Credit Risk Analysis

which machine learning model handles credit risk data the best?

# Background Story:

You are a Consultant Data Scientist employed at E Corp and your newest client, Piggy Bank, has come to you with a problem, they are trying to bring their banking systems into the 21st century by implementing a Machine Learning model, but they cannot decide on which model to implement.

To help you along your journey, Piggy Bank has supplied you with a sample data set.

With this data set, Piggy Bank would like you to advise them on what Machine Learning models they should consider implementing to help speed up the process of classifying the creditworthiness of individuals who are applying for a financial service, mainly in the form of loans and credit cards.

Contents

[Introduction: 2](#_Toc145448302)

[Data: 2](#_Toc145448303)

[**Link:** Click Here 2](#_Toc145448304)

[Dataset Columns: 2](#_Toc145448305)

[Visualization 3](#_Toc145448306)

[Histogram of Age Column: 3](#_Toc145448307)

[Bar chart of Number of Children: 4](#_Toc145448308)

[Number of Dependents 5](#_Toc145448309)

[Applicant Income 7](#_Toc145448310)

[Employment Status 8](#_Toc145448311)

[Performance Metrics 10](#_Toc145448312)

[Accuracy: 10](#_Toc145448313)

[Precision and Recall: 10](#_Toc145448314)

[F1-Score: 10](#_Toc145448315)

[AUC and ROC Curves 10](#_Toc145448316)

[AUC 10](#_Toc145448317)

[ROC 10](#_Toc145448318)

[ML Models 10](#_Toc145448319)

[Decision Tree: 11](#_Toc145448320)

[Logistic Regression: 11](#_Toc145448321)

[Naive Bayes: 11](#_Toc145448322)

[Decision Tree: Performance and Evaluation 12](#_Toc145448323)

[Naïve Bayes: Performance and Evaluation 12](#_Toc145448324)

[Logistic Regression: Performance and Evaluation 13](#_Toc145448325)

[Comparison 14](#_Toc145448326)

[Naive Bayes: 14](#_Toc145448327)

[Decision Tree: 14](#_Toc145448328)

[Logistic Regression: 14](#_Toc145448329)

[ROC Curve 15](#_Toc145448330)

[Naive Bayes 15](#_Toc145448331)

[Decision Tree 15](#_Toc145448332)

[Logistic Regression 15](#_Toc145448333)

[Best Model 16](#_Toc145448334)

**Credit Risk Analysis**

# Introduction:

For my final project, I decided to analyze loan applicant credit risk, this project is a data-driven initiative aimed at developing and implementing predictive models to assess and classify the creditworthiness of individuals applying for financial services, particularly loans or credit cards. By leveraging a diverse set of client attributes and utilizing machine learning algorithms such as Naive Bayes, Decision Tree, and Logistic Regression, this project seeks to improve the efficiency and accuracy of credit risk assessment for a bank or lending institution.

# Data:

This dataset comprises 1,225 rows and 15 columns, providing a comprehensive collection of client attributes sourced from GitHub, which was originally sourced from Kaggle. These attributes serve as crucial inputs for predictive modeling in the context of credit risk analysis.

