Analyzing Bank Marketing Dataset using Different Classification Algorithms

Dipesh Chand
Faculty of Computer Science
Østfold University College
1783 Halden,Norway
dipesh.chand@hiof.no

Bibek Acharya
Faculty of Computer Science
Østfold University College
1783 Halden,Norway
bibek.acharya@hiof.no

***Abstract*—This project applied different classification techniques to build the model to predict whether the customer will subscribe bank long-term deposit or not. A Portuguese retail bank collected data from 2008 through 2013. We 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 allows for understanding of the possible effects of the concentration on a wide variety of banking resources into a small group of national enterprises. This kind of study projects could be helpful to determine the likelihood of procurement of financial services.**

Keywords—Bank Marketing, Classification, Decision Tree, C5.0, Random Forest, XGBoost

I. INTRODUCTION

Businesses today often launch marketing campaigns to boost the sale of their products and services. While digital marketing becomes popular and have many advantages over traditional marketing, traditional marketing methods are still providing physical customer experience difficult to be offered by digital marketing and could not be completely replaced by digital marketing 1 . One of the drawbacks of traditional marketing is that it is typically more expensive than digital marketing since it involves one or multiple types of activities such as phone calls, customer visits, or physical prints etc. These activities often require significant efforts and investment from businesses. Therefore, for traditional marketing, it is important to target marketing activities towards desirable customers who are more likely to buy products and services than others. It will not be cost-effective if marketing campaign targets are simply randomly chosen without going through thorough review and selection.

Machine learning can provide a data-driven approach to help marketing campaign more targeted to desirable customers. In this project, using bank telemarketing as an example of traditional marketing, a machine learning model was developed and demonstrated its effectiveness in maximizing business return while minimizing marketing effort.

This paper is structured as follows:

Chapter 2 presents the problem statement and dataset description. Chapter 3 provides description about different related work. and Chapter 4 explains about the different models used to perform the classification on this project. Chapter 5 explores about the result of the classification model and compares them. Chapter 6 pointing forward to conclusion and several possibilities in this project aim which can be performed in near future.

II. PROBLEM STATEMENT AND DATASET DESCRIPTION

The most effective strategy to progress a business marketing at a least conceivable overhead is constantly seen as the fundamental issue by the supervisor. The foremost tremendous part of the challenge is to recognize the promising and potential clients of the displaying thing with limited data. Realizing that specific information which chooses the promising and potential client would empower executive to put more assets on positive portion towards the items and cut down the budget spent on non-promising client, so that to dispose of bottlenecks and make a progressively productive advancing way.

The dataset is related to direct marketing campaigns run by the Portuguese bank and

contains information on various features of interest for approximately 41,188 customers. The dataset has been taken from the UCI machine learning repository [1].

The features of interest can be broken down as follows:

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')

The initial impression that can be created using the dataset are as:

- Total 41188 records
- 10 numeric attributes : age, duration, campaign, pdays, previous, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed
- 10 Factors:
  - 10 multi-valued categorical attributes : job, marital, education, default, housing, loan, contact, month, day_of_week, poutcome
  - 1 target attribute y
- No missing values: Preprocessing should be easier.

III. RELATED WORK

Moro, Cortez, and Rita [2] has utilized semi-automatic modeling procedure. In their study they have picked information from July 2012 and utilized 22 highlights of data. They have compared about 4 distinctive sort of data mining model. They are "logistic regression, decision trees, neural network (NN) and support vector machine". There are two measurements utilized in the investigation known as AUC (area of the receiver operating characteristic curve) and ALIFT (area of LIFT cumulative curve). The testing model was utilized in advancement stage utilizing the most recent information of July 2012 and another model called a rolling window scheme. The research has shows the result of AUC as 0.8 and ALIFT as 0.7. This has permitted 79% of subscribers as half possible customers. There two extraction strategies are utilized which are sensitive analysis and DT. They were applied to NN. This uncovered several key traits which was tenable and important to telemarketing managers for the campaign.

Vajiramedhin and Suebsing [3] suggested that the performance of the predictive model with the number of smaller features can be improved. Their experiment on Direct Bank Marketing dataset can enhance the predictive model performance both of the TP rate and the ROC rate while it employs the smaller storage space, reduces the computation time and gains the higher predictive performance.

Another study Elsalamony and Elsayad [4], evaluate and compare the classification

performance of the two different techniques models Multilayer perceptron neural network (MLPNN) and C5.0 on the bank direct marketing dataset to classify for bank deposit subscription. In their study they used statistical measures; Classification accuracy, sensitivity and specificity and found that C5.0 has slightly better performance than MLPNN. And also Importance analysis has shown that attribute "Duration" in both models has achieved the most important attribute.

