**Predicting Term Deposit Subscriptions**

Meltem Yildirim, Shaoshao Yang, Xiyun Zou

December 10th, 2018

**Abstract**

Inthis project, we aim to access telemarketing efficiency by identifying the main factors that affect the success of a campaign and also predicting whether a customer will subscribe to term deposit or not. We implemented several machine learning algorithms: logistic regression, decision tree, bagging, bootsing, and random forest, through which we observed the relationship between successful campaign and multiple customer attributes as well as validate and compare the effectiveness of each predictive model.

### **1. Introduction**

A term deposit is a cash investment held at a financial institution. It is a deposit with a specified period of maturity and earns interest. Nowadays, telemarketing has been one of the most general marketing campaigns used by the bank to solicit prospective customers to subscribe to term deposit over the phone. It is beneficial for promoting products, narrowing down the range of of potential customers, and reducing the cost of marketing efficiently. Our objective is to build a classifier to predict whether or not a client will subscribe to a term deposit.

### **2. Data Source**

The dataset used for this project is published on UCI Machine Learning Repository. It is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phones calls. Often, more than one contact to the same client was required, in order to access if the product would be subscribed or not. There are in-total 17 attributes and 45211 observations.

### **3. Exploratory Data Analysis**

Dataset has more than 31,000 observation with 17 variables in training set, and more than 13,500 observation in the test set. All the attributes and brief information as follow. Our goal here to identify most important predictors for the model. One important note here is that the variable ‘duration’ represents last contact duration in second and it has highly effect on 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, we discarded this variable for intention is to have a realistic predictive model.

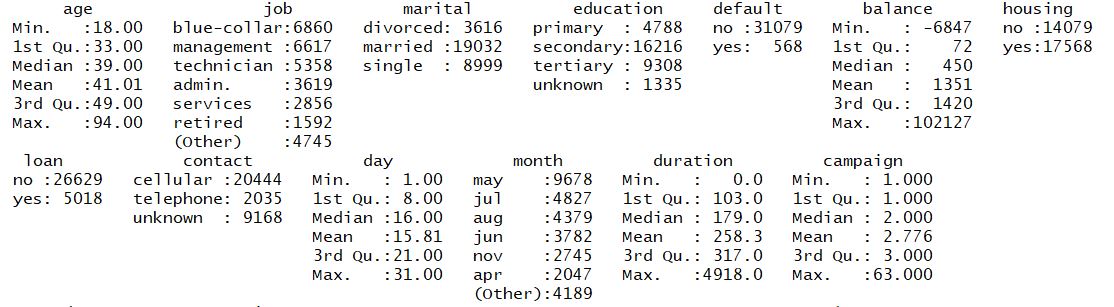
Attributes information;

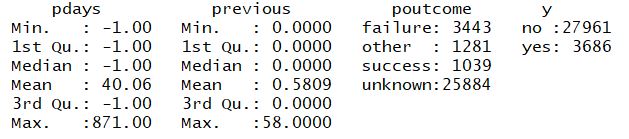
1. Age: Client age (numeric)
2. Job: Type of Job (Categorical: “ admin”, “unknown”, “unemployed”, “management”, “student”, “blue-color”, housemaid”, “entrepreneur”, self-employed”, “retired”, “technician”, “services”)
3. Marital : Marital status (categorical - "married","divorced","single"; note: "divorced" means divorced or widowed)
4. Education: Education Level (categorical: "unknown","secondary","primary","tertiary")
5. Default: Has a credit in default? (binary: "yes","no")
6. Balance: Amount of balance in the account (numeric)
7. Housing: Has a housing loan?(binary: "yes","no"l)
8. Loan: Has a personal loan?(binary: "yes","no" - related with the last contact of the current campaign:)
9. Contact: Contact communication type(categorical: "unknown","telephone","cellular")
10. Day: Last contact day of the week (numeric)
11. Month: Last contact month of the year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
12. Duration: Last contact duration in second (numeric)
13. Campaign:Number of contacts performed during this campaign(numeric)
14. pdays: Number of days that passed by after the client was last contacted from previous campaign(numeric)
15. previous: number of contacts performed before than this campaign (numeric)
16. poutcome: outcome of the previous marketing campaign(categorical: “unknown”, “other”,"failure","success")

Response variable

1. Y : Has the client subscribed a term deposit?(binary ‘yes’ and ‘no’)

Having a look at the rough distributions (min, max, mean, median, quartiles) of the variables in the dataset;



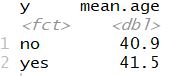


Percentage of people that subscribe term deposit in train set is 0.1164723 (11.65%)

We see that most of the clients in the dataset have not subscribed the term deposit .

