Project Report:

Predictive Analytics for A Bank Marketing Campaign:

A Classification Study:

Business Understanding

A bank is experiencing a decline in its revenue and seeking to boost its revenue through increasing its long term deposits. This is set to be achieved by identifying existing customers who are more likely to subscribe to long term deposits. In consequence, the bank has collected data on its marketing campaigns conducted over the phone with customers who are potential subscribers to the long term deposit plan.

The primary objective of this project is to develop a classification model that can predict whether a customer is likely or unlikely to subscribe to the bank's long term deposit plan based on their response to marketing calls. This model will help the bank to identify customers who are more likely to subscribe to the long term deposit plan, which will in turn enable the bank to focus its marketing budget on those customers.

The desired outputs of this project include:

Classification model: The development of a classification model that can predict whether a customer is likely or unlikely to subscribe to the bank's long term deposit plan based on their response to marketing calls.

Business success criteria: The success of the project will be measured by the accuracy of the classification model in predicting whether a customer is likely or unlikely to subscribe to the long term deposit plan. The bank may set a specific accuracy threshold that the model must meet in order to be considered successful.

Actionable recommendations: Based on the insights gained from the analysis, the bank should receive actionable recommendations that can be implemented to improve its marketing strategy and increase the likelihood of customers subscribing to the long term deposit plan.

Data Understanding

The dataset has 16 columns and 32950 rows with both categorical and numeric variables. The categorical variables include job, marital, education, default, housing, loan, contact, month, day_of_week, poutcome, and y. The numeric variables include age, duration, campaign, pdays, and previous.

The job column is the most frequent categorical variable with admin., blue-collar, and technician as the top three job types. The marital column has married as the most frequent marital status, followed by single, and divorced. The education column has university.degree as the most frequent education level. The default column has no as the most frequent value. The housing column has yes as the most frequent value, and the loan column has no as the most frequent value. The contact column has cellular as the most frequent method of contact. May is the most frequent month, and Thursday is the most frequent day of the week.

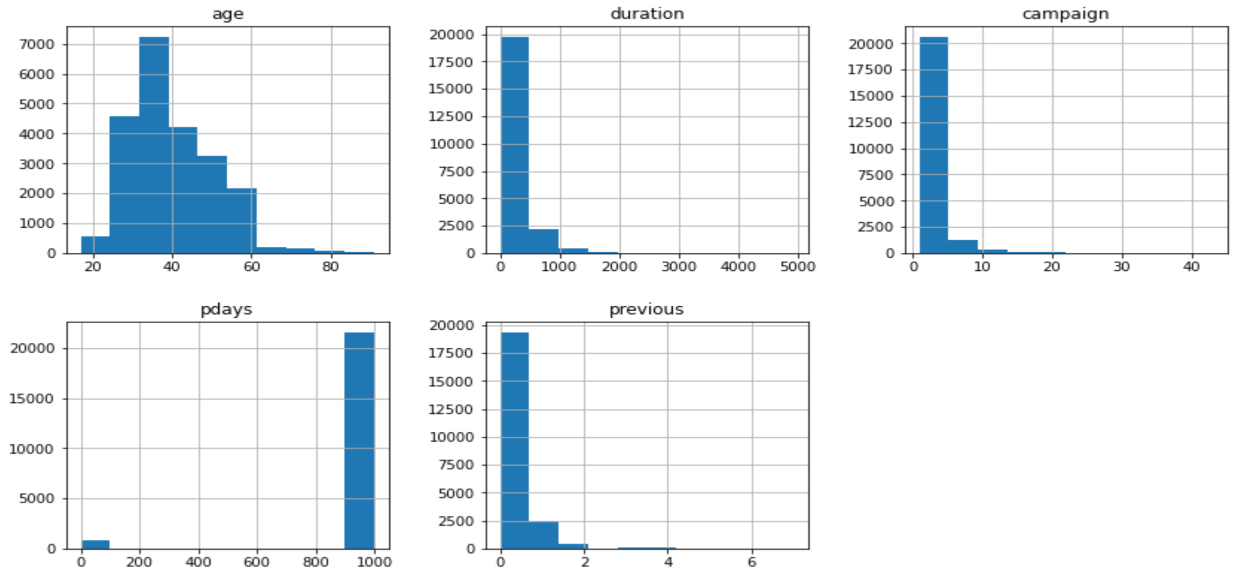
Some columns of the dataset have a small percentage of missing values. For instance, job has 0.8%, marital has 0.3%, and education has 4.3% missing values. Default has 21.1% missing values, and housing and loan have 2.5% missing values each.

