**Customer-Churn Analysis of A Multinational Bank**

**Github Link :**

<https://github.com/amireddy51/Customer-Churn-Analysis-of-A-Multinational-Bank>

**Team Members:**

Venkateswarlu AmireddySowjanya Devupalli

Abinay Goud KarnamNavaneeth Ragi

Sharmila Alikapati

**Abstract**

This project addresses the challenge of customer churn in a multinational bank through a comprehensive data analysis and predictive modelling approach. Recognizing the critical impact of customer retention on business growth, the project aims to identify customers at high risk of churn. Leveraging a dataset with key customer attributes, we conduct exploratory data analysis to understand underlying patterns and relationships. The project adopts Logistic Regression, a supervised machine learning technique, to classify customers based on their likelihood to churn. Evaluation metrics like confusion matrix and ROC-AUC curve are used to assess model performance. The outcome of this project is expected to provide actionable insights for the bank to enhance customer retention strategies and reduce churn rates, thereby contributing to sustainable business growth.

**Introduction**

Customer retention appears as a critical aspect for long-term growth and profitability in the banking sector's dynamic and competitive market. Customer churn, the phenomena in which customers discontinue their association with a bank, offers a substantial issue, resulting in revenue loss and higher marketing expenditures. This project investigates the essential issue of customer turnover at a global bank, with the goal of understanding, forecasting, and mitigating the problem.

The changing landscape of consumer expectations, as well as the increased competition in the banking business, highlight the necessity for such an examination. Customers now have more alternatives than ever before because to the rise of digital banking and fintech technologies. Understanding the reasons for client turnover becomes critical in this environment for banks to alter their strategy.

This project makes a contribution to this field by utilizing data analytics and machine learning. We want to uncover trends and predictors of turnover by analyzing customer data such as demographics, transaction history, and account information. To analyze and estimate the possibility of clients terminating their services, the project applies Logistic Regression, a rigorous statistical approach. The findings of this research are meant to help the bank develop focused retention efforts and improve client satisfaction.

In essence, this initiative aims to combine analytical methodologies with actual banking difficulties, providing a data-driven approach to improving client retention and fostering long-term business sustainability in the banking industry.

**Literature Survey**

1. **Decision Tree Technique in Electronic Banking Services**: A study applied for identifying characteristics of churned customers in electronic banking services. This method provided bank managers with insights to identify potential churners and develop strategies to retain them **[1]**.

1. **BiLSTM-CNN** (Scientific Reports): This research introduces a hybrid deep learning model combining (BiLSTM) and (CNN) to predict customer churn. This approach addresses some limitations of traditional ml methods and achieves an accuracy of 81% on benchmark datasets **[2]**.

2. **Importance of Customer Experience in Reducing Churn** (Qualtrics): Emphasizes that poor service is the primary reason for customer churn in banking. The study highlights that many customers believe their banks could have taken actions to retain them, but efforts to understand and address customer needs were lacking. The report also notes that attrition rates in financial services, particularly in organizations with non-binding contracts, can be as high as 25-30% **[3]**.

**Implementation**

Insights about the Dataset used.

1. 'customer\_id': An identifier unique to each customer in the collection.

2. 'credit\_score': A numerical score showing the customer's creditworthiness, which may influence their chance of churning.

3. 'nation': The nation in which the consumer has a bank account.

4. 'gender': The customer's gender.

5. 'age': The customer's age in years.

6. 'Tenure': The number of years the consumer has been a customer of the bank.

7. 'balance': The customer's current bank account balance.

'products\_number': The number of banking products used by the consumer.

9. 'credit\_card': Indicates whether or not the consumer has a credit card (1).

10. 'active\_member': Indicates whether or not the consumer is an active member (1).

11. 'estimated\_salary': The customer's estimated salary.

12. `**churn**`: Indicates customer has churned or not

**Independent Variables**: These are the features that are used to predict the target variable. In this dataset, the independent variables would be all columns except for `**churn**`. They include `**customer\_id**`, `**credit\_score**`, `**country**`, `**gender**`, `**age**`, `**tenure**`, `balance`, `products\_number`, `**credit\_card**`, `**active\_member**`, and `**estimated\_salary**`.

**Dependent Variable**: This is the target variable that the model aims to predict based on the independent variables. In this dataset, the dependent variable is `churn`, which signifies whether a customer has left the bank.

