Project Scope Statement

BANK MARKETING DATASET

Zezhen Liu, Zeju Li, Zonghong Yu
DSAN6700-GROUP-09
GEORGETOWN UNIVERSITY |

Table of Contents

1 OVERVIEW

The banking industry, historically rich in data, has consistently leveraged vast amounts of information to understand, predict, and enhance customer behavior (Zhuang, Yao, & Liu, 2018). Traditional models primarily focus on individual banking transactions - deposits, withdrawals, loans - forming a granular understanding of a customer's financial habits. However, in the evolving world of finance, there's a growing recognition: that personal financial decisions aren't made in a vacuum. They're often deeply intertwined with the broader socio-economic landscape.

The "Bank Marketing (with social/economic context)" dataset from the UCI repository exemplifies this shift in perspective. Beyond the conventional attributes like age, job, and education, it incorporates socio-economic factors, providing a more holistic view of potential bank clients. The added dimensions promise to unveil deeper layers of their decision-making processes, especially in areas like telemarketing, which relies heavily on understanding client motivations.

Central to our exploration is a pivotal question: ***Can a client's decision to subscribe to a term deposit be better predicted by harmonizing both individual banking data and overarching economic indicators?*** This paper endeavors to unravel this, hypothesizing that a comprehensive analysis, blending these two realms, will yield a more refined predictive accuracy. Our aim is to bridge the gap between micro-level banking analytics and macro-level economic insights, presenting a unified, more insightful lens to view and predict banking behaviors.

1.1 Project Background and Description

What is so interesting about this problem?

The banking sector, despite its long-standing existence, remains one of the most dynamic industries, adapting to new technologies and methodologies. This intersection of banking behavior with larger economic indicators is fascinating to the team as it offers a multifaceted view of customer decisions. We also find it interesting for it challenges the traditional boundaries of banking data analytics by suggesting that perhaps, a client's decision to subscribe to a term deposit is not just influenced by their personal banking history but also by the economic environment around them.

What are the proposed benefits of the solution (why do we need to solve this)?

The implications of an enhanced predictive model are manifold. For banks, this could mean a more targeted and efficient telemarketing strategy, leading to cost savings and increased revenue. For policymakers and economic researchers, understanding these correlations can offer insights into how national economic indicators trickle down and influence individual financial decisions.

What problem type are we evaluating (classification? regression, or something else?)

At its core, we are tackling a classification problem. *The objective is binary: predicting if a client will or will not subscribe to a term deposit.* However, the factors influencing this decision, as we hypothesize, could be multivariate, encompassing personal banking details and broader economic indicators.

1.2.0 Project Scope

The overarching goal is to navigate through the "Bank Marketing (with social/economic context)" dataset, understand its intricacies, and formulate predictive models that can capitalize on the enriched data. The project will employ a combination of generative and non-generative methods, further enhancing them with gradient boosting and bagging techniques. A significant focus will be on hyperparameter tuning, generative methods, and some other models, comparing the models to select the most effective one. The deliverable is not just a model but an understanding of the interactions between various attributes and their collective influence on a client's decision.

1.2.1 *Step 1. List of Project Tasks*

Task ID#	**Task to be completed**	**For Deliverable #**
1	Submit Project Charter: The outline of the project	1
2	Ensemble trained model with preliminary results of the training and testing.	2
3	Project Scope Report: The analysis and results from the model	3
4	Further analysis with grid search and app development	4
5	Final report of the project	5

1.2.2 *Step 2. Out of Scope*

This project **will NOT accomplish or include** the following:	Deep learning or neural network approaches might not be the best fit and are considered out of scope. Using the 'duration' attribute for real-time predictions is also out of scope due to its post-call nature.

Given the constraints of the dataset and the problem, deep learning or neural network approaches might not be the best fit and are considered out of scope. Using the 'duration' attribute for real-time predictions is also out of scope due to its post-call nature.

1.2.3 Step 3. Methodologies

The exploration will start with generative and non-generative methods as foundational analytical tools. As we delve deeper, techniques like gradient boosting and bagging will be employed. These methods will be tuned, with various parameters like n_estimator and learning rate being adjusted to optimize performance. Decision trees might also be explored for their interpretability and to understand feature importance.

1.1 Project Methods	1.2 Gradient Boost Evaluation	1.3 Bagging Evaluation
Traditional Exploration and Analysis: Focusing on predicting the likelihood of a client subscribing to a term deposit.	Gradient Boost Evaluation: Implement the gradient boosting algorithm on the dataset. Parameters will be varied for optimization, with n_estimator values set as [10, 25, 50, 75, 100, 125, 150] and learning rates ranging from 0.1 to 1.0.	Bagging Evaluation: Subsequent to gradient boosting, the bagging algorithm will be employed on the dataset. Similar to before, parameters will be varied, aiming to discern the most efficacious settings
1.1.1. Preliminary visual exploration and analysis of the data set (exploratory data analysis): Dive deep into the dataset using Exploratory Data Analysis (EDA) to visualize attributes like age distribution, education levels, and economic indicators. Histograms, scatter plots, or heatmap might play a pivotal role in this phase.	1.2.1. Result Comparison: After model training, a comparative analysis of the results across varied n_estimator and learning rate settings will be undertaken. The optimal parameter set will be identified and its implications for predicting a client's decision will be discussed.	Repeat 1.1-1.2 for the Non-Generative methods.
1.1.2. Propose a possible outcome of the analysis: Based on preliminary insights, we might hypothesize, for instance, that clients within a certain age range, coupled with specific economic conditions, are more likely to subscribe to a term deposit.	1.2.2. Feature Importance: After model training and comparison, a feature importance plot will be generated to provide analysis and result for us to determine which feature is relatively crucial.	Hyperparameter tuning and analysis will be performed based on the results of 1 and 2.

