Project Work In Machine Learning And Data Mining

Reza Shatery, 1083797

January 16, 2024

1 Introduction

This project is conducted with the aim of machine learning tools to visualize and measure the Bank Loan systems of relationships between customers and Bank by getting loan or not. I will show what the most influential metrics are, their collaborations and what kinds of metrics are most common to fulfill this project.

2 Problem And Motivation

This case is about a bank (Thera Bank) which has a growing customer base. Majority of these customers are liability customers (depositors) with varying size of deposits. The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans. In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9 percent success. This has encouraged the retail marketing department to devise campaigns to better target marketing to increase the success ratio with a minimal budget. The department wants to build a model that will help them identify the potential customers who have a higher probability of purchasing the loan. This will increase the success ratio while at the same time reduce the cost of the campaign

I aim to develop a classifier for identifying potential customers who are more likely to purchase a personal loan using the Thera-Bank dataset [Jac]. Thera-Bank is interested in expanding their loan business by converting liability customers into retail loan customers, while keeping them as depositors. The retail marketing department is developing campaigns with better target marketing to increase the success rate with a minimal budget.

3 Dataset

The data set includes 5000 observations with fourteen variables divided into four different measurement categories. The binary category has five variables, including the target variable personal loan, also securities account, CD account, online banking and credit card. The interval category contains five variables: age, experience, income, CC avg and mortgage. The ordinal category includes the variables family and education. The last category is nominal with ID and Zip code. The variable ID does not add any interesting information e.g. individual association between a person (indicated by ID) and loan does not provide any general conclusion for future potential loan customers. Therefore, it will be neglected in the examination.

4 Explorate Dataset

At first I should check the dataset. the dataset include 5000 rows and 14 columns. secondly, I check if there is any null in the dataset which I found that there is no null in the whole dataset. Then, I check what are the columns are categorical and what are not, so these columns are coategorical:
[Family, Education, Personal Loan, Securities Account, CD Account, Online, CreditCard]
and these are continuous :
[ID, Age, Experience, Income, ZIP Code, CCAvg, Mortgage]
Also I check duplicated rows, and there is no duplicated rows at this dataset.

5 Data Preprocessing

In data preprocessing for machine learning, several steps are typically performed to prepare the data for training and improve the performance of the machine learning model.

- Handling Missing Data: Identify and handle missing values in the dataset which is I do not have it in our dataset.

- Data Cleaning: Clean and correct errors in the dataset, such as typos, inconsistencies, and outliers. for the outliers I use standard deviation method, data that are more than 3 standard deviations away from the average data are known as outliers. I do have a bunch of outlier in Mortgage and CCAvg which can be removed from our dataset.

- Encoding Categorical Variables: Convert categorical variables into a numerical format, as most machine learning algorithms require numerical input. so I do this in our project.

- Remove useless coloumns: individual association between a person (indicated by ID) and loan does not provide any general conclusion for future potential loan customers. Therefore, it will be neglected in the examination.

- Attribute conversion: In this data set, CCAVG represents the average monthly credit card cost, but revenue represents the amount of annual revenue. To equalize the units of characteristics, I get the amount of monthly income These steps help create a clean, consistent, and informative dataset that enhances the performance and generalization ability of machine learning models.

6 exploratory data analysis (EDA)

exploratory data analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.primarily EDA is for seeing what the data can tell us beyond the formal modeling and thereby contrasts traditional hypothesis testing.

6.1 Correlation

Correlation analysis helps identify which features are most strongly related to the target variable. This information is useful for prioritizing features during feature engineering or model development.

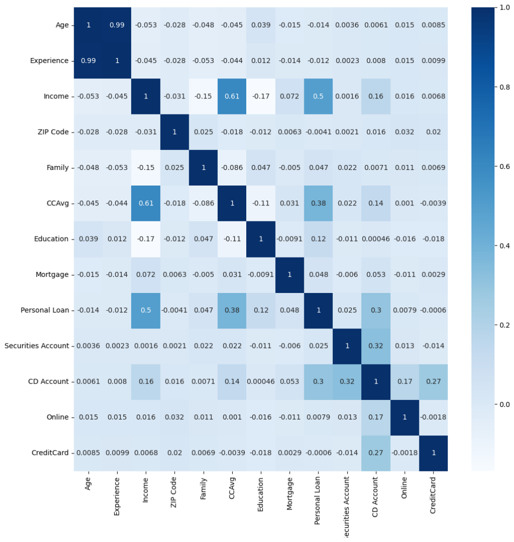

Figure 1: Correlation

If the correlation coefficient between two variables is greater than 0.7, it is considered as a strong correlation. If the correlation coefficient between two variables is between 0.3 and 0.7, it is considered as a moderate correlation. If the correlation coefficient between two variables is less than 0.3, it is considered as a weak correlation. Personal Loan is highly correlated with Income, CD-Account, CCAvg.Experience is highly correlated with Age

$$\rho = 1$$

CCAvg is correlated with Income to a good extent.

$$\rho = 0.6$$

Age and Experience features have very high correlation, 0.99. It is also intuitively understandable that experience increases as age increases. Correlated features degrade the learning performance and causes instability on the models

6.2 Data Visualization

sns.pairplot is a powerful tool for getting a visual overview of the relationships and distributions within a dataset. It is especially useful for exploratory data analysis (EDA) to identify patterns, trends, and potential areas of interest for further investigation. And also,it is a powerful tool for getting a visual overview of the relationships and distributions within a dataset.

