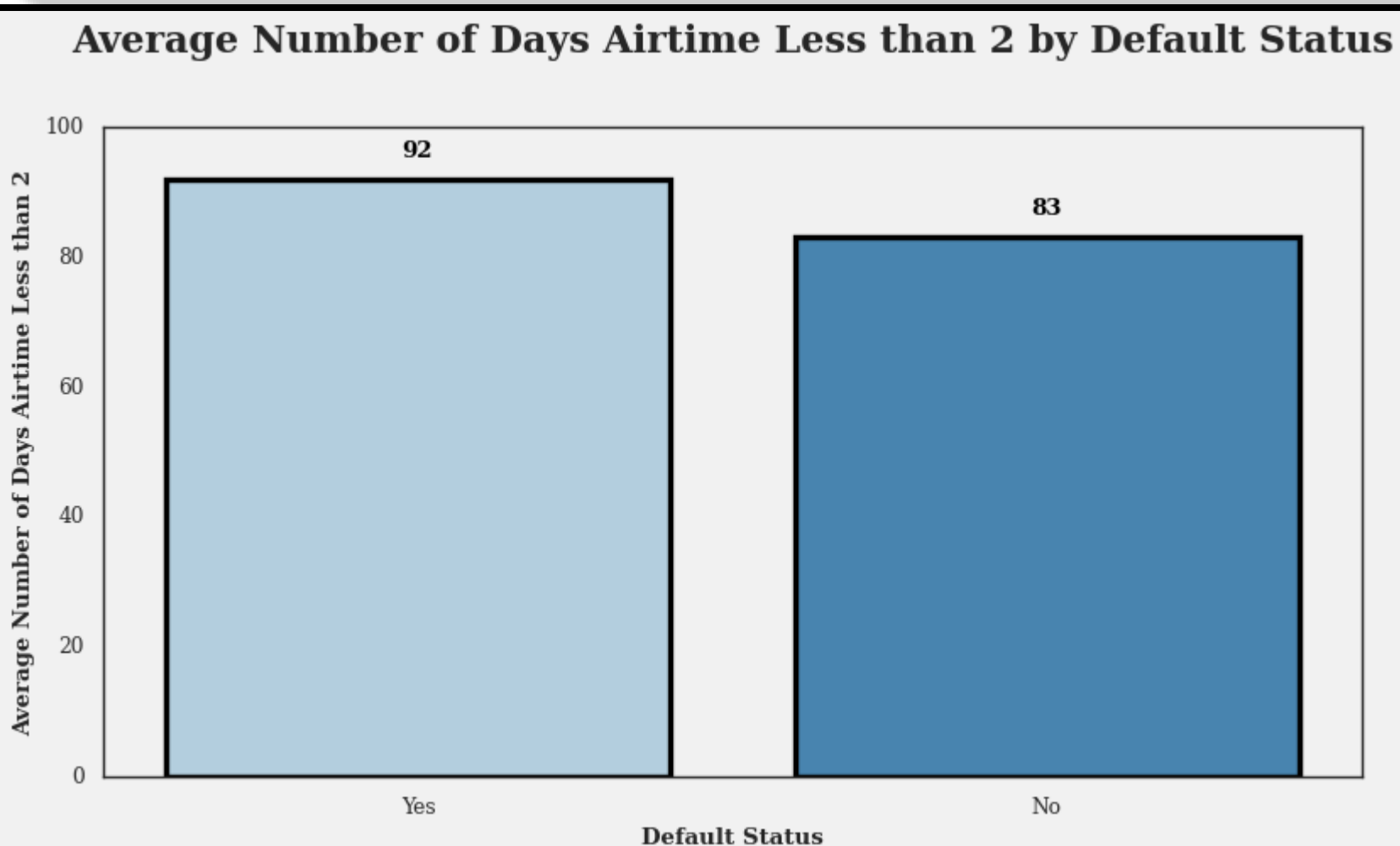
The bar graph shows the average amount of money transferred from the bank by customers who have defaulted or not defaulted on their loans. The graph indicates that customers who have not defaulted on their loans transfer more money from the bank than those who have defaulted. The difference between the two groups is about KES 1,828.

Average Number of Days Airtime Less than 2 units by Default Status



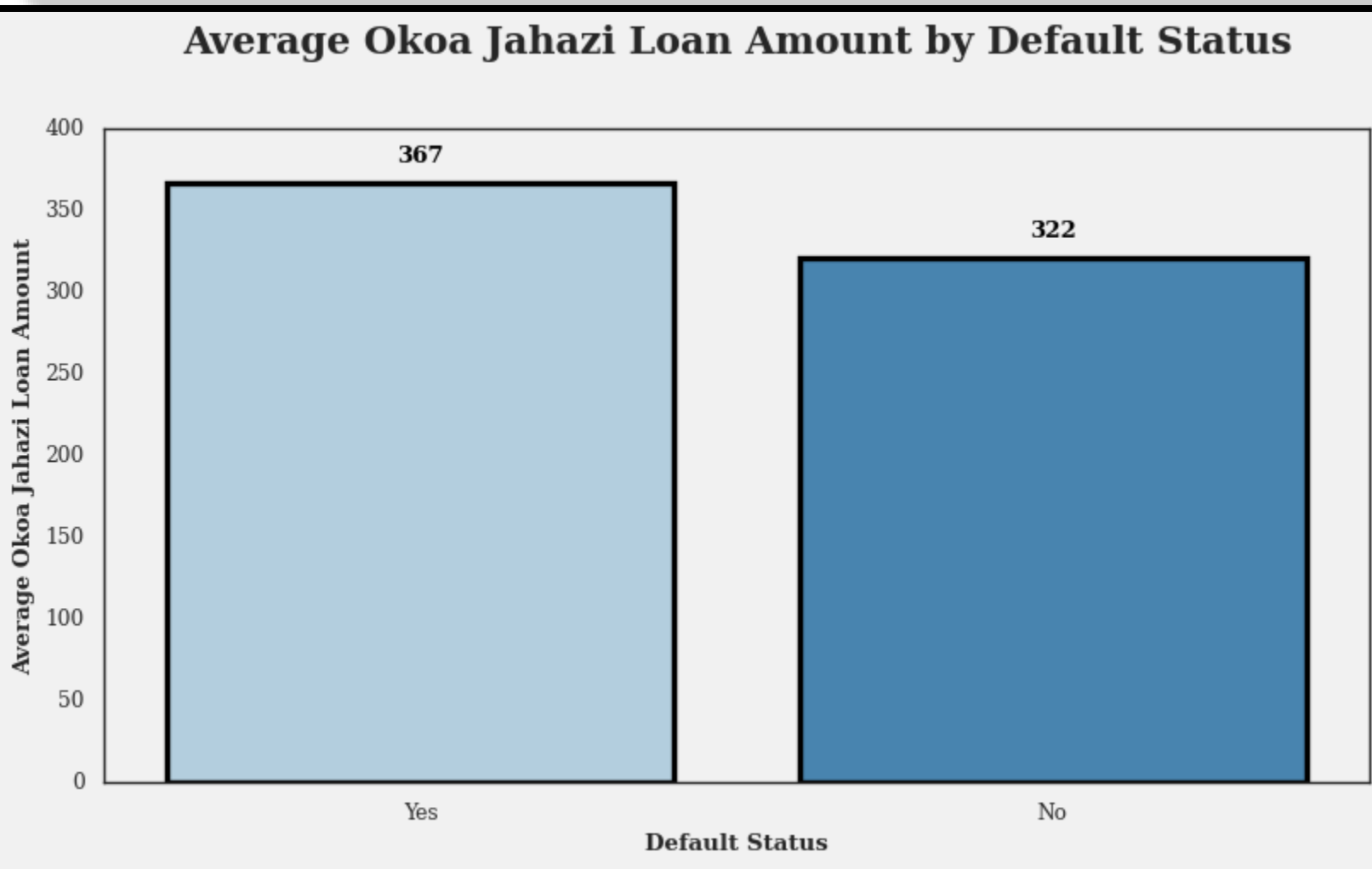
The graph suggests that customers who defaulted on their loans had more days with low airtime than those who did not. This could imply that customers who default on their loans are more likely to have financial difficulties or lower income. However, it is also possible that customers who default on their loans are more likely to use their airtime for mobile money transactions, which would cause them to run out of airtime more quickly.

