PROJECT REPORT

ISDS 574 – Data Mining for Business Applications

Wone Eui Jung, Tianqing Huang, Yiting Li,
Yuchen Bi, Sebastian Marx, Sohrab Gill

December 13, 2019

Table of Contents

1. Executive Summary / Introduction

In this project, we propose a data mining approach to predict and analyze the success of a Portuguese bank's telemarketing calls for selling long-term deposits (Moro, Cortez, & Rita, 2014). The source of the dataset is UCI machine learning. We used dplyr for data cleaning. The training dataset was imbalanced as there were much less y=1, so it was balanced by selecting 3,000 samples of each yes and no results. We cleaned up the data and analyzed the reduced features related to the product, bank client and social economic attributes. These features were classified according to different models like Logistic Regression, Decision Tree and kNN. Random forest and XGBoost was used to predict whether the client will subscribe to a term deposit or not. It was revealed that the key attributes are Euro bank interest rate (short term bank loan rate), duration of the call and bank agent experience. These attributes helped decide the success of a telemarketing call to sell the long-term bank deposit.

Because the data was data was studied between 2008 to 2013, the financial crisis had an impact on the decisions of the Portuguese bank. The bank was forced to have more capital requirement which means selling more long-term deposits to clients. We intend to find the best attributes to focus on when making these telemarketing calls to sell bank deposits.

2. Data Preprocessing

In the following section, we discuss data exploration, which is the main part in this section involving studying the data variables and figuring out what variables are more important than the other ones. This section also includes data cleaning and partition which is done to make our data analysis easier to understand.

2.1. Data Exploration

In the following section we explore the several variables and their effect on the outcome which is customer subscribing to the bank product (term deposit) or not.

In general, customers who are more tech savvy, educated and young tend to buy the term deposit more. Also, if customers are contacted more frequently and recently tend to buy the bank product more than the others.

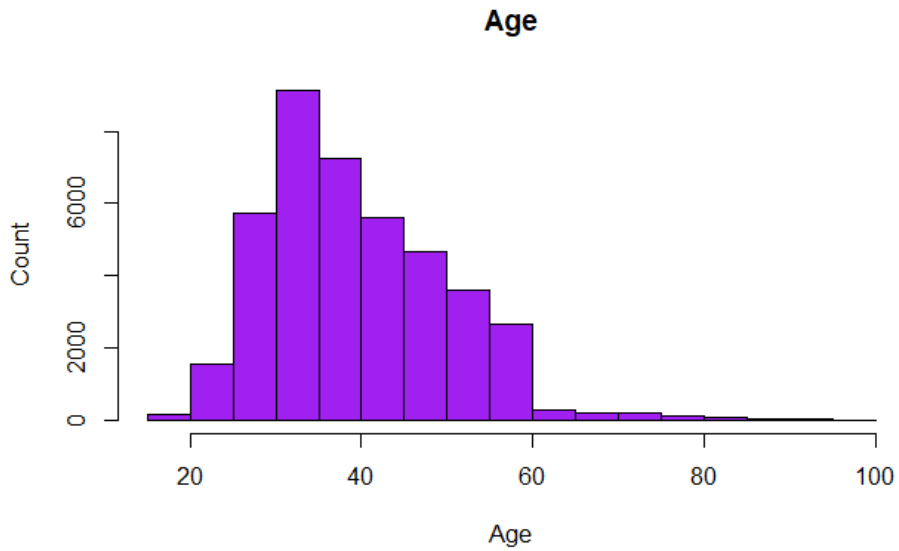

Figure 1 - Variable: Age

50percentageof Age data was concentrated from 32 to 47 years old (Min:17,1Q:32, Median:38, 3Q: 47, Max:98). There was no unknown data for Age variables.

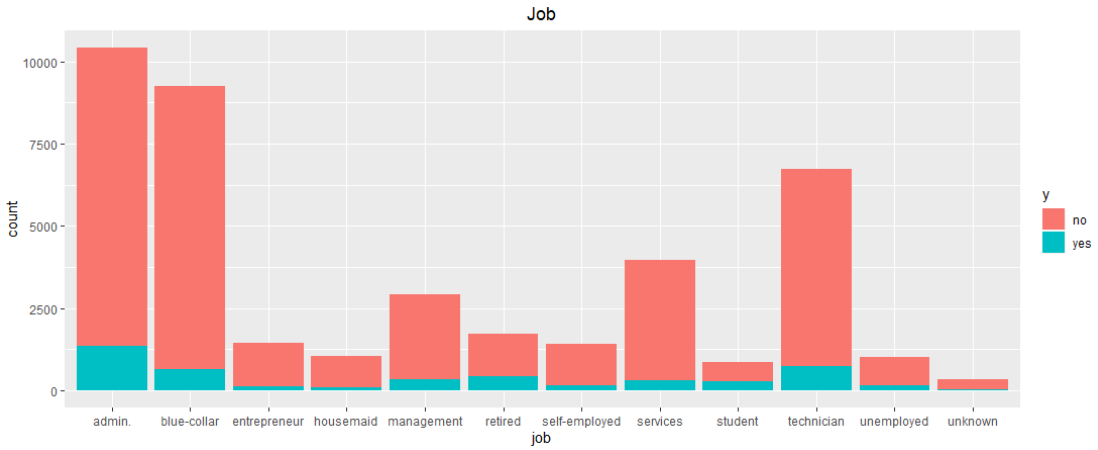

Figure 2 - Variable: Job

Variable	admin.	blue-collar	technician	services	management	retired	entrepreneur	self-employed	housemaid	unemployed	student	unknown	sum
Freq	10422	9254	6743	3969	2924	1720	1456	1421	1060	1014	875	330	41188
%	25.3%	22.5%	16.4%	9.6%	7.1%	4.2%	3.5%	3.5%	2.6%	2.5%	2.1%	0.8%	100.0%

Table 1 Variable: Job

The top five jobs (admin, blue-collar, technician, management) takes part 80percentageof data. Admin jobs responds much more. People in admin jobs tends to have more term deposit(y=yes) compare to blue-collar jobs. Admin group can be good target for marketing promotion.

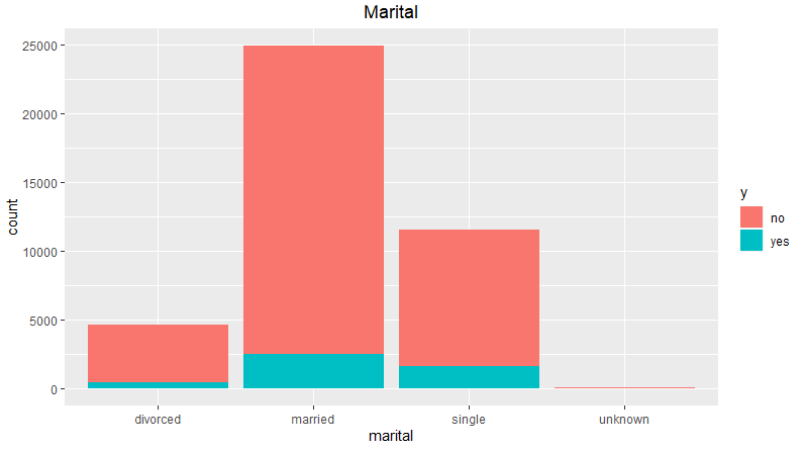

Figure 3 - Variable: Marital

	Count	no	yes	percentageof yes
divorced	4612	4136	476	10.32%
married	24928	22396	2532	10.16%
single	11568	9948	1620	14.00%
unknown	80	68	12	15.00%

Table 2 Variable: Marital

Single and unknown group tends to have higher rate for subscribing the term deposit than other groups. For the marketing, bank can put more effort for those groups.

