Bank Customer Churn Prediction Using Machine Learning Models

Submission Date: 28.11.2025

Submitted by

Group Name:

PySleuth

Submitted to

Farhana Akter Bina

Assistant Professor

Department of Statistics & Data Science

Jahangirnagar University

1. Introduction

1.1 Background

Customer churn-when clients discontinue a bank's services-poses a long-term threat to revenue stability and operational efficiency. In modern retail banking, churn prediction has become a critical analytical function due to increasing competition, evolving customer expectations, and the shift toward digital banking. Leveraging data science enables banks to anticipate at-risk customers, understand the underlying drivers of churn, and design retention strategies grounded in evidence rather than intuition.

This study focuses on **predicting customer churn using demographic variables, account activity metrics, and card usage behaviour**. By combining exploratory analysis with supervised machine learning (specifically **Multiple Linear Regression for explanatory insights** and **Logistic Regression for churn prediction**), the study aims to both interpret customer behaviour and build a functional predictive model.

1.2 Objectives

The major objectives of this project are:

1. To perform Exploratory Data Analysis (EDA) to describe customer demographics and behavioural features associated with churn.
2. To apply **Multiple Linear Regression (MLR)** to understand relationships among continuous usage variables.
3. To build and evaluate a **Logistic Regression model** for predicting customer churn.
4. To generate actionable insights for retention strategies based on model outcomes.

2. Methodology

2.1 Data Collection

This study uses the Credit Card Customers (BankChurners) dataset from Kaggle, which contains demographic, behavioral, and transaction-based attributes for thousands of credit-card customers (Goyal, 2020). Each record includes a mix of numerical and categorical features along with the binary target label Attrition_Flag, indicating whether the customer has closed their account.

The data consists of 10,000 customers with the following 21 features:

CLIENTNUM: Client number. Unique identifier for the customer holding the account
Attrition_Flag: Internal event (customer activity) variable - if the account is closed then 1 else 0
Customer_Age: Customer's Age in Years
Gender: M=Male, F=Female
Dependent_count: Number of dependents
Education_Level: Educational Qualification of the account holder (example: high school, college graduate, etc.)

Marital_Status: Married, Single, Divorced, Unknown
Income_Category: Annual Income Category of the account holder (< 40K, 40K - 60K, 60K− 80K, 80K− 120K, >120K)
Card_Category: Product Variable - Type of Card (Blue, Silver, Gold, Platinum)
Months_On_Book: Period of relationship with bank
Total_Relationship_Count: Total no. of products held by the customer
Months_Inactive_12_mon: No. of months inactive in the last 12 months
Contacts_Count_12_mon: No. of Contacts in the last 12 months
Credit_Limit: Credit Limit on the Credit Card
Total_Revolving_Bal: Total Revolving Balance on the Credit Card
Avg_Open_To_Buy: Open to Buy Credit Line (Average of last 12 months)
Total_Amt_Chng_Q4_Q1: Change in Transaction Amount (Q4 over Q1)
Total_Trans_Amt: Total Transaction Amount (Last 12 months)
Total_Trans_Ct: Total Transaction Count (Last 12 months)
Total_Ct_Chng_Q4_Q1: Change in Transaction Count (Q4 over Q1)
Avg_Utilization_Ratio: Average Card Utilization Ratio

2.2 Data Preprocessing

Robust preprocessing was conducted to ensure model reliability and minimize noise.

Data Cleaning

- Removed duplicate observations and records with inconsistent entries.
- Addressed missing values using mean/median imputation (numerical) and mode imputation (categorical).
- Handled outliers in variables such as transaction amount and credit limit using IQR filtering or winsorization when necessary.

Encoding

Categorical variables (e.g., Gender, Card Category, Income Category) were transformed using:

- **One-Hot Encoding** for non-ordinal variables.
- **Ordinal Encoding** where category order existed (e.g., education or income brackets).

Scaling

- StandardScaler or Min-Max Scaling was applied to continuous variables to ensure model stability.
- Scaling was especially important for Logistic Regression to avoid coefficient distortion.

The processed dataset was then split into **Training (80%)** and **Testing (20%)** partitions.

Implementation Workflow

The preprocessing logic was operationalized through a structured and containerized technical pipeline to ensure reproducibility, consistency, and environment isolation across different systems:

- **Python libraries** were imported to support data handling, statistical preprocessing, and machine learning operations.
- A **MySQL connector** established a secure and reliable interface for retrieving data from the database.
- The entire analytical environment- Python, dependencies, MySQL connectors, and preprocessing scripts-was **containerized using Docker**, ensuring that the preprocessing workflow executed identically across different machines without dependency conflicts.
- **Raw datasets** were stored in the MySQL database for centralized management, version control, and auditability.
- **Datasets were loaded and processed** separately for Linear Regression and Logistic Regression tasks, allowing each model to receive a tailored preprocessing pipeline.
- The **final processed datasets** were saved back into the MySQL database, enabling reproducible modeling and streamlined integration with future analytical stages.

2.3 Analysis Techniques

Exploratory Data Analysis (EDA)

The initial exploration highlights several patterns that shape our understanding of the dataset before we examine customer attrition directly. The demographic and account-usage variables show stable, predictable distributions, reflecting a relatively mature customer portfolio. Most customers fall within a mid-career age range and exhibit consistent account tenure, suggesting that the dataset captures established long-term relationships rather than newly opened accounts.

