Analyzing the Bank Marketing Data to Study and Understand If the Customer Will Subscribe a Term Deposit (Fix-Deposit)

Chiranjibi Ghimire
Hood College
Frederick, MD, USA
cg16@hood.edu

ABSTRACT

This project applied different data mining classification techniques to build the model to predict whether the customer will subscribe bank long-term deposit. A Portuguese retail bank was collected data from 2008 through 2013. I will analyze the small set of data related to the bank client based on telephone communication.

The Portuguese Bank had an issue of revenue declined, so they conducted a survey and campaign to identify existing clients that have higher chance to subscribe for term deposit and focus marketing effort of such customers. A customer-based analysis of banking services via Data mining, allows for understanding of the possible effects of the concentration on a wide variety of banking resources into a small group of national enterprises [13]. This kind of data mining projects could be helpful to determine the likelihood of procurement of financial services.

CCS CONCEPTS

• **Information systems** → *Wrappers (data mining)*; *Data mining*; *Data cleaning*; • **Theory of computation** → *Oracles and decision trees*; *Data modeling*; • **Computing methodologies** → *Neural networks*; • **Human-centered computing** → Heat maps; • **Hardware** → Data conversion;

KEYWORDS

Data Mining, Data Correlation, Classification Analysis, Decision Tree, Support Vector Machines, Random Forest Model, Data Visualization, Data Cleaning, Term Deposit, Bank Marketing, Telemarketing

1 MOTIVATION AND DESCRIPTION

There have been revenue decline for Portuguese bank. After Investigating, they found that the root cause- clients are as before not depositing as frequently. So they decided to conduct a marketing research on existing clients. The campaigning generally conducted on either mass campaign, which targets a large population or direct campaign, which targets specific clients. Some research shows that mass campaign is less effective. In contrast, the direct campaign focuses on specific potential clients and is often data driven, and more effective. In the era of big data, it is impossible to scale without data-driven technology and solution. We can make a good decision based on data analytics methodology to suggest marketing manager about effective client selection. The direct marketing research is effective but it has some drawbacks, such as clients worried about privacy and security. It is important to maintain security and privacy to conduct the telemarketing campaign.

The aim of this project is to build a suitable model to predict if the client will subscribe to a "term deposit". In other words, I am trying to recommend the best model with very high accuracy, because it can help the bank to filter clients and use available resources more efficiently to achieve the goal. This data analysis will benefit the business process and stakeholder of the banking and financial industry. The objective of this data analysis project is to demonstrate how the classification and predictive analytics study can be applied to produce real, tangible improvements in a company's business performance. Besides, I would try to identify some important factors that can be helpful to make a decision for managerial team.

The UCI data set of Direct Marketing campaign of a Portuguese Banking institution is used for this project. The dataset examined by this project was collected from a telemarketing campaign by a Portuguese banking institution. Occasionally, customers were contacted more than once, in order to attempt to sell Term Deposit subscriptions. The Bank Marketing dataset includes 4119 records, with 21 observations per record, including numerical and categorical data. Each record includes 20 explanatory observations about the client contacted, and 1 response observation of whether the customer subscribed to a Term Deposit. There are two datasets: 1) bank-additional-full.csv with all examples for 5-year period, 2008 through 2013. 2) bank-additional.csv with 10 percent of the examples (4119), randomly selected from *bank-additional-full.csv*. I am using *bank-additional.csv* data set for this project. The smallest dataset is provided to test more computationally demanding machine learning algorithms.

ExampleSet (41188 examples, 0 special attributes, 21 regular attributes) Filter (41,188 / 41,188 examples): no_missing_attrib...

Row No.	age	job	marital	educati...	default	housing	loan	contact	month	day_of...	duration	campai...	pdays	previous
1	56	housem...	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0
2	57	services	married	high.sch...	unknown	no	no	telephone	may	mon	149	1	999	0
3	37	services	married	high.sch...	no	yes	no	telephone	may	mon	226	1	999	0
4	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0
5	56	services	married	high.sch...	no	no	yes	telephone	may	mon	307	1	999	0
6	45	services	married	basic.9y	unknown	no	no	telephone	may	mon	198	1	999	0
7	59	admin.	married	professi...	no	no	no	telephone	may	mon	139	1	999	0
8	41	blue-co...	married	unknown	unknown	no	no	telephone	may	mon	217	1	999	0
9	24	technician	single	professi...	no	yes	no	telephone	may	mon	380	1	999	0
10	25	services	single	high.sch...	no	yes	no	telephone	may	mon	50	1	999	0

Figure 1: This figure describes the the meta data of the data set. It contains 4119 customer records and 21 columns including categorical and numerical columns.

