**ABSTRACT**

Data analysis helps data analysts to pull out meaningful results from the raw data and help their clients/users to make important business decisions by identifying various findings and patterns. It can be achieved by preparing reports, supporting analysis and presenting to management.

When it comes to processing and analyzing vast amounts of data to make sense of it, information technology in the twenty-first century is reaching new heights. Data analysis is essential in today's digital world for making sense of complex information and guiding well-informed decisions. The goal of this project was to increase fixed-term deposit product subscription rates. The study aims to provide information on how banks can effectively promote these items in the face of the overabundance of advertising.

By using Python libraries like numpy, pandas and matplotlib, this project handles a dataset of 41,188 clients of Portugal bank with diverse attributes. Categorical and numeric features including details of clients and their contact history, are conscientiously managed.

Results show significant factors influencing subscription rates. The impact of economic variables like Libor rates and client characteristics like age are among the key insights. The report makes useful suggestions for banks, like proactive marketing that is in line with economic trends and placing a focus on engagement during client contacts.

Banks now require a 360-degree perspective of their customers because, without one, they risk missing a competitive market advantage. This study emphasizes how effective data-driven insights can be in enhancing marketing plans for the banking industry. The study demonstrates how banking industries can use data analysis to improve marketing efforts to increase the subscription rates through Python's expertise. For visualizing the data, we have used Tableau tool and then deployed ML algos with the help of Python Colab notebook.

**TABLE OF CONTENTS**

| **TITLE** | **PAGE NO.** |
| --- | --- |
| **ABSTRACT** | **vi** |
| **LIST OF FIGURES** | **ix-xi** |
| **LIST OF TABLES** | **xii** |
| **LIST OF ABBREVIATIONS** | **xiii** |
| **CHAPTER - 1 INTRODUCTION**  **1.1 General Introduction**  **1.2 Tools and Technology Used** | **14-15**  **14-15**  **15** |
| **CHAPTER - 2 LITERATURE SURVEY**  **2.1 Summary of research papers studied**  **2.2 Problem Statement** | **16-21**  **16-20**  **20-21** |
| **CHAPTER - 3 REQUIREMENT ANALYSIS**  **3.1 Functional Requirements**  **3.2 Non-Functional Requirements** | **22-26**  **22-24**  **24-26** |
| **CHAPTER - 4 REPORT ON PRESENT INVESTIGATION**  **4.1 Ideation and Proposed Solution**  **4.2 Description of Project**  **4.3 Algorithms Used** | **27-34**  **27-29**  **29-32**  **32-34** |
| **CHAPTER - 5 RESULTS AND DISCUSSION**  **5.1 Implementation Details and Issues**  **5.1.1 Implementation Details**  **5.1.2 Implementation Issues**  **5.2 Evaluation Parameters**  **5.3 Observations and Results** | **35-73**  **35-40**  **35-39**  **39-40**  **40**  **40-73** |
| **CHAPTER - 6 CONCLUSION AND FUTURE WORK**  **6.1 Conclusion**  **6.1.1 Demographic Insights**  **6.1.2 Campaign and Subscription trends**  **6.1.3 Financial and Economic indicators**  **6.1.4 Machine Learning model insights**  **6.1.5 Project challenges and considerations**  **6.2 Future Work**  **CHAPTER - 7 REFERENCES** | **74-76**  **74-75**  **74**  **74**  **75**  **75**  **75**  **76**    **77-79** |
|  |  |

**LIST OF FIGURES**

| **Fig. No.** | **Name of Figure** | **Page No.** |
| --- | --- | --- |
| 4.2.1 | Block diagram showing the lifecycle of project | 30 |
| 5.1.1.1 | 1. CSV File Responses 2. CSV File Responses   (c) CSV File Responses  (d) CSV File Responses | 35  35  36  36 |
| 5.1.1.2 | 1. Dataframe uploaded in Project 2. Dataframe uploaded in Project | 36  37 |
| 5.1.1.3 | (a) Describing the dataset (Statistics)  (b) Describing the dataset (Statistics) | 37  37 |
| 5.1.1.4 | 1. Correlation Table 2. Correlation Table | 38  38 |
| 5.1.1.5 | Data Dashboard representing different parameters | 39 |
| 5.3.1 | Distribution of Age of clients | 41 |
| 5.3.2 | Distribution of Age Group of clients | 41 |
| 5.3.3 | Percentage of clients subscribed and unsubscribed to a term deposit | 42 |
| 5.3.4 | Percentage of clients having housing loan or not | 42 |
| 5.3.5 | Marital status of clients | 43 |
| 5.3.6 | Distribution of clients as per their job/profession | 43 |
| 5.3.7 | Distribution of clients as per their Education level | 44 |
| 5.3.8 | Duration of contact(calls) made to the clients (on a log scale) | 44 |
| 5.3.9 | Finding outliers using box plot in Duration variable | 45 |
| 5.3.10 | Count of different Employment variation rates | 45 |
| 5.3.11 | Percentage of different modes of contact to clients | 46 |
| 5.3.12 | Distribution of campaign days | 46 |
| 5.3.13 | Day-wise count of Last contact | 47 |
| 5.3.14 | Distribution of previous campaign outcome | 47 |
| 5.3.15 | (a) Word cloud for Campaign months  (b) Count of calls in different months of Campaign | 48  48 |
| 5.3.16 | Distribution of days from previous outcome to this campaign | 49 |
| 5.3.17 | Box plot of Consumer Price Index | 49 |
| 5.3.18 | Distribution of Consumer Confidence Index | 50 |
| 5.3.19 | Count of number of contacts before this campaign | 50 |
| 5.3.20 | Box plot representing Euribor 3 month rate | 51 |
| 5.3.21 | (a) Box plot for Number of employees  (b) Density plot for number of employees | 51  52 |
| 5.3.22 | Purchase frequency for different jobs | 53 |
| 5.3.23 | Relationship between Age(in years) and Consumer Confidence Index | 54 |
| 5.3.24 | Mean Age for different Education levels | 55 |
| 5.3.25 | Count of Marital status with different Education levels | 56 |
| 5.3.26 | Stacked bar chart of Marital status with their purchasing status | 57 |
| 5.3.27 | Distribution of Duration in different Age groups | 58 |
| 5.3.28 | Distribution of Job categories among different Marital status | 59 |
| 5.3.29 | Cross-tabulation of Education and default loan status | 60 |
| 5.3.30 | Cross-tabulation of Housing and Personal loan status | 61 |
| 5.3.31 | (a) Purchase frequency for different days of week  (b) Density plot for purchase frequency with respect to different duration | 61  62 |
| 5.3.32 | Line plot representing relationship between Month, Age group and their subscription status | 62 |
| 5.3.33 | Plot showing relationship between Age group, Campaign contacts and outcomes of previous campaign | 63 |
| 5.3.34 | Line plot showing relationship between pdays, duration of contact according to previous campaign outcome | 63 |
| 5.3.35 | Plot showing relationship between previous contact days and subscription status according to previous campaign outcomes | 64 |
| 5.3.36 | Plot showing relationship between Education, Personal loans and Default credit status | 65 |
| 5.3.37 | Mosaic plot showing subscription status according to Marital status | 66 |
| 5.3.38 | Month-wise proportion of housing loan by clients | 66 |
| 5.3.39 | Heat map between Housing, Personal and Default loan status | 67 |
| 5.3.40 | Correlation heat map between all numerical fields of Dataframe | 68 |
| 5.3.41 | (a) Accuracy percentage for different Supervised models in order  (b) Plot of accuracy percentage for different supervised models | 69  70 |
| 5.3.42 | (a) AUC ROC curve for the highest accuracy model (Gradient Boosting)  (b) Accuracy curve for the highest accuracy model (Gradient Boosting) | 71  71 |
| 5.3.43 | (a) Accuracy percentage for different unsupervised models  (b) Plot of accuracy percentage for different unsupervised models | 72  72 |

**LIST OF TABLES**

| **Table No.** | **Name of Table** | **Page No.** |
| --- | --- | --- |
| 2.1 | Summary of the research papers studied | 16-20 |

**LIST OF ABBREVIATIONS**

| * Algos | * Algorithms |
| --- | --- |
| * Colab | * Google Colaboratory |
| * CSV | * Comma Separated Values |
| * i.e. | * That is |
| * AI | * Artificial Intelligence |
| * ML | * Machine Learning |
| * AUC | * Area under the curve |
| * ROC | * Receiver Operating Characteristics |
| * CV | * Cross Validation |
|  |  |

**CHAPTER 1**

**INTRODUCTION**

**1.1 General Introduction**

Today's business environment places a high value on data analysis as a tool for strategic decision-making. The goal of this study was to increase fixed-term deposit product subscription rates. Due to the growing ubiquity of advertising in the digital age, businesses are looking for novel strategies to engage and attract customers. The research in this regard aims to answer a key question: How can banks optimize the marketing of fixed-term deposit products to ensure maximum effectiveness and success rates?

By using Python libraries like numpy, pandas and matplotlib, this project handles a dataset of 41,188 clients of Portugal bank with diverse attributes. Categorical and numeric features including details of clients and their contact history, are conscientiously managed. This study uses a systematic strategy to draw relevant insights from the data through careful preprocessing and meticulous analysis.

To tackle the challenge of predicting subscription rates, the project employs a range of machine learning models, including Logistic Regression, Decision Trees, Random Forests, CATBoost and Gradient Boosting, Stacking CV classifier, etc. These models had undergone rigorous training and evaluation using the fine-tuning of hyperparameters to enhance predictive accuracy.

In the final analysis, the research reveals useful information for the banking sector by highlighting the important factors that affect subscription rates. The research emphasizes the possibility of data-driven methods in transforming marketing campaigns and encouraging improved subscription outcomes by utilizing Python's capabilities and the power of data analysis. The findings of this analysis will provide actionable recommendations and insights to enhance decision-making processes within the banking industry and we can optimize resource allocation, refine marketing strategies, and improve customer satisfaction, thereby driving long-term success and growth in the ever-evolving market.

**1.2** **Tools and Technology Used**

* **Python**

Python is a high-level and general-purpose programming language. Python is dynamically-memory allocated. It supports multiple programming patterns, including structured , object-oriented and functional programming. Python is often seen as a "cell included" language due to its huge standard library [[1]](#1fob9te).

* **Machine learning**

Machine learning is a sub-part of AI that gives systems the capability to learn on their own and improve from experiences without the interference of humans. Machine learning focuses on the development of computer algorithms and accessing data and performing various functions [2].

* **Tableau**

Tableau is a visual analytics tool that is revolutionizing how we utilize data to address issues by enabling individuals and companies to maximize their data. It has consistently made unmatched investments in research and development, creating technologies to aid anyone working with data in finding answers more quickly and discovering unexpected insights [3].

* **Google Colab Notebook**

Colab is a product from Google Research. It performs the same as a jupyter notebook on the desktop [[4]](#2et92p0).

**CHAPTER 2**

**LITERATURE SURVEY**

**2.1 Summary of the research papers studied**

**Table 2.1 Summary of Literature Survey**

| **S.No** | **Name of Paper** | **Authors** | **Summary** | **Tools / Technologies** |
| --- | --- | --- | --- | --- |
| 1 | Big Data and Analytics in Retailing [5] | Venky [Shank](https://www.researchgate.net/scientific-contributions/Venky-Shankar-2163866383)ar | The importance of big data analytics in the retail industry, focusing on its role in shaping marketing strategies, understanding customer preferences, and making informed business decisions. | Big Data Analytics and ML |
| 2 | Retail Analytics: Driving Success in Retail Industry with Business Analytics [6] | Sudeep B. Chandramana | Big data in retail revolutionizes decision-making, enhances profitability, enables real-time insights, addresses data privacy challenges, leverages omnichannel behavior, and drives an AI-driven future for improved customer experiences, operations, and profitability. | Big Data Analytics, ML and AI |
| 3 | Retailing and retailing research in the age of big data analytics [7] | Marnik G. Dekimpe | This paper describes the allure of the retail sector as a research domain, emphasizing its size, multifaceted and dynamic nature, the availability of high-quality data, and coverage by business analysts. It also discusses the potential of big data analytics in the retail industry, including its impact on retail managers, retailing researchers, public policy makers, investors, and retailing educators. | Business Analytics |
| 4 | Big Data Analytics: A Literature Review Paper [8] | Elgendy, Nada & Elragal, Ahmed | This paper discusses that big data analytics allows organizations to analyze large volumes of data from various sources and uncover patterns, trends, and customer behaviors. Applications of big data analytics include customer intelligence, supply chain management, performance and quality management, risk mitigation, and fraud detection. It offers opportunities for organizations to enhance their decision-making processes, improve efficiency, and gain a competitive edge in various industries. | Big Data and its Analysis |
| 5 | Big Data Analysis in Banking Sector  [[9]](#2s8eyo1) | Rahul More,  Yash Moily | The definition of big data, its benefits for the banking industry, the data mining techniques used in big data analytics and some big data analytics applications in the business world, particularly in banking, are just a few of the topics covered in the paper. | Big Data, Data Analytics |
| 6 | Data Analytics Integration in Banking Industry [10] | Musa J. Jafar,  Jeffry Babb,  Amjda Abdullat | This provides a thorough investigation of data analysis in the banking sector, covering everything from data storage to predictive analytics and the usefulness of Python. The survey highlights the importance of data-driven decision-making and the opportunity for businesses to use data to their advantage and propel success in the volatile financial environment. | Data Analytics, Python |
| 7 | Walmart's Sales Data Analysis - A Big Data Analytics Perspective [11] | Singh, Manpreet & Ghutla, Bhawick & Jnr, Reuben & Mohammed, Aesaan & Rashid, Mahmood | The purpose is to learn more about consumer behavior, comprehend the variables that influence sales, and forecast future sales. By analyzing, we can improve processes, distribute resources wisely, and increase revenue. | Big Data  and  Analytics |
| 8 | Impact of big data analytics on sales performance in pharmaceutical organizations: The role of customer relationship management capabilities [12] | Shahbaz, Muhammad & Gao, Changyuan & Zhai, Lili & Shahzad, Fakhar & Luqman, Adeel & Zahid, Rimsha | This paper suggested the impact of big data analytics (BDA) on sales performance and customer relationship management (CRM) capabilities. The findings suggest that individual and organizational characteristics influence the salesforce's perception of BDA, leading to improved person-technology fit, CRM capabilities, and sales performance. | Big Data Analytics |
| 9 | A study on Financial Analysis of BSNL[[13]](#35nkun2) | J.Pavithra, Dilip Gurukrishnan | This paper focused on identifying the various assets of BSNL with respect to the Annual Reports ,studying the functioning of the finance department and then comparative analysis study of two years (2009-2010). | Statistical tools, Trend Analysis |

**2.2 Problem Statement**

**Inconsistent and Missing Data:** The project addresses the challenge of dealing with inconsistent and missing data in the dataset collected from a bank marketing campaign .

**Performance and Profitability:** The purpose of the study is to address the problem of varying sales performance and profitability related to fixed-term deposit products. The initiative tries to find patterns and connections that can lead to more efficient marketing tactics by utilizing data analysis and machine learning.

**Operational Efficiency:** The operational effectiveness of marketing campaigns for fixed-term deposit products is thoroughly explored in the study. It aims to determine the most effective ways for banks to distribute their resources, customize their methods, and reduce waste.

**Customer Satisfaction and Retention:** The project acknowledges the critical importance of client retention and satisfaction in the banking industry. In order to improve customer experience and loyalty for fixed-term deposit products, it aims to uncover consumer preferences, pain areas, and engagement patterns through data-driven insights.

**Market Analysis and Expansion:** The paper addresses the problem of conducting a thorough analysis of the fixed-term deposit market. It seeks to identify region-specific trends, variances in market demand, and potential expansion prospects using machine learning algorithms, empowering banks to make well-informed strategic decisions.

