

A Data Analytic Approach to Bank Marketing - Case Study

submitted by,

Avik Tarafdar

Class Roll No: 002210503034

Exam Roll: MCA244033

Master of Computer Application

under the supervision of

Prof. (Dr.) Subhadip Basu

in the

Department of Computer Science and Engineering

Faculty of Engineering and Technology

Jadavpur University

2024

**Faculty of Engineering and Technology
Jadavpur University**

CERTIFICATE OF RECOMMENDATION

To Whom It May Concern

I hereby recommend that the project report entitled “A Data Analytic Approach to Bank Marketing - Case Study” has been carried out by Avik Tarafdar (Reg. No.: 163626 of 2022-23, Roll No.: 002210503034, Exam Roll: MCA244033), under my guidance and supervision and be accepted in partial fulfilment of the requirement for the degree of Master of COMPUTER APPLICATION in DEPARTMENT of COMPUTER SCIENCE and ENGINEERING, JADAVPUR UNIVERSITY during the academic year 2023-2024.

Prof. (Dr.) Subhadip Basu
Project Supervisor
Dept. of Computer Science and
Engineering
Jadavpur University

Prof. (Dr.) Nandini Mukhopadhyay
Head of the Department
Dept. of Computer Science & Engineering
Jadavpur University

Prof. Dipak Laha
Dean, Faculty Council
of Engineering & Technology
Jadavpur University, Kolkata-032

Faculty of Engineering and Technology
Jadavpur University

CERTIFICATE OF APPROVAL

This is to certify that the project report entitled “**A Data Analytic Approach to Bank Marketing - Case Study**” is a bonafide record of work carried out by AVIK TARAFDAR, in fulfillment of the requirements for the award of the degree of Master of Computer Application in the Department of Computer Science and Engineering, Jadavpur University. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the project report only for the purpose for which it has been submitted.

Signature of Examiner 1

Date:

Signature of Examiner 2

Date:

Faculty of Engineering and Technology
Jadavpur University

**DECLARATION OF COMPLIANCE WITH
ACADEMIC ETHICS**

I hereby declare that this project report entitled "A Data Analytic Approach to Bank Marketing - Case Study" is an outcome of my genuine and independent work, as part of my Master of Computer Application program.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, all materials and results that are not original to this work have been fully cited and referenced.

Avik Tarafdar

Registration No.: 163626 of 2022-23

Examination Roll No.: MCA244033

Thesis Title: A Data Analytic Approach to Bank Marketing - Case Study

Signature with Date

ACKNOWLEDGEMENT

With my most sincere and gratitude, I would like to thank **Prof. (Dr.) Subhadip Basu, *Department of Computer Science & Engineering***, my supervisor, for his overwhelming support throughout the duration of the project. His motivation always gave me the required inputs and momentum to continue with my work, without which the project work would not have taken its current shape. His valuable suggestions and numerous discussions have always inspired new ways of thinking. I feel deeply honored that I got this opportunity to work under him.

I would like to express my sincere thanks to all my managers and senior employees of ***Hongkong and Sanghai Banking Corporation*** to allow me to work with them in this project.

I would like to thank all the faculty members of the ***Department of Computer Science & Engineering of Jadavpur University*** for their continuous support.

Last, but not the least, I would like to thank my batch mates for staying by my side when I needed them.

Avik Tarafdar

Class Roll No.: 002210503034

(Exam Roll No.: MCA244033)

ABSTRACT

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit. Total 41,188 Clients record are available in the dataset.

Key Data elements of the dataset are following –

1. Client Demographics (Age, Job, Marital & Education Status)
2. Client Financial Info (Default, Personal & Home Loan Status)
3. Campaign Attributes (Mode, Month/Day and Duration of contact)
4. Output variable / Desired target: Has the client subscribed a term deposit

The objective of the analysis is to assess client details subscribing the Product – Bank Term Deposit and provide an insight to run an efficient & successful marketing campaign in future.

We will analyze the average age of the clients subscribing the product and job type of the clients who subscribed the product. Then we will build a Machine Learning (Categorical) model which can predict whether the client will subscribe to a Term Deposit or not.

As the dataset is highly imbalanced, so we applied smote and random oversampling methods and then we used logistic regression and xgboost and compared the accuracy, precision and recall.

Finally, we saw how age and occupation can be useful in determining whether a client will subscribe a term deposit.

CONTENTS

1 Introduction	7
1.1 Overview to the Project	7
1.2 Data Descriptions and Types	9
1.3 Exploratory Data Analysis	11
2 Methodology	16
2.1 Data Preparation:	16
2.2 Applying classification models:	18
3 Results and Discussion:	19
3.1 Accuracy and Other metrics:	19
3.2 Alternative Solutions:	23
3.3 Probability of subscribing the product:	24
4 Conclusion	25
5 References	27

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW OF THE PROJECT:

The objective of this project is to develop a predictive model to determine whether a customer will subscribe to a term deposit based on various features provided in the bank marketing dataset. This report outlines the methodologies employed, analyses conducted, and insights gained throughout the project.

1. Data Exploration and Preprocessing:

- The dataset contains information about customers, their interactions with the bank, and the outcome variable indicating whether they subscribed to a term deposit.
- Initial exploration involved examining data types, checking for missing values, and understanding the distribution of features.
- Preprocessing steps included handling missing data, encoding categorical variables, and scaling numerical features to ensure compatibility with machine learning algorithms.
- Due to class imbalance, techniques such as SMOTE (Synthetic Minority Over-sampling Technique) and random oversampling were applied to balance the dataset.

2. Model Selection and Training:

- Two algorithms were chosen for model development: XG Boost and logistic regression.
- XG Boost, an ensemble learning algorithm, was selected for its ability to handle complex relationships and deliver high predictive performance.
- Logistic regression, a simpler algorithm, was chosen as a baseline model for comparison.
- The dataset was split into training and testing sets using a stratified approach to preserve the class distribution.
- Models were trained on the training data and evaluated using metrics such as accuracy, precision, recall, and F1-score.

3. Univariate and Multivariate Analysis:

- Univariate analysis involved studying the distribution of individual features to identify outliers and understand their characteristics.

- Multivariate analysis focused on exploring relationships between multiple variables to uncover patterns and correlations relevant to the predictive task.