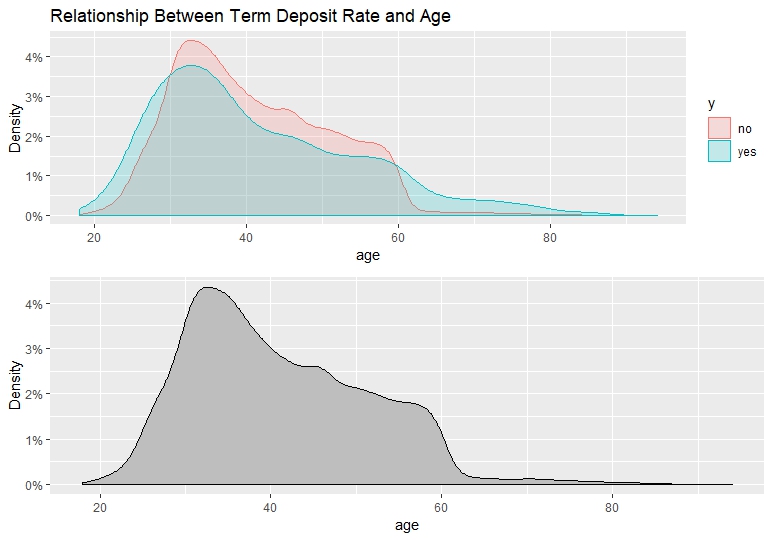
Here is closer look to relationship between outcome variable “y” with each variable;

**3.1. Age;**



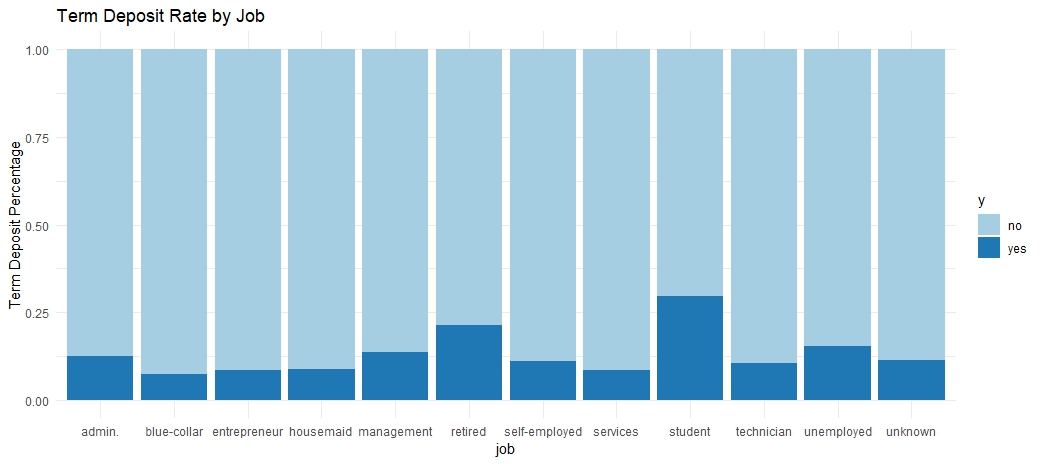
From this table, we see that there isn't much of a difference in the mean ages of clients who have a term deposit versus those who do not.

Probability density curve of the effect of age on outcome variable;

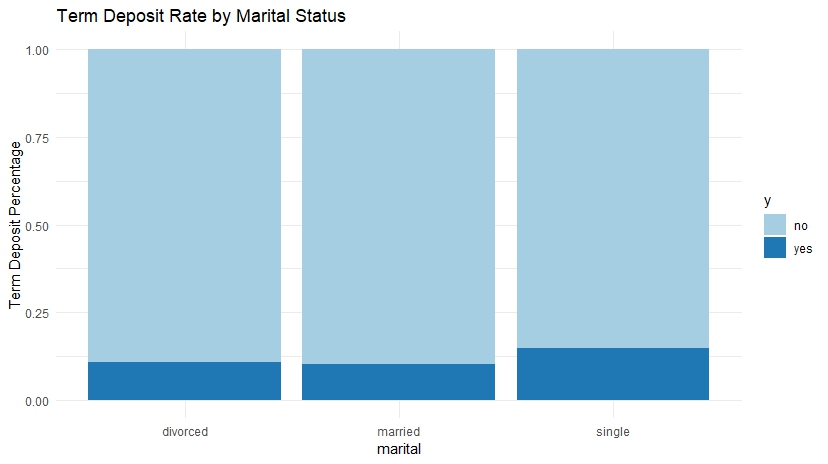


As we see in these graphs, the distribution of clients with and without term deposits is notably different after age 60, relative to other ages. After age 60, amount of people who subscribed to term deposit is significantly larger than the people who didn’t not. We also see that the distribution of clients with and without term deposits is more or less the same for middle ages.

