`ORIE 5741 Final Project Report

—— Predict If the Client Will Subscribe a Term Deposit

Fei Liang, Yizhou Yu

May 2023

Background

Bank marketing is a philosophy about the bank's business strategy and the way to make money. It is the business behavior of commercial banks to achieve their business goals by designing and providing financial products and services that better meet market demand, and by organizing a variety of direct marketing campaigns to reach the target market with such products and services.

In the 1980s and 1990s, the main marketing model of banks was direct marketing. Since the 21st century, with the concept of precision marketing the importance of data for marketing promotion has been further highlighted, and data is widely used in the whole process of insight into customer behavior, correlation analysis and marketing interaction.

Therefore, in the field of banking and finance, using the right tools and models for data analysis can better provide informative suggestions for marketing activities and achieve intelligent financial positioning.

Problem Statement

When banks hold marketing campaigns, they aim to focus on customer needs and their overall satisfaction, while marketing their own products. Marketing campaigns are based on phone calls. Usually, bank marketers need to make multiple contacts with the same customer. Through marketing campaigns, banks can collect information about their customers in order to see if the product (bank term deposits) will be subscribed

The success of a marketing campaign can be judged by whether the product is recognized and subscribed by the customer this time, and through data analysis, we can predict the sales of the product in a given campaign.We wanted to solve the problem - Can we predict a client's likelihood of subscribing to a term deposit based on their demographic, financial, and contact information by using machine learning algorithms.

Enhancing our ability to predict which clients are more likely to subscribe to a term deposit will enable us to focus our marketing efforts on the right audience, reduce costs, and improve customer satisfaction. This project aligns with our enterprise's goals to increase efficiency and optimize resources.

Data Description

The data used for this project was customer data from a marketing campaign of a Portuguese bank. The data set consists of 41188 examples and 20 input variables related to clients' demographic and financial information, contact details, and social and economic context. The output variable is whether or not the client subscribed to a term deposit. Following is some description of variables(full description see Appendix T1):

Campaign	Int	Number of contacts performed during this campaign
Pdays	Int	Number of days that passed by after the client was last contacted from a previous campaign
Previous	Int	Number of contacts performed before this campaign and for this client
Poutcome	String	Outcome of the previous marketing campaign
emp.var.rate	Int	Employment variation rate
euribor3m	Int	Euribor 3 month rate

The dependent variable y has 11.3% subscription and 88.7% non-subscription, which is an unbalanced dataset. Meanwhile, the numerical variables mostly have high variation and different scales, we will deal with this later.

Model and Approach

Overview

The present study aims to predict whether a potential customer will subscribe to the term deposit product. To achieve this goal, the study is divided into several key stages. Firstly, the dataset is subject to Data Pre-processing procedures to ensure its readiness for further analysis. Following this, the Exploratory Data Analysis step is undertaken, delving deeper into the independent variables to gain insights into patterns and trends that may inform subsequent analysis. The Feature Selection stage then proceeds, using rigorous statistical tests to select the most relevant features that will form the input dataset for the machine learning models. The models are then trained and validated, utilizing a variety of metrics to compare their performance, with the best model ultimately selected. Finally, the selected model is applied to the testing dataset to generate the final prediction results. These steps will be presented in detail in this section.

Data Pre-processing

Traditionally, the Data Pre-processing stage involves dealing with missing values and outliers. However, in the present project, the dataset does not contain any missing values. Therefore, the focus shifts to detecting outliers. It is worth noting that some categorical variables have "unknown" values. Although this does not affect the initial understanding of the dataset, these values will be dealt with during the subsequent Exploratory Data Analysis and Feature Selection stages.

To get a general understanding of the dataset, it is important to examine the unique values for categorical variables and the summary statistics for numerical variables. This information is critical

in identifying potential patterns and trends in the data that could inform the subsequent analysis stages. Therefore, the first step in the Data Pre-processing stage involves examining these variables to identify any potential issues that need to be addressed before proceeding to the subsequent stages.