The 20 explanatory observations contain 4 types of client data. 1) Customer data: age, job, marital status, education, default, housing and loan. 2) Telemarketing data: contact, month, day of the week, and duration. 3) Socioeconomic data: employment variation rate, consumer price index, consumer confidence index, 3 month Euribor rate, and number of employees. 4) Other data: campaign, past days, previous, and past outcome.

The Bank Marketing dataset contains numeric variables and categorical variables, which are useful for classification and predictive analytics. Most of the explanatory variables initially appear useful for prediction of future Term Deposit subscriptions. The output variable (desired targeted output) *Term Deposit* column is a binary *yes* or *no*.

As shown in Fig. 1, following is the definition of input variables:

age - Age of the client- (numeric)

job - Client's occupation - (categorical) (admin, bluecollar, entrepreneur, housemaid, management, retired, selfemployed, services, student, technician, unemployed, unknown)

marital - Client's marital status - (categorical) (divorced, married, single, unknown, note: divorced means divorced or widowed)

education - Client's education level - (categorical) (basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown)

default - Indicates if the client has credit in default - (categorical) (no, yes, unknown)

housing - Does the client as a housing loan? - (categorical) (no, yes, unknown)

loan - Does the client as a personal loan? - (categorical) (no, yes, unknown')

contact - Type of communication contact - (categorical) (cellular, telephone)

month - Month of last contact with client - (categorical) (January - December)

day of week - Day of last contact with client - (categorical) (Monday - Friday)

duration - Duration of last contact with client, in seconds - (numeric) For benchmark purposes only, and not reliable for predictive modeling

campaign - Number of client contacts during this campaign - (numeric) (includes last contact)

pdays - Number of days from last contacted from a previous campaign - (numeric) (999 means client was not previously contacted)

previous - Number of client contacts performed before this campaign - (numeric)

poutcome - Previous marketing campaign outcome - (categorical) (failure, nonexistent , success)

emp.var.rate - Quarterly employment variation rate - (numeric)

cons.price.idx - Monthly consumer price index - (numeric)

cons.conf.idx - Monthly consumer confidence index - (numeric)

euribor3m - Daily euribor 3 month rate - (numeric)

nr.employed - Quarterly number of employees - (numeric)

Output variable (desired target) - Term Deposit - subscription verified (binary: 'yes'or 'no')

2 METHODOLOGY

Since the data set contain both numerical and categorical data, we could use various data mining techniques. This project particularly utilizes the Data Classification Analysis technique to examine a dataset related to direct marketing campaign of a Portuguese banking institution. The objective of classification technique is to predict if the client will subscribe to a Term Deposit. In order to obtain more accurate and precise model to predict desired output, I will perform several classification techniques and model such as Decision Tree, Support Vector Machines, and Random Forest Model.

I will perform correlation analysis to see if there is any relationship between predicted attribute (client subscribe term deposit) and other explanatory attributes. The next method, classification model (decision tree), will be helpful to study the customer pattern and accuracy of the applied model.

After I perform all of the above techniques, I would be able to understand the data and suggest the best fit model for prediction of "customer term deposit" more accurately and precisely.

In order to perform the Data mining process, I use Cross-Industry Standard Process for Data Mining (CRISP-DM) process [17], which includes following steps;

1) Business Understanding

2) Data Understanding and Exploring: Data pre-processing, Cleaning, and Visualization

3) Data Preparation: Remove outliers, Sampling, and Scaling

4) Data Modeling: Splitting into Training and Testing Set

5)Model Evaluation

2.1 Business Understanding

The Portuguese Bank initiated the telemarketing campaign, (that provided the data examined by this document), contacted potential savings account depositors for a 5 year period, 2008 through 2013. This data therefore reflects the influence of the financial crisis of 2008. The aim of this project is to build a suitable model to predict whether a client will enroll for term deposit.

2.2 Data Understanding and Exploring

As described in the section motivation and description, I am using UCI data set of Direct Marketing campaign of a Portuguese Banking institution. It contained 4119 customer data with 21 variables (columns), including numeric and categorical columns. The sample of the data set is shown in Fig. 1.