4. Evaluation and Conclusion:

- Model performance was evaluated using a combination of metrics to assess predictive accuracy and generalization capability.
- XG Boost outperformed logistic regression, demonstrating superior predictive performance in terms of accuracy and other evaluation metrics.
- The report concludes with a summary of findings, limitations of the approach, and recommendations for future research or improvement.

In conclusion, this project successfully developed predictive models for bank marketing using the provided dataset. Through comprehensive data exploration, preprocessing, model training, and evaluation, valuable insights were gained into customer behaviour and factors influencing term deposit subscriptions. Further refinement and experimentation could enhance the predictive accuracy and applicability of the models in real-world scenarios.

1.2 DATA DESCRIPTIONS AND TYPES:

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The dataset has 41,188 rows corresponding to each customer and each row has 21 attributes.

Attribute Information:

1. Bank client data:

- **Age:** (numeric)
- **Job:** type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
- **Marital:** marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
- **Education:** (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
- **Default:** has credit in default? (categorical: 'no', 'yes', 'unknown')
- **Housing:** has housing loan? (categorical: 'no', 'yes', 'unknown')
- **Loan:** has personal loan? (categorical: 'no', 'yes', 'unknown')

2. Related with the last contact of the current campaign:

- **Contact:** contact communication type (categorical: 'cellular', 'telephone')
- **Month:** last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- **Day_of_week:** last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')
- **Duration:** last contact duration, in seconds (numeric).

3. Other attributes:

- **Campaign:** number of contacts performed during this campaign and for this client (numeric, includes last contact)
- **Pdays:** number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

- **Previous:** number of contacts performed before this campaign and for this client (numeric)
- **Poutcome:** outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')

4. Social and economic context attributes

- **Emp.var.rate:** employment variation rate - quarterly indicator(numeric)
- **Cons.price.idx:** consumer price index - monthly indicator (numeric)
- **Cons.conf.idx:** consumer confidence index - monthly indicator(numeric)
- **Euribor3m:** euribor 3month rate - daily indicator (numeric)
- **Nr.employed:** number of employees - quarterly indicator (numeric)

5. Output variable (desired target):

- **y** - has the client subscribed a term deposit? (binary: 'yes', 'no')

1.3 Exploratory Data Analysis:

Now we will perform univariate analysis to check whether there is any imbalance in the dataset and relation of each categorical variable with the output.

