# FEM11152 – Seminar Data Science for Marketing Analytics: Individual Assignment On Telemarketing campaign of a Portuguese Bank

Eleftherios Tranakos - EBB5 Team 5

2024-01-07

## *Introduction*

The purpose of this report is to forecast the success of Portuguese bank's telemarketing to its customers on whether they will open a term deposit after their last contact or not. The significance of this analysis for a bank is to understand which features of the telemarketing campaign lead to deposits, which features need to change in order to have the optimal result of a deposit and lastly to predict the amount of money that they expect to get after a term of a campaign. Thus, the research question that is being generated is the following:

Is telemarketing campaign successful for a bank term deposit and which factors influence the optimal outcome, as mentioned by the probability of a customer to open a deposit after the last phone contact?

## *Data Preparation*

In this report, the data was formulated from May 2008 - November 2010 from a Portuguese Bank institution and donated on 2/13/2012 for public research. It was published also by Sérgio Moro, P. Cortez, P. Rita. in 2014 in Decision Support Systems. The dataset is consisted of 41.188 observations and 21 variables included the response variable. Also, the research was conducted via phone calls on the weekdays in Portugal.

Describing the variables we have, both numeric and categorical, they are classified as demographic (ex.age,marital,education,loan), social-economic indicators(ex.cons.price.idx,euribor3m) and lastly are related with the last contact of the current campaign. More specifically, social-economic indicators, such as ex.cons.price.idx, which describes the inflation rate or changes in consumer prices, are related with external factors that possibly influence a client to open a bank deposit based on the financial power that person has by that time. In addition, the attributes of the last contact are represented by features such as the duration of the phone call, the frequency and the day of the contact.

Heading now to how we prepared our data for the analysis, we noticed that despite our data do not have NAs values, they have some ‘’unknown" values that could be treated as NAs as the actual value of these variables should be binary (yes-no). Thus, we converted the ‘’unknown" values into NAs and we deleted the rows of it since they were not too many. So, our observations decreased from 41.188 to 38.245. However, there was a specific variable "default" which had 8.000 NAs and we treated it differently. Dropping it out, would cause to lose a lot of information and for that reason we applied logistic regression with and without it in order to compare the mse and see how much it influencs our predictions. Before applying this method, we scaled the numeric data and encoded the categorical ones with one hot encoding method. After applying the function glm(), the difference between the two mse's was small, 0,006, so we decided to delete this variable from our dataset remaining now with 20 variables.

Exploring a bit more the numeric variables and plotting the correlation matrix, we noticed that there is a high correlation between the social-economic indicators. Despite the fact that the mothods we chose for our analysis could potentially handle the correlated variables but in order to make our models more robust,

we decided to proceed with lasso penalized regression model for variable selection. After the application of lasso, the 'euribor3m' economic indicator was selected for deletion. Thus, we remained with 19 variables and 38.245 observations.

Lastly, we realized that our dataset has a significant class imbalance for our response variable leading us to apply SMOTE technique from ROSE package in order to create synthetically data and oversample the minority and undersample the majority simultaneously. We proceed to this step after we had split the data into train - test, with 80-20%, respectively. Therefore, we applied the method on the train data consisting of almost 30.000 variables.

## *Methods*

In our analysis, we applied a variety of methods either to reduce the dimensionality of our data or in order to compare the different predictions and find the most optimal model for our data.

We firstly applied, **Logistic Regression** which is a statistical technique that can be used for classification tasks. We chose to use it as it can both handle numeric and categorical variables (our data) and can predict the response categorical variable(yes-no). Its application, helped us realized that "default" variable does not change the predictions (based on mse) in a great extend and for that reason was deleted.

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p)}}$$

Secondly, we proceeded with **Lassoo** penalized regression model in order to address high multicollinearity, automatically dropping out highly correlated variables such as "euribor3m" in our case. It does this by setting some coefficients $\beta$ to exactly zero, effectively removing those variables from the model.

The penalty parameter, $\lambda$, defines the weight of the penalty on each coefficient and can be found via recurrent cross-validation, in which the training dataset is divide into n folds, in the model trained on n-1 folds and tested on the remaining one fold n times for each fold, for each *lambda*.

