Bank Marketing Data Analysis and Model Evaluation Report

1. Short Story of Business/Organization Challenge

The dataset used in this project is related to direct marketing campaigns (phone calls) conducted by a Portuguese banking institution. The marketing campaigns aimed to promote term deposits among clients. The campaigns were carried out over a span of a few years, and data was collected on various attributes of the clients contacted during these campaigns.

The primary challenge for the bank was to predict whether a client would subscribe to a term deposit based on the information collected during the marketing campaigns. This prediction would help the bank to target their marketing efforts more efficiently and increase the success rate of their campaigns.

2. Problem Summary/Definition

Problem Statement

The primary problem faced by the bank was to predict whether a client would subscribe to a term deposit based on their demographic and historical data from previous campaigns. This prediction would help in:

- Targeting potential clients more effectively.
- Reducing marketing costs by focusing efforts on clients with a higher likelihood of subscription.
- Improving the overall success rate of marketing campaigns.

Dependent and Independent Variables:

- **Dependent Variable (Y):** The target variable is whether the client has subscribed to a term deposit (binary: 'yes' or 'no').
- **Independent Variables (X):** These include client demographic information and previous campaign data such as age, job, marital status, education, default status, balance, housing loan, personal loan, contact type, day of week, month, duration of last contact, number of contacts performed during this campaign, days since the client was last contacted from a previous campaign, and outcome of the previous campaign.

Dataset Overview

The dataset comprises 45,211 instances and 16 attributes, capturing client demographics, campaign-related information, and previous campaign outcomes. The key features (independent variables) and the target variable (dependent variable) are described below:

Independent Variables (X)

1. **Age**: Client's age (numeric).
2. **Job**: Type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', etc.).
3. **Marital**: Marital status (categorical: 'married', 'divorced', 'single').
4. **Education**: Education level (categorical: 'unknown', 'secondary', 'primary', 'tertiary').
5. **Default**: Has credit in default? (binary: 'yes', 'no').
6. **Balance**: Average yearly balance in euros (numeric).
7. **Housing**: Has housing loan? (binary: 'yes', 'no').
8. **Loan**: Has personal loan? (binary: 'yes', 'no').
9. **Contact**: Contact communication type (categorical: 'unknown', 'telephone', 'cellular').
10. **Day of Week**: Last contact day of the week (numeric).
11. **Month**: Last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec').
12. **Duration**: Last contact duration in seconds (numeric).
13. **Campaign**: Number of contacts performed during this campaign (numeric).
14. **Pdays**: Number of days since the client was last contacted from a previous campaign (numeric).
15. **Previous**: Number of contacts performed before this campaign (numeric).
16. **Poutcome**: Outcome of the previous marketing campaign (categorical: 'unknown', 'other', 'failure', ‘success').

3. Solution/Recommendations/Decisions

Data Preprocessing

- **Missing Values:** Missing values in the 'job', 'education', and 'contact' columns.
- **Encoding:** Categorical variables.
- **Scaling:** Standardize the dataset

Handling Missing Values

- Columns with missing values: `job`, `education`, `contact`, `poutcome`.
- Filled missing values in `job`, `education`, and `contact` using the mode.
- Dropped the `poutcome` column due to a high number of missing values.

Encoding Categorical Variables

- Applied one-hot encoding to categorical variables.
- The final dataset after encoding contained 36 features.

Standardizing the Data

- Used `StandardScaler` to standardize the dataset.