1.Univariate Analysis

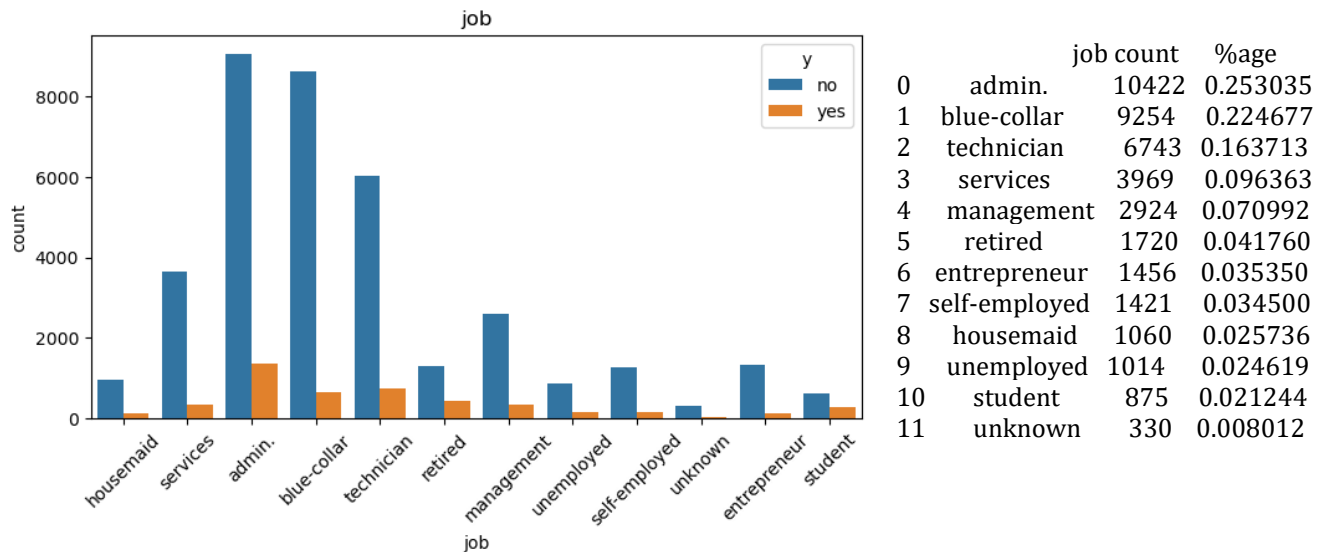


Fig1: job type Vs count of bank term deposit

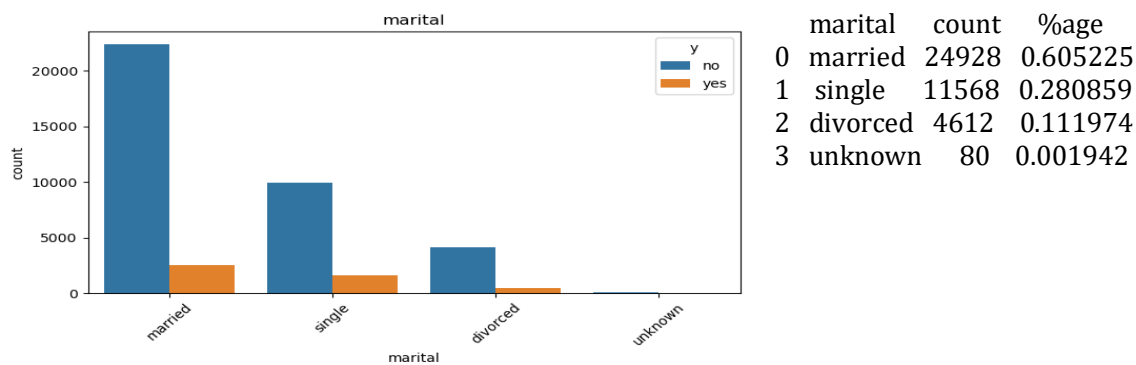


Fig2: marital status Vs count of bank term deposit

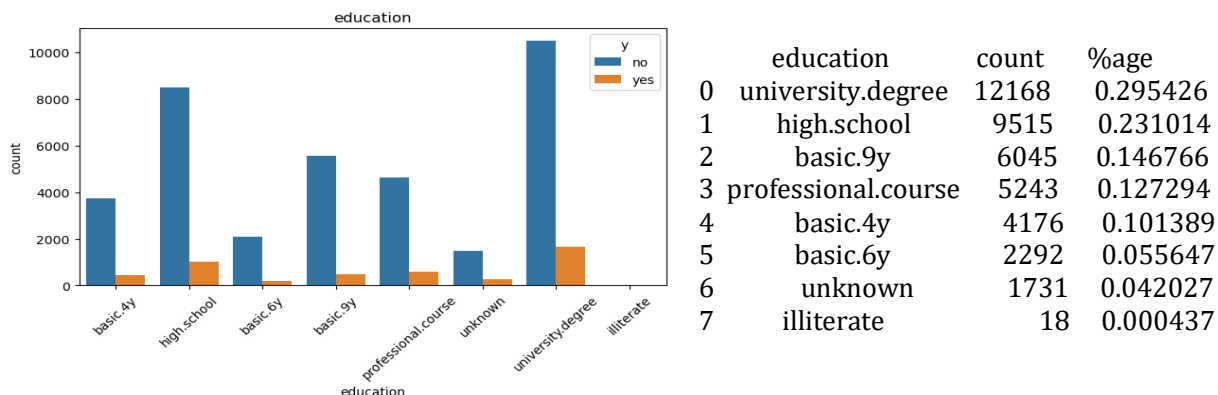
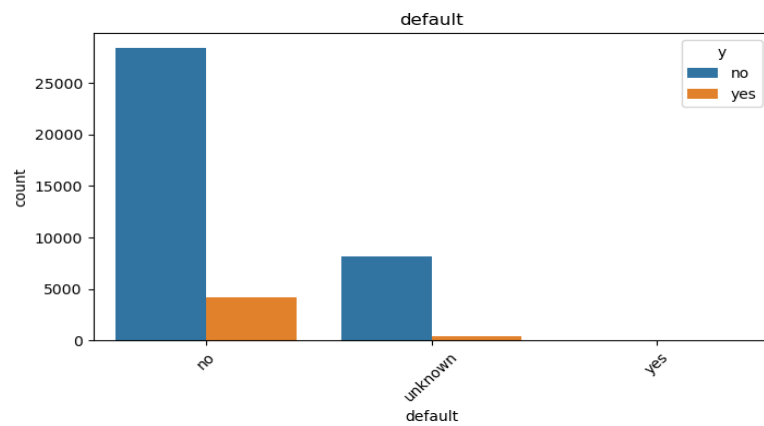
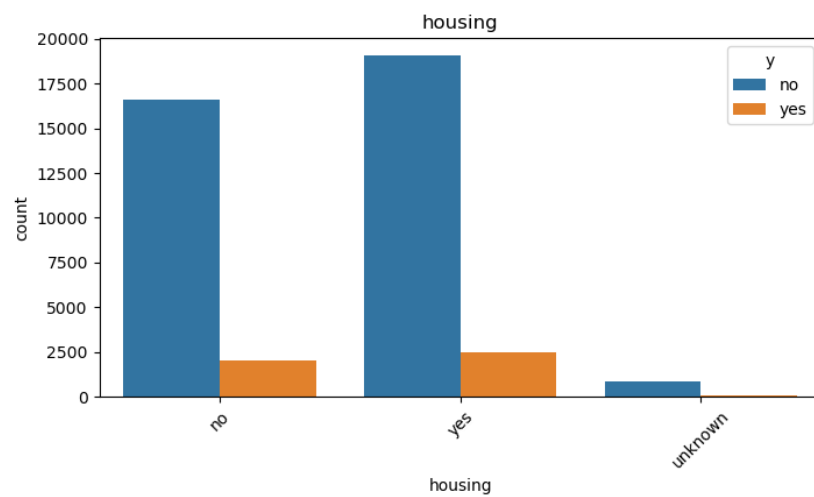


Fig3: education Vs count of bank term deposit



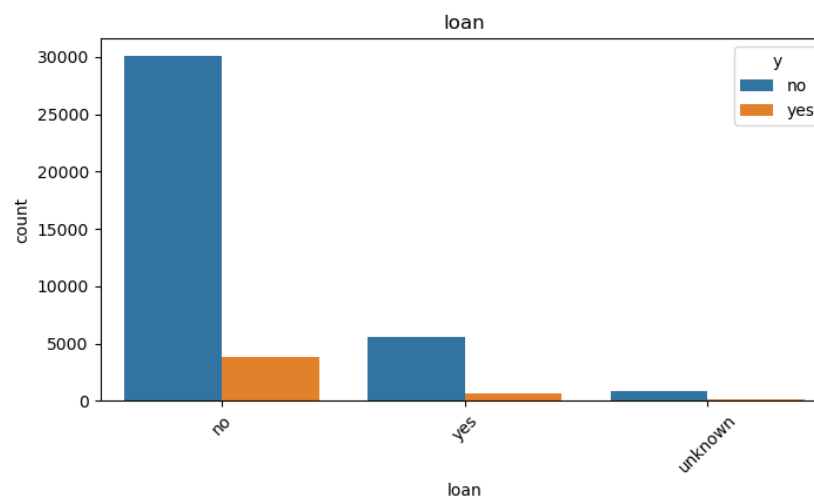
	default	count	%age
0	no	32588	0.791201
1	unknown	8597	0.208726
2	yes	3	0.000073

Fig4: loan default Vs count of bank term deposit



	housing	count	%age
0	yes	21576	0.523842
1	no	18622	0.452122
2	unknown	990	0.024036

