| | |
|---|---|
| UECS 3213 DATA MINING | |
| Programme(s) | Bachelor of Science (Honours) Applied Mathematics with Computing |
| Title | Bank Loan Approval |
| Submission Date | 14th April 2024 |

| Group Information | |
|---|---|
| Student Name | Programme |
| Loo Kar Kuan (Leader) | AM |
| Chong Ke Ying | AS |
| Lim Yi Zhe | AS |

# Table of Contents

# 1.0 Data Understanding

## Introduction of Dataset and Source

The dataset of Bank Loan Approval contains details of all bank loan applications and the bank loan default information. This dataset serves as the basis for the model training, testing and validation to predict the default status of new applications. The dataset is stored in a CSV file format, and it consists of 18 features recording the information of each unique application. The features are summarized in Table 1.1. The size of this dataset is 255327 rows and 18 columns.

Table 1.1: Summary of the data attributes.

| No. | Attribute | Description | Values |
|---|---|---|---|
| 1 | LoanID | Primary key to represent a unique application. | A ten-characters long combination of numbers and alphabets. |
| 2 | Age | Age of the loan applicant. | 18-69 |
| 3 | Income | Annual income of the loan applicant. | 15000-149999 |
| 4 | LoanAmount | Loan amount requested by the applicant. | 5000-249999 |
| 5 | CreditScore | A number from 300 to 850 that rates an applicant's creditworthiness. | 300-849 |
| 6 | MonthsEmployed | Number of months the applicant is employed. | 0-119 |
| 7 | NumCreditLines | Number of credit cards. | 1, 2, 3, 4 |
| 8 | InterestRate | Interest rate for the applying loan. | 2-25 |
| 9 | LoanTerm | The number of months the applicant needs to completely pay off the loan. | 12, 24, 36, 48, 60 |
| 10 | DTIRatio | Debt-to-income ratio. | 0.1-0.9 |

| | | | |
|---|---|---|---|
| 11 | Education | Education level of the applicant. | High School, Bachelor’s, Master’s, PhD |
| 12 | EmploymentType | Employment status of the applicant. | Unemployed, Self-employed, Part-time, Full-time |
| 13 | MaritalStatus | Marital status of the applicant. | Single, Married, Divorced |
| 14 | HasMortgage | The yes or no indicator for which whether the applicant has a mortgage. | Yes, No |
| 15 | HasDependents | The yes or no indicator for which whether the applicant has any dependents. | Yes, No |
| 16 | HasCoSigner | The yes or no indicator for which whether the applicant has a co-signer. | Yes, No |
| 17 | LoanPurpose | The purpose of the loan application. | Education, Auto, Business, Home, Others |
| 18 | Default | The failure to make required interest or principal repayments on a loan. | 0, 1 |

### Background and Context

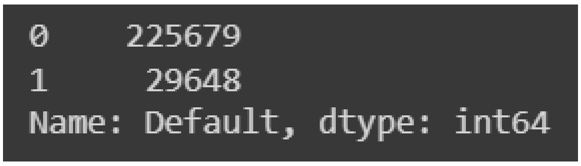
The dataset chosen is about determining the default status of the bank loan applicants based on the available information from their demographic data, their bank loan details and other supporting information such as their DTI ratio, number of employment months and number of credit lines. The core objective of this project is to evaluate the performance of various statistical-learning based and machine-learning based models in predicting the default status of the bank loan applicants based on their available information.

There are 18 features in total for this dataset; 9 of them are quantitative features while the rest are qualitative. The quantitative features include the bank applicants’ age, income, credit score, number of employment months, number of credit lines, loan amount, interest rate, loan term, and DTI ratio. DTI ratio stands for debt-to-income ratio which represents the ratio between the monthly debt payments and the gross monthly income of a person. It is a crucial information for the bank to measure the ability of a person to repay his or her debt on a monthly basis. Besides, the number of credit lines represents the number of credit cards owned by a person. If the person maxes out his or her credit cards, it may reflect that the person is at the edge of his or her finances. Furthermore, the qualitative features include loan ID, education level, employment status, marital status, presence of mortgage, presence of dependents, presence of co-signer, loan purpose and the default status of a bank loan applicant.

## Potential Issues and Limitation of the Dataset

### Class imbalance

The class distribution of the default status of the bank applicants is not equal. Based on Figure 1.1, there are 225679 observations with the default status of '0' and 29648 observations with the default status of '1'. This indicates that there are more applicants who did not default on the bank loan compared to those who did. In terms of percentage, there are 88.4% non-defaulted bank loan applicants and 11.6% defaulted bank loan applicants. In other words, the ratio of non-defaulted bank loan applicants to the defaulted bank loan applicants is approximately 13:1. Although the degree of imbalance is not extreme in this case, it may still affect the performance of the predictive models.



```
0    225679
1     29648
Name: Default, dtype: int64
```

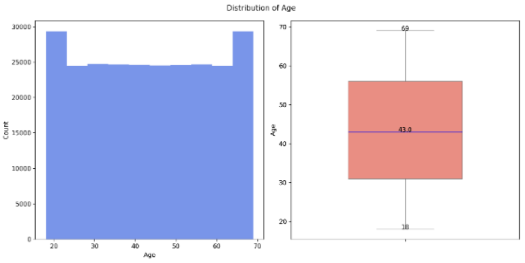
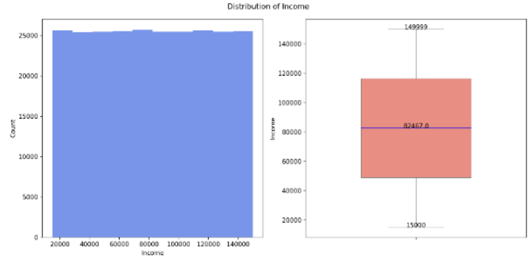
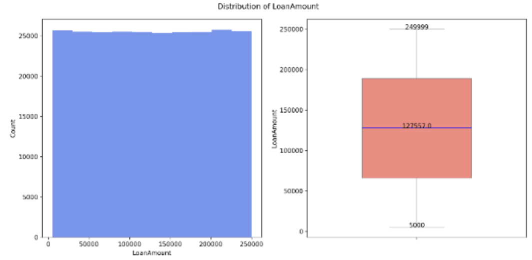
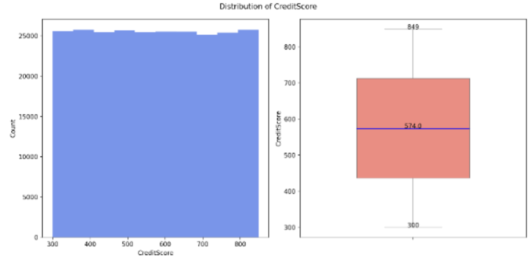
Figure 1.1: Class distribution of the default status.

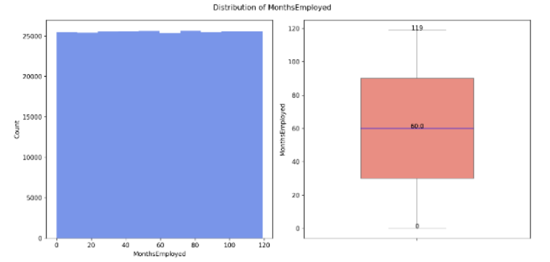
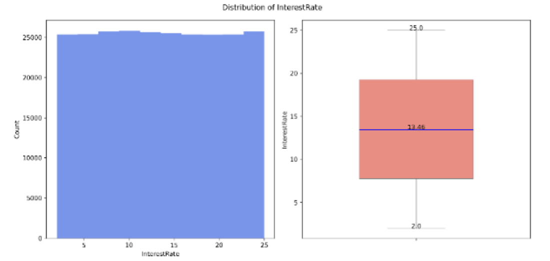
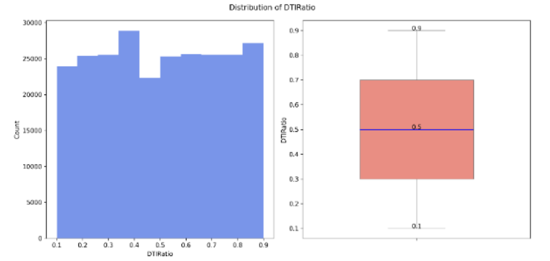
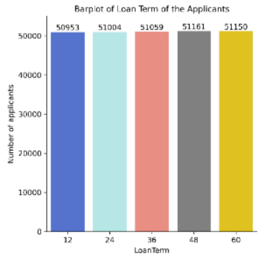
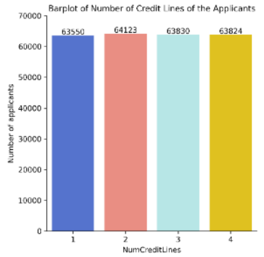
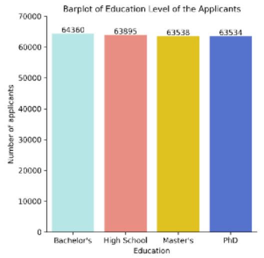
# 2.0 Data Description

