CSIS 5420 X1 S1 2025

Machine Learning for Business Analytics

Prem Kumar Chimakurthi

Professor: Dipak Biscuitwala

Master of Science in Business Analytics

Machine Learning Project Report

Project Title: Bank Marketing

Dataset Used:

https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset

1. Executive Summary:

The Bank Marketing Dataset consists of information about customers for a direct marketing campaign of a bank. The main purpose of this project is to examine customer properties and to model the behavior of a client as to whether they will sign up for a term deposit.

In this project, using RapidMiner, we preprocess the data, explore key insights, and fit decision tree, k-NN, Naive Bayes, and logistic regression models to the data to predict the customer behavior. The findings from this project are useful for banks to better design their marketing strategies and increase the effectiveness of their targeting efforts.

2.Introduction:

Banks use direct marketing campaigns (phone calls, emails, etc.) to reach potential customers. However, not all customers respond positively to these campaigns. Identifying which customers are more likely to subscribe to a term deposit can help banks focus their efforts more effectively.

This project aims to:

- **Analyze customer demographics and financial behavior.**
- **Identify key factors influencing term deposit subscription.**
- **Apply machine learning models to predict customer response.**
- **Provide insights to improve future marketing campaigns.**

3.Data Description:

The dataset consists of 11,162 observations with 17 attributes, including customer demographics, financial information, and campaign details. The target variable is "deposit", which indicates whether a customer subscribed to a term deposit (yes or no).

Key Features in the Dataset:

Feature	Type	Description
age	Integer	Age of the customer
job	Categorical	Type of job (e.g., admin, technician, services, etc.)
marital	Categorical	Marital status (married, single, divorced)
education	Categorical	Level of education (secondary, tertiary, primary, unknown)
default	Categorical	Has credit in default? (yes/no)
balance	Integer	Customer's average account balance

housing	Categorical	Has a housing loan? (yes/no)
loan	Categorical	Has a personal loan? (yes/no)
contact	Categorical	Contact communication type (cellular, telephone, unknown)
day	Integer	Last contact day of the month
month	Categorical	Last contact month of the year
duration	Integer	Last contact duration (in seconds)
campaign	Integer	Number of contacts performed during this campaign
pdays	Integer	Days since the client was last contacted (-1 means never)
previous	Integer	Number of contacts before this campaign
poutcome	Categorical	Outcome of the previous marketing campaign
deposit (Target)	Categorical	Whether the client subscribed to a term deposit (yes/no)

4.Dataset Overview:

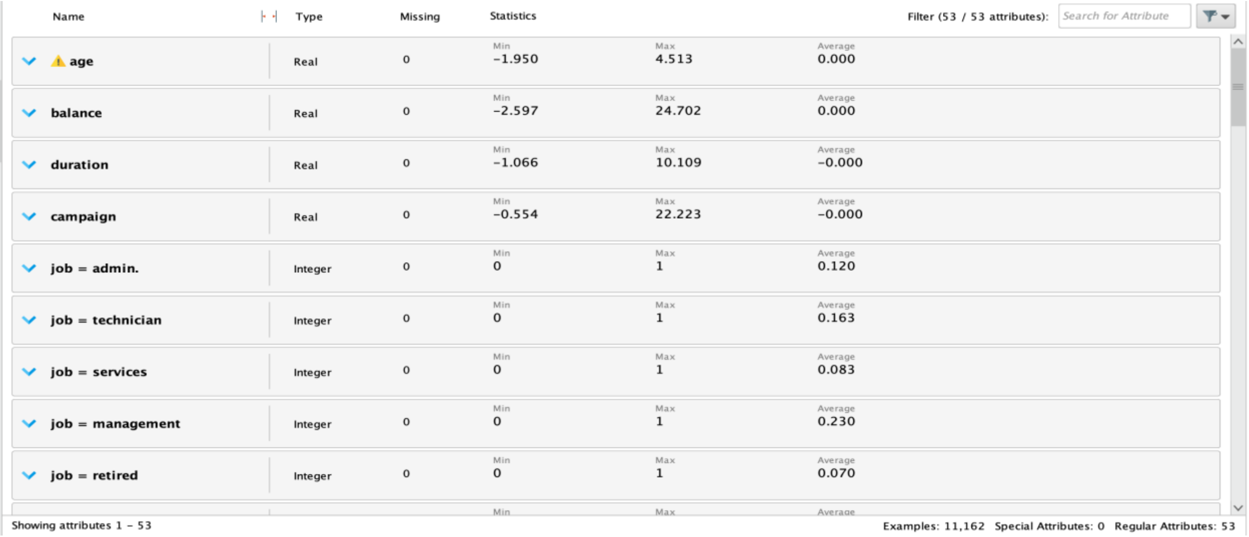
- The dataset consists of both numerical and categorical variables.
- There are no missing values in the dataset, but some values like ‘job’, ‘education’, and ‘poutcome’ contain values like ‘unknown’.
- The target variable ‘deposit’ is imbalanced; there are more ‘no’ than ‘yes’ responses.
- Some numerical variables like balance and duration have outliers that need to be dealt with.

5.Data Preprocessing:

1. Handling Missing & Unknown Values:

- The dataset does not have null values, but categorical variables have "unknown" as a category.

- ❖ These unknown values can be replaced with mode (most frequent value) or kept as a separate category.



Filter (53 / 53 attributes): Search for Attribute

Name	Type	Missing	Statistics		
			Min	Max	Average
age	Real	0	-1.950	4.513	0.000
balance	Real	0	-2.597	24.702	0.000
duration	Real	0	-1.066	10.109	-0.000
campaign	Real	0	-0.554	22.223	-0.000
job = admin.	Integer	0	0	1	0.120
job = technician	Integer	0	0	1	0.163
job = services	Integer	0	0	1	0.083
job = management	Integer	0	0	1	0.230
job = retired	Integer	0	0	1	0.070

Showing attributes 1 – 53

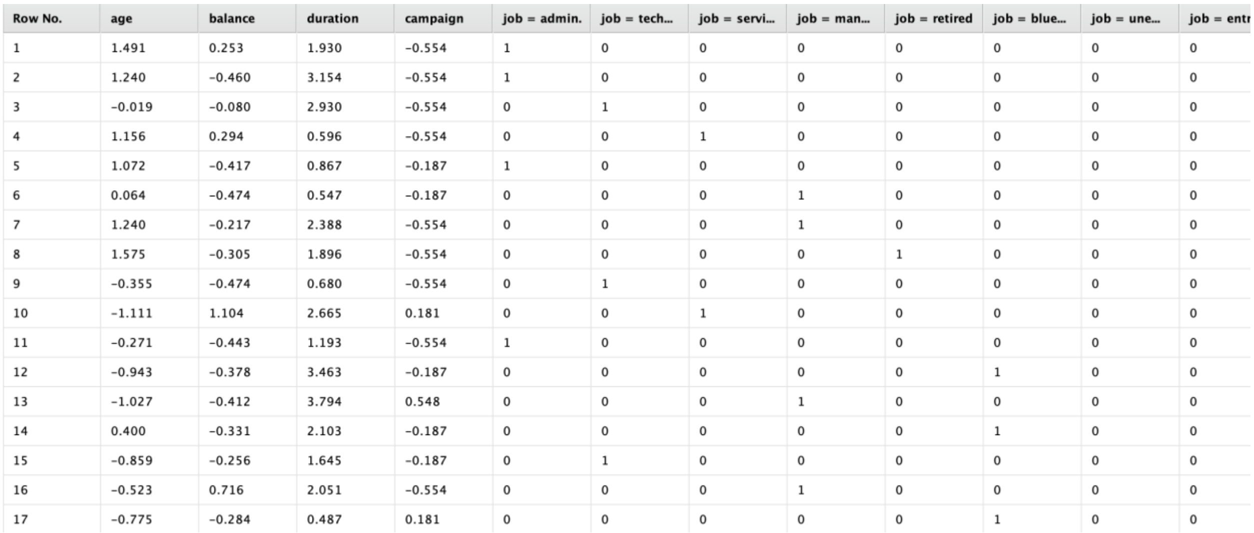
Examples: 11,162 Special Attributes: 0 Regular Attributes: 53

2. Encoding Categorical Variables:

- ❖ Converted categorical variables (**job, marital, education, etc.**) into numerical values using **One-Hot Encoding (Dummy Variables)**.
- ❖ This ensures models like **Decision Trees, k-NN, and Logistic Regression** can process the data.

