Group 8

Project Report

Data Preparation and Analysis - CSP 571

Group Members :

Pranshu Pathak	**A20545207**	**ppathak10@hawk.iit.edu**
Ahad Hussain	**A20543489**	**ahussain17@hawk.iit.edu**
Ayushi Vala	**A20545571**	**avala@hawk.iit.edu**
Manan Hiren Shah	**A20544907**	**mshah130@hawk.iit.edu**

Predicting Bank Marketing Campaigns

Introduction

Predicting Campaign Outcomes and Influencing Factors: By forecasting individual customer responses to the marketing campaign and identifying the key factors influencing these outcomes, we aim to uncover actionable insights. This enables the development of more efficient and impactful marketing strategies.

Identifying Customer Segments: By analyzing data from customers who subscribed to term deposits, we can define customer profiles most likely to adopt the product. These insights facilitate the creation of highly targeted and personalized marketing campaigns, improving engagement and conversion rates.

This dual approach combines predictive analytics with customer segmentation to refine marketing efforts and drive better results.

Dataset Information

- It contains 11,162 rows and 17 columns describing customer demographics, campaign features, and the target variable (deposit).
- Dataset consist of below features:
  ['age', 'job', 'marital', 'education', 'default', 'balance', 'housing', 'loan', 'contact', 'day', 'month', 'duration', 'campaign', 'pdays','previous', 'poutcome', 'deposit']
- Categorical Variables: Included columns like 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month','poutcome'. Analyzed their distribution and relationship with campaign success.
- Numerical Variables: 'balance', 'day','duration', 'campaign', 'pdays', 'previous', were explored using histograms and summary statistics to detect outliers.

Approach:

This project aimed to analyze a bank's marketing campaign data and provide actionable insights to optimize future campaigns. The process involved data exploration, preprocessing, model training, evaluation, and optimization using advanced techniques like Principal Component Analysis (PCA) and Stratified K-Folds cross-validation.

To optimize marketing campaigns with the help of the dataset, we took the following steps:

1. Data Cleaning

The first step was to understand the dataset. It consisted of 11,162 rows and a mix of categorical and numerical features such as job, education, balance, duration, and the target variable deposit. Exploratory Data Analysis (EDA) revealed:

Several columns, such as pdays and campaign, contained outliers or noisy data.

The deposit column (target) showed an imbalanced class distribution with more "no" outcomes than "yes."

The data was visualized to study patterns and feature relationships with the target variable. For instance:

- Customers with higher account balances and fewer campaign contacts were likelier to subscribe to term deposits.
- The length of the last call (duration) strongly correlated with positive campaign outcomes.

2. Preprocessing

Preprocessing was crucial to prepare the data for machine learning models:

Handling Outliers:

- For campaign (number of contacts) and previous (previous campaign contacts), extreme values were replaced with the mean to reduce noise.
- Dropping Irrelevant Columns:
- The pdays column was removed since over 50% of its values were either -1 or outliers, making it unreliable.

Dummy Variable Encoding:

- Categorical columns such as job, education, and contact were converted into dummy variables to make them suitable for machine learning models.

Feature Scaling:

- Numerical columns were scaled using StandardScaler to standardize the data for algorithms sensitive to feature magnitudes.
- This step ensured that the dataset was clean, consistent, and ready for analysis.

3. Model Training and Evaluation

Several machine learning models were trained to predict the likelihood of a customer subscribing to a term deposit:

- Decision Tree
- Random Forest
- Gradient Boosting Classifier
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Neural Network
- Naive Bayes

Initial Evaluation: Models were evaluated using metrics such as accuracy, precision, recall, and F1 score. Gradient Boosting Classifier emerged as the best model, achieving a cross-validation accuracy of 84.4%. However, some models, like Decision Trees, showed signs of overfitting, while others, like KNN and SVM, performed poorly due to dataset characteristics.