**CHAPTER 3**

**REQUIREMENT ANALYSIS**

**3.1 Functional Requirements**

**Customer Data Management:**

* The system should securely store and manage customer data, including personal information, contact details and account history.
* It should allow creating, updating and deleting customer profiles while ensuring data accuracy and integrity.
* The system should provide role-based access to customer data, with appropriate security measures to prevent unauthorized access.

**Account Management:**

* The system should facilitate the creation and management of different types of bank accounts, such as savings, checking and fixed deposits.
* It should allow customers to open new accounts, close existing ones and update account information.
* The system should enable account linkage for joint account holders or beneficiaries.

**Transaction Processing:**

* The system should support various types of transactions, including deposits, withdrawals, fund transfers and bill payments.
* It should ensure real-time transaction processing, accurate balance updates and transaction history maintenance.
* For security, the system should implement multi-factor authentication for sensitive transactions.

**Loan and Credit Management:**

* The system should handle loan and credit applications, approvals and disbursements.
* It should calculate loan interest rates, repayment schedules and ensure adherence to regulatory guidelines.
* The system should monitor credit limits, track outstanding loans and send reminders for upcoming payments.

**Security and Fraud Prevention:**

* The system should implement robust security measures, including encryption, firewall protection and intrusion detection.
* It should continuously monitor transactions for suspicious activities and flag potential fraudulent transactions.
* The system should enable customers to set transaction alerts and notifications for account activities.

**Customer Support and Communication:**

* The system should provide channels for customer inquiries, support requests and issue resolution.
* It should offer communication options like secure messaging, chat support and helpline numbers.
* The system should maintain a record of customer interactions and support responses.

**Reporting and Analytics:**

* The system should generate reports on account balances, transaction histories, interest accruals and loan statuses.
* It should offer data analytics tools to identify customer behavior patterns, popular services and potential market trends.
* Reports and analytics should be accessible to authorized bank personnel for decision-making.

**Regulatory Compliance:**

* The system should adhere to relevant banking regulations and data protection laws.
* It should generate required reports for regulatory bodies, taxation, and audits.
* The system should implement procedures to ensure data privacy and customer consent.

**3.2 Non-Functional Requirements**

**Performance:**

* The system should ensure responsive query times and efficient data processing, even for extensive data sets.
* It must maintain optimal performance under heavy concurrent user loads, avoiding noticeable slowdowns.
* The system's architecture should prioritize efficient data retrieval and analysis for timely insights.

**Scalability:**

* The system must be designed to handle increased data volume, user activity and accommodate future growth.
* It should allow for the seamless addition of new banking products, services and expanding customer base.
* Scalability options such as horizontal scaling should be available to accommodate varying demand levels.

**Reliability:**

* Data integrity and consistency must be maintained, with minimal occurrences of data errors or discrepancies.
* The system should include regular data backup and recovery mechanisms to prevent data loss due to system failures.
* High availability is crucial to minimize disruptions, ensuring uninterrupted customer access and service.

**Security:**

* The system should employ robust access controls, user authentication, and authorization mechanisms to safeguard sensitive data.
* Encryption protocols should be implemented to protect data during transmission and storage.
* Compliance with relevant data protection regulations and industry standards is essential to ensure customer data privacy.

**Usability:**

* The system should offer an intuitive, user-friendly interface for easy navigation and interaction.
* Visualizations, reports, and dashboards should be clear and meaningful to facilitate effective data analysis.
* Customization features should enable users to tailor the interface to their preferences, enhancing user experience.

**Maintainability:**

* The system's modular structure and comprehensive documentation should support easy maintenance and updates.
* Data updates and modifications should be performed efficiently without causing disruptions to ongoing operations.
* Effective troubleshooting and debugging should be possible through a clear separation of components.

**Compatibility:**

* The system should be compatible with various operating systems, browsers, and devices, ensuring flexibility for users.
* Integration capabilities with external systems, such as payment gateways or regulatory databases, should be seamless.
* Compliance with data interoperability standards will facilitate data exchange and integration with third-party applications.

**CHAPTER 4**

**REPORT ON PRESENT INVESTIGATION**

**4.1 Ideation and Proposed Solution**

Based on the Portugal Bank’s marketing campaign data , here are some ideation areas and proposed solutions that can be explored:

**Customer Engagement Enhancement:**

* Create targeted marketing techniques to promote the financial products or services that are in more demand.
* Create consumer segments based on their spending habits and preferences to create more specialized offers and incentives.
* Use upselling and cross-selling tactics to get consumers to investigate related financial services.

**Operational Efficiency Improvement:**

* To decrease operational bottlenecks and improve processing times, analyze banking transaction data.
* Automate document submission and identification verification to improve the customer onboarding process.
* Use digital document management programs to simplify paperwork and minimize human mistakes.

**Risk Mitigation and Fraud Prevention:**

* Using machine learning techniques to find odd patterns in transaction data that can signal probable fraud.
* To spot trends that might point to security breaches, track and analyze client activity.
* Create real-time alerts and triggers to alert security personnel to questionable activity so they can take action right away.

**Credit Scoring Enhancement:**

* Create predictive models using past credit data to evaluate the creditworthiness of your consumers.
* To improve the assessment of credit risk, including additional data sources such as online activity or social media activity.
* Create flexible credit scoring systems that can adjust to shifting market conditions and consumer trends.

**Digital Banking Experience Enhancement:**

* Examine how users engage with digital banking services to spot problems and potential solutions.
* Create simple, user-friendly interfaces for seamless account administration and navigation.
* Use chatbots with AI to answer common questions and offer real-time customer care.

**Financial Planning and Advisory Services:**

* Utilize data insights to provide individualized financial guidance and investment suggestions.
* Create financial planning tools that assist clients in establishing and achieving their financial objectives.
* Give customers financial portfolios and growth trajectories as interactive dashboards.

**Regulatory Compliance and Reporting:**

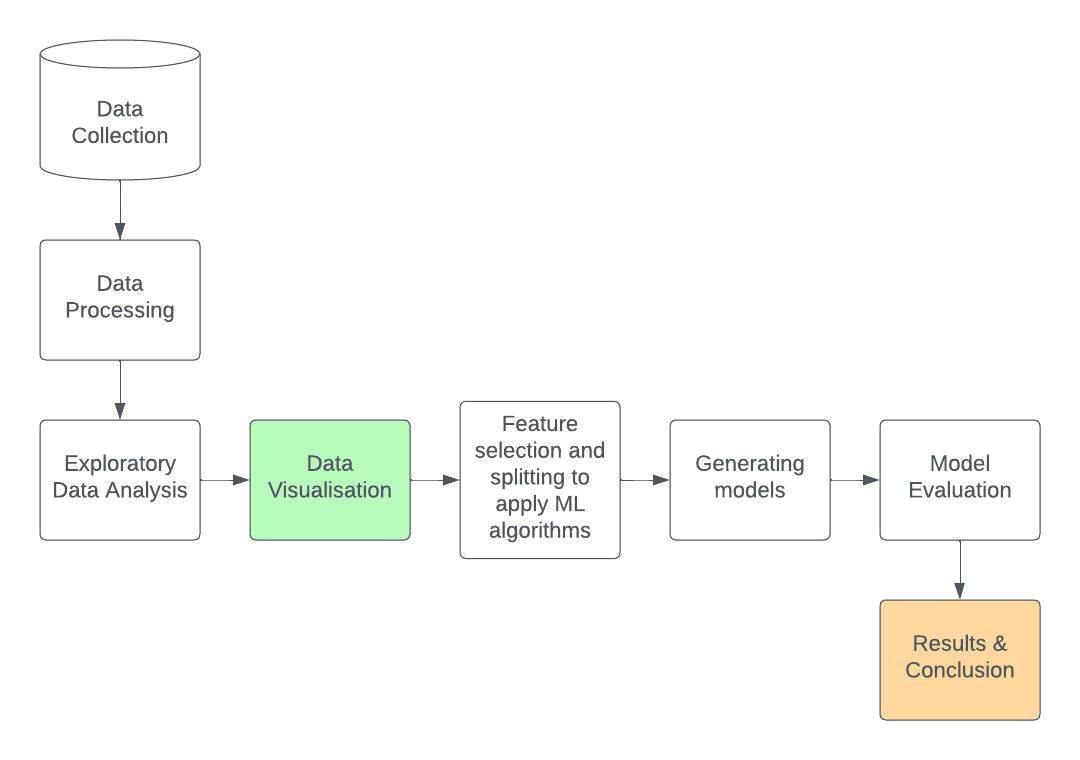
* Identify potential infractions using data analytics to ensure compliance with legal standards.
* Reduce human labour by automating the production of regulatory reports and disclosures.
* Secure access controls and data encryption should be used to protect sensitive client information.

**Branch Optimization and Customer Segmentation:**

* In order to improve branch locations and services, analyze consumer demographics and transaction trends.
* To provide specialized services, segment customers depending on variables such as income, age, and transaction frequency.
* Create focused marketing initiatives to draw particular customer segments to certain branches.

**4.2 Description of Project**

**Data Collection** : We searched and downloaded the dataset of the Bank marketing campaign which captured data of a marketing campaign run by Banco de Portugal and the records were saved in the form of a csv file. The file consisted of 41188 rows and 21 columns namely, ‘age’, ‘job’, ‘marital’, ‘education’, ‘default’, ‘ housing’, ‘loan’, ‘contact’, ‘month’, ‘day\_of\_week’, ‘duration’, ‘campaign’, ‘pdays’, ‘previous’, ‘poutcome’, ‘emp.var.rate’, ‘cons.price.idx’, ‘cons.conf.idx’, ‘euribor3m’, ‘nr.employed’ and ‘y’.



**Fig 4.2.1 Block Diagram showing the Lifecycle of Project**

**Data Processing (in python):**

1. Importing python libraries
2. Loading the data (.csv) in dataframe
3. Dropping the irrelevant columns
4. Renaming the columns
5. Replacing the values with the new values
6. Extracting required columns (new)
7. Removing duplicate values

After preprocessing, we finally included the following 22 columns with their respective values -

1. age : age of clients (numeric).
2. job : type of job (categorical) having values “admin.” , “blue-collar”, “entrepreneur”,“housemaid”,“management”,“retired”,“self-employed”, “services”, “student”, “technician”, “unemployed” and “unknown”.
3. marital : marital status of clients (categorical) having values “divorced”, “married”, “single” and “unknown”.
4. education : education level of clients (categorical) having values “basic” [aggregate form of “basic.4y”, “basic.6y” and “basic.9y”], “high.school”, “illiterate”, “professional.course”, “university.degree” and “unknown”.
5. default : does he/she have credit in default or not (categorical) having values “yes”, ”no” and “unknown”.
6. housing : does he/she have a housing loan or not (categorical) having values “yes”, ”no” and “unknown”.
7. loan : does he/she have a personal loan or not (categorical) having values “yes”, ”no” and “unknown”.
8. contact : contact communication type (categorical) having values “cellular” and “telephone”.
9. month : last contact month of year (categorical) having values “jan”, “feb”, “mar”, “apr”, “may”, “jun”, “jul”, “aug”, “sep”, “oct”, “nov” and “dec”.
10. day\_of\_week : last contact day of the week (categorical) having values “mon”, “tue”, “wed”, “thu” and “fri”.
11. duration : last contact duration, in seconds (numeric).
12. campaign : number of contacts performed during this campaign and for this client (numeric).
13. pdays : number of days that passed by after the client was last contacted from a previous campaign (numeric ; -1 means client was not previously contacted).
14. previous : number of contacts performed before this campaign and for this client (numeric).
15. poutcome : outcome of the previous marketing campaign (categorical) having values “failure”, “nonexistent” and “success”.
16. emp.var.rate : employment variation rate - quarterly indicator (numeric)
17. cons.price.idx : consumer price index - monthly indicator (numeric)
18. cons.conf.idx : consumer confidence index - monthly indicator (numeric)
19. euribor3m : euribor 3 month rate - daily indicator (numeric)
20. nr.employed : number of employees - quarterly indicator (numeric)
21. subs : has the client subscribed a term deposit? (binary) having values “yes” and “no”.
22. Age\_Group : Grouping of “age” column in the following list of values "Under 15", "16-20", "21-25", "26-30", "31-35", "36-40", "41-45", "46-50", "51-55", "56-60", "61-65", "66-70", "71-75", "76-80", "81-85", "86-90", "91-95" and "Above 95".

There were 41176 rows in the dataframe after final pre-processing.

**Exploratory Data Analysis and Data Visualisation:** We performed analysis through visualizations using Tableau and Python.

**Feature Selection and Splitting to apply ML algorithms :** In this, we selected the and splitted the categorical and numerical features required for analysis.

**Generating Models:** In this, we applied different machine learning algorithms on the field ‘y’ , renamed as ‘subs’ having values as ‘yes’ and ‘no’ which tells whether the client has signed up for the offer or not and generated the models for evaluation.

**Model Evaluation:** In this, we checked the performance of different models (including ensemble models) and which model has the highest accuracy.

**Applying the Model:** In this, we chose the final model to be applied and implemented as a predictive model for this problem.

**Results and Conclusion:** We have discussed the outcomes drawn from our analysis and concluded our project.

**4.3 Algorithms Used**

* **Random forest:** Random forest is a type of supervised ML algo. that is used in classification as well as regression problems. It works on making decision trees on different samples and takes majority votes for classification problems and average for regression problems [[14]](#44sinio).
* **K-nearest neighbour:** KNN or K-Nearest Neighbour, is a type of supervised machine learning algo. that can be used to solve both classification and regression problems. It is also known as “Lazy Learning”. On the basis of “K” , the new data point is given the position after calculating distance to the nearest neighbourhood. It is used for non linear data [[15]](#2jxsxqh).
* **Gradient boosting:** It is one of the most powerful boosting algo. of ML. It works on the principle of comparing previous models and giving the best possible next model to minimise errors [[16].](#z337ya)
* **Decision tree:** It is a supervised learning algo. that can be used for both classification and regression problems, but mostly preferred for classification problems. It makes a tree-based structure based on splitting of data in two or more homogeneous sets/groups based on the most significant splitter input variable [[17].](#3j2qqm3)
* **SVM:** Support Vector Machine or SVM is a classification algorithm (in most cases) which works on a mathematical technique called kernel functions (these allow to map data into higher dimensional space) and then find a hyperplane that separates the two classes of data. This hyperplane is called a decision boundary [[18].](#1y810tw)
* **Logistic regression:** This is a process of fitting data into a logit function and predicting probability of an event. The most common logistic regression model gives a binary output. Examples- Yes/No, 0/1 [[19].](#4i7ojhp)
* **CATBoost:** This is a variant of gradient boosting that can handle both categorical and numerical features. It does not require any feature encodings techniques like One-Hot Encoder or Label Encoder to convert categorical features into numerical features [20].
* **XGBoost:** This combines the predictions of multiple weak models to produce a stronger prediction. XGBoost can efficiently handle missing values, which allows it to handle real-world data with missing values without requiring significant pre-processing [21].
* **Naive Bayes:** Supervised ML which is based on Bayes theorem. Using Bayes theorem, find the probability of A happening, given that B has occurred where A is the hypothesis and B is the evidence. Basically, it is used for classification problems. It considers each predictor variable to be independent of each other and hence it is called “naive” [22].
* **Agglomerative clustering:** It is also known as hierarchical agglomerative clustering (HAC) or the bottom-up technique. a structure that provides more useful information than the flat clustering method's unstructured set of clusters. The number of clusters does not need to be predetermined for this clustering process. Bottom-up algorithms initially consider each piece of data as a singleton cluster before combining pairs of clusters until there is only one cluster left that contains all of the data [23].
* **K means:** One of the most straightforward and well-liked unsupervised machine learning algorithms is K-means clustering. Finding underlying patterns by combining comparable data points is K-means straightforward goal. K-means searches a dataset for a predetermined number (k) of clusters in order to accomplish this goal. Finding the centroid is what "means" in the K-means algorithm indicates: averaging the data [24].
* **Stacking CV Classifier:** Stacking is an ensemble learning strategy that uses a meta-classifier to integrate various classification models. Following the fitting of the meta-classifier using the outputs or meta-features of the individual classification models in the ensemble, the individual classification models are trained using the entire training set. Either the ensemble probabilities or the projected class labels can be used to train the meta-classifier [26].

