Final Project Report on Exploratory Data Analysis of Bank Marketing Campaign

Group Name: Data Adventurers
Name: Shatabdi Pal
Email: shatabdi@pdx.edu
Country: USA
College: Portland State University
Specialization: Data Science

Abstract

Project Title: Predicting Customer Subscription to Term Deposits

Problem description:
ABC Bank aims to optimize its marketing efforts to promote a new term deposit product by developing a predictive model. The model will analyze customer data from past interactions to forecast the likelihood of each customer subscribing to the term deposit. By accurately identifying potential buyers, the bank can focus its marketing resources on high-probability customers, enhancing efficiency and reducing costs. This project involves creating and evaluating models with and without the 'duration' feature, a critical element that will ensure practical applicability and performance.

Business understanding:
ABC Bank strategically leverages machine learning to streamline its marketing campaigns for a term deposit product. The goal of increasing conversion rates and minimizing resource expenditure is to develop a predictive model that identifies customers with a higher subscription probability. This will allow us to implement targeted marketing strategies through telemarketing and email. The bank's focus on high-potential customers will save time and resources, ultimately reducing marketing costs. Understanding and converting machine learning metrics into business terms is crucial for aligning the technical outcomes with the bank's strategic objectives.

Tools & Technologies used:

Programming Language: Python

IDE: Jupyter Notebook, Google Colaboratory

Visualization: Python (Matplotlib and Seaborn)

Models: Logistic Regression, Ensemble, Ada Boost

Libraries: NumPy, Pandas, Seaborn, Matplotlib, Scikit-learn

Dataset: https://archive.ics.uci.edu/dataset/222/bank+marketing

1. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) involves visually and statistically examining a dataset to extract meaningful insights and understand its underlying patterns and characteristics. EDA provides a foundation for further analysis and decision-making processes by summarizing key features and relationships within the data.

A. Data Understanding and Preprocessing

1. Dataset Description:

This dataset is related to the marketing campaigns of a Portuguese banking institution. The campaigns were based on direct phone calls, and the goal was to predict whether a client would subscribe to a term deposit.

Input Variables

- age: Age of the client (numeric)
- job: Type of job (categorical)
  Possible values: "admin.", "unknown," "unemployed," "management," "housemaid," "entrepreneur," "student," "blue collar," "self-employed", "retired," "technician," "services"
- marital: Marital status (categorical)
  Possible values: "married," "divorced," "single."
- Note: "Divorced" includes both divorced and widowed clients
- education: Level of education (categorical)
  Possible values: "unknown," "secondary," "primary," "tertiary."
- default: Is credit in default? (binary)
  Possible values: "yes", "no"
- balance: Average yearly balance in euros (numeric)
- housing: Has a housing loan? (binary)
  Possible values: "yes", "no"
- loan: Has a personal loan? (binary)
  Possible values: "yes", "no"
- contact: Contact communication type (categorical)
  Possible values: "unknown," "telephone," "cellular."
- day: Last contact day of the month (numeric)
- month: Last contact month of year (categorical)
  Possible values: "Jan," "Feb," "mar,"..., "nov," "dec"
- duration: Last contact duration in seconds (numeric)
- campaign: Number of contacts performed during this campaign for this client (numeric)
- pdays: Number of days that passed since the client was last contacted from a previous campaign (numeric)
  Note: 1 means the client was not previously contacted
- previous: Number of contacts performed before this campaign for this client (numeric)
- poutcome: Outcome of the previous marketing campaign (categorical)
  Possible values: "unknown," "other," "failure," "success."

Output Variable

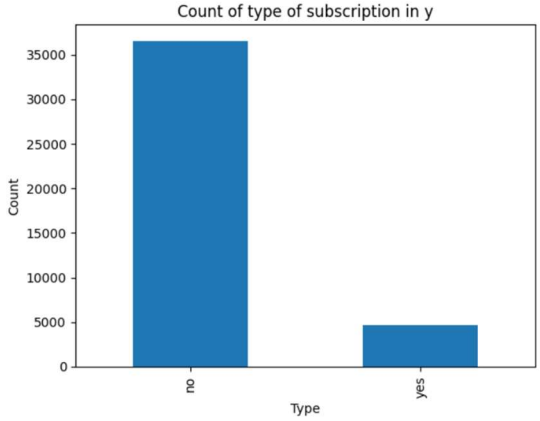
- y: Has the client subscribed to a term deposit? (binary)
  Possible values: "yes", "no"