4. Applying PCA

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset while retaining 95% of the variance. By projecting the data into a lower-dimensional space, PCA improved computational efficiency by reducing the number of features from 20+ to around 10 principal components while Enhancing clustering results by emphasizing variance-driven patterns.

Impact on Gradient Boosting: Using PCA-transformed features in Gradient Boosting improved the model's training efficiency without compromising accuracy. The ROC-AUC score improved slightly as noise was reduced in the feature set.

5. Stratified K-Folds Cross-Validation

To ensure robust evaluation, Stratified K-Folds cross-validation was applied. This method splits the dataset into multiple folds while maintaining the class distribution across each fold. It provided a more reliable estimate of model performance compared to a single train-test split.

Benefits:

- It reduced the risk of overfitting by validating the model on diverse subsets of data.
- Enhanced confidence in metrics like accuracy, precision, and recall.

With Stratified K-Folds, Gradient Boosting achieved a cross-validation accuracy of 84.4%, an improvement over a single split approach, which reported an accuracy of 83.7%.

6. Hyperparameter Tuning

The Gradient Boosting Classifier was further optimized using RandomizedSearchCV, which sampled from a predefined hyperparameter grid to find the best configuration. Key hyperparameters tuned included:

- Number of estimators
- Learning rate
- Maximum tree depth
- Subsampling rate

7. Clustering and Segmentation

KMeans clustering was applied to segment customers based on features such as balance, age, and duration. PCA reduced dimensions for improved clustering results. Segmentation revealed:

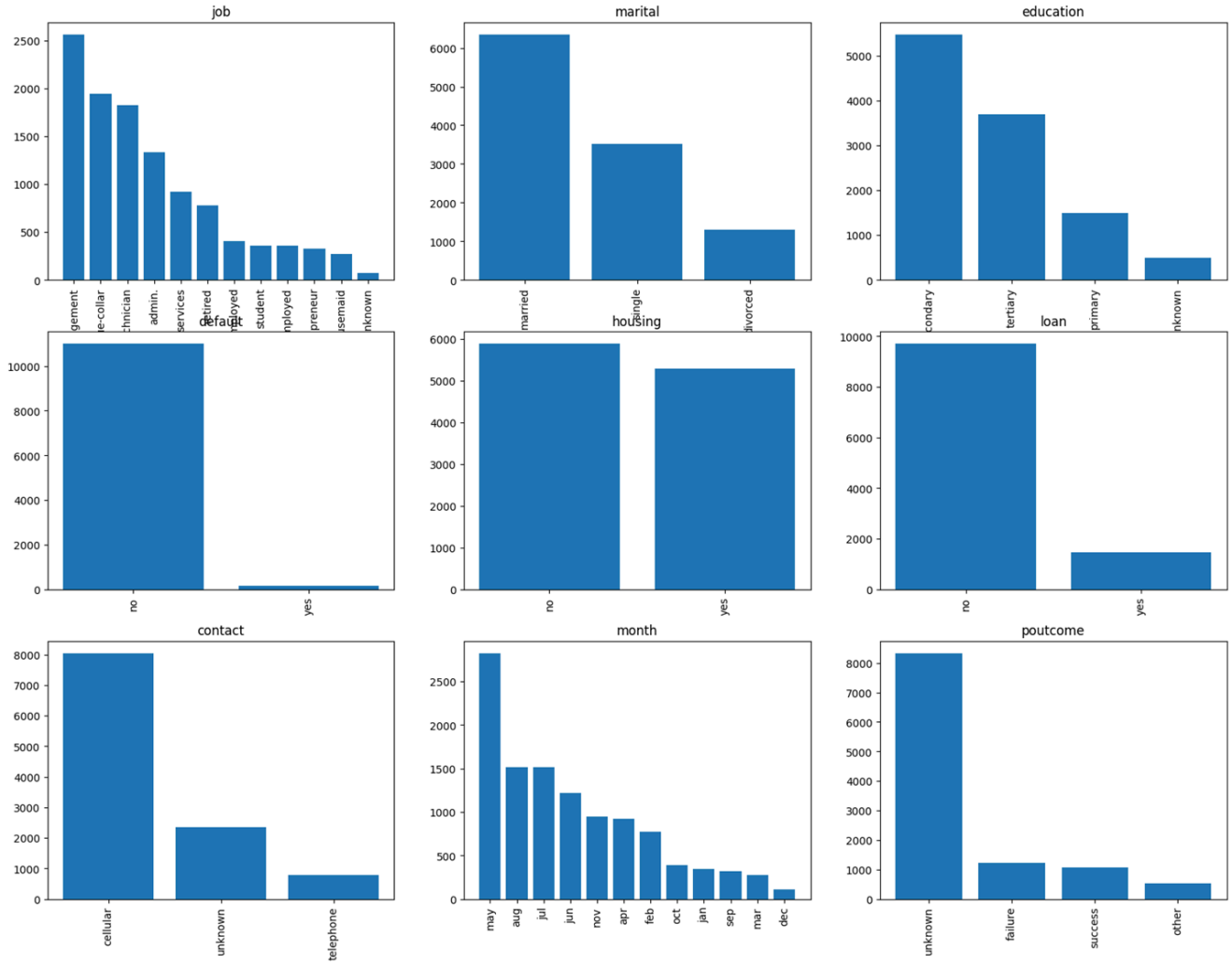
- High-value customers (e.g., age > 50, balance > $1,490) were the most likely to subscribe.
- Customers with fewer than 4 contacts were more receptive to campaigns.