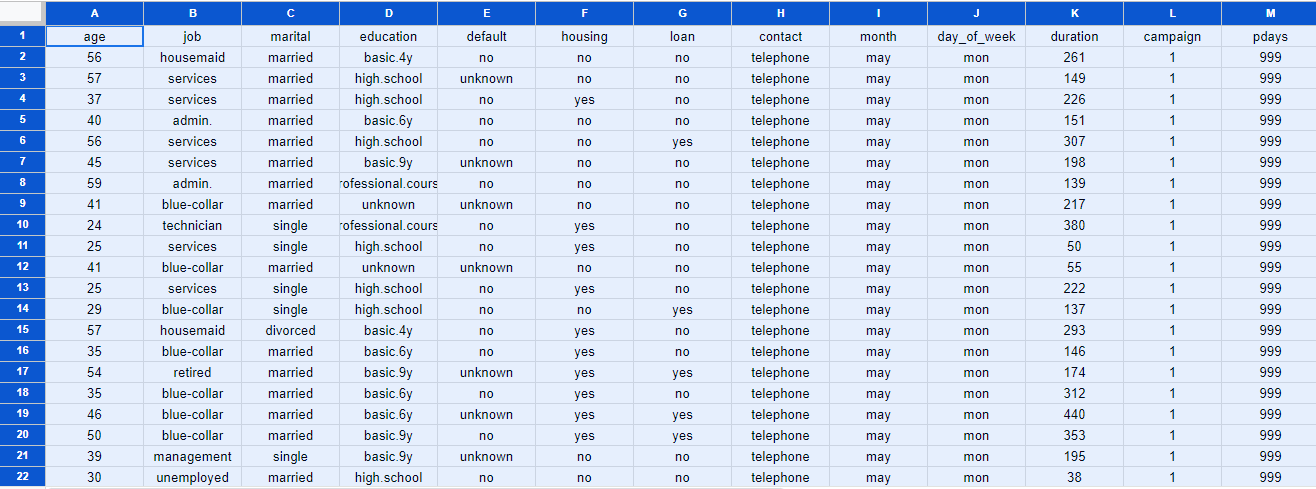
**CHAPTER 5**

**RESULTS AND DISCUSSIONS**

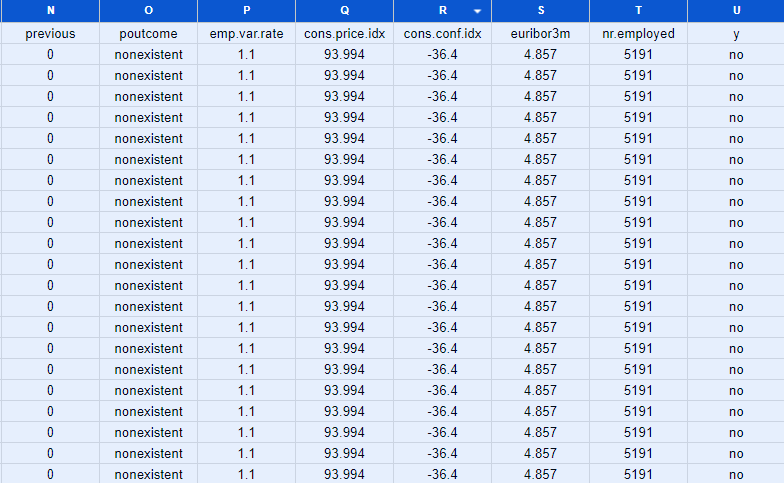
**5.1 Implementation Details and Issues**

**5.1.1 Implementation Details**

**Step -1 Gathering of data :** We gathered the data of Portugal Bank marketing campaign in the form of .csv file format.Fig 5.1.1.1(a) to Fig 5.1.1.1(d) represents the first 21 responses stored in the original CSV .

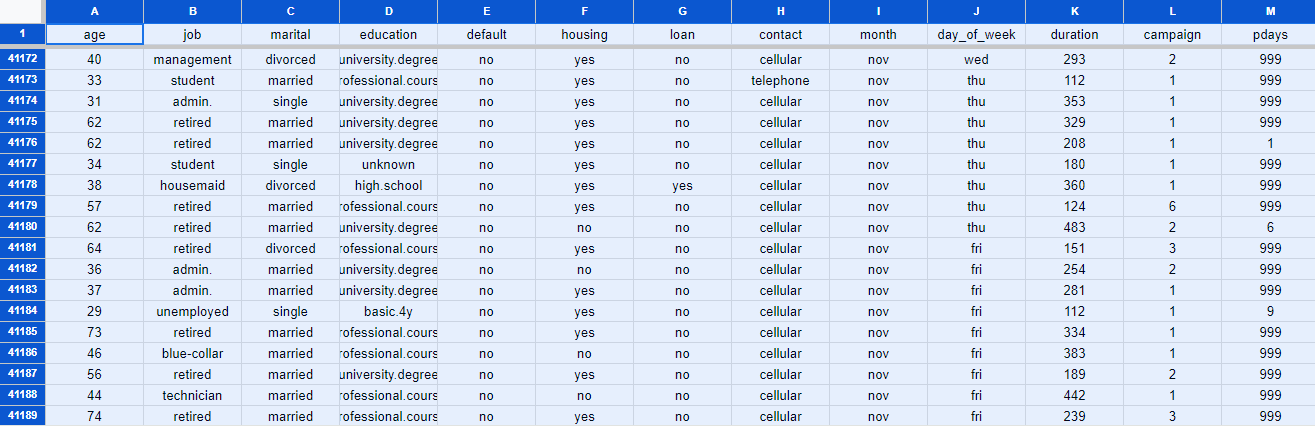
****

**Fig 5.1.1.1(a) CSV File Responses**

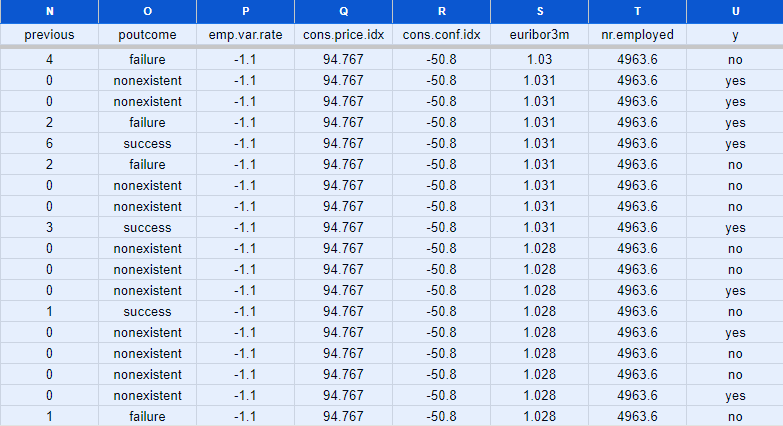
****

**Fig 5.1.1.1(b) CSV File Responses**

Fig 5.1.1.1(c) and Fig 5.1.1.1(d) represent the last 18 records of the responses collected.

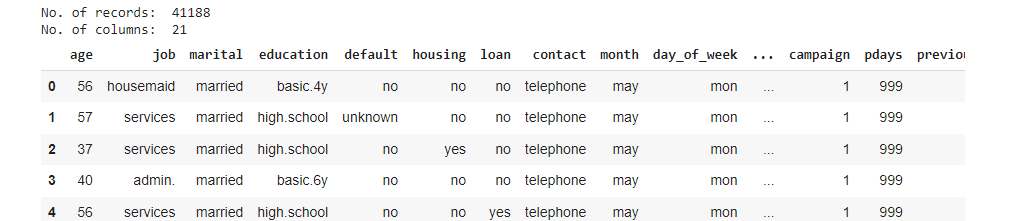


**Fig 5.1.1.1(c) CSV File Responses**

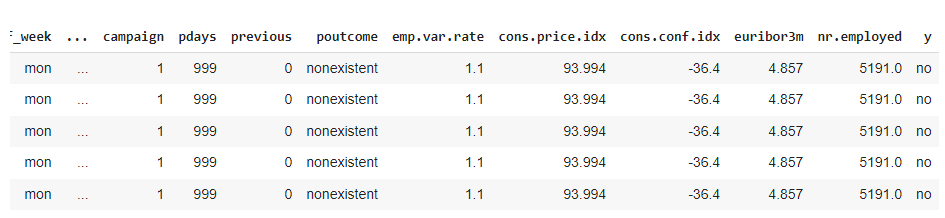
****

**Fig 5.1.1.1(d) CSV File Responses**

**Step -2 Processing of Data :** Fig 5.1.1.2(a) and Fig 5.1.1.2(b) represent the data frame after uploading it in the python notebook.



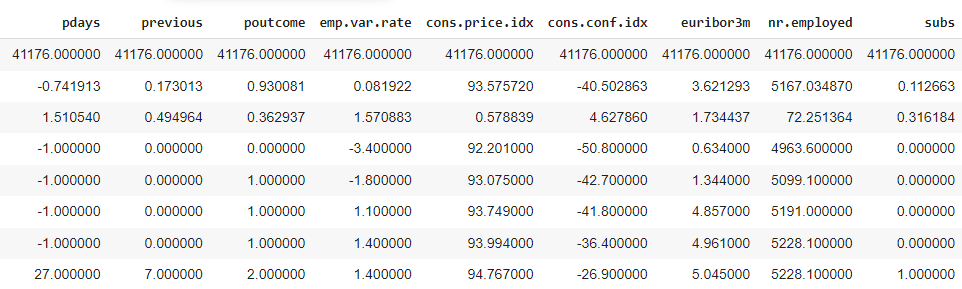
**Fig 5.1.1.2(a) Dataframe uploaded in project**

****

**Fig 5.1.1.2(b) Dataframe uploaded in project**

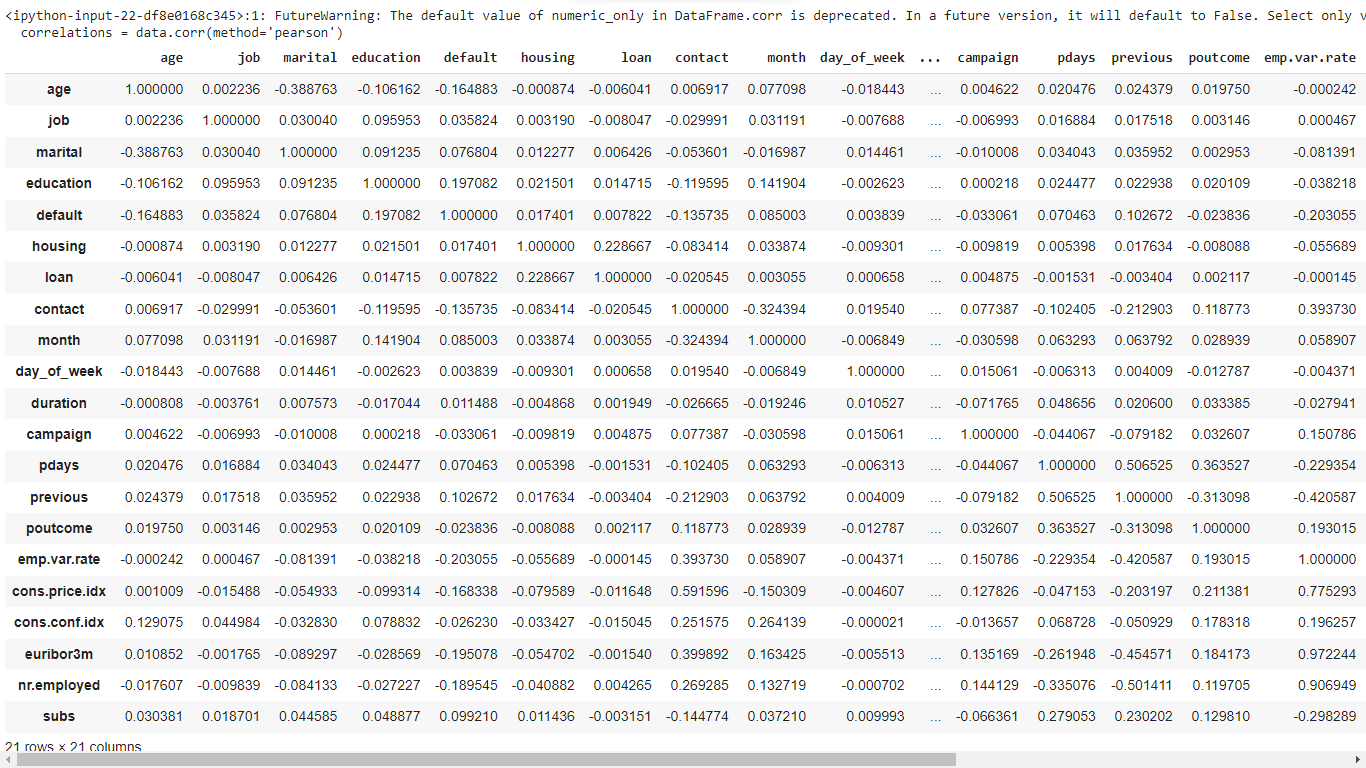
**Step -3 Describing the Dataframe :** After uploading and pre-processing of CSV file, Fig 5.1.1.3(a) and Fig 5.1.1.3(b) depicts the measures of central tendency in statistics between the attributes. We observed that the mean age of clients is around 40 and the maximum age of the client is 98 years, whereas the minimum age is 17 years. The mean value for subscription status is 0.11, meaning most clients have not subscribed yet.

**Fig 5.1.1.3(a) Describing the dataframe (Statistics)**

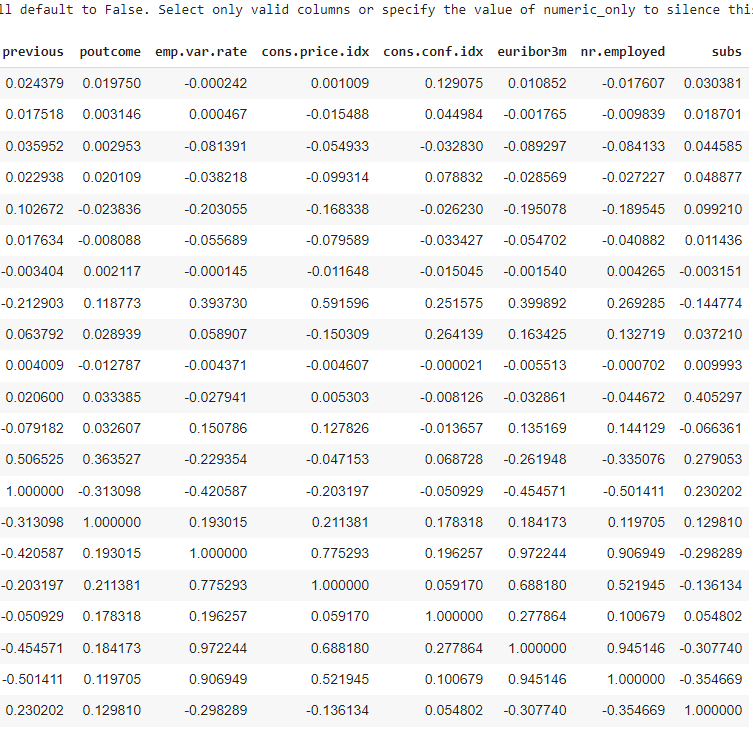
****

**Fig 5.1.1.3(b) Describing the dataframe (Statistics)**

**Step -4 Finding out Correlation of columns :** Fig 5.1.1.4(a) and Fig 5.1.1.4(b) depicts the first few records of the correlation table which has the values of coefficient of correlation (Karl Pearson’s method) and shows how the fields are correlated to each other. It is observed that there is a strong positive correlation between the emp.var.rate and euribor3m columns.

