Machine Learning Application Deployment for Predicting Term Deposit Subscriptions

Group 10: Jingda Yang, Haiyu Xiao, Tingsong Li

Bank Marketing Dataset

Table of Contents:

1. Project Overview

In today's competitive banking landscape, retaining existing customers and converting potential leads into clients is more crucial than ever. With the introduction of various marketing channels and campaigns, banks are aggressively promoting their term deposit schemes among their clients. However, not all promotions convert into actual subscriptions, resulting in wasted resources and missed opportunities. This is where predictive analytics can offer an advantage. The Bank Marketing Dataset aims to predict if a client will subscribe to a term deposit, based on a plethora of attributes ranging from personal details, economic indicators to interaction history with the bank.

1.1 - Project Background and Description

Background	This problem captures attention for several reasons. First, it offers a route to resource optimization, enabling banks to allocate their marketing resources more efficiently by focusing on clients most likely to subscribe to a term deposit. Second, the dataset's wide variety of features allows for in-depth customer segmentation, presenting an opportunity to personalize marketing strategies for higher conversion rates. Adding another layer of complexity, the inclusion of macroeconomic indicators like employment variation rates and consumer price indices provides a unique challenge. It allows for the exploration of how external economic factors might influence individual financial decisions. Overall, the problem is not just a real-world issue faced by banks but also a fascinating challenge from a data science perspective.
Description	Solving this classification problem comes with several tangible benefits. At the forefront is the potential for increased revenue: by effectively targeting the right candidates, the bank can substantially improve its term deposit subscription rates. Additionally, narrowing down the focus to likely subscribers can considerably cut down on the overall marketing costs, increasing the campaign's ROI (Return on Investment). Beyond financial metrics, avoiding the irritation that comes with irrelevant offers could result in improved customer satisfaction and loyalty. Adopting a predictive model for this task can also serve as a steppingstone towards a more comprehensive data-driven decision-making approach within the organization.

1.2- Project Questions

Problem 1.	How can we accurately predict whether a client will subscribe to a term deposit based on various personal, economic, and interaction-based attributes?
Problem 2.	What insights can we gain about the influence of individual and macroeconomic factors on a client's decision to subscribe to a term deposit, and how can these insights optimize marketing strategies and resource allocation?

1.3- Project Methods

Model selection phase	Begin with a straightforward approach by employing logistic regression as a baseline model. Following this, explore more intricate models such as decision trees and random forests to capture intricate relationships within the data. Additionally, harness the power of gradient boosting algorithms like XGBoost and LightGBM to address complex patterns effectively. If the dataset is substantial and computational resources are available, considering the utilization of support vector machines (SVMs) and neural networks can be beneficial.
Model evaluation	When it comes to model evaluation, ensure the dataset is divided into distinct training and testing sets to assess model performance accurately. Employ cross-validation to gauge the generalization ability of the models, and meticulously scrutinize key evaluation metrics such as accuracy and ROC-AUC. The selection of these metrics should be based on their relevance to the specific business objectives, ensuring that the chosen metrics align with the project's ultimate goals.

1.4- Project Deliverables

Problem identification	Define the project's problem statement and objectives.
Method Evaluations	Systematically assess and compare various machine learning methods for binary classification.

Solution by hyperparameter optimization	Choose a set of optimal hyperparameters for the selected model and get the result.
Deployment pipeline/platform	Integrate our trained models for real-world use.
Project demonstration in class	

1.5- Out of Scope

This project **will NOT accomplish or include** the following:	Utilizing neural networks for predictive modeling falls outside the scope of this project, as the primary focus is on evaluating simpler machine learning algorithms. Complex neural network architectures are not within the project's predefined objectives. Also, as the feature engineering part is done, some macroeconomic factors (e.g., GDP, inflation rates) might stand out from others. Analyzing macroeconomic factors that might influence banking decisions could be out of scope unless directly relevant to the project's objectives.

