Customer-Churn Analysis of A Multinational Bank

Github Link :

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

Team Members:

Venkateswarlu Amireddy
Abinay Goud Karnam
Sharmila Alikapati
Sowjanya Devupalli
Navaneeth Ragi

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 the decision tree technique to build a model 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 Hybrid Model for Customer Churn Prediction** (Scientific Reports): This research introduces a hybrid deep learning model combining Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Network (CNN) to predict customer churn. This approach addresses some limitations of traditional machine learning 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

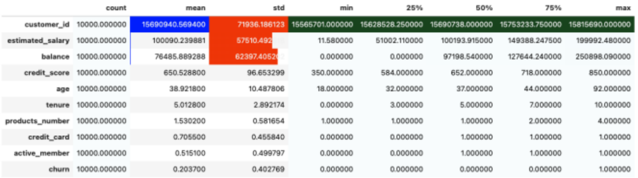
Insights about the Dataset used.

1. `**customer_id**`: A unique identifier for each customer in the dataset.
2. `**credit_score**`: A numerical score indicating the creditworthiness of the customer, which may affect their likelihood of churning.
3. `**country**`: The country where the customer's bank account is located.
4. `**gender**`: The gender of the customer (Female/Male).
5. `**age**`: The age of the customer in years.
6. `tenure`: The number of years the customer has been with the bank.
7. `**balance**`: The current balance in the customer's bank account.
8. `**products_number**`: The number of banking products the customer is using.
9. `**credit_card**`: Indicates whether the customer has a credit card (1) or not (0).
10. `**active_member**`: Indicates whether the customer is an active member (1) or not (0).
11. `**estimated_salary**`: The estimated salary of the customer.
12. `**churn**`: Indicates whether the customer has churned (1) or not (0).

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.



	count	mean	std	min	25%	50%	75%	max
customer_id	10000.000000	15690940.569400	71936.186123	15565701.000000	15628528.250000	15690738.000000	15753233.750000	15815690.000000
estimated_salary	10000.000000	100090.239881	57510.492	11.580000	51002.110000	100193.915000	149388.247500	199992.480000
balance	10000.000000	76485.889288	62397.4052	0.000000	0.000000	97198.540000	127644.240000	250898.090000
credit_score	10000.000000	650.528800	96.653299	350.000000	584.000000	652.000000	718.000000	850.000000
age	10000.000000	38.921800	10.487806	18.000000	32.000000	37.000000	44.000000	92.000000
tenure	10000.000000	5.012800	2.892174	0.000000	3.000000	5.000000	7.000000	10.000000
products_number	10000.000000	1.530200	0.581654	1.000000	1.000000	1.000000	2.000000	4.000000
credit_card	10000.000000	0.705500	0.455840	0.000000	0.000000	1.000000	1.000000	1.000000
active_member	10000.000000	0.515100	0.499797	0.000000	0.000000	1.000000	1.000000	1.000000
churn	10000.000000	0.203700	0.402769	0.000000	0.000000	0.000000	0.000000	1.000000

Figure 1: 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.