2. Problem in Dataset:

- **Missing Value:** The categorical variables job, marital, education, default, housing, and loan have 'unknown.' values, which can be considered missing.

```
Percentage of 'unknown' values in job: 0.80%
Percentage of 'unknown' values in marital: 0.19%
Percentage of 'unknown' values in education: 4.20%
Percentage of 'unknown' values in default: 20.88%
Percentage of 'unknown' values in housing: 2.40%
Percentage of 'unknown' values in loan: 2.40%
```

- **Outliers:** Numeric features like 'age,' 'duration,' 'campaign,' 'days,' and 'previous' have outliers that could affect the model's performance.
- **Skewed Data:** Features like 'duration,' 'campaign,' and 'previous' have skewness greater than 1.
- **Imbalanced Data:** The target variable 'y' is imbalanced, meaning there are significantly more 'no' than 'yes' responses.



3. Handling of Unknown Values of Dataset:

Categorical variables with 'unknown' values were imputed with the most frequent category. The random forest method was also applied to imputing those unknown values. Then, the KS (Kolmogorov-Smirnov) KS statistic and P value were used to compare the two imputation techniques. I found that the KS statistic was 0.0, and the p-value was 1.0 for each column with unknown values. This result suggests that both imputation methods (most frequent and Random Forest Classifier) produce imputed statistically indistinguishable values in their distribution. Therefore, these specific columns and the chosen imputation methods are equally effective in approximating the original data distribution. During the data cleaning process, categorical variables with 'unknown' values were carefully imputed with the most frequent category, ensuring the integrity of the dataset.

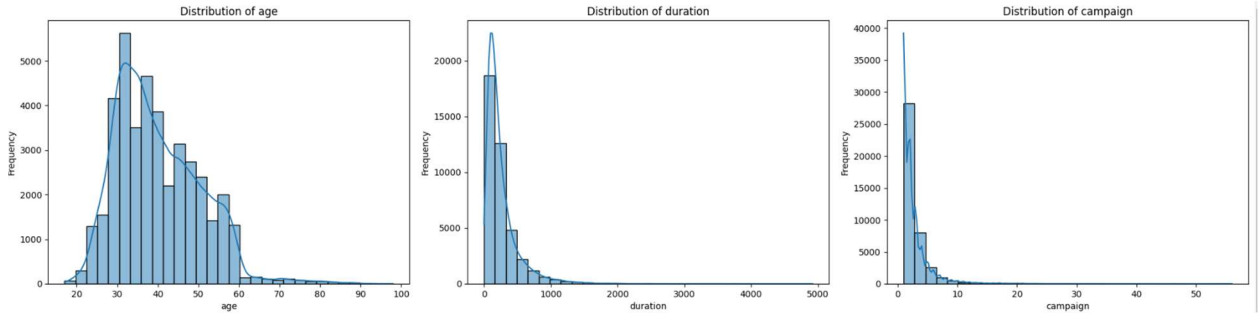
Summary Statistics:

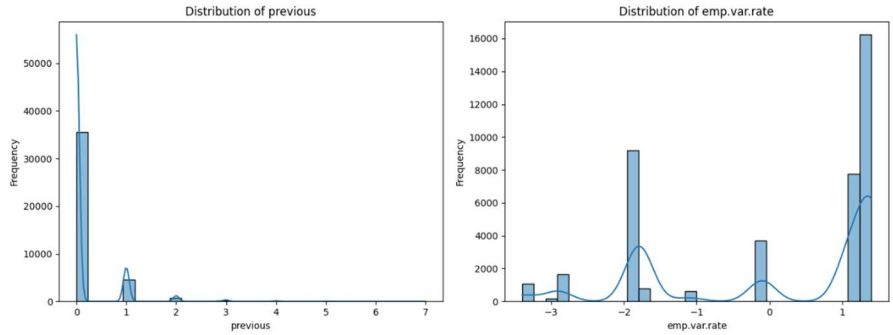
	age	duration	campaign	pdays	previous
count	41176.00000	41176.000000	41176.000000	41176.000000	41176.000000
mean	40.02380	258.315815	2.567879	962.464810	0.173013
std	10.42068	259.305321	2.770318	186.937102	0.494964
min	17.00000	0.000000	1.000000	0.000000	0.000000
25%	32.00000	102.000000	1.000000	999.000000	0.000000
50%	38.00000	180.000000	2.000000	999.000000	0.000000
75%	47.00000	319.000000	3.000000	999.000000	0.000000
max	98.00000	4918.000000	56.000000	999.000000	7.000000

