**BANK LOAN MODELING**

**Context:**

This case is about a bank (Thera Bank) which has a growing customer base. Most of these customers belong to the liability customers category or they are also called depositors. The total number of customers who also borrow is very small. They are also called asset customers. The bank wants to expand its business rapidly in order for the increase of the loan business and so during the process, earn more through the interest on loans. Particularly, the management wants to explore ways in which it is able to convert its customers (liability) to personal loan customers (while also retaining them as depositors). The bank ran a campaign last year to check for liability customers.

The conversion rate is of 9% (or 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.

The department wants to build a model that can be used for the identification of potential customers, those who have a greater probability of taking the loan. This in turn will increase the success ratio alongside by decreasing the costs that are made for campaigning.

**AIM:**

To predict the likelihood of a liability customer (all the customer data given is assumed to be the of the liability customers) buying personal loans.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data Description:** |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ID | Customer ID | |  |  |  |  |  |  |
| Age | Customer's age in completed years | | | |  |  |  |  |
| Experience | #years of professional experience | | | |  |  |  |  |
| Income | Annual income of the customer ($000) | | | | |  |  |  |
| ZIPCode | Home Address ZIP code. | | |  |  |  |  |  |
| Family | Family size of the customer (1,2,3,4) | | |  |  |  |  |  |
| 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 use a credit card issued by UniversalBank? | | | | | | |  |

**Relating the features, their types and their significance to the problem:**

**ID:** Categorical Nominal discrete (not a factor)

**Age:** Numerical Ratio Continuous(Significant factor) (Scale: years) (Range: 26yrs to 65yrs)

**Experience:** Numerical Ratio Continuous (Significant factor) (Scale: years) (Range: 0 to 41)

**Income:** Numerical Ratio Continuous (Significant) (Scale: years) (Range: 60000$ to 203000$)

**Zip Code:** Categorical Nominal Discrete (Invalid factor) (Scale: years)

**Family:** Numerical Ratio Discrete (Significant Factor) (Scale: years) (Range: 1,2,3,4)

**CCAvg:** Numerical Ratio Continuous (Significant Factor) (Scale: years) (Range: 0$ to 10,000$)

**Education:** Categorical Ordinal Discrete (Significant Factor) (level) (Range: 1,2,3)

**Mortgage:** Numerical Ratio Continuous (Significant Factor) (Scale: years) (Range: 0$ to 635,000$)

**Personal Loan:** Binary **(Classification variable to be predicted)** (Scale: years) (Range:1,0)

**Securities Account:** Binary (nominal) (Significant Factor) (Scale: years) (Range: 1,0)

**CD Account:** Binary (nominal) (Significant Factor) (Scale: years) (Range: 1,0)

**Online:** Binary (nominal) (Could be a significant factor) (Scale: years) (Range: 1,0)

**CreditCard:** Binary (nominal) (Significant Factor) (Scale: years) (Range: 1,0)

#### **On the dataset:**

The data is collected for 5000 customers. It includes details about customers demographic information such as age and income, the relationship of the customer with the bank (what type of account? mortgage, securities account, etc.), and also the response of the customer to the campaign (Personal Loan).

It is given that among these 5000 customers, there are 480 (= 9.6%) who took the personal loan that was offered in the last campaign.

#### **Information on the features or attributes:**

The attributes can be divided accordingly :

* The variable **ID** does not add any interesting information as it is just a unique number with no significant meaning. Hence it is observed that there is absence of an association between a person's customer ID and loan. It also does not give a general conclusion which is useful for future potential customers who are interested in personal loan. We can neglect this information for our model prediction.

The binary category have five variables as below:

* Personal Loan – Whether the customer accepted the personal loan that was offered in the bank’s last campaign? **This is our target variable**
* Securities Account – Whether the customer has a securities account registered with the bank?
* CD Account – Whether the customer has a certificate of deposit (also called CD) account registered with the bank?
* Online - Are internet banking facilities used by the customer?
* Credit Card - Whether the customer use a credit card issued by UniversalBank?