IV. METHODOLOGY

Since the dataset 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 this classification technique is to predict if the client will subscribe to a Term Deposit and to improve the performance of model and increase the classification accuracy using various approaches. In order to obtain more accurate and precise model to predict desired output, we will performed several classification techniques and model such as Decision Tree using rpart and C5.0, Random Forest Model and eXtreme Gradient Boosting (XGBoost). We will perform correlation analysis to see if there is any relation- ship 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 we perform all of the above techniques, we would be able to understand the data and suggest the best fit model for prediction of "customer term deposit" more accurately and precisely. We perform the following steps in our study:

1. Data Understanding and Exploring: Cleaning, and Visualization
2. Data Pre-processing
3. Data Modeling

A. DATA UNDERSTANDING AND EXPLORING

We used R language with RStudio IDE in this project for analysis. The first step is to load the dataset into a data frame for easy manipulation and exploration. The initial findings of this dataset are like:

- Most of the clients have never been contacted since contact is unknown for 28.79%.
- 86.34% of the time outcome of previous marketing campaign is unknown.
- Duration seems to have lot more variation, it may be a good predictor.
- Data is very imbalanced, only 11.26% yes in outcome.

These insights were the very basic details found by just looking at the summary of the datasets, The Fig. 2 displays the summary and clear picture of the dataset.

```
> summary(dataset)
      age                job             marital
 Min.   :17.00   admin.     :10422   divorced: 4612
 1st Qu.:32.00   blue-collar: 9254   married :24928
 Median :38.00   technician : 6743   single  :11568
 Mean   :40.02   services   : 3969   unknown :   80
 3rd Qu.:47.00   management : 2924
 Max.   :98.00   retired    : 1720
                 (Other)    : 6156
           education        default         housing          loan
 university.degree  :12168   no     :32588   no     :18622   no     :33950
 high.school        : 9515   unknown: 8597   unknown:  990   unknown:  990
 basic.9y           : 6045   yes    :    3   yes    :21576   yes    : 6248
 professional.course: 5243
 basic.4y           : 4176
 basic.6y           : 2292
 (Other)            : 1749
    contact          month       day_of_week    duration
 cellular :26144   may    :13769   fri:7827    Min.   :   0.0
 telephone:15044   jul    : 7174   mon:8514    1st Qu.: 102.0
                   aug    : 6178   thu:8623    Median : 180.0
                   jun    : 5318   tue:8090    Mean   : 258.3
                   nov    : 4101   wed:8134    3rd Qu.: 319.0
                   apr    : 2632               Max.   :4918.0
                   (Other): 2016
   campaign          pdays          previous            poutcome
 Min.   : 1.000   Min.   :  0.0   Min.   :0.000   failure    : 4252
 1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.000   nonexistent:35563
 Median : 2.000   Median :999.0   Median :0.000   success    : 1373
 Mean   : 2.568   Mean   :962.5   Mean   :0.173
 3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.000
 Max.   :56.000   Max.   :999.0   Max.   :7.000

  emp.var.rate      cons.price.idx  cons.conf.idx     euribor3m
 Min.   :-3.40000   Min.   :92.20   Min.   :-50.8   Min.   :0.634
 1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.7   1st Qu.:1.344
 Median : 1.10000   Median :93.75   Median :-41.8   Median :4.857
 Mean   : 0.08189   Mean   :93.58   Mean   :-40.5   Mean   :3.621
 3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.4   3rd Qu.:4.961
 Max.   : 1.40000   Max.   :94.77   Max.   :-26.9   Max.   :5.045

 nr.employed     y
 Min.   :4964   no :36548
 1st Qu.:5099   yes: 4640
 Median :5191
 Mean   :5167
 3rd Qu.:5228
 Max.   :5228
```

Fig.1. *Summary of the Dataset*

1) Data Cleaning

Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. Although it is clearly mentioned that there are no missing values in the dataset but we didn't want to take risk so we performed missing value check and found that the dataset is free from missing value. So we don't have to perform any task to manage the missing value.

