Data Analysis and Modeling Report: Bank Marketing Dataset

# Introduction

This report documents the analysis and modeling of a dataset from a bank marketing campaign. The goal is to understand the factors that influence a customer's decision to subscribe to a bank term deposit and to develop a predictive model.

# Project plan (Staff, 2020) , (Smith, 2023)

## Phase 1: structure of the project and Business understanding

Timeline: The allocated timeframe for project completion.

Business understanding: Comprehend the objectives behind implementing the dataset and the techniques to be employed, identify the key influencers involved in the project.

## Phase 2: Data Collection and Understanding

Data Acquisition: Gather historical data on bank marketing campaigns, customer demographics, transaction histories, and economic indicators.

https://www.kaggle.com/datasets/henriqueyamahata/bank-marketing/data

Data Quality Assessment: Evaluate the quality, completeness, and relevance of the collected data.

Initial Data Exploration: Perform preliminary analysis to understand data structure, variables, and potential challenges (like missing values or class imbalances).

## Phase 3: Exploratory Data Analysis (EDA)

Detailed Analysis: Conduct thorough EDA to uncover trends, patterns, and correlations.

Report Findings: Document initial insights, anomalies, and potential hypotheses about customer behavior and campaign effectiveness.

## Phase 4: Data Preprocessing

Data Cleaning: Handle missing values, outliers, and errors in the data.

Feature Engineering: Create new features that could enhance model performance.

Data Transformation: Perform necessary transformations like normalization, scaling, and encoding categorical variables.

## Phase 5: Model Development

Model Selection: Choose appropriate machine learning models for classification (e.g., Logistic Regression).

Model Training and Validation: Train models on the processed dataset and validate using cross-validation techniques.

Hyperparameter Tuning: Optimize model parameters for best performance.

## Phase 6: Model Evaluation and Selection

Performance Metrics Evaluation: Assess models using metrics such as accuracy, precision, recall, ROC-AUC.

Model Interpretation: Evaluate the interpretability of the models and the significance of different features.

Model Selection: Choose the best-performing model based on evaluation metrics and business relevance.

## Phase 7: Reporting and Presentation

Presentation: Detail how we intend to present the project's results and the strategies we will employ to ensure that the project is easily understandable to the audience.

Final Report: Prepare a comprehensive report detailing the analysis, model development and findings.

Future recommendations: Provide guidance on potential improvements aimed at enhancing accuracy, recall, and other relevant metrics. Emphasize that if the initial results fall short of expectations, a return to the initial phase for a more optimized implementation is a viable approach.

## Phase 8: Project Review and Closure

Project Review: Conduct a post-project review to evaluate successes, challenges, and learnings.

Conclusions: Write a conclusion of the project

Project Closure: Formally close the project and release resources.

# Phase 1: structure of the project and Business understanding

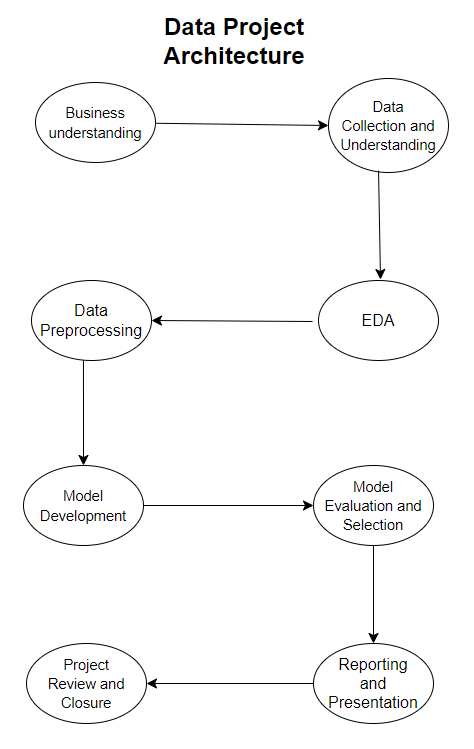


Figure 1 Data project phases

## Timeline:

Overview of Project's Timeline

Month 1: structure of the project and Business understanding

Month 2: Data Collection and Understanding

Month 3: Exploratory Data Analysis (EDA)

Month 4: Data Preprocessing

Month 5: Model Development

Month 6: Model Evaluation and Selection

Month 7: Reporting and Presentation

Month 7: Project Review and Closure

## Business Understanding: Bank Marketing Campaign Analysis

The project revolves around a dataset from a bank, specifically focusing on a marketing campaign related to term deposit subscriptions. Understanding customer behavior and predicting the likelihood of a customer subscribing to a term deposit is crucial for the bank's marketing strategy.