1.2.4 *Step 4. Project Deliverable*

Deliverable ID#	Description
1	Project Charter
2	Ensemble trained model with preliminary results of the training and testing.
3	Project Scope Report
4	Further modeling analysis with gridsearch and app development.
5	Final report of the project in addition to the cloud (heroku) deployed link.

To be more specific, the first deliverable aims to allow readers to quickly understand the problem, expected outcomes, proposed solution, and high-level approach. For the second deliverable, models will be conducted by performing the following steps: Exploratory Data Analysis; Training and testing most of the algorithms over the given dataset; Using the best algorithms to generate a stacked model; and Providing feature importance figures based on the results. The third step is to make an analysis and write a report based on the ensemble-trained model with preliminary results of the training and testing.

General Deliverables	Generative Methods-Based Analysis of the dataset	More Evaluation
1.1.1. Data definition and normalization practices: The dataset offers a detailed description of each feature. It comprises attributes like 'age', 'job', 'education', and socio-economic indicators. Each attribute provides a facet of the client's profile and is instrumental in predicting if they'd subscribe to a term deposit. However, raw data often requires preprocessing. Normalization techniques, such as Min-Max scaling, will ensure that these features are on a similar scale, aiding in better model training and prediction.	2.1.1. Preliminary visual exploration and analysis of the data set (exploratory data analysis): Using visual tools like histograms, scatter plots, and heatmaps, we'll decipher patterns, correlations, and outliers. This visual exploration aids in understanding feature distributions and their relationships.	2.2.1 Use gradient boost to evaluate the dataset. Alter the parameters for the boosting procedure (for n_estimator use [10, 25, 50, 75, 100, 125, 150]; for learning rate use [0.1, till 1.0]).
1.1.2. Feature definition and training models: Features like 'education', 'job', and 'housing' provide insights into the socio-economic standing of the client. Moreover, broader economic indicators might reflect the overall economic health and its potential impact on individual financial decisions. After preprocessing, these features will serve as inputs to our machine-learning models. Models like Decision Trees, Logistic Regression, and Support Vector Machines might be employed initially.	2.1.2. Propose a possible outcome of the analysis: From the preliminary analysis, we might discover that variables such as 'education' and certain economic indicators have strong correlations with the likelihood of a client subscribing to a term deposit.	2.2.2. Compare the results: Upon training, models will be evaluated, and a comparative analysis will shed light on the best parameter combinations. This will be pivotal in understanding the most influential features and their impact on the outcome.
1.1.3. Processing pipelines and optimization methods: The analysis will follow a structured pipeline: data preprocessing, feature engineering, model training, and evaluation. Preliminary models provide a baseline. However, ensemble methods, such as Gradient Boosting and Bagging, will be integrated to enhance prediction accuracy. These techniques bring the benefit of pooling knowledge from multiple models, ensuring robust predictions.	1.2.2. Feature Importance: After model training and comparison, a feature importance plot will be generated to provide analysis and result for us to determine which feature is relatively crucial.	Hyperparameter tuning and analysis will be performed based on the results of 1 and 2.

1.2.5 *Step 5. Project Assumptions*

#	Assumption
1	The economic indicators and the bank marketing data will have affected the clients' decision to subscribe to a term deposit.

1.2.6 *Step 6. Project Constraints & Success Criteria*

The dataset, while rich, comes with its limitations. Categorical attributes with "unknown" labels can introduce ambiguity. The "duration" attribute, which holds significant predictive power, presents a paradox: while it's crucial for prediction, in a real-world scenario, its value isn't known until after a call, rendering it unusable for proactive predictions. Also, the dataset is a snapshot from May 2008 to November 2010, which might not reflect current trends or economic conditions.	A model that can predict with high accuracy. In addition, the model can show demonstrable evidence that the integration of social and economic indicators enhances prediction accuracy. A thorough understanding of which attributes hold the most predictive power, offering insights into client behavior.

2 Analysis of the dataset and Trained Model

With the bank marketing dataset that we obtained from the UCI repository, we can train different models to get the results such that better models can be evaluated and compared. This modeling section involves data handling such as data cleaning, exploratory analysis, and feature selections. By understanding the dataset, we can further the process and make an analysis to train the model more effectively. Based on the results from models, the decision-making process of customers regarding term deposits can be determined by getting the feature importance. We can acquire deep insights into these factors.

In addition, this study delves into a comprehensive dataset detailing customer interactions and their outcomes related to term deposits. To start with data preprocessing to ensure that the data is fit and balanced for analysis. In this way, we can have a basic understanding of correlations and relationships between features. Then a series of classification models are employed. Based on the comparison, the stack model with base and meta-models will combine the predictive capabilities of several models for evaluation. Through this, we aim to predict customer behavior regarding term deposits and gain insight into the features that affect the decision.