3. Normalizing & Scaling:

- ❖ Balance, duration, and campaign variables have large numerical ranges, which may affect model performance.
- ❖ Using Min-Max Scaling to normalize values between 0 and 1.

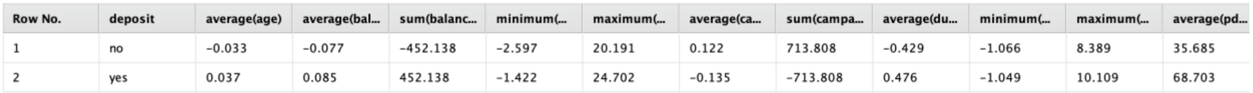


Row No.	age	balance	duration	campaign	job = admin.	job = tech...	job = servi...	job = man...	job = retired	job = blue...	job = une...	job = entr
1	1.491	0.253	1.930	−0.554	1	0	0	0	0	0	0	0
2	1.240	−0.460	3.154	−0.554	1	0	0	0	0	0	0	0
3	−0.019	−0.080	2.930	−0.554	0	1	0	0	0	0	0	0
4	1.156	0.294	0.596	−0.554	0	0	1	0	0	0	0	0
5	1.072	−0.417	0.867	−0.187	1	0	0	0	0	0	0	0
6	0.064	−0.474	0.547	−0.187	0	0	0	1	0	0	0	0
7	1.240	−0.217	2.388	−0.554	0	0	0	1	0	0	0	0
8	1.575	−0.305	1.896	−0.554	0	0	0	0	1	0	0	0
9	−0.355	−0.474	0.680	−0.554	0	1	0	0	0	0	0	0
10	−1.111	1.104	2.665	0.181	0	0	1	0	0	0	0	0
11	−0.271	−0.443	1.193	−0.554	1	0	0	0	0	0	0	0
12	−0.943	−0.378	3.463	−0.187	0	0	0	0	0	1	0	0
13	−1.027	−0.412	3.794	0.548	0	0	0	1	0	0	0	0
14	0.400	−0.331	2.103	−0.187	0	0	0	0	0	1	0	0
15	−0.859	−0.256	1.645	−0.187	0	1	0	0	0	0	0	0
16	−0.523	0.716	2.051	−0.554	0	0	0	1	0	0	0	0
17	−0.775	−0.284	0.487	0.181	0	0	0	0	0	1	0	0

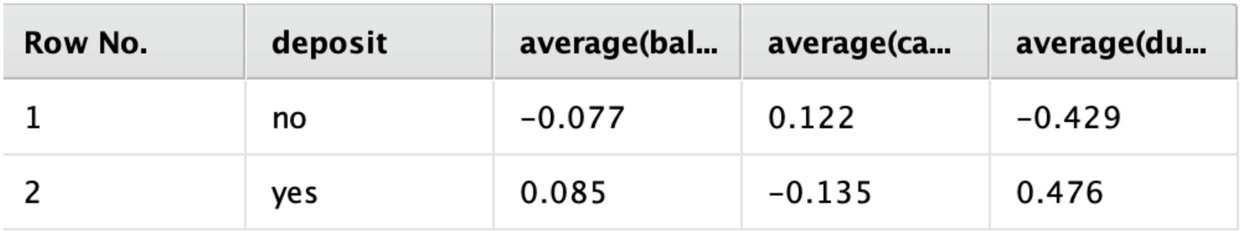
6.Exploratory Data Analysis (EDA):

Univariate Analysis:

- **Goal:** Understand the distribution and summary statistics of each variable.



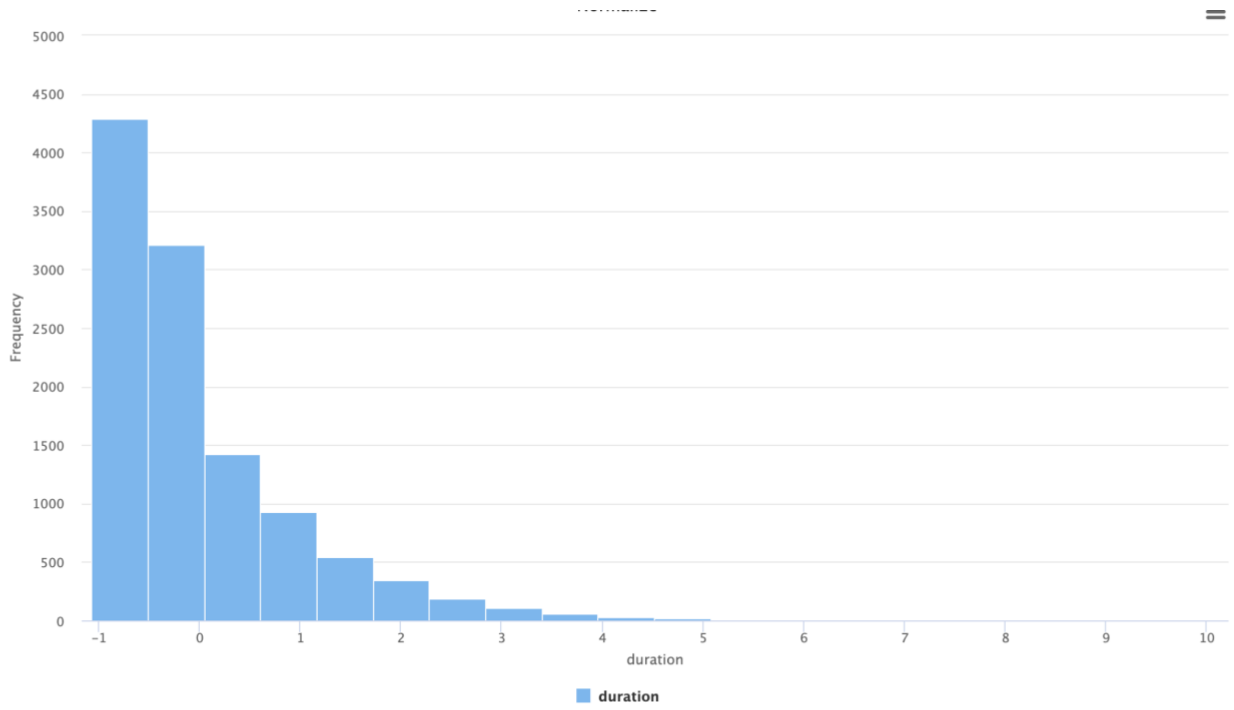
Row No.	deposit	average(age)	average(bal...	sum(balanc...	minimum(...	maximum(...	average(ca...	sum(campa...	average(du...	minimum(...	maximum(...	average(pd...
1	no	−0.033	−0.077	−452.138	−2.597	20.191	0.122	713.808	−0.429	−1.066	8.389	35.685
2	yes	0.037	0.085	452.138	−1.422	24.702	−0.135	−713.808	0.476	−1.049	10.109	68.703