The response variable is categorical and data is very structured. The decision tree model or random forest model can be more appropriate to explore features. I will also perform support vector machines for classification. The Fig. 2 displays the summary and clear picture of the data set.

Since this data set is highly biased and imbalanced, we have a challenge to model this data. In general, most of the telemarketing data set are based on human emotion and it is hard to predict. In our data set, about 81% of the time outcome of previous marketing campaign is unknown (nonexistent) and it has only around 11% *yes* in outcome.

2.2.1 Correlation Analysis. We can compare the correlation plot with other plots to emphasize the important of variable [11]. The correlation plot can tell if predictor is a good predictor or not a good predictor as shown in Fig. 5. This analysis can help us to decide if we can drop some columns (predictors) depending upon its correlation with the output variable.

In order to understand more about relation between different variable with projected output, the correlation plot is very helpful.

```
> summary(mydata)
      age                job           marital                    education
 Min.   : 0.00   admin.     :1012   divorced: 446   university.degree  :1264
 1st Qu.:32.00   blue-collar: 884   married :2509   high.school        : 921
 Median :38.00   technician : 691   single  :1153   basic.9y           : 574
 Mean   :39.39   services   : 393   unknown :  11   professional.course: 535
 3rd Qu.:47.00   management : 324                   basic.4y           : 429
 Max.   :69.00   retired    : 166                   basic.6y           : 228
                 (Other)    : 649                   (Other)            : 168
    default        housing          loan           contact         month
 no     :3315   no     :1839   no     :3349   cellular :2652   may    :1378
 unknown: 803   unknown: 105   unknown: 105   telephone:1467   jul    : 711
 yes    :   1   yes    :2175   yes    : 665                    aug    : 636
                                                              jun    : 530
                                                              nov    : 446
                                                              apr    : 215
                                                              (Other): 203
 day_of_week     duration        campaign         pdays          previous
 fri:768      Min.   :  0.0   Min.   :0.000   Min.   :  0.0   Min.   :0
 mon:855      1st Qu.: 84.0   1st Qu.:1.000   1st Qu.:999.0   1st Qu.:0
 thu:860      Median :157.0   Median :2.000   Median :999.0   Median :0
 tue:841      Mean   :189.2   Mean   :1.932   Mean   :960.2   Mean   :0
 wed:795      3rd Qu.:262.0   3rd Qu.:3.000   3rd Qu.:999.0   3rd Qu.:0
              Max.   :638.0   Max.   :6.000   Max.   :999.0   Max.   :0

       poutcome     emp.var.rate      cons.price.idx  cons.conf.idx
 failure    : 454   Min.   :-3.40000   Min.   :92.20   Min.   :-50.8
 nonexistent:3523   1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.7
 success    : 142   Median : 1.10000   Median :93.75   Median :-41.8
                    Mean   : 0.08497   Mean   :93.58   Mean   :-40.5
                    3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.4
                    Max.   : 1.40000   Max.   :94.77   Max.   :-26.9

   euribor3m      nr.employed     y
 Min.   :0.635   Min.   :4964   no :3668
 1st Qu.:1.334   1st Qu.:5099   yes: 451
 Median :4.857   Median :5191
 Mean   :3.621   Mean   :5166
 3rd Qu.:4.961   3rd Qu.:5228
 Max.   :5.045   Max.   :5228
```

Figure 2: Summary of the dataset. It display the mean and median of attributes. Since the predicted output *y* has lots of *no* in compare to *yes*, this data set is imbalanced and biased

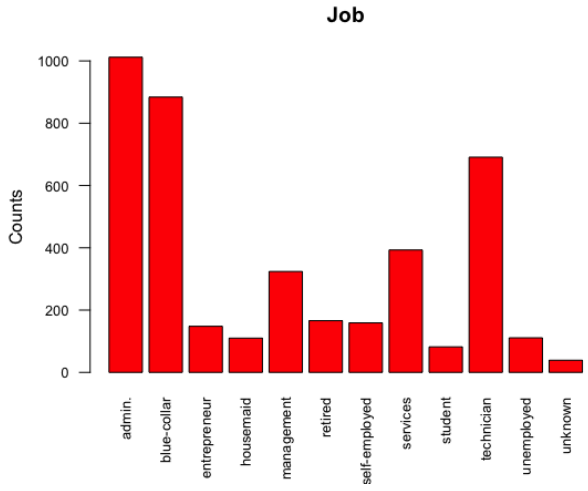

Figure 3: This plot shows the number of clients Vs job category. The highest number of clients are from the job category "admin" followed by blue-color category. Similarly, there are less students involved in the telemarketing campaign.

