Parisha Desai

Prof. Adolfo Coronado

Data Analytics in Business using R

4th December 2023

Bank Loan Default Prediction

Table of Contents

1. Abstract 3
2. Introduction 3
3. Purpose of the Project 3
4. Audience of the Project 4
5. Research Methodology 4
6. Understanding the dataset and Data Pre-processing 4
7. Exploratory Data analysis 5
8. Classification Model Analysis 6
   8.1 Evaluating Classification model performance basis Confusion Matrix 6
   8.2 Evaluating Classification model performance basis performance metric 7
   8.3 Comparing Classification model performance and best model selection 10
9. Conclusion 11

1. Abstract

The 'Bank Loan Default Prediction' project aims to develop models for predicting whether a customer will likely default on a loan based on various applicant profile attributes and loan details. The project's significance lies in enabling banks to make informed decisions when granting loans, potentially reducing business losses. The project aims to empower banking personnel by using predictive models that will aid in decision making during loan sanctioning process. The research methodology involves data analysis, exploratory data analysis (EDA), and the creation of various prediction model to assess the likelihood of customer default. The project approach aims to enhance the accuracy and precision of loan default predictions.

2. Introduction

The project's primary focus is determining various factors that can help predict whether a customer will default on a loan and creating models to help predict the likelihood of a 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 empowers 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 banks can widely adopt 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 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. The primary objective of utilizing this dataset is to develop predictive models for assessing the likelihood of loan default. Notably, the focus is not on conducting an analysis for a specific time frame or location. Instead, our aim is to concentrate on studying model performance, avoiding the drawing of historical trends that might be referenced for future decision-making.

3. 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 perform key decision-making as to whether they should approve a loan and, at the same time, ensure that a person eligible for a loan is not denied loan sanctioning. The likelihood model of the customer defaulting on a loan if is provided by the bank in form

of assessment tool to the customer; it can also be used by the customer to understand the risk involved in taking that loan and then deciding on whether he/she should proceed with the loan sanctioning process.

4. Audience of the Project

The audience using this analysis will be banking personnel who can use the work captured in this project and further apply the models and study the results obtained in order to assist them in better decision making in loan sanctioning process.

5. 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'** and **'Can we predict the likelihood of customer defaulting or not loan payment'.**

Our approach towards this project has been to initially understand the dataset, tidying, and transform the dataset for ease of handling and interpretation, handle the missing values, and then conduct Exploratory data analysis by performing different univariate and bivariate analyses and drawing correlations between the impacting parameters. We have drawn various graphs and statistics for the EDA using univariate and bivariate analysis. This will help us determine the impact of the individual variables when considered independently and, when considered in pairs, which variables strongly influence the probability of default.

Basis the parameters identified, we have created three prediction model wherein we have split the dataset into training and testing datasets and have then computed the accuracy and precision of each model in predicting the probability of default. We have then analyzed the performance metric of each model and have then selected the best model for predicting the probability of default.

6. Understanding the dataset and Data Pre-processing

We have understood that the dataset consists of 3,07,511 observations and 122 variables. After analyzing the datatypes and the values of different variable it has been understood that the dataset is mix of categorical and continuous variables and that the dataset would need certain amount of handling for missing and duplicate data (ex. Variables like occupation type and housing type in our dataset had blank values for some records) as well as data transformation would be required (few variables like age and no. of years of service were in negative and had to be converted into positive value and we had to convert some continuous variables like income amount, credit amount, days since birth into categorical variable like income amount range, credit amount range and age range); so that the data is more interpretable. We also identified outliers in continuous variables like income amount and credit amount. We handled these outliers by replacing outliers with the values in value contained in 90% of data. It is recommended that banking personnel using the models captured in this report also first check for the data tidiness and perform necessary data cleaning and transformation exercises so that the data is interpretable.

7. Exploratory Data analysis

We have performed exploratory data analysis to study the dataset in the form of univariate and bivariate analysis. While performing the univariate analysis for variables like gender, age, credit amount, income type, education type, family status, housing type and occupation type we have found the following:

- The maximum number of customers involved in loan payments are females aged 30-40 with a low income, wherein the loan amount belongs to a lower bracket.
- The highest number of customers making 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 in our dataset is unknown.

Thus, through the univariate analysis we were able to identify the customer segments that the bank should focus on targeting for selling loans as products.

