STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

CPE 695-APPLIED MACHINE LEARNING

PROJECT REPORT

Enhancing Bank Marketing with Machine Learning

Team Members

1. Arkat Charan Varma
2. Sravanthi Erukulapati
3. Keerthi Reddy Kotha
4. Ruthura Meedimale

ABSTRACT

European banks have faced significant challenges due to internal competition and financial crises, necessitating innovative strategies to enhance financial assets. In response to the limitations of traditional marketing strategies, which have shown minimal impact in increasing bank business, one Portuguese banking institution embarked on a phone-based direct marketing campaign to promote long-term deposits with attractive interest rates. Despite the potential of such campaigns, the extensive effort required to contact numerous potential customers yields relatively low success rates. To address this, the bank has provided a dataset from these marketing efforts to explore more effective tactics using machine learning. This study applies five machine learning models—k-nearest neighbors, logistic regression, support vector machine, random forest and decision trees—to analyze the dataset and develop a predictive model. The goal is to improve marketing efficiency by reducing the number of calls needed while increasing the success rate of securing term deposit subscriptions. The project's findings aim to assist the bank in refining its marketing strategies, thereby enhancing customer engagement and contributing to the bank's profitability through increased long-term deposits.

DATA DESCRIPTION

With a Dataset containing records from these phone-based marketing efforts, this project aims to use machine learning to increase the efficiency and success rate of these campaigns. The dataset we used is taken from Kaggle[1].

- The dataset has 56373 entries and 17 features.
- There are no missing values in any of the columns.

Given below are the column names and their description:

Column	**Data Type**	**Description**
age	Numerical	Age of the customer.

job	Categorical	Type of job of the customer (e.g., admin, technician, services).
marital	Categorical	Marital status of the customer (e.g., married, single).
education	Categorical	Educational qualification of the customer (e.g., secondary, tertiary).
default	Categorical	Indicates if the customer has credit in default ('yes', 'no').
balance	Numerical	Account balance of the customer.
housing	Categorical	Indicates if the customer has a housing loan ('yes', 'no').
loan	Categorical	Indicates if the customer has a personal loan ('yes', 'no').
contact	Categorical	Type of communication contact (e.g., cellular, unknown).
day	Numerical	The day of the month on which the customer was contacted.
month	Categorical	The month of the year during which the customer was contacted (e.g., may, june).
duration	Numerical	Duration of the contact in seconds.

campaign	Numerical	Number of contacts performed during this campaign for this client.
pdays	Numerical	Number of days that passed by after the client was last contacted from a previous campaign (999 means client was not previously contacted).
previous	Numerical	Number of contacts performed before this campaign for this client.
poutcome	Categorical	Outcome of the previous marketing campaign (e.g., unknown, success).
deposit	Categorical	Indicates if the client has subscribed to a term deposit ('yes', 'no').

DATA PREPROCESSING : Enhancing Data Quality

1.Outlier Detection and removal

2.Label Encoding

3.Handling imbalanced Data

1.Outlier Detection and Removal:

Outliers are data points that deviate from the rest of the data in a dataset. They can occur due to various reasons such as measurement errors, data entry errors, natural variation, or rare events. Outliers can have a significant impact on accuracy of machine learning models so detecting them and removing them is important.

In the dataset,there are certain records where the balance of the customer is negative and they subscribed to term deposit.Such records are removed because those scenarios are not so common and they might affect the learning of the classifiers.

2. Label Encoding:

One-Hot Encoding: Represents each category as a binary vector where each element corresponds to a unique category.

All the categorical columns are converted to numerical columns.

3. Handling Imbalanced Data:

Handling Imbalanced data is crucial as it may lead to inaccurate results.Such imbalances can bias the predictive model towards the majority class, leading to poor classification performance on the minority class, which is often of greater interest.

The dataset contains 45795 records whose deposit column value is ‘no’.And there are 10578 records with deposit column value ‘yes’ which leads to data imbalance.

