DePaul University

Final Report

Personal Loan Campaign

Amy Aumpansub

DSC 478: Programming Machine Learning Applications

Professor Aleksandar Velkoski

March 19, 2019

Abstract

Personal loans have lower fundraising costs and are more adaptability to individual needs. Bank plans to expand their personal loan business by launching the campaign to specially offer personal loans to their target customers. During the last campaign, bank contacted 5,000 customers to offer the personal loan. However, the last campaign showed only 9.6% success rate, in which only 480 customers accepted the personal loan offered by the bank. So, the bank needs to improve their campaign and better identify potential customers to increase the success rate of their campaign. This paper is a study in the application of data mining to build strong classification models to identify the important attributes of customers and to predict whether or not the customers will accept the personal loan offered by the bank during a campaign period. Using classification analysis as a framework, 13 characteristics of customers including age, education level, family size, credit card spending, value of house mortgage were considered in this study. To develop the strong predictive model, we performed various supervised learning algorithms including Decision Tree, Naive Bayes (Gaussian) classifier, Linear Discriminant Analysis, and K-Nearest Neighbors Classification. Several Python libraries including Pandas, Matplotlib, NumPy, sklearn, and imblearn were utilized in different stages of the KDD process. All developed models have a high accuracy rate, measured by index of balanced accuracy and balanced accuracy score which take an imbalanced class into account. The final model was selected based on its high accuracy and ability to predict the campaign's outcome. Our study found that from 13 characteristics of customer, the income and educational level attributes have more influence on whether or not the customer will accept the personal loan offer after getting contacted during the campaign period.

Objective

The goal of this project is to develop strong predictive models to predict campaign outcomes which determine whether or not the customers will respond to bank's personal loan campaign by accepting the loan offered during the campaign period. Additionally, several classification models are developed to predict campaign outcomes and to identify the important characteristics of bank's customers who are likely to accept the personal loans offered by the bank.

Problem Description

The bank performed a marketing campaign to offer their customers a personal loan. However, the study shows that the last campaign had a low success rate. Therefore, bank needs to better identify important attributes of clients who are likely to accept their offer, so that they can improve the success rate of their campaign and focus their marketing efforts on those specific customers. We analyze data to address the following questions:

- Which types of customers are more likely to accept the personal loans offered by the bank?
- Which independent attributes have a high influence on target variable?
- What are the potential campaign outcomes predicted by our classification models?

Dataset

The "Bank Personal Loan" dataset contains 5,000 observations, retrieved from https://www.kaggle.com/itsmesunil/bank-loan-modelling. The dataset has 13 independent attributes including 7 numeric attributes such as age, income, and the average monthly spending on credit cards, and 2 categorical attributes such as education level and 4 binary attributes.

Dependent Variable: Did client accept the personal loan offer? (binary variable: yes, no)

Independent Variables:

	Name	Attribute	Descriptions
1	ID	numeric	Customer ID
2	Age	numeric	Customer's age in completed years
3	Experience	numeric	#years of professional experience
4	Income	numeric	Annual income of the customer ($000)
5	Area	categorical	Rural/Urban Areas
6	Family	numeric	Family size of the customer
7	CCAvg	numeric	Avg. spending on credit cards per month ($000)
8	Education	categorical	1: Undergrad; 2: Graduate; 3: Advanced/Professional
9	Mortgage	numeric	Value of house mortgage if any. ($000)
10	Securities Account	binary	Does the customer have a securities account with bank?
11	CD Account	binary	Does the customer have a CD account with the bank?
12	Online	binary	Does the customer use internet banking facilities?
13	CreditCard	binary	Does the customer use a credit card issued by bank?

Methodology

The classification algorithms were used to solve this supervised learning problem to predict target (binary) attribute. The models were developed using Decision Tree, Naive Bayes (Gaussian) classifier, Linear Discriminant Analysis, and K-Nearest Neighbors Classification. Pandas, Matplotlib, NumPy, sklearn, and imblearn were utilized in this study. The following steps were implemented for this project:

- Explored the distribution and basic statistics and visualize data of all attributes
- Preprocessed data by dropping some attributes, normalizing numeric attributes, and creating dummy variables for categorical attributes
- Performed Principal Component Analysis and Correlation Analysis to identify the association among customers' characteristics and target attribute
- Split data into a training set (80%) for fitting model and a test set (20%) for evaluation
- Developed several models and perform model evaluation using metrics that take an imbalanced class into account to select the final model for making a prediction