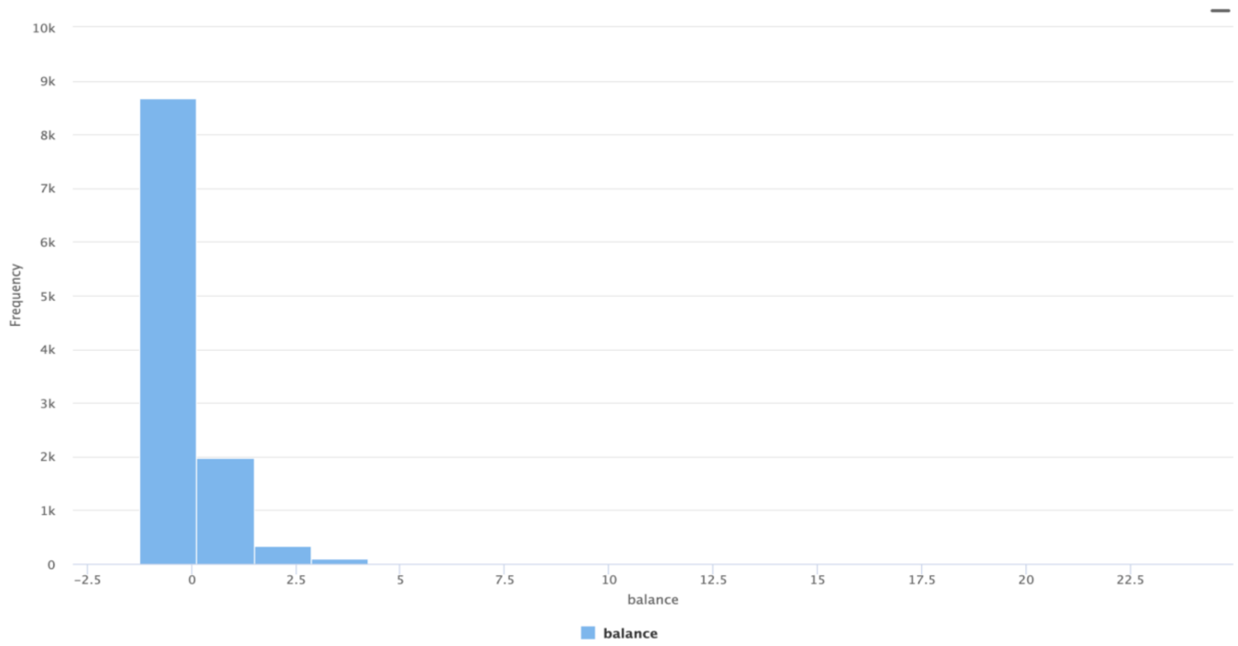
Row No.	deposit	average(bal...	average(ca...	average(du...
1	no	−0.077	0.122	−0.429
2	yes	0.085	−0.135	0.476

-

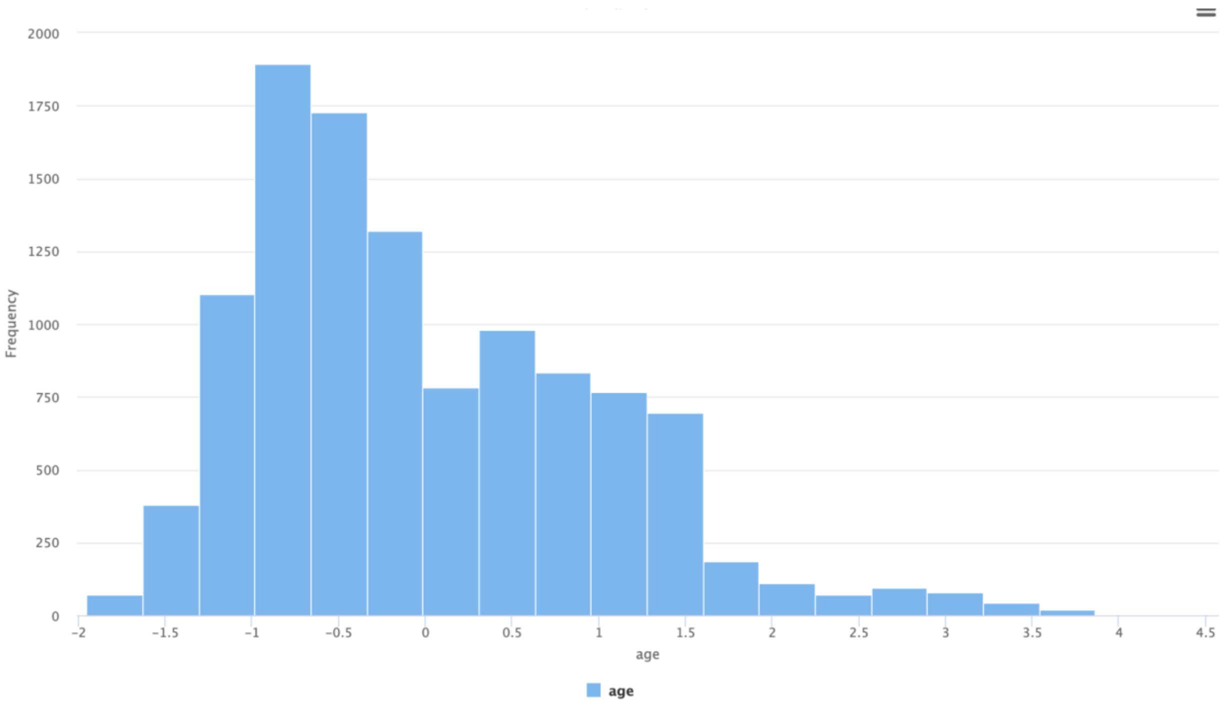
Bivariate Analysis:

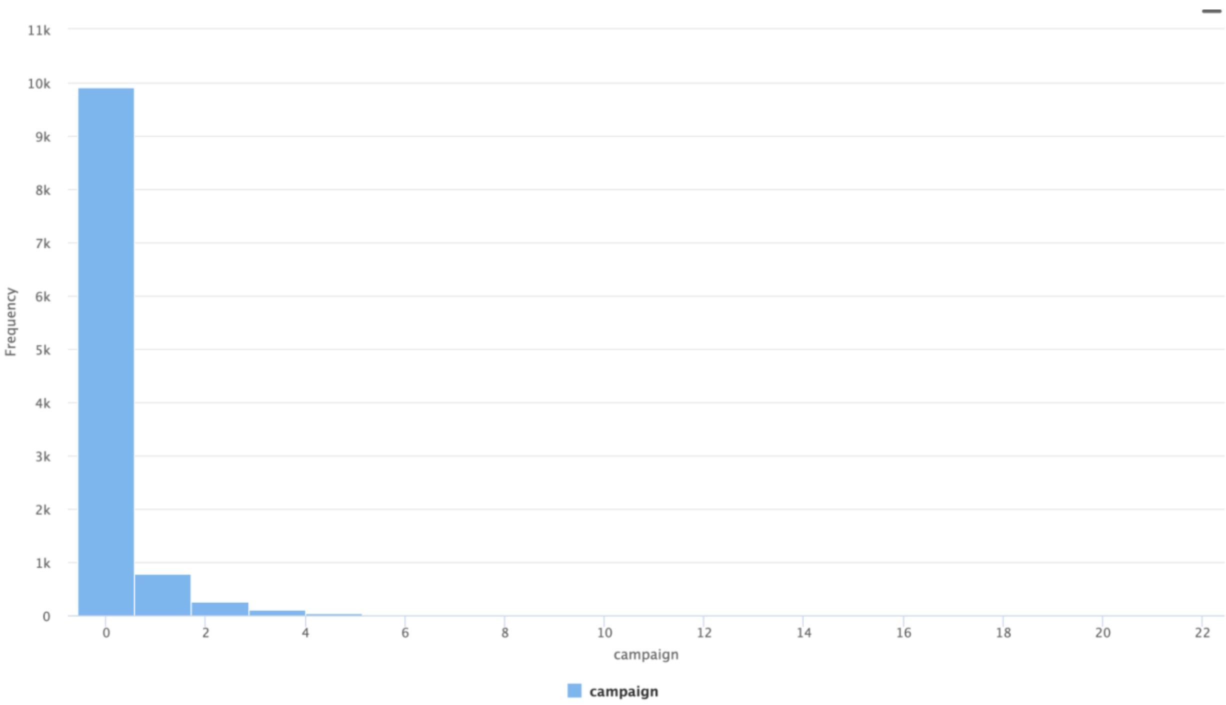


5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
Frequency
−1
0
1
2
3
4
5
6
7
8
9
10
duration
duration



10k
9k
8k
7k
6k
5k
4k
3k
2k
1k
0
Frequency
−2.5
0
2.5
5
7.5
10
12.5
15
17.5
20
22.5
balance
balance





Duration Histogram:

- Right - skewed, meaning most calls were short, but there are a few long calls.
- **Insight**: Longer calls may be strong predictors of deposit subscription.

