Team 3

DATS 6103: Summary Report

Professor Ning Rui

April 28, 2022

Data Mining Final Project: Leads Conversion Data Analysis

Introduction

The rapidly evolving digital environment has had a profound impact on the banking industry, presenting both opportunities and challenges. Banks are increasingly focusing on improving their customer acquisition strategies through digital channels such as search engines, display advertising, email campaigns and affiliate partners. However, these channels often generate different pools of leads with different conversion rates. By identifying high-converting prospect segments, banks can effectively target prospects and optimize marketing strategies.

This study aims to address this issue by analyzing a dataset containing customer credit histories and basic information. The dataset obtained from Kaggle has 30,038 rows and 22 columns, providing a rich data source for analysis. Various machine learning models including exploratory data analysis (EDA), simple linear regression (SLR), multiple linear regression (MLR), T-test, logistic regression, SMOTE, decision tree or random to extract valuable insights from this data.

SMART Questions

1. What is the best predictor or combination of predictors for loan approval?
2. Is there any correlation between the monthly income and the interest rate for the applicant's loan? What about loan amount and interest rate? Loan period and interest rate?
3. Are there any variables that indicate how much an applicant will be approved for?

Literature Review

1. Digital Channels and Customer Acquisition

Digital channels have played an increasingly important role in customer acquisition, especially for financial institutions. Phd. Anber(2022) explained how to conduct digital marketing actions to improve customer loyalty. In the age of digitization, organizations should think of advanced strategies to increase their competitiveness and market share by employing the potential of digital content and enhancing their digital capabilities. Based on his research result, Digital content marketing and digital Marketing ability has a significant impact on the success of digital marketing, based on the views of Jordanian online restaurant customers.

2. Loan Prediction Model

There are several factors to predict whether one borrower would default his/her loan. In Lifeng Zhou and Hong Wong's research(2012), they try to make loan default predictions on imbalanced data sets with an improved random forests approach which employs weighted majority votes in tree aggregation. The weights assigned to each tree in the forest are based on OOB(out-of-bag) errors which are easy to obtain during the forest construction process. From their result, they tried to tune sample size and SMOTE sampling method to compare SVM, KNN and C.4.5. Experiments show that the weighted majority approach in tree aggregation improves random forest performance in terms of overall accuracy and balanced accuracy.

3. Machine Learning Techniques in Banking

Machine learning techniques are also frequently used in the prediction of detecting the default behaviors of their customers. In this paper, T. Aditya Sai Srinivas, Somula Ramasubbareddy & K. Govinda(2022) offered a new machine learning approach to fix this problem using KNN, decision tree, SVM and logistic regression and demonstrating that these models could effectively segment and target potential customers.

Finally, the literature highlights the growing importance of digital channels in customer acquisition for the banking sector, the development of loan prediction models, and the successful application of machine learning techniques to improve conversion rates and decision-making. Our study builds on these findings by examining a comprehensive set of machine learning models to identify factors that contribute to improved conversion rates and to inform digital customer acquisition strategies in the banking industry. It is intended to build upon**.**

Data Description

The dataset used in this study comprises various anonymized features related to loan applicants and their financial background. The data includes 22 variables, which can be categorized into demographic information, financial details, and loan-related attributes. The following is a brief description of each variable in the dataset:

Gender: The gender of the loan applicant (Male/Female).

DOB: The date of birth of the loan applicant.

Lead_Creation_Date: The date on which the loan lead was created.

City_Code: An anonymized code representing the city of the applicant.

City_Category: An anonymized categorical variable representing the city feature.

Employer_Code: An anonymized code representing the employer of the applicant.

Employer_Category1: An anonymized categorical variable representing the first employer feature.

Employer_Category2: An anonymized categorical variable representing the second employer feature.

Monthly_Income: The monthly income of the applicant in US dollars.

Customer_Existing_Primary_Bank_Code: An anonymized code representing the applicant's primary bank.

Primary_Bank_Type: An anonymized categorical variable representing the type of primary bank.

Contacted: A binary variable indicating whether the applicant's contact information has been verified (Y/N).

Source: A categorical variable representing the source of the loan lead.

Source_Category: A categorical variable representing the type of lead source.

Existing_EMI: The equated monthly installment (EMI) of the applicant's existing loans in US dollars.

Loan_Amount: The requested loan amount in US dollars.

Loan_Period: The requested loan period in years.

Interest_Rate: The interest rate of the submitted loan amount.

EMI: The equated monthly installment (EMI) of the requested loan amount in US dollars.

Var1: An anonymized categorical variable with multiple levels.

Approved: The target variable indicating whether a loan is approved (1) or not (0).

This dataset provides a comprehensive set of variables, which enables us to explore various factors that may influence the conversion rate of leads and their loan approval probability. By analyzing this data, we aim to identify the key features that contribute to higher conversion rates and inform digital customer acquisition strategies in the banking industry.

Data Cleaning

Prior to the Exploratory Data Analysis, several data cleaning steps were taken. This included checking for missing values:

```
Number of missing values:
 Gender                                    0
DOB                                       15
Lead_Creation_Date                         0
City_Code                                814
City_Category                            814
Employer_Code                           4018
Employer_Category1                      4018
Employer_Category2                      4298
Monthly_Income                             0
Customer_Existing_Primary_Bank_Code     9391
Primary_Bank_Type                       9391
Contacted                                  0
Source                                     0
Source_Category                            0
Existing_EMI                              51
Loan_Amount                            27709
Loan_Period                            27709
Interest_Rate                          47437
EMI                                    47437
Var1                                       0
Approved                                   0
dtype: int64
```