2.0 Data Preprocessing

After loading the raw dataset, the data preprocessing step is crucial for us to prepare for further analysis. Since the dataset contains imbalanced columns, we use one-hot encoding to create dummy variables for each categorical column and drop the unknown columns. Then, we drop out some redundant columns. The primary objective of preprocessing is to transform the data into a format that will be more easily and effectively processed for the purpose at hand. By making the features more understandable, we can generate results that can help the organization understand and gain insight more effectively. After the transformation, we can utilize exploratory analysis and feature selection to prepare for the modeling process.

2.0.1 Exploratory Analysis and Visualization

During the data preprocessing step, we need to perform exploratory analysis. The following table shows the basic information of the cleaned dataset. The dataset does not have any missing values.

Table 1 Summary of the Variables

Parameters	age	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
count	41188	41188	41188	41188	41188	41188	41188	41188	41188	41188
mean	40.024	2.568	962.475	0.173	0.0819	93.576	-40.502	3.621	5167.0359	0.113
std	10.421	2.770	186.911	0.495	1.571	0.579	4.628	1.734	72.252	0.316
min	17	1	0	0.000	-3.400	92.201	-50.800	0.634	4963.600	0.000
25%	32	1	999	0.000	-1.800	93.075	-42.700	1.344	5099.100	0.000
50%	38	2	999	0.000	1.100	93.749	-41.800	4.857	5191.000	0.000
75%	47	3	999	0.000	1.400	93.994	-36.400	4.961	5228.100	0.000
max	98	56	999	7.000	1.400	94.767	-26.900	5.045	5228.100	1.000

Then, a correlation matrix is used to represent pairwise correlation coefficients between variables in the dataset to have a visualization of the correlations among features. This heatmap aims to provide us with a view of the relationships between variables. In addition, it can help the organization to have a clear and basic understanding of the relationship between features which can also let us first have a clue about which features could possibly be the cause to affect the subscription of term deposit.