We take the $\lambda$ average accuracy metric (mse) for each n trials. The model is trained with parameter $\lambda = \lambda_{\min}$ to extract an optimal model, and the one with lowest mse is chosen ($\lambda_{\min}$).

$$min(\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \lambda \sum_{i=1}^{n} |\beta_i|)$$

Heading now to our main methods, we started with **Boosting** which is an ensemble machine learning technique used to improve prediction's accuracy by training the weak learners(decision trees) in order to correct the errors of the previous models. The main purpose is to work more on the instances that were misclassified. We chose to use this method as it is well-known for its robustness, it is less prone to overfit, handle complex relationships and generalize quite well to unseen data.

Boosting includes repeatedly adding models, $f_t(x)$, to minimize the loss function, $L(y, F(x))$, where $y$ is the real value and $F(x)$ is the prediction.

$$F_t(x) = F_{t-1}(x) + \alpha_t f_t(x)$$

Here, $F_{t-1}(x)$ is the ensemble model at repetition $t-1$, $f_t(x)$ is the new model added at iteration $t$, and $\alpha_t$ is the learning rate. The procedure is repeated in order to improve $F(x)$, the final predictive model.

Continuing with **Classification and Regression Trees(CART)** in our methods, it is a supervised machine learning models used for classification and regression. The algorithm constructs binary trees by dividing the train data into subsets based on an attribute value test. These steps are repeated continuously on each built subset. When the algorithm understand that no further splits can originate more homogeneous subsets, the

repetition is finished. In our case we made use of this method for classification purposes and it tends to maximize the purity of each node, using measures like Gini impurity:

$$Gini(t) = 1 - \sum_{i=1}^{c} p_i^2$$

Gini impurity quantifies the likelihood of misclassifying a random instance from a labeled subset based on the majority class. Lower Gini impurity indicates greater purity of the subset.

We chose to use it as CART can be interpreted relatively easy, is versatile and can handle mixed data types as we have in our case.

Completing the methods part, the last method we used is **Random Forest** which is an ensemble machine learning model used for classification and regression tasks. It is a specific type of bagging( Aggregate Boosting) that decreases overfitting by introducing randomness as we construct the tree.

It choose randomly subsets of variables at each split and combine the predictions of many decision trees in order to improve the accuracy and decrease the variance.

$$RF(x) = \text{mode}\{T_1(x), T_2(x), \ldots, T_B(x)\}$$

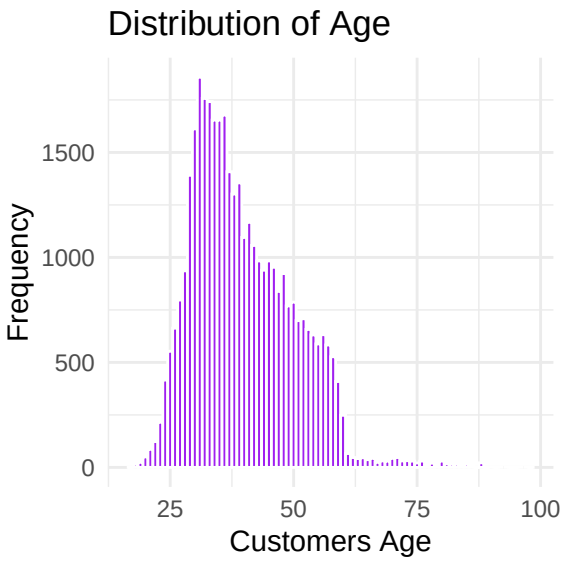
We decided to use this method as it is robust against overfitting, can handle large datasets with high dimensionality(as in our case) and often leads to high accuracy.