When performing bivariate analysis we have tried to observe the frequency of person defaulting or not basis various factors like age, gender, marital status, income type, credit amount, income amount, type of education and type of housing and we found the following:

- Most female customers have no issues in paying the loan, and customers aged 30-40 are most able to make payments on time. Additionally, people in the age group of 20-30 have a higher possibility of default than people belonging to the age group of 50-60. Thus the bank can employ stringent risk assessment measure for people belonging to age group of 20-30. People under the age group of 20 have no payment visibility given that this is a younger age group with minimal to no income and thus bank can consider filtering out the dataset of people belonging to this age group and excluding them from any analysis.
- Customers with less credit, low income, and belonging to the working group category are most likely to make payments. Thus the banking personnel can have less stringent risk assessment for these customer group and focus their marketing efforts in targeting this customer group. The distribution of people able to repay the loan among the low, medium, and high credit loan amount groups is similar. Thus, credit amount is not key differentiating factor that could provide us any distinct insights.
- 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. This indicates that there is some strong correlation between marital status, education type and housing type that can be explored and studied further.
- 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. The bank personnel can thus study the reasons of customers not being able to pay the annuity fees on time and work on measures that could help the customers to do timely payment.

Thus, through the bivariate analysis we were able to identify the variables that would have some correlation and at the same time we were able to draw insights on likelihood of loan payment and defaulting basis key attributes.

As last step of EDA we have drawn interpretations from correlation matrix and have found that:

- When considering income amount, credit amount and rate at which the credit was given it is found that those who have paid the loan amount on/within time are likelier to get higher credit amount than those who didn't pay/made late payments. People who have got loan on good rate and made payments on time have been offered higher credit amount than those with good rates but didn't pay loans.
- People with higher education students have higher credit amounts and are more likely to make timely payments. However the people with secondary and secondary special education are less likely to make payments on time. Thus education type has strong correlation to the amount of credit given and likelihood of loan payment

Thus the findings derived from EDA can help bank personnel to focus on key attributes – marital status, age, gender, income type, education type, housing type, income amount, credit amount and credit rate which will help them target the customer segment to which they should sell loan products basis their strong likelihood to repay loan in timely manner and at the same time they can build stringent risk assessment models for those customer segment that are more likely to default.

8. Classification Model Analysis

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

We have created a training and testing dataset wherein the training dataset contains the 70% of the data, and the testing dataset contains 30% of the dataset. We have created three classification models – Logistic Regression, Naïve Bayes and Decision Tree Classification model wherein first we have trained the model on training dataset and then tested the model's predictive ability to correctly classify an individual would default or not by running it on testing dataset and have verified model's performance in order to determine the best model basis its predictive ability.

8.1 Evaluating Classification model performance basis Confusion Matrix

Lets first understand the model performance basis the confusion matrix created for each of the models

Confusion Matrix		
Logistic Regression	Naives Bayes	Decision Tree
Predicted Actual 0 1 0 58338 26455 1 2434 5013	Predicted Actual 0 1 0 54056 30737 1 3113 4334	Predicted Actual 0 1 0 67441 17352 1 4225 3222

The Decision Tree:

- The model correctly predicted that 3222 users would default (actual default cases) and 67441 users would not default (actual non-default cases).

- There are 17352 cases where the model incorrectly predicted that the user would default when they did not (false alarms) and 4225 cases where the model incorrectly predicted that the user would not default when they actually did (missed opportunities). This suggests that the bank may end up approving loan applications wrongly for such applicants.

Naive Bayes:

- The model correctly predicted that 4334 users would default (actual default cases) and that 54056 users would not default (actual non-default cases).
- There are 30737 cases where the model incorrectly predicted that the user would default when they did not (false alarms) and there are 3113 cases where the model incorrectly predicted that the user would not default when they actually did. This suggests potential missed opportunities in terms of the number of customers (30737) whose loan application could get rejected if this model was used.

Logistic Regression:

- The model correctly predicted that 5013 users would default (actual default cases) and that 58338 users would not default (actual non-default cases).
- There are 26455 cases where the model incorrectly predicted that the user would default when they did not (false alarms) and there are 2434 cases where the model incorrectly predicted that the user would not default when they actually did.