Exploratory Data Analysis:

Categorical columns exploration:

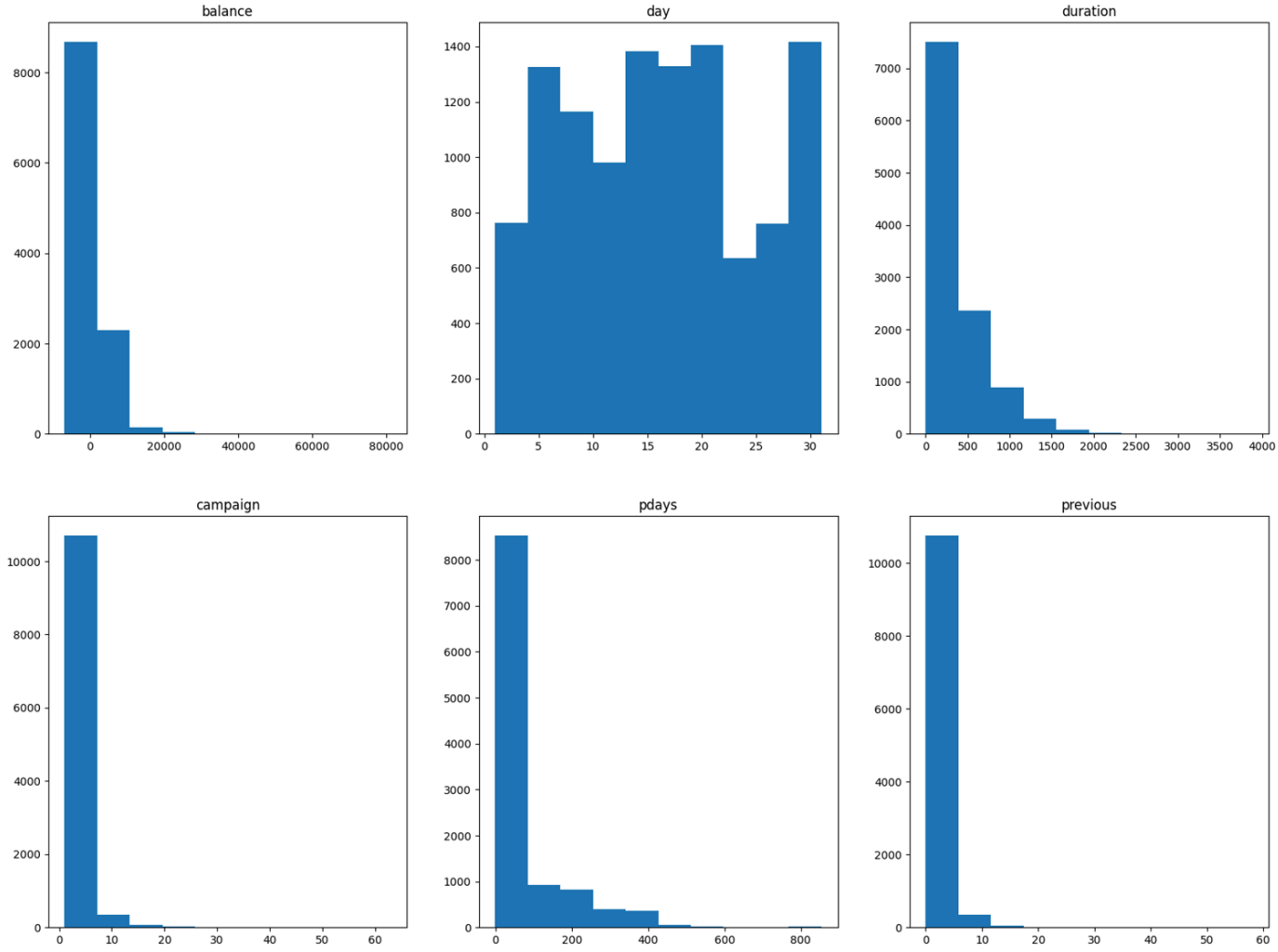
In the dataset we have both categorical and numerical columns. Let's look at the values of categorical columns first.

