https://doi.org/10.58419/gbs.v11i1.1112520

EXPLORATORY DATA ANALYSIS (EDA) FOR BANKING AND FINANCE: UNVEILING INSIGHTS AND PATTERNS

Ankur Agarwal
Department of Computer Science and Information Technology,
Sam Higginbottom University of Agriculture, Technology and Sciences,
Prayagraj, Uttar Pradesh, India.
22phcoa101@shiats.edu.in

Shashi Prabha
Department of Computer Science and Information Technology,
Sam Higginbottom University of Agriculture, Technology and Sciences,
Prayagraj, Uttar Pradesh, India

Raghav Yadav
Department of Computer Science and Information Technology,
Sam Higginbottom University of Agriculture, Technology and Sciences,
Prayagraj, Uttar Pradesh, India

ABSTRACT
This paper explores the application of Exploratory Data Analytics (EDA) in the banking and finance domain, specifically focusing on analyzing customer churning. The report presents a comprehensive step-by-step analysis of banking and other data using various EDA techniques, including descriptive statistics, data visualization, and correlation analysis. Furthermore, the study investigates the critical issue of customer churning in the banking sector by using three datasets. Churn rate analysis uncovers the proportion of customers who discontinue with the bank within a given period. By identifying the churn rate, financial institutions can assess customer retention performance and set benchmarks for improvement. The study further investigates the factors driving customer churn, including customer demographics, transaction history, and customer satisfaction levels. These insights enable banking and finance professionals to make data-driven decisions, formulate targeted marketing strategies, and design effective customer retention initiatives. The findings from this analysis contribute to the advancement of the banking and finance sector's ability to enhance customer satisfaction, optimize credit card services, and ultimately drive profitability.

Keywords: *Exploratory Data Analytics (EDA), customer churning, descriptive statistics, data visualization, correlation analysis, data-driven decision-making, customer retention.*

1. INTRODUCTION

In today's data-driven world, organizations and researchers are inundated with vast amounts of data. **Data analysis** is the process of inspecting, cleaning, transforming, and modeling data to uncover useful information, patterns, and insights. It involves applying various statistical, mathematical, and computational techniques to understand the data and draw meaningful conclusions [1].

Data analysis is crucial in the **banking and finance** sector, as it helps institutions make data-driven decisions, manage risks, detect fraud, assess creditworthiness, optimize operations, and gain insights into customer behavior. It enables banks and financial institutions to leverage the vast amount of data they collect to improve efficiency, identify opportunities, and enhance overall performance. **Data analysis** can be further divided into two types: **Classical Data Analysis (CDA)** and **Exploratory Data Analysis (EDA)** [2].

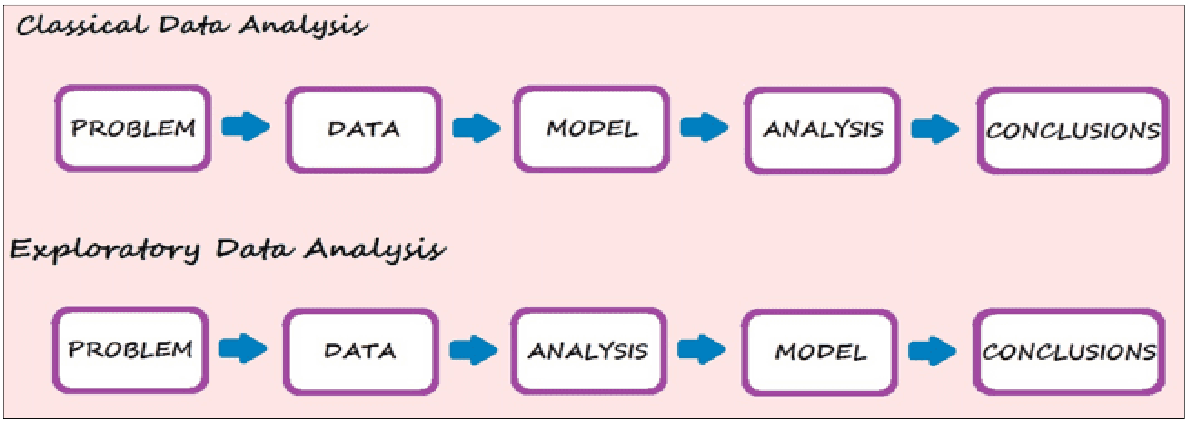
Classical Data Analysis (CDA) refers to the traditional approach of analyzing data using well-established statistical techniques and models. It follows a structured and hypothesis-driven approach, where specific hypotheses are formulated and tested using statistical tests and models. Classical data analysis involves making formal inferences, drawing conclusions, and making decisions based on the results of these statistical tests. It emphasizes hypothesis testing, parametric assumptions, and statistical significance.

On the other hand, **Exploratory Data Analysis (EDA)** is an approach to analyzing data that focuses on exploring and understanding the data without preconceived notions or specific hypotheses. It involves **visual exploration, summary statistics, data transformations**, and other techniques to gain insights and discover patterns in the data.

EDA is more flexible, intuitive, and iterative compared to CDA. Its primary goal is to **summarize and visualize** the data, detect patterns, identify relationships, and generate hypotheses for further investigation. EDA is particularly useful when dealing with **messy, incomplete, or unstructured data**.

Figure 1 illustrates the key differences between Classical Data Analysis and Exploratory Data Analysis [3].

Figure 1: Difference between Classical Data Analysis and Exploratory Data Analysis



While classical data analysis has its own merits and is suitable for hypothesis testing and formal inference, **Exploratory Data Analysis (EDA)** provides a more exploratory and intuitive approach to understanding the data, generating insights, and formulating research questions [4].

This paper focuses on Exploratory Data Analysis and its applications in **Banking and Finance**. **Section 2** of the paper reviews prior research and introduces the fundamental concepts and objectives of Exploratory Data Analysis (EDA) in the context of banking and finance, setting the stage for deeper exploration. **Section 3** and 4 discusses the methodology and various techniques used in EDA, including data cleaning, preprocessing, and statistical methods that are essential for analyzing complex datasets effectively.

In **Section 5**, the paper explores how EDA is applied to specific banking functions, illustrating its role in enhancing operational effectiveness and strategic decision-making. **Section 6** focuses on the application of EDA to analyze customer churn, detailing how data analysis can identify patterns and factors influencing customer retention and presents real-world insights from industry reports on bank customer churn. Actionable recommendations for financial institutions is presented in section 7.

Section 8 addresses the challenges and critical considerations in implementing EDA, such as data quality, privacy, and compliance with industry regulations, underscoring the complexities of deploying EDA in a regulated environment. Finally, **Section 9** provides concluding remarks that summarize the insights derived from the application of EDA and highlights its value in fostering data-driven strategies within the banking sector.

1.1. Related Work

Numerous research papers exist on the use of EDA in banking and finance, focusing on gaining insights, identifying patterns, and understanding the underlying dynamics of financial data to support informed decision-making [5]. **Exploratory Data Analysis (EDA)** serves as a cornerstone in the banking and finance sectors, playing a pivotal role in shaping accurate forecasts, assessing risks, and detecting anomalies with robustness.

John W. Tukey's groundbreaking work, published in 1977, laid the foundational bedrock for EDA by introducing concepts and techniques that advocate for an iterative approach of plotting, modeling, and validating, ultimately revealing hidden data structures and outliers [6]. This iterative methodology, pioneered by Tukey, forms the backbone upon which subsequent advancements in EDA have been built. Expanding upon Tukey's seminal contributions, William S. Cleveland's *Visualizing Data* (1993) underscored the indispensable nature of visual tools in uncovering patterns, particularly crucial for dissecting the intricate distributions and anomalies inherent in financial datasets [7]. Cleveland's emphasis on visualization not only enhanced the interpretability of data but also facilitated a deeper understanding of complex financial phenomena, thereby empowering analysts to make informed decisions.

In *Data Science for Business,* Foster Provost and Tom Fawcett (2013) further expound on the practical applications of EDA within business contexts, particularly within banking and finance, elucidating how a nuanced comprehension of data can catalyze decision-making processes [8]. By bridging the gap between theoretical concepts and real-world applications, Provost and Fawcett's work underscores the transformative potential of EDA in driving strategic initiatives and fostering innovation within the financial landscape.

Moreover, EDA plays a pivotal role in unraveling customer behavior and preferences within the banking and finance sectors, thus facilitating targeted marketing strategies aimed at enhancing customer satisfaction and fostering long-term loyalty. A plethora of research endeavors delve into various methodologies and insights in this domain, exemplified by Gupta et al.'s exploration of machine learning techniques for customer segmentation [9]. Additionally, Wibowo et al. leverage predictive analytics to discern consumer behavior patterns, thereby optimizing marketing strategies and bolstering retention efforts [10]. Ahmed et al.'s application of data mining methods further augments these efforts by enabling the prediction of customer behaviors, thereby facilitating more effective marketing campaigns and service enhancements [11]. Königstorfer et al., in their research work, explored applications of Artificial Intelligence in commercial banks and investigated behavioral finance to understand retail banking customers' decisions [12]. Abdolvand focused on customer lifetime value modeling to refine marketing strategies and enhance profitability [13]. These studies provide a comprehensive view of how EDA and associated analytical methods can be leveraged to improve understanding and engagement with banking customers.

Exploratory Data Analysis (EDA) significantly enhances decision-making processes in banking and finance, influencing areas such as product development, pricing strategies, portfolio management, investment decisions, and resource allocation. Many researchers delve into how EDA techniques improve financial decision support systems, aiding in more nuanced investment and financial product decisions [14]. Tatsat et al., in their recent work, discussed the benefits of EDA in optimizing portfolio strategies and managing investment risks [15]. Furthermore, EDA extends its influence into diverse realms of financial decision-making, permeating areas such as product development, pricing strategies, portfolio management, investment decisions, and resource allocation. Hasan et al.'s exploration of big data analytics, including EDA, sheds light on how these methodologies contribute to the development of competitive pricing strategies and the creation of innovative product offerings [16].

In addition to strategic decision-making, EDA also serves as a catalyst for technological innovation within the financial landscape, as exemplified by Levenberg et al.'s exploration of text mining and sentiment analysis for predicting economic indicators [17]. Similarly, Kim et al.'s proposition of a financial time series forecasting model utilizing Support Vector Machines underscores the efficacy of advanced machine learning techniques in augmenting traditional forecasting methodologies [18].

A comprehensive review of data mining techniques used in financial markets is provided by Kumar et al. [19]. Masini extends machine learning applications beyond traditional time series analysis in financial predictions [20]. Collectively, many of these research endeavors underscore the multifaceted nature of EDA’s impact on financial markets, transcending traditional boundaries to drive innovation and foster informed decision-making [21][22].

In conclusion, the literature review presented herein underscores the indispensable role of **Exploratory Data Analysis (EDA)** in revolutionizing decision-making processes within the banking and finance sectors. From its foundational principles laid down by visionaries like John W. Tukey and William S. Cleveland to its contemporary applications elucidated by scholars such as Foster Provost, Tom Fawcett, and numerous others, EDA has emerged as a cornerstone for understanding, interpreting, and leveraging complex financial data. Through its iterative approach, emphasis on visualization, and integration with advanced analytical techniques, EDA not only facilitates the dissection of customer behavior and preferences but also drives strategic initiatives ranging from product development to portfolio management. Moreover, the synthesis of diverse methodologies—from machine learning to sentiment analysis—underscores EDA’s transformative potential in shaping the future landscape of banking and finance. As the financial ecosystem continues to evolve, EDA stands as a beacon of innovation and insight, empowering stakeholders to navigate the complexities of an ever-changing market with confidence and precision.

2. METHODOLOGY

This section outlines the data source, preprocessing techniques, and statistical tools employed in this study to apply Exploratory Data Analysis (EDA) in the banking and finance domain, particularly focusing on customer churn prediction.

Dataset Description and Source: The study utilizes three publicly available datasets to analyze customer churn across different domains. The Kaggle Churn dataset comprises 10,000 banking customer records with demographic and financial features, where the target variable indicates account closure. The Iranian Churn dataset includes 3,150 telecom customers, focusing on behavioral factors like complaints, call failures, and subscription length. The Bank Marketing dataset from a Portuguese bank contains 45,211 records, capturing demographic and campaign interaction details, with the target indicating term deposit subscription. Together, these datasets offer a comprehensive foundation for cross-sector churn analysis.