Average Number of Days Airtime Less than 2 units by Default Status



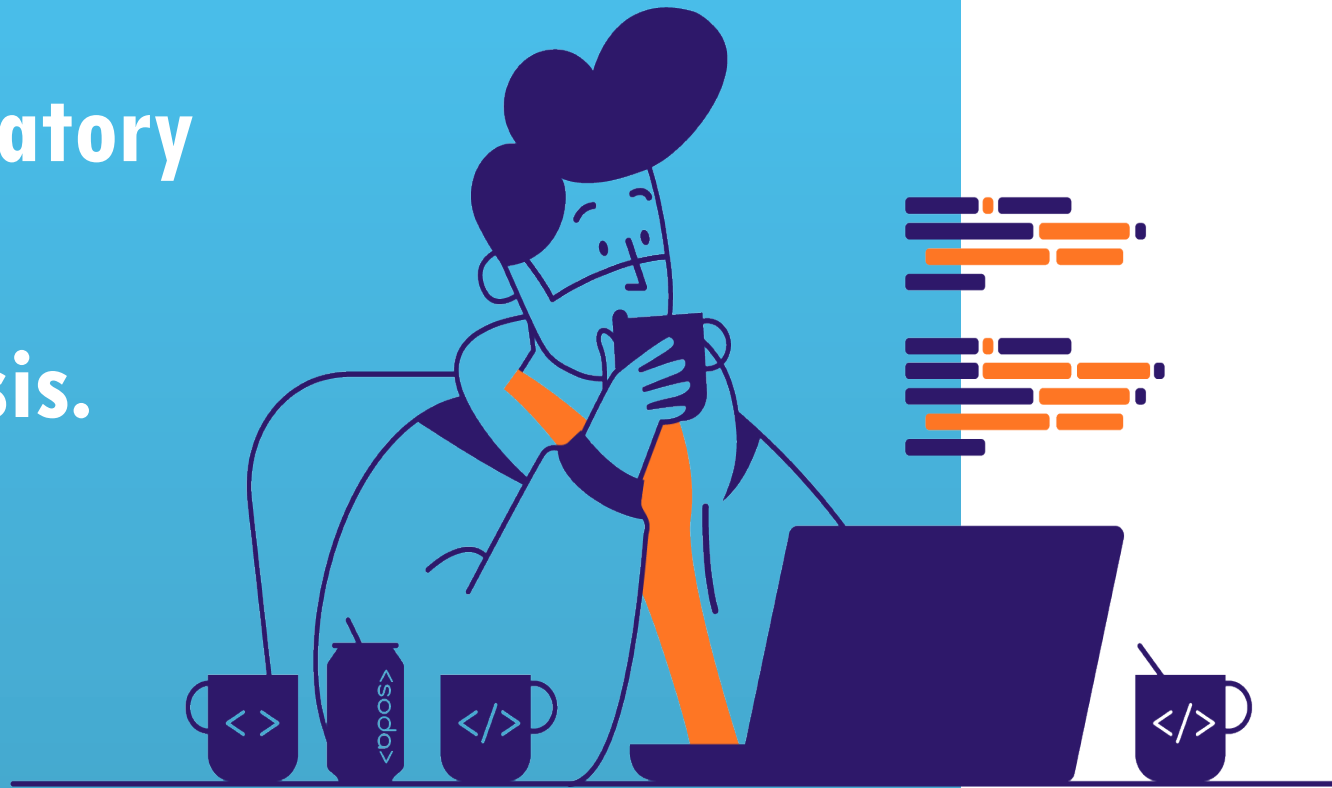
The graph suggests that customers who defaulted on their loans had more days with low airtime than those who did not. This could imply that customers who default on their loans are more likely to have financial difficulties or lower income. However, it is also possible that customers who default on their loans are more likely to use their airtime for mobile money transactions, which would cause them to run out of airtime more quickly.

Average Okoa Jahazi Loan Amount by Default Status



The bar graph that shows the average Okoa Jahazi loan amount for customers who defaulted or not on their loans. The graph indicates that customers who defaulted had a slightly higher average loan amount than those who did not. The difference is about 45 shillings.

Exploratory Data Analysis.



This process of examining and summarizing data sets using various techniques such as visualization and descriptive statistics to help to identify patterns and relationships in the data.