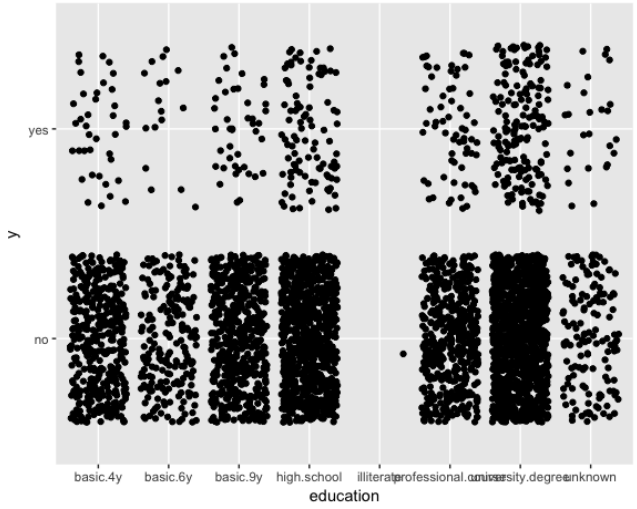

Figure 4: This figure demonstrate the distribution of *yes/no* based on education level of the clients. The people with college degree says more yes than other categories. The overall ration of *yes* is very low for all education level.

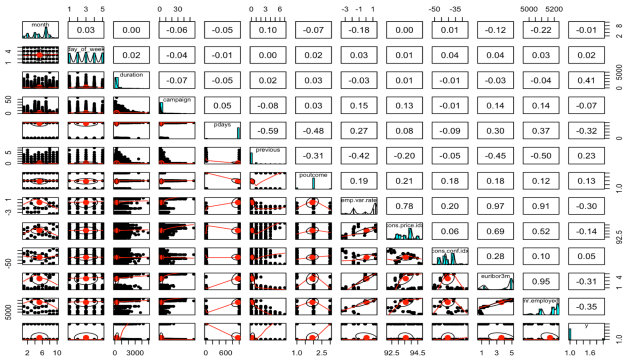

Figure 5: This plot demonstrates the correlation between different variables. There is no strong relation between predictors and predicted output variable *y*, however, there is some relationship between predictor variable *duration* and predicted output *y*.

2.3 Data Preparation

In order to get unbiased and accurate result, the data need to be cleaned. The data pre-processing and cleaning method is the most important steps in data mining. It consumes about 80% of total time.

2.3.1 Missing Value and Outliers Check. The missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. Initially, there was no missing values in the data set. After visualizing the data, it is examined that there are some outliers in the data set. There was some missing values once I removed the outliers as shown in Fig. 6. After that I replaced those NAs by 0. The right plot of Fig. 6 displays the comparison of presence and absence of outliers.

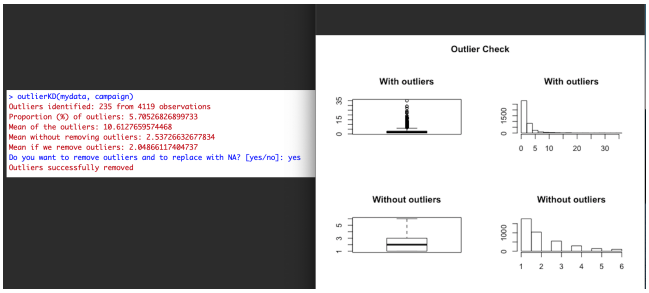

Figure 6: This figure shows the outliers removal process. The left figure is the process of outlier removal and then replaced by NAs (After removing outliers, NAs was replaced by 0). The right side of the figure represents with-outliers and without outliers plots.

```
> set.seed(1234)
> TrainingDataIndex <- createDataPartition(mydata_sub$y, p=0.75, list = FALSE)
> train <- mydata_sub[TrainingDataIndex,]
> test <-mydata_sub[-TrainingDataIndex,]
> prop.table(table(train$y))

       no       yes
0.8902913 0.1097087
> # We can see imbalancing has been taken care of or not
> table(train$y); table(test$y)

  no  yes
2751  339

 no yes
917 112
```

Figure 7: This figure shows the process of splitting the data set into training and testing set with 75% and 25% respectively.

