**Smart banking customer targeting using ML Ensemble for improved business operational efficiency through reduced cost per call (CPC)**

Sujeeth

Department of Psychology, The George Washington University

PSYC 3170: Clinical Psychology

Dr. Tia M. Benedetto

April 19, 2023

Contents

[1. Project Background & Overview 4](#_Toc133316456)

[2. Problem Statement 4](#_Toc133316457)

[3. Literature Review 5](#_Toc133316458)

[4. Machine Learning Techniques used for Current Work 5](#_Toc133316459)

[Decision Tree 6](#_Toc133316460)

[Random Forest 6](#_Toc133316461)

[XGboost 6](#_Toc133316462)

[5. Methodology 7](#_Toc133316463)

[5.1 Overview 7](#_Toc133316464)

[5.2 Data Source 8](#_Toc133316465)

[Description 8](#_Toc133316466)

[Collection Process 9](#_Toc133316467)

[5.3 Data Cleaning and Transformation 9](#_Toc133316468)

[Missing Value Treatment 9](#_Toc133316469)

[Outlier Treatment 9](#_Toc133316470)

[Outlier Treatment for Age 9](#_Toc133316471)

[Outlier Treatment for Call Duration 10](#_Toc133316472)

[Outlier Treatment for Balance 10](#_Toc133316473)

[Outlier Treatment for Campaigns 11](#_Toc133316474)

[Train and Test Data Preparation 11](#_Toc133316475)

[6. Feature Selection and Engineering 11](#_Toc133316476)

[Addressing Multicollinearity 11](#_Toc133316477)

[Treating Categorical Features 12](#_Toc133316478)

[7. Model Selection and Evaluation 12](#_Toc133316479)

[Decision Tree – Parameter Tuning 12](#_Toc133316480)

[Random Forest – Parameter Tuning 13](#_Toc133316481)

[XGboost Parameter Tuning 14](#_Toc133316482)

[8. Model Results and Discussion 16](#_Toc133316483)

[Descriptive Analysis of the Data - Summary 16](#_Toc133316484)

[Decision Tree - Loss Metrics & Model Final Model Results Discussion 17](#_Toc133316485)

[Random Forest Tree - Loss Metrics & Model Final Model Results Discussion 18](#_Toc133316486)

[XGboost Tree - Loss Metrics & Model Final Model Results Discussion 19](#_Toc133316487)

[9. Discussion on Key Findings 20](#_Toc133316488)

[Ensemble of Decision Tree, Random Forest and XGboost 20](#_Toc133316489)

[Bias Validation 20](#_Toc133316490)

[Cost of ML Model Development 21](#_Toc133316491)

[Model Results Summary Discussion 21](#_Toc133316492)

[10. Conclusion 23](#_Toc133316493)

[11. Reference 24](#_Toc133316494)

# 1. Project Background & Overview

Traditionally businesses reach out to prospect customer for encashing potential opportunities (can be in terms of cross sell or upsell). Though this process of targeting customer for potential cross sell or upsell in observed in many industries, it is very wide and frequent seen in banking industry were a customer holding a with a bank will be targeted for cross selling opportunities like loans, fixed deposits and term deposits etc., Banks traditionally uses various channels to reach their customer and one of the major such channel is Telemarketing. Though telemarketing seems like a very easy way to reach a customer at the same time it is very costly. As per industry standard typical Cost per Call (CPC) is around $2.7 to $5.6 1 (it might difference from business to business), based this statistic we can estimate the possible impact of targeting a wrong prospect and importance of accurate targeting strategy. In the current project we plan to address this problem of high operating cost due to inaccuracy customer targeting using machine learning.

# 2. Problem Statement

banking industry, traditionally use various channels to reach out to their customers and cross-sell or upsell products like loans, fixed deposits, and term deposits. One of the major channels used by banks is telemarketing, which can be costly, with a typical cost per call ranging from $2.7 to $5.6. Accurate customer targeting strategy is very vital for avoiding prospect and which otherwise will result in high operating costs. We propose the use of machine learning to address this problem and reduce costs associated with inaccurate customer targeting.

In the current project we plan to address this problem of high operating cost due to inaccuracy customer targeting using machine learning. As part of the analysis, we will be using banking telemarketing call data for predicting the propensity of a customer opting for cross selling, which can later be used by banking businesses for making better call plan as well customer target list.