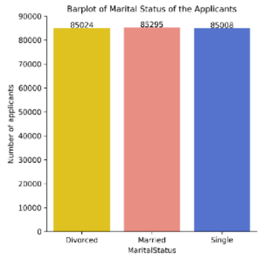
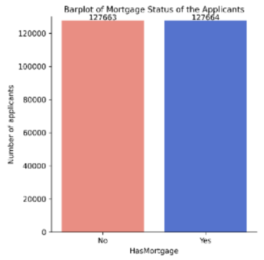
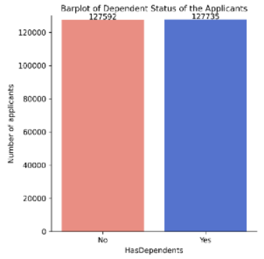
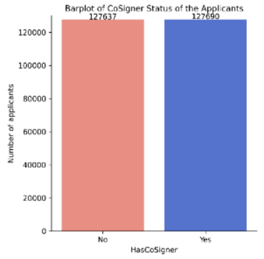
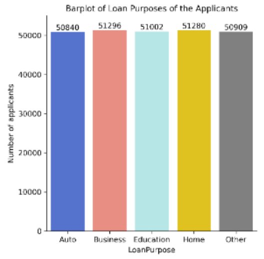
### Univariate Analysis

This section provides univariate analysis for 15 explanatory features and 1 target feature which include both qualitative and quantitative features. The feature “LoanID” is excluded in this section.

### Features

| Feature | Information | Outlier | Graphs |
|---|---|---|---|
| Age | Boundaries:<br>Upper:- 69<br>Lower:- 18<br>Median:- 43 | No |  |
| Income | Boundaries:<br>Upper:- 149,999<br>Lower:- 15,000<br>Median:- 82,467 | No |  |
| LoanAmount | Boundaries:<br>Upper:- 249,999<br>Lower:- 5,000<br>Median:- 127,557 | No |  |
| CreditScore | Boundaries:<br>Upper:- 849<br>Lower:- 300<br>Median:- 547 | No |  |

| | | | |
|---|---|---|---|
| MonthsEmployed | Boundaries:<br>Upper:- 119<br>Lower:- 0<br>Median:- 60 | No |  |
| InterestRate | Boundaries:<br>Upper:- 25<br>Lower:- 20<br>Median:- 13.46 | No |  |
| DTIRatio | Boundaries:<br>Upper:- 0.9<br>Lower:- 0.1<br>Median:- 0.5 | No |  |
| LoanTerm | Largest class: 48-month<br>Smallest class: 12-month | - |  |
| NumCreditLines | Largest class: 2<br>Smallest class: 1 | - |  |
| Education | Largest class: Bachelor’s<br>Smallest class: PhD | - |  |

|  |  |  |  |
|---|---|---|---|
| MaritalStatus | Largest class: Married<br>Smallest class: Divorced | - |  |
| HasMortgage | Larger class: Yes<br>Smaller class: No | - |  |
| HasDependents | Larger class: Yes<br>Smaller class: No | - |  |
| HasCoSigner | Larger class: Yes<br>Smaller class: No | - |  |
| LoanPurpose | Larger class: Business<br>Smaller class: Auto | - |  |

**Relationship between Feature and Target**

1. Age – Default

   The box plot below illustrates the relationship between age and default status. It indicates that the median age of non-defaulted bank loan applicants is greater than that of those who defaulted. Also, the distribution of defaulted bank loan applicants is slightly more skewed than that of non-defaulted applicants.

Overall, the distribution of defaulted bank loan applicants appears slightly right skewed, indicating a higher concentration of younger individuals among this group.

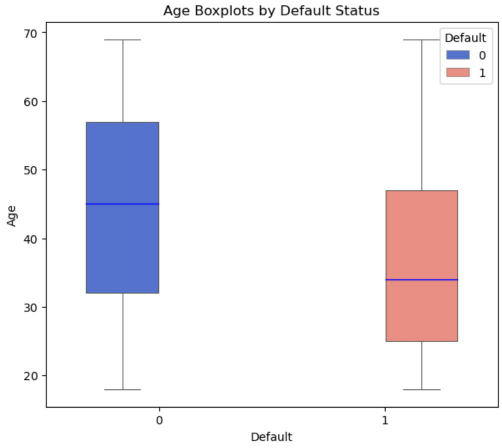


Figure 2.1: Age Boxplots by Default Status

2. Income – Default

The box plot below illustrates how income relates to default status, revealing that non-defaulted bank loan applicants typically have higher median incomes compared to defaulted applicants. Moreover, the distribution of defaulted applicants appears slightly more skewed than that of non-defaulted applicants.

Overall, the distribution of defaulted applicants is slightly right-skewed, indicating a higher concentration of lower-income individuals in this group.

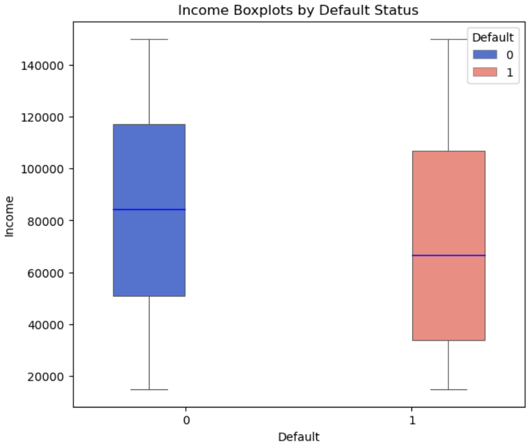


Figure 2.2: Income Boxplots by Default Status

3. Loan Amount – Default

The box plot below illustrates how loan amount relates to default status, showing that defaulted bank loan applicants tend to have higher median loan amounts compared to non-defaulted applicants. Additionally, the distribution of defaulted applicants is slightly skewed, while that of non-defaulted applicants appears more symmetric.

Overall, the distribution of defaulted applicants is slightly left-skewed, indicating a higher concentration of individuals with higher loan amounts in this group.

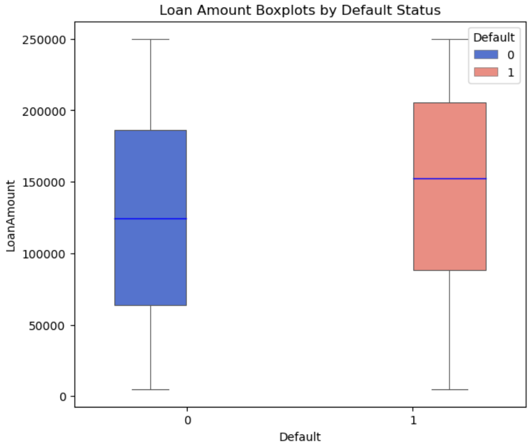


Figure 2.3: Loan Amount Boxplots by Default Status

4. Credit Score – Default

The box plot provided below illustrates the correlation between credit score and loan default status, revealing that non-defaulted bank loan applicants typically boast slightly higher median credit scores compared to their defaulted counterparts. Both the distributions of defaulted and non-defaulted applicants appear symmetric.

In summary, the median credit score of defaulted applicants tends to be slightly lower than the median of the credit score distribution among non-defaulted applicants.