## **Link:** [Click Here](https://github.com/deepanshu88/Datasets/blob/master/CreditData/Loan%20Data.csv)

## Dataset Columns:

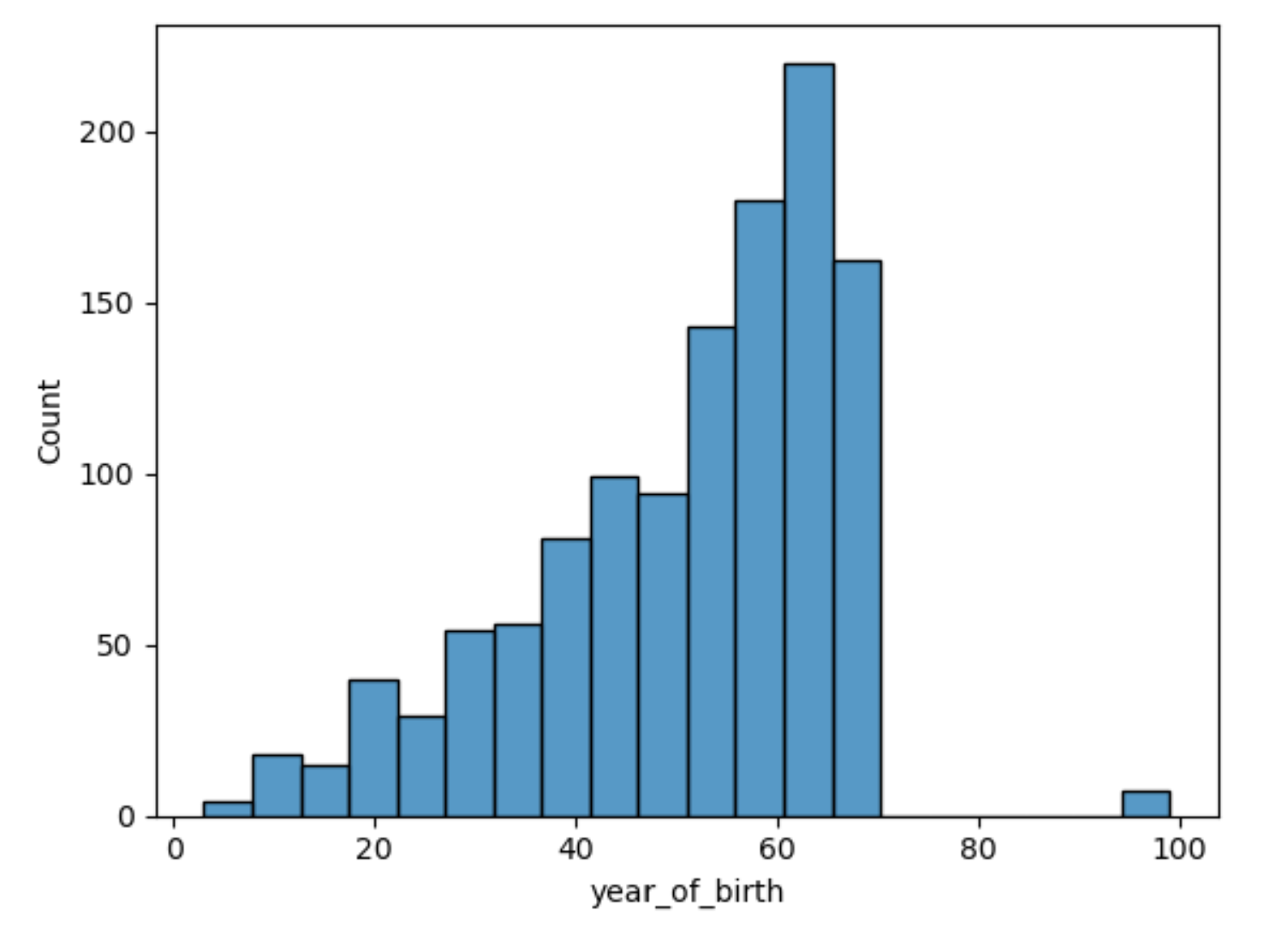
1. **Year of Birth (yob):** The applicant's birth year, used for age calculation.
2. **Number of Children (nkid):** The count of dependent children.
3. **Number of Other Dependents (dep):** The count of other dependents besides children.
4. **Home Phone (phon):** Indicates whether the applicant has a home phone.
5. **Spouse's Income (sinc):** The income of the applicant's spouse if applicable.
6. **Employment Status (aes):** Categorizes the applicant's employment status.
7. **Applicant's Income (dainc):** The income of the applicant.
8. **Residential Status (res):** Indicates the type of residential status the applicant has.
9. **Value of Home (dhval):** The value of the applicant's home if they are an owner.
10. **Mortgage Balance Outstanding (dmort):** The outstanding mortgage balance, if applicable.
11. **Outgoings on Mortgage/Rent (doutm):** Monthly outgoings on mortgage or rent.
12. **Outgoings on Loans (doutl):** Monthly outgoings on loans.
13. **Outgoings on Hire Purchase (douthp):** Monthly outgoings on hire purchase.
14. **Outgoings on Credit Cards (doutcc):** Monthly outgoings on credit cards.
15. **Good/Bad Indicator (Bad):** Binary indicator of creditworthiness (1 represents bad, 0 represents good).
16. **Target Variable (good\_or\_bad):** The variable to be predicted by the models.

The project aims to build predictive models that use the provided data to determine whether an applicant is a good or bad credit risk, ultimately aiding in making informed lending decisions. These models will undergo training and evaluation using the machine learning algorithms mentioned, enabling a bank to automate and optimize its credit assessment process.

By employing these models, a bank can enhance its risk management practices, reduce the chances of default, and ensure that credit is extended to individuals with a higher likelihood of repayment. This not only benefits the financial institution but also supports responsible lending practices and provides better financial opportunities to eligible applicants.