Keywords: Business optimization, Machine Learning for CPC reduction [2,3], Call Center Optimization

# 3. Literature Review

“Machine Learning Approaches for Marketing Campaign in Portuguese Banks” paper by Jennifer Alexandra; Kristina Pestaria Sinaga (2021) 4 show the implementation of Decision Tree, Naïve Bayes, and Random Forest as well as clustering techniques for predicting. Afore mentioned paper evaluated each of these models in isolation. Our approach stands out in comparison with other in the through the usage of model ensemble [5, 6, 7] instead of building and evaluating individual models.

# 4. Machine Learning Techniques used for Current Work

The following stack ML model will be test for the current business problem and an ensemble of individual model recommendation will be used for final prediction

* Decision Tree
* Random Forest
* XGBoost

## Decision Tree [8,9]

A decision tree is a machine learning algorithm that uses a tree-like model to make decisions based on input parameters. It starts with a single node and branches out into subsequent decisions and outcomes. The goal is to reach a final outcome that maximizes accuracy or profit. Decision trees are used for classification and regression problems to identify critical factors and provide insights. However, they can be complex and prone to overfitting, which can be addressed through pruning and ensemble methods.

## Random Forest [10, 11]

Random forest is an ensemble machine learning algorithm that combines multiple decision trees to enhance accuracy and reduce overfitting. It randomly selects subsets of the training data and features for each tree, then aggregates their results to make a final prediction. It can handle both categorical and continuous data, and is frequently used in classification and regression problems. Random forest can handle missing values and noisy data, making it a popular choice for real-world applications. However, tuning hyperparameters such as the number of trees and maximum depth is essential, and it can be computationally expensive. Overall, random forest is a versatile algorithm that can generate highly precise predictions while avoiding overfitting.

## XGboost [12, 13]

XGBoost is a powerful and scalable open-source gradient boosting machine learning library designed to improve accuracy and speed of traditional gradient boosting algorithms. It can handle both regression and classification problems, and uses a combination of multiple decision trees and gradient boosting techniques to generate highly accurate models. XGBoost can handle missing values and provides support for parallel processing, and can be used with various programming languages. It offers several hyperparameters for tuning the model, but may be computationally expensive and requires careful tuning to avoid overfitting.

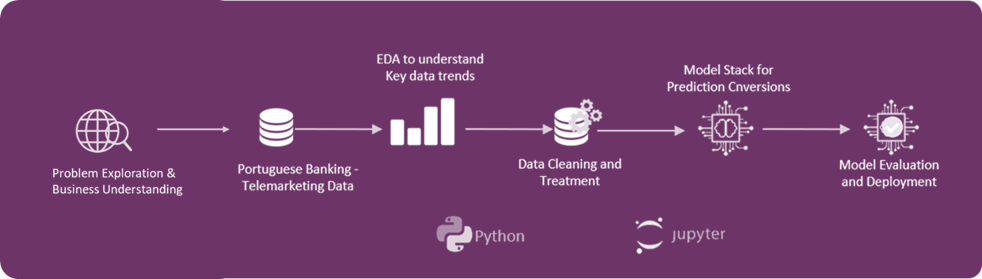
# 5. Methodology

## 5.1 Overview

As the objective of the current solution is to predict the customer propensity, we need to build a classification model for the current problem. Also, for this kind of problem, businesses would be keen to look at the drivers of recommendation as it helps them in planning better and deep learning model due to their black box nature might be not ideal for implementation. Hence, we will be using traditional machine learning models as they offer better visibility into the ‘why?’ part of recommendations [Refer to Figure 1 below]. The following stack ML model will be test for the current business problem and an ensemble of individual model recommendation will be used for final prediction

* Logistic Regression (baseline model)
* Random Forest
* XGBoost

Based on the initial look at the data we found the data suffers from class imbalance problem hence appropriate class balancing technique like over or under sampling will be using for addressing this problem. Since the problem we are dealing with is classification problem MAPE would be used as primary measure for model performance (on test, validation and train datasets).



