****

**AIB504**

**Machine Learning in Business**

**Group-based Assignment**

**July Semester 2023**

**Group Members:**

**Leader: THNG CHEE SING (B2270173)**   
**SER JIE HUI (Y0805238)   
HENG MING EN (WANG MING'EN) (B2272487)   
LIU KAILIANG (B2372684)**

**Submission Date: 24 October 2023**

**A. Introduction - Bank Marketing**

**1. Background and Motivation:**

EU banks experienced a deposit surge from June 2017 to June 2021, particularly post-COVID. The deposit increase was driven by increased savings and reduced spending due to uncertainty about the economy's future. (*Europe Bank Deposits After COVID-19*, n.d.).

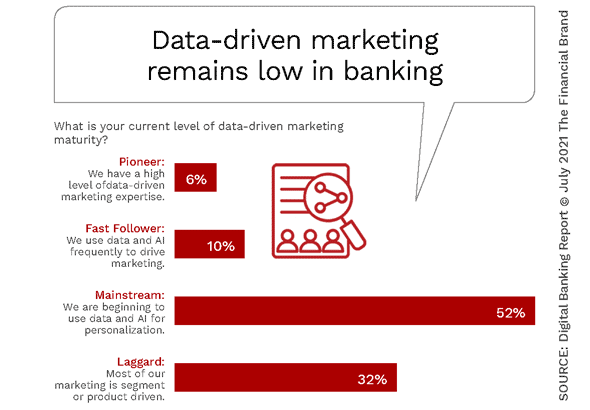
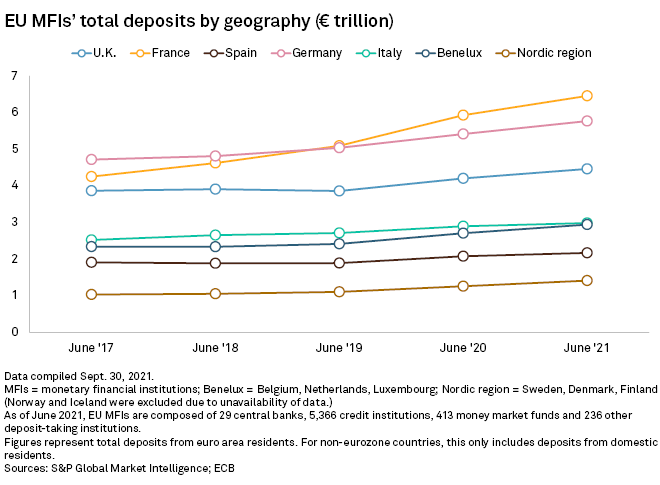


Fig 1 Fig 2

Fig. 1. EU deposit trend by geography regarding trillion (*Europe Bank Deposits After COVID-19*, n.d.)

Fig 2. Data-driven marketing maturity statistics (Marous, 2022) illustrate that the banking sector needs to catch up in adopting data-driven marketing.

Banks continue to utilise telemarketing for product and service promotion, but this method is often inefficient, time-consuming, and ineffective in acquiring new clients.

**2. Business Problem:**

Portuguese banks must adapt to new strategies and utilise machine learning for predictive insights to enhance their ability to promote term deposits and compete with the rest of EU banks.

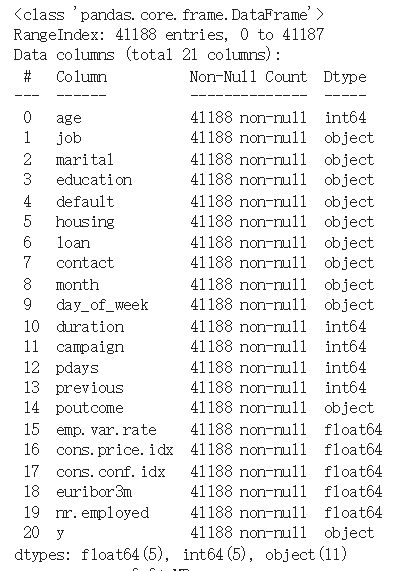
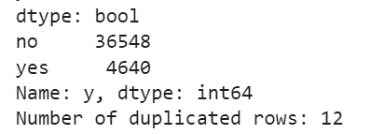
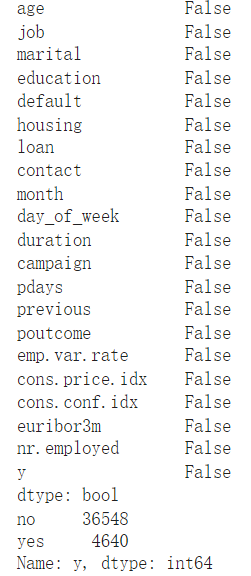
1. **Targeted Marketing:** Banks can improve conversion rates by targeting clients based on model predictions rather than a general approach.
2. **Optimisation of resources:** Predictive predictions are crucial for optimising bank resources, particularly telemarketing personnel, to enhance cost efficiency and return on investment by focusing on likely subscribers.
3. **Personalisation of Client Interactions:** Machine learning enables banks to identify client data patterns and adjust their interactions, enhancing their offerings' appeal.

Thus, the bank's promotional activities must be enhanced to increase efficiency and effectiveness by devising a machine-learning model to identify clients' responses to term deposit marketing campaigns.

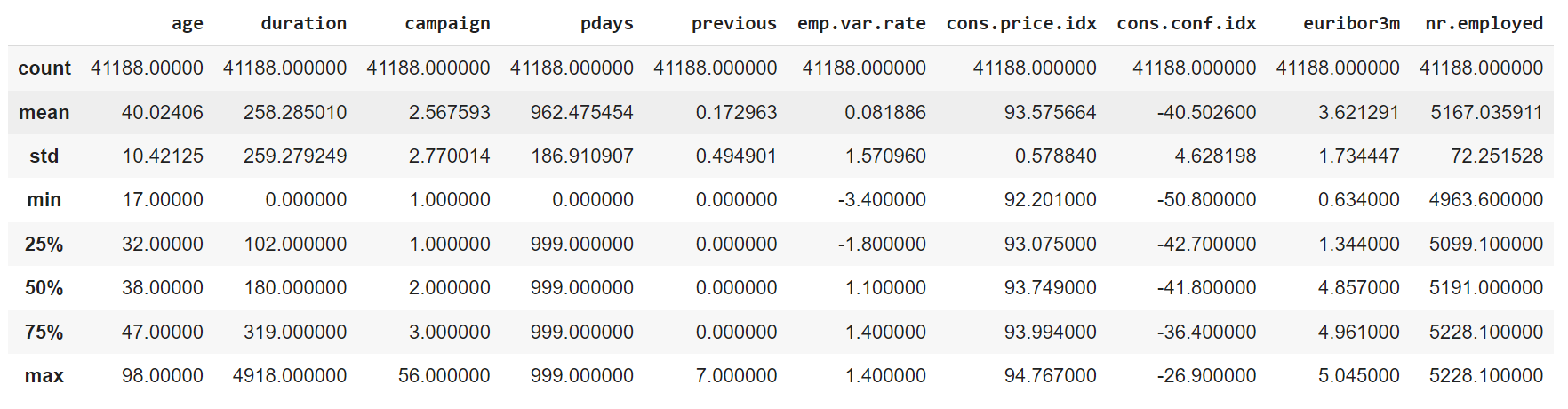
**B. Data Understanding and Preparation**

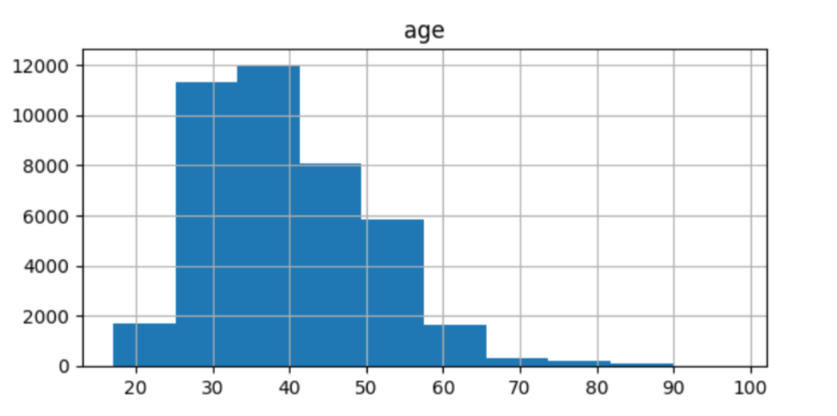
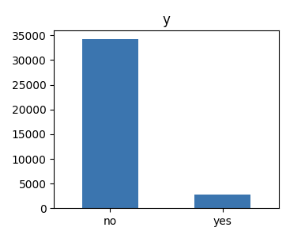
**1. Data Characteristics:**

The dataset contains 41,188 bank client records, including 20 input attributes, with a target variable labeled "y" indicating term deposit subscription with no missing values but 12 duplicate rows have been identified. The dataset comprises both numeric and categorical attributes.