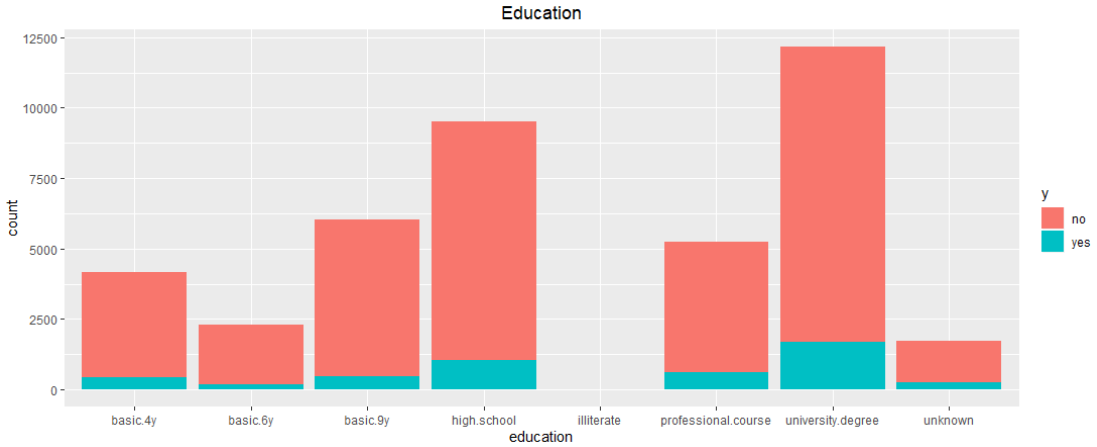

Figure 4- Variable: education

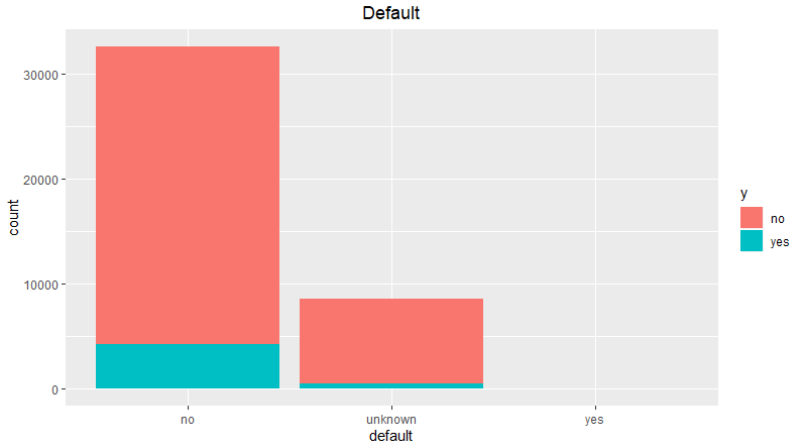

Figure 5 - Variable: default

Figure 6- Variable: housing

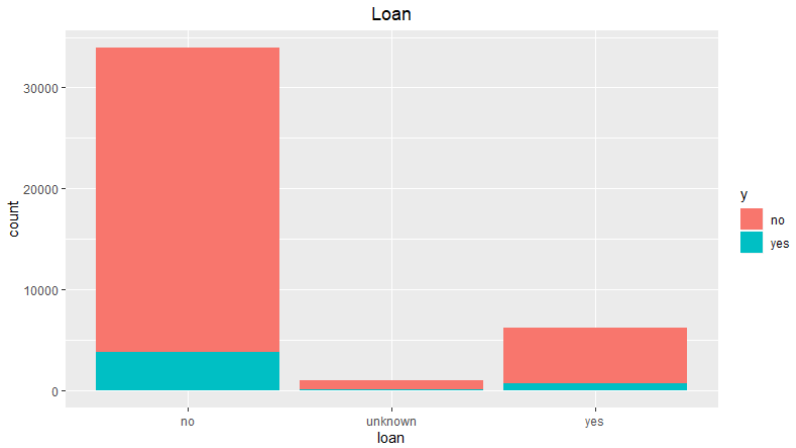

Figure 7 - Variable: loan

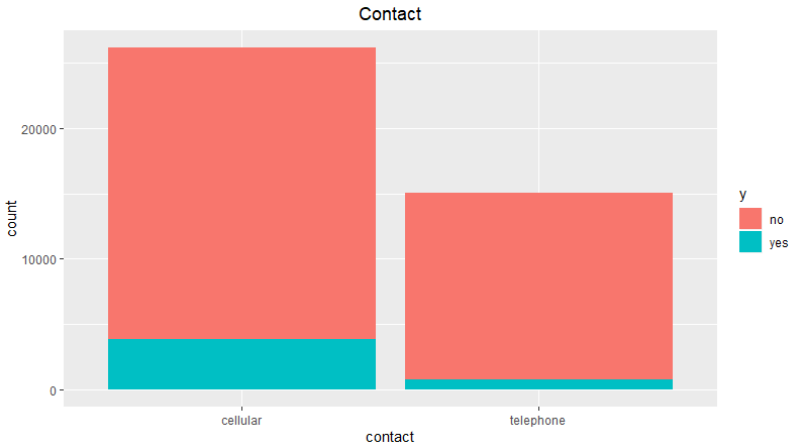

Figure 8 - Variable: Contact

In the chart above, we observed that people with cellular devices tend to buy the bank product more than the ones using simple telephones. Customers with cell phones take 65percentageof the data which can be attributed towards technological advances as more and more people have cell phones. It is also easy to reach out to people with cell phones because of the portable nature of the phone, therefore leading to higher success for selling term deposits.

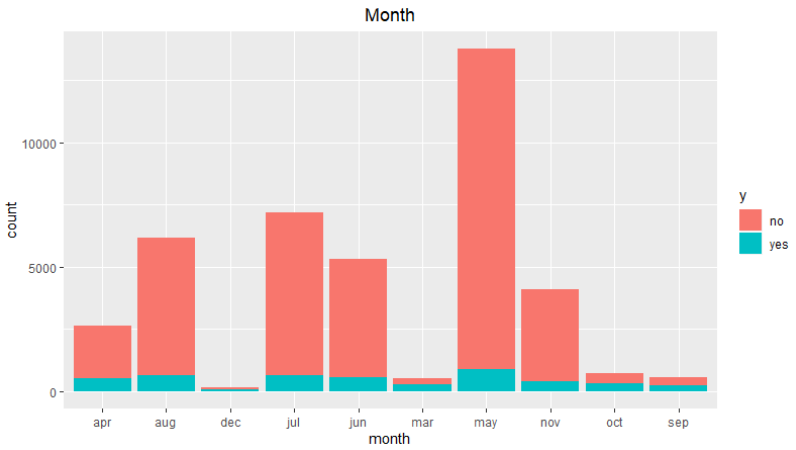

Figure 9 - Variable: Month

In the above charts, we describe the number of term deposits sold on the base of different months during the year. The month of May saw the most number of counts and December saw the least. The percentage of products sold to the total number of counts in December is the highest.