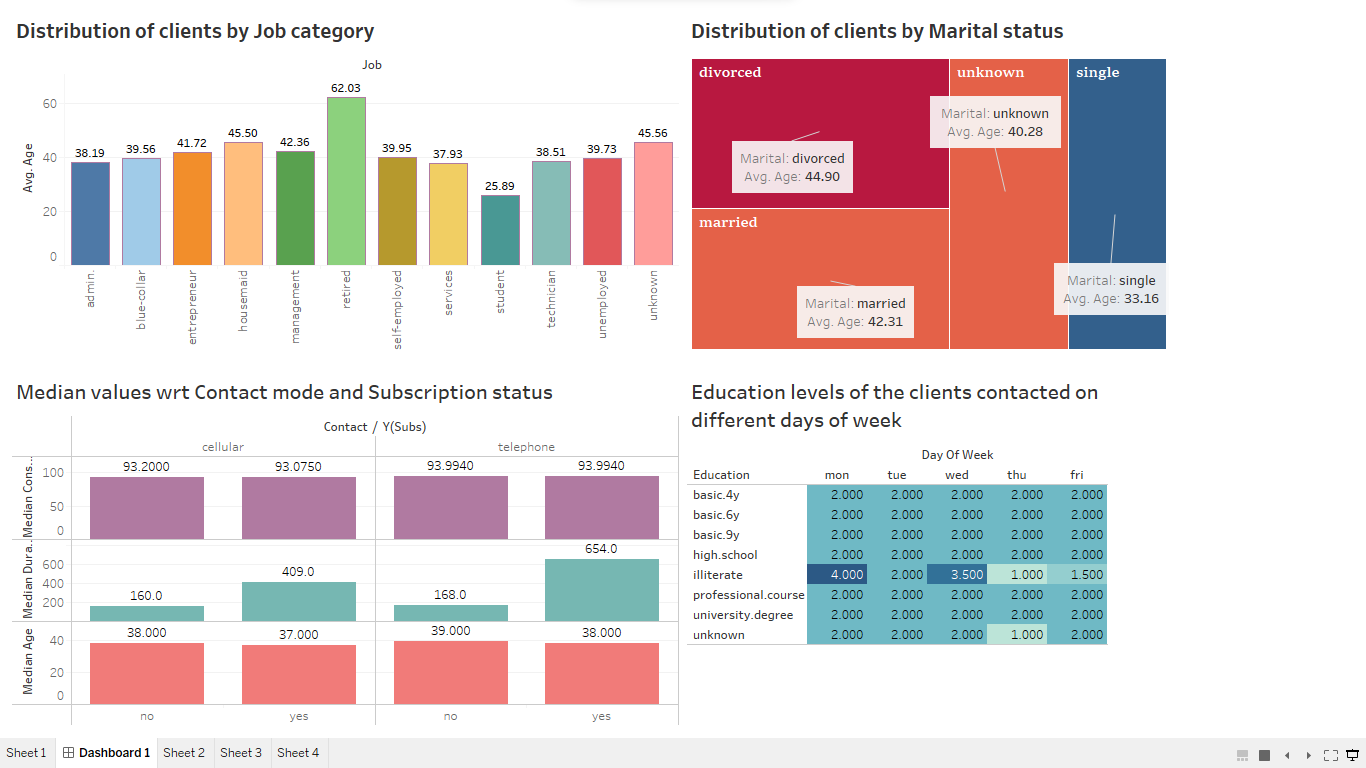


**Fig 5.1.1.4(a) Correlation table**



**Fig 5.1.1.4(b) Correlation table**

**Step - 5 Visualising data :** Fig 5.1.1.5 depicts the Tableau Data exploration platform.



**Fig 5.1.1.5 Data Dashboard representing different parameters**

**Step - 6 Generating Models and Evaluation :** After analysing graphically on Tableau, we used some classification models in python and predicted “subs” of the clients. We then evaluated the models and finally came up with the conclusion that our “Gradient Boosting” model has the highest accuracy. Also, we made a predictive model at the end and predicted the output using the user input values.

**5.1.2 Implementation Issues**

While working with the Bank marketing campaign dataset, there are several implementation issues that need to be considered. These issues may impact the practical application and utilization of the dataset for analysis and decision-making.

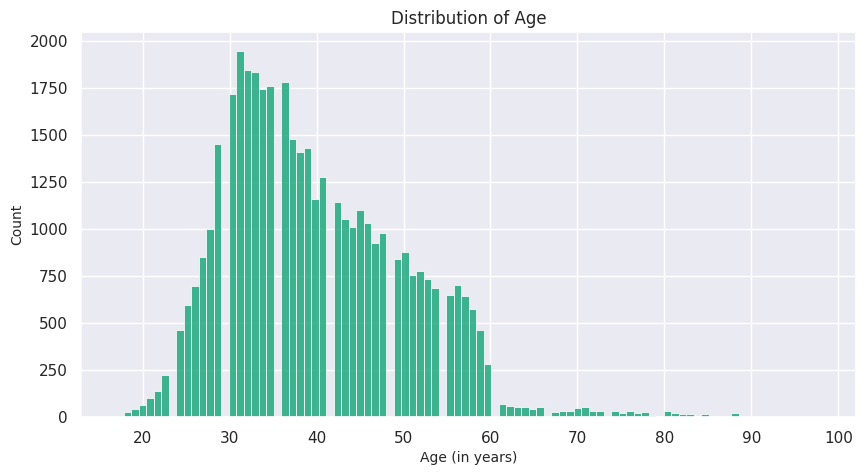
* **Data Quality and Cleansing:** The dataset may contain inconsistencies, duplicates or missing values that need to be identified and addressed through data cleansing techniques to ensure accurate and reliable analysis results.
* **Data Integration and Preprocessing:** The dataset may consist of multiple sources or formats, requiring data integration and preprocessing to consolidate and transform the data into a usable format to ensure data consistency and compatibility.
* **Data Security and Privacy:** The dataset may contain sensitive information such as customer details or order transactions that need to be protected from unauthorized access or breaches.
* **System Infrastructure and Resources:** Adequate computing resources, including processing power and storage capacity, should be available to handle the volume and complexity of the dataset.
* **Analytics Tools and Technologies:** Selecting appropriate analytics tools and technologies based on the dataset size, analysis complexity, and organizational capabilities is crucial. Consideration should be given to the availability of tools for data exploration, visualization, statistical analysis, and machine learning.

**5.2 Evaluation Parameters (for Models)**

* Test/Train split of data - 80% training and 20% testing
* Classification metrics - Confusion matrix, Accuracy, Mean Square Error and AUC-ROC curve
* Detailed study of plots and correlation heat map

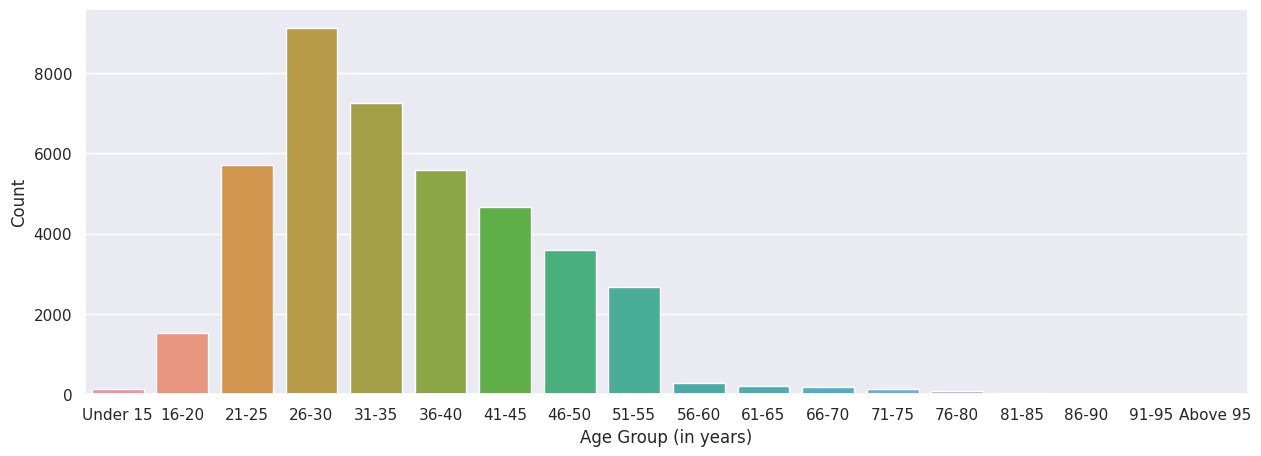
**5.3 Observations and Results**

After visualizing in the colab notebook, we came up with the following observations and results :



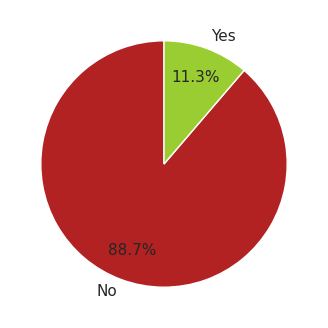
**Fig 5.3.1 Distribution of Age of clients**

Fig 5.3.1 depicts the distribution of clients’ age that were contacted by the organisation and the majority (97.5%) of the clients fall in age range of 20 - 60 years.



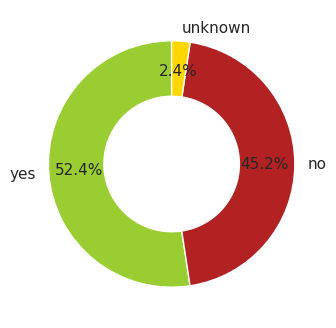
**Fig 5.3.2 Distribution of Age Group of clients**

Fig 5.3.2 represents the age distribution of clients and it was found that 8000+ contacts were made to the age group of 26-30 and the least contacted age group were ‘Under 15 years’ and ‘76-80 years’.



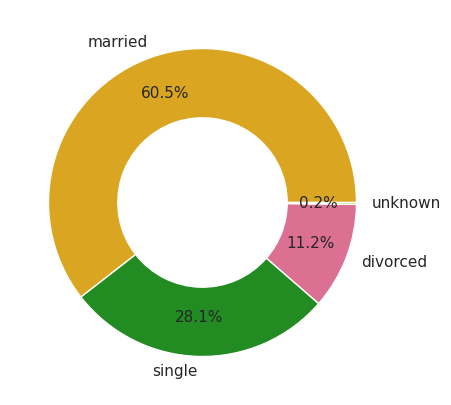
**Fig 5.3.3 Percentage of clients subscribed and unsubscribed to a term deposit**

Fig 5.3.3 depicts that 88.7% of clients were not subscribed to the term deposits and 11.3% were subscribed to the term deposits.



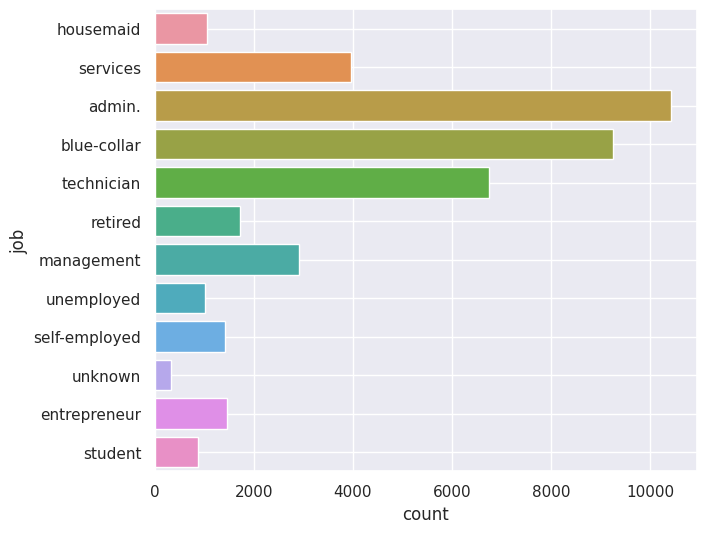
**Fig 5.3.4 Percentage of clients having housing loan or not**

Fig 5.3.4 depicts that 52.4% of clients already had the housing loan while 45.2% had not. Also, status of 2.4% of the population was not known.



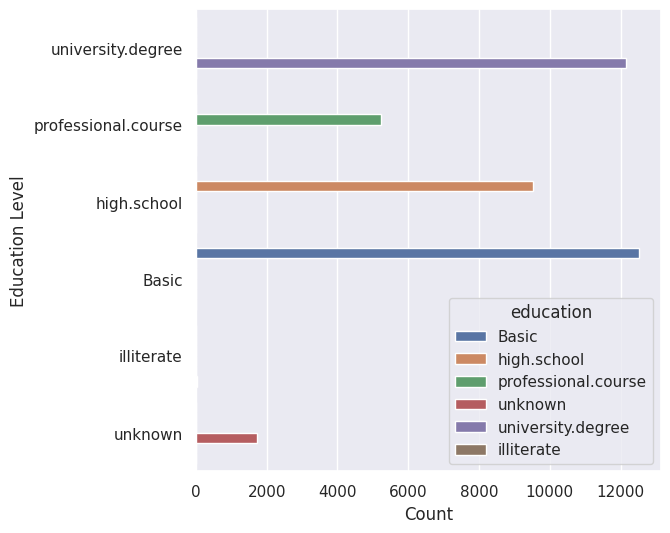
**Fig 5.3.5 Marital status of clients**

Fig 5.3.5 depicts that the majority of clients were married (60.5%) , followed by 28.1% single, 11.2% divorced and 0.2% unknown.



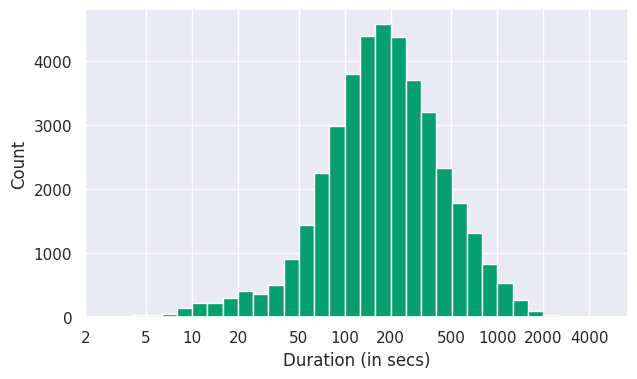
**Fig 5.3.6 Distribution of clients as per their job/profession**

Fig 5.3.6 depicts the count of clients in different job sectors. Majority of clients were doing admin. job and students were less than 2000 in the campaign.



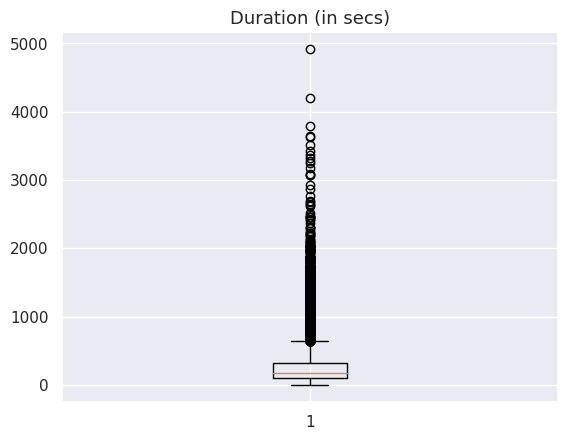
**Fig 5.3.7 Distribution of clients as per their Education level**

Fig 5.3.7 depicts that majority of the population had completed their basic education level , very few were illiterate. More than 12000 clients completed their University degree. Majority of clients have completed university degree & high school. For 1730 clients, education status is unknown.