1.6- Constraints

The **constraints** for this Project	Regarding constraints, the primary limitation is the dataset itself. While the Bank Marketing data offers a broad range of attributes from personal details to macroeconomic indicators, it's important to acknowledge that real-world conditions and customer behaviors can change over time. Therefore, the model might need regular retraining or fine-tuning to account for these shifts. Moreover, there could be crucial external factors or variables not present in the dataset that influence a client's decision, potentially leading to unaccounted biases or variances in predictions.

1.7- Success Criteria

Quantitative	Quantitatively, the primary metric would be the model's accuracy in predicting whether a client would subscribe to a term deposit. Other metrics like precision, recall, and the F1-score might also be relevant, especially considering the potential imbalances between subscribers and non-subscribers in the dataset.
Qualitative	Qualitatively, the project would be considered a success if the results and insights derived can be seamlessly integrated into the bank's marketing strategy, leading to more personalized and efficient campaigns, and whether there is an evident improvement in customer satisfaction due to fewer irrelevant offers.

2. Analysis of the dataset

2.1- Exploratory Analysis and Visualization

Before moving forward, it is important to perform exploratory analysis. Since the project is about classification on whether a client will subscribe to a term deposit, a pie chart will be used to show the distribution of this classification variable:

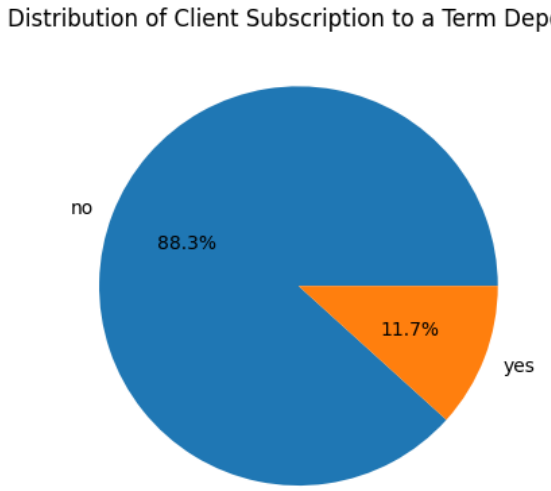


Figure 1: Distribution of the Dependent Variable (Client Subscribes to a Term Deposit)

Figure 1 shows how the dependent variable is distributed: 88 percent of the clients do not subscribe to a term deposit, and 12 percent of the clients do. This suggests that the dataset is highly unbalanced. Table 1 shows the summary of all the variables after converting the categorical columns into numerical columns.

	age	balance	day	duration	campaign	pdays	previous	job_label	marital_label
mean	40.936	1362.272	15.806	258.163	2.764	40.198	0.58	4.34	1.168
std	10.619	3044.766	8.322	257.528	3.098	100.129	2.303	3.273	0.608
min	18	-8019	1	0	1	-1	0	0	0
50%	39	448	16	180	2	-1	0	4	1
max	95	102127	31	4918	63	871	275	11	2

Table 1: Summary of the Variables

	education_label	default_label	housing_label	loan_label	contact_label	month_label	poutcome_label	y
mean	1.225	0.018	0.556	0.16	0.64	5.523	2.56	0.117
std	0.748	0.133	0.497	0.367	0.898	3.007	0.989	0.321
min	0	0	0	0	0	0	0	0
50%	1	0	1	0	0	6	3	0
max	3	1	1	1	2	11	3	1