**Figure 1:** *Model Architecture*

## 5.2 Data Source

Data for current project has been procured from UCI Machine learning repository. Following is the description of repository (as per the website) “*The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. The archive was created as an ftp archive in 1987 by David Aha and fellow graduate students at UC Irvine. Since that time, it has been widely used by students, educators, and researchers all over the world as a primary source of machine learning datasets.*”

### Description

We are using “Bank Marketing Dataset” hosted in UCI repository for the current analysis. Data actually belongs to a Portuguese bank where existing bank customers are targeted for term deposit subscription over phone calls. The data provided has information related to customers and past behavior when targeted with marketing campaign.

### Collection Process

Data is directly downloaded from UCI website. As required information is available as single file there is no need for data merges. Data is available as CSV file which can be directly loaded into python for model development.

## 5.3 Data Cleaning and Transformation

### Missing Value Treatment

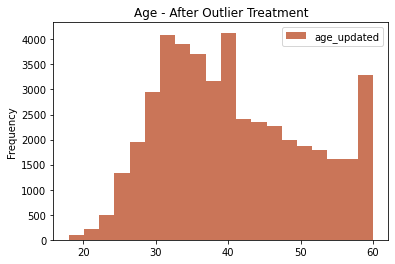
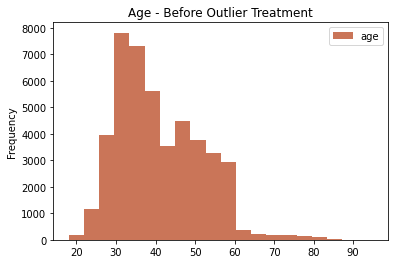
No missing value treatment performed on the data as procured data doesn’t contain any missing values

### Outlier Treatment

Of all features analyzed Age, Call duration, Balance and Campaign had relatively high degree of outliers and have been treated. We have applied very tight threshold for outlier treatment and detection as outlier can be key for prediction of subscription due to low class ratio. Following is the approach followed and distribution before and after outlier treatment

### Outlier Treatment for Age

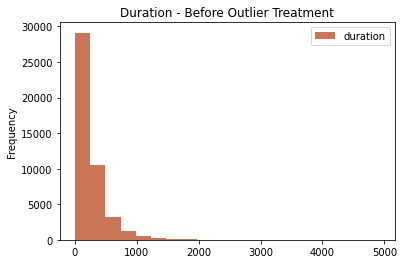
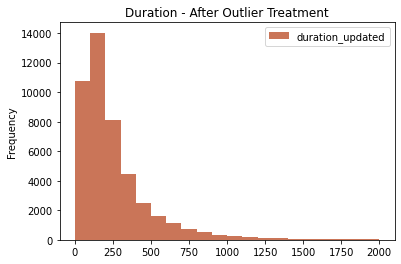
Based on the visual observation of age, we see that there are very few customers with age > 60 and it is really rare for someone to opt for term plan with age greater than 60. Hence, we have capped age to 60 i.e., if age is greater than 60, we cap it to 60. Refer to Chart 9 below



**Figure 2:** *Distribution of age before and after capping*

### Outlier Treatment for Call Duration

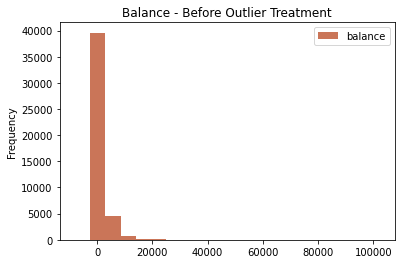
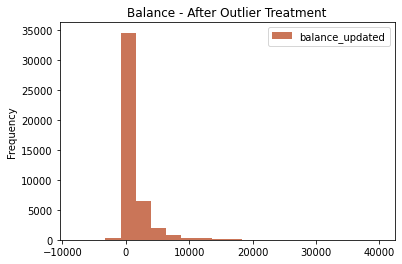
We observe that the call duration is ranging from 0 to 5000 sec in the dataset. A call duration of 2000 sec (33Min) is very high for a customer to interact with agents over phone. Hence, we treated anything above 2000 sec as outlier and capped it. Refer to Chart 10

