Bank Marketing

Hasan Ahmed Shaik
School of Computing

Shoaib Mohammed
School of Computing

Nikhila Veera Chandini Pothula
School of Computing

Venkata Subba Rao Are
School of Computing

***Abstract* — In the contemporary banking industry, targeted marketing is crucial for enhancing customer engagement and reducing operational inefficiencies. This project offers a data-centric method to forecast success of telemarketing initiatives designed to promote long-term deposit subscriptions. Utilizing an actual dataset from a Portuguese retail bank, we implemented a Decision Tree model to pinpoint customers who are likely to make term deposits. Through data preprocessing, feature encoding, and model assessment using metrics such as accuracy, precision, recall, classification report, ROC-AUC, and confusion matrices, the Decision Tree model achieved a prediction accuracy of 76.4%. Our findings indicate that smart customer targeting can considerably boost marketing effectiveness, lower call volumes, and enhance conversion rates. This study establishes a basis for future developments in hybrid modeling, real-time prediction systems, and responsible AI implementation in the financial sector.**

Keywords-Bank Marketing, Term Deposit, Prediction, Machine Learning, Targeting, Telemarketing Optimization, Predictive Modeling, Classification Algorithms, SMOTE, Precision, Accuracy, Recall, F1-Score, EDA, ROC-AUC, Random Forest, Classification Report, Data Preprocessing, Model Evaluation Metrics, Feature Encoding, Data Science in Banking

I. Introduction

In the competitive banking sector, acquiring and maintaining customers are crucial elements of an effective marketing strategy. Telemarketing continues to be a commonly utilized avenue for reaching out to customers. Nevertheless, random cold-calling not only causes customer dissatisfaction but also leads to resource waste and low conversion rates. This project tackles these issues by employing data-driven methods to enhance the efficiency of telemarketing efforts for term deposit subscriptions.

We suggest a predictive modeling technique that makes use of historical customer and campaign information to pinpoint potential customers who are likely to respond favorably to marketing calls. By doing this, we intend to minimize unnecessary customer interactions and boost conversion rates. Our method employs several machine learning algorithm - Decision Tree assesses their effectiveness in forecasting customer reactions.

II. Ease of Use

The created predictive system is intended with practical usability as a priority, making it possible for banking professionals with limited technical expertise to effectively use the model. The design is modular, permitting smooth incorporation into pre-existing banking software systems or call center applications.

A. User Interface (optional)

Although the existing version is based on code, it can be easily transformed into a web interface by utilizing frameworks such as Flask or Streamlit, facilitating user-friendly engagement through straightforward inputs and one-click predictions.

B. Input Simplicity:

The model needs merely a subset of customer data, including age, occupation, marital status, education, prior contact outcome, and campaign-related information. These data points are already gathered during telemarketing activities, ensuring minimal disruption to current processes.

C. Fast Predictions:

After being trained, the model generates real-time predictions with minimal delays, making it ideal for immediate decision-making during customer interactions.

D. Interpretability:

While the Neural Network offers the highest level of accuracy, more straightforward models like Decision Trees are also available, delivering interpretable results for users who value transparency.

E. Scalability:

The model can scale larger datasets or be incorporated into cloud-based systems with slight adjustments. This emphasis on user-friendliness guarantees that the solution is not only technically robust but also practical and prepared for application in actual banking scenarios.

III. RELATED WORKS

Several methods have applied machine learning techniques to improve marketing effectiveness. Below, we group prior work thematically and highlight gaps in our approach addresses.

A. Neural-Based Approaches

Moro et al. [1] showed how neural networks performed better than traditional classifiers such as logistic regression and decision trees when supplemented with feature selection, highlighting the importance of economic indicators and call-length features; however, they didn’t consider class imbalance. Subhani [2] used multilayer perceptron models with feature reduction and imputation approaches; however, he reported overfitting in smaller data samples, highlighting a need for more effective resampling methodologies.

