|  |
| --- |
|  |
| Bank Loan Default Prediction |
| |  | | --- | | Prepared By: Parisha Desai and Swetha Bommireddy  10-30-2023 |   Purdue University Fort Wayne |

Table of Contents

[1. Abstract 2](#_Toc149588968)

[2. Introduction 2](#_Toc149588969)

[3. Purpose of the Project 2](#_Toc149588970)

[4. Audience of the Project 3](#_Toc149588971)

[5. Research Methodology 3](#_Toc149588972)

[6. Data Pre-processing 3](#_Toc149588973)

[6.1 Understanding the Dataset 3](#_Toc149588974)

[6.2 Handling missing values 4](#_Toc149588975)

[6.3 Data Transformation 5](#_Toc149588976)

[7. Exploratory Data analysis 6](#_Toc149588977)

[7.1 Univariate analysis 6](#_Toc149588978)

[7.2 Bivariate analysis 8](#_Toc149588979)

[7.3 Correlation matrix 11](#_Toc149588980)

[8. Classification Model 13](#_Toc149588981)

[9. Future Actionable 15](#_Toc149588982)

## Abstract

The 'Bank Loan Default Prediction' project aims to develop models for predicting whether a customer is likely to default on a loan, based on various applicant profile and loan details. The project's significance lies in enabling banks to make informed decisions when granting loans, potentially reducing business losses. By identifying key factors and correlations, the project seeks to empower financial analysts, banking personnel, data scientists, and data analytics students. The research methodology involves data analysis, exploratory data analysis (EDA), and the creation of a Logistic Regression prediction model to assess the likelihood of customer default. As the project progresses, further exploration and study of additional models will be considered based on the performance of the initial Logistic Regression model. This approach aims to enhance the accuracy and precision of loan default predictions.

## Introduction

The primary focus of the project is determining various factors that can help predict whether a customer will default on a loan and creating models that will help in predicting the likelihood of customer defaulting. The project's importance lies in developing diverse models to accurately assess the likelihood of a user defaulting on a loan. This endeavor aims to empower banks to proactively take corrective actions before granting a loan, primarily based on the applicant's profile information. By doing so, the bank can make informed decisions, potentially refusing a loan to individuals at high risk of default while approving loans for customers more likely to repay promptly, thereby mitigating potential business losses. This project seeks to establish a set of models that can be widely adopted by banks for future loan applications.

The dataset used for the analysis has been obtained from Kaggle. This is the URL where the dataset has been placed [parisha-homework-acs560 / data-analytics-using-r — Bitbucket](https://bitbucket.org/parisha-homework-acs560/data-analytics-using-r/src/master/) and our project is located at <https://github.com/swetha-r13/Bank-Loan-Default-Prediction> . The dataset contains various information about the applicants currently applying for loans such as their loan details – loan amount, loan interest rate, loan term, and customer demographics. The dataset also has information about the previous applicants, their loan details, demographics, and the decisions taken on those applications.

## Purpose of the Project

The project aims to identify the key factors that contribute to correctly determining whether a customer will default on paying the loan by deriving different correlations and creating models with the best accuracy and precision that can predict the likelihood of loan default given the characteristics of the loan applicant. Once these factors are determined and different correlations drawn it can be inferred whether a particular customer with a certain demographic detail may or may not default on repaying the loan amount. Predicting this customer behavior will help banks in performing key decision-making as to whether they should approve a loan and at the same time ensure that a person who is eligible for a loan is not denied loan sanctioning. The likelihood model of the customer defaulting on a loan can also be used by the customer to understand the risk involved in taking that loan and then deciding on whether the customer can proceed with the loan sanctioning process.

## Audience of the Project

The audience using this analysis will be financial analysts, banking personnel, data scientists, and students studying data analytics. They can use the work captured in this project and further research and study the predictions and conclusions drawn as part of this project.

## Research Methodology

The key research question we are trying to answer is ‘**What are the important factors that help determine whether a customer will default or not.’**

Our approach towards this project has been with initially understanding the dataset, tiding, and transforming the dataset for easy of handling and interpretation, handling the missing values and then conducting Exploratory data analysis by performing different univariate and bivariate analyses and drawing correlations between the impacting parameters. Basis the parameters identified we have created a Logistic Regression prediction model wherein we have split the dataset into training and testing dataset and have then computed the accuracy and precision of the model in predicting the probability of default.

For the EDA we have drawn various graphs and statistics using univariate and bivariate analysis as this will help us determine the impact of the individual variables when considered independently as well as when considered in pair which variables strongly influence probability of default.

We have selected Logistic Regression as our model for the purpose of prediction as our expected output is binary in nature wherein a customer may have defaulted or would have not defaulted and for such cases Logistic regression is the most suitable model.