### Objective

Identify Key Influencers: Determine which factors (demographic, economic, or campaign-related) most significantly influence a customer's decision to subscribe to a term deposit.

Predictive Modeling: Develop a predictive model to forecast the likelihood of a customer subscribing to a term deposit. This model can be used to target potential subscribers more effectively, optimizing resources and improving campaign success rates.

Customer Segmentation: Understand different customer segments and their responsiveness to the campaign, allowing for more tailored and effective marketing approaches.

# Phase 2: Data Collection and Understanding

## Data Understanding

This phase focuses on data collection, exploration, and initial data analysis. It is crucial for understanding the available data, its quality, and its suitability for the project. We will collect and clean a large dataset of historical bank data.

The dataset contains 41,188 entries with 21 features, including customer demographics, economic indicators, and campaign details. Key features include age, job, marital status, education, housing, default, loan, contact, month,day of the week, duration, campaign, Pdays, previous, Poutcome, Emp.var.rate, cons.price.idx, cons.confi.idx, Euribor3m, Nr.emplyed, Y.

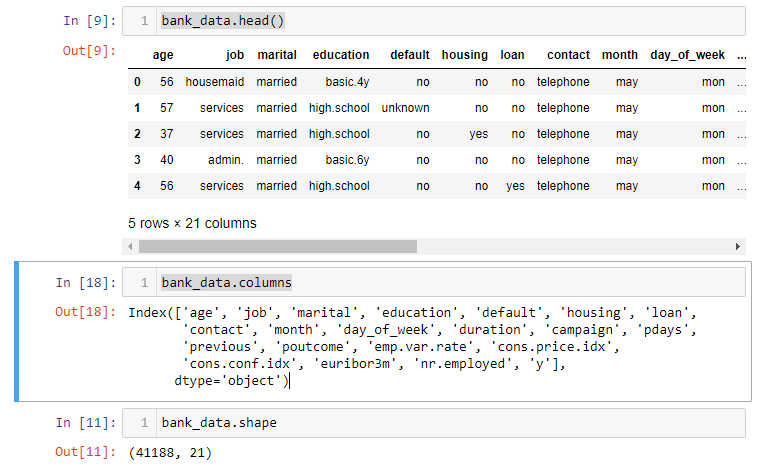


Figure 2 Visualizing the name of the columns

In our dataset, while we are familiar with the rows and columns, it's crucial to also understand the significance of each column. Some columns have obvious meanings, but others can be more complex and require detailed analysis. Let's analysis to the complexities columns for have a better understand of our data:

Duration: The duration of the last contact, in seconds

Pdays: The number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted).

Emp.var.rate: Employment variation rate ,quarterly indicator this rate indicate the influence customers financial stability.

Cons.price.idx: Consumer price index(CPI) ,monthly indicator that measures the average change over time in the prices paid by consumers for a market basket of consumer goods and services. (www.cso.ie, n.d.)

Cons.conf.idx: Consumer confidence index(CCI) ,monthly indicator that reflects the degree of confidence individual households have in the performance of the economy. (Investopedia, n.d.)

Euribor3m: Euribor 3 month rate ,daily indicator is the interest rate that refers to various financial products, including mortgages and savings accounts.

“Euribor is the acronym for the Euro Interbank Offered Rate. This is the interest rate at which credit institutions lend money to each other, which is often referred to as “the price of money” (Bankinter, n.d.).

Nr.employed: Number of employees.

Y: The target variable indicating whether the client has subscribed to a term deposit (binary: 'yes','no').

Step that we are going to implements:

“Collect initial data, Describe data, Explore data, Verify data quality” (Hotz, 2022).

# Phase 3: Exploratory Data Analysis (EDA)

## EDA

Proceeding to create the next steps for we can understand the data:

- Histograms for numerical variables to understand their distribution.

- Bar charts for categorical variables to see the distribution of different categories.

- A correlation matrix to observe any potential correlations between numerical variables.

