

77778//

**Classification Analysis Report**

Overview

# **1.Objective**: This study's goal is to use regression to predict a target variable. Numerous attributes pertaining to Exposure Categories are present in the World Risk Report dataset. Data preparation, EDAs, model creation, cross-validation, hyperparameter adjustment, feature selection, and model building are all covered in this study.

# **2. Dataset Description**

**Dataset Name:** Bank Marketing Dataset

**Source:** Kaggle

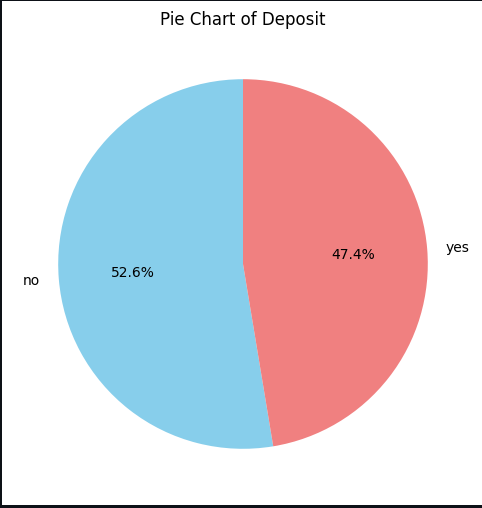
Description: There are more than 11959 rows and 15 columns in this bank report collection.

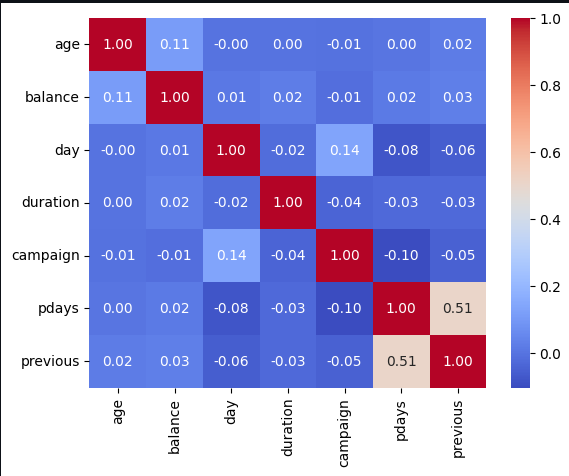
The report explores the bank marketing dataset, predicts term deposit subscriptions, analyses features, and evaluates model performance for insights.

# **3.Data Preprocessing**

**Data Cleaning:**

Data cleaning include addressing missing values, fixing mistakes, eliminating duplicates, and assuring consistency in order to prepare data for analysis.





**Feature Transformation:**

# The numerical characteristics were normalized using Standard Scaler, while categorical variables were encoded with Label Encoder and One-Hot Encoder.

# **4.Exploratory Data Analysis (EDA)** Histograms, bar charts, pie chart = where no is (52.6% and yes is 47.4%) and a heatmap were generated to analyse feature distributions and identify patterns. Insights: Accuracy, precision, recall, and f1 scores were used to train and assess the models.

# **5. Model Building**

Logistic Regression from scratch

Logistic Regression from sklearn

Gradient boosting classifier

Training Process:

Models were trained and evaluated based on Accuracy score, precision score, recall score and f1 score

# **6. Model Evaluation**

Performance

**Logistic Regression Implementation (Sigmoid)**

Accuracy: 0.7789

Precision: 0.7916

Recall: 0.7226

F1 Score: 0.7556

Train Loss: 0.4990

**Primary Model:**

**the Logistic Regression model:**

Accuracy on training set: 0.7902340687646993

**Class 0**:

* Precision: 0.78
* Recall: 0.83
* F1 Score: 0.81
* Support: 4707

**Class 1**:

* Precision: 0.80
* Recall: 0.75
* F1 Score: 0.77
* Support: 4222

Accuracy on test set: 0.7819077474249888

**Class 0:**

Precision: 0.78

Recall: 0.81

F1 Score: 0.79

Support: 1166

**Class 1:**

Precision: 0.78

Recall: 0.75

F1 Score: 0.77

Support: 1067

**Gradient boosting classifier model:**

**Training Set Metrics:**

* **Accuracy:** 0.85
* **Class 0 (Precision):** 0.87, **Recall:** 0.84, **F1 Score:** 0.86, **Support:** 4707
* **Class 1 (Precision):** 0.83, **Recall:** 0.86, **F1 Score:** 0.85, **Support:** 4222

**Test Set Metrics:**

* **Accuracy:** 0.8204
* **Class 0 (Precision):** 0.84, **Recall:** 0.81, **F1 Score:** 0.82, **Support:** 1166
* **Class 1 (Precision):** 0.80, **Recall:** 0.84, **F1 Score:** 0.82, **Support:** 1067

# **7.Hyperparameter** The Logistic Regression and Random Forest Classifier models' hyperparameters were optimized using Grid Search CV.

# **8. Feature Selection**

**Techniques used:**

Select K best and Recursive Feature Elimination were used to select features.

# **9. Conclusion**

**Final Model:**

**Gradient boosting classifier after hyperparameter:**

Final Model Training Metrics:

* Training Accuracy: 0.8531
* Training Precision (Class 0): 0.87, Training Precision (Class 1): 0.83
* Training Recall (Class 0): 0.84, Training Recall (Class 1): 0.86
* Training F1 Score (Class 0): 0.86, Training F1 Score (Class 1): 0.85

Final Model Testing Metrics:

* Testing Accuracy: 0.8204
* Testing Precision (Class 0): 0.84, Testing Precision (Class 1): 0.80
* Testing Recall (Class 0): 0.81, Testing Recall (Class 1): 0.83

s F1 Score Testing (Class 0): 0.82; F1 Score Testing (Class 1): 0.82 With an accuracy of 0.8531 on the training set and 0.8204 on the test set, the final Gradient Boosting Classifier demonstrated good performance. Excellent generalization was suggested by the balanced precision, recall, and F1 scores between the two courses. The model is appropriate for practical uses in bank marketing since it predicts deposit subscriptions with accuracy.

**Limitation:**

The model's performance, while strong, may still be affected by data imbalances, particularly in predicting minority classes. Additionally, it might overfit the training set due to the complexity of the Gradient Boosting Classifier. Further tuning and testing on more diverse data could help improve generalization and reduce the risk of overfitting, enhancing real-world applicability.

**Future Work:**

To increase accuracy, future research could concentrate on fine-tuning hyperparameters and experimenting with more sophisticated models like XGBoost or Neural Networks. Furthermore, model performance may be improved by resolving class imbalance by methods like class weighting. Predictions for bank marketing efforts could be improved by testing the model on fresh, untested data and investigating feature engineering approaches.