2.4 Data Modeling

Data modeling aims to identify all entities that have data. It then defines a relationship between these entities. Data models can be conceptual, logical or physical data models. The Conceptual models are typically used to explore high level business concepts in case of stakeholders. The Logical models are used to explore domain concepts. While Physical models are used to explore database design [2].

2.4.1 Splitting the Data Set into Training and Testing. :

Since we have imbalanced data, as discussed already in the earlier section, we need to split these data set consistently. I used *CreateDataPartition* method present in caret package to split in such a way that training and testing data will have same ratio of target variable. The data set is split into training set and testing set with 75% for training and 25% for testing as shown in Fig. 7.

In order to model the data, I am performing three data-mining classification techniques, 1) Support Vector Machines (SVM) 2) Decision Tree Model 3) Random Forest Model

2.4.2 Decision Tree Model. The decision tree shows the possible outcomes of the model with conditional control statement [1]. The decision tree typically starts with single nodes and it has several possible branches as shown in figure Fig. 8. After partitioning the data to train and test, I used a 10 fold cross validation repeated 5

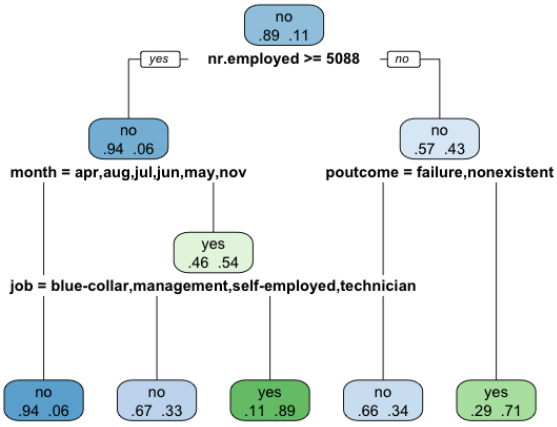

Figure 8: This figure represents the decision tree structure. For example, If no. of employed is greater than 5088, then this client is belongs to *NO* category with 94% of probability. That means the client is more likely to say *NO*.

```
> confusionMatrix(DTPredictions, test$y)
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  894  69
       yes  23  43

               Accuracy : 0.9106
                 95% CI : (0.8915, 0.9273)
    No Information Rate : 0.8912
    P-Value [Acc > NIR] : 0.02318

                  Kappa : 0.4378
 Mcnemar's Test P-Value : 2.711e-06

            Sensitivity : 0.9749
            Specificity : 0.3839
         Pos Pred Value : 0.9283
         Neg Pred Value : 0.6515
             Prevalence : 0.8912
         Detection Rate : 0.8688
   Detection Prevalence : 0.9359
      Balanced Accuracy : 0.6794

       'Positive' Class : no
```

Figure 9: This figure illustrates decision tree model for testing data set using C5.0 algorithm. Based on confusion matrix for test data, using the decision tree model we have correctly classified 894+43 = 937 observations and misclassified 69+23 = 92 representing a 91% accuracy.

times to evaluate the model. I applied C5.0 algorithm [16] to create decision tree model as shown in figure Fig. 9

2.4.3 Support Vector Machines (SVM). The Support Vector Machines (SVM) model is another classification method that can be

```
> confusionMatrix(SVMPredictions, test$y)
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  873  93
       yes  44  19

               Accuracy : 0.8669
                 95% CI : (0.8446, 0.887)
    No Information Rate : 0.8912
    P-Value [Acc > NIR] : 0.9936

                  Kappa : 0.1506
 Mcnemar's Test P-Value : 4.115e-05

            Sensitivity : 0.9520
            Specificity : 0.1696
         Pos Pred Value : 0.9037
         Neg Pred Value : 0.3016
             Prevalence : 0.8912
         Detection Rate : 0.8484
   Detection Prevalence : 0.9388
      Balanced Accuracy : 0.5608

       'Positive' Class : no
```

