Module 4: Final Project Report

Bank Marketing Campaigns

Class ALY6040.80439: Data Mining

Prof. Justin Grosz

Team X members:

Anuj Khanna,

Hang Wu,

Ivan Todorov,

Fatima Nurmakhamadova &

Supreeth Murugesh

May 18, 2022

Table of Contents

I. Introduction

The dataset “Bank marketing campaigns" that we have chosen for the final project was obtained from Kaggle (Gavrysh, 2020). It is derived from a bank in Portugal that describes the results from the organization’s marketing campaigns, proposed by the phone calls. We have been hired as an external consulting service to evaluate the effects from the marketing campaign. To do so we will incorporate various data mining techniques to find trends and patterns of the current marketing practices. Furthermore, we will assess a range of predictive analytics that will allow us to forecast the likelihood of an individual opening a deposit account, given the circumstances. Finally, we will be providing a recommendation to the Portuguese Bank and the Telemarketing campaign of how these practices can be improved to maximize the effectiveness of campaigns of similar nature in the future to ultimately bring additional revenue to the bank.

II. About the data

To be able to identify the business problem, we had to make sure that the data at hand is in shape to conduct an analysis on and make sense of each variable that the telemarketing campaign has recorded. We first identified all data types within the dataset that turned out to consist of 7 numerical, including the index, and 14 string (object) type variables. Nextly, we wanted to know if missing data would affect our analysis but it turned out that neither of the features had any missing values. Furthermore, to ensure that we had to dive deeper to identify any values that could have been used as a substitute to a missing value, using data cleaning strategies.

Data Cleaning

A) Checking the data type and missing Values

First, we have checked the data type, and found that most of the variables were objects. Thus, we have converted them into categorical types as they contain categorical values. Next, to assess how the numerical variables are distributed, we did a summary of the statistics table identifying the IQR, maximum and minimum values among others. What stood out was the extremely high values (999) of pdays which is the variable that indicated the number of days since the individual was last contacted. We found out that this was a subsite value for if an individual has never been contacted before. Hence, we replaced these values with 0. Additionally, we parsed through the entire dataset to identify values marked as "unknown" and pinpointed 6 variables with unknown values - occupation (job), marital status (marital), education, credit default (default), housing loan and loan history in general. For best practices, we removed all unknown values which accounted for less than 5% of all values of the respected variables. The only exception we had was the credit default value which was 20.87% of all values because it is a form of a felony, however just 3 individuals had committed to it and therefore we have replaced the unknown values with the mode.

B) Identifying outliers

After handling the missing values, we moved to check the outliers shown in Graph 1 in the Appendix in the numerical variables and found outliers in all of them. But a significant amount of outliers was in Age, Duration, and Campaign columns. Initially, we used the Interquartile range method to remove the outliers which are out of the range of Q1/Q3 +/- 1.5. IQR. We have checked the outlier values in all columns, and all of them were correct and justified. For example, the age ranges from 17 to 98, where the average age is 39.8 years, while the upper limit is 47 years. Thus, we cannot remove the customers of age between 48 to 98 years as they are also an important part of the research. The total number of outliers was about 12,875 rows. This is less than a half of the total dataset, thus we decided to leave them because they can very well be key identifiers to assess the potential ideal target audience of the bank in the future as well as best practices of the telemarketing strategy going forward.

Further, we conducted an exploratory data analysis (EDA) to gather additional insights into the marketing campaign, demographic analysis of the audience reached, and the external factors that have influenced the decision of individuals to open a deposit account at the bank in Portugal.

C) Exploratory data analysis (EDA)

Visualizations are a powerful way to take this exploration of the campaign's data further. Firstly, looking into some of the distinguishing demographic factors would help us get an insight of the underlying demographics of the current audience. Graph 2 in the Appendix, shows how the age ranges are distributed and the respecting level of education of individuals. The average age of the customers is 40 years, while the minimum is 17, while the oldest is 98. The most frequent age range is individuals between 30 and 40.The population sample has a wide range of educational levels, having a university degree prevailing. Additionally, referring to graph 6 in the appendix, the majority of the audience targeted are employed with 11 distinctive occupations. Administrative positions are the most frequent occurrence followed by technicians and blue - collar positions. On the other hand, the number of unemployed and students is proportionally rather low. However, the percentage of students who have opened a bank account is around 25%. This could be something that the bank should account for in the future when assessing their potential customer base.

Next, understanding the mode of telemarketing campaign would aid us insights of how the marketing team leveraged each strategy. Graph 3 in the Appendix hand side shows the proportions of each outreach approach,with cellular being the more widely used one. Percentage wise, cellular has been also significantly more successful in converting potential clients to customers with 5.6% and 16.9% respectively. This can be due to the majority of the population being between 30 and 40 who are using primarily cellular networks as a mode of communication.

Duration of the calls is another interesting variable as an indicator of identifying successful marketing practices. Graph 4 in the Appendix is a scatter plot which shows the effectiveness of outreaches and their respective duration in seconds. The majority of the calls have been completed within 16 minutes (1000/60). The red color has a stronger color but we can notice the density of the blue indicators showing the rejection to open a deposit account by the target customer. Additionally, we can see that with the increase of time taken on the call, we can see that the distribution of successful and unsuccessful calls is quite evenly distributed. Therefore, we will keep the outliers for the duration factor as it might dilute our analysis going forward.

Apart from demographical and marketing strategy insights, the external factors affecting the client's decision should also be taken into consideration. For example, employment variability (emp.var) showed that people who tend to have more stable jobs are less likely to open an additional deposit account at another bank (Graph 5 in the Appendix). Similarly, the consumer confidence index (CCI) can be a strong indicator of how financially content an individual is and therefore their openness to open a bank account. Despite both variables supposedly having an impact on the predicting variable, by looking deeper in that relationship, as well as the one with one another, we found that they are highly correlated. Graph 6 in the appendix shows the correlation matrix of all numerical variables within the data sample. The color palette clearly signifies these relationships. The bottom right of the graph shows very strong dependency between individual features including employment variability (emp.var.rate), number of employees at the workplace (nr.employed), the consumer price index (CPI) (cons.price.idx) and euribor rate (euribor3m). To solidify these findings we used the variance inflation factor (VIF). The findings indicated a significant multicollinearity for CPI and number of employees as initially with VIF scores of 18,785 and 20,680 respectively. To mitigate lower prediction accuracy for our modeling, in the next section, as a result of multicollinearity we have removed the two variables from our features list.