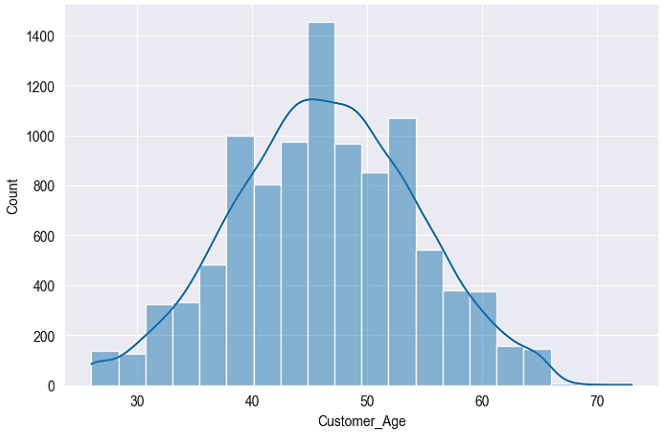

Fig 1: Customer Age Distribution

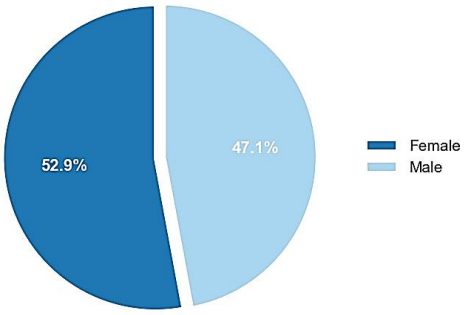

Fig 2: Gender Distribution

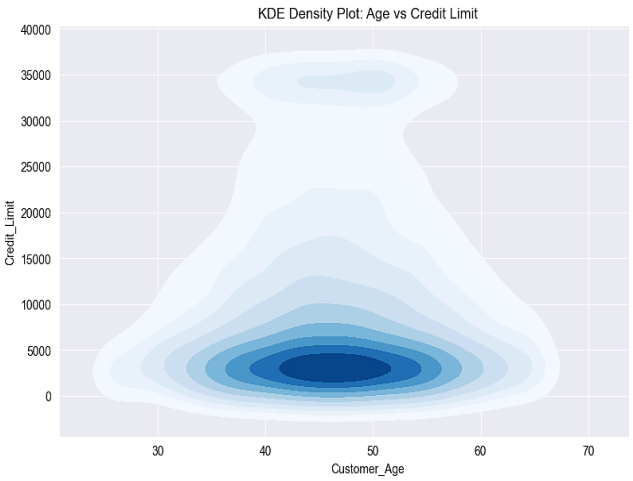

Fig 3: Age vs Credit Limit

Fig 4: Credit Limit Distribution by Gender

The KDE plot (Fig. 3) depicting the joint distribution of customer age and credit limit indicates that the highest concentration of customers lies within the age range of approximately 35–50 years with moderate credit limits between 2,000 and 6,000. This age group constitutes the bank's core customer base in terms of both volume and credit exposure. Customers with very high credit limits (above 30,000) are relatively few and are predominantly concentrated within the middle-aged segment, suggesting that high-value clients represent a small but financially significant portion of the portfolio. Younger and older customers are comparatively underrepresented in higher credit limit categories.

The violin plot (Fig. 4) illustrating credit limit distribution by gender reveals noticeable differences between male and female customers. Male customers exhibit a wider dispersion of credit limits and account for the majority of high-limit observations, whereas female customers are primarily concentrated in lower to moderate credit limit ranges. The central tendency of credit limits also appears slightly higher for male customers than for female customers.

From a churn prediction perspective, these findings suggest that middle-aged customers with moderate credit limits represent the segment where churn would have the greatest aggregate impact due to their numerical dominance. Conversely, churn among high-limit customers, although less frequent, may result in disproportionately large financial losses. The observed gender-based variation in credit limit distribution further implies that demographic characteristics interact with financial behaviors and should therefore be considered as relevant explanatory variables in the churn prediction model.