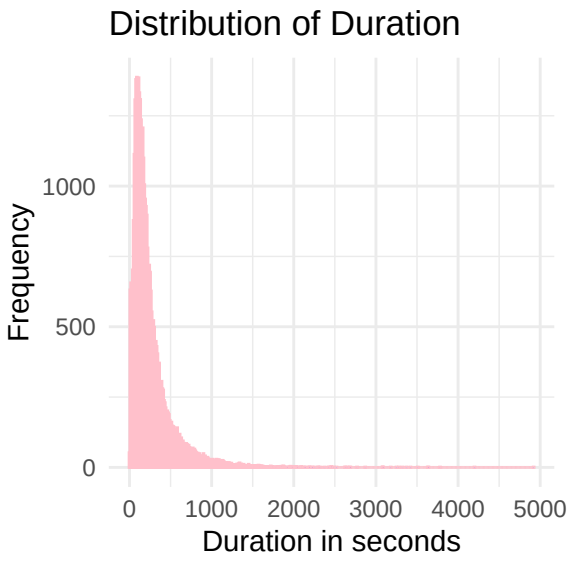
## *Analysis*

Heading now to the analysis part, before we explain the results from the methods and the best one, we will dive a bit into the **exploratory analysis** of our data.

Taking a look on the numeric variables first and more precisely on the plot below, we can extract information about our client's age which is between 17 and 98 years old with the mean to be 40 years old. The higher concentration of the client's age is gathered between 30 and 40 years old.

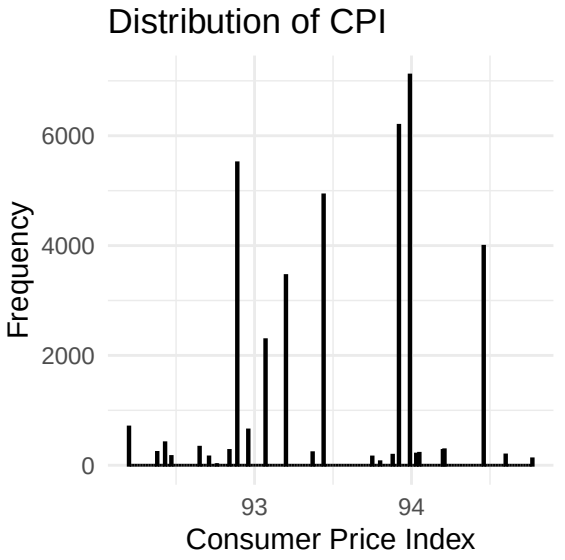
On the next plot, we can explore the duration of the last call in seconds that the bank's employers had with their customers. As we can see the range is from 0 to 4918 seconds and an average call last around 258 seconds. Converting into minutes, a call usually last maximum 4.3 minutes with a regular call to be around 1.2 minutes. This can be explained from the feature's display which its distribution is right-skewed distribution, indicating that most calls are relatively short.

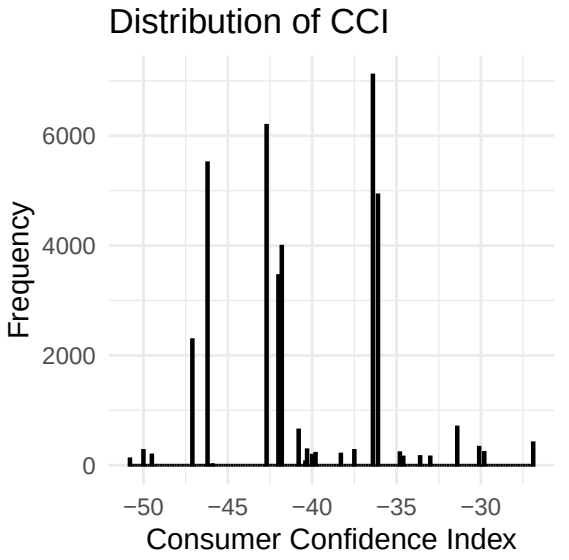



Heading further, it is time to have a short look on some social-economic indicators by the time that the research was conducted. Starting with "cons.price.inx" which is a consumer price index(CPI), measure of inflation and economic health. Changes in CPI influence people's financial power and decisions including term deposits. When CPI is high the real value of money fall which may encourage customers to open a term deposit, and proceed to investment moves and vice verca. From the bank's aspects, they often adjust their interest rates based on CPI. In our case, the range of CPI is from 92.201 to 94.767 with an average to be around 93.576. The distribution of this index represents many peaks, possibly indicating specific periods where the index value was more popular.