**Figure 3:** *Distribution of call duration before and after capping*

### Outlier Treatment for Balance

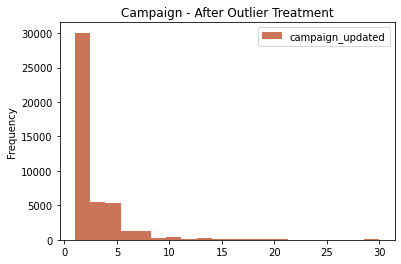
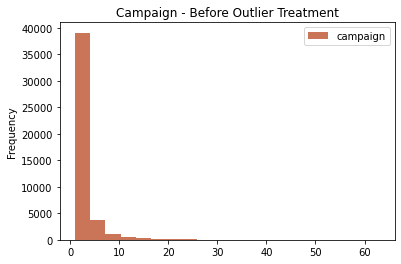
There are very few data point with balance greater than 40K. Hence, records with greater than 40K balance are capped at 40K. Refer to Chart 11 below for distribution before and after

**Figure 4**: *Distribution of balance before and after capping*

### Outlier Treatment for Campaigns

Number of campaigns is ranging from 0 to 60 in the dataset, which means that a customer reached anywhere between 0 to 60 times is the current campaign. Based on distribution of data we few data point have campaign value >30. Hence, all data point with campaign value > 30 are capped at 30. Refer to Chart 12 below for distribution before and after



**Figure 5**: *Campaign value before and after capping*

### Train and Test Data Preparation

Data has been split into train and test in 4:1 ration through random sampling. Model will be trained on train dataset and performance will be evaluated on test dataset.

# 6. Feature Selection and Engineering

## Addressing Multicollinearity

Of all the correlation observed Poutcomes is very highly correlated with Pdays both statistically and in business sense and hence Pdays is been excluded from the modeling. Though there are other correlated variables we are not excluding then as we are using tree-based models for model building and are robust to outliers and has no assumptions on multicollinearity

## Treating Categorical Features

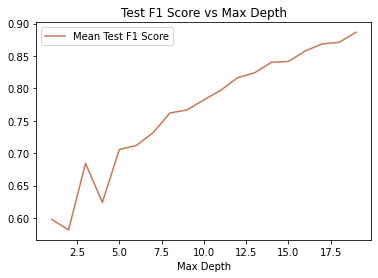
Data has lot categorical features like Education, Job etc., which cannot be used as is. Categorical variables are dummy coded for making data model ready

# 7. Model Selection and Evaluation

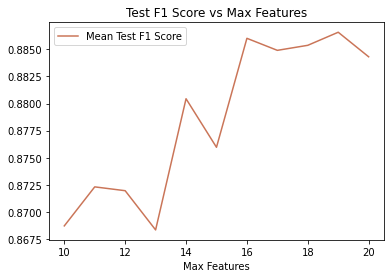
## Decision Tree – Parameter Tuning

As per the proposal decision tree model was developed as baseline model for current project. We have implemented and tuned decision trees using Grid Search and performance has been evaluated on Test and Train datasets. Following are the initial model results.

Decision tree model has been majorly tuned on max\_depth and max\_features parameter. Gridsearch with cross validation is used for parameter tuning. Max depth parameter is search in the range of 1 to 20 while max\_features is search in the range of 10 to 20. Best model was found at max depth = 19 and max\_features = 19. [Refer to Chart 13 and 14 below for understanding model performance with change in max\_feature and max\_depth]