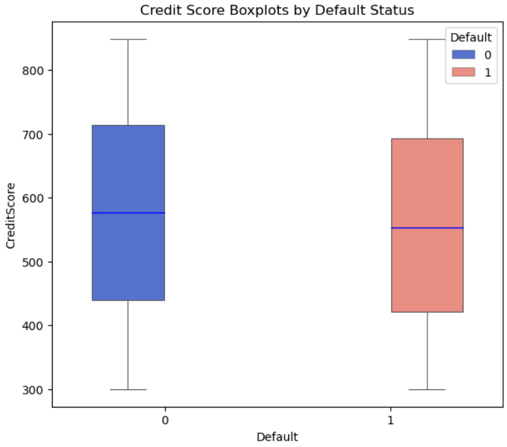


Figure 2.4: Credit Score Boxplots by Default Status

5. Months Employed – Default

The box plot below displays the comparison between months employed and loan default status, indicating that non-defaulted bank loan applicants generally have slightly higher median months employed than their defaulted counterparts. Both distributions of defaulted and non-defaulted applicants seem to be symmetric.

It can be concluded that the median months employed of defaulted applicants tends to be slightly lower than the median of the months employed distribution among non-defaulted applicants.

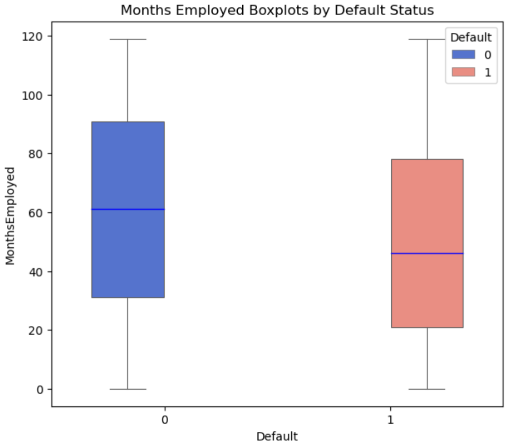


Figure 2.5: Months Employed Boxplots by Defaults Status

6. Num Credit Lines – Default

The bar chart below illustrates the distribution of NumCreditLines (number of credit cards). The categories include 1, 2, 3, and 4 for the number of credit cards.

It is evident that there are 56,863 non-defaulted bank loan applicants and 6,687 defaulted bank loan applicants in the category of having 1 credit card. Subsequently, 57,033 non-defaulted bank loan applicants and 7,090 defaulted bank loan applicants are found in the category of having 2 credit cards. In the category of having 3 credit cards, there are 56,220 non-defaulted bank loan applicants and 7,610 defaulted bank loan applicants. Lastly, in the category with the most credit cards, which is 4 credit cards, there are 55,563 non-defaulted bank loan applicants, while 8,261 are defaulted bank loan applicants.

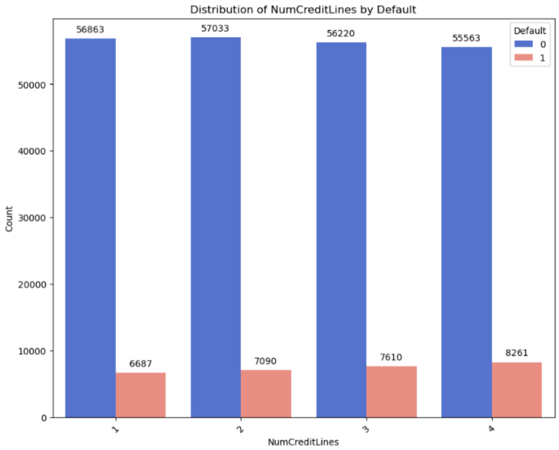


Figure 2.6: Distribution of NumCreditLines by Default

7. Interest Rate – Default

The box plot below displays the relationship between interest rate and loan default status, indicating that defaulted bank loan applicants generally have higher median interest rate than their non-defaulted counterparts. The distribution of non-defaulted applicants is symmetric while the distribution of defaulted applicants is slightly skewed.

It can be concluded that the distribution of defaulted bank loan applicants appears slightly left skewed, indicating a higher concentration of higher interest rate individuals among this group.

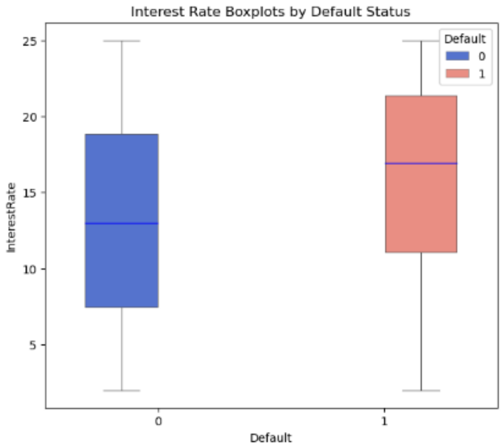


Figure 2.7: Interest Rate Boxplots by Default Status

8. Loan Term – Default

The bar chart below illustrates the distribution of loan term. The loan term is distributed into 5 parts which are 12, 24, 36, 48, and 60.

The data reveals a similar distribution among each loan term based on default status. However, it is evident that individuals with a 48-month loan term have the highest number of non-defaulted bank loan applicants, totalling 45,240, compared to 5,921 defaulted bank loan applicants. Conversely, the lowest number of non-defaulted bank loan applicants (45,033) is observed in the 12-month loan term category, while the second lowest number of defaulted bank loan applicants (5,950) is also seen in this category. Additionally, the highest number of defaulted bank loan applicants (5,983) is found in the 60-month loan term category, while simultaneously having the second highest number of non-defaulted bank loan applicants (45,167).

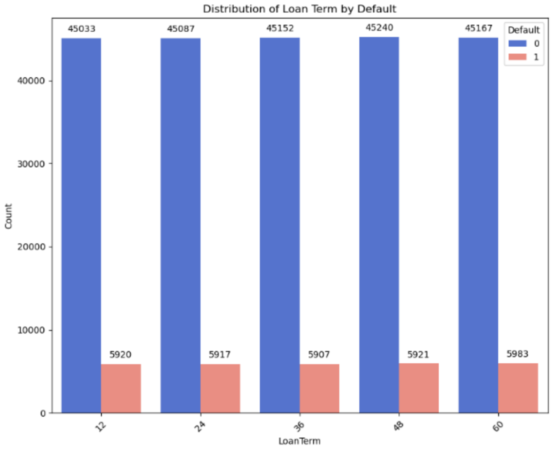


Figure 2.8: Distribution of Loan Status by Default

9. DTI Ratio – Default

The box plot below displays the relationship between DTI Ratio and loan default status indicating that defaulted bank loan applicants generally exhibit slightly higher median DTI ratios than their non-defaulted counterparts. Both distributions of defaulted and non-defaulted applicants appear symmetric.

Overall, it can be inferred that the median DTI ratio of non-defaulted applicants is lower than the median DTI ratio of defaulted applicants.

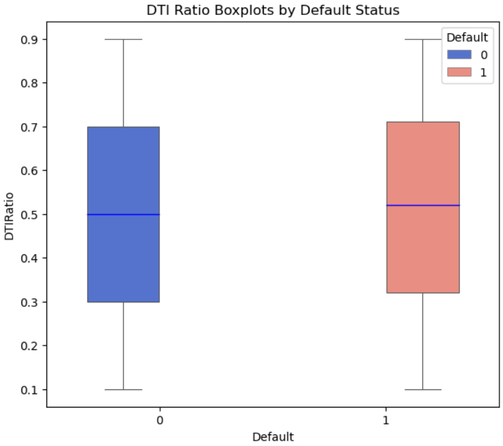


Figure 2.9: DTI Ratio Boxplots by Default Status

10. Education – Default

The bar chart below illustrates the distribution of education level. The categories of education level are High School, PhD, Master’s, and Bachelor’s.

The data reveals that among applicants with a high school education as their highest level, there are 55,668 non-defaulted bank loan applications compared to 8,227 defaulted ones. At the Bachelor’s level, we observe 56,572 non-defaulted bank loan applications and 7,788 defaulted ones. For the Master’s level, there are 56,331 non-defaulted bank loan applications and 6,907 defaulted ones. Remarkably, the PhD level exhibits the highest number of non-defaulted bank loan applications at 56,808, while the lowest number of defaulted bank loan applications stands at 6,726.