To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) is employed.

SMOTE is an advanced over-sampling method that creates synthetic samples rather than simply duplicating existing samples in the minority class. Here’s how SMOTE works:

1. Choosing Samples : SMOTE selects a minority class sample and considers its k-nearest minority class neighbors. The number of neighbors, k, is usually chosen as part of the model hyperparameter tuning process.

2. Synthesizing Samples : For each chosen sample and its neighbors, SMOTE synthesizes new examples that are added to the dataset. This is done by:

- Taking the difference between the feature vector (sample) and its nearest neighbor.

- Multiplying this difference by a random number between 0 and 1.

- Adding this result to the feature vector, thus creating a new sample point that lies on the line segment joining the two specific features in the feature space.

METHODOLOGY :

Following historical data, our study employed a variety of machine learning classifiers to predict whether bank customers will enroll on deposit. K-nearest neighbors (KNN), logistic regression, random forest, support vector machine (SVM), and decision tree classification are the models that we tested. These models were chosen based on a range of techniques for analyzing data patterns, which offer a thorough understanding of the model's performance in various algorithmic approaches:

K-Nearest Neighbors (KNN): Selected for its non-parametric approach, which determines a class based on the feature space's nearest training instances. Because KNN works well in situations with very irregular decision boundaries, it was chosen.

Logistic Regression: Used because it's effective in binary classification jobs and can yield outcome probabilities. This was a good choice because of its prominence in financial modeling because of its easily interpreted coefficients that may measure the impact of different features.

Support Vector Machine (SVM): This model was chosen because of the kernel technique, which makes it adept at managing non-linear boundaries. It works especially well in high-dimensional areas, which makes complex datasets like those found in banking perfect for it.

Decision Tree: Used because of its simple interpretability, which is essential for comprehending the significance of features and making decisions in a commercial setting. Decision trees are fairly easy to explain and can handle qualitative aspects without the need of dummy variables.

Random Forest: An ensemble technique that constructs several decision trees and combines them to produce a prediction that is more reliable and accurate. Random Forest is essential in a banking setting where predictability and dependability are essential since it corrects the decision trees tendency of overfitting to their training set.

We utilized the effectiveness of hyperparameter tuning to optimize the performance of machine learning models. Since it has an immediate effect on the model's ability to accurately predict results based on the input data. By adjusting the model's parameters, it also aids in increasing model accuracy, allowing us to greatly enhance the model's

predictive ability on unobserved data. By balancing the bias-variance tradeoff and preventing overfitting and underfitting, proper hyperparameter selection helps ensure that the model is neither overly complex nor too simple. The model's complexity is controlled by a few parameters, which also have an impact on training time and resource usage. Model training can be made more effective by adjusting variables, particularly when working with big datasets.

The most common method for hyperparameter tuning is Grid Search, which involves testing a wide range of parameter values until the optimal combination is found that yields the highest model performance according to a specified assessment criterion. This methodical search makes sure that the model configuration that is chosen is best suited to the peculiarities and difficulties of the particular dataset being used.

Our project's target variable is whether or not a consumer subscribes to a term deposit (yes or no). It was designed as a binary classification problem. The elements included contact details, marketing results from prior campaigns, and demographic data about the customer.

RESULTS:

EDA ANALYSIS :

From exploratory data analysis on the dataset ,we can conclude a few points.

Customers who belong to job categories like Management ,blue collar,technicians are approached more compared to other job category customers.

More term deposits are done by people who belong to the age group 30-40.

Ratio of number of customers to term deposits is high among students and retired.

Largest number of customers are approached in the month of may.The Month of april has a good ratio of number of customers and term deposits subscribed.

The data contains 59.7% married customers,28.9% single customers and 11.4% of divorced.

Overall percentage of customers who secured secondary and tertiary education is 89.9% ,and the number of term deposits by them is high.

Ratio of the number of customers approached to the number of term deposits is high in case of single customers.