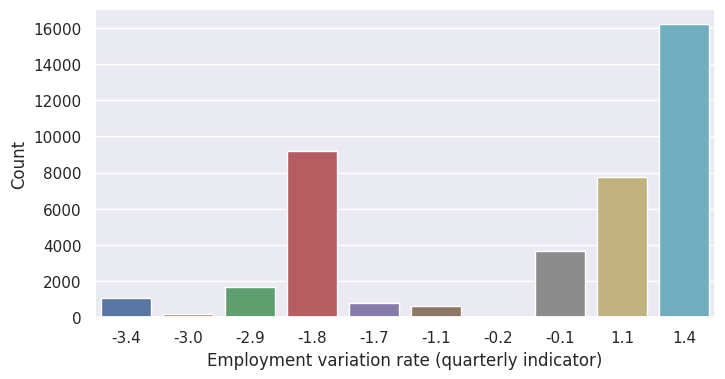
**Fig 5.3.8 Duration of contact(calls) made to the clients (on a log scale)**

Fig 5.3.8 represents the duration of calls for the campaign in seconds (on a log scale for a more simplified view). According to this plot, the call was between 100-150 seconds in most of the cases.

****

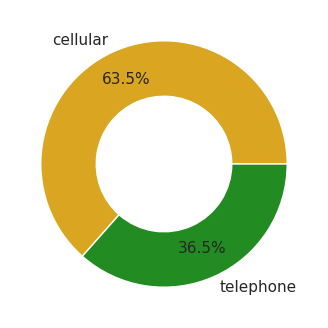
**Fig 5.3.9 Finding outliers using box plot in Duration variable**

Fig 5.3.9 depicts the process of finding outliers in the last contact duration column in the dataset. Most of the values lie between 0 to 1000.



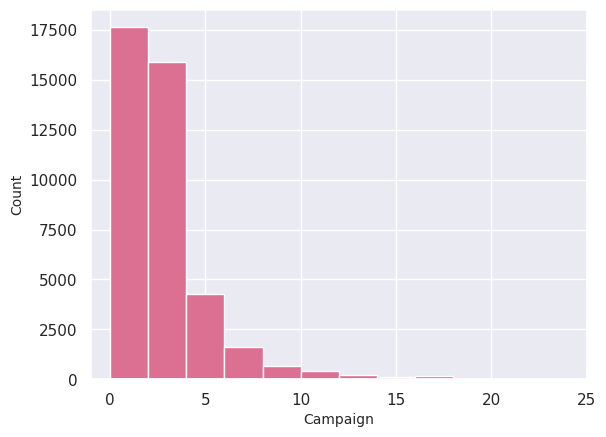
**Fig 5.3.10 Count of different Employment variation rates**

Fig 5.3.10 depicts the different employment variation rate with their counts. This was a quarterly indicator in the dataset. Majority value was 1.4 and the least were -0.2 and -3.0.

****

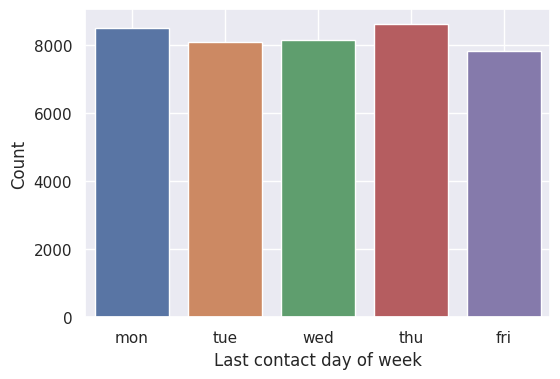
**Fig 5.3.11 Percentage of different modes of contact to clients**

Fig 5.3.11 depicts that the majority of contacts (63.5%) were made using cellular mode.



**Fig 5.3.12 Distribution of campaign days**

Fig 5.3.12 depicts the number of contacts performed during this campaign and for that client. It also included the last contact made to the client.



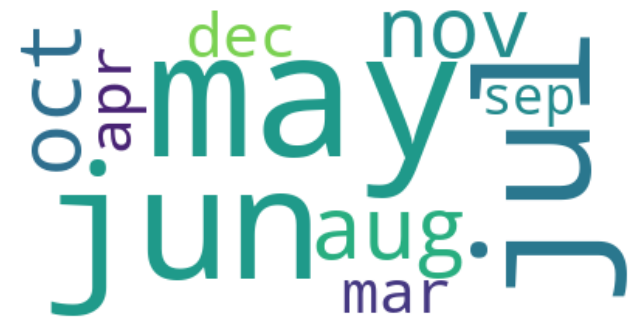
**Fig 5.3.13 Day-wise count of last contact**

Fig 5.3.13 depicts the day-wise last contact day of the week. ‘Thursday’ was the most last contact day of the week.

****

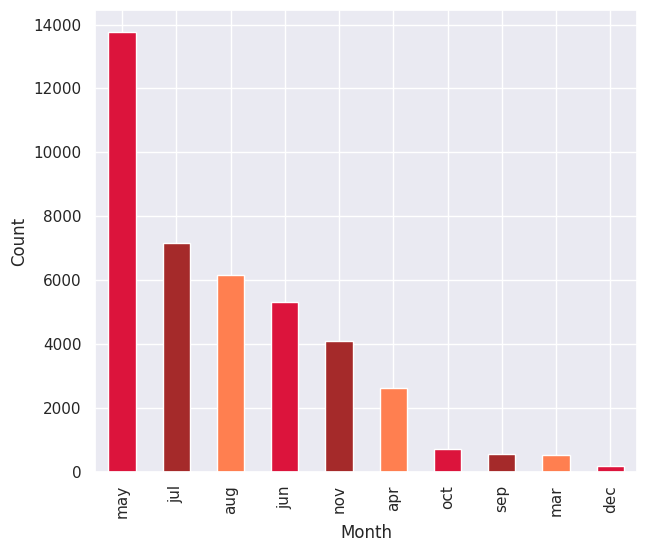
**Fig 5.3.14 Distribution of previous campaign outcome**

Fig 5.3.14 depicts the outcome of the previous marketing campaign. It was found that only 3.3% of previous outcomes were successful. Previous campaign outcome was known for 13.6% of the customers only.



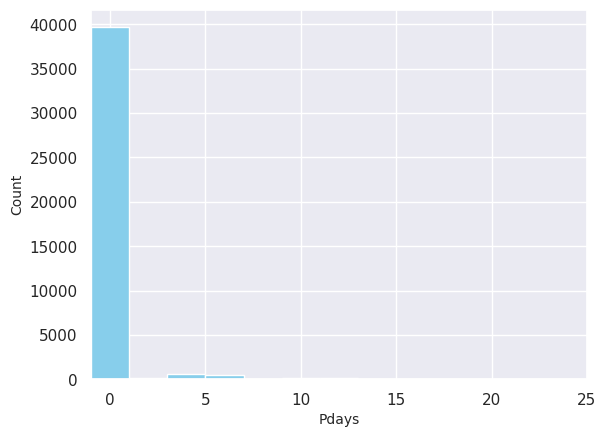
**Fig 5.3.15(a) Word cloud for Campaign months**

Fig 5.3.15(a) depicts the word cloud for the months of the campaign. It was observed that ‘may’ was the most contacted month , followed by ‘july’ and ‘august’. ‘December’ was the least contacted month with the clients.



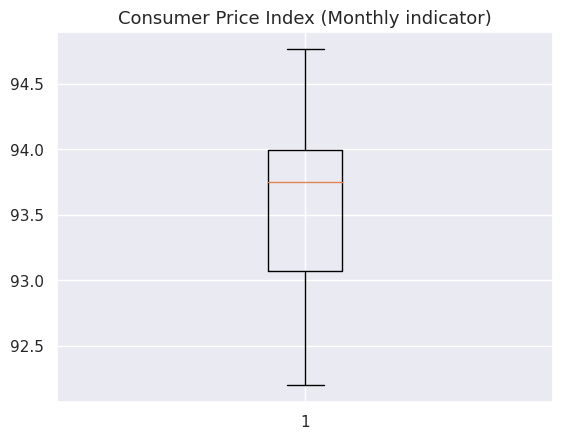
**Fig 5.3.15(b) Count of calls in different months of campaign**

Fig 5.3.15(b) depicts the count of calls observed that ‘may’ was the most contacted month with a count of 13767 , followed by ‘july’ with the count of 7169 and ‘august’ with the count of 6176. ‘December’ was the least contacted month with the clients with the count of 182.



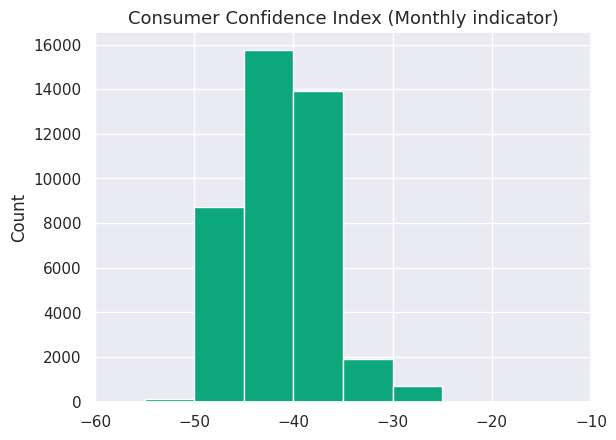
**Fig 5.3.16 Distribution of days from previous outcome to this campaign**

Fig 5.3.16 depicts the number of days that passed by after the client was last contacted from a previous campaign. It was observed that the majority of clients were not previously contacted.



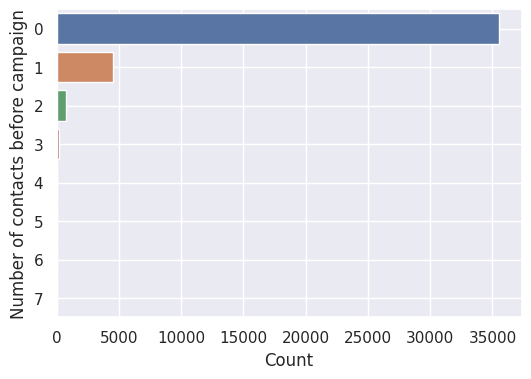
**Fig 5.3.17 Box plot of Consumer Price Index**

Fig 5.3.17 depicts minimum and maximum values for Consumer Price Index were 92.756 and 93.994. The median value was 93.749.



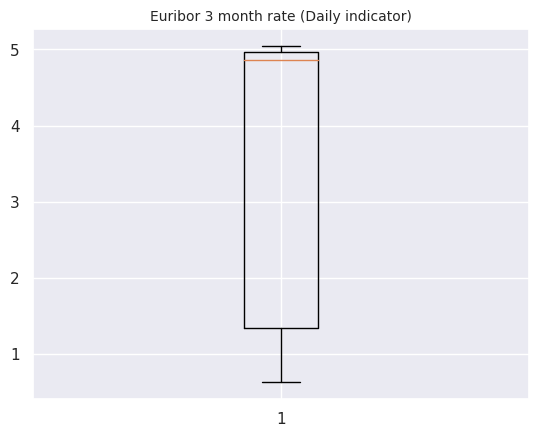
**Fig 5.3.18 Distribution of Consumer Confidence Index**

Fig 5.3.18 depicts that the maximum consumer confidence index was -36.4 with 7762 counts and minimum value was -45.9 with 10 counts. The mean value for this was -40.50.

****

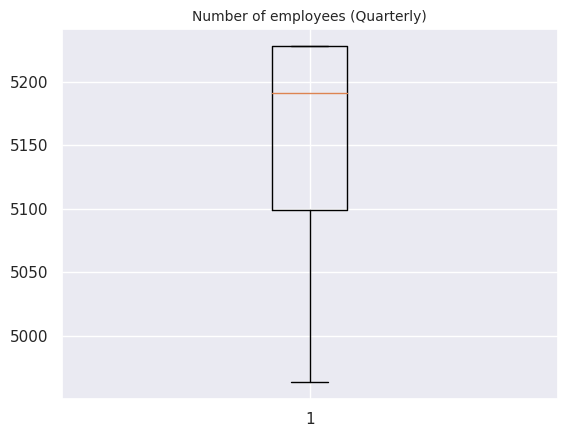
**Fig 5.3.19 Count of number of contacts before this campaign**

Fig 5.3.19 depicts the distribution of the number of contacts performed before this campaign and for that client. It was 0 for the majority of population, i.e.,there were no contacts before this campaign for the clients.



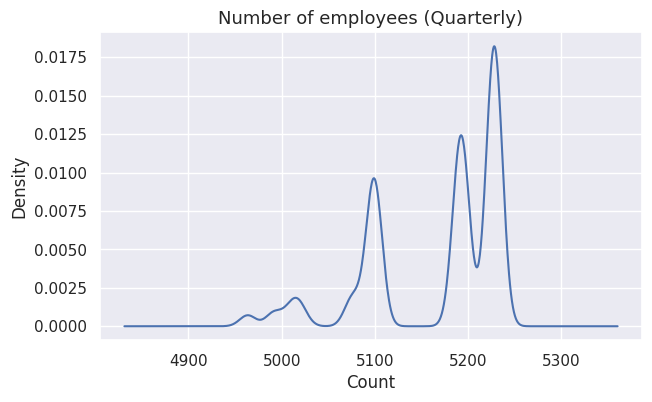
**Fig 5.3.20 Box plot representing Euribor 3 month rate**

Fig 5.3.20 depicts euribor 3 month rate which was a daily indicator. The mean value for this parameter was 3.621. The maximum value is near 5 and median is above 4.55.



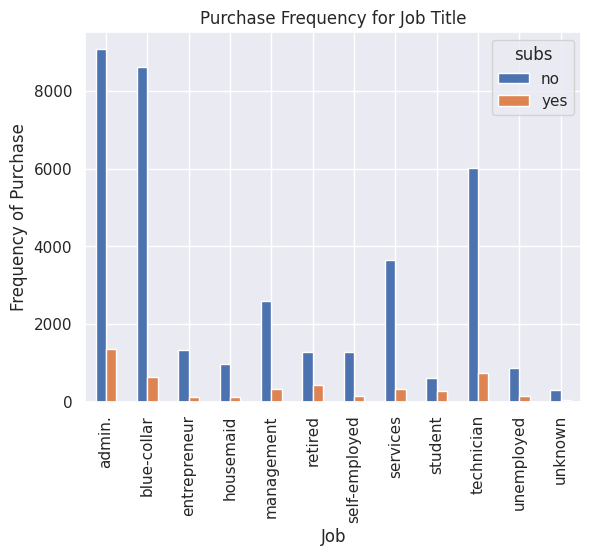
**Fig 5.3.21(a) Box plot for number of employees**

Fig 5.3.21(a) depicts the number of employees on a quarter basis. The mean number of employees were 5167 and the majority of the number of employees were around 5228 with 16228 count.



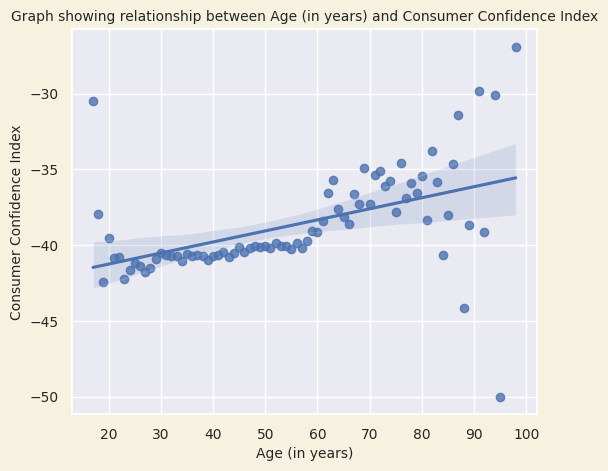
**Fig 5.3.21(b) Density plot for number of employees**

Fig 5.3.21(b) depicts the density plot for the number of employees. There is a high density around the range 5200-5300.