Fig5: housing loan Vs count of bank term deposit



	Loan	count	%age
0	no	33950	0.824269
1	yes	6248	0.151695
2	unknown	990	0.024036

Fig6: personal loan Vs count of bank term deposit

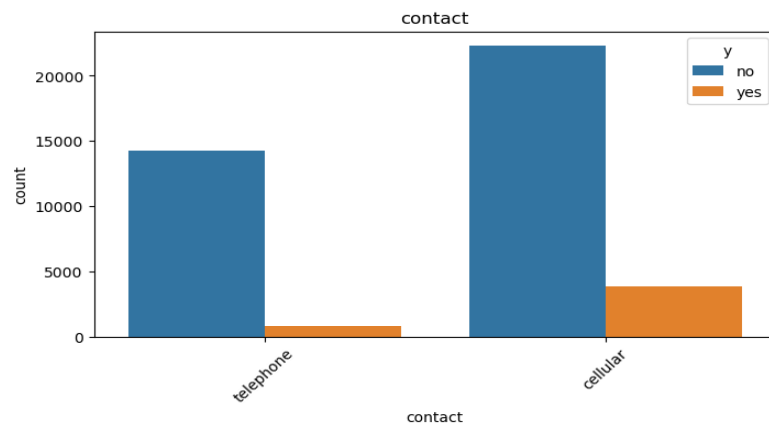


Fig7: contact communication type Vs count of bank term deposit

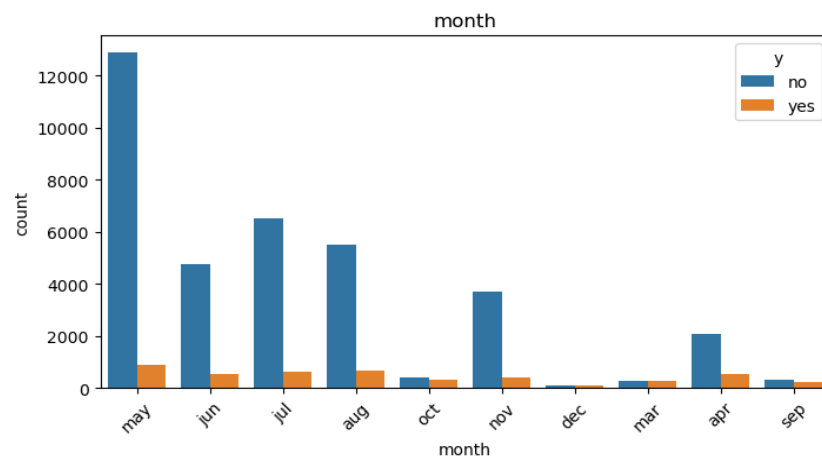
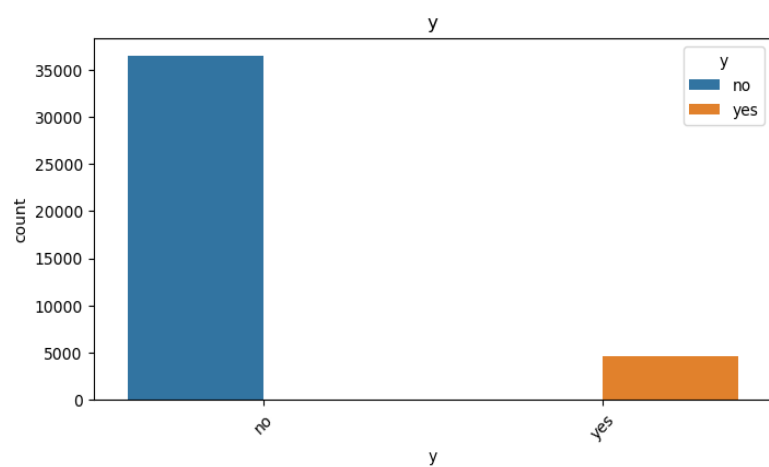


Fig8: month Vs count of bank term deposit



y	count	%age
0 no	36548	0.887346
1 yes	4640	0.112654

Fig9: count of bank term deposits

As we can see from the above univariate analysis that the dataset is highly imbalanced and therefore if we use any machine learning algorithm there will be some bias.

To overcome this situation, we will use two procedure -

- **SMOTE (Synthetic Minority Oversampling)**
- **Random resampling**

1. **SMOTE (Synthetic Minority Oversampling):** SMOTE works by generating synthetic examples of the minority class to balance the class distribution. It does this by creating new instances along the line segments joining existing minority class instances. Specifically, for each minority class instance, SMOTE selects one or more of its k-nearest neighbors and generates synthetic instances along the line segments connecting them.
2. **Random resampling:** In random resampling, the dataset is manipulated to achieve a more balanced class distribution by either oversampling the minority class, under sampling the majority class, or a combination of both.

2. Multivariate analysis

For multivariate analysis we will take all the numerical features along with a new feature `y_yes` which indicates whether a client has subscribed the term deposit. We can see with duration `y_yes` has good correlation 0.42 which indicates duration of the last call has good impact in determining subscription of term deposit.

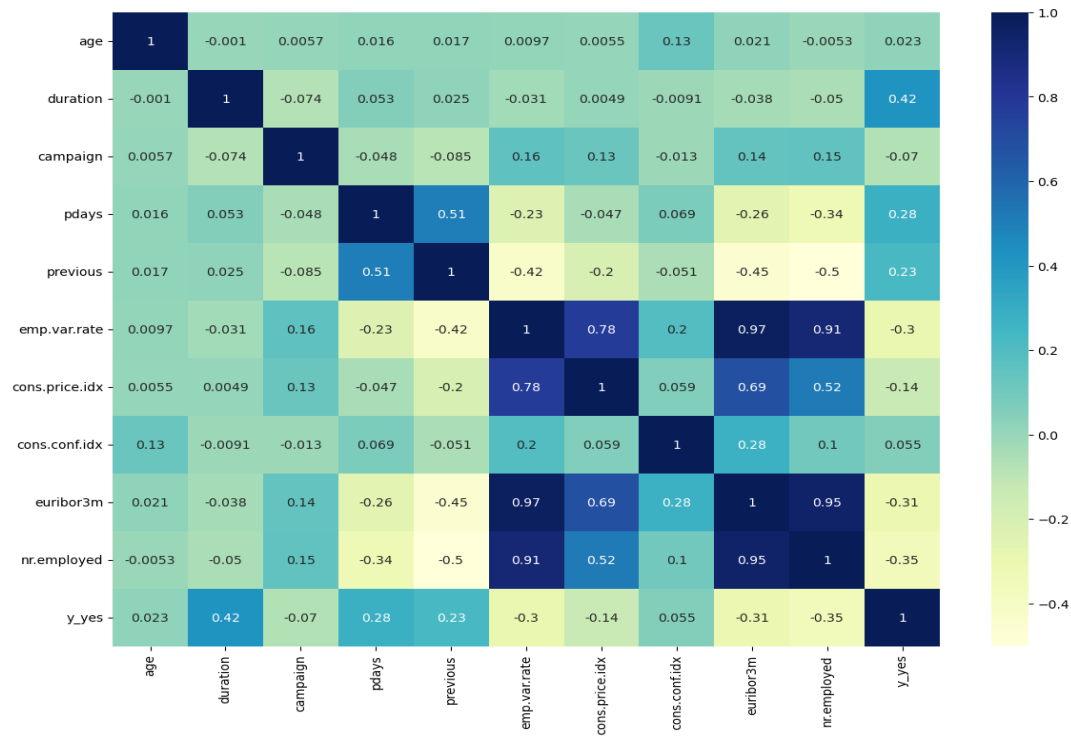


Fig10: heatmap of numerical columns

We found a high correlation between 4 economic variables based on the Heat Map Method “emp.var.rate”, “cons.price.idx”, “euribor3m”, “nr.employed”. So we will remove those variables and we will get the following heatmap after removing those variables.