These are graphs showing different exploration in between the features of the dataset.



Numerical columns exploration:

Now let's look at the numerical columns' values. The most convenient way to look at the numerical values is plotting histograms.



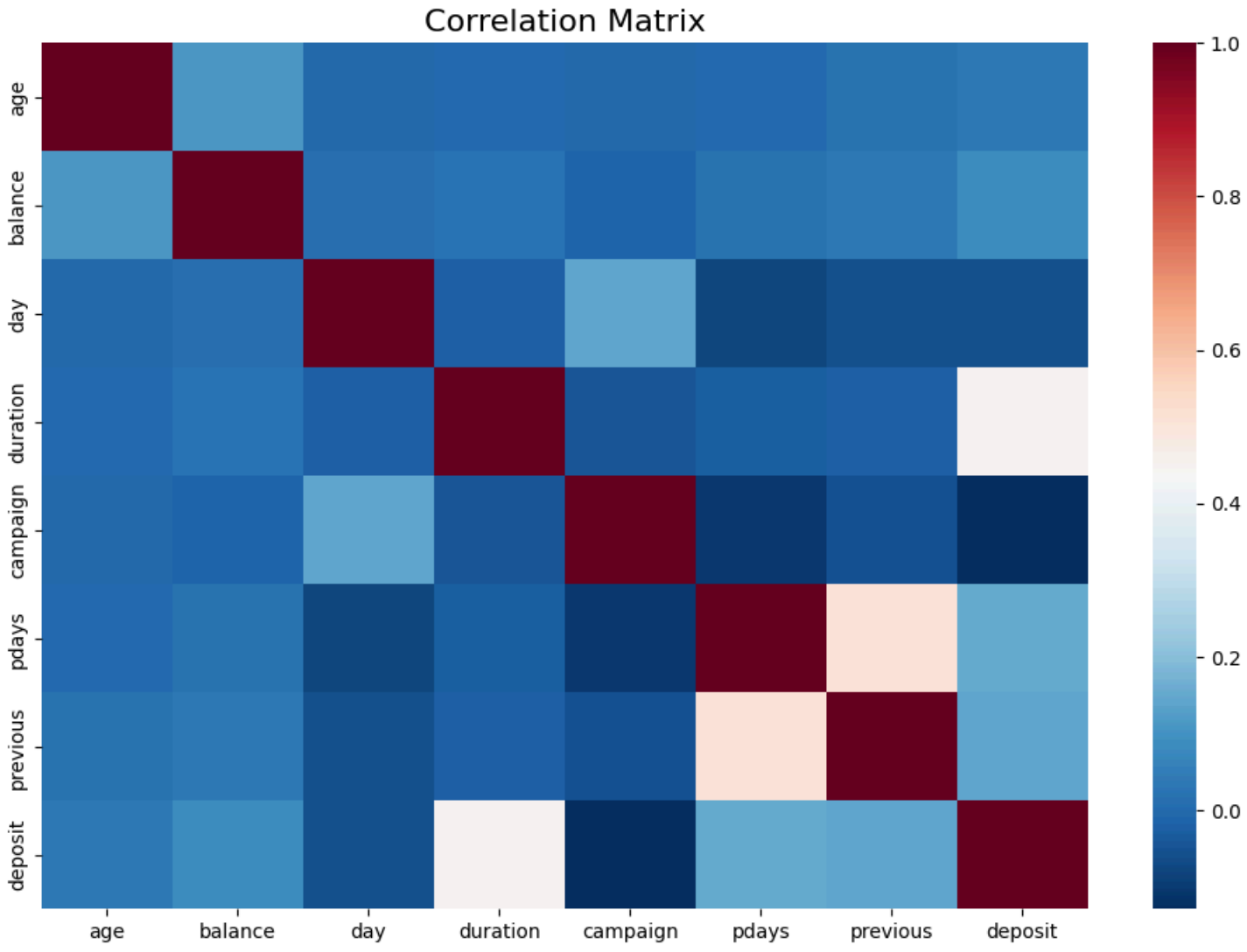
We can see that numerical columns have outliers (especially 'pdays', 'campaign' and 'previous' columns). Possibly there are incorrect values (noisy data), so we should look closer at the data and decide how do we manage the noise.

Findings:

- only 1.2% of values above 400. They are possibly outliers, so we should consider imputing something (possibly mean value) instead of these values.
- -1 possibly means that the client wasn't contacted before or stands for missing data.