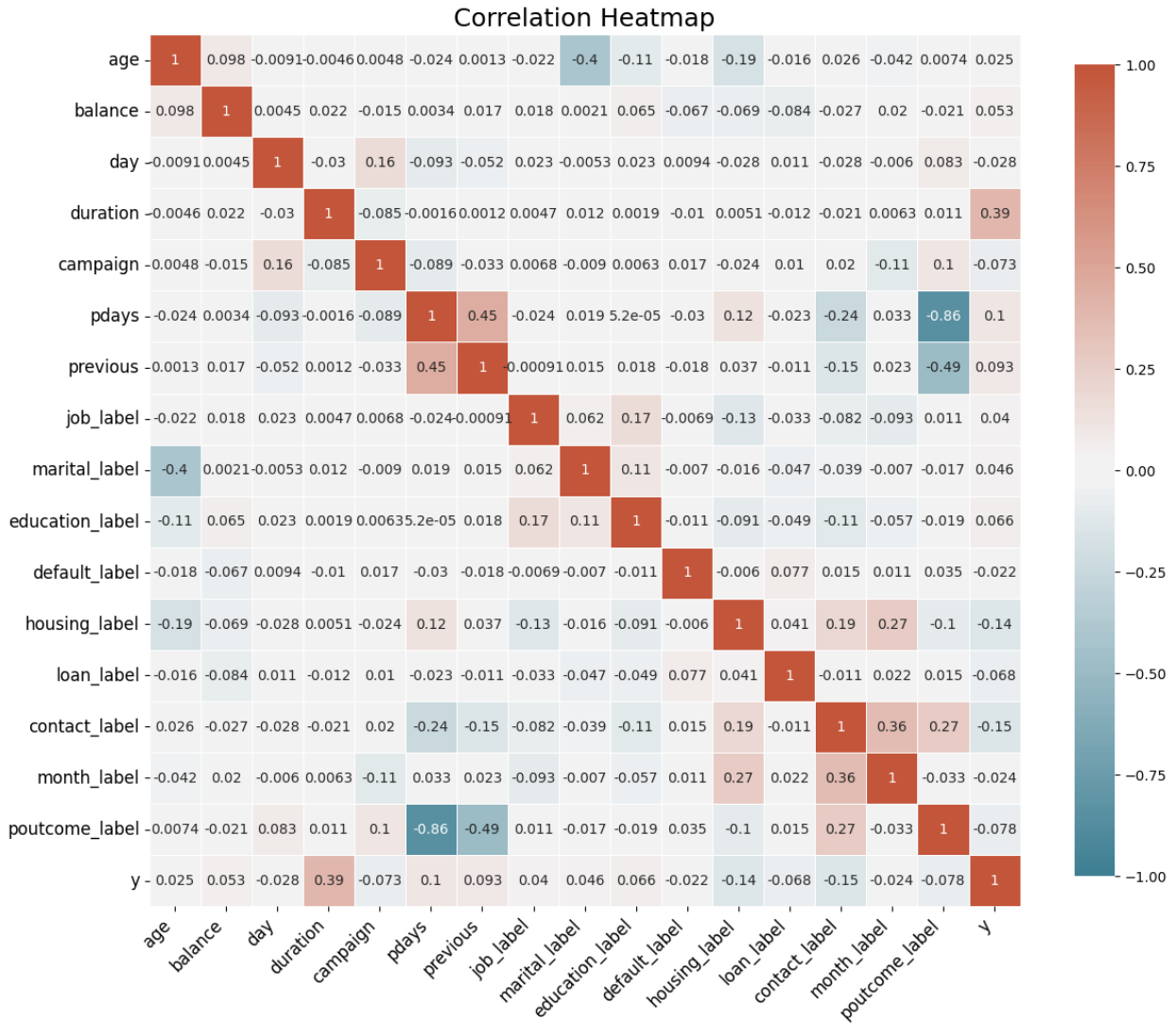


Figure 2: Correlation Heatmap of All the Variables

Figure 2 shows the correlation heatmap among all the variables after converting some of the categorical ones into the numerical variables. We can deduce that between variables 'pdays' and 'poutcome_label' there is a strong negative correlation, and no two variables show strong positive correlation.

3. Model Selection

In the context of a machine learning project, one of the most crucial tasks is to select an appropriate machine learning model for the dataset that could maximize the accuracy of predictions. Before testing the model, we converted all categorical variables to one-hot encoding format. Besides the baseline model logistic regression, we will also be examining the following models: K-Nearest Neighbors, Decision Tree, SVM, Gaussian Naive Bayes, Random Forest, Bagging, Gradient Boosting, and XGBoost. To compare the accuracy of the aforementioned models, we conducted cross-validation for each model, plotting the results from all folds as a box plot.