****

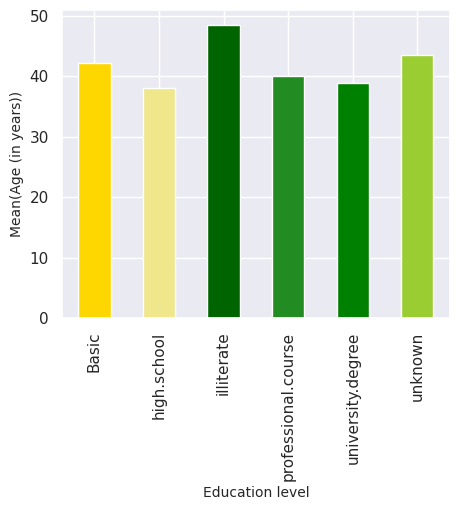
**Fig 5.3.22 Purchase frequency for different jobs**

Fig 5.3.22 depicts the purchasing pattern of term deposits according to the type of jobs that the clients were doing.



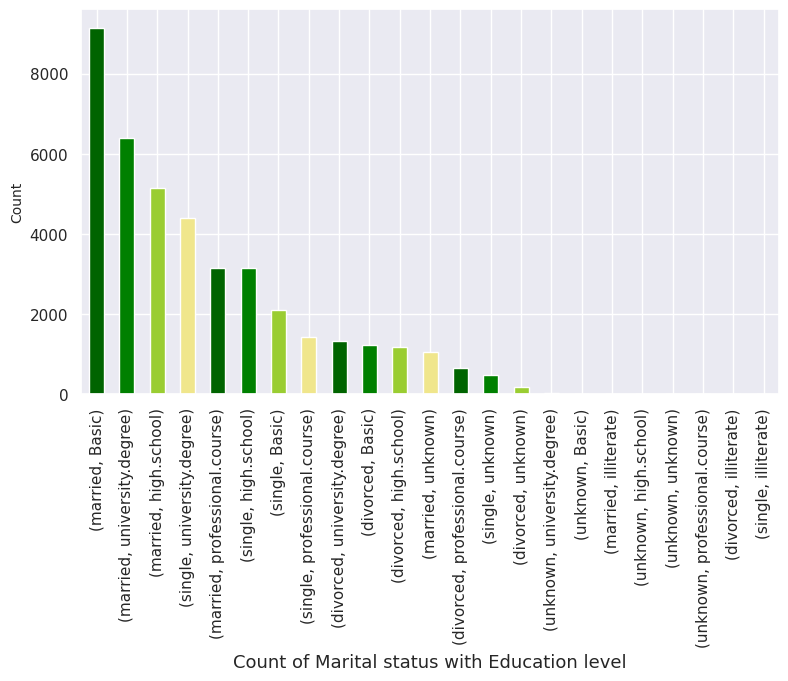
**Fig 5.3.23 Relationship between Age (in years) and Consumer Confidence Index**

Fig 5.3.23 depicts the graph showing a positive correlation between age and consumer confidence index, meaning that as people get older, they tend to be more confident in the economy. Consumers aged 50-64 had the highest levels of consumer confidence and consumers aged 25-34 had the lowest levels of consumer confidence. Consumer confidence tends to increase with age, up to a certain point. After age 65, consumer confidence began to decline slightly.

****

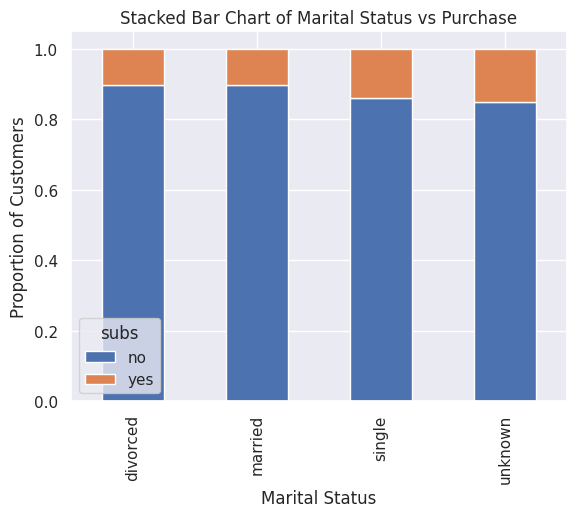
**Fig 5.3.24 Mean age for different Education levels**

Fig 5.3.24 provides a valuable snapshot of the educational attainment of the population. This information can be used by governments, businesses, and other stakeholders to develop policies and programs that promote education and lifelong learning.



**Fig 5.3.25 Count of Marital status with different Education levels**

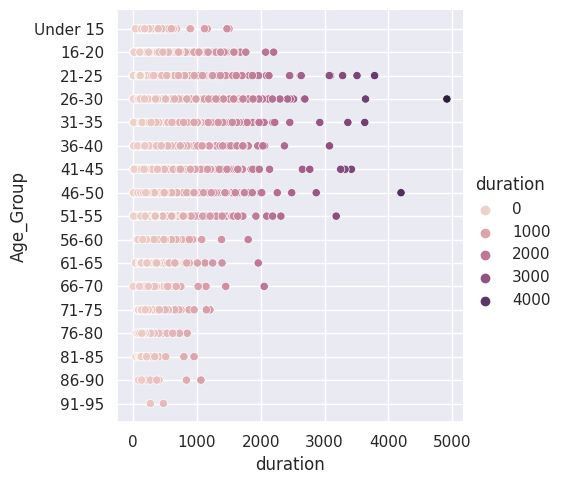
Fig 5.3.25 depicts that married people were more likely to have a high school education than people with other marital statuses, while single people were more likely to have a professional course or university degree than people with other marital statuses. Divorced people are more likely to have a basic education than others.



**Fig 5.3.26 Stacked bar chart of Marital status with their purchasing status**

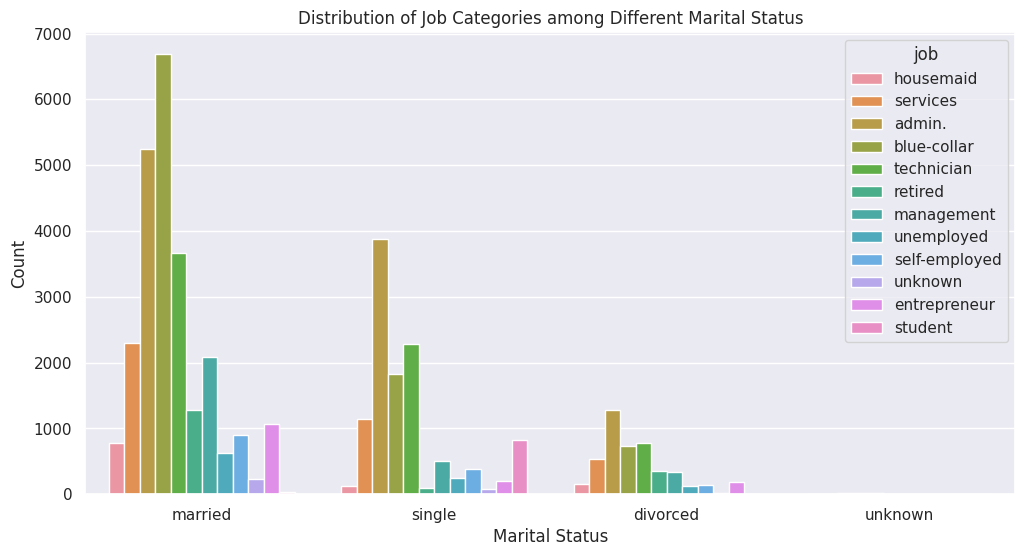
Fig 5.3.26 depicts that single people were the most likely to make purchases, followed by married people, divorced people and unknown people. The proportion of married people making purchases was highest for subscriptions, followed by one-time purchases. The proportion of single people making purchases wass highest for one-time purchases, followed by subscriptions.

.



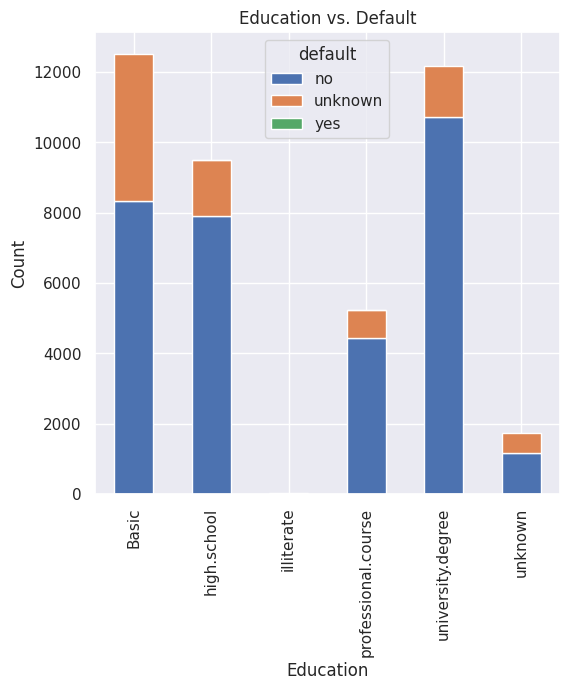
**Fig 5.3.27 Distribution of duration in different age groups**

Fig 5.3.27 depicts that positive correlation between duration and Age group. This means that as the duration increases, the age group also increases up to a certain age. The relationship is not linear, meaning that the rate of increase in Age group slows down as the duration increases.

****

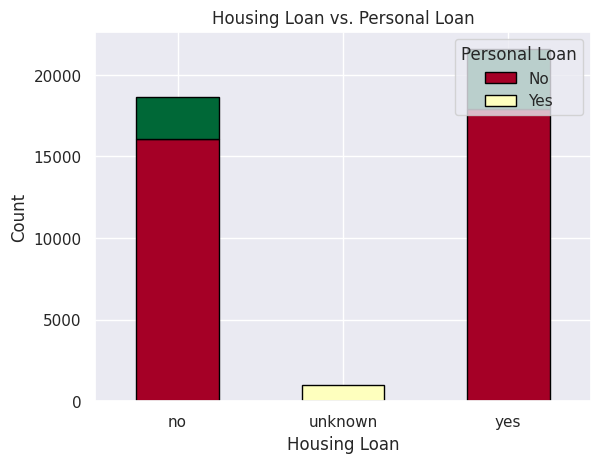
**Fig 5.3.28 Distribution of Job Categories among Different Marital Status**

Fig 5.3.28 depicts the observations that while married individuals were most likely to hold managerial positions, single individuals predominantly worked in service roles. Interestingly, both divorced and widowed individuals primarily hold service jobs, with retired and unemployed positions following respectively.

****

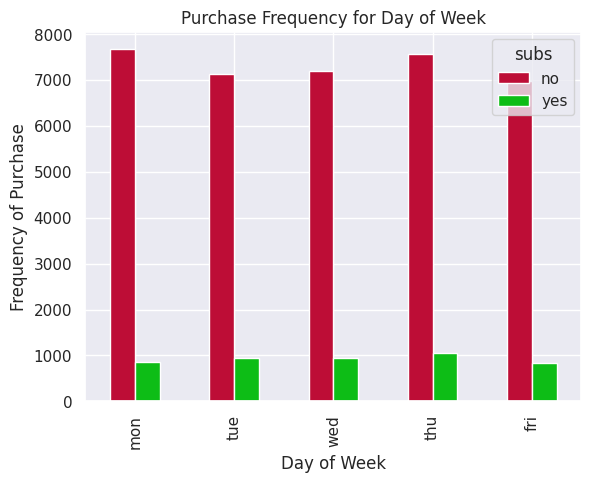
**Fig 5.3.29 Cross-tabulation of Education and Default loan status**

Fig 5.3.29 depicts the cross-tabulation of education and default, revealing a negative correlation between the two variables. This means that people with higher levels of education were less likely to get default loans.

****

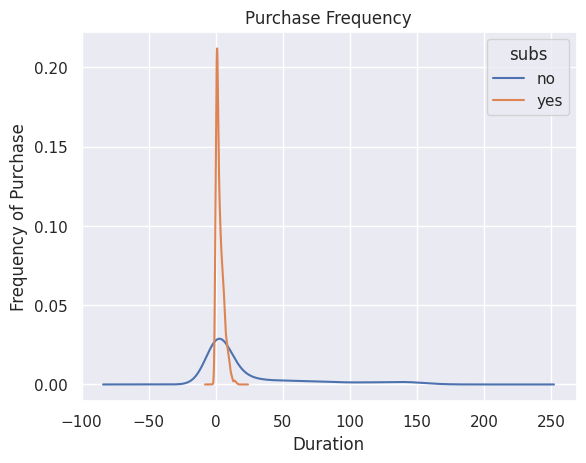
**Fig 5.3.30 Cross-tabulation of Housing and Personal loan status**

Fig 5.3.30 depicts the cross-tabulation of Housing and Personal loan, showing that people with housing loans were more likely to have personal loans as well.



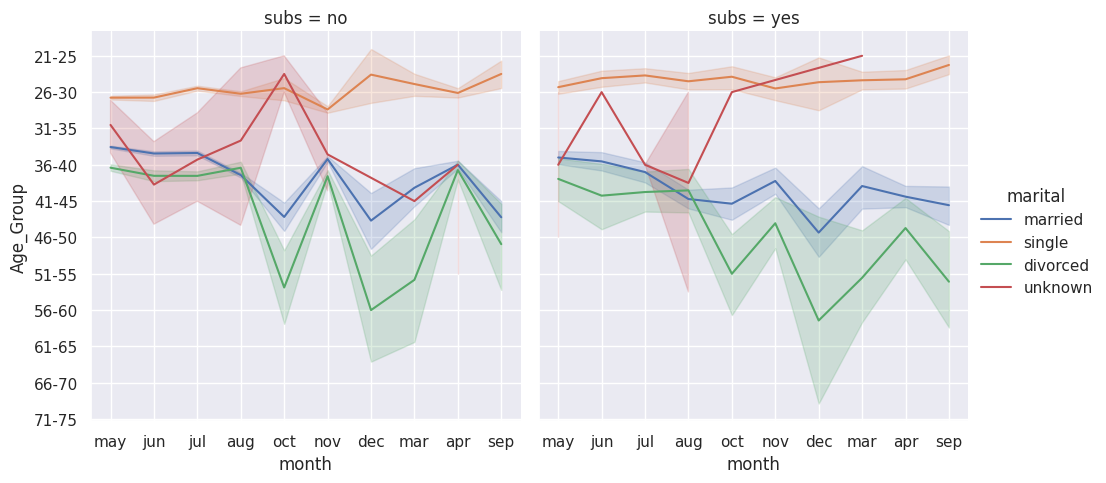
**Fig 5.3.31(a) Purchase frequency for different days of week**

Fig 5.3.31(a) depicts the purchase frequency of clients and it was highest on Monday and Tuesday, and lowest on friday. This trend was likely due to a number of factors, including the fact that people are more likely to have interest after weekends.