2) Data Visualization

The number of plots were created to visualize the data so that it would be easy to understand the information provided by the dataset. and more over we also plot combination of variable to the outcome to demonstrate the relation between them. By examining the histogram of numerical features, although most of features are not

distributed normally, there are few features with extremely skewed distribution except 'pdays'. However, UCI Repository intentionally used '999' for 'pdays' to arbitrarily represent clients not previously contacted [2] and is not really outliers. Thus, no data transformation was performed.
This dataset is heavily biased for class label ('y') 88.7% of samples (36548) are labeled as "no"and only 11.3% (4640) are labeled as "yes".

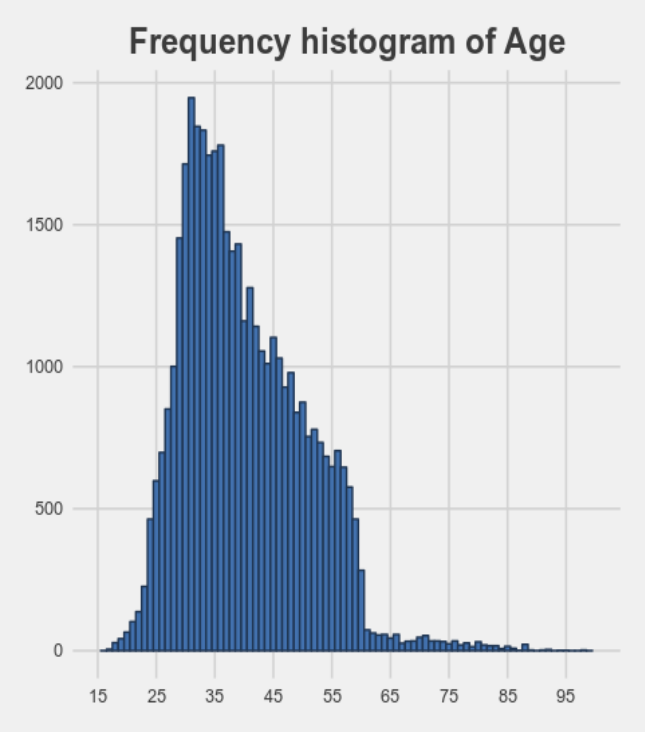

Fig.2. Histogram showing frequency of Age

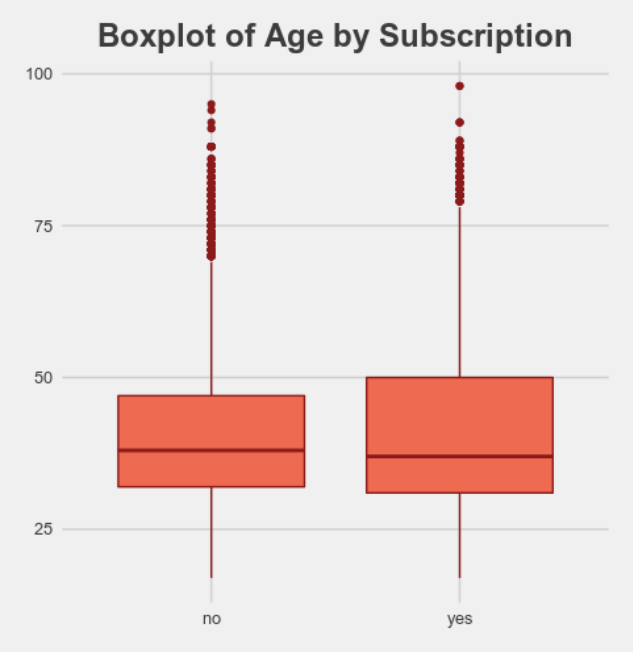

Fig.3. Box plot of Age by subscription

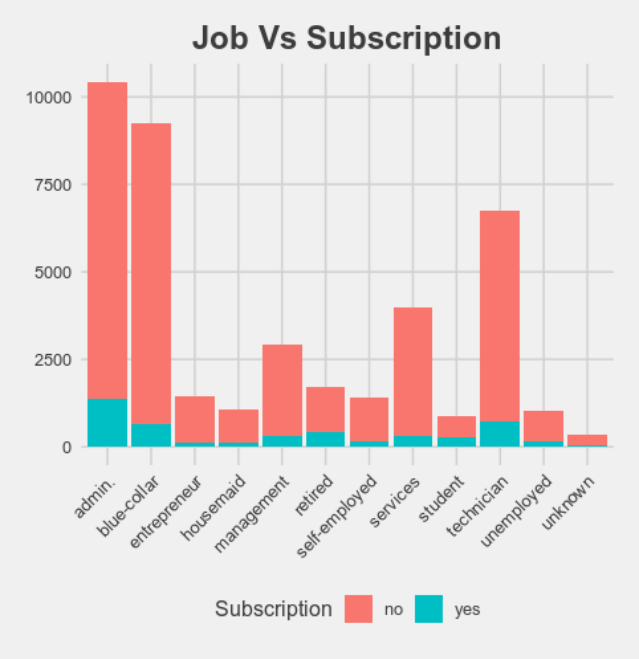

Fig.4. Plot demonstrating the distribution of Yes/ No based on Job attribute

B. DATA PRE-PROCESSING

1) Feature Binning

The Age variable is widely scattered, shown in Fig. 2 and doing classification on this type of data is very challenging, so binning of data is the best option. We divided the data into four category as
0-19→ "Teens",
20-35→ "Young Adults",
36-60→ "Adults",
61-100→ "Senior Citizens"

By binning, it may improve accuracy of the predictive models by reducing the noise or non-linearity.

2) Feature Selection

Checking the predictor variables that are highly correlated with each other. Two metrics are used - Correlation factor and VIF.
At first we created the correlation Matrix of the dataset and then find attributes that are highly correlated, here we set cutoff as 0.75. We got euribor3m and emp.var.rate attributes highly correlated. Then we applied variance inflation factor (VIF). VIF quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.
Again we found euribor3m variable have greater value. So analyzing Correlation factor and VIF it

is decided to remove the 'euribor3m' variable before performing any further activities.

The duration variable is also removed even though it was a strong predictor based on the following dataset author's notes.

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.* [4]

3) Data Splitting

The dataset is divided into training data and test data with the intention of using the training data to find the parameters of the particular model being used (fitting the model on the training data) and then applying this to the test data to determine the model's performance and to draw conclusions about its predictive capability. This can be done with a sample.split function call by specifying split ratio. In this project the dataset is split into training set and testing set with 80% for training and 20% for testing.

4) Managing Imbalance Data