B. Ensemble and Kernel Methods

Tang and Zhu [3] investigated ensemble learners, reporting that Random Forests and support vector machines (SVM) achieved accuracies above 90% after extensive hyperparameter tuning and data normalization. They achieved notable gains in revenue uplift, yet their work assumed balanced classes and did not evaluate precision–recall trade-offs for the minority “subscribe” class.

C. *Imbalance Handling and Hybrid Models*

Sonkar et al. [4] tackled class imbalance using SMOTE and integrated Naive Bayes with clustering, achieving robust performance metrics such as a G-mean of 0. 743 and 80. 8% accuracy.

D. *Positioning our Work*

While prior research has shown the promise of neural, ensemble and hybrid methods, few studies paired the decision tree with resampling to address interpretability and class-imbalance. Our Study implements Decision Tree Classifer enhanced by SMOTE and feature encoding. Evaluation Metrics like Accuracy, Precision, Recall,F1-Score, ROC-AUC and confusion matrix.

IV. About the Dataset

The dataset used in this project is from a Portuguese retail bank and it has been sourced from the UCI Machine Learning Repository focusing on promotions related to term deposits in the bank. There are four dataset files provided for this project and two are for training and two are testing purposes which are given below:

bank-additional-full.csv:41188 examples with 20 inputs recorded from May 2008 and November 2010

bank-full.csv: dataset with 17 inputs, ordered by date (older version of this dataset with less inputs).

bank-additional.csv: dataset with 10% of the examples (4119), randomly selected from 1), and 20 inputs.

bank.csv: dataset with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

We have merged the dataset by removing the additional features present in each other. The resultant merged dataset consists of 86399 observations with 15 input features common across the datasets provided. It also includes data specific to the campaigns, such as the type of contact, duration of calls, and results from prior campaigns.

Features:

- *Demographic:* age, job, marital, education
- *Financial:* default, housing, loan
- *Communication:* contact, month, duration
- *Campaign Details:* campaign, pdays, previous, poutcome
- *Target:* y – whether the customer subscribed to a term deposit

The dataset's target variable “y” indicates whether a client subscribed to a term deposit (yes/no). A key feature is its class imbalance most clients didn’t subscribe which means accuracy alone won’t cut it when evaluating models. This makes it a solid test case for classification algorithms and alternative metrics like precision or recall.

The data is clean, with no missing values, and mixes categorical and numerical features. After basic preprocessing, it’s ready for machine learning workflows.

V. Exploratory Data Analysis (EDA)

The goal of EDA is to grasp the dataset’s structure and key characteristics before applying machine learning models. This phase helps identify trends, patterns, anomalies, and relationships among variables.

Key Insights:

A. Target Variable Distribution:

The target variable (y) has a clear imbalance: about 90% of clients didn’t subscribe to a term deposit (class “0”), while under 10% did (class “1”). This uneven split called for SMOTE to balance classes when building the model, since relying only on accuracy could be misleading. After applying SMOTE, both classes were balanced to an even split (approx. 76k samples each).

B. Univariate Analysis:

- Age: Most clients fall within the 30–50 age range , with a peak around 30–40. Outliers exist beyond 70 years old.
- Marital Status: “Married ” is the most common category followed by university degree” (~12,000), and smaller groups like “basic” or “illiterate.”
- Education: “Secondary ” education dominates
- Contact Method : Over 50,000 clients were reached via cellular , compared to ~20,000 via telephone and ~10,000 with unknown contact methods.
- Default/Housing/Loan Status : Most clients have no default (75,000), no housing loan (45,000), and no personal loan (~70,000).

C. Bivariate Analysis:

- Subscription Rates by Contact Method : Clients contacted via cellular showed notably higher subscription rates compared to telephone or unknown methods
- Previous Contact Success : Clients who had a successful prior contact (“poutcome = success”) were more likely to subscribe.
- Call Duration : Longer call durations (duration) strongly correlated with subscription likelihood, suggesting persistence in outreach efforts pays off.