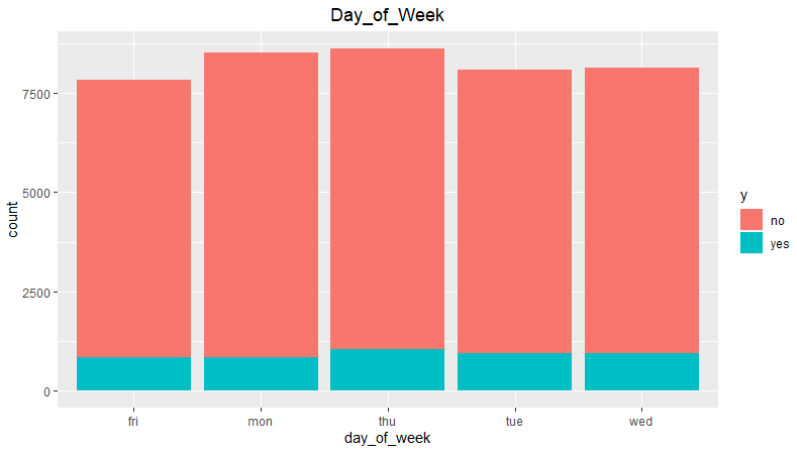

Figure 10 - Variable: Day of Week

Different days of the week represent last contact day of the week. This variable has a minor effect on the outcome as most days are pretty much on the same level. Most number of counts was on Thursday and the least was for Friday.

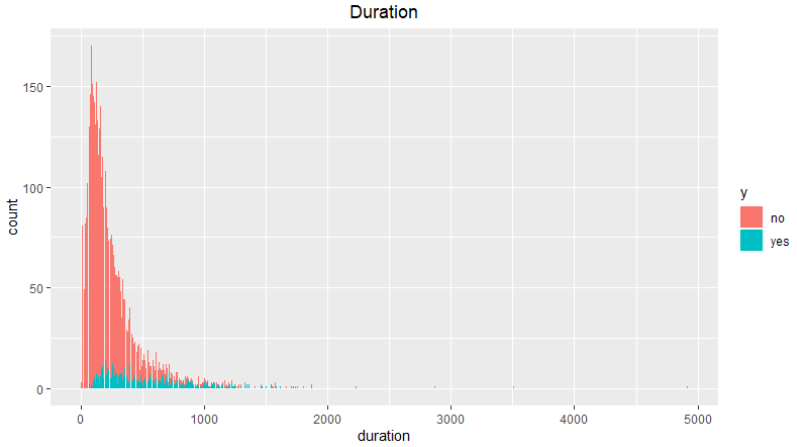

Figure 11 - Variable: Duration

Here we observe the bank products sold according the duration of the last phone calls. The more duration (seconds) had the highest number of counts which explains that most calls required longer duration due to explaining the bank products to customers. Longer duration also probably indicates higher complications with the application. However, lesser duration calls have higher percentage of products sold to products not sold.

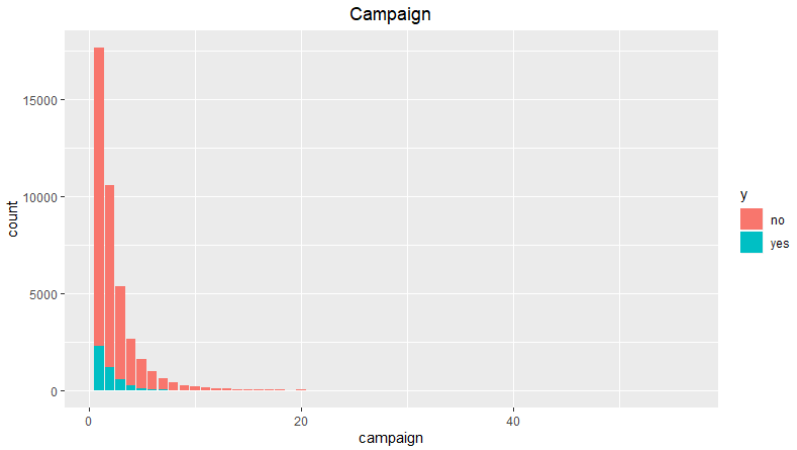

Figure 12 - Variable: Campaign

The greatest number of contacts were performed during the first few campaigns. Towards the later campaigns, the number of calls dropped as most people in the later campaigns were not interested in getting the bank product. If anyone is interested in investing in fixed deposit, they would do it in the beginning and would not require more campaigning for the product.

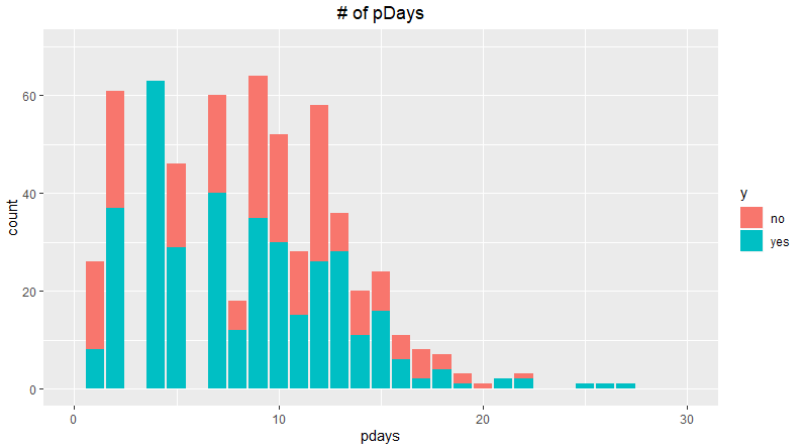

Figure 13 - Variable: pdays (Part 1)

The above and below variable graphs explain the number of pdays passed since the client was last contacted from previous campaigns. The maximum number of products were sold when the client was called within a week or so. The calls significantly dropped after two weeks. The below graph depicts 999 days which means the client was never contacted before. The percentage of products sold were very low in the case when client was never called before which means clients need some time to think about investing in a fixed deposit.

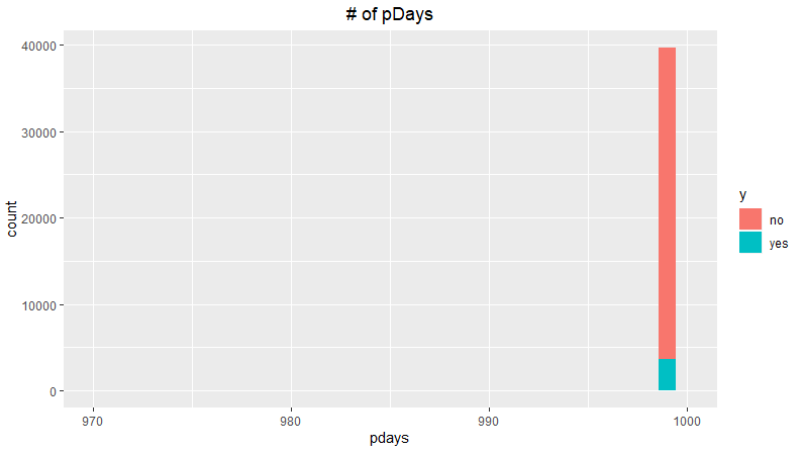

Figure 14 - Variable: pdays (Part 2)