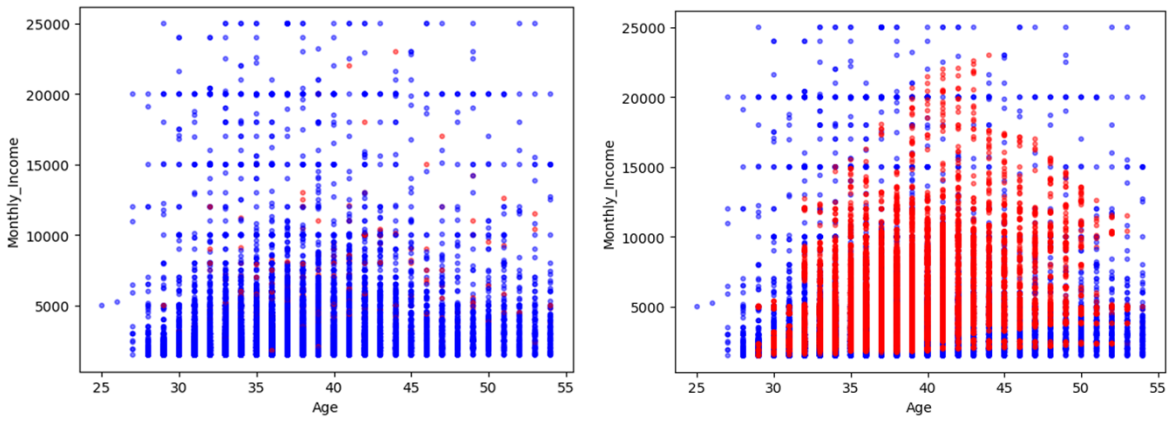
```
Number of missing values:
 Gender                                    0
City_Code                                 0
City_Category                             0
Employer_Code                             0
Employer_Category1                        0
Employer_Category2                        0
Monthly_Income                            0
Customer_Existing_Primary_Bank_Code       0
Primary_Bank_Type                         0
Source                                    0
Source_Category                           0
Existing_EMI                              0
Loan_Amount                               0
Loan_Period                               0
Interest_Rate                             0
EMI                                       0
Var1                                      0
Approved                                  0
age                                       0
lead_years                                0
dtype: int64
Shape of dataset: (13525, 20)
```

We also removed duplicate values, renamed variables and converted data types such as DOB and Lead_Creation_Date which we converted into years to present. After cleaning we ended up with 13,525 observations and 20 columns.

Exploratory Data Analysis

SMOTE

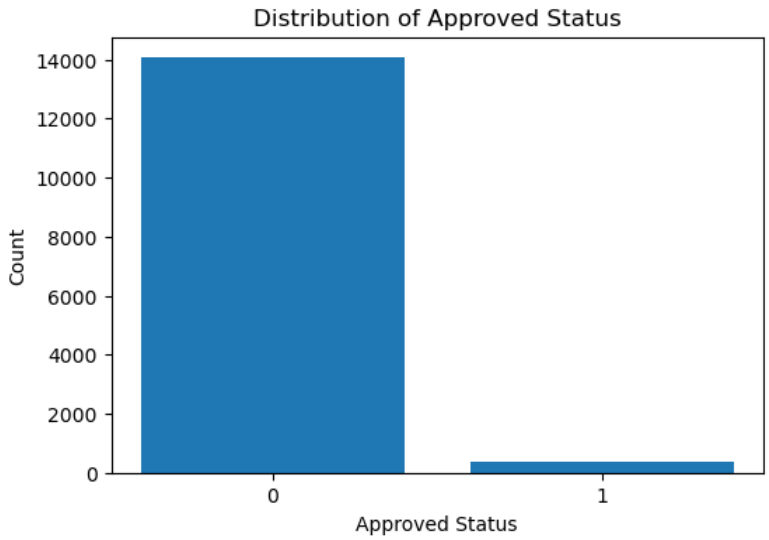
The target Variable was highly imbalanced after data cleaning, there were 12708 applicants not approved and only 325 applicants approved. The imbalance affected the performance of machine learning for classification as having a poor prediction performance on the minority class. One popular technique used to address class imbalance was called Synthetic Minority Over-sampling Technique (SMOTE). The technique uses KNN or K-nearest neighbors algorithm to generate synthetic or artificial values of the minority or underrepresented target variable. While it can be effective in improving the performance of classification models, it also has some limitations: it can also introduce some bias into the dataset and potentially decrease the generalizability of the model. SMOTE can also lead to overfitting, especially if the number of synthetic data points generated is too large.



The two plots above show the distribution of target variables before (left) and after (right) the SMOTE technique. SMOTE generated a balanced number of approved applicants to 12708, the balanced sample size was almost double the imbalanced sample size. Data points generated were large, introducing some bias such as gender: In the imbalanced sample size, there were about 87% male and 13% female, and in the EDA part, it was shown that gender was an influential variable for approval. However, in the balanced sample size, the amount of male and female were almost the same, erasing this distinction when doing model fitting. In the subsequent analysis, it would be explained why this technique was necessary for prediction.

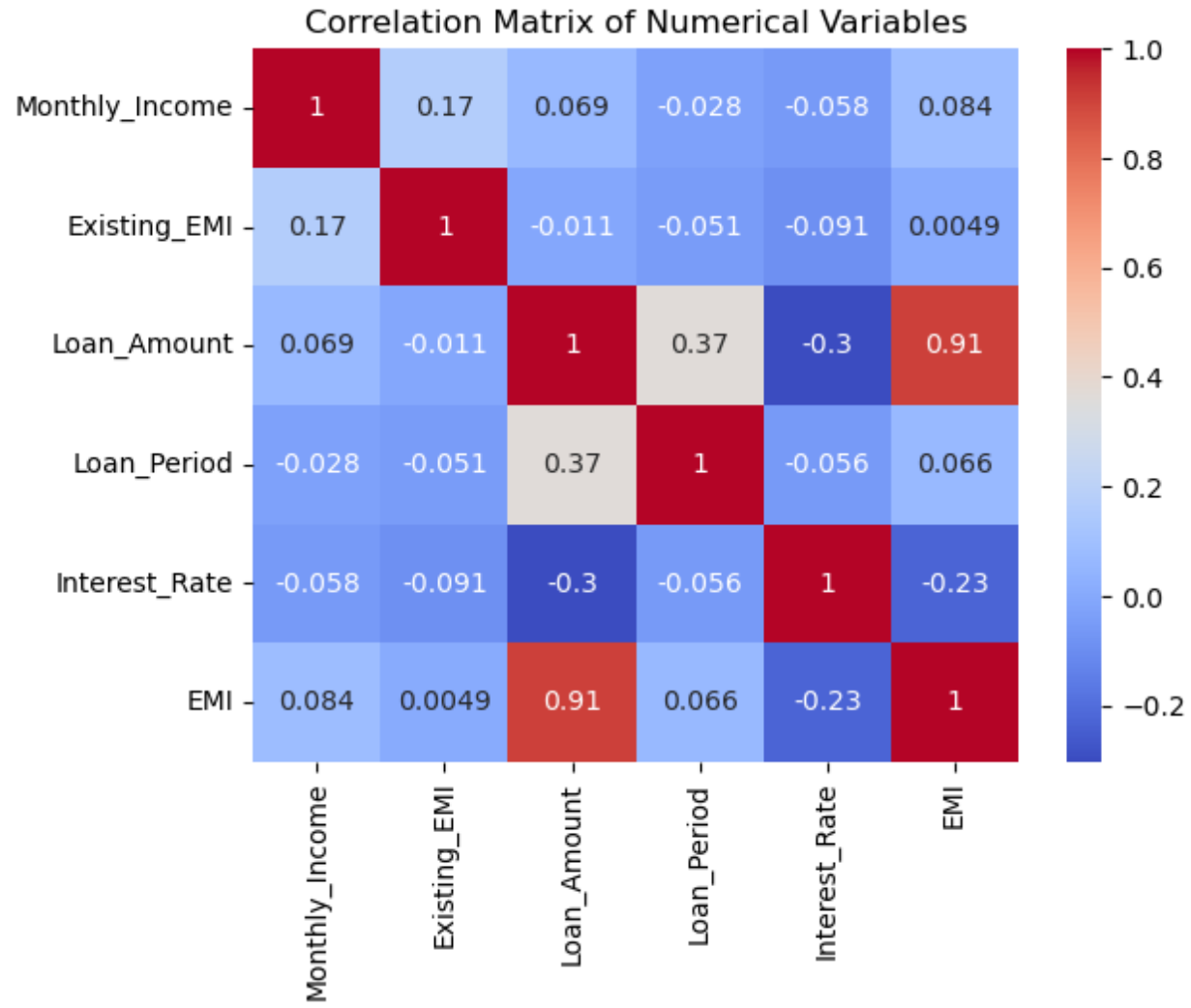
Bar plot of approved:

Understanding the distribution of approved conversion status in the dataset is important for analyzing the performance. The chart shows that the majority of entries in the dataset had a low number of approved conversions, while a few entries had a high number of approved conversions. 0 represents the not approved status and 1 represents the approved.



Correlation Matrix of Numeric Variables

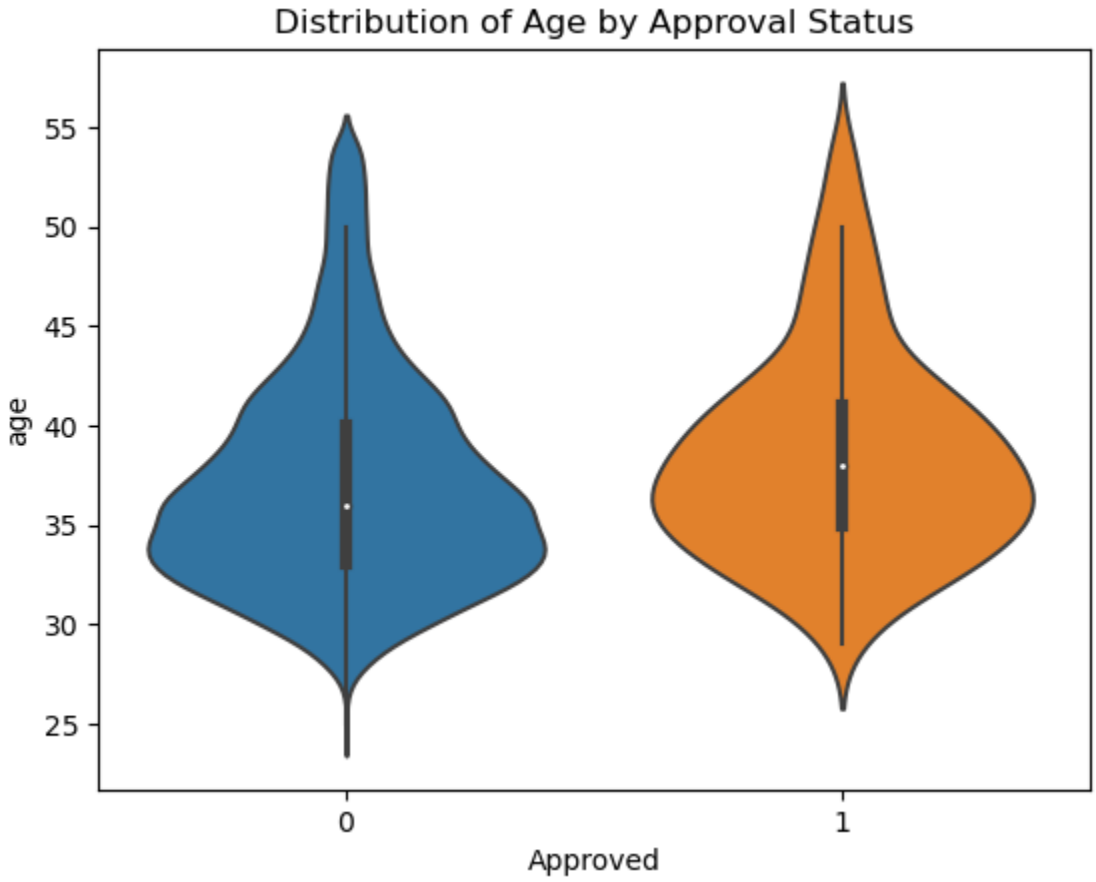
The correlation analysis of numerical variables reveals a strong positive correlation between Loan_Amount and EMI, with moderate correlation between Loan_Amount and Interest_Rate. Monthly_Income has weak positive correlation with Loan_Amount and EMI, while Existing_EMI has weak positive correlation with EMI. These insights can also help to predict loan eligibility and assess the repayment capacity.