D. Correlation Analysis:con

- Duration & Target Variable : Call duration (duration) exhibits a strong positive correlation (0.40) with the target variable (y), indicating its predictive power.
- Campaign & Previous Contacts : The number of contacts per campaign (campaign) is highly correlated with pdays (days since last contact), suggesting repeated outreach strategies.
- Economic Indicators : While not directly visualized here, external economic factors like euribor3m (interest rates) and nr. employed (employment numbers) are known to influence subscription decisions and should be explored further.
  Figure is correlation matrix of the input features

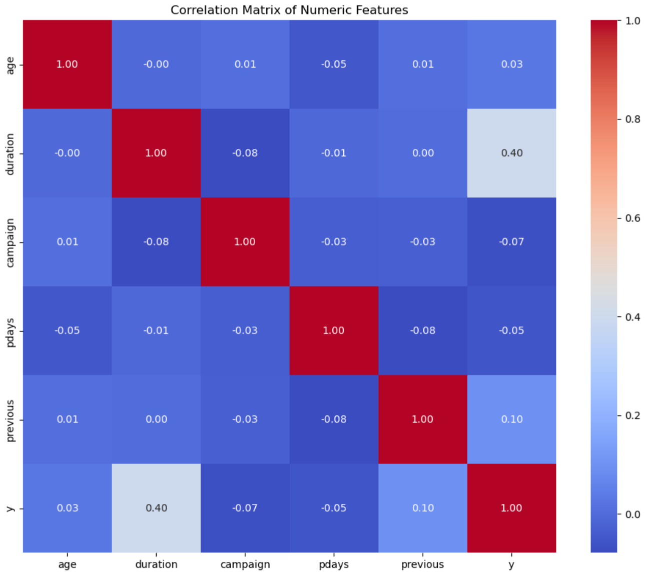
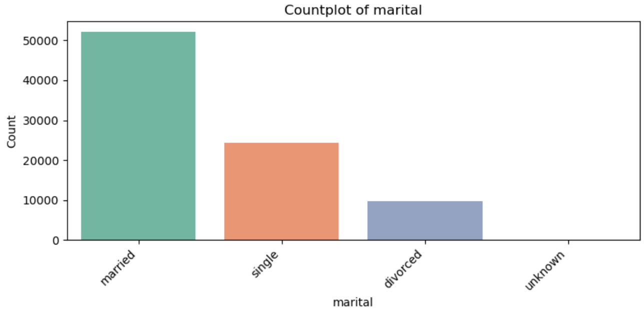
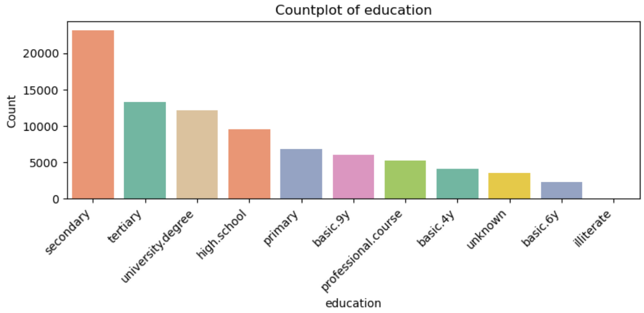
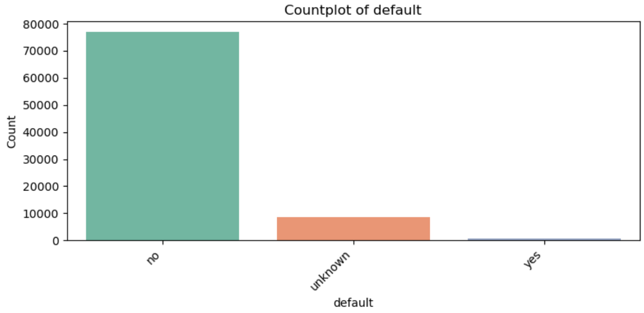