Fig11: heatmap of numerical columns after removing 4 economic variables

CHAPTER 2

METHODOLOGY

2.1 DATA PREPERATION:

For numerical variables we plotted boxplots to check for outliers and then we removed the outliers by removing extreme values by replacing them with 99 percentile value.

For categorical variables we imputed the unknown values with mode of the distribution and then we applied one-hot-encoding.

1. Outlier Detection and removal

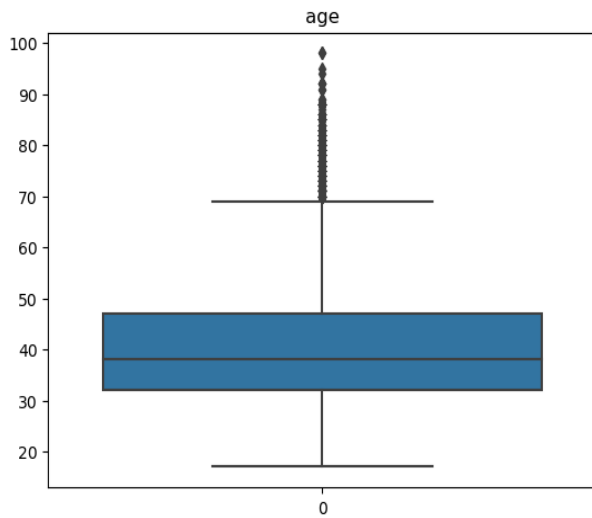


Fig12: Boxplot for age before removing outlier

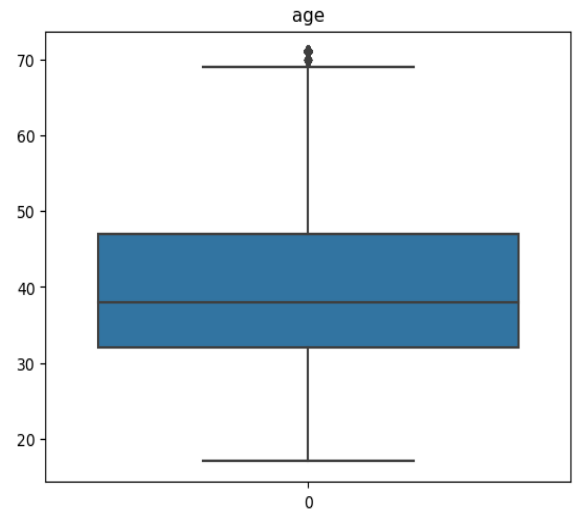


Fig13: Boxplot for age after removing outlier

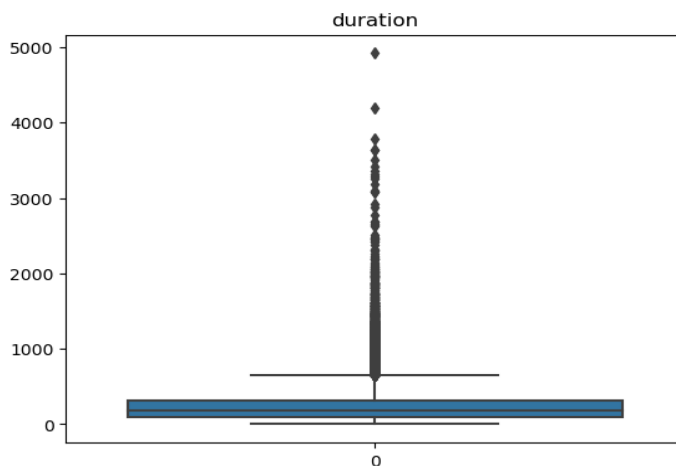


Fig14: Boxplot of duration before removing outlier

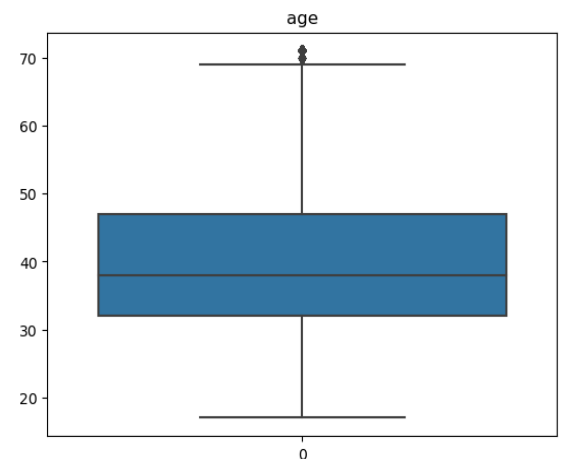


Fig 15: Boxplot of duration before removing outlier

2. Treatment for categorical variables

One-Hot Encoding

We applied One Hot Encoding method to categorical data with using get dummies function. One hot encoding is a process by which categorical variables are converted into matrix form with 1 and 0 values.

Data Normalization

We normalized our numerical data as data preparation. The goal of normalization is to change the values of numeric columns in the data set to use a common scale, without distorting differences in the ranges of values or losing information. We used Standard Scaler function.

Unknown Values

There isn't Null values but we have unknown values. We replaced unknown values with mode of the categorical variable.

