Purwadhika Digital Technology School

Bank Marketing Campaign Effectiveness Prediction Model

Presented By:

Lintang Rizki Ramadhani

JCDS 2504-011

Contents

Contents

Contents

Contents

Context

- Term deposits are a popular financial product where customers deposit money in a bank for a fixed period .
- Term deposit investments usually carry short-term maturities ranging from one month to a few years and will have varying levels of required minimum deposits.
- To stay competitive and attract new customers, banks often run marketing campaigns to offer this term deposit product.

Problem Statement

A bank wants to **improve the effectiveness of its marketing campaigns.**

Goal

Predict whether a customer will invest to a term deposit (yes/no) after being contacted by the bank.

Analytical Approach

- **Analyze** customer personal information and marketing information.
- **Build a classification model** that can accurately predict whether a customer will invest to a term deposit (yes/no)

Evaluation Metric

		Predicted: Positive	Predicted: Negative
Actual	Positive Invest Term Deposit	True Positive TP	False Negative FN
	Negative Not Invest Term Deposit	False Positive FP	True Negative TN

F1 Score

F1 score is an evaluation metric that combines precision and recall in one value, to assess the performance of a classification model.

F1 Score helps minimizing the False Positives and False Negatives the accuracy and sensitivity of the model, ensuring that marketing campaigns can be effectively targeted toward potential subscribers.

Contents

Contents

Dataset

1 Personal Information

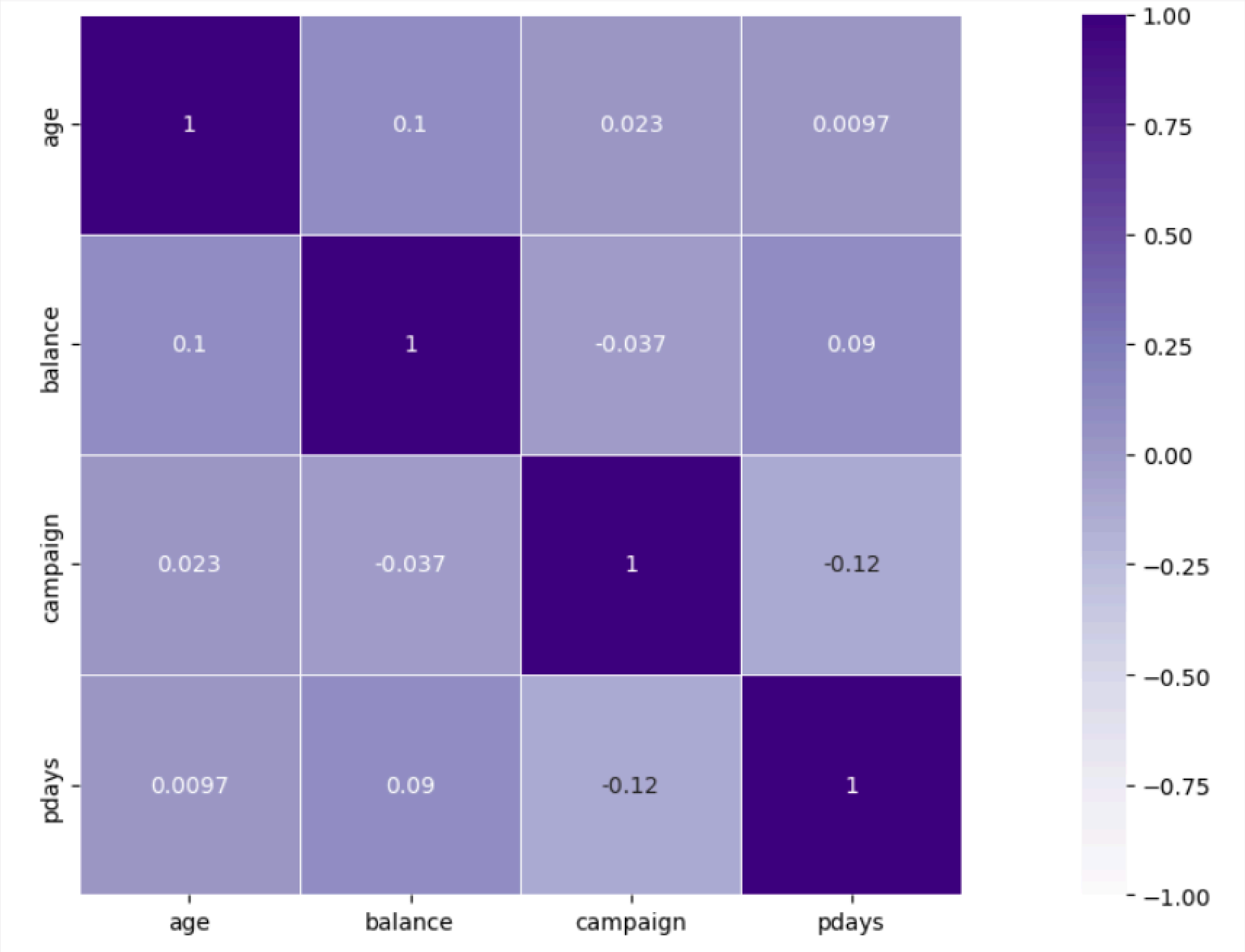
Column	Data Type	Description
age	integer	The age of the customer.
job	object	The customer's occupation (e.g., admin., self-employed, services, housemaid).
balance	integer	The account balance (in monetary units) of the customer.
housing	object	Whether the customer has a housing loan (yes/no).
loan	object	Whether the customer has a personal loan (yes/no).