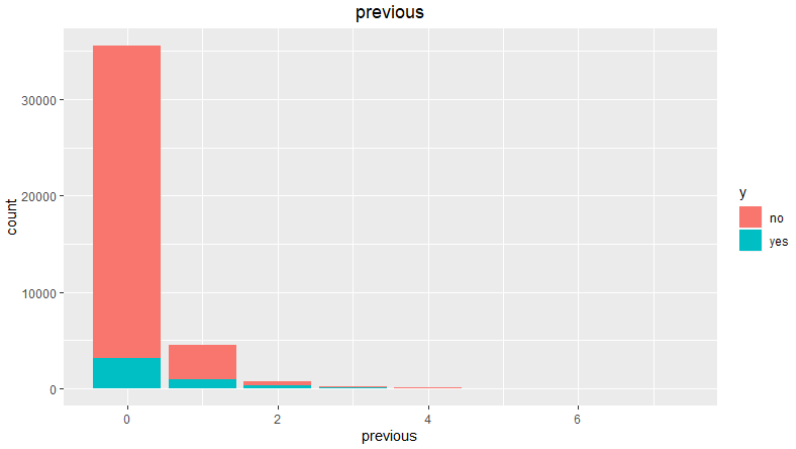

Figure 15 - Variable: previous

The above graph represents the number of contacts performed before this campaign. Mostly no contact was made before a campaign, but only a few of them were made. Products were sold highest in the case where no contact was performed.

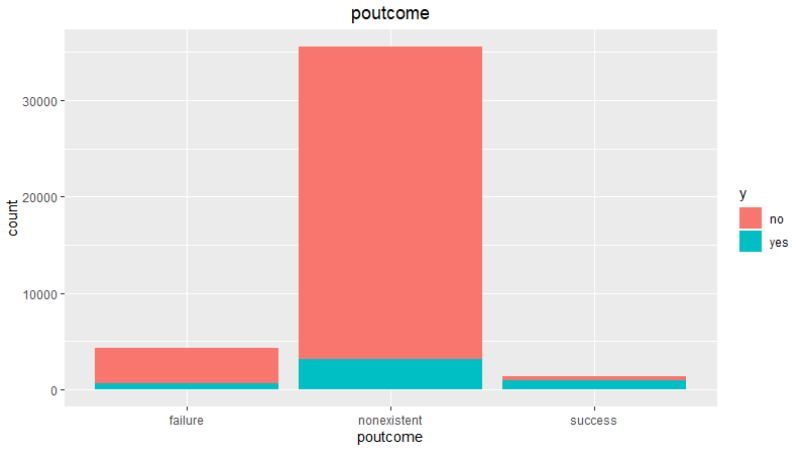

Figure 16 - Variable: poutcome

The above graph represents the outcome of the previous marketing campaign. Mostly the data was nonexistent which is the outcome was unknown. There were more success outcomes than failures.

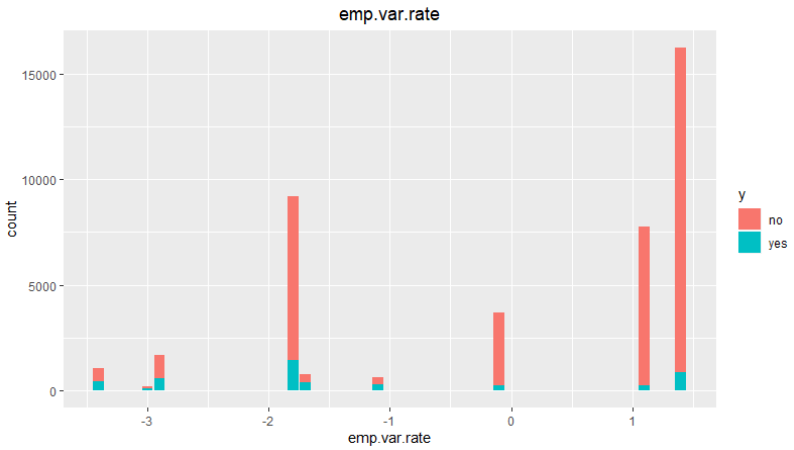

Figure 17 - Variable: emp.var.rate

Employment variation rate measures the variation of employment terminations and new-hires due to the shift of the economy.

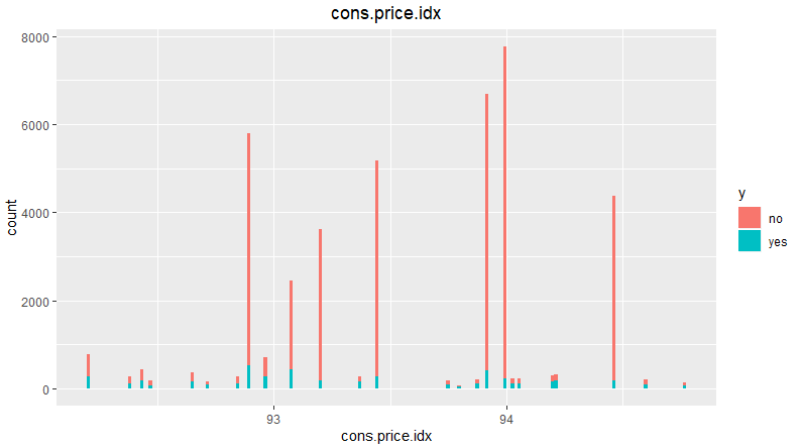

Figure 18 - Variable: cons.price.idx

Variable consumer price index of 94 has a very high rate of calls made. Most of the bank products were sold when the CPI was 93 which indicates the bank's fixed deposit rates might be higher.

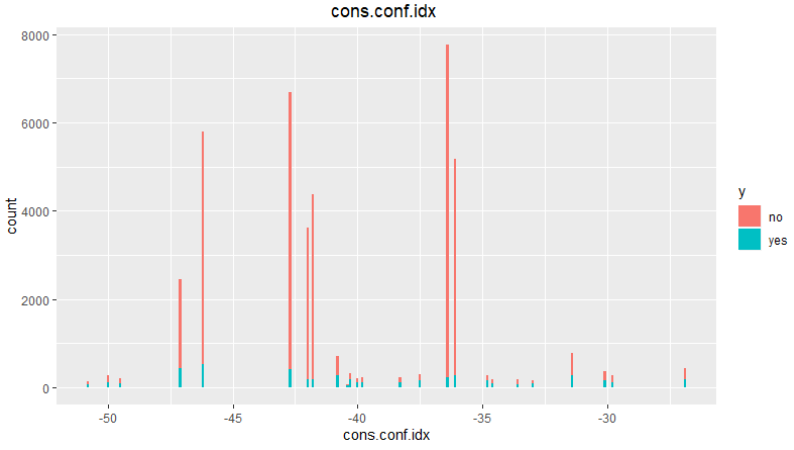

Figure 19 - Variable: cons.conf.idx

When the consumer confidence rate was –36, that is when a lot of offers were made to invest in fixed deposit. The counts drastically dropped when the consumer confidence index was –35 and above.