Respectively, the next graph which represents the "cons.conf.idx", consumer confidence index(CCI), is an index which measures the consumer's confidence about their financial situation and the state's economy. A higher CCI reveals, that consumers are optimist and vice verca. This situation, may lead consumers to consume more but it also may lead to invest/save money for future use, as in bank deposits. In this case, the range of this index is from -50.8 to -26.9 with an average to be around -40.503. The negative numbers is a result of how the index was constructed and scaled.

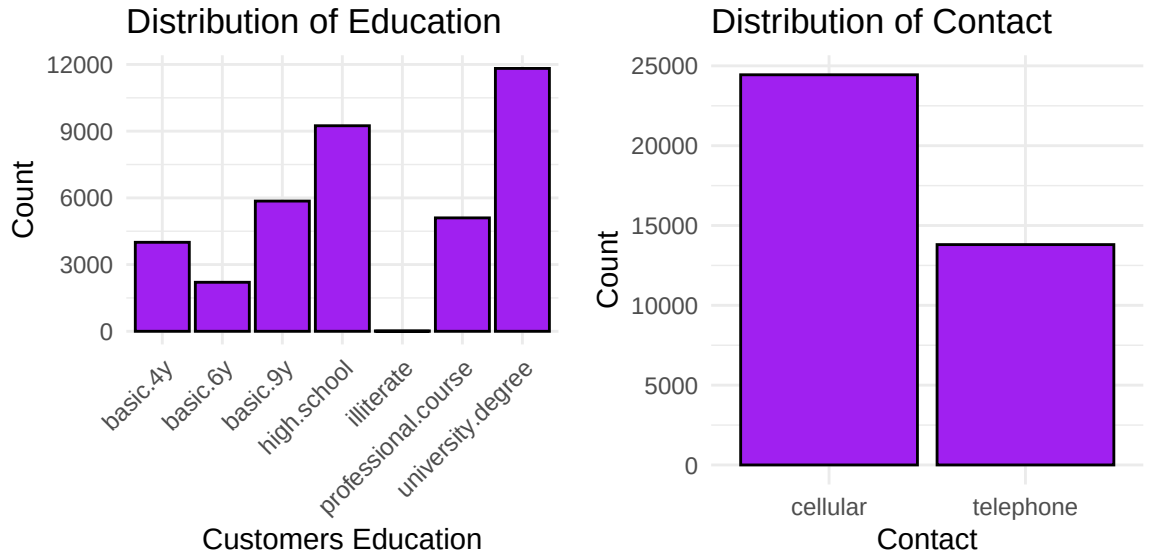



Passing now to the categorical variables of our dataset, on the below plot we can see the educational level of

the customers were called. The dominate level was university degree with around 12.000 people belonging to this category almost one third of the whole population followed by high school level consisted of around 10.000 customers. It is worth mentioning that a minor percentage and only 18 people belong to the illiterate category.

Continuing with the way that the clients were contacted, the next plot indicates that the majority of the people was call on their cellular phone, a number almost doubled with the ones that contacted on their telephone equals to almost 15.000 people.




In addition, on the plot below we can study the distribution across the days of the week the bank's customers were called. It is slightly obvious that the dominant day is Thursday, followed by Monday while Friday is the least common day among all.

Lastly, taking a look on the last exploratory plot to our target variable,below, we notice that the majority of the clients do not open a deposit term to the bank, indicating at the same time the low success for the bank's telemarketing campaigns. Undoubtless, there is a class imbalance on our response variable and this is why we proceed to balance the class applying techniques from ROSE package creating synthetically data for the minority class.