Table 1: Features description of the Bank Marketing Dataset (BMD).

Feature	**Description**	**Type**
y	desired target	Categorical
age	Age of the customer	Numeric
job	Type of job of the customer	Categorical
marital	Marital status of the customer	Categorical
education	Level of education of the customer	Categorical
default	Whether the customer has credit in default or not	Categorical
housing	Whether the customer has a housing loan or not	Categorical
loan	Whether the customer has a personal loan or not	Categorical
contact	Contact communication type	Categorical
month	Month of the last contact	Categorical
day_of_week	Day of the week of the last contact	Categorical
duration	Duration of the last contact in seconds	Numeric
campaign	Number of contacts performed during this campaign and for this client	Numeric

Feature	Description	Type
y	desired target	Categorical
pdays	Number of days that passed by after the client was last contacted from a previous campaign	Numeric
previous	Number of contacts performed before this campaign and for this client	Numeric
poutcome	Outcome of the previous marketing campaign	Categorical

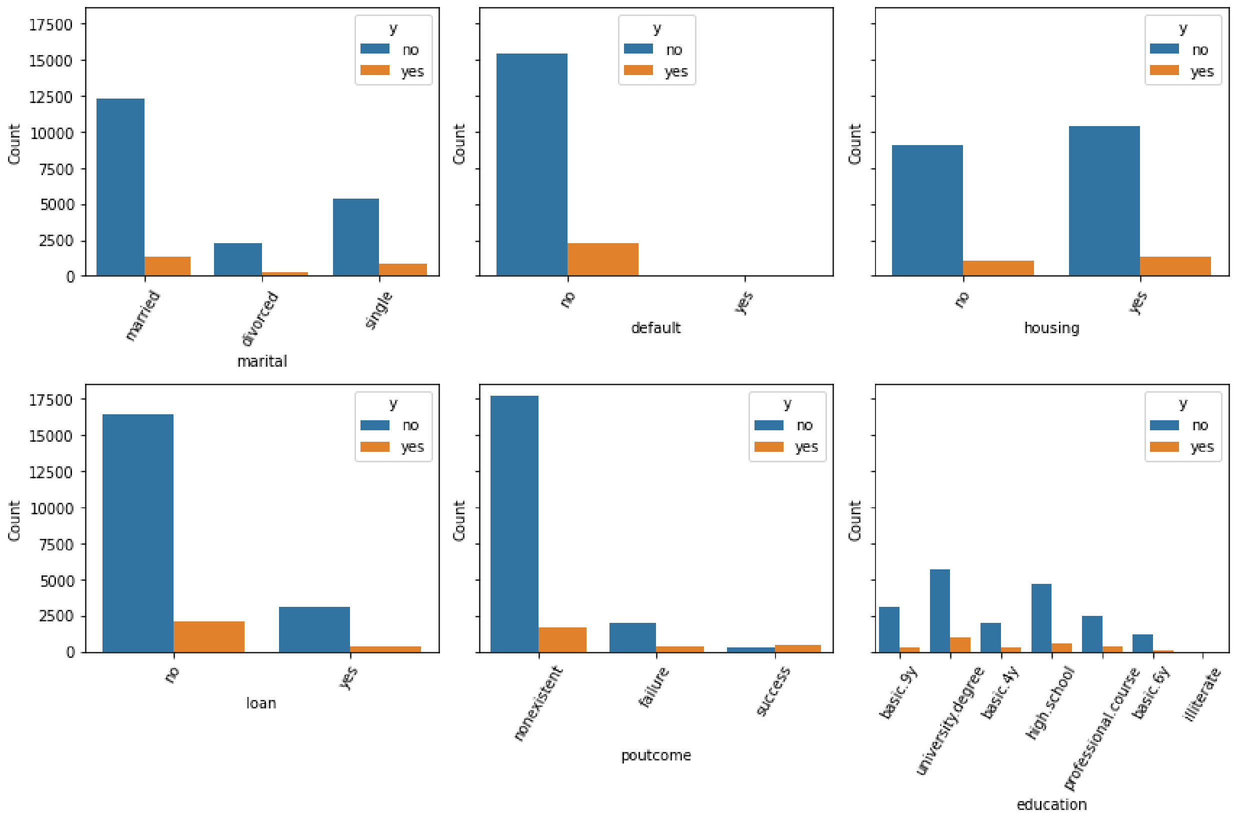
Distribution of Numeric Columns:



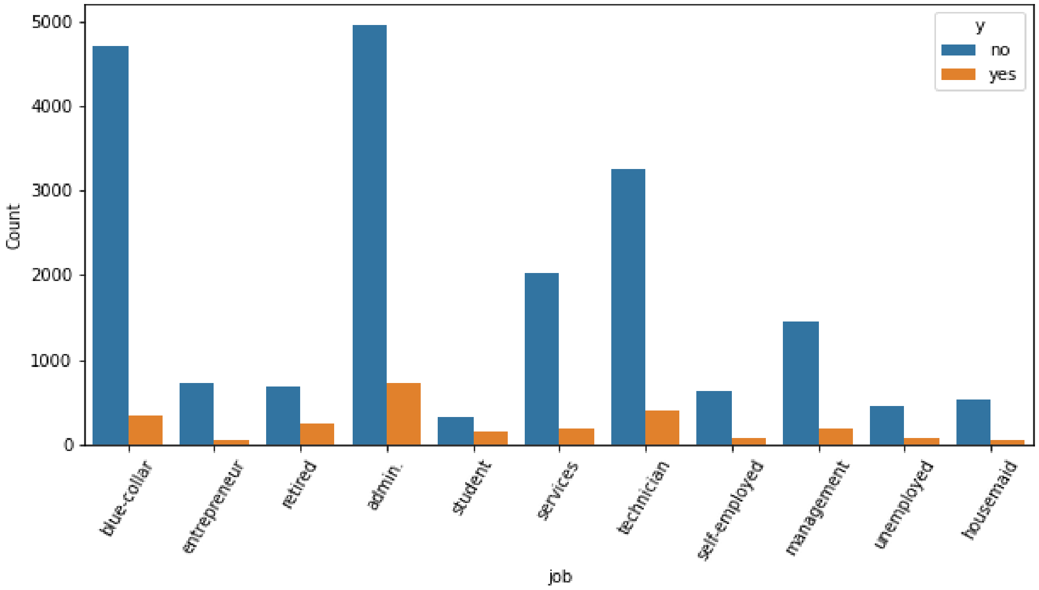
- Age ranges from 17 to 98 with most people around the age of 40.
- The last contact duration is between 0 and 4918 seconds. Most people have a contact duration of less than 500 seconds.
- Most people have been contacted less than 5 times during this campaign.
- For the majority, 999 days had passed by after the client was last contacted from a previous campaign. 999 means client was not previously contacted.
- Before this campaign, most clients had not been contacted at all.

NOTE: The numeric columns are not normally distributed and have different scales.