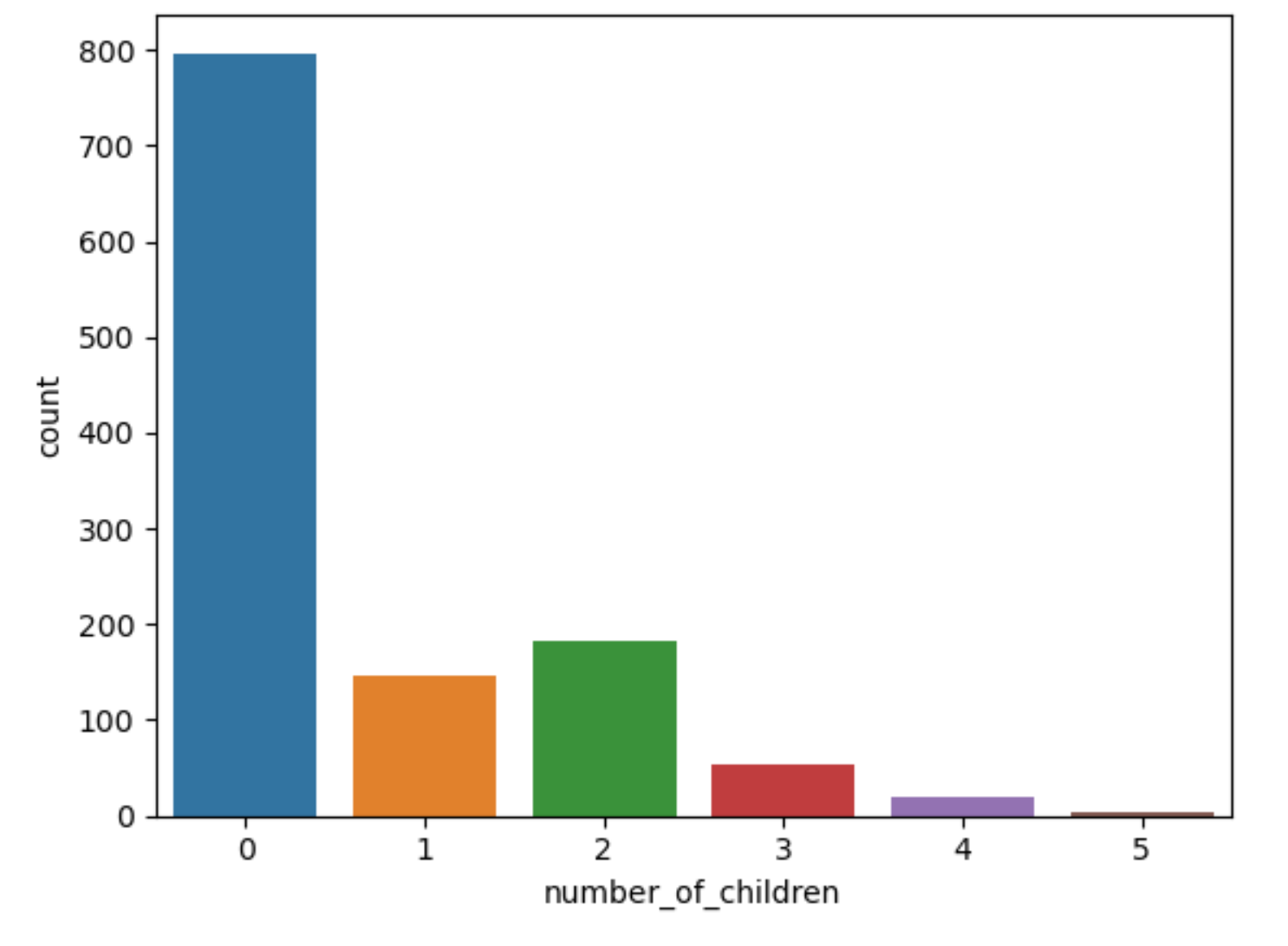
# Visualization

## Histogram of Age Column:

****

This Chart shows that most of the clients are of higher age. 200 or more clients are over 60 years of Age. Teenagers’ clients are less numerous as compared to old clients. This age disparity underscores the importance of recognizing and accommodating the diverse age groups within the client base.

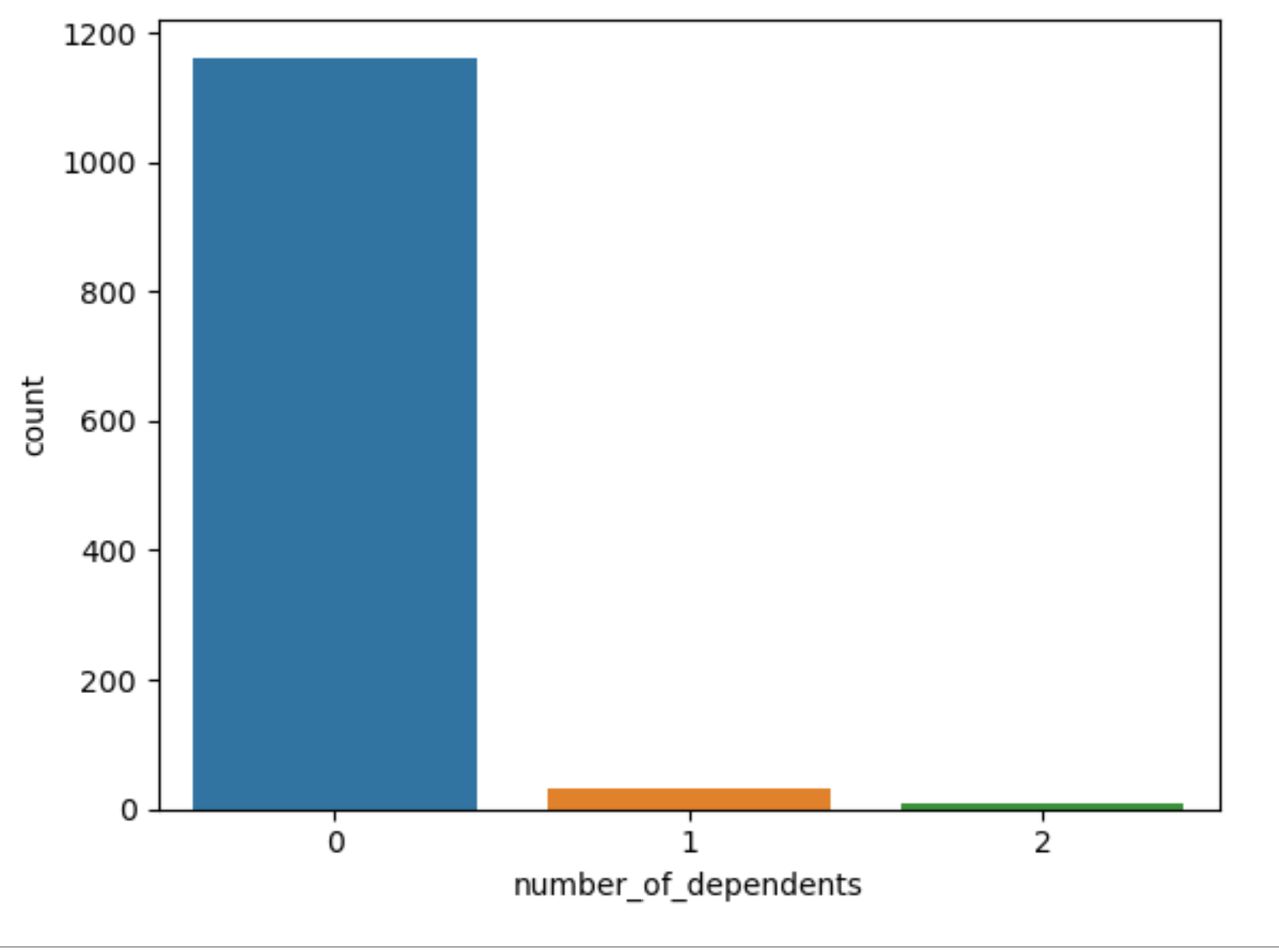
## Bar chart of the Number of Children:



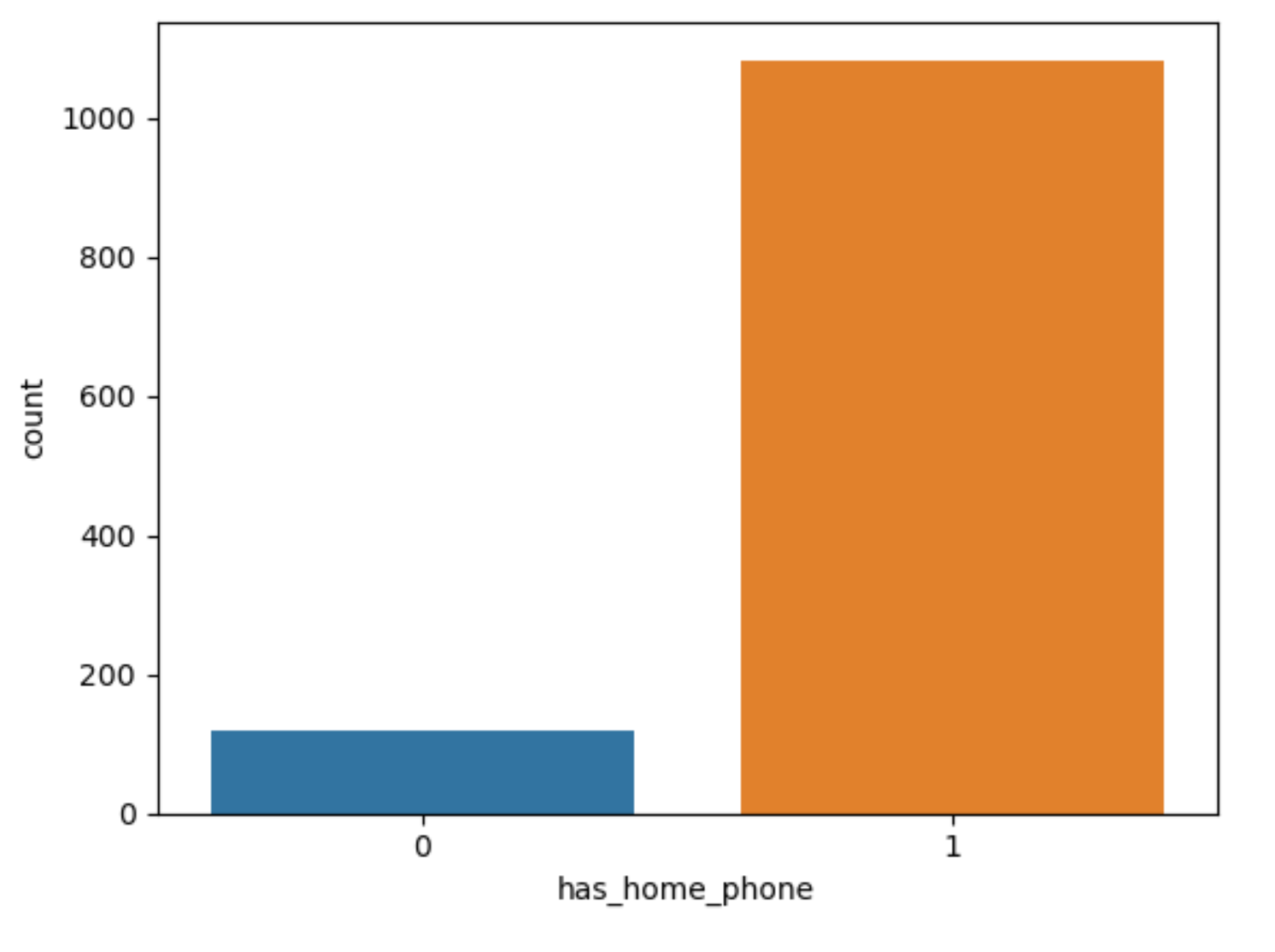
The bar chart offers a concise snapshot of clients and their number of children. Notably, 790 clients have no children, indicating a significant childless client segment. In contrast, 140 clients have one child, while 180 or more have two children, reflecting moderate-sized family demographics. Additionally, 50 clients are categorized as having three children. This chart provides valuable insights into the family composition within the client base.

## 

## Number of Dependents



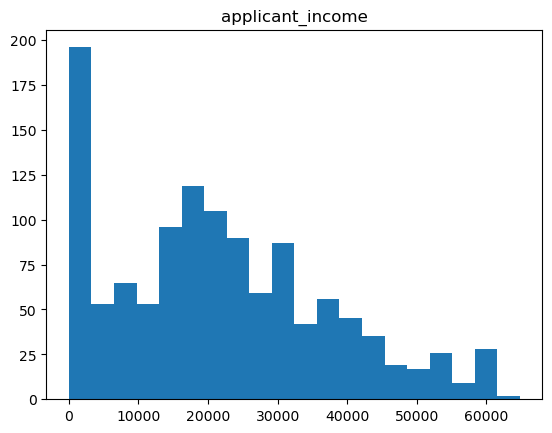
This chart shows how many people rely on others financially. Most of our clients, around 1,150 of them, don't depend on anyone else. Only a few clients, maybe just a handful, have one person they support financially. This means most of our clients are financially independent, and only a small number have someone they need to take care of. It's important to understand these differences to provide the right kind of financial services and support for everyone.