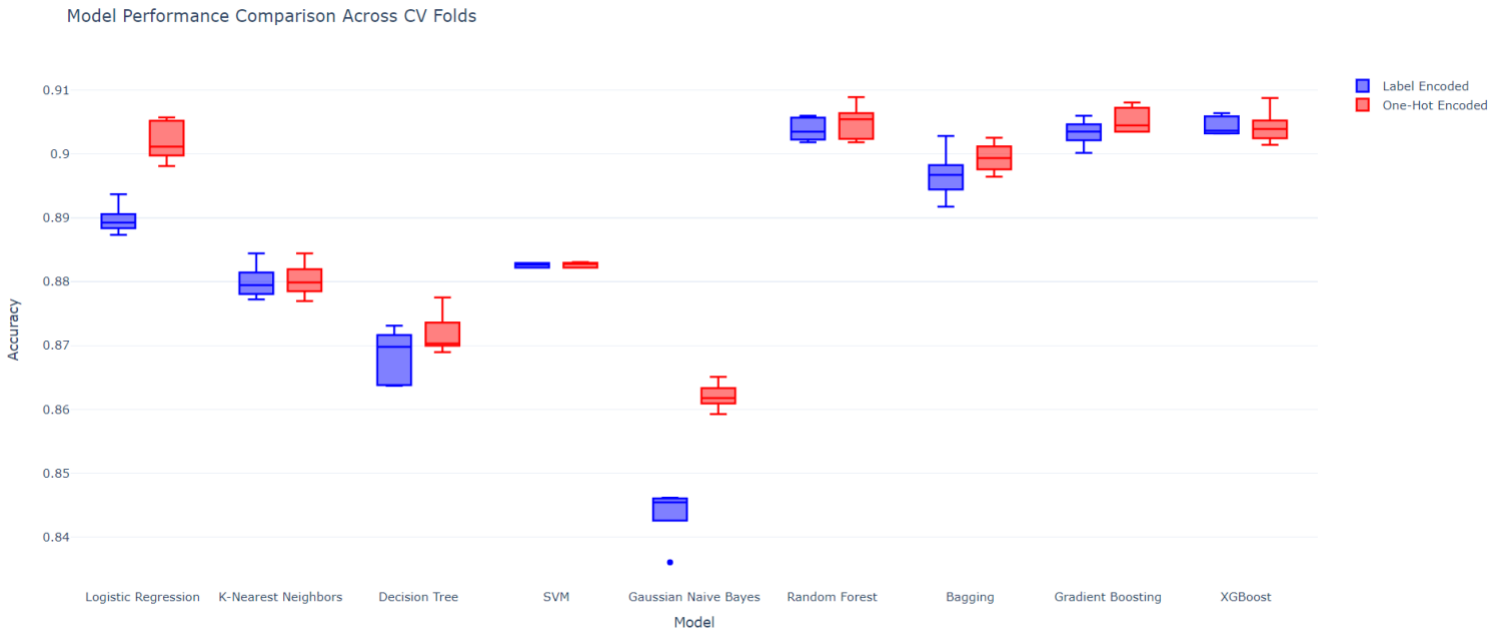


Figure 3: Model Performance Comparison

Based on the box plot, we found that Random Forest, Bagging, Gradient Boosting, and XGBoost stand out from the other machine learning models. To enhance the accuracy of predictions, we attempted to build a stacked model. In the stacked model, we included Random Forest, Bagging, Gradient Boosting, and XGBoost, and we used the baseline model logistic regression as the final layer. The box plot of the stacked model, as well as its component models, is presented below.

Figure 4: Model Comparison with Stacked Model

According to Figure 4, the stacked model exhibits the best performance compared to the other models. In this case, the stacked model would be selected.