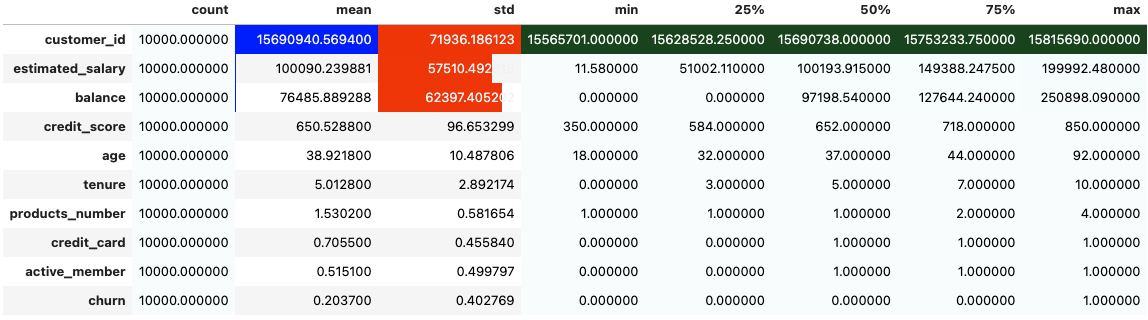


Figure : Statistics about the dataset columns

Overall, this dataset reflects a diverse customer base with varying credit scores, balances, and product usage. The churn rate is a critical figure, as it shows that about one-fifth of the customers in this dataset have churned.

