Campaign for selling personal loans

Problem Statement

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% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with minimal budget.

Objective

The department wants to build a model that will help them identify the potential customers who have higher probability of purchasing the loan. This will increase the success ratio while at the same time reduce the cost of the campaign.

Dataset

The csv file provided contains data on 5000 customers. The data includes customer demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign.

Expectations

1. What are the insights that you can infer from studying the various attributes? Are there any specific points that you would like to bring attention to?

2. What are the approaches that can be used to model the likelihood of a customer buying personal loans?

3. Of the approaches tried, which model is the best when it comes to predictive performance in this case? Are their any drawbacks in your chosen approach? How would you adjust for it?

4. Which are the most important attributes which influence a customer’s decision to go for a loan? What actions can you recommend in order to improve the campaign to sell more loans?

5. Are their any improvements from a data collection perspective that you would like to suggest?

6. Submit your technical analysis in a python notebook. Also create a summary report explaining your insights and recommendations in plain and simple words such that nonanalyst can understand it.

Solution/Approach

Steps

1. Reading csv file and then checking dimensions, few rows to know little about data etc

2. Variable identification to know about **continuous or categorical in nature**, checking any missing value, removing ID column

3. Univariate analysis

A. Continuous Variables

Tabular and graphical method is used to know about **mean,median,mode,Q1,Q3, skewness ,outliers** etc

B. Categorical Variables

Tabular and graphical method is used to know about number of times each value occuring in that particular columns

4. Bivariate analysis

Tabular and graphical method is used to know about relation between continuous-continuous variables using correlation, categorical-continuous variables using grouping and categorical-categorical variables using crosstab.

5. Missing Value Treatment:

This step is not required as there are no missing values.

6. Outliers Treatment

Using **box plot**, i am able to find outliers then using **Q1-1.5(Q3-Q1)** and **Q3+1.5(Q3-Q1)**, i am replacing that outliers with median.

7. Variable Transformation

Using StandardScaler from scikit learn , transforming all columns values in such a manner that it will lie in **-3 to +3 std**

8. Modeling

a. Splitting the dataset into training and testing

8.1 Logistic Regression (Accuracy : 93%,F1_score=46%)

8.2 KNN(Accuracy : 92%,F1_score=53%)

8.3 Random Forest(Accuracy : 97%,F1_score=85%)

b. Finally compare the **accuracy, F1_score** of each of the model.

INSIGHTS

1. What are the insights that you can infer from studying the various attributes? Are there any specific points that you would like to bring attention to?

1 There is **no missing value** in any of the columns.

2. Column **'ID'** is appearing exactly like **primary key**(as in DBMS),so each row value differs

3. **Experience** column contains some **negative value** like -1,-2 and -3

4. Family, Education, Personal loan, Securities Account, CD Account, Online and Credit Card are **categorica**l in nature but these **all present in numerical form**

5. There is much fluctuation in values of different columns so i had transformed these values in transformation section

6. Most of the values in **Age** column lie between 30 and 60

7. These three columns **CCAvg,Income and Mortgage** values are **left skewed**

8. These three columns **CCAvg,Income and Mortgage** contains **outliers**

9. Age and Experience are highly correlated(0.977182) so i had drop Experience column

10. Personal Loan with Income and Mortgage(with value 1) are more affected

11. personal loan with value 0 are playing important role with all other categorical variables.

12. **The target column that is Personal Loan, value is not in balance form so accuracy measure is not the good choice so I had preferred to go to f1_score.**

Models/Algorithms

2. What are the approaches that can be used to model the likelihood of a customer buying personal loans?

3. Of the approaches tried, which model is the best when it comes to predictive performance in this case? Are their any drawbacks in your chosen approach? How would you adjust for it?

1. I had first tried simple classification algorithm that is **Logistic Regression** to form the model and I had used **75% data for the training purpose** and **25% data for the testing purpose** so that I am able to know how good our model in terms of generalization.

2. The **accuracy** which i got is **93%** but **accuracy is not the good measure due to unbalanced** value in target column so I had used f1_score and value I got is only 46%

3. I had moved to next simple model to improve f1_score and the model is **KNN**

4. First I had used **K=5** to form model (keeping 25% data for testing) but the f1_score remain 46%

5. Then I had tried to tune hyperparameters that is K using **Elbow technique** and I had used K=1 to K=30 but at K=3 the **f1_score** is highest that is **53%**

6. Then I had moved to some famous algorithm which generally work well during solving Kaggle type problem or hackathon problem. The algorithm is **Random Forest**.

7. I had used Random Forest with same percentage of dataset that is 75% for training dataset and 25% for test dataset and I had only given 21 estimator and there is huge improvement in f1_score that is 85%

8. After forming these three models, Random Forest is best and giving **accuracy of 97%** and **f1_score is 85%.**

	Accuracy	**F1_Score**
Logistic Regression	0.9296	0.463415
KNN	0.9272	0.525140
Random Forest	0.9752	0.847291

4. Which are the most important attributes which influence a customer's decision to go for a loan? What actions can you recommend in order to improve the campaign to sell more loans?

→ Income and Mortgage are two attributes which influence a customer's decision most to go for a loan.

Some advance techniques like feature engineering, neural network, I can use to improve the campaign to sell more loans.

5. Are their any improvements from a data collection perspective that you would like to suggest?

→ If we add these attributes like **Gender, Married, Dependents, Loan Amount, Loan Amount Term** etc then there might be chance that model will be more better.