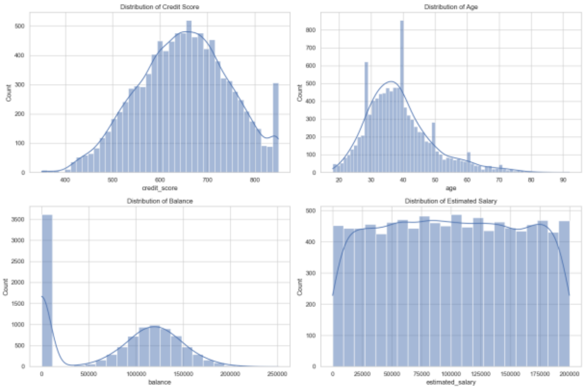


Figure 2 distribution credit scores, Age, balance, salary

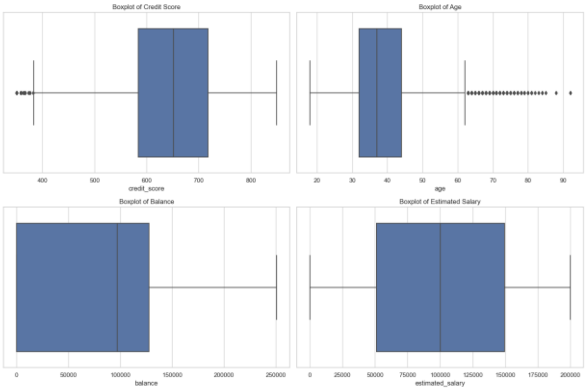


Figure 3 Box plots for age, salary, balance

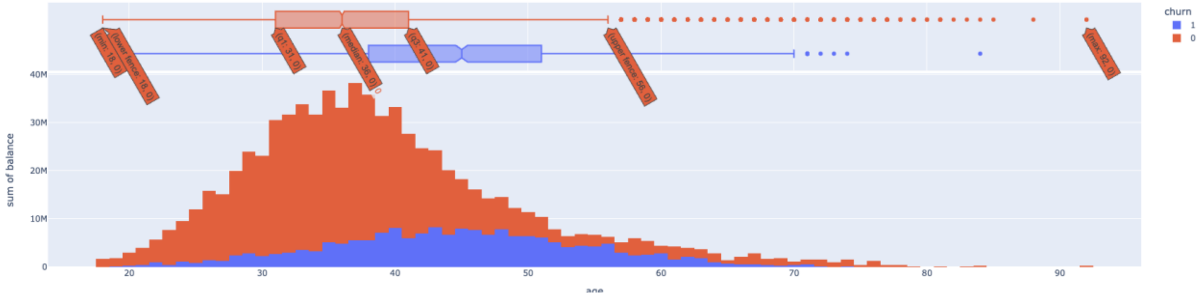


Figure 4: histogram plot for balances sum

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 with statistical measures like mean, median, or mode; and deletion, where rows or columns with missing values 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.

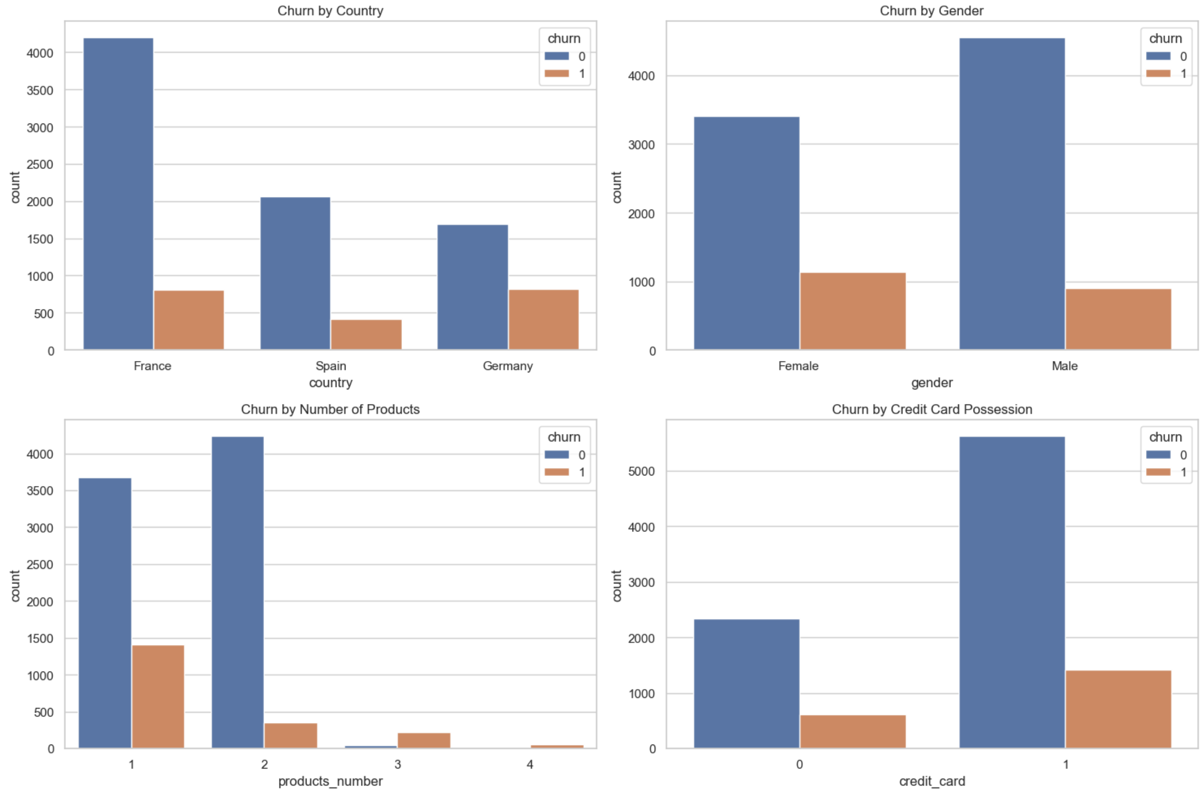


Figure 5 : churn analysis by country, gender, credit card possession from correlation heatmap