## Data Pre-processing

### Understanding the Dataset

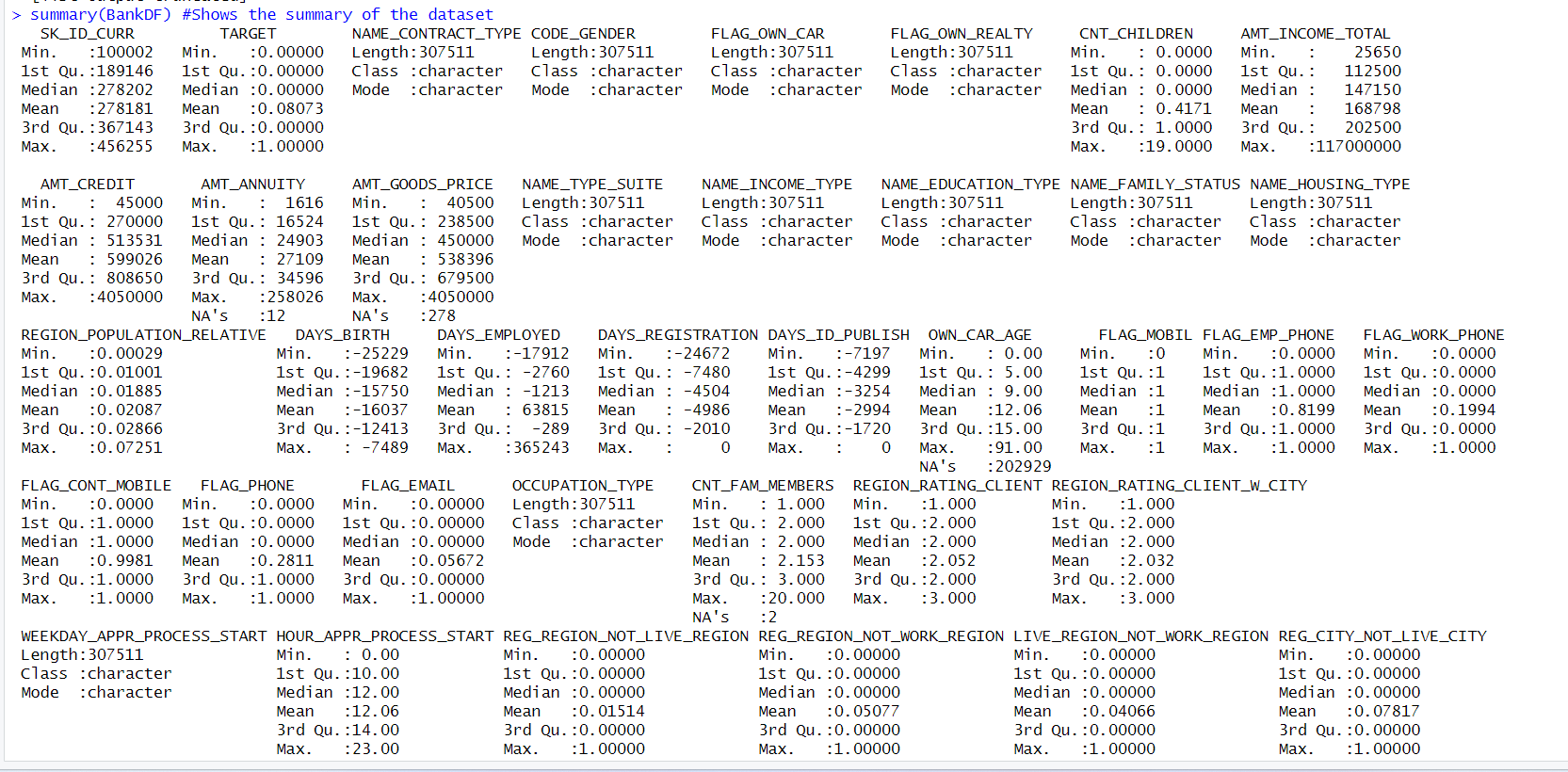
We have used dim () and str () functions to understand the dataset in order to know the number of rows and columns as well as the datatypes and the kind of value available in individual column.



A screen shot of a computer

Description automatically generated

In order to get a fair understanding of the range of values in the dataset we have used summary () function so that we can know the kind of data handling and transformation that shall be required so that we can accordingly take the steps to perform same.



### Handling missing values

We are using the is.na function to identify the missing values at column level and across the entire dataset.



A group of text on a white background

Description automatically generated

Since the number of records in dataset is around 3,00,000 with the large volume of data, we have taken step to remove those rows that have more than 50% of missing values. At the same time there are columns containing more than 40% missing value we have dropped such columns from our dataset.