Data Collection and Import: The dataset was downloaded as a .csv file from respective official websites and imported into the Python environment using the pandas library. The initial inspection involved examining the data types, shape, column names, and basic statistics using functions like .info(), .describe(), and .head().

Preprocessing Techniques: Several data cleaning and transformation steps were conducted to ensure data quality and suitability for analysis:

- **Handling Missing Values:** The dataset was examined for null or missing entries. As no missing values were detected, no imputation was required.

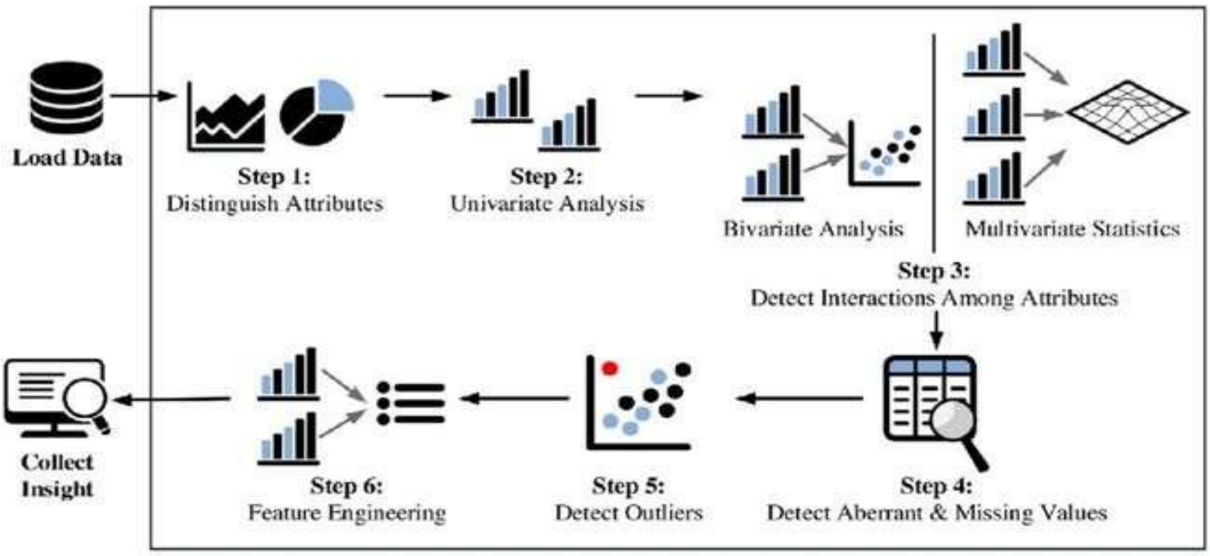
- **Outlier Detection and Handling:** Outliers in numerical features such as CreditScore and Balance were identified using **box plots** and statistical methods like **Z-score** and **Interquartile Range (IQR).**
- **Encoding Categorical Variables:** Categorical features such as Geography and Gender were encoded using **one-hot encoding** to make them compatible with analysis tools.
- **Feature Scaling:** Continuous variables were scaled using **standardization** to ensure that all features contributed equally to the analysis, particularly for correlation and dimensionality reduction.

Exploratory Techniques and Tools: The analysis utilized Matplotlib and Seaborn for visualizations, and computed descriptive statistics using pandas and scipy. Pearson correlation measured relationships among features, while PCA reduced dimensionality. Feature engineering, including age bins and tenure groups, enhanced segmentation and pattern discovery.

3. KEY TECHNIQUES IN EXPLORATORY DATA ANALYTICS

Exploratory Data Analysis (EDA) encompasses several key techniques that aid in uncovering insights and patterns within data. Important techniques that are employed during EDA to achieve its objectives are indicated in figure 2:

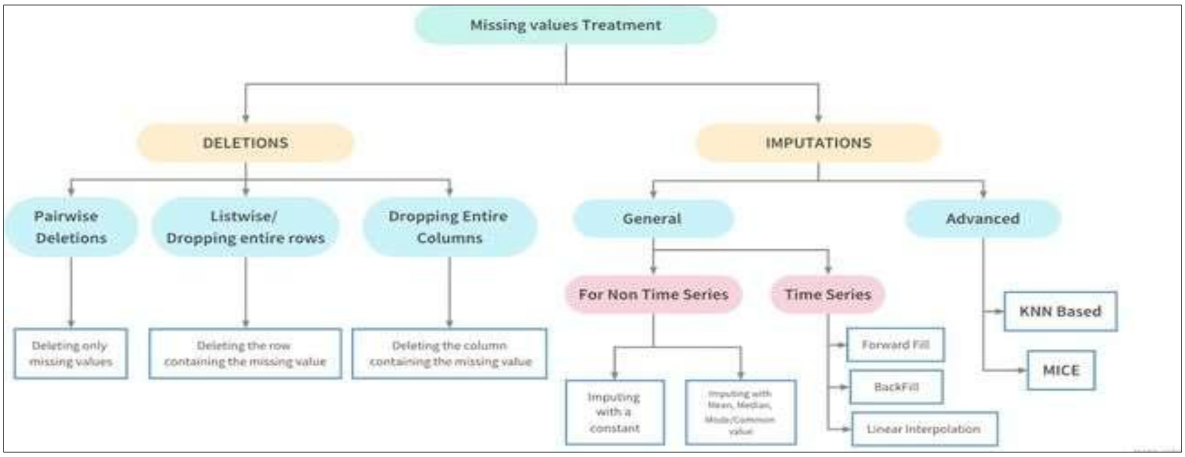
Figure 2: Key Techniques in Exploratory Data Analytics



3.1. Data Cleaning and Preprocessing

Data cleaning and preprocessing are crucial steps in Exploratory Data Analysis (EDA) that involve identifying and handling issues in the dataset to ensure its quality and suitability for analysis. The main objectives of data cleaning and preprocessing are to address missing values, handle outliers, resolve inconsistencies, and transform the data into a usable format. Here are the key steps involved in data cleaning and preprocessing [6]:

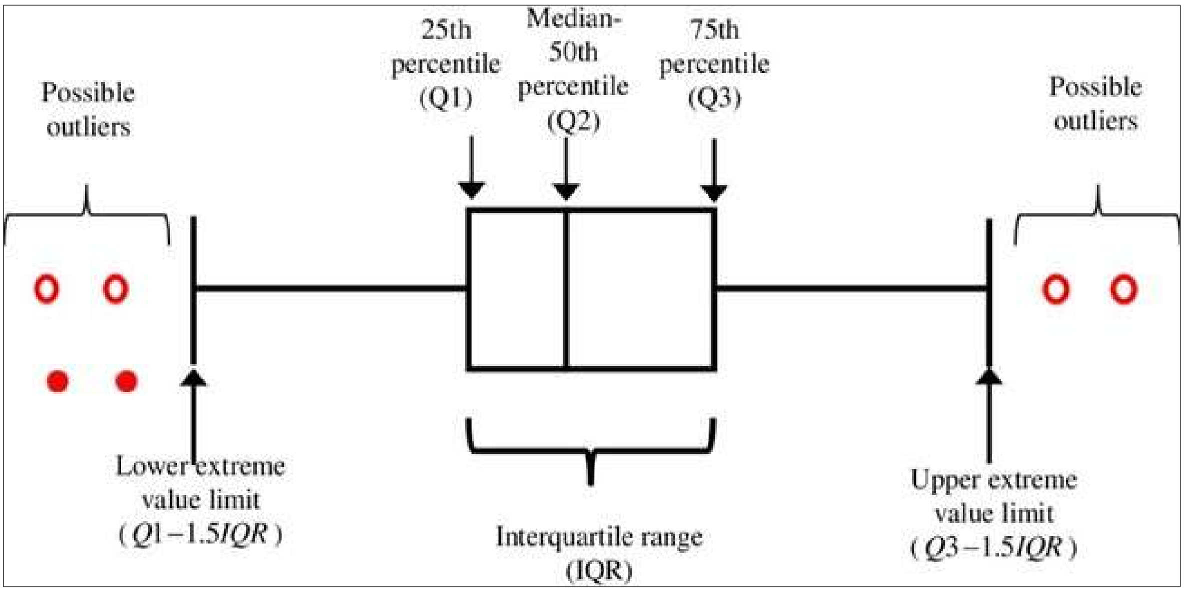
Figure 3: Techniques for handling missing values



Handling Missing Values: Missing values are problematic because they may introduce bias and affect the accuracy of the exploratory data analysis. Techniques for handling missing values, as indicated in Figure 3, include **imputation,** where missing values are filled in using statistical methods such as **mean, median,** or **regression-based imputation.** Alternatively, data points with missing values can be removed, but this should be done carefully to avoid significant data loss.

Outlier Detection and Handling: Outliers, as highlighted in Figure 4, are extreme or unusual data points that deviate significantly from the majority of the data. They can be caused by errors, data entry mistakes, or genuine anomalies. Outliers can impact the analysis, so it is important to identify and handle them appropriately. Outliers can be detected using statistical methods such as the z-score, interquartile range (IQR), or visualization techniques like box plots. Handling outliers may involve replacing them with more reasonable values or removing them entirely if they are deemed invalid [7].

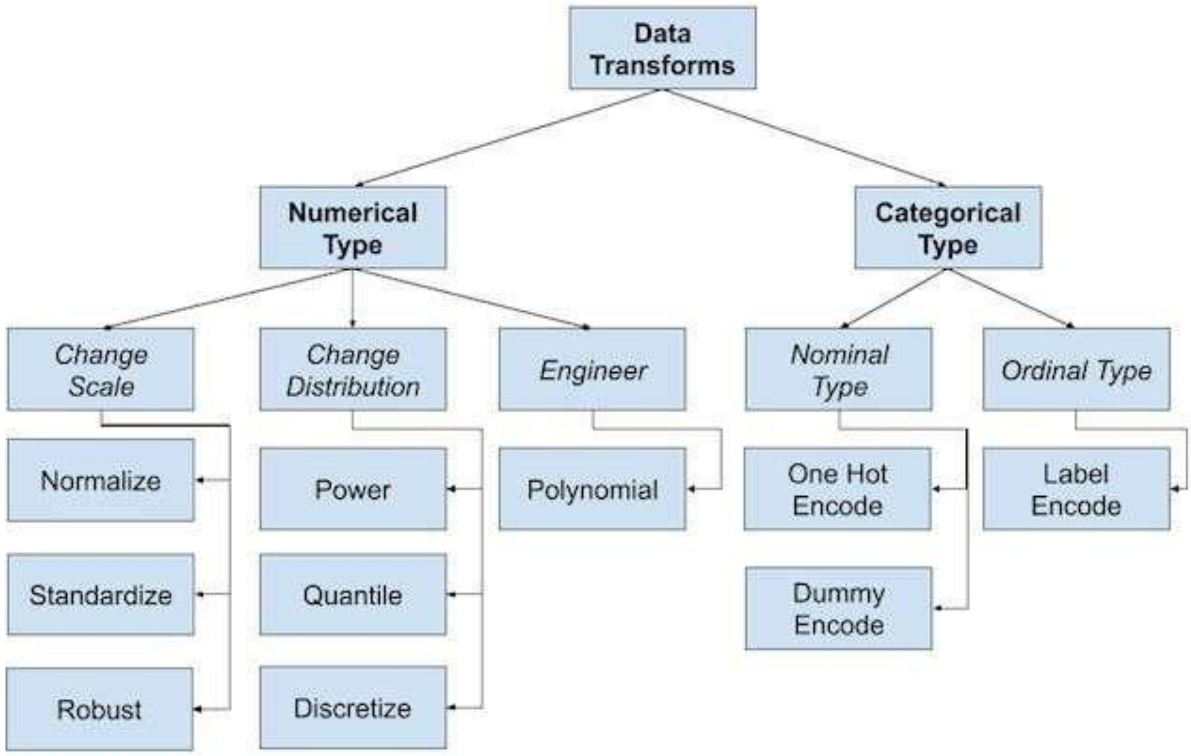
Figure 4: Outlier detection techniques



Resolving Inconsistencies: Inconsistencies in the data can arise from data entry errors, different data formats, or inconsistencies in data coding. It is important to identify and resolve these inconsistencies to ensure data accuracy and reliability. This may involve standardizing variables, correcting data entry errors, or converting data into a consistent format.