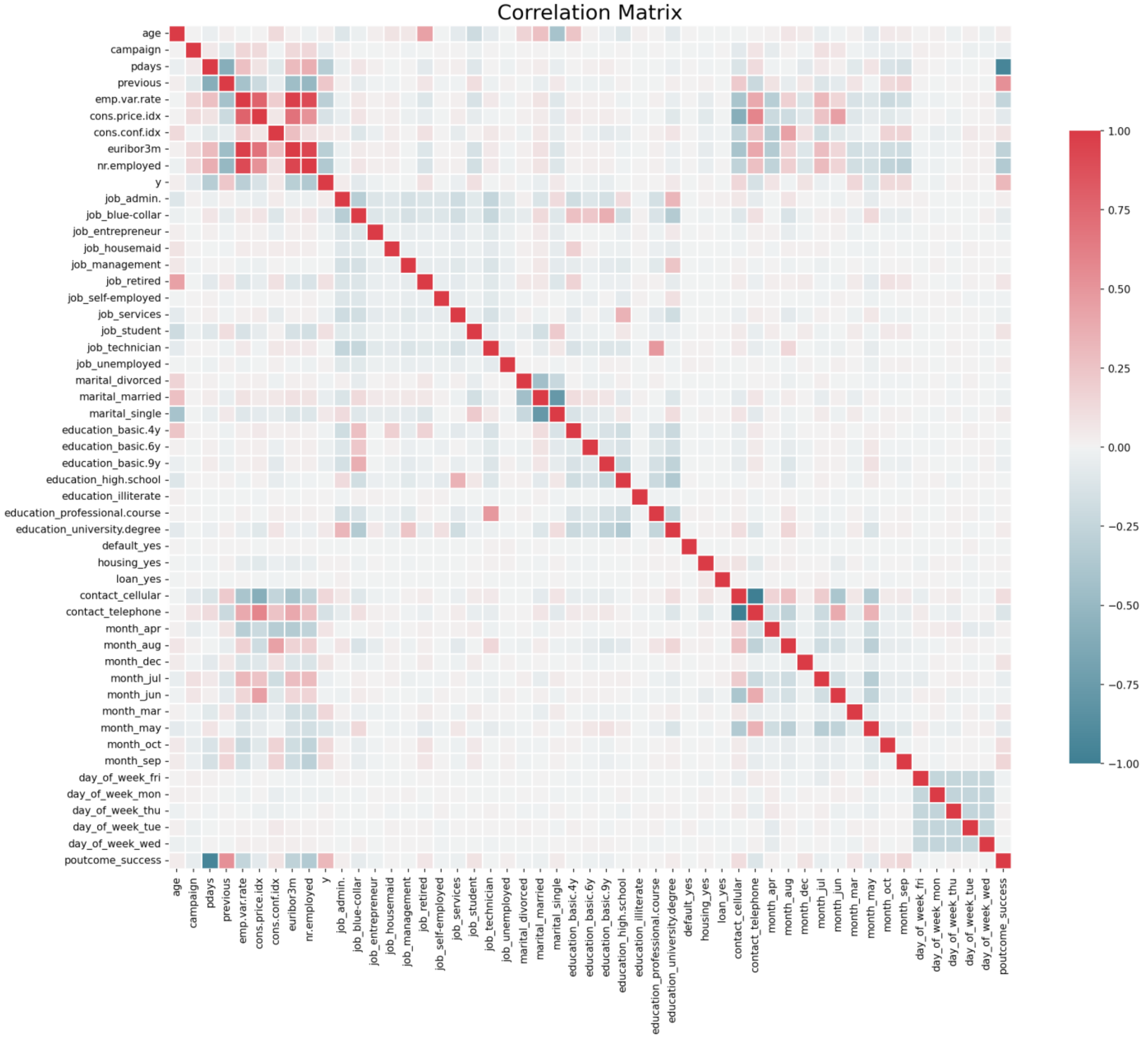


Figure 1: Correlation Matrix of The Cleaned Dataset

Based on Figure 1, we can see that there are relatively strong correlations among the following variables such as No. of Employees (nr.employed); Euribor 3M Rate (euribor3m); number of contacts in the current campaign (campaign); and Consumer Confidence Index (cons.conf.idx). In addition, some variables, particularly those related to contact (telephone/cellular) and month, exhibit blue shades indicating negative correlations. For instance, the use of cellular contact might be negatively correlated with certain months, suggesting that during those months, telephone contact might be more prevalent.

2.0.2 Feature Selection

Based on the previous steps of exploratory data analysis, we can see that there are some features that do not have certain correlations. Therefore, feature selection is an essential step for us to select the features to improve the efficiency and effectiveness of a predictive model. Specifically, we do feature selection on the scaled training set to prevent data leakage. Computes the K-best between each feature and the target variable for classification tasks, reducing features to half of all features (top 25) to decrease the dimensionality. This significant reduction helps not only speed up the model training process but also potentially enhances generalization by reducing the chances of overfitting which will save future operation costs for our organization.

2.1 Model Selection: Base Model & Meta Model

After the data preprocessing, we utilized different models to first try out the classification process such that we could choose the baseline model and the meta-model based on the initial performance. The utilizations of models are the following: Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVC), Gaussian Naive Bayes, Random Forest, Bagging, Gradient Boosting, and XGBoost. By comparing the performance of the model accuracy with box plot visualization, we can further make an analysis with base and meta models.

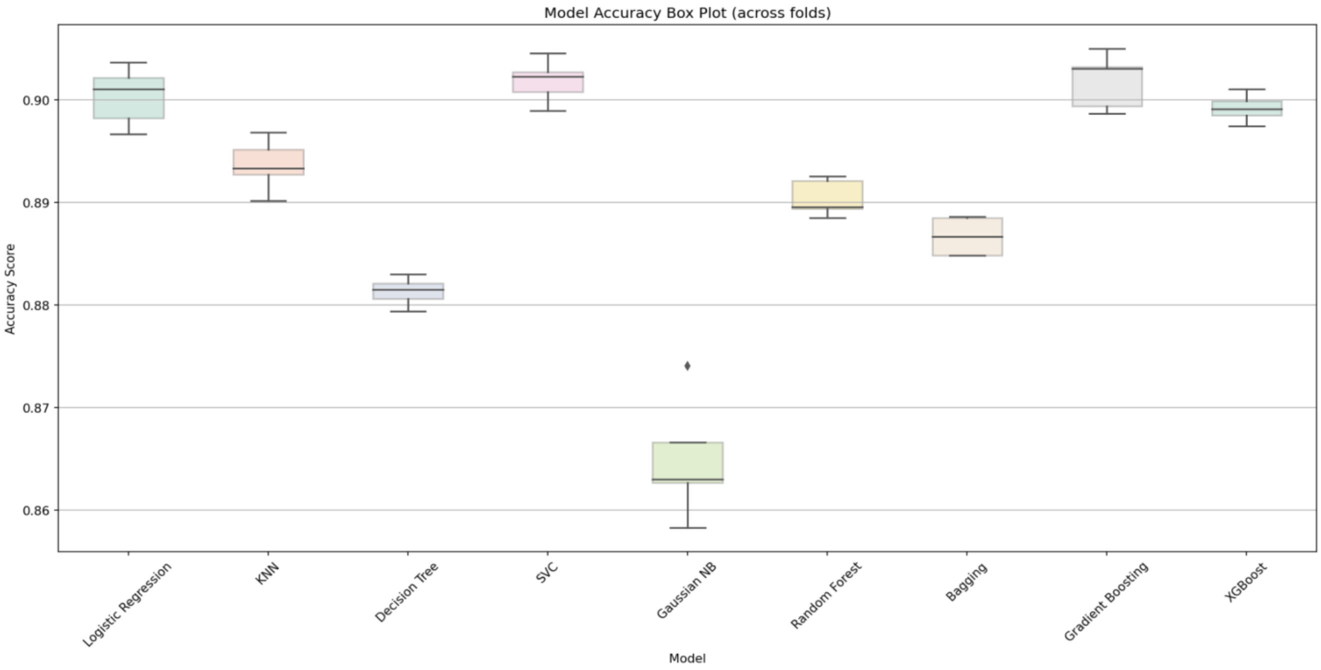


Figure 2: The Model Accuracy Box Plot.

Based on the accuracy box plot, we can see that the Logistic Regression model appears to perform very well, with the median accuracy score nearing 0.90 and a relatively narrow interquartile range (IQR), suggesting consistent performance across different folds. In addition, the knn model also shows relatively good performance. Its median accuracy is not far behind Logistic Regression. On the other hand, the

Gaussian NB model seems to have the lowest median accuracy score. Other models like SVC, Decision Tree, Random Forest, Bagging, Gradient Boosting, and XGBoost have their median accuracies clustered closely between 0.88 and 0.89, suggesting comparable performances. The spread of their IQRs is also relatively similar, which can allow us to utilize them as base models. In the next step, we utilize the Random Forest, Bagging, and Gradient Boosting as the base models. Use XGBoost as the metal model.