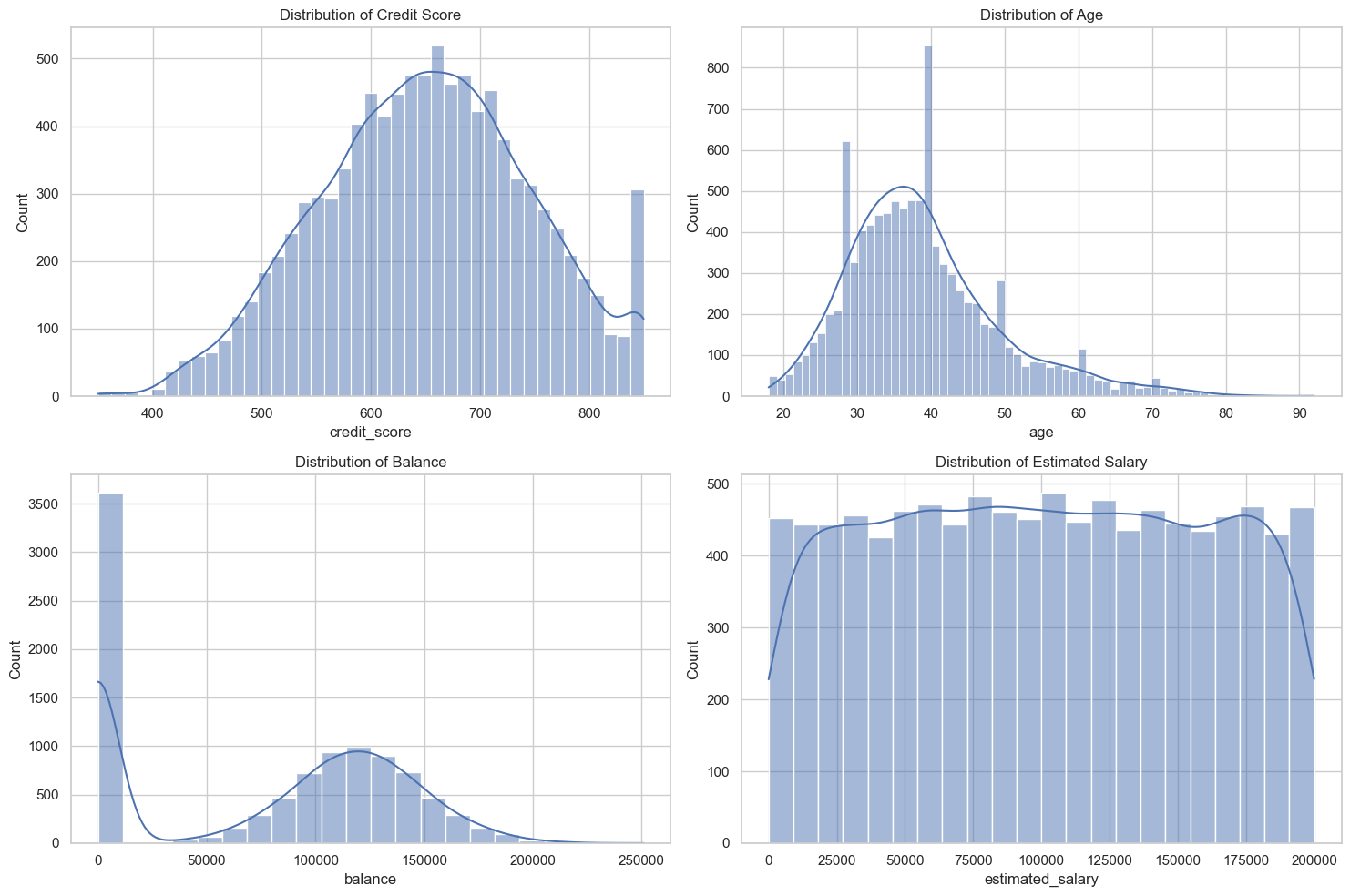
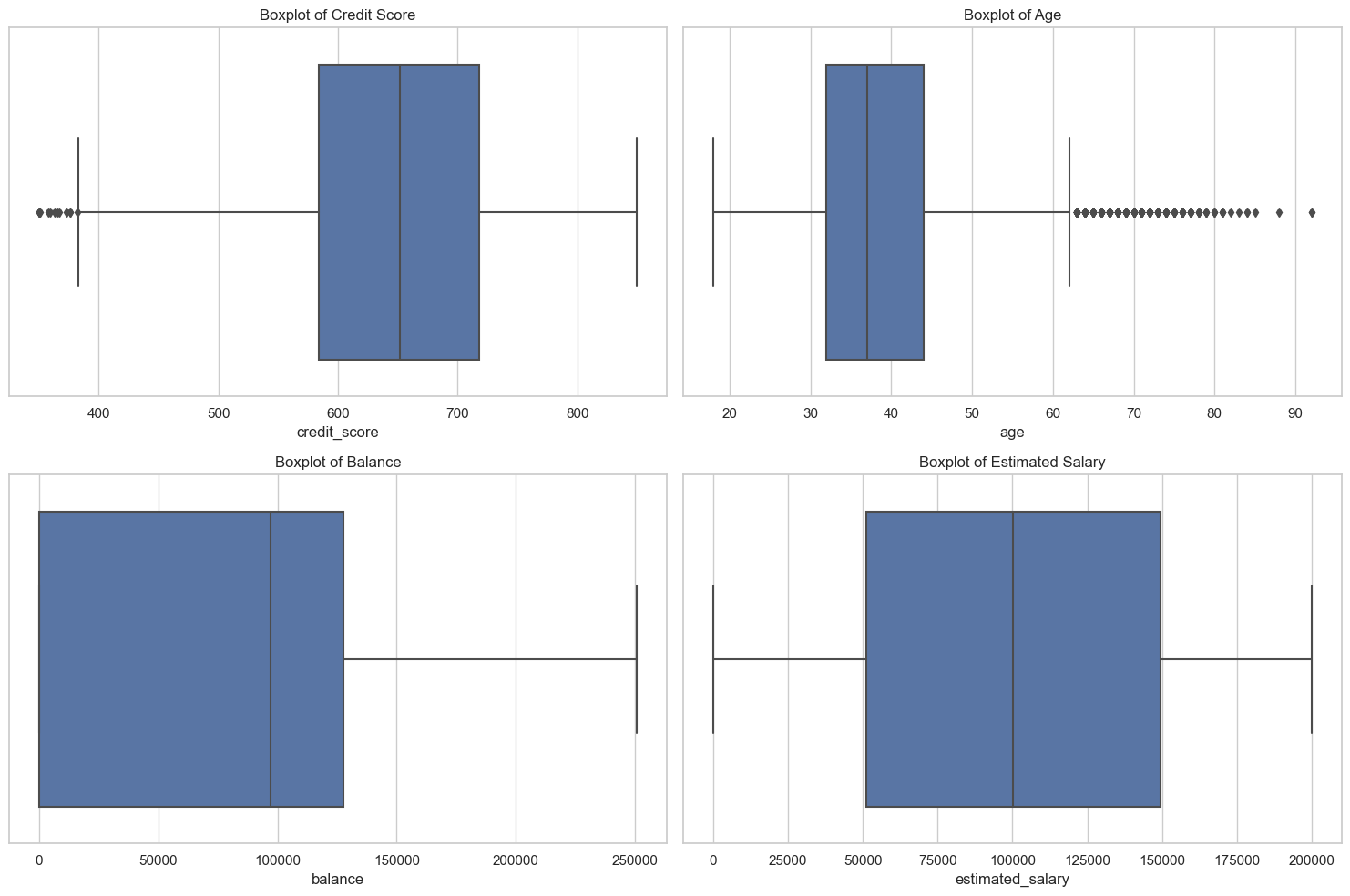
 