Data Transformation: Data transformation techniques are used to normalize the data or adjust its distribution. These techniques help improve the performance of statistical models and enhance interpretability. Figure 5 illustrates various data transformation techniques commonly applied during preprocessing.

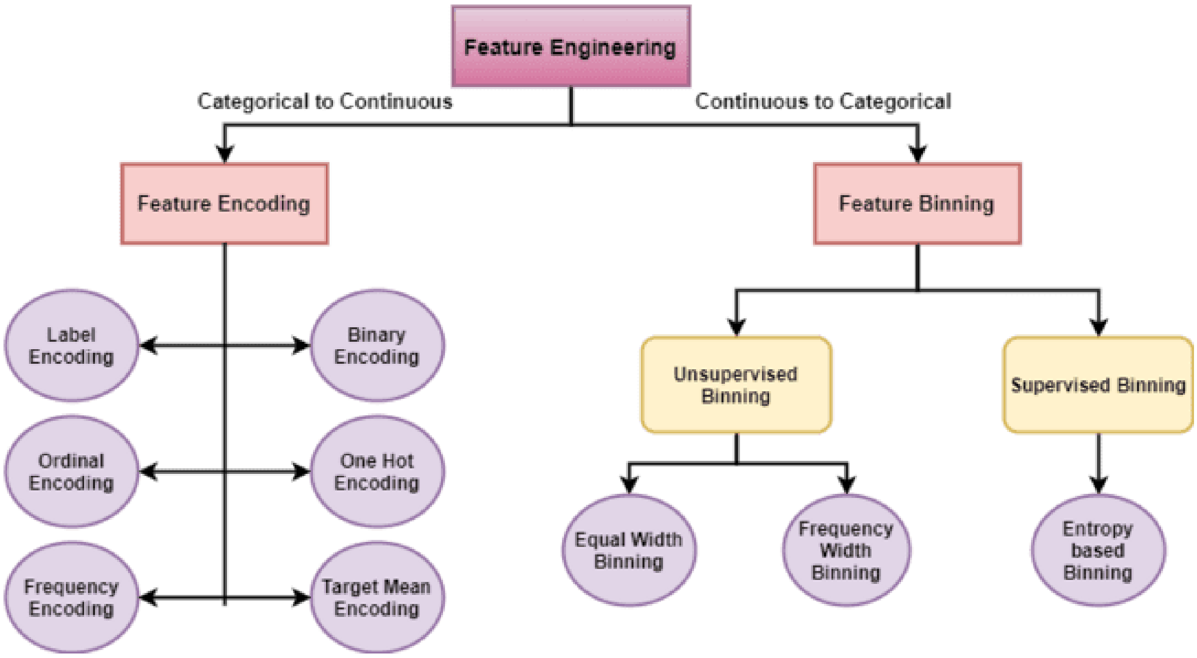
Figure 5: Data Transforms techniques



Common transformations include logarithmic transformation, square root transformation, or normalization to scale the data between a specific range. These transformations can help address issues such as skewed distributions or different scales among variables. Ensuring consistent formatting and encoding of data is also important for effective analysis. This includes converting variables into the appropriate data types (e.g., numeric, categorical, date/time) and encoding categorical variables using techniques like one-hot encoding or label encoding.

Feature Engineering: Feature engineering involves creating new features or transforming existing ones to extract relevant information and enhance the predictive power of the dataset. This may include creating interaction terms, polynomial features, or binning continuous variables into categorical variables [9]. Figure 6 illustrates various feature engineering techniques.

Figure 6: Feature engineering techniques



3.2. Summary Statistics

Summary statistics play a fundamental role in **Exploratory Data Analysis (EDA)** by providing key insights into the **central tendency, spread,** and **distribution** of the data. They help in understanding the overall characteristics of the dataset without delving into complex statistical models. Below are some important types of summary statistics:

Measures of Central Tendency: These summary statistics measures the center or average of the data include the mean, median, and mode. The mean represents the arithmetic average of the data, the median is the middle value when the data is sorted, and the mode is the most frequently occurring value.

Measures of Dispersion: Summary statistics that quantify the spread or variability of the data include the range, variance, and standard deviation. The range is the difference between the maximum and minimum values in the dataset. The variance measures the average squared deviation from the mean, while the standard deviation is the square root of the variance, representing the average distance of data points from the mean [8]. Figure 7 illustrate Measure of central tendency.

Percentiles: Percentiles divide the data into equal portions, providing information on how values are distributed across the dataset. For example, the **median** represents the **50th percentile,** dividing the data into two equal halves. The **quartiles** further divide the data into four equal parts: the **first quartile (Q1)** represents the **25th percentile,** the **second quartile (Q2)** is the **median or 50th percentile,** and the **third quartile (Q3)** represents the **75th percentile**. These percentile values are useful for understanding data distribution and detecting potential outliers. **Figure 8** illustrates the concept of percentiles and quartile division.

Figure 7: Measure of central tendency

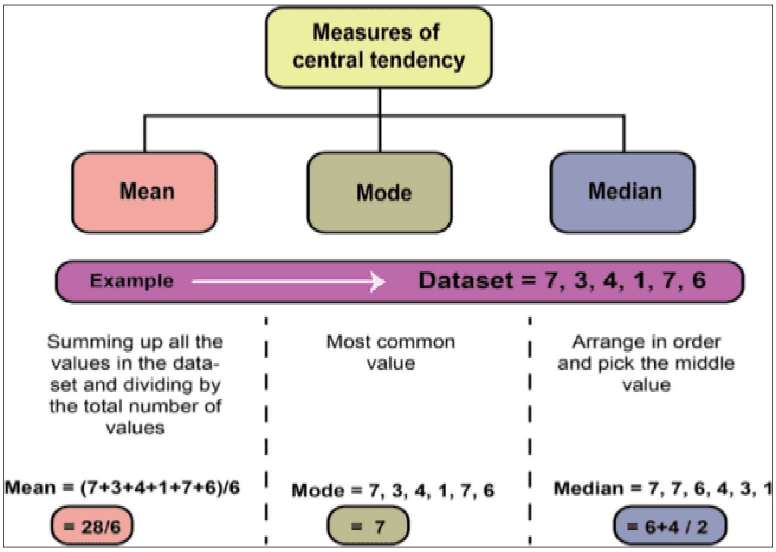
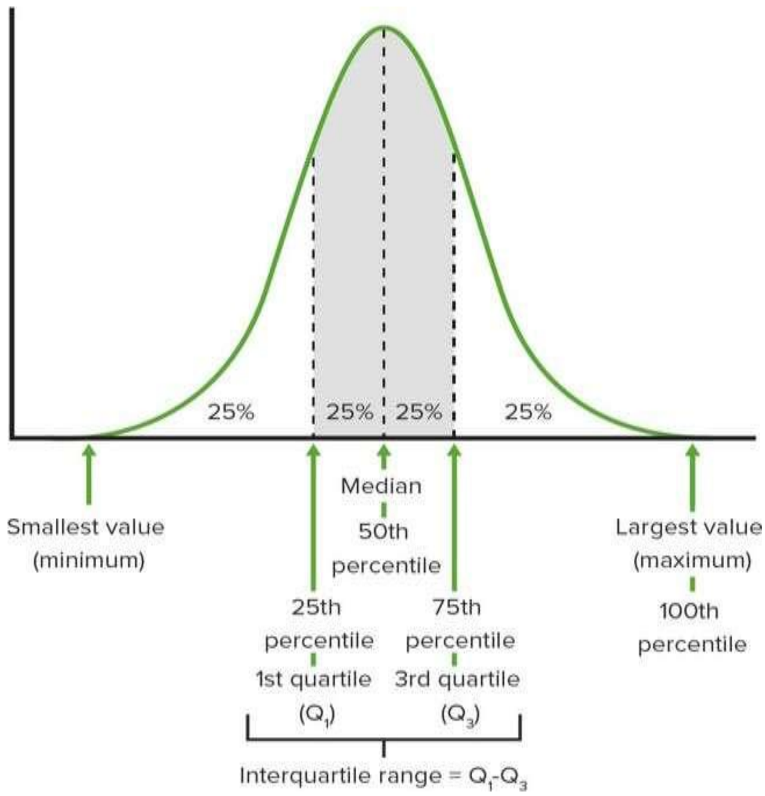


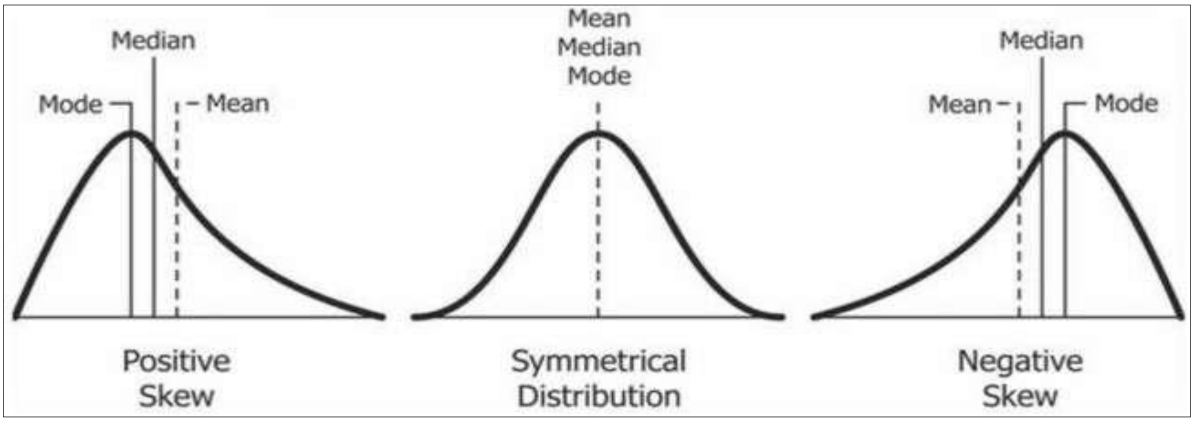
Figure 8: Percentiles description



Skewness and Kurtosis: Skewness measures the asymmetry of the data distribution, indicating whether it is skewed to the left (negative skew) or to the right (positive skew) as shown in figure 9. Kurtosis measures the degree of peakedness or flatness of the data distribution, highlighting whether it has heavy tails or is more concentrated around the mean.

$$Skewness = \frac{3\,(Mean - Median)}{Std\ \ Deviation}$$

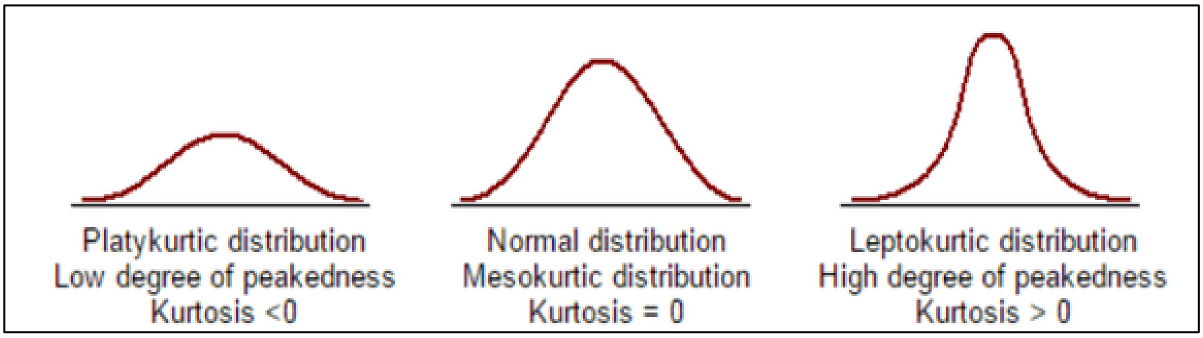
Figure 9: Types of Skewness



Kurtosis — Kurtosis describes the whether the data is light tailed (lack of outliers) or heavy tailed (outliers present) when compared to a Normal distribution. There are three kinds of Kurtosis as indicated in figure 10:

- Mesokurtic — This is the case when the kurtosis is zero, similar to the normal distributions.
- Leptokurtic — This is when the tail of the distribution is heavy (outlier present) and kurtosis is higher than that of the normal distribution.
- Platykurtic — This is when the tail of the distribution is light (no outlier) and kurtosis is lesser than that of the normal distribution

Figure 10: Types of Kurtosis



Summary statistics provide a snapshot of the data and help analysts gain an initial understanding of its characteristics. They can be used to identify outliers, assess the data's normality, and make comparisons between different datasets or subgroups. However, it's important to note that summary statistics alone do not provide a comprehensive analysis and should be used in conjunction with other exploratory techniques, visualizations, and hypothesis testing for a deeper understanding of the data.