A dataset is imbalanced if the distribution of output classes is not uniform. Here we have imbalanced data, as discussed already in the earlier section, to be precise class 1 ('not subscribed') has 89% of instances but class 2 ('subscribed') has only 11% of instances. By using this type of imbalance dataset problem may occurs such as:

1. The algorithms will be biased towards majority class. So it will predict the majority class irrespective of the predictors.
2. The test Accuracy of the model will be more than 90% but it won't serve the our objective because the test data will also have more than 90% of the data pertaining to majority class. So, here, the accuracy won't be a good representation of the model performance.
3. The algorithm assumes that errors obtained from both classes have same cost. But in this case, the Type II error (False Negative) error is very serious than the Type I error (False Positive). Because we should not miss the chance or the opportunity of identifying a potential customer. But we can afford to get a few false positive, because comparatively it won't do much harm.

So to avoid these type of issues we need to deal with imbalance data. For this project, Synthetic Data Generation method is used. In regards to synthetic data generation, Synthetic Minority Oversampling Technique" (SMOTE) - a powerful and widely used method is applied. SMOTE algorithm creates artificial data based on feature space (rather than data space) similarities from minority samples. It generates a random set of minority class observations to shift the classifier learning bias towards minority class. To generate artificial data, it uses bootstrapping and k-nearest neighbors. By using this technique we make sets balance up to 55% No and 45% Yes.

C. DATA MODELING

In order to model the data, we used four classification algorithms in this project to predict the term deposit subscription.

1. Decision Tree Model with rpart
2. Decision Tree Model with C5.0
3. Random Forest Model
4. eXtreme Gradient Boosting (XGBoost).

1) Decision Tree Model with rpart

The decision tree shows the possible outcomes of the model with conditional control statement [5]. The decision tree typically starts with single nodes and it has several possible branches as shown in figure Fig. 6. Decision trees are a simple yet effective method for classification. Using a tree structure, this algorithm splits the data set based on one feature at every node until all the data in the leaf belongs to the same class. The criterion used for splitting is called information gain, which is based on a purity measure called entropy, a measure of disorder. The set with the highest impurity will have higher entropy whereas the set which has higher purity will have lower entropy. Information gain measures the change in entropy due to the amount of information added. The higher the information gain, the more information that feature provides about the target variable.

By default, the decision tree grows deep and complex until every leaf is pure and hence it is prone to over fitting.