There are no deposits made by customers who have credit default.

Number of term deposits is low for customers who took housing loans or personal loans.

More number of Credit defaults are seen among customers with job categories like Management,blue collar,technician,admin.
The average duration of time taken by customers who said no to deposit is 200 seconds,whereas the average amount of time taken by customers who subscribed to deposit is 500 seconds.

CLASSIFICATION RESULTS:

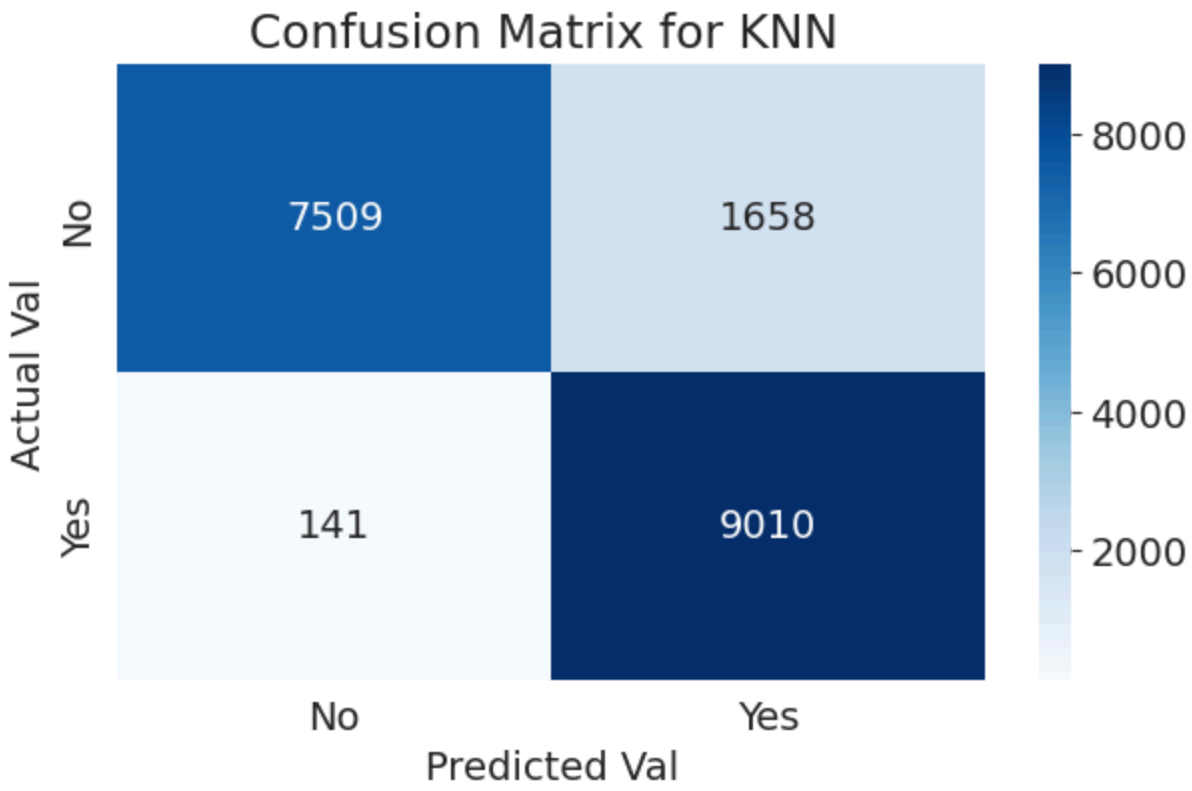
K-Nearest Neighbors (KNN):

Cross-validation score and best params
The best parameters are {'n_neighbors': 5, 'weights': 'uniform'}
Cross-validation score: 0.8943525245055823

KNN Test data accuracy Score 0.9017905884921935
KNN Train data accuracy Score 0.9280898569712851
---

	precision	recall	f1-score	support
0	0.98	0.82	0.89	9167
1	0.84	0.98	0.91	9151
accuracy			0.90	18318
macro avg	0.91	0.90	0.90	18318
weighted avg	0.91	0.90	0.90	18318