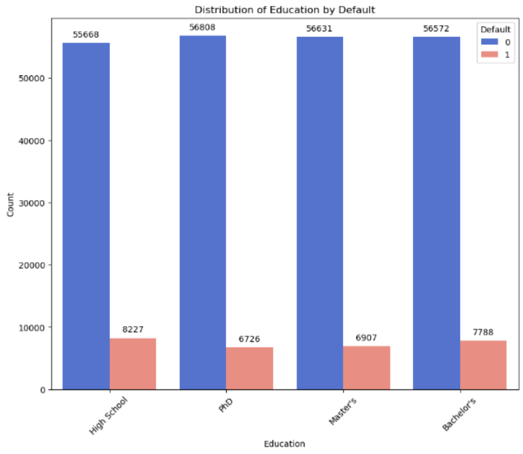


Figure 2.10: Distribution of Education by Default

11. Employment Type – Default

The bar chart below illustrates the distribution of employment type. The categories of employment type are unemployed, full-time, part-time, and self-employed.

In the provided dataset, significant variations are evident in bank loan applications based on employment status. Among unemployed applicants, non-defaulted bank loan applications total 55,171, contrasting with 8,648 defaulted ones. Notably, the highest number of non-defaulted bank loan applicants, reaching 57,626, is observed among full-time employees, accompanied by the lowest number of defaulted applicants, standing at 6,024, within the same category. Part-time employees follow with the second-highest defaulted applications at 7,675, along with 56,481 non-defaulted ones. Lastly, self-employed individuals contribute to 56,401 non-defaulted bank loan applicants and 7,301 defaulted bank loan applicants.

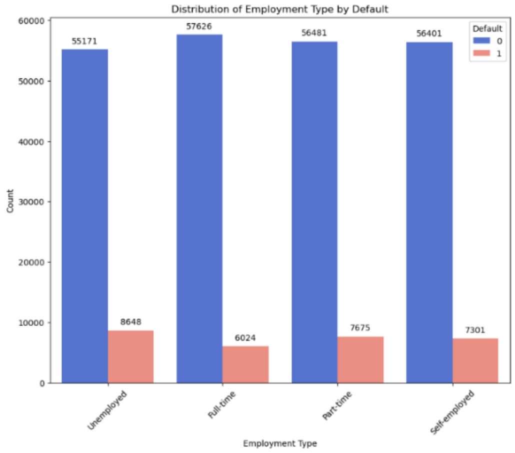


Figure 2.11: Distribution of Employment Type by Default

12. Marital Status – Default

The bar chart provided illustrates the distribution of defaulted and non-defaulted bank loan applicants based on marital status, categorized as married, divorced, and single.

The dataset reveals that the highest number of defaulted bank loan applicants falls under the marital status of divorced, with 74,371 non-defaulted bank loan applicants, the lowest among the other marital statuses. Among married applicants, there is the greatest number of non-defaulted bank loan applicants, totalling 76,426, while 8,869 applicants are defaulted. Lastly, among single applicants, 10,126 are defaulted, with 74,882 being non-defaulted.

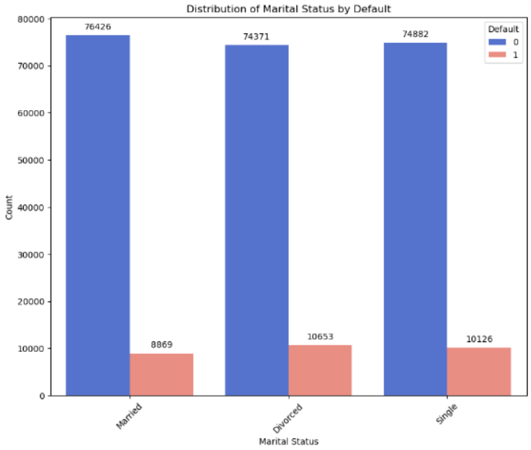


Figure 2.12: Distribution of Marital Status by Default

13. Has Mortgage – Default

The bar chart below shows the number of non-defaulted and defaulted loan applicants under 2 conditions, which are having mortgage and not having mortgage.

The dataset clearly indicates that there are 113,776 non-defaulted bank loan applicants and 13,888 defaulted bank loan applicants among those with mortgages. Conversely, there are 111,903 non-defaulted bank loan applicants and 15,760 applicants among those without mortgages. It clearly shows that applicants with mortgages have a greater number of non-defaulted bank loan applicants.

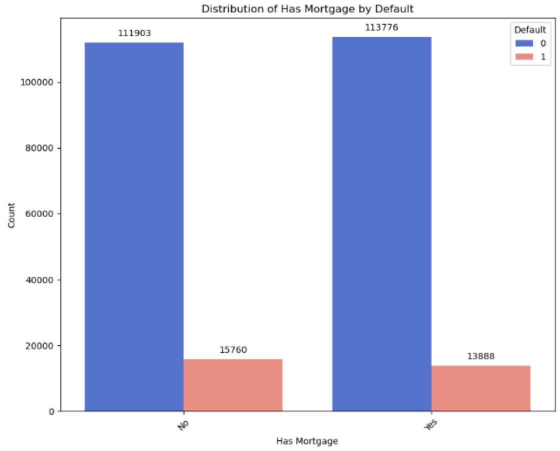


Figure 2.13: Distribution of Has Mortgage by Default

14. Has Dependents – Default

The bar chart below shows the number of non-defaulted and defaulted loan applicants under 2 conditions, which are having dependents and not having dependents.

The dataset reveals that when applicants have dependents, there are higher numbers of non-defaulted bank loan applicants (114,320) and fewer defaulted bank loan applicants (13,425). Conversely, applicants without dependents show fewer non-defaulted bank loan applicants (111,359) and a higher number of defaulted bank loan applicants (16,233).

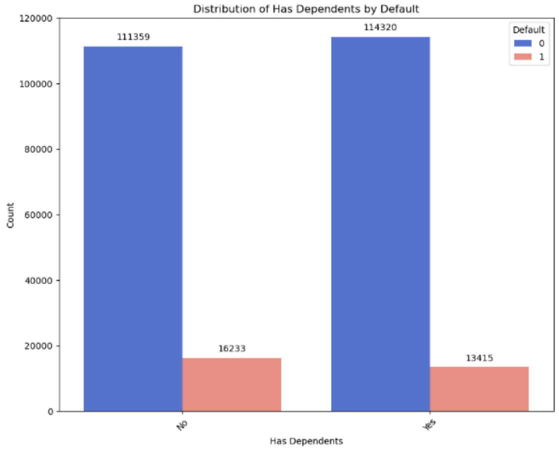


Figure 2.14: Distribution of Has Dependent by Default

### 15. Loan Purpose – Default

The bar chart below illustrates the distribution of loan purpose. The categories of loan purpose are education, auto, home, business, and other.

The dataset reveals that the highest number of non-default bank loan applicants (46,033) occurs when the loan purpose is home-related, while the lowest number of defaulted bank loan applicants (5,247) is associated with this purpose. Conversely, the lowest number of non-default bank loan applicants (44,800) is observed when the loan purpose is auto with defaulted bank loan applicants numbering 6,040. Additionally, the highest number of defaulted bank loan applicants (6,322) occurs when the loan purpose is business-related, while the second-highest number of non-defaulted bank loan applicants (44,974) is associated with this purpose.

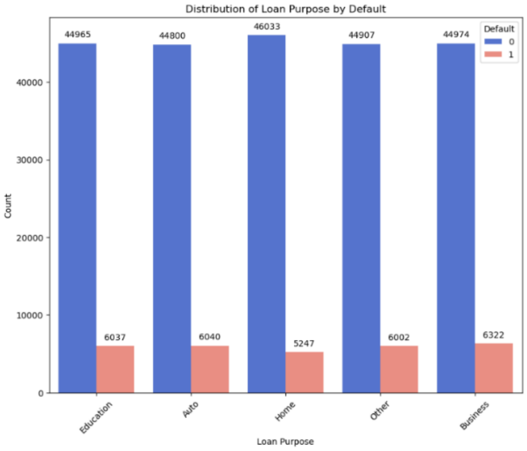


Figure 2.15: Distribution of Loan Purpose by Default

### 16. Has Co-Signer – Default

The bar chart below shows the number of non-defaulted and defaulted loan applicants under 2 conditions: (1) with co-signer or (2) without a co-signer.