Based on the results from previous sections, we utilize the Random Forest, Bagging, and Gradient Boosting as the base models and use XGBoost as the metal model. Utilizing ensemble techniques like Random Forest, Bagging, and Gradient Boosting as base models, while using XGBoost as the meta model, essentially constructs a Stacking Classifier. The idea behind stacking is to combine the outputs of several base models and use our meta-model to make a final prediction and comparison. This helps in leveraging the strengths of individual models and potentially yields better performance than any single model. In this way, we can help the organization with more promising results by identifying factors that affect the subscriptions of the clients.

Table 2: result evaluation of each model

	Accuracy	Precision	Recall
Logistic Regression	0.900334	0.662472	0.235195
KNN	0.893627	0.551825	0.297955
Decision Tree	0.881305	0.458154	0.293913
SVC	0.901851	0.660067	0.264824
Gaussian NB	0.864917	0.409378	0.446928
Random Forest	0.890410	0.522396	0.317891
Bagging	0.886677	0.495552	0.307654
Gradient Boosting	0.901851	0.671985	0.251897
XGBoost	0.899181	0.614323	0.282332

2.2 Model Performance Evaluation

Then, by using the base models with the meta model, we also evaluate the accuracy of the models. Figure 3 shows the model results, and we can see that the Stacked Classifier, Bagging, Gradient Boosting, and XGBoost, perform the best based on the model selection.

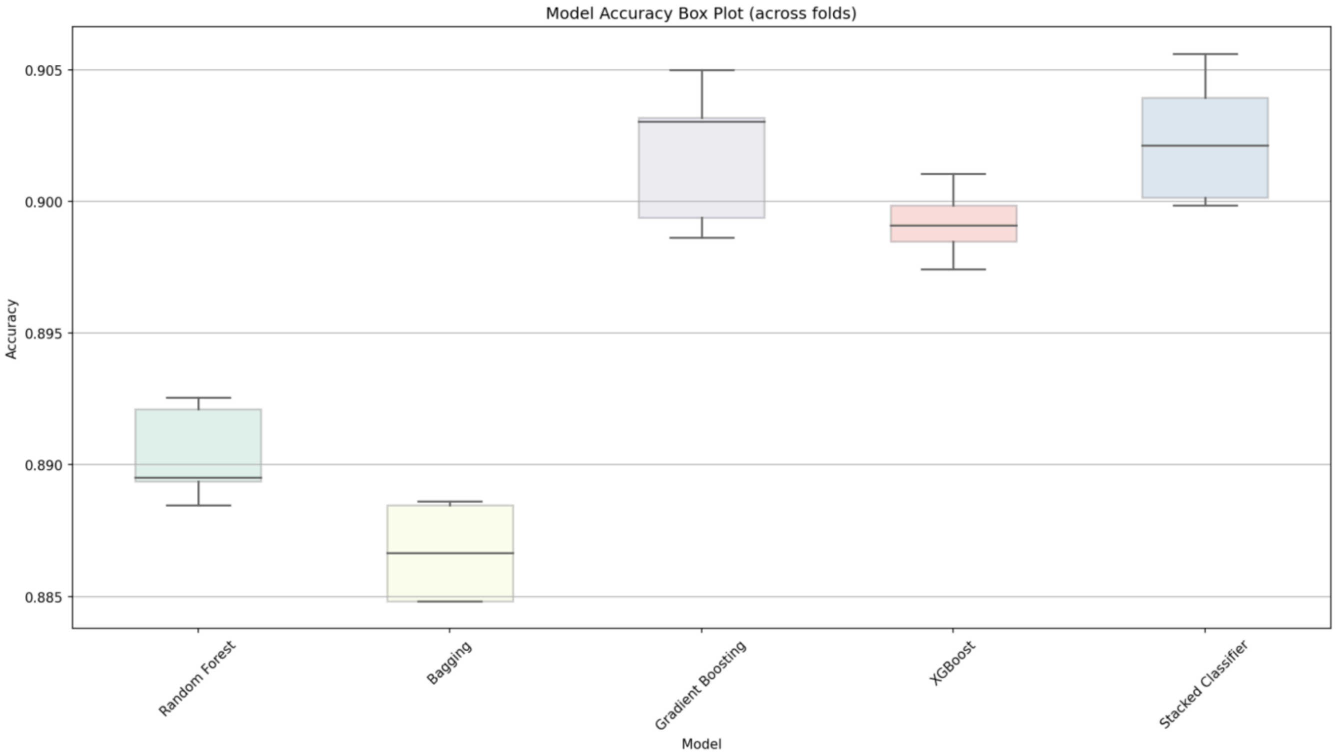
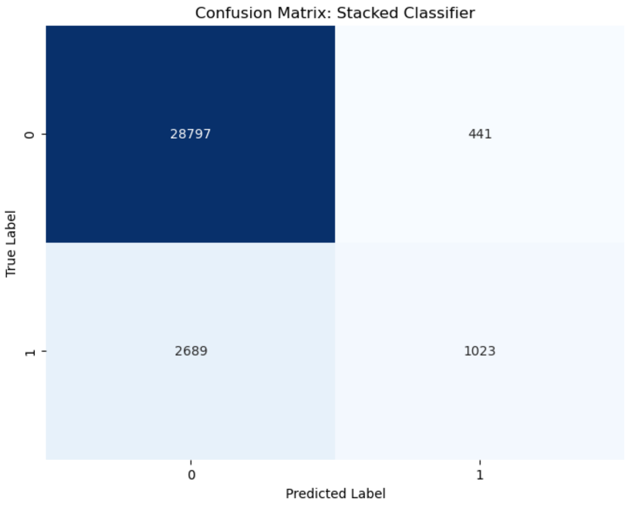
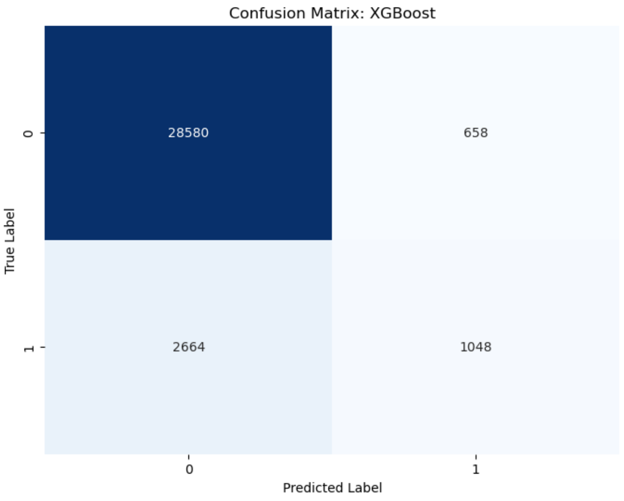


