Project Protocol

Bank Marketing Targets

Almog Cohen

Introduction:

Business Problem:
The objective of this project is to predict whether a customer will subscribe to a term deposit (binary classification) based on various personal, financial, and campaign-related features. The ability to accurately predict this response can help the business optimize its marketing strategies and improve customer targeting efficiency.

Dataset Summary:
The dataset used for this project is the UCI Bank Marketing Dataset. It contains approximately 50,000 rows, each representing a customer.

- The target variable is y, indicating whether the client subscribed to a term deposit.
- The dataset is **imbalanced**, with about 88% of the responses being "no" and only 12% being "yes".

Data Preparation:

The dataset features were grouped into three logical clusters for better clarity and understanding: General Customer Info, Campaign Contact Details, and Other Attributes. Below are the detailed summaries of each cluster.

Customer Demographics and Profile

Column	Explanation	Data Type	Numeric/Categorical
age	Age of the customer	int	Numeric
job	Type of job	string	Categorical
marital	Marital status	string	Categorical
education	Education level	string	Categorical
default	Has credit in default?	bool	Categorical
housing	Has housing loan?	bool	Categorical
loan	Has personal loan?	bool	Categorical

Campaign Information:

Column	Explanation	Data Type	Numeric/Categorical
contact	Contact communication type (last campaign contact)	string	Categorical
day	Last contact day of the month	int	Numeric
month	Last contact month of the year	string	Categorical
duration	Last contact duration (seconds)	int	Numeric
campaign	Number of contacts during this campaign	int	Numeric
pdays	Days since last contact from previous campaign	int	Numeric
previous	Number of contacts before this campaign	int	Numeric
poutcome	Outcome of previous marketing campaign	string	Categorical

Target Variable:

Column	Explanation	Data Type	Numeric/Categorical
y	Has the client subscribed a term deposit?	bool	Categorical

Exploratory Data Analysis (EDA):

With the data properly prepared and features typed, we move on to exploratory data analysis. This phase involves:

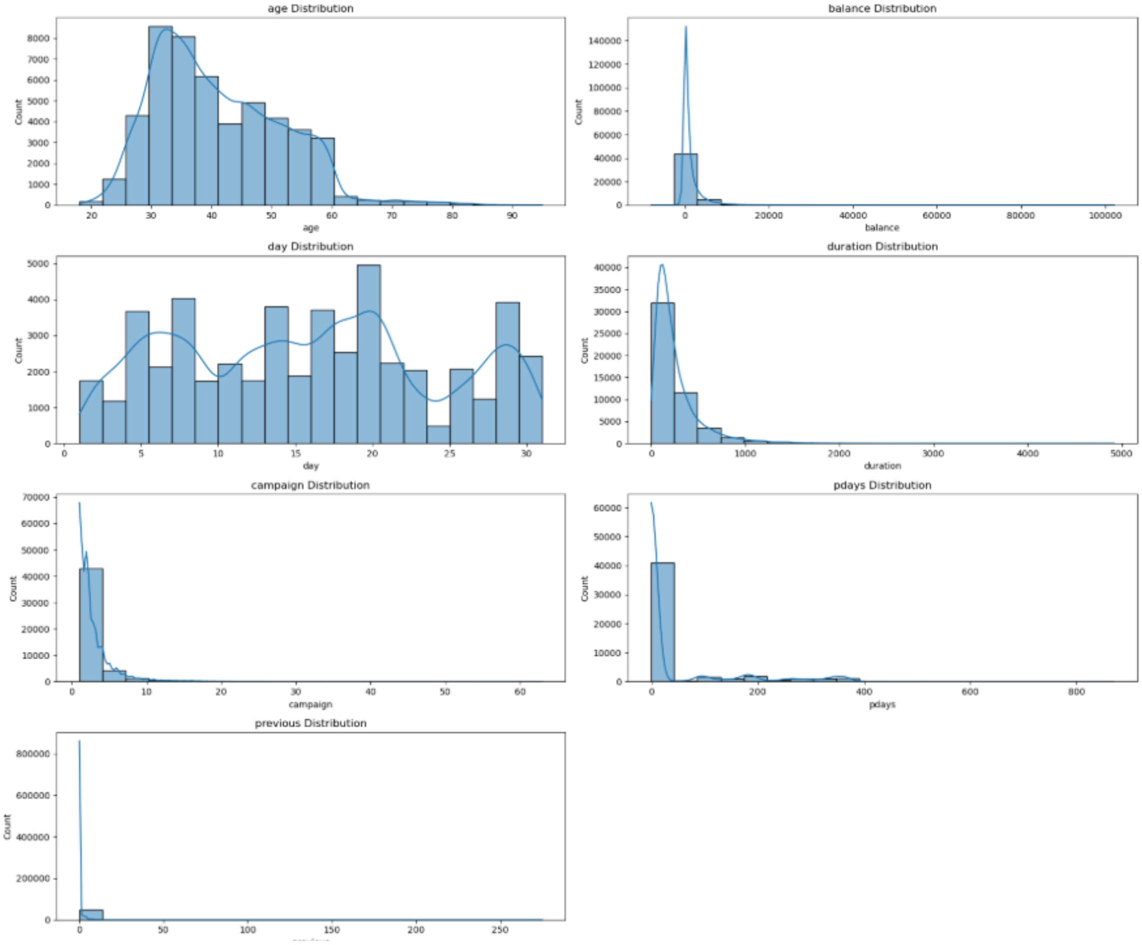
- Understanding the distribution of numeric features
- Examining relationships between variables
- Identifying potential outliers and missing values
- Visualizing categorical feature frequencies and interactions with the target variable

