BA 723 – Business Analytics Capstone

Data-Driven Optimization of Loan Approval Process

Submitted by:

Miguel Polanco Sanchez

Submitted to:

David Parent

Due Date:

August 15, 2024

**Table of Contents**

[**Executive Summary** 4](#_Toc174547915)

[**1.** **Introduction** 5](#_Toc174547916)

[**1.1.** **Background** 5](#_Toc174547917)

[**1.2.** **Problem Statement** 6](#_Toc174547918)

[**1.3.** **Objectives and Measurement** 7](#_Toc174547919)

[**1.3.1.** **Business Objective** 7](#_Toc174547920)

[**1.3.2.** **Analytics Objective** 7](#_Toc174547921)

[**1.3.3.** **Measurement** 8](#_Toc174547922)

[**1.4.** **Assumptions and Limitation** 8](#_Toc174547923)

[**2.** **Data Sources** 9](#_Toc174547924)

[**2.1.** **Data Set Introduction** 9](#_Toc174547925)

[**2.2.** **Exclusions** 10](#_Toc174547926)

[**2.3.** **Initial Data Cleansing or Preparation** 10](#_Toc174547927)

[**2.3.1.** **Review of the dataset and Data Dictionary** 10](#_Toc174547928)

[**2.3.2.** **Removal of Exclusions on the data** 12](#_Toc174547929)

[**3.** **Data Exploration** 12](#_Toc174547930)

[**3.1.** **Data Structure** 13](#_Toc174547931)

[**3.1.1.** **Conversion of variables to Category** 13](#_Toc174547932)

[**3.2.** **Data Visualization** 13](#_Toc174547933)

[**3.2.1.** **Visualization of data before cleaning** 14](#_Toc174547934)

[**3.2.2.** **Visualization of Missing Values** 15](#_Toc174547935)

[**3.3.** **Data Cleansing** 16](#_Toc174547936)

[**3.3.1.** **Missing Values Cleaning** 16](#_Toc174547937)

[**3.3.2.** **Management of outliers** 17](#_Toc174547938)

[**3.3.3.** **Visualization of data after cleaning** 18](#_Toc174547939)

[**3.3.4.** **Correlation** 21](#_Toc174547940)

[**3.3.5.** **Distribution of the target variable** 23](#_Toc174547941)

[**4.** **Data Preparation** 25](#_Toc174547942)

[**4.1.** **Skewness** 25](#_Toc174547943)

[**4.2.** **Transformation** 25](#_Toc174547944)

[**5.** **Feature Selection** 27](#_Toc174547945)

[**5.1.** **Preparing data for modeling** 27](#_Toc174547946)

[**5.1.1.** **One-hot encoding** 27](#_Toc174547947)

[**6.** **Model Exploration** 29](#_Toc174547948)

[**6.1.** **Data partition and target data frames creation** 29](#_Toc174547949)

[**6.2.** **Modeling** 29](#_Toc174547950)

[**6.2.1.** **Decision Tree** 30](#_Toc174547951)

[**6.2.1.1.** **Model Parameters** 31](#_Toc174547952)

[**6.2.1.2.** **Model Performance** 31](#_Toc174547953)

[**6.2.1.3.** **Visualization of the tree** 32](#_Toc174547954)

[**6.2.1.4.** **Feature Importance** 32](#_Toc174547955)

[**6.2.2.** **Full Misclassification Tree** 33](#_Toc174547956)

[**6.2.2.1.** **Model Parameters** 33](#_Toc174547957)

[**6.2.2.2.** **Model Performance** 33](#_Toc174547958)

[**6.2.2.3.** **Visualization of the tree** 34](#_Toc174547959)

[**6.2.2.4.** **Feature Importance** 35](#_Toc174547960)

[**6.2.3.** **GridSearch Decision Tree** 35](#_Toc174547961)

[**6.2.3.1.** **Model Parameters** 36](#_Toc174547962)

[**6.2.3.2.** **Model Performance** 36](#_Toc174547963)

[**6.2.3.3.** **Feature Importance** 37](#_Toc174547964)

[**6.2.4.** **Random Forest** 38](#_Toc174547965)

[**6.2.4.1.** **Model Parameters** 38](#_Toc174547966)

[**6.2.4.2.** **Model Performance** 38](#_Toc174547967)

[**6.2.4.3.** **Feature Importance** 39](#_Toc174547968)

[**6.2.5.** **Random Forest with only best features** 40](#_Toc174547969)

[**6.2.5.1.** **Model Parameters** 41](#_Toc174547970)

[**6.2.5.2.** **Model Performance** 41](#_Toc174547971)

[**6.2.5.3.** **Feature Importance** 42](#_Toc174547972)

[**6.2.6.** **Gradient Boost** 43](#_Toc174547973)

[**6.2.6.1.** **Model Parameters** 43](#_Toc174547974)

[**6.2.6.2.** **Model Performance** 44](#_Toc174547975)

[**6.2.6.3.** **Feature Importance** 44](#_Toc174547976)

[**6.2.7.** **Full Logistic Regression** 45](#_Toc174547977)

[**6.2.7.1.** **Model Parameters** 46](#_Toc174547978)

[**6.2.7.2.** **Model Performance** 46](#_Toc174547979)

[**6.2.7.3.** **Feature Importance** 47](#_Toc174547980)

[**6.2.8.** **Forward Logistic Regression** 48](#_Toc174547981)

[**6.2.8.1.** **Model Parameters** 48](#_Toc174547982)

[**6.2.8.2.** **Model Performance** 48](#_Toc174547983)

[**6.2.8.3.** **Feature Importance** 49](#_Toc174547984)

[**6.2.9.** **Backward Logistic Regression** 50](#_Toc174547985)

[**6.2.9.1.** **Model Parameters** 50](#_Toc174547986)

[**6.2.9.2.** **Model Performance** 50](#_Toc174547987)

[**6.2.9.3.** **Feature Importance** 51](#_Toc174547988)

[**6.2.10.** **Stepwise Logistic Regression** 52](#_Toc174547989)

[**6.2.10.1.** **Model Parameters** 52](#_Toc174547990)

[**6.2.10.2.** **Model Performance** 52](#_Toc174547991)

[**6.2.10.3.** **Feature Importance** 53](#_Toc174547992)