Distribution of age by approval status

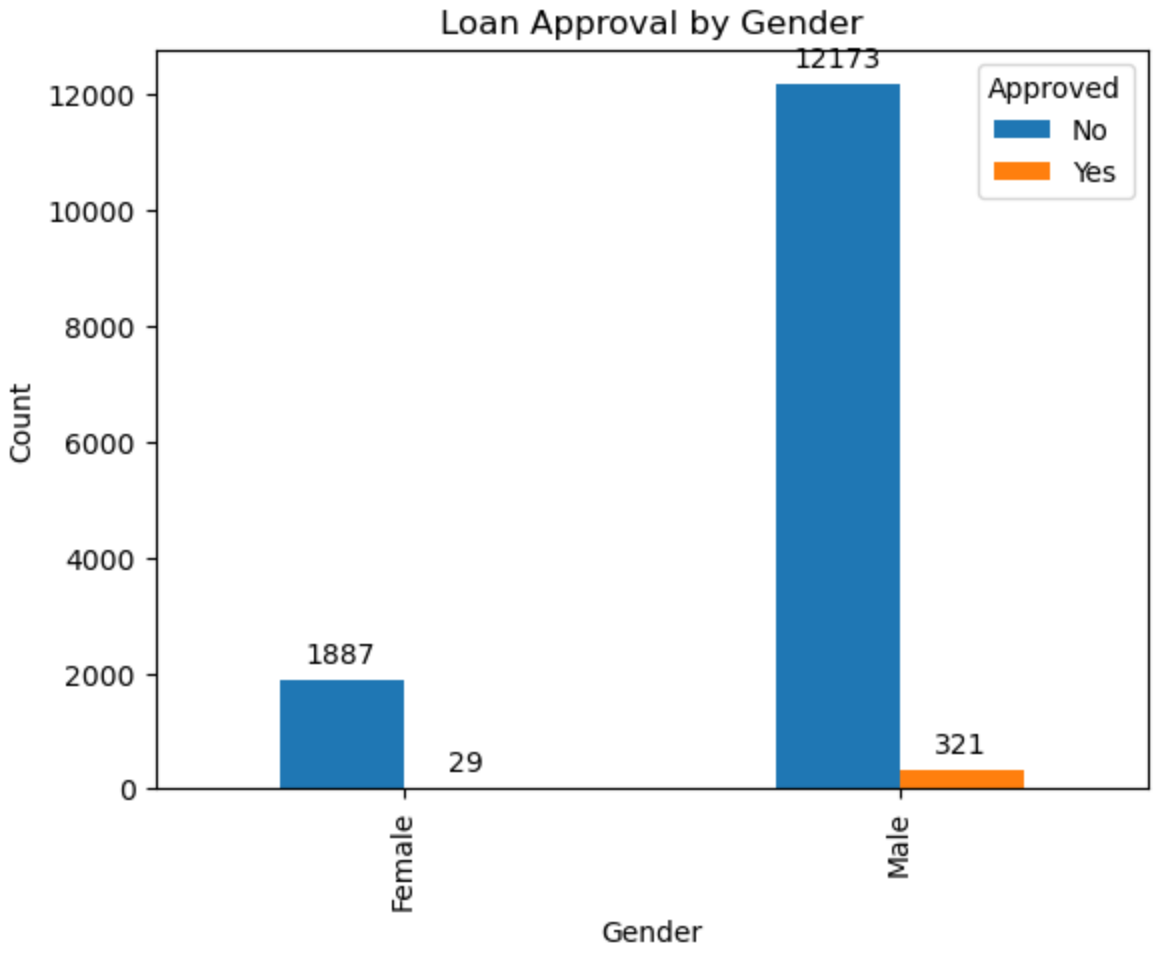
This visualization we utilized is a violin plot to display the distribution of age between approved and non-approved loan applications. The x-axis indicates the loan approval status, while the y-axis represents the age of loan applicants. It shows that most of the approved

applicants' age group are around 35-40 years and most of the not approved applicants' age group are around 33-37 years.



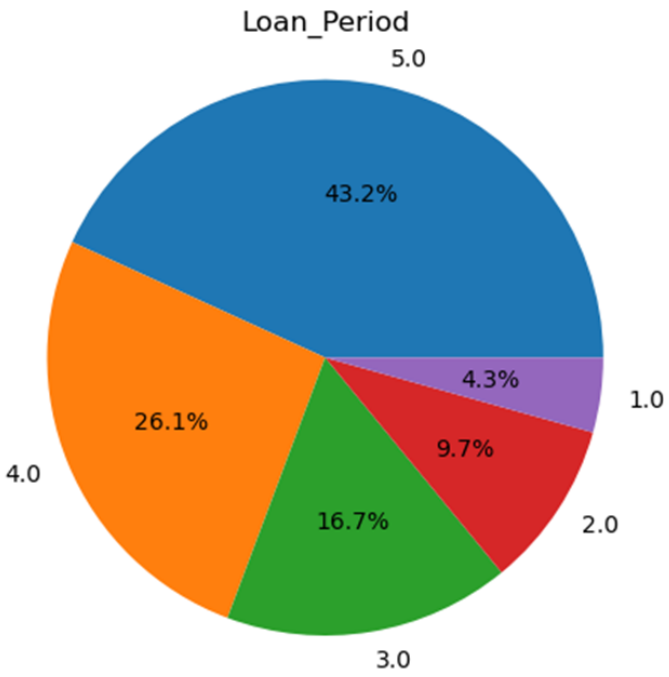
Loan approval by Gender

The barchart displays two bars for each gender, one for approved loan applications and the other for non-approved loan applications. The x-axis would represent the gender categories and the y-axis would represent the count of loan applications. It Results that there are 29 approved and 1887 not approved female applicants & 321 approved and 12173 not approved male applicants.



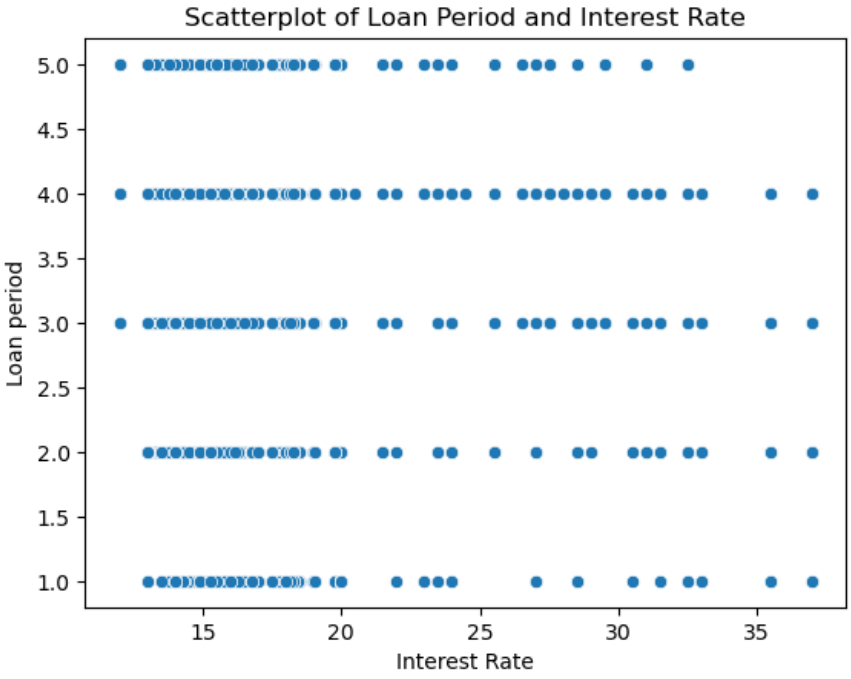
Pi chart for loan period

A pie chart was created to show the distribution of loan periods in the dataset. The chart shows that the majority of loans have a period of 12 months (4.3%), followed by 24 months (9.7%), followed by 36 months (16.7%) , followed by 48 months (26.1%) and 60 months(43.2%).