Variable	Age	Duration	Campaign	Pdays	Previous
Mean	40.02	258.29	2.57	962.48	0.173
STD	10.42	259.28	2.77	186.91	0.495

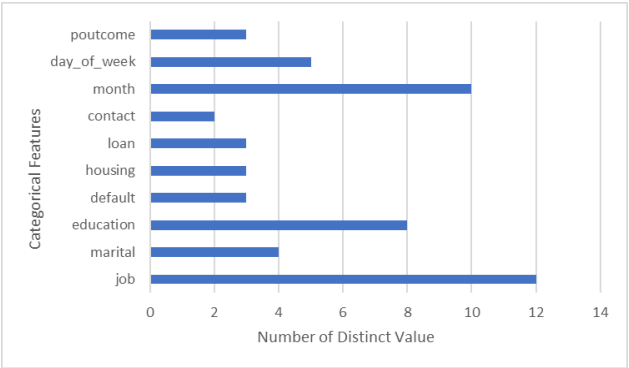
Variable	emp.var.rate	cons.price.idx	euribor3m	nr.employed
Mean	0.08	93.58	3.62	5167.04
STD	1.57	0.58	1.73	72.25

Table: Summary Statistics for Numerical Variables

From the Summary Statistics for Numerical Variables table, we noticed that variable “Duration”, “Campaign”, and “Previous” have relatively large deviations. It is also very important to remember that for variable “Pdays”, the value “999” means the customer was not contacted before. So this value needs to be deal with carefully in the next step. We also find out that the three rate variables have only a few unique values. Even if they look lika a numerical variable, we can still treat them as categorical variables.

Upon examining the Summary Statistics for Numerical Variables table, it was observed that variables such as "Duration," "Campaign," and "Previous" exhibit relatively large deviations. Additionally, it is critical to note that a value of "999" for the variable "Pdays" indicates that the customer was not contacted previously. This value warrants careful consideration during subsequent analysis stages.

Furthermore, the three rate variables demonstrate only a few unique values. Despite their numerical appearance, it is appropriate to treat them as categorical variables.



Graph: Number of Unique Values for Categorical Variables

The graph depicting the number of unique values for each categorical variable suggests that these variables are well-controlled, with no apparent issues that require special attention.

With the initial analysis results, a general picture of the dataset is depicted. Exploratory Data Analysis will provide a more detailed look into each variable and their relationship.

Exploratory Data Analysis

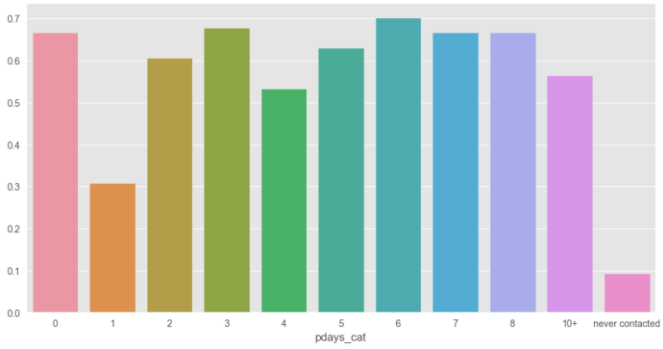
In this section, the relationship between each independent variable and the dependent variable was analyzed to gain a comprehensive understanding of the dataset. To facilitate this analysis, visualizations were utilized to identify any patterns and trends that may be present. These visualizations provided a deeper insight into the dataset and helped in uncovering the underlying relationships among the variables.

During the previous stage of analysis, it was observed that a significant proportion of the categorical variables in the dataset contained "unknown" values. To optimize the utilization of the dataset and retain as much information as possible, different strategies were employed to address the different types of "unknown" values.

For the first type of "unknown" variable, it was noted that the customer may choose not to disclose certain information due to personal reasons. To preserve this information, the "unknown" value was retained and incorporated into the subsequent analysis stages. This approach was adopted for variables such as "job", "education", "default", "housing", and "loan".

The second type of "unknown" variable was characterized by a small number of "unknown" values, which likely resulted from data collection issues. For these variables, imputation was performed by replacing the "unknown" value with the most frequent value. The "marital" variable was handled in this way.

Finally, particular attention was given to the "Pdays" variable, which takes the value "999" when the customer has not been contacted previously. To address this variable, a separate variable was created to capture the information. Values below 10 were treated as a separate category, while values above 10 but not equal to 999 were combined into another category. The value 999 was assigned to a separate category named "never contacted".



Graph: Subscribe Rate of Pdays variable translated to categorical

After handling the "unknown" values, we proceeded to investigate each variable's relationship with the dependent variable and used visualizations to aid our analysis. We identified variables

that appeared to have a stronger association with the dependent variable, while those with less apparent correlations were placed in an "ambiguous" list for further evaluation during the modeling step. Additionally, we examined the correlation among independent variables. As a result, we found that "emp.var.rate," "cons.price.idx," "euribor3m," and "nr.employed" had high correlations with one another. After further examination, we determined that the first three variables could be represented by one variable. We took note of this and planned to handle these variables with caution in the Feature Selection step.