[**7.** **Model Recommendation** 54](#_Toc174547993)

[**7.1.** **Model Scores** 54](#_Toc174547994)

[**7.2.** **Model Selection** 54](#_Toc174547995)

[**8.** **Conclusion and Recommendations** 57](#_Toc174547996)

[**8.1.** **Impacts on Business Problem** 57](#_Toc174547997)

[**8.2.** **Recommended Next Steps** 58](#_Toc174547998)

[**9.** **References** 60](#_Toc174547999)

[**9.1.** **References** 60](#_Toc174548000)

# 

# **Executive Summary**

A recent initiative by a financial institution has recognized the challenges posed by the traditional loan approval processes, which often fail to capture the full spectrum of applicant profiles, leading to inefficiencies and potential risks. This project aims to address these challenges by developing a predictive model that leverages various features, including applicant demographics, employment details, loan specifics, and credit history. The primary objective is to enhance the accuracy and efficiency of loan approval decisions, thereby reducing the likelihood of bad loans and optimizing customer offerings.

This exploration was conducted through the application of multiple machine learning models, including Decision Tree, Logistic Regression, Gradient Boost and Random Forest. The Random Forest model emerged as the top performer, achieving an ROC-AUC score of 82% and an accuracy rate of 77%. The analysis highlighted two critical features, credit history and income levels, that significantly influence loan approval outcomes.

The findings suggest that incorporating these key variables into the loan approval process can substantially improve decision-making, benefiting both the institution by mitigating financial risks and the customers by offering more tailored loan products. This approach not only enhances operational efficiency but also fosters greater financial inclusion, aligning with the institution’s strategic goals.

# **Introduction**

## **Background**

The financial services sector, particularly banking, plays a crucial role in economic development by providing capital to individuals and businesses. One of the key functions of banks is to lend money to qualified applicants, which involves assessing the risk of lending.

Traditionally, the loan approval process involves several steps. Applicants submit their personal, employment, and financial information. Loan officers then manually verify the authenticity of the documents provided, such as income statements, employment records, and credit reports. Credit scores are assessed based on past financial behavior, often using standard scoring models. Loan officers evaluate the risk of default based on the applicant's financial stability, credit history, and other factors. Finally, a decision is made whether to approve or reject the loan application, involving a combination of automated rules and human judgment. (American Bankers Association, 2021)

This traditional approach is time-consuming, as the manual verification and assessment process can lead to long waiting times for applicants. It is also susceptible to human bias and error, resulting in inconsistent decisions and potential biases based on the loan officer's personal experiences.

As the number of loan applications increases, scaling the manual process becomes difficult without a corresponding increase in manpower. With the advent of big data and machine learning technologies, there is a significant opportunity to automate and enhance the loan approval process.

## **Problem Statement**

In the contemporary financial landscape, efficient and accurate loan approval processes are critical for both banking institutions and their customers. However, traditional loan approval methods, which rely heavily on manual document verification and subjective judgment, are plagued with inefficiencies and risks. These methods are time-consuming, prone to human error, and susceptible to biases, leading to inconsistent and potentially unfair decisions. Furthermore, the manual nature of the process incurs significant operational costs and is difficult to scale when applications start to increase. (Koch & MacDonald, 2014)

One of the major issues with the traditional approach is the prolonged turnaround time for loan approvals. Applicants often face delays as loan officers manually review and verify each piece of information submitted. This slow process not only frustrates customers but also impacts on the bank's ability to serve a larger client base efficiently. In a competitive market, banks that cannot provide quick and efficient service risk losing customers to more agile competitors. (American Bankers Association, 2021)

Operational costs are another significant concern. The manual processing of loan applications requires substantial human resources and administrative overhead. From verifying documents to assessing risk, the traditional approach involves multiple stages of manual intervention, each contributing to the overall cost. These costs are further exacerbated when the volume of loan applications increases, as banks need to hire more staff to manage the workload. This lack of scalability makes it challenging for banks to expand their loan portfolios without incurring additional expenses. (Kagan, 2024)

## **Objectives and Measurement**

### **Business Objective**

The primary objective of this project is to develop a machine learning model that can predict loan approvals. This model will analyze various features related to the applicant's personal information, financial status, and loan details to determine the likelihood of loan approval. By implementing this solution, the banking institution can streamline its loan approval process, improve the accuracy of credit assessments, and enhance overall customer satisfaction.

### **Analytics Objective**

The objectives for this project are the following:

1. Get insight of the data presented from the clients to reveal the key variables that are key to predicting the approval of loans.
2. Create different predictive models and identify the best model that will accurately predict the approval of loans for the bank.
3. Enhance the performance of the machine learning model to ensure the final model delivers the most effective and efficient results for the bank, while also reducing the cost of the approval process.

The project will seek to test the following hypothesis: *How the creation of a loan application model will help banks to reduce the cost and time spent on the approval process of a loan application?*

Key questions:

1. What factors most significantly influence loan approval decisions?
2. How will the implementation of a predictive loan approval model impact the bank’s revenue?
3. What cost savings can be achieved through automation of the loan approval process?

### **Measurement**

The best model will be selected based on three metrics, ROC-AUC score, Average Squared Error (ASE), and Accuracy. Given the small size of the dataset, the ROC-AUC score will be the primary metric, as it provides a detailed view of model performance even with limited data. ASE will be used as the secondary metric to evaluate the average prediction error, while Accuracy will be the final metric to assess the overall correctness of the model, regardless of class distribution.

## **Assumptions and Limitation**

This project assumes that the dataset is accurate representative, and free from significant errors or biases. The features included in the dataset (demographic, employment, loan details, and credit history) are assumed to have predictive power for loan approval. Each loan application is assumed to be independent of others, with no hidden dependencies affecting the predictions. The patterns and relationships in the historical data are assumed to be stable and applicable to future loan applications.