In the columns containing categorical value (like house type and occupation type) and having missing value we have handled the missing value by inserting Unknown for all such data

In the columns containing numerical value (like Amount of credit, Amount of loan amount, Annuity) and having missing value we have handled the missing value by replacing the missing value with mean for all such data.

### Data Transformation

We have found inconsistencies in the dataset wherein few columns like DAYS\_BIRTH, DAYS\_EMPLOYED, DAYS\_REGISTRATION, DAYS\_ID\_PUBLISH, DAYS\_ID\_PUBLISH, DAYS\_LAST\_PHONE\_CHANGE have negative values and as the data in these type of columns cant be negative we have converted the negative values into positive values

We have transformed a few continuous variables into categorical variables as it will help us draw better insights into customer characteristics impacting the likelihood of default. We have bin AMT\_CREDIT variable to AMT\_CREDIT\_RANGE, AMT\_INCOME\_TOTAL variable to AMT\_INCOME\_RANGE and DAYS\_BIRTH variable to AGE\_RANGE.

We have included only distinct rows in our dataset so to avoid any duplicity that may have been introduced after handling missing values.

After doing all the necessary data transformation we have checked for the number of missing values and duplicate rows whose output indicates that the dataset is clean

## Exploratory Data analysis

### Univariate analysis

We have used the variables below for performing univariate analysis and deriving inferences on the loan payment frequency.

* CODE\_GENDER,
* NAME\_INCOME\_TYPE,
* NAME\_EDUCATION\_TYPE,
* NAME\_HOUSING\_TYPE,
* OCCUPATION\_TYPE,
* ORGANIZATION\_TYPE
* AMT\_CREDITRANGE
* AMT\_INCOME\_RANGE
* AGE\_RANGE

A group of different colored bars

Description automatically generated

From the above bar graph, it can be derived that the maximum number of customers involved in loan payment are females that are belonging to an age group of 30-40 with a low income wherein the loan amount is belonging to lower bracket.

A graph of different colored rectangular shapes

Description automatically generated

A graph with different colored bars

Description automatically generated

From the bar graph it is known that the highest number of customers making the loan payments are married employees, owning a house or an apartment with a secondary education where the main source of income is through work. However, the occupation for these customers is not known.

Thus, the bank should focus on targeting these kinds of customers belonging to above mentioned customer segment with respect to selling loan as product.

### 7.2 Bivariate analysis

1. AGE & Gender vs TARGET

A graph of different colored rectangles

Description automatically generated

It can be inferred that most of the female customers have no issues in paying the loan and customers belonging to age group 30-40 are most able to make payment on time. Additionally, people in the age group of 20-30 have a higher possibility of default than people belonging to age group of 50-60. People below age group of 20 have no payment visibility given that this is younger age group with minimal to no income and hence is understood.

2. Amount of credit(loan), Income type, and Amount of Income vs TARGET

A group of red squares

Description automatically generated

Customers with less credit, low income and belonging to the working group category are most likely to make payment. Customers having medium and high credit can also be considered while lending the loan as the target customer group. The distribution of people able to repay the loan among the low, medium, and high credit loan amount group is similar.

1. Types of Education, Marital Status & Types of Housing vs TARGET

A group of red squares

Description automatically generated

Married Customers with secondary education and staying in a house or an apartment are most likely to make payments when compared to customers with academic degree.

1. AMT\_CREDIT, AMT\_ANNUITY TARGET, AMT\_GOODS\_PRICE vs TARGET

A graph of a number of different numbers

Description automatically generated with medium confidence

Most customers have paid their amount of credit and have good price on time. However, they are unable to pay the annuity fees on time.

### 7.3 Correlation matrix

A graph with a dotted line

Description automatically generated

Based on the correlation matrix, we have selected a few variables AMT\_INCOME\_TOTAL, AMT\_CREDIT, AMT\_GOODS\_PRICE, AMT\_INCOME\_RANGE, CODE\_GENDER, NAME\_EDUCATION\_TYPE FOR Exploratory data analysis.

AMT\_INCOME\_TOTAL vs AMT\_CREDIT, AMT\_GOODS\_PRICE vs AMT\_CREDIT

A comparison of a graph

Description automatically generated

Those who have paid the loan amount on/within time are more likely to get higher credits than those who didn’t pay/did late payments. People who have higher goods prices and have made payments on time have higher credits than those with higher goods price but didn’t pay loan.

AMT\_INCOME\_RANGE vs CODE\_GENDER

A screenshot of a graph

Description automatically generated

We can see that Females with low income don’t have any payment issues.

AMT\_CREDIT vs NAME\_EDUCATION\_TYPE

A graph of colorful bars

Description automatically generated with medium confidence

From the boxplot above:  
(1) Some of the highly educated, married person are having credits higher than those who have done lower secondary education.  
(2) Those with higher education have higher credits and are more likely to make payments on time.  
(3) More number of outliers are seen in higher education.  
(4) The people with secondary and secondary special education are less likely to make payments on time.