- 'campaign' holds the number of contacts performed during this campaign and for this client (numeric, includes last contact) Numbers for 'campaign' above 34 are clearly noise, so I suggest to impute them with average campaign values while data cleaning.
- 'previous' holds the number of contacts performed before this campaign and for this client (numeric) Numbers for 'previous' above 34 are also really strange, so I suggest to impute them with average campaign values while data cleaning.

Correlation Matrix:

- Generated correlation matrices between features, heatmap showing relationships among features (e.g., age, balance, duration)
- With that we can determine if duration has an influence on term deposits.

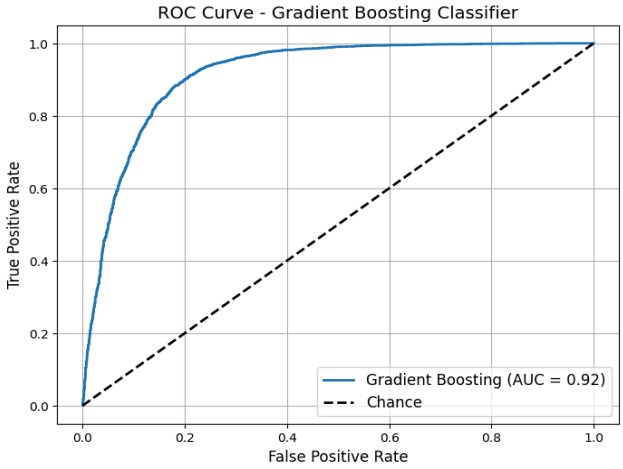


Models Evaluation and Predictive Analysis:

Impact of PCA and Stratified K-folds:

We used a gradient-boosting classifier as an evaluation model to evaluate the performance before and after tuning.