Balance Histogram:

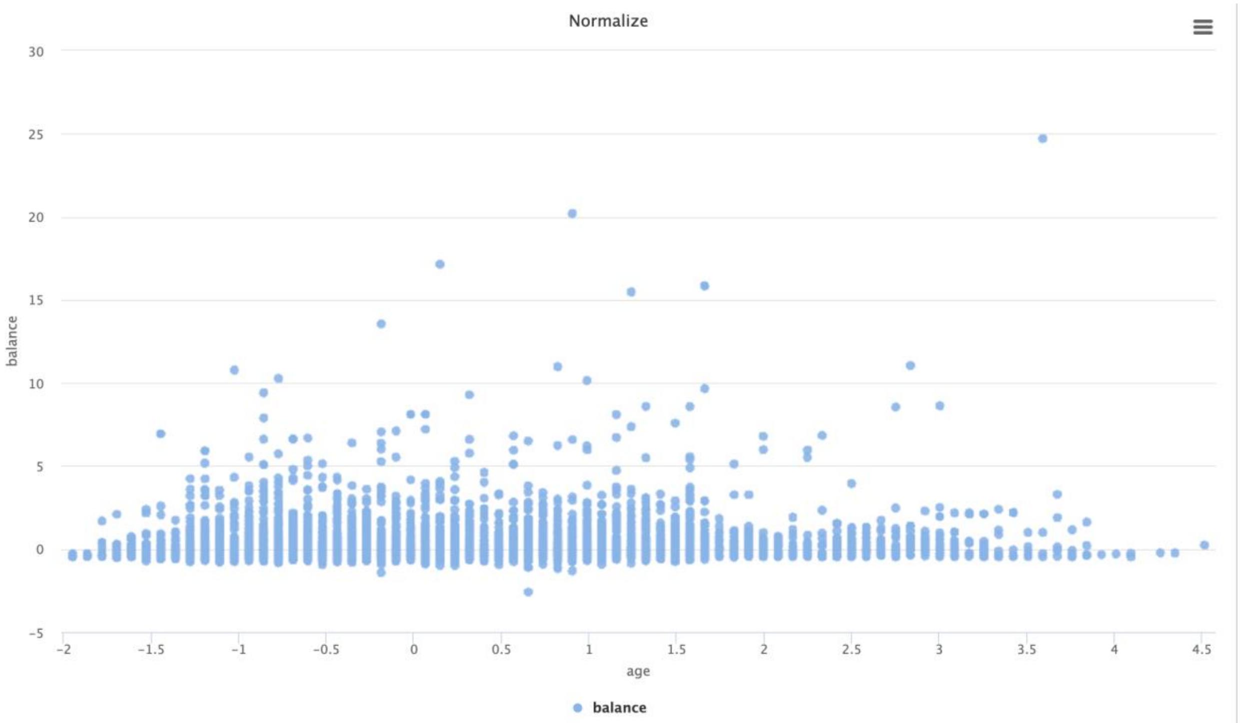
- Highly skewed, with most balances near zero.
- **Insight**: Many customers have low balances, which may impact their likelihood of subscribing.

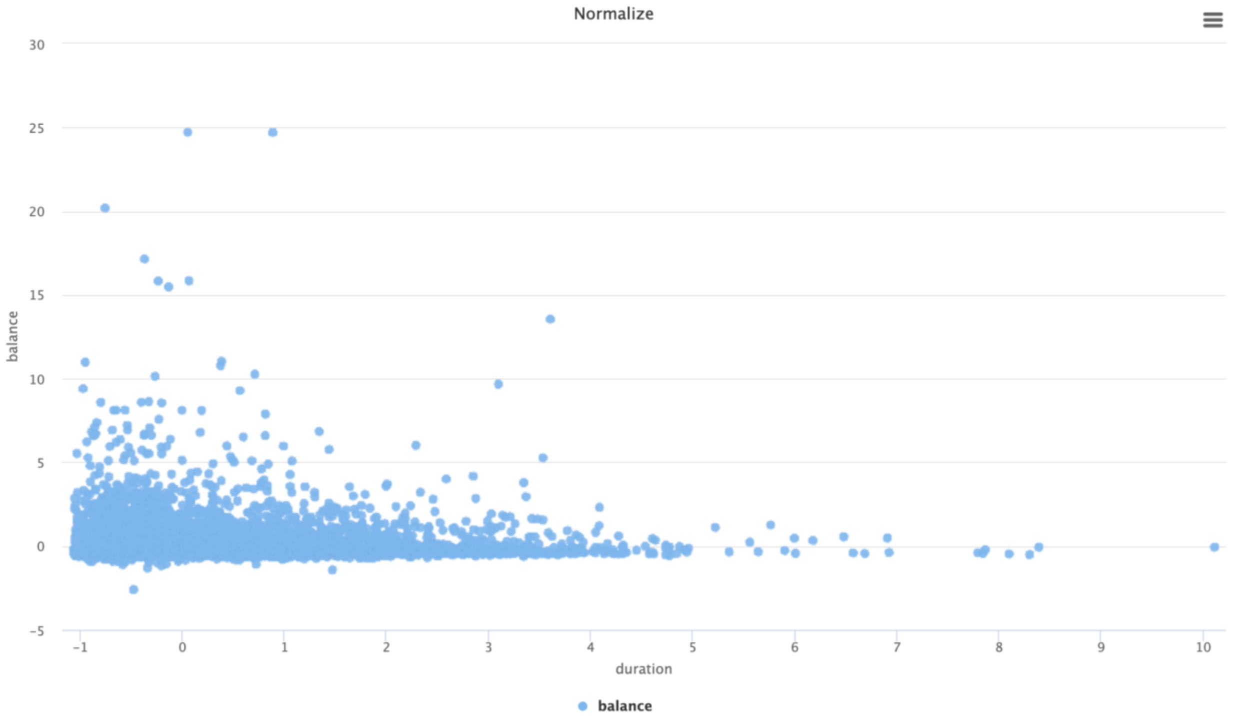
Age Histogram:

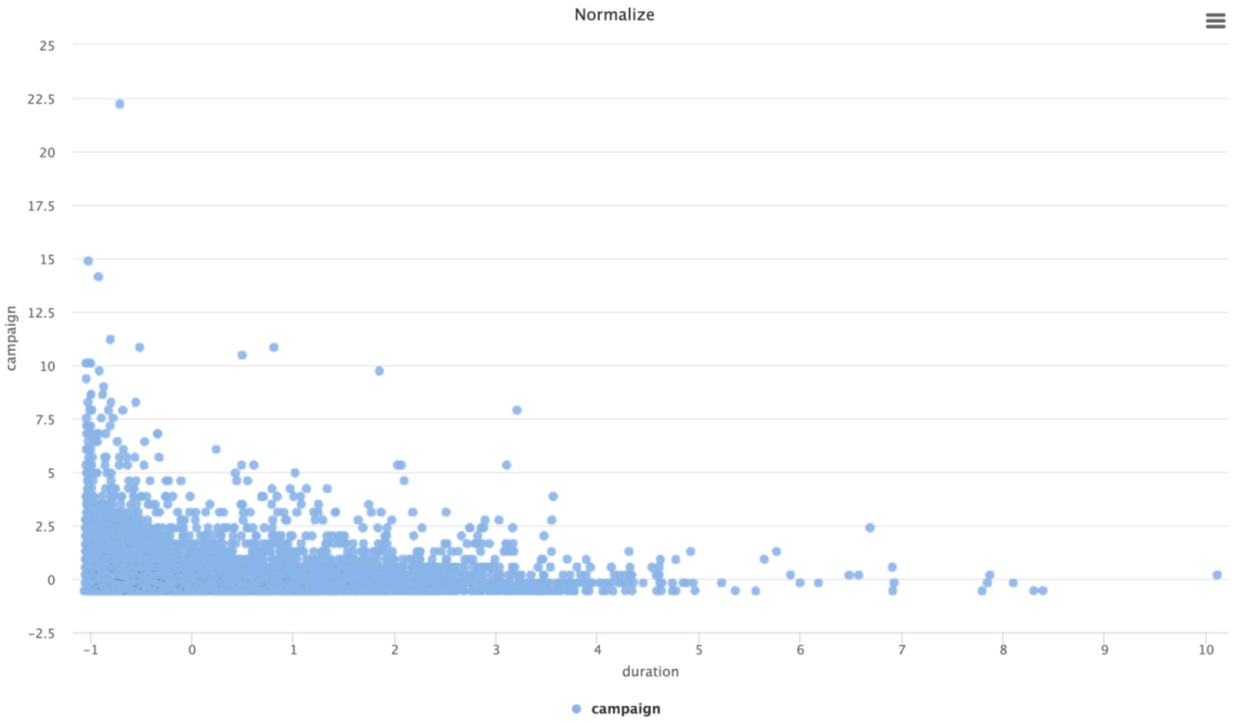
- Slightly normal distribution, but some peaks in middle-aged groups.
- **Insight**: Targeting certain age groups may increase marketing effectiveness.

Campaign Histogram:

- Most customers were contacted only once or twice.
- **Insight**: Higher contact frequency does not always mean better conversion.


Normalize
30
25
20
15
10
5
0
-5
balance
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
age
● balance


Normalize
30
25
20
15
10
5
0
-5
balance
-1
0
1
2
3
4
5
6
7
8
9
10
duration
● balance



Scatter Plot: Age vs. Balance:

- The distribution shows no strong correlation between age and balance.
- However, some customers have high balances at different ages, suggesting that age alone does not determine a customer's financial status.

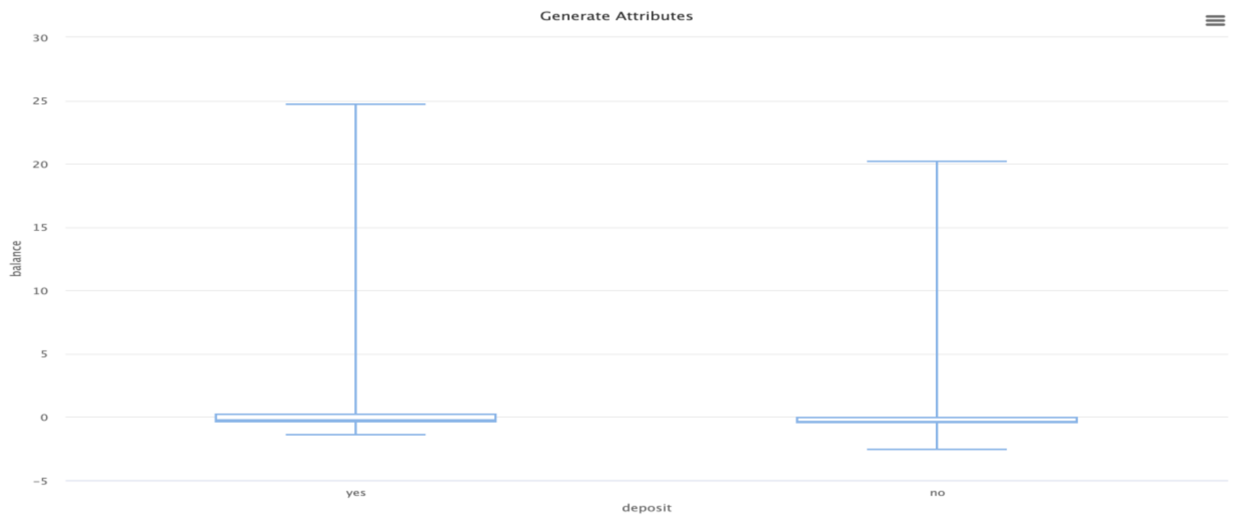
Scatter Plot: Duration vs. Balance:

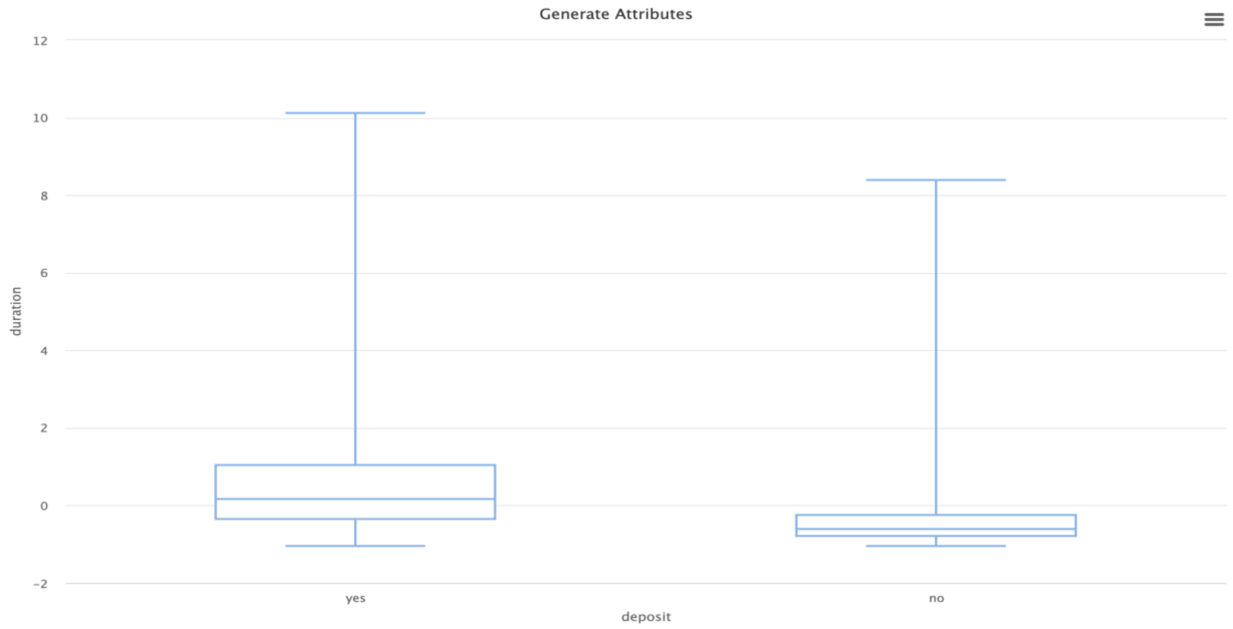
- No clear trend, but some customers with higher balances tend to have longer call durations.
- Insight: Longer conversations may indicate higher engagement, which could lead to a successful deposit subscription.