We use two different ways to create decision tree first one is rpart for creating decision tree and

confusion Matrix for analyzing. Figure 5 shows the prediction of rpart :

```
> table(pred_TA)
pred_TA
   no   yes
7397   841
>
```

Fig.5. Prediction table of rpart

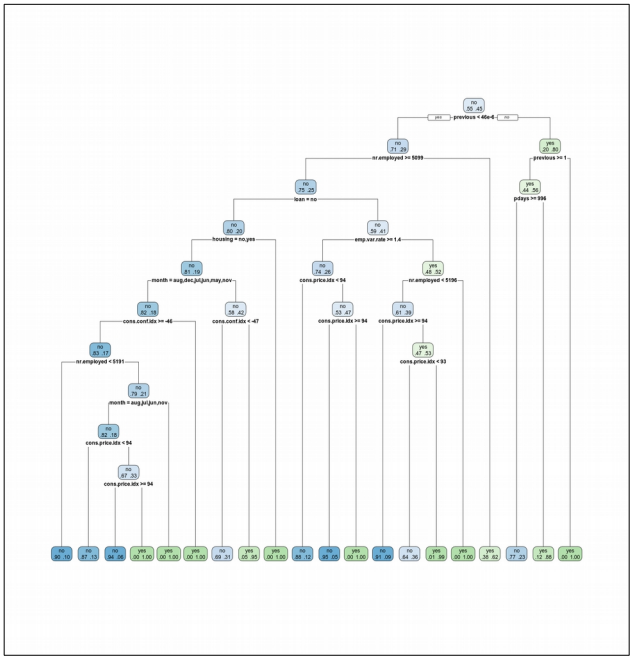

Fig.6. Figure representing the decision tree

2) Decision Tree Model with C5.0

While there are various implementations of decision trees, one of the foremost well-known is the C5.0 algorithm. The C5.0 algorithm has become the industry standard for creating decision trees, since it does well for most sorts of problems directly out of the box. Compared to more advanced and sophisticated machine learning models (e.g. Neural Networks and Support Vector Machines), the decision trees under the C5.0 algorithm generally perform nearly as well but are much easier to understand and deploy.

A C5.0 model works by splitting the sample based on the field that provides the maximum information gain. Each sub sample defined by the first split is then split again, usually based on a different field, and the process repeats until the sub samples cannot be split any further. Finally, the lowest-level splits are reexamined, and those that do not contribute significantly to the value of the model are removed or pruned. [6]. We also implement the boosting up to 100 trials to improve the performance of the model. Figure 7 and Figure 8 shows the prediction of C5.0 algorithm before and after boosting:

```
> table(pred_C50)
pred_C50
   no   yes
7622   616
>
```

Fig.7. Prediction table of C5.0 (initial)

```
> table(predict_C50_imp)
predict_C50_imp
   no   yes
7480   758
>
```

Fig.8. Prediction table of C5.0 after boosting

3) Random Forest Model

Random forest classifiers are one of the ensemble learning methods for classification. 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. Random forest method builds multiple decision trees and merges them together to get a more accurate and stable prediction. We use Random Forest method to build the best fit model to predict an output. Figure 9 shows the prediction of Random Forest Model:

```
> table(pred_rf)
pred_rf
   no   yes
7544   694
>
```

Fig.9. Prediction table of Random Forest Model

4) eXtreme Gradient Boosting (XGBoost).

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. The objective of any supervised learning algorithm is to define a loss function and minimize it.

We want our predictions, such that our loss function (MSE) is minimum. By using gradient descent and updating our predictions based on a learning rate, we can find the values where MSE is minimum.

So, we are basically updating the predictions such that the sum of our residuals is close to 0 (or minimum) and predicted values are sufficiently close to actual values.

Since XGBoost only works with numeric vectors so we have to handle our data differently than our others classifiers. We use One Hot Encoding to convert categorical variable into numeric vector. Figure 10 shows the prediction of XGBoost Model:

```
> table(pred.resp)
pred.resp
  no  yes
7760  478
>
```

Fig.10. Prediction table of XGBoost Model

V. Model Evaluation and Result

We performed four different classification models to classify whether a customer would open a bank account or not. We consider Confusion Matrix metrics for evaluation which is a breakdown of predictions into a table showing correct predictions (the diagonal) and the types of incorrect predictions made (what classes incorrect predictions were assigned). Figure 11, Figure 12, Figure 13, and Figure 14 shows the details of confusion matrix and statistics of different classification models we used.

```
> confusion_TA
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  6866  531
       yes  444  397

               Accuracy : 0.8816
                 95% CI : (0.8745, 0.8885)
    No Information Rate : 0.8874
    P-Value [Acc > NIR] : 0.950290

                  Kappa : 0.3827
 Mcnemar's Test P-Value : 0.005884

            Sensitivity : 0.9393
            Specificity : 0.4278
         Pos Pred Value : 0.9282
         Neg Pred Value : 0.4721
             Prevalence : 0.8874
         Detection Rate : 0.8335
   Detection Prevalence : 0.8979
      Balanced Accuracy : 0.6835

       'Positive' Class : no
```