Test-Train Data Split

First step is splitting data into two and selected randomly 70% of data as Training set and 30% of data as Test Set on Python.

```
## Tretment of categorical data with labels as unknown - populating it with Mode
labels_rem = ['job', 'marital', 'education', 'default', 'housing', 'loan']
for variables in labels_rem:
    df[variables] = np.where(df[variables]!='unknown', sts.mode(df[variables]), df[variables])
```

2.2 APPLYING CLASSIFICATION MODELS:

Logistic Regression

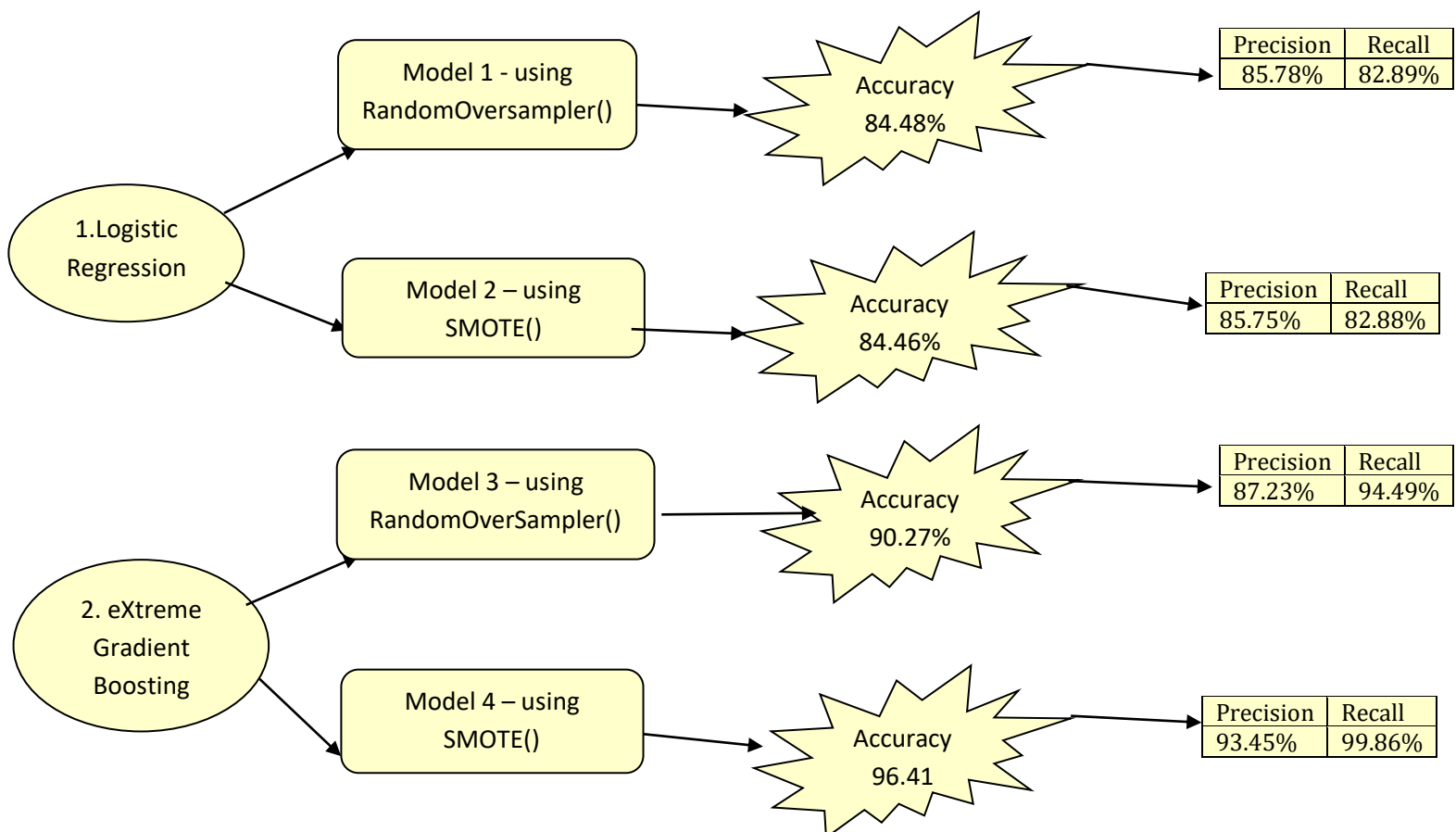
We applied Logistic Regression algorithm which is the most commonly used algorithm for solving all classification problems.

XGBOOST

We applied Extreme Gradient Boosting algorithm which is the most commonly used algorithm for solving all classification problems.

Model Building

Two different algorithms were used to built 4 different models with tuned hyperparameters which can predict whether the client will subscribe to a term deposit or not.



CHAPTER 3

RESULTS AND DISCUSSION

3.1 ACCURACY AND OTHER METRICS:

MODEL 4 USING SMOTE OVER SAMPLING AND USING XGBOOST IS GIVING VERY HIGH ACCURACY AND WE HAVE ALSO CREATED THE CONFUSION MATRIX FOR THE TEST DATASET. NOW WE WILL LOOK INTO THE PRECISION, RECALL, F1 SCORE AND ROC CURVE GENERATED FROM MODEL 4.

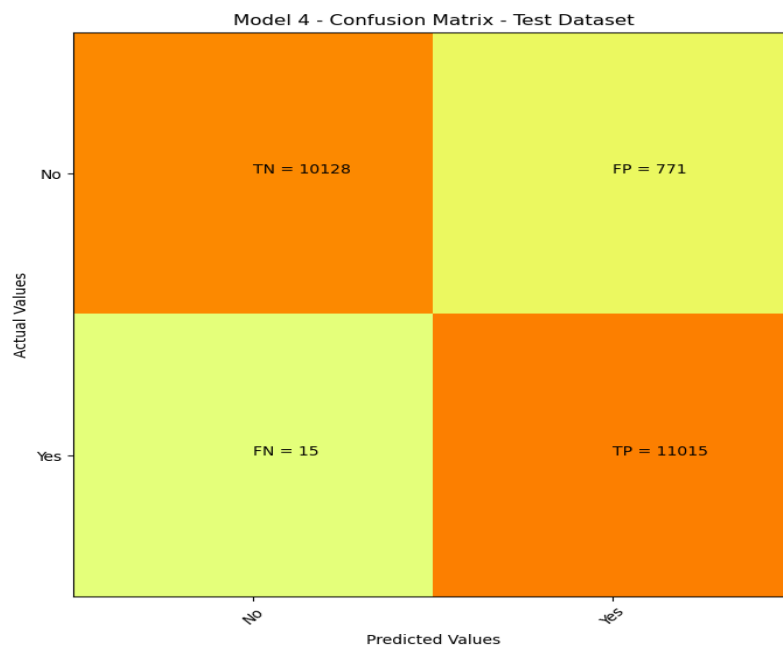


Fig15: CONFUSION MATRIX FOR XGBOOST USING SMOTE

TN <- true negative indicates actual and predicted values both are negative.