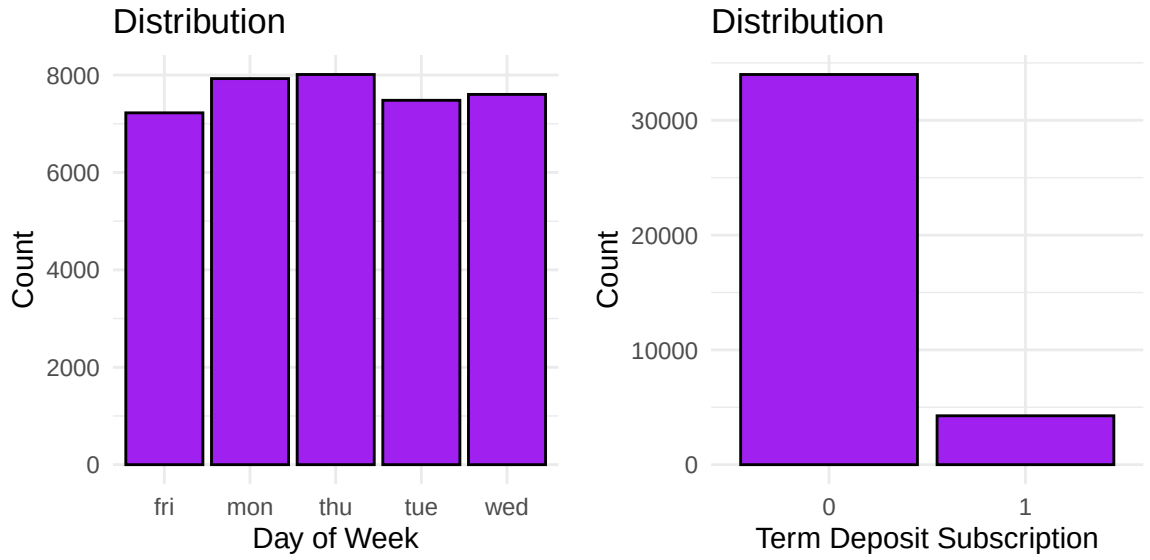


Table 1: Confusion Matrix for CART Model

| | 0 | 1 |
|---|---|---|
| 0 | 5706 | 61 |
| 1 | 1119 | 763 |

Table 2: Model Performance Metrics

| Model | Sensitivity | Specificity | Kappa | Balanced_Accuracy |
|---|---|---|---|---|
| LightGBM | 0.877 | 0.879 | 0.545 | 0.878 |
| CART | 0.926 | 0.836 | 0.487 | 0.881 |
| Random Forest | 0.472 | 0.985 | 0.553 | 0.728 |

Introducing our report to the main analysis, we first applied boosting and more precisely *Light Gradient Boosting*. Since we have a large dataset with 38.000 observations, we applied this method for fast training speed and higher accuracy. For classification purposes, we set objective parameter equals to "binary" and the metric equals to "binary_error" which is used for the model's evaluation. The number of leaves that was chosen is equals to 20 as we wanted a large number in order to increase the accuracy but not that large to avoid overfitting. The last parameter that we tuned is learning_rate which influences how the model learns during training. A small number of it requires more iterations but give us better results and a more general model. In our case the number was chosen is 0.05. Lastly, our model has been through one thousands iterations.

The next method we used is *CART* (Classification and Regression Trees) which is the most optimal one as we will see above. The tree is constructed by continuously partitioning the data into subsets using rpart function, which makes it easier to interpret. For model validation, we used 10 fold cross validations with cp from 0.001 to 0.1. CP stands for model complexity, a small number creates larger trees while a large number leads to smaller trees. Based on the cross-validation results we found out the most optimal model which balance tree complexity and accuracy. Our model has been through 1000 iterations to extract the results.

The last model that we used was *Random Forest* which did not perform as good as the rest above. On the training, the number of trees was set up to 500, as we increase the number of trees, the performance and the robustness of our model is being increased simultaneously at the cost of computational time. Since, we had around 38.000 observations we decided to choose this number of iterations in order to save time. The parameter "nodesize" was given the number 5 which equilibrate the accuracy and model's complexity. Lastly, the replace parameter turned into TRUE meaning that it it allows sampling of the training dataset with replacement to build each tree, bootstrapping.