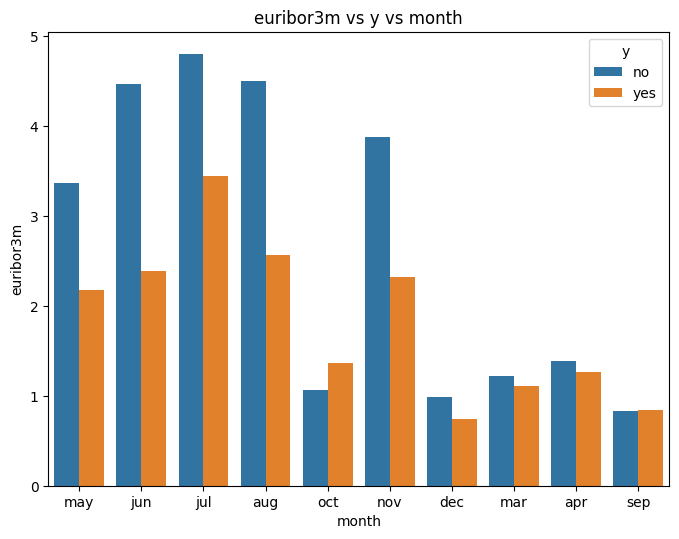
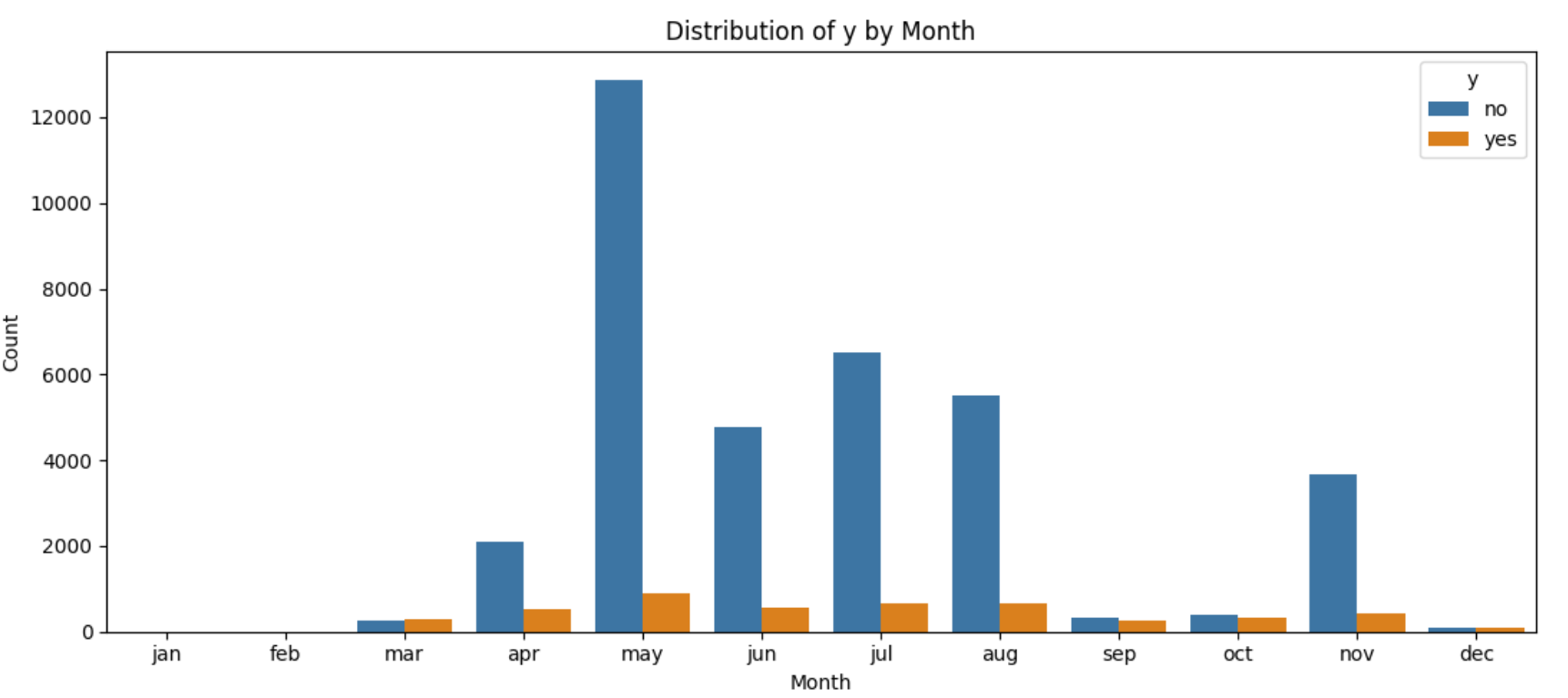
 

**2. Data Exploration:**

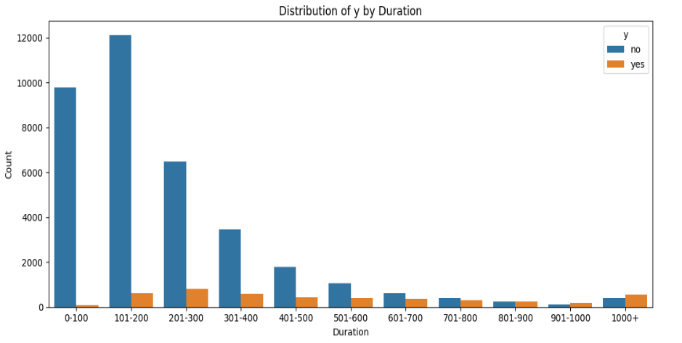




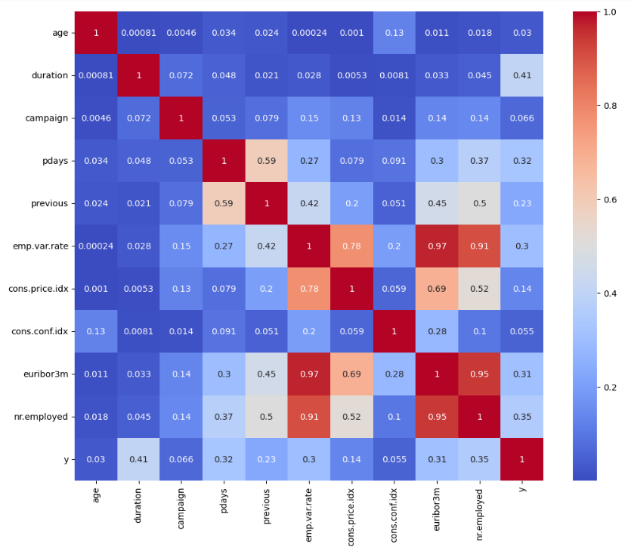
The dataset is imbalanced, with most values being "no" in column 'y'. Clients are mostly in their 20s to 40s, whereas younger individuals lack investment experience, and older ones are hesitant to invest in risky investments.



The histogram of the distribution of y by month indicates that the telemarketing campaign should focus on April to August for optimal marketing success, as Portuguese bonuses are typically paid in June, aligning with the histogram of euribor3m vs y vs month. (Dzhambazova, 2023)



The histogram of the distribution of y by duration reveals that telemarketing calls typically last 101-200 seconds, with calls lasting 201-300 seconds having the highest success in attracting clients, aligning with the mean duration of 258 seconds, suggesting that telemarketers should aim for 201-300 seconds.