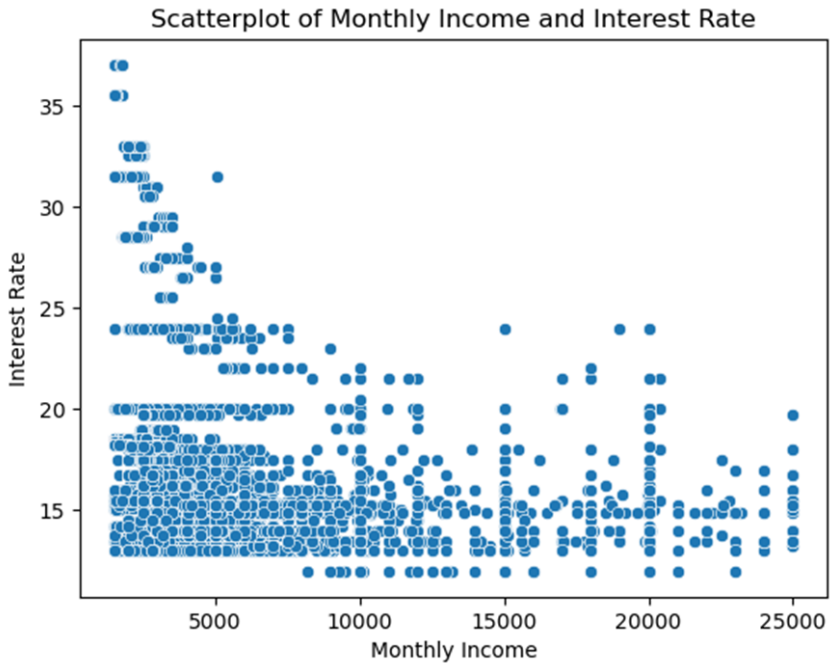
Scatterplot for Loan period & Interest rate

The scatter plot shows a weak negative correlation between "Loan_Period" and "Interest_Rate" variables in the dataset. Both the Longer and shorter loan periods tend to have a combination of lower interest rates and higher interest rates.



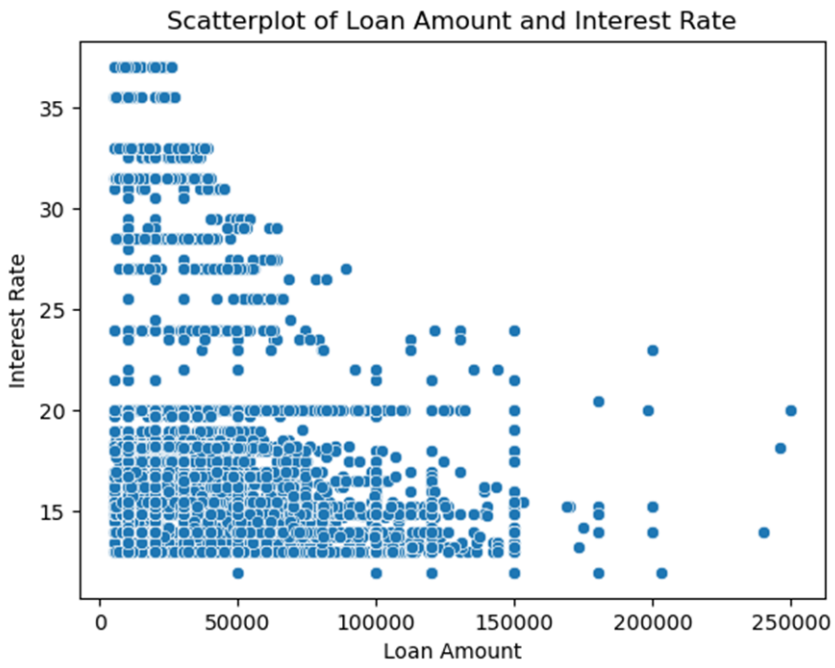
Scatter plot for monthly income and interest rate

This scatter plot displays the relationship between "Monthly_Income" and "Interest_Rate". It suggests a weak negative correlation between the two variables, indicating that higher income borrowers tend to be charged lower interest rates and most of the lower income borrowers tend to be charged higher interest rates.



Scatter plot for Loan amount & Interest rate

The scatter plot displays the relationship between the "Loan_Amount" and "Interest_Rate" variables in the dataset. Each data point represents a loan application and is plotted based on the loan amount and interest rate associated with that application. The x-axis represents the loan amount, and the y-axis represents the interest rate. It indicates that loans with higher loan amounts tend to have lower interest rates, and loans with lower loan amounts tend to have higher interest rates.



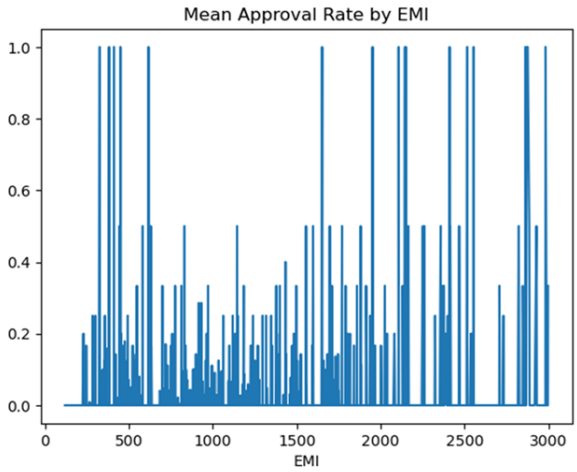
Logistic Regression

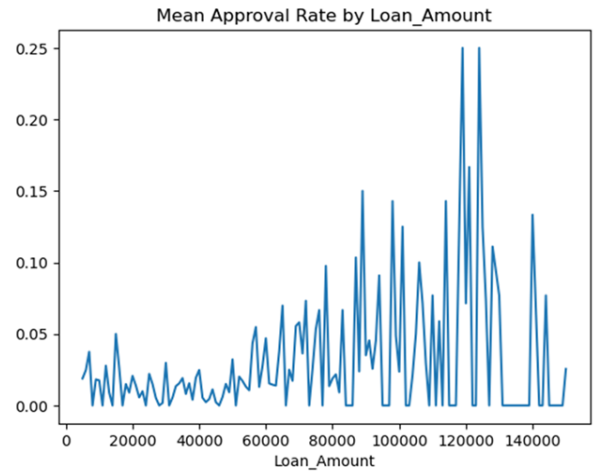
Feature Selection

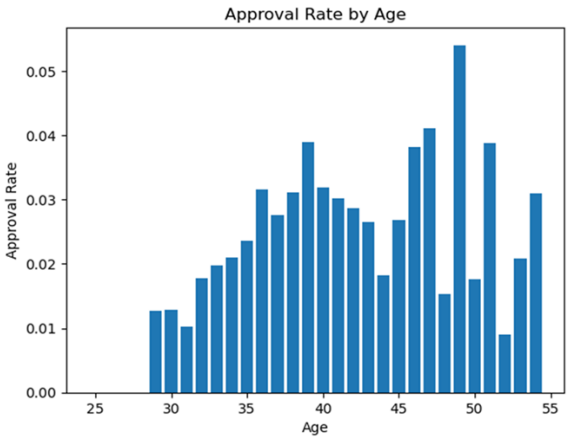
Based on visualization and exploratory data analysis, some numeric variables showed very high multicollinearity and some categorical variables tended to be dominated by one level. Variance Inflation Factor (VIF) measures the extent to which the variance of an estimated regression coefficient is increased due to multicollinearity. After converting into dummy variables, VIF is also explanatory on categorical variables. Usually a high VIF value (e.g., > 10) may indicate multicollinearity.