Figure 3: Base Models & Meta Model Results

In addition to the model accuracy box plot, we have also included some confusion matrices for us to evaluate the differences among models. Figure 4 shows the performance of the standalone models and stacked models.





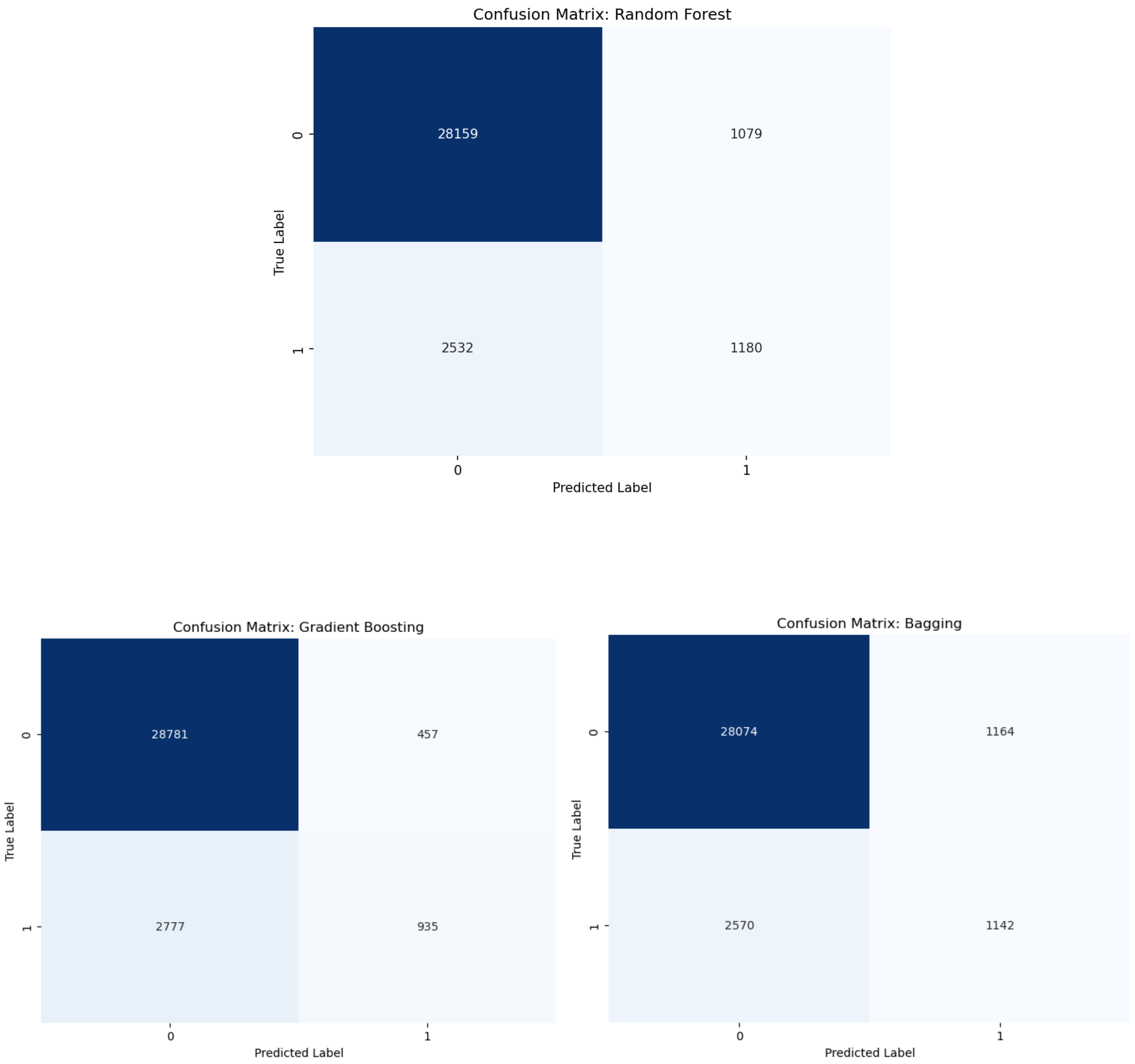


Figure 4: Confusion matrix for the models and the resulting stacked classifier

Based on the confusion matrices, while the stacked classifier leverages the power of multiple models, including Gradient Boosting, its performance appears to be a trade-off between Bagging and Gradient Boosting. In terms of reducing false positives, the Stacked Classifier performs relatively effectively.