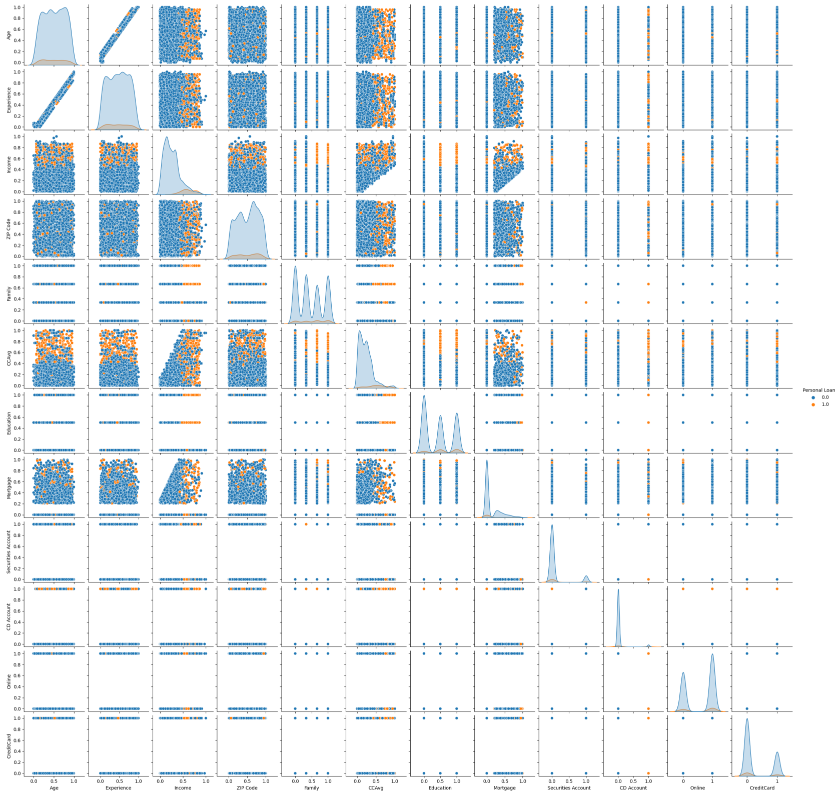

Figure 2: pairplot

By examining both the histogram and boxplot for each column, I can gain insights into the distribution, skewness, and presence of outliers in the data. This is useful for visualizing the univariate distribution of each column, helping you identify patterns and potential issues in the data

6.3 Categorical Variables

Explore the distribution of categorical variables using bar charts. Check for class imbalances, especially in the target variable for classification problems

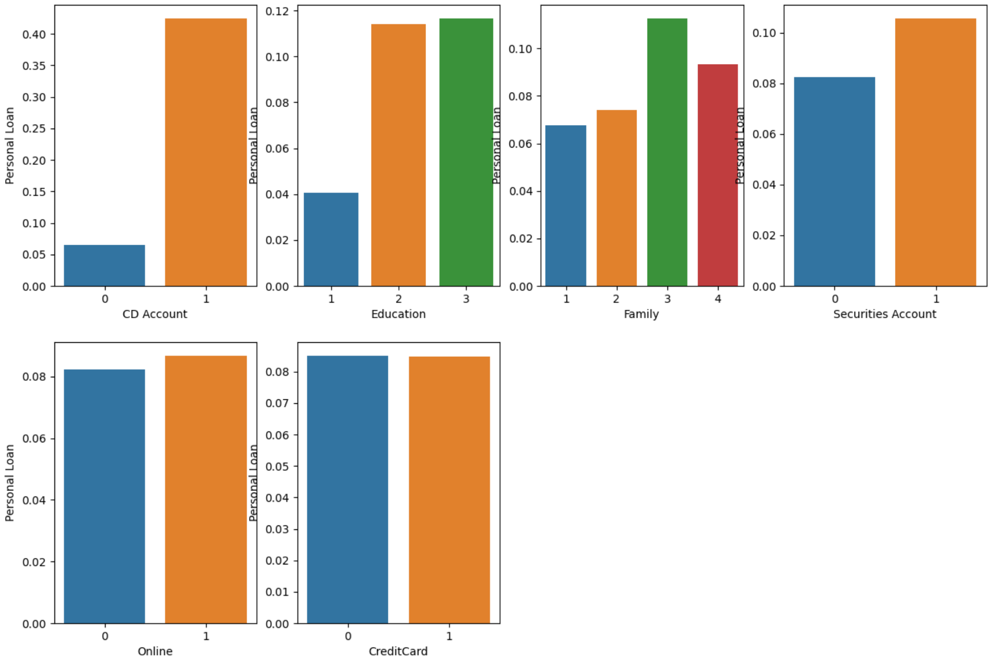

Figure 3: categorical distribution

figure 3 shows that: Customers with family size equal to 3 have more chances of having Personal Loan. Customers with Undergraduate degree have less chances of having Personal Loan as compared to other customers having Graduate or Advanced/Professional degree. Customers with CD Account and Securities Account have more chances of having Personal Loan. Customers with Online and Credit Card is more likely to have Personal Loan than others don't have a one.