Dataset

2 Marketing Information

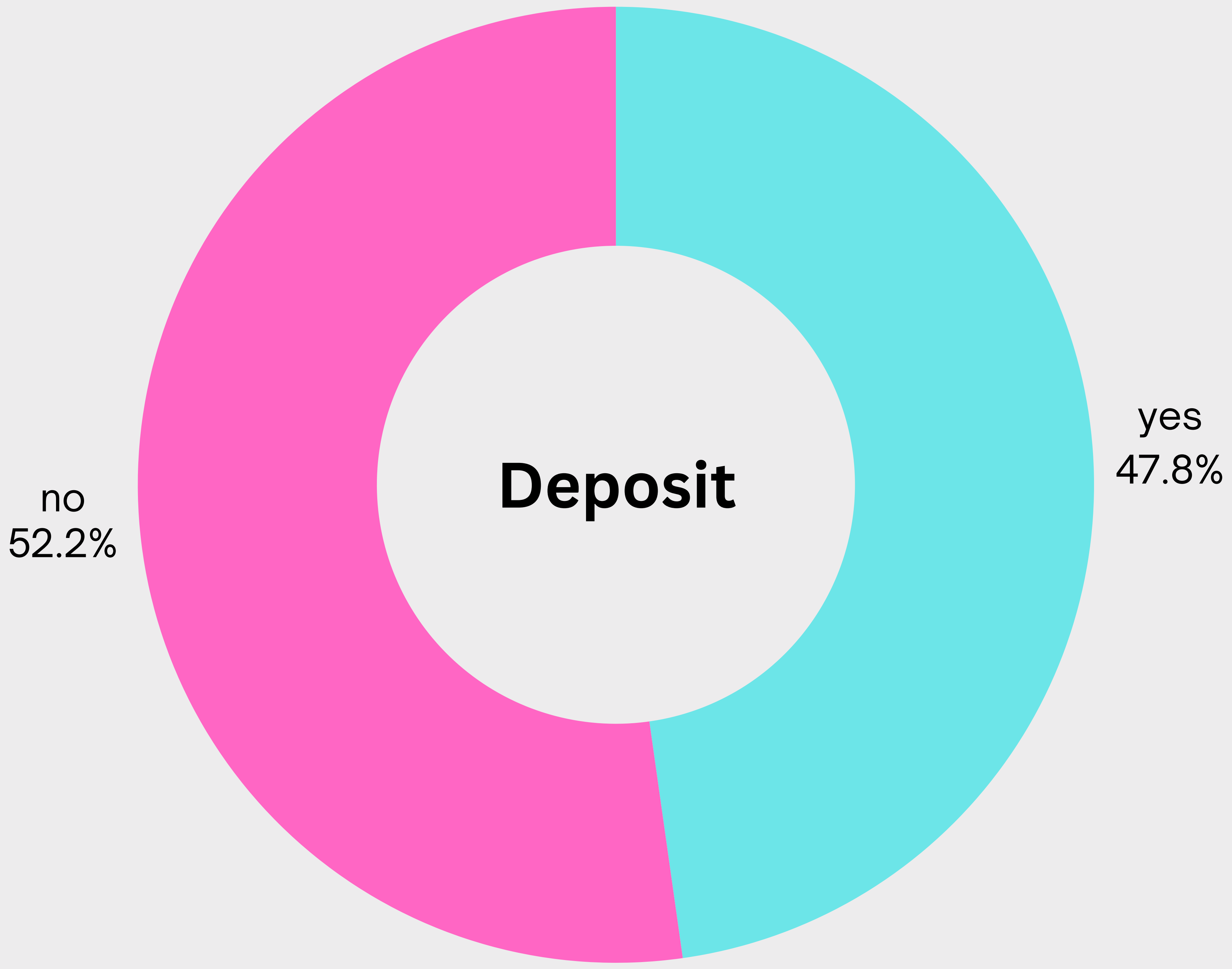
Column	Data Type	Description
contact	object	The contact communication type used in the campaign (e.g., cellular, telephone).
month	object	The last contact month of the year during the marketing campaign.
campaign	integer	The number of contacts performed during this campaign for this customer.
pdays	integer	The number of days since the customer was last contacted in a previous campaign (-1 indicates no prior contact).
poutcome	object	The outcome of the previous marketing campaign (e.g., success, failure, other, unknown).
deposit	object	Whether the customer subscribed to a term deposit (yes/no).