The dataset shows that there were 111,217 approved loan applications and 16,420 defaulted loan applications when the applicants did not have a co-signer. Conversely, when applicants had at least one co-signer, the number of approved loan applications rise to 114,462, whereas the number of defaulted loan applications decreased to 13,228.

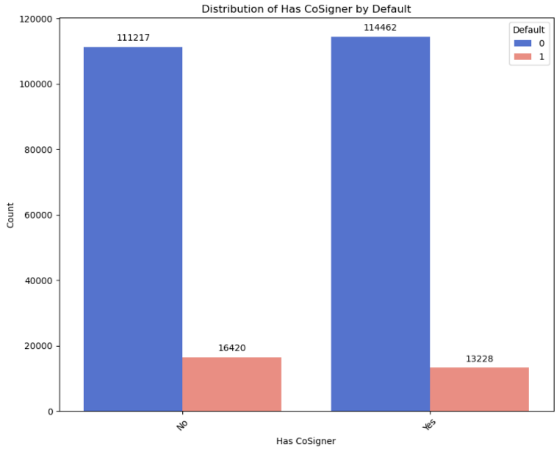


Figure 2.16: Distribution of Has CoSigner by Default

**Relationship between Feature and Feature**

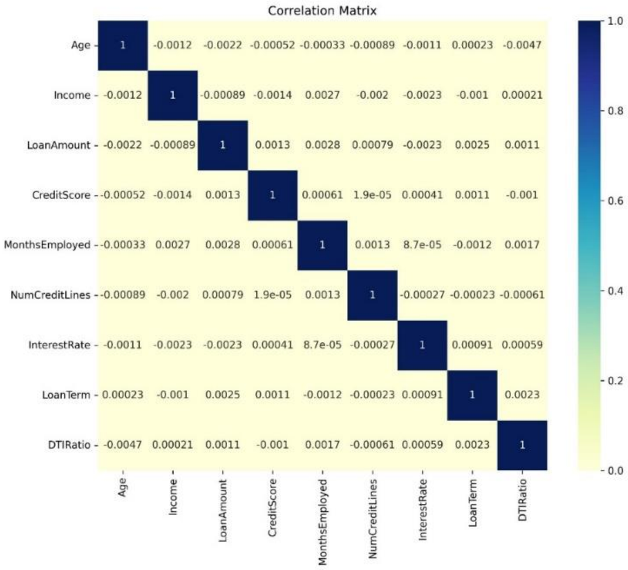
1. Correlation Matrix

   The correlation matrix below shows the correlation between the numerical features in the dataset. The positive and negative signs of the correlation coefficient represent the direction of the relationship. Hence, we use the absolute value to measure the strength of the linear relationship between the features.

   The matrix reveals that the credit score and number of credit lines have the correlation coefficient with smallest absolute value, which is $1.9X10^{-5}$. Conversely, the correlation coefficient between age and DTI ratio is $-0.0047$, which has the greatest absolute value of 0.0047.

   As the above discussion suggests, there is almost no linear relationship between the variables as the correlation coefficient with the greatest absolute value is only $-0.0047$ and very close to 0. This indicates that changes in one numerical feature have very little influence on the pattern of other features, since they contain a low level of same information. In other words, they are not correlated to each other and therefore have distinct effects on predicting the dependent variable.

In such instances, we conclude that the problem of collinearity or multicollinearity does not exist in this dataset. All features have their own different impacts on predicting the dependent variables instead of other features.

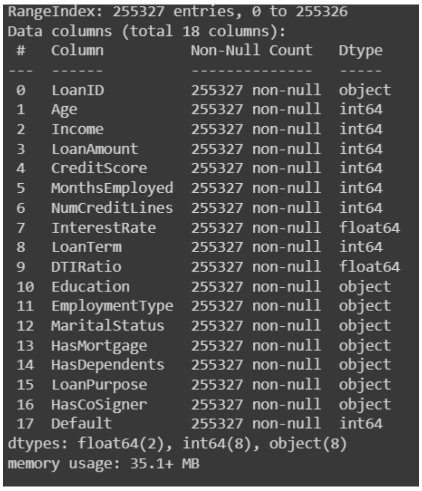


# 3.0 Data Preprocessing

## Data Cleaning

### Check for Missing Value

The dataset is complete without any missing values, eliminating the possibility of bias or inaccurate results that may arise from mishandling incomplete data.

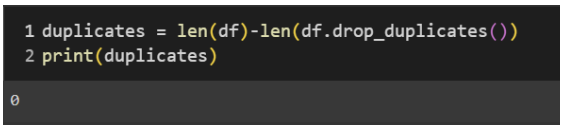


```
RangeIndex: 255327 entries, 0 to 255326
Data columns (total 18 columns):
 #   Column          Non-Null Count   Dtype  
---  ------          --------------   -----  
 0   LoanID          255327 non-null  object 
 1   Age             255327 non-null  int64  
 2   Income          255327 non-null  int64  
 3   LoanAmount      255327 non-null  int64  
 4   CreditScore     255327 non-null  int64  
 5   MonthsEmployed  255327 non-null  int64  
 6   NumCreditLines  255327 non-null  int64  
 7   InterestRate    255327 non-null  float64
 8   LoanTerm        255327 non-null  int64  
 9   DTIRatio        255327 non-null  float64
 10  Education       255327 non-null  object 
 11  EmploymentType  255327 non-null  object 
 12  MaritalStatus   255327 non-null  object 
 13  HasMortgage     255327 non-null  object 
 14  HasDependents   255327 non-null  object 
 15  LoanPurpose     255327 non-null  object 
 16  HasCoSigner     255327 non-null  object 
 17  Default         255327 non-null  int64  
dtypes: float64(2), int64(8), object(8)
memory usage: 35.1+ MB
```

Figure 3.1: Examining null values in the dataset.

### Check for Duplicated Values

As shown in Figure 3.2, there is no duplicated values found in this dataset. Hence, the dataset is free from the redundancy issue.



```
duplicates = len(df)-len(df.drop_duplicates())
print(duplicates)
```

```
0
```

Figure 3.2: Examining duplicated values in the dataset.

### Check for Misspelled Categories

The qualitative features in the dataset have been thoroughly inspected by examining the category names displayed in the bar chart generated in Chapter 2. No misspelled category names were detected, ensuring the reliability of all features for further analysis.

Detection of Outliers

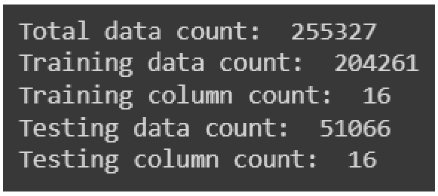
In Chapter 2, outliers in each numerical feature were investigated using the boxplot method. Data points falling outside the upper and lower whiskers of the boxplot were considered outliers. No outliers were found in any of the features, indicating a clean dataset free from extreme values.

Deletion of irrelevant feature

The feature "LoanID" has been identified as irrelevant to the predictive model development process, as it solely represents the identity number of the applicant without contributing analytical value. Consequently, this feature has been removed from the dataset.

### Data Splitting

The dataset is split into 80% of training dataset and 20% of testing dataset by using the stratified train-test split to preserve the same proportions of target classes. Figure 3.4 depicts the number of observations in each of the datasets after splitting the original dataset.



```
Total data count:  255327
Training data count:  204261
Training column count:  16
Testing data count:  51066
Testing column count:  16
```

Figure 3.3: Information about the training and testing dataset.

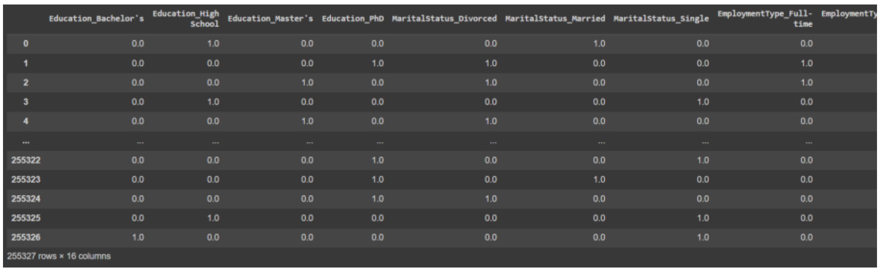
### Data Transformation

Standardization

In this project, the qualitative features are standardized using the z-score method using the StandardScaler function.

