Identifying Customer Subscription Using Predictive Data Mining Techniques

By Ashly Mary Abraham

Prepared for: Dr. Zekiye Erdem

CIND 820 Big Data Analytics Project

GitHub: <https://github.com/frontal24/capstone>

Table of Contents

[Introduction 3](#_Toc57983948)

[Literature Review 3](#_Toc57983949)

[Data Set 4](#_Toc57983950)

[Approach 5](#_Toc57983951)

[Summary 25](#_Toc57983952)

[Conclusion 25](#_Toc57983953)

[References 26](#_Toc57983954)

# Introduction

In this modern world, converting the raw data into useful information has become inevitable in all the businesses. Data mining is one such process where hidden patterns (knowledge) are explored in large datasets amalgamating statistics, ML and databases [1].

It is widely used in businesses, education and research. One of the most benefited sectors is the banking sector where understanding the customer behaviour/data plays a crucial role in expanding the customer base and retaining prevailing customers.

Customers are considered an asset in banks. Customer satisfaction and customer retention goes hand in hand. When it comes to introducing a new product to the customers, existing customers play an important role. Banks have a huge collection of data about existing customers. One can learn how customers responded to different schemes (term deposits/loans/saving accounts etc.) and identify the demographic factor/factors (and any other variable) that influenced their decision making in the past. With effective understanding of customer data and incorporating technology one can not only get new customers but also satisfy and retain the existing customers.

The aim of this project is to identify the group of customers who will most likely open (positive response) a term deposit in a direct marketing campaign. To achieve the goal, data used as the source is associated with the direct marketing campaign of a Portuguese bank. The campaign was mostly done by calls. After the initial preparation of the dataset, classification methods like decision tree and naive bayes will be used on the source to identify customers who responded positively. Decision tree is a common and popular data mining technique that uses the tree hierarchy for data classification and rule inductions [2]. Naive bayes classifier is based on Bayes’ rule to compute conditional probabilities and assumes that attributes are independent [3].

## Literature Review

* The authors explain data mining prediction using the bank marketing dataset as this project. They use Rough set theory and decision trees to achieve the goal. Both are used and comparison is done to understand and explain which works best. They point out that though the decision tree is easy to implement, the rough set theory is more efficient in spite of the issues like memory space needed to build, complexity time for the model, and building Pawlak matrix [2].

* The authors use the Portuguese bank marketing data set for feature selection by using data mining techniques. They have used chi-square and information gain feature selection method to arrive at the best few attributes that could contribute to better accuracy, precision, F1 score. The baseline model is created using a naive Bayes classifier and then they applied the feature selection methods to compare and study the result. They have concluded that the performance of both the feature selection methods is almost the same for the dataset [3].

* The author addresses the problem of class imbalance in the bank marketing dataset. Various classification algorithms are used in WEKA for building model, and to explain the class imbalance problem, the author under samples the majority class, oversamples the minority class and SMOTE techniques are used. Then the author compares various class balancing methods with the baseline model (without addressing the class imbalance) and concludes that SMOTE technique gives better precision, recall, F1 score and ROC area than other techniques and concludes the need for balancing the class in data mining for classification [4].
* The authors uses the same dataset of bank marketing .They use multiple classifiers like Random forest, Support vector machines, K. nearest neighbors, artificial neural network, naïve bayes and class balancing techniques and class balancing techniques like SMOTE, Adaptive synthetics minority (ADASYN) ,Random over sampling techniques(ROS), Adjusting the direction of the synthetic minority class (ADOMS) , The selective preprocessing of imbalanced data (SPIDER) Agglomerative hierarchical clustering (AHC) .After analysis, authors concluded that AHC followed by SMOTE outperformed other class balancing techniques. Naïve bayes and SVM provided better performance metrics than other classifiers and Naïve bayes was chosen among the two because of its simplicity and maintainability. [5]

### Data Set

Data was taken from the University of California, Irvine repository. It has 4119 instances and 21 attributes including the class variable. [6]

