**Business Case Study on Bank Loan**

By

Lavina D

Makarand B

Shiva A

**Problem Statement**

A small business loan issuer wants to improve the acquisition of customers to his personal loan scheme from his current 5000 customer base.

**Business Analysis**

* After further research and analysis of the problem, we determined that we were primarily concerned with retention of existing customers and acquisition of new customers to the loan scheme.
* We further divided customer acquisition and customer retention into issues and sub-issues using an issue tree.
* Using the pyramid model (what? , why? , how?), we provided recommendations on how each sub-issue can be resolved. We used the ‘ghost pack’ approach for this.
* To view the presentation, please view ‘Business Case Study - Loan’ PowerPoint file.

**About the Data**

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 to better target marketing to increase the success ratio with a minimal budget.

Out of the **5000** customers, **480 customers** were took the loan.

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.

|  |  |
| --- | --- |
| 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) |
| ZIPCode | 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? |
| CreditCard | Does the customer uses a credit card issued by UniversalBank? |

**Information on the features:**

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 have 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 uses a credit card issued by UniversalBank?

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

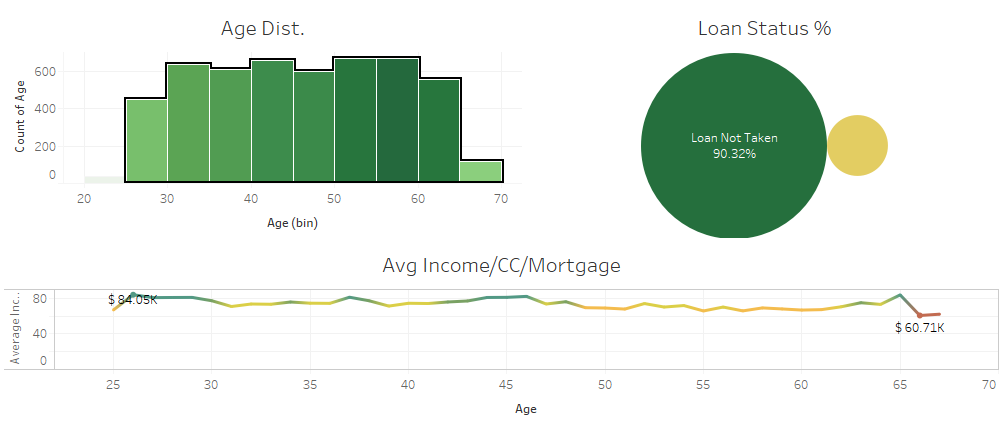
Kaggle Dataset Link: <https://www.kaggle.com/itsmesunil/bank-loan-modelling>

**EDA using Tableau**

We chose to do exploratory data analysis using Tableau as it is more dynamic and user friendly than Python for data visualization.

Below are the dashboards that we made and insights we found using Tableau.

Monetary Factors:



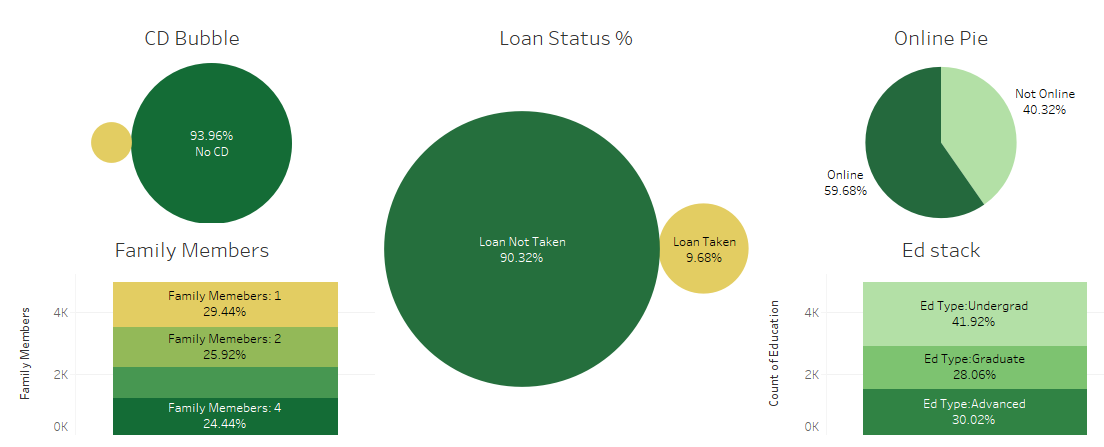
Insights from Monetary Factors:

1. People between the ages 25 yrs - 65 yrs were primarily targeted.

2. Across target age groups, people having high avg. income were targeted.

3. Across target age groups, people having high credit card spending on average were targeted.

Other Factors:



Insights from Other Factors:

1.No . Family members does not have any influence in taking the personal loan.