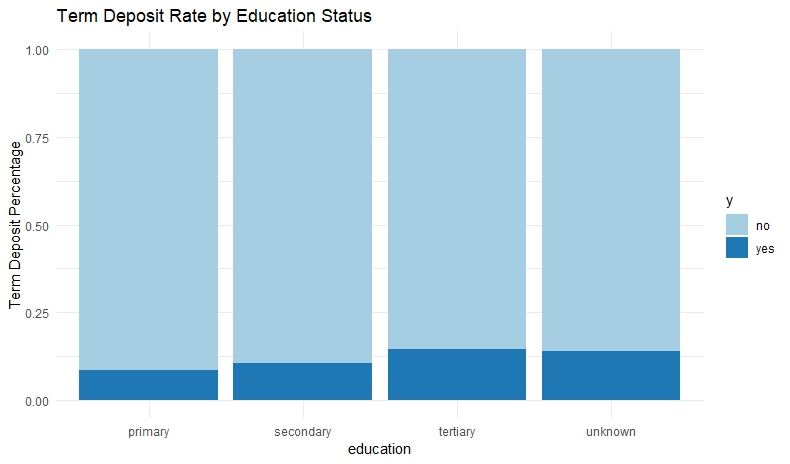
**3.2. Job;**



Job is categorical variable with 13 level; Administrator positions, blue color position, entrepreneur, management roles, retired, self-employed , service , student, technician, unemployed, unknown. In the histogram above we see that students have the highest chance of having term deposits followed by retired clients. Clients who are unemployed or have a job in management or administration also have a higher likelihood of having term deposits comparing to other positions.

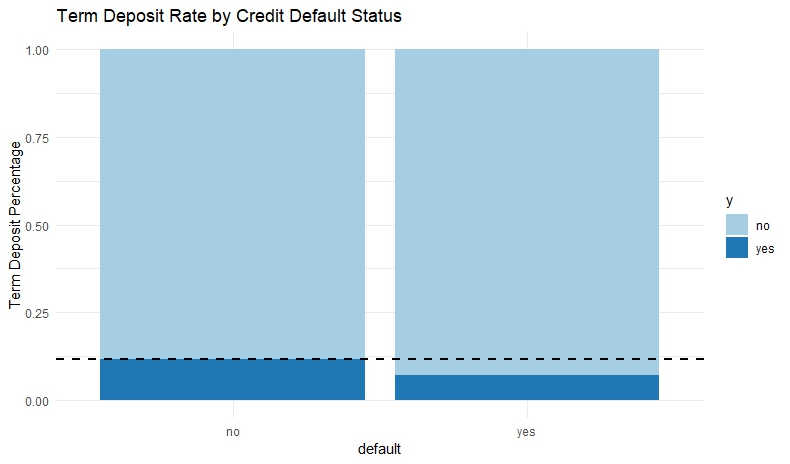
**3.3. Marital status;** 

We have four levels in this variable ; single, married, divorced (includes widowed) and unknown which is not specified. Although we see that ratio of single clients is slightly higher than others , yet difference might not be statistically significant.

**3.4. Education;**

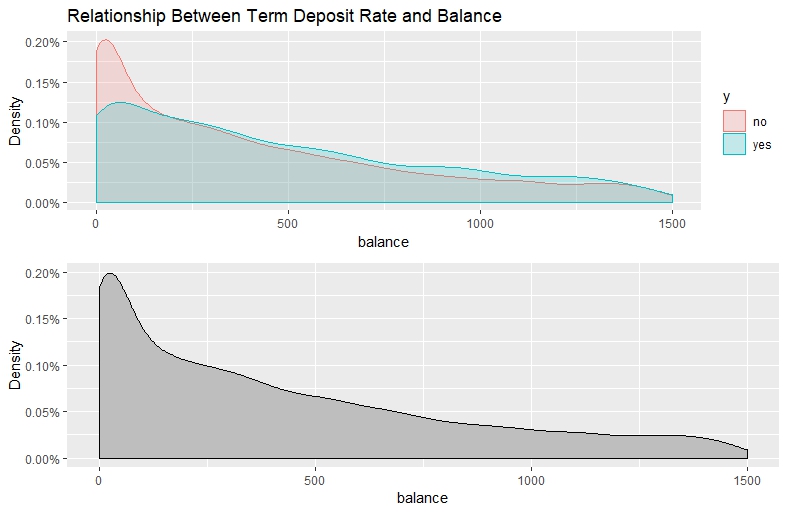
Education has also four levels ; primary education, secondary, tertiary and unknown. Clients with a tertiary education status have a slightly higher rate of term deposits. However, as it is in marital status, the difference might not be statistically significant.