The details of the attributes are as follows.

|  |  |  |  |
| --- | --- | --- | --- |
| No | Attribute Name | Type | Description |
| 1 | Age | Numeric(integer) | Age |
| 2 | Job | Categorical (Nominal) | admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, ‘unknown |
| 3 | Marital | Categorical (Nominal) | married, divorced(widow/divorce), single, unknown |
| 4 | Education | Categorical (Nominal) | basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown |
| 5 | Default | Categorical (Binary) | Any default in credit (N/Y) |
| 6 | Housing | Categorical (Binary) | Any H. Loan(Y/N) |
| 7 | Loan | Categorical (Binary) | Any personal loan availed(Y/N) |
| 8 | Contact | Categorical (Nominal) | Telephone, cellular |
| 9 | Month | Categorical (Nominal) | Last contacted month in the year |
| 10 | Day\_of\_week | Categorical (Nominal) | Last contacted day in a month |
| 11 | Duration | Numeric(seconds) | Time spend during the call made for the campaign |
| 12 | campaign | Numeric | Number of contacts performed during this campaign for the particular client |
| 13 | Pdays \*\* | Numeric | Number of days passed by after the client was contacted from prev.campaign |
| 14 | previous | Numeric | Number of contacts performed before this campaign |
| 15 | poutcome | Categorical | Outcome of the previous marketing campaign |
| 16 | emp.var.rate | Numeric | Employment variation rate(quarterly) |
| 17 | cons.price.idx | Numeric | Consumer price index (monthly) |
| 18 | cons.conf.idx | Numeric | Consumer confidence index(monthly) |
| 19 | euribor3m | Numeric | Euribor 3-month rate (daily indicator) |
| 20 | nr.employed | Numeric | Number of employees(quarterly) |
| 21 | y | Categorical (Binary) | Y/N - Did the client open a term deposit or not |

\*\*     Pdays: 999 is mentioned in the dataset indicating that the client was not contacted previously

#### Approach

Initial data preparation and analysis

Exploratory Analysis

Building and evaluating classification model

Conclusion

**Step1: Initial Data preparation and Analysis**

As described above about the dataset, most of the time, the raw data will require some improvement from the way how it is stored. The inaccurate, missing, irrelevant, and extreme values in the dataset are considered as noise and it has to be identified and analyzed if it is real data or misrepresented or it is typological error and accordingly it has to be corrected.