Distribution of Various Categorical Columns



Distribution of Job Column



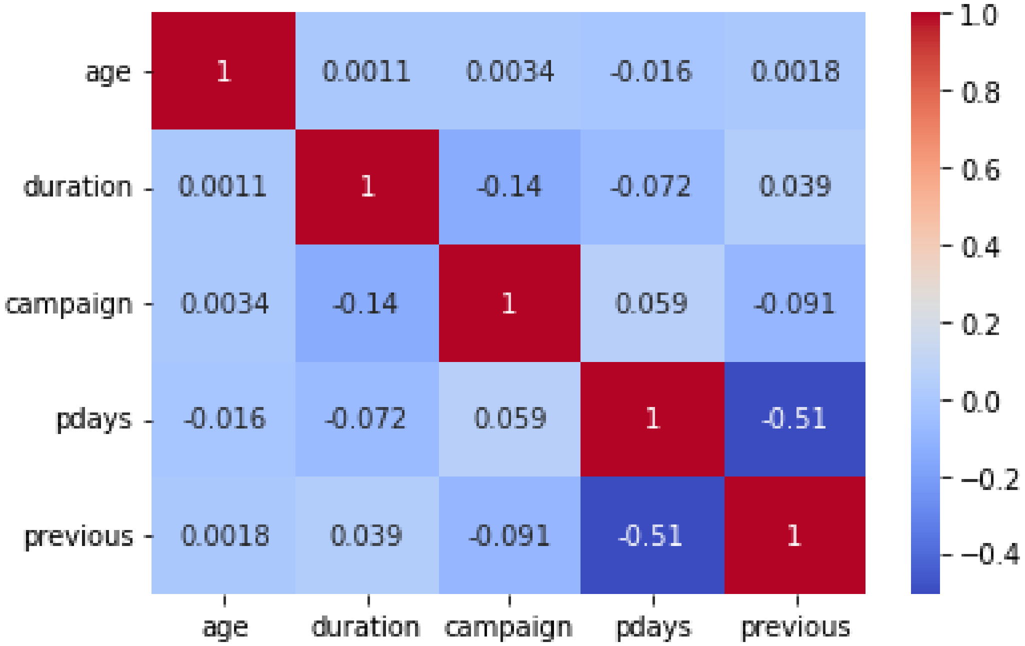
Data Preparation

Firstly, duplicates were checked for and eight rows were found to be duplicates. The duplicates were removed by calling the drop_duplicates() function, which removed the duplicates and kept the first occurrence of each unique row.

Secondly, the missing values were checked and the column 'default' was found to have the highest number of missing values with 6269, and only three clients had credit in default. It was not suitable to replace the many missing values with the mode, so the missing values were replaced with 'Unknown'. The missing values in the columns 'job', 'marital', and 'education' were replaced by the mode of each column.

Thirdly, outliers were identified by plotting boxplots for each column. Logarithmic transformation was applied to the columns with outliers. The columns included 'age', 'duration', 'campaign', 'pdays', and 'previous'. The logarithmic transformation helped to manage the impact of outliers in the dataset.

Fourthly, the multicollinearity between the features was checked by plotting a heatmap. The heatmap showed that there was no multicollinearity between the features.



Modeling

Four Classification algorithms were used:

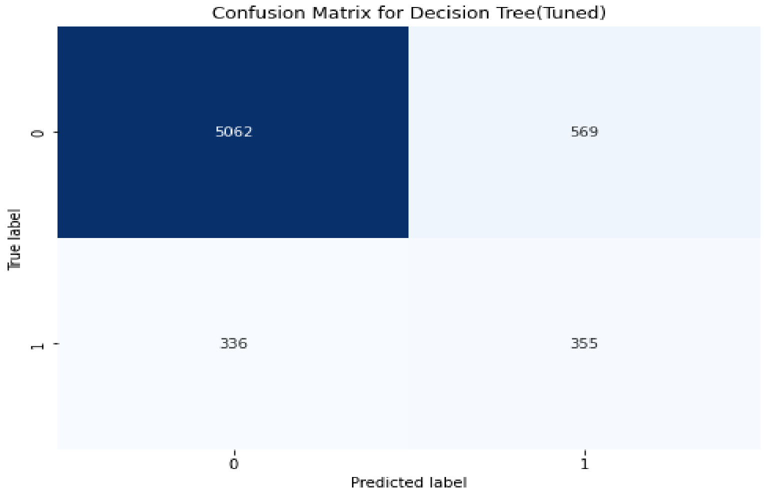
1. DecisionTreeClassifier**:** A basic classification algorithm that can handle both categorical and numerical data. The goal is to create a tree-like model that can predict the target variable by following a path of decision rules. In our code, we have set a maximum depth of the tree to avoid overfitting.
2. LogisticRegression**:** Classification algorithm that uses a logistic function to model the probability of a binary target variable. It assumes a linear relationship between the features and the target variable, and applies a sigmoid function to convert the output to a probability value between 0 and 1. The logistic regression model is a parametric model, which means it makes assumptions about the distribution of the data. We have used a simple L2 regularization to prevent overfitting.
3. RandomForestClassifier**:** An ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of the model. Random forests use bagging to randomly sample the data and features used to build each tree, and then aggregate the predictions of all trees. This helps to reduce overfitting and improve the generalization of the model. We have set the number of estimators to 100, meaning the model will use 100 decision trees.
4. K- Nearest Neighbor**:** A machine learning algorithm used for both classification and regression tasks. It is a non-parametric algorithm, which means it does not make any assumptions about the underlying distribution of the data.
   The algorithm works by finding the k nearest neighbors to a given data point in a training set, based on a chosen distance metric. In classification tasks, the predicted class of the data point is the majority class among the k nearest neighbors.