**Figure 6**: *Model F1 Score vs Max Depth*

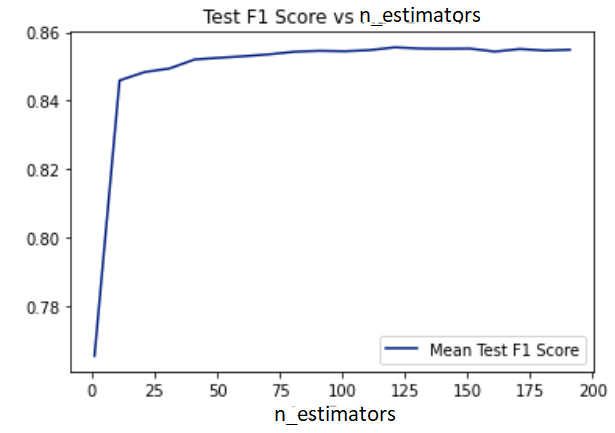
**

**Figure 7**: *Model F1 Score vs Max Features*

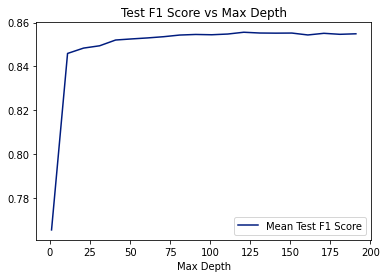
## Random Forest – Parameter Tuning

As per the initial proposal Random Forest model was developed as second model in ensemble of 3 model stack for current project. We have implemented and tuned random forest using Grid Search and performance has been evaluated on Test and Train datasets. Following are the initial model results.

Random Forest model has been majorly tuned on max\_depth and n\_estimators parameters. Gridsearch with cross validation is used for parameter tuning. Max depth parameter is search in the range of 10 to 20 while n\_ estimators is search in the range of 1 to 200. Best model was found at max depth = 15 and n\_ estimators = 101. [Refer to Chart 16 and 17 below for understanding model performance with change in n\_estimators and max\_depth]



**Figure 8**: *Model F1 Score vs Number of Estimators*

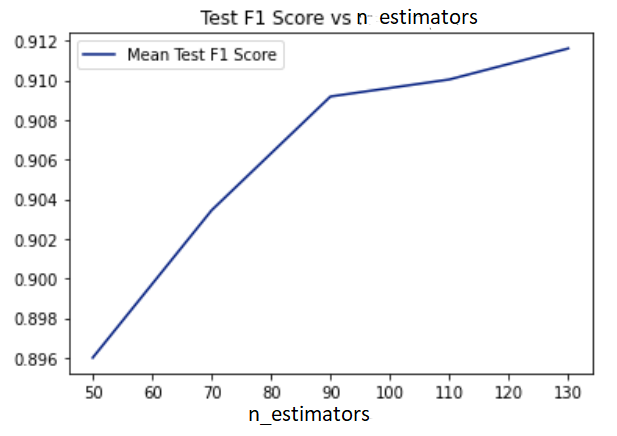


**Figure 9**: *Model F1 Score vs Max Depth*

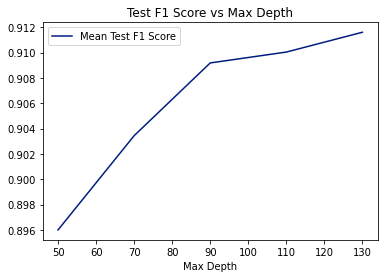
## XGboost Parameter Tuning

XGboost model was developed as third model in ensemble of 3 model stack for current project. We have implemented and tuned Xgboost using Grid Search and performance has been evaluated on Test and Train datasets. Following are the initial model results.

XGboost model has been majorly tuned on max\_depth and n\_estimators parameters. Gridsearch with cross validation is used for parameter tuning. Max depth parameter is search in the range of 10 to 20 while n\_ estimators is search in the range of 50 to 150. Best model was found at max depth = 12 and n\_ estimators = 130. [Refer to Chart 19 and 20 below for understanding model performance with change in n\_estimators and max\_depth]