The data set contains 4119 instances and 21 attributes including the class variable. The dataset has a mix of numeric and categorical variables. Below is the table with the changes done in the attribute for analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | Description | Changes done on the attribute | Missing values  Missing: NONE  Total unknown: 1029 |
| Age | Age | Age is widely spread from 18 till 88 hence the age is sub categorised to:   * Teen * younger adult * middle aged * older adult * retired   And made as new column  “age group “  It is then converted to factor for easy visualisation | There is no missing or unknown values in this attribute |
| Job | admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, ‘unknown | The rows with unknown values in (2 attributes),   * Job- default (18 nos) * job- education (13 nos)   are deleted,  Rest of the unknown values are categorised as,   * admin.   Self-employed and entrepreneurs are combined together as “ self -employed “  Unemployed and retired as ‘   unemployed ‘  Services and house maid are combined as ‘services’  This attribute is converted to factor.  With just 7 levels.   * Admin * Blue-collar * Self -employed * Management * Unemployed * Services * student | There are 39 unknowns. |
| Marital | married, divorced(widow/divorce), single, unknown | There are 11 unknown values. Since most of the instances belong to the married category, all the 11 are considered as married.   |  |  |  | | --- | --- | --- | | divorce | married | single | | 446 | 2520 | 1153 | | Unknown: 11   |  |  |  |  | | --- | --- | --- | --- | | divorced | married | single | unknown | | 446 | 2509 | 1153 | 11 | |
| Education | basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown | The categories   * Basic.4y, * Basic.6y, * Basic.9y      are combined together as “basic”  Since most of the instances in this attribute belong to basic education and university degrees,  The unknown values are split into these two groups rather than deleting and losing so many data.  And is converted to factor | There are 157 unknown values. |
| Default | Any default in credit (N/Y) | The unknown values are considered as “No” since many do not have any default.   |  |  | | --- | --- | | no | yes | | 4087 | 1 | | There are 783 unknown values   |  |  |  | | --- | --- | --- | | no | unknown | yes | | 3304 | 783 | 1 | |
| Housing | Any H. Loan(Y/N) | Since both the Y/N categories have a good number of instances, the unknown rows are divided and combined to both the Y/N categories.   |  |  | | --- | --- | | no | yes | | 1878 | 2210 | | There are 105 unknown values   |  |  |  | | --- | --- | --- | | no | unknown | yes | | 1826 | 105 | 2157 | |
| Loan | Any personal loan availed(Y/N) | Most of the instances do not have any loans. Hence all the unknown values are moved to the  “no “category.   |  |  | | --- | --- | | no | yes | | 3454 | 665 | | There are 105 missing value     |  |  |  | | --- | --- | --- | | no | unknown | yes | | 3349 | 105 | 665 | |
| Contact | Telephone, cellular | Changed as factor | ----- |
| Month | Last contacted month in the year | Changed as factor | ----- |
| Day\_of\_week | Last contacted day in a month | Changed as factor | ------ |
| Duration | Time spend during the call made for the campaign | Duration is given in seconds.  For easy understanding, its converted to minutes and imputed as new column   * duration\_min | -------- |
| Pdays \*\* | Number of days passed by after the client was contacted from prev.campaign | This attribute has ‘999’ as values in many instances. Hence, they are changed to -1.  And converted to factor | -------- |
| Y  Target Variable | Y/N - Did the client open a term deposit or not | Converted as factor | --------------- |

A picture containing text, receipt

Description automatically generated

The descriptive statistics of the prepared data is given above. It is a summary of all the features and its statistical measures. It has 21 attributes with no unknown and missing values.

We do have some numerical variables and all the other categorical variables are converted as factors for understanding. The attributes age (imputed as age\_group) and duration (imputed as duration \_min) are deleted from dataset.

**Step:2 Exploratory analysis**

Analysis of attributes is essential for gaining in-depth knowledge about the dataset which comprises visualization and summary statistics. This includes univariate analysis to learn about individual attributes, bivariate analysis to know the relationship between the attribute and the target variable.

Below is the visual study of each attributes.

Class Variable: y

|  |  |
| --- | --- |
| no | yes |
| 3643 | 445 |

Chart, bar chart

Description automatically generated

Based on the above bar graph, it is clear that the data set has most of the customers responding to term deposit campaigns negatively and very few have subscribed towards the deposit.

This indicates an imbalance in the class variable which should be addressed before building the model.

Marital:

Chart, bar chart

Description automatically generated

Many in this data set are married followed by single, and married couples have responded positively for the term deposit followed by single.

It could be because they think for future saving

Job:

Chart, bar chart

Description automatically generated

 Many are in administrative, blue collar jobs and followed by technician and services.

Some administrative job holders and technicians have responded yes to the term deposit campaign call.

 Education

Chart, bar chart

Description automatically generated

Most of the customers are university degree holders and also basic education. Basic education comprises of 4 year, 6 year and 9 year education. The university degree holders have responded well to the campaign

Chart, bar chart

Description automatically generated

Almost all the customers do not have any default in credit, and some have responded yes to the term deposit.

Housing Loan:

Chart, bar chart

Description automatically generated

The customers in the data set have and do not have housing loan and there is no significant difference between both the categories ‘responses for term deposit.

Personal:

Chart, bar chart

Description automatically generated

Many customers in the dataset do not have personal loan. And they have responded better in the campaign than the people who have loan. It could be because of the less liability and extra cash available .