# **Data Sources**

## **Data Set Introduction**

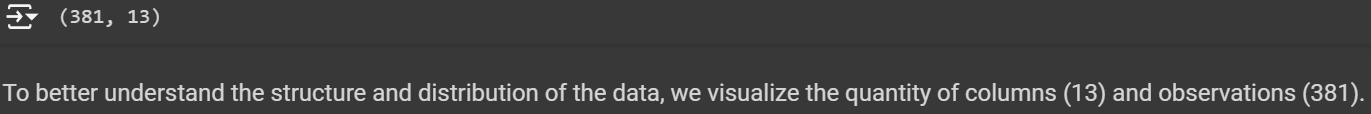
The dataset used in this project is prepared by Willian Oliveira and it is published in Kaggle. The dataset has a main table called loan\_data new.csv, this dataset includes comprehensive information on loan applicants, which is used to predict whether a loan application will be approved or rejected. The dataset consists of 381 observations and 13 variables. Each observation represents one loan application in the dataset. The target variable is in this dataset and is called Loan Status which is categorized by Y for approval and N for rejection.

## **Exclusions**

For this project, no records were excluded and only the Loan\_ID variables will be excluded as it is not significant for the objective of the project.

## **Initial Data Cleansing or Preparation**

The dataset was uploaded to Python which recorded the following dimensions, 381 observations and 13 variables.



### **Review of the dataset and Data Dictionary**

This table shows the name of the column, the data type and description of the column. To ensure the dataset is correctly structured, it is needed to visualize the data types of each column. This will help identify any columns that might have incorrect data types and need a conversion.

|  |  |  |
| --- | --- | --- |
| Variables | Data Type | Description |
| Loan\_ID | Object | Unique Identifier for each loan application. |
| Gender | Object | Gender of solicitor. |
| Married | Object | Applicant marital status. |
| Dependents | Float | Applicant number of dependents. |
| Education | Object | Applicant education level. |
| Self\_Employed | Object | Indicates if applicant is self-employed. |
| ApplicantIncome | Integer | The income of the applicant. |
| CoapplicantIncome | Object | The income of the Co-applicant. |
| LoanAmount | Float | The amount of the loan requested by applicant. |
| Loan\_Amount\_Term | Float | The term of the loan in days. |
| Credit\_History | Float | The credit history of the applicant. (1= Have Credit History, 0= Don’t have Credit History) |
| Property\_Area | Object | The area where the property is located. (Urban, Semiurban, Rural) |
| Loan\_Status | Object | The outcome of the loan application. (Y = Approved, N = Rejected) |

### **Removal of Exclusions on the data**

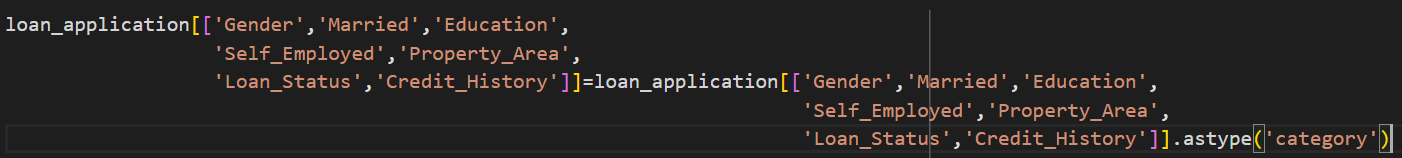
As it is stated in point 2.2 Exclusions, the variable Loan\_ID will be removed from the modeling. The number of variables is now down to 12.

# **Data Exploration**

## **Data Structure**

### **Conversion of variables to Category**

The conversion of the variables GENDER, MARRIED, EDUCATION, SELF\_EMPLOYED, PROPERTY\_AREA, LOAN\_STATUS and CREDIT\_HISTORY to category is necessary because these variables represent discrete categories rather than continuous numerical values, which aligns with their inherent nature. Additionally, categorical encoding is essential for many machine learning algorithms that require or perform better with properly categorized data, ensuring more accurate and meaningful model training and predictions.



|  |  |
| --- | --- |
| Variables | Data Type |
| Loan\_ID | Object |
| Gender | Category |
| Married | Category |
| Dependents | Float |
| Education | Category |
| Self\_Employed | Category |
| ApplicantIncome | Integer |
| CoapplicantIncome | Object |
| LoanAmount | Float |
| Loan\_Amount\_Term | Float |
| Credit\_History | Category |
| Property\_Area | Category |
| Loan\_Status | Category |

## **Data Visualization**

### **Visualization of data before cleaning**

The visualizations provide a comprehensive overview of all variables, allowing to identify those requiring cleaning or adjustment.

A screenshot of a graph

Description automatically generatedA group of graphs showing different sizes and shapes

Description automatically generated with medium confidence

By looking at the variables visually it is easier to identify the ones that need to be cleaned:

GENDER: Requires correction for any missing or incorrect values.

SELF\_EMPLOYED: Needs adjustment for missing or incorrect entries.

COAPPLICANTINCOME: Should be examined for potential outliers and adjusted accordingly.

### **Visualization of Missing Values**

There are a total of 2 numerical columns with missing values and 9 records with missing values which represent less than 2.5% of the total records.

A screenshot of a computer

Description automatically generated

Also, there are some categorical variables that contain missing values that are represented by zeros. Specifically, the variables GENDER has 5 missing values, while SELF\_EMPLOYED has 21 missing values. Adding this to the before mentioned missing values, the dataset has a total of 35 missing values which represents a 9.2% of the total records.

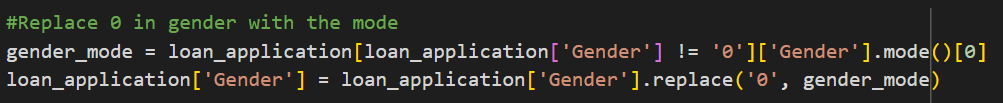
A screenshot of a computer screen

Description automatically generated

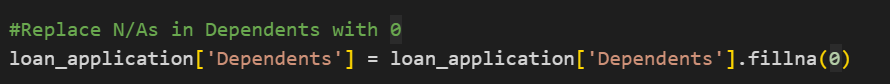
## **Data Cleansing**

### **Missing Values Cleaning**

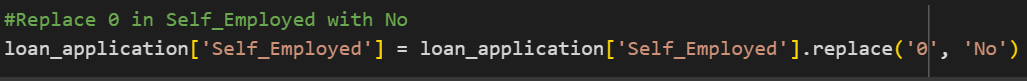
For the variable GENDER the missing values are replaced with the mode as it’s the representation of the most frequently value used so the dataset remains as representative as possible of the original population



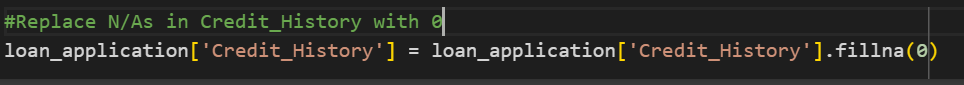
For DEPENDANTS the missing values are replaced with zeros based on assumptions that no answer to this question indicates the absence of dependents.