Figure 2 distribution credit scores, Age, balance, salary Figure 3 Box plots for age, salary, balance

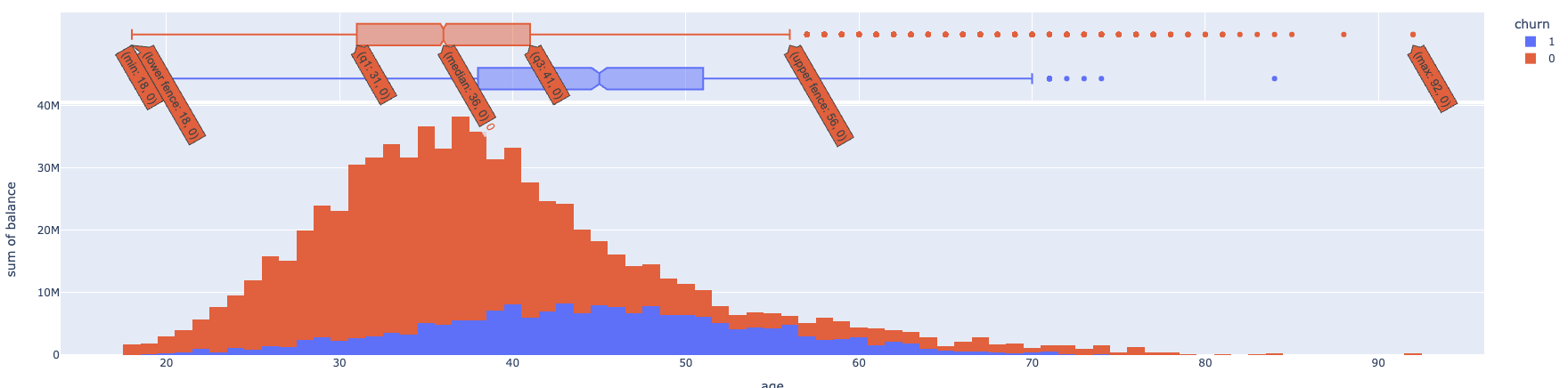


Figure 4: histogram plot for balances sum

**Handling Missing Values**

Before analyzing the correlation heatmap, it's essential to handle missing values in the dataset. Missing data can skew the results and lead to inaccurate interpretations. The common approaches to handle missing values include imputation, where missing data are filled are removed from the dataset. The chosen method should consider the data distribution and the percentage of missing values to preserve the integrity of the dataset.

**Observations from the Heatmap Correlation Diagram**

From the provided correlation heatmap, we observe that `**age**` appears to have a positive correlation with `**churn**`, indicating that older customers may have a higher tendency to churn. Conversely, `**active\_member**` has a negative correlation with `**churn**`, suggesting that active members are less likely to churn. There is also a notable negative correlation between `**balance**` and `**products\_number**`, implying that customers with more products tend to have lower balances. Most variables show little to no correlation with `**customer\_id**`, which is expected as it is just an identifier. Variables like `**credit\_score**`, `**tenure**`, and `**estimated\_salary**` show very low correlation with `**churn**`, indicating that they may have less predictive power in determining the churn status of a customer.