Contact:

Chart, bar chart

Description automatically generated

Many customers seem to have updated their cellular number with the bank and they have been responding yes to term deposit compared to the people with telephone. It is better to obtain everyone cellular number which is easier to reach and update in the database.

Month:

Chart, bar chart

Description automatically generated

The maximum number of customer calls were done in the month of May, followed by July, August, June and November

The customers in the month of may respond well to the term deposit comparatively with the other months mentioned above.

Day of the week:

Chart, bar chart

Description automatically generated

Customers are contacted all the weekdays and the response does not vary significantly based on the day of the call.

Age:

Chart, bar chart

Description automatically generated

The data set has any younger adult followed by middle aged and older adults.

The response from the younger adult seems to be little more compared with the other age group.

Duration (minutes):

Chart, histogram

Description automatically generated

The time taken by the customers for the campaign call varies between 0 - 35 and some beyond and many people have given their response within the 35 min.

Campaign

Chart

Description automatically generated

The number of calls made for this campaign for a particular client ranges from 1 - 20 and some beyond, many have responded positively within 5-6 calls.

Pdays:

Chart, waterfall chart

Description automatically generated

Many customers are non-existent in the previous campaign and they have responded positively to the campaign calls compared to other categories. They are maybe new customers to the bank or did not partake in previous campaign.

Previous:

Chart, bar chart

Description automatically generated

Many are not contacted in the previous campaign and they have responded better towards this current campaign than others. It could be that they are new to the bank.

pOutcome:

Chart, bar chart

Description automatically generated

This indicates the outcome of the previous campaign. Many are non-existent in the previous campaign and they have turned positively towards the current campaign.

Emp.var.rate:

Chart, bar chart

Description automatically generated

Employee variation rate is a quarterly indicator. The majority of positive response are from groups where the rate is between -1 and -2 and some between 1 and 2 .

Cons.price.indx

Chart, histogram

Description automatically generated

The consumer price index is a monthly indicator that ranges from 92- 95 in the dataset and majority of the positive response occurs when the range is between 93-94

Consumer.confi.index

Chart, bar chart

Description automatically generated

Consumer confidence index is a monthly indicator and explains how confident are the people in investing their hard earned money and positive response where given by some when the index was around -46 ,-42.5.

Euribor3m :

Chart, waterfall chart

Description automatically generated

The is a 3month rate daily indicator, explaining the currency value and conversion and some have responded positively towards the campaign when the rate is around 0-2 and 4-5.

nr.employed :

Chart

Description automatically generated

This is the quarterly indicator of the number of employees participating in the current campaign. Some have said yes for the campaign when the value is around 5100, above 5200.

**Correlation analysis:**

 It is a statistical tool that is used to analyse and understand the relationship between two variables, i.e., numeric variables. In short, it is a bivariate analysis. Higher the correlation value, the relationship between the variable is more and vice versa. It is measured on scale between -1 and 1. The attributes are converted to numeric and the analysis is conducted and visualized to determine the variables with higher correlation and hence eliminated accordingly to avoid multicollinearity.

Chart

Description automatically generated

Upon analyzing the above plot, it is observed that the correlation is higher among,

* Emp.var.rate
* Euribor3m
* Duration\_min (based on the below mentioned description from the dataset in repository)

*“Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.”*

The above 3 attributes are removed from the dataset.

**Outlier detection:**

Chart, box and whisker chart

Description automatically generated

Outliers are the datapoints which lies vary far away from the other values. It could be normal value, or it could be a data entry error, or it could be an abnormal value in the dataset.

The box plot indicates the presence of the outliers in the numeric attributes. There are many outliers in the campaign attribute which indicates the number of times a customer is contacted during this campaign for opening the term deposit. Not all outliers are removed from the dataset since its acceptable that some clients need more calls for understanding/response.

The outliers that fall beyond 98 % are removed from the dataset which comprises some of the negative responders.