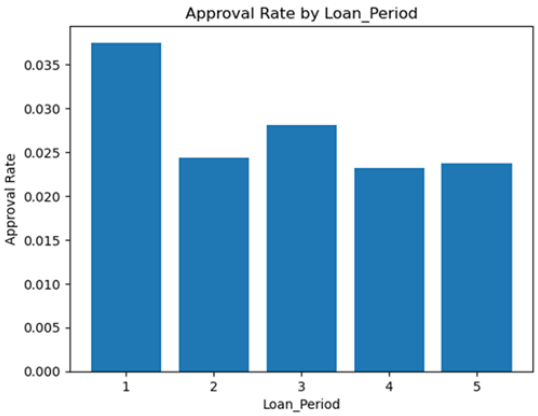
In terms of mean approval rate, barplots were helpful to visually observe whether there is a direct relationship.

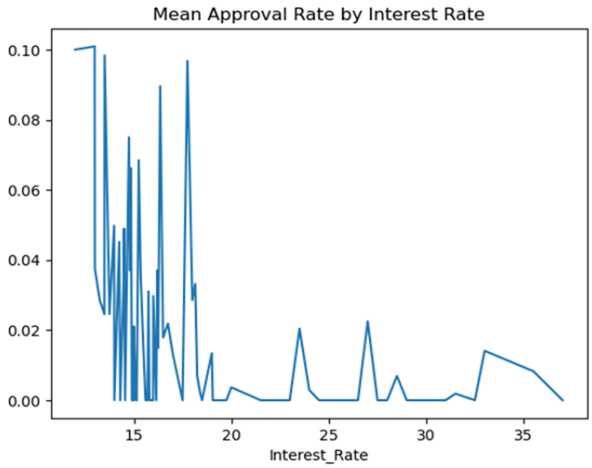
	Variable	VIF
5	EMI	80.814408
2	Loan_Amount	77.277122
7	age	47.349627
3	Loan_Period	41.694302
4	Interest_Rate	26.005073
0	Monthly_Income	13.079771
6	Var1	12.574982
9	Gender_Male	6.620683
8	lead_years	3.999074
17	Primary_Bank_Type_P	2.928022
1	Existing_EMI	2.541275
18	Source_Category_C	1.731703
12	Employer_Category1_B	1.494416
21	Source_Category_G	1.462831
13	Employer_Category1_C	1.362270
11	City_Category_C	1.187102
10	City_Category_B	1.131317
14	Employer_Category2_one	1.096036
16	Employer_Category2_two	1.059769
15	Employer_Category2_three	1.046466
20	Source_Category_F	1.040304
19	Source_Category_E	1.033369

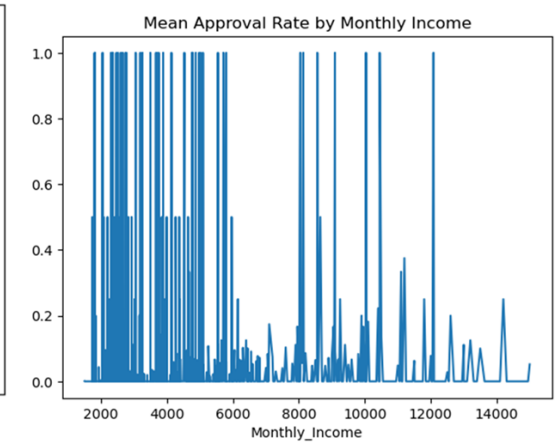

Mean Approval Rate by EMI
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
3000
EMI

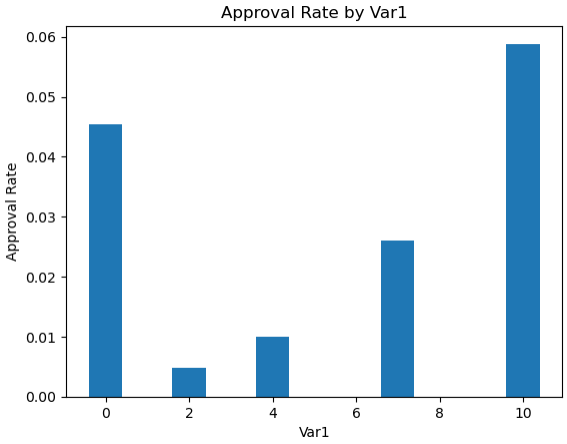

Mean Approval Rate by Loan_Amount
0.25
0.20
0.15
0.10
0.05
0.00
0
20000
40000
60000
80000
100000
120000
140000
Loan_Amount


Approval Rate by Age
Approval Rate
0.05
0.04
0.03
0.02
0.01
0.00
25
30
35
40
45
50
55
Age


Approval Rate by Loan_Period
Approval Rate
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
1
2
3
4
5
Loan_Period


Mean Approval Rate by Interest Rate
0.10
0.08
0.06
0.04
0.02
0.00
15
20
25
30
35
Interest_Rate


Mean Approval Rate by Monthly Income
1.0
0.8
0.6
0.4
0.2
0.0
2000
4000
6000
8000
10000
12000
14000
Monthly_Income



Noise existed in bar plots of mean approval rate by EMI and by monthly income. The distribution of approval rate by age was multimodal, these variables were dropped. In boxplots of categorical variables, different levels were distinct. Mean approval rate dropped at an interest rate of 20% and more fluctuant with increasing loan amount.