## One-Hot Encoding

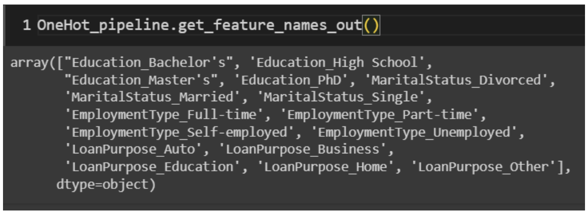
Four qualitative features, namely Education, MaritalStatus, EmploymentType, and LoanPurpose, have been transformed using the one-hot encoder. Figure 3.5 presents a portion of the transformed dataset, while Figure 3.6 displays a list of the transformed feature names, demonstrating the effectiveness of the one-hot encoding approach.



| | Education_Bachelor's | Education_High School | Education_Master's | Education_PhD | MaritalStatus_Divorced | MaritalStatus_Married | MaritalStatus_Single | EmploymentType_Full-time | EmploymentTy |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | |
| 1 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | |
| 2 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | |
| 3 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | |
| 4 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 255322 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | |
| 255323 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | |
| 255324 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | |
| 255325 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | |
| 255326 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | |

255327 rows × 16 columns

Figure 3.4: Dataset transformed by the one-hot encoder.



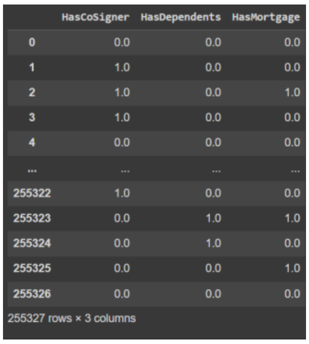
```
OneHot_pipeline.get_feature_names_out()
```

```
array(["Education_Bachelor's", 'Education_High School',
       "Education_Master's", 'Education_PhD', 'MaritalStatus_Divorced',
       'MaritalStatus_Married', 'MaritalStatus_Single',
       'EmploymentType_Full-time', 'EmploymentType_Part-time',
       'EmploymentType_Self-employed', 'EmploymentType_Unemployed',
       'LoanPurpose_Auto', 'LoanPurpose_Business',
       'LoanPurpose_Education', 'LoanPurpose_Home', 'LoanPurpose_Other'],
      dtype=object)
```

Figure 3.5: Output feature names of the one-hot encoder.

## Ordinal Encoding

The ordinal encoder has been utilized to replace the "Yes" and "No" classes in the "HasDependents", "HasCoSigner", and "HasMortgage" features with 1 and 0, respectively. This transformation was achieved using the OrdinalEncoder function. Figure 3.7 displays the dataset with the transformed features following the application of ordinal encoding.



| | HasCoSigner | HasDependents | HasMortgage |
|---|---|---|---|
| 0 | 0.0 | 0.0 | 0.0 |
| 1 | 1.0 | 0.0 | 0.0 |
| 2 | 1.0 | 0.0 | 1.0 |
| 3 | 1.0 | 0.0 | 0.0 |
| 4 | 0.0 | 0.0 | 0.0 |
| ... | ... | ... | ... |
| 255322 | 1.0 | 0.0 | 0.0 |
| 255323 | 0.0 | 1.0 | 1.0 |
| 255324 | 0.0 | 1.0 | 0.0 |
| 255325 | 0.0 | 0.0 | 1.0 |
| 255326 | 0.0 | 0.0 | 0.0 |

255327 rows × 3 columns

Figure 3.6: Dataset transformed by the ordinal encoder.

Preprocessing pipeline

After configuring the above scaler and encoders, they are combined into a single preprocessing pipeline using the ColumnTransformer function to provide code reusability. Figure 3.8 shows the pipeline architecture.

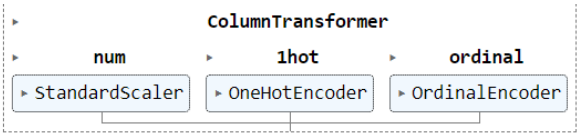


Figure 3.7: Creation of preprocessing pipeline.

### Data Oversampling

In this project, data imbalance issue is found in the target feature, which is "Default". According to Figure 3.8, there is 88.4% of default case and 11.6% of non-default case in the training dataset. The data might be slightly skewed to the default class since it is the bigger class. To address this issue, oversampling is employed to improve model performance and ensure better generalization across all classes.

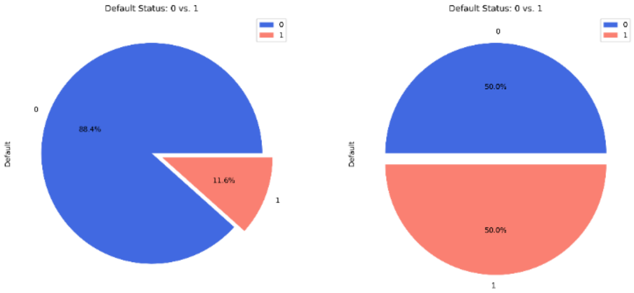


Figure 3.8: Default status before and after oversampling.

Considering the level of skewness is not extreme, the decision of employing the oversampling method into the model is left to be justified in the later part of this project. The performance of a model that employs the oversampling method is compared with the performance of a model that utilizes only the original data. If the performance of the former model is better than the latter model, then the oversampling method shall be employed.

# 4.0 Model Interpretation

This project assessed the efficiency and effectiveness of seven models in predicting the default status of the bank loan applicants. The models include logistic regression, Random Forest (RF), Naïve Bayes, Support Vector Machine (SVC), K-Nearest Neighbours (KNN), Gradient Boosting Machine (GBM), decision trees, and multilayer perceptron (MLP). Additionally, it aimed to explore the feasibility of the random search algorithm and Principal Component Analysis (PCA). Four evaluation metrics were employed to assess model performance under various settings, namely accuracy, precision, sensitivity and F1-score.

### The Performance of the Models

| Model | Accuracy | | Precision | | Sensitivity | | F1-Score | |
|---|---|---|---|---|---|---|---|---|
| | SMOTE | w/o SMOTE | SMOTE | w/o SMOTE | SMOTE | w/o SMOTE | SMOTE | w/o SMOTE |
| Logistic | 0.680414 | 0.885795 | 0.220278 | 0.642442 | 0.689882 | 0.037268 | 0.333932 | 0.070449 |
| RF | 0.782830 | 0.884698 | 0.262955 | 0.808824 | 0.482631 | 0.009275 | 0.340431 | 0.018339 |
| Naïve Bayes | 0.690440 | 0.885462 | 0.220233 | 0.600998 | 0.655649 | 0.040641 | 0.329715 | 0.076133 |
| KNN | 0.655242 | 0.874026 | 0.175149 | 0.321505 | 0.529848 | 0.076391 | 0.263270 | 0.123450 |
| GBM | 0.551796 | 0.886402 | 0.167281 | 0.638116 | 0.718887 | 0.050253 | 0.271408 | 0.093169 |
| Decision Tree | 0.722751 | 0.801884 | 0.176114 | 0.198386 | 0.377234 | 0.232209 | 0.240125 | 0.213969 |
| MLP | 0.876004 | 0.885521 | 0.409050 | 0.549528 | 0.152445 | 0.078583 | 0.222113 | 0.137504 |

Table 4.1: Performance of default models.