	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000
mean	0.081922	93.575720	-40.502863	3.621293	5167.034870
std	1.570883	0.578839	4.627860	1.734437	72.251364
min	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	1.400000	94.767000	-26.900000	5.045000	5228.100000

B. Visualization: (Relationship between variables)

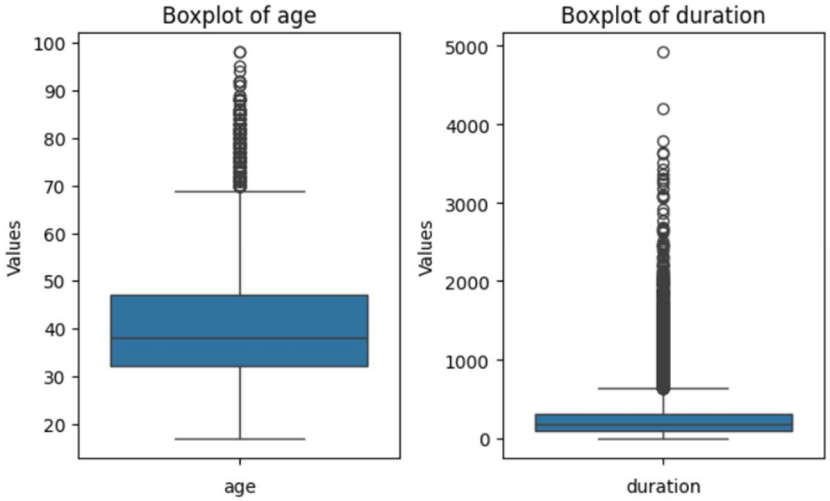
Distribution of Variables (Numerical Features):

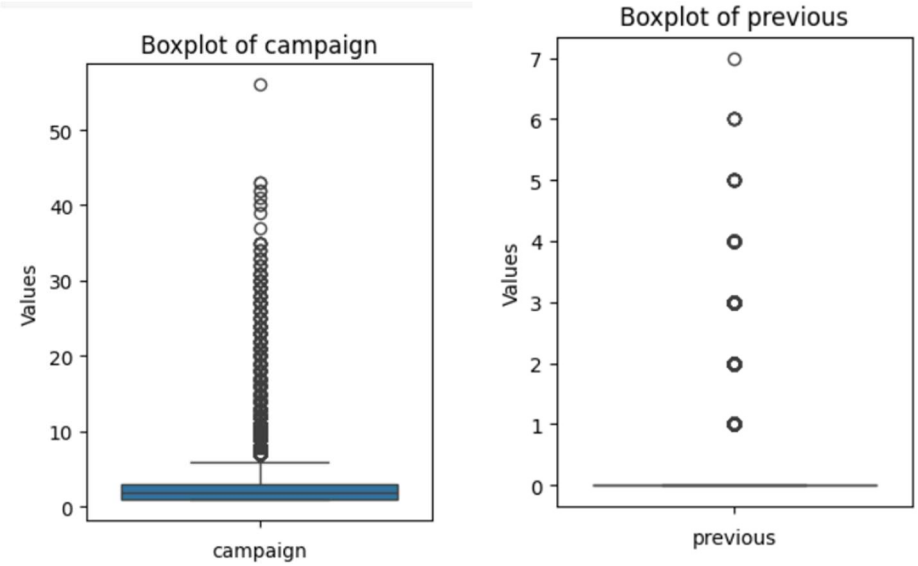




The duration, campaign, and previous features had more significant skewness greater than 1. This indicates that the feature's distribution is significantly skewed to the right (positively skewed). A skew value greater than 1 means that the tail on the right side of the distribution is longer or fatter than the left. By addressing the skewness, we can ensure that the features contribute effectively to the predictive models and that the insights derived from the data are reliable and meaningful.