Feature Selection

During the Feature Selection process, we aimed to select the most relevant variables for our model. Firstly, we removed the variables "emp.var.rate" and "cons.price.idx" due to their high correlation with the variable "euribor3m", which was retained as it had a stronger relationship with the dependent variable. Next, we analyzed the variables that were previously classified as "ambiguous" in the Exploratory Data Analysis step, which included "housing", "loan", "day_of_week", and 'age'. We performed a chi-square analysis between these categorical variables and the dependent variable. The results showed that the p-values of "loan" and "housing" were greater than 0.05, indicating that there was insufficient evidence to reject the null hypothesis that these variables were not significantly correlated with the dependent variable. Therefore, we decided to exclude these variables from our model. For numerical variables, we tested the f-score between them and the dependent variable. The results indicated that all of the variables had a p-value less than 0.05, suggesting that they were statistically significant and should be retained in the model.

Once the variables were selected for the model, we proceeded to transform the categorical variables using the one-hot encoding method. Additionally, as noted during the pre-processing step, the numerical variables varied in scale and range. In order to keep them in the same scale and facilitate the modeling process, we standardized the numerical variables. This step is particularly important for models that are not scaling invariant.

We first combined the two transformed datasets and added an additional offset column to create a final dataset with a dimension of 41158*70. Next, we divided this dataset into training, validation, and testing datasets in a 6:2:2 ratio. Since our dataset is highly imbalanced and we are more interested in the records that will subscribe to our product, which only represents 11.3% of the total data points, we explored methods to improve the model's performance on this unbalanced dataset. We discovered that oversampling and undersampling techniques are effective. (Shelke, 2017) In oversampling, additional sample points belonging to the minor category are generated, while in undersampling, some of the points in the majority category are removed. We utilized SMOTE and ENN methods to address this issue.(Mohammed, 2020)

Model Training and Validation

This section involved the selection of six distinct machine learning models, which were trained using the training dataset. The next step involved evaluating the performance of these models using the validation dataset. The results of the model performance are presented in the table below.

Model	Logistic Regression	Decision Tree	Random Forest	Gradient Boosting	XGBoost	SVM	ANN
F1-score on Training Set	0.92/0.94	1.00/1.00	1.00/1.00	0.96/0.97	0.99/1.00	0.96/0.97	0.98/0.98
F1-score on Validation Set	0.89/0.55	0.93/0.59	0.93/0.61	0.92/0.59	0.93/0.62	0.91/0.56	0.94/0.52
Weighted F1-score on Validation Set	0.85	0.89	0.89	0.88	0.90	0.87	0.89

Table: Model Results

The results above indicate that the training dataset has a higher score than the validation dataset. However, this is not necessarily a sign of overfitting, as we applied oversampling and undersampling only to the training dataset, unlike what is commonly taught in class. We then selected the top three performing models, XGBoost, Random Forest, and Decision Tree, and fine-tuned their hyperparameters using the grid search method. Ultimately, XGBoost was the best performing model with the following hyperparameters: (colsample_bytree: 0.75, gamma: 0.5, max_depth: 6, min_child_weight: 1, subsample: 0.8). XGBoost, which stands for eXtreme Gradient Boosting, is highly efficient and scalable, and is capable of handling large-scale data problems. It uses a more regularized model formalization to control overfitting, leading to improved performance. (Chen, 2016; Morde, 2019)

Testing and Results

In this project, we embarked on a journey to predict the likelihood of a client subscribing to a term deposit. We implemented a Train-Validation-Test split and applied several machine learning models, ultimately selecting XGBoost, which achieved the highest weighted F1-score.

Model Performance

The XGBoost model demonstrated robust performance on the test data, with precision scores of 0.98 for class 0 (clients not subscribing to a term deposit) and 0.5 for class 1 (clients subscribing to a term deposit). It showed high recall rates of 0.89 for class 0 and 0.87 for class 1. F1-scores were 0.93 for class 0 and 0.63 for class 1, leading to a commendable weighted F1-score of 0.90.