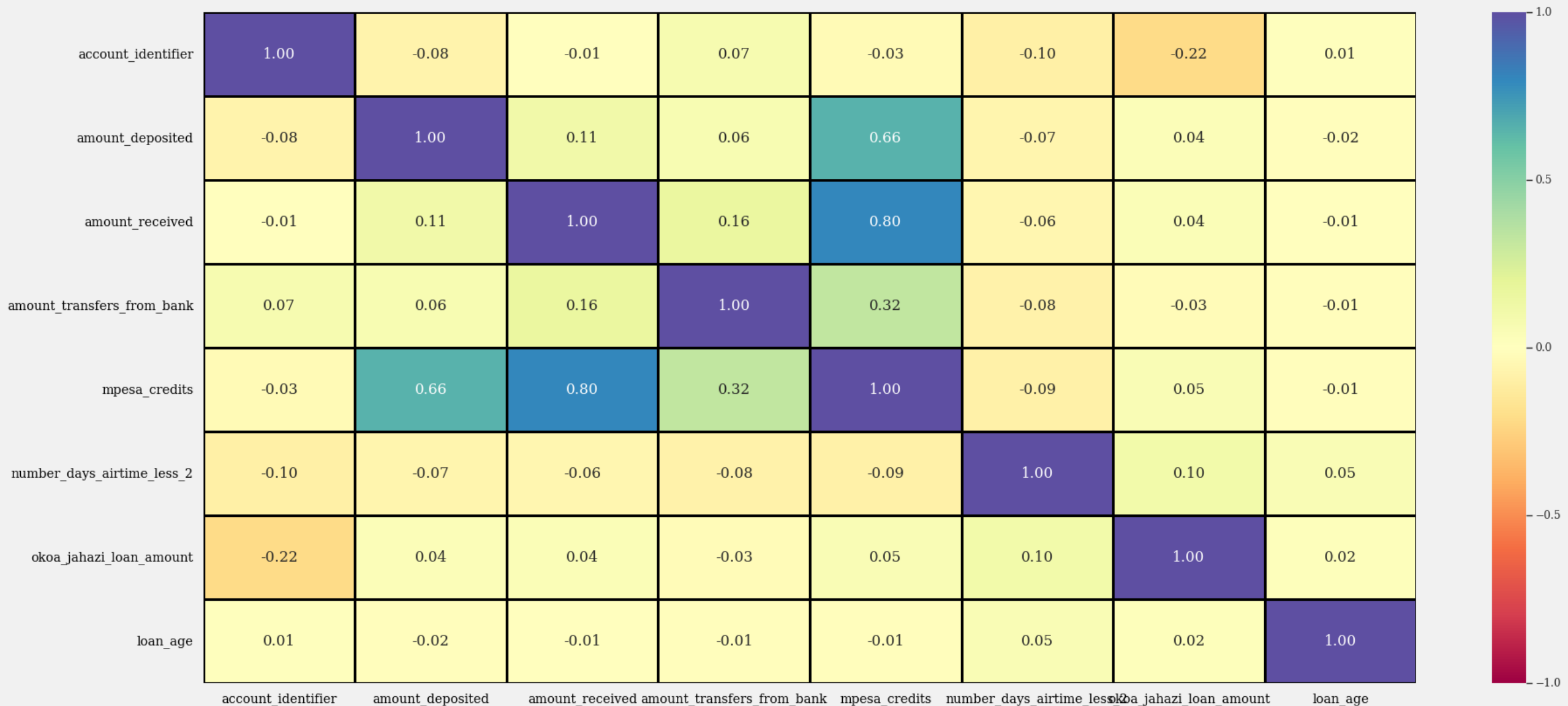
Univariate Data Analysis.



Bivariate Data Analysis.



Multivariate Data Analysis.



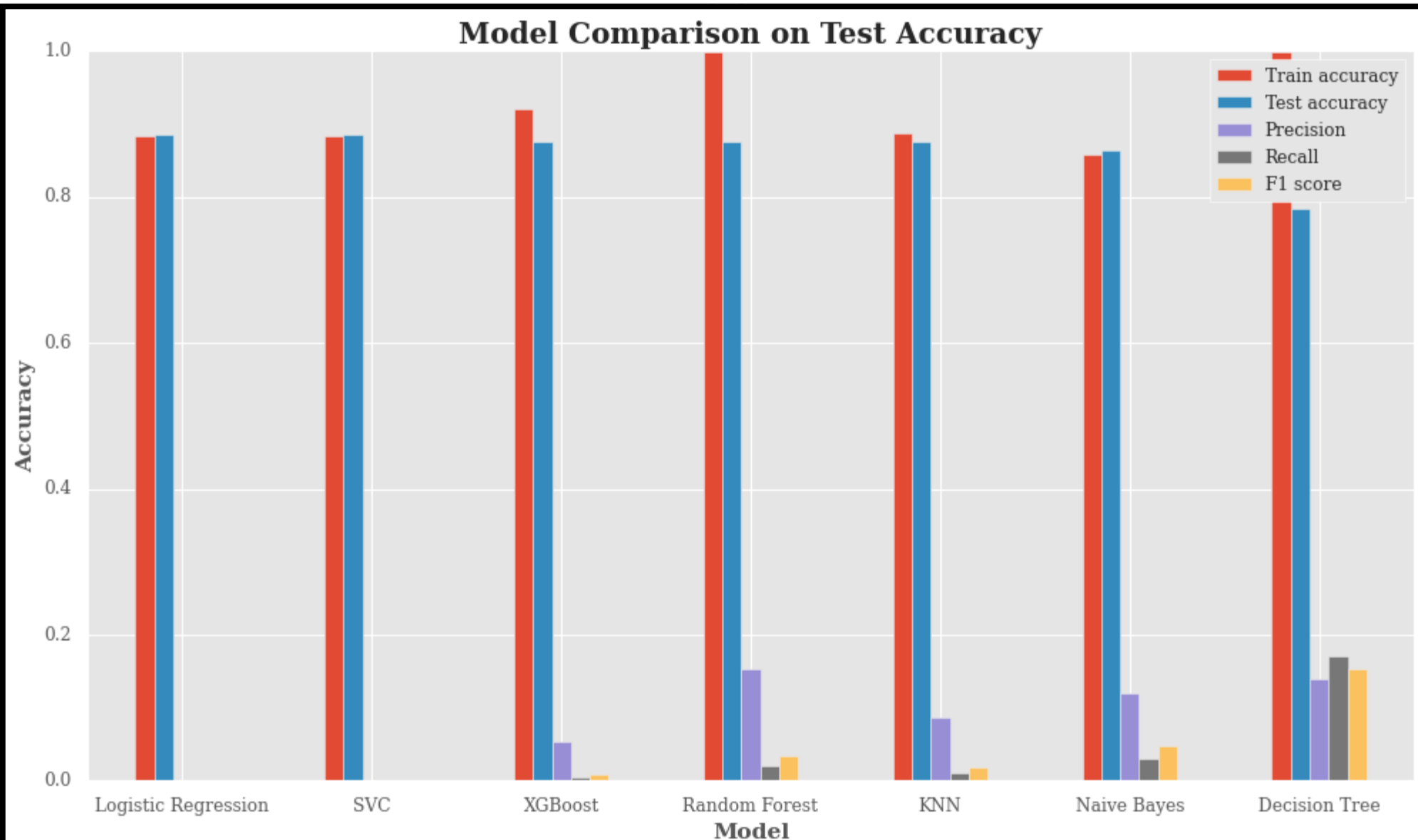
We can note that there are two strong positive correlations between `mpesa_credits` and `amount_received` and `amount_deposited`. We also note a weak positive correlation between `mpesa_credits` and `amount_transfers_from_bank`.

Question:

What predictive
models are we going
to build?



Vanilla Models



Based on the evaluation plot, it seems that none of the models are performing very well on the classification task. All of them have very low precision, recall, and F1 score, which indicate that they are not able to correctly identify the default and non-default cases. The accuracy metric is also misleading, as it does not account for the class imbalance in the data. A model that always predicts non-default will have a high accuracy, but a very poor performance on the default cases.