Selected model Summary

Logistic Regression Summary Table

	Coefficients	Standard Errors	p-values	Odds Ratios
const	-3.0351	0.6215	0.0	0.0481
Existing_EMI	0.0005	0.0001	0.0	1.0005
Interest_Rate	-0.1609	0.0279	0.0	0.8513
Var1	0.1994	0.0279	0.0	1.2206
Gender_Male	0.4993	0.2057	0.0152	1.6476
Employer_Category2_one	-0.5729	0.2218	0.0098	0.5639
Source_Category_C	0.4223	0.1642	0.0101	1.5254
AIC	2706.63	2706.63	2706.63	2706.63
BIC	2758.96	2758.96	2758.96	2758.96
Log-Likelihood	-1346.31	-1346.31	-1346.31	-1346.31
Pseudo R-squared	0.11	0.11	0.11	0.11

The summary table indicated existing EMI, interest rate, var1, gender, and some employer or source types were influential for approval. As an interpretation, for example, for every one percentage increase in interest rate, the log odds of approval versus not approval decreases by 0.1609, as interest rate is a tradeoff of risky loan. Male versus female, increases the log odds of approval by 0.4993.

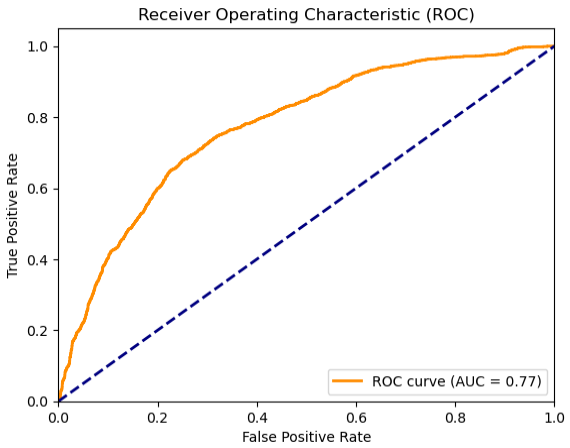
A model based on balanced data was also made. However, all variables were significant except in gender and lead years. This model had a pseudo R-squared of 0.4, better than 0.11 in

the selected model, however, it also had extremely large AIC and BIC value, indicating much larger information lost with the SMOTE technique.

Prediction

By splitting the dataset into a 80% training set and a 20% testing set, we trained the selected model to make predictions on the testing set. The accuracy rate was 0.97, however the high accuracy rate was considered as overfitted because of bias in classification, as no testing point was assigned into the predicted set.

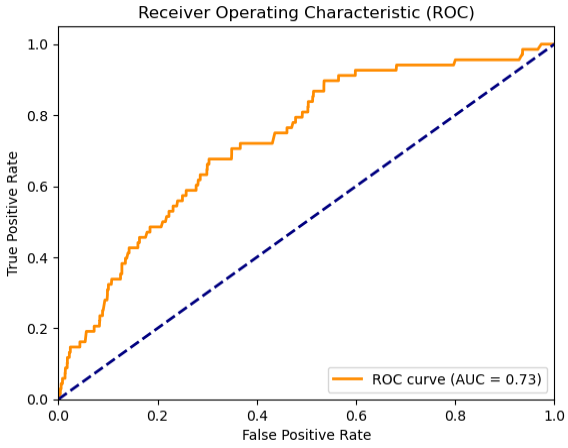
```
Accuracy Score: 0.9739163789796701
Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.99      2539
           1       0.00      0.00      0.00        68

    accuracy                           0.97      2607
   macro avg       0.49      0.50      0.49      2607
weighted avg       0.95      0.97      0.96      2607

            Predicted No  Predicted Yes  Total
Actual No           2539              0   2539
Actual Yes            68              0     68
Total               2607              0   2607
```



To find a more reliable prediction, the balanced dataset was also splitted as a 80% training set and a 20% testing set for machine learning. The accuracy score was 0.713, however the interpretability of this classification increased: As 20% of total balanced data, there were 2430 assigned into not approved and 2654 assigned into approved.

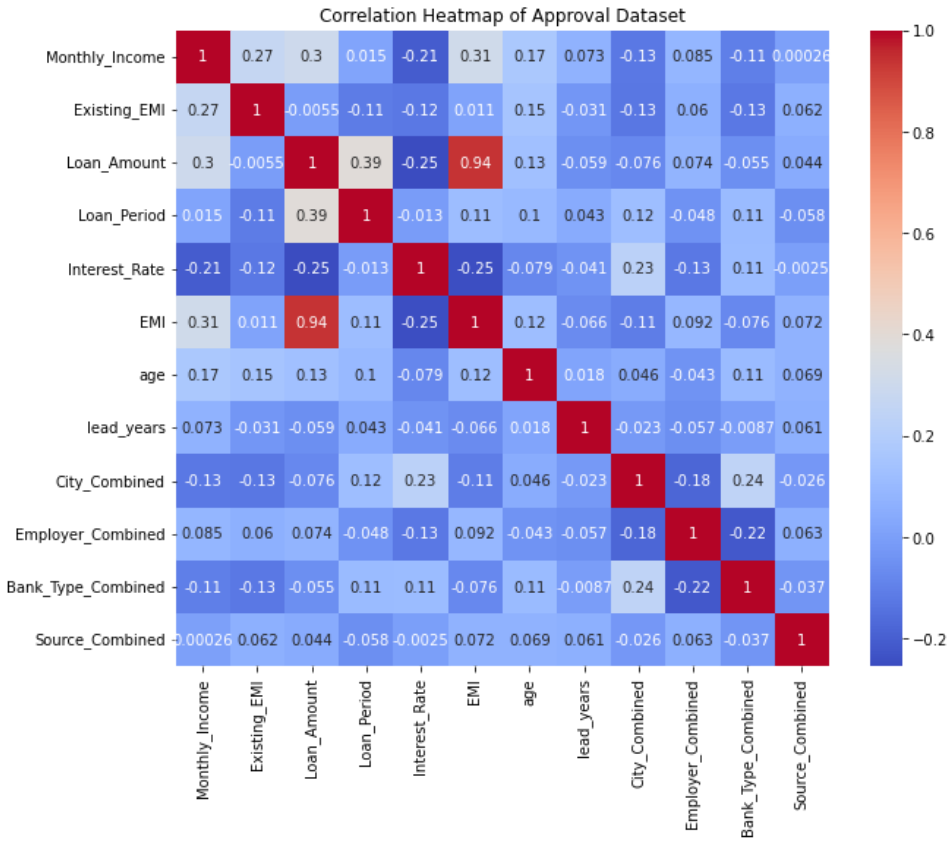
```
Accuracy Score: 0.7128245476003147
Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.69      0.71      2580
           1       0.70      0.74      0.72      2504

    accuracy                           0.71      5084
   macro avg       0.71      0.71      0.71      5084
weighted avg       0.71      0.71      0.71      5084

            Predicted No  Predicted Yes  Total
Actual No           1775            805   2580
Actual Yes           655           1849   2504
Total               2430           2654   5084
```

In the ROC curve, area under the curve was 0.73, some distance from what is considered a good classification and if we would like to improve this classification, minimized false positive rate while true positive rate increasing should be considered, as false positives usually represent a risk of default.