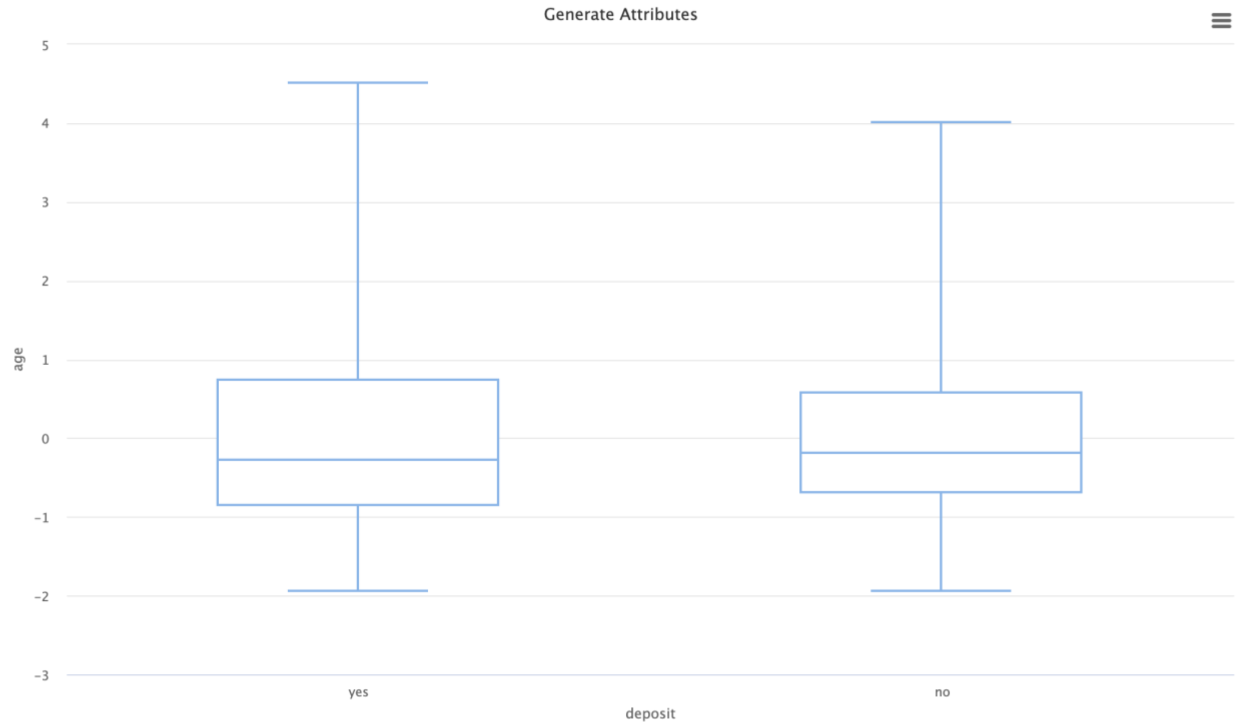
Scatter Plot: Duration vs. Campaign:

- There are many low durations calls with high campaign numbers, suggesting that frequent calls do not necessarily mean longer engagement.
- Insight: Too many short calls may annoy customers, reducing the chances of a deposit.

Box Plot Variables (For Outlier Detection):







Balance vs. Deposit:

- The median balance for those who made a deposit (yes) is slightly higher than those who didn't.
- Both categories have many outliers, indicating that some customers have extremely high balances.
- Most customers have low balance values, as seen in the small interquartile range (IQR).

Insight: Customers with higher balances may be slightly more likely to subscribe, but extreme balances should be examined separately.

Duration vs. Deposit:

- Customers who subscribed ("yes") had longer average call durations compared to those who did not.
- Interquartile range (IQR) is wider for "yes", meaning higher variation in call durations.
- The maximum duration for "yes" deposits is much larger.

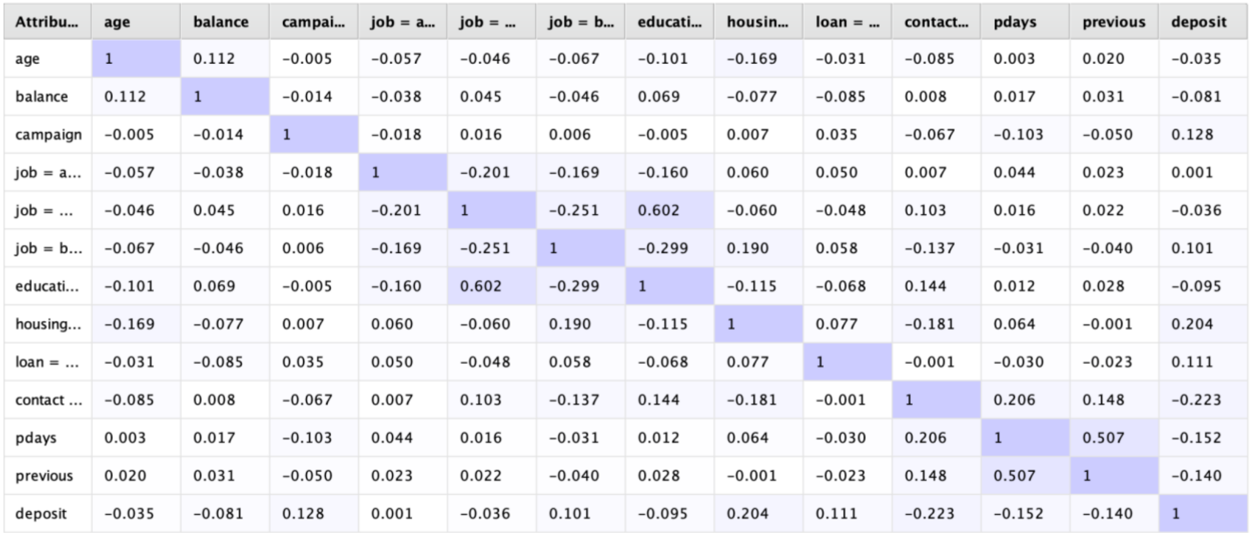
Insight: Longer calls correlate with a higher likelihood of making a deposit, suggesting that effective customer engagement is key.

Age vs. Deposit:

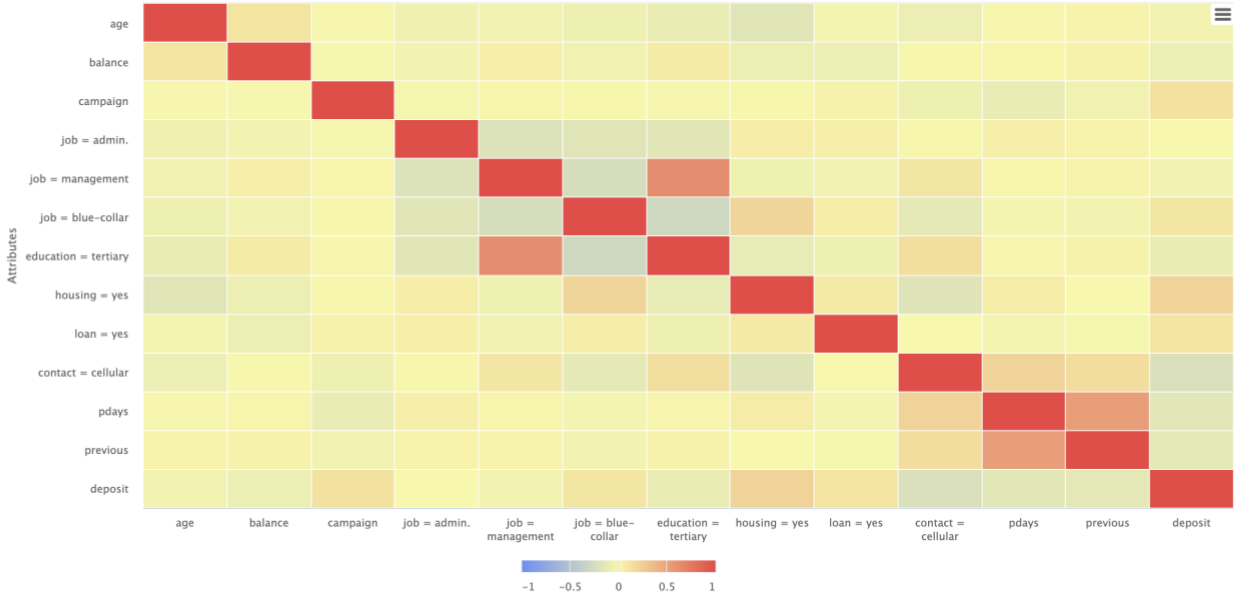
- The age distributions for both "yes" and "no" categories are very similar.
- Slightly higher median age for customers who subscribed.
- Outliers exist, but they do not seem to significantly affect the trend.