3.3. Univariate, Bivariate and Multivariate Analysis

Univariate Analysis: Univariate analysis focuses on examining individual variables in isolation. It involves exploring the distribution, central tendency, dispersion, and shape of a single variable. Histograms, box plots, bar charts, and summary statistics are commonly used to visualize and analyze a single variable's characteristics [10].

Distribution: Distribution describes the pattern of values a variable takes on. It can be visualized using histograms, density plots, or kernel density estimations.

Central Tendency: Central Tendency measures the center or average of a variable's values. Common measures include the mean (average), median (middle value), and mode (most frequent value).

Dispersion: Dispersion describes the spread or variability of a variable's values. Measures such as variance and standard deviation quantify dispersion.

Shape: Shape represents the form or pattern of a variable's distribution. It can be characterized as symmetric, skewed (positively or negatively), or bimodal.

Bivariate Analysis: Bivariate analysis examines the relationship between two variables. It aims to understand how changes in one variable relate to changes in another variable. Scatter plots, line graphs, correlation analysis, and contingency tables are often used to explore associations, dependencies, and correlations between variables.

Scatter Plot: Scatter Plot is a graphical representation that displays the relationship between two continuous variables. It plots data points on a graph, where each point represents the values of the two variables.

Correlation: Correlation is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is typically represented by the correlation coefficient, such as Pearson's correlation coefficient.

Contingency Table: Contingency Table is also known as a cross-tabulation table. It displays the relationship between two categorical variables. It presents the counts or percentages of observations falling into different categories for each variable.

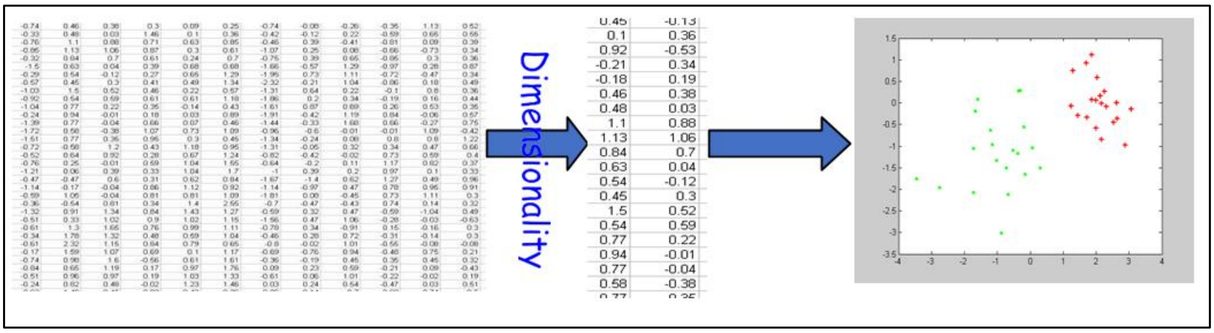
Covariance: Covariance measures the relationship between two variables. It indicates how changes in one variable are related to changes in another. Positive covariance indicates a direct relationship, while negative covariance indicates an inverse relationship.

Multivariate Analysis: Multivariate analysis explores the relationships between three or more variables simultaneously. It allows for a more comprehensive understanding of complex interactions and dependencies. Techniques such as principal component analysis (PCA), factor analysis, and multidimensional scaling are used to reduce dimensionality and visualize high-dimensional datasets.

3.4. Dimensionality Reduction

Dimensionality reduction is a technique used in exploratory data analysis (EDA) to reduce the number of variables or features in a dataset while preserving important information. High-dimensional datasets with numerous variables can pose challenges in analysis and interpretation. Dimensionality reduction aims to simplify the dataset by transforming it into a lower-dimensional space, making it more manageable and facilitating data visualization and modelling [11]. Figure 11 shows an example of dimensionality reduction.

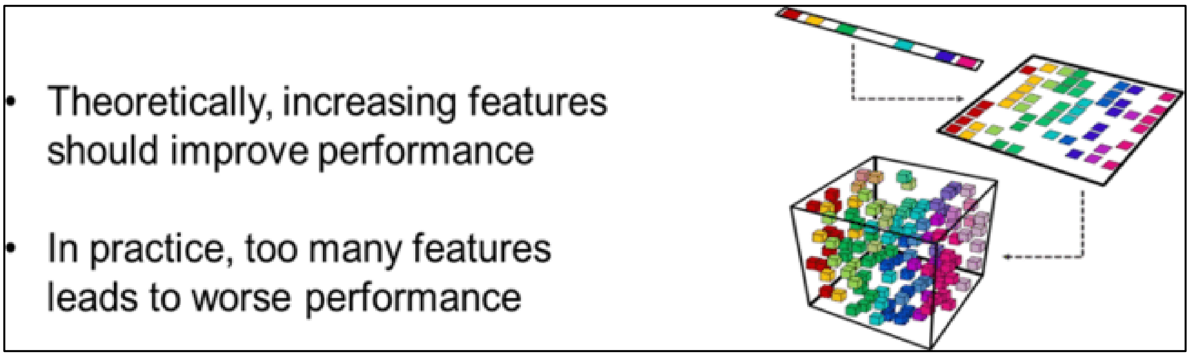
Figure 11: Dimensionality Reduction



The main objectives of dimensionality reduction are as follows:

Curse of Dimensionality: High-dimensional datasets can suffer from the curse of dimensionality, where the amount of data required to adequately cover the featurespace grows exponentially with the number of dimensions. This can lead to sparse data, increased computational complexity, and difficulty in identifying meaningful patterns. Dimensionality reduction helps in mitigating these issues by reducing the dimensionality of the dataset. Figure 12 illustrates Curse of Dimensionality:

Figure 12: Curse of Dimensionality



3.5. Approaches to Dimensionality Reduction

Feature Selection: This approach selects a subset of the original features based on certain criteria, such as statistical significance, importance, or correlation with the target variable. It aims to retain the most informative features while discarding redundant or irrelevant ones.

Feature Extraction: This approach transforms the original features into a new set of lower-dimensional features. Principal Component Analysis (PCA) is a widely used feature extraction technique that linearly transforms the data to orthogonal components, known as principal components, capturing the maximum amount of variance in the data.

Other popular dimensionality reduction techniques include **t-Distributed Stochastic Neighbor Embedding (t-SNE),** which emphasizes preserving the local structure of the data, and **Linear Discriminant Analysis (LDA),** which focuses on maximizing class separability. Dimensionality reduction helps in simplifying complex datasets, improving visualization and interpretability, and enhancing the performance of subsequent analyses and models. It is a valuable tool in EDA to gain insights and extract meaningful information from high-dimensional data.

4. APPLICATION OF EDA IN BANKING AND FINANCE

Exploratory Data Analysis (EDA) plays a crucial role in the banking and finance industry. Some key applications of EDA in this domain are shown in **Figure 13:**

Figure 13: Application of EDA in Banking and Finance

5. EDA ON CHURNING OF CUSTOMERS IN BANKING AND FINANCIAL SECTORS

In the banking and financial sectors, **customer churn** refers to the phenomenon where customers discontinue their relationship with a bank or stop using a particular product or service offered by the bank, such as credit card services. Customer churn poses a significant challenge for banks as it can lead to a loss of revenue, market share, and customer loyalty. Customer churn in the banking sector can occur due to various reasons, including dissatisfaction with service quality, high fees, better offers from competitors, changing financial needs, or a negative customer experience. Identifying and predicting customer churn is crucial for banks to implement proactive retention strategies and minimize customer attrition. By analyzing customer data and employing techniques such as **Exploratory Data Analysis (EDA),** banks can gain insights into the factors influencing customer churn. EDA allows banks to explore the data, identify patterns, and uncover key drivers that contribute to customer attrition. This analysis can involve examining customer demographics, transactional behavior, account activity, customer feedback, and other relevant data points. Once the potential churn factors are identified, banks can develop **targeted retention initiatives** to retain at-risk customers. This may involve personalized communication, tailored offers, improved customer service, loyalty programs, or incentives to encourage customers to stay with the bank.

5.1. EDA using Multiple Churn Datasets

To enhance the empirical depth and generalizability of our study, we incorporated three real-world datasets. These datasets differ in domain, geography, and customer profiles, thereby enabling a broader exploration of customer churn patterns through Exploratory Data Analysis (EDA). The details of various datasets are as follows:

Kaggle Churn Modelling Dataset

- **Source:** Kaggle
- **Size:** 10,000 customers × 14 features
- **Geography:** France, Spain, Germany
- **Target Variable:** Exited (binary)
- **Notable Features:** CreditScore, Age, Balance, NumOfProducts, IsActiveMember

This dataset focuses on customer attrition in the retail banking sector and has been extensively used to study financial churn behavior. As covered earlier in the manuscript, this dataset served as our baseline.

Iranian Churn Dataset

- **Source:** Telecom domain (open repository)
- **Size:** 3,150 customers × 13 features
- **Geography:** Iran
- **Target Variable:** Churn (yes/no)
- **Notable Features:** Call Failure, Complaints, Subscription Length, Charge Amount, Customer Value

This dataset provides a contrasting view from the telecom sector. Despite being in a different industry, the EDA showed interesting parallels such **as longer subscription length** and **higher charge amount** correlating with **lower churn probability,** while frequent **complaints** and **call failures** were key churn indicators.

Bank Marketing Dataset

- **Source:** UCI Machine Learning Repository
- **Size:** 45,211 customers × 17 features
- **Geography:** Portugal
- **Target Variable:** y (Subscribed to term deposit: yes/no)
- **Notable Features:** Job, Marital, Education, Contact, Duration, Campaign, Previous

This dataset includes customer responses to direct marketing campaigns. Though the target variable measures marketing success rather than explicit churn, the binary classification setup and behavioral attributes make it highly relevant for analyzing attrition risk in response to customer engagement strategies. EDA revealed that **contact duration** and **previous outcomes** were significant indicators of a customer's likelihood to subscribe (or churn from engagement).

5.2. EDA steps on Bank Churning dataset

The dataset used in this project comprises comprehensive information about the bank customers. The dataset includes various attributes related to each customer, with the target variable is represented as a binary value indicating whether the customer has closed their bank account or remains an active customer. This target variable is a crucial for analyzing customer churn within the banking sector. The dataset is taken from Kaggle repository [27]. There are 14 attributes in the dataset. The information of the dataset is shown as follows in figure 14:

Figure 14: Dataset Description

RowNumber — the record (row) number and has no effect on the output.
CustomerId — contains random values and has no effect on customer leaving the bank.
Surname — the surname of a customer has no impact on their decision to leave the bank.
CreditScore — can have an effect on customer churn, since a customer with a higher credit score is less likely to leave the bank.
Geography — a customer's location can affect their decision to leave the bank.
Gender — it's interesting to explore whether gender plays a role in a customer leaving the bank. We'll include this column, too.
Age — this is certainly relevant, since older customers are less likely to leave their bank than younger ones.
Tenure — refers to the number of years that the customer has been a client of the bank. Normally, older clients are more loyal and less likely to leave a bank.
Balance — also a very good indicator of customer churn, as people with a higher balance in their accounts are less likely to leave the bank compared to those with lower balances.
NumOfProducts — refers to the number of products that a customer has purchased through the bank.
HasCrCard — denotes whether or not a customer has a credit card. This column is also relevant, since people with a credit card are less likely to leave the bank. (0=No,1=Yes)
IsActiveMember — active customers are less likely to leave the bank, so we'll keep this. (0=No,1=Yes)
EstimatedSalary — as with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries.
Exited — whether or not the customer left the bank. This is what we have to predict. (0=No,1=Yes)

To perform operations, exploratory data analysis (EDA), visualization, and analysis on the Churn Modelling dataset from Kaggle, we will utilize various libraries in Python. Firstly, we need to import essential libraries such as Pandas and Matplotlib. Pandas provides powerful data manipulation and analysis tools, while Matplotlib aids in creating visualizations to gain insights from the data. By using these libraries, we can extract meaningful information, identify patterns, and understand customer churn in the banking industry. Our main aim is to anticipate whether clients will opt to terminate their banking relationship in the near future. We have access to historical data that encompasses customer behavior patterns and instances of contract terminations with the bank.