For the variable SELF\_EMPLOYED the missing values are replaced with No as it is considering that by having zeros in this variable it represents that these individuals are not self-employed.

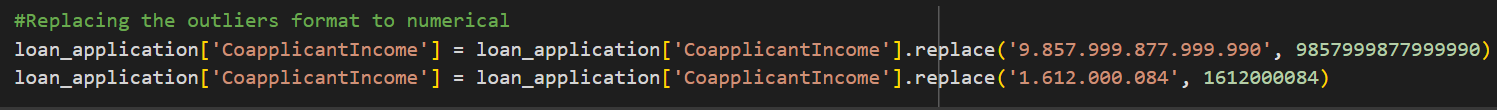


For the variable CREDIT\_HISTORY missing values are replaced with a designation indicating no credit history.

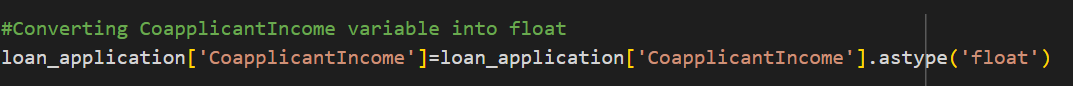


### **Management of outliers**

For the potential outliers on COAPPLICANTINCOME variable it was discovered that these values were computed with the wrong format, so they were converted into numerical format to address correctly these outliers.

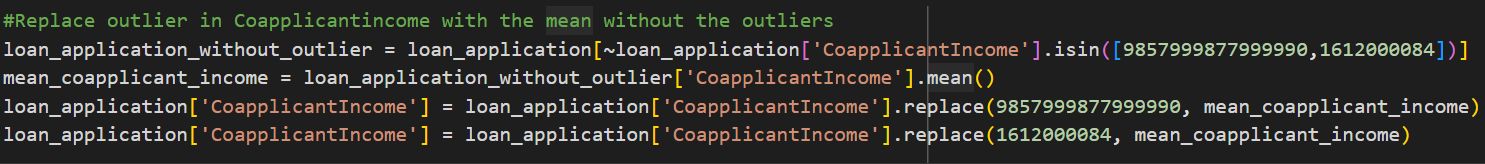


After this conversion, the variables which were an object because of these format errors are now converted into a float so it can be recognized as a continuous numerical data variable.



|  |  |
| --- | --- |
| Variables | Data Type |
| Loan\_ID | Object |
| Gender | Category |
| Married | Category |
| Dependents | Float |
| Education | Category |
| Self\_Employed | Category |
| ApplicantIncome | Integer |
| CoapplicantIncome | Float |
| LoanAmount | Float |
| Loan\_Amount\_Term | Float |
| Credit\_History | Category |
| Property\_Area | Category |
| Loan\_Status | Category |

Now that the variable is in the correct format it is possible to manage the outliers. To manage the outliers, first a new data frame is created by excluding the outliers from the variable. Then another data frame is created to calculate the mean of the variable without the outliers. Finally, the outlier’s values are replaced with the calculated mean without the outliers.



### **Visualization of data after cleaning**

With the clearing of missing values from categorical and numerical variables it is possible to visualize a data set without any missing values.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Also, after the cleaning of the remaining affected variables, it is now possible to visualize the effects of these changes in the graphics of each variable. The most notable change in the graphics is in the variable COAPPLICANTINCOME where the removal of the outliers has resulted in a more clearly defined chart curve for the variable.

A group of blue squares

Description automatically generated A group of graphs showing different sizes and numbers

Description automatically generated with medium confidence

For further identification of outliers, the boxplot aids in the identification of outliers. As can be seen in the boxplot, the variable APPLICANTINCOME contains some outliers, while the variable COAPPLICANTINCOME still has some outliers after the cleaning. By analyzing the context of these variables, it can be concluded that although outliers are present, their values are not excessively anomalous when compared to realistic expectations.

A graph with numbers and a green and orange bar chart

Description automatically generated

### **Correlation**

The correlation coefficient is a statistical measure of the strength of a linear relationship between two variables. So, variables that are highly correlated with other variables will create unstable models and can lead to inaccurate predictions. Therefore, it is important to verify the correlation coefficient between the numerical variables, anything over 0.6 will be considered a strong correlation.

For the numerical variables in the dataset a correlation matrix was run, and no strong correlations were found between variables. As shown in the correlation matrix the highest correlation observed is considered to be a weak correlation. Therefore, no further changes are required.

A red and blue squares with white text

Description automatically generated

For the categorical variables in the dataset a correlation matrix using dython library in python was run, and no strong correlations were found between variables. As shown in the correlation matrix the highest correlation observed is considered to be a moderate correlation. Therefore, no further changes are required.

A graph of a number of people

Description automatically generated with medium confidence

### **Distribution of the target variable**

This dataset features a binary target variable, where 1 indicates the approval of a loan, and 0 indicates the rejection of a loan. The distribution reveals a somewhat imbalanced target variable, with 71.1% of clients getting approved compared to 28.9% who got rejected. Such little imbalances are typical in this kind of dataset.

A graph of a graph with a bar and a number of blue squares

Description automatically generated with medium confidence

# **Data Preparation**

## **Skewness**

Having a skewed distribution poses risks when training a machine learning model. To address skewed data, one effective method is to apply a log base e transformation. Upon examining the skewness of the final dataset, it is evident that 2 variables have a skewness greater than 2, indicating the need for transformation.

A screenshot of a computer

Description automatically generated

## **Transformation**

After applying transformation to the COAPPLICANTINCOME variable, a new variable is created with the log values of the original variable. The new variable now called COAPPLICANTINCOME\_LOG shows a skewness of -0.127637 which is below the targeted levels.

For the variable LOAN\_AMOUNT\_TERM different methods of transformation were tested but the skewness of the variable was unable to be reduced. The variable will be tested in the modeling stage to see if it affects the model sufficiently to be removed.