3.1 - Model Performance Evaluation

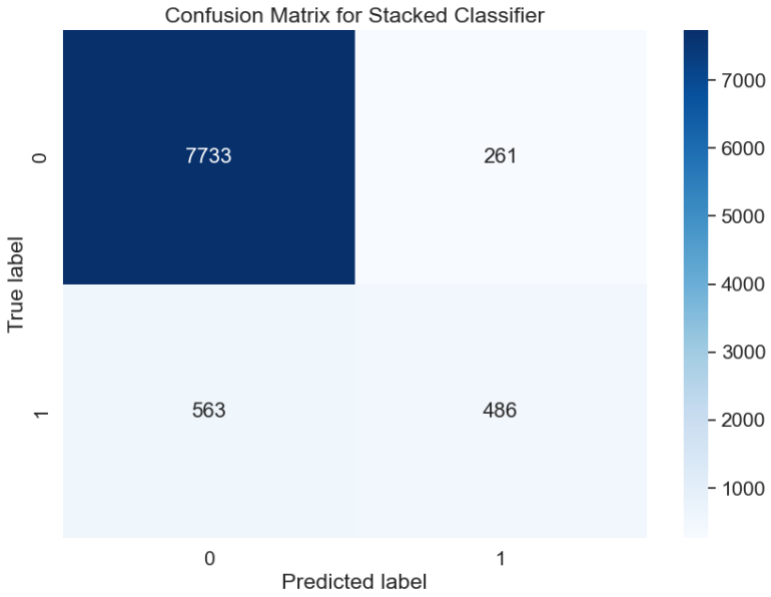


Figure 5: Confusion Matrix for Stacked Model

4. Initial Deployment

In this project, we decided to use Heroku for the final deployment. In terms of deployment, there are two side-by-side figures displayed on the website. The first figure displays the relationship between the dependent variable and the two most important independent variables. The second figure displays the input independent variables along with the predicted label, in addition to the first graph.

In preparing to visualize the dataset and simplify the model, we derived the feature importance from the random forest model using impurity decrease. As a result, the top 5 most important features are "Duration," "Balance," "Age," "Day," and "Pdays." In this case, these 5 variables will be used as independent variables for prediction in this part of the project. The figures displayed on the site will include "Duration" and "Balance."

4.1 - Screen 1

Figure 6 displays the distribution of the dependent variable, indicating whether a customer will subscribe to a term deposit, in relation to the two most important features: "Duration" (referring to how long the last contact took) and "Balance" (representing the yearly average balance). Users can input new data points with feature vectors into the data using the text box below the figure. Afterward, they can click on "Predict" to view the input features and the predicted result on a