Several key analyses are required:

- Determine the target feature and the labels in the target.
- Analyze the correlation between the target and discrete/continuous features.
- Analyze the correlation among different features.
- Determine the number of missing values.
- Analyze and determine the possible outlier data.
- Plan to use appropriate strategies to handle missing values and outliers.

The details of the steps taken are shown from figure 15 to figure 23:

STEP 1: Importing required libraries

Figure 15: Screenshots of various libraries.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings


sns.set()
warnings.simplefilter('ignore')
```

Getting the data

```
[ ] data = pd.read_csv('Churn_Modelling.csv')
    df = data.copy()
    df.head()
```

The data contains 10000 rows and 14 columns.

STEP 2: Perform Data wrangling to convert raw data in to a usable form such as removing the unimportant field and understanding the missing values, null values and taking appropriate measures.

Figure 16: Screenshots of dataset

```
df.drop(['RowNumber', 'CustomerId', 'Surname'], axis=1, inplace=True)
df.head()
```

	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

```
[ ] df.isnull().sum().to_frame('No. of Nulls')
```

Once data is clean and not null. We may perform the Exploratory Data Analysis.

STEP 3: Exploratory Data Analysis

In this dataset, we aim to understand whether a customer will continue banking with us or decide to leave. From the statistics provided above, we observe that the majority of customers are from France, and a significant portion of them are male.

Figure 17: Graph between Customer count and geography

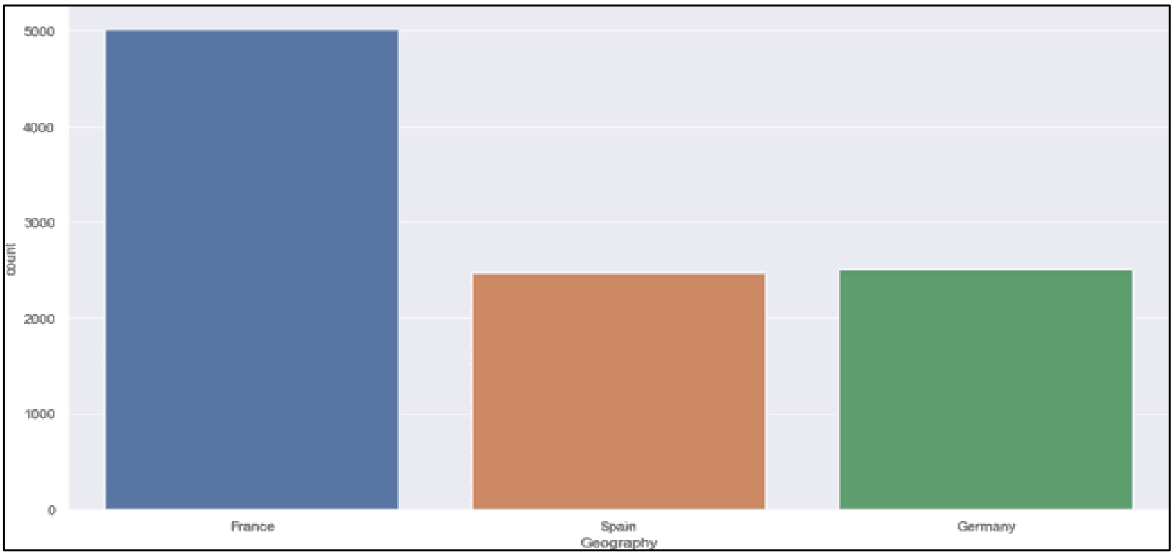
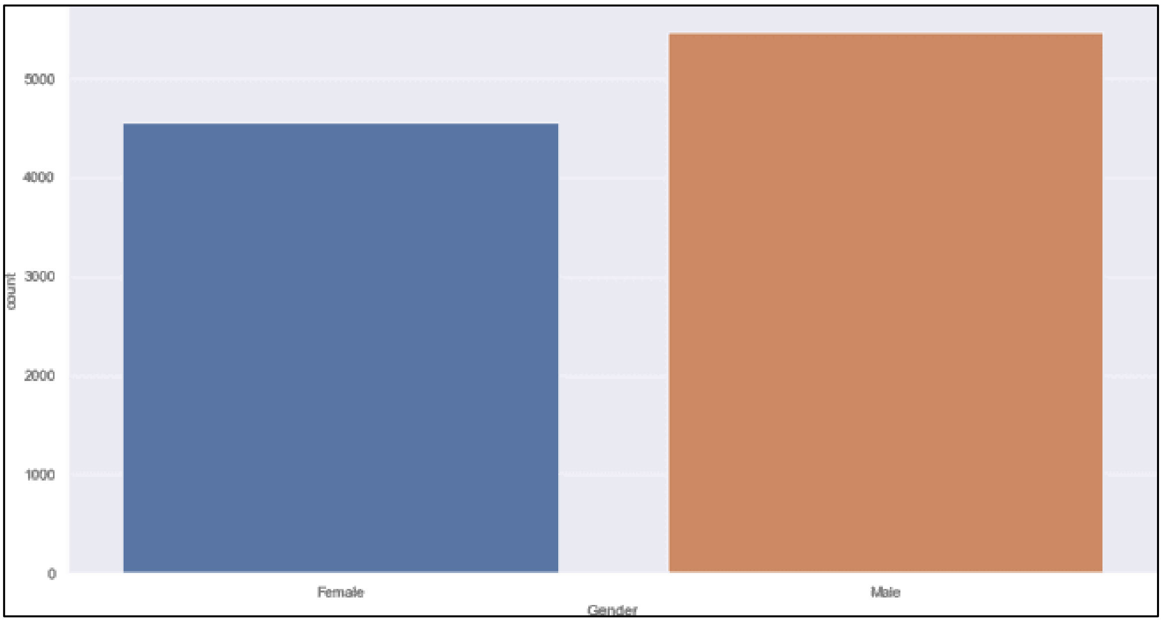
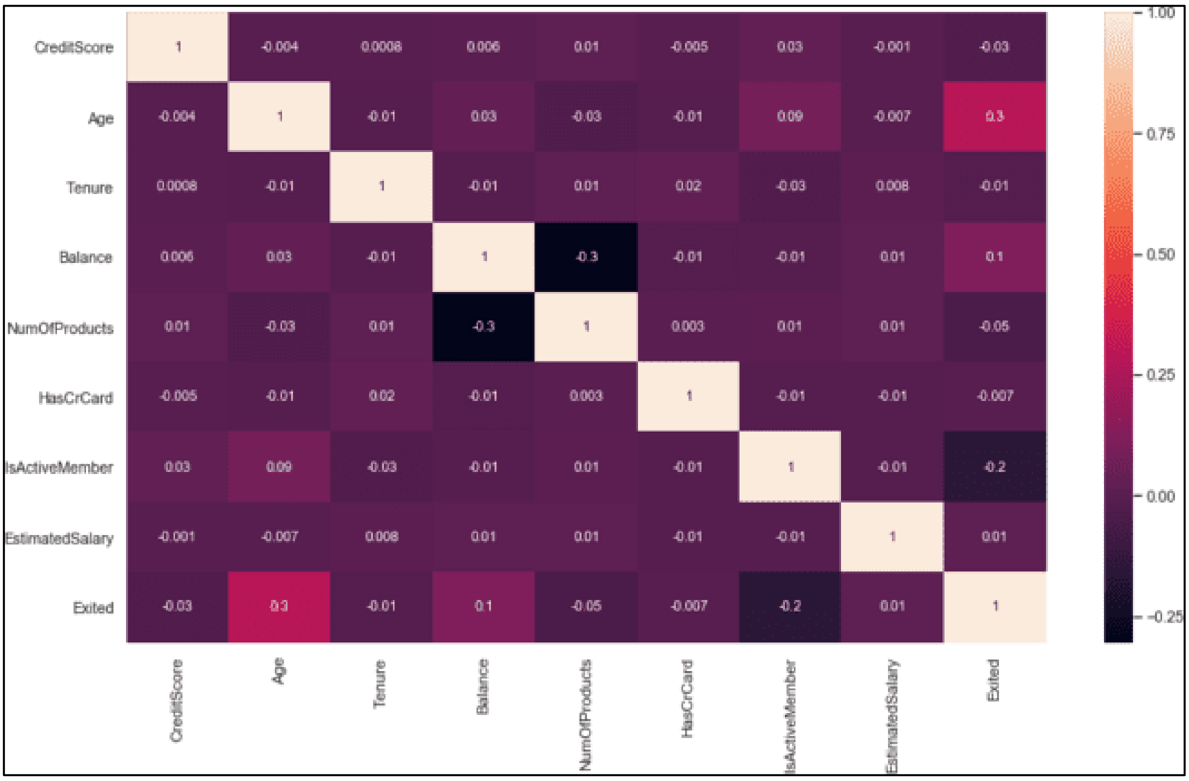


Figure 18: Graph between Customer count and Gender



We have also plotted the heatmap to get information about the correlation among different columns.

Figure 19: Heatmap



We have also plotted the distribution of Credit Score. Here, most of the distributions are between 600 and 700 and the distribution is normal (i.e. not skewed).

Figure 20: Distribution of Credit Score

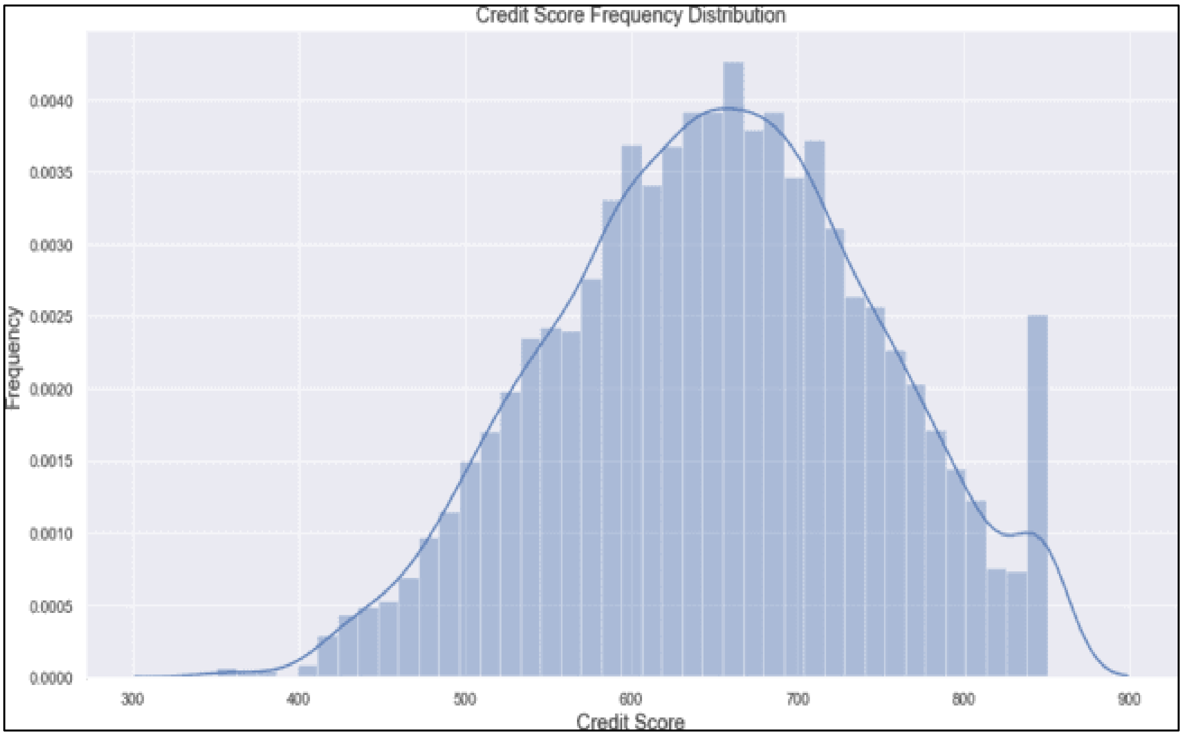
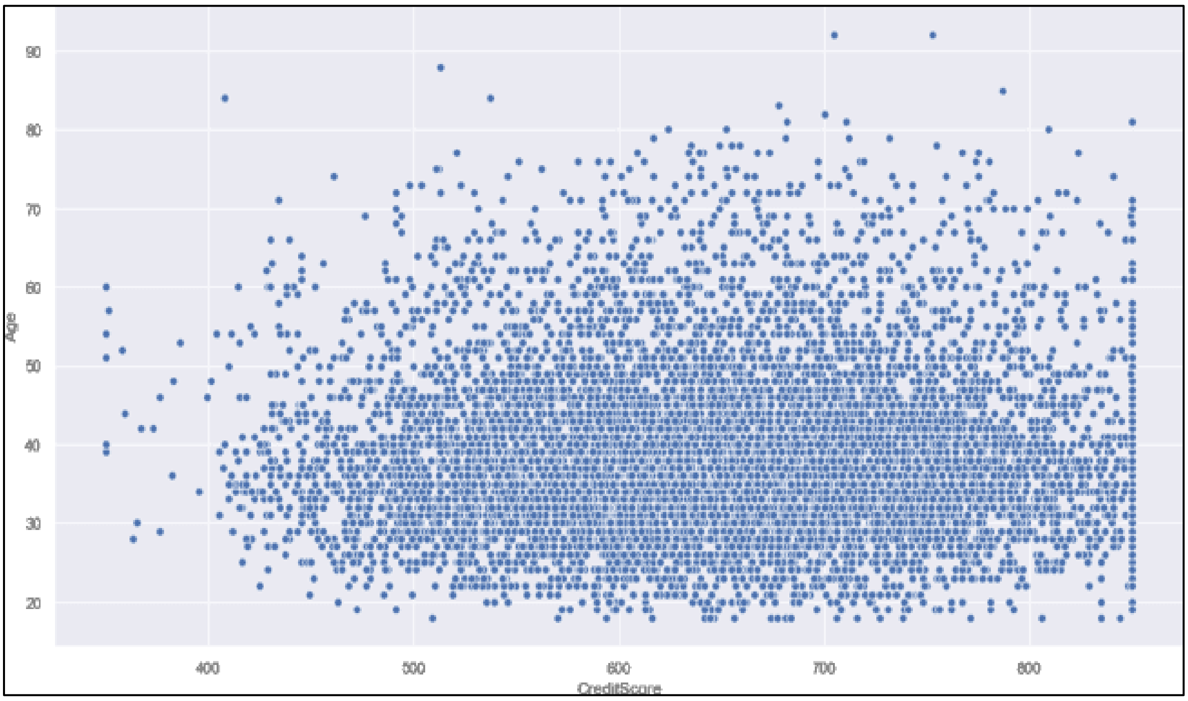
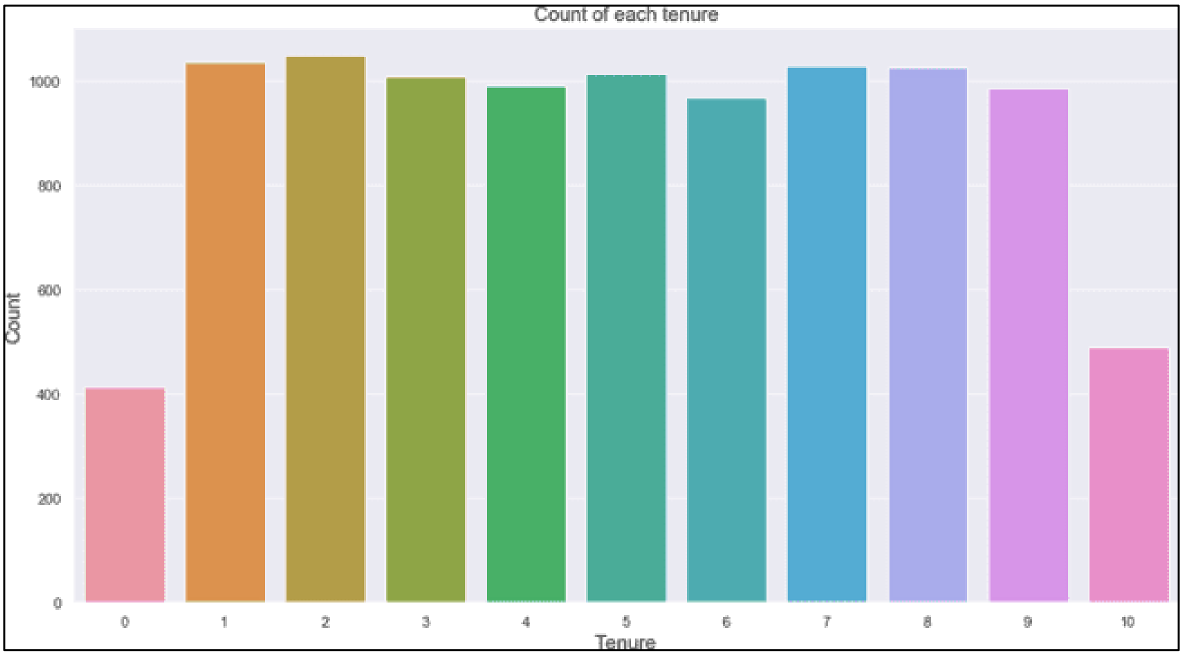


Figure 21: Graph between Credit Score and Age



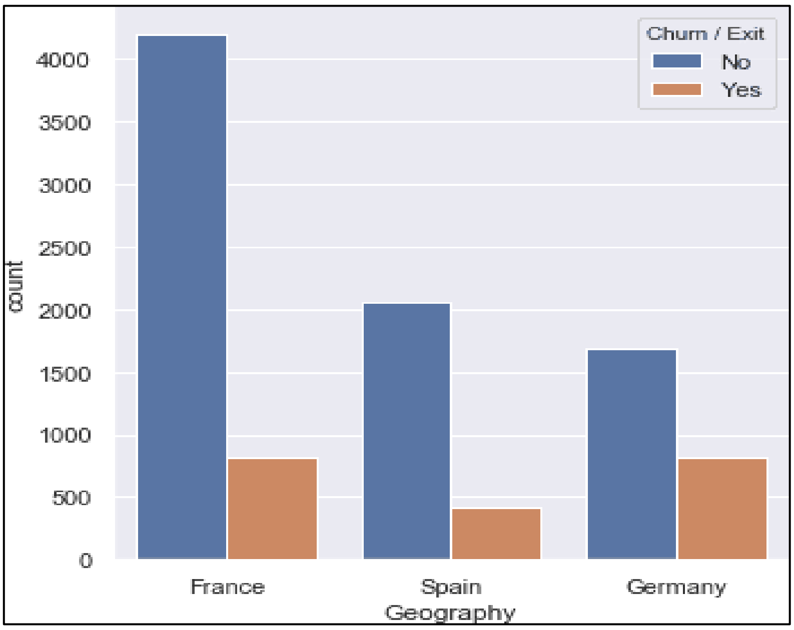
After plotting the credit score with age. It is evident that there is no correlation among age and credit score. The customer tenure with the bank was also studied and it is found that the people have similar years of tenure. However, there are a few people with less than a year or 10 years.

Figure 22: Graph between Customer Count and Tenure



The graph below represents the relationship of Geography with Churn/Exited. It is evident that all the countries are having a similar pattern of exiting.

Figure 23: Relationship of Geography with Churn/Exited

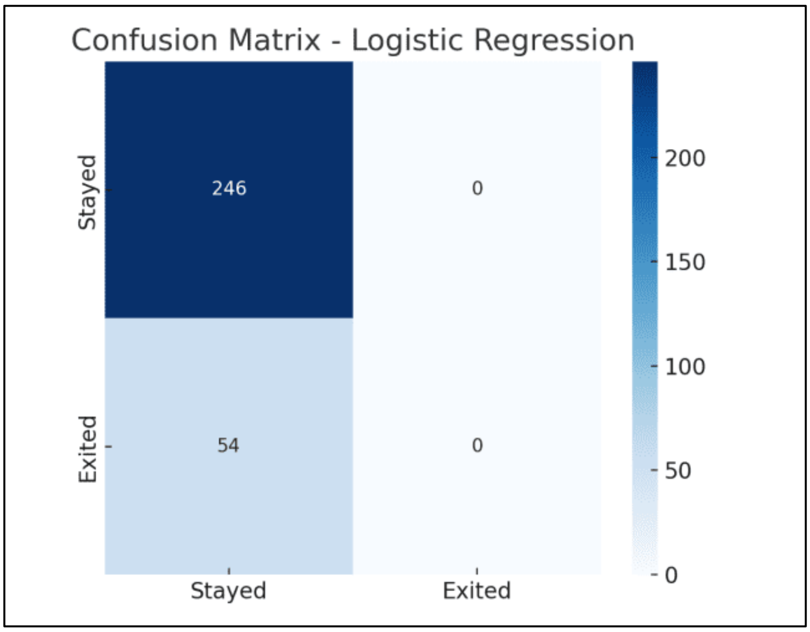


To enhance the depth of analysis, we have integrated clustering and predictive modeling techniques on the Kaggle Churn dataset. These techniques provide additional insight into customer segmentation and the predictive power of features identified during EDA. We applied K-Means clustering on standardized features including Credit Score, Age, Balance, and Activity Status. As shown in figure 24, three distinct clusters were identified, helping categorize customers into groups with similar churn-related behavior.

Figure 24: Customer clusters visualized based on Age and Balance.

To model churn prediction, we used Logistic Regression. The model was trained on selected features and evaluated using classification metrics. The confusion matrix in figure 25, shows the performance of the model on the test set.

Figure 25: Confusion matrix of Logistic Regression predictions



In conclusion, a comprehensive analysis of the dataset has revealed several key observations. The ‘CreditScore’ feature exhibits a distribution that is close to normal, with an exceptional outlier at the maximum value of 850, suggesting the presence of the highest possible credit rating within the bank. The ‘Age’ feature follows a normal distribution with a right skew, while notable outliers are observed at values that are multiples of 10, specifically 30, 40, 50, 60, and 70. Additionally, the ‘Tenure’ feature demonstrates a distribution that is approximately uniform. The ‘Balance’ feature follows a normal distribution, albeit with a significant outlier at zero, indicating a considerable number of customers who do not maintain funds in their accounts. Moreover, the majority of customers in the dataset possess 1 or 2 products, while the ‘EstimatedSalary’ feature exhibits a uniform distribution.

It is noteworthy that the bank has an almost equal number of male and female clients, as well as an equitable distribution between active and inactive clients. Interestingly, 71% of the customers hold a bank credit card, while approximately 20% of the customers have discontinued their usage of the bank's services. These insights provide a valuable understanding of the dataset, enabling further exploration and the potential development of targeted strategies for customer retention and satisfaction.

5.3. EDA Steps on Iranian Churn Dataset

This dataset includes customer usage patterns and service interactions from an Iranian telecom provider. Our goal is to understand which telecom customers are most likely to terminate their subscription [28]. Key Analyses Performed:

- Identify the target feature (Churn) and class distribution.
- Analyze correlation between churn and features such as Subscription Length, Charge Amount, Complaints, and Call Failures.

- Examine feature inter-correlations.
- Check for missing values.
- Identify and treat potential outliers.

Step 1: Importing Required Libraries: Libraries such as pandas, matplotlib, and seaborn were imported for analysis.

Step 2: Data Wrangling: The dataset consists of 3,150 rows and 13 features. We confirmed there were no missing values. Numerical columns were examined for outliers using box plots and z-score techniques. Categorical columns were label-encoded.

Step 3: Exploratory Data Analysis: Most customers had a subscription length between 12–24 months. Customers with frequent complaints and call failures had a significantly higher likelihood of churn. The feature Charge Amount showed a skewed distribution with a few high-paying customers standing out as outliers. Figure 26 and 27 explain various characteristics of the dataset.