A screen shot of a computer program

Description automatically generatedA screenshot of a computer

Description automatically generated

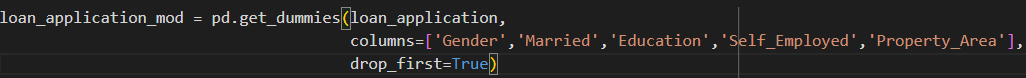
# **Feature Selection**

## **Preparing data for modeling**

Before starting the modeling, it is important to follow some steps that will help for smooth training of the machine learning models.

### **One-hot encoding**

As regressions are going to be part of the modeling process, it is important to do a one-hot encoding to ensure that the machine learning algorithm can handle the non-numerical categorical variables. In this case the variables GENDER, MARRIED, EDUCATION, SELF\_EMPLOYED and PROPERTY\_AREA were affected. As the drop\_first parameter was included in the code the variables selected were substituted with the name of the variable underscore the first result.



As a result, these are the variables to be used for the modeling process.

|  |  |
| --- | --- |
| Variables | Data Type |
| Loan\_Status | Binary (Target) |
| Gender\_Male | Bool |
| Married\_Yes | Bool |
| Dependents | Float |
| Education\_Not Graduate | Bool |
| Self\_Employed\_Yes | Bool |
| ApplicantIncome | Integer |
| CoapplicantIncome | Float |
| LoanAmount | Float |
| Loan\_Amount\_Term | Float |
| Credit\_History | Category |
| Property\_Area\_Semiurban | Bool |
| Property\_Area\_Urban | Bool |
| CoapplicantIncome\_log | Float |

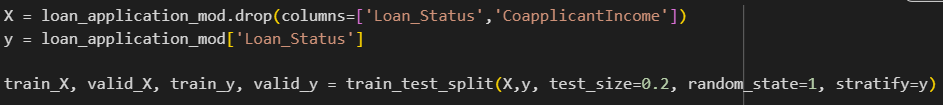
# **Model Exploration**

## **Data partition and target data frames creation**

Data partitioning involves splitting the dataset into training and validation sets. The training set is utilized to train the model and identify patterns and relationships within the data. In contrast, the validation or holdout set is employed to fine-tune the model by adjusting hyperparameters and selecting the optimal model. The purpose of the validation or holdout set is to mimic real-world scenarios where the model is tested on new, unseen data.

The data will be split by 80:20 and a stratified sampling (stratify=y) will be used to ensure that each split maintains the same class ratio. Since the target variable is a little bit imbalanced, this approach helps guarantee that the models are trained and evaluated on data that reflects the original dataset's distribution.

Also, creating separate data frames for features and the target variable enables the model to differentiate between the input (independent) variables and the target (dependent) variable. These distinct data structures are used to train the models to predict the target variable based on the input features. For the X data frame, the variable COAPPLICANTINCOME is removed to only leave the transformation available for the modeling.



## **Modeling**

Multiple models are used in this project such as decisions tree, linear regression, logistic regression, random forest and gradient boost. The hyperparameters were tuned for the best models to compare which model is the best in predicting approval of loans.

The following metrics were used to evaluate the models:

1. The ROC-AUC score is the primary metric used to evaluate the model because it is less influenced by imbalanced class distributions and assesses how effectively the model distinguishes between approvals and rejections. The score ranges from 0 to 1, with 0.5 indicating random guessing and 1 indicating perfect performance.

* 0.5: Random performance. The model's predictions are no better than random guessing.
* 0.5 - 0.6: Weak performance. The model has limited ability to discriminate between classes.
* 0.6 - 0.7: Fair performance. The model is better than random guessing but could be improved.
* 0.7 - 0.8: Good performance. The model has a reasonably strong ability to differentiate between classes.
* 0.8 - 0.9: Very good performance. The model effectively discriminates between classes.
* 0.9 - 1.0: Excellent performance. The model has an almost perfect ability to differentiate between classes.

2. Accuracy indicates how correct the model predictions are regardless of the class.

### **Decision Tree**

A decision tree is a widely used machine learning algorithm for classification problems due to its simplicity and interpretability. It visualizes predictions using a tree-like structure where each internal node represents a feature, each branch represents a decision rule, and each leaf node (terminal node) indicates the class label outcome. Besides offering results that are highly interpretable even to non-technical audiences, decision trees have several advantages: they can automatically ignore irrelevant features, are robust to outliers, and can capture non-linear relationships in the data.

#### **Model Parameters**

The parameters for this model were the default settings for this one.

A black and white sign with white text

Description automatically generated

#### **Model Performance**

|  |  |  |
| --- | --- | --- |
| Model | ROC-AUC Score | Accuracy |
| Decision Tree ASE | 0.67 | 0.69 |

Evaluating the performance of the model, the ROC-AUC score is only 0.67 which indicates fair performance. The model has a better-than-random ability to distinguish between the classes but still leaves room for improvement. The accuracy score shows that the model correctly predicts 69% of the instances.

A graph of a line with blue and orange lines

Description automatically generated

#### **Visualization of the tree**

This tree accounts for 68 splits and 69 terminal nodes. The starting point for the splits is the variable CREDIT\_HISTORY followed by the variables APPLICANTINCOME and COAPPLICANTINCOME.

A diagram of a flowchart

Description automatically generated

#### **Feature Importance**

By ranking the important features of this tree, it is possible to visualize that the most important variables are APPLICANTINCOME, CREDIT\_HISTORY and LOANAMOUNT. The importance level can be visualized below.

A screenshot of a computer

Description automatically generated

### **Full Misclassification Tree**

Similar to the decision tree, which by default takes into account the averaged squared error, this tree will take into account the aspect of a full misclassification tree.

#### **Model Parameters**

To define a full misclassification tree, it is necessary to define the criterion='entropy’ as a parameter.



#### **Model Performance**

|  |  |  |
| --- | --- | --- |
| Model | ROC-AUC Score | Accuracy |
| Full Misclassification Tree | 0.61 | 0.64 |

Evaluating the performance of the model, the ROC-AUC score is only 0.61 which indicates fair performance. The model has a better-than-random ability to distinguish between the classes but still leaves room for improvement. The accuracy score shows that the model correctly predicts 64% of the instances.