### Histogram

The histograms provide (matplotlib.org, n.d.) insights into the distribution of the numerical variables:

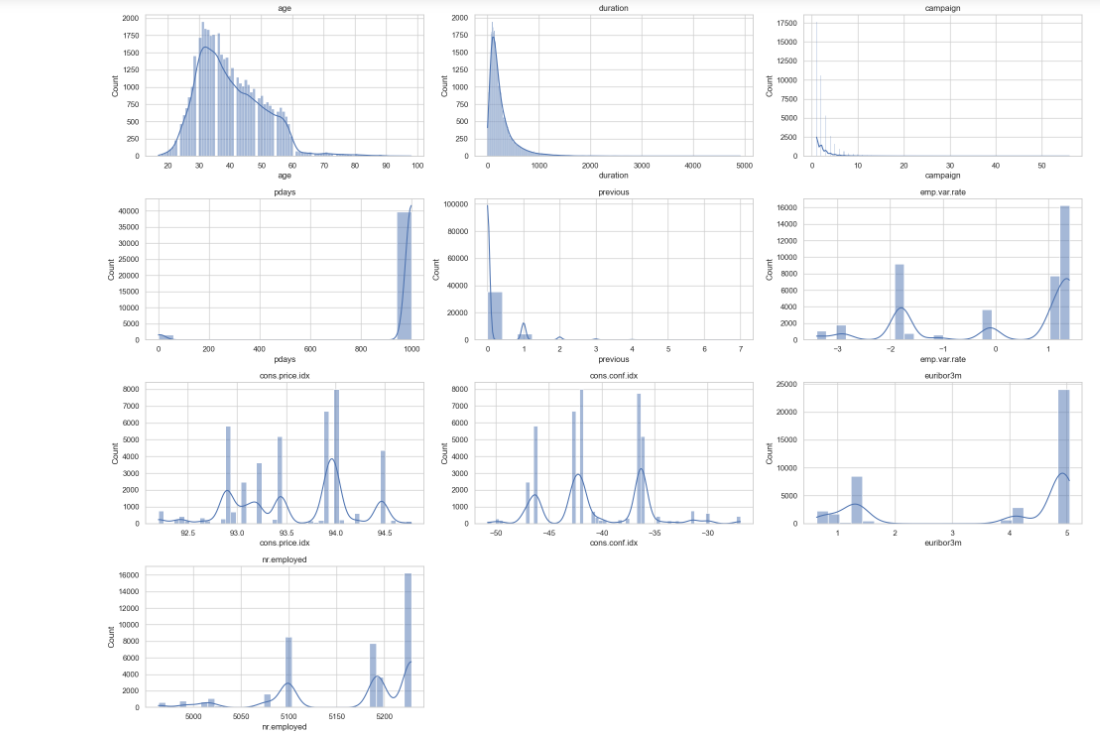


Figure 3 Histogram of the distribution of the numeric values

Age: Most customers are in the 30-40 age range, with a right-skewed distribution.

Duration: This shows a right-skewed distribution, with most calls being relatively short.

Campaign: Most customers were contacted a few times, with the distribution heavily skewed to the right.

Pdays: There's a spike at 999, which likely represents a 'not contacted' category. This needs to be considered in any analysis.

Previous: Similar to pdays, most customers were not previously contacted, as indicated by the peak at 0.

Economic Indicators (emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed): These show various distributions, some are multimodal, reflecting different economic conditions during the data collection period.

### Bar chart

Next, let's plot bar charts for the key categorical variables to understand their distribution. We will focus on 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day\_of\_week', 'poutcome', and the target variable 'y'.