The total instances of the dataset are reduced to 4002 and 18 attributes (from 4119 and 21 attributes)

**Feature selection:**

The feature selection is a very important step in data analysis, it’s always important to build a model with higher accuracy with the minimum number of features contributing to the response variable. Boruta is a feature selection algorithm. It is a wrapper built around the random forest classification algorithm. It ties to obtain all the important, interesting features with respect to an outcome variable in the data set.

The algorithm goes as follows,

* First, it creates duplicates the dataset and shuffles the value in each column called shadow features
* Then it trains the random forest classifier on the above and evaluate the importance of each attribute
* At each iteration, it checks if the attribute has higher importance (z score) than higher z score of its shadow attributes and removes it if considered unimportant.
* This continues till all the attributes are either rejected or confirmed.

Below is the plot describing the chosen attributes.

**BORUTA FEATURE SELECTION**

Chart, waterfall chart

Description automatically generated

Based on the above, the green box plots indicate the selected attributes, Yellow box plot indicates tentative (one can choose or reject) and red are rejected features, Blue ones are for the shadow variables. Here we have confirmed(green), rejected(red), shadow(blue). We consider all the accepted attributes that are in green.

There are 11 + (1 class variable) attributes selected as important

They are,

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Job | Marital | Contact | Month | Pdays | previous |
| education | poutcome | cons.price.idx | cons.conf.idx | nr.employed | y(target variable) |

**Normalization:**

Normalization is the process of rescaling the attributes in the range of 0 and 1 without varying the difference in the range of values. The minimum value will be 0 and the maximum value will be 1. This is done in our dataset to all the selected attributes. For this purpose, the data set is converted into numeric.

x(norm) = (x - x(min) / (x(max) - x(min)

**Class Balancing:**

As mentioned before, the class variable in the dataset is severely imbalanced. Imbalanced class means one class(Y) exceeds the other (N) by a larger amount. Many ML algorithms results depend on the type of the dataset loaded for building a model. Minority class is the class of interest and we intend better results. With an imbalance dataset, the algorithm does not get all the required information regarding the minority class (in our case, term deposit: yes) to make good predictions and thereby reducing the performance. Most of the predictions pertain to majority class.

Hence it is vital to address the above problem to maximize the performance of the model. The solution that is adhered for this problem is SMOTE. SMOTE is Synthetic Minority Oversampling TEchnique. It generates artificial data by using bootstrapping and K nearest neighbors by creating synthetic observations based on minority class observations.

The SMOTE () technique is applied on the dataset and the final class variable distribution is,

|  |  |
| --- | --- |
| No | Yes |
| 4435 | 4440 |

**Step 4: Building and evaluating classification Model:**

Once the data is clean, analysed, features selected, rescaled(normalized) and balanced, the next step is to build the classification model to address the goal of the project. Two classification algorithms are used here for the analysis. They are decision tree and naive bayes classifiers.

**Splitting the data set:**

 The normalized, balanced dataset is split as training (75 %) set for building the model and test set (25%) for prediction and evaluation. The test and train set are independent of each other, both will not have the same instances. Once split, the training set is used to create the model.

**Decision tree classification:**

Decision Trees (DTs) are a supervised learning method used for classification and also regression. The goal is to make a model that predicts the value of a target variable by learning decision rules concluded from the data features. A decision tree is built top to down from a root node and involves splitting the data into subsets that contain instances with similar values (homogeneous). The root and the internal node contains the attribute test conditions (Y/N) to separate the instances based on the values.

The decision tree is simple and easy to understand, and it has the ability to handle numeric and categorical data.

**Naive bayes classification**

Naïve bayes is a probabilistic machine learning algorithm. It is a classification technique that is based on bayes theorem with the assumption that each feature makes an independent contribution to the target variable. That is, features are not dependent on each other and they contribute equally to the outcome.