A graph of a positive rate

Description automatically generated with medium confidence

#### **Visualization of the tree**

Similarly to the ASE tree, this tree accounts for 69 splits and 70 terminal nodes. The starting point for the splits is the variable CREDIT\_HISTORY followed by the variables APPLICANTINCOME and COAPPLICANTINCOME.

A diagram of a person

Description automatically generated

#### **Feature Importance**

By ranking the important features of this tree, it is possible to visualize that the most important variables are APPLICANTINCOME followed by LOANAMOUNT and CREDIT\_HISTORY as well as COAPPLICANTINCOME\_LOG. The importance level can be visualized below.

A screenshot of a computer

Description automatically generated

### **GridSearch Decision Tree**

For the decision tree a grid search is conducted to optimize the model.

A screenshot of a computer program

Description automatically generated

#### **Model Parameters**

The GridSearch results show that the max\_depth of 5, min\_impurity\_decrease of 0.05 and a min\_samples\_split of 2 are needed to train the model.

**A screenshot of a computer

Description automatically generated**

#### **Model Performance**

|  |  |  |
| --- | --- | --- |
| Model | ROC-AUC Score | Accuracy |
| GridSearch Decision Tree | 0.71 | 0.74 |

Evaluating the performance of the model, the ROC-AUC score is 0.71 which indicates a good performance. The model has a reasonably strong ability to discriminate between the classes. The accuracy score shows that the model correctly predicts 74% of the instances. Overall, this model is better than the original model.

A line graph with blue and orange lines

Description automatically generated

#### **Feature Importance**

Unfortunately, this model is only considering the variable CREDIT\_HISTORY making very difficult the interpretation of the model.

A screen shot of a computer

Description automatically generated

### **Random Forest**

Random Forest is an ensemble learning method in machine learning designed to enhance prediction accuracy by aggregating the results of multiple individual decision trees, each trained on different random subsets of the data and features. This approach is robust against overfitting and capable of capturing complex relationships within the data.

#### **Model Parameters**

For the Random Forest, the parameters were set to a max\_depth of 5, a min\_samples\_split of 25 and a n\_estimators of 50. This set of parameters seems to be the best manually searched set for this model.



#### **Model Performance**

|  |  |  |
| --- | --- | --- |
| Model | ROC-AUC Score | Accuracy |
| Random Forest | 0.82 | 0.75 |

Evaluating the performance of the model, the ROC-AUC score is 0.82 which indicates a very good performance. The model has a strong ability to discriminate between the classes. The accuracy score shows that the model correctly predicts 75% of the instances.

A graph of a curve

Description automatically generated with medium confidence

#### **Feature Importance**

By ranking the important features of this model, it is possible to visualize that the most important variables are CREDIT\_HISTORY followed by APPLICANTINCOME and COAPPLICANTINCOME\_LOG as well as LOANAMOUNT. The importance level can be visualized below.

A screenshot of a computer screen

Description automatically generated

### **Random Forest with only best features**

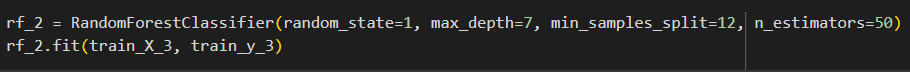
To experiment further with the best model found, the model will be run again but only with the most important features collected in the original model. In this case the variables SELF\_EMPLOYED\_YES, DEPENDENTS, PROPERTY\_AREA\_URBAN, GENDER\_MALE, MARRIED\_YES and EDUCATION\_NOT GRADUATE will be removed from the dataset and the model will be runed with only the important variables on it.

A computer screen with text and numbers

Description automatically generated

#### **Model Parameters**

For this model, the parameters were set to a max\_depth of 7, a min\_samples\_split of 12 and a n\_estimators of 50. This set of parameters seems to be the best manually searched set for this model.



#### **Model Performance**

|  |  |  |
| --- | --- | --- |
| Model | ROC-AUC Score | Accuracy |
| Random Forest with BF | 0.82 | 0.77 |

Evaluating the performance of the model, the ROC-AUC score is 0.82 which indicates a very good performance. The model has a strong ability to discriminate between the classes. The accuracy score shows that the model correctly predicts 77% of the instances. Overall, this model is better than the original model.

A graph of a line graph

Description automatically generated with medium confidence

#### **Feature Importance**

By ranking the important features of this model, it is possible to visualize that the most important variables are CREDIT\_HISTORY followed by APPLICANTINCOME and LOANAMOUNT as well as COAPPLICANTINCOME\_LOG. The importance level can be visualized below.

A screenshot of a graph

Description automatically generated

### **Gradient Boost**

Gradient Boosting is an ensemble learning technique used for regression and classification tasks. It builds models sequentially, with each new model correcting the errors made by the previous ones. The method involves training models on the residual errors of the previous models, optimizing performance through gradient descent.

#### **Model Parameters**

The parameters for this model were the default settings for this one.

A black background with white text

Description automatically generated

#### **Model Performance**

|  |  |  |
| --- | --- | --- |
| Model | ROC-AUC Score | Accuracy |
| Gradient Boost | 0.78 | 0.73 |

Evaluating the performance of the model, the ROC-AUC score is 0.78 which indicates a good performance. The model has a reasonably strong ability to discriminate between the classes. The accuracy score shows that the model correctly predicts 73% of the instances.

A graph of a curve

Description automatically generated with medium confidence

#### **Feature Importance**

By ranking the important features of this model, it is possible to visualize that the most important variables are CREDIT\_HISTORY followed by APPLICANTINCOME and LOANAMOUNT as well as COAPPLICANTINCOME\_LOG. The importance level can be visualized below.

A screenshot of a computer

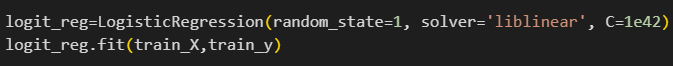
Description automatically generated

### **Full Logistic Regression**

Logistic regression is employed for both binary and multiclass classification tasks. It utilizes the logistic (or sigmoid) function to model the relationship between input features and the probability of the target variable belonging to a specific class.

#### **Model Parameters**

For the full logistic regression model, the solver will be liblinear as the data set is not too big this solver is the best one for this data set. Also, the liblinear solver employs the coordinate descent approach to optimize the objective function, updating one coefficient at a time while keeping the remaining coefficients fixed.