Correlation Analysis



There are no multicolinearity between any features.

Target Variable/Label Analysis



Contents

Contents

Split Data

Train Dataset

80 %

Model Fitting

Test Dataset

20 %

Evaluation

Data Cleaning & Feature Engineering

1

Handling Duplicated Data

Dropping duplicated data since there are only 5 duplicates.

2

Handling Missing Values

There are no null/missing values found in the data but there are ‘unknown’ value in several categorical columns such as job, contact and poutcome. We could treat the 'unknown' values as missing values.

‘Unknown’ value → NaN value → SimpleImputer (strategy = ‘most_frequent’)

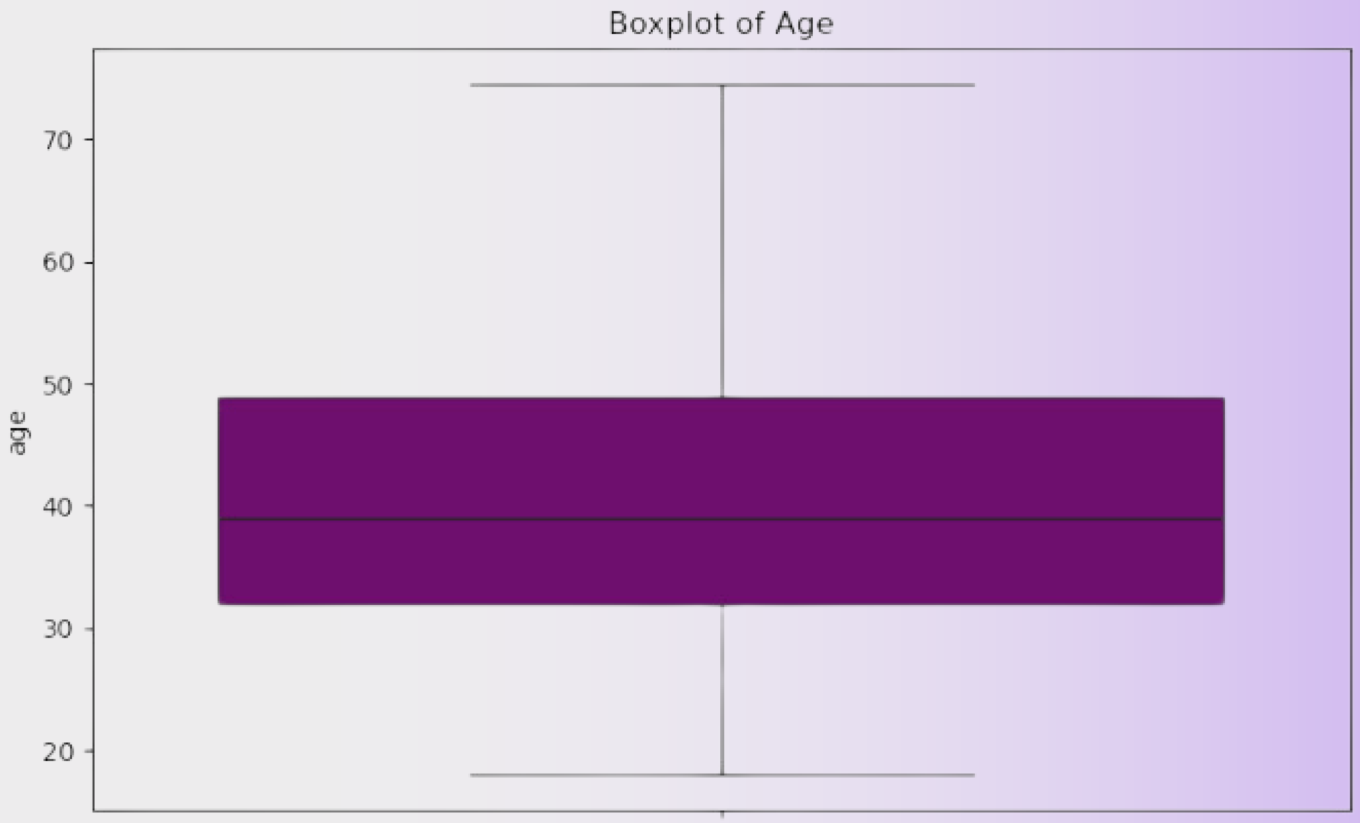
Data Cleaning & Feature Engineering

3 Handling Outliers

Using Inter Quartile Range (IQR) for removing outliers from numerical features

Before Handling Outliers

After Handling Outliers



Data Cleaning & Feature Engineering

4

Handling Categorical Variables

One Hot Encoder →

- job
- housing
- loan
- contact
- poutcome
- month

5

Features Scaling

Robust Scaler →

- age
- balance
- campaign
- pdays

Transformation Pipelines

A machine learning (ML) pipeline is a series of steps that process and model data to streamline the process of working with ML models

```
categorical_features = ['job', 'housing', 'loan', 'contact', 'poutcome','month']