Among the models, the logistic regression and SVC models have the highest test accuracy, but they also have zero precision and recall, which means that they never predict default. This is not desirable, as we want to identify the default cases and take appropriate actions. The random forest model has the highest precision, recall, and F1 score among the models, but they are still very low compared to the ideal values. The random forest model also has a high train accuracy, but a lower test accuracy, which suggests that it is overfitting the training data and not generalizing well to the test data.

As these are vanilla models, then my next step of action would be to try some of the approaches below to see if I can improve the performance of the models:

- **Feature engineering**
- **Hyperparameter tuning**
- **Ensemble methods**
- **Resampling techniques**
- **Cross-validation techniques**

Model Evaluation



Question:

What model did we
pick?



Random Forest Classifier

Resampling - Random Over-Sampling

Calculating Baseline Accuracy – 88.30%

Building a Random Forest Classifier

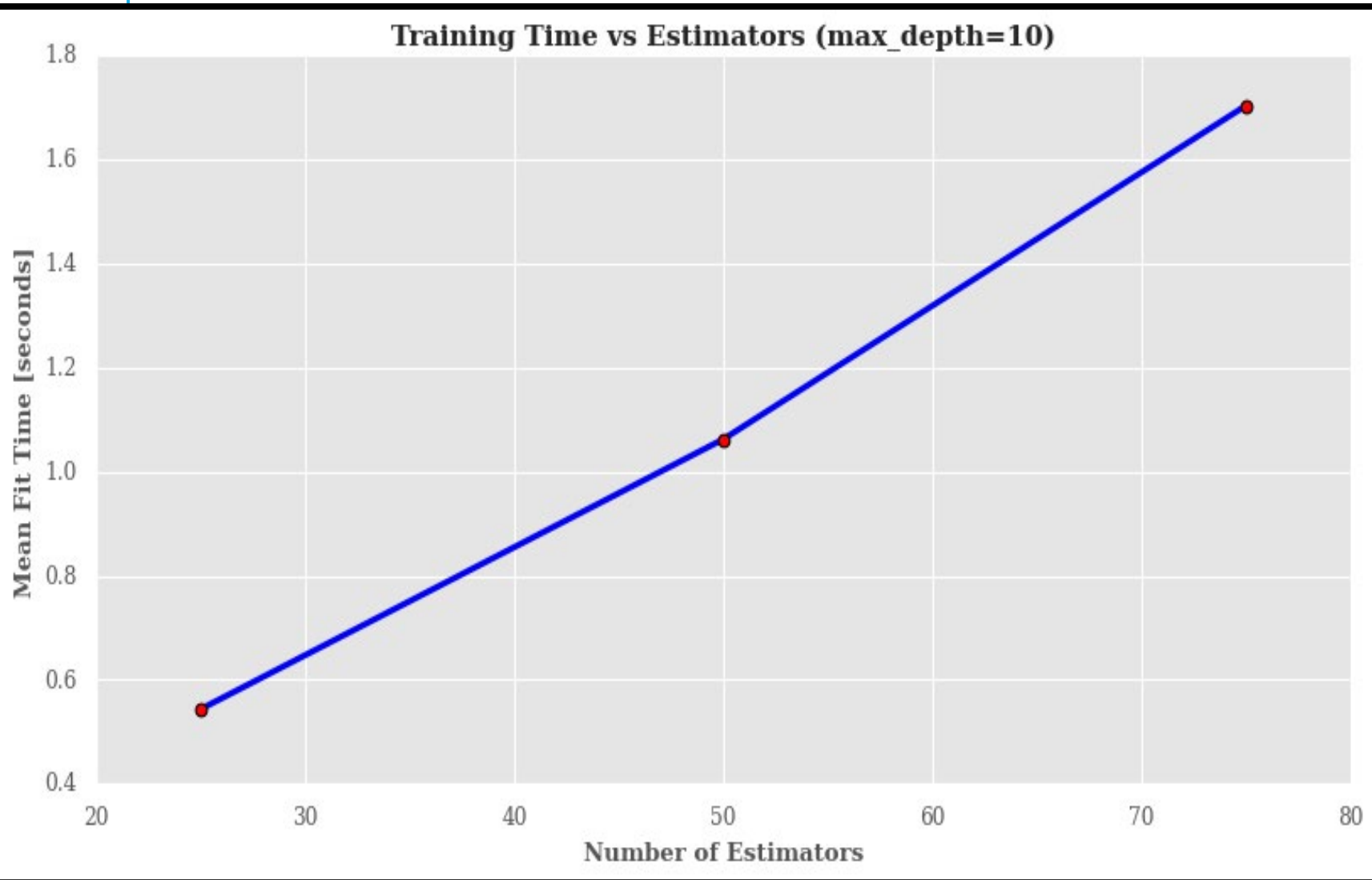
Perform cross-validation using over-sampled data

Training the model

- **There are 12 candidate models being evaluated. These could be 12 different algorithms, 12 different hyperparameter configurations, etc.**
- **For each candidate model, 5-fold cross-validation is being used to evaluate it.**
- **In 5-fold cross-validation, the data is split into 5 folds or subsets. Each fold is held out in turn as a validation set, while the remaining folds are used as the training set. A model is fitted on the training set and then tested on the validation set.**
- **So for each candidate model, it is fitted 5 times on different combinations of training and validation sets.**
- **With 12 candidates and 5 folds per candidate, this results in $12 * 5 = 60$ total fits or training procedures of the models.**

Random Forest Classifier.

