FINAL REPORT - HOMEWORK 5

PROJECT TITLE : " COMPREHENSIVE ANALYSIS OF CUSTOMER CHURN"
Integrating Clustering and Classification Techniques for Predictive Insights

A) Data Gathering and Integration

Source and Nature of Data: The data for the project was sourced from a file named "Churn_Modelling.csv". This suggests that the focus was on modeling customer churn.

Data Dimensions: The dataset contained 10,000 rows and 14 columns, likely representing individual customers and a range of attributes pertinent to churn analysis. These attributes might have included demographic information, account details, and transaction history.

Integration Aspects: Although not detailed in the brief summary, integration could have involved combining this dataset with other relevant data sources to enrich the analysis, ensuring consistency in format and scale.

B) Data Cleaning and Preprocessing

Addressing Unique and NA Values: The initial step in data preprocessing involved identifying and handling unique value columns and any NA (null or missing) values, which is crucial for maintaining data quality.

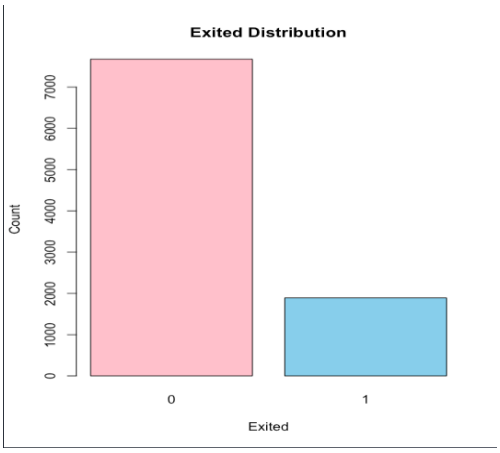
Outlier Detection and Removal: Outliers were detected using the Interquartile Range (IQR) method, particularly in 'CreditScore', 'Age', and 'NumOfProducts' columns. This step is essential to prevent skewed analysis due to extreme values.

Categorical Data Handling: Categorical variables like 'Geography' and 'Gender' were transformed into factor variables. This conversion is necessary for machine learning algorithms to process categorical data effectively.

C) Target Variable Analysis

Analysis of 'Exited' Variable: The target variable 'Exited' presumably indicates whether a customer has churned (left the company). The analysis involved understanding its distribution across the dataset.

Visualization and Distribution: A bar chart was used to visualize the distribution, and the percentages of 'Yes' (churned) and 'No' (retained) were calculated. This step is crucial for understanding the imbalance in the dataset.

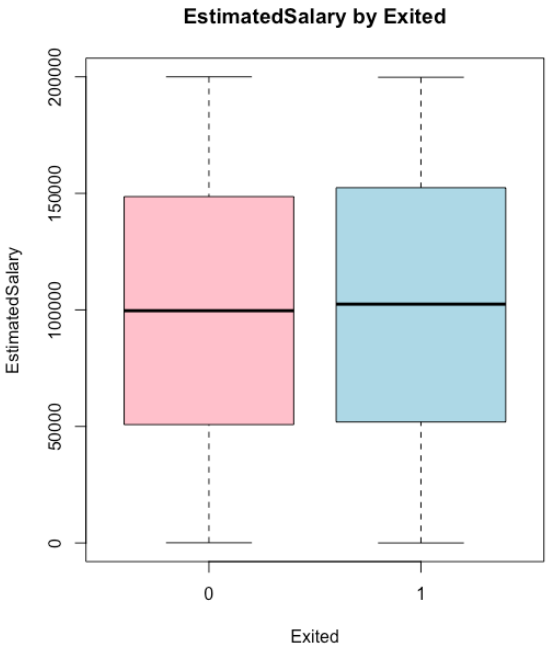


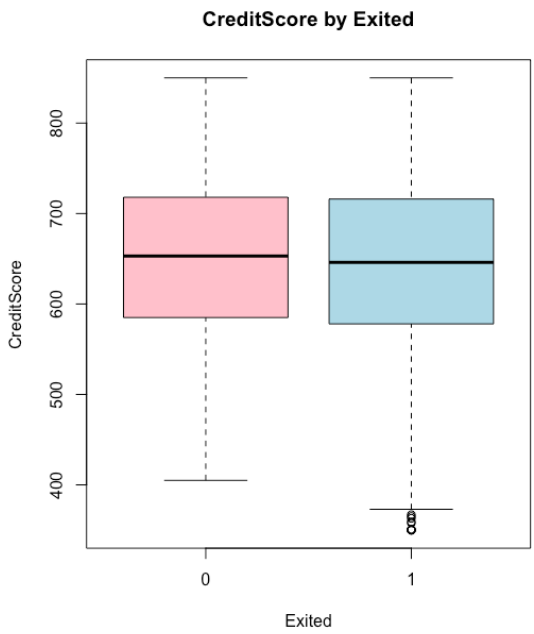
The bar chart shows a higher number of customers staying (represented by '0' in pink) compared to those who have exited (represented by '1' in blue), indicating more customer retention in the dataset.