categorical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('encoder', OneHotEncoder(handle_unknown='ignore'))])

numeric_features = ['age','balance', 'pdays','campaign']
numeric_transformer = Pipeline(steps=[
    ('scaler', RobustScaler())])
```

```
# Combine transformers into a single preprocessor
transformer = ColumnTransformer(
    transformers=[
        ('numerical', numeric_transformer, numeric_features),
        ('categorical', categorical_transformer, categorical_features)
    ]
)
```

Contents

Contents

Modelling

Default Hyperparameters & Cross Validation

Model	Mean F1 Score	Std. F1 Score
Random Forest Classifier	0.645798	0.006234
XG Boost Classifier	0.633459	0.006669
Logistic Regression	0.624448	0.011982
Decision Tree Classifier	0.594932	0.011540

Modelling

Hyperparameters Tuning Using Cross Validation

Parameters	Mean F1 Score	Std. F1 Score
RandomForestClassifier(class_weight = 'balanced', criterion='entropy', max_depth=6, min_samples_split=10, random_state=42)	0.649915	0.015513
RandomForestClassifier(class_weight = 'balanced', criterion='gini', max_depth=8, min_samples_split=20, random_state=42)	0.649061	0.019361
RandomForestClassifier(class_weight = 'balanced', criterion='gini', max_depth=6, min_samples_split=10, random_state=42)	0.648595	0.015344
RandomForestClassifier(class_weight = 'balanced', criterion='entropy', max_depth=6, min_samples_split=8, random_state=42)	0.648530	0.015998