Figure 10: This figure shows the SVM model of test data set. As evident, the SVM classfication method gives a 86% accuracy predicting only 44 instances of false positives, which is less than decision tree model accuracy.

```
> confusionMatrix(predicted, test$y)
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  917   4
       yes   0 108

               Accuracy : 0.9961
                 95% CI : (0.9901, 0.9989)
    No Information Rate : 0.8912
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9796
 Mcnemar's Test P-Value : 0.1336

            Sensitivity : 1.0000
            Specificity : 0.9643
         Pos Pred Value : 0.9957
         Neg Pred Value : 1.0000
             Prevalence : 0.8912
         Detection Rate : 0.8912
   Detection Prevalence : 0.8950
      Balanced Accuracy : 0.9821

       'Positive' Class : no
```

Figure 11: This figure shows the random forest model of the test data set. This classification algorithm gives a 99% accuracy, which is the best model for our data set to predicting the client subscription.

used to predict if a client falls into either *yes* or *no* class. The SVM model performs classification defined by separating hyperplane.

2.4.4 Random Forest Model. The Random Forest model is a supervised learning algorithm. It builds a group of Decision Trees, most of the time trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result. I use Random Forest method to build the best fit model to predict an output.

3 MODEL EVALUATION AND CONCLUSION

I performed three different classification models to classify whether a customer would open a bank account or not. Based on the model build for this project, Decision Tree and Random Forest model are more accurate to predict the output. The Random Forest model is a recommended model for this classification problem.

The important variables, shown in figure Fig. 13, are agree with the decision tree structure as shown in figure Fig. 8.

This project implemented Data Visualization, Data Correlation, and classification modeling techniques in order to verify that it is possible to predict Term-Deposit successes with Bank Marketing client data. Correlation Analysis is helpful to identify the predictor variables that have the most correlation with Term Deposit.

Since I have been using different data mining techniques, I am expecting the proposed classification models are powerful to predict the output. However, the proposed methods has some limitations. It is not feasible to study all the variables in detail, which might be interesting to predict the output, because of time limitation.

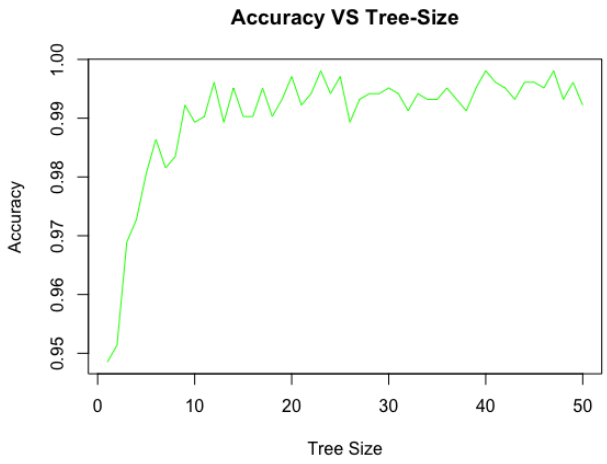

Figure 12: This figure shows Effect of increasing tree count on accuracy in Random Forest Model.

4 FUTURE PLAN

The Logistic Regression, Neural Network, and Naive Bayes method (machine learning method) may be appropriate for this type of data set. I could apply all these methods if I get time to continue this project in the future.

We can also perform a principal component analysis (PCA), a dimensionality reduction technique, to compare the different models with/without PCA. Since the data set is highly biased, we

```
> importance(model)
                  MeanDecreaseGini
job                     71.105837
education               49.708383
default                  8.315837
housing                 16.495925
loan                    12.066024
contact                 10.159236
month                   55.131596
day_of_week             47.743123
duration               210.635629
campaign                35.797608
pdays                   27.086523
previous                 0.000000
poutcome                34.205422
emp.var.rate            31.714938
cons.price.idx          36.704984
nr.employed             73.838370
is_divorced              7.066843
is_single               11.747816
is_married              11.624288
```

Figure 13: This figure displays the most important variables in predicting an output using Random Forest Model algorithm. According to the figure, variables called *duration*, *nr.employed*, *job*, and *month* are more significant to predict whether a customer will open a fix deposit account.

could use Synthetic Minority Over-sampling Technique (SMOTE) model to deal with the imbalance dataset; which could led biased prediction and misleading accuracy.

5 ACKNOWLEDGMENT

I would like to thank some special peoples who helped me a lot on this project.

First my family, In particular my wife Bandana Dahal, from whom I get lots of love and support towards my study and commitment, and my sister-in-law Sadhana Adhikari, who helped me to create an environment and support to do this project.