Presence of Mobile Phone  


This graph makes it clear that most of our clients have mobile phones at home. About 1,100 clients own mobile phones at their residences. On the other hand, only a small number of clients do not have mobile phones at home. This data emphasizes the widespread use of mobile phones among our clients, with only a few exceptions.

## 

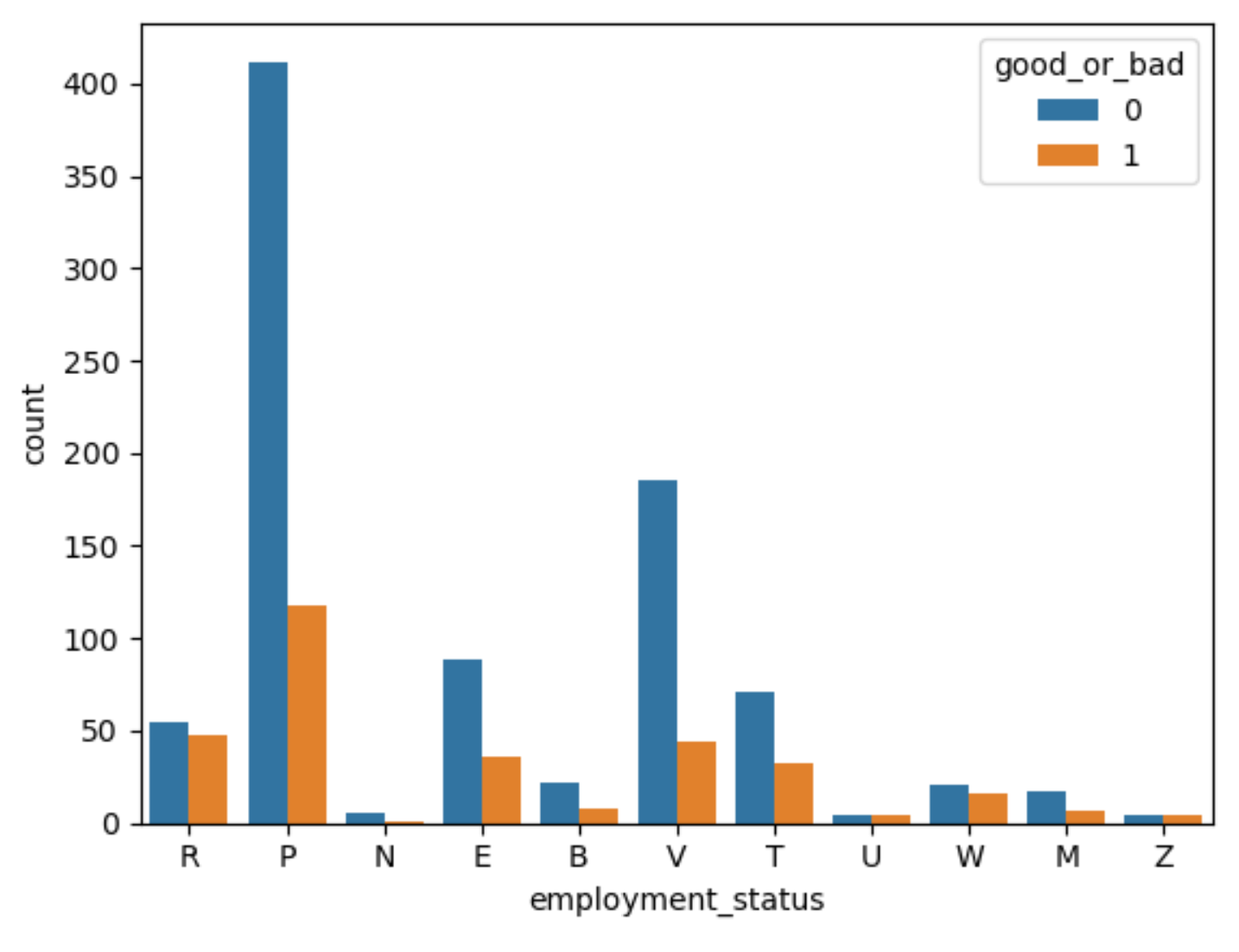
## Applicant Income



The distribution of client income is summarized in this data. Approximately 200 clients report having no income, while 125 clients indicate an income of $20,000. Moreover, 100 clients report earnings of $21,000, 80 clients have an income of $22,000, and 78 clients earn $30,000. Additionally, 55 clients have an income of $10,000. This income data provides insights into the diverse financial situations of our clients, ranging from no income to varying levels of earnings.

## 

## Employment Status



This bar graph helps us see how different types of jobs relate to whether a client's credit is good or bad.