**3.5. Credit Default;**



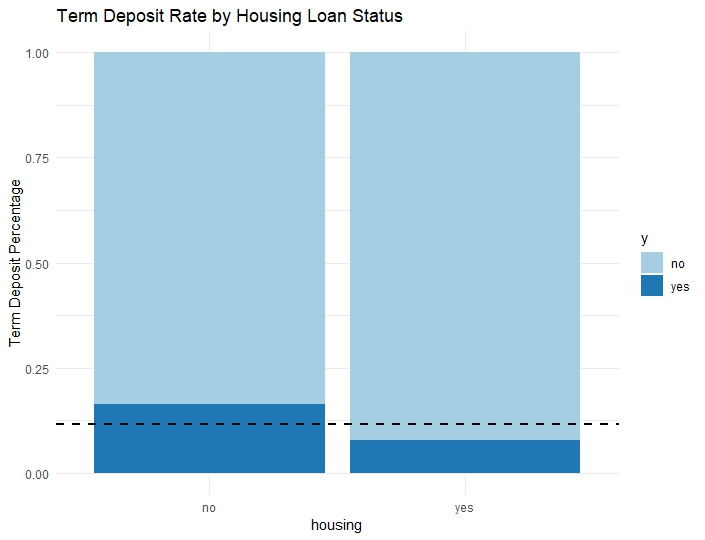
The black dash- line represents the overall ‘yes’ percentage of clients who want to subscribe to term deposit . We see that clients who have a credit in default have a lower likelihood rate to subscribe term deposits than the clients who don’t have credit default .

**3.6. Balance;**



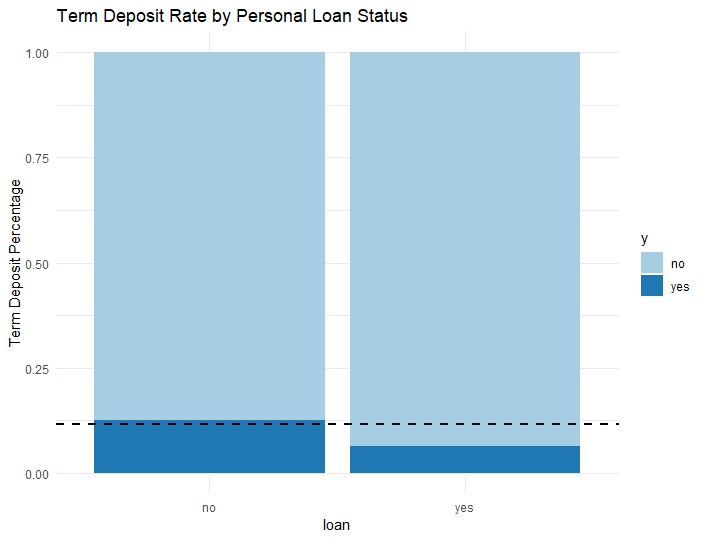
From this graphs above, we see that there is a difference in the mean balance of clients who want to subscribe a term deposit versus those who don’t want. In second plot we restrict outliers and draw the distribution between the 1st and 3rd quartiles approximately. The distributions for both groups closely mimic the overall distribution. No appreciable difference is found except the ‘yes’ cohort's distribution offsetting the overall distribution by a small amount when the balance get above 1000.

**3.7. Housing;**



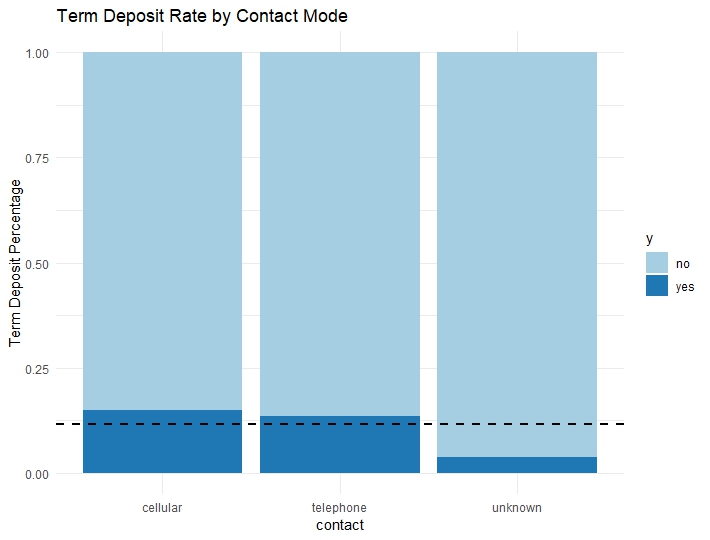
According the graph clients without a housing loan have a greater rate of term deposits, and the ones with a loan have a lower rate.