III. Analysis

A) Logistic Regression

Logistic regression model is applied to the cleaned dataset and the statistical summary of the model is as shown in figure. There are around 17 variables that have values close to 0 and less than the statistically significant value of 0.05. Snapshot of the statistical summary of the logistic model as shown in Figure-2 is attached in the appendix. Ordering them with the highest absolute value of coefficients is the effect of the variable on the outcome variable. Poutcome_success - a successful outcome of the previous marketing campaign has the highest coefficient value which indicates that for *every 1 unit increase in the coefficient value of the Poutcome_success, it is highly likely that the client is likely to open a deposit account with the bank again*. Another interesting factor in the logistic model is the strong positive relationship between retired individuals and their likelihood using the baking services. The model gave an accuracy of **91.16 %.** The model performed well, that is **98%** of the times the model was successful in predicting if a client did not open a term deposit. However, due to the class imbalance, the model predicted only **38%** of the times correctly that a client will open a term deposit. The confusion matrix (Figure1) and the classification report (Figure-3) are attached in the appendix. The result tells us that the model predicted 6652 of True Negative and + 321 of True Positive = 6973 values correctly, and 143 of False Positive + 533 of False Negative = 676 values incorrectly. This means that the model predicted that 143 clients opened a deposit account while they did not, and that 533 clients did not place the deposit while they did. The result is skewed and not reliable as the model is not able to predict mostly if the client opened the deposit or not. This is a fine example of an **Imbalanced classification problem** as there were no sufficient data points of successful opening of a term depository in the training data for the model.

B) Decision Tree

The decision tree model facilitates decision-making by breaking down a problem into a series of questions. It is most commonly used to solve difficult challenges. We must reduce entropy to reduce prediction uncertainty and make the necessary judgments to solve the problem. Gini signifies purity or impurity, as well as entropy, in the decision tree. The lower the Gini value, the better, as it helps us to determine how frequently a randomly selected element is incorrectly detected, allowing the tree to split and identify the next characteristic to be selected for the tree's internal node.

In our situation, we created a decision tree to handle a classification problem in which the target variable has two different classes that indicate whether the customer will open a fixed deposit, or not, considering all the variables of interest. The Gini values for each tree level node decrease with each level, and samples denote how many samples were taken into account while making decisions, whilst class denotes the node's decision.“Duration”, “Poutcome_success”, “Contact_telephone”, and “euribor3m” are the features examined for this model. The accuracy of the model is 91%, which is somewhat similar to the logistic regression. The decision tree model also performed well, that is **96%** of the times the model was successful in predicting if a client did not open a term deposit while only 56% of the times a customer will open a term deposit.

Looking at the confusion matrix in Figure 5, we can see our model results with all variables. The result tells us that the model predicted 6495 of True Negative, + 480 of True Positive resulting in 6975 correct predictions. Moreover, the model predicted 300 of False Positive and + 374 of False Negative with a total of 674 false predictions. This means that the model predicted that 143 clients placed the deposit while they did not, and that 533 clients did not place the deposit while they did. Here we can notice that the decision tree predicted more True Positive values than the Logistic regression model.

C) XgBoost (Extreme Gradient Boosting)

The last predictive analytic methodology that we will use in the attempt to reach an even higher level of prediction accuracy for opening a deposit account as a result of a telemarketing campaign is the XgBoost. It is an algorithm designed to be highly efficient, flexible and portable that performs under the Gradient Boosting framework. Given the nature of our predictive variable, we have used the supervised learning branch designed to predict the outcome of the marketing campaign using classification. The only parameters used when constructing the model is the reduction of the learning rate of 0.5, as it aided the highest accuracy for training and nearly for testing as well, and the maximum depth of 4. The number of estimators were kept at the default (100).

Looking at the confusion matrix of XgBoost in figure 6, we can see the ensemble learning of the XgBoost can comfortably predict under what circumstances the client will not be opening an account at the bank (True Positives) in comparison to the instance when the customer will open an account (False Negatives). The algorithm seems to have just as high of an accuracy as the False Positives for True Negative value (407). Instances where the the profile of the client and the external factors point out to a high likelihood of not not opening an account, the algorithm falsely predicts that they will. This could very well be a product of an imbalance classification problem.

On figure 7, we can see the feature importance ranked in a descending order as a result of the XgBoost model. Duration, Euribor rate and consumer confidence index are the primary factors that would influence the client's decision of opening an account. Employment variability and the outcome of the previous campaign also have a significant weight in their decision making.

V. Interpretations

After gathering the results from our predictive analytics, it's time to evaluate the key takeaways and important findings. All three models have almost the same results, the further interpretation will be given based on the most frequent findings. Although the accuracy level was above 90% in all of them, their prediction accuracy of the profile of an individual who will open an account is significantly lower. This resulted in a high amount of false predicted values about 674 clients' decisions, having more of False Negative values. Thus, this might be a realistic challenge of providing a recommendation of why they would, however it creates a lot of room for suggestions of why individuals wouldn't open an account.

The initial question was to find the key factors that affected the client's decision to place a deposit in order to understand whether the marketing campaign works or not. Overall, we saw that the most influential variable that determines the client's decision is call duration. More clients agreed to place a deposit if the call duration was less than 827.5 seconds (13.7minutes) and the client was contacted by telephone, and their euribor rate was less than 4.965. While if the call duration was less than 166.5 seconds (2.7 minutes), then the majority of the clients still agreed to place the deposit if the previous contact outcome with them was successful.

Xgboost performs better than Decision Tree and Logistic Regression in almost every category and it's unlikely to cause overfit like the rest of the models. However realistically speaking since the data pipeline is expected to train the model with new data every week or so, it’s likely that we will use Decision Tree which is much faster to train and have stable accuracy. In terms of the Confusion Matrix comparison, 6640 True Negative (TN) records in the Logistic Regression Model (Figure1) are shown compared to the 6495 of the Decision Tree and 6568 of the XgBoost . Xgboost has a better True Negative/Nonsubscription prediction accuracy than Decision tree, and less than the LR model. The LR model has more False Negative predictions, 521, compared to the 374 of the Decision Tree and 401 of the XgBoost Model. The 39% increase of the FN predictions of the Decision Tree model indicates that LR model will overlook 39% of the subscribed users for terminal deposit as if they won’t subscribe. We would in turn lose 39%

of the users due to wrong prediction and therefore the bank wouldn't capitalize on this unrealized revenue. On the other hand, LR performs very well for the False Positive records, having 155 in comparison to the 300 of the Decision Tree and the 227 of the XgBoost. Since False Positive records are records that we predict will subscribe but they won't. We certainly want to minimize that in order to minimize the number of the total error calls. In this case LR performs 49% better than xgboost.

In terms of Confusion Matrix comparison, LR works slightly better than xgBoost and it is good at predicting FP and TN cases. Meaning that it's better at predicting the people who won't subscribe.