The correlation heatmap reveals duration as the most correlated feature with target variable y, followed by nr.employed, suggesting increased client confidence in a bank with more employees, and euribo3m, indicating the importance of deposit interest.

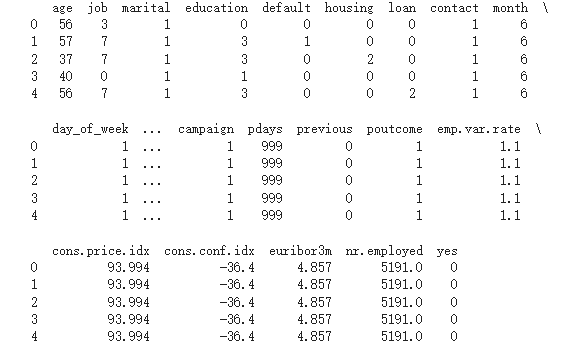
**3. Preparation**:

**Delimiter:** To separate the semicolon character and split the values in each row into separate columns.

**Drop duplicates:** To remove duplicate rows.

**LabelEncoder:** Categorical variables need to be converted into dummy variables for machine learning models. Columns\_to\_encode include 'job','marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', and 'day\_of\_week'. Converting them into dummy variables expands the number of columns to around 50. A label encoder labels each category within each column, maintaining consistency.

**Train test split**: The dataset is split into a training and testing set at 80%-20% ratios, using SMOTE oversampling and RandomUnderSampler to balance the training set. The numerical features are standardised using StandardScaler.



**C. Model Design and Result Illustration**

**1. Model Design:**

Given that our target variable is binary ("yes" or "no"), we implemented a classification approach using binary variables, including Logistic Regression, Decision Trees, Random Forest, and Support Vector Machine, Multi-layer Perceptron, and Voting Classifier, to efficiently identify new clients likely to invest in term deposits, optimising resources, enhancing its promotion and competitiveness with EU banks.

**a) Logistic Regression (LR):**

The algorithm is a widely used model suitable for the banking sector for binary classification problems, predicting client acceptance of term deposits and enhancing promotional strategies.

* **Predictive Output:** LR provides a probability score between 0 and 1 for each customer, showcasing their likelihood to subscribe.

**b) Decision Trees (DT):**

The algorithm effectively classifies datasets by dividing them into subsets based on feature values, resulting in a tree that can be easily visualized and interpreted to show the relationships.

* **Predictive Output:** Binary outputs based on subscription likelihood and also offer probability scores to gauge the prediction's certainty.

**c) Random Forest (RF):**

Random Forest, a combination of multiple decision trees, effectively prevents overfitting and offers more accurate predictions than a single decision tree.

* **Predictive Output:** Similar to DT, binary outputs are based on subscription likelihood and also offer probability scores to gauge the prediction's certainty.

**d)**  **Support Vector Machines (SVM):**

Due to their strength in high-dimensional spaces, SVMs are effective in binary classification of diverse, non-linear data like telemarketing. We hope they can identify hidden patterns by finding the hyperplane that separates the datasets.

* **Predictive Output:** Determines the hyperplane that separates the data and classifies clients based on this boundary.

**e) Multi-layer Perceptron (MLP):**

MLP can discover non-linear patterns and complex relationships in the data, which is helpful for telemarketing data.

* **Predictive Output:** A neural network's prediction on whether a customer will subscribe or not based on the learned patterns in the data.

**f) Voting Classifier:**

Ensemble models, like voting classifiers, enhance bank predictions by combining multiple models' strengths, reducing potential shortcomings and providing more reliable predictions.

* **Predictive Output:** A collective decision based on the predictions from the ensemble of models.

**2. Model Construction:**

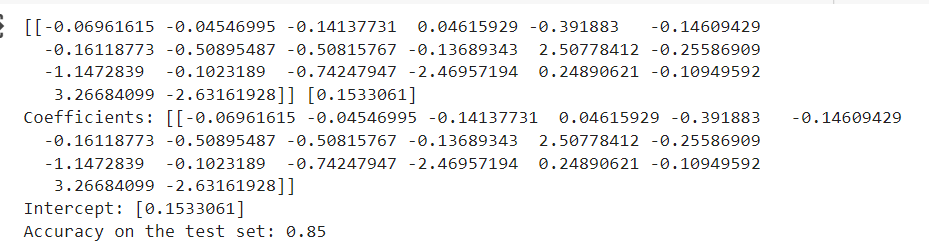
All models were constructed using Google Colab, ensuring a seamless environment for data processing and model building. Here are the parameter settings for each:

* **LR:** max\_iter: 1000 (We had increased it to try to ensure convergence due to the large size of the dataset), random\_state: 42 (For consistent results across the runs)
* **DT:** full tree and pruned tree max 5 at 42 default settings with Fig 40,40.
* **RF:** n\_estimators=100, max\_features = "sqrt", random\_state=42
* **SVM:** kernel='rbf', random\_state=42
* **MLP:** random\_state=1, hidden\_layer\_sizes= (20,10), activation = 'relu', max\_iter=2000
* **Voting Classifier:** Comprised of the above models with soft voting to ensure probabilities were considered.

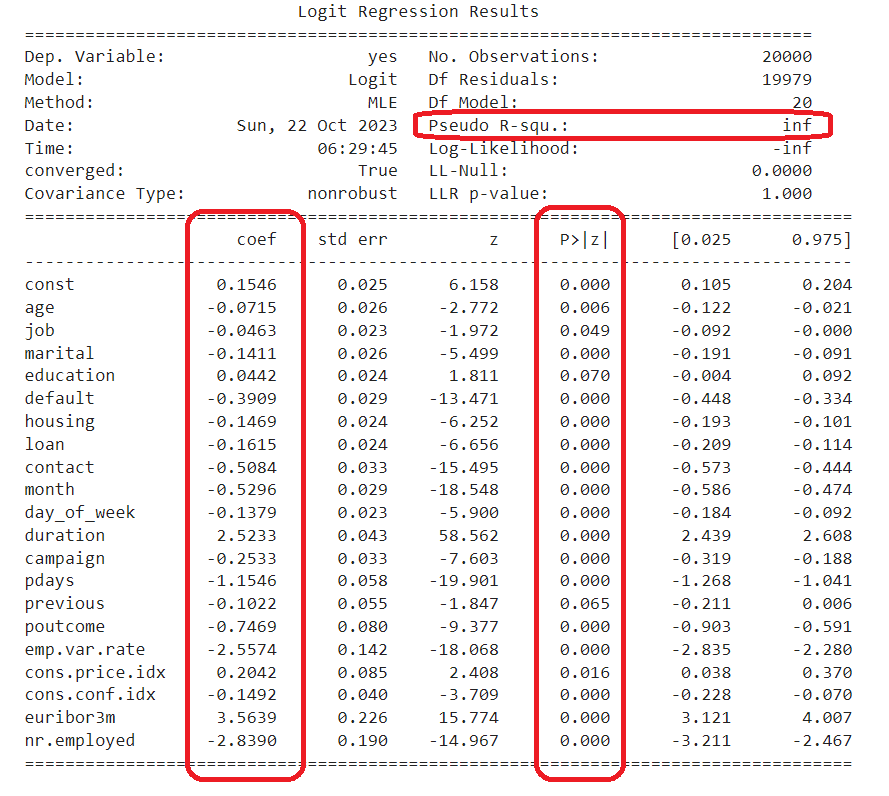
**3. Model Evaluation and Interpretation:**