Data Exploration

Basic Statistics

The dataset contains 13 attributes of customer's characteristics including age, education level, years in work experience, family size, averaged monthly spending on credit card, value of house mortgage, and the use of online services, etc. We examine the basic statistics as following:

- The basic statistics table are shown in Appendix A.1.1. The range of customers' ages is 23 - 67 years old with the mean of 45.33 years and standard deviation of 11.46 years.
- The range of years in work experience is 0 to 43 years with the mean of 11.44 years.
- The income (in $ thousand) ranges from $8 thousand to $224 thousand. The income has the greater mean of $73 thousand than the median of $64 thousand. The distribution is right-skewed. The standard deviation of income is $11.46 thousand.
- The range of family size is 1 to 4 members with the mean of 2.3 members and the standard deviation of 1.14 members

Data Distribution

Numeric variables

In addition to the basic statistics, we examine the data distribution by creating histograms with 9 bins as shown in Appendix A.1.2.1.

- The Age distribution is normal. The range of ages is 23 - 67 years with mean of 45.33 years.
- The Education attribute has multimodal distribution with several peaks.
- The income distribution is right-skewed, as majority of customers have income less than the median and the mean or less than the average income of $73 thousand.
- The CCAvg distribution is also right-skewed, so most people spend less than the average spending on credit card of $1.93 thousand.
- The value of house mortgage distribution is right-skewed, so most people have less value of mortgage than the average value of $56 thousand.

Categorical and Binary Variables

We examine these attributes by plotting frequency of data in each category (Appendix A.1.2.2).

- The majority of customers about 97% live in urban areas.
- 41% of customers have bachelor's degree and 30% of customers have Advanced/Professional degree. Only 28% customers have master's degree.
- 89% of customers do not have securities account and 93% do not have CD account.
- The majority of customers about 59% use internet banking facilities.
- The majority of customers about 70% do not use a credit card issued by the bank.

Compare and contrast the subsets of customers with target variable

The cross tabulation (Appendix A.1.2.3) was performed to compare the subsets of customers with our target variable (Personal Loan).

- There are greater number of customers who didn't accept personal loan or about 90.4% of total bank customers. Only 9.6% of customers accepted the offer. Thus, we have an imbalanced class, which will be considered when measuring accuracy of models.
- The region graph shows similar trend regarding Personal Loan. The majority of customers who accepted and did not accept Personal Loan Offer live in the urban areas.
- The majority of customers who accepted the personal loan offer have a master's degree.
- The majority of customers who did not accept the offer have a bachelor's degree.
- The Securities and CD Account graphs show similar trend. The majority of customers who accepted and did not accept personal loan offer do not have Securities and CD Account.
- The majority of customers who accepted and did not accept personal loan offer use internet banking facilities. The majority of customers who accepted and did not accept personal loan offer do not use a credit card issued by the bank.

Data Preprocessing

Data Transformation

- The original data does not contain missing values.
- The original categorical attributes including “Area” and “Education” were transformed to dummy variables using get_dummies function from Pandas (Appendix A.2.1).
- 6 numeric attributes were transformed using min-max normalization. After Normalization, all values of 6 numeric attributes are in a range between 0 and 1.
- The correlation analysis in Appendix A.2.2 shows that age and work experience variables are strongly correlated ($p = 0.99$), so we dropped “Experience” variable as it can cause multicollinear problem (with Age) and repeated info.
- The ID attribute was dropped as it did not provide the useful information for analysis.
- The PCA was also performed (Appendix A.3.4.2) to reduce dimension, which 9 components explain 95.5% of variance. We used these 9 features to fit the KNN model 2.

Training and Test Sets

- We separated x and y attributes from original data (Appendix A.2.3).
- The original data was spitted into 80% training set and 20% test set using train_test_split function from sklearn.model_selection with the random_state of 33 (Appendix A.2.4).
- The training set contains 4000 samples and the test set contain 1000 samples. We used the training set to fit the classification models in the next section and used the test set for calculating accuracy and evaluating our models.

Model Development I

Decision Tree

Total observations were spitted into a training set and a test set. A training set contains 4000 records or 80% of the total records. A training set is used to build 3 Decision Trees using DecisionTreeClassifier function from sklearn.tree module (Appendix A.3.1). A test set contains 1,000 records or 20% of total records, used for model evaluation. The accuracy was calculated using "balanced_accuracy_score" from sklearn and "classification_report_imbalanced" from imblearn library. Both metrics take an imbalanced class into account.