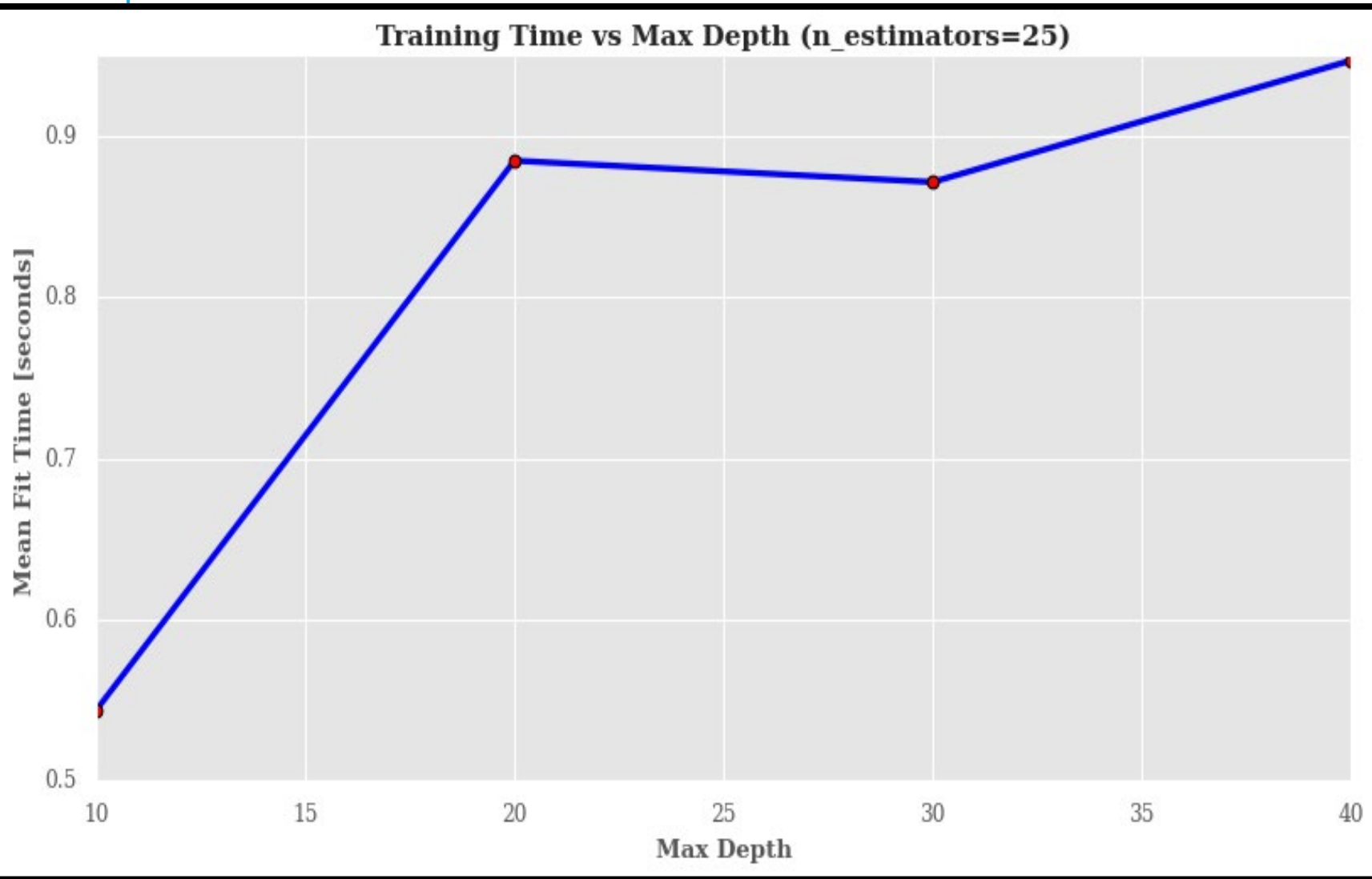
Training Time versus n_estimators.



There is a clear relationship between the number of estimators and the training time. As the number of estimators increases, the training time increases. This makes sense, as the model has to fit more trees.

Random Forest Classifier.

Training Time versus `n_estimators`.



There is a general upward trend between `max_depth` and training time. As `max_depth` increases, the training time increases.

Question:

How well does our
model perform?



Model Evaluation

	Accuracy Scores (%)
Train accuracy	99.87%
Test accuracy	84.76%

The scores you have provided tell us that the model has high training accuracy but low test accuracy. This is a sign of overfitting. Overfitting occurs when a model becomes too complex and learns the training data too well, including the noise in the data. As a result, the model is not able to generalize to new data and perform well on the test set.



Model Evaluation (Cont)

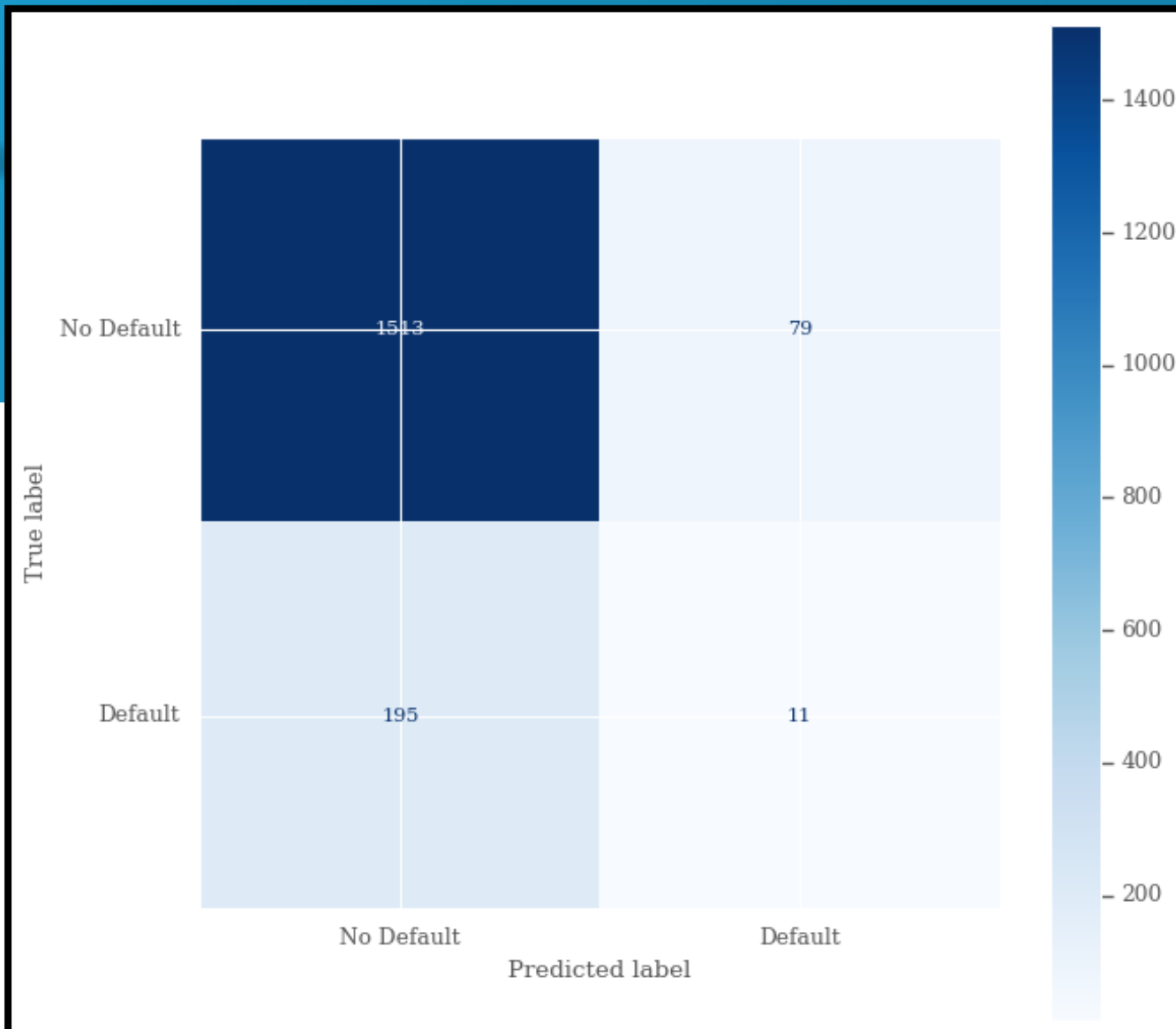
Confusion Matrix

The top left quadrant is the number of true negatives (TN), which means that the model correctly predicted that the value was negative . In this case, there are 1513 TNs.