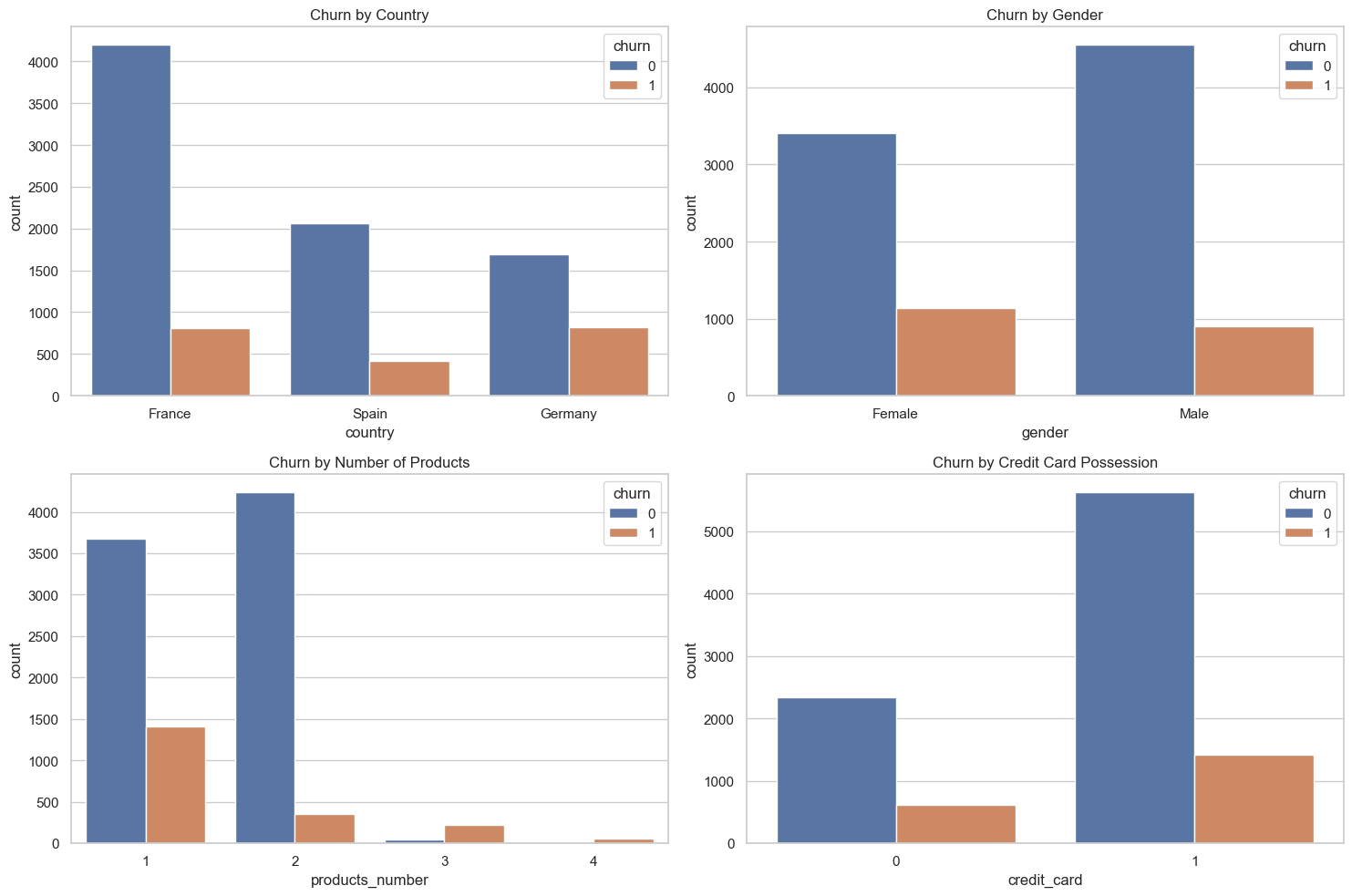


Figure 5 : churn analysis by country, gender, credit card possession from correlation heatmap