My terrific professor, Dr. Xinlian Liu, encouraged me to start this project. His ideas and suggestions are always valuable for me not only for this project but also for my further career in data science. Finally, I am very thankful to my entire CS 522- Data Mining class for their feedback and encouragement.

REFERENCES

[1] E. Konukoglu A. Criminisi, J. Shotton. 2017. *Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning.* Technical Report 5. 1–151 pages. https://doi.org/10.1136/tc.2009.033175

[2] K Chitra and B Subashini. 2013. Data Mining Techniques and its Applications in Banking Sector. *International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com ISO Certified Journal* 9001, 8 (2013), 2–8. https://doi.org/ISSN2250-2459

[3] Alvin Choong, Frank Devlin, Mudit Gupta, Tan Wei-chyin, and Kate Chen. [n. d.]. Predictive Analytics in Marketing A Practical Example from Retail Banking. 1, Oct 2017 ([n. d.]), 1–17.

[4] Elsayad M. Alaa Elsalamony A. Hany. 2018. Bank Direct Marketing Based on Neural Network. *Advanced Energy Materials* 8, 25 (2018), 1–9. https://doi.org/10.1002/aenm.201800466

[5] Mansi Gera and Shivani Goel. 2015. Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity. *International Journal of Computer Applications* 113, 18 (2015), 2–8. https://doi.org/10.5120/19926-2042

[6] Paolo Giudici. 2005. Applied Data Mining. *Statistical Methods for Business and Industry* August (2005), 1–67.

[7] Jiawei Han, Micheline Kamber, and Jian Pei. 2011. *Data Mining: Concepts and Techniques.* 1–740 pages. https://doi.org/10.1016/B978-0-12-381479-1.00001-0 arXiv:arXiv:1011.1669v3

[8] Hossein Hassani, Xu Huang, and Emmanuel Silva. 2018. Digitalisation and Big Data Mining in Banking. *Big Data and Cognitive Computing* 2, 3 (2018), 1–13. https://doi.org/10.3390/bdcc2030018

[9] Rajni Jain. 2009. Introduction to Data Mining Techniques. *India* (2009), 1–11. http://www.iasri.res.in/ebook/expertsystem{_}o/c-10DataMiningtechniques.pdf

[10] Vikas Jayasree and Rethnamoney Vijayalakshmi Siva Balan. 2013. A review on data mining in banking sector. *American Journal of Applied Sciences* 10, 10 (2013), 1–6. https://doi.org/10.3844/ajassp.2013.1160.1165

[11] Vincent Lemaire, Orange Labs, Pierre Marzin, Lannion Cedex, Carine Hue, G F I Informatique, Louis De Broglie, and Olivier Bernier. 2010. Correlation Analysis in Classifiers. 19.

[12] Breiman Leo, Friedman H Jorome, Olshen A. Richard, and Stone J. Charles. 1984. *Classification and Regression Trees.* Vol. 136. 1–31 pages.

[13] Sérgio Moro, Paulo Cortez, and Paulo Rita. 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems* 62 (2014), 5–10. https://doi.org/10.1016/j.dss.2014.03.001

[14] Muneeb A. [n. d.]. Predicting the Success of Bank Telemarketing Using Various Classification Algorithms. ([n. d.]), 30.

[15] Andrew Ng. 2012. CS229 Lecture notes: Margins and Intuition. x (2012), 1–25.

[16] Rutvija Pandya and Jayati Pandya. 2015. C5. 0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. *International Journal of Computer Applications* 117, 16 (2015), 4. https://doi.org/10.5120/20639-3318

[17] Thomas Khabaza Pete Chapman, Julian Clinton, Randy Kerber. 2000. CRISP-DM 1.0 Step-by-step. *ASHA presentation* (2000), 76. https://doi.org/10.1109/ICETET.2008.239 arXiv:arXiv:1011.1669v3

[18] Brian Ramsay and Esther Van Der Knaap. [n. d.]. Confusion Matrix-based Feature Selection Sofia Visa. ([n. d.]), 8.

[19] R. Vaidehi. 2016. Predictive Modeling to Improve Success Rate of Bank Direct Marketing Campaign. *Ijmbs* 9519 (2016), 1–3.

[20] Fabricio Voznika and Leonardo Viana. [n. d.]. DATA MINING CLASSIFICATION. ([n. d.]), 1–6.