The top right quadrant is the number of false positives (FP), which means that the model incorrectly predicted that the value was positive when it was actually negative (0). In this case, there are 79 FPs.

The bottom left quadrant is the number of false negatives (FN), which means that the model incorrectly predicted that the value was negative when it was actually positive .In this case, there are 195 FNs.

The bottom right quadrant is the number of true positives (TP), which means that the model correctly predicted that the value was positive. In this case, there are 11 TPs.



Model Evaluation (Cont)

Classification Report

	Precision	Recall	F1-Score	Support
0	0.89	0.95	0.92	1592
1	0.12	0.05	0.07	206
Accuracy				1798
Macro Avg	0.50	0.50	0.50	1798
Weighted Avg	0.80	0.85	0.82	1798

The model achieved an overall accuracy of 0.85. For the Non-Default class:

- Precision is 0.89, meaning 89% of examples the model predicted as Non-Default were actually Non-Default.
- Recall is 0.95, meaning the model correctly identified 95% of true Non-Default examples.
- F1-score is 0.92, indicating a good balance of precision and recall.

For the Default class:

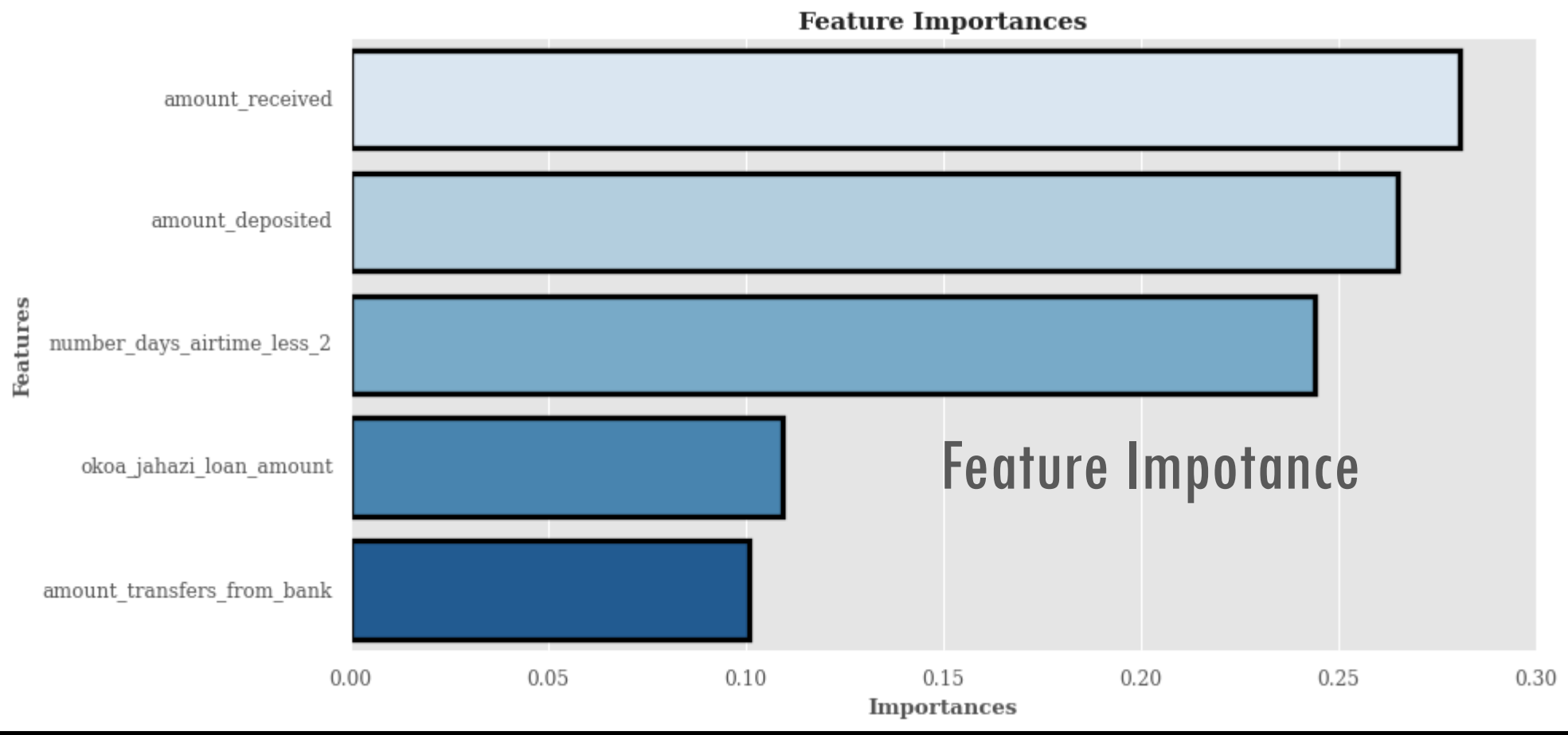
- Precision is low at 0.13, meaning only 13% of examples predicted as Default were truly Default.
- Recall is very low at 0.05, meaning the model only detected 5% of true Default examples.
- F1-score is 0.08, indicating poor performance on the Default class.

The macro average scores consider both classes equally and show that overall, precision and recall are balanced at 0.51 and 0.50. The weighted average scores account for class imbalance, showing higher weighted precision of 0.80 and recall of 0.85.

Question:

What are our top features?





The features are related to the amount of money received, deposited, transferred, or borrowed by customers, as well as the number of days they have less than 2 units of airtime. The graph indicates that the most important feature is `amount_received`. The second most important feature is `amount_deposited`. The third most important feature is `number_days_airtime_less_2`. The least important features are `okoja_jahazi_loan_amount` and `amount_transfers_from_bank`. The graph also shows that the feature importances sum up to 1, which means that they are normalized and represent the relative contribution of each feature to the prediction model.

What next?

Summary of Findings

“ A summary of the key findings, organized by research questions or objectives, and supported by evidence from the data.

Please move to the next page



There are significantly more accounts that have been repaid on time than accounts that have defaulted. This is a good thing for lenders, as it means that they are more likely to get their money back. It is also a good thing for borrowers, as it means that they are more likely to be able to qualify for loans in the future. This also raises the issue of class imbalance between the negative and positive class.**

The distribution of loan age is skewed right, with most customers having loans less than 90 days old. This means that most customers have recently taken out loans. This could be because the company is new and is still growing its customer base, or it could be because the company offers short-term loans.