The insights gathered here will guide feature engineering and model selection for the project.

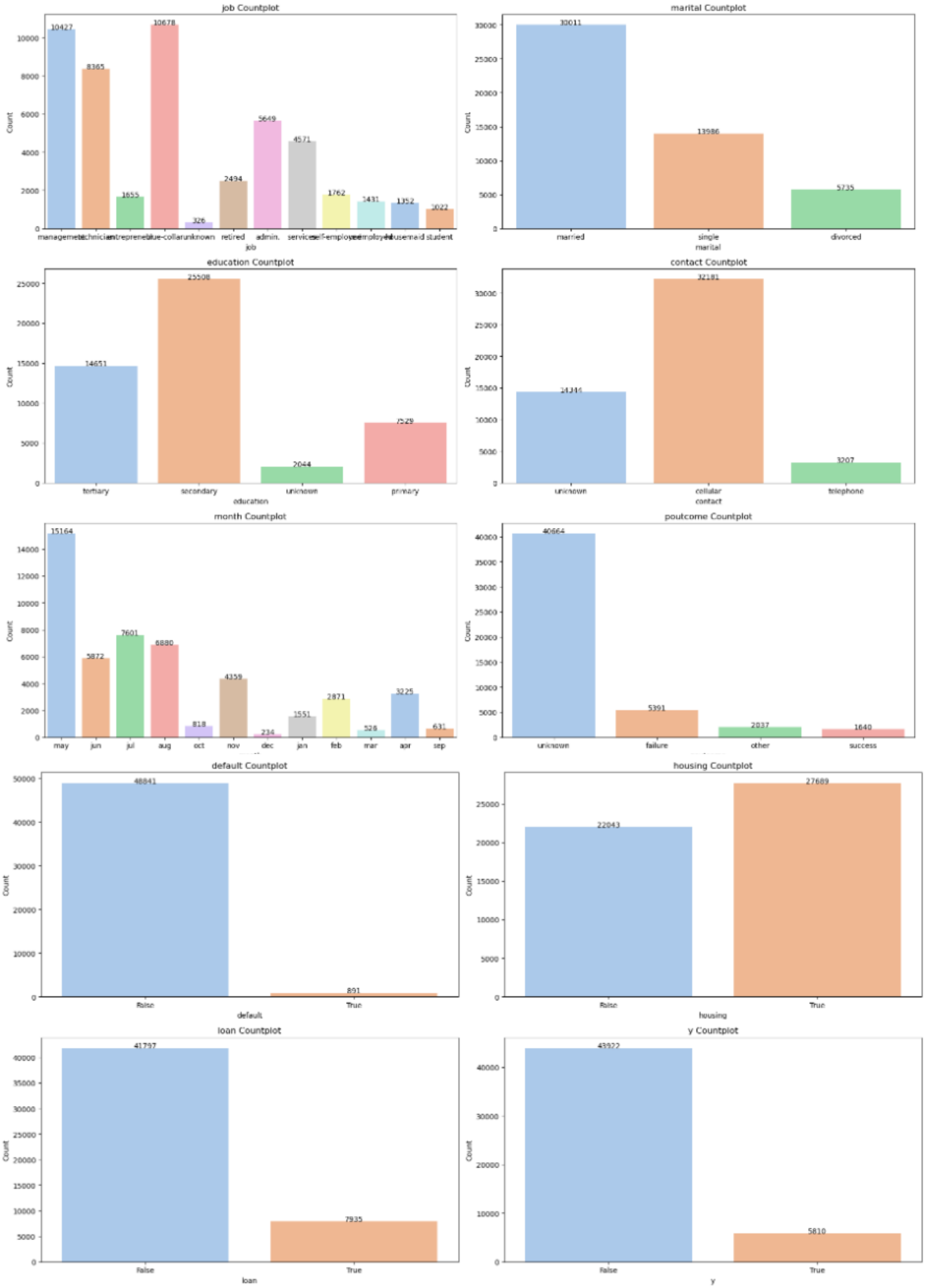
Feature Visualizations

Before diving into the statistical analysis, we explored the data visually to understand feature distributions and relationships.

- **Numeric Features:** Histograms and boxplots were used to observe the distribution, spread, and presence of outliers for key numeric variables such as age, balance, duration, and campaign.



- **Categorical Features:** Bar charts and count plots helped visualize the frequency of categories within features like job, marital status, education, and contact type.


job Countplot
Count
0
2000
4000
6000
8000
10000
10427
8365
1655
10678
326
2494
5649
4571
1762
1431
1352
1022
management
technician
entrepreneur
blue-collar
unknown
retired
admin.
services
self-employed
unemployed
housemaid
student
job
marital Countplot
Count
0
5000
10000
15000
20000
25000
30000
30011
13986
5735
married
single
divorced
marital
education Countplot
Count
0
5000
10000
15000
20000
25000
14651
25508
2044
7529
tertiary
secondary
unknown
primary
education
contact Countplot
Count
0
5000
10000
15000
20000
25000
30000
14344
32181
3207
unknown
cellular
telephone
contact
month Countplot
Count
0
2000
4000
6000
8000
10000
12000
14000
15164
5872
7601
6880
818
4359
234
1551
2871
526
3225
631
may
jun
jul
aug
oct
nov
dec
jan
feb
mar
apr
sep
poutcome Countplot
Count
0
5000
10000
15000
20000
25000
30000
35000
40000
40664
5391
2037
1640
unknown
failure
other
success
default Countplot
Count
0
10000
20000
30000
40000
50000
48841
891
False
True
default
housing Countplot
Count
0
5000
10000
15000
20000
25000
22043
27689
False
True
housing
loan Countplot
Count
0
5000
10000
15000
20000
25000
30000
35000
40000
41797
7935
False
True
loan
y Countplot
Count
0
10000
20000
30000
40000
43922
5810
False
True
y

Skewness Analysis

Before applying correlation tests, we assessed the distribution shape of numeric features:

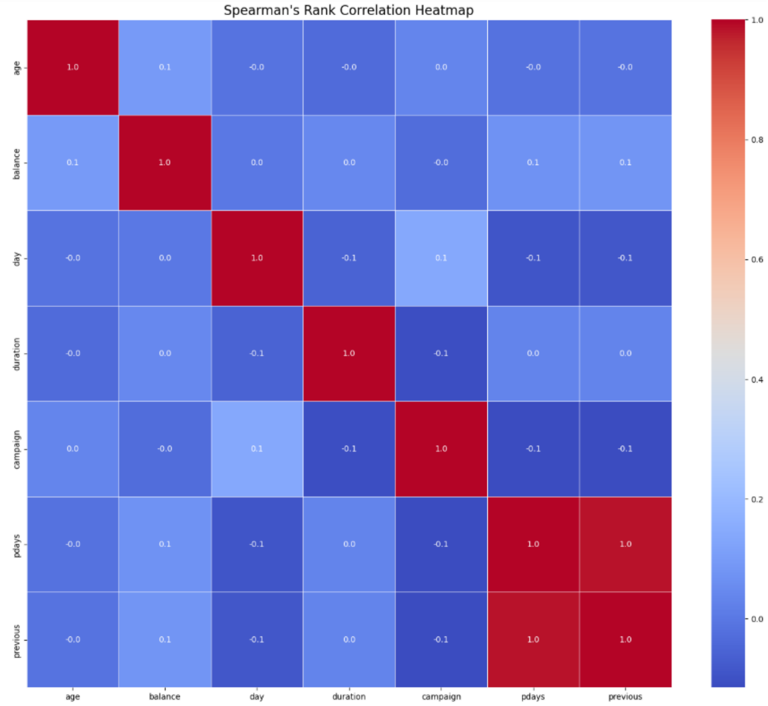
- **Key Insight:**
  Out of 7 numeric features, 6 were found to be non-normally distributed, exhibiting moderate to high skewness.
- **Decision:**
  Due to the prevalent non-normality, we opted to use Spearman correlation, a non-parametric measure suitable for detecting monotonic relationships in skewed or ordinal data.

Spearman Correlation Analysis

- **Strong Correlation:**
  - pdays and previous have an extremely strong positive correlation (~0.99), reflecting their close relationship (number of previous contacts and days since last contact).
- **Weak Correlations:**
  - A few numeric feature pairs show weak correlations, such as day & campaign (0.14), balance & pdays (0.07), and balance & previous (0.08).
- **No Correlation:**
  - Most other numeric pairs have very low correlation (absolute values < 0.1), indicating they provide mostly independent information.

Insight:

Due to the near-perfect correlation, pdays and previous may be redundant. Other numeric features contribute unique signals.



Chi-Square Test for Categorical Variables

- The majority of categorical feature pairs exhibit **statistically significant associations** (p-value < 0.05), meaning these features are dependent.
- The only exception is the pair default and housing (p-value = 0.2891), which are independent.
- Using Cramér's V to measure association strength:
  - Strong associations (≥ 0.25) found between pairs like contact & month, housing & month, and education & job.
  - Moderate associations (0.15 to 0.25) include contact & poutcome and job & marital.
  - Most other pairs have weak associations (< 0.15).

Insight:
Dependent categorical features may carry overlapping information, so multicollinearity and redundancy should be monitored.

Outlier Analysis Conclusion

We conducted an outlier detection using the Interquartile Range (IQR) method on all numeric features to assess their impact on the data distribution and relationship with the target variable.

Several features—such as balance, duration, and campaign—exhibited noticeable outliers. However, since these features also demonstrated meaningful correlations with the target, we decided **not to remove any outliers** at this stage. This approach preserves potential predictive signals carried by extreme values.

Missing Values

No missing values were detected in the dataset; hence, no imputation or data cleaning for missingness was required.

Feature Engineering:

Encoding of Categorical Variables

- **One-hot encoding:** Applied to nominal categorical features without ordinal relationships:
  job, marital, contact, month, and poutcome were transformed into binary indicator columns to enable model interpretability and compatibility.

- **Ordinal encoding:** Applied to the education feature using a meaningful order:
  'primary' = 1, 'secondary' = 2, 'tertiary' = 3, 'unknown' = 4
  This preserves the natural ranking in education levels, which can enhance model performance.

- **Binary mapping:** Features with boolean-like values, including 'default', 'housing', 'loan', and target variable 'y', were mapped from text or boolean to numeric 0/1 encoding to streamline processing.