**3.8. Loan;**



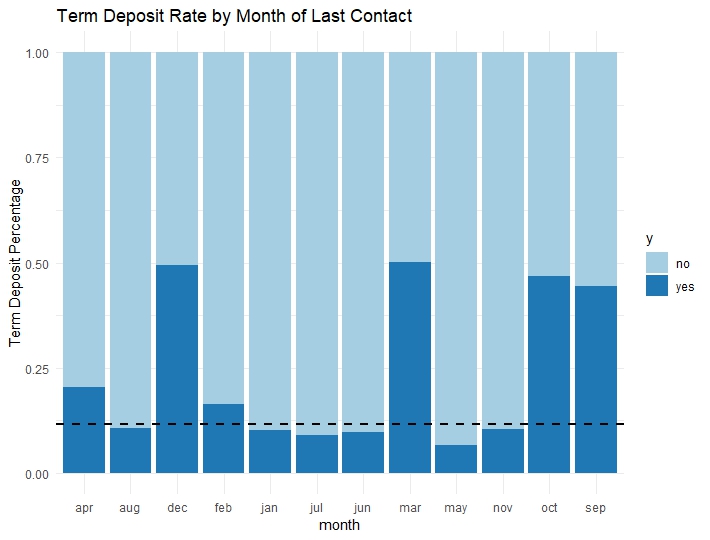
The ones with a personal loan are expected to have lower likelihood of term deposits. However, the effect is less pronounced than those with housing loans.

**3.9. Contact;**



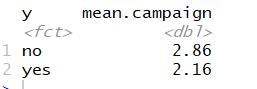
As we see in the graph , the clients who have unknown type of contact communication have very low chance to subscribe to term deposits. Telephone and Cellular type of contact communication rate are above than average rate.

**3.10. Months;**

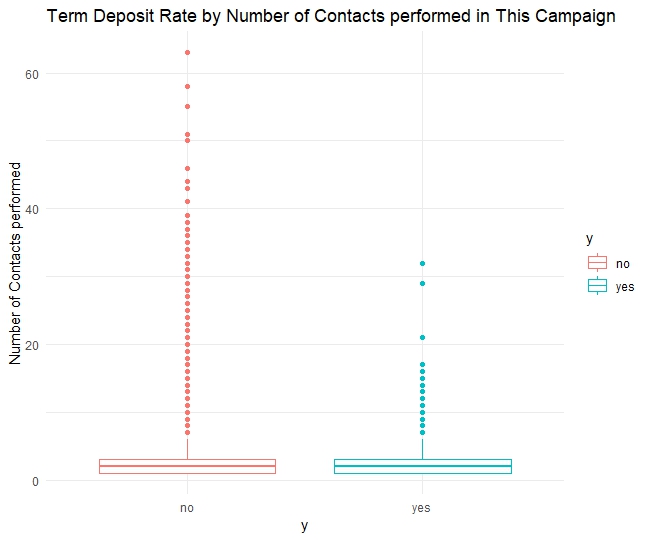


People who were last contacted in December, March, October and September have a pretty high (~0.50) likelihood of subscribing to term deposits.

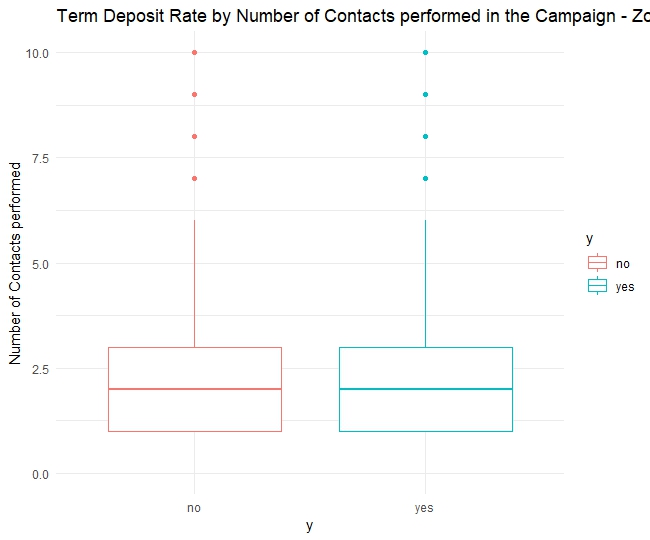
**3.11 Campaign;**



Mean differences tell us that there isn't a significant difference between the average number of contacts performed for the ‘yes’ and ‘no’ cohorts at a first glance.