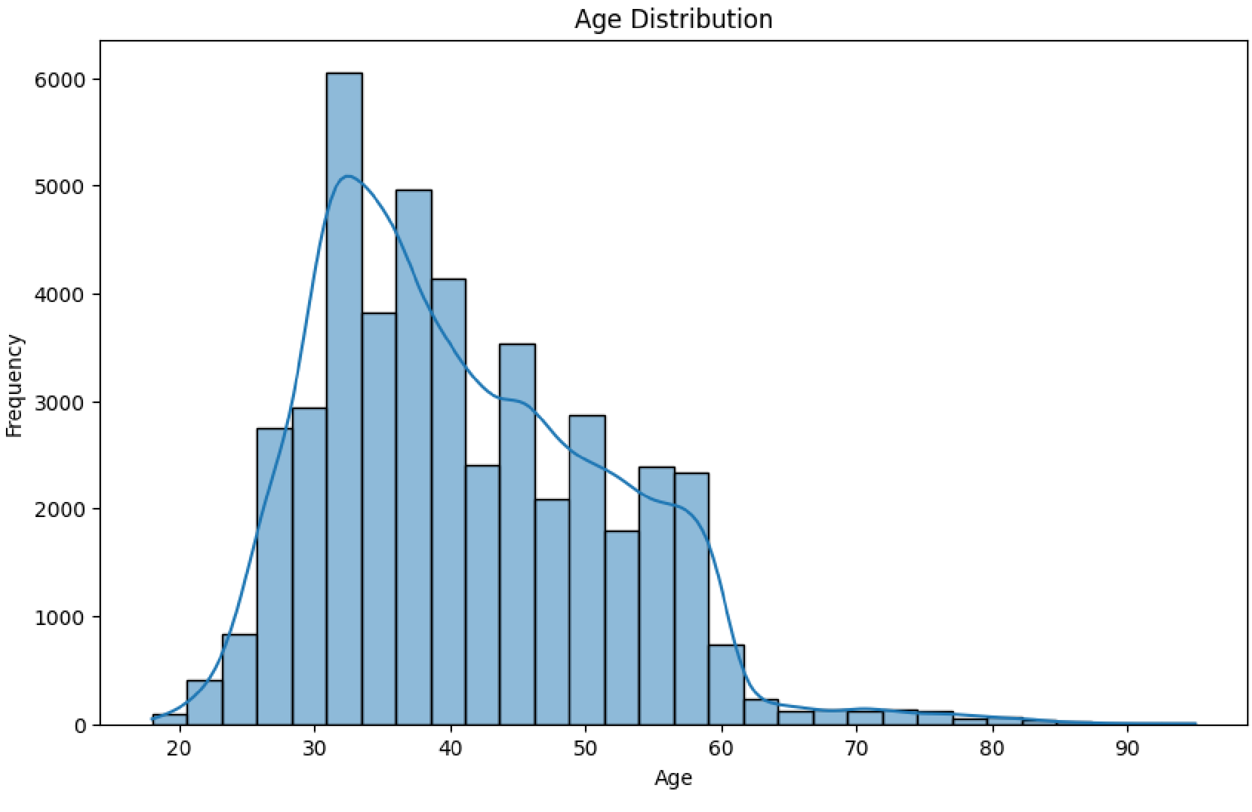
Balancing the Dataset

- Applied SMOTE (Synthetic Minority Over-sampling Technique) to balance the training set.