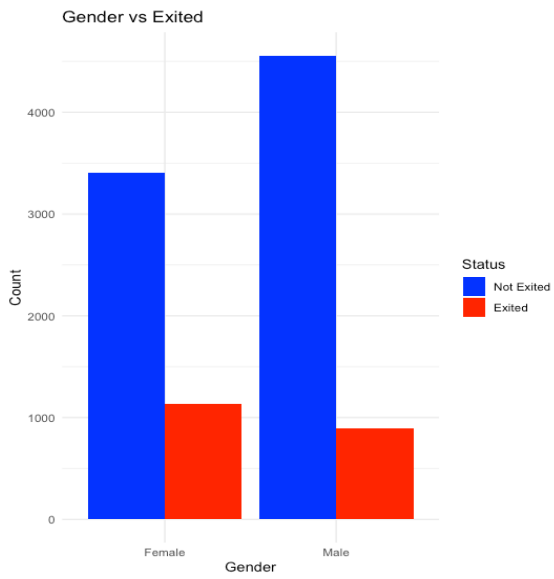
D) Data Exploration - Exploratory Data Analysis (EDA)

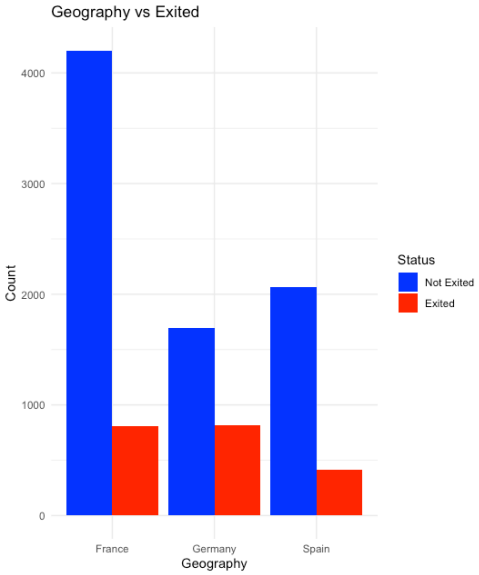
Separation of Data Types: Categorical and numerical data were analyzed separately, which is standard in EDA to tailor the analysis approach to the data type.

Use of Various Plots: Boxplots, histograms, and bar plots were utilized to explore relationships between features (like salary, credit score, balance, age, gender, geography) and the target variable. These visualizations help in identifying trends, patterns, and anomalies.









Correlation Analysis: A correlation matrix was created to identify how different numerical features are related to each other. This step helps in understanding multicollinearity and the potential influence of features on each other.

E) Clustering

Determining Optimal Clusters: The project likely used methods like the elbow method to determine the optimal number of clusters for K-means clustering.

Application of K-means Clustering: This unsupervised learning technique was used to group similar data points together, which can reveal inherent groupings in the data.

Model Evaluation: The model was evaluated using a confusion matrix, accuracy, precision, sensitivity, specificity, ROC curves, and AUC. These metrics provide a comprehensive understanding of the model's performance.

F) Classification

Decision Tree and SVM Models: Both these models were employed for classification. Decision Trees are known for their ease of interpretation, while SVMs are effective in high-dimensional spaces.

Model Training and Prediction: Each model underwent a process of training on a subset of data and then predicting the target variable on another set.

Evaluation Metrics: Accuracy, precision, sensitivity, specificity, and AUC were calculated for each model. ROC curves were plotted to visually assess model performance.

G) Evaluation

Comparative Analysis: A summary table compared the performance of K-means Clustering, Decision Tree, and SVM models across various metrics.

	Model	Accuracy	Precision	Sensitivity	Specificity	AUC
1	K-means Clustering	0.8457919	0.7871287	0.3868613	0.9713715	0.5693359
2	Decision Tree	0.8547241	0.7662062	0.3812797	0.9713430	0.7290228
3	SVM Model	0.8457919	0.7871287	0.3868613	0.9713715	0.8012846

Model Performance: The Decision Tree model showed the highest accuracy, indicating its effectiveness in correctly classifying cases. The SVM model excelled in AUC, suggesting its strength in distinguishing between classes.

This detailed exploration highlights a structured and comprehensive approach to analyzing customer churn, encompassing data preparation, exploratory analysis, clustering, classification, and rigorous evaluation of different models.

H) Reflection

In the "Fundamentals of Data Science" class, I have gained comprehensive knowledge in key areas of data science. I learned about data preprocessing, which involves cleaning and preparing data for analysis, an essential step to ensure accuracy in results. The course also covered advanced machine learning techniques such as Support Vector Machines (SVM) and various clustering algorithms, equipping me with the skills to identify patterns and group data effectively. Additionally, I delved into classification techniques, including the nuances of decision trees and random forests, gaining insights into their applications and strengths in predictive modeling. A significant part of my learning was focused on model evaluation, where I learned to assess the performance of machine learning models, ensuring their reliability and effectiveness in real-world scenarios. This course has been instrumental in building a solid foundation in data science, enhancing both my theoretical understanding and practical skills in this dynamic field.

Reflecting on this course, the project on customer churn analysis has been a pivotal experience in understanding the depth and applicability of data science. It taught me the criticality of thorough data preparation and the power of different analytical techniques like clustering and classification in deriving meaningful insights. This hands-on project bridged the gap between theoretical knowledge and practical application, enhancing my analytical skills and highlighting the importance of storytelling with data.

Furthermore, the project reshaped my perspective on the data scientist's role, emphasizing not just technical proficiency but also ethical responsibility. The experience underscored the importance of clear visualizations and unbiased model creation, instilling a deeper appreciation for the nuanced responsibilities in data science. This journey has equipped me with crucial skills and a profound understanding of the impact and ethical considerations of data-driven decision-making.