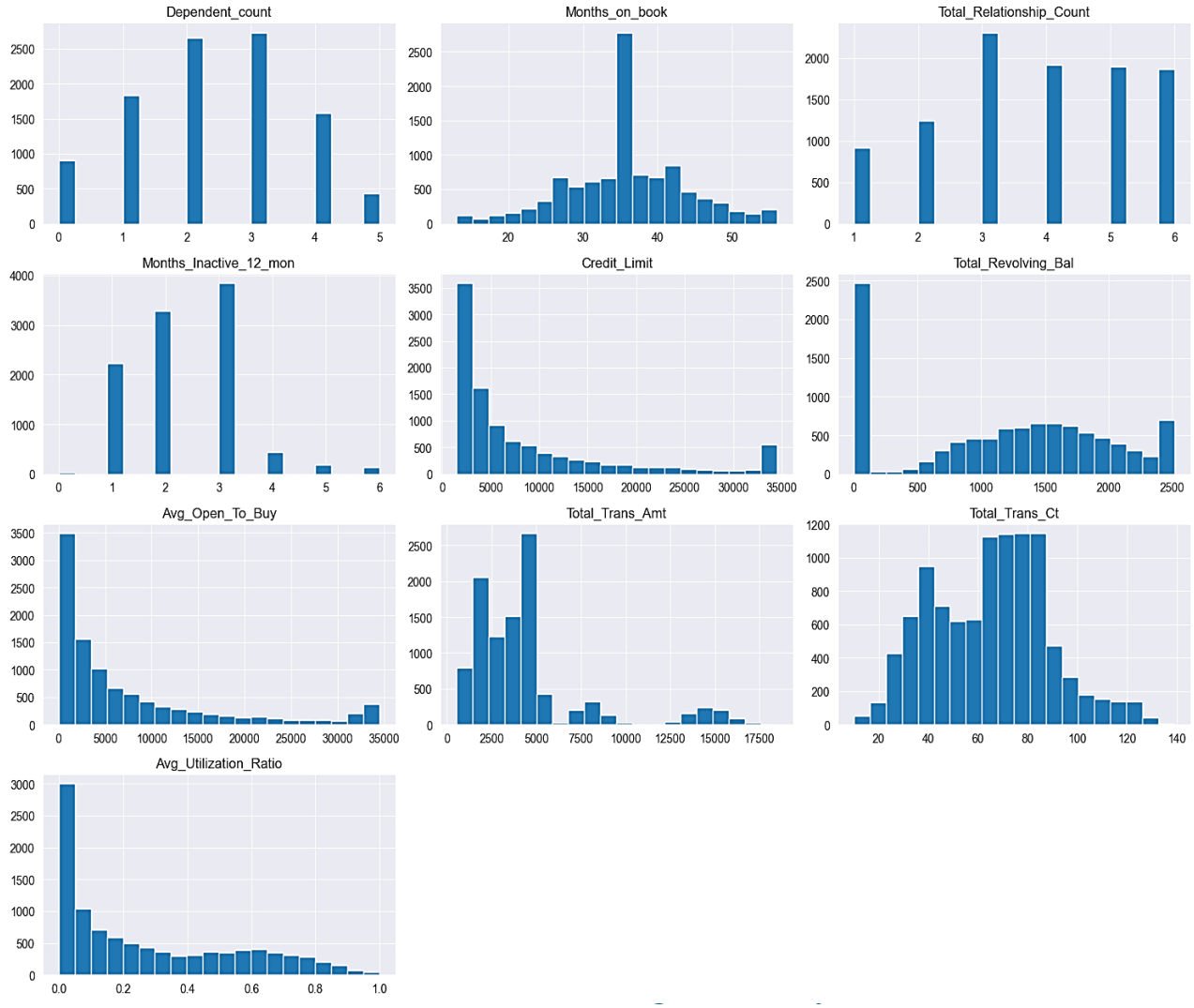

Fig 5: Distribution of all Numeric Features

From Fig. 5, it can be observed that behavioral and transactional features show meaningful variability across the customer base. Transaction counts and transaction amounts follow familiar financial patterns, with most customers exhibiting moderate activity and a smaller

group showing higher engagement. This variation provides a useful signal for distinguishing differences in spending and account-usage behavior.

Credit limits and utilization ratios also range widely, indicating diverse spending capacities and financial habits. These features are often correlated with transactional activity and may play an important role in understanding which customers remain engaged versus those who disengage over time.

Attrition_Flag	Count	Percentage (%)
Existing Customer	8500	83.93
Attrited Customer	1627	16.07

Table 1: Attrition_Flag Distribution

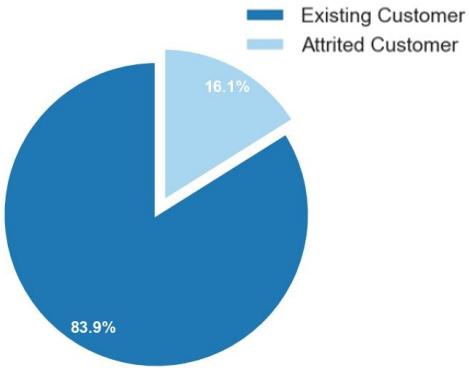

Fig 6: Customer Attrition Flag Distribution

The Attrition_Flag distribution shows that most customers in the dataset remain active, while a smaller but meaningful portion has attrited. This imbalance is expected in customer retention datasets, where churn events occur less frequently than continued account activity. The majority class represents existing customers, accounting for roughly 84 percent of the data, while attrited customers make up the remaining 16 percent.