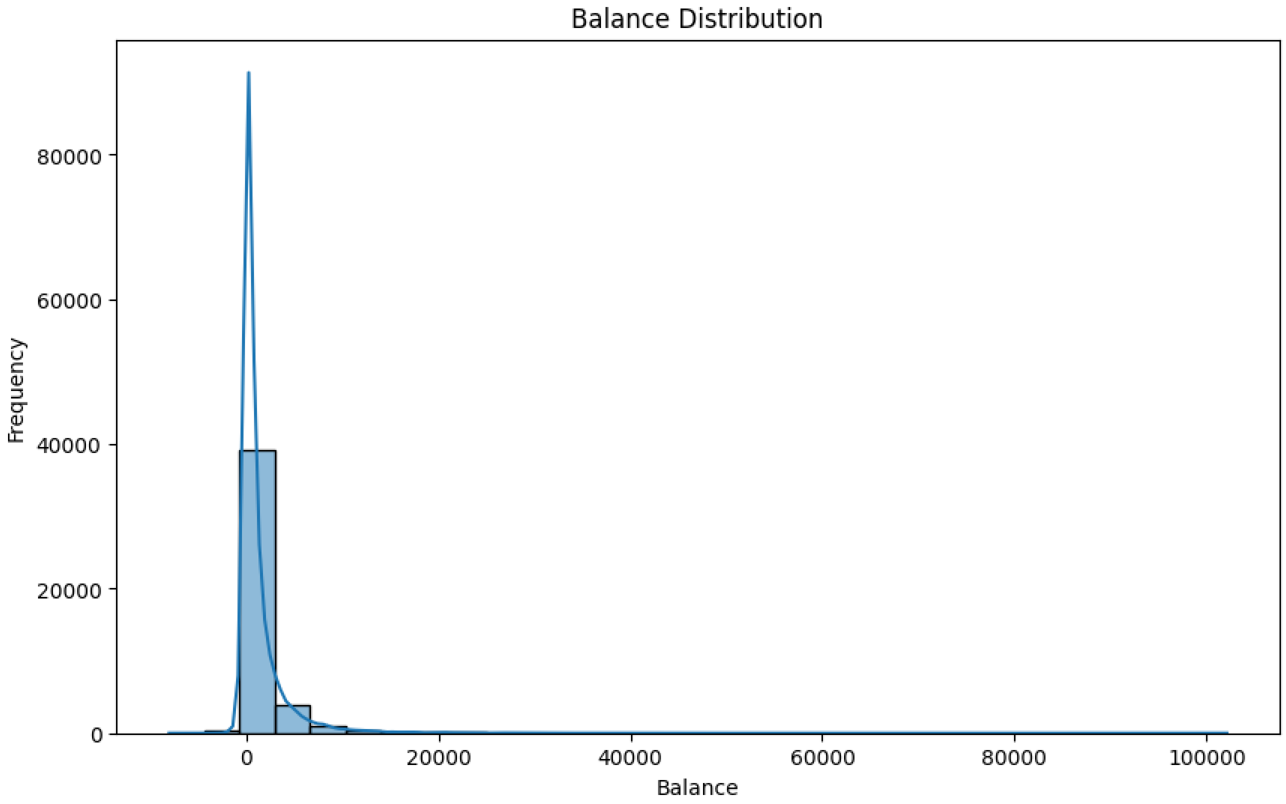
Visualizations After Data preprocessing

Distribution Analysis

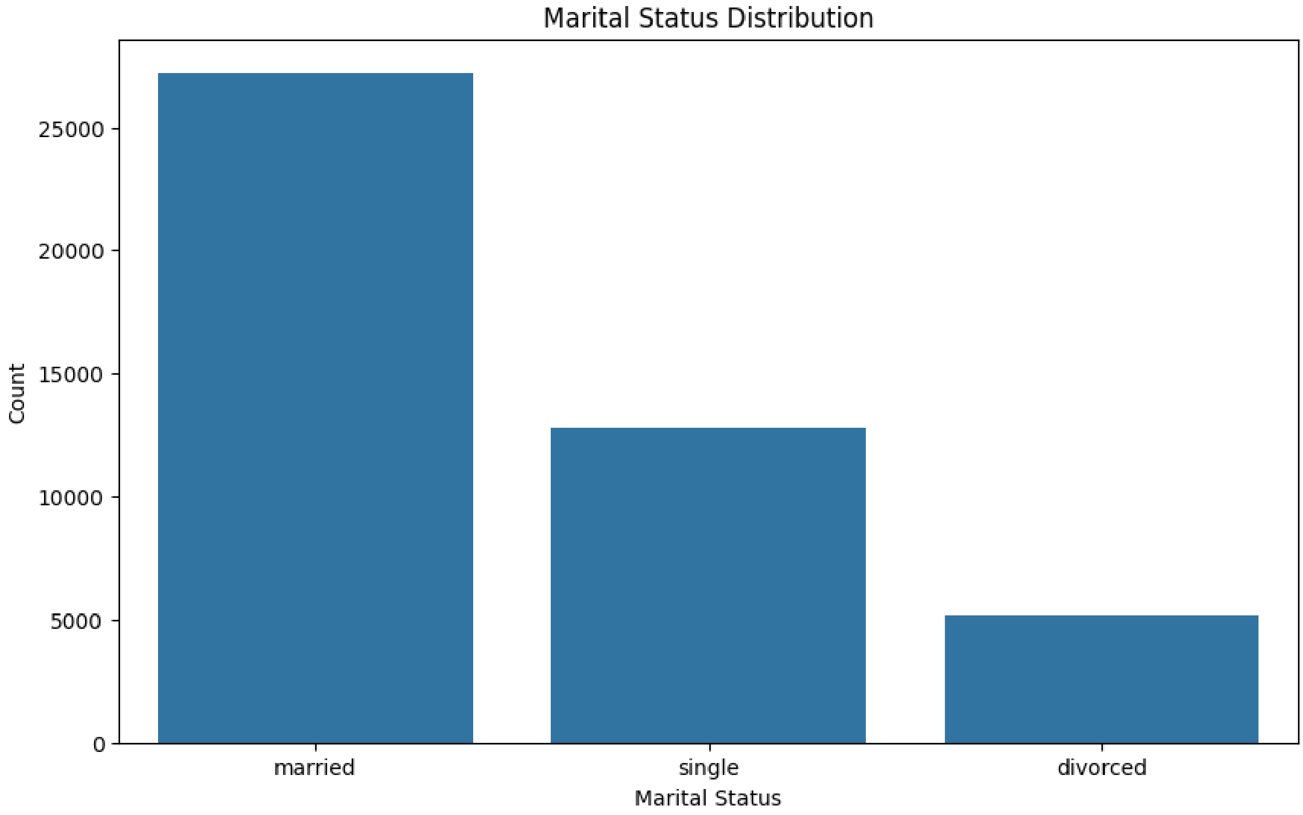
Age Distribution



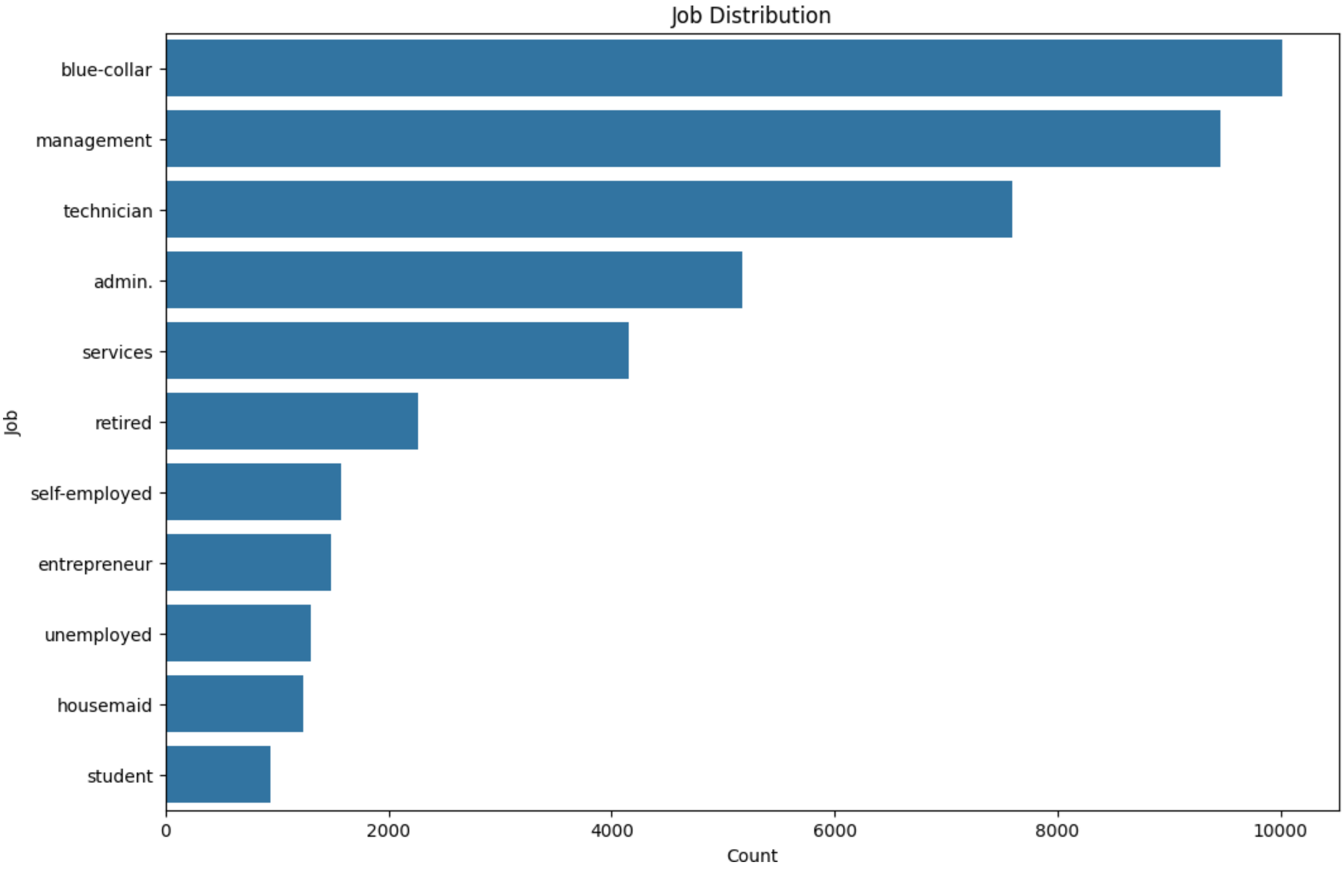
Balance Distribution



Marital Status Distribution



Job Distribution



Model Training and Evaluation

Predictive Analytical Methods

Three machine learning models were trained and evaluated:

1. **Logistic Regression**
2. **Random Forest Classifier**
3. **Support Vector Machine (SVM) with RBF Kernel**