**Figure 10**: *Model F1 Score vs Number of Estimators*

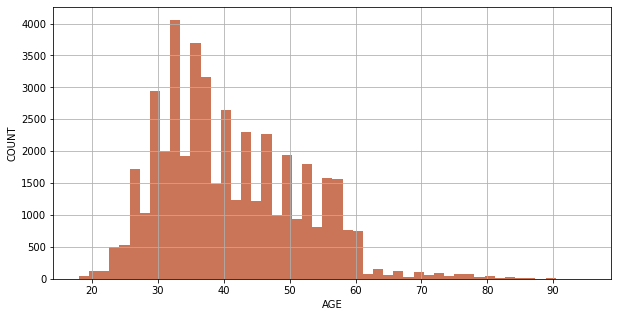


**Figure 11**: *Model F1 Score vs Max Depth*

# 8. Model Results and Discussion

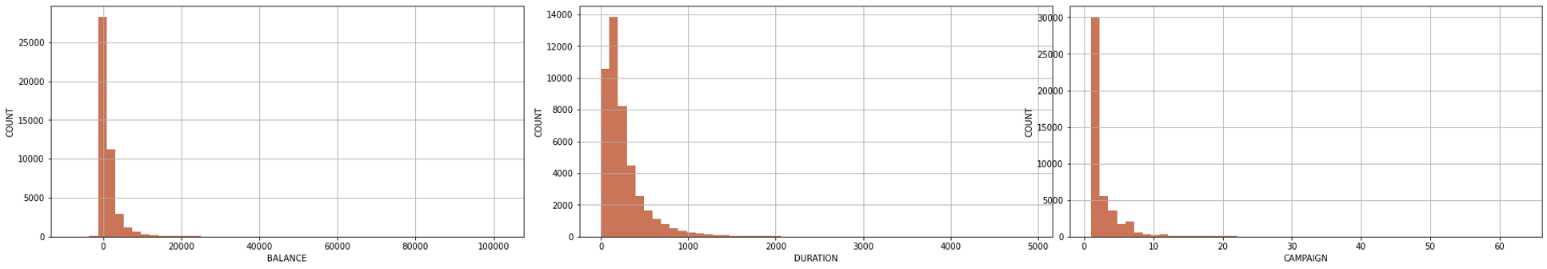
## Descriptive Analysis of the Data - Summary

1. Data has no missing values hence no need for missing value treatment
2. Certain outliers have been observed for age variable which can be treated during preprocessing stage



**Figure 12**: *Histogram of Customer Age*

1. Majority for the numeric variables are showing skewed distribution, it is suggested to avoid methods like linear regression as data is violating modeling assumptions

****

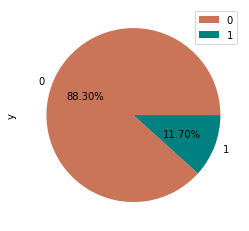
**Figure 13**: *Right Skewed Distribution of Balance, Duration & Campaign*

1. High Balance, High Duration, customer targeted in previous campaign, customer last targeted, customers without house, customer without existing loans and customers without any defaults are characters showing high tendency to opt for term deposits – (Observation made from initial EDA). Hence these can be import variables for model development
2. poutcome – pdays, Education secondary - Education tertiary and Married\_single - marital\_married are very highly correlated with each other hence they to be appropriately treated while modeling
3. poutcome\_unknown, month\_may, day and marital\_married feature have > 5 VIF [14, 15] suggesting a strong presence of multicollinearity



**Figure 14**: *Variables with High VIF*

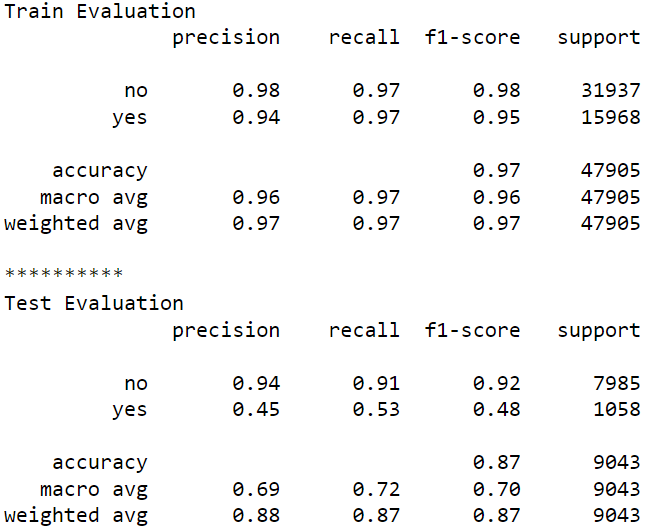
1. Data shows the presence of slight class imbalance issue which has to be addressed using appropriate class balancing techniques like over sampling