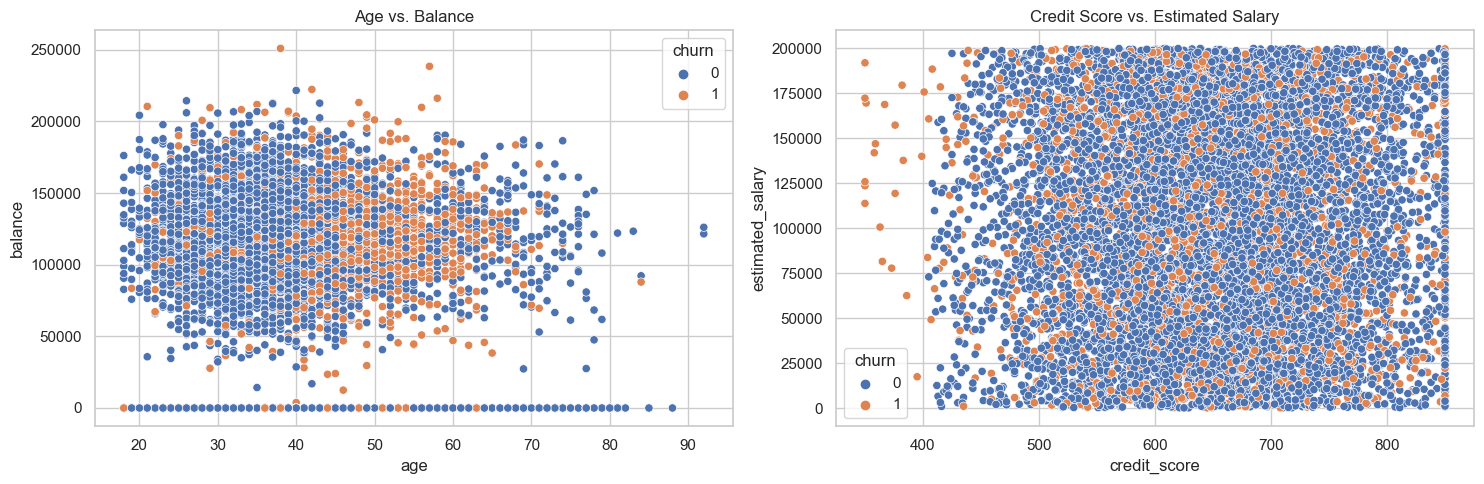


Figure 6: Scatter plots for AgeVsBalance and CreditScoreVsSalary

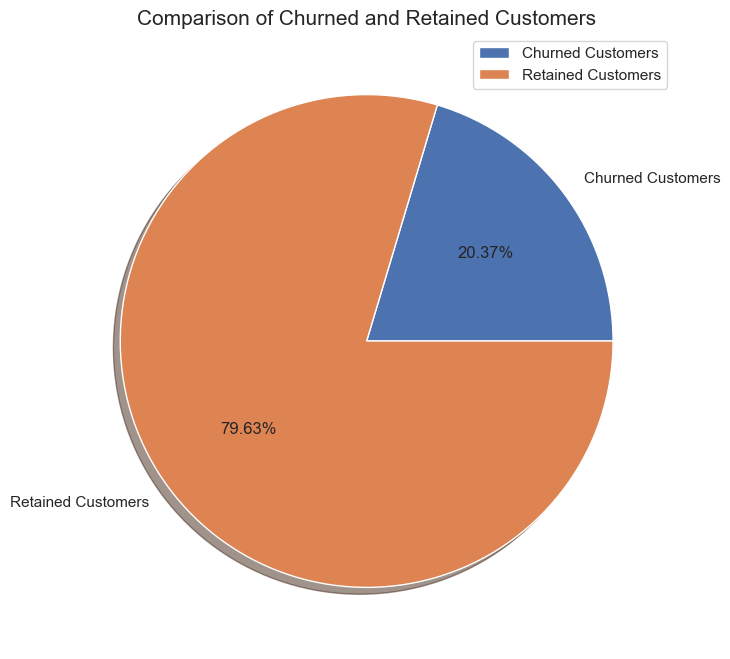
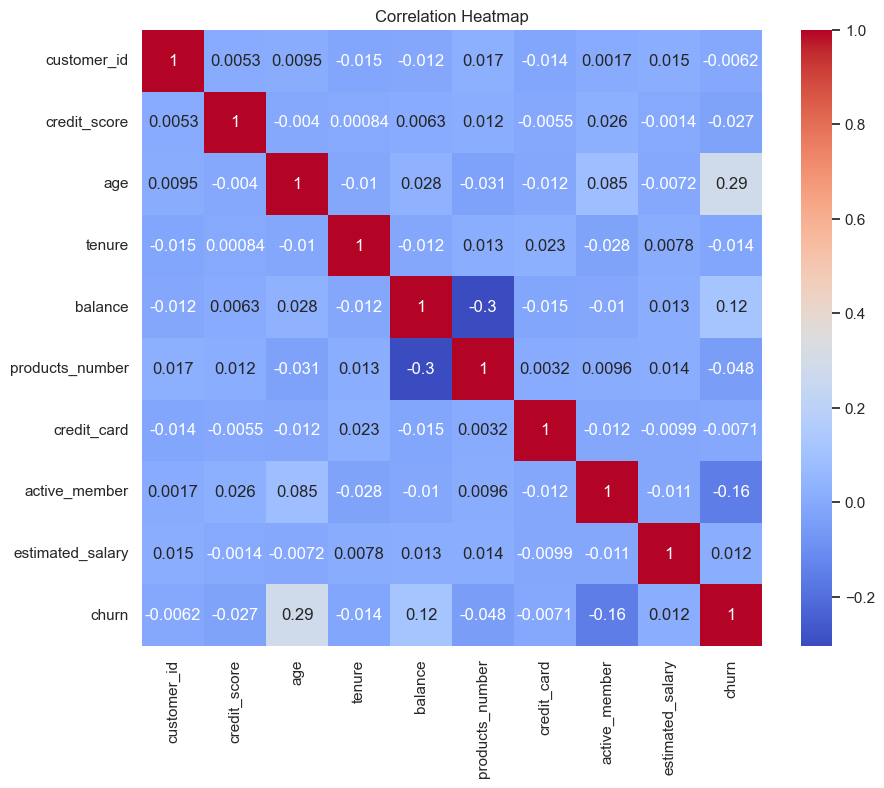
 