Ordinary Least Squares Regression

To answer our third SMART question: Are there any variables that indicate how much an applicant will be approved for? We used ANOVA test, T-test, Tukey's HSD test for statistical analysis and applied Multi-Linear regression to predict the amount of Loan money after the customer got the approval. After subsetting the dataset, we only have 350 rows of data. For the approval group, we did correlation heatmap and shows as below:



Since EMI is highly correlated with Loan_Amount, we only select Loan_Period, Interest_Rate, Monthly_Income and Age as our independent variables. First of all, we conducted ANOVA-test for different Loan_period categories.

ANOVA test

F-statistic: 17.924811035109016

p-value: 2.2470564142307515e-13

The ANOVA test result shows an F-statistic of 17.36 and a very low p-value. This suggests that there is a significant difference in the mean 'Loan_Amount' between at least two 'Loan_Period' categories in the group of approved loans.

To determine which airs of 'Loan_Period' categories have significantly different mean 'Loan_Amount' values. We used Perform post-hoc tests (Tukey's HSD test). The result shows below:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
1.0	2.0	8484.8485	0.8573	-14888.7263	31858.4232	False
1.0	3.0	24818.1818	0.0109	3912.221	45724.1426	True
1.0	4.0	42654.1275	0.0	22388.7645	62919.4904	True
1.0	5.0	42465.0098	0.0	22979.7455	61950.2741	True
2.0	3.0	16333.3333	0.099	-1771.7598	34438.4265	False
2.0	4.0	34169.279	0.0	16807.8235	51530.7345	True
2.0	5.0	33980.1613	0.0	17535.9913	50424.3313	True
3.0	4.0	17835.9457	0.0043	3973.9351	31697.9562	True
3.0	5.0	17646.828	0.0015	4952.5014	30341.1546	True
4.0	5.0	-189.1177	1.0	-11798.2376	11420.0022	False

Group 1 is different with Group 3, 4, 5

Group 2 is different with Group 4 ,5

Group 3 is different with Group 5

After statistical analysis, we conduct an OLS regression for dependent variable Loan_amount. Result shows as below:

OLS Regression Results

Dep. Variable:	Loan_Amount	R-squared:	0.362
Model:	OLS	Adj. R-squared:	0.349
Method:	Least Squares	F-statistic:	27.48
Date:	Sun, 23 Apr 2023	Prob (F-statistic):	7.67e-30
Time:	20:09:48	Log-Likelihood:	-4033.1
No. Observations:	347	AIC:	8082.
Df Residuals:	339	BIC:	8113.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.105e+04	1.2e+04	3.422	0.001	1.75e+04	6.46e+04
Monthly_Income	4.1597	0.532	7.817	0.000	3.113	5.206
Interest_Rate	-1445.1118	562.553	-2.569	0.011	-2551.647	-338.577
age	-378.0413	301.278	-1.255	0.210	-970.650	214.568
Loan_Period_1.0	-1.877e+04	5334.138	-3.518	0.000	-2.93e+04	-8274.772
Loan_Period_2.0	-8406.5399	4857.557	-1.731	0.084	-1.8e+04	1148.208
Loan_Period_3.0	1.05e+04	4059.443	2.588	0.010	2519.159	1.85e+04
Loan_Period_4.0	2.921e+04	3782.261	7.723	0.000	2.18e+04	3.67e+04
Loan_Period_5.0	2.85e+04	3480.968	8.189	0.000	2.17e+04	3.54e+04

Omnibus:	33.165	Durbin-Watson:	2.152

The result shows that Monthly Income, Interest Rate, Loan Period are significant variables And R-squared 0.362 indicates these predictors only explained a partial effect. From the slopes of independent variables, with the increase of Interest_Rate and Age, the loan_amount that can be approved is likely to decrease. Whereas with the increase of Loan_period,the loan_amount that can be approved is likely to increase. In order to test the performance of OLS regression, we performed 5-fold cross-validation with RMSE: 27610.54, lower RMSE usually indicates a better fitted model.

Conclusion

Based on the coefficients provided, we can make the following conclusions and insights:

The variable with the largest positive coefficient is Var1, indicating that higher values of this variable are associated with a higher likelihood of loan approval.

The variable with the largest negative coefficient is Interest_Rate, indicating that higher interest rates are associated with a lower likelihood of loan approval.

Among the categorical variables, the employer category and source category appear to have the largest coefficients, with negative coefficients for certain categories indicating that they are associated with a lower likelihood of loan approval.

Gender (male) has a small negative coefficient, indicating that being male is associated with a slightly lower likelihood of loan approval.

Existing_EMI has a small positive coefficient, indicating that higher values of these variables are associated with a slightly higher likelihood of loan approval.

Overall, these coefficients suggest that Var1 and Interest_Rate are the most important predictors of loan approval, followed by the employer category and source category variables. However, it is important to note that the coefficients only provide a partial picture of the relationship between the predictor variables and the outcome variable, and other factors such as interactions between variables and model fit should also be taken into consideration. In this case, we went with SMOTE due to the imbalance data set that this provided us.

In the future, our group thought that it would be worth doing two data sets, one with SMOTE and one without to create a better comparison and see the effects that SMOTE had on our conclusions.

Reference

Mohammad, A. (2022) *The impact of digital marketing success on customer loyalty anber ..., THE IMPACT OF DIGITAL MARKETING SUCCESS ON CUSTOMER LOYALTY.*

Available at: https://armgpublishing.com/wp-content/uploads/2022/10/A633-2022-09_Anber.pdf (Accessed: April 30, 2023).

Zhou, L. and Wang, H. (2012) *Loan default prediction on large imbalanced data using random forests*, *Loan Default Prediction on Large Imbalanced Data Using Random Forests*. Available at: https://www.researchgate.net/publication/267864165_Loan_Default_Prediction_on_Large_Imbalanced_Data_Using_Random_Forests (Accessed: April 30, 2023).

T. Aditya Sai Srinivas, Kharisma Govinda and Somula Ramasubbareddy (2022) *The impact of digital marketing success on customer loyalty anber ...*, *Loan Default Prediction Using Machine Learning Techniques*. Available at: https://armgpublishing.com/wp-content/uploads/2022/10/A633-2022-09_Anber.pdf (Accessed: April 30, 2023).