TP <- true positive indicates actual and predicted values both are positive.

FP <- false positive indicates actual value is no, but predicted value is wrongly classified as yes.

FN <- false negative indicates actual value is yes, but predicted value is wrongly classified as no.

Now accuracy is defined as - $(TP+TN)/(TN+TP+FP+FN)$

Precision is defined as - $TP/(TP+FN)$

Recall is defined as - $TP/(TP+FN)$

F1 score is defined as - $(TP+TN)/(TN+TP+FP+FN)$

Result

Accuracy - $(TN+TP)/(TN+TP+FP+FN)$: 0.9641570523051667

Precision - $TP/(TP+FP)$: 0.93458340403869

Recall - $TP/(TP+FN)$: 0.9986400725294651

F1 Score - $2*((P*R)/(P+R))$: 0.9655504908835905

ROC Curve

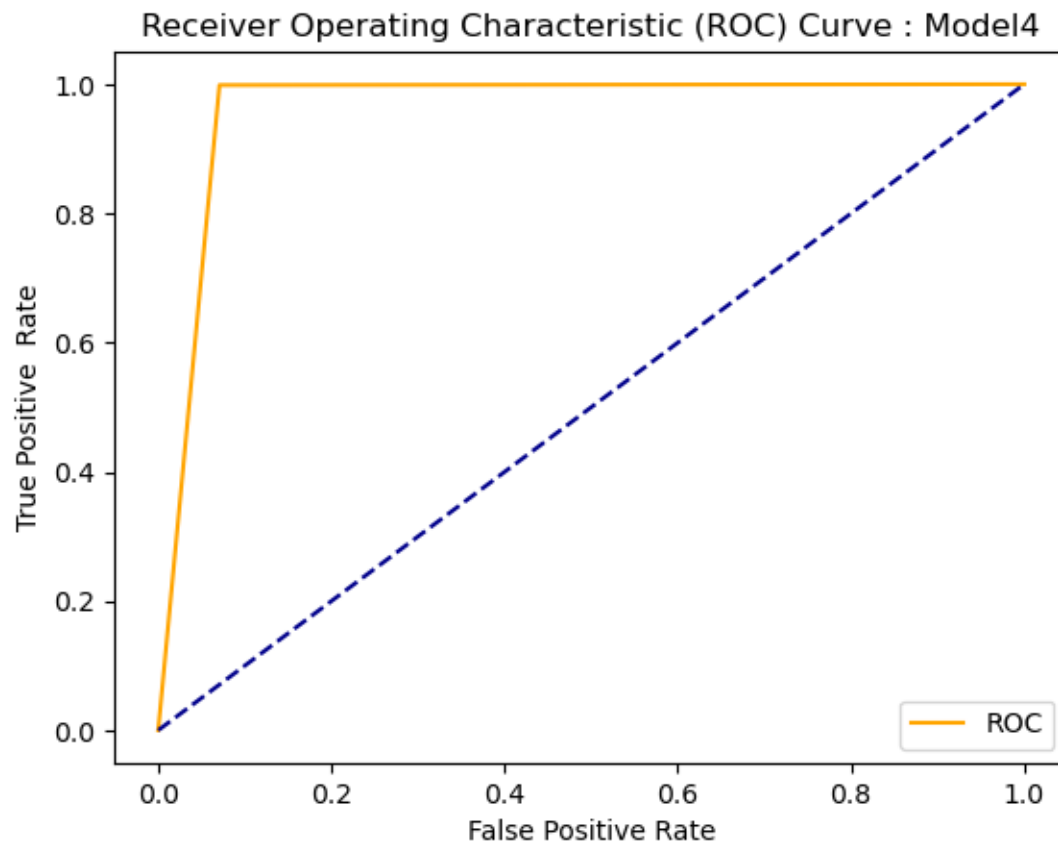


Fig16: ROC CURVE FOR XGBOOST USING SMOTE

Roc curve (Receiver Operating Characteristic curve) is the graph where we plot False positive rate along x axis and True positive rate along y axis.

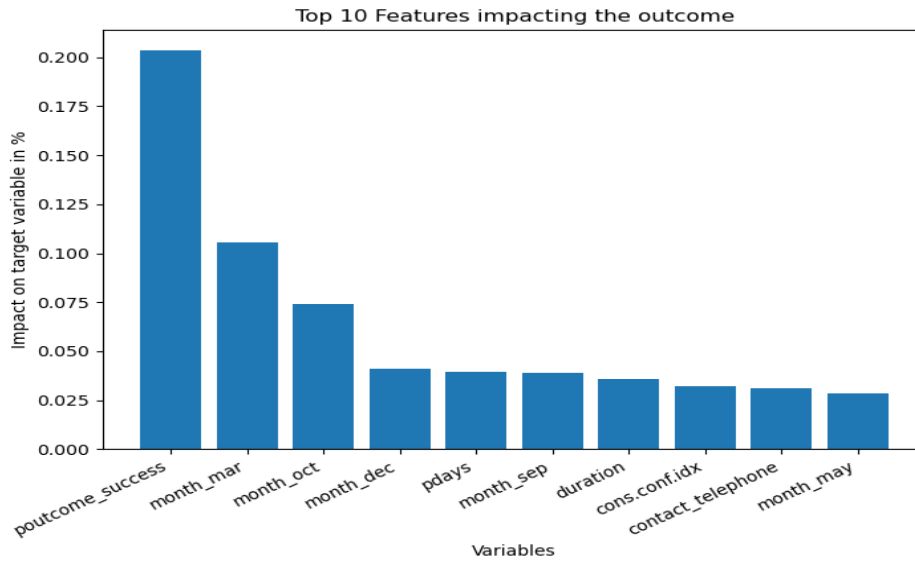


Fig17: Impact of the variables in %

These are the bar chart of the Impact of the top 10 variables which impacted the outcome.