****

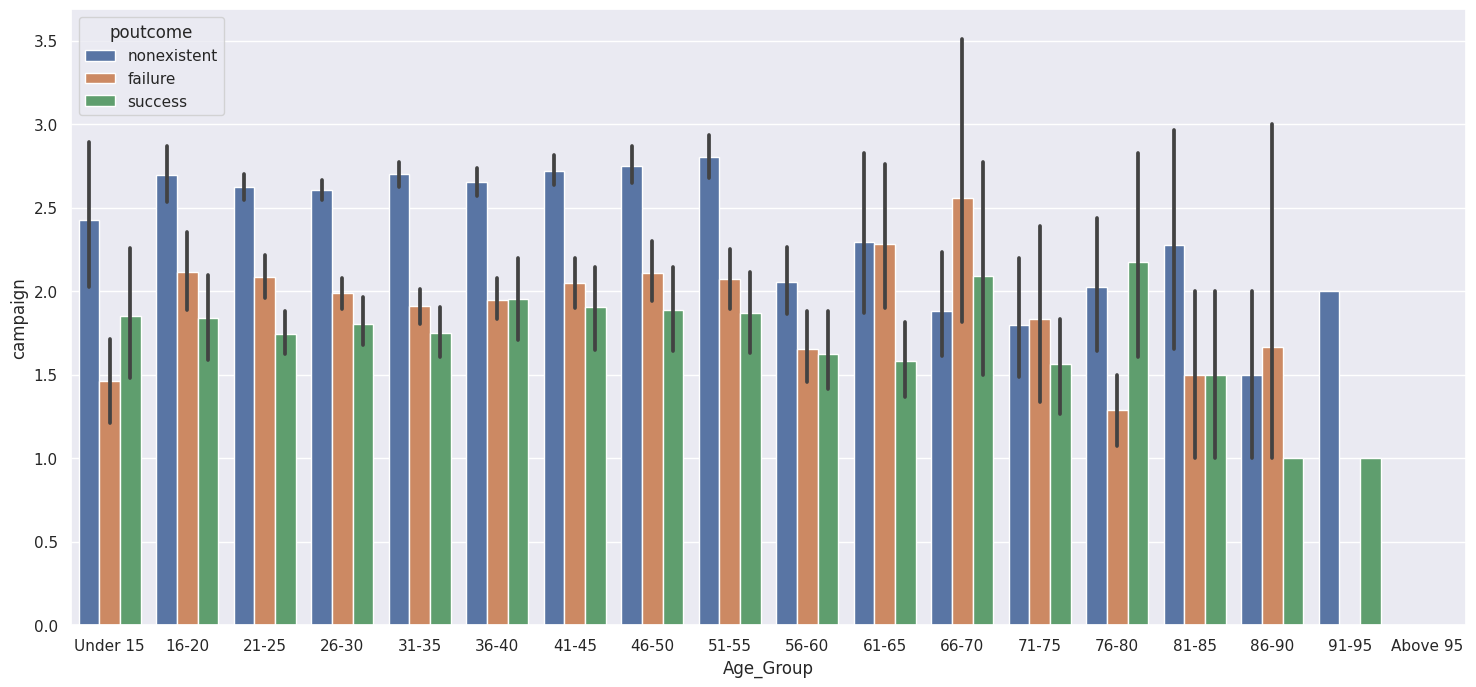
**Fig 5.3.31(b) Density plot for purchase frequency with respect to Duration**

Fig 5.3.31(b) depicts the density plot showing the purchase frequency for different durations. This trend may be due to a number of factors, for example, people are more likely to try out a shorter subscription before committing to a longer subscription.

****

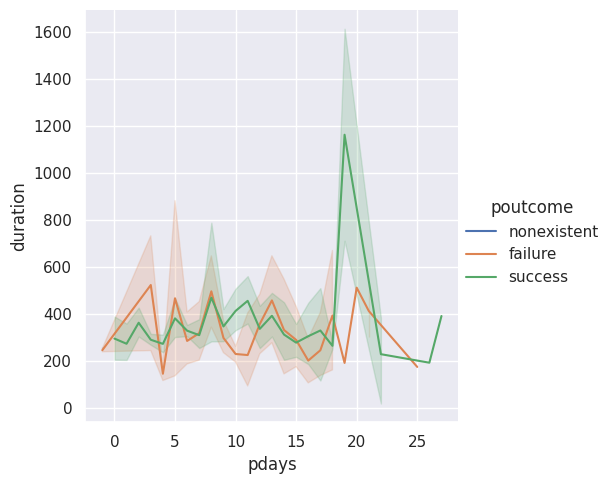
**Fig 5.3.32 Line plot representing relationship between month, Age group and their subscription status**

Fig 5.3.32 depicts the subscription rate for the service was highest among married people and lowest among divorced people. The subscription rate was also highest in the summer months and lowest in the fall months.

****

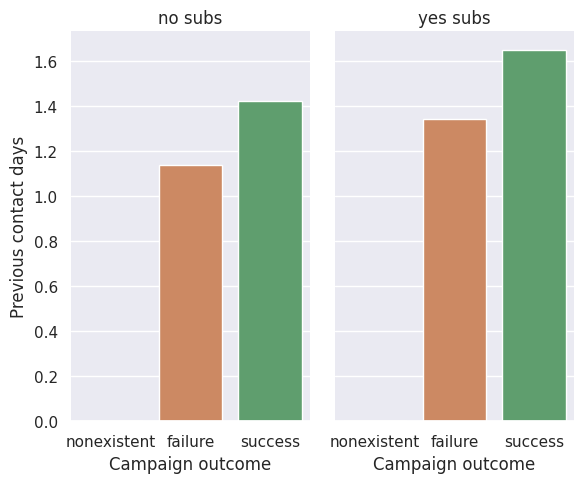
**Fig 5.3.33 Plot showing relationship between Age group, Campaign contacts and outcomes of previous campaign**

Fig 5.3.33 depicts the bar plot showing that the successful campaign was with the age group 76-80 whereas most of the age groups had non-existent previous campaign outcome and those groups were contacted this time.

****

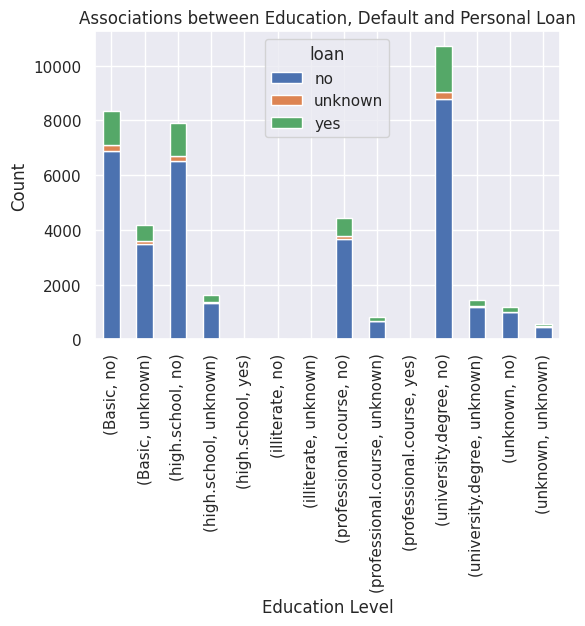
**Fig 5.3.34 Line plot showing relationship between pdays, duration of contact according to previous campaign outcome**

Fig 5.3.34 depicts the number of failures and successes in a project over time, based on the duration of the project and the number of pdays and suggests that the project team should focus on reducing the duration and complexity of projects in order to improve the success rate.



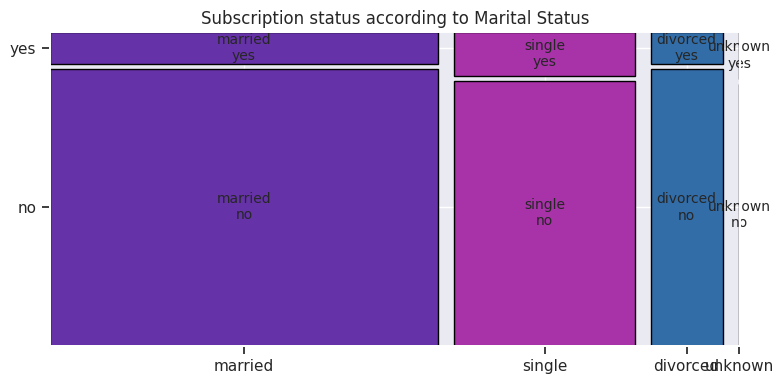
**Fig 5.3.35 Plot showing relationship between Previous contact days and subscription status according to previous campaign outcomes**

Fig 5.3.35 depicts that having subscribers was a positive factor for campaign success. This was likely because subscribers were more likely to be engaged with the campaign and more likely to take the desired action.

****

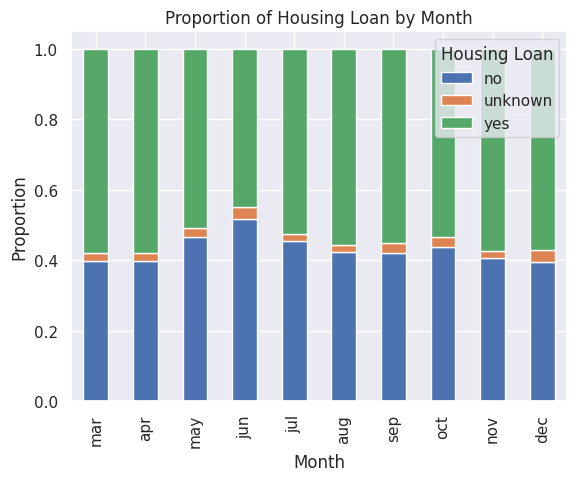
**Fig 5.3.36 Plot showing relationship between Education, Personal loan and Default credit status**

Fig 5.3.36 depicts that education was an important factor in determining the likelihood of defaulting on a personal loan. People with higher levels of education were less likely to take loans by default, while people with high school degrees and professional courses were more likely to.

****

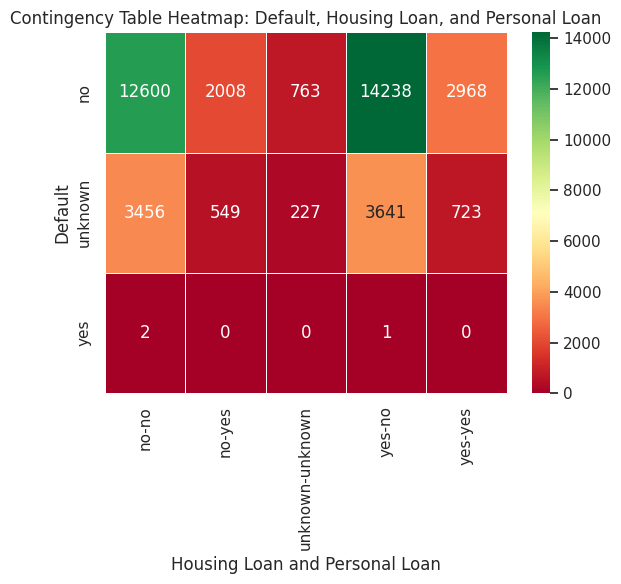
**Fig 5.3.37 Mosaic plot showing Subscription status according to Marital status**

Fig 5.3.37 depicts the subscription status of customers according to their marital status. Married people were more likely to subscribe than single people and divorced people were less likely to subscribe than married or single people. The unknown marital status category had the lowest subscription rate.

****

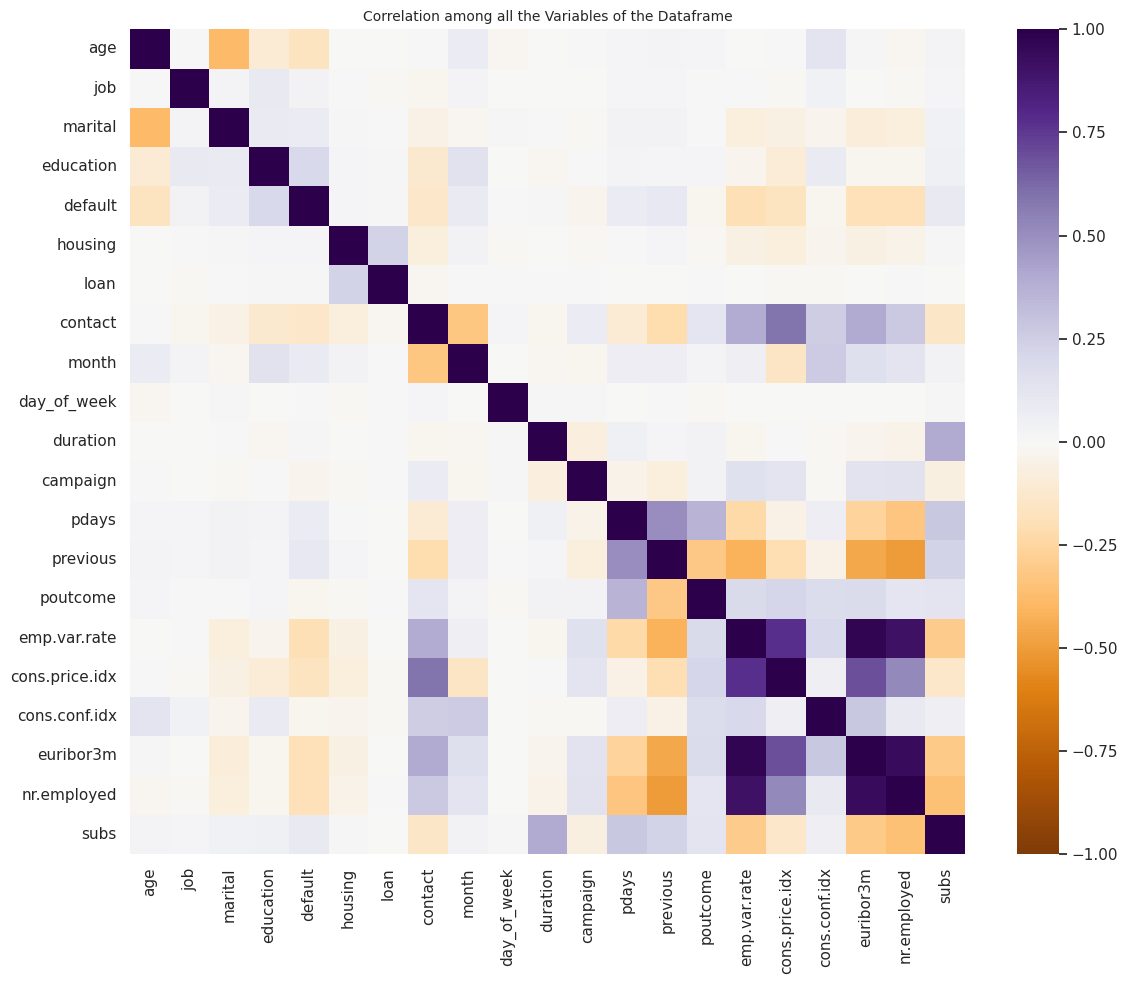
**Fig 5.3.38 Month-wise proportion of Housing Loan by clients**

Fig 5.3.38 depicts the proportion of housing loan varies throughout the year, with the highest proportion occurring in March and April. There were a number of factors that could explain this variation, such as the start of the new financial year, the availability of tax benefits, the onset of the monsoon season, and the increased likelihood of financial difficulties during the end of the year.

****

**Fig 5.3.39 Heatmap between Housing , Personal and Default loan status**

Fig 5.3.39 depicts that there was a high probability that when default status of loan was ‘no’ , the housing loan status was ‘yes’ whereas personal loan status was ‘no’. Around 3641 entries were labelled as unknown in Default loan with a ‘yes’ in housing loan and ‘no’ in personal loan.