[[7509 1658]
[141 9010]]



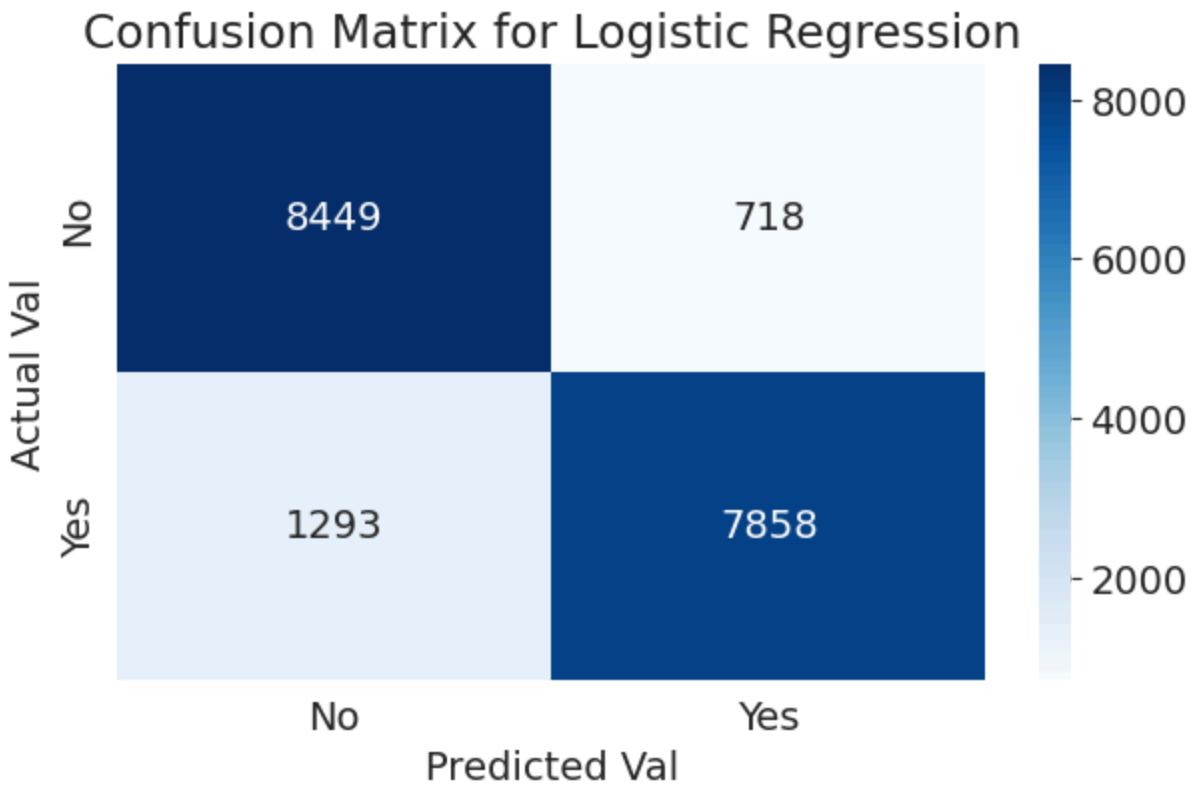
Logistic Regression:

Cross-validation score and best params
The best parameters are {'max_iter': 50}
Cross-validation score: 0.8883202169787763

Logistic Regression Test data accuracy Score 0.8902172726280162
Logistic Regression Train data accuracy Score 0.8886477781417186
---

	precision	recall	f1-score	support
0	0.87	0.92	0.89	9167
1	0.92	0.86	0.89	9151
accuracy			0.89	18318
macro avg	0.89	0.89	0.89	18318
weighted avg	0.89	0.89	0.89	18318

[[8449 718]
[1293 7858]]