Insight: Age alone is not a strong predictor for deposit subscriptions, but it might be useful in combination with other factors.

Correlation:



Attribu...	age	balance	campai...	job = a...	job = ...	job = b...	educati...	housin...	loan = ...	contact...	pdays	previous	deposit
age	1	0.112	-0.005	-0.057	-0.046	-0.067	-0.101	-0.169	-0.031	-0.085	0.003	0.020	-0.035
balance	0.112	1	-0.014	-0.038	0.045	-0.046	0.069	-0.077	-0.085	0.008	0.017	0.031	-0.081
campaign	-0.005	-0.014	1	-0.018	0.016	0.006	-0.005	0.007	0.035	-0.067	-0.103	-0.050	0.128
job = a...	-0.057	-0.038	-0.018	1	-0.201	-0.169	-0.160	0.060	0.050	0.007	0.044	0.023	0.001
job = ...	-0.046	0.045	0.016	-0.201	1	-0.251	0.602	-0.060	-0.048	0.103	0.016	0.022	-0.036
job = b...	-0.067	-0.046	0.006	-0.169	-0.251	1	-0.299	0.190	0.058	-0.137	-0.031	-0.040	0.101
educati...	-0.101	0.069	-0.005	-0.160	0.602	-0.299	1	-0.115	-0.068	0.144	0.012	0.028	-0.095
housing...	-0.169	-0.077	0.007	0.060	-0.060	0.190	-0.115	1	0.077	-0.181	0.064	-0.001	0.204
loan = ...	-0.031	-0.085	0.035	0.050	-0.048	0.058	-0.068	0.077	1	-0.001	-0.030	-0.023	0.111
contact ...	-0.085	0.008	-0.067	0.007	0.103	-0.137	0.144	-0.181	-0.001	1	0.206	0.148	-0.223
pdays	0.003	0.017	-0.103	0.044	0.016	-0.031	0.012	0.064	-0.030	0.206	1	0.507	-0.152
previous	0.020	0.031	-0.050	0.023	0.022	-0.040	0.028	-0.001	-0.023	0.148	0.507	1	-0.140
deposit	-0.035	-0.081	0.128	0.001	-0.036	0.101	-0.095	0.204	0.111	-0.223	-0.152	-0.140	1



Deposit Correlation:

- "Deposit" has a weak correlation with most variables, meaning no single feature strongly influences deposit decisions.
- "Campaign" shows the highest positive correlation (0.128), suggesting a slight impact of the number of contacts on deposit.
- "Previous" and "Pdays" have small negative correlations with "Deposit," indicating that past campaign effectiveness may have some minor impact.

Feature Relationships:

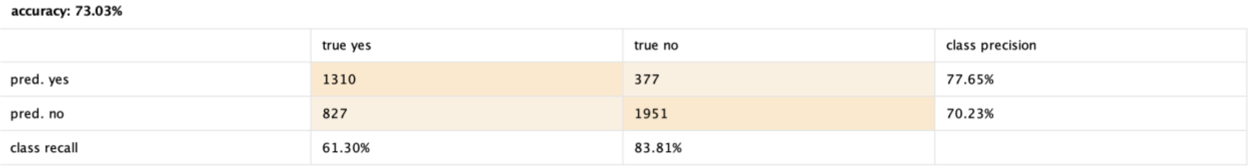
- Strong positive correlations exist within categorical variables like "Job = Management" and "Education = Tertiary" (0.602).
- "Pdays" and "Previous" are moderately correlated (0.507), meaning clients with prior contact history tend to be contacted again.

Heatmap Validation:

- The heatmap visually confirms these relationships.
- Darker red areas indicate positive correlations, while blue shades indicate negative relationships.

Model Evaluation:

1.Logistic Regression:



accuracy: 73.03%

	true yes	true no	class precision
pred. yes	1310	377	77.65%
pred. no	827	1951	70.23%
class recall	61.30%	83.81%	

Model Performance Summary for Our Prediction:

Logistic Regression model for this dataset, and below are the key performance metrics from prediction results:

Accuracy: 73.03%:

- Our model correctly predicted **73.03%** of cases, indicating a good classification performance.

2. Confusion Matrix Breakdown:

Prediction	Actual "Yes"	Actual "No"
Predicted "Yes"	**1,310** (True Positives)	**377** (False Positives)
Predicted "No"	**827** (False Negatives)	**1,951** (True Negatives)

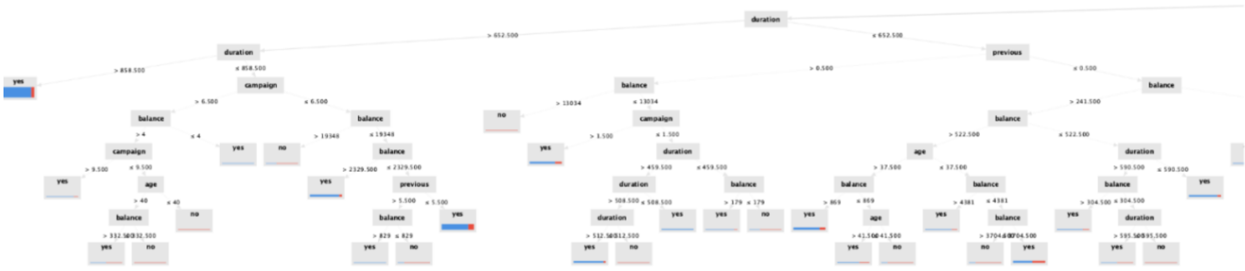
- **Recall (Sensitivity): 61.30%**
  - Out of all actual "Yes" cases, **61.30%** were correctly predicted as "Yes".
- **Precision: 77.65%**
  - When the model predicted "Yes," **77.65%** of the time it was correct.
- **F1-score: 68.51%**

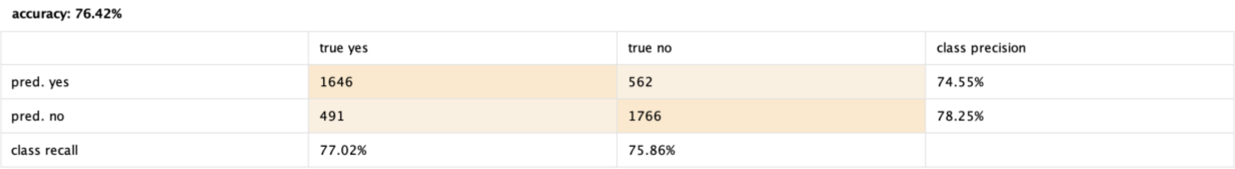
- A harmonic mean between **Precision** and **Recall**, showing an overall balanced performance.

- **AUC Score: 0.819**
  - This means our model has a good ability to distinguish between the two classes (Yes/No).

This summary provides a clear understanding of our **prediction results** and the effectiveness of our **Logistic Regression model**.