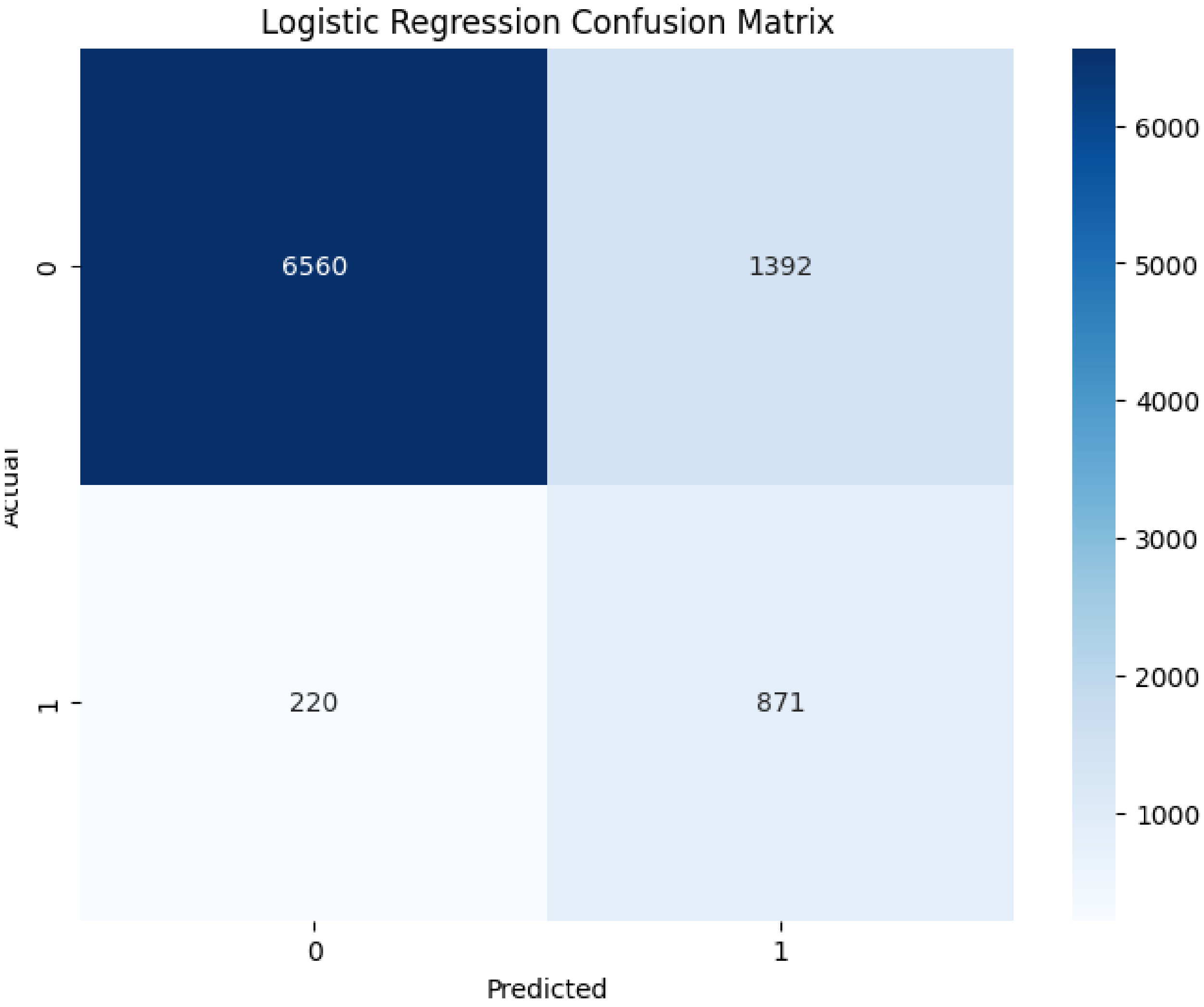
Logistic Regression

Performance Metrics:

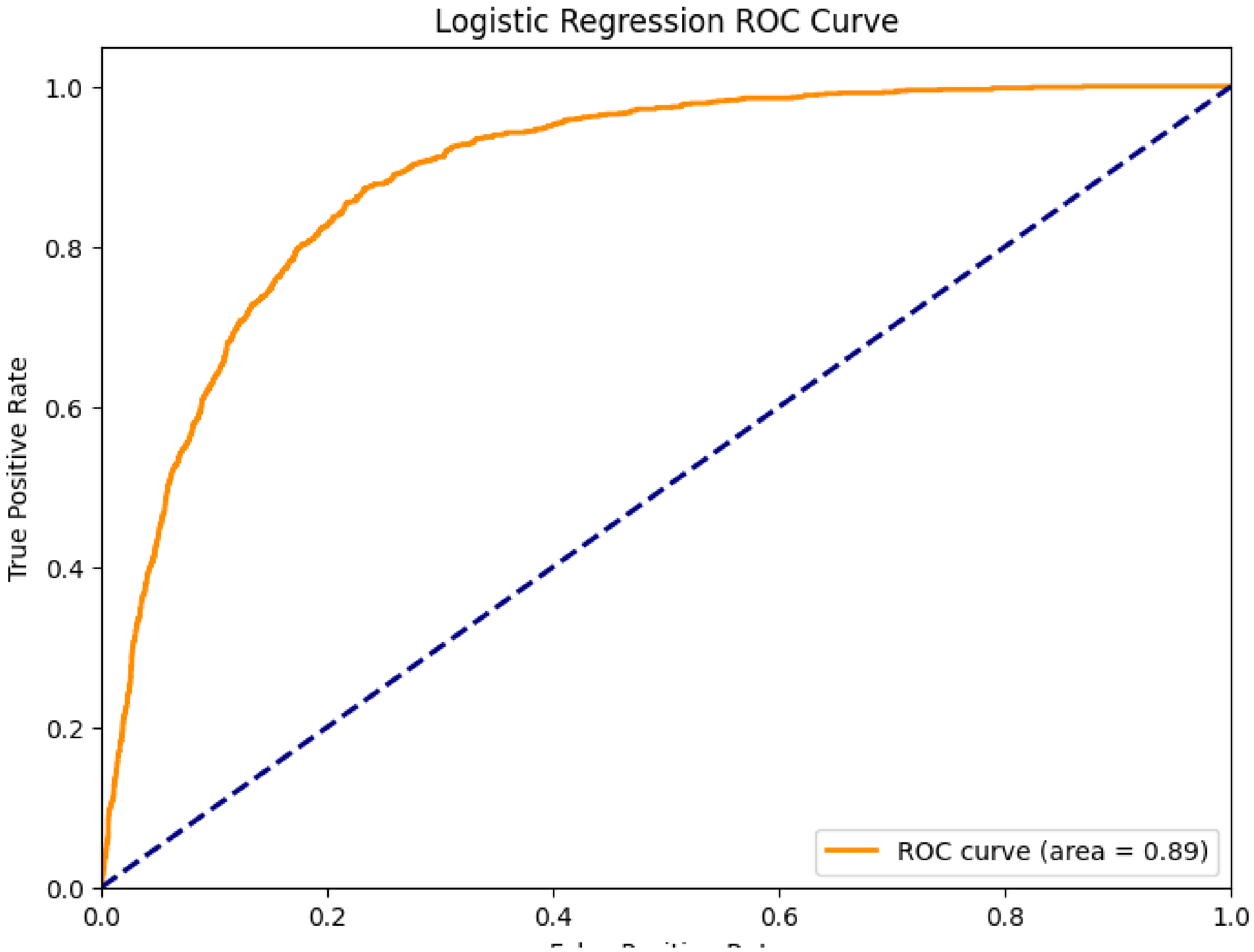
- **Accuracy:** 82%
- **Precision (no):** 0.97
- **Recall (no):** 0.82
- **F1-score (no):** 0.89