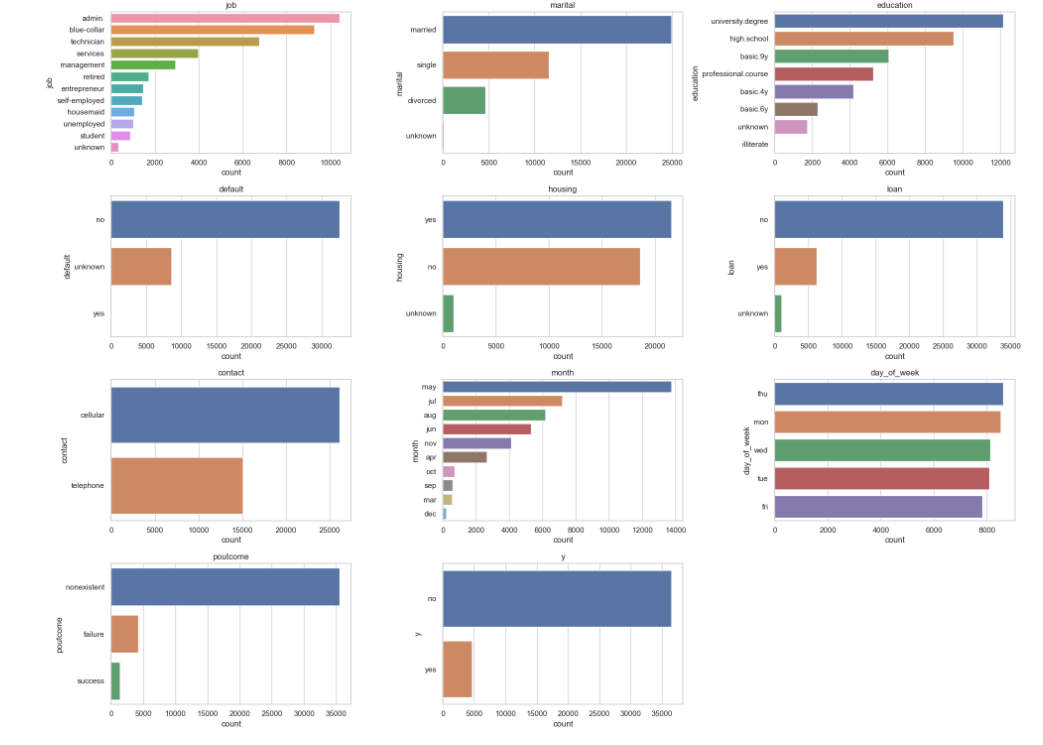


Figure 4 bar chart of the distribution of the categorical variables

The bar charts provide a visual representation of the distribution of categorical variables:

Job: The dataset includes a variety of job categories, with 'admin.', 'blue-collar', and 'technician' being the most common.

Marital: Most customers are married, followed by single and divorced.

Education: The highest frequency is for university degree, followed by high school education.

Default: Majority of customers are recorded as 'no' for default, with a significant portion marked as 'unknown'.

Housing: A fairly even distribution among 'yes', 'no', and 'unknown'.

Loan: Most customers do not have a personal loan.

Contact: 'Cellular' is the most common method of contact.

Month: There's a clear seasonality in contact, with May being the most common month.

Day of Week: Distribution is relatively even across different days of the week.

Poutcome: Majority of outcomes from the previous marketing campaign are 'nonexistent', indicating no prior contact.

Target Variable (y): There are significantly more 'no' responses than 'yes', indicating a lower subscription rate.

Finally, let's examine the correlation between numerical variables using a correlation matrix and a heatmap. This will help us understand if any variables are strongly related to each other

## Correlation matrix

The correlation matrix, visualized as a heatmap, shows the relationships between numerical variables:

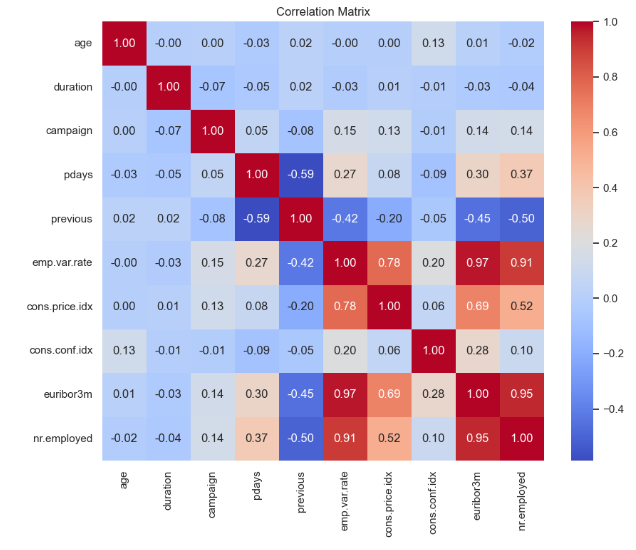


Figure 5 Correlation matrix

Strong Correlations: Some pairs of variables exhibit strong correlations. For example:

emp.var.rate and euribor3m.

nr.employed and euribor3m.

emp.var.rate and nr.employed.

These strong correlations are logical as these economic indicators often move together in response to overall economic conditions.

Negative Correlations: There are also negatively correlated pairs such as:

pdays and previous.

emp.var.rate and cons.conf.idx.

The negative correlation between pdays and previous might indicate that the more a client was contacted in the past (previous), the more recent their last contact was (pdays).

Weak or No Correlation: Many variables show weak or no significant correlation with each other.

Insights and Further Analysis

The EDA has provided valuable insights into the distribution and relationships within the data.

The next steps could include feature engineering (like handling the '999' in pdays), outlier detection, and preparing the data for modeling if the goal is predictive analysis.

For a more detailed analysis, we could delve into specific hypotheses, like understanding which factors most strongly predict the target variable 'y'.