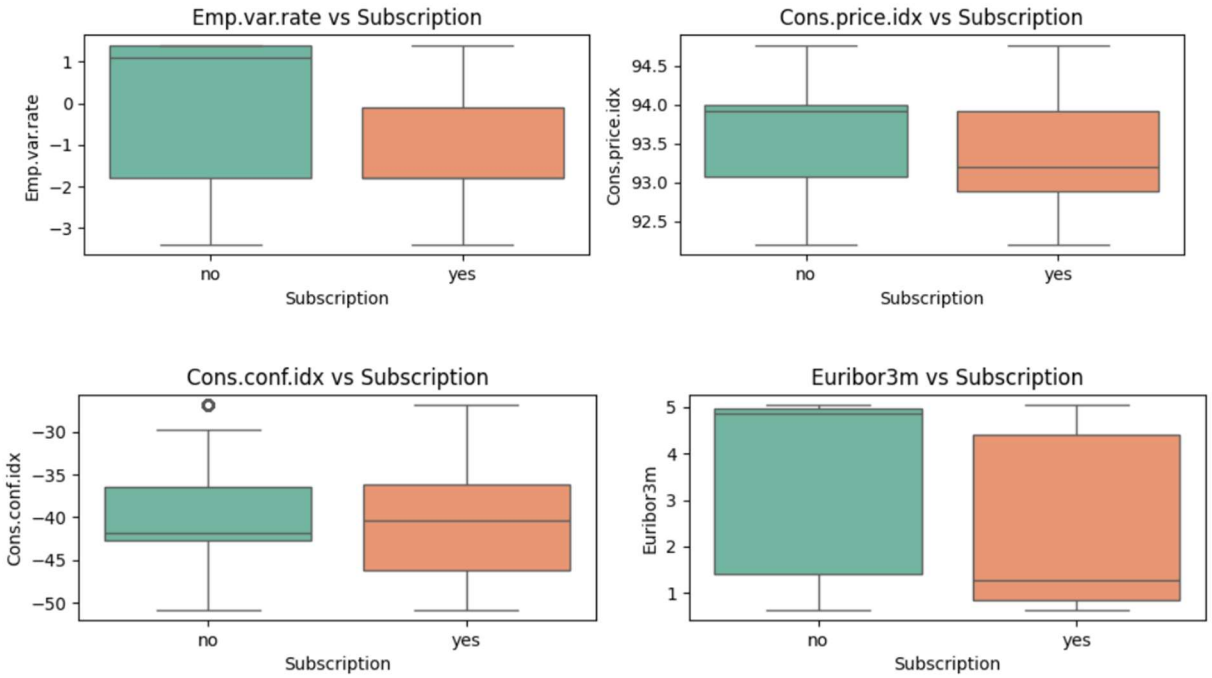
Presence of Outliers:

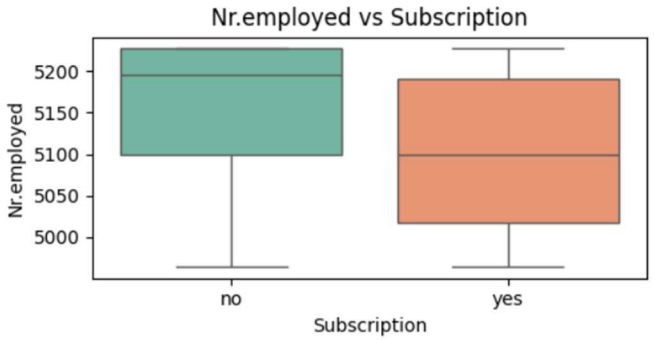




Certain numeric features, such as 'age,' 'duration,' 'campaign,' 'days,' and 'previous,' exhibit outliers that could impact the model's performance. We employed a comprehensive approach to detecting outliers to ensure our data's reliability. Box plots and statistical methods, such as the Z-score and IQR (Interquartile Range), were used to identify outliers in numerical features.