Analysis:

- The false positive rate is lowest for Decision Tree classification model amongst all models thus the Decision Tree model is most preferred model considering the probability of bank losing business because of wrongly rejecting a loan to an individual who would have not defaulted would be least amongst all the models. In this regard, Naïve Bayes is the least preferred model because of the high number of false positives.
- The highest number of true positives are indicated by Logistic regression model thus indicating the ability of the model to correctly predict a loan default when the customer would actually default. The number of true positive and true negatives are more well balanced in logistic regression model as compared to other models

8.2 Evaluating Classification model performance basis performance metric

Now let's understand each model's performance basis the accuracy, precision, recall and F1 score metric

1. **Decision Tree Model**

	accuracy	precision	recall	f1
0	0.77	0.94	0.80	0.86
1	0.77	0.16	0.43	0.23

- **Accuracy:**

- Overall accuracy of the model is 77%, indicating that 77% of the predictions are correct.

- **Precision:**
  - For class 0 (non-default): Precision is 94%. This means that when the model predicts a non-default, it is correct 94% of the time.
  - For class 1 (default): Precision is 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.
- **Recall (Sensitivity):**
  - For class 0 (non-default): Recall is 80%, meaning the model correctly identifies 80% of the actual non-default cases.
  - For class 1 (default): Recall is 43%, indicating that the model correctly identifies 43% of the actual default cases.
- **F1 Score:**
  - For class 0 (non-default): The F1 score is 86%, which is a balance between precision and recall for class 0.
  - For class 1 (default): The F1 score is 23%, suggesting that there's a trade-off between precision and recall for class 1.

Interpretation:

- The model performs very well in predicting non-default cases with high precision (94%) and recall (80%).
- However, the model struggles with predicting default cases. The low precision (16%) indicates a high rate of false positives, and the relatively low recall (43%) indicates that the model misses a substantial portion of actual default cases.
- The F1 score for class 0 is high, reflecting a good balance between precision and recall. However, the F1 score for class 1 is lower, indicating the challenges in achieving a good trade-off between precision and recall for default cases.

2. Naïve Bayes Model

	accuracy	precision	recall	f1
0	0.63	0.95	0.64	0.76
1	0.63	0.12	0.58	0.20

- **Accuracy:**
  - Overall accuracy of the model is 63%, indicating that 63% of the predictions are correct.
- **Precision:**

- For class 0 (non-default): Precision is 95%. This means that when the model predicts a non-default, it is correct 95% of the time.
- For class 1 (default): Precision is 12%. This suggests that when the model predicts a default, it is correct only 12% of the time. In other words, the model has a high rate of false positives for class 1.

- **Recall (Sensitivity):**
  - For class 0 (non-default): Recall is 64%, meaning the model correctly identifies 64% of the actual non-default cases.
  - For class 1 (default): Recall is 58%, indicating that the model correctly identifies 58% of the actual default cases.
- **F1 Score:**
  - For class 0 (non-default): The F1 score is 76%, which is a balance between precision and recall for class 0.
  - For class 1 (default): The F1 score is 20%, suggesting that there's a trade-off between precision and recall for class 1.

Interpretation:

- Similar to the Decision Tree model, Naive Bayes performs well in predicting non-default cases with high precision (95%) and reasonable recall (64%).
- However, the model struggles with predicting default cases. The low precision (12%) indicates a high rate of false positives, and the relatively moderate recall (58%) indicates that the model misses a significant portion of actual default cases.
- The F1 score for class 0 is high, reflecting a good balance between precision and recall. However, the F1 score for class 1 is lower, indicating challenges in achieving a good trade-off between precision and recall for default cases.

3. **Logistic Regression Model**

	accuracy	precision	recall	f1
0	0.69	0.96	0.69	0.80
1	0.69	0.16	0.67	0.26

- **Accuracy:**
  - Overall accuracy of the model is 69%, indicating that 69% of the predictions are correct.
- **Precision:**
  - For class 0 (non-default): Precision is 96%. This means that when the model predicts a non-default, it is correct 96% of the time.

    - For class 1 (default): Precision is 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.

- **Recall (Sensitivity):**
    - For class 0 (non-default): Recall is 69%, meaning the model correctly identifies 69% of the actual non-default cases.
    - For class 1 (default): 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.

Interpretation:

- The Logistic Regression model performs well in predicting non-default cases (class 0) with high precision (96%) and reasonable recall (69%).
- However, similar to the Naive Bayes model, the Logistic Regression model struggles with predicting default cases (class 1). The low precision (16%) indicates a high rate of false positives, and the relatively moderate recall (67%) indicates that the model misses a portion of actual default cases.
- The F1 score for class 0 is high, reflecting a good balance between precision and recall. However, the F1 score for class 1 is lower, indicating challenges in achieving a good trade-off between precision and recall for default cases.