In terms of Classification Report for f1-score, Xgboost and Decision Tree perform better than Logistic Regression in terms of F1-score. Xgboost performs slightly better than Decision Tree in terms of precision prediction for 1 (0.67 vs 0,62) Figure 6 . The f1-score for the overall accuracy for the Xgboost is 0.92 which is 0.01 more than the Decision Tree.

The Feature Importance Diagrams give us an overview of the feature weights in the model during the fitting process. If the top feature dominates the FI diagram, then a model is likely to overfit. Conversely, the model would see multiple Features with distinctive importance, which would be a better fit. In this case, Xgboost has a more uniform Feature Importance than the Decision Tree and the LR. In fact, the Logistic Regression model sees a much stronger coefficient of the credit default category(-15.67) than the other features. This indicates a potentially far more sensitive model than the XgBoost model.

V. Conclusion

The marketing campaign has had moderate success in successfully converting potential customers to actual customers after the outreach with a success rate of a little over 12%. This disproportionate distribution of the data however, did make all of our models learn more instances of the unsuccessful attempts rather than the desired one, opening a deposit account, which we would consider the actual success. From a technical standpoint, to mitigate that imbalance classification model in data in the future, the bank can use over or under sampling of the data to achieve that balance. By doing so, we will be able to improve the recall score of our models.

Nevertheless, our models did prioritize certain features over others to be more statistical significance. Retired, previously contacted individuals and duration of the call are among the most important determinants of whether the person would be placing a deposit or not. In fact, a low duration call of less than 13 minutes, being deposited in the precious campaign and having an euribor rate less than 5 are strong descriptors of the perfect profile of a target customer and good telemarketing practices in maximizing the chances of conversion.

When it comes to telemarketing campaign strategy, we would recommend making a few additional changes to the approach for a higher call conversion rate. Based on the current analysis, changing the contact communication type solely to cellular would aid better final results. Since cell phones have wireless connections, most people carry them everywhere compared to the telephone which is tied to one place. Thus, the bank could reach customers anywhere. Another action is to set a call duration limit from 3 to 10 minutes, which will be an optimal duration to spark interest in the client, giving necessary information while keeping the call concise. Lastly, the company should build the relationship and discover the customers that agreed to place deposits in the previous campaigns as they would be more lenient to do this again.

VI. References

Gavrysh, V. (2020, January 12). Bank Marketing Campaigns Dataset: Opening deposit. Kaggle. Retrieved April 16, 2022, from https://www.kaggle.com/datasets/volodymyrgavrysh/bank-marketing-campaigns-dataset?resource=download

S. Moro, P. Cortez and P. Rita. “A Data-Driven Approach to Predict the Success of Bank Telemarketing.” (“Machine Learning Case Study: A data-driven approach to predict the success ...”) Decision Support Systems (2014), doi:10.1016/j.dss.2014.03.001.

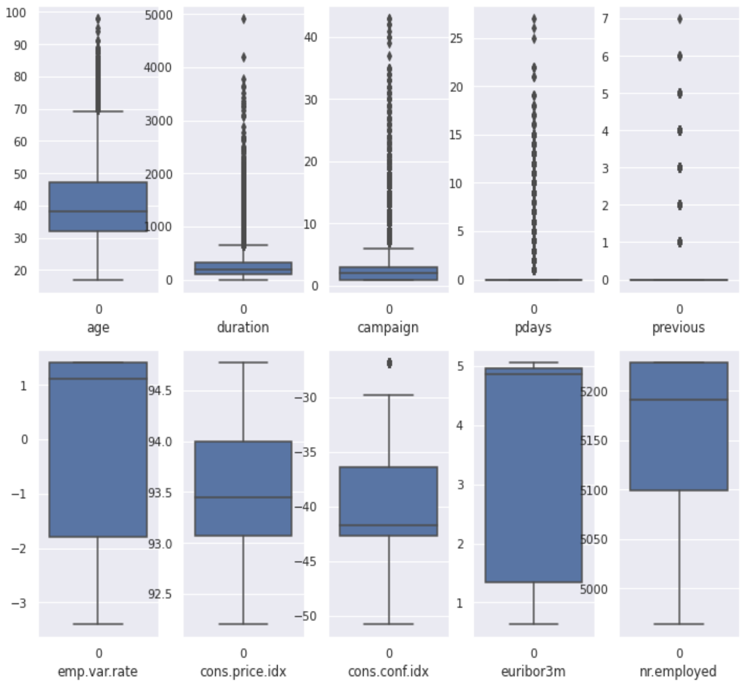
Dominitz, Jeff, and Charles F. Manski. 2004. "How Should We Measure Consumer Confidence?" *Journal of Economic Perspectives*, 18 (2): 51-66.DOI: 10.1257/0895330041371303

Sorokowski, P., Randall, A. K., Groyecka, A., Frackowiak, T., Cantarero, K., Hilpert, P., Ahmadi, K., Alghraibeh, A. M., Aryeetey, R., Bertoni, A., Bettache, K., Błażejewska, M., Bodenmann, G., Bortolini, T. S., Bosc, C., Butovskaya, M., Castro, F. N., Cetinkaya, H., Cunha, D., … Sorokowska, A. (2017). Marital satisfaction, sex, age, marriage duration, religion, number of children, economic status, education, and collectivistic values: Data from 33 countries. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01199

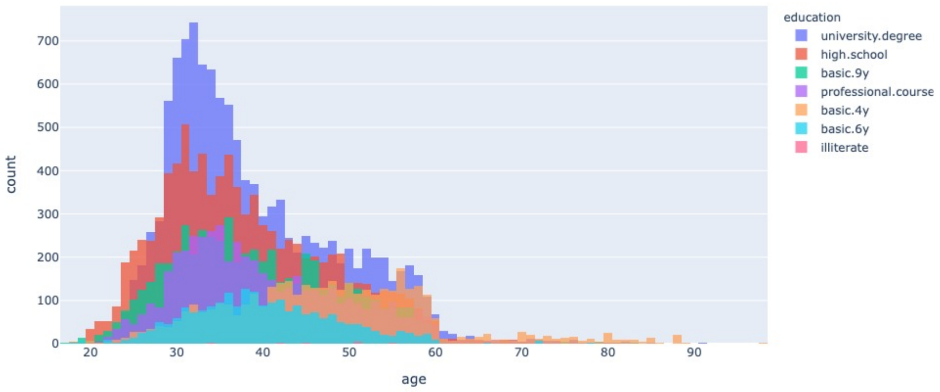
Notter, M. (2022, February 18). *Advanced exploratory data analysis (EDA) with python*. Medium. Retrieved April 24, 2022, from https://medium.com/epfl-extension-school/advanced-exploratory-data-analysis-eda-with-python-536fa83c578a