- The decision trees were built using “Personal Loan” as a target attribute (Did client accept the personal loan offered by the bank?). The independent variables include Age, Income, Family, CCAvg, Mortgage, SecuritiesAccount, CDAccount, Online, CreditCard, Area_Rural, Area_Urban, Education_1, Education_2, and Education_3.
- The three different decision tree models were developed by changing the number of minimum records per node, the maximum level of depth, and the criterion. The performance of each tree is presented in model evaluation metrics below.
- All three models have a high accuracy rate and specificity rate, indicating that the models can accurately recognize the “no” class. However, the balanced accuracy score and index of balance accuracy are lower because it penalizes score for an imbalanced class.

Model 1

The first model (Appendix A.3.1.1) was developed using the default values which using “gini” index. The default values did not set the minimum records per nodes and did not specify the maximum level of depth, so the first model can overfit the data. The first tree indicates that income attribute is the most influential attribute. The balanced accuracy rate of 93% and the index of balance accuracy of 88% imply that the model 1 has a good performance in prediction.

Model 2

The second model (Appendix A.3.1.2) was developed using “entropy” as criterion with the maximum depth of 3, and minimum records per nodes of 20 to avoid over fitting the data. The second tree also indicates that income attribute is the most influential attribute. The balanced accuracy rate of the second tree is relatively high, but slightly lower than the first and third tree.

Model 3

The third model (Appendix A.3.1.3) was developed using entropy as criterion with the maximum depth of 4, and minimum records of 20 to avoid over fitting the data. The third tree indicates that income attribute is the most influential attribute. The model also has high balanced accuracy rate.

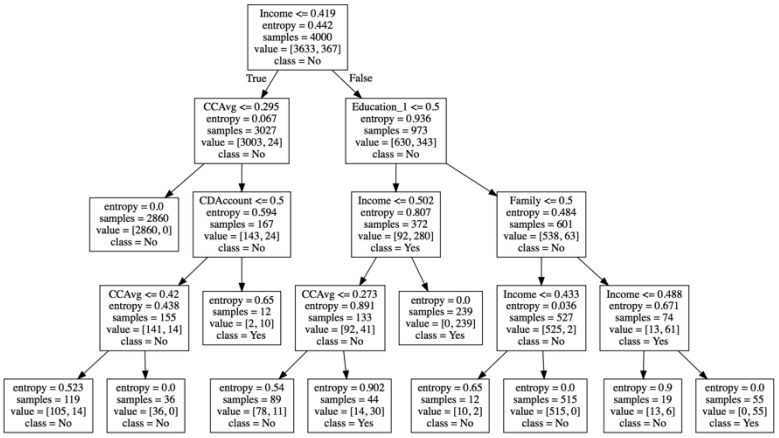
Decision Tree	Precision	Recall	Specificity	f 1	Balanced Accuracy Score	Index of Balanced Accuracy
1	98.0%	98.0%	90.0%	98.0%	94.0%	88.0%
2	97.0%	97.0%	83.0%	97.0%	90.0%	82.0%
3	98.0%	98.0%	87.0%	98.0%	92.0%	86.0%

Figure 1: Decision Tree Model Evaluation

Decision Tree (Final Model)

The final Decision Tree model is Model 3 because it has the high balanced accuracy rate and index of balanced accuracy. It was built with the 'entropy' criterion, the minimum number of records of 20, and the maximum level of depth of 4 (Appendix A.3.1.3).

- The dependent variable is a binary variable (Did client accept the loan offer?). The independent variables include Age, Income, Family, CCAvg, Mortgage, SecuritiesAccount, CDAccount, Online, CreditCard, Area_Rural, Area_Urban, Education.
- The most three important attributes are Income, Education Level, and Family Size.
- The least important attributes are Age, Online, Area, and Credit Card.
- The confusion matrix (figure 3) shows that final model has a good performance in prediction.
- The balanced classification matrix is shown in figure 4. The final model has a high balanced accuracy rate of 92 % and a higher index of balanced accuracy for class 0, indicating that it can better recognize the negative records, but the model has a high (balanced) accuracy rate (83%), so it is also good at predicting class 1. This rate takes an imbalanced class into account. Overall (balanced) accuracy rate is 92.3% implying it is a strong predictive model.
- The final model contains 21 nodes, which 11 nodes are terminal nodes which generate 11 classification rules (figure 2). For example, Rule 1: if the income is less than 0.419 ($93.8K) and the average monthly spending on credit card is less than 0.295 ($2.95K), the customer will not accept the personal loan offer.