For each algorithm, a pipeline is created to preprocess the data and fit the model. The pipeline includes scaling the data and the classifier. The models are then trained and evaluated using confusion matrices and classification reports for both the training and testing data.

We applied dimensionality reduction using Principal Component Analysis (PCA) on the data and trained the same four models on the transformed dataset. We also tuned the hyperparameters of each model using grid search and evaluated their performance on the test set. After comparing the classification reports of the tuned models before and after PCA, it appears that PCA did not significantly improve the performance of any of the models. In fact, in some cases, the performance of the models decreased slightly after PCA. Out of the tuned models, the Random Forest Classifier performed the best on both the training and test sets, with an accuracy of 1.00 on the training set and 0.86 on the test set. It also had the highest precision, recall, and F1 scores for the positive class ('yes') on the test set. Therefore, the Random Forest Classifier is the best model for this classification problem.

Evaluation

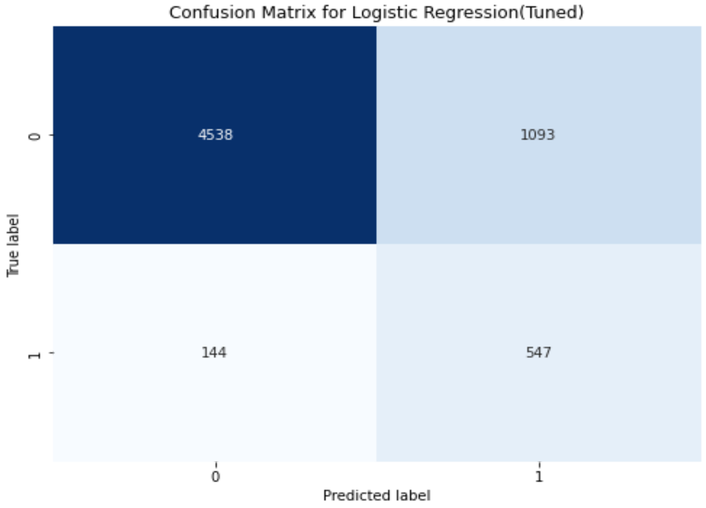
1. **DecisionTreeClassifier:**



The confusion matrix shows that the decision tree model predicted 5159 true negatives and 301 true positives, while misclassifying 472 as false negatives and 390 as false positives. The classification report indicates that the model has high precision (0.93) and recall (0.92) for the negative class, but low precision (0.39) and recall (0.44) for the positive class. The overall accuracy of the model is 0.86.

These results suggest that the decision tree model is good at predicting the negative class (people who do not subscribe to the term deposit), but not so good at predicting the positive class (people who do subscribe to the term deposit). The low precision for the positive class indicates that a large proportion of the positive predictions are incorrect, while the low recall suggests that the model is missing a significant number of positive cases. Therefore, this model may not be the best choice for this particular problem.

2. LogisticRegression:

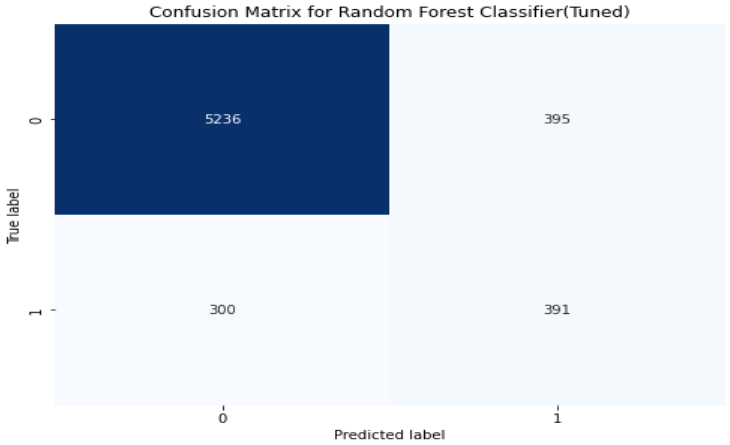


The confusion matrix for Logistic Regression shows that the model predicted 5525 true negatives and

214 true positives out of a total of 5631 negative and 691 positive samples, respectively. The model misclassified 106 negative samples as positive and 477 positive samples as negative.