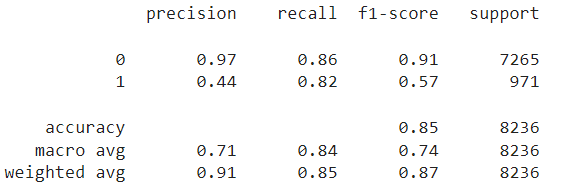
**a) LR:**

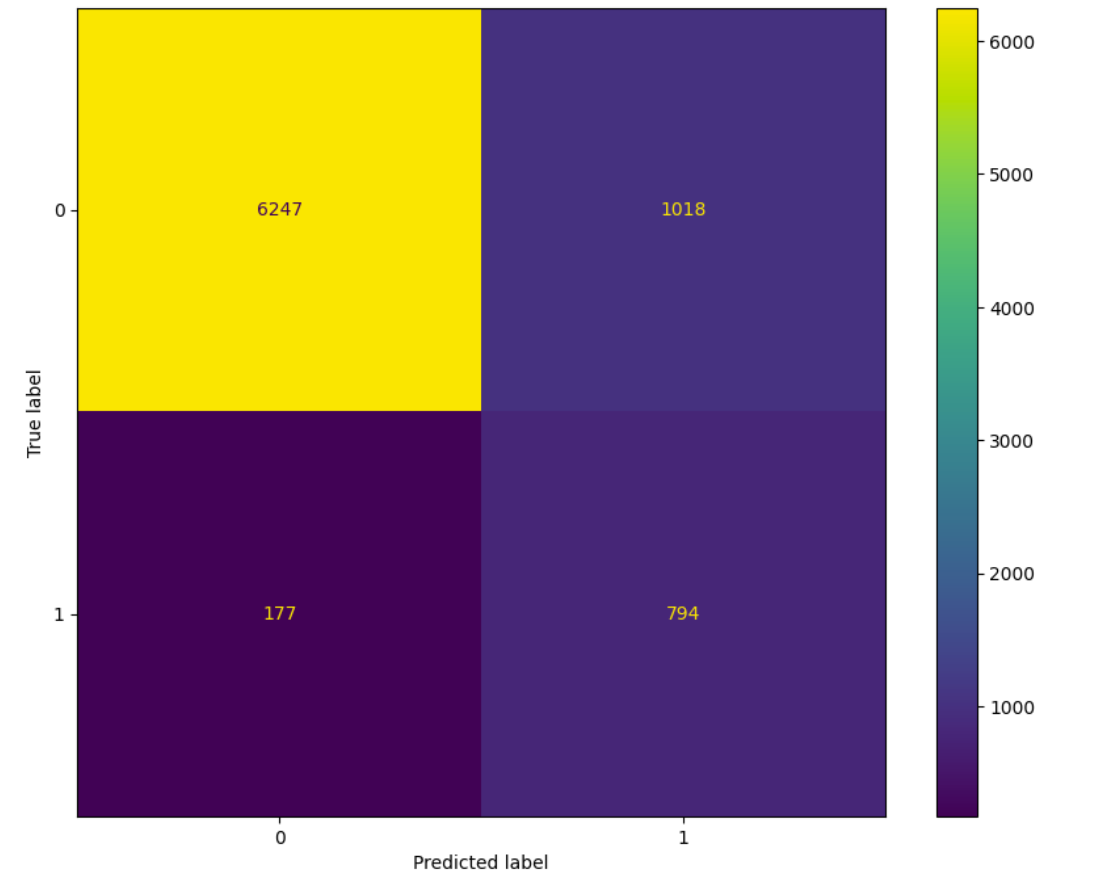
* **Performance:** Achieved an accuracy of 85% on the test set.
* **Interpretation:** The logistic regression model's coefficients reveal which features significantly influence predictions, such as "euribom3", suggesting higher interest increases the likelihood of accepting term deposits, suggesting banks should telemarket more during high-interest periods.
* **Model results:**
* There seems to be more features with a negative influence on the target variable. This can be seen by more features having negative coefficients.
* The intercept, which is the baseline prediction, is 0.1533061.



* **Statsmodel:** The statsmodels Logit model was initialised using the default parameters.
  + Pseudo R-squared: It measures the goodness of fit. Our result is an infinity value, which is one of the challenges which we had faced.
  + Coefficients: The majority of the features have negative coefficients
  + P>|z|: Most variables have p-values less than 0.05, indicating statistical significance, while 'education' and 'previous' have p-values higher than 0.05, suggesting they may not be significant predictors.

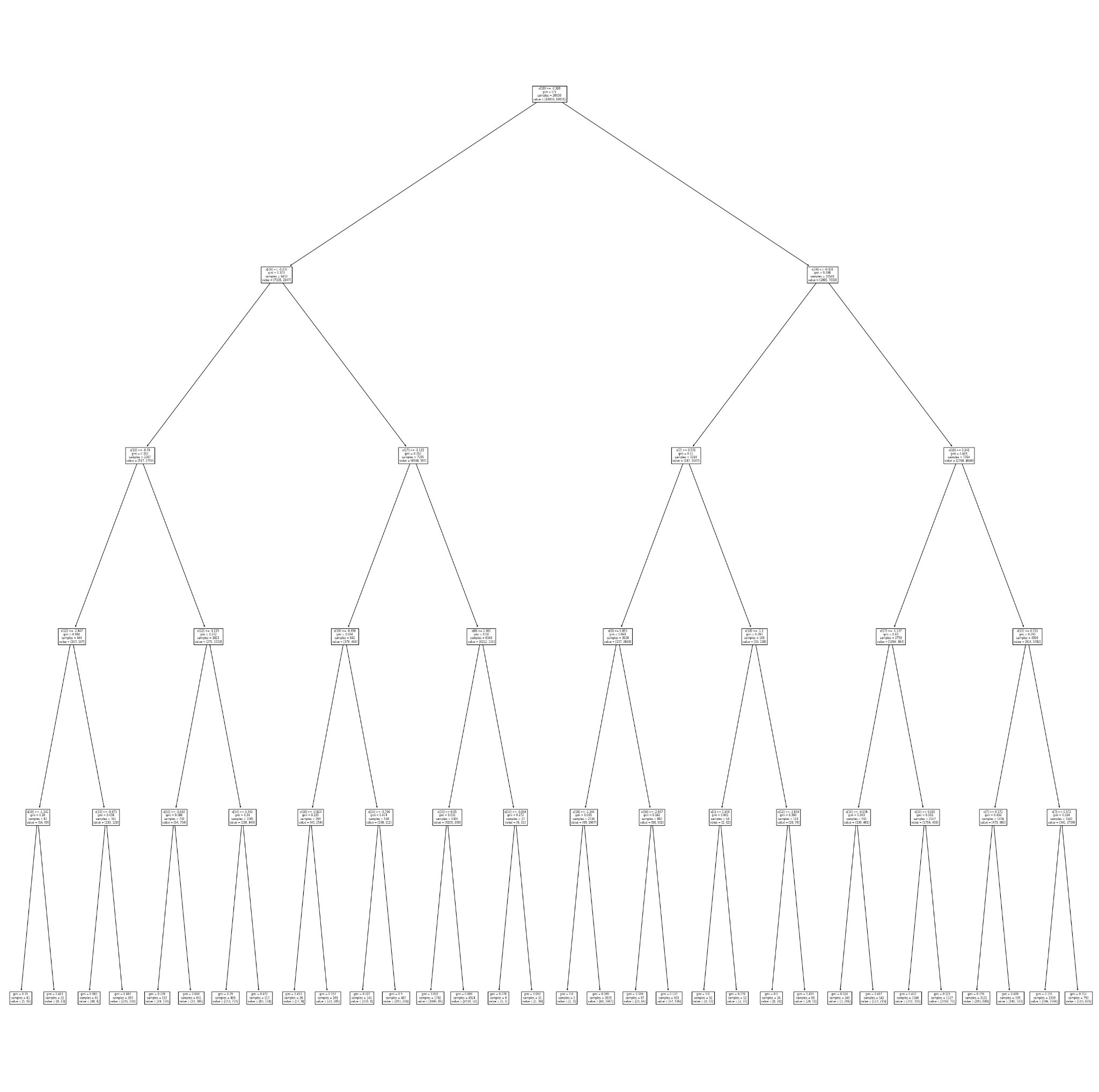
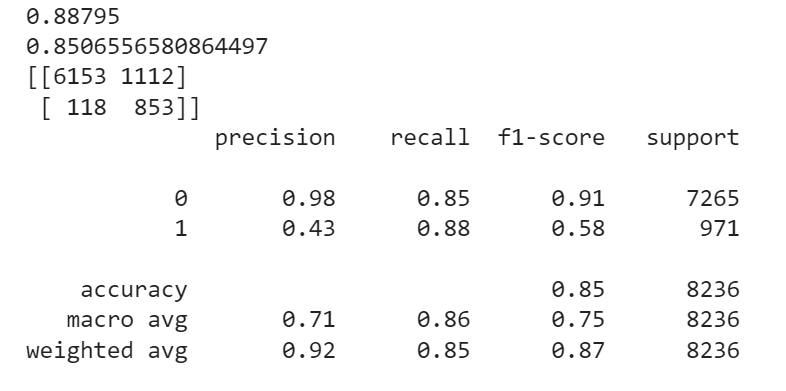
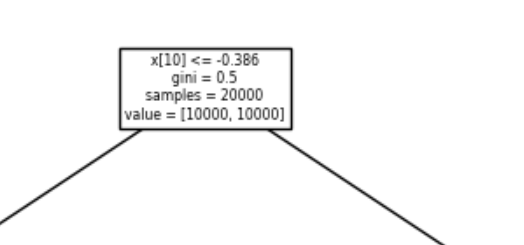