#### **Model Performance**

|  |  |  |
| --- | --- | --- |
| Model | ROC-AUC Score | Accuracy |
| Full Logistic Regression | 0.81 | 0.75 |

Evaluating the performance of the model, the ROC-AUC score is 0.81 which indicates a very good performance. The model has a strong ability to discriminate between the classes. The accuracy score shows that the model correctly predicts 75% of the instances.

A graph of a line graph

Description automatically generated with medium confidence

#### **Feature Importance**

The most important variables based on odds ratio are showed below where the most important variables are CREDIT\_HISTORY followed by PROPERTY\_AREA\_SEMIURBAN and MARRIED\_YES.

A screenshot of a graph

Description automatically generated

For the interpretation of the best 3 variables, it can be said that:

* **Credit\_History**: Applicants who have credit history are 10.57 times more likely to get approved for a loan than applicants without credit history.
* **Property\_Area\_Semiurban**: Applicants who are looking for a property in a semiurban area are 1.81 times more likely to get approval for a loan than applicants looking for properties in rural areas.
* **Married\_Yes**: Applicants who are married are 1.56 times more likely to get approval for a loan than applicants that are not married.

### **Forward Logistic Regression**

#### **Model Parameters**

For the forward logistic regression, a best\_forw\_model parameter is implemented to get the best forward logistic regression.

A screen shot of a computer code

Description automatically generated

#### **Model Performance**

|  |  |  |
| --- | --- | --- |
| Model | ROC-AUC Score | Accuracy |
| Forward Logistic Regression | 0.71 | 0.74 |

Evaluating the performance of the model, the ROC-AUC score is 0.71 which indicates a good performance. The model has a reasonably strong ability to discriminate between the classes. The accuracy score shows that the model correctly predicts 74% of the instances.

A line graph with blue and orange lines

Description automatically generated

#### **Feature Importance**

The most important variable based on the odds ratio is only the CREDIT\_HISTORY. Unfortunately, there are not any other variables to interpret for this model. For this variable it can be said that applicants who have a credit history are 11.71 times more likely to get approved for a loan than applicants without a credit history.

A screenshot of a computer

Description automatically generated

### **Backward Logistic Regression**

#### **Model Parameters**

For the backward logistic regression, a best\_back\_model parameter is implemented to get the best backward logistic regression.

A screen shot of a computer code

Description automatically generated

#### **Model Performance**

|  |  |  |
| --- | --- | --- |
| Model | ROC-AUC Score | Accuracy |
| Forward Logistic Regression | 0.80 | 0.77 |

Evaluating the performance of the model, the ROC-AUC score is 0.80 which indicates a very good performance. The model has a strong ability to discriminate between the classes. The accuracy score shows that the model correctly predicts 77% of the instances.

A graph of a line graph

Description automatically generated with medium confidence

#### **Feature Importance**

The most important variables based on odds ratio are showed below where the most important variables are CREDIT\_HISTORY followed by PROPERTY\_AREA\_SEMIURBAN and MARRIED\_YES.

A screenshot of a graph

Description automatically generated

For the interpretation of the best 3 variables, it can be said that:

* **Credit\_History**: Applicants who have a credit history are 13.41 times more likely to get approved for a loan than applicants without a credit history.
* **Property\_Area\_Semiurban**: Applicants who are looking for a property in a semiurban area are 1.99 times more likely to get approval for a loan than applicants looking for properties in rural areas.
* **Married\_Yes**: Applicants who are married are 1.95 times more likely to get approval for a loan than applicants that are not married.

### **Stepwise Logistic Regression**

#### **Model Parameters**

For the Stepwise logistic regression, a best\_step\_model parameter is implemented to get the best stepwise logistic regression.

A computer screen with text

Description automatically generated

#### **Model Performance**

|  |  |  |
| --- | --- | --- |
| Model | ROC-AUC Score | Accuracy |
| Stepwise Logistic Regression | 0.71 | 0.74 |

Evaluating the performance of the model, the ROC-AUC score is 0.71 which indicates a good performance. The model has a reasonably strong ability to discriminate between the classes. The accuracy score shows that the model correctly predicts 74% of the instances.

A graph of a line

Description automatically generated with medium confidence

#### **Feature Importance**

The most important variable based on the odds ratio is only the CREDIT\_HISTORY. Unfortunately, there are not any other variables to interpret for this model. For this variable it can be said that applicants who have a credit history are 11.71 times more likely to get approved for a loan than applicants without a credit history.

A screenshot of a computer

Description automatically generated

# **Model Recommendation**

## **Model Scores**

|  |  |  |
| --- | --- | --- |
| Model | ROC-AUC Score | Accuracy |
| [Random Forest with BF](#_Random_Forest_only) | 0.82 | 0.77 |
| [Random Forest](#_Random_Forest) | 0.82 | 0.75 |
| [Full Logistic Regression](#_Full_Logistic_Regression) | 0.81 | 0.75 |
| [Forward Logistic Regression](#_Forward_Logistic_Regression) | 0.80 | 0.77 |
| [Gradient Boost](#_Gradient_Boost) | 0.78 | 0.73 |
| [Stepwise Logistic Regression](#_Stepwise_Logistic_Regression) | 0.71 | 0.74 |
| [Forward Logistic Regression](#_Forward_Logistic_Regression) | 0.71 | 0.74 |
| [GridSearch Decision Tree](#_GridSearch_Decision_Tree) | 0.71 | 0.74 |
| [Decision Tree ASE](#_Decision_Tree) | 0.67 | 0.69 |
| [Full Misclassification Tree](#_Full_Misclassification_Tree) | 0.61 | 0.64 |

## **Model Selection**

For selecting the best model, ROC-AUC is used as the primary metric because it provides a more comprehensive view of the model performance. This is especially important for the slightly imbalanced dataset, as relying solely on accuracy could lead to misleading conclusions. Lastly, accuracy is used as the third metric to determine how accurate the model is, regardless of the class.

Based on these metrics, the best model is the Random Forest with only the best features which has a ROC-AUC score of 0.82 and can accurately predict the data 77% of the time. This model represents the best features of the Random Forest and is a re-run of the original model but with less variables. The reduction in variables led the model to achieve a 2% gain in accuracy. The reduction in variables also will lead to a reduction in costs and time saving as less information is needed in the application process for loan approval.