Box plot of Features:

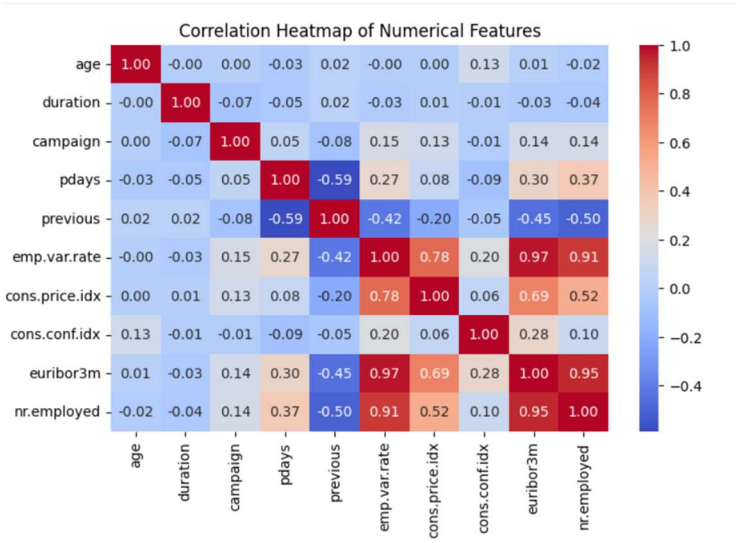




Overall Insights:

The employment variation rate (`Emp.var.rate`) and consumer confidence index (`Cons.conf.idx`) show higher medians for non-subscribers and broader ranges, indicating they may influence subscription decisions. The consumer price index (`Cons.price.idx`) has minimal impact on subscription decisions, while the Euribor 3-month rate (`Euribor3m`) shows a slight difference in medians, suggesting limited but potential relevance. The number of employed individuals (`Nr.employed`) shows similar ranges and medians for both groups, suggesting it may not significantly impact subscription decisions.

Correlation between variables:



Strong Positive Correlations:

The strong positive correlations between emp. var. rate, euribor3m, and nr. employed suggest that these economic indicators are closely related.

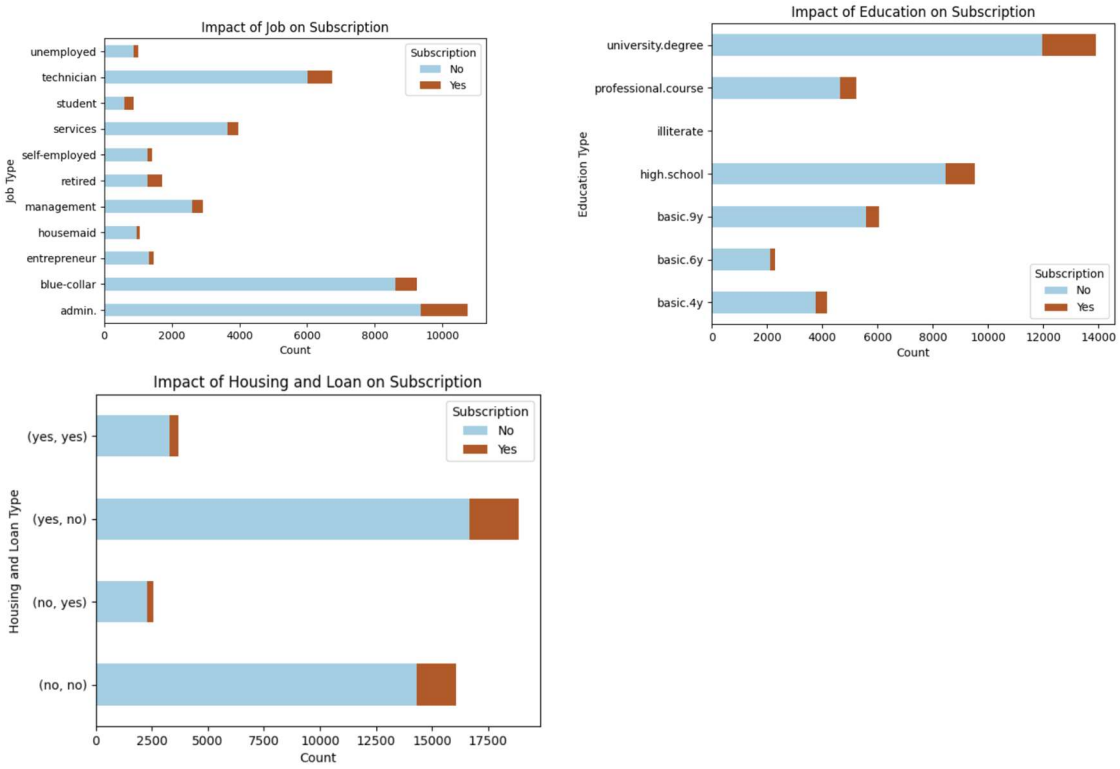
Negative Correlations

The negative correlation between emp.var.rate and cons. conf. idx indicates that higher employment variation rates might be associated with lower consumer confidence.