Figure 26: Boxplot of Charge Amount for churned and non-churned users

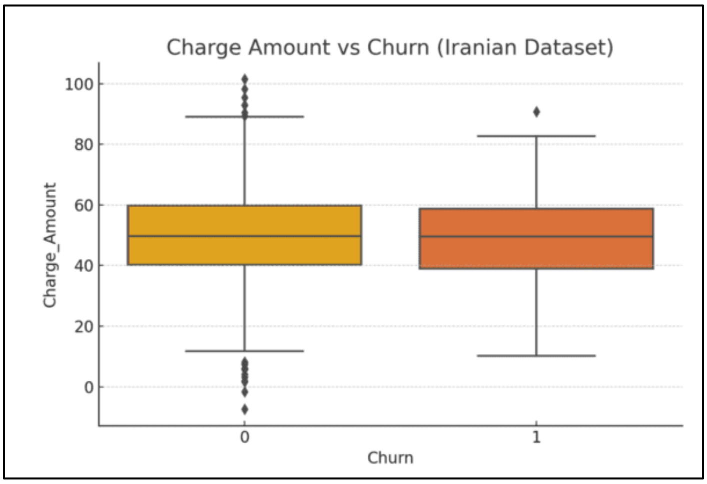
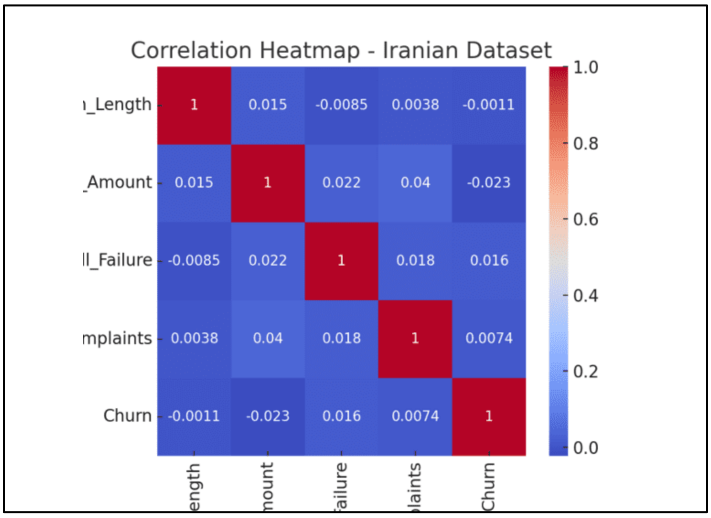


Figure 27: Correlation heatmap showing relationships among key features and churn



The analysis suggests that service quality is a major churn driver. Features such as complaint frequency and call reliability are critical indicators, emphasizing the need for telecom companies to monitor these metrics for retention efforts.

5.4. EDA Steps on Bank Marketing Dataset

This dataset, from a Portuguese bank [29], captures outcomes of direct marketing campaigns. While the target variable is y (whether the client subscribed to a term deposit), it indirectly reflects retention-oriented behavior.

Key Analyses Performed:

- Define the target (y) and inspect class imbalance.
- Explore how campaign success varies by job, education, and contact method.
- Evaluate correlation between features.
- Identify missing values (none found).
- Detect outliers in features such as Duration.

Step 1: Importing Required Libraries: We used pandas, seaborn, and matplotlib for visual and statistical analysis.

Step 2: Data Wrangling: The dataset contains 45,211 entries and 17 columns. Data cleaning revealed no missing entries. Categorical features like education and contact were encoded. The feature duration was normalized due to wide variance.

Step 3: Exploratory Data Analysis: Campaign duration (duration) was highly indicative of subscription—longer calls led to more positive responses. Prior campaign interaction (previous) also increased subscription likelihood. Class imbalance was visualized using a count plot showing only 11% of customers subscribed.

Figure 28: Barplot of Duration across Education Levels by Subscription Outcome.

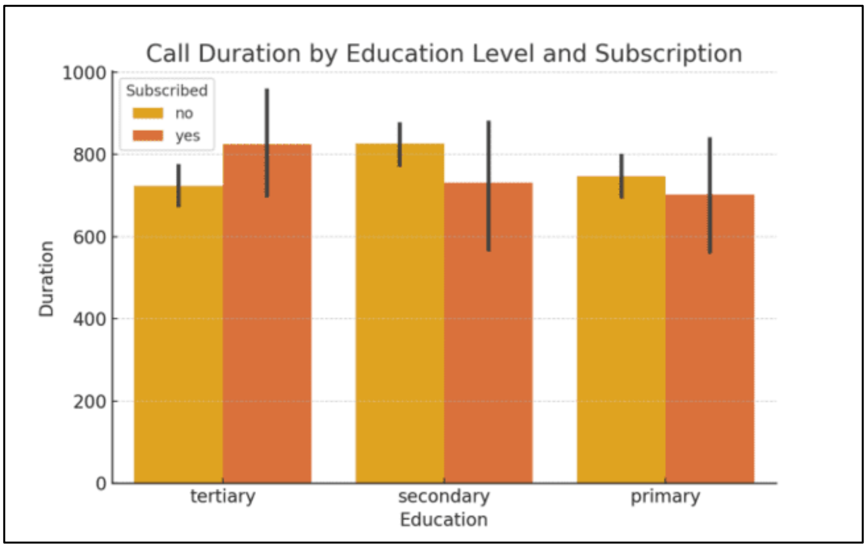
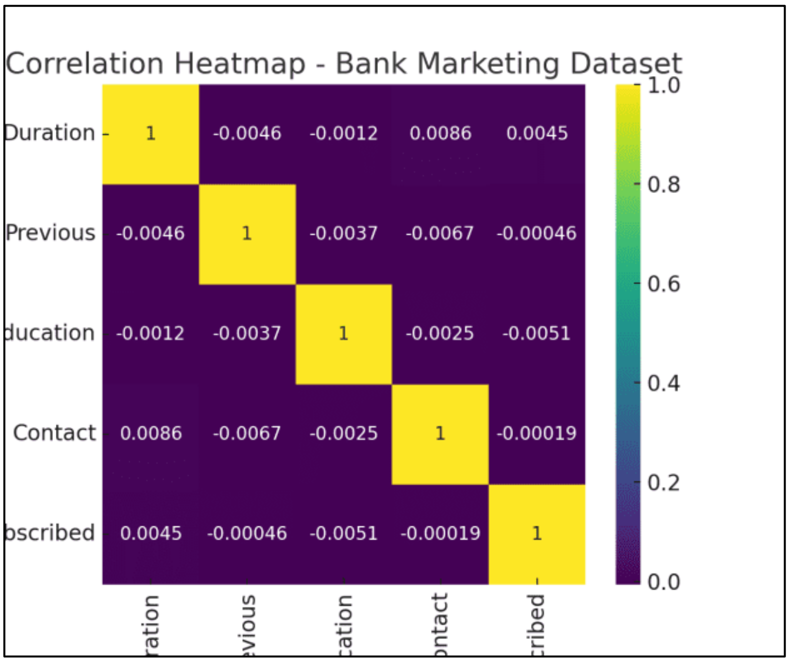


Figure 29: Correlation Heatmap of Encoded Numerical Features



EDA on this dataset revealed that personalized communication, prior engagement, and sustained contact time are strong predictors of positive marketing outcomes. These insights could help banks tailor outreach strategies more effectively.

5.5. Comparison of Churn Datasets

This section provides a comparative analysis of three different churn-related datasets used in the study: Kaggle Churn Dataset, Iranian Churn Dataset, and Bank Marketing Dataset. The comparison is based on attributes such as domain, geographic origin, number of records, feature count, target variable characteristics, and key indicators as indicted in table 1.

Table 1: Comparison of Churn Datasets

Attribute	Kaggle Churn Dataset	Iranian Churn Dataset	Bank Marketing Dataset
Dataset Name	Kaggle Churn	Iranian Churn	Bank Marketing
Domain	Banking	Telecom	Banking (Marketing)
Geography	France, Spain, Germany	Iran	Portugal
Records	10000	3150	45211
Features	14	13	17
Target Variable	Exited	Churn	y (Subscribed)
Class Distribution	Balanced (80:20)	Imbalanced (80:20)	Highly Imbalanced (~11% yes)
Missing Values	No	No	No
Outliers Present	Yes (e.g., Balance = 0)	Yes (Charge Amount, Complaints)	Yes (Duration variance)
Key Features	CreditScore, Age, Balance, IsActiveMember	Subscription Length, Complaints, Call Failure	Education, Contact Type, Duration, Previous Outcome

The visualizations shown in figure 30 and 31, summarize the dataset sizes and feature counts across the three datasets.

Figure 30: Record Count Comparison across datasets

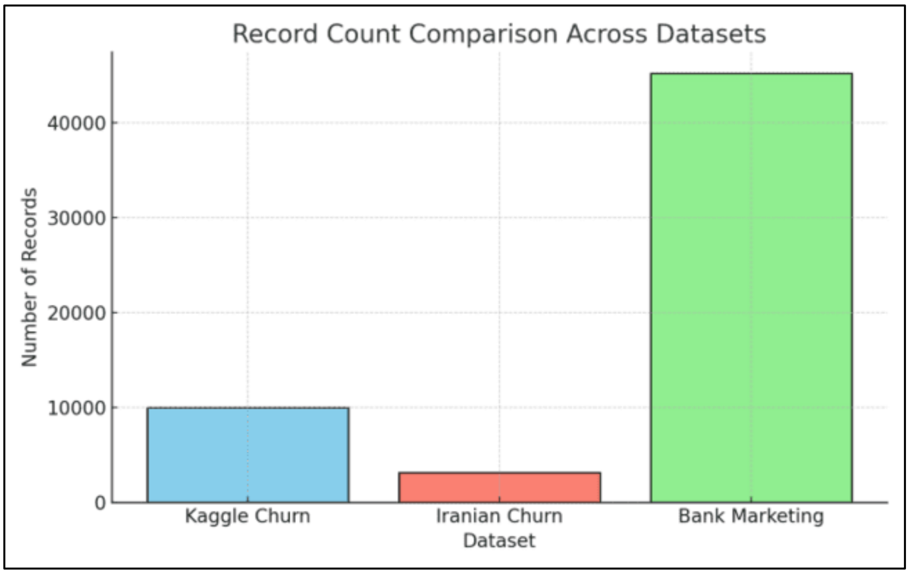
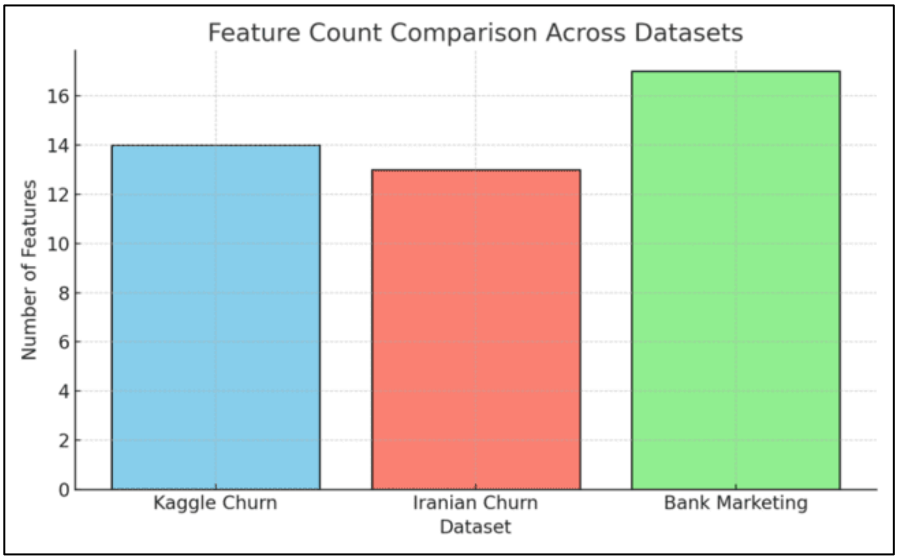


Figure 31: Feature Count Comparison across Datasets



In the Kaggle Churn dataset, customers were analyzed based on their geographic location (France, Spain, Germany) and gender. Bar plots revealed a slightly higher churn rate in Spain compared to France and Germany, while male and female customers showed nearly balanced churn rates, with a marginally higher attrition observed among males. A heatmap analysis indicated that although Geography and Tenure individually showed weak correlations with churn, their influence became more significant when combined with financial metrics such as balance and activity status. In the Bank Marketing dataset, education level and contact method were examined for their impact on term deposit subscriptions. Tertiary-educated customers showed the highest subscription rates, cellular contact methods led to better campaign outcomes, and prior contact history improved responsiveness in follow-ups.

These trends, supported by grouped bar plots and correlation analysis, offer practical insights for optimizing marketing strategies. Although the Iranian Churn dataset lacked explicit demographic attributes, behavioral segmentation was performed using subscription length, number of complaints, and call failures. Box plots and heatmaps demonstrated that customers with shorter subscription periods and higher frequencies of complaints or call failures were more likely to churn. This behavioral segmentation served as a proxy for demographic analysis, highlighting usage-based churn patterns.

6. REAL-WORLD INSIGHTS FROM INDUSTRY REPORTS ON BANK CUSTOMER CHURN

To complement our exploratory data analysis (EDA) findings, we examined recent industry reports and whitepapers that shed light on customer churn in the banking sector. These real-world insights underscore the importance of data-driven strategies in understanding and mitigating churn.

6.1. McKinsey & Company: The Impact of Analytics on Customer Retention

McKinsey's research emphasizes the transformative role of advanced analytics in banking. For instance, a European bank employed machine learning algorithms to predict which active customers were likely to reduce their business. By targeting these customers with personalized campaigns, the bank achieved a 15% reduction in churn. Similarly, a U.S. bank analyzed discount patterns offered by private bankers, leading to an 8% revenue increase after correcting unnecessary discounts [24].

6.2. Deloitte: Leveraging AI for Enhanced Customer Experience

Deloitte highlights the potential of artificial intelligence (AI) in predicting customer churn and enhancing personalization. By analyzing customer profiles and transaction data, banks can estimate customer lifetime value and identify those at risk of leaving. Implementing AI-driven strategies enables banks to deliver hyper-personalized experiences, thereby improving customer retention [24].