The variable `number_days_airtime<2` displays characteristics of a mixture distribution. This means that the distribution of this variable is likely made up of two or more different distributions. For example, imagine a distribution of the number of days that customers have gone without purchasing airtime. This distribution might be a mixture of two distributions: one distribution for customers who regularly purchase airtime and another distribution for customers who rarely purchase airtime.

The values for the variables ``amount_deposited``, ``amount_received``, ``amount_transfers_from_bank``, ``mpesa_credits``, and ``okoa_jahazi_loan_amount`` are centered around zero and skew to the right side of the distribution. This means that the majority of values for these variables are close to zero, but there are more large positive values than large negative values. For example, imagine a distribution of customer account balances, where most customers have balances close to zero, but there are a few customers with very large balances. This distribution would be skewed to the right.

Mobile money transactions are more common on weekdays than on weekends, with Tuesdays having the highest number of transactions. This means that people are more likely to send and receive money using their mobile phones during the workweek, when they are more likely to be making purchases or paying bills. On the other hand, mobile money transactions are less common on weekends, when people are more likely to be relaxing or spending time with their families.

Here are some possible explanations for why mobile money transactions are more common on weekdays than on weekends:**

- **People are more likely to be working or running errands on weekdays. This means that they are more likely to need to send or receive money to pay for goods and services.**
- **Businesses are more likely to be open on weekdays. This means that people have more opportunities to use mobile money to make payments.**
- **People may be more likely to use mobile money to pay for bills on weekdays. This is because many bills are due on or around the first of the month, which typically falls on a weekday.**

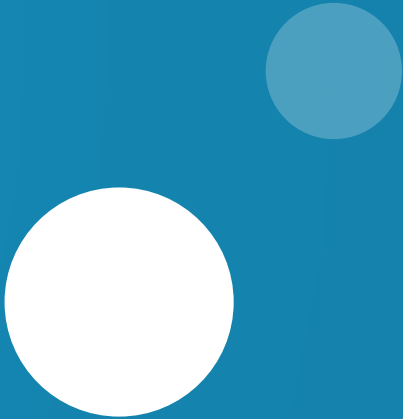
The daily average amount of transfers for the period February shoes an overall upward trend. There are a few possible explanations for this upward trend. One possibility is that the bank was running a promotion or offering a special incentive for customers to make transfers. Another possibility is that the bank's customers were simply becoming more comfortable with making transfers online or through mobile banking. It is also possible that there was an increase in the overall number of transactions taking place during this period.

The average loan amount by loan age, for both loans that have defaulted and loans that have not defaulted shows that the average loan amount for both types of loans increases with loan age. This is likely because borrowers are able to qualify for larger loans as they build up their credit history and become more established financially.

However, the average loan amount for defaulted loans is consistently higher than the average loan amount for non-defaulted loans. This suggests that borrowers who take out larger loans are more likely to default. This may be because they are more likely to be overextended financially and less likely to be able to withstand unexpected financial shocks.

The graph also shows that the gap between the average loan amount for defaulted loans and the average loan amount for non-defaulted loans widens with loan age. This suggests that the risk of default increases for larger loans over time.

Overall, the graph suggests that lenders should be careful when making large loans, especially to borrowers with limited credit history. Lenders should also carefully consider the borrower's financial situation and ability to repay the loan before making a decision.

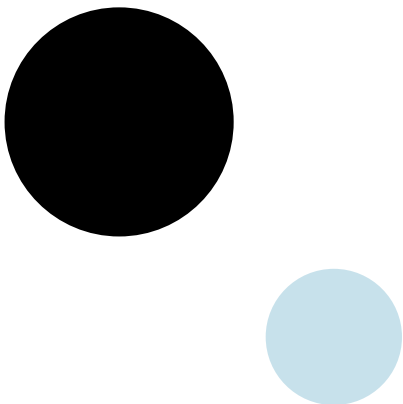


The average amount deposited by customers who have defaulted or not on their loans shows that customers who have not defaulted have deposited more money on average than those who have defaulted. The difference between the two groups is about KES 8,000. This is a large difference, so this variable could be a good predictor of whether or not a customer will default on their loan.

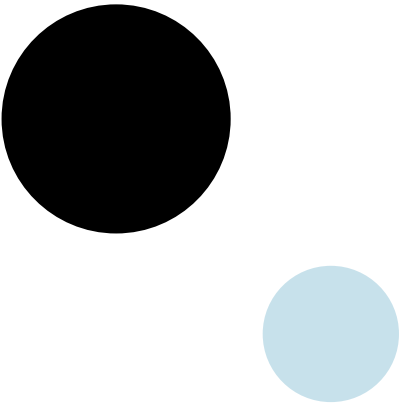
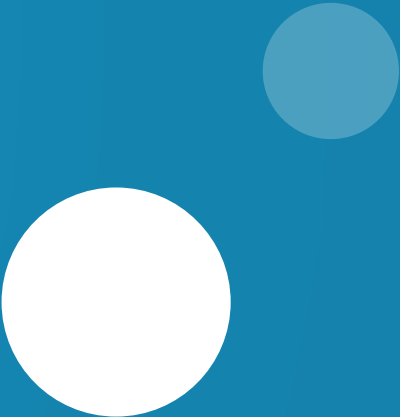
The average number of days that customers have less than 2 units of airtime left on their phones, based on whether they defaulted on their loans or not suggests that customers who defaulted on their loans had more days with low airtime than those who did not. This could imply that customers who default on their loans are more likely to have financial difficulties or lower income. However, it is also possible that customers who default on their loans are more likely to use their airtime for mobile money transactions, which would cause them to run out of airtime more quickly.