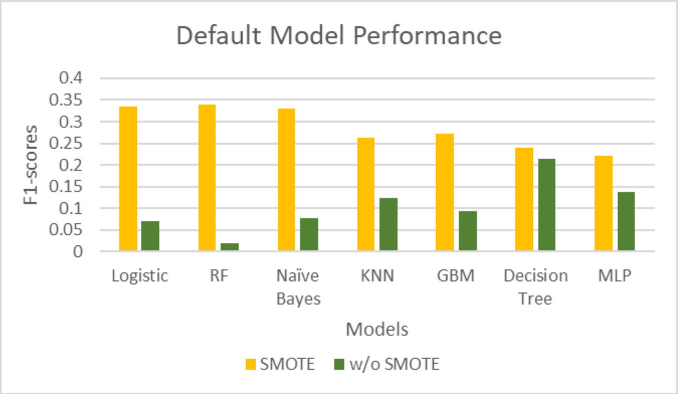


Figure 4.1: F1-scores of default models with and without SMOTE oversampling

Based on Table 4.1, models trained with oversampled data consistently outperformed those trained with the original imbalanced data using default hyperparameters. Consequently, the SMOTE oversampling method was chosen for model development. Hyperparameters were tuned using RandomizedSearchCV, with results detailed in Table 4.2. Due to computational constraints, the number of iterations in RandomizedSearchCV varied based on search process time. For logistic regression, Naïve Bayes, and decision trees, a maximum of 50 iterations was set. For models with longer computational requirements like RF and GBM, iterations were limited to 10 to reduce computational time. Unfortunately, the random search for GBM still failed to complete after several hours. Hence, further analysis on the GBM's performance will be based on its default hyperparameter setting. Besides, the maximum search iteration for MLP is set as 50 but better model configuration hyperparameter was found when the search iteration was limited to 10. Furthermore, the KNN model's maximum search iterations are capped at 40 due to system errors encountered with 50 iterations. The SVC model was eliminated from the project in this stage as the default SVC model took more than 2 hours and did not reach its completion eventually. Additionally, the random search failed to perform search on the SVC's hyperparameters after several attempts which took more than three hours each time. Figure 4.1 shows the performance of the default models with and without SMOTE oversampling in terms of their F1-scores.

| Model | Tuned Parameters | Accuracy | Precision | Sensitivity | F1-Score |
|---|---|---|---|---|---|
| Logistic | C: 4.16339e-05<br>penalty: 'l2'<br>solver: 'liblinear' | 0.673031 | 0.217298 | 0.697808 | 0.331398 |
| RF | max_depth: 18,<br>bootstrap: True | 0.835135 | 0.304784 | 0.327656 | 0.315807 |
| Naïve Bayes | var_smoothing: 0.1 | 0.884659 | 0.750000 | 0.010118 | 0.065248 |
| KNN | n_neighbors: 9,<br>weights: 'distance' | 0.616751 | 0.174019 | 0.613997 | 0.271180 |
| Decision Tree | max_depth: 10,<br>min_samples_leaf: 4,<br>min_samples_split: 15 | 0.800611 | 0.260586 | 0.390219 | 0.312492 |
| MLP | hidden_layer_sizes: 10, | 0.768809 | 0.251679 | 0.496228 | 0.333973 |

| | | | | | |
|---|---|---|---|---|---|
| | alpha: 0.01 | | | | |

Table 4.2: Optimized models' performance with SMOTE oversampling.

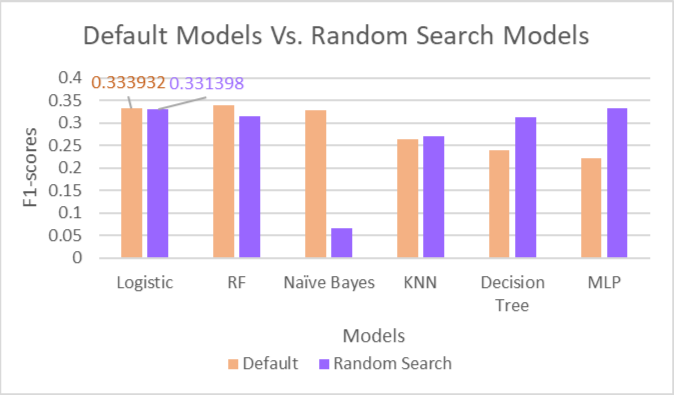


Figure 4.2: Comparison of F1-scores between the default models and the tuned models.

The evaluation metric scores of the tuned models serve as indicators of the RandomizedSearchCV function's effectiveness in optimizing model performance. F1-score is the primary metric used to assess the feasibility of RandomizedSearchCV in model optimization, as reflected in the scores. The hyperparameter settings for the models are summarized in Table 4.3 based on these scores. Using these optimized settings, model performance is further analysed through stratified k-folds validation, with results presented in Table 4.4. However, the stratified k-fold of the RF model consumed unexpected long time and did not complete at the end. Hence, the result for it was excluded from Table 4.4. Apart from that, the average scores from stratified k-folds validation closely align with those in Tables 4.1 and 4.2, demonstrating consistency based on each model's hyperparameter settings. The F1-scores of the default models and the tuned models were compared and depicted in Figure 4.2.

| Model | Hyperparameter Setting |
|---|---|
| Logistic | Default |
| RF | Default |
| Naïve Bayes | Default |
| KNN | Tuned |
| GBM | Default |
| Decision Tree | Tuned |
| MLP | Tuned |

Table 4.3: Hyperparameter settings of the models.

| Model | Accuracy | Precision | Sensitivity | F1-Score |
|---|---|---|---|---|
| Logistic | 0.682442 | 0.219926 | 0.681128 | 0.332494 |
| RF | - | - | - | - |
| Naïve Bayes | 0.693129 | 0.219936 | 0.645048 | 0.328028 |
| KNN | 0.613406 | 0.171089 | 0.605808 | 0.266822 |
| GBM | 0.885786 | 0.556246 | 0.081355 | 0.141936 |
| Decision Tree | 0.789544 | 0.251977 | 0.412676 | 0.312816 |
| MLP | 0.752094 | 0.247520 | 0.551979 | 0.340883 |

Table 4.4: Average performance of models in stratified k-fold validation.

| Model | Accuracy | Precision | Sensitivity | F1-Score |
|---|---|---|---|---|
| | PCA | PCA | PCA | PCA |
| Logistic | 0.574453 | 0.083021 | 0.265261 | 0.126462 |
| RF | 0.422453 | 0.107218 | 0.542327 | 0.179039 |
| Naïve Bayes | 0.405123 | 0.083282 | 0.411973 | 0.138555 |
| KNN | 0.654780 | 0.180303 | 0.556324 | 0.272341 |
| GBM | 0.744625 | 0.246488 | 0.582968 | 0.346480 |
| Decision Tree | 0.743410 | 0.218640 | 0.469983 | 0.298442 |
| MLP | 0.737301 | 0.244138 | 0.602192 | 0.347424 |

Table 4.5: The performance of models with the use of PCA.

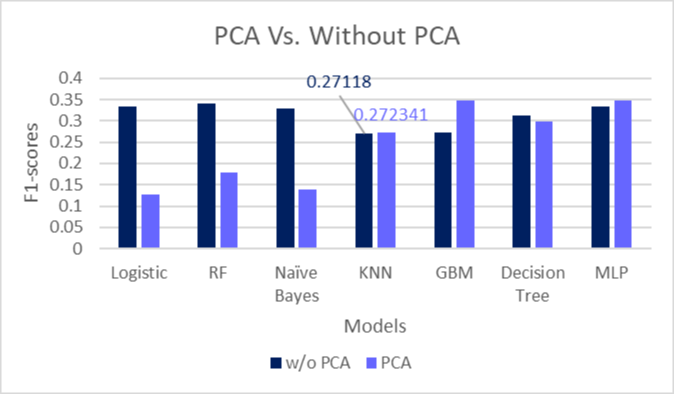


Figure 4.3: Comparison of F1-scores between the models with and without PCA.