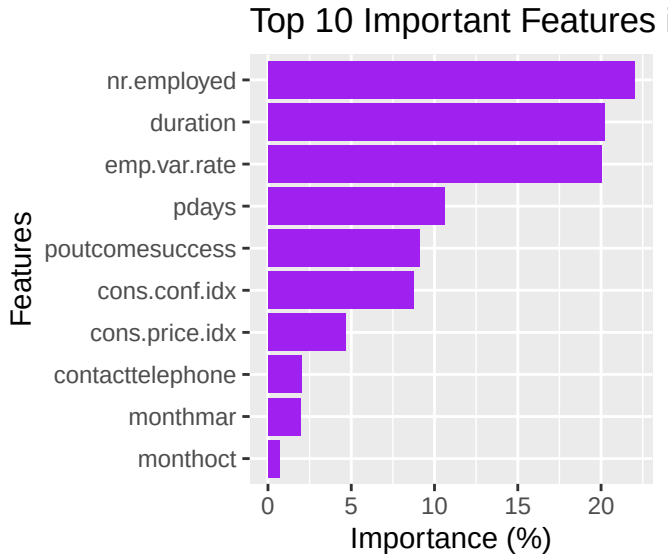
For models' evaluation we used **confusion matrix** method which allows us to understand the accuracy of the model, showing the number of correct and incorrect classifications compared to the actual values. On the below table, we can see the confusion matrix of the most optimal model, CART,which had the highest accuracy. Our model correctly forecast 5706 (True Negative) cases of negative class (0) and incorrectly predicted 1119 cases (False Negative) as the negative class when they were truly positive. On the other hand, the the CART model correctly predicted 763 cases (True Positive) of the positive class(1) but incorrectly forecast 61 cases(False Positive) as the positive class when they were negative in reality.

On the below table, we can inspect the model's comparison based on several metrics. In order to determine which model is the best we used **balanced accuracy** (average of specificity and sensitivity) as the performance metric particularly because our dataset was imbalanced and this metric is more accurate for imbalanced datasets in comparison with the traditional accuracy that might be misleading. The balanced accuracy of CART model is slightly better than LightBoosting having 0.881 in comparison with the later 0.878. In addition CART's sensitivity of 0.926 means it successfully identifies the 92.6% of the actual positive cases. Thus, CART's high sensitivity and balanced accuracy, together with its simplicity and interpretability, are the reasons to consider it as the best model.

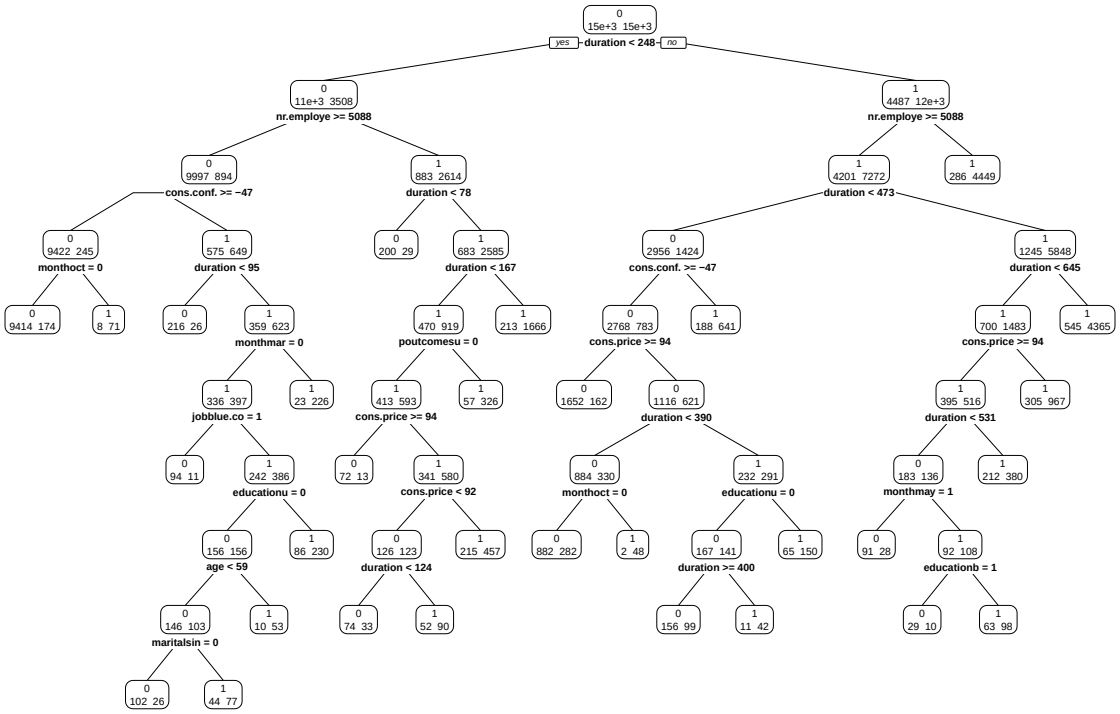
Having decided the optimal model of our analysis, it is time now to see the most important features that influenced our target variable. On the below plot we can see the 10 more influential variables. The nr.employed, duration, and emp.var.rate are the most dominant variables in contrast to the rest. More specifically, the number of employers represents the total number of employers in the country and could be interpreted as the higher is this number the higher the consumer's confidence which potentially could lead to more bank's deposits. Regarding the duration, the longer a call lasts, the higher the possibilities of a term deposit, as the consumer might show high interest. As for, emp.var.rate it is a indicator that provide insights about job market dynamics and economic activity. Indicating a growing economy, can positively influence consumers to subscribe a bank deposit.