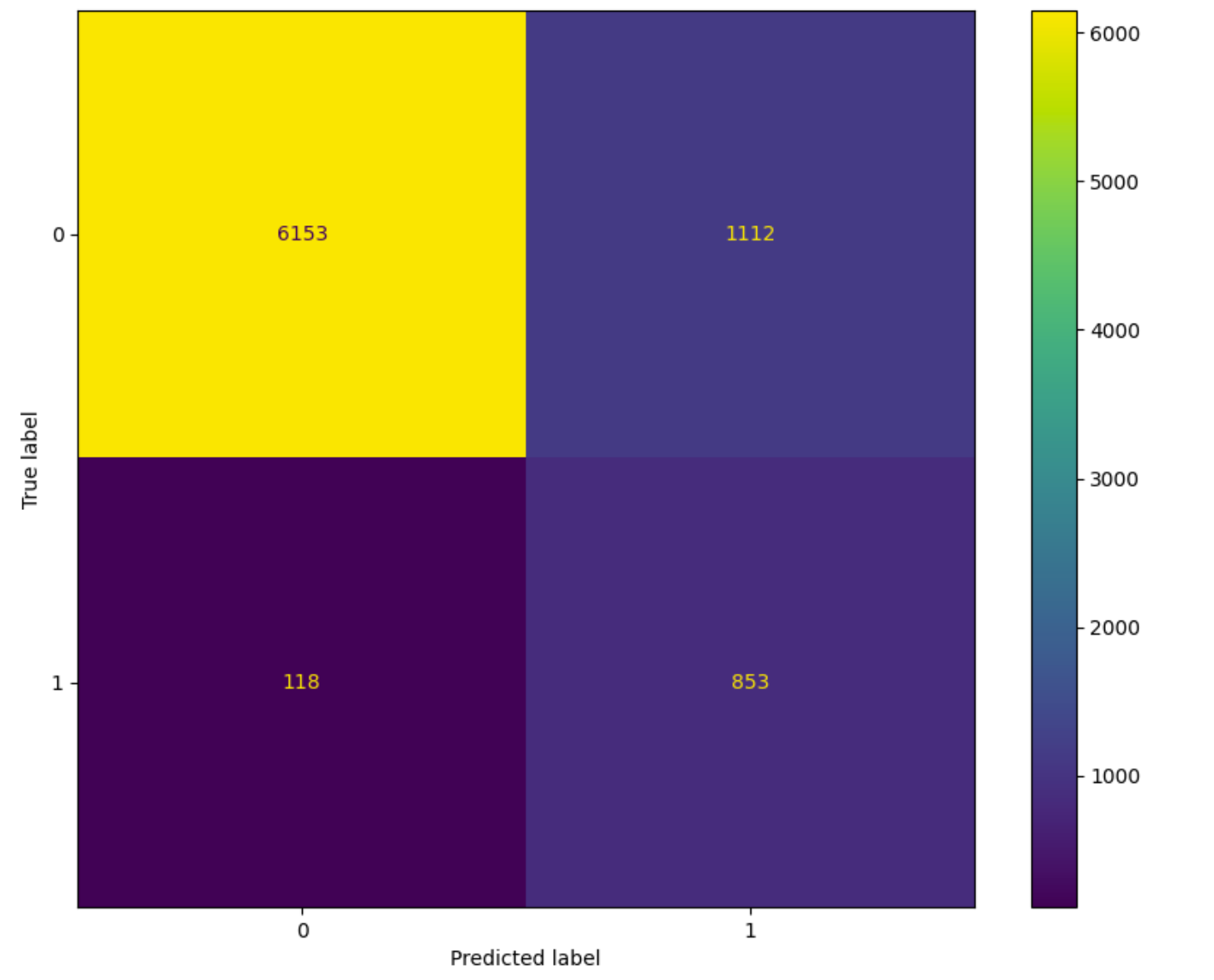
* **Model Performance Evaluation:** The logistic regression model excels in identifying Class 0 ('No') with an F1-Score of 91 but struggles to distinguish Class 1 ('Yes') with an F1-Score of 57%, despite undersampling and oversampling was carried out to try to balance the data.
* Confusion Matrix Graph: The graph shows similar information where the model struggles to identify Class 1 where there are more False Positives than True Positives.