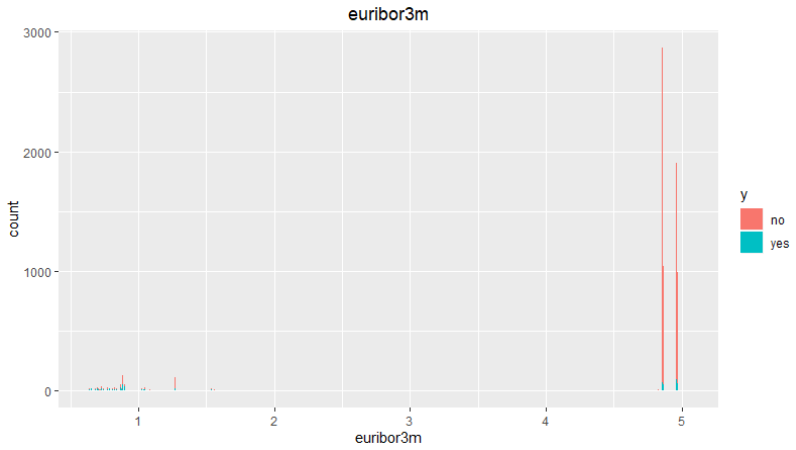

Figure 20 - Variable: euribor3m

Euribor3m is the interest rate decided by the Euro Interbank, it is the interest rate at which the bank lends money to an individual/entity with a maturity of three month.

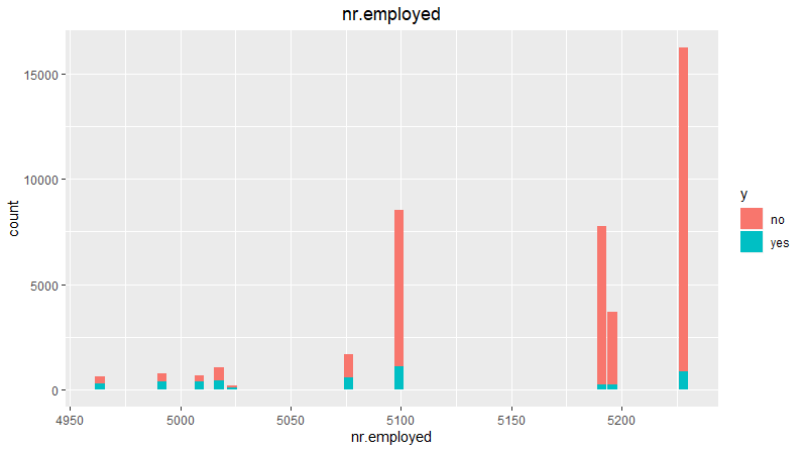

Figure 21 - Variable: nr.employed

The more the number of employees, the higher rates of counts. When the number of employees were 5250, the maximum number of counts were experienced.

2.2. Data Cleaning

Figure 22 – Correlation of variables and outcome

For numerical variables, euribor3m (euribor 3 month rate) and emp.var.rate (employment variation rate) showed 0.94 correlation, nr.employed (number of employees) and emp.var.rate (employment variation rate) showed 0.94 correlation, and finally nr.employed (number of employees) and euribor3m (euribor 3 month rate) showed 0.93 correlation. In short, 3 variables (euribor3m, emp.var.rate, nr.employed) were highly correlated each other, we can use only one variable among them. For our project we will use euribor3m.

2.3. Data partition

Original data set had 41188 records and 21 variables. Among the sample only 11.3 percentage had target value of yes. Since our target is to predict the success(yes) of marketing calls, in order to overcome the imbalance of the data for model prediction, we used 50 percentage eyes data for training data. For this, we divided the data into yes and no data by the target(y). And then we randomly select 3000 from yes data and 3000 from no data, 6000 in total for training data. We used balanced training sample (50 percentage yes, 50 percentage no for target) for model building and used rest of 35188 rows data for validation (4.66 percentage yes, 95.34 percentage no for target).

Target(y)	raw data	Training data	Validation data
yes	4640	3000	1640
no	36548	3000	33548
total	41180	6000	35188

Table 3 Data partition

3. Data Classification

3.1. Logistic Regression

Logistic Regression is a powerful model-based classification model. It is similar to Linear Regression, except the dependent variable is categorical and is a logit function of Y. Predictors can be either categorical or continuous numeric. The logit can be mapped back to a probability of the occurrence of a particular category.

3.1.1. Variable Selection

In Logistic Regression, correlated predictors introduce bias in the method. As in the data cleaning part, we took out two other variables that have a high correlation with euribor3m. Also, overly complicated models have the danger of overfitting. Thus, we used three variable selection methods: Forward, Backward, and Stepwise. The Forward and Stepwise approach chose the same variables for outcome logit Y. And the Backward approach gave slightly different ones.

Forward/Stepwise Selection

After running forward selection, the model chose 19 significant variables, including 13 dummy variables created earlier.

	OR	SE	95% CI, lower	95% CI, upper	p value
duration	1.0067145	0.0002003089	0.0070820213	0.0070820213	5.539511e-248
euribor3m	0.4594111	0.0165983896	-0.7069969025	-0.7069969025	8.465018e-103
month_may	0.2796825	0.0304442648	-1.0607522645	-1.0607522645	1.204651e-31
poutcome_success	3.3513169	1.5570767318	2.1199847476	2.1199847476	9.243694e-03
age_E	2.7355804	0.5668702835	1.4124897436	1.4124897436	1.195549e-06
education_university.degree	1.4666238	0.1263508627	0.5518155345	0.5518155345	8.778499e-06
poutcome_nonexistent	1.6464654	0.2019875163	0.7390781821	0.7390781821	4.813537e-05
cons.conf.idx	1.0342510	0.0080437749	0.0489209267	0.0489209267	1.489861e-05
cons.price.idx	1.2606951	0.1075820178	0.3989177031	0.3989177031	6.632873e-03
default_no	1.3956774	0.1715922960	0.5743486962	0.5743486962	6.695804e-03
age_A	3.9062533	2.0829074339	2.4076782265	2.4076782265	1.060782e-02
age_B	1.3131235	0.1272679603	0.4623684156	0.4623684156	4.944090e-03
marital_unknown	0.1711651	0.1337843058	-0.2331998261	-0.2331998261	2.392557e-02
campaign	0.9552307	0.0205497999	-0.0036378846	-0.0036378846	3.324877e-02
month_nov	0.7815818	0.1140946160	0.0396783733	0.0396783733	9.138140e-02
day_of_week_wed	1.2127588	0.1207244048	0.3880029414	0.3880029414	5.264888e-02
pdays	0.9992681	0.0004491921	0.0001489019	0.0001489019	1.033726e-01
job_technician	1.1961029	0.1294955492	0.3912633083	0.3912633083	9.812921e-02
month_Other_Month	1.2113268	0.1456976607	0.4274595901	0.4274595901	1.109529e-01

Figure 23 - Logistic Regression Output Table (Forward/Stepwise)

After running backward selection, the model chose 20 significant variables, including 14 dummy variables created earlier.