Interval variables are as below:

* Age – The age of the customer
* Experience – Number of years of experience the customer has
* Income - Annual income given in US dollars
* CCAvg - Credit card spending on an average in US dollars
* Mortage - Value of House Mortgage

Ordinal Categorical Variables are:

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

The nominal variable is :report

* ID
* Zip Code

**PROCESS:**

NOTE: Refer to the image “data eyeballing.png” for the below points:

* The statistical points such as mean , standard deviation, min, max, 3 quartiles are obtained for a better understanding of the scale of the data.
* The data is then processed to find the unique values in each column

NOTE: Refer to the image “pair plot.png” for the below points:

* **Age** feature is normally distributed with majority of customers falling between 30 years and 60 years of age. We can confirm this by looking at the describe statement above, which shows **mean** is almost equal to **median**
* **Experience** is normally distributed with more customer having experience starting from 8 years. Here the **mean** is equal to **median**. There are negative values in the **Experience**. This can be an input error because it is not possible for the years of experience to be negative. We can delete these values, because we have 3 or 4 records from the sample.
* **Income** variable is observed to be positively skewed. Majority of the customers are observed to have income between 45K and 55K. We can confirm this by saying the **mean** is greater than the **median**
* **CCAvg** is also a positively skewed variable and average spending is between 0K to 10K and majority spends less than 2.5K
* **Mortgage:** It is observed that 70% of the individuals are having a mortgage which is less than 40,000. Note that the max value is observed to be 635,000.
* The variables “family” and “education” fall under ordinal variable category. The distribution of families is evenly distributed.
* The records with negative experiences are improper values for the experience attribute which are 52 in number. They are replaced by the median positive value.

**OBSERVATIONS:**

* It seems the customers whose education level is 1 is having more income. However customers who has taken the personal loan have the same income levels. **(“income and education on personal loan.png”)**
* It is observed that customer who did not take personal loan and the ones that have taken it have high mortgage**. (“personal loan and high mortgage.png”)**
* Also, majority of bank customers who did not take the loan have securities accounts with the bank **(“loan and securities account.png”)**
* Family size, when observed with respect to personal loan, does not have any impact. But it appears that those families with a size of three are more likely to opt for loan. When considering future campaign this might be good association. **(“Family size and loan.png”)**
* It is also observed that most of the customers who has a CD account has taken loan as well. **(“CD account and loan.png”)**
* The graph show persons who have personal loan have a higher credit card average. Credit card spending (Average) with an observed median of 3800 dollar indicates personal loan probability to be higher. At the lower end, Credit card spending (average) with a median of 1400 dollars has less probability to opt loan. This information is due noted.**(“ CCAvg.png”)**
* The above plot show with experience and age have a positive correlation. As experience increase age also increases. Also the colors show the education level. There is gap in the mid forties of age and also more people in the under graduate level. **(“Exp vs Age.png”)**
* Income and CCAvg is moderately correlated.
* Age and Experience is highly correlated **(“Corr with heat map.png”)**
* A score of 0.8666666667 is achieved on naïve bayes model
* On Random Forest the significance of the variables is given as shown below:

A screenshot of a cell phone

Description automatically generated

Also, a score of 0.89933333 is achieved on random forest.

* Whereas 0.910607071 is achieved on KNN

**Model Comparison:**

A screenshot of a cell phone

Description automatically generated

### Conclusion:

The Universal bank has the aim to convert their liability customers into loan customers. In order to plan for a new campaign they need information about the relation between the variable in the collected data. Four classification algorithms were used in this study and each of them are provided with a score. They are also compared with respect to better predict (which is the actual result). From the above graph , it is evident that the **Decision Tree** algorithm shows the highest accuracy when compared to a random prediction and other models used, and we can opt for the same as our final model to predict better the personal loans.

RESOURCES:

<https://www.utc.edu/>

https://dataminingworld.wordpress.com/