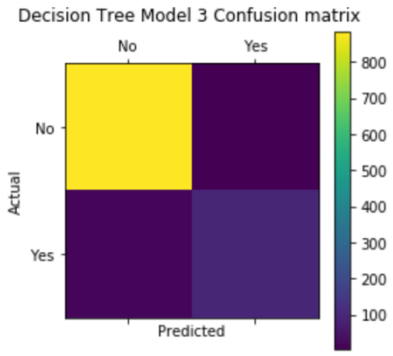


Figure 2: Decision Tree Model 3 (Final Model)

Figure 3: Confusion Matrix (Model 3)

	pre	rec	spe	f1	geo	iba	sup
0	0.98	1.00	0.85	0.99	0.92	0.86	887
1	0.97	0.85	1.00	0.91	0.92	0.83	113
avg / total	0.98	0.98	0.87	0.98	0.92	0.86	1000

Overall Accuracy Rate on Test Set : 0.9231

Figure 4: The Balanced Classification Matrix (Model 3)

Model Development II

Naive Bayes classifier (Gaussian)

The Naive Bayes model was built with Gaussian distribution using naive_bayes function from sklearn module (Appendix A.3.2)

- The model was built using "Personal Loan" as a target attribute (Did client accept the personal loan offered by the bank?). The independent variables include Age, Income, Family, CCAvg, Mortgage, SecuritiesAccount, CDAccount, Online, CreditCard, Area_Rural, Area_Urban, Education_1, Education_2, and Education_3.
- Taken an imbalanced target class into account, we used "balanced_accuracy_score" from sklearn and "classification_report_imbalanced" from imblearn to measure performance.
- As shown in figure 5, the Naive Bayes model has a moderate accuracy rate (balanced) of 75% and relatively low index of balanced accuracy (iba) of 56%. This implies that the model does not have a good performance in prediction. This is because the model underfitted the data and assumed normal distribution despite the fact that several attributes have right-skewed distribution. Thus, this model is not be suitable for our analysis.

	pre	rec	spe	f1	geo	iba	sup
0	0.95	0.91	0.60	0.93	0.74	0.56	887
1	0.45	0.60	0.91	0.52	0.74	0.53	113
avg / total	0.89	0.87	0.64	0.88	0.74	0.56	1000

Overall Accuracy Rate on Test Set : 0.7547

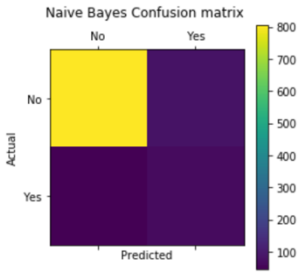


Figure 5: The Balanced Classification and Confusion Matrix (Naive Bayes)

Linear Discriminant Analysis

The LDA model was built using LinearDiscriminantAnalysis function from sklearn module. (Appendix A.3.3). The model was built using same training and test sets as other classification models. The balanced accuracy rate of 80% is higher than NB model but still lower than Decision Tree and KNN models. Its index of balanced accuracy (iba) of 64% is relatively low. From balanced classification matrix in figure 6, we can notice that this LDA model does not perform well in predicting class "1". So, this model was not selected as our final model.

	pre	rec	spe	f1	geo	iba	sup
0	0.95	0.98	0.64	0.97	0.79	0.64	887
1	0.79	0.64	0.98	0.71	0.79	0.60	113
avg / total	0.94	0.94	0.68	0.94	0.79	0.64	1000

Overall Accuracy Rate on Test Set : 0.8079

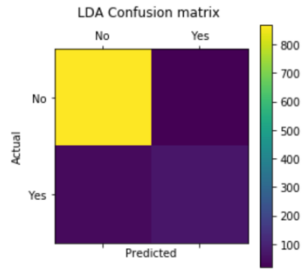


Figure 6: The Balanced Classification and Confusion Matrix (LDA)

Model Development III

K-Nearest Neighbors

KNN Model 1

Total observations were spitted into a training set and a test set. A training set contains 4000 records or 80% of the total samples. A training set is used to build KNN model using KNeighborsClassifier function from sklearn.neighbors module (Appendix A.3.4.1). A test set contains 1,000 records or 20% of total records, used for model evaluation. The accuracy was calculated using "balanced_accuracy_score" from sklearn and "classification_report_imbalanced" from imblearn library. Both metrics take an imbalanced class into account.