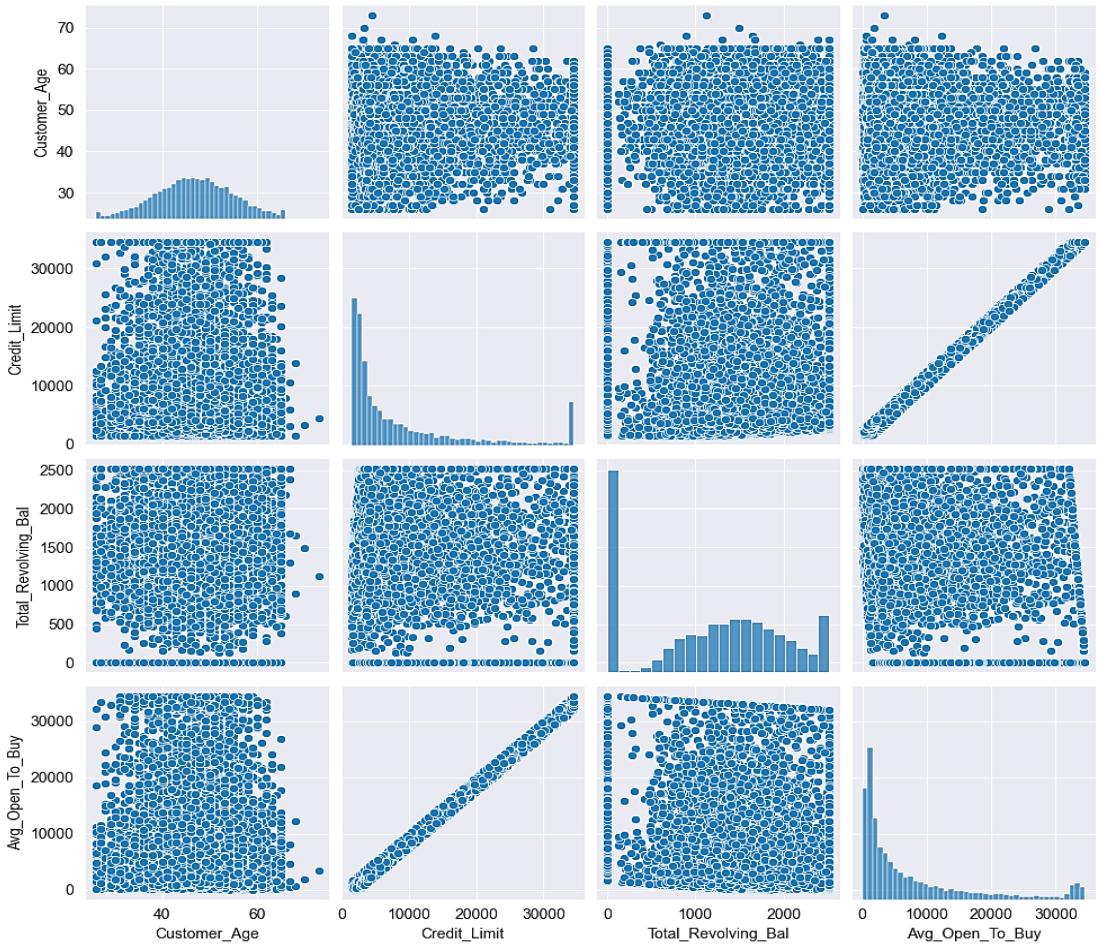

Fig 7: Pairplot of Customer_Age, Credit_Limit, Total_Revolving_Bal, Avg_Open_To_Buy

The pairplot (Fig. 7) presents a comprehensive visualization of the relationships between key continuous variables in the study.

The diagonal histograms reveal the distributional characteristics of each variable. Customer_Age exhibits an approximately normal distribution centered around 45-50 years, suggesting a mature customer base. Credit_Limit and Avg_Open_To_Buy display right-skewed distributions with substantial concentration at lower values, indicating that most customers maintain relatively modest credit limits. Total_Revolving_Bal shows a pronounced spike near zero, suggesting a significant proportion of customers carry minimal revolving balances.

The scatter plots in the off-diagonal cells illustrate several critical relationships. A perfect linear correlation exists between Credit_Limit and Avg_Open_To_Buy, which is mathematically expected since available credit is derived from the credit limit minus the revolving balance. The scatter plots between Customer_Age and other variables show relatively uniform, cloud-like patterns with no visible relationships, indicating that age alone may not be a strong predictor of credit behavior or churn. Similarly, Total_Revolving_Bal exhibits weak correlations with most other variables, though it shows slight clustering patterns that warrant further investigation. Therefore, we need to eliminate the features Avg_Open_To_Buy and Total_Revolving_Bal to prevent the potential data leakage.

Notably, the absence of strong bivariate relationships for most variable pairs suggests that churn prediction may require more sophisticated multivariate modeling approaches, as simple linear relationships between individual predictors appear limited. This observation underscores the necessity of employing a logistic regression model. However, initially, we will use a Multiple Linear Regression model to gather insights to build a more sophisticated logistic regression model.

Multiple Linear Regression – Insight Model

As discussed in the EDA, if we keep the leaking features and perform linear regression, we will get the coefficient/ feature importance as shown in (Fig 8).

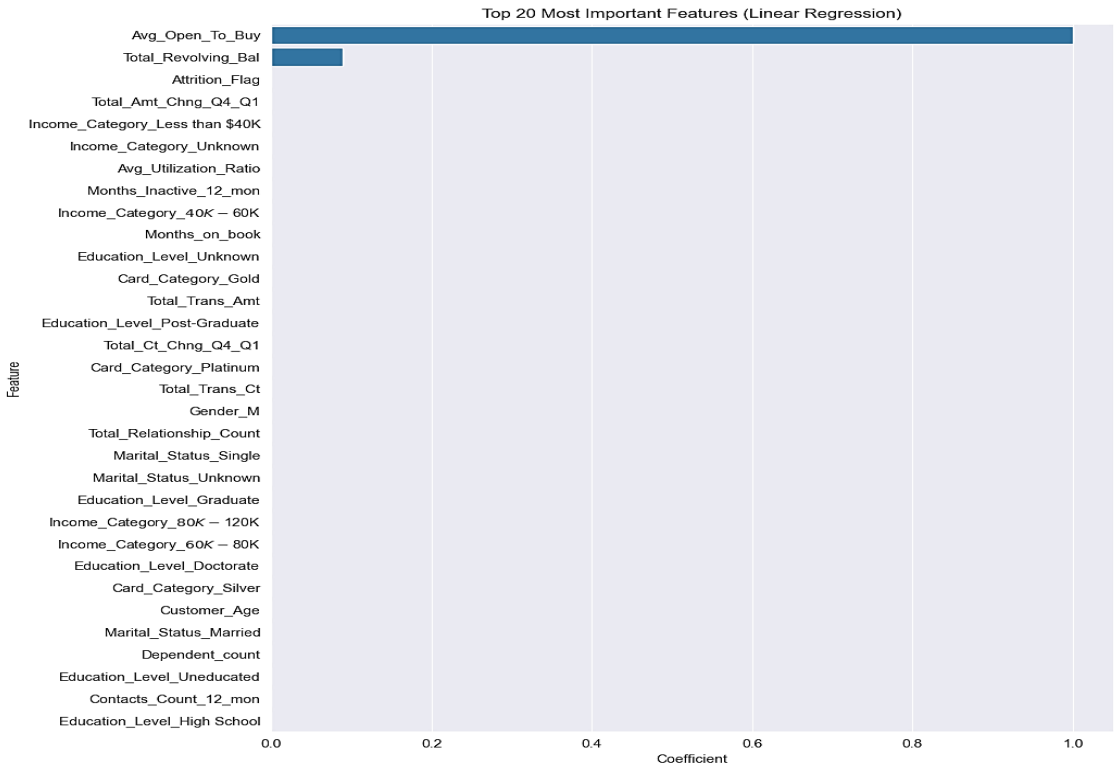

Fig 8: 20 Most Important Features (Linear Regression.)