# Phase 4: Data Preprocessing

### Data preparation

After perform a EDA we handling with special values: The '999' values in the pdays column were replaced with '-1' to differentiate clients not previously contacted.

We procced to separate the target variable, and we scale our data, by identifying the column type, numerical and categorical columns. Standardizes features like 'age' and 'balance' by removing the mean and scaling to unit variance for example: If the average age in the dataset is 40 with a standard deviation of 10, then the standardized value for an age of 30 would be (30 - 40) / 10 = -1.

Transforming the categorical values into a one-hot numeric array for 'job', it creates binary columns for each job type ('admin.', 'entrepreneur', 'blue-collar'),for 'education', it creates binary columns for each education level ('university', 'high.school', 'primary').

Scaling and encoding process is important for preparing the dataset for machine learning analysis. It ensures that numerical data is on a comparable scale and that categorical data is represented in a format suitable for modeling. The result is a transformed dataset where all features can be effectively used by our ML model.

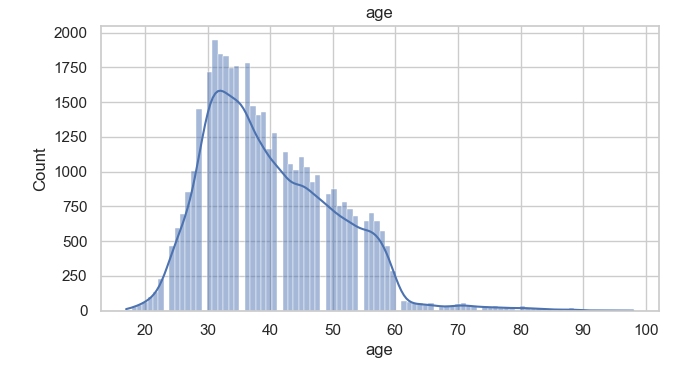


Figure 6 Histogram of Unscaled Age Data

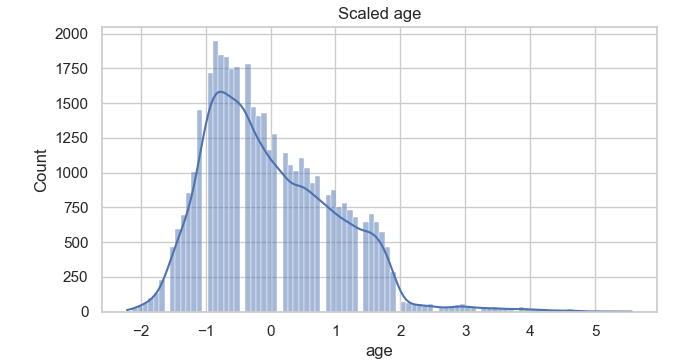


Figure 7 Histogram of Scaled Age Data

# Phase 5: Model Development

## Training the model

A logistic regression model was trained using a dataset comprising both numerical and categorical variables after scaling. The rationale of the choosing model is four our data set of bank marketing campaign, the objective is to predict whether a customer will subscribe to a term deposit, which is inherently a binary decision. The model's predictive performance was evaluated using a hold-out test set, and the results were summarized in a confusion matrix.

## Confusion Matrix Interpretation

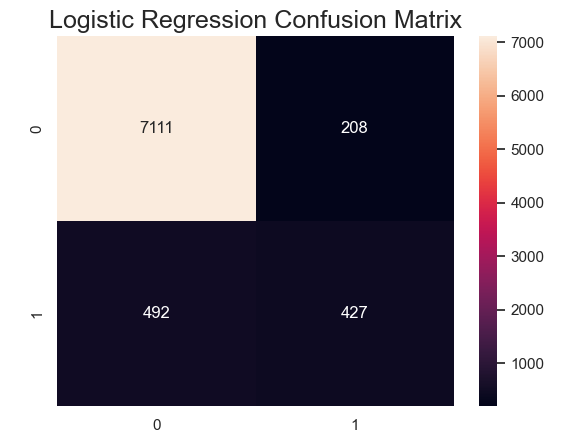


Figure 8 Confusion matrix of Logistic Regression

The confusion matrix provided a detailed breakdown of the model's predictions compared to the actual values:

The model showed a high number of true negatives, suggesting it is effective at identifying customers who are not likely to subscribe.

The true positives count indicated successful identification of actual subscribers, though this number was relatively low.

False negatives and false positives were present, revealing instances where the model's predictions deviated from the actual customer behaviors.