2.Decision Tree:





accuracy: 76.42%

	true yes	true no	class precision
pred. yes	1646	562	74.55%
pred. no	491	1766	78.25%
class recall	77.02%	75.86%	

Decision Tree Performance:

- **Accuracy**: **76.42%** (Higher than Logistic Regression)
- **Confusion Matrix**:
  - **True Positives (Actual Yes, Predicted Yes)**: **1,646**

  - **True Negatives (Actual No, Predicted No)**: **1,766**
  - **False Positives (Actual No, Predicted Yes)**: **562**
  - **False Negatives (Actual Yes, Predicted No)**: **491**
- **Recall (Sensitivity)**: **77.02%** (Higher than Logistic Regression)
  - The model correctly identified **77.02%** of actual "yes" cases.
- **Precision**: **74.55%**
  - Out of all cases predicted as "yes," **74.55%** were correct.
- **F1-score**: **75.77%** (Better than Logistic Regression)
  - Shows better balance between precision and recall.
- **AUC Score**: **0.822**
  - Slightly better than Logistic Regression, indicating improved classification ability.

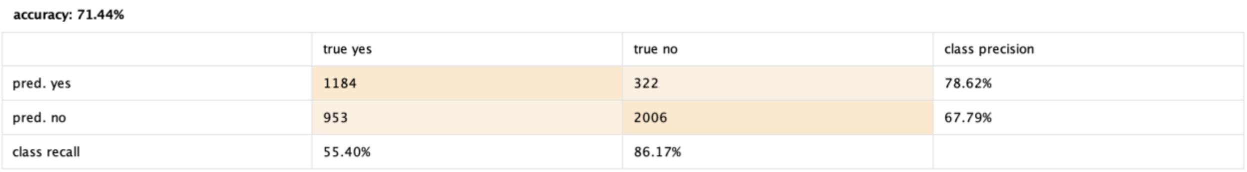
The Decision Tree model outperformed the Logistic Regression model in accuracy (76.42% vs. 73.03%) and recall (77.02% vs. 61.30%).

Logistic Regression showed higher precision (77.65% vs. 74.55%), meaning fewer false positives.

The AUC score for Decision Tree was slightly better (0.822 vs. 0.819), meaning it differentiates better between classes.

Overall, the Decision Tree model provides better recall and accuracy, making it a stronger choice for this dataset.

Naive Bayes:



accuracy: 71.44%

	true yes	true no	class precision
pred. yes	1184	322	78.62%
pred. no	953	2006	67.79%
class recall	55.40%	86.17%	

Accuracy: 71.44%

- **Confusion Matrix:**
  - True Yes (TP): 1,184
  - True No (TN): 2,006
  - False Positive (FP): 322
  - False Negative (FN): 953
- **Recall: 55.40%**
  - Lower than both Logistic Regression and Decision Tree.
- **Precision: 78.62%**
- **F1-score: 65.00%**
- **AUC Score: 0.796**
  - Performs worse in distinguishing between classes.

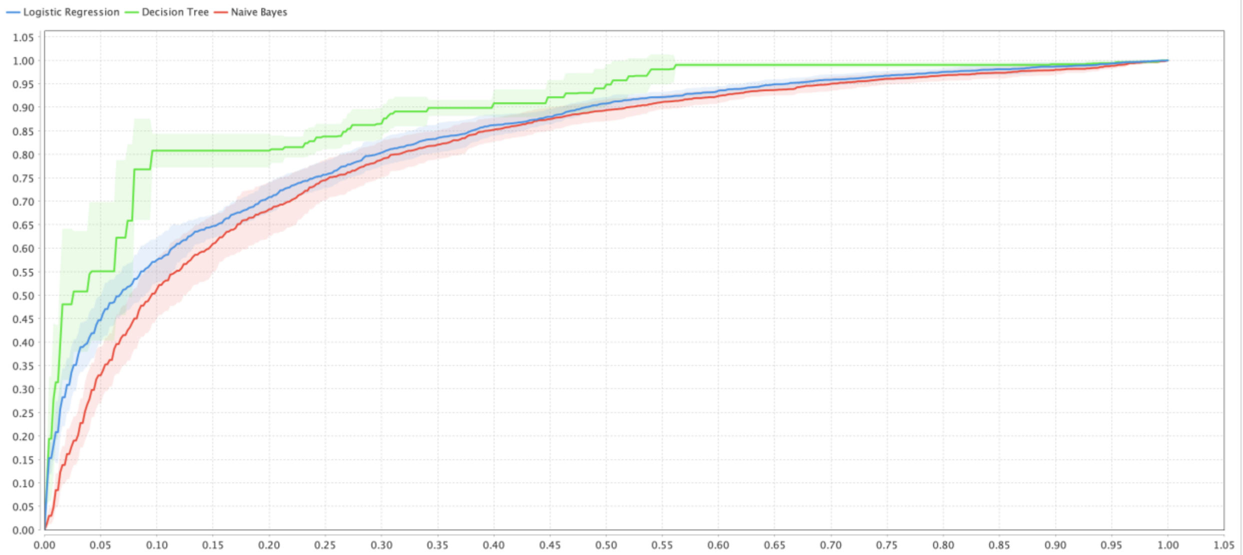
Compare ROC:

Analysis of ROC Curve:

Comparison of Model Performances:

Metric	Logistic Regression	Decision Tree	Naïve Bayes
Accuracy	73.03%	76.42%	71.44%
Precision (Yes)	77.65%	74.55%	78.62%

Recall (Yes)	61.30%	77.02%	55.40%
F1-Score (Yes)	68.51%	75.77%	65.00%
AUC Score	0.819	0.822	0.796



Logistic Regression (Blue Curve)

- The ROC curve for **Logistic Regression** follows a steady upward trend.
- It indicates a balanced trade-off between **true positives and false positives**.
- **AUC value is decent (~0.80)**, meaning the model is good at distinguishing between classes.

1. Decision Tree (Green Curve)

- The **Decision Tree model has the highest AUC** and the steepest rise initially.
- It shows that **Decision Tree makes strong confident predictions**, but it may **overfit**.
- The curve has more sharp steps, which means it has **harder decision boundaries**.

2. Naïve Bayes (Red Curve)

- The **Naïve Bayes curve is the lowest** among the three.

- **AUC is lower (~0.75 or below)**, which suggests it is not as good at distinguishing between positive and negative cases.
- Naïve Bayes assumes **independent features**, which might not hold for this dataset.

Conclusion:

The goal of this project was to analyze customer behavior and predict whether a client would subscribe to a term deposit based on the Bank Marketing Dataset. Using different machine learning models—**Logistic Regression, Decision Tree, and Naïve Bayes**—evaluated the predictive performance of each and identified the most effective approach for customer classification.

From analysis, observed that **Decision Tree** outperformed both Logistic Regression and Naïve Bayes in terms of accuracy (**76.42%**), recall (**77.02%**), and overall class separation (AUC **0.822**). This suggests that Decision Tree is the most suitable model for understanding and predicting customer behavior in this dataset.

Key Insights & Takeaways:

1. **Customer Behavior Patterns:**
    - Features such as **campaign contact frequency, duration, balance, and previous interactions** had a significant impact on whether a client would subscribe to a term deposit.
    - Customers who had **longer call durations** and **previous successful interactions** were more likely to subscribe.
2. **Model Performance & Business Decision Impact:**

- **Decision Tree performed best** and can be useful for bank marketing teams to **identify potential customers with a higher likelihood of subscribing**.
- **Logistic Regression had stable performance (73.03% accuracy, AUC 0.819)** and can still be used for **interpretable insights** into key influencing factors.
- **Naïve Bayes had the lowest performance (71.44% accuracy, AUC 0.796)**, indicating that it may not be the best choice for this dataset due to its assumptions of feature independence.

3. **Targeted Marketing Strategy:**
   - By leveraging the Decision Tree model, banks can focus their marketing efforts on **customers with a high probability of subscription**, optimizing call strategies and reducing unnecessary costs.
   - The **classification model can help prioritize leads**, enabling banks to personalize their approach and improve customer conversion rates.

Final Thoughts:

Finally, Analysis successfully demonstrated that customer properties play a crucial role in predicting term deposit subscriptions. The **Decision Tree model stands out as the most effective approach**, providing actionable insights for bank marketing teams to **increase customer engagement and improve marketing efficiency**.