From (Fig. 8), we can see that Avg_Open_To_Buy and Total_Revolving_Bal have the highest coefficient, while the other coefficients are essentially zero. This happened because in the dataset, Credit_Limit = Avg_Open_To_Buy + Total_Revolving_Bal, which is not a behavioral or predictive feature, but instead a direct mathematical derivative of the target variable.

This confirms that no other feature has any meaningful predictive power once the two leakage variables are included. The model is not learning customer behavior, income, age, transaction patterns, etc. — it's just reconstructing the target from its own components.

This is why we need to drop these two leakage features, which are causing the leaks in our data and retrain our model accordingly.

After retraining our model without the leaks, we evaluated the new model's performance metrics like R-squared, Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The values of these metrics are shown below, and the coefficients of this model are shown in Fig.9,

MAE: 0.4488
MSE: 0.3819
RMSE: 0.6180
R^2: 0.6190

The R-squared value of 0.6190 indicates that approximately 61.9% of the variance is explained by the included predictor variables, suggesting a reasonably good model fit. However, this also implies that nearly 38% of the variance remains unexplained, indicating the presence of additional factors or non-linear relationships not captured by the linear model.

The error metrics provide insight into prediction accuracy. The Mean Absolute Error (MAE) of 0.4488 suggests that, on average, predictions deviate from actual values by approximately 0.45 units. The Root Mean Squared Error (RMSE) of 0.6180 is notably higher than the MAE, suggesting the presence of some larger prediction errors that disproportionately influence model performance. The Mean Squared Error (MSE) of 0.3819 further corroborates these findings.

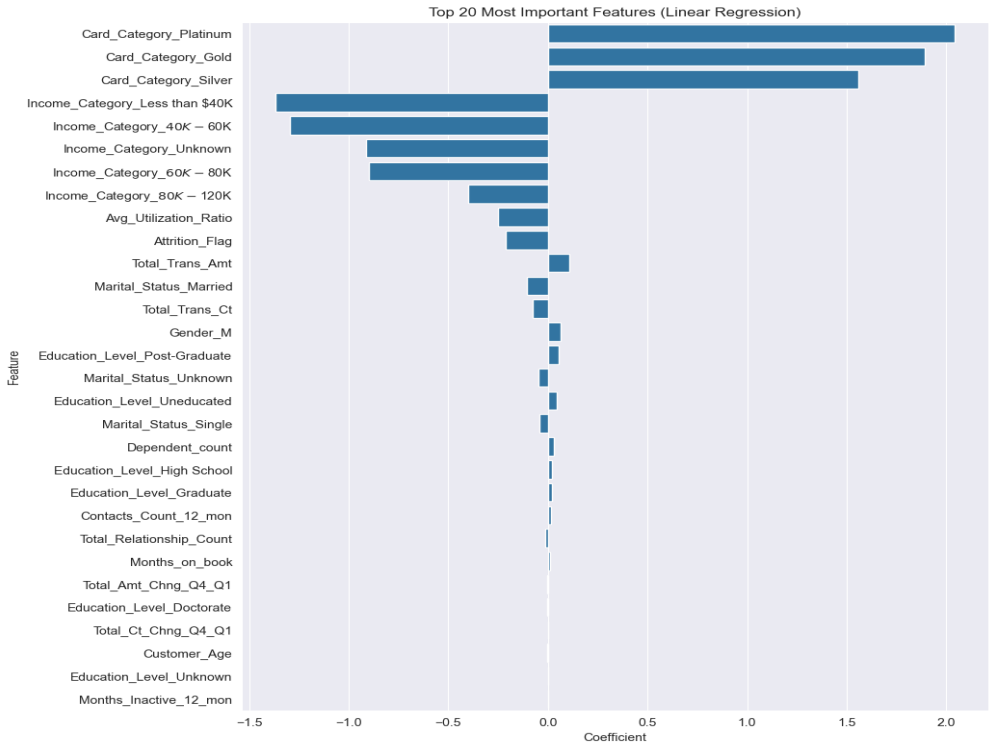

Fig 9: 20 Most Important Features (Linear Regression.)