2.3 *Stacked Model Analysis*

After getting the results, we want to make predictions on whether customers will subscribe to a term deposit or not through another confusion matrix since the confusion matrix provides insights into the performance of a classification model by comparing its predictions with actual labels.

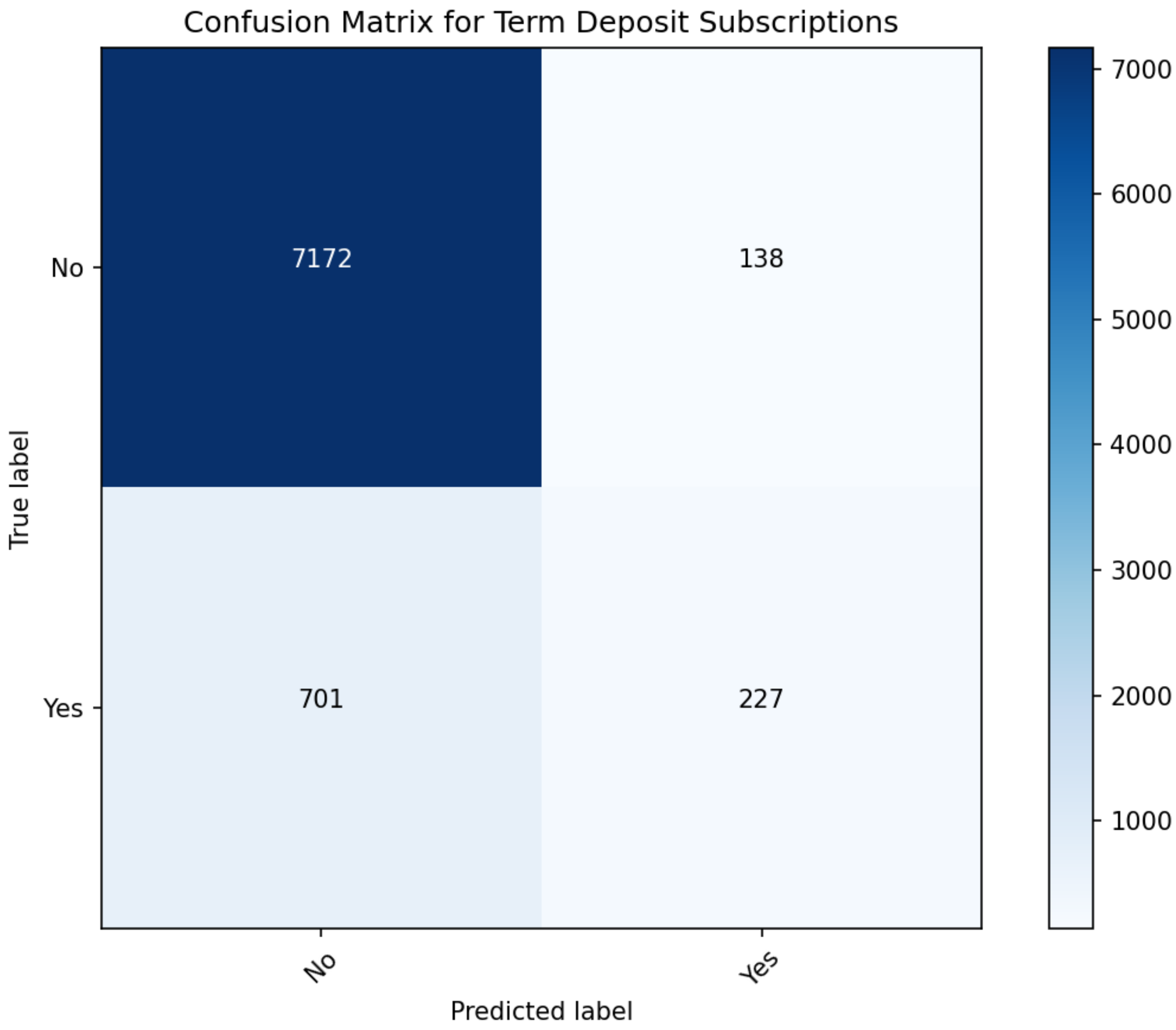


Figure 5: Confusion matrix for the model for our objective feature

Based on Figure 5, here are some key insights that we can obtain from the numbers. The model correctly identified 227 instances of customers who subscribed to the term deposit. This means that it can accurately predict those who are interested in the financial product. In addition, the model correctly identified 7,172 instances of customers who did not subscribe to the term deposit. This also accurately recognizes those who are not interested. However, the model mistakenly predicted 138 instances as customers who subscribed to the term deposit when they did not. The model missed 701 instances of customers who actually subscribed to the term deposit, but it classified them as non-subscribers.