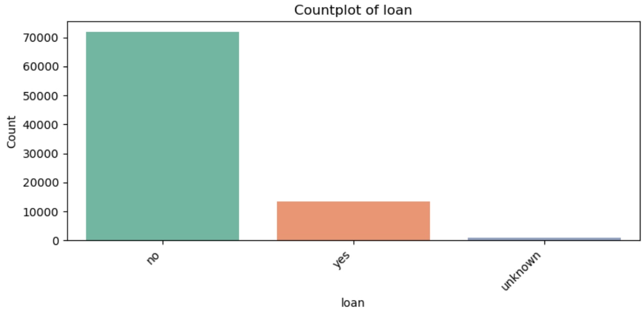
Figure : Correlation matrix of input features in the dataset

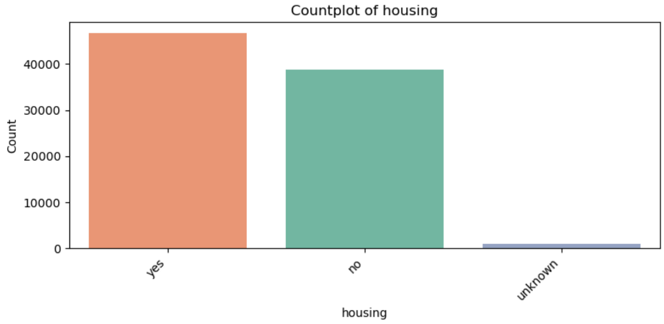
Below is the visualization of different variables in the datasets separated for categorical and numerical features in the dataset

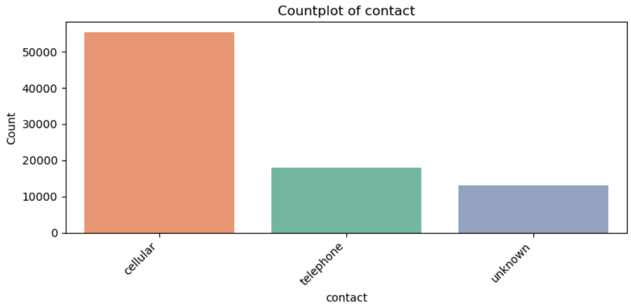


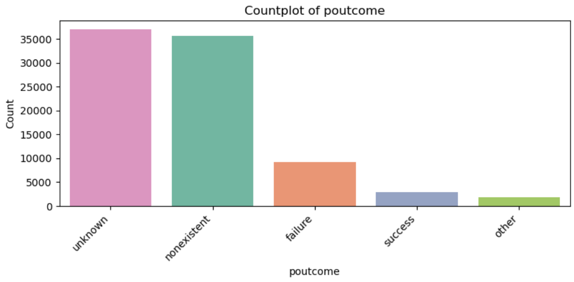












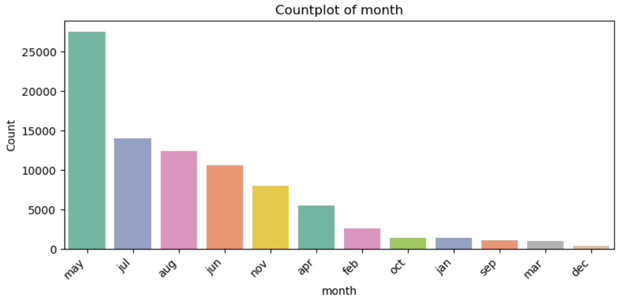


Figure: Count plot of Numerical Columns present in the dataset

For Numerical Columns we used the boxplot and also the distribution in the dataset shown below