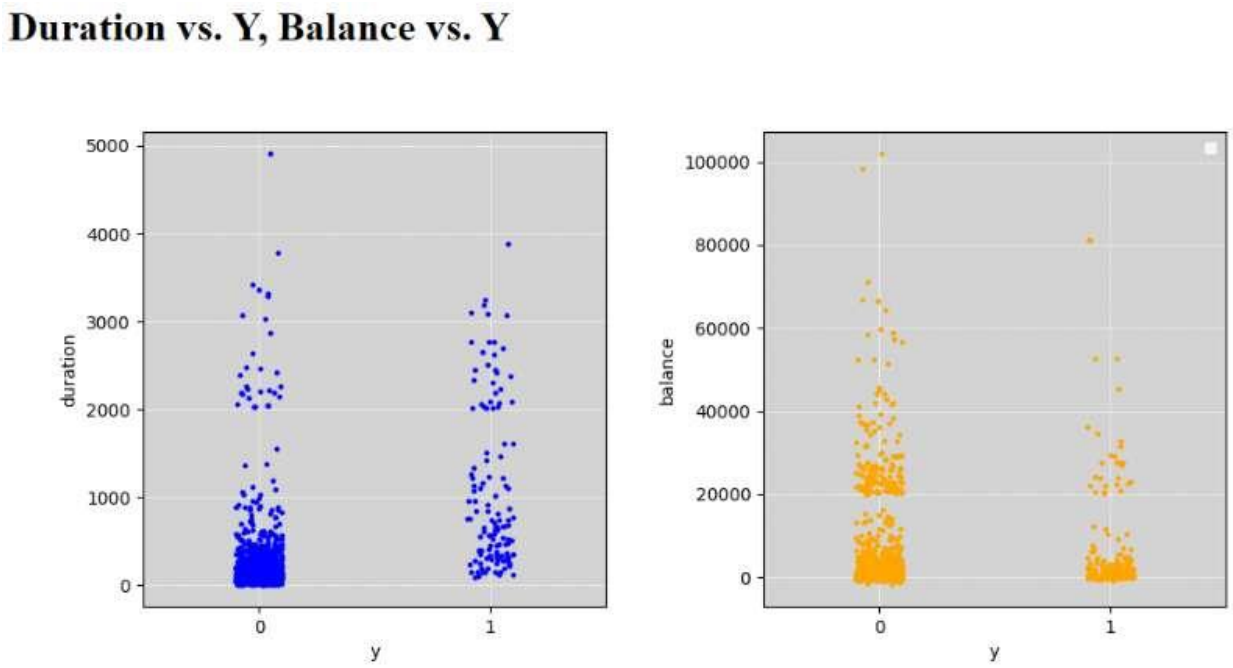


Figure 6: Screenshot of Unpredicted Dataset

new feature. Afterward, they can click on "Predict" to view the input features and the predicted result on a new feature.

4.2 - Screen 2

Figure 7 displays two features inside the feature vector entered by the user as well as the predicted label based on the embedded stacked model. The feature vector is 200, 1000, 35, 10, -1. Here are 5 example feature vectors users could use:

1. 200, 1000, 35, 10, -1
2. 2000, 30000, 40, 10, -1
3. 500, 35000, 25, 28, 100
4. 10, 20, 30, 31, 40
5. 2000, 40000, 40, 10, -1