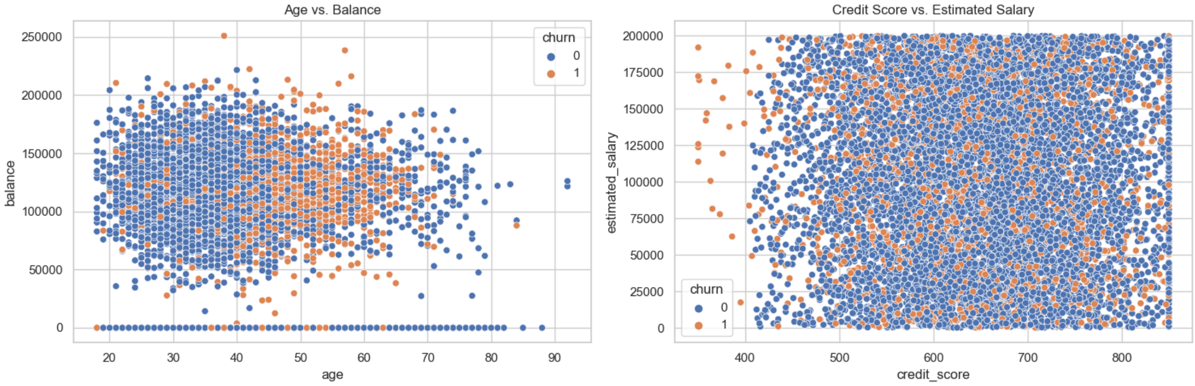
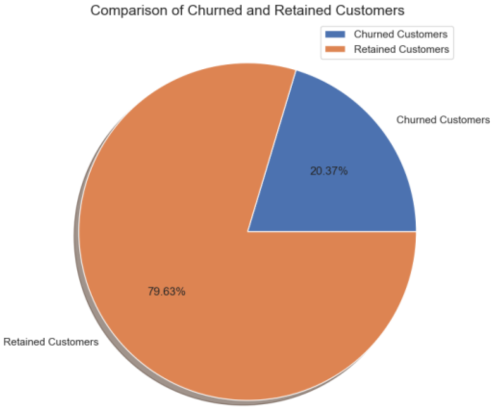


Figure 6: Scatter plots for AgeVsBalance and CreditScoreVsSalary



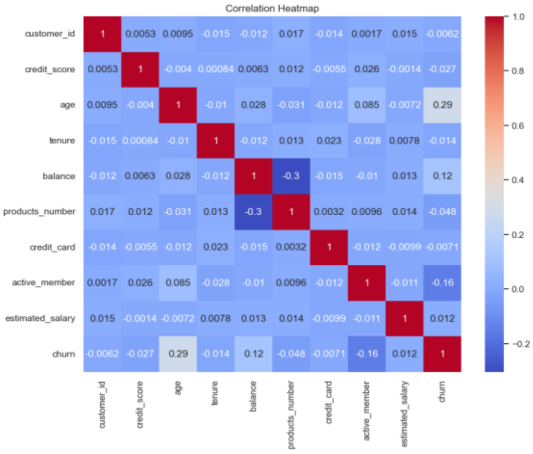


Figure 7: Comparing churned and retained customers

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 a mean of 0 and a standard deviation of 1, which is important for many machine learning models that 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**: You calculate the accuracy of the model, which is the proportion of correct predictions out of all predictions made. The `accuracy_score` function compares the predicted values `y_pred` with the actual values `y_test`.

- **Confusion Matrix**: The `confusion_matrix` provides a summary of correct and incorrect predictions broken down by each class. It shows the number of true positives, false positives, true negatives, and false negatives.

- **Classification Report**: The report gives a more detailed look at the performance of the model, including metrics such as precision, 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).

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.

Future Work:

Model Initialisation and fitting by using the Logistic Regression, as it is classification problem after getting accuracy.

We are doing Quantitative Analysis for our problem. So we also want to find the model fit by accuracy, precision, R squared, LLR. And also want to give scope for future works like Advanced Modelling, Model Interpretability, Customer Segmentation.

Contribution:

Venkateswarlu Amireddy – 11677056: 20% - He will engage fully in exploratory data analysis and data preprocessing. He will support in the documentation.

Sowjanya Devupalli – 11717210: 20% - She will develop the model and do the hyper parameter tunning on the sample data set. She will assist in documentation.

Navaneeth Ragi -11564953: 20% - He will check the goodness of the overall model and do the significance tests (G-test) on overall model and z-test on each independent variable. Also, will do the Odds ratio analysis.

Abinay Goud Karnam – 11665110: 25% - He will perform the evaluation metrics on the model developed. Will help in documentation.

Sharmila Alikapati – 11702541: 15% - She will technically contribute to the hyper parameter tunning. She will also generate necessary visualizations for the project. And compiles all the codes does the documentation by taking assistance from others.

References :

[1] Developing a prediction model for customer churn from electronic banking services using data mining https://jfin-swufe.springeropen.com/articles/10.1186/s40854-016-0029-6

[2] Customer churn prediction using composite deep learning technique https://www.nature.com/articles/s41598-023-44396-w

[3] Reducing customer churn for banks and financial institutions. Article link