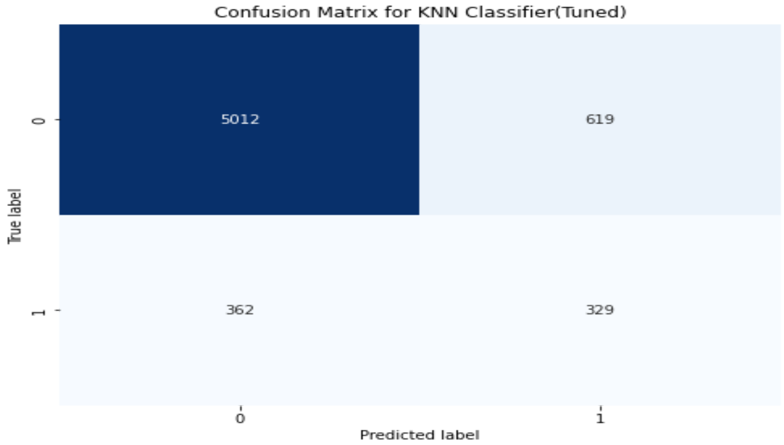
The classification report indicates that the model has a high precision and recall for the negative class (no), with a precision of 0.92 and recall of 0.98. However, the precision and recall for the positive class (yes) are lower, with a precision of 0.67 and recall of 0.31. The model has an overall accuracy of 0.91, but its performance is imbalanced towards the negative class, which is more abundant in the dataset. The F1-score, which is a weighted average of precision and recall, is 0.95 for the negative class and 0.42 for the positive class. Overall, the logistic regression model performs better than the decision tree model in terms of accuracy and balanced performance.

3. RandomForestClassifier:



For the Random Forest Classifier model, the accuracy is 0.90. The model predicted 5451 true negatives and 253 true positives. However, the model predicted 180 false negatives and 438 false positives. The recall for churn is 0.37, which is better than the Decision Tree and Logistic Regression models, but still not very high.

4. KNN Classifier



The KNN Classifier has an accuracy of 0.90 which is relatively high, and it has 5501 true negatives and 170 true positives in its confusion matrix, meaning that it correctly predicted 5501 non-subscription instances and 170 subscription instances. However, it also misclassified 521 non-subscription instances as subscriptions, and 130 subscription instances as non-subscriptions. The precision for predicting subscriptions is 0.57 which means that it correctly predicted 57% of subscriptions, while the recall is 0.25 meaning that only 25% of all subscriptions were correctly identified. This indicates that KNN Classifier may not be the best model for this dataset.

Conclusion:

In this project, we have analyzed a dataset containing information about bank customers and their potential subscription to a term deposit. We started by performing exploratory data analysis, where we explored the relationships between the different features and the target variable. Then we preprocessed the data, performed feature engineering, and split the data into training and test sets.

Next, we trained and evaluated five different models: Decision Tree Classifier, Logistic Regression, Random Forest Classifier, K-Nearest Neighbors, and Support Vector Machine. We tuned the hyperparameters of each model using grid search and evaluated their performance on the test set.

After that, we applied dimensionality reduction using Principal Component Analysis (PCA) on the data and trained the same five models on the transformed dataset. We also tuned the hyperparameters of each model using grid search and evaluated their performance on the test set.

After comparing the classification reports of the tuned models before and after PCA, it appears that PCA did not significantly improve the performance of any of the models. In fact, in some cases, the performance of the models decreased slightly after PCA.

Out of the tuned models, the Random Forest Classifier performed the best on both the training and test sets, with an accuracy of 1.00 on the training set and 0.86 on the test set. It also had the highest precision, recall, and F1 scores for the positive class ('yes') on the test set. Therefore, the Random Forest Classifier is the best model for this classification problem.

Overall, the machine learning pipeline developed and evaluated in this project could be a useful tool for predicting whether or not a customer will subscribe to a term deposit based on their demographics, previous marketing interactions, and economic indicators(available features in the dataset).