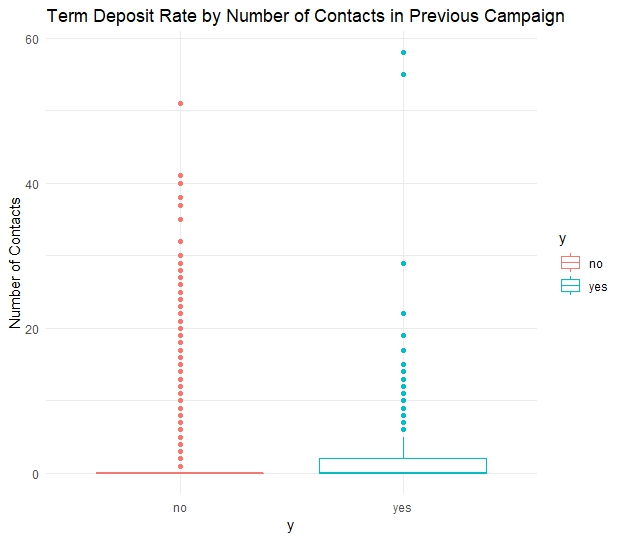


A box plot gives us some insight into the outliers. We find that almost all of the outliers above 20 belongs to ‘no’ category.

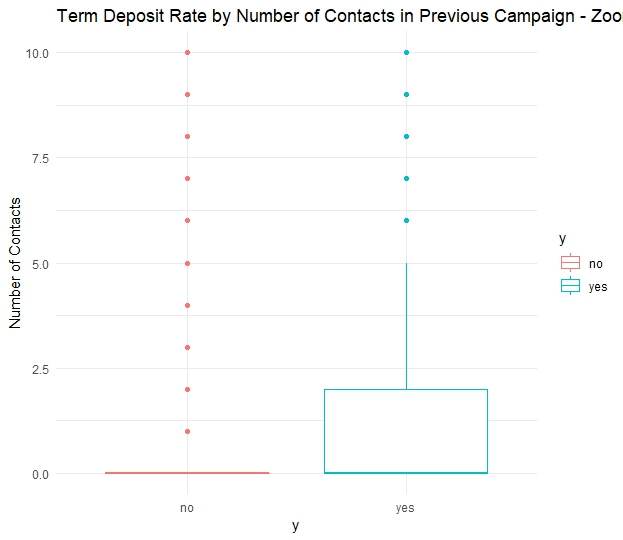


When we limit the outliers and take a closer look at the boxplot, we see that no appreciable difference now. So, number of contacts above 20 end up being in ‘no’ category.

**3.12 Previous**

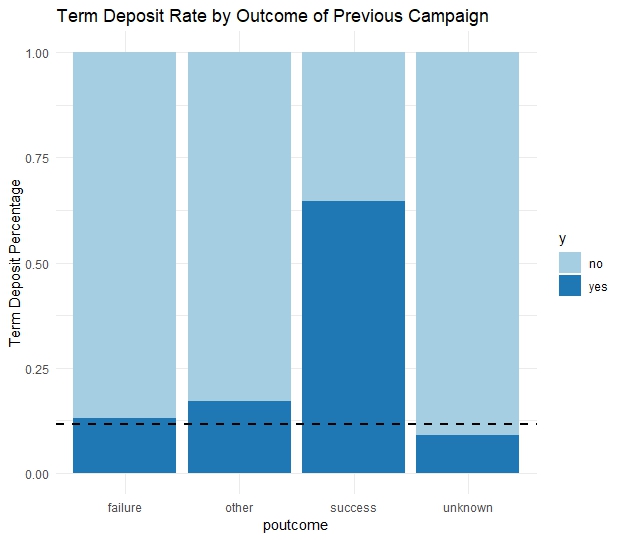


Again, we find that almost all of the outliers above 20 belong to the ‘no’ category.



When we limit the outliers and take a closer look at the boxplot , it doesn't tell us much except that the distribution of the ‘yes’ clients is wider than the ‘no’ ones.

**3.13 Putcomes**



As expected, clients with a success in the previous marketing campaign have a high likelihood of having a term deposit. Also we that the rate of likelihood of those have failure value is not very different from the overall average.

**3.14 Conclusion:**

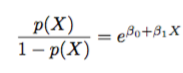
Based on our EDA, we found that the outcome of the previous marketing campaign (potcome) is the most suitable predictor for model.

### **4. Methodology**

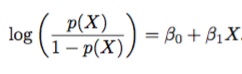
Predicting whether or not a client will subscribe to term deposit is a binary classification problem. Classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known. There is a lot of classification algorithms available now. For this project, we will be focusing on logistic regression, decision tree, bagging, boosting, and random forest.