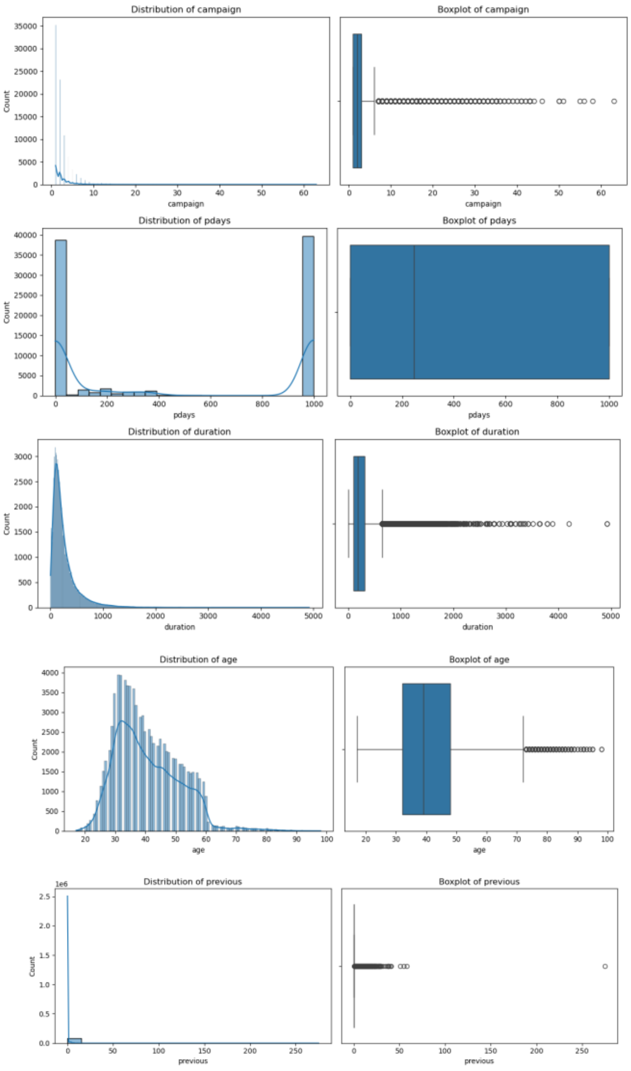


Figure: Distribution of the Numerical columns in the dataset

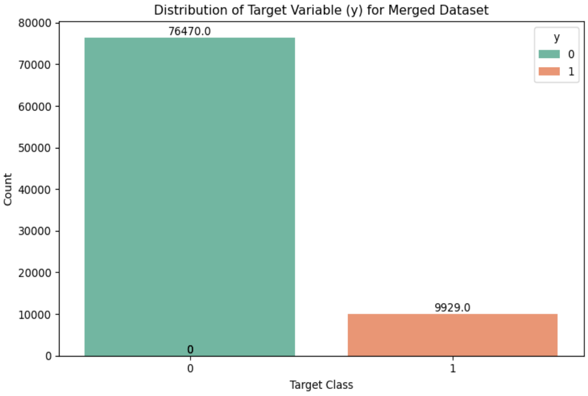


Figure 1: Target variable y Distribution of merged dataset

VI. EXPERIMENTS

A. Resampling Strategy

To correct the skewed distribution between “subscribe” and “no subscribe” classes, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to the training data only. Using imbalanced-learn’s default SMOTE (`sampling_strategy='auto'`) with `k_neighbors=5`, we generated synthetic minority-class samples until both classes were equally represented. This resampling step preceded model fitting and improved Decision Tree’s ability to identify potential subscribers without altering the original test set distribution.

Below is the distribution of output variable “y” after SMOTE

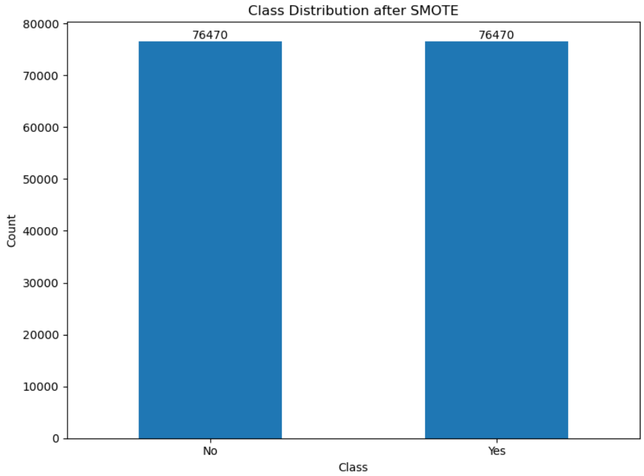


Figure : Distribution of variable “y” after SMOTE

B. Feature Scaling

All numeric predictors were standardized to zero mean and unit variance using scikit-learn’s `**StandardScaler**`. We fit the scaler on the training set only, then applied the learned transformation to both training and test data. This ensured that features with larger ranges did not unduly influence the Decision Tree’s splits and helped stabilize downstream performance metrics.