The ROC curve(see Appendix G1) further confirmed the model's performance with an AUC of 0.878 for the test data, 0.870 for the validation data, and 0.995 for the train data.

A key factor behind our choice of the XGBoost model was the strong performance of the model in terms of recall. A high recall implies that the model has a strong ability to correctly identify clients who are likely to subscribe to a term deposit. This can help the bank to preserve most of its prospective customers.

Additionally, by narrowing down the bank's sets of target customers, we can significantly reduce costs associated with broad, less-targeted marketing campaigns. Instead, efforts and resources can be concentrated on the clients who are most likely to respond positively.

Confidence in Results and Production Readiness

The high performance of the model on both the validation and test data sets provides confidence in the results. The high AUC score of the ROC curve indicates that the model has a strong discriminatory capacity between clients who will subscribe to a term deposit and those who will not.

However, we need to be mindful of the model's precision for class 1 (subscribed to a term deposit) at 0.5. This suggests that the model may sometimes incorrectly predict that a client will subscribe to a term deposit. The effect of this limitation on business outcomes should be carefully evaluated.

Despite this, considering the high recall and the nature of the problem, we would be willing to use these results in a production environment to inform the bank's marketing strategies. The potential cost savings and efficiencies gained by targeting likely subscribers more accurately are significant benefits that outweigh the noted limitation.

Future Work

Future work includes conducting a thorough analysis of our preferred customers' characteristics to gain deeper insights into their specific needs and preferences. We can also apply cluster analysis algorithms to guide the bank's sales staff's future working focus, ensuring resources are utilized optimally.

Furthermore, examining optimal call frequency and connection dates can help refine our customer interaction strategies, leading to potentially higher subscription rates to the term deposit.

In conclusion, we believe that the results of this project provide valuable insights that can positively influence our company's decision-making process, with the potential to substantially improve marketing outcomes and customer satisfaction.

Discussion

Weapon of Math Destruction

The term "Weapon of Math Destruction" was coined by mathematician and data scientist Cathy O'Neil to describe algorithms or mathematical models that can have harmful or destructive effects on individuals or society as a whole. These models are typically used in contexts such as finance, employment, insurance, criminal justice, and education. A WMD often arises when a mathematical model or algorithm is flawed or misapplied, leading to biased or unfair outcomes.

Let's discuss whether our project might produce a Weapon of Math Destruction in the following three ways：

1. Are outcomes hard to measure?

The outcome in this project is whether or not a client will subscribe to a term deposit. This can be directly measured by observing the client's subsequent actions after receiving marketing communications. If the client subscribes to a term deposit, the outcome is a success; if not, it's a failure. Therefore, the outcome is not hard to measure in this case, which reduces the risk of it being a WMD.

2. Could its predictions harm anyone?
   The predictions themselves, which aim to identify the likelihood of a client subscribing to a term deposit, don't seem to be harmful on the surface. However, potential harm could arise depending on how these predictions are used. For example, if the model results in certain demographic groups being unfairly targeted or excluded from marketing campaigns, it could contribute to discrimination or unequal access to banking services. This would make the model a WMD. It's essential to ensure that the model is fair and doesn't disproportionately disadvantage or discriminate against certain client groups.

3. Could it create a feedback loop?
   A feedback loop could potentially occur if the results of the model are used to continually refine the target audience for the bank's marketing campaigns. If the model predicts that certain clients are unlikely to subscribe to a term deposit, and these clients are then excluded from future marketing efforts, they may indeed become less likely to subscribe simply due to lack of exposure to the offer. This could reinforce the model's predictions and result in a self-fulfilling prophecy, which could unfairly disadvantage certain clients. This scenario would classify the model as a WMD.

In conclusion, while our project doesn't inherently seem to be a WMD, there are potential risks and ethical considerations to keep in mind. Ensuring fairness, avoiding discrimination, and preventing harmful feedback loops are key to ensuring our model doesn't turn into a WMD.

Fairness

Fairness is an essential criterion to consider when choosing a model for our project, especially when we're dealing with people's financial data and decisions that could potentially impact their access to financial services. (see Appendix D1).

Reference

Morde, V. (2019, April 7). XGBoost Algorithm: Long May She Reign! Towards Data Science. https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d. Accessed 20 Aug 2019.

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). ACM

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. In 2020 11th International Conference on Information and Communication Systems (ICICS) (pp. 243-248). Irbid, Jordan. https://doi.org/10.1109/ICICS49469.2020.239556.