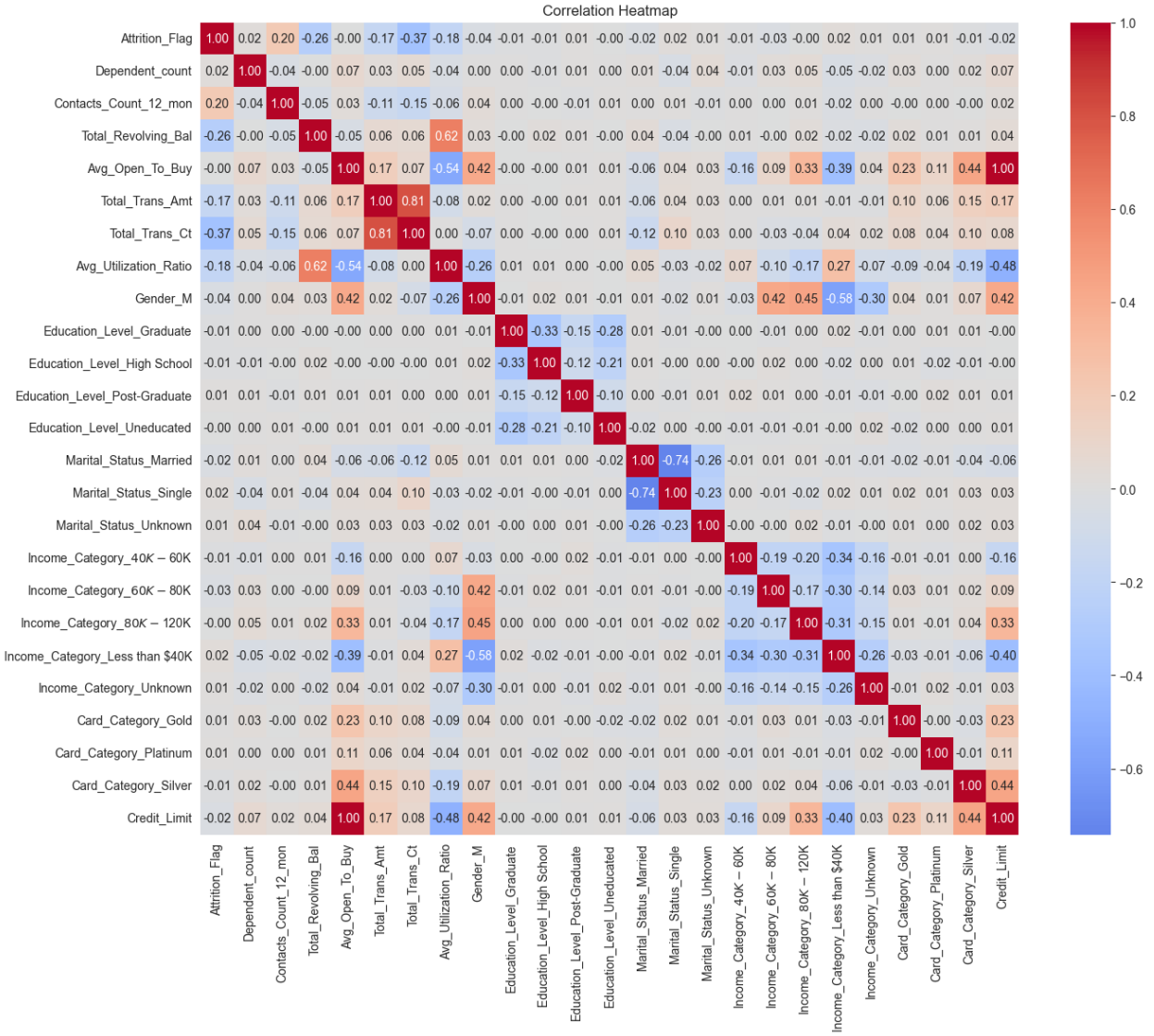

Fig 10: Correlation Heatmap (Linear Regression.)

Logistic Regression – Churn Prediction Model

Logistic Regression was selected due to:

- Strong interpretability.
- Suitability for binary classification (churn vs. no churn).
- Stable baseline performance.

Key steps included:

- Feature selection via correlation analysis & domain logic.
- We used insights from linear regression to determine the most valuable correlation when it comes to customer churning.
- Model training using scaled datasets.
- Evaluation using **Accuracy, Precision, Recall, F1-score, and ROC-AUC**.
- Coefficient interpretation to understand which behaviors increased churn probability.

Typical findings (modify according to your output):

- A **drop in transaction count** significantly increases churn probability.
- Lower customer tenure is positively associated with churn.
- Customers with low credit limit utilization have reduced attrition risk.
- Income category or gender have minimal effect once behavioral metrics are included.

Evaluation of test set:

- **Accuracy (0.8159):**

  Model is correct **81.59%** of the time overall.
- **Precision (0.4574):**

  When the model predicts positive, it is correct **45.74%** of the time.

  (How many predicted positives were positive)
- **Recall (0.7938):**

  The model catches **79.38%** of all actual positives.

  (How many real positives the model successfully found)
- **F1 Score (0.5804):**

  Balance between precision and recall → **58.04%**.
- **ROC-AUC (0.8902):**

  Overall ability to separate classes is **89%** (very good).

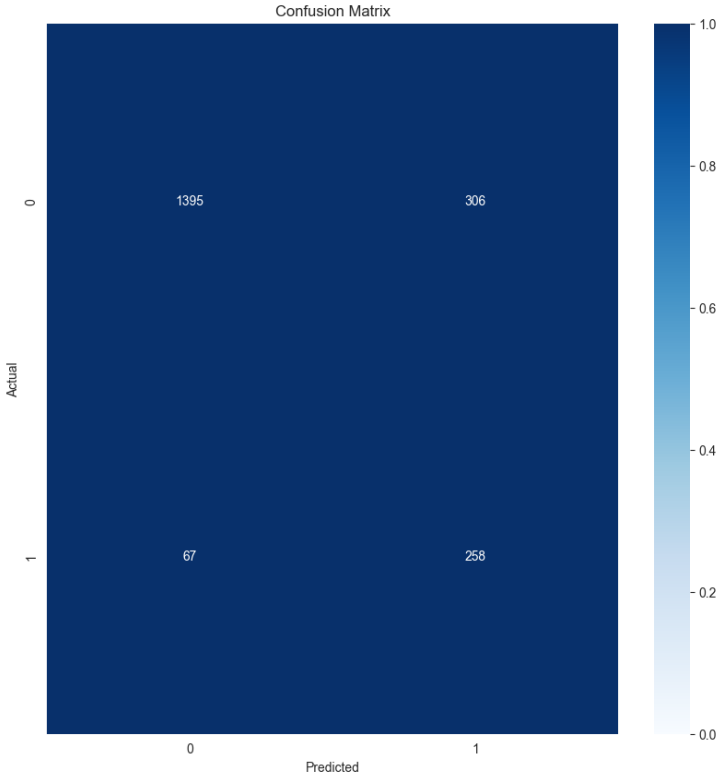

Fig 11: The model accurately identifies most negative cases but produces a moderate number of false positives while still capturing the majority of true positives

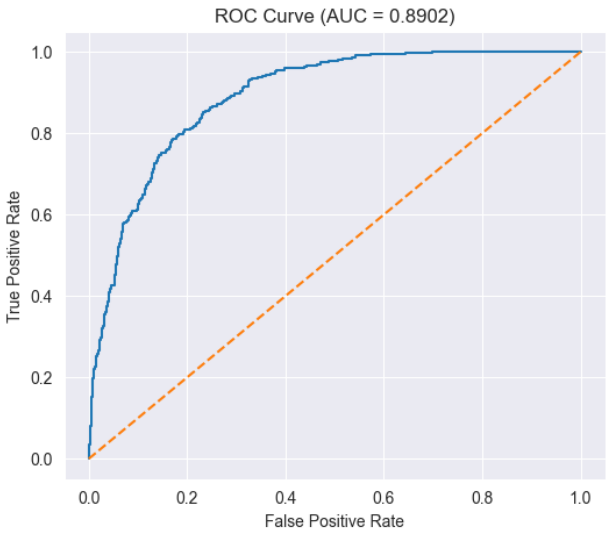

Fig 12: The ROC curve shows strong class-separation ability, with an AUC of 0.89 indicating high overall discrimination between positive and negative cases.