- **Precision (yes):** 0.38
- **Recall (yes):** 0.80
- **F1-score (yes):** 0.52

Confusion Matrix:



ROC Curve:

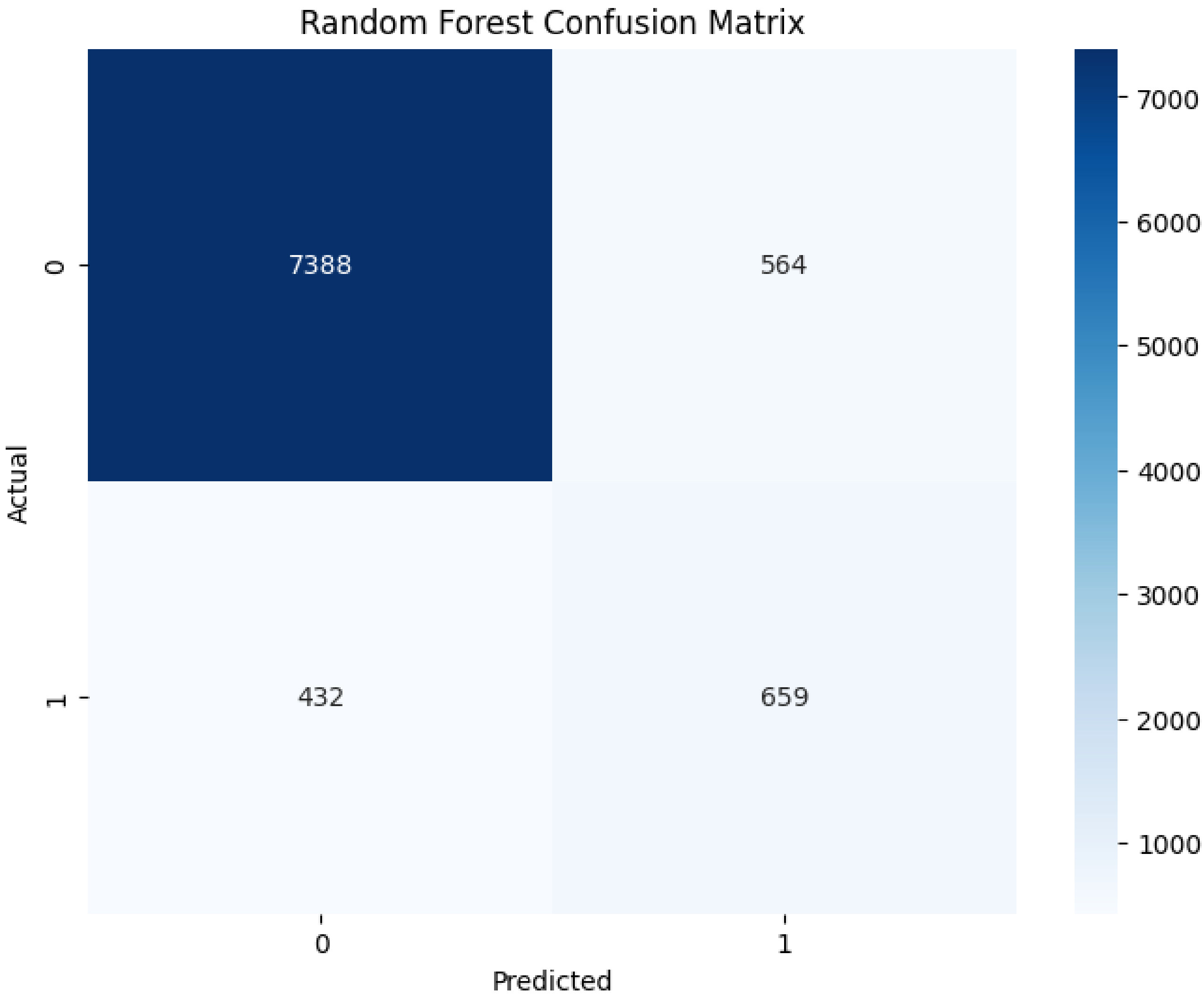


Random Forest Classifier

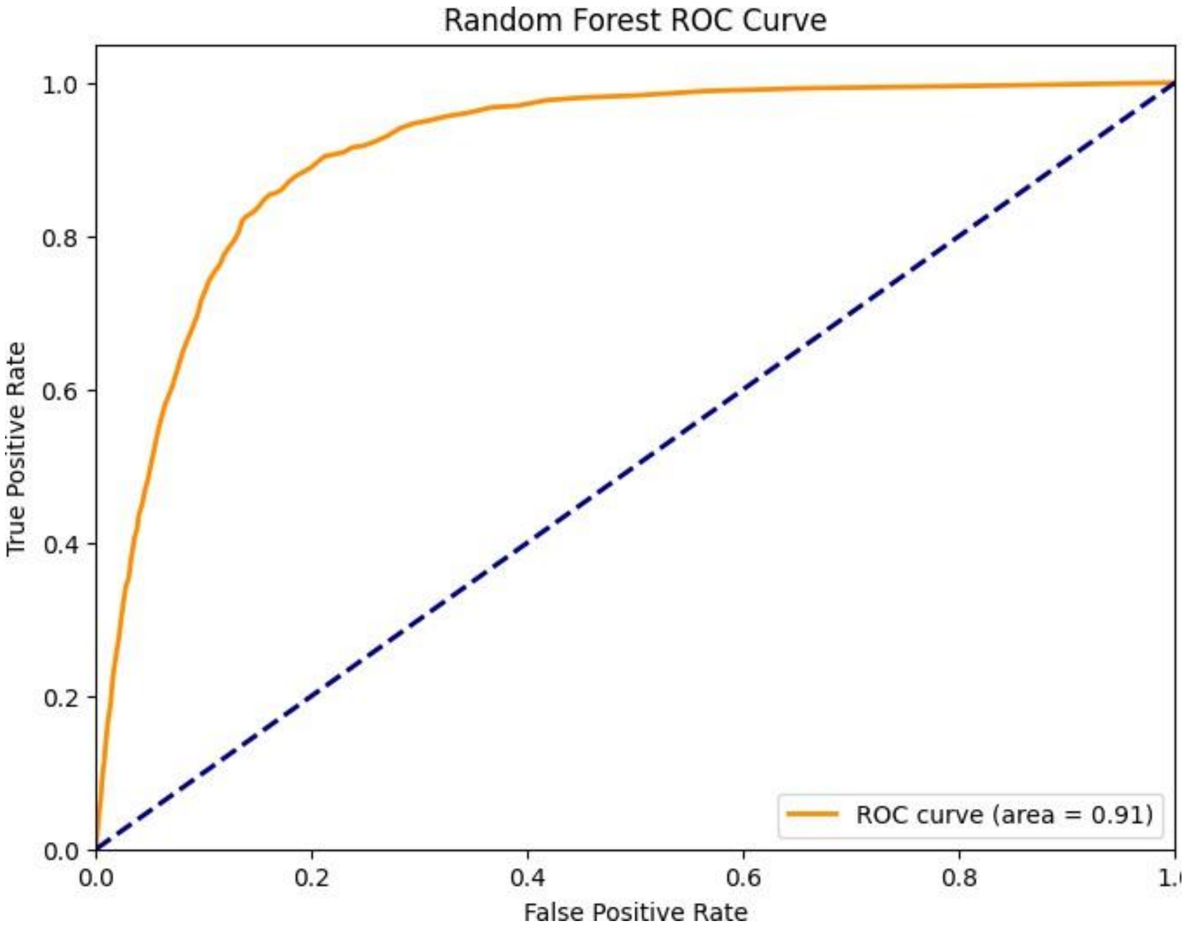
Performance Metrics:

- **Accuracy:** 89%
- **Precision (no):** 0.94
- **Recall (no):** 0.93
- **F1-score (no):** 0.94
- **Precision (yes):** 0.54
- **Recall (yes):** 0.60
- **F1-score (yes):** 0.57

Confusion Matrix:



ROC Curve:

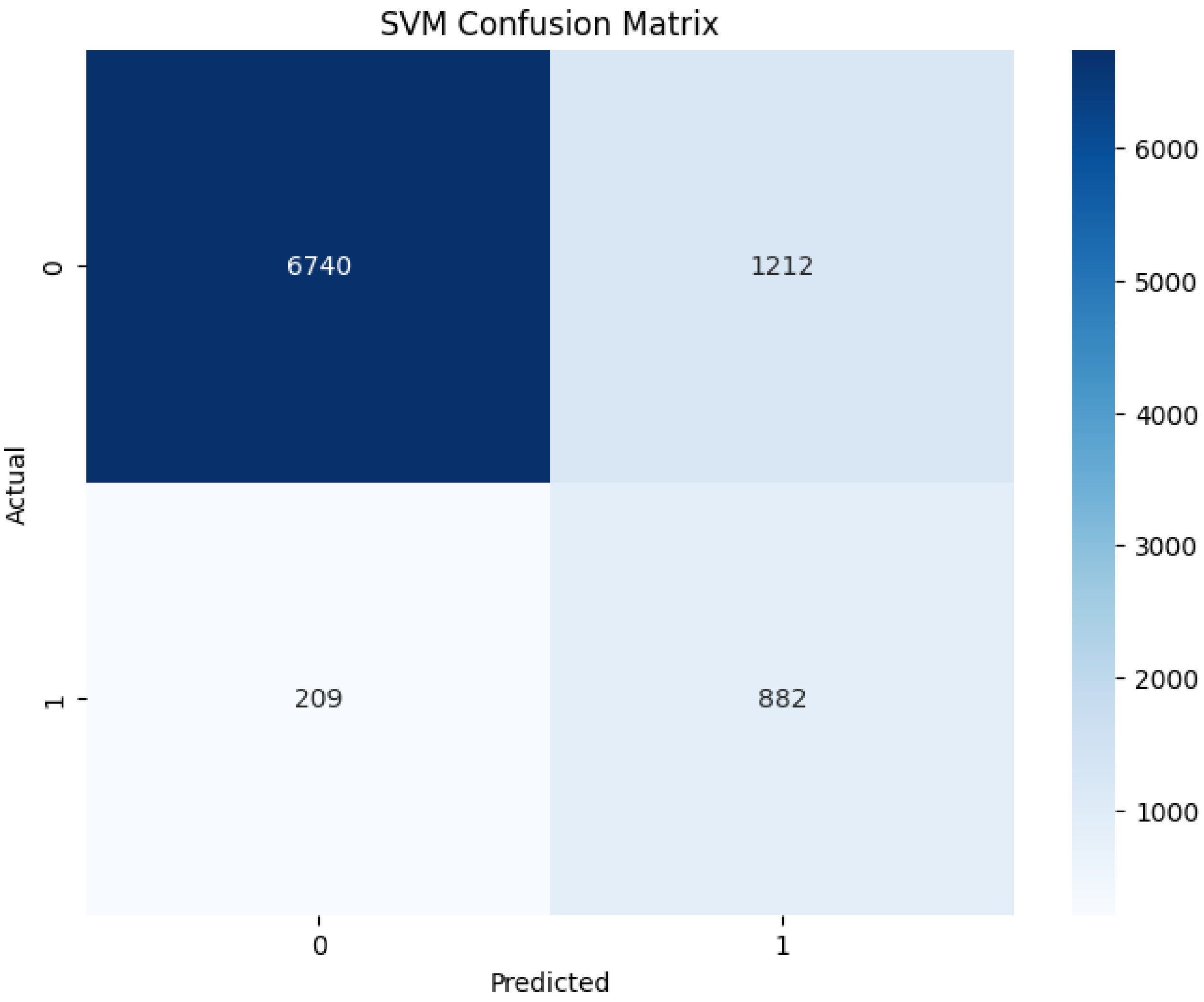


SVM (RBF Kernel)

Performance Metrics:

- **Accuracy:** 84%
- **Precision (no):** 0.97
- **Recall (no):** 0.85
- **F1-score (no):** 0.90
- **Precision (yes):** 0.42
- **Recall (yes):** 0.81
- **F1-score (yes):** 0.55

Confusion Matrix:



Recommendations

Based on the evaluation metrics and ROC curves, the Random Forest Classifier performed the best with an accuracy of 89%, followed by the SVM with 84%, and Logistic Regression with 82%. Random Forest also had a good balance between precision and recall for both classes.

Logistic Regression showed high precision but lower recall for the positive class, indicating that it missed a significant number of 'yes' responses.

SVM had a balanced performance but took longer to train and evaluate.

4. Follow Up & Evaluation Plan

Further Steps

For future work, the following steps are recommended:

1. **Hyperparameter Tuning:** Further tuning of hyperparameters for all models, especially Random Forest and SVM, to improve performance.
2. **Ensemble Methods:** Exploring ensemble techniques such as stacking or boosting to combine the strengths of multiple models.
3. **Feature Engineering:** Investigating additional feature engineering techniques to uncover more predictive features.
4. **Cross-Validation:** Implementing k-fold cross-validation to ensure that the model's performance is consistent across different subsets of the data.

Recommendations:

Based on the evaluation, the Random Forest model demonstrates the highest overall performance with a balanced trade-off between precision and recall. Therefore, we recommend implementing the Random Forest model for predicting future client subscriptions.

Action Plan:

1. **Model Deployment:**
   - Deploy the Random Forest model in a real-time prediction environment.
   - Integrate the model with the bank's marketing system to assist in campaign decisions.
2. **Performance Monitoring:**
   - Continuously monitor the model's performance using real-time data.
   - Regularly update the model with new data to ensure its accuracy and reliability.
3. **Campaign Strategy:**
   - Use the predictions to target clients with a higher likelihood of subscribing to term deposits.
   - Customize marketing strategies based on client profiles predicted by the model.
4. **Feedback Loop:**
   - Collect feedback on the campaign outcomes.
   - Refine the model and marketing strategies based on feedback and performance metrics.

Conclusion

By implementing predictive analytics through the Random Forest model, the bank can significantly improve its marketing efficiency and success rate. The comprehensive evaluation and visualization of the models provide a solid foundation for data-driven decision-making in future marketing campaigns.

This report outlines the detailed process from data preprocessing to model evaluation and the actionable insights derived from the analysis, fulfilling the requirements of the project.

Berk Apak

Google colab link:

https://colab.research.google.com/drive/1AyxtXMeAZ9L9up75C-zJgxA7EhVdRwaP?usp=sharing