6.3. Global Banking Trends: Embracing Digital Transformation

The McKinsey Global Banking Annual Review 2024 discusses the industry's shift toward digitalization and the adoption of AI to improve customer experiences. Banks that effectively integrate these technologies can better understand customer needs, personalize offerings, and proactively address factors contributing to churn [25]. These industry findings align with our EDA results, reinforcing the significance of leveraging advanced analytics and AI to comprehend customer behavior and reduce churn. Incorporating such real-world data enhances the robustness of our study and provides actionable insights for financial institutions aiming to improve customer retention.

7. ACTIONABLE RECOMMENDATIONS FOR FINANCIAL INSTITUTIONS

Drawing from our exploratory data analysis, clustering, and predictive modeling across the Kaggle, Bank Marketing, and Iranian Churn datasets, we offer a set of actionable recommendations for financial institutions aiming to reduce customer churn and enhance engagement.

First, institutions should focus on identifying and targeting high-risk customer segments. Our clustering analysis revealed that customers with low activity levels, short tenure, or frequent complaints are more likely to exit. These individuals require personalized retention strategies, such as loyalty incentives, tailored communication, or early engagement through customer support.

Second, personalized outreach based on demographic characteristics can significantly improve customer responsiveness. Insights from the Bank Marketing dataset showed that customers with higher education levels and those contacted via cellular channels responded more positively to marketing campaigns. Additionally, prior interactions played a crucial role in influencing current decisions. Financial institutions should consider these factors when designing and executing customer outreach strategies.

Service quality monitoring emerged as another critical factor influencing churn, particularly in the Iranian Churn dataset. Customers with a high number of complaints or call failures exhibited a stronger likelihood of discontinuing services. Institutions must proactively track such service metrics, resolve issues promptly, and implement feedback loops to prevent dissatisfaction from escalating into attrition.

Moreover, insights from the Kaggle dataset emphasized the need to optimize credit and product offerings. Customers with specific patterns—such as low credit scores, underutilized products, or high estimated salaries—exhibited complex churn behaviors that warrant closer attention. Financial institutions should use this information to recommend personalized financial solutions and credit optimization programs.

Finally, predictive modeling, particularly using logistic regression, demonstrated the potential to forecast customer churn with reasonable accuracy. Integrating such models into existing CRM systems can enable real-time alerts, allowing relationship managers and customer support teams to act swiftly and prevent customer loss. These recommendations translate data insights into practical strategies, enabling institutions to transition from reactive churn management to proactive customer retention planning.

8. CHALLENGESAND CONSIDERATIONS

Performing Exploratory Data Analysis (EDA) in the banking and finance industry comes with its own set of challenges and considerations [24]. Here are some key factors to keep in mind:

Data Quality and Availability: One of the primary challenges in EDA is ensuring the quality and availability of data. In banking and finance, data sources may be vast and complex, with different data formats and data quality issues. Missing data, inconsistencies, and errors can impact the accuracy and reliability of the analysis. Employing data cleansing and preprocessing techniques is necessary to address these challenges.

Data Privacy and Security: Banks manage sensitive customer information, financial transactions, and compliance-related data. Maintaining data privacy and security throughout the EDA process is crucial. Compliance with regulations, such as the General Data Protection Regulation (GDPR), and ensuring data anonymization or pseudonymization are essential to protect customer privacy.

Data Integration: Data in banking and finance often reside in various systems and databases. Integrating data from multiple sources and ensuring data consistency and compatibility can be challenging. Proper data integration techniques are needed to create a unified and comprehensive dataset for analysis.

Dimensionality and Scale: Banking and finance datasets can be voluminous and high-dimensional, containing numerous variables and attributes. Handling large datasets and dealing with high dimensionality pose challenges in terms of computational complexity and visualization. Dimensionality reduction techniques, such as feature selection or extraction, may be necessary to reduce complexity and improve analysis efficiency.

Complex Relationships and Dependencies: Financial data often exhibits complex relationships and dependencies that simple statistical techniques may not capture. Non-linear patterns, time-varying relationships, and hidden dependencies can pose challenges in uncovering meaningful insights. Advanced analytics techniques, such as machine learning algorithms, may be required to address these complexities.

Interpretability and Explainability: It is crucial for the banking and finance industry to have transparent and interpretable analysis results. Stakeholders need to understand the rationale behind the findings and decisions based on EDA. Ensuring the interpretability and explainability of models and analysis outputs is essential for gaining trust and making informed business decisions.

Regulatory and Compliance Considerations: The banking and finance industry is heavily regulated, and compliance with regulatory requirements is paramount. EDA should adhere to regulatory guidelines and consider the legal and ethical implications of the analysis. Compliance with data protection laws, anti-money laundering regulations, and risk management frameworks should be taken into account.