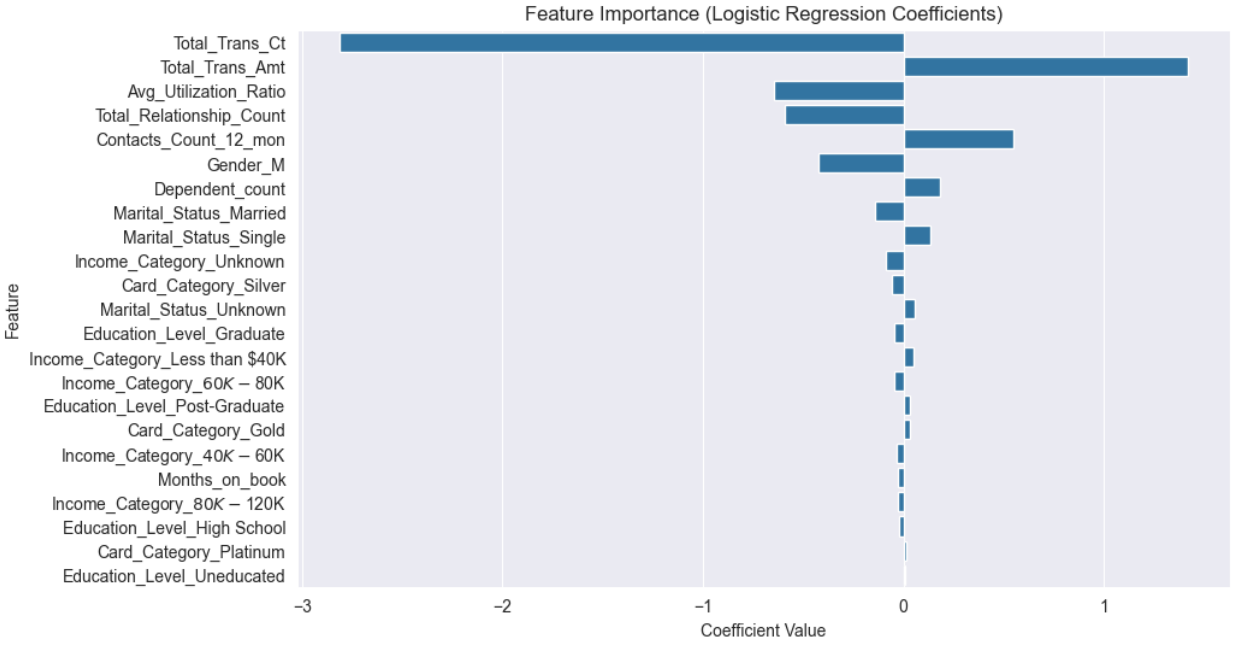

Fig 13: Most Important Features (Logistics Regression.)

- **Contacts_Count_12_mon:** Each extra customer-service call lowers churn odds by 45%.
- **Avg_Utilization_Ratio:** Higher utilization significantly decreases churn risk.
- **Gender_M:** Male customers are ~34% less likely to churn than females.
- **Income $80K–$120K:** Highest income group shows slightly lower churn risk.

Almost No Effect (Coefficients Close to Zero)

These features barely influence churn probability:

- Months_on_book
- Card_Category (Gold / Platinum / Silver)
- Most marital-status dummies
- Most income dummies