Bayes’ Theorem finds the probability of an event occurring given the probability of another event that has already occurred (conditional probability). Bayes’ theorem is stated mathematically as the following equation:

P(A|B) = (P(B|A) \* P(A)) / P(B)

A, B = events

P(A), P(B) = independent probabilities

P(B|A) = conditional probability, likelihood of event B occurring given A true/occurred.

 P(A|B) = conditional probability, likelihood of event A occurring given B true/occurred.

**Model Evaluation:**

Both the models are built using the training dataset (75 %) and now the performance of the model that’s built will be evaluated using few performance metrics on the test set (25 %).

The main metric used for evaluating the model are,

* Confusion matrix and accuracy
* Precision
* Recall
* F1 score
* ROC and AUC

**Confusion Matrix:**

Confusion matrix is a performance measurement of a classification problem with a table of combination of actual and predicted classes.

|  |  |  |
| --- | --- | --- |
| Predicted values | Actual values  positive (0) | Actual values  Negative (1) |
| Positive (0) | True Positive | False positive |
| Negative (1) | False negative | True negative |

|  |  |
| --- | --- |
| Decision Tree | Naïve Bayes |
| |  |  |  | | --- | --- | --- | | Prediction | Reference  0 | Reference  1 | | 0 | 1026 | 74 | | 1 | 82 | 1036 | | |  |  |  | | --- | --- | --- | | Prediction | Reference  0 | Reference  1 | | 0 | 954 | 317 | | 1 | 158 | 793 | |

Based on the above table, we could conclude that Decision tree predicts the test data better than the Naïve bayes classifier. The number of customers who have responded positively is better in decision tree classifier. But lets confirm the same based on the performance metrics.

**Performance metrics Evaluation and comparison:**

Accuracy: Accuracy is the ratio of the number of correct predictions in the entire sample to sum of all the input.

Precision: The denotes the predictive ability of the classifier, i.e.,

Precision = TP /TP +FP

Recall: This means the models ability to find all positives, i.e.,

Recall = TP/TP + FN

F1 score: It is the measure of test’s accuracy [1]. In other words, it is the harmonic mean of models precision and recall.

F1 score = (2 \* precision \* recall) / (precision + recall)

Below is a table of all the 5 metrics for both the decision tree and naive bayes classification models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | precision | Recall | F1 score | AUC |
| Decision Tree | 0.9296 | 0.9266 | 0.9333 | 0.9299 | 0.93 |
| Naive Bayes | 0.7876 | 0.8373 | 0.7144 | 0.7710 | 0.88 |

Choosing a model based on precision and recall sometimes may not give reliable result. Hence F1 score plays an important role and tells the models accuracy, It's the harmonic mean of precision and recall.

Based on the above metrics and the values, it is clear that decision tree classifier performs much better than the naïve bayes classifier.

**Decision Tree:**

Timeline

Description automatically generated

The tree has 9 nodes including the root and 10 leaves. Its observed that campaign and socio economic attributes are the contributors towards predicting the customer subscription.

ROC and AUC:

ROC is Receiver operator characteristic, an evaluator for classification models It is a very good evaluation technique when dealing with imbalance dataset. It’s a probability curve plotted between True Positive rate and false positive rate. AUC is the area under the curve. If AUC is greater than 0.5 and less than 1, this means the model has the ability to classify the positive and the negative class values correctly. Based on our both model, the AUC though both the vales are above 0.5 , the best model is the decision tree classifier.

We have plotted ROC and AUC for both the classifiers,

Chart, line chart

Description automatically generated

The ROC curve of both the plot indicates that decision tree is better performing.