Figure 7: Comparing churned and retained customers Figure 8: Correlation Heatmap

**The model preparation and Preprocessing Steps**

1. **Feature Selection**: Initially, the features for the model are determined by dropping irrelevant columns such as `customer\_id`, which is not predictive of churn, and the target variable `churn` itself.

2. **Target Variable**: The target variable is isolated into its own series. In this case, it's the `churn` column indicating whether a customer has churned.

3. **Categorical and Numeric Column Identification**: The dataset is divided into numeric and categorical features. This distinction is necessary because different types of data require different preprocessing techniques.

4. **Numeric Transformation**: Numeric data may contain missing values and will likely have varying scales. To address this, a two-step pipeline is created:

- **SimpleImputer**: Fills missing values with the median value of the column, which is robust to outliers.

- **StandardScaler**: Scales the numeric data to have are sensitive to the scale of input data.

5. **Categorical** Transformation: Categorical features also undergo a two-step pipeline:

- **SimpleImputer**: Fills any missing values with a placeholder string, ensuring that all data points can be used in the model.

- **OneHotEncoder**: Converts categorical variables into a series of binary (0 or 1) variables for each category, a necessary step since most machine learning models cannot handle categorical data directly.

6. **Column Transformer**: The numeric and categorical transformers are combined into a `**ColumnTransformer**`, which applies the appropriate transformations to each column in the dataset.

7. **Data Splitting**: The dataset is split into training and testing sets using `train\_test\_split`, with 80% of the data allocated for training and 20% for testing. This allows for the evaluation of the model on unseen data.

8. **Preprocessing** Application: The `preprocessor` is fitted to the training data, learning any necessary transformations such as the medians of columns and the categories in the categorical data. These transformations are then applied to both the training and testing sets.

9. **Dataset Shapes**: After preprocessing, the shapes of the training and testing sets are displayed, showing that there are 8,000 samples for training and 2,000 for testing, with each having 13 features after preprocessing.

**Model Initialisation and fitting**

In this section we have employed Logistic Regression to model and predict customer churn. Logistic Regression is a commonly used statistical method for binary classification problems like churn prediction.

1. **Model Initialization**: initializing the `LogisticRegression` model with a random state to ensure reproducibility. The random state controls the randomness of the algorithm's initialization.

2. **Model Training**: fitting the logistic regression model to the training data using `fit()`, which trains the model on the preprocessed features `X\_train` and the target variable `y\_train`.

3. **Prediction**: After the model has been trained, you use it to predict the target variable for the test set, generating `y\_pred`.

4. **Evaluation Metrics**:

- Accuracy: The 'accuracy\_score' function compares the predicted 'y\_pred' values to the actual 'y\_test' values.

- Confusion Matrix: The 'confusion\_matrix' offers a breakdown of accurate and wrong predictions by class. The amount of true positives, false positives, true negatives, and false negatives is displayed.

- Classification Report: The classification report provides a more extensive look into the model's performance, including metrics like as accuracy, recall, and f1-score for each class.

5. **Accuracy Result**: The printed accuracy of the model is 0.811, which means that the model correctly predicted whether a customer would churn or not approximately 81.1% of the time on the test data.

This process gives us a quantitative measure of our model's performance. However, looking beyond just accuracy, especially for imbalanced classes, which is common in churn datasets, the confusion matrix and classification report can give more insight into how well the model is identifying the churned customers (sensitivity) and how well it is identifying the non-churned customers (specificity).

**Quantitative Analysis**

In this phase , we have utilized the `**statsmodels**` library to reconstruct the logistic regression model for a more detailed statistical analysis. The `statsmodels` approach is particularly useful for interpretability, as it provides extensive statistics about the model coefficients and their significance.

we started by adding a constant to `X\_train` to account for the intercept in the logistic regression equation. The logistic model is then built with `Logit`, which is fit to the training data. Upon fitting the model, the summary2 method is used to generate a detailed statistical summary of the model's performance.