Model Accuracy

The model achieved an accuracy of approximately 91.5%, which at first glance appears to be excellent. However, this high accuracy may be partly due to an imbalanced dataset, where one class outnumbers the other significantly.

# Phase 6: Model Evaluation and Selection

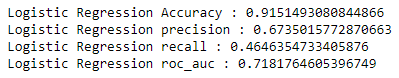


Figure 9 Results of the ML Logistic Regression

The model correctly predicts whether a customer will subscribe to a term deposit about 91.51%(Accurancy) of the time. While this number seems high, accuracy is not always the best measure of performance. When the model predicts that a customer will subscribe, it is correct approximately 67.35% (Precision)of the time. This is a relatively moderate precision rate, indicating that when the model predicts a positive outcome (subscription), it is reasonably reliable.

As well the model correctly identifies 46.46%(Recall) of all actual subscription cases. This means that it misses more than half of the potential subscribers, which could be a disaster to the bank’s marketing.

An AUC of 71.82% is considered good and suggests that the model has a reasonable measure of separability. It indicates that there is a 71.82% chance that the model will be able to distinguish between a subscriber and a non-subscriber.

This marks the initial implementation of our first model. In the subsequent phases of our project, we will explore a variety of models to determine the best fit for our data. While logistic regression shows promise, it may not be the optimal choice.

# Phase 7: Reporting and Presentation

## Future recommendations:

Explore whether additional features or feature transformation could improve model performance, particularly in terms of recall or trying to use the technique resampling or stratified cross-validation with hyperparameter tuning

Experiment with different models, such as decision trees, random forests, or gradient boosting machines, which might offer better performance, especially on an imbalanced dataset.

### Presentation:

# To effectively present the results of a data analysis project in a PowerPoint presentation, keeping it understandable for a diverse audience, including those not familiar with the data analysis process

# Final Report:

# Writing a comprehensive report detailing all aspects of the data analysis project. This report should encompass the project's objectives, the motivation behind its execution, a thorough presentation of the results, and a well-justified explanation of the chosen model. Additionally, we should provide detailed insights into specific sections of the code, incorporating advanced technical terminology and techniques employed throughout the project.

# Phase 8: Project Review and Closure

After we finish all the phases of our project to evaluate successes, challenges, and learnings; we write a conclusion of about the project and for close the project we put all the references that we implement in our project

After the completion of all project phases, where we evaluate the achievements, confront challenges, and acquire valuable insights, we compose a comprehensive project conclusion. To formally conclude the project, we compile list of references that we implemented in the project.

## Conclusion

We have structured our project, outlining each step in great detail. We began by selecting a dataset and conducting an in-depth Exploratory Data Analysis (EDA) to uncover patterns and insights. Following this, we processed the data to prepare it for Machine Learning (ML) techniques and analyzed the results.

To further enhance our project's outcomes, we are committed to improving the results by implementing additional Machine Learning techniques. In the upcoming phases, we will delve into steps such as presentation and closure, ensuring a comprehensive and well-rounded project delivery.

# Reference

Staff, G.P. (2020). *Understanding the Lifecycle of a Data Analysis Project*. [online] Graduate Blog. Available at: https://graduate.northeastern.edu/resources/data-analysis-project-lifecycle/.

Smith, A. (2023). 7 Fundamental Steps to Complete a Data Analytics Project. [online] blog.dataiku.com. Available at: <https://blog.dataiku.com/fundamental-steps-data-project-success>.

www.cso.ie. (n.d.). What is the CPI - CSO - Central Statistics Office. [online] Available at: https://www.cso.ie/en/interactivezone/statisticsexplained/consumerpriceindex/whatisthecpi/#:~:text=The%20Consumer%2 Investopedia. (n.d.). Understanding the Consumer Confidence Index. [online]

Available at: https://www.investopedia.com/insights/understanding-consumer-confidence-index/#:~:text=The%20Consumer%20Confidence%20Index%20(CCI.0Price%20Index%20or [Accessed 13 Dec. 2023].

Bankinter. (n.d.). What is Euribor and how does it affect me? | FAQs. [online] Available at: https://www.bankinter.com/banca/en/faqs/mortgages/what-is-euribor-and-how-does-it-affect-me#:~:text=Euribor%20is%20the%20acronym%20for [Accessed 13 Dec. 2023].

matplotlib.org. (n.d.). matplotlib.pyplot.subplot — Matplotlib 3.4.3 documentation. [online] Available at: https://matplotlib.org/stable/api/\_as\_gen/matplotlib.pyplot.subplot.html.