Newly Created Features

Feature Name	Description	Business Rationale
Wealth Indicator	Categorizes customers based on bank account balance into four groups: negative, low, medium, and high wealth.	Enables segmentation based on financial strength, which can affect responsiveness to marketing offers. Targets can be customized based on wealth profiles.
Campaign Intensity	Ratio of current campaign contacts (campaign) to previous contacts (previous + 1).	Quantifies customer contact frequency in campaigns. High values may signal potential contact fatigue; low values indicate untapped engagement opportunities. Helps optimize marketing outreach strategies.

Feature Selection:

Feature Selection Methodology

To identify the most relevant features for our predictive model, we applied multiple feature selection methods, each offering a unique perspective on feature importance:

- **Lasso Regression (L1 penalty)**
- **Gradient Boosting**
- **Random Forest**
- **Ridge Regression (L2 penalty)**
- **XGBoost**

Each method scored features based on their importance or coefficient magnitude. We combined these results into a consensus table where each feature received a score from 0 to 5, representing how many models selected it as important.

Feature Selection Summary and Decision Criteria

- Features with **low consensus scores (2 or 3)** were dropped due to inconsistent importance across models, which helps reduce noise and complexity.
- Features with **borderline scores (4)** were also dropped to simplify the model further and reduce potential multicollinearity or overfitting risks.
- Only features with **full consensus (score 5)** — selected by all methods — were retained to ensure strong, stable predictive power and model interpretability.

This rigorous selection approach strikes a balance between predictive performance and simplicity, focusing the model on the most consistently valuable features.

Dropping the duration Feature

The duration feature records the length of the last contact call during the campaign. While it is highly predictive of the target variable, it represents **post-outcome information**—known only after the interaction happens.

Including duration introduces **target leakage**, allowing the model to leverage future information unavailable at prediction time, which leads to over-optimistic performance estimates and poor generalization.

To maintain model integrity and ensure predictions are based solely on information available **before contacting the client**, the duration feature was excluded prior to feature selection and model training.

Model Selection:

Model Selection Strategy

To identify the best-performing model for predicting customer responses, we trained and evaluated the following six classification models:

- **Logistic Regression**
- **Support Vector Machine (SVM)**
- **Random Forest Classifier**
- **Gradient Boosting Classifier**
- **K-Nearest Neighbors (KNN)**
- **XGBoost Classifier**

All models were evaluated using a comprehensive set of metrics including Accuracy, Precision, Recall, F1 Score, and AUC (Area Under the ROC Curve).

Potential Overfitting Detected

Initial evaluation showed **very high accuracy scores** across all models but **relatively lower AUC scores**, suggesting possible **overfitting** — where a model learns patterns in the training data too well but fails to generalize to unseen data.

To assess this, we:

1. Trained each model on the training set
2. Calculated AUC for both the **training** and **test** sets
3. Compared the AUC values to detect generalization issues

A significant gap between training and test AUC indicates that the model is likely overfitting.

Final Model Selection: Logistic Regression & Gradient Boosting

Based on the AUC comparison, we observed overfitting in several models, particularly in ensemble-based classifiers like Random Forest and XGBoost. However, two models stood out for their generalization ability:

- **Logistic Regression**
    - **Train AUC:** 0.714
    - **Test AUC:** 0.692
    - Strong baseline with balanced and stable performance
    - Highly interpretable and efficient
- **Gradient Boosting Classifier**
    - **Test AUC:** 0.707 (best among all models)
    - Minor overfitting, but excellent overall performance
    - Known to improve with careful fine-tuning

These two models were selected for **hyperparameter tuning** as finalists for the project

Final Model Selection: Gradient Boosting Classifier

After fine-tuning both Logistic Regression and Gradient Boosting, we selected **Gradient Boosting** as the final model.

Why Gradient Boosting?

Despite the class imbalance in our dataset, Gradient Boosting consistently outperformed Logistic Regression:

Metric	Gradient Boosting	Logistic Regression
Recall	**0.159**	0.081
F1 Score	**0.25**	0.14
AUC	**0.71**	0.69

Although the absolute values are modest (due to class imbalance), **Gradient Boosting demonstrated better ability to identify true positives**, making it more suitable for our business goal: predicting likely responders to marketing campaigns.

Thank You!