**Business Case Study - Bank Loan**

By: Sakshi Jain

**Context:**

This case is about a bank (Thera Bank) whose 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 a minimal budget

**Objective:**

The classification goal is to predict the likelihood of a liability customer buying personal loans.

|  |  |
| --- | --- |
| ***Column Name*** | ***Description*** |
| ID | Customer ID |
| Age | Customer's age in completed years |
| Experience | No of years of professional experience |
| Income | Annual income of the customer ($000) |
| ZIP Code | Home Address ZIP code |
| Family | Family size of the customer |
| CCAvg | Avg. spending on credit cards per month ($000) |
| Education | Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional |
| Mortgage | Value of house mortgage if any. ($000) |
| Personal Loan | Did this customer accept the personal loan offered in the last campaign? |
| Securities Account | Does the customer have a securities account with the bank? |
| CD Account | Does the customer have a certificate of deposit (CD) account with the bank? |
| Online | Does the customer use internet banking facilities? |
| Credit Card | Does the customer use a credit card issued by Thera Bank? |

**Information on the features:** We have 13 independent variables and 1 dependent variable i.e. ‘Personal Loan’ in the data set. Also, we got 5000 rows which can be split into test & train datasets:The attributes we divided as below:

The variable **ID** and **ZIP Code** does not add any interesting information. There is no association between a person's customer ID and loan; also it does not provide any general conclusion for future potential loan customers. Hence we are **dropping** this information for our model prediction.

The binary category has five variables as below:

* Personal Loan - Did this customer accept the personal loan offered in the last campaign? **This is our target variable**
* Securities Account - Does the customer have a securities account with the bank?
* CD Account - Does the customer have a certificate of deposit (CD) account with the bank?
* Online - Does the customer use internet banking facilities?
* Credit Card - Does the customer use a credit card issued by Universal Bank?

Interval variables are as below:

* Age - Age of the customer
* Experience - Years of experience
* Income - Annual income in dollars
* CCAvg - Average credit card spending
* Mortgage - Value of House Mortgage

Ordinal Categorical Variables are:

* Family - Family size of the customer
* Education - education level of the customer

The nominal variable is:

* ID
* Zip Code

**Insights from Other Factors:**

* No Family members have any influence in taking the personal loan.
* Majority of the users having the online banking were given loans.
* Loans were not given to those with Education type as undergraduate.
* Customers who does not have CD account, does not have loan as well. This seems to be the majority.

**Data Pre-processing:**

**Data Cleaning: Experience** feature is normally distributed with more Customers having experience starting from 8 years. There are negative values in the **Experience**. This could be a data input error as in general it is not possible to measure negative years of experience. So we have replaced the negative year with the positive year.

Ex: - -1 year to +1 year.

**Data Transformation:** As we have different features with different unit values to build our models we need to scale it at same level. We made use of **StandardScaler ()** and made it unit less. This was done only for the **five** numerical features.

**Applying Models**

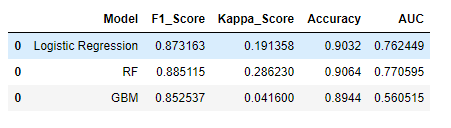
We have split the data into train and test as 70% and 30% respectively.

**Models**

* Logistic Regression: Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables.
* Random Forest: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees
* Gradient Descent: Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting

**Model Performance Measures**

The metrics we chose to compare the model performance were - F1-score, Cohen’s Kappa score, Accuracy and Area under the Curve (AUC)

****

**Conclusion:**

* Let’s have a look at the base models - Log Reg, Random Forest & Gradient Descent. RF gives the best performance amongst all the base models.