Domain Expertise: EDA in banking and finance requires a deep understanding of the industry, financial products, risk management, and regulatory frameworks. Domain expertise is essential to ask relevant questions, interpret the results, and derive actionable insights from the analysis. By addressing these challenges and considerations, EDA in banking and finance can provide valuable insights, support decision-making processes, and help in risk management, compliance, customer satisfaction, and overall business performance.

9. CONCLUSION

Exploratory Data Analytics (EDA) serves as a vital tool for understanding and gaining insights from complex datasets. By employing various techniques such as data cleaning, descriptive statistics, data visualization, correlation analysis, dimensionality reduction, and clustering analysis, analysts can uncover patterns, relationships, and anomalies within the data. EDA finds application in diverse domains, including business analytics, healthcare, finance, social sciences, manufacturing, and environmental studies. However, analysts must be aware of challenges such as data quality, bias, confounding factors, overfitting, and ethical considerations throughout the EDA process. By leveraging EDA effectively, organizations and researchers can make data-informed decisions, drive innovation, and unlock hidden insights that can contribute to better outcomes in various fields.

REFERENCES

1. Morgenthaler, S. (2009). Exploratory data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 1*(1), 33–44.
2. Vigni, M. L., Durante, C., & Cocchi, M. (2013). Exploratory data analysis. In *Data Handling in Science and Technology* (Vol. 28, pp. 55–126). Elsevier.
3. Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory data analysis using Python. *International Journal of Innovative Technology and Exploring Engineering, 8*(12), 4727–4735.
4. Chakri, P., Pratap, S., Gouda, S. K., et al. (2023). An exploratory data analysis approach for analyzing financial accounting data using machine learning. *Decision Analytics Journal, 7*, 100212.
5. Liu, Q. (2019). An application of exploratory data analysis in auditing: Credit card retention case. In *Rutgers Studies in Accounting Analytics: Audit Analytics in the Financial Industry* (pp. 3–15). Emerald Publishing Limited.
6. Tukey, J. W. (1993). *Exploratory data analysis: Past, present, and future*. Defense Technical Information Center.
7. Cleveland, W. S. (1993). *Visualizing data*. Hobart Press.
8. Provost, F., & Fawcett, T. (2013). *Data science for business*. O'Reilly Media.
9. Gupta, R., Jain, T., Sinha, A., & Tanwar, V. (2023). Review on customer segmentation methods using machine learning. In *International Conference on IoT, Intelligent Computing and Security: Select Proceedings of IICS 2021* (pp. 397–411). Springer.

10. Wibowo, L. A., & Ariyanti, M. (2023). Utilization of artificial intelligence systems to predict consumer behavior. *Journal of Jabar Economic Society Networking Forum, 1*, 45–53.

11. Al Ayub Ahmed, A., Rajesh, S., Lohana, S., Ray, S., Maroor, J. P., & Naved, M. (2022). Using machine learning and data mining to evaluate modern financial management techniques. In *Proceedings of Second International Conference in Mechanical and Energy Technology (ICMET 2021)* (pp. 249–257). Springer.

12. Königstorfer, F., & Thalmann, S. (2020). Applications of artificial intelligence in commercial banks: A research agenda for behavioral finance. *Journal of Behavioral and Experimental Finance, 27*, 100352.

13. Abdolvand, N., Albadvi, A., & Koosha, H. (2021). Customer lifetime value: Literature scoping map, and an agenda for future research. *International Journal of Management Perspective*.

14. Jha, A., & Prajapati, B. (2023). Machine learning and exploratory data analysis in cross-sell insurance. In *Encyclopedia of Data Science and Machine Learning* (pp. 651–685). IGI Global.

15. Tatsat, H., Puri, S., & Lookabaugh, B. (2020). *Machine learning and data science blueprints for finance*. O'Reilly Media.

16. Hasan, M. M., Popp, J., & Oláh, J. (2020). Current landscape and influence of big data on finance. *Journal of Big Data, 7*(1), 21.

17. Levenberg, A., Pulman, S., Moilanen, K., Simpson, E., & Roberts, S. (2014). Predicting economic indicators from web text using sentiment composition. *International Journal of Computer and Communication Engineering, 3*(2), 109–115.

18. Kim, K.-j. (2003). Financial time series forecasting using support vector machines. *Neurocomputing, 55*(1–2), 307–319.

19. Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in the financial domain. *Knowledge-Based Systems, 114*, 128–147.

20. Masini, R. P., Medeiros, M. C., & Mendes, E. F. (2023). Machine learning advances for time series forecasting. *Journal of Economic Surveys, 37*(1), 76–111.

21. Gavrikova, E., Volkova, I., & Burda, Y. (2020). Strategic aspects of asset management: An overview of current research. *Sustainability, 12*(15), 5955.

22. Agrawal, S. S., Rose, N., PrabhuSahai, K., et al. (2024). The fintech revolution: AI's role in disrupting traditional banking and financial services. *Decision Making: Applications in Management and Engineering, 7*(1), 243–256.

23. Tae, K. H., Roh, Y., Oh, Y. H., Kim, H., & Whang, S. E. (2019). Data cleaning for accurate, fair, and robust models: A big data-AI integration approach. In *Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning* (pp. 1–4).

24. Choi, S., Kalra, N., Sharma, A., & Tandon, A. (2021). Analytics in banking: Time to realize the value. *McKinsey & Company*. Retrieved from https://www.mckinsey.com/industries/financial-services/our-insights/analytics-in-banking-time-to-realize-the-value

25. Deloitte Insights. (2023). *AI and the future of financial services*. Deloitte United States. Retrieved from https://www2.deloitte.com/us/en/insights/industry/financial-services/ai-in-banking.html

26. McKinsey & Company. (2024). *Global banking annual review 2024: Embracing a new era of digital transformation*. Retrieved from https://www.mckinsey.com/industries/financial-services/our-insights/global-banking-annual-review

27. Mouna, A. (2017). *Churn modelling*. Kaggle. Retrieved from https://www.kaggle.com/datasets/shubhendra7/customer-churn-modelling

28. UCI Machine Learning Repository. (2020). *Iranian churn dataset*. Retrieved from https://archive.ics.uci.edu/dataset/563/iranian%2Bchurn%2Bdataset

29. UCI Machine Learning Repository. (2012). *Bank marketing dataset*. Retrieved from https://archive.ics.uci.edu/dataset/222/bank%2Bmarketing