3.2 ALTERNATIVE SOLUTIONS:

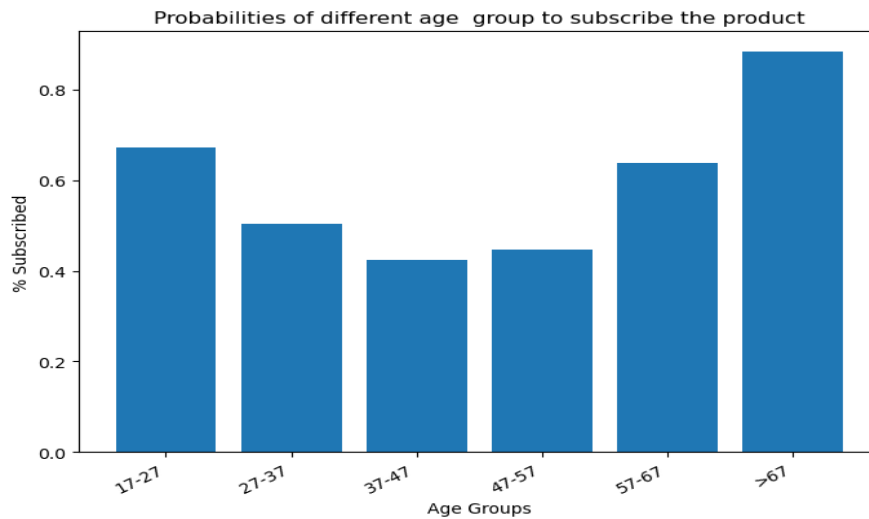


Fig18: probabilities of different age group to subscribe the product

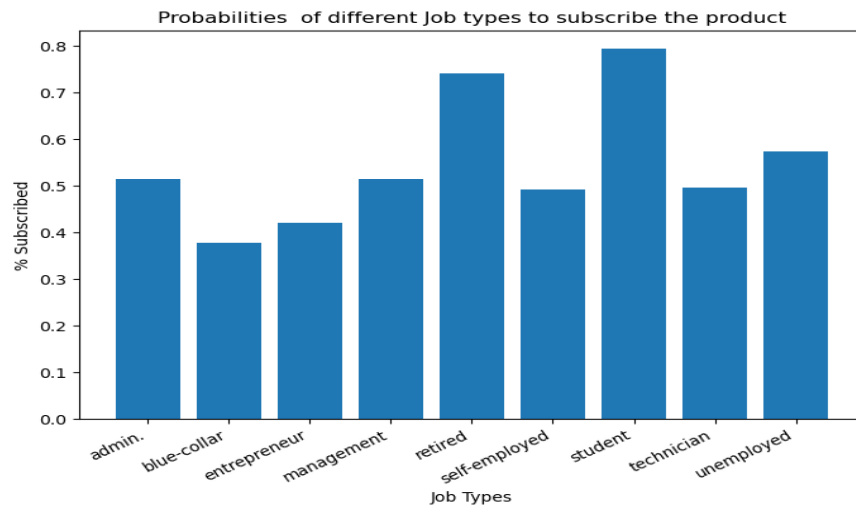


Fig19: probabilities of different job types to subscribe the product

As the above bar graphs Fig18. shows the clients more than 67 years of age are highly likely to subscribe the product ->85%.

From Fig19. we can see that the clients with profession retired and student are highly likely to subscribe the product >65%.

3.3 PROBABILITY OF SUBSCRIBING THE PRODUCT:

Generation of the probability scores for each of the records/clients whether subscribing the product or not

day_of_w	day_of_w	day_of_w	day_of_w	poutcome	poutcome	0	1
1	0	0	0	1	0	0.059727	0.940273
0	1	0	0	1	0	0.976573	0.023427
1	0	0	0	1	0	0.012787	0.987213
0	0	0	0	1	0	0.993912	0.006088
1	0	0	0	1	0	0.00764	0.99236
0	0	1	0	0	0	0.024338	0.975662
1	0	0	0	1	0	0.036346	0.963654
0	0	0	0	1	0	0.011489	0.988511
1	0	0	0	1	0	0.999109	0.000891
0	0	0	0	1	0	0.001882	0.998118
0	0	1	0	0	0	0.00517	0.99483
0	1	0	0	0	1	0.032439	0.967561
1	0	0	0	1	0	0.999976	2.36E-05
1	0	0	0	1	0	0.005905	0.994095
0	0	1	0	1	0	0.010892	0.989108

Fig20: probabilities score to subscribe the product for each clients

- Each row above represents a client
- Column 0 shows the probability scores of a client not subscribing the product
- Column 1 shows the probability scores of a client subscribing the product

CHAPTER 4

CONCLUSION

The culmination of the project marks a significant achievement in harnessing data analytics and Machine learning techniques to gain valuable insights into the Bank marketing procedure and through a meticulous and stepwise approach, the project has successfully achieved its objectives and delivered tangible outcomes for future marketing strategies. Let us elaborate on the conclusion of each step and tie them together:

- **Understanding of features & Problem statement:** The initial phase of the project involved understanding about the features, the outcome needed and the classification algorithms to choose.
- **Exploratory Data Analysis:** This part provided the detailed insight about the data and its relation with the outcome. we have performed univariate and multivariate analysis.
- **Outlier Detection and Data Cleaning:** This stage involves detecting the potential outliers and removing them and for categorical values removing unknown values to get higher accuracy.
- **Scaling:** This step involves encoding categorical variables with one-hot-encoding and removing features that are not needed. We also Scaled the other numerical features so the dataset can be used efficiently for building Machine Learning Models.
- **Hyperparameter Tuning:** Hyperparameter tuning is a crucial step in machine learning model development where we search for the best set of hyperparameters to optimize a model's performance. We used GridSearchCV for finding the best set of hyperparameters.

- **Model Building:** Extensive testing and feedback gathering ensured that the dashboard met the needs and expectations of its intended users. Iterative design updates addressed any identified usability issues, resulting in a user-friendly and intuitive analytical tool.
- **Alternate Solutions:** We have also shown how the age and profession can affect in subscription. Particularly clients having age more than 67 years and (retired and students) are very highly likely to subscribe.
- **Future Enhancements:** The completion of the project has already begun to make a positive impact on asset management and maintenance practices within the electrical distribution network. As stakeholders continue to leverage the insights provided by the dashboard, opportunities for further enhancements and refinements will emerge, driving continuous improvement in operational efficiency and reliability.

In conclusion, the successful completion of the project underscores the transformative potential of data analytics and machine learning in finding the best marketing strategies.

REFERENCES

- <https://www.kaggle.com/code/souritrabanerjee/eda-pandas-bank-marketing/input>
- <https://numpy.org/doc/stable/user/basics.html>
- https://scikit-learn.org/stable/user_guide.html
- <https://pandas.pydata.org/docs/>