According to Table 4.1, 4.2 and 4.5, the use of random search and PCA with the models were determined. The selection of 'n_components' for PCA was determined through trial and error, aiming to achieve the best F1-score for each model. After experimenting with different 'n_components', the optimal setting was identified, resulting in the highest score. The finalized models with their corresponding results were listed in Table 4.6 for a clearer comparison. Figure 4.1 illustrates a bar chart for comparing the F1-scores of the models based on their corresponding model settings. The highest F1-score was achieved by the MLP model at 0.347424. Another alternative model for this project would be the logistic regression which also achieved a high F1-score at 0.333932. The models' performance with and without the implementation of PCA are presented in Figure 4.3 in terms of F1-scores.

| **Model** | **Random Search** | **PCA** | **Accuracy** | **Precision** | **Sensitivity** | **F1-Score** |
|---|---|---|---|---|---|---|
| Logistic | No | No | 0.680414 | 0.220278 | 0.689882 | 0.333932 |
| RF | No | No | 0.782830 | 0.262955 | 0.482631 | 0.340431 |
| Naïve Bayes | No | No | 0.690440 | 0.220233 | 0.655649 | 0.329715 |
| KNN | Yes | Yes | 0.654780 | 0.180303 | 0.556324 | 0.272341 |
| GBM | No | Yes | 0.744625 | 0.246488 | 0.582968 | 0.346480 |
| Decision Tree | Yes | No | 0.800611 | 0.260586 | 0.390219 | 0.312492 |
| MLP | Yes | Yes | 0.737301 | 0.244138 | 0.602192 | 0.347424 |

Table 4.6: The evaluation metric scores of the finalized models.

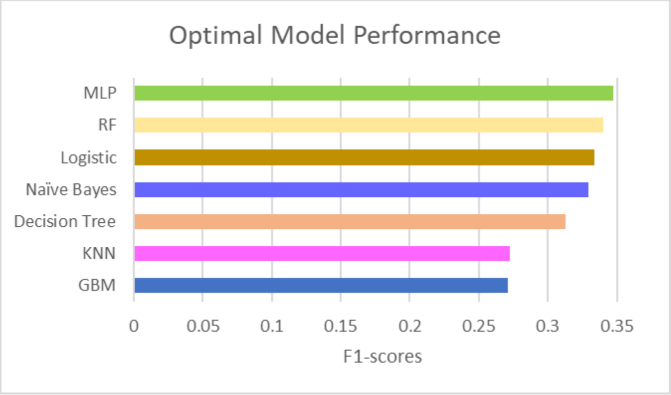


Figure 4.4: The F1-score performance of models with their optimal hyperparameter setting.

Table 4.6 summarises the optimal hyperparameter setting for each model and their corresponding evaluation scores. The F1-scores of the models are compared in Figure 4.4 and the MLP secured the first place in terms of the model's performance in this project with a F1-score of 0.347424 and Figure 4.5 illustrates the confusion matrix of the MLP.

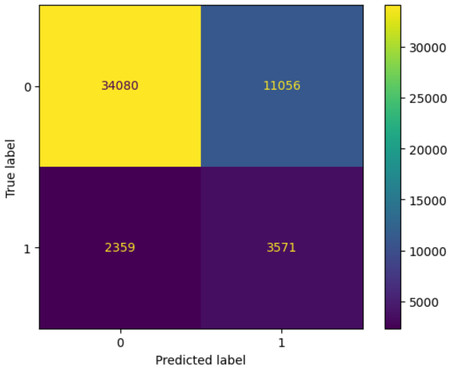


Figure 4.5: Confusion matrix of the MLP.

**Importance of Features**

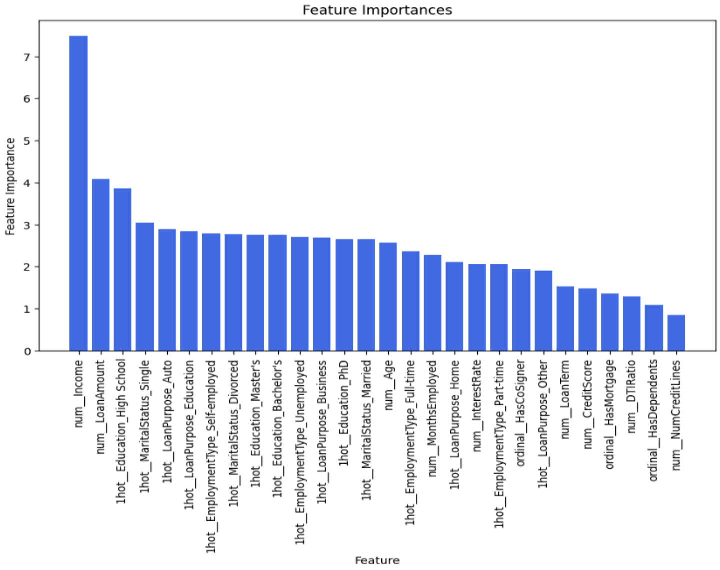


Figure 4.6: Importance of Features for MLP

According to Table 4.6, the MLP model yields the highest F1-score, making it the best-performing model. Therefore, feature importance is determined based on the operations of MLP. As illustrated in Figure 4.6, income serves as the most influential feature in determining the output, surpassing all others with a scale exceeding 7. Following behind, we have loan amount and high school education level with scales of 4.079620 and 3.855391 respectively.

Subsequently, the following 19 features exhibit scales ranging between 3.1 and 1.9. Their scales gradually decrease, so they are grouped and considered to have a moderate impact on the results. The last 6 features have an obvious gap with the moderate group as their scales range from 1.60 to 0.8, indicating a lower impact on the results.

### How the features affecting the model's results.

Based on Figure 4.1, there is a significant gap in scale difference between the top 3 important features and other features.

1. Income

   Lenders typically evaluate an applicant's income to assess their ability to repay the loan. For instance, applicants with higher incomes often have lower DTI ratios, instilling confidence in lenders when considering loan approval. Thus, individuals with higher income are more likely to get their loan approval.

2. Loan Amount

   Lenders also assess the risk associated with a loan by considering the loan amount relative to other factors like applicant's income and creditworthiness. Generally, a higher loan amount may indicate a greater financial risk for the lender, especially when the applicant's credit profile does not support the repayment of a large loan.

3. High School Education Level

   Some lenders may require applicants to meet a minimum education level as part of their eligibility criteria. A high school diploma may fulfil this requirement by indicating that the individual has attained a basic level of education and literacy.

### New Applicants' Prediction

| LoanID | Default |
|---|---|
| A01 | 0 |
| A02 | 0 |
| A03 | 1 |
| A04 | 0 |
| A05 | 0 |

| LoanID | Default |
|---|---|
| B01 | 0 |
| B02 | 0 |
| B03 | 0 |
| B04 | 0 |
| B05 | 0 |

| | |
|---|---|
| A06 | 1 |
| A07 | 0 |
| A08 | 0 |
| A09 | 0 |
| A10 | 0 |

Table 4.7

| | |
|---|---|
| B06 | 1 |
| B07 | 1 |
| B08 | 0 |
| B09 | 1 |
| B10 | 0 |

Table 4.8

Table 4.7 and 4.8 illustrate the prediction outcomes of the MLP model for the 20 new applicants. Due to space constraints, these tables exclusively display the “LoanID” and “Default” columns. However, the remaining features are available in the “((GAssign) NewApplicants.csv” file for reference. The result indicates that only 5 loan applications have been defaulted.

# 5.0 Conclusion

In this study of predicting the approval of bank loan, selecting the F1-score as the evaluation metric is the best choice for several reasons. Firstly, in many real-world scenarios, the dataset can be imbalanced, meaning one class (e.g., approved loans) may dominate the other (e.g., defaulted loans). In such cases, accuracy alone can be misleading because a model can achieve high accuracy by simply predicting the majority class.

Moreover, misclassifying loan approvals or rejections can have significant consequences for both the bank and the loan applicants. False negatives (approving a loan that should be rejected) can lead to financial losses for the bank, while false positives (rejecting a loan that should be approved) can result in missed business opportunities and customer dissatisfaction. The F1-score considers both types of errors, making it a suitable metric for evaluating the overall performance of the model in minimizing these costs.

We aim to optimize each model by evaluating the effectiveness of SMOTE oversampling method, RandomizedSearchCV function and PCA. Among the best performance of each model, MLP achieved the highest F1-score of 34.7424%, and an accuracy score of 73.7301% which is the median among these 7 models. Therefore, it exhibits the best performance in balancing the trade-off between all the metrics.

In short, this analysis considers all the 16 features that affect a loan application, subjecting them to pre-processing for suitability in analysis. Subsequently, each model is trained based on these processed data, and their results are compared. The selection of the best model prioritizes the impact of loan applications outcome on the bank, rather than solely focusing on accuracy. In such instances, the MLP model (neural network) emerges as the best-performing model in this analysis.