Furthermore, the "pdays" represents the days that passed since the last contact with the customer indicating that a frequent contact again with them might lead to a bank deposit. Also, the poutcomesuccess reveal that if the previous marketing campaign was successful, the client maybe more open to give a bank deposit. In addition, as the cons.conf.idx and cons.price.idx was discussed on the beginning of our analysis, it worth mentioning that contacttelephone is a significant variable which influence client's decisions. As we saw on the beginning more people were contacted via cellular a thing that we should take it into account. Lastly, the significance of monthmar and monthoct feature might reveal a season factor or financial cycles that take place on that time.

As a conclusion of the variable's importance, the social-economic indicators and the direct engagement such as duration and recent contact(pdays) play a main role on influencing customer's decisions to place a bank deposit.




Lastly, the below plot represents a decision tree revealing the decision-making process to each node represented by a decision based on the value of a specific variable every time. As we can see, there are some patterns of specific features that are used often to make the splits. More specifically, features such as duration, nr.employed, cons.conf.idx and cons.price.idx are the ones that influence the model's predictions. These 4 features, as we saw above on the variable's importance plot, belong to the most important ones which influence the target variable. Thus, the frequency of certain features indicate the importance of it on the final outcome




## *Conclusion*

In conclusion, answering the question that was formulated on the introduction, the bank telemarketing campaign was not that successful as the majority of the people, based on the confusion matrix, did not open a term deposit after their last contact with the bank. Nevertheless, we found out which variables influence the most a customer to open a deposit which are nr.employed, duration, and emp.var.rate. , pdays, poutcomesuccess etc. Some of these variables are based on external factors that a bank cannot predict or intervene such as the social-economic indicators. However, for a future marketing campaign, marketing leaders could possibly build a strategy based on the rest of the most important features that they can intervene.

As about the limitations of our analysis, it has been more than 10 years since the conduction of this research and nowadays most marketing campaigns take place online as they influence the customers more. Thus, a campaign nowadays about bank deposit could be possible more successful. Furthermore, the data was

gathered from a Portuguese's Bank, which is not representative for the rest Banks in Portugal. One of the most important limitation is that, the research was created in a period where there was an ongoing financial global crisis and not a lot of people would open a bank deposit. Also, limitations relating to our analysis, the class of our target variable was imbalanced and we synthetically created new data with SMOTE that are not representative. In addition, If we were not worry about computational time, we could have applied RandomForest 1000 times instead of 500.

Last but not least, as a future developments, we could propose to apply techniques like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to dive even deeper on the interpretation of our models. In addition, we could have applied neural network models to possible increase prediction accuracy. Lastly, conducting segmentation analysis will help us to understand the age groups and aim specific marketing strategies on specific age groups that are more possible to open a bank deposit.

## *References*


1. Moro,S., Rita,P., and Cortez,P.. (2012). Bank Marketing. UCI Machine Learning Repository. https://doi.org/10.24432/C5K306.
2. Casu, B., Girardone, C., & Molyneux, P. (2006). Introduction to Banking. Prentice Hall
3. Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Journal of Artificial Intelligence Research.