	OR	SE	95% CI, lower	95% CI, upper	p value
duration	1.0067140	0.0002003190	0.0070815749	0.0070815749	6.350292e-248
campaign	0.9560630	0.0205447067	-0.0028141135	-0.0028141135	3.653516e-02
pdays	0.9986649	0.0001796377	-0.0009834193	-0.0009834193	1.109931e-13
cons.price.idx	1.2787948	0.1070690362	0.4100190225	0.4100190225	3.312347e-03
cons.conf.idx	1.0338475	0.0080242220	0.0484995683	0.0484995683	1.796725e-05
euribor3m	0.4598690	0.0165516920	-0.7062701807	-0.7062701807	2.603822e-103
age_B	0.4659822	0.0999999270	-0.3429988919	-0.3429988919	3.732925e-04
age_C	0.3595987	0.0719874275	-0.6304049353	-0.6304049353	3.238262e-07
age_D	0.3868327	0.0794405606	-0.5472617965	-0.5472617965	3.748991e-06
marital_married	5.5402672	4.2338473421	3.2098384177	3.2098384177	2.507014e-02
marital_single	5.9900128	4.5811575948	3.2890726350	3.2890726350	1.925249e-02
marital_divorced	5.3528284	4.1329750850	3.1909338973	3.1909338973	2.979700e-02
education_Basic_education	0.6586551	0.0699682836	-0.2093501706	-0.2093501706	8.469615e-05
education_high.school	0.6679135	0.0695482921	-0.1995100696	-0.1995100696	1.061975e-04
education_professional.course	0.8030458	0.1014079553	0.0281590452	0.0281590452	8.239184e-02
default_no	1.4077866	0.1748436669	0.5854414220	0.5854414220	5.890282e-03
month_may	0.2980666	0.0311373663	-1.0056916778	-1.0056916778	4.791755e-31
month_Other_Month	1.2829725	0.1485706467	0.4761471728	0.4761471728	3.141508e-02
day_of_week_wed	1.2140540	0.1208247727	0.3890242417	0.3890242417	5.129862e-02
poutcome_failure	0.5692807	0.0678019995	-0.3299476898	-0.3299476898	2.242124e-06

Figure 24 -Logistic Regression Output Table (Backward)

We can tell from the p-value from both variable selection approaches that those predictors are significantly contributing to the outcome variable, especially "duration" and "euribor3m."

Odds Ratio (OR) plays an essential role in the logistic model. For categorical predictors, the Odds Ratio is the odds of A when B divided by the odds of A when not B. If the OR is greater than 1, A and B are positively correlated. The presence of B raises the odds of A. On the other hand, if the OR is less than 1, the occurrence of B reduces the odds of A. For example, the OR of p outcome_success is 3.3513. That means the odds of subscribing a term deposit for a person having success previous campaign outcome is 3.3513 times as the odds of subscribing a term deposit for a person not having success previous campaign outcome. For continuous predictors like "euribor3m",
the odds of subscribing a term deposit decreases by 45.94percentageif the interest rate with a maturity of three months increases by 1.

3.1.2. Classification Using Cutoff

The model produces an estimated probability of being a "1", in this case, P(subscribe a term deposit). Now we can convert it into classification by establishing a cutoff level. If estimated P > cutoff, classify as "1". Thus, the model helps in classification as well as predicting the probability of subscribing a term deposit.
The Error Rate / Confusion Matrix / Sensitivity / Specificity at cutoff=0.5 are shown as below:

- Forward / Stepwise

```
> errfwd  [1] 0.1422644
> table(yhatfwd.class, test$y)
             yhatfwd.class       0          1
                         0         28761      219
                         1         4787        1421
> senfwd(test$y, yhatfwd.class)  [1] 0.8664634
> spefwd(test$y, yhatfwd.class)  [1] 0.8573089
```

- Backward

```
> errbwd  [1] 0.1423781
```

```
> table(yhatbwd.class, test$y)
        yhatbwd.class    0       1
                    0    28741   203
                    1    4807    1437
> senbwd(test$y, yhatbwd.class)   [1] 0.8762195
> spebwd(test$y, yhatbwd.class)   [1] 0.8567128
```

The Error Rate / Confusion Matrix / Sensitivity / Specificity at cutoff=0.3 are shown as below:

- Forward / Stepwise:

```
> errfwd  [1] 0.2190235
> table(yhatfwd.class, test$y)
        yhatfwd.class   0       1
                    0   25912   71
                    1   7636    1569
> senfwd(test$y, yhatfwd.class) [1] 0.9567073
> spefwd(test$y, yhatfwd.class) [1] 0.7723858
```

- Backward

```
> errbwd   [1] 0.217972
> table(yhatbwd.class, test$y)
        yhatbwd.class    0       1
                    0    25950   72
                    1    7598    1568
> senbwd(test$y, yhatbwd.class)   [1] 0.9560976
> spebwd(test$y, yhatbwd.class)   [1] 0.7735185
```

We can see that the results table for forward, backward, and stepwise is very similar. The Error Rate of both models is lower at cutoff=0.5. We want to focus more on the classified "1" because those are the number of people who will subscribe to a term deposit. More importantly, although cutoff=0.3 gives higher sensitivity, the effort put into the contact is much higher when giving similar actual positives. The Logistic Regression performed well, but we will investigate other models and determine which ones are better.

3.2. kNN

We first ran with all variables. Our best K=1 with cutoff= 0.5. The confusion matrix is as shown below:

		Y test	
Y predict		0	1
	0	22738	626
	1	10810	1014

Table 4 kNN Confusion matrix cutoff=0.5

Sensitivity	0.6182927
Specificity	0.6777751

Table 5 KNN - Cutoff 0.5 - Sensitivity & Specificity

When we changed cutoff to 0.3, best k=3. The confusion matrix is as shown below:

	Y test		
Y predict		0	1
	0	12129	207
	1	21419	1433

Table 6 kNN Cutoff 0.3 Confusion Matrix

Sensitivity	0.8737805
Specificity	0.3615417

Table 7 KNN - Cutoff 0.5 - Sensitivity & Specificity

From correlation graph (see *fig. 22*), we took out 3 variables that are much less correlated to the outcome. They are age, campaign and consumer confidence index. After which, we ran KNN again with cutoff=0.5. Our best K=1. The confusion matrix is as shown below:

	Y test		
Y predict		0	1
	0	25291	496
	1	8257	1144

Table 8 kNN Confusion matrix cutoff=0.5

Sensitivity	0.697561
Specificity	0.753875

Table 9 KNN - Cutoff 0.5 - Sensitivity & Specificity

We then changed the cutoff to 0.3. Our best K=3. The confusion matrix is as shown below:

	Y test		
Y predict		0	1
	0	19963	148
	1	13585	1492

Table 10 kNN - Cutoff 0.3 - Confusion Matrix

Sensitivity	0.9097561
Specificity	0.5950578

Table 11 KNN - Cutoff 0.5 - Sensitivity & Specificity

From our confusion matrixes, we can see there's minimum improvement before and after variables removal, and the different cutoffs. We get the best result when we ran with all variables and cutoff=0.3. However, the specificity is lowest in that scenario.
Given our data is heavily imbalanced, with only 1640 "1" and more than 30,000 "0", KNN isn't the best method for this data set.