Metric	Before Tuning	After Tuning	Improvement
Precision	82.3%	83.1%	+0.8%
Recall	85.3%	87.3%	+2.0%
F1 Score	83.8%	85.1%	+1.3%
Test Acurracy	84.0%	85.4%	+1.4%
ROC AUC	91.56%	92.01%	+0.45%



ROC curve before hyperparameter tuning

The ROC curve visualizes the trade-off between sensitivity (True Positive Rate) and specificity (False Positive Rate). After tuning:

- The curve moved closer to the top-left corner, indicating better discrimination between classes.
- The AUC increased from 91.56% to 92.01%, confirming enhanced model performance.

Confusion Matrix Analysis

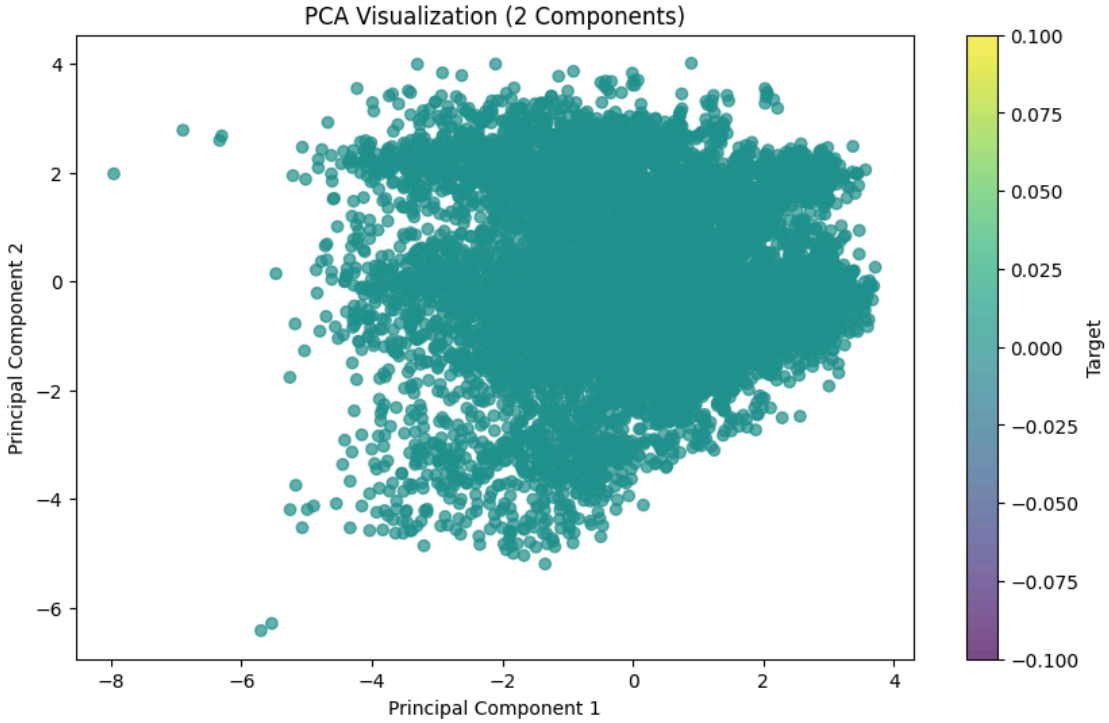
Before Tuning:

- Model correctly classified most of the "No" and "Yes" cases but had some false positives and negatives.

After Tuning:

- The number of false negatives decreased, as indicated by the improved recall.
- The number of false positives also reduced slightly, leading to higher precision.