#### **4.1 Logistic Regression**

Logistic regression a form of binomial regression which estimates the parameters of a model that has a binary dependent variable with two possible values. The odds and log-odds of the response taking a particular value are modeled based on a combination of one or more independent variables:



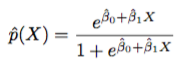
and



The quantity log(p(x)/[1-p(x)]) is called the odds, and can take on any value between 0 and 1. Logistic regression chooses parameters that maximize the likelihood of observing the response taking a particular value:



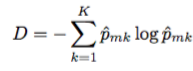
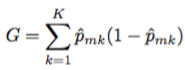
The estimates beta-0 and beta-1 are chosen to maximize this likelihood function. Once the coefficients have been estimated, the probability of the response taking a particular value is calculated:



The resulting probability is the predicted probability and is further used to classify the observed instance to one of the classes. Usually, the cutoff value is 0.5.

#### **4.2 Decision Tree**

A classification tree if used to predict a qualitative response. It predicts that each observation belongs to the most commonly occuring cass of training observations in the region to which it belongs. The task of growing a classification tree involves using recursive binary splitting. Either the Gini index or the cross-entropy are typically used to evaluate the quality of a particular split.



Where G (Gini index) is a measure of total variance across the K classes. A small value indicates that a node contains predominantly observations from a single class (pure node). D (cross-entropy) is very similar to Gini index such that it take on a small value if the mth node is pure. Another measure used in growing the tree is the classification error rate which is the fraction of the training observations in that region that do not belong to the most common class:

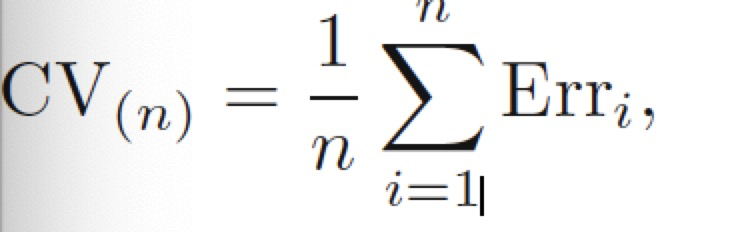


**4.3 Tree-Based Modeling -**

Bootstrapping approach: Bagging, Random Forest, and Boosting method

* **Parameters turning in models training -**

In order to choose the model parameters that results in the best model performance, k-fold cross-validation was applied in the training dataset. 10-fold cross-validation were used initially. However, the model training was really time consuming due to heavy computation processes on sampled training and validation set. Instead of using 10-fold cross-validation, 5-fold cross-validation were used in choosing the optimal parameters for all the following models in this project.



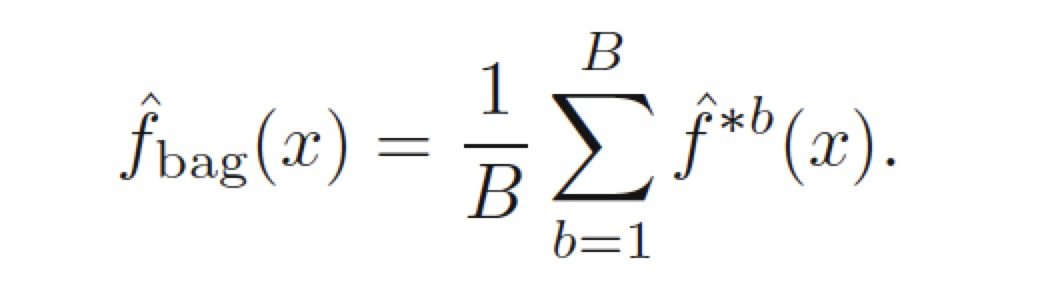
Where, n=3, Err is the classification error.

* **Upsampling samples V.S Original samples -**

Given that the ‘yes’ class data contributed about 10% to the total dataset, the composition of target variable y in the given dataset is imbalance. Due to the concern of imbalance data may turn in less accurate models, in addition to using the original training data for models training, upsampling method was also applied to the training dataset while building the models in hope of a better overall models performance in the test dataset. That is, when choosing the final model, models built with upsampling method are also taken into consideration.

* **Bagging -**

Bootstrap aggregation(bagging), generating repeated samples from the training data set, to train the method on each dataset, and finally combine models by averaging all the predictions, to obtain the bagging function:



* **Random Forest -**

Random forests provide an improvement over bagged trees by averaging the trees to reduces the variance. When building a number of decision trees on the bootstrapping training samples, each time a split in a tree is considered, a random sample of m predictors is chosen as split from p predictors. Typically the number of predictors considered at each split is approximately equal to the square root of the total number of predictors.