A screenshot of a graph

Description automatically generated

The most important variables of this model are primarily CREDIT\_HISTORY which is if the applicant have a credit history or not, which have the highest level of importance with 0.37, followed by APPLICANTINCOME which is the income of the primary applicant, with 0.21 of level of importance, LOANAMOUNT which is the amount of the loan requested, with 0.17 of level of importance and COAPPLICANTINCOME\_LOG which is the log of the Co Applicant Income with 0.16 of level of importance.

A screenshot of a graph

Description automatically generated

Now, taking the odds ratio of the next best regression model, the full logistic regression, it is possible to analyze a little further the interpretation of the most important variables of the Random Forest with only the best features.

For the CREDIT\_HISTORY variable is possible to say that applicants who have a credit history are 10.57 times more likely to get approved for a loan than applicants without credit history.

Also, it is reflected that for APPLICANTSINCOME, for every unit increase in the applicant’s income, the odds of the applicant getting approved for a loan increase by 0.0005%. This is a very small effect, implying that applicant income does not significantly impact the odds of the applicant getting approved for a loan in this model, but the variable has some sort of importance in the Random Forest with only the best features.

For the LOANAMOUNT variable, the odds ratio tells that for every unit increase in the loan amount, the odds of the applicant getting approved for a loan decrease by approximately 0.32%. This suggests that higher loan amounts are slightly associated with lower odds of getting approved for a loan, so as loan amounts increase the likelihood of getting approved decreases.

Finally, for the COAPPLICANTINCOME\_LOG variable, the odds ratio says that for every unit increase in the log of the co-applicant income, the odds of the applicant getting approved for a loan increase by approximately 7.41%. This means that as the log of the co-applicant income increases, the likelihood of the applicant getting a loan approved increases moderately.

# **Conclusion and Recommendations**

In evaluating the loan approval process through advanced machine learning techniques, the analysis has provided key insights that can drive more effective business decisions. By focusing on the most impactful variables and refining the model to enhance accuracy, the company is well-positioned to optimize its loan approval operations. The following conclusions outline how these insights can be translated into actionable strategies that not only improve efficiency but also contribute to risk management, targeted marketing, and long-term growth.

After developing different predictive models, the Random Forest model with only the best features was selected as the best model for predicting the appropriate approval of loans. The selection of the model was made taken into consideration primary metrics like ROC-AUC which was 0.82 for the selected model, and accuracy which was 0.77 for the selected model. The analysis identified key variables as crucial for accurate model predictions, including the applicant's credit history, the income of both the applicant and co-applicant, and the requested loan amount.

## **Impacts on Business Problem**

Based on the model selection and results, several actionable insights can be applied to enhance the loan approval process for housing loans.

* **Operational Efficiency for the Loan Application Process**: The Random Forest model demonstrates that focusing on key features like credit history and income levels can reduce the need for collecting excessive information from applicants. The reduction in features used in the model, leading to a 2% gain in accuracy for the Random Forest model, suggests that the business can focus on fewer but more impactful data points in its decision-making processes. This reduction in complexity not only speeds up the decision-making process but also cuts down on the operational costs associated with data collection and analysis.
* **Risk Management and Decision-Making**: The findings highlight that applicants with a solid credit history are significantly more likely to be approved for loans. This insight can guide the business in implementing stricter credit history checks, thereby reducing the risk of approving loans to high-risk clients. By prioritizing applicants with favorable credit history, the company can lower default rates and improve its loan portfolio's overall health.
* **Targeted Marketing**: Understanding that higher loan amounts slightly decrease the likelihood of approval can help with marketing strategies. The business can target lower-risk customers with tailored loan products that match their income levels and repayment capabilities, improving approval rates and customer satisfaction.
* **Customer Segmentation:** Recognizing the positive impact of co-applicant income on loan approval can encourage the promotion of joint applications, particularly to customers with lower individual incomes but higher combined household incomes, leading to better approval rates and higher loan disbursements.

## **Recommended Next Steps**

To implement the business solutions effectively, several key actions should be taken across different areas of the loan approval process:

* **Implementation of Automation for Loan Application Process**: A review of the current loan processing operations is needed to identify areas where reducing the number of variables can lead to efficiency gains. Software updates that automate the use of fewer, more impactful features in the loan approval process should be implemented, ensuring that decisions are both quick and accurate. Continuous monitoring of processing times and costs will be essential to assess the impact of this streamlined approach, with adjustments made as necessary to maintain and improve efficiency.
* **Integration of Stricter Credit History Protocols**: It is crucial to develop and integrate stricter credit history assessment protocols into the loan approval workflow, leveraging the insights gained from the model selected. Monitoring and analyzing default rates of approved loans over time will allow for the fine-tuning of credit history thresholds and approval criteria. A periodic review process should be established to regularly assess and adjust risk management strategies based on updated data and model predictions, ensuring that the company's approach remains effective and responsive to changing conditions.
* **Marketing Strategies for Joint Loans and Low-Risk Applicants**: Campaigns should be designed to target low-risk applicants, promoting loan products that align with their income levels and financial profiles. Joint loan applications should be encouraged, particularly for segments where combined household income can improve approval odds, using personalized communication strategies to increase engagement. Customer responses to these targeted marketing efforts should be analyzed to refine segmentation strategies and maximize both loan approvals and overall customer engagement.

The insights from this project can be used to refine long-term loan approval strategies, ensuring that the business focuses on high-quality applicants. By continually analyzing and refining the models based on new data, the company can adapt to changing market conditions and customer behaviors, maintaining a competitive edge in the financial services industry.

# **References**

## **References**

Kaggle. Federal Trade Commission (FTC). <https://www.kaggle.com/datasets/willianoliveiragibin/federal-trade-commission-ftc>

American Bankers Association. (2021). *The State of Digital Lending.* Washington: Accenture.

Kagan, J. (2024, February 28). *Investopedia*. Retrieved from What Is a Loan, How Does It Work, Types, and Tips on Getting One: https://www.investopedia.com/terms/l/loan.asp

Koch, T., & MacDonald, S. (2014). *Bank Management.* Cengage Learning.