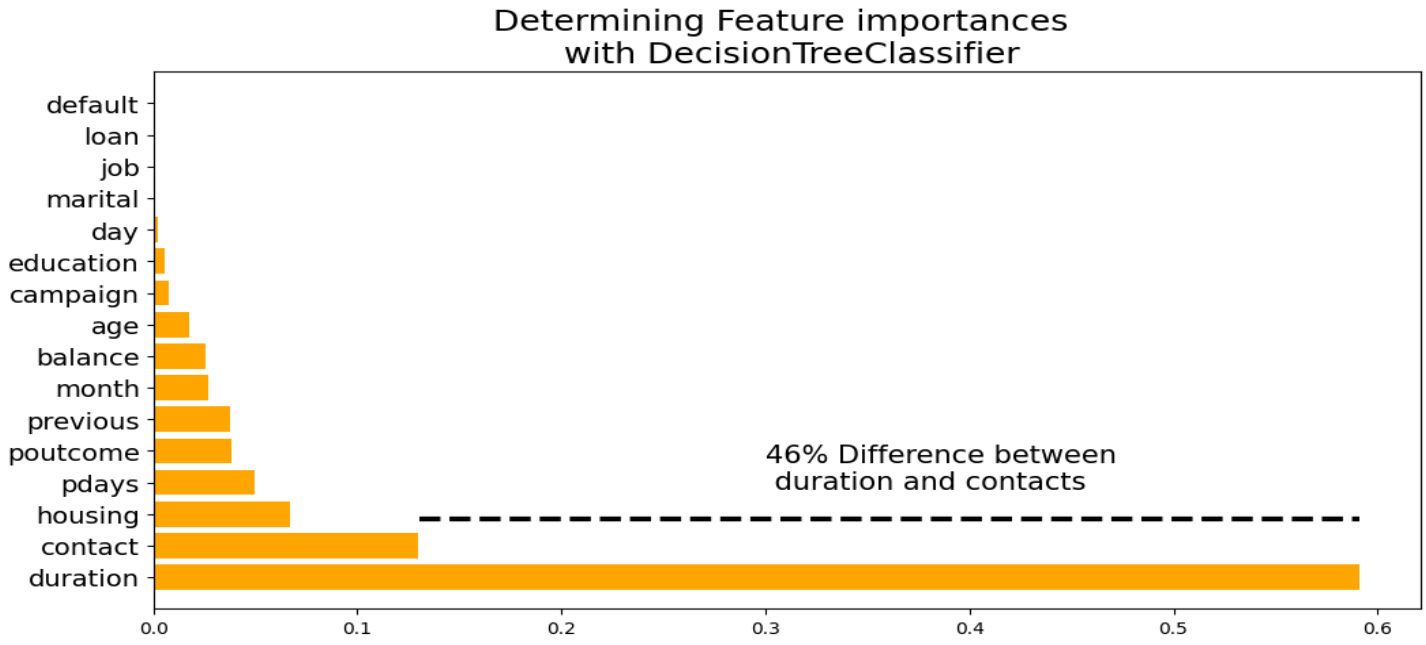
Principal Component Analysis (PCA):

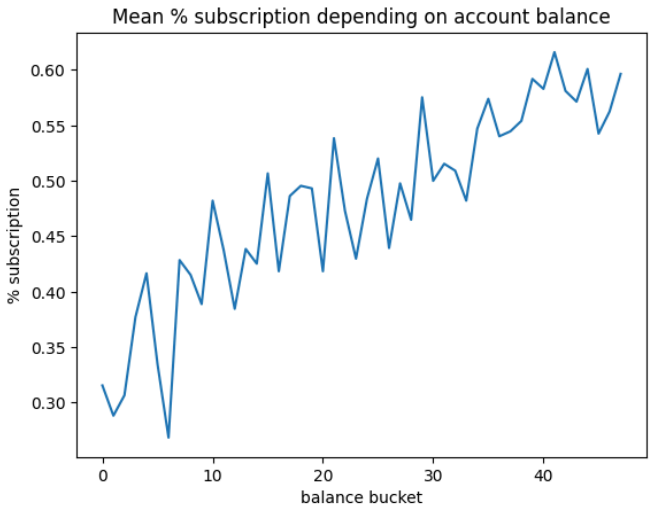


After performing PCA we saw a slight difference in our number of optimal clusters and cluster visualization.