C. Feature Importance Analysis

After fitting the DecisionTreeClassifier, we extracted the feature_importances array to quantify each predictor’s contribution to reducing node impurity. These values were normalized so that their sum equals one, allowing direct comparison across features. We then sorted features in descending order of importance and visualized the top ten in a horizontal bar chart. This analysis revealed which customer attributes—such as call duration, previous campaign outcome, and economic indicators drove the model’s predictions most strongly.

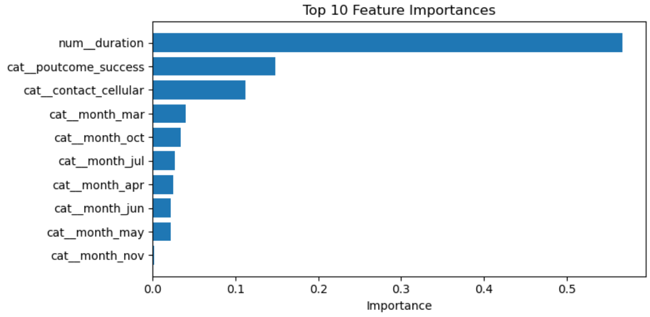


Figure : Feature importance graph for Decision Tree Model

VII. METHODOLOGY AND RESULTS

The methodology details the sequential process utilized to create a predictive model that can determine if a customer will opt for a term deposit based on their personal characteristics and features related to the campaign. This project adheres to a supervised machine learning pipeline, commencing with data cleaning and concluding with model evaluation. The main steps included are:

A. Model Selection

Decision Tree Classifier: A non-linear model that can capture intricate patterns, with increased interpretability through visualizations.

These models were chosen to evaluate linear versus tree-based approaches on this dataset.

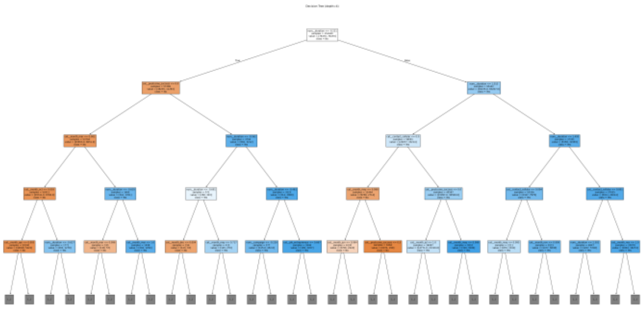


Figure :Plot Diagram of Decision Tree

B. Model Training

Both models were trained on balanced and preprocessed training data. Pipelines were utilized to simplify the preprocessing and training processes, ensuring transformations were applied consistently. Hyperparameters were set manually (or adjusted, where applicable) to enhance model performance.

C. Evaluation

The model underwent assessment utilizing the subsequent metrics:

- Accuracy: Assesses the ratio of correct predictions.
- Precision and Recall: Crucial because of the class imbalance present in the dataset.
- F1-Score: Balances the compromise between precision and recall.
- ROC-AUC Score: Illustrates the model's capacity to distinguish the classes.

Visual instruments such as the Confusion Matrix and ROC Curve were additionally employed to evaluate performance.

VIII. RESULTS

The performance of the Decision Tree Classifier was evaluated using multiple classification metrics to assess its effectiveness on the imbalanced test data. The model achieved a good balance between correctly predicting both subscribed and non-subscribed customers after applying SMOTE.

Key Results:

- Accuracy: 0.7635

Indicates that approximately 76% of all predictions were correct.

- Precision: 0.3073

Among all customers predicted to subscribe, around 30.7% actually did.

- Recall: 0.8786

The model successfully identified about 88% of actual subscribers, which is critical in marketing campaigns where missing a potential customer is more costly than targeting a non-interested one.