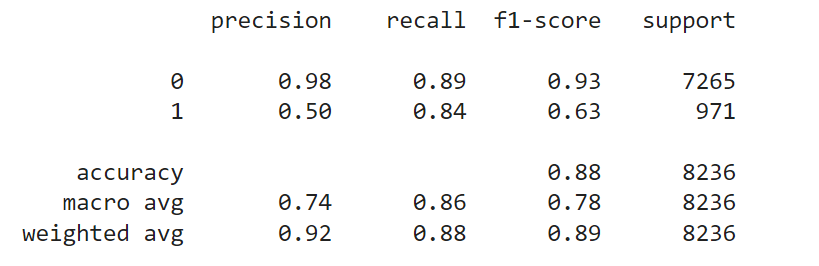
**b) DT:**

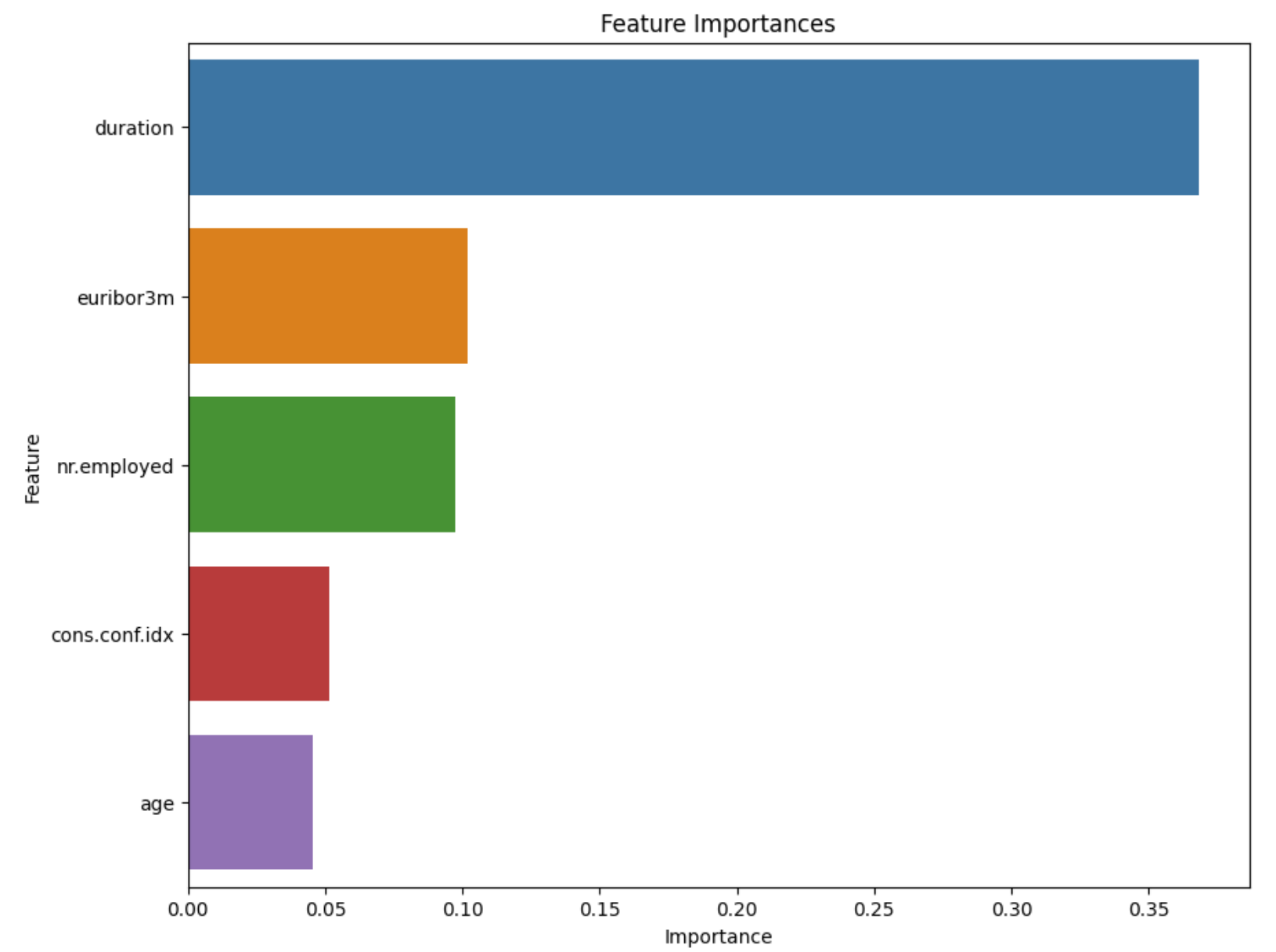
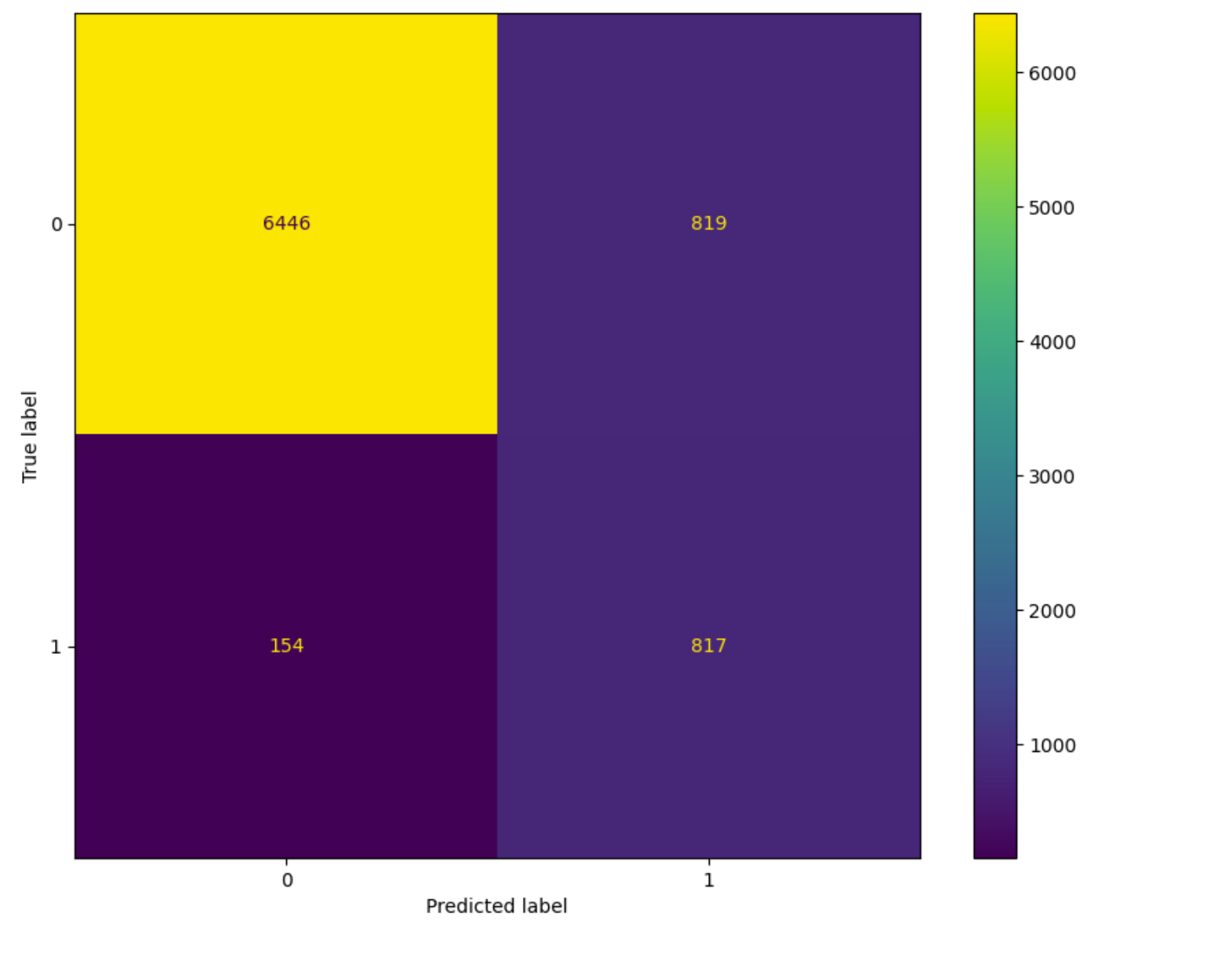
* **Performance:** The pruned tree reached an accuracy of 88.795% on training data and an accuracy of 85.066% for testing data. It is noted that DT struggles to identify true positives.
* **Interpretation:** Decision trees aid banks in understanding decision-making processes, with critical factors like "duration" / "x[10]" at the top node. This insight suggests that banks should train telemarketers to engage clients longer, as highlighted in data exploration.