**Figure 15**: *Distribution customers subscribing for term deposit*

## Decision Tree - Loss Metrics & Model Final Model Results Discussion

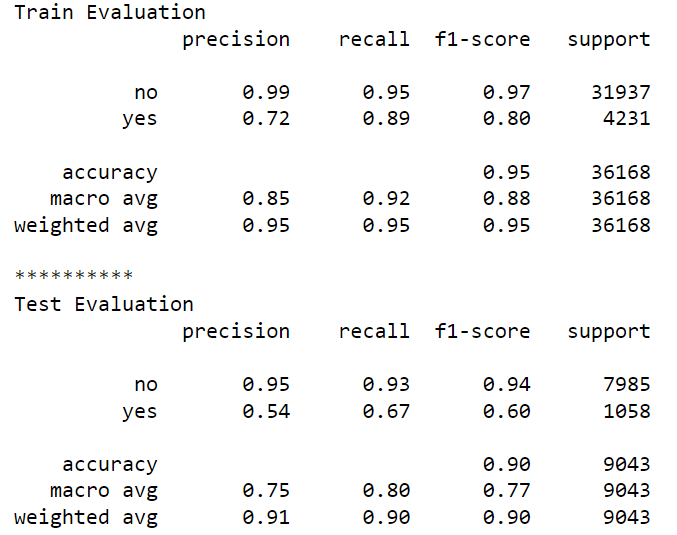
* Decision tree model has exhibited a good train and test accuracy of 97% and 87% respectively on support ratio of 50% (post class balancing)
* Though model is exhibiting a good accuracy of more than 90% but due to presence of class imbalance it is not good enough (especially in test)
  + When we look at recall score, 97 percent of the term plan subscribers are being identified during testing while only 53% are identified during testing. Which suggest that model is biased towards non-subscribers due to data size limitations [Refer to Table 1 below]



**Table 1**: *Decision Tree – Train and Test Results*

## Random Forest Tree - Loss Metrics & Model Final Model Results Discussion

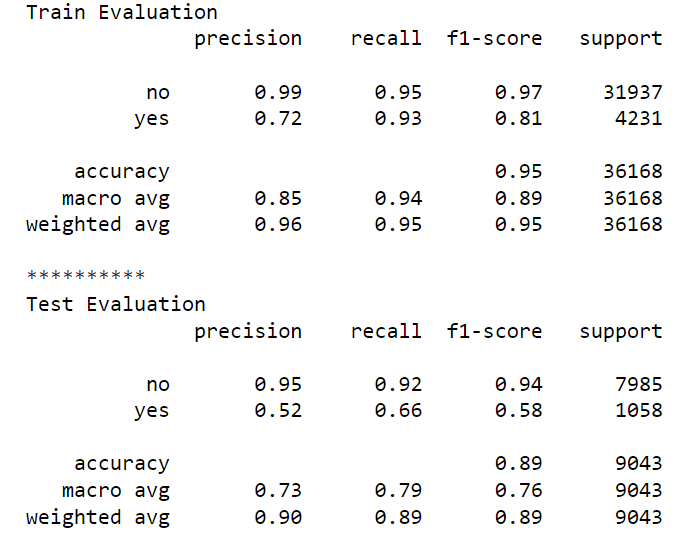
* Random Forest model has exhibited a good train and test accuracy of 95% and 90% respectively on support ratio of 50% (post class balancing)
* Though model is exhibiting a good accuracy of more than 90% (in test) but exhibited lower precision and recall
  + It can be observed from recall scores that 93 percent of the term plan subscribers are being identified during testing while only 67% are identified during testing. [Refer to Table 2 below]
* Random forest model shows better performance than that of Decision tree in identifying possible customers purchasing term plan
  + Decision tree showed an recall score of 53% while random forest exhibited 67% (an increment of 14%)



**Table 2**: *Random Forest – Train and Test Results*

## XGboost Tree - Loss Metrics & Model Final Model Results Discussion

* XGboost model has exhibited a good train and test accuracy of 95% and 89% respectively on support ratio of 50% (post class balancing)
* Though model is exhibiting a good accuracy of more than 95% (in test) but exhibited lower precision and recall
  + It can be observed from recall scores that, 92 percent of the term plan subscribers are being identified during testing while only 66% are identified during testing. [Refer to Table 3 below]
* XGboost model showed similar performance to that of Random Forest in identifying



**Table 3**: *XGboost – Train and Test Results*

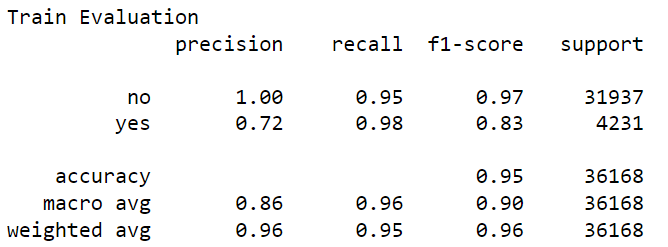
# 9. Discussion on Key Findings

## Ensemble of Decision Tree, Random Forest and XGboost

Initial idea of the proposal was to use the ensemble of all the three models but decision tree showed very different performance than that of random forest hence ensemble of Random Forest and XGboost is used as final model.