- F1 Score: 0.4553

A balance between precision and recall; it reflects the model's ability to minimize both false positives and false negatives.

- ROC-AUC Score: 0.8732

Shows the model has a strong ability to distinguish between the two classes (subscribed vs not subscribed).

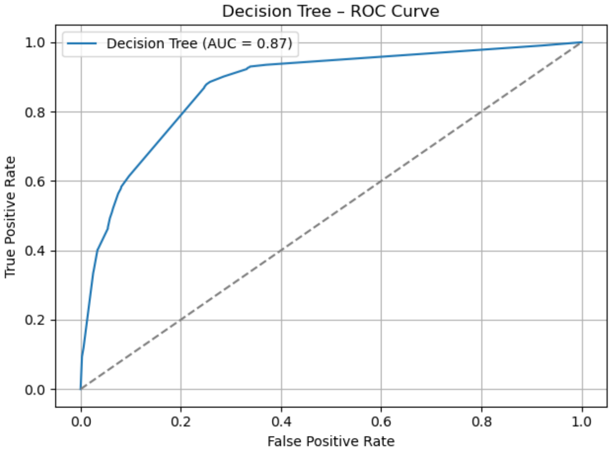


Figure : ROC - AUC Curve of Decision Tree Model

Confusion Matrix or truth table of the decision tree model shown below

Label	Actual Class	Predicted Class
No	5743	1925
Yes	118	854

Table : Confusion Matrix of Decision Tree Model

Below is the classification Report of Decision Tree Model

	Precision	Recall	F1-score	Support
No	0.98	0.75	0.85	7668
Yes	0.31	0.88	0.46	972
Accuracy	-	-	0.76	8640
macro avg	0.64	0.81	0.65	8640
weighted	0.90	0.76	0.80	8640

Table : Classification Report of Decision Tree Model

IX. Discussion

The Decision Tree classifier established a baseline performance (76.4% accuracy, strong ROC-AUC) for predicting term-deposit subscriptions. Key drivers included call duration and prior campaign outcomes , aligning with expectations that longer, targeted conversations improve conversion. Applying SMOTE balanced class distribution, enhancing recall for "yes" responses but introducing minor false positives in the majority class.

While interpretable, the model underperformed compared to ensemble methods (>90% accuracy in literature), likely due to its rigid structure limiting nuanced feature interactions and SMOTE's synthetic data limitations. Future steps include:

- Testing Random Forest/XGBoost for accuracy gains while retaining interpretability.
- Implementing cost-sensitive learning or threshold adjustments to prioritize ROI over raw accuracy.
- Validating via A/B testing on live campaigns to ensure offline improvements drive real-world efficiency and conversion rates.

X. Conclusion

This project aimed to develop a predictive model to identify customers who are likely to subscribe to a term deposit based on their personal and campaign-related attributes. By applying a structured machine learning work, including data preprocessing, handling class imbalance with SMOTE, and using a Decision Tree classifier we were able to achieve meaningful results.

The model achieved an accuracy of 76.35% and a recall of 87.86% for the subscribed class, making it a strong candidate for marketing applications where capturing as many potential responders as possible is critical. While the model's precision (30.73%) indicates some false positives, this trade-off is acceptable in many real-world marketing contexts.

Overall, the project demonstrates that even a simple, interpretable model like a Decision Tree can be effectively used to guide data-driven marketing strategies. Further improvements can be explored through ensemble models, hyperparameter tuning, or cost-sensitive learning approaches.

References

[1] Moro, S., Laureano, R., & Cortez, P. (2014). Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In Proceedings of the European Simulation and Modelling Conference. https://archive.ics.uci.edu/ml/datasets/bank+marketing

[2] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://scikit-learn.org/

[3] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[4] Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer. ISBN: 978-1-4614-6848-6

[5] Raschka, S., & Mirjalili, V. (2019). Python Machine Learning (3rd ed.). Packt Publishing.

[6] Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies, 2(1), 37–63.

[7] Brown-lee, J. (2020). Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning. Machine Learning Mast