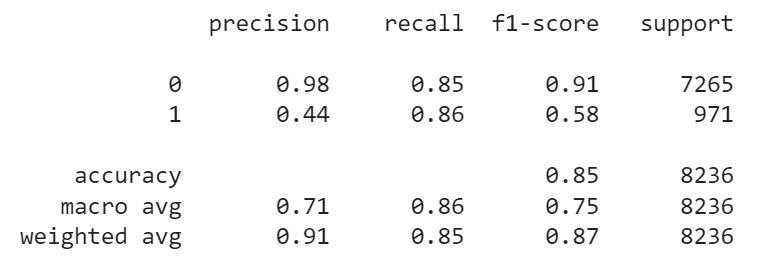


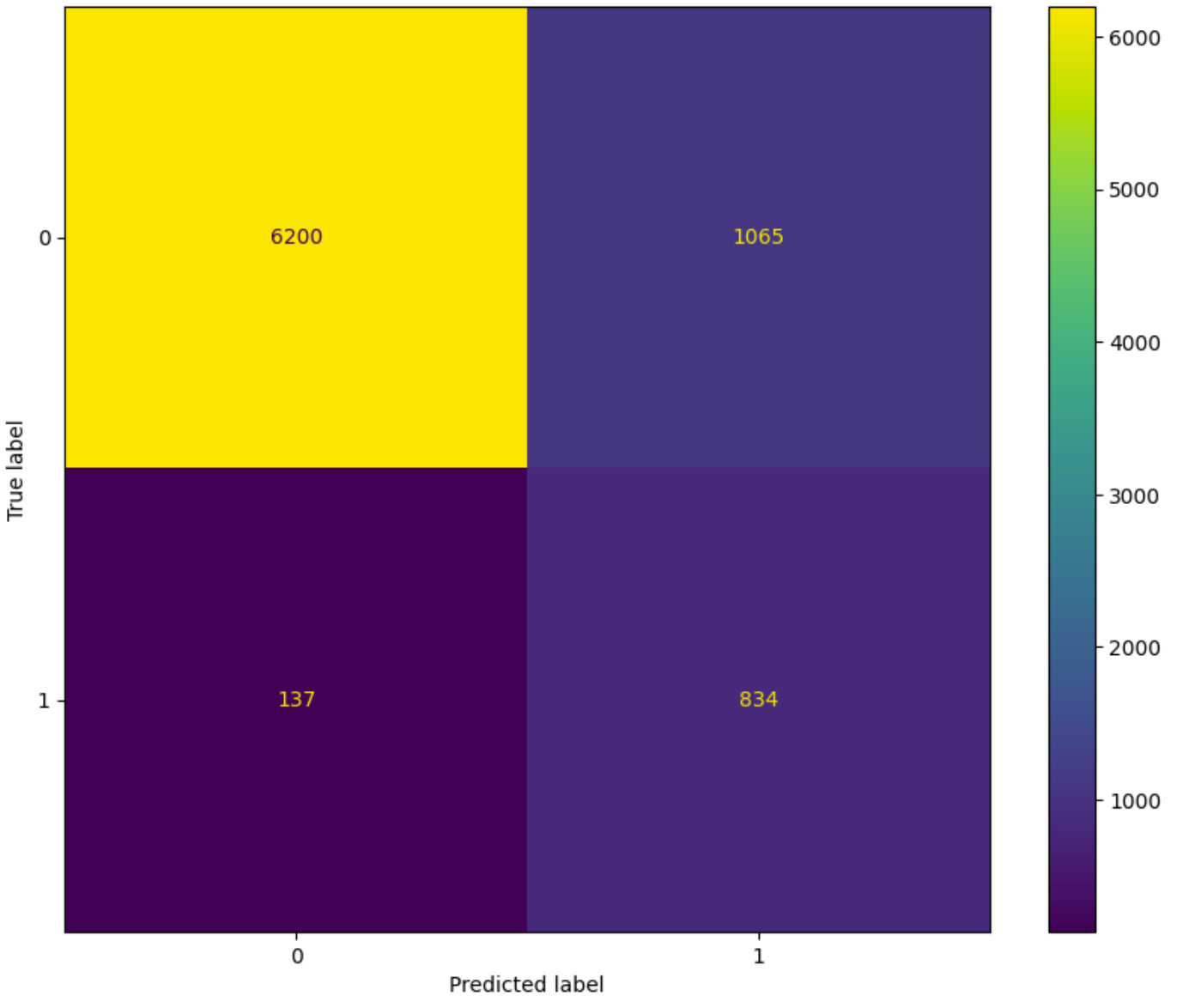
**c) RF:**

* **Performance:** Accuracy of 88%. RF handles slightly better in identifying true positives.
* **Interpretation:** Random forests also give a feature importance metric. In this instance, it is the same as the decision tree, where 'duration' is the most important.

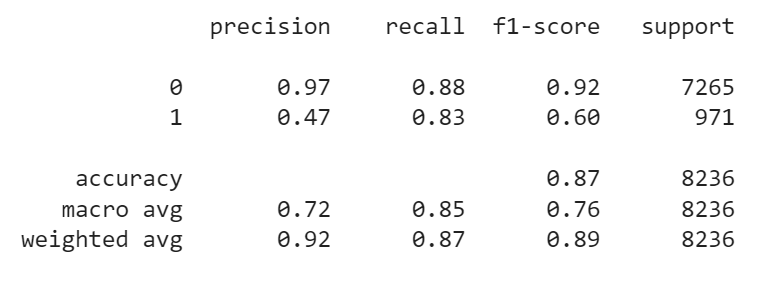
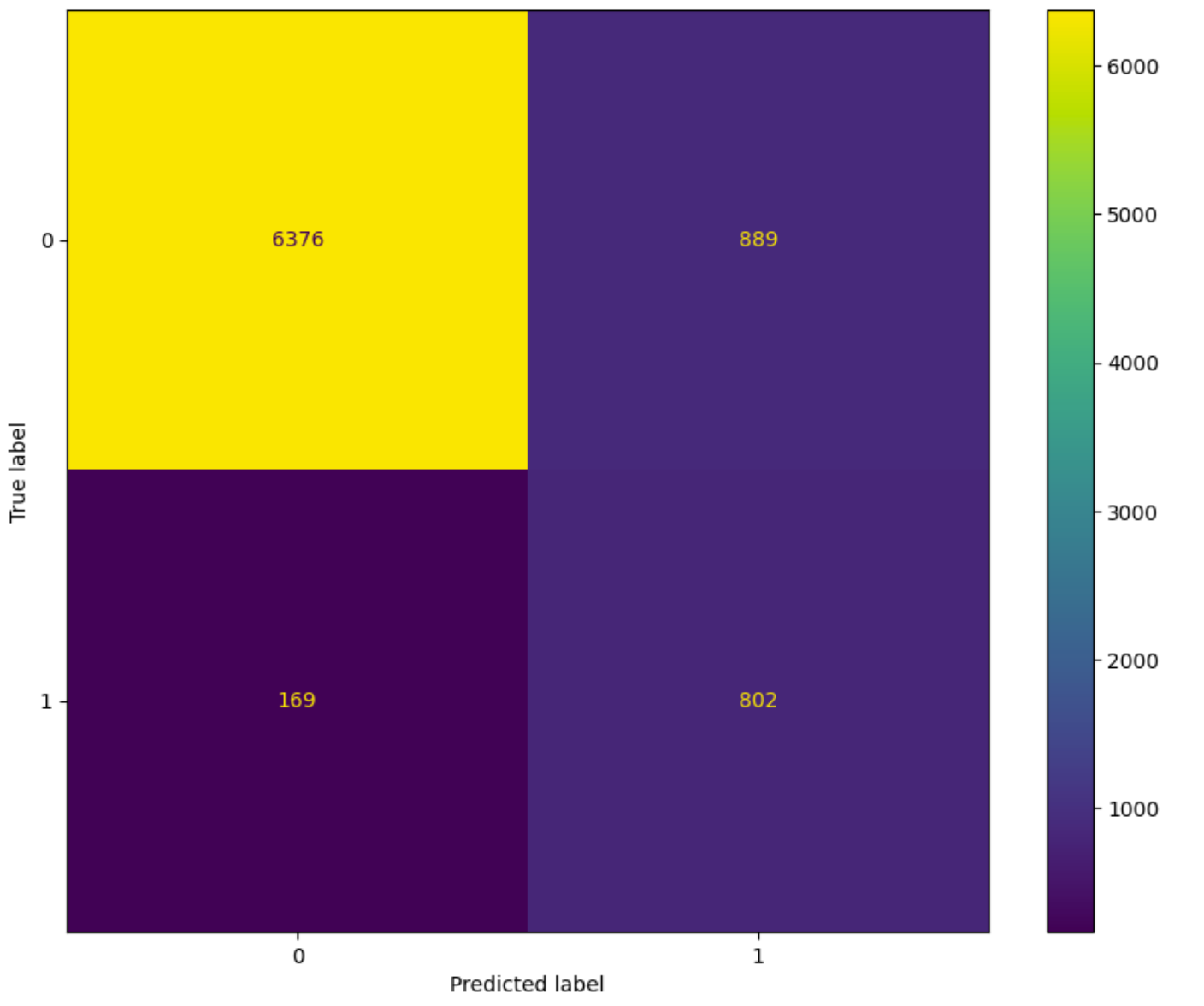
**d)**  **SVM:**

* **Performance:** SVM has an accuracy of 85% but similarly struggles with true positives.
* **Interpretation:** SVMs, while powerful, are less interpretable directly. However, their strength lies in working efficiently with higher-dimensional data, which means they can handle many features and interactions effectively. Their success might indicate complex relationships in the data, suggesting banks incorporate more nuanced strategies.