VII. Appendix

Graph 1: Boxplot of numerical variables **Return↶**



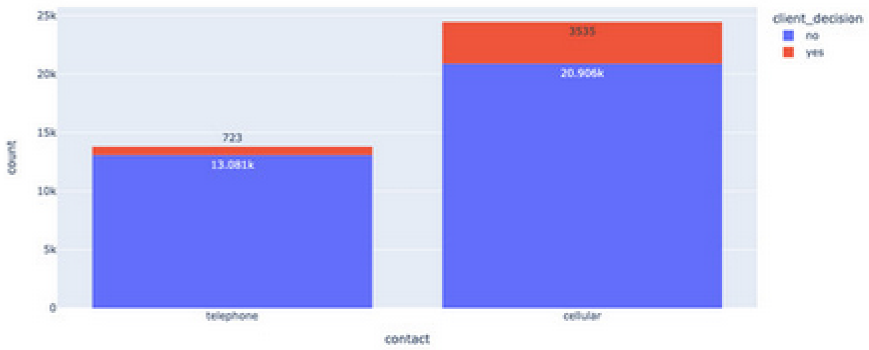
Return↶



Graph 1: Age distribution of the population by education

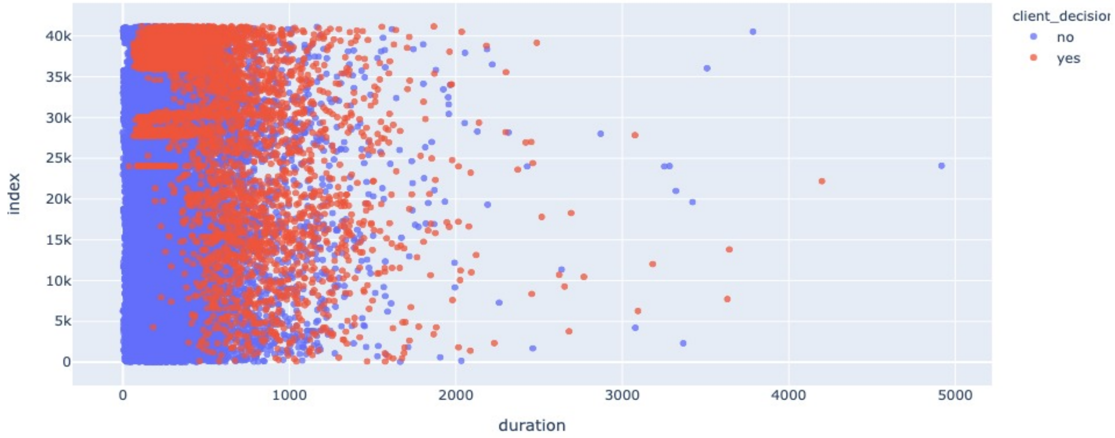
Return↶

Graph 2: Proportions of communications type used in the marketing campaign

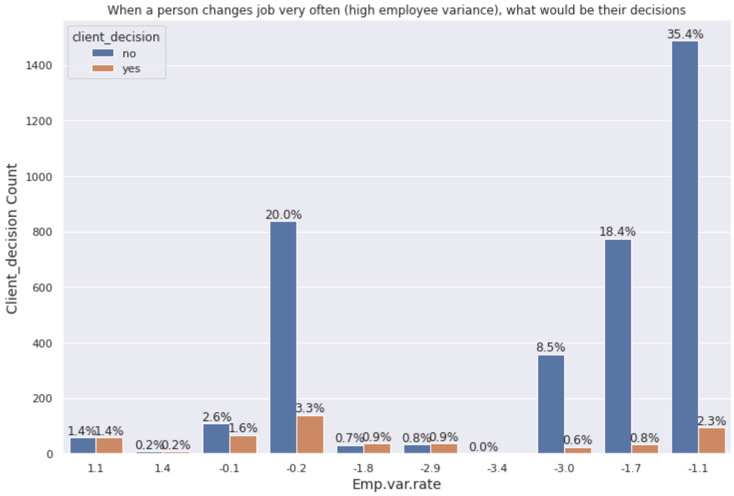


Return↩

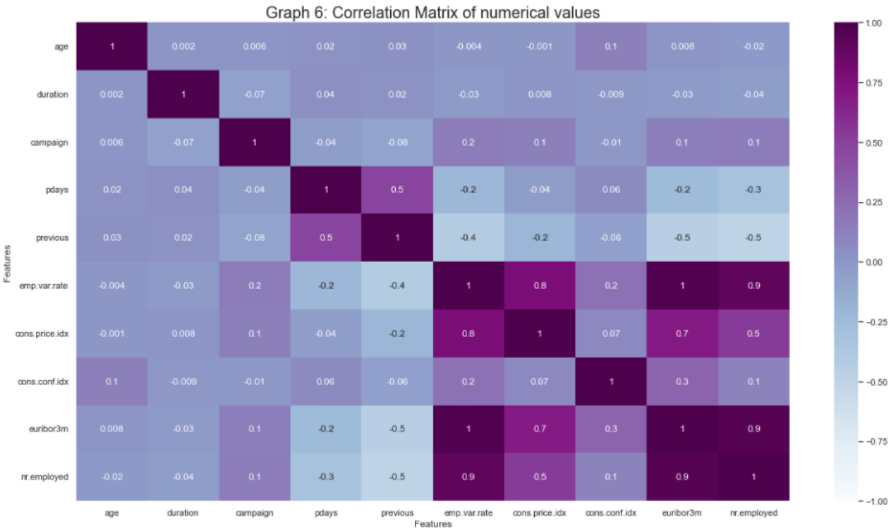
Graph 3: Effects on client decision based on call duration