From the results:

- **Model Fit**: The model has successfully converged after six iterations, indicating that the optimization algorithm was able to find a solution.

- **Pseudo R-squared**: A value of 0.151 is indicative of the model's goodness of fit. While not directly comparable to R-squared from linear regression, this suggests a moderate fit to the data.

- **LLR (Likelihood Ratio Test)**: The LLR p-value is significant, indicating that the model as a whole fits the data better than a model with no predictors.

- **Coefficients**: The table provides the coefficients for the intercept and each feature. Significant predictors have p-values less than 0.05. For example, `x2` has a highly significant positive coefficient, suggesting it's a strong predictor of churn.

- **Standard Errors and Confidence Intervals**: For each coefficient, the standard error and confidence intervals are given, which provide information about the precision of the estimates and the range within which the true parameter values lie.

A graph of a bar graph

Description automatically generated with medium confidence

Figure 9: Impact of each coefficient on the churn

A graph of a bar graph

Description automatically generated

Figure 10: Odds Ratio of variables

**Conclusion**

This project embarked on the task of predicting customer churn for a multinational bank using a data-driven approach. Through exploratory data analysis, we gained insights into the factors influencing customer behavior. The dataset was meticulously preprocessed, handling missing values and encoding categorical variables to ensure a robust input for model training.

Methodology Used We employed logistic regression, a sophisticated statistical method for binary classification issues, to estimate churn probability.

The model was trained on numeric and categorical data using a standardized workflow that comprised feature scaling and transformation. The statsmodels package allowed for sophisticated statistical analysis, offering a comprehensive overview of model performance and the significance of different attributes.

The outcomes were achieved The logistic regression model has an accuracy of about 81 percent. 1% on the test set, indicating a strong ability to forecast turnover.

According to statistical analysis, various predictors have specific impacts, as indicated by significant coefficients.

Variables like "age" and "active membership status" revealed substantial associations with turnover scores.

The pseudo R-squared values revealed a decent fit, and likelihood ratio tests verified the model's overall validity vs the null model.

The existence of 'Nan' values in the summary result, on the other hand, necessitates careful assessment of data quality and model assumptions.

**Future Work Scope**: Future work may look into many methods to enhance this project. Advanced Modeling: Mechanisms such as Random Forests and Gradient Boosting are examples of advanced modeling.

The use of learning algorithms can increase prediction performance, particularly when nonlinear interactions are present.

- Feature Engineering: Creating new features or altering current features adds insight and increases your model's prediction capabilities.

- Model Interpretability: Tools like SHAP can analyze complicated models and deliver meaningful business insights.

- Correcting class imbalances: Techniques like as SMOTE (Synthetic Minority Oversampling Technique) and target cost functions have the ability to successfully handle Churn distribution imbalances.

- Customer Segmentation: Segmentation analysis reveals certain customer segments who are more likely to defect, resulting in tailored retention measures.

- Temporal Dynamics: Taking into account temporal factors such as consumer behavior over time gives a dynamic viewpoint on churn prediction.

In conclusion, this research establishes a solid basis for predicting customer churn and identifies clear development possibilities that may be followed to further improve the model and its applicability in real-world settings.

**Contribution:**

11677056 - Venkateswarlu Amireddy: 20% - He will completely participate in exploratory data analysis and data preparation. He will assist with the documents.

11717210 - Sowjanya Devupalli: 20% - She will create the model and tune the hyperparameters on the sample data set. She will help with documents. Helped about the work to be done according to scope for future works.

11564953 - Navaneeth Ragi: 20% - He will assess the overall model's quality and perform significance tests (G-test) on the overall model and z-tests on each independent variable. Will also do an odds ratio analysis.

11665110 - Abinay Goud Karnam: 20% - He will apply the assessment measures to the constructed model and Quantitative analysis. This will aid with documenting.

11702541 - Sharmila Alikapati: 20% - She will technically help with hyper parameter tuning. She will also create the project's essential visuals.

**References :**

[1] Using data mining, create a prediction model for client turnover from electronic banking services.

<https://jfin-swufe.springeropen.com/articles/10.1186/s40854-016-0029-6>

[2] Prediction of customer turnover using a composite deep learning approach <https://www.nature.com/articles/s41598-023-44396-w>

[3] Reducing customer churn for banks and financial institutions. [Article link](https://www.qualtrics.com/blog/customer-churn-banking/#:~:text=,reason%20for%20bank%20customer%20churn)