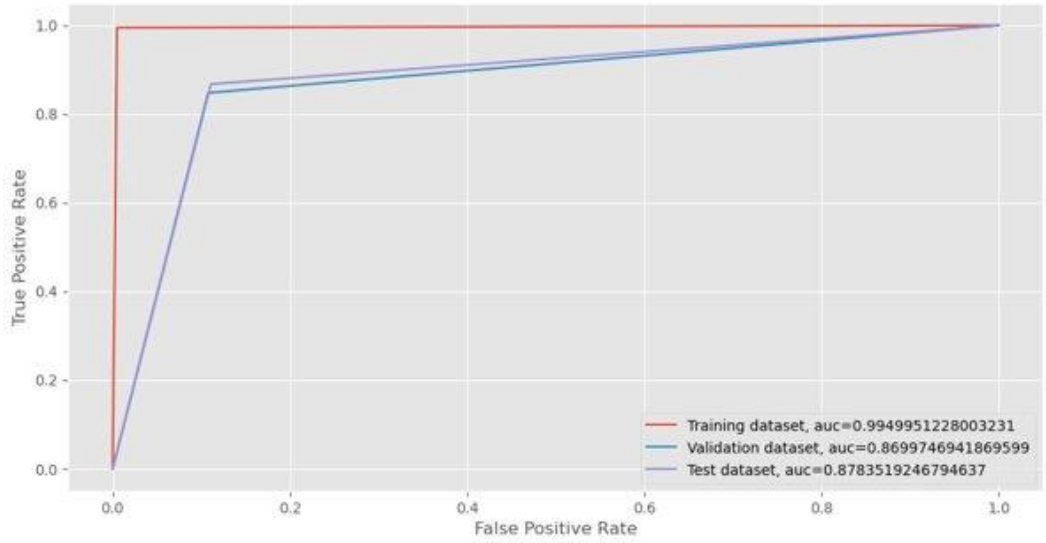
Shelke, M.M., Deshmukh, D.P., & Shandilya, V.K. (2017). A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique.

Appendix:

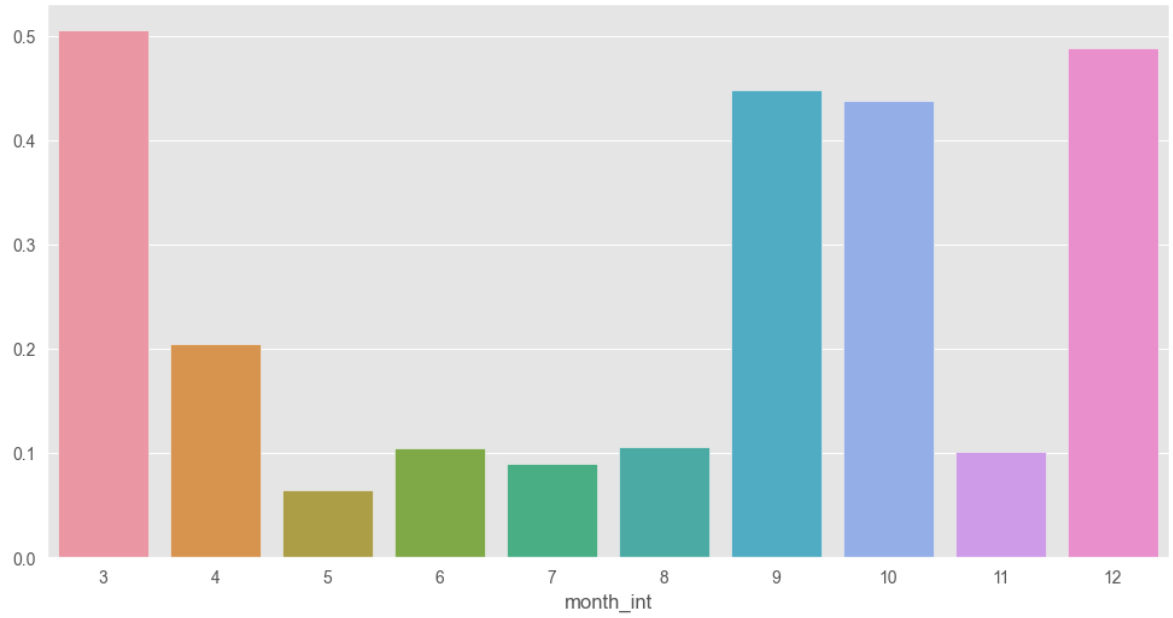
T1. Description Table of DataVariables:

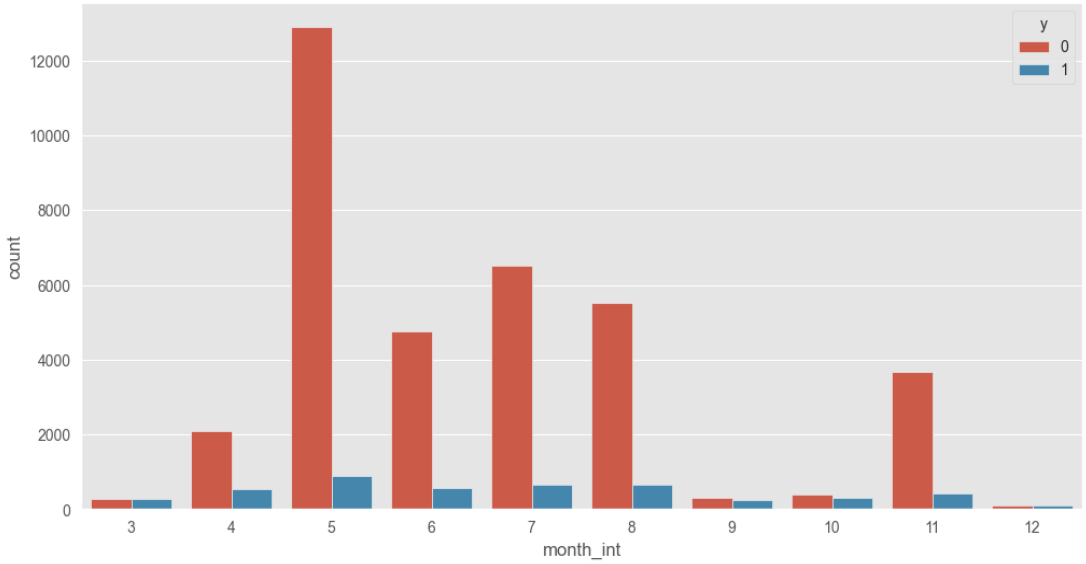
1	age	Int	Age of client
2	job	String	Type of job
3	marital	String	Marital status
4	education	String	Education level
5	default	String	Has credit in default?
6	housing	String	Have a housing loan?
7	loan	String	Have a personal loan
8	contact	String	Contact communication type
9	month	String	Last contact month of year
10	day_of_week	String	Last contact day of the week
11	duration	Int	Last contact duration
12	campaign	Int	Number of contacts performed during this campaign
13	pdays	Int	Number of days that passed by after the client was last contacted from a previous campaign
14	previous	Int	Number of contacts performed before this campaign and for this client
15	poutcome	String	Outcome of the previous marketing campaign
16	emp.var.rate	Int	Employment variation rate
17	cons.price.idx	Int	Consumer price index
18	cons.conf.idx	Int	Consumer confidence index
19	euribor3m	Int	Euribor 3 month rate
20	nr.employed	Int	Number of employees
21	y	Int	Has the client subscribed to a term deposit?

G1. ROC Curve of XGBoost Testing Data:



G2: Example of Count Plot and Subscribe Rate Plot





D1. Discussion of Fairness:
Fairness is an essential criterion to consider when choosing a model for our project, especially when we're dealing with people's financial data and decisions that could potentially impact their access to financial services. Following are reasons why I think fairness matters:

1. Prevent Discrimination: If a model inadvertently learns to make decisions based on sensitive attributes like race, gender, or age (which could correlate with some of the variables in your dataset), it could lead to discriminatory practices. For example, if the model learns that older people are less likely to subscribe to a term deposit and then starts to exclude them from being targeted for marketing campaigns, this could be seen as age discrimination.

2. Public Perception and Trust: A model that is perceived as unfair can damage the bank's reputation and undermine public trust in its services.

3. Long-term Customer Relationship: Unfair practices can alienate customers and potentially lead to loss of business in the long run. On the other hand, fairness can contribute to customer satisfaction and loyalty.

To ensure fairness, it's important to test the model for potential bias during the model selection and validation process. This could involve techniques like disparate impact analysis (to identify whether the model's decisions disproportionately harm a certain group) or fairness-aware machine learning algorithms that explicitly aim to minimize discrimination.