3.3. Classification Tree

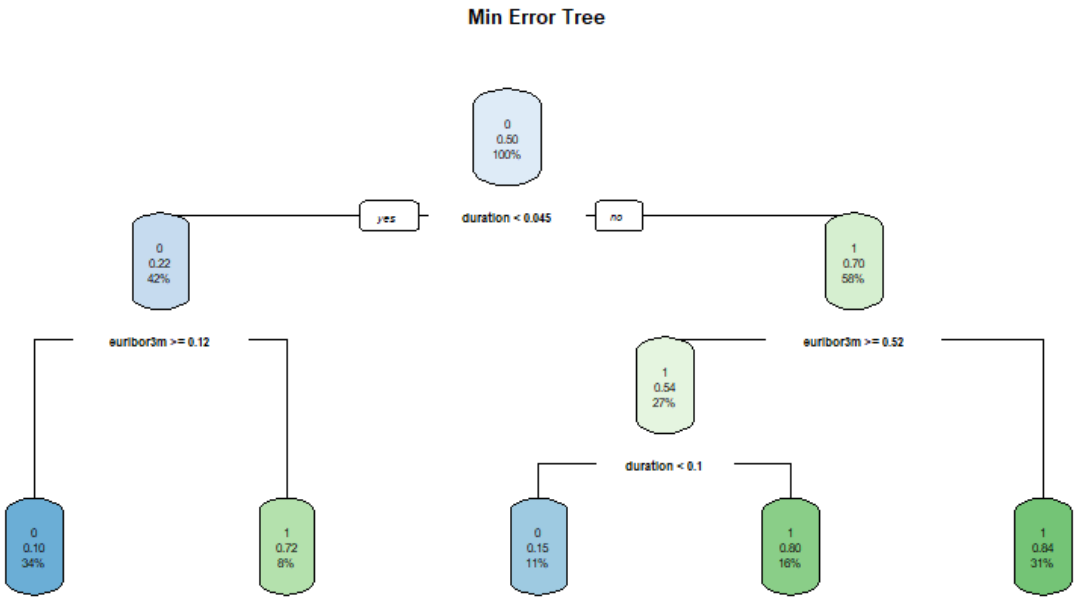

Figure 25 - Classification Tree - Minimum Error Tree

In the above chart, we have classified the data with a help of a classification tree with a minimum error rate. The duration of the phone calls is the first variable. The second variable analyzed in the chart is the Euribor3m which is the short-term loan rate of the bank to individuals and entities. The first step is about the duration of the phone call. If the calls are less than 0.045 (left side of the tree) the bank term deposit is not sold. Under this scenario, the telemarketing call is a success only if the Euribor3m rate is more than 12%, which can be attributed to higher long-term deposit rate making it more attractive. If the duration is more than 0.045, the bank deposit is sold in every case thereafter except when the duration of the phone call is less than 0.1. Longer duration of call may be associated with the bank agent getting more time to convince the customer. This decision tree explains that it is important to have a higher Euribor3m rate and also longer duration of call is desirable to have a successful telemarketing call.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 26889   180
         1  6659  1460

               Accuracy : 0.8056
                 95percentageCI : (0.8015, 0.8098)
    No Information Rate : 0.9534
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2403
```

```
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8015
            Specificity : 0.8902
         Pos Pred Value : 0.9934
         Neg Pred Value : 0.1798
             Prevalence : 0.9534
         Detection Rate : 0.7642
   Detection Prevalence : 0.7693
      Balanced Accuracy : 0.8459