Impact on Target

The relationship between pdays and previous can give insights into how previous contacts impact the time between contacts, which might help understand customer behavior and optimize campaign strategies.

Bar Plot of Features:

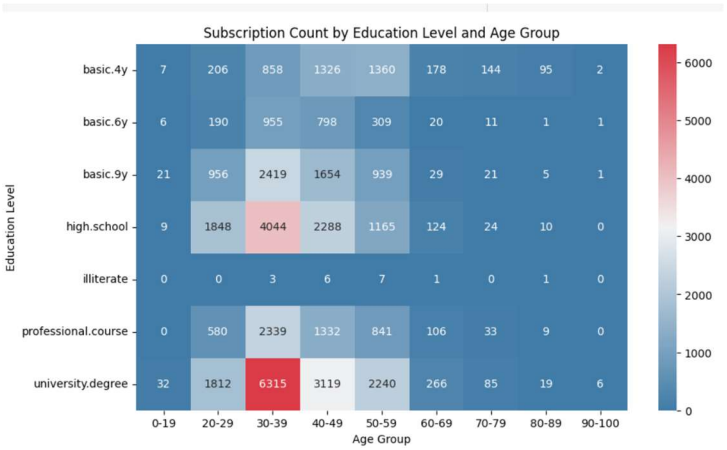


Overall insight from the above bar plots is as follows:

- For most job types shown, the number of people not subscribing is higher than those who subscribe.
- Students and retired individuals have a relatively higher subscription rate than other job types.
- For most other education types, the number of people who do not subscribe is higher than those who do.

- University degree holders dominate in subscribing and not subscribing, underscoring the need for personalized content to cater to their diverse needs.
- Having a housing or personal loan may impact subscription behavior.
- Individuals without loans (housing and personal) are more likely to subscribe (as indicated by higher counts).
- Those with both loans (housing and personal) are less likely to subscribe.
- The highest count for 'No' subscriptions is in the no housing and no loan category, followed by housing and no loan.

Multivariable Relationship:



The following conclusion can be drawn from the above plot:

- The highest concentration of subscriptions (indicated by darker red shades) occurs in the age group 30-39.
- University degree holders have the highest subscription count, followed by high school graduates and those with professional courses.

Business Recommendation Based on EDA

- **Optimize Contact Strategy:** Limit the number of contact attempts to prevent customer fatigue and increase subscription rates.
- **Target-Specific Job Types:** Focus on students and retired individuals due to their relatively higher subscription rates.
- **Personalize Content for Education Levels:** Tailor marketing messages for university degree holders to cater to their diverse needs.
- **Address Loan Impacts:** To improve the likelihood of subscribing, offer budget-friendly plans to those with housing and personal loans.

- **Leverage Economic Indicators:** Time campaigns during favorable economic conditions indicated by Emp.var.rate and Cons.conf.idx.
- **Employment Status:** Consider other influencing factors, as the number of employed individuals shows minimal impact on subscriptions.
- **Demographic Focus:** Focus marketing efforts on the 30-39 age group with the highest subscription concentration.
- **Analyze Non-Subscribers:** Investigate reasons behind high non-subscription rates in individuals without loans and address their concerns.

Implementing targeted, personalized strategies based on job type, education level, loan status, and economic indicators, optimizing contact frequency, and focusing on critical demographics can significantly enhance subscription rates.

2. Machine Learning Prediction

2.1. Feature Engineering: Transformations, Scaling, or encoding methods.

Transformation:

We implemented the Z-score and IQR methods to deal with the outliers. The comparison of both methods using Cohen's Kappa coefficient, which resulted in 1.0, indicates that both methods are equally effective and reliable in identifying outliers for the specific set of data points or indices being compared. The impact of outliers on the model was considered, and the IQR method was used to handle those. Features like 'duration,' 'campaign,' and 'previous' have a skewness greater than 1. A square root transformation is applied to normalize skewed features.

Encoding:

Age Features: Binning of age features are also done in this part

Scaling:

First, the summary statistics for the numerical features are calculated to understand their distributions. Then, we visualized these distributions using histograms to assess their spread and skewness. After identifying variations in scale among the features, we performed feature scaling using MinMaxScaler to normalize the data to a range of [0, 1]. This normalization process involved subtracting the minimum value and dividing by the range for each numerical variable, ensuring consistency in scale for machine learning algorithms.

