Exploratory Data Analysis (EDA) and Data Preprocessing

Overview

This project focuses on applying **Exploratory Data Analysis (EDA)** and data preprocessing to the `marketing_data.csv` dataset. The goal is to clean and prepare the data for a **binary classification model** that predicts whether a customer will subscribe to a term deposit.

Dataset Summary

- **Total Rows:** 43,097
- **Total Columns:** 17
- **Target Variable:** `y` (Subscription Status: `yes` = 1, `no` = 0)

Column Overview

Feature	Type	Description
`age`	Numeric	Age of the customer
`job`	Categorical	Type of job
`marital`	Categorical	Marital status
`education`	Categorical	Level of education
`default`	Categorical	Has credit in default?
`balance`	Numeric	Average yearly balance in euros
`housing`	Categorical	Has housing loan?
`loan`	Categorical	Has personal loan?
`contact`	Categorical	Contact communication type
`day`	Numeric	Last contact day of the month
`month`	Categorical	Last contact month of the year
`campaign`	Numeric	Number of contacts performed during this campaign
`pdays`	Numeric	Days since last contact (-1 if not previously contacted)
`previous`	Numeric	Number of contacts before this campaign
`Location`	Categorical	Customer’s location
`poutcome`	Categorical	Outcome of the previous campaign
`y`	Binary Target	Subscribed (`yes` = 1, `no` = 0)

Data Cleaning & Preprocessing

1. Handling Missing Data

- `age` had **23 missing values**, replaced with the **median**.
- `contact` had **58 missing values**, replaced with the **mode**.
- `poutcome` had **10 missing values**, replaced with the **mode**.

2. Removing Duplicates

- Found **3 duplicate rows** and removed them.

3. Handling Outliers

Applied the **Interquartile Range (IQR) method** to cap outliers for numeric columns:

- `age`
- `balance`
- `campaign`
- `pdays`
- `previous`

4. Encoding Categorical Variables

- Used **one-hot encoding** for categorical variables (`job`, `marital`, `education`, etc.).
- **Dropped first category** in each one-hot encoding to avoid multicollinearity.
- Converted **target variable (`y`) into binary format**: `yes` → 1, `no` → 0.

5. Feature Scaling

- Standardized **numerical features** using `StandardScaler()` to ensure equal weightage in the classification model.

Final Processed Dataset

- **Shape After Cleaning:** `(43,094, X)` (after handling missing data and duplicates)
- **Missing Values After Cleaning:** `0`
- **Encoded categorical variables** and **scaled numerical features** for model training.

Next Steps

- Apply feature selection techniques.
- Train classification models (Logistic Regression, Random Forest, etc.).
- Evaluate model performance using Precision, Recall, and F1-score.

Repository Structure

```
├── data/                  # Raw and processed datasets
├── notebooks/             # Jupyter notebooks for EDA and preprocessing
├── models/                # Trained classification models
├── README.md              # Project documentation (this file)
└── requirements.txt       # Dependencies
```

Acknowledgments

This project is inspired by **bank marketing campaigns** to improve targeted customer outreach. The dataset originates from real-world financial marketing efforts.