

Bank Marketing

Objective

The primary objective of this project is to develop a machine-learning model to predict whether a client will subscribe to a term deposit based on data collected from direct marketing campaigns (e.g., phone calls). The dataset contains various features related to the client's demographic, financial, and interaction history with the bank. The target attribute, Subscribed, is binary (yes or no), indicating whether the client subscribed to a term deposit.

Learning Method

For this project, we utilized a **Logistic Regression model** as the classification algorithm. Logistic regression is suitable for binary classification tasks and provides a probabilistic framework to predict the likelihood of class membership.

Data Preprocessing:

1. Check for missing values
2. Unknown values
3. Replace unknown values with mode values, if any
4. Label encoding
5. one hot encoding
6. scaler - for Standardization
7. Train with logistic regression

Model Training:

Used logistic regression to train the model on the trainset.csv dataset.

Testing:

Evaluated the trained model on testSet.csv.

Used metrics such as precision, recall, F1-score, and accuracy for performance evaluation.

Results

The performance of the logistic regression model was assessed using the classification reports for both the training and test sets.

Train Set Results:

Accuracy: 96%

No (Unsubscribed):

Precision: 95%, Recall: 100%, F1-score: 98%

Yes(Subscribed):

Precision: 99%, Recall: 60%, F1-score: 75%

The model performs exceptionally well for unsubscribed clients but struggles with identifying subscribed clients.

Test Set Results:

Accuracy: 93%

No (Unsubscribed):

Precision: 95%, Recall: 97%, F1-score: 96%

Yes (Subscribed):

Precision: 72%, Recall: 65%, F1-score: 68%

The model generalizes well for class 0(Unsubscribed) but has moderate difficulty correctly identifying subscribed clients.

Key Insights:

Class imbalance has a significant effect, as the model performs much better on the majority class (Unsubscribed -> 0).

The overall performance, measured by accuracy, is high, but recall and F1-score for the minority class (Subscribed -> 1) indicate room for improvement.

Discussion and Conclusion

The logistic regression model provided reliable predictions for the majority class (unsubscribed) but struggled with the minority class (subscribed). This is likely due to the imbalanced nature of the dataset, where the no class significantly outweighs the yes class. Despite this limitation, the model achieved high accuracy overall and demonstrated consistent performance on the test set, indicating minimal overfitting.

Challenges:

Class imbalance negatively impacted the model's ability to recall subscribed clients.

Logistic regression may not fully capture complex relationships in the data.