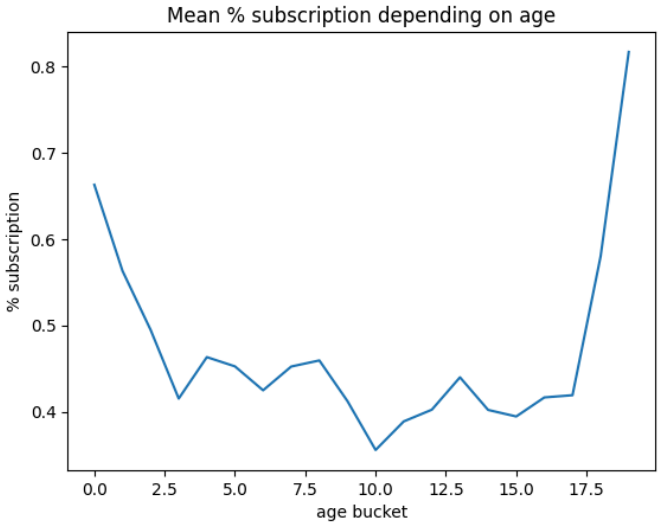
Results:

The top three most important features for our classifier are Duration (how long it took the conversation between the sales representative and the potential client), contact (number of contacts to the potential client within the same marketing campaign), and house.

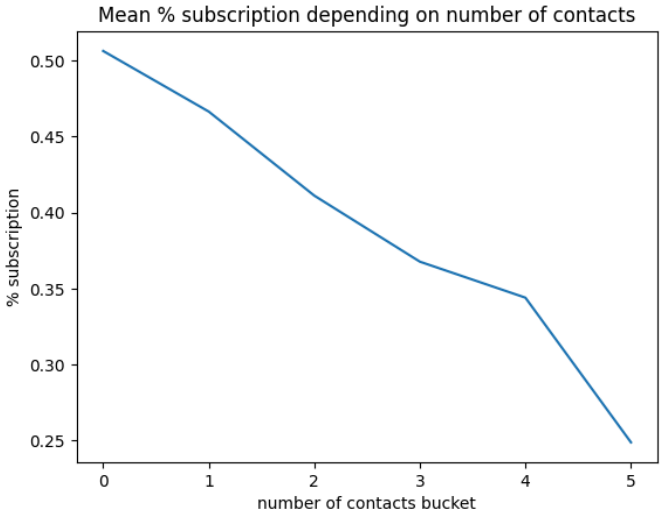




We found that marketing campaigns should concentrate on customers with account balance greater than 1490$.



So we see that the average subscription rate tends to be higher for customers below 31 years old or above 56 years old.



From the plot above we see that the average subscription rate is below 50% if the number of contacts during the campaign exceeds 4.

Conclusion:

This project successfully used machine learning to analyze bank marketing campaigns. Gradient Boosting emerged as the best-performing model, achieving an ROC-AUC of 91.6%. Clustering revealed actionable customer segments, providing a foundation for future targeted campaigns.

The methodology highlights the importance of:

- Data preprocessing and outlier handling.
- Model optimization through hyperparameter tuning.
- Dimensionality reduction techniques like PCA for clustering.

The results offer banks a data-driven roadmap to optimize their marketing strategies and improve customer acquisition rates.

Key outcomes of the analysis are the recommendations for future marketing campaigns:

- The customer's account balance has a huge influence on the campaign's outcome. People with account balance above 1490$ are more likely to subscribe for term deposit, so future addresses those customers.
- The customer's age affects the campaign outcome as well. Future campaigns should concentrate on customers from age categories below 30 years old and above 50 years old.
- Number of contacts with the customer during the campaign is also very important. The number of contacts with the customer shouldn't exceed 4.

Reference:

- https://github.com/ahad-02/CSP571-Project