In addition, for deposit marketing, based on the confusion matrix, the false positives may include unnecessary follow-up with customers who are not interested, potentially causing inconvenience. Furthermore, false negatives represent missed opportunities to engage with interested customers, potentially resulting in lost business. By taking account of these potential factors, the organization can have a better view of the customers.

Also, we think that further analysis such as grid search and random search will be conducted for later sections in order to have more effective results for us to gain insights.

2.4 Feature Importance Analysis

After the modeling results, we wish to know which feature is more crucial or important for the term deposit subscription. Therefore, we utilized the aggregate results from the models to provide the feature importance plot.

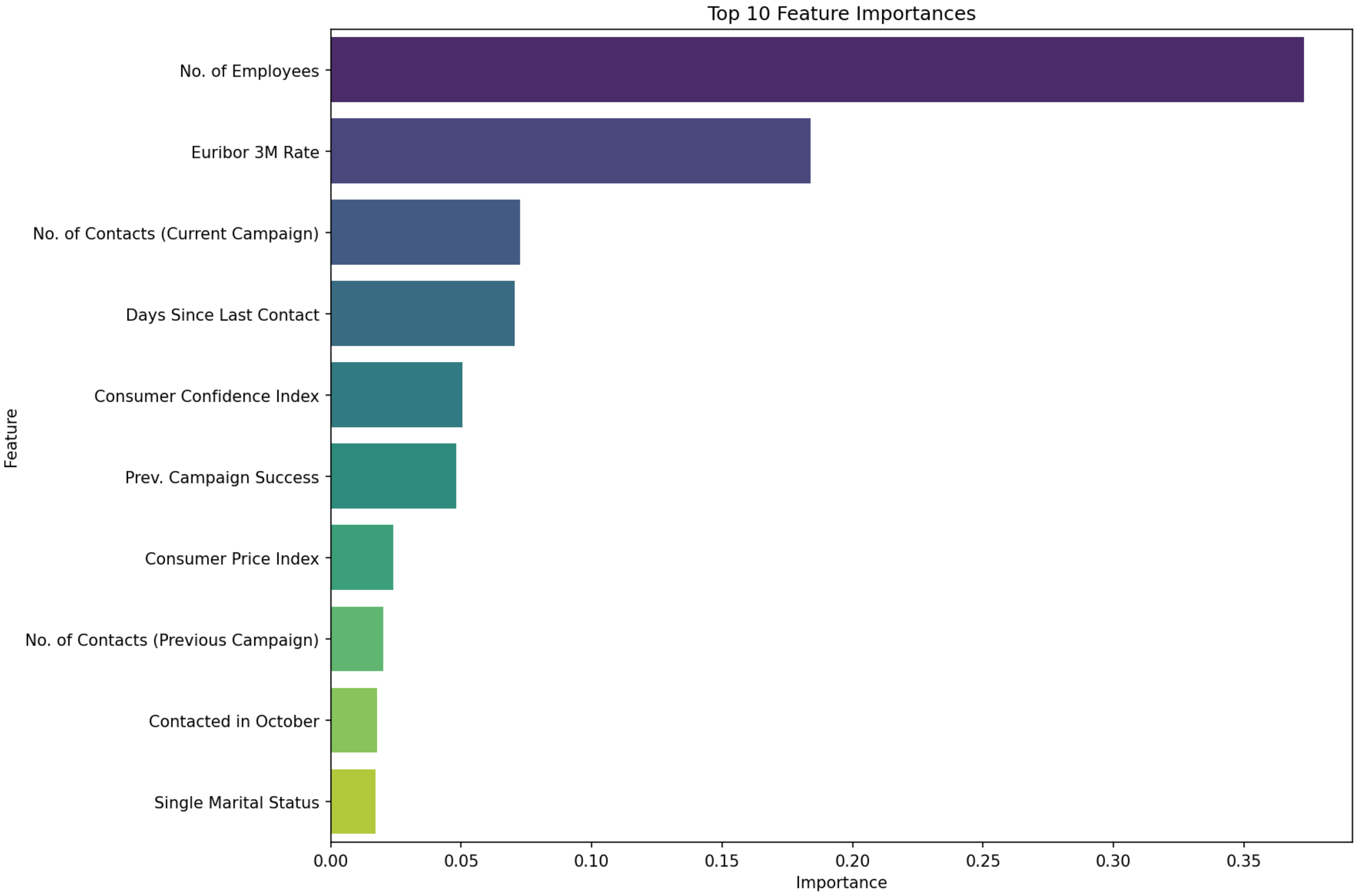


Figure 6: Aggregated Feature Importance Plot

Based on Figure 6, we can see that the No. of Employees continues to serve as a significant quarterly economic indicator. This metric's prominence underscores the broader economic environment's role, possibly linked to the health of the banking sector or prevailing employment trends, in influencing clients' decisions to opt for a term deposit. Then, the second most important feature is the Euribor 3M Rate, which is an economic indicator, suggesting that shifts in this rate could impact clients' financial product decisions. Banks must be proactive and adjust their strategies in tandem with these shifts. Other macroeconomic factors such as the Consumer Confidence Index or Consumer Price Index are also important features.

In summary, macroeconomic factors, such as the number of employees, exert a pronounced influence on the results. Incorporating such factors into the dataset can enhance predictive accuracy, providing a more holistic understanding of clients' decision-making dynamics.