- The KNN model was built using “Personal Loan” as a target attribute. The numeric variables were normalized using min-max normalization to remove an effect of the differences in scale and prevent distance measures from being dominated by larger values.
- To select the best k, the KNN models were developed with both Euclidean and Manhattan distances selected range of k from 1 to 20 where k is the number of the nearest neighbors used to classify the new records. The best k from our experiment is k = 3 from Euclidean distance. It has the highest balanced accuracy rate of 0.811905 (figure 7).
- The KNN model 1 was fit with the 12 features, k = 3, and metric = Euclidean. The Overall Accuracy Rate on Test Set is 81% and the index of balanced accuracy is 64%. The balanced accuracy score and index of balance accuracy are lower because it penalizes score for an imbalanced class. From the balanced classification matrix in figure 9, we can notice that this KNN model 1 does not perform well in predicting class “1”.
- Comparing to other classification models, this KNN model 1 performs poorer than Decision Tree models, but better than Naïve Bayes and LDA models.

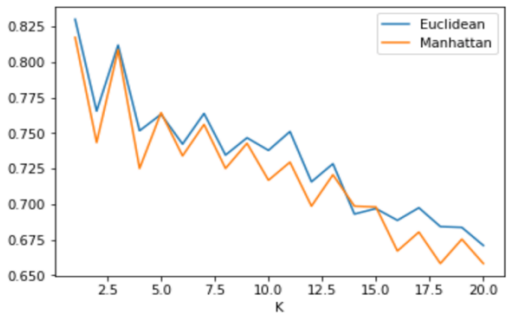


Figure 7: Balanced Accuracy Rate (k = 1 to 20)

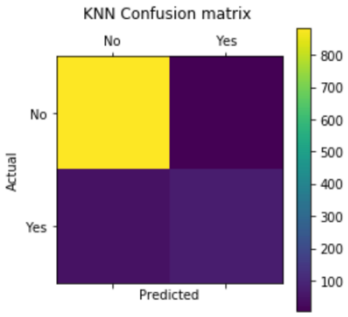


Figure 8: Confusion Matrix (KNN Model 1)

	pre	rec	spe	f1	geo	iba	sup
0	0.95	1.00	0.63	0.97	0.79	0.65	887
1	0.95	0.63	1.00	0.76	0.79	0.60	113
avg / total	0.95	0.95	0.67	0.95	0.79	0.64	1000

Overall Accuracy Rate on Test Set : 0.8119

Figure 9: The Balanced Classification (KNN Model 1)

KNN Model 2

- The PCA was also performed (Appendix A.3.4.2) to reduce dimension, which 9 components explain 95.5% of variance. We used these 9 features to fit the KNN model 2.
- The model was built using "Personal Loan" as a target attribute (Did client accept the personal loan offered by the bank?). The independent variables include 9 PCs. The model was built using a training set (80% of total records) and evaluated with a test set (20% of total records)
- To select the best k, the KNN models were developed with both Euclidean and Manhattan distances selected range of k from 1 to 20 where k is the number of the nearest neighbors used to classify the new records. The best k from our experiment is k = 5 from Manhattan distance. It has the highest balanced accuracy rate of 0.74 (figure 10).
- The KNN model 2 was fit with the 9 features from PCA, k = 5, and metric = Manhattan. Overall Accuracy Rate on Test Set is 74% and the iba is 52%. From the balanced classification matrix in figure 11, we can notice that this KNN model 2 perform poorer than KNN model 1.

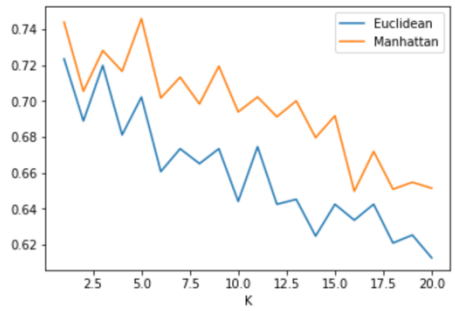


Figure 10: Balanced Accuracy Rate (k = 1 to 20)

	pre	rec	spe	f1	geo	iba	sup
0	0.94	0.97	0.52	0.96	0.71	0.53	887
1	0.69	0.52	0.97	0.59	0.71	0.48	113
avg / total	0.91	0.92	0.57	0.91	0.71	0.52	1000