Creating New Features and Dropping Redundant Ones:

Besides the transformation processes, some redundant columns were removed after creating new features such as Total Contacts, Economic Indicators Ratio, and Interaction between features.

Feature Selection:

Excluding the 'duration' feature from the final model is not practical for real-time prediction, and its inclusion would lead to unrealistic performance expectations. Focus on features related to client demographics, past interactions, and economic indicators.

The screenshot below shows the final features and target for model training:

```
Data columns (total 22 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   age                             41176 non-null  float64
 1   job                             41176 non-null  object
 2   marital                         41176 non-null  object
 3   education                       41176 non-null  object
 4   default                         41176 non-null  object
 5   housing                         41176 non-null  object
 6   loan                            41176 non-null  object
 7   contact                         41176 non-null  object
 8   month                           41176 non-null  object
 9   day_of_week                     41176 non-null  object
 10  pdays                           41176 non-null  float64
 11  poutcome                        41176 non-null  object
 12  y                               41176 non-null  object
 13  age_group                       41176 non-null  category
 14  duration_sqrt                   41176 non-null  float64
 15  campaign_sqrt                   41176 non-null  float64
 16  previous_sqrt                   41176 non-null  float64
 17  total_contacts                  41176 non-null  int64
 18  emp_var_rate_euribor3m_ratio    41176 non-null  float64
 19  cons_price_conf_idx_ratio       41176 non-null  float64
 20  nr_employed_emp_var_rate        41176 non-null  float64
 21  nr_employed_cons_price_idx      41176 non-null  float64
dtypes: category(1), float64(9), int64(1), object(11)
```

2.2 Model Choice:

After evaluating various models, logistic regression, ensemble model using random forest classifier, gradient boosting classifier, voting classifier, and Adaboost using decision tree classifier can be used for implementation.

Here are the reasons for model choice:

- **Logistic Regression:** Simple and interpretable, effective for linearly separable data, providing probabilistic outputs.
- **Random Forest Classifier:** Robust to overfitting, handles non-linear relationships, and ranks feature importance well.
- **Gradient Boosting Classifier:** This classifier builds models sequentially to correct errors. It is effective for complex data and often yields high accuracy.
- **Voting Classifier**: Combines multiple models to improve generalizability and robustness, leveraging the strengths of each.
- **Adaboost using a Decision Tree Classifier:** This classifier focuses on misclassified instances, enhances weak learners, and improves prediction accuracy.

Addressing Class Imbalance with SMOTE

- **Synthetic Minority Over-sampling Technique (SMOTE)**: This technique balances the dataset by generating synthetic examples for the minority class (yes response), improving model performance.
- **Improved Performance**: This ensures that models learn characteristics of both yes and no responses, leading to more accurate and reliable predictions.

2.3 Split the data to train and test:

The data is split by separating the features (X) from the target variable (y). Following this, the classes are balanced using SMOTE, with synthetic examples generated for the minority class. Subsequently, the balanced data is divided into training and testing sets using a 70-30 ratio, whereby 70% of the data is allocated for training the models, and 30% is reserved for testing their performance.

2.4 Performance Metrics

The model's performance is evaluated using a combination of various metrics, such as accuracy, precision, recall, F1 score, and ROC-AUC.

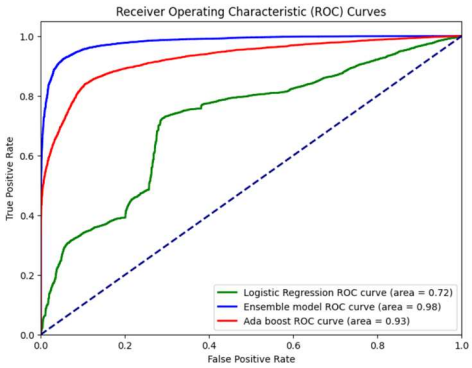
- Accuracy: The ratio of correctly predicted instances to the total cases.
- Precision: The ratio of true positive predictions to the total predicted positives.
- Recall (Sensitivity): The ratio of true positive predictions to the total actual positives.
- F1 Score: The harmonic mean of precision and recall.
- ROC (Receiver Operating Characteristic) Curve and AUC (Area Under the Curve): The ROC curve plots the true positive rate (recall) against the false positive rate at various threshold settings, and AUC represents the area under this curve.