Graph 5: Employment variability and final decision **Return↶**



Return↶



Graph 6: Final client decision by occupation **Return↺**

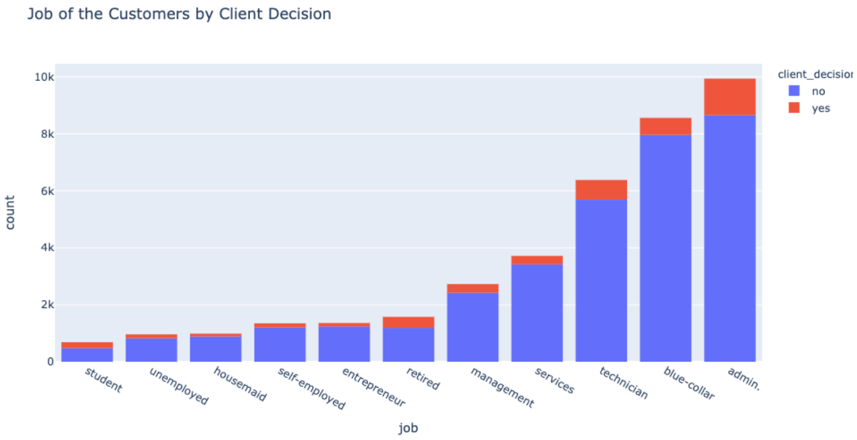


Figure 1: Confusion matrix of Logistic Regression **Return⎌**

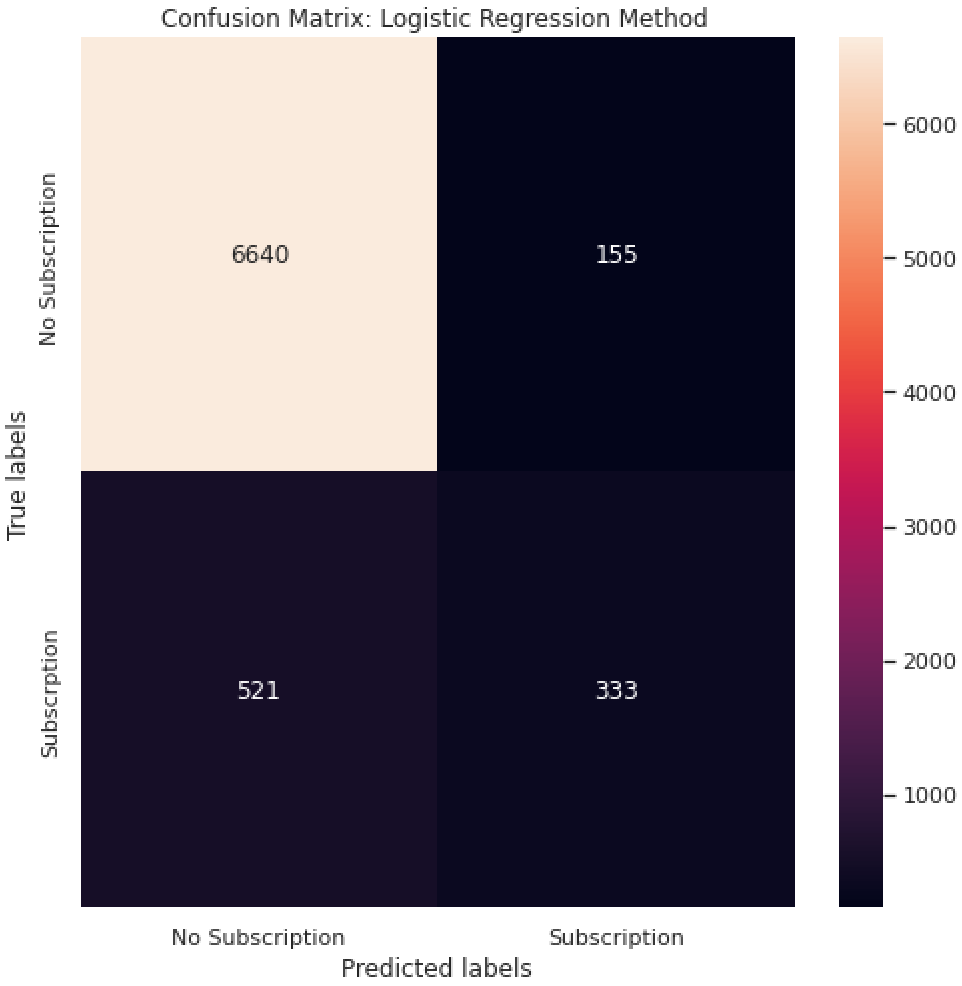
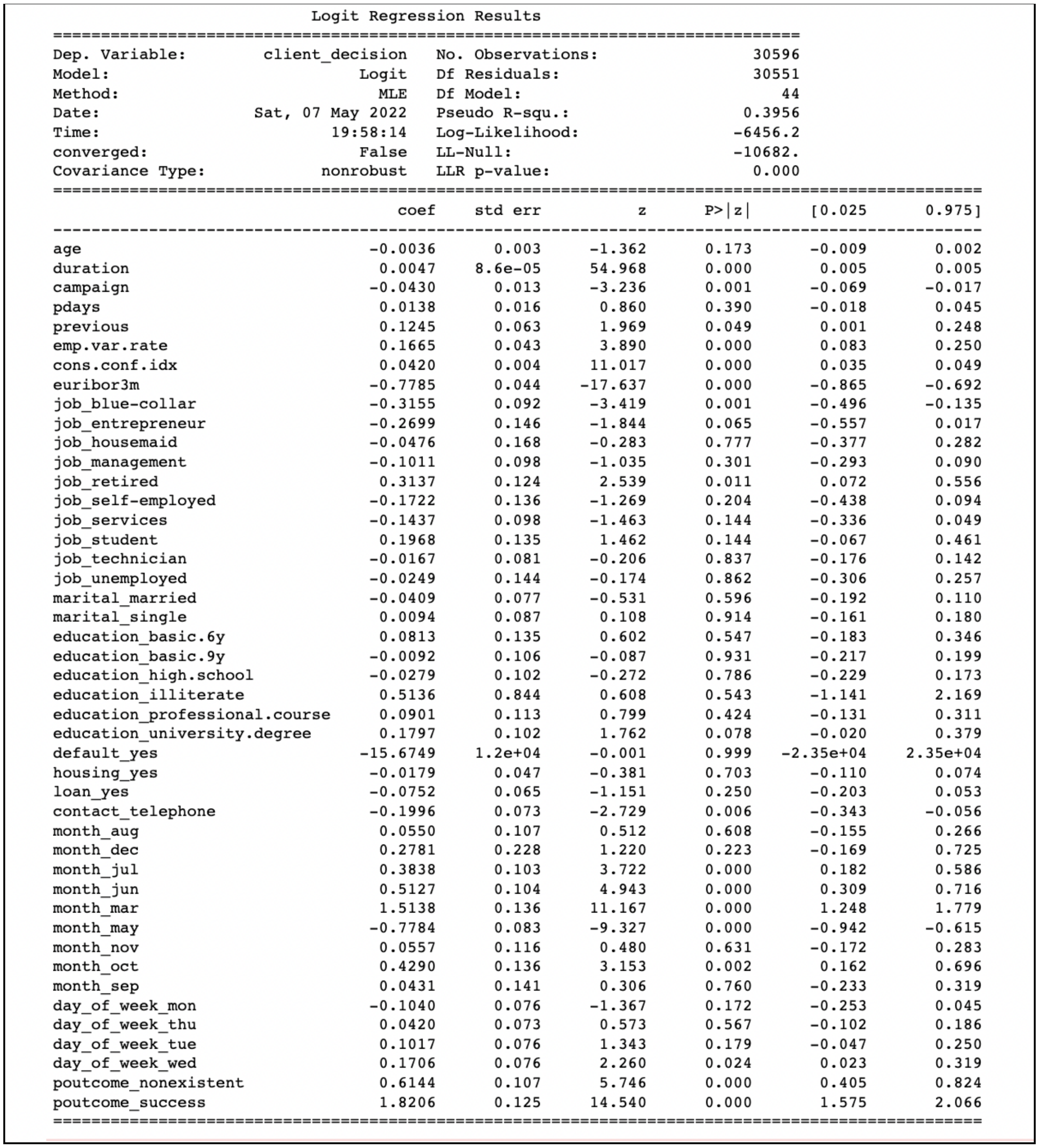
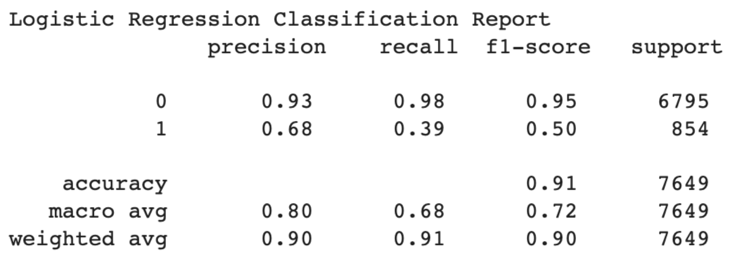


