Nosipho Mfusi

Customer Segmentation and Marketing Campaign Optimization Report

Introduction

This project aimed to leverage the Bank Marketing Dataset to identify customer segments for personalised marketing campaigns. By segmenting customers and analysing the characteristics of each cluster, the project provides actionable insights to improve campaign effectiveness and target the most promising customer groups.

Data Preprocessing

The dataset underwent extensive preprocessing to prepare for clustering and classification:

- **Feature Encoding:** Ordinal encoding was applied to ordered categorical features (e.g., education), while one-hot encoding was used for nominal categorical features.
- **Scaling:** Numerical features were standardised using `StandardScaler` to ensure uniformity.
- **Handling Missing Data:** Missing values were addressed by imputing median values for numerical features.

Exploratory Data Analysis (EDA)

Key insights from EDA include:

- **Distribution of Numerical Features:** Histograms revealed the underlying distributions of numerical attributes such as age, balance, and the duration of calls.
- **Boxplots:** Balance and call duration were explored against the subscription outcome, highlighting significant differences between customers who subscribed ("yes") and those who did not ("no").
- **Correlation Analysis:** A heatmap displayed correlations between numerical features, helping to identify multicollinearity and relationships.

Clustering Analysis

Methodology

KMeans clustering was applied to the scaled dataset to identify customer segments. The elbow method suggested three optimal clusters, which were further validated by profiling their characteristics.

Cluster Characteristics

The clusters were summarised based on the mean values of numerical features:

Cluster	Average Age	Average Balance	Days Passed Since Last Contact	Times Contacted This Campaign
0	42.29	1338.62	230.10	2.84
1	33.38	1288.12	225.81	2.65
2	55.99	1795.42	178.20	2.58

Cluster Sizes

- **Cluster 0:** 28,838 customers
- **Cluster 1:** 12,635 customers
- **Cluster 2:** 3,738 customers

Cluster Insights

Cluster 2 exhibited the highest average balance and the lowest days passed since the last contact. These attributes, combined with the age distribution, suggest that this cluster may include customers more likely to respond positively to campaigns. Further analysis should focus on Cluster 2 to refine marketing strategies.

Classification Model

A Random Forest Classifier was trained to predict subscription outcomes based on the identified clusters and customer features.

Model Performance

- **Accuracy:** 88.42%
- **Classification Report:**
    - Precision for "yes": 52%
    - Recall for "yes": 21%
    - F1-score for "yes": 29%

- **Confusion Matrix:**
    - True positives: 336
    - True negatives: 11,598
    - False positives: 368
    - False negatives: 1,262

Feature Importance

The top 10 features contributing to the model were identified, highlighting the critical role of balance, age, and days since the last contact in predicting subscription outcomes.

Silhouette Score

The silhouette score for the clustering was 0.083, indicating modest separation between clusters. While this score is relatively low, the distinct cluster profiles provide valuable insights for targeted marketing.

Recommendations

1. **Focus on Cluster 2:** Given its high average balance and short time since the last contact, Cluster 2 should be prioritised for personalised marketing campaigns. Strategies could include premium offers or financial products tailored to older customers with substantial balances.
2. **Improve Data Collection:** Enhancing feature granularity and reducing noise could lead to better cluster separation and model accuracy.
3. **Experiment with Advanced Clustering Techniques:** Techniques such as DBSCAN or Gaussian Mixture Models may improve cluster quality by accounting for non-linear relationships.
4. **Refine Campaign Messaging:** Tailor messages to the unique characteristics of each cluster, leveraging the insights gained from profiling.

Conclusion

The project successfully identified three customer segments, with Cluster 2 emerging as the most promising group for targeted campaigns. The use of clustering and classification provided a comprehensive framework for analysing customer behaviour and improving marketing outcomes. Future efforts should build on these findings to enhance predictive accuracy and campaign effectiveness.