* **Boosting -**

For Boosting, the trees are grown sequentially: each tree grown using information from previously grown trees.It fits a tree using the current residuals, rather than the outcome Y, then add the new decision tree into the fitted function to update the residuals. By fitting small tree to the residuals, we slowly improve the predictive function in area where it does not perform well. Output the boosted model,



### **5. Results**

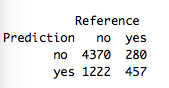
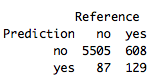
**5.1 Logistic Regression**

During data preprocessing, we noticed that there are 11.6% “yes” and 88.4% “no” in the training set. The presence of imbalanced class distribution in our data may distort the algorithms and its predicting performance, so we performed up-sampling inside of 5-fold cross-validation by specifying the subsampling parameter as “up” when using the train function.

After we got a tuned model, we tested its performance on the test set. Since we are more interested in correctly identifying people who will subscribe to term deposit, we will not be using the traditional measurement, which is the accuracy, to assess the model performance. Instead, we will be using sensitivity and AUC to evaluate the model. We got an accuracy of 76.27% , a specificity of 78.15%, a sensitivity of 62.01%, and an AUC of. We also tried another model with 5-fold cross validation, but without up-sampling. The true positive rate is only 17.5%. Therefore, by performing up-sampling, we are able to greatly increase the true positive rate.

**Confusion matrix**:

Without up-sampling With up-sampling



**Importance ranking**:



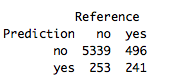
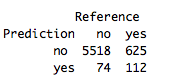
The most important factor that contributed the most in predicting whether or not a client will subscribe is that the previous campaign was successful. The coefficient of poutcome success is 2.305, meaning that if the previous outcome is successful, the odds of client subscribing to term deposit is expected to increase by 10%.

**5.2 Decision Tree**

For decision tree, we performed up-sampling for the minority class inside a 3-fold cross validation when training the model. After we got a tuned model, we tested its performance on the test set. We got an accuracy of 88.17%, a sensitivity of 32.7%, a specificity of 95.48%, and an AUC of. In terms of true positive rate, it is doing worse than the logistic regression.

**Confusion Matrix**:

With up-sampling Without up-sampling



We also tried 3-fold cross validation without up-sampling the minority class. The result is even worse. Only 15% of the clients was correctly identified as subscribers.

#### **5.3 Bagging, Random Forest and Boosting**

After using the Bagging, Random Forest and Boosting model with 5-folds cross-validation on model training with the training dataset (80%), following are the model accuracy and AUC values from predicted outcome of the testing dataset (20%) -

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | AUC | Accuracy\_upsampling | AUC\_upsampling |
| Bagging | 0.909320695 | 0.911493735 | 0.896524487 | 0.906249735 |
| Boosting | 0.908056872 | 0.911580606 | 0.839020537 | 0.91336847 |
| RandomForest | 0.909004739 | 0.922956827 | 0.890521327 | 0.928321702 |

Since AUC is calculated based on the ROC curve, which is take almost all the classifiers threshold into consideration.Therefore, the AUC can be used to compare the performance of the three models. In upsampling, it will be not necessary to look at the balanced accuracy since it will be same as the accuracy. Based on the AUC, it seems like that models with upsampling has better performance and Random Forest has the highest AUC. That is, Random Forest with upsampling method was chosen as the best model from tree-based models.

#### **6. Conclusion**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | **AUC** | Accuracy\_upsampling | **AUC\_upsampling** |
| Bagging | 0.909320695 | **0.911493735** | 0.896524487 | **0.906249735** |
| Boosting | 0.908056872 | **0.911580606** | 0.839020537 | **0.91336847** |
| RandomForest | 0.909004739 | **0.922956827** | 0.890521327 | **0.928321702** |

Among all those models we tested out, models trained using up-sampling data have better performance. Specially, resampling tree-based models performed the better than regular decision and logistics models. Out of those 3 resampling tree-based models, RandomForest is the one with the best model performance in term of AUC.

**8. Future work and limitation**

In modeling processes, we tested out 3-fold and 5-fold cross validation. It seems like 5-fold cross validation output a better model. In general, it makes sense to test out 10-fold cross validation. However, we were not able to run the 10 fold cross validation because of the limit computation power on our own laptops. In addition, there is a chance to get better model performance if we spend time working on feature selections and engineering .

**9. References**

## Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, ISLR: An Introduction to Statistical Learning with Applications in R, 2017

Jiong Chen, Yucen Han, Zhao Hu, Yicheng Lu, Mengni Sun, “Who Will Subscribe A Term Deposit?”, 2014

|  |
| --- |
|  |

## 