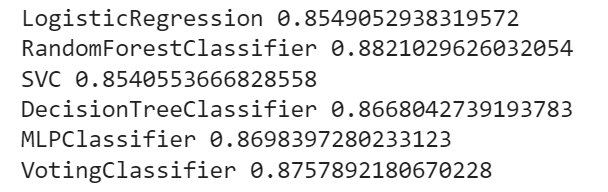
**e) MLP:**

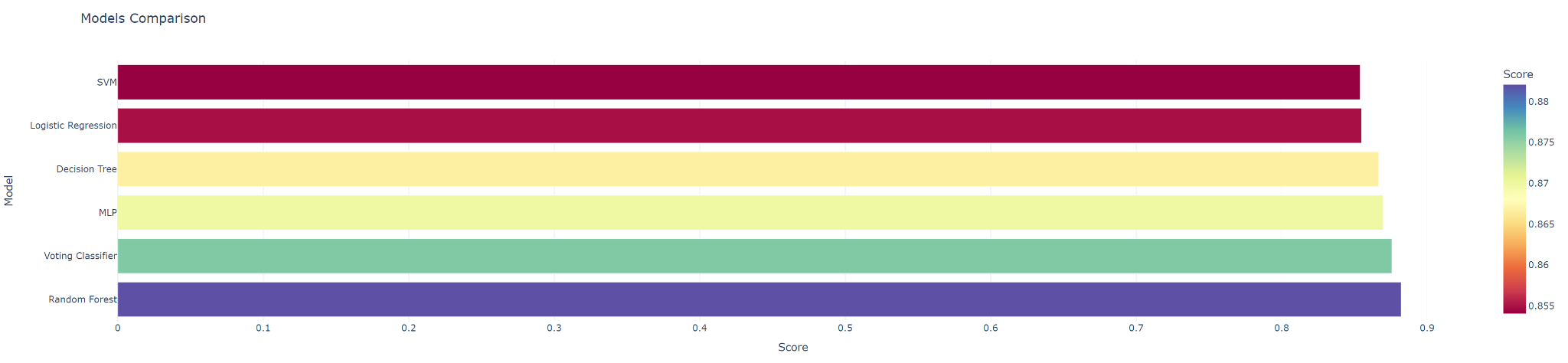
* **Performance:** MLP, with its deep architecture, achieved an accuracy of 87%.
* **Interpretation:** The neural network has three layers: an input layer, a first hidden layer with 20 neurons, and a second hidden layer with 10 neurons. Banks can invest more to analyse more in-depth data to unearth these hidden patterns and refine their strategies.

**f) Voting Classifier:**

* **Performance:** Surpassing most models, except RF, achieved an accuracy of 87.86%.
* **Interpretation:** The ensemble method combines the strengths of various models, ensuring a more robust prediction. Its success underscores the idea that a multi-faceted approach to understanding customer behaviour can be more effective, drawing insights from different models. However, it is noted that RF has the highest accuracy.





**Overview**

Random forest, with the highest accuracy score, is the most effective model for identifying potential clients to subscribe to term deposits. It estimates key variables for classification, crucial for targeted marketing, resource optimisation, and client interaction personalisation. The model's accuracy on the test set is approximately 88.2%.

**D. Discussion & Concluding Remark**

**1. Summary and Implications:**

**Overview:** Our research commenced with the challenge faced by Portuguese banks where conventional marketing avenues are inefficient in promoting term deposits, especially post-COVID-19. To address this, we employed a suite of machine learning algorithms, from the fundamental Logistic Regression to ensemble methodologies like the Voting Classifier.

**Patterns and Discoveries:**

* **Influential Variables:** The research uncovered a significant relationship between variables and the likelihood of a client subscribing to a term deposit. For instance, attributes like call duration, euribo3m interest rates, and number of employees in the bank held pronounced influence in our models.
* **Model Variances:** Each model has its unique strengths and shortcomings. For instance, while Decision Trees provided a transparent, intuitive understanding of decisions, they required pruning to avoid overfitting. The Voting Classifier, conversely, consolidated strengths from various models. However, its accuracy was below Random Forest, where it also has the highest F1-score for identifying true positives.

**Implications and Deployment Suggestions:**

* **Tailored Marketing:** Banks can design personalised campaigns by understanding call duration and training telemarketers for more extended conversations or using digital platforms for interactive engagement with potential clients.
* **Resource Optimisation:** Our predictive models enable the bank to use machine learning algorithms to identify potential clients, prioritise leads, improve resource allocation, minimise wasted efforts, and increase objectives.

**2. Limitations and Concluding Remarks:**

**Limitations:**

* **Data Imbalance:** The inherent imbalance in the dataset, with fewer subscribers, poses challenges. Although we employed techniques like SMOTE, synthetic data might not capture all nuances of the minority class.
* **External Factors:** While our models account for several influencing variables, significant events like global economic downturns, pandemics, war, or geopolitical events are beyond the predictive scope of our current models.
* **Scope of Data:** The data is specific to Portuguese banks and might not cater to global banking network, highlighting the need for region-specific datasets and models when expanding scope.
* **Relevancy:** Data and patterns observed are bound by time, and economic, social, and technological landscapes changes over time. The data collected from 2008 to 2013 might be less relevant today.

**Concluding Remarks and Further Steps:**

In the intricate dance of numbers, patterns, and predictions, what stands out is the transformative capability of data. Our research, grounded in rigorous analysis, paves the way for Portuguese banks to rethink and reshape their marketing strategies.

Future endeavours could explore:

* **Global Analysis:** The Machine Learning model is currently limited to Portugal's banks, but it could be applied globally for financial product telemarketing, considering cultural norms and country regulations, for example in Singapore the PDPA and Do-Not-Call registry, among others. (PDPC | Do Not Call Registry and Your Business, n.d.)
* **Feedback Loop Integration:** Integrating a feedback loop where models are retrained based on their prediction outcomes for a truly adaptive system could ensure continuous learning and adaptation can help ensure the banks stay relevant and updated.

The data-driven approach is not a one-time venture. It requires consistent updating, re-evaluation, and fine-tuning to stay relevant.

**References:**

*Europe bank deposits after COVID-19*. (n.d.). S&P Global Market Intelligence. Retrieved October 21, 2023, from <https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/europe-bank-deposits-continue-to-increase-after-covid-19-spike-66729421>

Marous, J. (2022, October 3). *Banking Needs To Prepare For Marketing’s Data Arms Race*. The Financial Brand. <https://thefinancialbrand.com/news/data-analytics-banking/banking-financial-marketing-data-arms-race-ai-maturity-trends-119312/>

Moro, S., Cortez, P., & Rita, P. (2014, June 1). *A data-driven approach to predict the success of bank telemarketing*. Decision Support Systems; Elsevier BV. <https://doi.org/10.1016/j.dss.2014.03.001>

H. (2020, July 10). *Bank Marketing + Classification + ROC,F1,RECALL. . .* Kaggle. <https://www.kaggle.com/code/henriqueyamahata/bank-marketing-classification-roc-f1-recall>

S. (n.d.). *GitHub - sukanta-27/Predicting-Success-of-Bank-Telemarketing: Given the Bank customer relationship data, predict whether a customer will subscribe to a term deposit or not.* GitHub. <https://github.com/sukanta-27/Predicting-Success-of-Bank-Telemarketing>

H. (2018, June 8). *Bank Marketing - Imbalanced Dataset 94%*. Kaggle. <https://www.kaggle.com/code/henriqueyamahata/bank-marketing-imbalanced-dataset-94>

Rajawat, A. S. (2023, July 7). *Voting Classifiers in Machine Learning - Amit Singh Rajawat - Medium*. Medium. <https://medium.com/@imamitsingh/voting-classifiers-in-machine-learning-a532935fe592>

*Bank Telemarketing Analysis : Predicting customers’ responses to future marketing campaigns*. (n.d.). <https://yfsui.github.io/Bank-Telemarketing-ML-Project/>

Dzhambazova, I. (2023, July 24). *Benefits*. Boundless. <https://boundlesshq.com/guides/portugal/benefits/>

R, A. (2023, October 6). *How Long Should A Cold Call Last? [Original Study]*. Klenty Blog. <https://www.klenty.com/blog/how-long-should-a-cold-call-last/>

*PDPC | Do Not Call Registry and Your Business*. (n.d.). https://www.pdpc.gov.sg/overview-of-pdpa/do-not-call-registry/business-owner/do-not-call-registry-and-your-business