Fig.11. Confusion Matrix and Statistics of rpart model

```
> confusion_improved
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  6943  537
       yes  367  391

               Accuracy : 0.8903
                 95% CI : (0.8833, 0.8969)
    No Information Rate : 0.8874
    P-Value [Acc > NIR] : 0.2068

                  Kappa : 0.4034
 Mcnemar's Test P-Value : 1.9e-08

            Sensitivity : 0.9498
            Specificity : 0.4213
         Pos Pred Value : 0.9282
         Neg Pred Value : 0.5158
             Prevalence : 0.8874
         Detection Rate : 0.8428
   Detection Prevalence : 0.9080
      Balanced Accuracy : 0.6856

       'Positive' Class : no
```

Fig.12. Confusion Matrix and Statistics of C5.0 model

```
> confusion_rf
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  6983  561
       yes  327  367

               Accuracy : 0.8922
                 95% CI : (0.8853, 0.8988)
    No Information Rate : 0.8874
    P-Value [Acc > NIR] : 0.08369

                  Kappa : 0.3941
 Mcnemar's Test P-Value : 5.326e-15

            Sensitivity : 0.9553
            Specificity : 0.3955
         Pos Pred Value : 0.9256
         Neg Pred Value : 0.5288
             Prevalence : 0.8874
         Detection Rate : 0.8477
   Detection Prevalence : 0.9158
      Balanced Accuracy : 0.6754

       'Positive' Class : no
```

Fig.13. Confusion Matrix and Statistics of Random model

```
> confusionMatrix(pred.resp, new_test$y)
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  7109  651
       yes  201  277

               Accuracy : 0.8966
                 95% CI : (0.8898, 0.9031)
    No Information Rate : 0.8874
    P-Value [Acc > NIR] : 0.003922

                  Kappa : 0.3438
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9725
            Specificity : 0.2985
         Pos Pred Value : 0.9161
         Neg Pred Value : 0.5795
             Prevalence : 0.8874
         Detection Rate : 0.8630
   Detection Prevalence : 0.9420
      Balanced Accuracy : 0.6355

       'Positive' Class : no
```

Fig.14. Confusion Matrix and Statistics of XGBoost model

The undermentioned table demonstrates the prediction accuracy of the five machine learning models we have used in our project using R language.

Classification Model	Prediction Accuracy
Decision Tree Model with rpart	0.8816
Decision Tree Model with C5.0	0.8903
Random Forest Model	0.8922
eXtreme Gradient Boosting (XGBoost).	0.8966

Table.1. Prediction Accuracy of different classification models

From the table, it can be seen that Random Forest and XGBoost are the top performing algorithms. Even though the differences in all the metric values are infinitesimal, we will choose XGBoost in terms of predicted accuracy and Random Forest as the best model for our prediction because it ranks first in 2 out of the 3 considered metrics .

VI. CONCLUSION

The results explained in the previous chapter concludes the study. We managed to get results from the given data set and compare algorithms, which was our primary goal for this project. While programming, interpreting the results and studying the results, whole research group got a hands-on experience with modern tools for computational data analytic with a real business case.

The next step for this project would be the optimization of each algorithm and implementing neural network. In this project, default settings of each algorithm was used. Adjusting settings to get better results would require a more in-depth study of each algorithm to know how the performance could be improved.

REFERENCES

[1] "UCI Machine Learning Repository: Bank Marketing Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing. [Accessed: 14-Mar-2019].

[2] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decis. Support Syst.*, vol. 62, pp. 22–31, Jun. 2014.

[3] C. Vajiramedhin and A. Suebsing, "Feature selection with data balancing for prediction of bank telemarketing," *Appl. Math. Sci.*, vol. 8, pp. 5667–5672, 2014.

[4] H. A. Elsalamony and A. M. Elsayad, "Bank Direct Marketing Based on Neural Network and C5.0 Models," vol. 2, no. 6, p. 10.

[5] A. Criminisi, E. Konukoglu, and J. Shotton, "Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning," Oct. 2011.

[6] "C5.0 Node," 24-Oct-2014. [Online]. Available: undefined. [Accessed: 19-Mar-2019].