Figure 2: Statistical summary of Logistic Regression **Return⎌**



```
                           Logit Regression Results
==============================================================================
Dep. Variable:        client_decision   No. Observations:                30596
Model:                          Logit   Df Residuals:                    30551
Method:                           MLE   Df Model:                           44
Date:                Sat, 07 May 2022   Pseudo R-squ.:                  0.3956
Time:                        19:58:14   Log-Likelihood:                -6456.2
converged:                      False   LL-Null:                       -10682.
Covariance Type:            nonrobust   LLR p-value:                     0.000
=================================================================================================
                                    coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
age                              -0.0036      0.003     -1.362      0.173      -0.009       0.002
duration                          0.0047    8.6e-05     54.968      0.000       0.005       0.005
campaign                         -0.0430      0.013     -3.236      0.001      -0.069      -0.017
pdays                             0.0138      0.016      0.860      0.390      -0.018       0.045
previous                          0.1245      0.063      1.969      0.049       0.001       0.248
emp.var.rate                      0.1665      0.043      3.890      0.000       0.083       0.250
cons.conf.idx                     0.0420      0.004     11.017      0.000       0.035       0.049
euribor3m                        -0.7785      0.044    -17.637      0.000      -0.865      -0.692
job_blue-collar                  -0.3155      0.092     -3.419      0.001      -0.496      -0.135
job_entrepreneur                 -0.2699      0.146     -1.844      0.065      -0.557       0.017
job_housemaid                    -0.0476      0.168     -0.283      0.777      -0.377       0.282
job_management                   -0.1011      0.098     -1.035      0.301      -0.293       0.090
job_retired                       0.3137      0.124      2.539      0.011       0.072       0.556
job_self-employed                -0.1722      0.136     -1.269      0.204      -0.438       0.094
job_services                     -0.1437      0.098     -1.463      0.144      -0.336       0.049
job_student                       0.1968      0.135      1.462      0.144      -0.067       0.461
job_technician                   -0.0167      0.081     -0.206      0.837      -0.176       0.142
job_unemployed                   -0.0249      0.144     -0.174      0.862      -0.306       0.257
marital_married                  -0.0409      0.077     -0.531      0.596      -0.192       0.110
marital_single                    0.0094      0.087      0.108      0.914      -0.161       0.180
education_basic.6y                0.0813      0.135      0.602      0.547      -0.183       0.346
education_basic.9y               -0.0092      0.106     -0.087      0.931      -0.217       0.199
education_high.school            -0.0279      0.102     -0.272      0.786      -0.229       0.173
education_illiterate              0.5136      0.844      0.608      0.543      -1.141       2.169
education_professional.course     0.0901      0.113      0.799      0.424      -0.131       0.311
education_university.degree       0.1797      0.102      1.762      0.078      -0.020       0.379
default_yes                     -15.6749    1.2e+04     -0.001      0.999   -2.35e+04    2.35e+04
housing_yes                      -0.0179      0.047     -0.381      0.703      -0.110       0.074
loan_yes                         -0.0752      0.065     -1.151      0.250      -0.203       0.053
contact_telephone                -0.1996      0.073     -2.729      0.006      -0.343      -0.056
month_aug                         0.0550      0.107      0.512      0.608      -0.155       0.266
month_dec                         0.2781      0.228      1.220      0.223      -0.169       0.725
month_jul                         0.3838      0.103      3.722      0.000       0.182       0.586
month_jun                         0.5127      0.104      4.943      0.000       0.309       0.716
month_mar                         1.5138      0.136     11.167      0.000       1.248       1.779
month_may                        -0.7784      0.083     -9.327      0.000      -0.942      -0.615
month_nov                         0.0557      0.116      0.480      0.631      -0.172       0.283
month_oct                         0.4290      0.136      3.153      0.002       0.162       0.696
month_sep                         0.0431      0.141      0.306      0.760      -0.233       0.319
day_of_week_mon                  -0.1040      0.076     -1.367      0.172      -0.253       0.045
day_of_week_thu                   0.0420      0.073      0.573      0.567      -0.102       0.186
day_of_week_tue                   0.1017      0.076      1.343      0.179      -0.047       0.250
day_of_week_wed                   0.1706      0.076      2.260      0.024       0.023       0.319
poutcome_nonexistent              0.6144      0.107      5.746      0.000       0.405       0.824
poutcome_success                  1.8206      0.125     14.540      0.000       1.575       2.066
=================================================================================================
```

Figure 3: Classification report of Logistic Regression



```
Logistic Regression Classification Report
              precision    recall  f1-score   support

           0       0.93      0.98      0.95      6795
           1       0.68      0.39      0.50       854

    accuracy                           0.91      7649
   macro avg       0.80      0.68      0.72      7649
weighted avg       0.90      0.91      0.90      7649
```

