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

Chiranjibi Ghimire
Hood College, Frederick , Maryland

Abstract

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 from 2008 through 2013 to identify the existing customers that have higher chance to subscribe for term deposit and focus marketing effort of such customers .

The aim of this project is to build the model by performing various data-mining classification techniques to predict whether the customer will subscribe bank long-term deposit. This project will analyze the data set related to the bank client based on telephone communication. 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.

Data Description

The UCI data set of direct marketing campaign of a Portuguese banking institution is used for this project. The dataset examined in 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 columns as shown in Fig.1. Each record includes 20 explanatory observations about the client contacted, and 1 response observation of whether the customer subscribed to a Term Deposit.

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.

AGE	JOB	MARITAL	DEF	HOUSING	LOAN	CONTACT	EMP_VAR_RATE	CPI	CCI	EURIBOR	NUM_EMP	Y
56	housemaid	married	no	no	no	telephone	1.1	93.994	-36.4	4.857	5191	no
57	services	married		no	no	telephone	1.1	93.994	-36.4	4.857	5191	no
37	services	married	no	yes	no	telephone	1.1	93.994	-36.4	4.857	5191	no
40	admin.	married	no	no	no	telephone	1.1	93.994	-36.4	4.857	5191	no
56	services	married	no	no	yes	telephone	1.1	93.994	-36.4	4.857	5191	no
45	services	married		no	no	telephone	1.1	93.994	-36.4	4.857	5191	no
59	admin.	married	no	no	no	telephone	1.1	93.994	-36.4	4.857	5191	no
41	blue-collar	married		no	no	telephone	1.1	93.994	-36.4	4.857	5191	no
24	technician	single	no	yes	no	telephone	1.1	93.994	-36.4	4.857	5191	no

Figure 1. This figure illustrates the metadata of the data set. The last column, with red color, is the predicted output and other columns are predictors.

Data Mining Steps and Data Modeling

Since the data set contain both numerical and categorical columns, I used Data Classification methods to build best fit model. In order to perform the Data mining, I use Cross-Industry Standard Process for Data Mining (CRISP-DM) process, which includes, *Business Understanding*, *Data Understanding and Exploring*, including Data pre-processing, Cleaning, and Visualization, *Data Preparation*, *Data Modeling*, and *Model Evaluation*.

In order to model the data set, I split the data into 75% for training and 25% for testing. I applied three data mining classification technique, **Decision Tree model**, **Random Forest Model**, and **Support Vector Machines (SVM**,) to build the best fit model to predict the output.

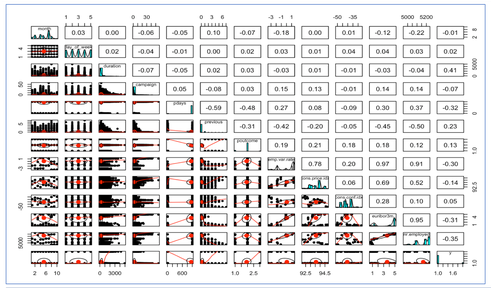

Figure 2. This figure represents the correlation between different variables. There is no strong relation between predictors and resulted output variable (*y*).

Results

1) Decision Tree Model

```
> 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
```

91 % accuracy

Fig 3. This figure represents the confusion matrix for Decision Tree model using C5.0 algorithm with 91% of accuracy.

2) Support Vector Machines

```
> 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
```

86 % accuracy

Fig 4. This figure represents the confusion matrix for SVM with 86% of accuracy.

3) Random Forest Model

```
> 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
```

99 % accuracy

Fig 5. This figure represents the confusion matrix for Random Forest Model with 99% of accuracy. This model is the best fit model for this data set to predict the client subscription.

For more detail about this project and source code, please visit: https://github.com/cghimire/CS-522_Bank-Marketing-Project

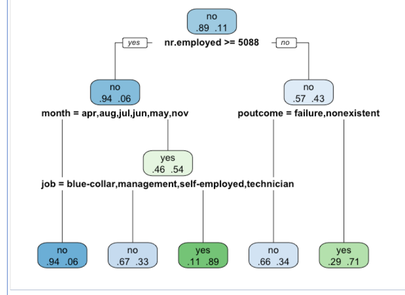

Figure 6. This figure represents the Decision Tree Model. It shows that the nr.employed, month, job, and poutcome variables are more significant.

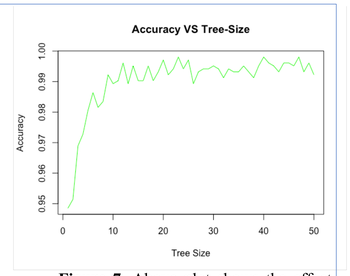

Figure 7. Above plot shows the effect of increasing tree count on accuracy with Random Forest Model

```
> 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 8. This figure displays the importance variables using Random Forest Model.

Future Works

The Logistic Regression, Neural Network, and Naive Bayes method (machine learning method) may be appropriate for this kind of data set.

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

Acknowledgement

I would like to thank some special peoples, in order of appearance, 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 instructor, Dr. Xinlian Liu, encouraged me a lot throughout this project. His ideas and suggestions are always valuable to me not only for this project but also for my further data science career. Finally, I am very thankful to my entire Data Mining (CS 522) class for their feedback and encouragement.

Contact

Chiranjibi Ghimire
Hood College, Department of Information Technology
Frederick , MD
cg16@hood.edu
443-657-0838

References

[1] 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
[2] 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.
[3] 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
[4] 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
[5] Paolo Giudici. 2005. Applied Data Mining. Statistical Methods for Business and Industry August (2005), 1–67.