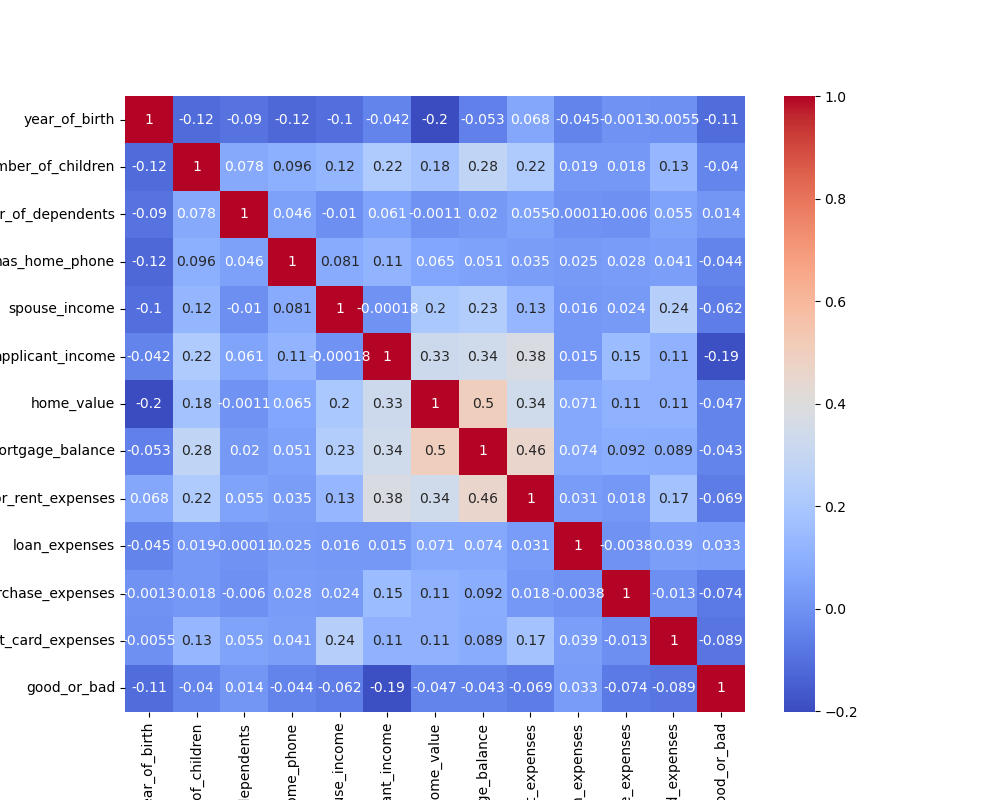
For clients working in the private sector (P), there are 400 clients with bad credit (good\_or\_bad = 0) and 120 with good credit (good\_or\_bad = 1). This shows that most private sector employees in this dataset have bad credit.

The "N" category, which represents other types of employment, has fewer clients, and they tend to have both good and bad credit in lower numbers.

For government employees (V), 180 of them have bad credit, while 50 have good credit. This suggests that government employees are more likely to have bad credit.

In summary, this graph helps us understand the connection between job types and credit quality. It shows that private sector employees generally have not better credit, and government employees also tend to have bad credit. This information can be useful for making decisions about loans and credit for clients in different job categories.

Correlation



These correlation values help us understand how different factors relate to each other in our dataset.

Firstly, we see a positive correlation of 0.5 between mortgage balance and home value. This means that as the value of homes goes up, the amount of money clients owe on their mortgages also tends to increase. It suggests that clients with more valuable homes often have larger mortgage balances.

Secondly, there's a positive correlation of 0.46 between mortgage or rent expenses and mortgage balance. This means that clients who spend more on their mortgage or rent tend to have higher mortgage balances. It's a logical connection, as larger expenses typically result in larger outstanding balances.

Lastly, there's a moderate positive correlation of 0.38 between mortgage or rent expenses and applicant income. This suggests that clients with higher incomes tend to spend more on their mortgages or rent. It's a reasonable finding since people with higher incomes may afford more expensive housing.

Additionally, we also notice a small negative correlation of 0.2 between the year of birth and home value. This indicates that as the year of birth (age) increases, home values may slightly decrease. Lastly, there's a small positive correlation of 0.3 between applicant income and home value, suggesting that clients with higher incomes may have homes with slightly higher values.

# Performance Metrics

## Accuracy:

Accuracy measures the overall correctness of a model's predictions, providing a straightforward assessment of its performance. It's essential to gauge how often the model makes correct predictions.

## Precision and Recall:

Precision evaluates the accuracy of positive predictions, ensuring that when the model predicts a positive outcome, it's correct. Recall, on the other hand, measures the model's ability to capture all relevant positive instances. These metrics are vital for understanding the model's performance in differentiating between classes, such as identifying clients with good or bad credit.

## F1-Score:

The F1-score combines precision and recall into a single metric, providing a balanced assessment of a model's overall accuracy, especially when dealing with imbalanced datasets. It helps strike a balance between minimizing false positives and false negatives.

These measures collectively help us evaluate and fine-tune machine learning models to ensure they meet specific objectives, such as correctly identifying credit risk in this context.

# AUC and ROC Curves

## AUC

The **AUC (Area Under the Curve)** is a number that tells us how good a model is at distinguishing between different things, like good and bad credit.

## ROC

The **ROC curve (Receiver Operating Characteristic)** is a graph that shows how well a model can tell the difference between good and bad things, depending on how picky or lenient we want it to be. We use these to see how well our model works and choose the best way to use it.

# ML Models

In addressing the problem at hand, three distinct machine learning algorithms have been employed: **Decision Tree, Logistic Regression, and Naive Bayes**. Each of these algorithms plays a unique role in the context of credit risk analysis:

## 

## Decision Tree:

Decision Tree is a versatile algorithm known for its interpretability and ability to handle both categorical and numerical data. In the context of credit risk analysis, a Decision Tree can be used to create a tree-like structure of decision rules based on client attributes such as age, income, and employment status. By recursively splitting the data into subsets, it can classify clients into different credit risk categories, providing a clear, understandable decision path. Decision Trees are useful for identifying key factors that contribute to credit risk, making them valuable for risk assessment.