## Classification Model

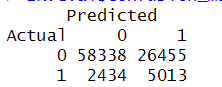
Our problem statement is to create model that will predict whether a individual will default or not basis set of given customer characteristics since our expected output is binary in nature i.e. either customer will default or customer will not default it is essential for us to use a classification model for purpose of classification to classify customer into the right group.

We have used a Logistic Regression model for the purpose of classification as it helps in classifying outputs that are binary in nature.

We have created a training and testing dataset wherein training dataset contains the 70% of the data and testing dataset contains 30% of the dataset we have trained the model using glm () function with family=” binomial” to identify the probability of likelihood of customer defaulting. Once the model is trained, we have tested the model by running the model on test data and have verified the performance of the model to predict the output when model run on testing dataset using predict () function.

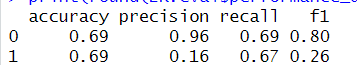
We have then created a confusion matrix to predict the data and have computed model performance using parameters like precision, recall and accuracy.

**Confusion Matrix**



* **True Positives (Actual 1, Predicted 1)**: There are 5013 cases where the model correctly predicted that the user would default (actual default cases).
* **True Negatives (Actual 0, Predicted 0)**: There are 58338 cases where the model correctly predicted that the user will not default (actual non-default cases).
* **False Positives (Actual 0, Predicted 1)**: There are 26455 cases where the model incorrectly predicted that the user would default when they did not (false alarms).
* **False Negatives (Actual 1, Predicted 0)**: There are 2434 cases where the model incorrectly predicted that the user will not default when they actually did (missed opportunities). This number indicates that if this model is used then the bank may end up approving loan applications wrongly for such applicants wherein the loan application instead should have been rejected as there was probability that such applicants would default.

**Performance Metrics**



* **Accuracy**: The overall accuracy of the model for predicting default as well as non-default is 69%. This means that 69% of the model's predictions are correct.
* **Precision**:
  + For class 0 (non-default): The precision is high at 96%. This means that when the model predicts a non-default, it is correct 96% of the time.
  + For class 1 (default): The precision is lower at 16%. This suggests that when the model predicts a default, it is correct only 16% of the time. In other words, the model has a high rate of false positives for class 1 which can be inferred even from the confusion matrix. This percentage indicates that if the model is used then the bank may end up rejecting loan applications of multiple applicants wrongly wherein the loan application instead should have been approved.
* **Recall** (Sensitivity):
  + For class 0 (non-default): The recall is 69%, meaning the model correctly identifies 69% of the actual non-default cases.
  + For class 1 (default): The recall is 67%, indicating that the model correctly identifies 67% of the actual default cases.
* **F1 Score**:
  + For class 0 (non-default): The F1 score is 80%, which is a balance between precision and recall for class 0.
  + For class 1 (default): The F1 score is 26%, suggesting that there's a trade-off between precision and recall for class 1.

Overall, the model is not able to correctly capture the likelihood of person default as is indicated by low precision rate for defaulting class. Moreover, the number of high false positives indicates if bank were to use this model; the cost of loss in business that the bank may face by rejecting the loan applications wherein they should have approved the loan applications would be significant. Hence, we can either fine tune focusing on different set of parameters or choosing selected few parameters amongst the ones that have been chosen for this model or using a different model are the possible options.

## Future Actionable

* Basis the output received from predictive logistic regression model we understand that the accuracy of the model is not the best and the cost tradeoff for the false positives is quite high hence we plan on including a descriptive model such as Logistic Regression with Feature Importance Analysis wherein we will analyze the coefficients of the logistic regression model, and gain insights into the importance of each variable in predicting loan defaults. This can help us understand the factors contributing to defaults in a more interpretable manner.
* The other predictive model that we plan to consider in our analysis is the Decision Tree Classification model as it is capable of handling a mix of data and since our dataset has a mix of numerical and categorical variables, we believe this model will help us handle the versatility of the data and classify default likelihood correctly.
* Another model that we plan to explore is the Naïve Bayes classification model since our problem statement is focused on determining whether an individual defaults or not, we can consider this model at the same time we will have to account for the independence between the variables that are assumed as part of Naïve Bayes model.
* We plan to compare the performance of the models mentioned above with the logistic regression model implemented in this phase and derive the model with the best performance and accuracy for predicting correctly whether an individual would default or not given their individual characteristics and attributes.

1. **References**
2. <https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12544&context=theses>
3. <https://medium.com/geekculture/mastering-loan-default-prediction-tackling-imbalanced-datasets-for-effective-risk-assessment-8e8dfb2084d0>
4. <https://www.scaler.com/topics/data-science/loan-default-prediction/>