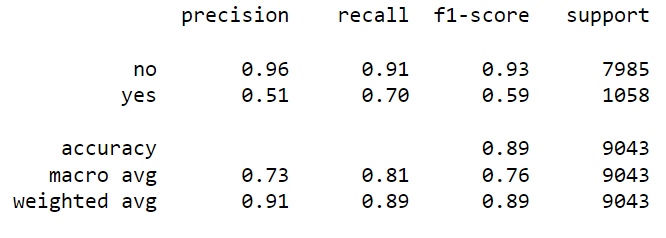
### Bias Validation

Train ensemble model showed good accuracy similar to that of Random Forest and XGboost. Train accuracy of ensemble model is 92%. [Refer to Table 4 below]



**Table 4**: *Ensemble – Train Results*

Test ensemble model showed better performance that of Random Forest and XGboost. Train accuracy of ensemble model is 89% and positive class recall of 70%. [Refer to Table 5 below]



**Table 5**: *Ensembl – Test Results*

No significant model bias has been observed during modeling as train and test results are more or less similar.

### Cost of ML Model Development

Model Train Time: XGboost – 2Sec; Random Forest – 1Sec

Model Prediction Time: Negligible

### Model Results Summary Discussion

Final model evaluation on test data showed an recall of 70% which is better than random forest with 67% and XGboost 66%. Which basically means our model could successfully identify 70% of total customers who would opt for term plan. Also, this 70% recall rate is observed at a precision of 51% which means for every one positive customer identified model is falsely identifying two customers who in actual won’t opt for term plan.

# 10. Conclusion

“Machine Learning Approaches for Marketing Campaign in Portuguese Banks” paper by Jennifer Alexandra; Kristina Pestaria Sinaga (2021) show the implementation of Decision Tree, Naïve Bayes, and Random Forest as well as clustering techniques for predicting. Afore mentioned paper evaluated each of these models in isolation our approach stands out in comparison with other in the through the usage of ensemble.

Ensemble model performed well, achieving a 92% train accuracy, and showed better performance than the individual Random Forest and XGboost models, with a 70% recall rate on identifying potential term plan customers. No significant model bias observed, and the model had a negligible prediction time. Precision of the model is a bit low, with the model falsely identifying two customers for every one positive customer identified as a potential term plan customer.

# 11. Reference

[1] "Cost Per Call (CPC) Definition: What is CPC in Customer Support?" LiveAgent, 2021, https://www.liveagent.com/customer-support-glossary/cost-per-call/.

[2] Li, M., Yang, W., & Zhou, S. (2017). A Machine Learning Approach to Cost-per-Click Optimization in Sponsored Search. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 687-696). ACM.

[3] Zhang, J., Huang, W., Zhao, Y., & Wang, Y. (2019). Cost-per-click prediction in online advertising with gradient boosting decision tree. Journal of Ambient Intelligence and Humanized Computing, 10(1), 159-169.

[4] Jennifer Alexandra and Kristina Pestaria Sinaga. (2021). Machine Learning Approaches for Marketing Campaign in Portuguese Banks. International Journal of Advanced Science and Technology, 30(9), 7749-7758.

[5] Brownlee, J. (2020). Ensemble machine learning: methods and applications. Machine Learning Mastery.

[6] Zhou, Zhi-Hua. (2012). Ensemble Methods: Foundations and Algorithms. CRC Press.

[7] Dietterich, T. G. (2000). Ensemble methods in machine learning. Multiple classifier systems, 1-15.

[8] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. CRC press.

[9] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

[10] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[11] Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. R news, 2(3), 18-22

[12] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[13] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in neural information processing systems (pp. 4765-4774).

[14] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). Multivariate data analysis (7th ed.). Pearson Education.

[15] Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied linear statistical models (5th ed.). McGraw-Hill Irwin.