## Logistic Regression:

Logistic Regression is a widely used algorithm for binary classification problems, such as determining whether a client's credit is good or bad. It models the relationship between the client's attributes (independent variables) and the probability of having good or bad credit (dependent variable). Logistic Regression provides a probabilistic interpretation, making it possible to estimate the likelihood of a client falling into a specific credit category based on their attributes. This algorithm is valuable for assessing the probability of creditworthiness and understanding the impact of each attribute on the outcome.

## Naive Bayes:

Naive Bayes is a probabilistic classification algorithm that is particularly useful for text classification and categorical data. In credit risk analysis, Naive Bayes can be employed to calculate the probability of a client having good or bad credit given their attribute values. It relies on the assumption of independence between attributes, which is why it's called "naive." Despite this simplification, Naive Bayes often performs well in practice, especially when dealing with categorical data like employment status or residential status. It provides a straightforward way to estimate probabilities and make credit risk predictions.

Each of these algorithms has its strengths and limitations, and their choice depends on the specific characteristics of the dataset and the objectives of the credit risk analysis. By applying a combination of Decision Tree, Logistic Regression, and Naive Bayes, a comprehensive approach to credit risk assessment can be achieved, with the potential to improve the accuracy and robustness of the predictive models.

# Decision Tree: Performance and Evaluation

In assessing the performance of the Decision Tree algorithm for credit risk analysis, we conducted cross-validation to evaluate its predictive accuracy. The cross-validation scores, which represent how well the model generalizes to unseen data, ranged from approximately 59.5% to 70.2%. On average, the Decision Tree achieved a cross-validation score of approximately 63.5%, indicating its ability to make reasonably accurate predictions.

Furthermore, the Decision Tree's classification report sheds light on its precision, recall, and F1-score. In this context, precision measures the accuracy of positive predictions, recall assesses the model's ability to capture all relevant instances, and the F1-score balances precision and recall. For clients with good credit (0), the model exhibited a precision and recall of 75%, while for clients with bad credit (1), these metrics were at 31%. This suggests that the Decision Tree performs better at identifying clients with good credit but is less effective at recognizing clients with bad credit.

Overall accuracy, which quantifies the model's overall correctness in its predictions, stood at approximately 62.9%. In summary, the Decision Tree demonstrates reasonable predictive capabilities for credit risk analysis, with some areas of improvement needed, particularly in correctly identifying clients with bad credit.

# Naïve Bayes: Performance and Evaluation

When evaluating the performance of the Naïve Bayes classification model for credit risk analysis, we found that it achieved an accuracy rate of 70%. For clients with good credit (class 0), the precision was 0.77, indicating that when the model predicted good credit, it was accurate around 77% of the time. However, for clients with bad credit (class 1), the precision was 0.43, showing that the accuracy of predictions for this group was lower.

In terms of recall, which measures the ability to identify all relevant instances, the model's performance was lower for class 1, with a recall rate of 0.33. This affected the F1-score, which combines precision and recall and stood at 0.37 for class 1.

During 10-fold cross-validation, the Naïve Bayes model exhibited an average accuracy of approximately 65.5%, with a standard deviation of about 10.9%. The area under the ROC curve (AUC), a measure of the model's ability to distinguish between classes, was 0.585.

In summary, the Naïve Bayes model demonstrates moderate performance in credit risk analysis, with higher accuracy in identifying clients with good credit and room for improvement in identifying clients with bad credit.

# Logistic Regression: Performance and Evaluation

In the assessment of the Logistic Regression model's performance for credit risk analysis, we found some noteworthy results. The model achieved an overall accuracy of around 74.52%, suggesting that it accurately predicts the credit outcomes for most clients.

When it comes to precision, which measures the accuracy of positive and negative predictions, the model displayed high precision (75%) for clients with good credit (class 0). However, for clients with bad credit (class 1), the precision was lower at 67%, indicating some misclassification of clients as having good credit when they don't.

Recall, which gauges the model's ability to capture all relevant instances, was relatively low at 10% for class 1 (clients with bad credit). This means the model struggled to correctly identify a substantial portion of clients with bad credit, which can be a concern.

The F1-score, which combines precision and recall, resulted in a value of 0.18 for class 1, highlighting the challenges in correctly identifying clients with bad credit.

During cross-validation, the Logistic Regression model demonstrated consistency with mean scores of approximately 73.01%. Overall, the model's performance suggests effective identification of clients with good credit but room for improvement in identifying clients with bad credit.

# 

# Comparison

In this section, we will compare the performance of the three machine learning models—Decision Tree, Naïve Bayes, and Logistic Regression—in the context of credit risk analysis. Each model has its strengths and weaknesses, and a comparative analysis will help us determine which model is best suited for the specific task.

We will consider key metrics such as **accuracy, precision, recall,** and **F1-score** for each model, focusing on their ability to correctly classify clients into good and bad credit categories. By evaluating and contrasting these metrics, we can make informed decisions about the most suitable model for credit risk assessment.

We've summarized the performance of three machine learning models: Naive Bayes, Decision Tree, and Logistic Regression. Here's a quick overview of their performance:

## Naive Bayes:

This model achieved an accuracy of approximately 70.36% with good precision and recall for class 0 (good credit). However, its performance for class 1 (bad credit) was relatively lower.