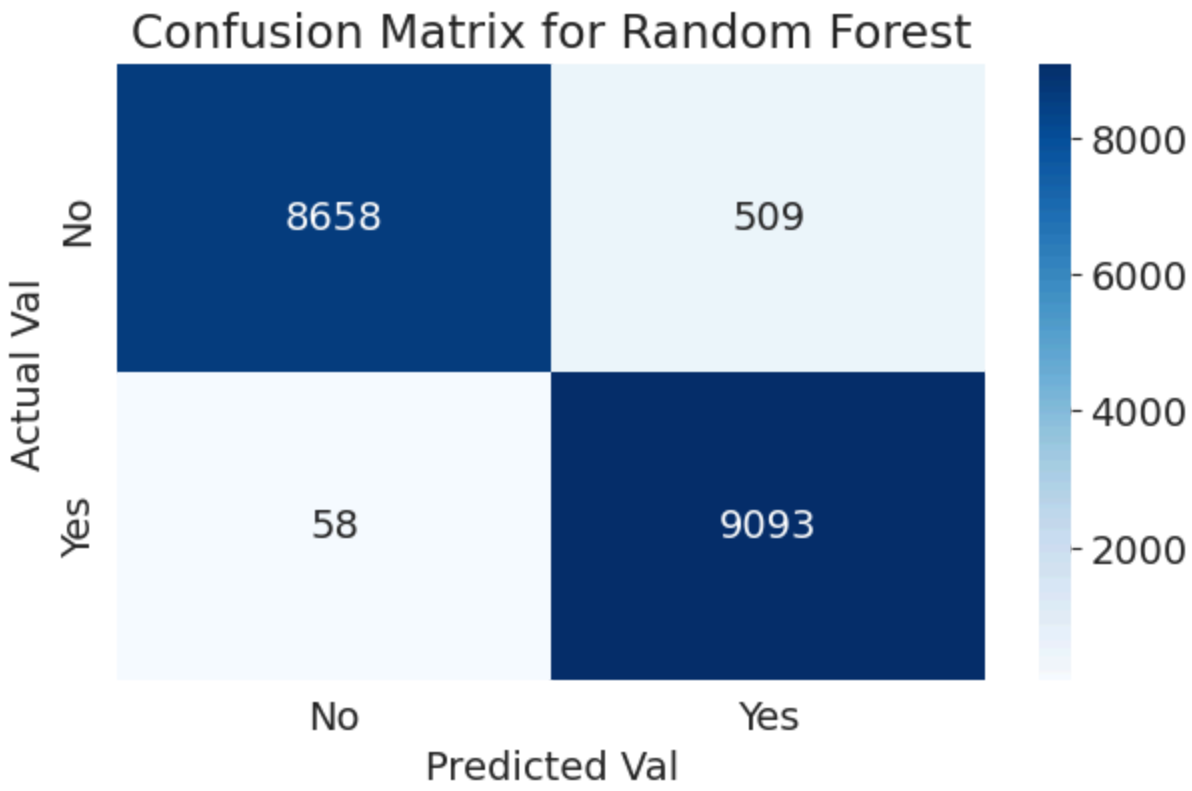
Random Forest:

Cross-validation score and best params
The best parameters are {'max_depth': 20, 'min_samples_split': 2, 'n_estimators': 200}
Cross-validation score: 0.9564362921716345

Random Forest Test data accuracy Score 0.9690468391745823
Random Forest Train data accuracy Score 0.9937220220548095
---

	precision	recall	f1-score	support
0	0.99	0.94	0.97	9167
1	0.95	0.99	0.97	9151
accuracy			0.97	18318
macro avg	0.97	0.97	0.97	18318
weighted avg	0.97	0.97	0.97	18318

[[8658 509]
[58 9093]]