Each metric highlights different aspects of model performance, which is crucial for understanding and optimizing marketing efforts in the bank marketing campaign project. These metrics ensure that the models are not only accurate but also effective in handling imbalanced data and minimizing both false positives and false negatives.

2.5 Interpretation of the results of these metrics for model effectiveness:

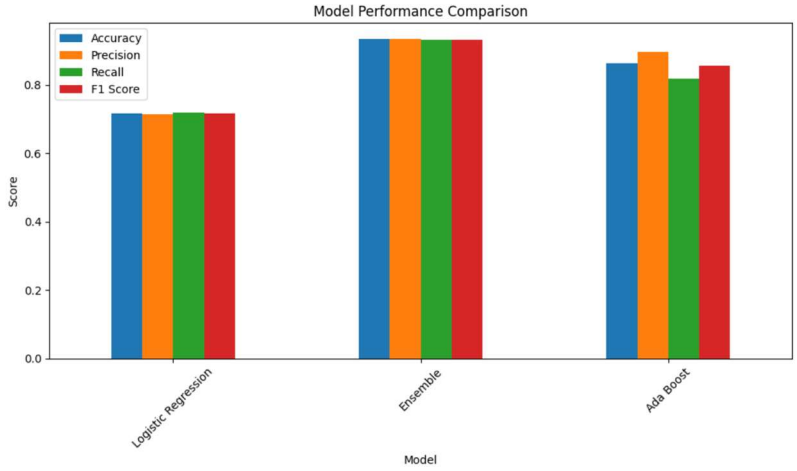
	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.716325	0.713336	0.718243	0.715781
1	Ensemble	0.933266	0.934223	0.931395	0.932807
2	Ada Boost	0.863340	0.896818	0.819499	0.856417

Logistic Regression shows moderate performance across all metrics. The accuracy of 0.716325 indicates that the model correctly predicts about 71.6% of the cases. The precision (0.713336) and recall (0.718243) are pretty balanced, resulting in an F1 score of 0.715781. This model is adequate but not optimal for the task at hand. The Ensemble model significantly outperforms Logistic Regression, with an accuracy of 0.933266. Its precision (0.934223) and recall (0.931395) are very high, leading to an F1 score of 0.932807. This indicates that the Ensemble model is excellent at correctly identifying positive and negative cases, making it the most effective model among the three. Ada Boost also performs well, with an accuracy of 0.863340. It has a high precision (0.896818), which is very good at correctly identifying true positives. However, its recall (0.819499) is slightly lower than its precision, leading to an F1 score of 0.856417. While it performs better than Logistic Regression, it does not reach the performance level of the Ensemble model.



The Ensemble model has the highest AUC (0.98), indicating that it performs best among the three models in distinguishing between positive and negative classes. This makes it a highly effective model for this classification task. The Ada Boost model performs very well, with an AUC of 0.93. It is slightly less effective than the Ensemble model but still performs strongly. Logistic Regression has the lowest AUC (0.72) among the three models, suggesting that while it can still perform the task, it is less effective than the other two models in distinguishing between classes.

2.6 Performance comparison chosen model.



The following insights can be drawn from the above bar plot of model performance metrics:

- **Consistency Across Metrics:** All models perform consistently across the four metrics, suggesting that they are balanced and that there is no significant trade-off between precision and recall.
- **Model Comparison:** The Ensemble model shows the highest performance across all metrics, making it the most effective model for this problem. Logistic

Regression and Ada Boost also perform well but slightly below the Ensemble model.

- **Balanced Metrics:** The similar values for precision, recall, and F1 score indicate that the models do not suffer from high false positives or false negatives.

The Ensemble model is the most effective for predicting customer subscription, with the highest scores in all performance metrics.

Conclusion:

Implementing targeted strategies based on job type, education level, loan status, and economic indicators while optimizing contact frequency can significantly enhance subscription rates. Focus marketing efforts on students, retired individuals, and the 30-39 age group with tailored messages for university degree holders. Offering budget-friendly plans to those with housing and personal loans will further boost subscription likelihood. The Ensemble model is recommended for deployment due to its superior performance across all metrics, with Ada Boost as a strong alternative. Though more straightforward, Logistic Regression is less effective for this specific application. Leveraging these insights will improve the bank's marketing efficiency and customer conversion rates.