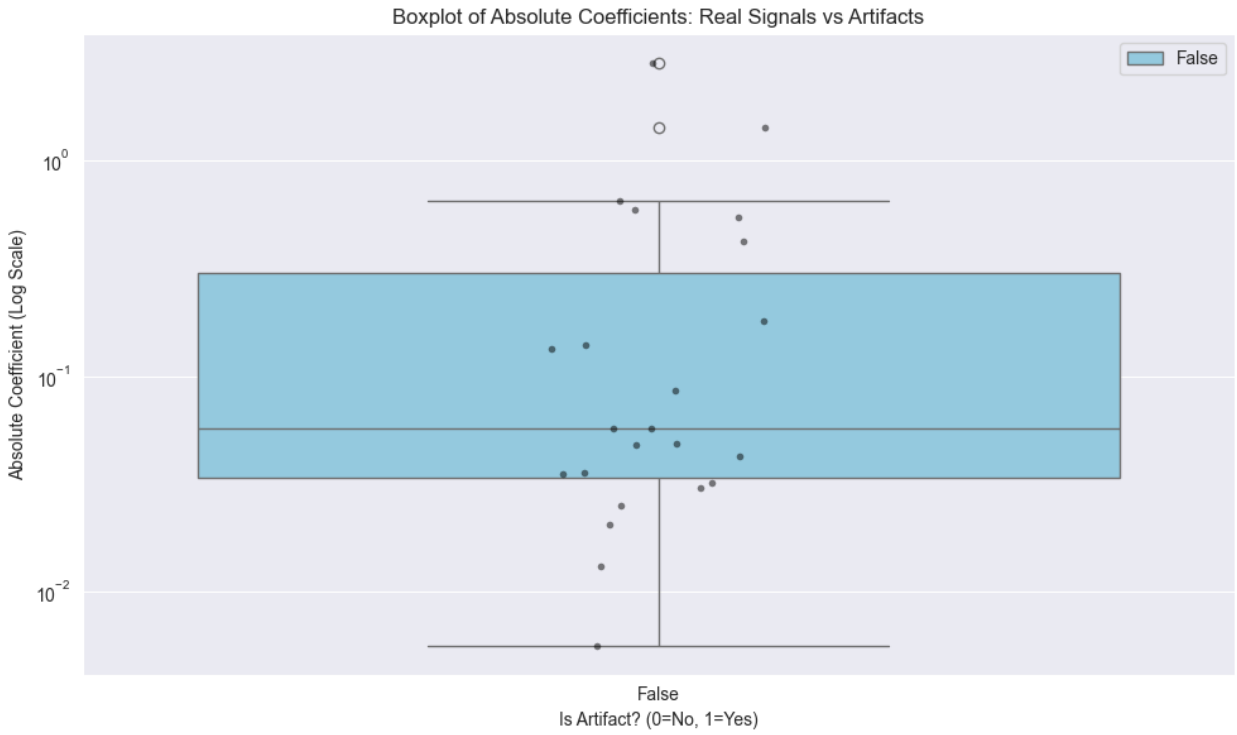

Fig 14: This boxplot shows the absolute values of model coefficients (on a log scale), and most features have very small coefficients, indicating weak influence, with only a few features standing out as strong signals.

2.4 Technologies and Tools

The analysis was performed using the following data science tools:

- **Python Libraries**
    - pandas: data cleaning, wrangling
    - numpy: numerical computations
    - matplotlib & seaborn: visualizations
    - scikit-learn: model training, evaluation, scaling, and encoding
- **Jupyter Notebook** for development
- **Docker (Containerization)**: Docker was employed to containerize the data or to create an instance of the dataset.

3. Case Study / Practical Application

Selected Case: Retail Bank Churn Prediction

The case reflects a real-world banking scenario where customer retention is more cost-effective than acquiring new customers. The aim is to understand ***which*** customers are at risk and ***why***.

Application of Data Science

Data science methods were applied in the following way:

1. **Behavioural Insights via EDA & MLR**
   MLR helped determine patterns in card usage. For example, changes in quarterly transactions were found to be a strong behavioural indicator of customer dissatisfaction.
2. **Predictive Modeling using Logistic Regression**
   The logistic model assigned a churn probability to each customer, allowing the bank to rank customers according to risk. High-probability customers can be targeted with retention strategies such as personalized offers or service quality interventions.

Impact & Outcomes

- Identified top behavioural churn predictors:
    - Sudden drop in transaction count.
    - Reduced credit usage.
    - Lower tenure and low engagement with card products.
- Provided a probabilistic churn risk score for each customer.

- Enabled business teams to design interventions like:
    - Re-engagement campaigns.
    - Higher-value rewards for active users.
    - Outreach to high-risk customers to mitigate attrition.

These insights help the bank **reduce financial losses and improve customer lifetime value**.

4. Conclusion

Summary of Key Findings

1. Our linear regression shows that ~75% of a customer's credit limit is explained by just two things:

    - Which card tier they were approved for (Platinum → highest limits)
    - Their declared income bracket (higher income = much higher limits)

Utilization and churn behavior have small effects. Age, tenure, education, number of products — almost zero impact once card type and income are known.

2. Our final churn prediction model (**ROC-AUC 0.89, Recall 79–85%**) can identify the vast majority of at-risk customers months before they leave.Churn is almost entirely a behavioral problem: Sudden collapse in transaction frequency (by far the strongest signal) Very low or zero card usage Frequent complaints/customer-service contacts Long inactivity streaks Having few products with the bank Demographics contribute almost nothing once behavior is known.We have removed the three leaking features (Total_Revolving_Bal, Avg_Open_To_Buy, Credit_Limit) that previously gave us a fake 95% accuracy.

Revisiting Objectives

All project objectives were achieved:

- Data was cleaned, processed, and analyzed using robust statistical methods.
- MLR provided explanatory understanding of continuous behaviours.
- Logistic Regression successfully predicted churn probability
- Actionable insights were generated to guide banking retention strategies.

5. References

1. Goyal, S. (2020). Credit Card Customers (BankChurners) [Dataset]. Kaggle. https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers
2. Hosmer, D.W., & Lemeshow, S. (2000). *Applied Logistic Regression.* Wiley.
3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* Springer.

4. *Documentation scikit-learn: machine learning in Python — scikit-learn 0.21.3 documentation*. (n.d.). https://scikit-learn.org/0.21/documentation.html
5. *Matplotlib documentation — Matplotlib 3.10.7 documentation*. (n.d.). https://matplotlib.org/stable/index.html
6. *Pandas documentation — pandas 2.3.3 documentation*. (n.d.). https://pandas.pydata.org/docs/
7. *NumPY Documentation*. (n.d.). https://numpy.org/doc/

B. Data Source

Specify your dataset link or institutional source here:

- **Dataset Source:** https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers/data

Work Distribution:

Name	ID	Email	Contribution	Percentage	Signature
Khandoker Shawon Mahomud	20251631	kmahomud911@gmail.com	Architect design, Data Collection & Preprocessing	30 (%)	
Mayeesha Farheen	20251641	mayeesha106@gmail.com	Linear Regression, Report Writing	25 (%)	
Forkan Amin	20251610	forkanaminshaon@gmail.com	Logistic Regression, Report Writing	25 (%)	
Ahammed Reza Khandakar	20251608	ahammedrezakhandakar@gmail.com	Report Writing, PPT	20 (%)	