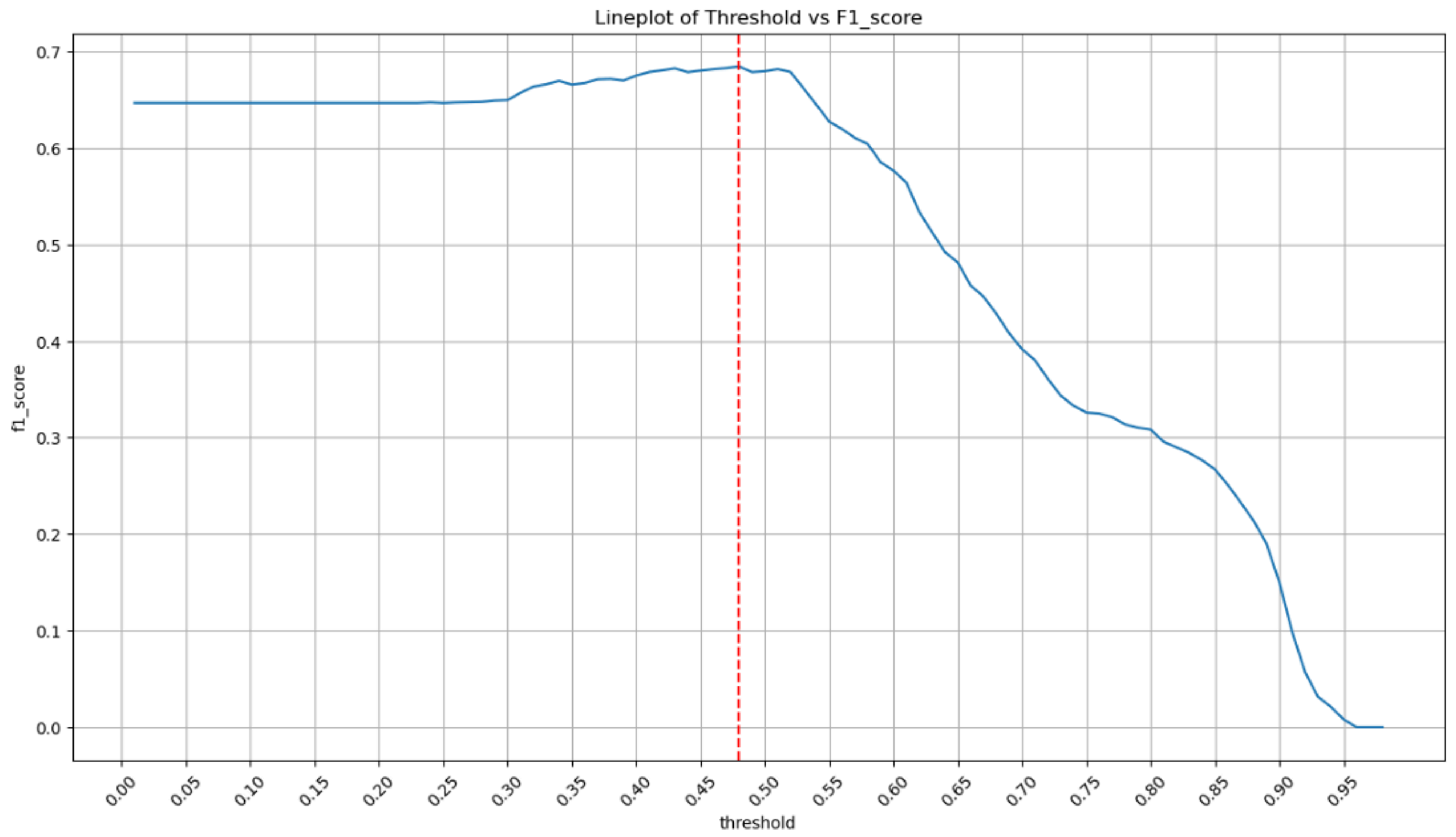
Final Model Based On Evaluation Metric (F1 Score)

Tuning	Best Model	Mean F1 Score	Std. F1 Score
Before Tuning	Random Forest Classifier (random_state=42)	0.645798	0.006234
After Tuning	RandomForestClassifier(class_weight = 'balanced', criterion='entropy', max_depth=6, min_samples_split=10, random_state=42)	0.649915	0.019361