Figure 5: Confusion matrix of Decision Tree

Return⮌

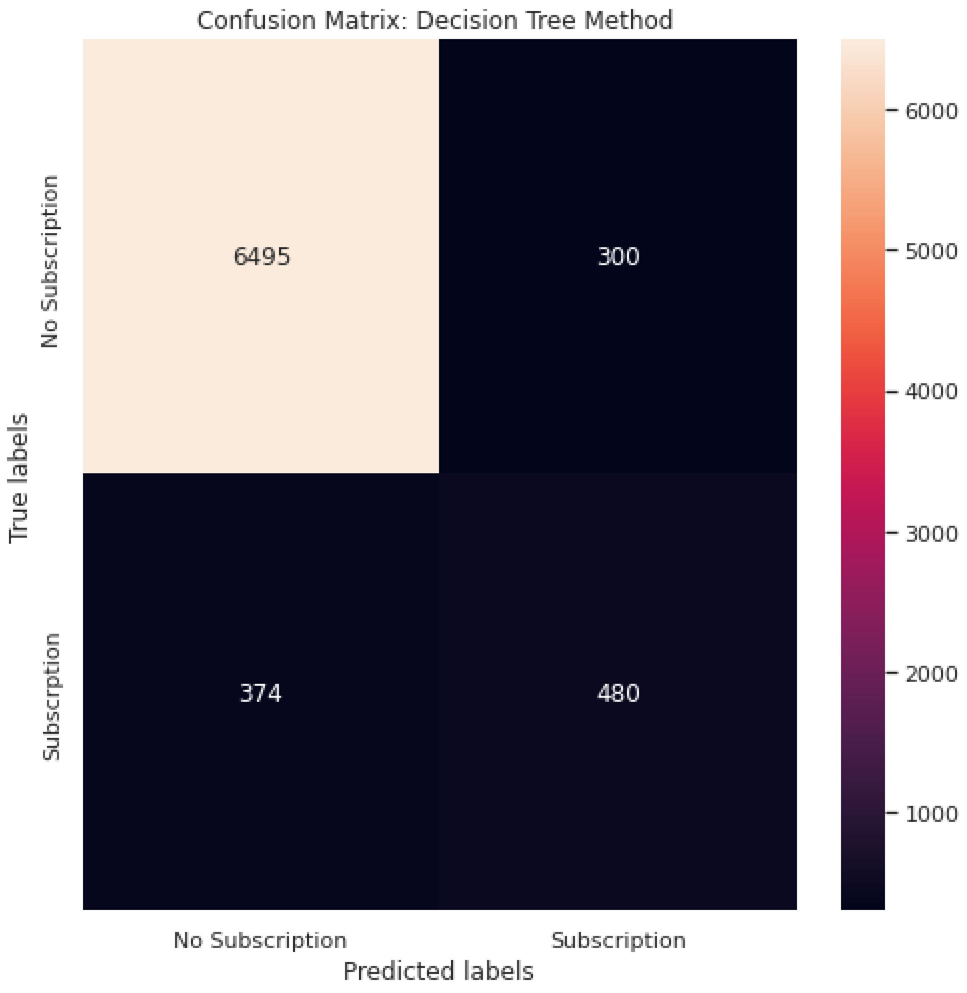


Figure 6: Confusion Matrix XgBoost **Return↶**

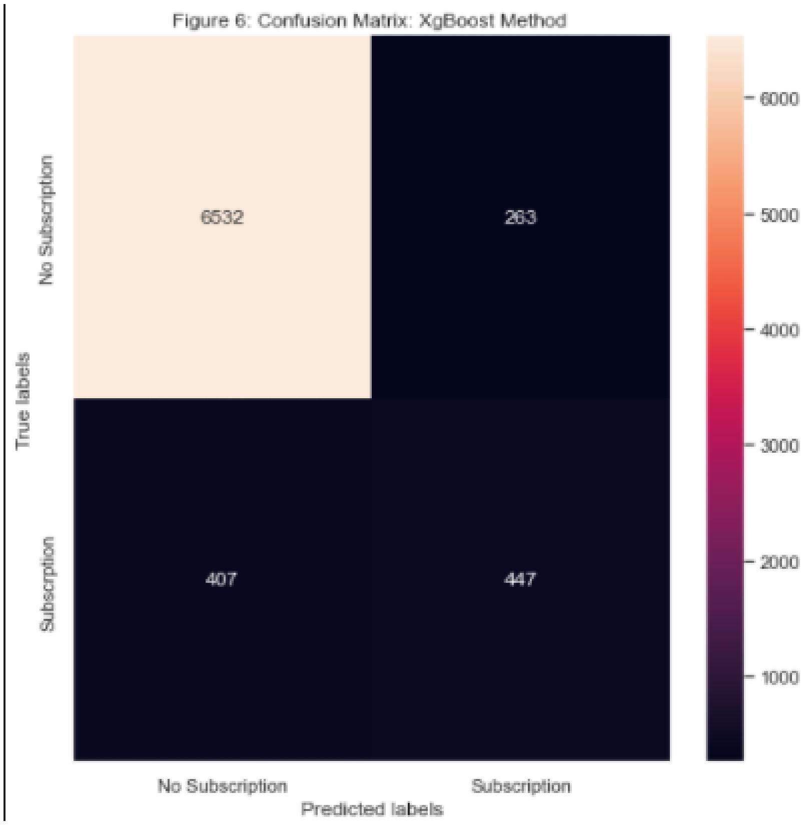


Figure 4: Decision Tree

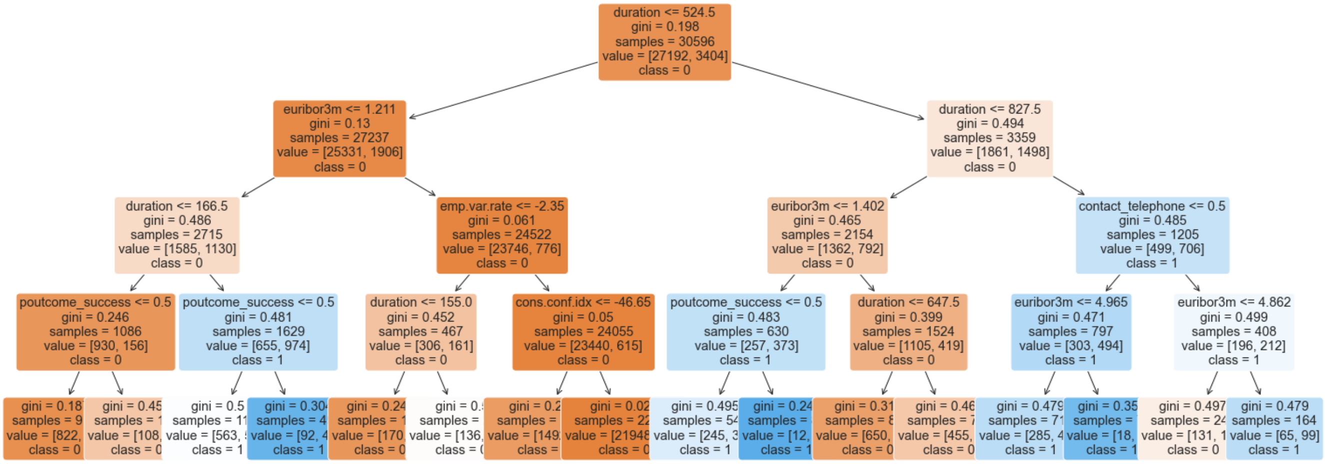
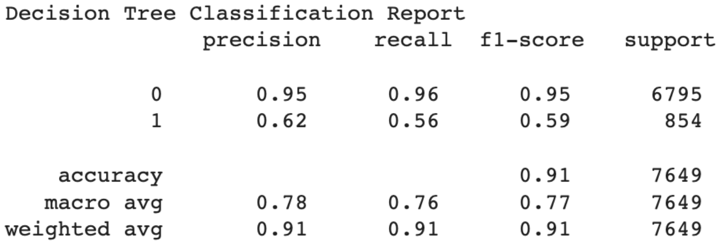