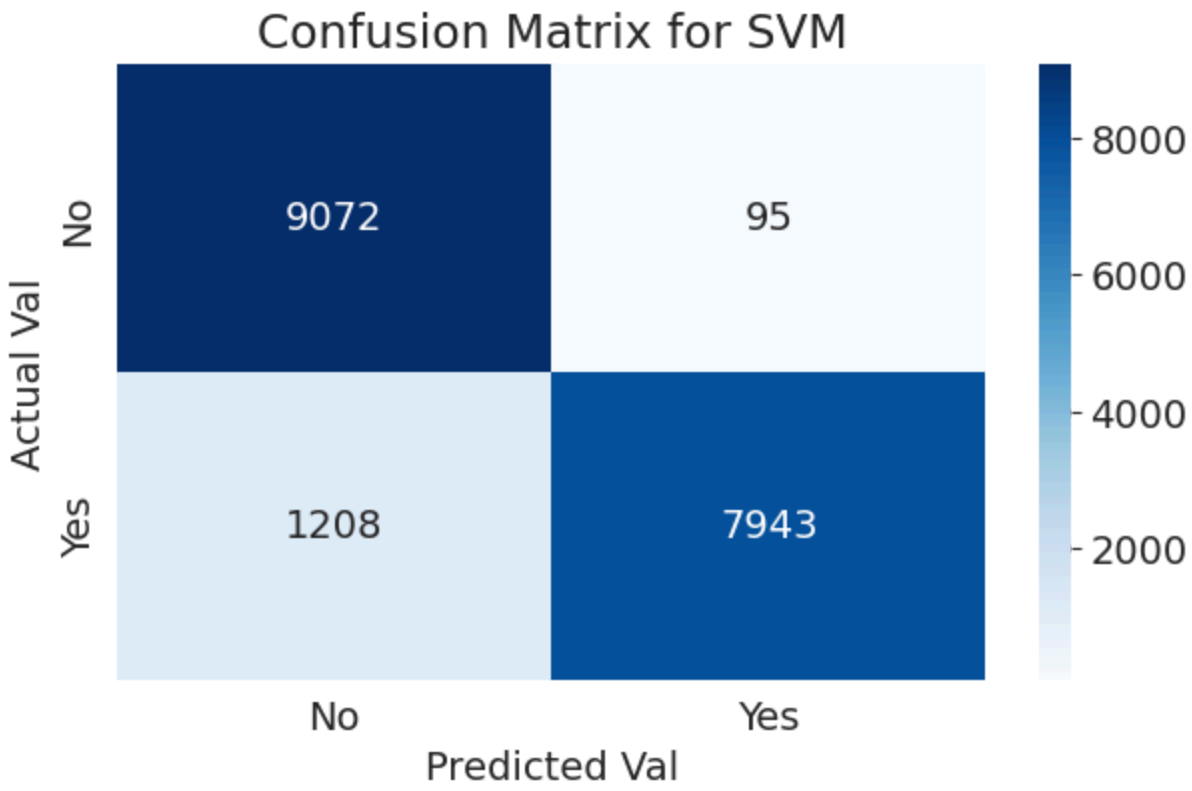
Support Vector Machine (SVM):

Cross-validation score and best params
The best parameters are {'C': 10, 'gamma': 'auto', 'kernel': 'rbf'}
Cross-validation score: 0.8989382028605744

SVM Test data accuracy Score 0.928867780325363
SVM Train data accuracy Score 1.0
---

	precision	recall	f1-score	support
0	0.88	0.99	0.93	9167
1	0.99	0.87	0.92	9151
accuracy			0.93	18318
macro avg	0.94	0.93	0.93	18318
weighted avg	0.94	0.93	0.93	18318

[[9072 95]
[1208 7943]]