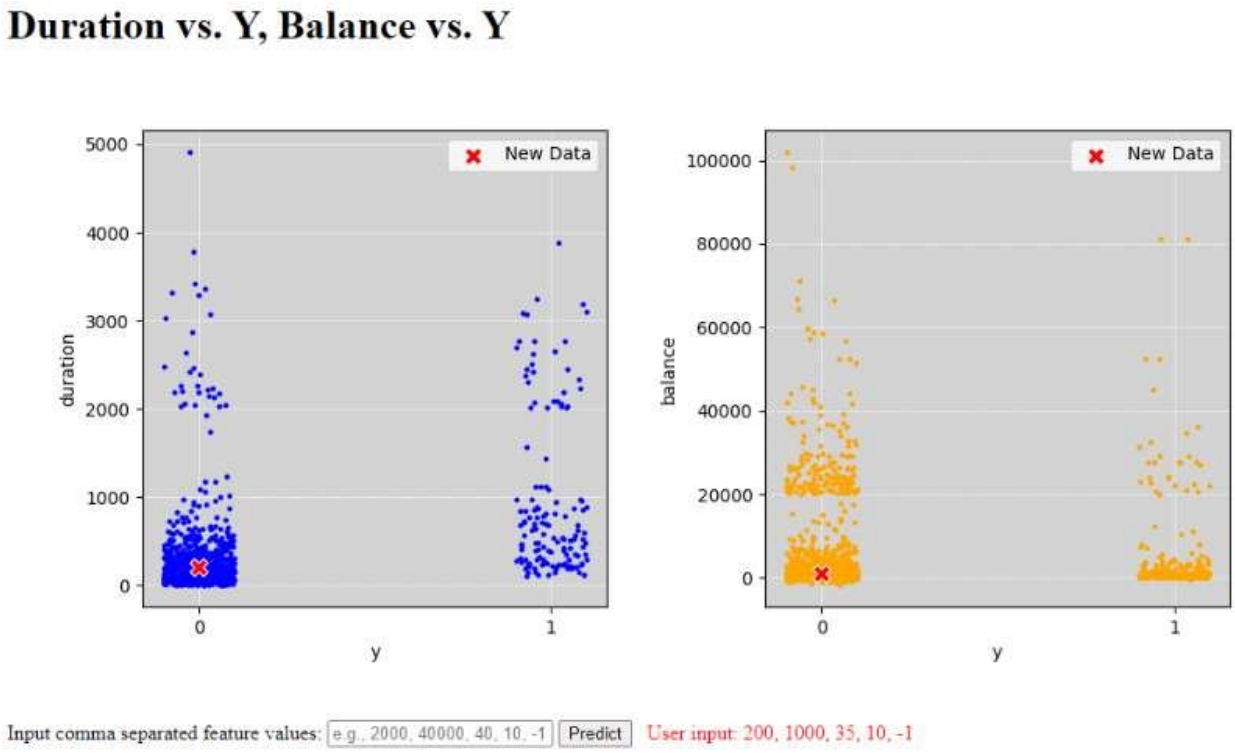
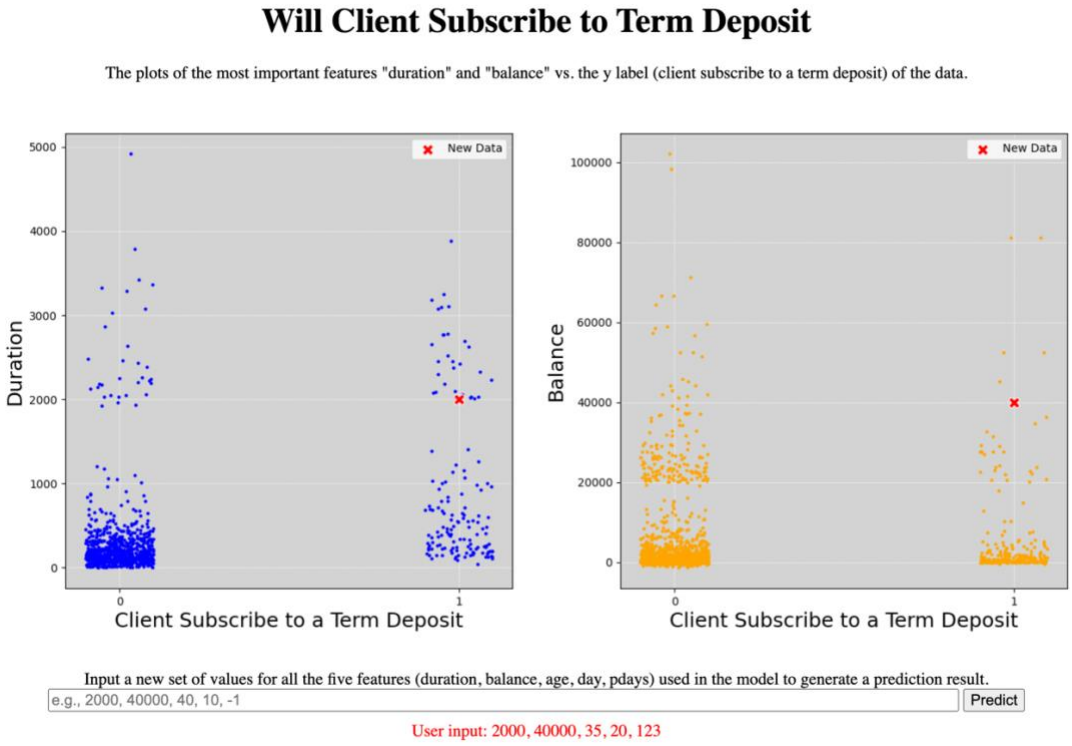


Figure 7: Predicted Vector

5. Heroku Application

The stacked model above was deployed on Heroku: https://bank-model-prediction-84db523114a3.herokuapp.com/.

In this app, input a new set of values for all the five features (duration, balance, age, day, pdays) used in the model to generate a prediction result. The result will be shown in both plots.



6. Conclusion

In retrospect, we are attempting to create a machine learning application that predicts whether a customer will subscribe to a term deposit. This prediction is valuable for banks to identify potential subscribers and allocate marketing resources efficiently. To achieve this, we initially cleaned the data, converting categorical data into one-hot encoding. We then compared multiple classification models and combined the four best models with a logistic regression model to create a stacked model, which outperformed the other models. For model deployment, we used the stacked model with the five most significant variables to generate predicted labels based on user-provided feature vectors. Additionally, figures displaying the original data distributions and predicted results help users understand the dataset comprehensively.

In the process, we focused on model selection and built a robust stacked model for classification. The application's user-friendly interface allows easy input of new data points for instant predictions, promoting accessibility and practicality for both banking professionals and clients.