6.4 Applying scaling

Min-max scaling is beneficial when numerical features have different scales, ensuring that each feature contributes equally to the model. The code is useful for preprocessing data, especially when preparing it for machine learning algorithms that are sensitive to feature scales.
Note: Min-max scaling transforms the data such that the minimum value becomes 0, the maximum value becomes 1, and values in between are linearly scaled. The transformation is column-wise, and each feature is scaled independently

6.5 Data Splitting

I should divide the dataset into training and testing sets to assess the model's performance on unseen data. I use 80 percent for training and 20 percent for testing with the random state equal to 42.

7 Models

I want use models for classify personal loan. I use criteria such as accuracy and confusion matrix to find our best models. I examine six models namely, SVM, DecisionTreeClassifier, RandomForestClassifier, LogisticRegression, and Gaussian Naive Bayes. after examining of these models I have found that DecisionTreeClassifier with 0.9812 accuracy and RandomForestClassifier with 0.98956 accuracy are the best models to use for classification. and for the sake of learning I used GaussianNB to just check what happened and just work on it.

7.1 Decision Tree Model

First I want to find that what is the maximum depth is. The choice of the maximum depth is a hyperparameter that can be adjusted during the model training process to find the optimal balance between capturing patterns and avoiding overfitting. The maximum depth of the tree fitted on X-train is 14. in Decision Tree classifier this result indicates that, when training a Decision Tree classifier on the training data (X-train), the depth of the resulting decision tree is set to 14. after I examine it on test set the result is: The accuracy on test set tuned with cross validation is 98.5 percent with depth of the tree 4.

class	precision	recall	f1-score
class 0	0.99	0.99	0.99
class 1	0.94	0.89	0.91

Table 1: Decision Tree Model Results

The balance between precision and recall for Class 0 and 1 are very good. The high precision and recall for both classes suggest that the model is performing very well on the given dataset. The F1-score provides a balanced measure of performance, considering both precision and recall.

7.2 Random Forest Model

Firstly,I perform a randomized search for hyperparameter tuning on a RandomForestClassifier. I get this results:Fitting 3 folds for each of 100 candidates, totalling 300 fits.

- folds: Refers to the number of folds used in cross-validation. In this case, the dataset is split into 3 folds for cross-validation.
- candidates : Indicates that the randomized search is trying out hyperparameter combinations for the RandomForestClassifier from a total of 100 random combinations.
- totalling 300 fits: Suggests that each candidate is fit on each of the 3 folds, resulting in a total of 300 fits (100 candidates * 3 folds)

the accuracy on test set is 0.98 which is very good. and this is the reuslts:

class	precision	recall	f1-score
class 0	0.99	1.0	0.99
class 1	0.95	0.85	0.90

Table 2: Random Forest Model Results

High precision indicates that when the model predicts a certain class, it is likely to be correct. High recall indicates that the model can capture a large proportion of actual instances of a class. F1-score is a harmonic mean of precision and recall, providing a balance between the two metrics.

7.3 Gussian Naive Bayes Model

this is the results of GussianNB:

class	precision	recall	f1-score
class 0	0.96	0.93	0.95
class 1	0.45	0.60	0.51

Table 3: GussianNB Results

High precision for Class 0 indicates that when the model predicts Class 0, it is likely to be correct. The lower precision for Class 1 suggests that the model may have some false positives for this class. The recall for Class 1 suggests that the model captures more than half of the actual instances of Class 1. The F1-score provides a balanced measure of performance, considering both precision and recall. Gaussian Naive Bayes (GaussianNB) is a popular machine learning and it is particularly suitable for situations where the features (attributes) are continuous and assumed to follow a Gaussian (normal) distribution. as a reuslts because are features are not continous the results are not good enough.

8 Conclustion

in this work, I use the dataset of Thera Bank. the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans.I built models that will help them identify the potential customers who have a higher probability of purchasing the loan. This will increase the success ratio while at the same time reduce the cost of the campaign. Firstly, I get some information form dataset using the functions of data frame form pandas library, to see what are the columns and if there is null or error in the dataset that I should cope with that. Secondly, I have done some data preprocessing for cleaning the data, encoding categorical variables, remove useless columns, and some attribute conversion. Thirdly, I have done the data analysis, such as correlation, data visualization, check the categorical variables, applying scaling and data splitting. And Finally, I use 3 models based on our metrics, 1. Decision Tree Model 2. Random Forest Model 3. Gussian Naive Bayes Model. Based on our work the Decision Tree Model is the best one and after that the Random Forest Model. Of course, I could use other models but in this case I have done with these models.

References

[Jac] Sunil Jacob. Bank loan modelling. *https://www.kaggle.com/datasets/itsmesunil/bank-loan-modelling/data*.