Overall Accuracy Rate on Test Set : 0.7458

Figure 11: Balanced Classification (KNN Model 2)

Model Evaluation

Model	Precision	Recall	Specificity	f 1	Balanced Accuracy Score	Index of Balanced Accuracy
Tree 1	98.0%	98.0%	90.0%	98.0%	94.0%	88.0%
Tree 2	97.0%	97.0%	83.0%	97.0%	90.0%	82.0%
Tree 3	98.0%	98.0%	87.0%	98.0%	92.0%	86.0%
Naive Bayes	89.0%	87.0%	64.0%	88.0%	74.0%	56.0%
LDA	94.0%	94.0%	68.0%	94.0%	79.0%	64.0%
KNN 1	95.0%	95.0%	67.0%	95.0%	79.0%	64.0%
KNN 2	91.0%	92.0%	57.0%	91.0%	71.0%	52.0%

The balanced accuracy score and iba take an imbalanced class into account. The Naïve Bayes and LDA perform poorer than KNN and decision tree models. The KNN 2 using PCA does not predict the class 1 well. Comparing to all models, the Decision Tree Model 3 performs the best and not overfitting the data, so it was selected as the final model.

Final Model

Final Model is from Decision Tree Model 3 (Appendix A.3.1.3). It is developed by using DecisionTreeClassifier function with inputs as criterion='entropy', max_depth= 4, min_samples_split=20. The final model has the highest overall Accuracy Rate on Test Set of 92% with an Index Balanced Accuracy is 86%. This rate is relatively high as we've already taken out imbalanced target class into account by using "balanced_accuracy_score" from sklearn and using "classification_report_imbalanced" from imblearn library

- The most three important attributes are Income, Education Level, and Family Size (number of members).
- The least important attributes are Age, Online, Area, and Credit Card.
- The final model contains 21 nodes, which 11 nodes are terminal nodes which generate 11 classification rules (figure 2). For example, Rule 1: if the income is less than 0.419 ($93.8K) and the average monthly spending on credit card is less than 0.295 ($2.95K), the customer will not accept the personal loan offer.

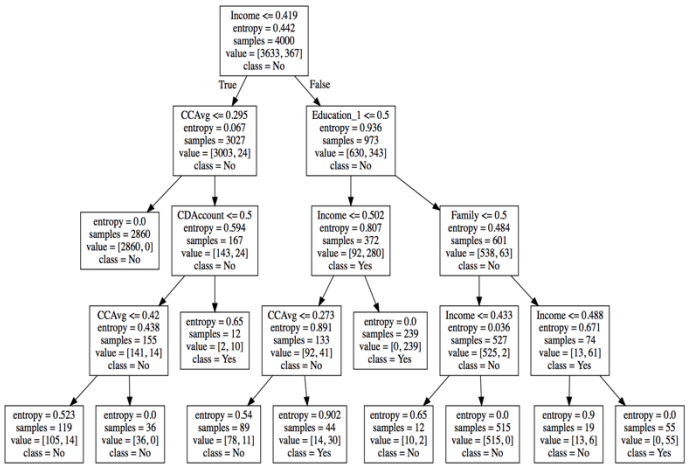


Figure 12: Final Model

Summary

Data mining can provide help in marketing strategies, and its applications can be influential when having complex data and large procedures. In this paper, we use the dataset which collected during the Bank's campaign. The target variable is a binary variable which indicates whether or not the clients accepted the personal loan offered by the bank during the campaign period. The several classification models were built using both Decision Tree, Naïve Bayes, Linear Discriminant Analysis, and K-Nearest neighbors method. All models were evaluated based on their performance on the test set using the index of balanced class and balanced accuracy score. The final model was selected from decision tree which contains 11 terminal nodes. The final model exhibits a high accuracy. We will use the final model to predict the campaign outcome for the new records and to identify important attributes of client's characteristic.

Future Studies

Our study considers only client's characteristics in determining whether or not the clients will accept the personal loan offered by the bank. However, the bank should consider the economic circumstances and economic attributes such as the interest rates, GDP, and unemployment rates. These economic indicators have useful information for our prediction. Moreover, the bank should tailor their campaign to increase the success rate of their campaign which will increase the chance of getting greater number of clients in their personal loan business.