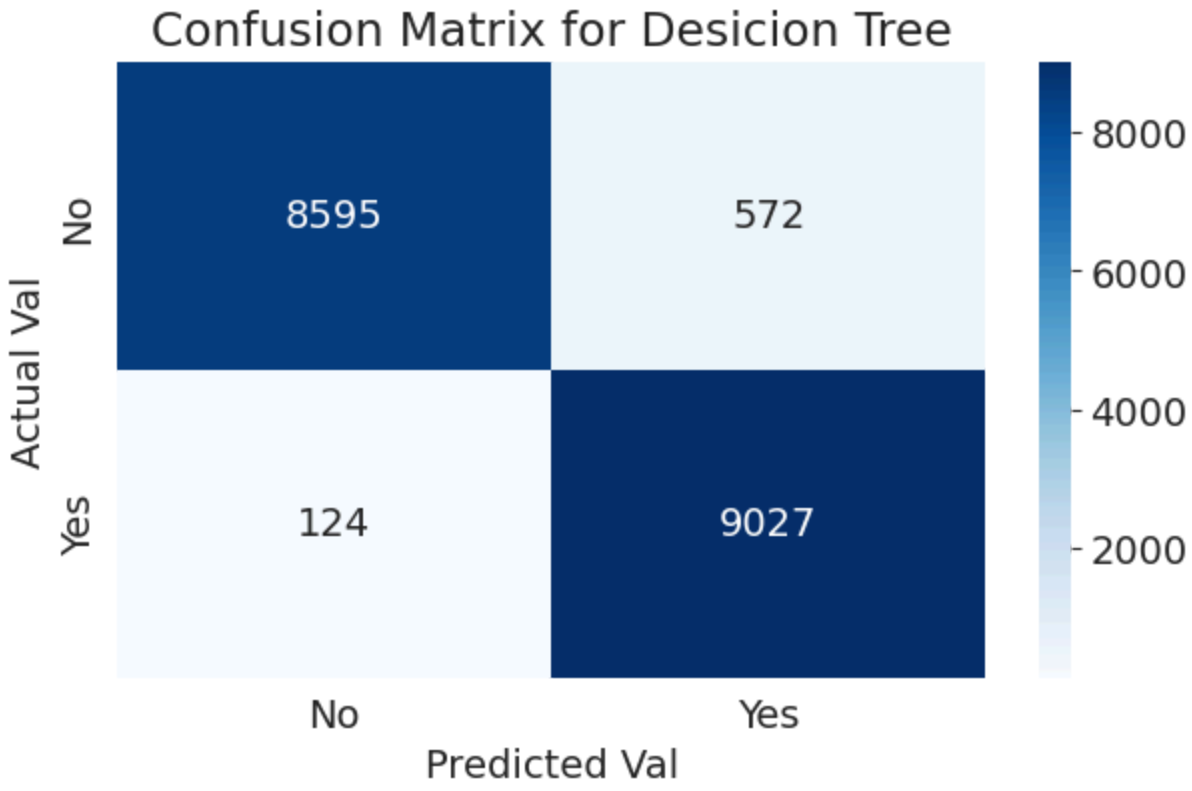
Decision Tree:

Cross-validation score and best params
The best parameters are {'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2}
Cross-validation score: 0.9433071295993013

Desicion Tree Test data accuracy Score 0.9620045856534556
Desicion Tree Train data accuracy Score 0.9996997488808822
---

	precision	recall	f1-score	support
0	0.99	0.94	0.96	9167
1	0.94	0.99	0.96	9151
accuracy			0.96	18318
macro avg	0.96	0.96	0.96	18318
weighted avg	0.96	0.96	0.96	18318

[[8595 572]
[124 9027]]



CONCLUSION:

Our study led us to conclude that the random forest and decision tree models were the most accurate of the models we tested ,with accuracy of around 96% on test data and should be further used to predict the term deposit behavior of customers. To improve prediction accuracy, future research could investigate deeper learning models and more complex ensemble methods. Possible improvements could include integration of complementary data sources, use of advanced machine learning models, and real-time analytics to change dynamic marketing strategies More research and ethical considerations integration into automated continuous learning systems is needed over time Strengthened.

REFERENCES:

[1]https://www.kaggle.com/datasets/seanangelonathanael/bank-target-marketing/data