2. Majority of the users having the online banking were given loans.

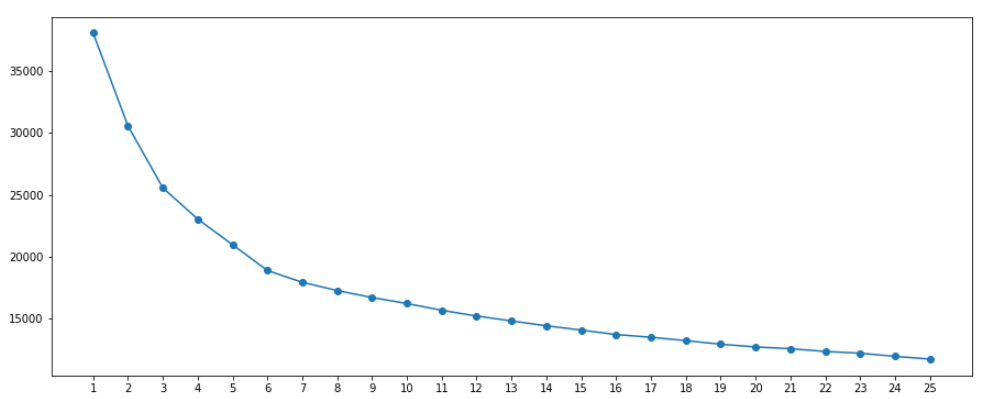
3. Loans were not given to those with Education type as undergraduate.

4.Customers who does not have CD account, does not have loan as well . this seems to be the majority.

Tableau Data Visualization Link: <https://public.tableau.com/profile/makarand3796#!/vizhome/Book1_15605198955130/Story1>

We further continued EDA. We wanted to use unsupervised learning and split the customers into clusters and then check how many customers from each cluster were targeted.

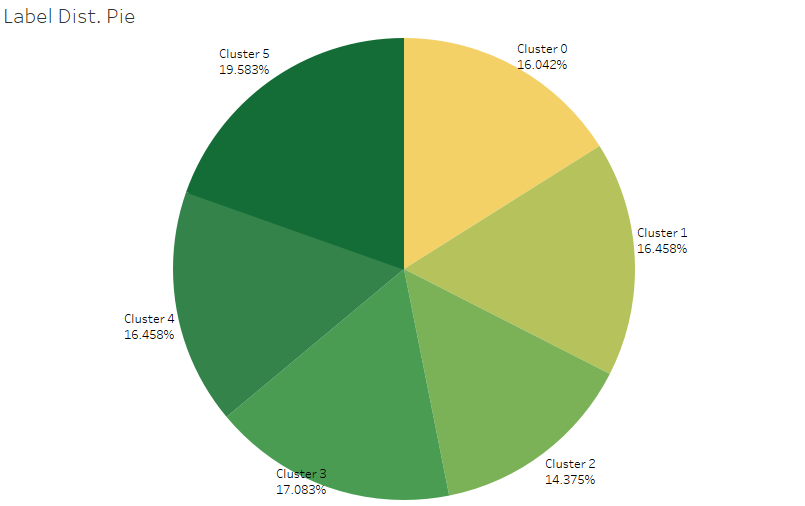
We used elbow method to find out the optimum number of clusters. Below is the screenshot of elbow plot visualisation.



We could see that for k=6 (no. of clusters) there was a shift in error and beyond k=6 error became constant.

We used KMeans Clustering to segregate the data into **six clusters**.

Using Tableau, we found out that each cluster was similarly targeted. Each cluster had approximately **15%** of target customers according to our previous insights for monetary and other factors.

****

**Data Preprocessing**

**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.

**Feature Engineering**

* In our data set we have 7 categorical variable existing.With the help of feature engineering we have extracted new features on the basis of existing dataset.
* As we have a wide spread of data features such as(Age,Experience,CCAvg,Income) we have created categories for each feature so as to understand the data and to predict the output.
  + **Age** - Young Adult, Middle Aged, Senior Citizen
  + **Experience** - Low, Medium, High,Very High.
  + **CCAvg-** Low, Medium, High
  + **Income-** Low ,Medium, High
* To compare the effect of feature engineering on the performance of the model, we have built certain model with and without feature engineering.

**Note:** This was done only for comparison. Therefore, this comparison was followed only for the base models and not for ensemble techniques.

**Applying Models**

We have split the data into train and test as 70% and 30% respectively with the random state as 100 .

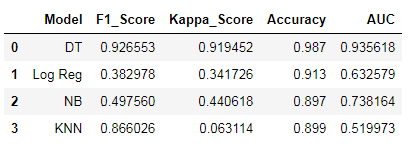
**Models**

**Without Feature Engineering**

1. Decision Tree
2. Logistic Regression
3. Naives Bayes
4. K-Nearest Neighbors

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

Below is the comparison of model performances **without feature engineering**

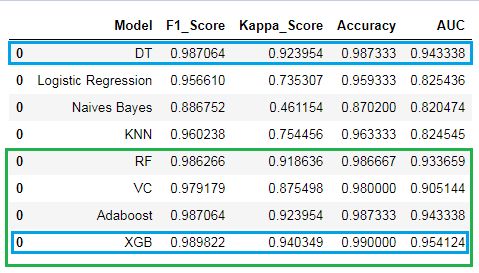


**With Feature Engineering**

1. Decision Tree
2. Logistic Regression
3. Naives Bayes
4. K-Nearest Neighbors
5. Random Forest
6. Voting Classifier
7. Adaboost
8. Extreme Gradient Boosting (XGB)

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

Below is the comparison of model performances **with feature engineering**

****

**Conclusion**

* Compared to the models without feature engineering, models using data with feature engineering perform better.
* Let’s have a look at the base models - DT, Log Reg, KNN and NB. DT gives the best performance amongst all the base models. Coming to the ensemble models - RF, VC, Adaboost and XGB, we can see that XGB gives the best performance.
* In between DT and XGB, we would recommend using XGB. The DT model built does not take care of variance error as we are not sure of its performance in production. Hence, we would recommend using XGB. Being an ensemble model, XGB takes care of bias and variance error. This would give the business confidence in deploying this model in production as it would perform relatively better compared to DT.