Best Model

Modelling

Threshold Adjustment



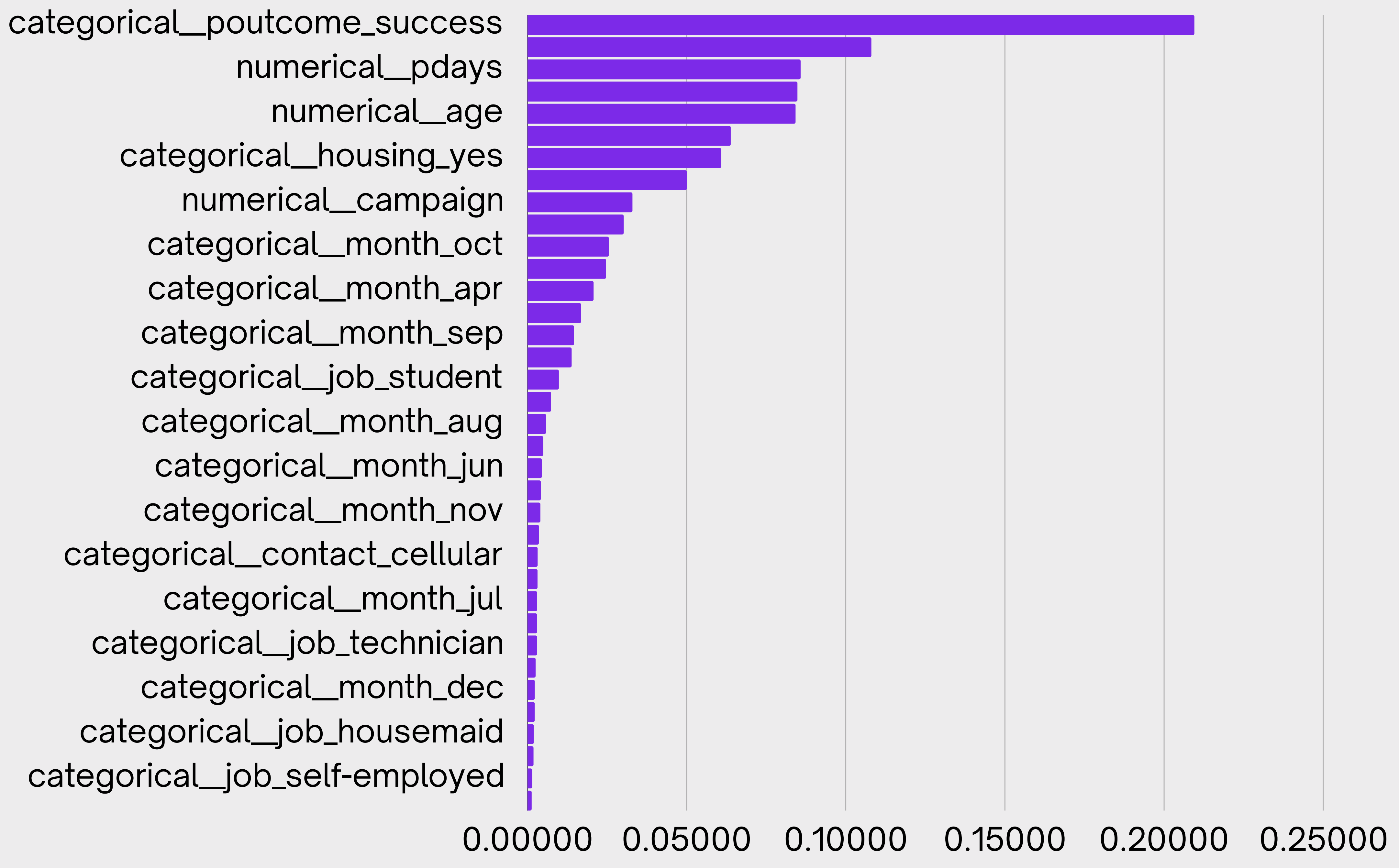
	precision	recall	f1-score	support
0	0.71	0.73	0.72	816
1	0.70	0.67	0.68	747
accuracy			0.70	1563
macro avg	0.70	0.70	0.70	1563
weighted avg	0.70	0.70	0.70	1563

Feature Importances

Top 3 Most Important Features

- The success or failure of previous marketing outcomes is the most important feature, indicating that customers with a successful past outcome are much more likely to invest, and those with a failed outcome are less likely.

- The number of days since a customer was last contacted is the third most important feature, with a higher influence, likely reflecting that recently contacted customers are more likely to invest.



Contents

Contents

Conclusion

Random Forest Classifier Model with Hyperparameter Tuning and Threshold Adjustment has the best performance (highest F1-Score) in predicting whether a customer will make a deposit or not after being contacted by the bank.

Contents

Contents

Recommendation

Model Improvement

- **Generate New Features**

  Some features like job, month, and contact show minimal importance to the model.

- **Avoid filling data with 'unknown' values**

  poutcome feature contains a lot of ‘unknown’ values which we treated as a missing values in this modelling, meaning that a lot of values in poutcome features were missing.

- **Explore another model**

  Implement bagging or other boosting models like Gradient Boosting, AdaBoost, or CatBoost. Or even use Deep Learning.

Thank You