       'Positive' Class : 0
```

We have calculated the accuracy measures of the decision tree. The accuracy is 80.56 percentage which means error rate is 19.44%. There was a total of 35,188 cases with '0' as not a success and '1' as success in the telemarketing calls. There were 26,889 cases and 1,460 cases which were correctly classified as not a success and success respectively. There were 180 cases which were a success but were predicted as not a success. Similarly, there were 6,659 cases which are predicted as success but actually they were not a success.

The sensitivity measures the percentage of "C1" class correctly classified, which is 80,15%. Specificity is the percentage of "C0" class correctly classified which is 89.02%, a number higher than sensitivity and is a rare case. Positive prediction value is the percentage of C1 that were not actually C1. more than the Negative prediction value which clearly highlights that C1 is not as properly classified as C0.

The cutoffs of 0.5 and 0.3 produced same results as it is seen in the confusion matrix. It produces the same outcome variable irrespective of the cutoff between 0.3 and 0.5. Below the information is provided in a table format:

Cutoff = 0.5

	Actual Class 0	**Actual Class 1**
Predicted Class 0	28201	144
Predicted Class 1	5347	1496

Table 12 Confusion matrix

Cutoff 0.3

	Actual Class 0	**Actual Class 1**
Predicted Class 0	25379	37
Predicted Class 1	8169	1603

Table 13 Confusion matrix

3.4. Random forest

Random forest is using the ensemble method, which combines several models to enhance the prediction performance. Random forest selects random samples with replacement and fit separate model for the sample and average them for the final prediction. Unlike a single tree, the random forest cannot plot dendrograms since it is like a black box model that does not show what went through between input and output. It provides a variable importance graph showing the contribution of variables; the mean decrease in the Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes.

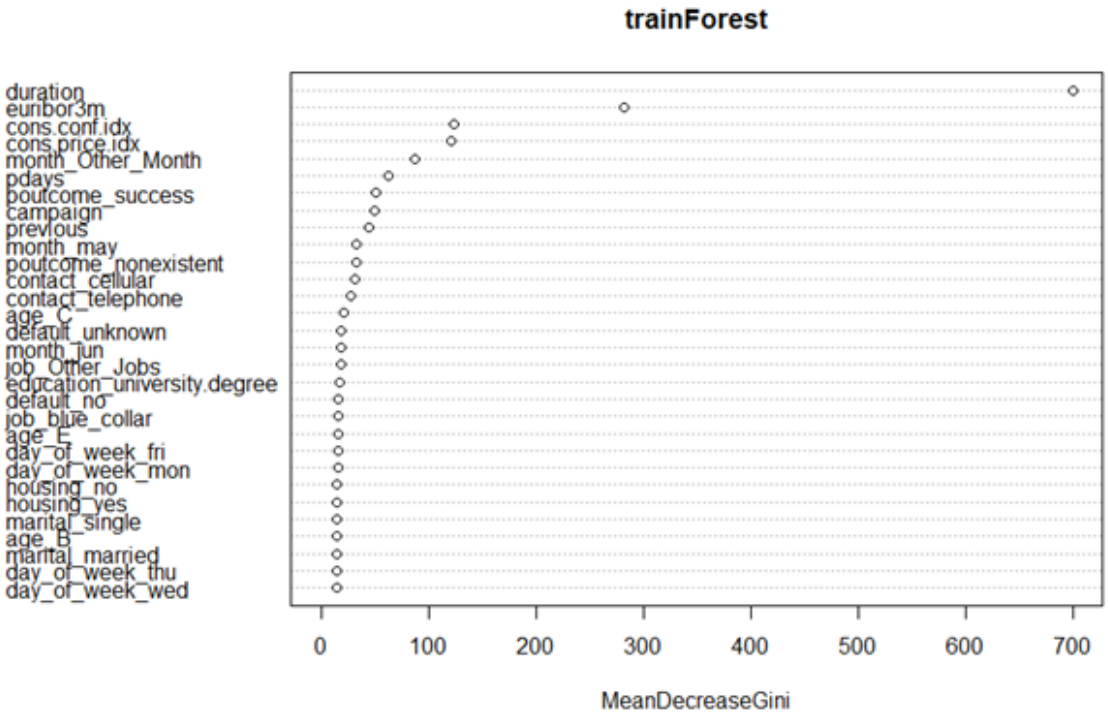

Figure 26 - Mean Decrease Gini

From the chart above, the duration and Euribor 3M interest rate were the most contributing variables for the random forest model. It does not mean it has a positive relationship with the increase of success of marketing results, but their value changes determine trees. The result for the cutoff=0.5 sensitivity was 91.2%. When the cutoff=0.3, sensitivity was 97.7%, which was about 6.2percentagehigher than cutoff=0.5. If the bank thinks to get more customers (107 more success) is more beneficial compared to promoting for more people (2822 more people), then they can choose the cutoff =0.3.

Cutoff = 0.5

- Sensitivity = 0.912
- Specificity =0.841

	Actual Class		
Predicted Class		0	1
	0	28201	144
	1	5347	1496

Table 14 Confusion matrix

Cutoff = 0.3

- Sensitivity =0.977
- Specificity =0.756

	Actual Class		
Predicted Class		0	1
	0	25379	37
	1	8169	1603

Table 15 Confusion matrix

3.5. XGBoost

XGBoost is using the boosting model, which gives higher selection probabilities to misclassified records to enhance the prediction. XGBoost can manage spare data; it stores data without string zeros, which can save memory and calculation time. It uses the quantile sketch, which transforms the data by weighting algorithm based on accuracy. Parallel computing is also possible for XGBoost that using multiple threading the phase of the search for the best split. XGBoost can be used both for prediction and classification.

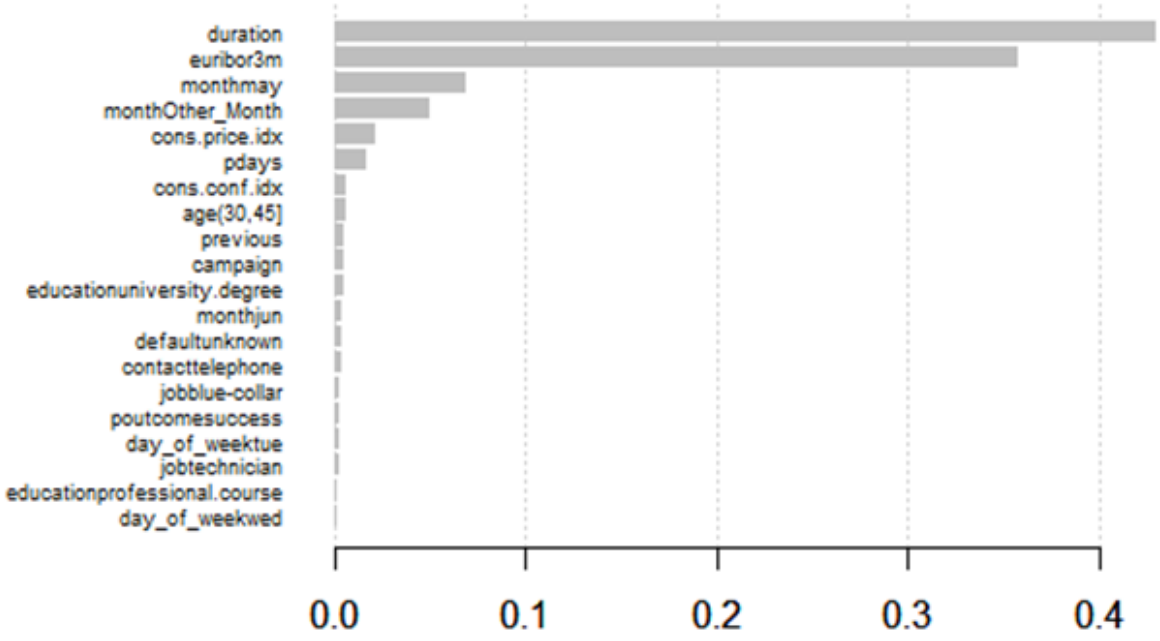

Figure 27 - Prediction and classification

From the chart above, the duration and Euribor 3M interest rate were the dominant contributing variables for the random forest model. The result for the cutoff=0.5 sensitivity was 93.8%. When the cutoff=0.3, sensitivity was 97.5%, which was about 3.7percentagehigher than cutoff=0.5. If the bank thinks to get more customers (60 more success) is more beneficial compared to promoting for more people (1741 more people), then they can choose the cutoff =0.3.

Cutoff = 0.5

- Sensitivity = 0.938
- Specificity=0.838

	Actual Class		
Predicted Class		0	1
	0	28100	101
	1	5448	1539

Table 16 Confusion matrix

Cutoff = 0.3

- Sensitivity= 0.975
- Specificity = 0.786

	Actual Class		
Predicted Class		0	1
	0	26359	41
	1	7189	1599

Table 17 Confusion matrix

4. Results

Models		Logistic regression			kNN		Classification tree		Random Forest	XGBoost
Methods		Forward	Backward	Stepwise	w/ all	Removed 3	Best pruned	Min Error		
cutoff=.5	Valid ER	0.143	0.142	0.143	0.325	0.249	0.194	0.194	0.156	0.158
	Sensitivity	0.876	0.876	0.876	0.618	0.698	0.890	0.890	0.912	0.938
	Specificity	0.856	0.857	0.856	0.678	0.754	0.802	0.802	0.841	0.838
cutoff=.3	Valid ER	0.219	0.218	0.219	0.615	0.249	0.194	0.194	0.233	0.206
	Sensitivity	0.957	0.956	0.957	0.874	0.698	0.890	0.890	0.977	0.975
	Specificity	0.772	0.774	0.772	0.362	0.754	0.802	0.802	0.756	0.786

Table 18 Result

From the result table, the reader can get the following information: Although when cutoff at 0.3 has a high sensitivity if the bank wants to detect more response will cost a lot. Cutoff equals to 0.5 is the best cost-efficiency cutoff value for all models. For cutoff at 0.5, the XGboost has the highest sensitivity. The specificity is higher for the random forest. If the bank chooses to increase 43 of the term deposit subscription, they must increase the number of telemarketing by 101 times. If the benefit from new followers is higher than the cost of 101 times telemarketing, XGboost is the best model for the bank. In contrast, the benefit is lower than the cost, and the bank should choose the random forest model.

5. Conclusion

Facing increasing pressure, the bank wants to make more money and reduce the cost. In that case, optimizing targeting for telemarketing is essential. Banks can use data mining based on a data-driven model to predict the result of telemarketing selling long-term deposit by phone call. Data mining also a useful tool for the bank to choose clients. Five data mining methods are compared in the table. They are logistic regression, KNN, decision trees, XGboost, and random forest. These models are compared using three metrics: error rate, sensitivity, and specificity.

Based on the analysis above, there are two optional models for the bank. The bank can choose the model based on the real situation. We determine these two models not only because it has a higher sensitivity and specificity, but also it is a useful tool when dealing with the big size data.

References

Hearty, J; Sjardin, B; Prateek. Python: Real World Machine Learning by Luca Massaron; Alberton Boschetti; Joshi Published by Packt Publishing, 2016.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31. https://doi.org/10.1016/j.dss.2014.03.001

Shmueli.G , Peter C. Bruce, Yahav. I, Nitin R. Patel, and Kenneth, Data mining for business analytics concepts techniques and applications in r answers, page 229-232.

Cover: https://d1qq9lwf5ow8iz.cloudfront.net/live-images-1/ImageDetail_fd3463c4-f15b-4991-9353-627cbc4e0274_Medium

Table of Figures

List of Tables