Figure 6: Classification report of Decision Tree

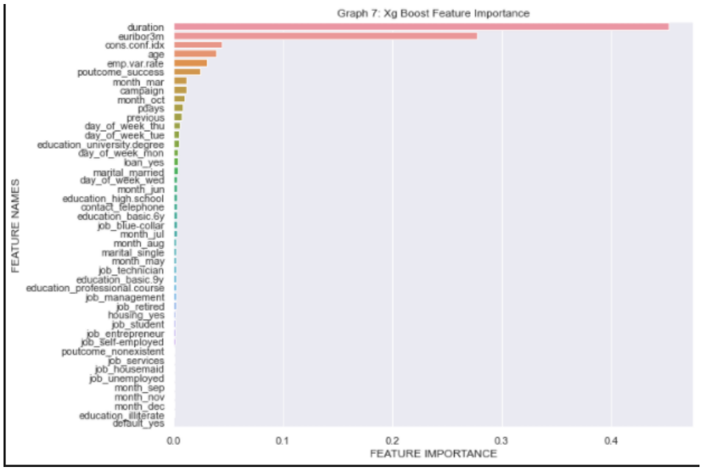


```
Decision Tree Classification Report
              precision    recall  f1-score   support

           0       0.95      0.96      0.95      6795
           1       0.62      0.56      0.59       854

    accuracy                           0.91      7649
   macro avg       0.78      0.76      0.77      7649
weighted avg       0.91      0.91      0.91      7649
```

Figure 7: Feature importance XgBoost **Return⎌**



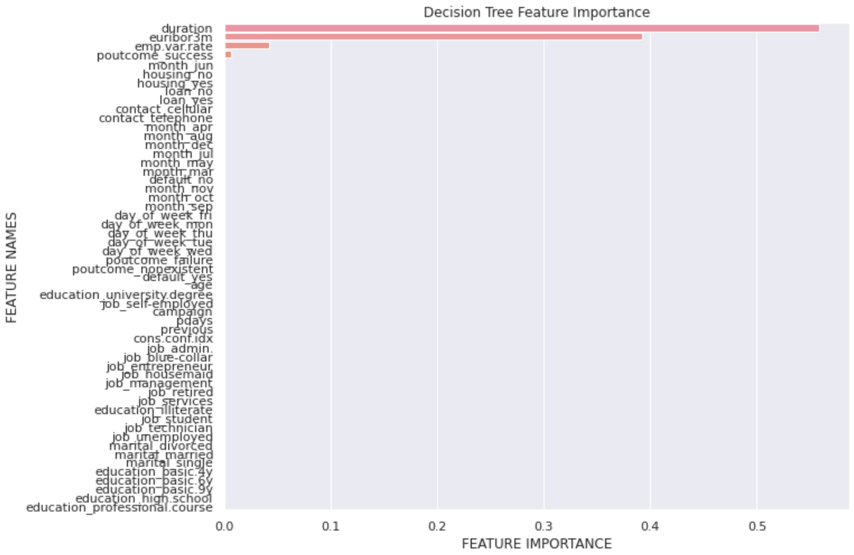
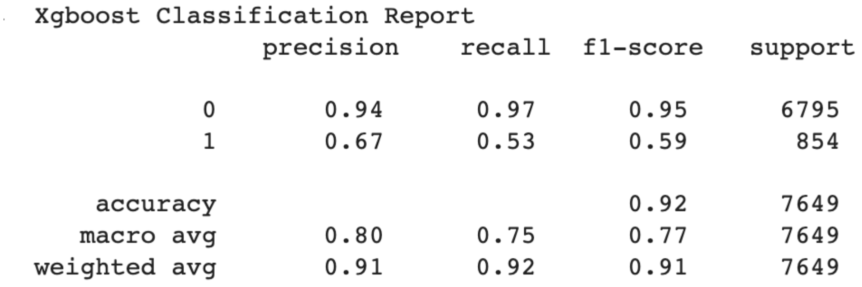


Figure 8: Xgboost Classification Report

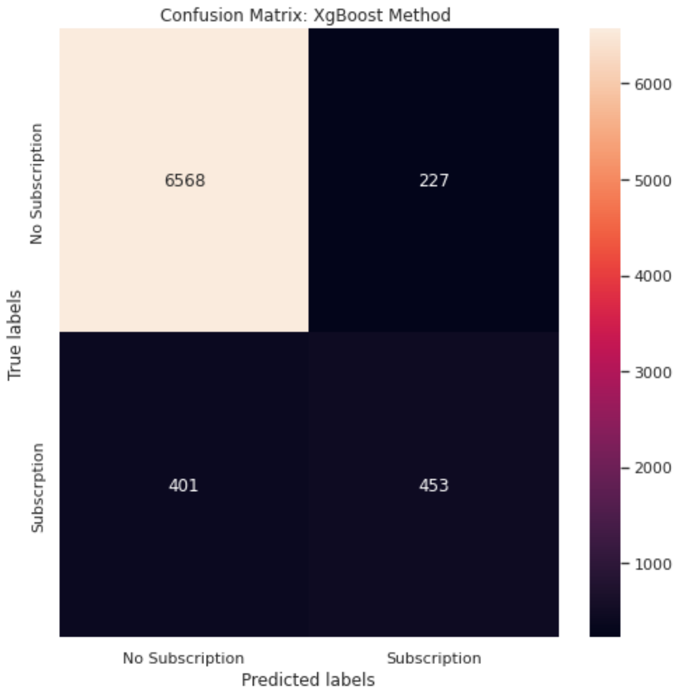


```
Xgboost Classification Report
              precision    recall  f1-score   support

           0       0.94      0.97      0.95      6795
           1       0.67      0.53      0.59       854

    accuracy                           0.92      7649
   macro avg       0.80      0.75      0.77      7649
weighted avg       0.91      0.92      0.91      7649
```

Figure 9:Xgboost Confusion Matrix


Confusion Matrix: XgBoost Method
True labels
No Subscription
Subscrption
6568
227
401
453
No Subscription
Subscription
Predicted labels
6000
5000
4000
3000
2000
1000