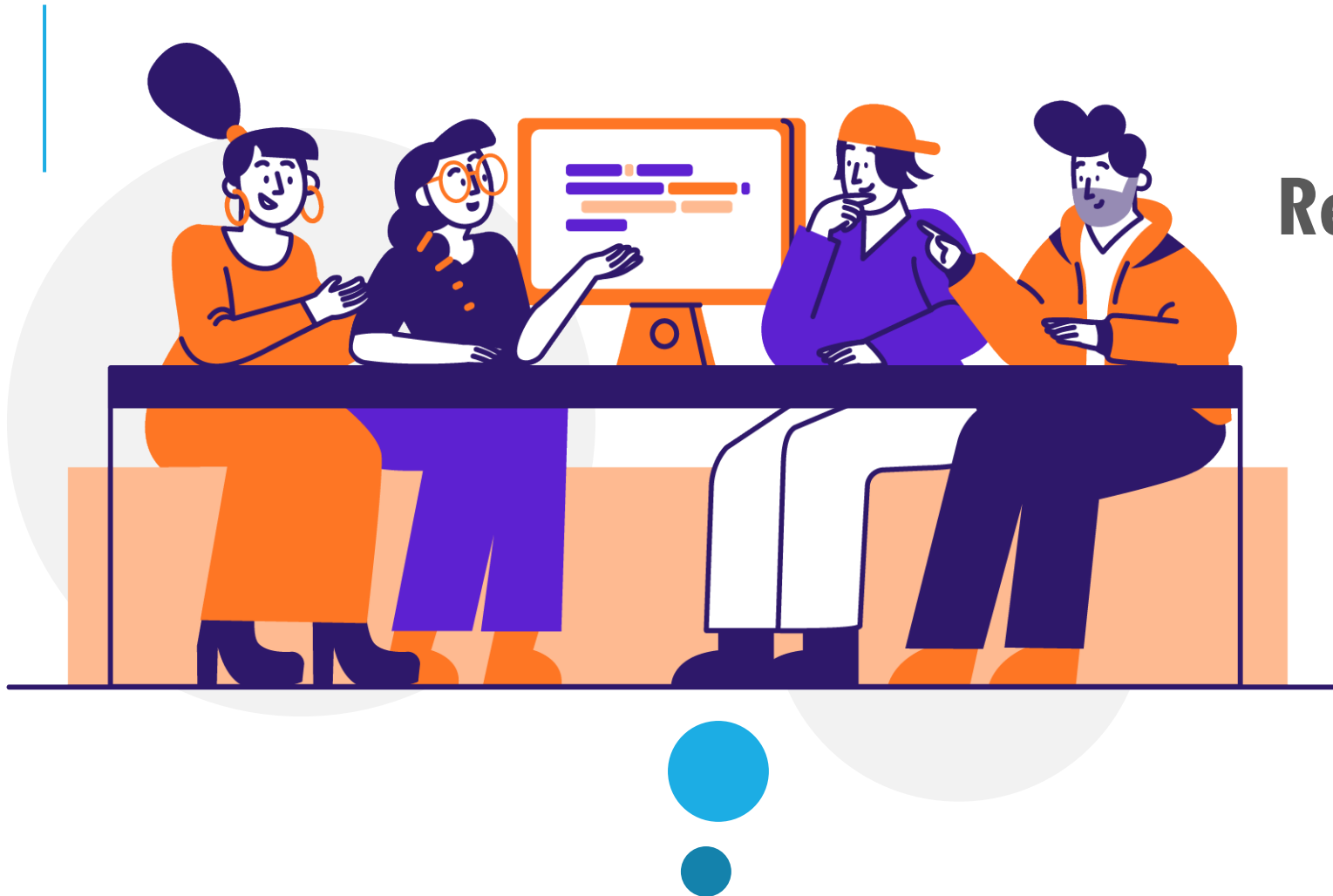
Customers who have not defaulted on their loans tend to receive more money than customers who have defaulted on their loans. This could be because customers who receive more money are more likely to be able to pay back their loans. It could also be because customers who receive more money are more likely to be able to qualify for larger loans.

The average amount of money transferred from the bank by customers who have defaulted or not defaulted on their loans indicates that customers who have not defaulted on their loans transfer more money from the bank than those who have defaulted. The difference between the two groups is about KES 1,828.



The average Okoa Jahazi loan amount for customers who defaulted or not on their loans indicates that customers who defaulted had a slightly higher average loan amount than those who did not. The difference is about 45 shillings.





Recommendations

My actionable insights for the stakeholders at CBA relevant to the business problem supported by evidence from the data and aligned with the project objectives and scope.

Lenders should be careful when making large loans, especially to borrowers with limited credit history. This is because the data shows that borrowers who take out larger loans are more likely to default, especially over time.

Lenders should also carefully consider the borrower's financial situation and ability to repay the loan before making a decision. This means looking at factors such as the borrower's income, debt-to-income ratio, and credit score.

Lenders could use the average amount deposited by customers as a predictor of whether or not a customer is likely to default on their loan. Customers who have not defaulted on their loans have deposited more money on average than those who have defaulted.

Lenders could also use the average number of days that customers have less than 2 units of airtime left on their phones as a predictor of whether or not a customer is likely to default on their loan. Customers who defaulted on their loans had more days with low airtime than those who did not. This suggests that customers who default on their loans may be more likely to have financial difficulties or lower income.

Lenders could also consider the average amount of money transferred from the bank and the average Okoa Jahazi loan amount as predictors of whether or not a customer is likely to default on their loan. Customers who have not defaulted on their loans transfer more money from the bank and have slightly lower Okoa Jahazi loan amounts than

Offer financial education programs to borrowers. This can help borrowers to better understand their finances and how to manage their debt.

Provide loan repayment assistance programs to borrowers who are struggling to repay their loans. This could include things like loan modification or forbearance.

Partner with other financial institutions to share data and insights on borrowers. This can help lenders to make more informed decisions about who to lend to and how much to lend.

What next?

Challenging my solution

“ A summary of the key findings, organized by research questions or objectives, and supported by evidence from the data.

Please move to the next page



CLASS IMBALANCE:

The dataset contains a large number of non-default cases and a small number of default cases. This means that the model will be biased towards predicting non-default cases, which may lead to poor performance on the default cases. To address this issue, you can use resampling techniques to balance the classes in the dataset. This can be done by either oversampling the minority class or undersampling the majority class. Some common resampling techniques include SMOTE, ADASYN, and Random Oversampling.

Use cost-sensitive learning algorithms. These algorithms take into account the cost of misclassifying each class and try to minimize the overall cost. Some common cost-sensitive learning algorithms include cost-sensitive SVM and cost-sensitive decision trees.

Use ensemble learning. Ensemble learning algorithms combine the predictions of multiple machine learning models to improve the overall accuracy. Some common ensemble learning algorithms include random forests, gradient boosting machines, and XGBoost.

SKEDWED DISTRIBUTION OF LOAN AGE:

Use transformation techniques to normalize the distribution of the loan age variable. This can be done by using a logarithmic transformation or a square root transformation. Use machine learning algorithms that are robust to skewed distributions, such as decision trees and random forests.

MIXED DISTRIBUTION OF number days airtime<2:

Use clustering to identify the different groups of customers in the data. Once the customers have been clustered, you can train a separate machine learning model for each cluster. Use model ensembles to combine the predictions of multiple machine learning models, each trained on a different subset of the data. This can help to improve the accuracy of the overall prediction.

SKEWED DISTRIBUTIONS of amount deposited, amount received, amount transfers from bank, mpesa credits, and okoa jahazi loan amount:

Use transformation techniques to normalize the distributions of these variables. This can be done by using a logarithmic transformation or a square root transformation. Use machine learning algorithms that are robust to skewed distributions, such as decision trees and random forests.

MORE MOBILE MONEY TRANSACTIONS ON WEEKDAYS THEN WEEKENDS:

Use time series analysis to identify patterns in the mobile money transaction data. This can help you to understand how the number of transactions varies over time, including on weekdays versus weekends. Use machine learning algorithms to predict the number of mobile money transactions on a given day. This can be useful for businesses that need to forecast demand for their products and services.

Questions??

Made with  by Richard Taracha
taracharichard@gmail.com