#### 

#### Summary

The bank marketing dataset was obtained from the UCI repository, and studied in-depth about the dataset, their initial data types and its dimensions. Then each attribute distribution is visualized to understand the spread and its relationship with the target variable. The unknown values are deleted/imputed. Outliers are addressed by removing them beyond 98th percentile. It's understood that the attributes month, campaign, previous, poutcome, duration shows customers who are new to this current campaign talking less than 20 minutes between May and July of the year seems to respond to this current campaign for opening the term deposit positively. Though duration plays an important role, it's removed since the outcome is known at the end of the call. The other variables like euribor3m and emp.var.rate are also dropped from the dataset for modelling after correlation analysis because of high correlation coefficient values. For further studies, a feature selection algorithm is applied to obtain the important features to build the model. The final (11 +1) 12 features are selected . Another problem that is addressed is for the class imbalance. Hence SMOTE is applied to the dataset with 12 attributes after feature selection and normalized.

Once the balanced data set is obtained, the train and test set are created for model building and evaluation. The performance of both the models are good but the decision tree seems to perform very well in this dataset with F1 score of 0.9299.

#### Conclusion

Based on the above two models, though both decision tree and naive bayes classifier works really well, the decision tree classifier classifies the response variable more accurately and helps in identifying the customers who will open the term deposit or not.

Of all the 20 features, few economic features and campaign related attributes seem to play a significant role in the identification of customers' response in the decision tree classifier. The attributes are,

* Customer.conf.idx
* Customer price.idx
* Nr.employed
* Poutcome
* previous
* Pdays

Consumer confidence index speaks about how confident the customer(person) in the economy to make investment. This factor cannot be controlled during the campaign.

Consumer price index indicates the variation in the price of items that we use on a daily basis (groceries, transportation medical, etc.) It indicates the ability of the customer to invest in the term deposit. Favorable price index, more savings in the account and customers think about investments but can’t be controlled by the bank employees.

Nr.employed is the number of employee participating in this campaign. This too cannot be controlled by the bank employees as management decides how many employees per branch or in a particular department.

Next, the campaign related features are previous campaign outcome(poutcome) and number of days passed since last call made to that client (pdays) and number of contacts performed before current campaign (previous).

The results of the model indicate that many customers are new to this campaign either new to the bank or did not participate in previous campaign have responded well in subscription. Hence during the conversation with such clients (nonparticipant in previous campaign/new customers), their needs must be understood and explain the details of the term deposit clearly and patiently. In this manner, not only does bank get a new term deposit subscriber but also a long-lasting client for the bank.

It is observed that though the feature selection algorithm considered demographic attributes like marital, job, education as important in this model, it does not play vital role in the prediction of current campaign and term deposit subscription.

In the future,

* Cluster analysis could be applied on the dataset and results can be compared with the above and see if any new feature contributes to the result.
* Different feature selection algorithm like information gain and chi square can be applied and compare the current results.
* Classification can be done with and without feature selection, normalization to see the performance variation

#### References

* [1] <https://en.wikipedia.org/wiki/Data_mining>
* [2] Deposit subscribe Prediction using Data Mining Techniques based Real Marketing Dataset by Safia Abbas, Computer Science Department, International Journal of Computer Applications (0975 – 8887) Volume 110 – No. 3, January 2015
* [3] Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data by Tuba Parlar1, Songul Kakilli Acaravci, Vocational School of Antakya, Department of Computer Technology, Mustafa Kemal University, Antakya, Hatay, Turkey, International Journal of Economics and Financial Issues, 2017, 7(2), 692-696.
* [4] Evaluation of Classification Algorithms with Solutions to Class Imbalance Problem on Bank Marketing Dataset using WEKA by Archit Verma, MTech, Computer Science and Engineering, from United College of Engineering and Research, Prayagraj Uttar Pradesh, India, International Research Journal of Engineering and Technology (IRJET), Volume: 06 Issue: 03 | Mar 2019.
* [5] Visualization and Analysis in Bank Direct Marketing Prediction by Alaa Abu-Srhan1, Sanaa Al zghoul3 Department of Basic Science, The Hashemite University Zarqa, Jordan, Bara’a Alhammad2, Rizik Al-Sayyed4 Computer Science Department University of Jordan Amman, Jordan *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 7, 2019*
* [6] <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>