## Decision Tree:

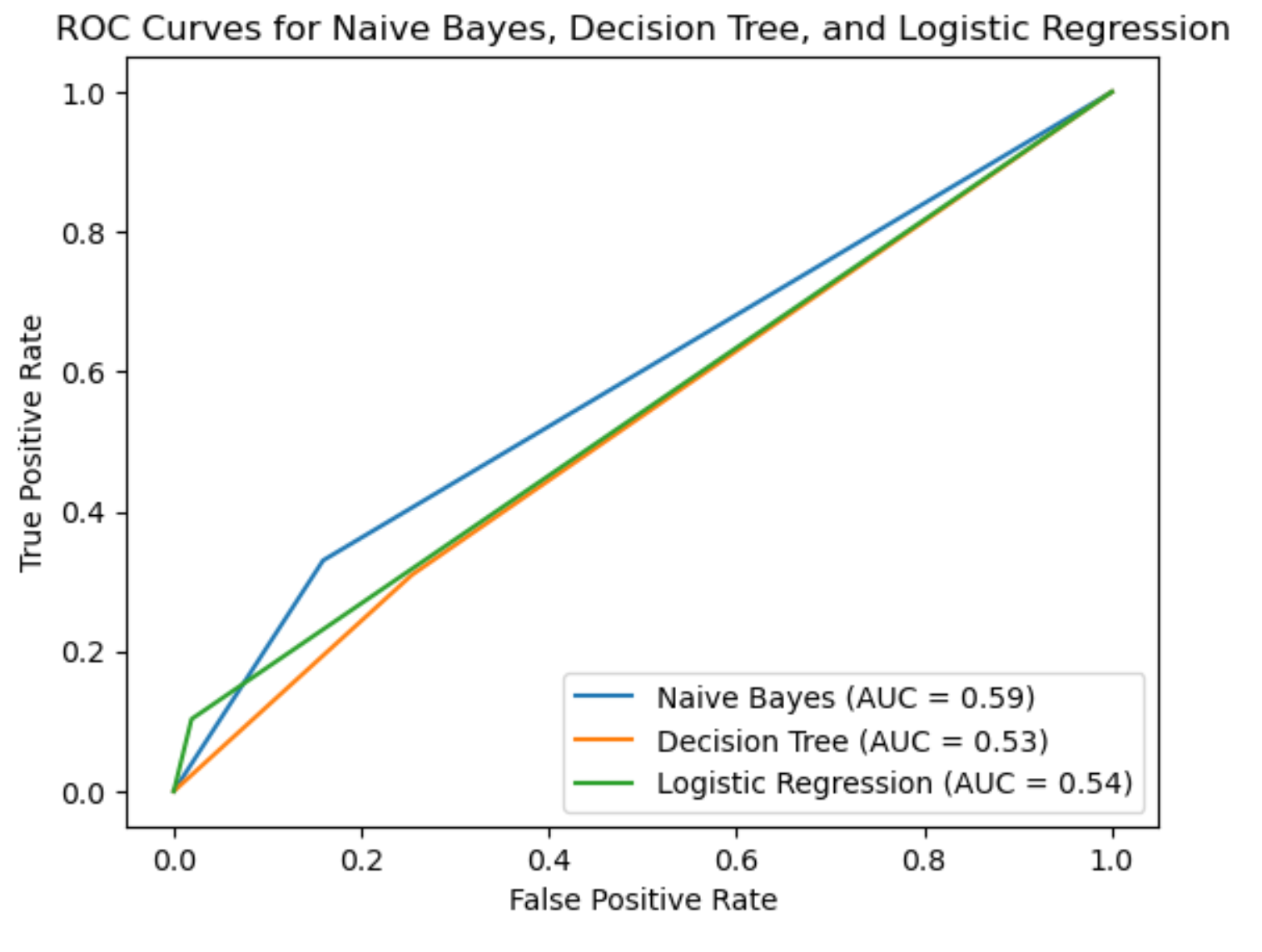
The Decision Tree model had an accuracy of around 62.88% with balanced precision and recall for both classes. It performed reasonably well in identifying both good and bad credit cases.

## Logistic Regression:

Logistic Regression emerged as the top performer with an accuracy of about 74.52%. It demonstrated high precision for class 0 (good credit) and excellent recall for both classes, making it a promising choice for this credit risk analysis.

Based on these results, **Logistic Regression** appears to be the most effective model for the task of credit risk assessment, as it achieves a good balance between precision and recall for both credit categories.

# ROC Curve

****

## Naive Bayes

(AUC = 0.59): Naive Bayes demonstrates moderate discriminative power, with an AUC of 0.59. This means that, on average, the model is better than random chance at distinguishing between clients with good and bad credit. However, it's important to consider other metrics to get a complete picture of its performance.

## Decision Tree

(AUC = 0.53): The Decision Tree model has a slightly lower AUC of 0.53. This indicates that its ability to separate clients with good and bad credit is not as strong as Naive Bayes’. However, it's crucial to remember that AUC is just one part of the performance evaluation.

## Logistic Regression

(AUC = 0.54): Logistic Regression falls in a similar range with an AUC of 0.54, indicating reasonable discrimination power. Like Naive Bayes, it's better than random guessing but may not be highly effective in distinguishing between the two credit classes.

# Best Model

When selecting the best model, it's essential to consider a holistic view of all relevant metrics, including accuracy, precision, recall, and F-measure, as mentioned. While Naive Bayes had the highest AUC, other factors like precision and recall also played a crucial role. In my case, I found that **Logistic Regression** performed best when considering all these metrics. This suggests that it strikes a good balance between correctly identifying good and bad credit clients, making it a more suitable choice for credit risk analysis.