****

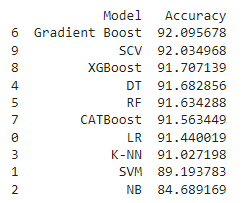
**Fig 5.3.40 Correlation heat map between all numerical fields of dataframe**

Fig 5.3.40 depicts that there is a positive correlation between the following fields :

1. emp.var.rate and nr.employed
2. emp.var.rate and cons.price.idx
3. emp.var.rate and euribor3m (0.97)
4. cons.price.idx and euribor3m
5. nr.employed and euribor3m
6. housing and loan
7. cons.price.idx and contact
8. duration and subs

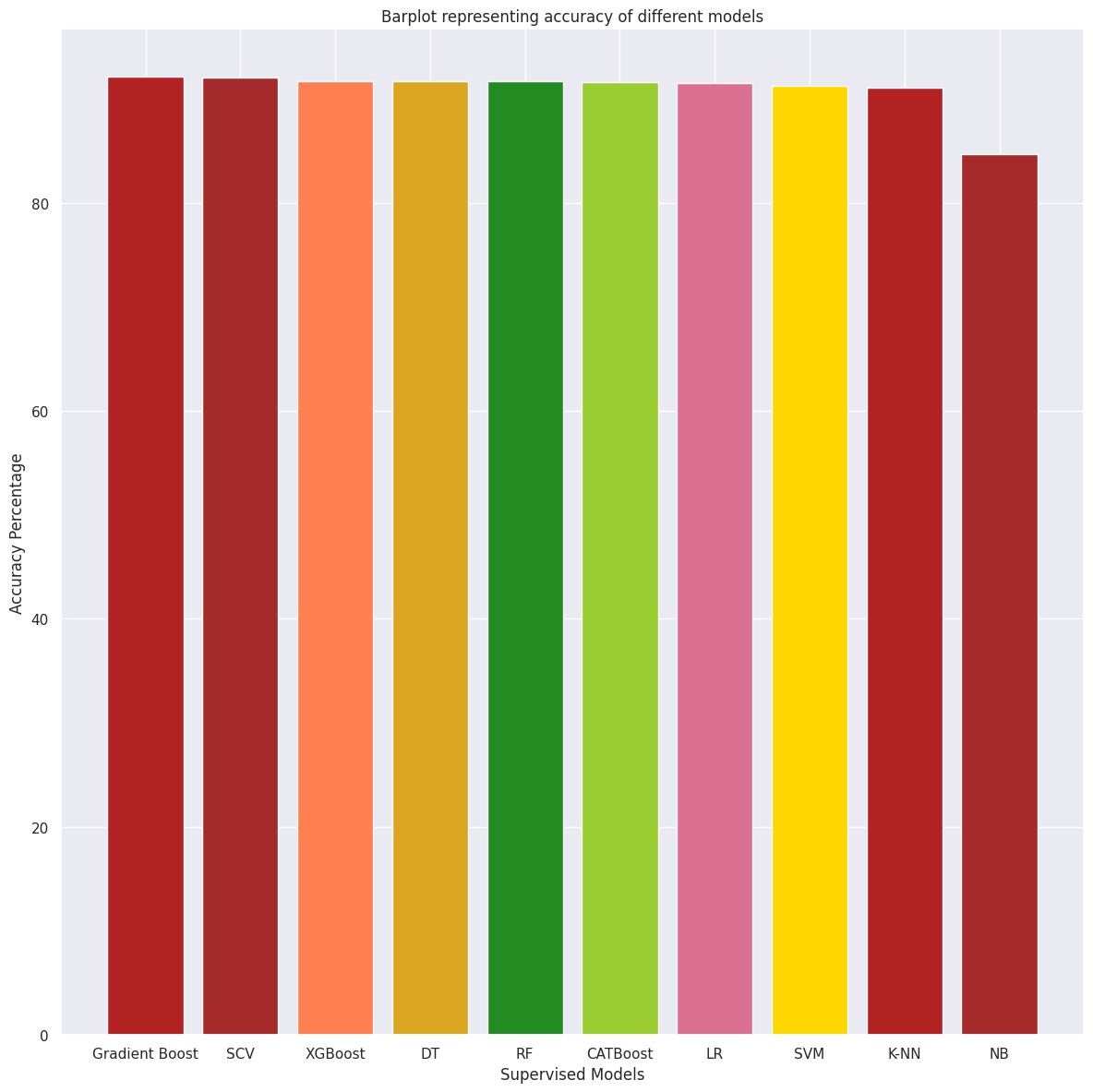
A slight correlation between default and education, month and education, previous and default, outcome and contact, contact and nr.employed, contact and campaign, default and previous, pdays and subs.

There is a negative correlation seen in the fields : marital and age, month and contact, emp.var.rate and previous, nr.employed and previous, age and default, euribor3m and previous.

****

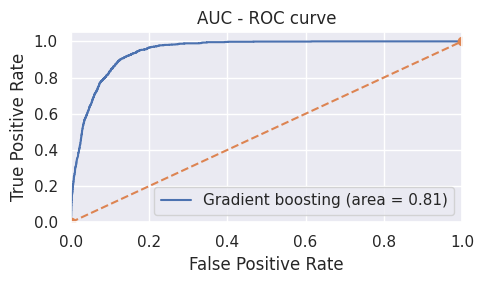
**Fig 5.3.41(a) Accuracy percentage for different supervised models in order**

Fig 5.3.41(a) depicts the accuracy percentage for different supervised algorithms deployed in the project. Gradient boost has the highest accuracy with 92.095% and Naive Bayes has the lowest accuracy with 84.689%.



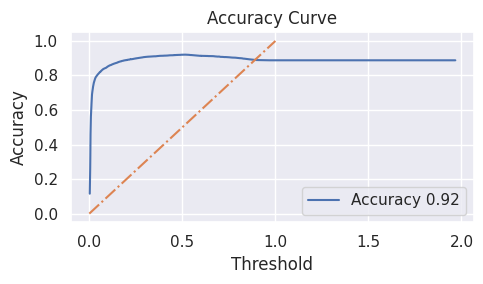
**Fig 5.3.41(b) Plot of accuracy percentage for different supervised models**

Fig 5.3.41(b) depicts the accuracy of different supervised classification models. The Gradient Boost model has the highest accuracy (92.09%), followed by the Stacking CV Classifier with 92.04% accuracy and the Naive Bayes model has the least accuracy (84.69%).



**Fig 5.3.42(a) AUC ROC curve for the highest accuracy model (Gradient Boosting)**

Fig 5.3.42(a) depicts the AUC ROC curve and area under the curve was found to be 0.81.



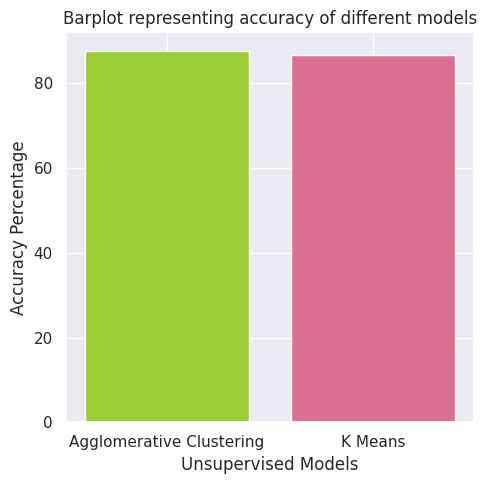
**Fig 5.3.42(b) Accuracy curve for the highest accuracy model (Gradient Boosting)**

Fig 5.3.42(b) depicts the Accuracy curve pattern for the Gradient Boosting algorithm.



**Fig 5.3.43 (a) Accuracy percentage for different unsupervised models**

Fig 5.3.43(a) depicts the accuracy percentage for different unsupervised algorithms deployed in the project. Agglomerative Clustering has 87.63% accuracy followed by K means with 86.59% accuracy.

****

**Fig 5.3.43 (b) Plot of accuracy percentage for different unsupervised models**

Fig 5.3.43(b) depicts the accuracy of different unsupervised classification models. The Agglomerative clustering model has the highest accuracy (87.63%) and the K means model has the least accuracy (86.59%).

This research focused on analyzing the Bank Campaign marketing dataset to identify the impacts of different aspects on sales of term deposits schemes and to understand customer behavior trends. It involved tools like Tableau for data visualization and Python with ML algorithms for deploying machine learning models and more visualizations.With the help of this project, Portugal Bank will have practical insights on how to better target future marketing campaigns, increase subscription rates, and comprehend customer engagement dynamics. Data-driven decision-making is made possible by thorough analysis and predictive modeling.

**CHAPTER 6**

**CONCLUSION AND FUTURE WORK**

**6.1 Conclusion**

The completion of the Portugal Bank marketing campaign analysis project has provided valuable insights into customer behavior, campaign performance, and the factors influencing subscription to term deposits. The extensive exploration and evaluation of the dataset involved various data processing, visualization, and machine learning techniques. The following key findings and conclusions emerge from the project:

**6.1.1 Demographic Insights**

* The concentration of clients within the 20 to 60 years age range suggests that the bank's marketing efforts effectively target the economically active population. Further exploration into subgroups within this range could unveil nuanced preferences and behaviors.
* The extensive outreach to the age group of 26-30 implies a strategic focus on a younger demographic, possibly due to their potential as long-term clients.
* Marital status analysis not only indicates the prevalent status but also hints at potential differences in financial needs and decision-making dynamics across marital categories.

**6.1.2 Campaign and Subscription Trends**

* The overwhelming non-subscription rate underscores the challenges in convincing clients to commit to term deposits. A deeper dive into unsuccessful campaigns could unveil patterns or areas requiring refinement.
* The popularity of cellular communication for contacts suggests the significance of mobile platforms in contemporary marketing strategies.
* The low success rate from the previous campaign emphasizes the necessity of reevaluating and optimizing outreach strategies.

**6.1.3 Financial and Economic Indicators**

* The examination of economic indicators provides a contextual backdrop for campaign performance. Understanding how external factors impact customer behavior is crucial for strategic planning.
* Employment variation rates and the Euribor 3-month rate emerge as pivotal financial indicators affecting customer decisions. A more granular analysis of these indicators may offer insights into their specific influence.

**6.1.4 Machine Learning Model Insights**

* The selection of Gradient Boosting as the most accurate supervised model indicates the effectiveness of ensemble methods in predicting subscription outcomes with an accuracy of 92.09%. Fine-tuning parameters and experimenting with alternative models could potentially enhance predictive capabilities.
* Agglomerative Clustering achieved the highest accuracy (87.63%) among unsupervised models. Agglomerative Clustering's success suggests discernible patterns within the dataset. Further exploration of cluster characteristics and their alignment with demographic or behavioral traits could be enlightening.

**6.1.5 Project Challenges and Considerations**

* The project's success hinges on the meticulous handling of data quality issues. Establishing robust data cleansing protocols and constant vigilance for anomalies are imperative.
* The choice of analytics tools and technologies played a pivotal role. Regular evaluations of emerging technologies ensure the project remains at the forefront of analytical capabilities.
* Ongoing considerations for data security, privacy, and infrastructure readiness are non-negotiable components of successful data analytics projects.

In conclusion, this project equips the Portugal Bank with actionable insights to optimize future marketing campaigns, improve subscription rates, and understand the dynamics of customer engagement. The comprehensive analysis, along with the predictive modeling, lays the foundation for data-driven decision-making. As the financial landscape evolves, continuous monitoring and adaptation of strategies will be essential for sustained success.

**6.2 Future Work**

* Enhanced predictive modeling can benefit from feature engineering strategies, such as incorporating temporal trends, sentiment analysis from customer interactions, and integrating external datasets for a more holistic view.
* A deeper dive into income-based segmentation could provide tailored insights for personalized marketing strategies.
* Real-time data integration and continuous monitoring mechanisms should be implemented to keep the models robust and adaptable to evolving customer behaviors.

**CHAPTER 7**

**REFERENCES**

[1] ["About Python"](https://www.python.org/about). Python Software Foundation. [Archived](https://web.archive.org/web/20120420010049/http://www.python.org/about/) from the original on 20 April 2012. Retrieved 24 April 2012., second section "Fans of Python use the phrase "batteries included" to describe the standard library, which covers everything from asynchronous processing to zip files." Available : https://www.python.org/about

[2] Jay Selig. “What is Machine Learning? A definition. “ . Published 14 March 2022. [Online]. Available: https://www.expert.ai/blog/machine-learning-definition/

[3] Tableau from Salesforce, “What is Tableau?,”. Last edited in 2023. [Online]. Available: https://www.tableau.com/why-tableau/what-is-tableau.

[4] Google. “Colaboratory Frequently Asked Questions”. [Online]. Available: https://research.google.com/colaboratory/faq.html

[5] Shankar, Venky. “Big Data and Analytics in Retailing”. Published : 2019. NIM Marketing Intelligence Review. 11. 37-40. DOI : 10.2478/nimmir-2019-0006.

[6] Chandramana, Sudeep. “Retail Analytics: Driving Success in Retail Industry with Business Analytics”. Published : 2017. Volume 7. DOI : 10.6084/m9.figshare.13323179.

[7] Marnik G. Dekimpe. “Retailing and retailing research in the age of big data analytics”. International Journal of Research in Marketing, Volume 37, Issue 1, Published : 2020. Pages 3-14, ISSN 0167-8116. Available : https://doi.org/10.1016/j.ijresmar.2019.09.001.

[8] Elgendy, Nada & Elragal, Ahmed.” Big Data Analytics: A Literature Review Paper”.Published : 2014. Lecture Notes in Computer Science. 8557. 214-227. DOI : 10.1007/978-3-319-08976-8\_16.

[9] Rahul More, Yash Moily. “Big Data Analysis in Banking Sector” International Journal of Engineering Research and Applications. ISSN: 2248-9622, Vol. 11, Issue 4, (Series-II) April 2021, pp. 01-05.

[10] Ismayilova, Aygun. (2019). Theme: “Data Analytics Integration in Banking Industry”. 10.13140/RG.2.2.22127.28327.

[11] Singh, Manpreet & Ghutla, Bhawick & Jnr, Reuben & Mohammed, Aesaan & Rashid, Mahmood. “Walmart's Sales Data Analysis - A Big Data Analytics Perspective”. Published : 2017. Pages : 114-119. DOI : 10.1109/APWConCSE.2017.00028.

[12] Shahbaz, Muhammad & Gao, Changyuan & Zhai, Lili & Shahzad, Fakhar & Luqman, Adeel & Zahid, Rimsha.” Impact of big data analytics on sales performance in pharmaceutical organizations: The role of customer relationship management capabilities”. Published : 2021. PLoS ONE. 16. 1-22. DOI : 10.1371/journal.pone.0250229.

[13] J.Pavithra, Dilip Gurukrishnan. ”A STUDY ON FINANCIAL ANALYSIS OF BSNL”. International Journal of Pure and Applied Mathematics. Volume 119 No. 12 2018, 1471-1489. ISSN: 1314-3395 (on-line version). Available : http://www.ijpam.eu.

[14] [Sruthi E R](https://www.analyticsvidhya.com/blog/author/sruthi94/). “ Understanding Random Forest ”. Published 17 June 2021. [Online]. Available : https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[15] Shivam Sharma . “K-Nearest Neighbour: The Distance-Based Machine Learning Algorithm.”. Published 15 May 2021. [Online]. Available : <https://www.analyticsvidhya.com/blog/2021/05/knn-the-distance-based-machine-learning-algorithm/>

[16] Nitesh Trabani.”How the Gradient Boosting Algorithm works?”. Published 19 April 2021. [Online]. Available : <https://www.analyticsvidhya.com/blog/2021/04/how-the-gradient-boosting-algorithm-works/>

[17] V[ijay Choubey](https://medium.com/@vijay-choubey?source=post_page-----adf1bc246254--------------------------------). “Decision Tree-End to End Implementation” . Published 28 October 2020. [Online]. Available : <https://medium.com/analytics-vidhya/decision-tree-end-to-end-implementation-adf1bc246254>

[18] Rohith Gandhi. “Support Vector Machine — Introduction to Machine Learning Algorithms ”. Published 7 June 2018. [Online]. Available : https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[19] ”Logistic Regression in Machine Learning”. Published [Online]. Available : https://www.javatpoint.com/logistic-regression-in-machine-learning

[20] Akshi Saxena. “CatBoost in Machine Learning”. Published [Online] . Last Updated : 12 May, 2023 . Available : https://www.geeksforgeeks.org/catboost-ml/

[21] Pawan. “XGBoost”. Published [Online]. Last Updated : 06 Feb, 2023. Available : https://www.geeksforgeeks.org/xgboost/

[22] Rohith Gandhi. “Naive Bayes Classifier”. Published 5 May 2018. [Online]. Available : https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c.

[23] Debomit Dey. “Hierarchical Clustering in Machine Learning”. Published and last updated : 05 May , 2023. [Online]. Available : https://www.geeksforgeeks.org/ml-hierarchical-clustering-agglomerative-and-divisive-clustering/.

[24] Education Ecosystem (LEDU). “Understanding K-means Clustering in Machine Learning”. Published : 13 Sep, 2018. [Online]. Available : https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.

[25] Sebastian Raschka. “Stacking Classifier : Simple stacking” . [Online]. Available : https://rasbt.github.io/mlxtend/user\_guide/classifier/StackingClassifier/.