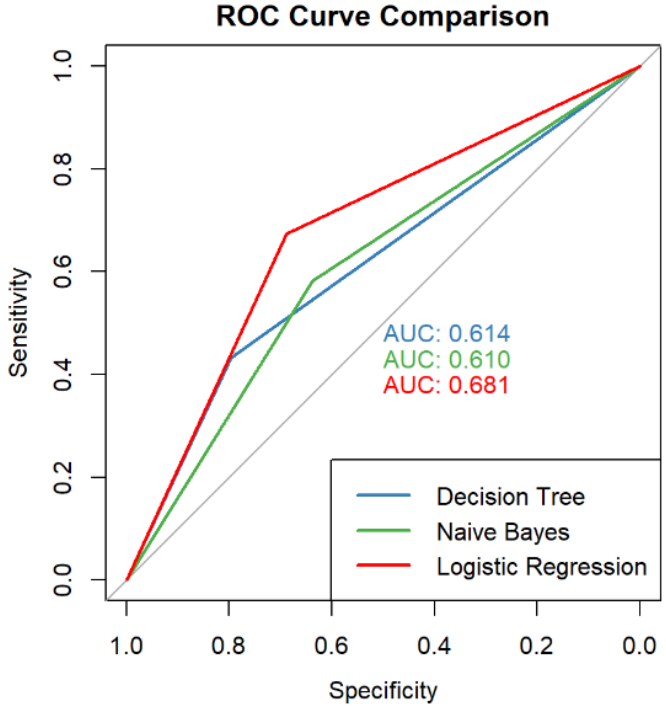
8.3 Comparing Classification model performance and best model selection

The table captures comparison of Models performance based on key performance metric

Method	Accuracy	Precision	Recall	FScore	ROC
Decision Tree	0.77	0.55	0.61	0.55	0.61
Naive Bayes	0.63	0.53	0.61	0.48	0.61
Logistic Regression	0.69	0.56	0.68	0.53	0.68

- Decision Tree achieves the highest accuracy among the three models but we need to also consider that it works well in predicting non-defaulters but less effectively for defaulters which is known by the true positive rate which is lowest amongst all models.

- Decision Tree and Naive Bayes have similar ROC scores, indicating comparable performance in distinguishing between classes.
- Logistic Regression has the highest ROC score, suggesting a better ability to distinguish between positive and negative instances. It also strikes a good balance between precision and recall.



Considering the ROC and overall performance of Logistic Regression model we suggest that the Logistic Regression model is the best choice in predicting if a person has the ability to repay the loan.

9. Conclusion

- We have been able to identify that following are key variables that help determine the probability of individual defaulting – marital status, age, gender, income type, education type, housing type, income amount, credit amount and credit rate. The bank personnel are thus suggested to further research and study past trends based on these parameters.
- We have been able to identify Logistic Regression as best classification model that helps classify the probability of loan default considering various performance criteria like accuracy, precision, F1 score and ROC.
- There have been few limitations in the dataset that we couldn't use a historical data across different time period to draw trends and further fine tune our analysis in determining the key factors that help determine whether an individual will default or not. Additionally, the dataset used is not location specific or doesn't include details of the region from which data is collected; had this information been included we could have a more achieved a more solid and grounded analysis specifying that the observations are universal or only applicable to selected geography. Presence of parameters like credit scores or FICO scores further would have aided in refining analysis and

observing predictions of likelihood of default and draw correlation between default possibility and credit scores.

- In the future scope of the project the amount that can be sanctioned to an individual without the risk of the individual defaulting can also be explore using regression model. This would be of key benefit to bank personnel as it would aid them in better decision making while sanctioning the loan amount.

10. References

1. Ali Abdullatif Ali Albastaki, ” Loan Default Prediction System” RIT Scholar Works, Feb 2022. https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12544&context=theses
2. Utkarsh Lal,” Mastering Loan Default Prediction: Tackling Imbalanced Datasets for Effective Risk Assessment”, Medium Blog, April 2023 https://medium.com/geekculture/mastering-loan-default-prediction-tackling-imbalanced-datasets-for-effective-risk-assessment-8e8dfb2084d0
3. Utkarsh,” Loan default Prediction “, Scaler Topics, May 2023 https://www.scaler.com/topics/data-science/loan-default-prediction/