DEPAUL UNIVERSITY

DEMOGRAPHIC IMPACT ON CREDIT CARD USAGE

JUNAITH SHAIK
SARAH HASHMI
PRADEEP SHANE
UDAY KIRAN LINGA
SPARSHIKA AJMAAN DINESH KUMAR

1. **Background:** In a fiercely competitive banking industry, where it often costs more to acquire a new customer than to retain an existing one, customer retention has gained importance as a key success factor. Credit card product retention is more important since frequent users of this product create revenues from fees, interest payments, and volumes of transactions. The proposed study will research customer retention in the credit card user base of a bank with an objective to discover real causes for loyalty of some of the customers and attrition of others through analyzing a range of customer information.

2. **Significance:** The bank should learn what factors affect customer retention in order to reduce attrition and maintain an exclusive customer base. Identification of concrete factors related to consumer demographics, credit usage patterns, and financial behavior corresponding to retention or attrition would empower the bank to tailor its services more closely to client needs, thus driving customer satisfaction and sustaining competitiveness in the market. This data will provide banks with a roadmap on how to proactively enhance retention efforts under increasing competition and changing customer expectations.

3. **Importance**: Analyzing the variables that have an effect on the amount of credit card transactions, using insights that will improve customer retention, targeting, and risk assessment is important. This project identifies the important predictors and interactions—like transaction frequency and credit limit—for data-driven techniques in personalized marketing and effective credit management. Advanced modeling techniques that can be applied in decision making enhance accuracy, hence better financial planning and resource allocation. This will, in turn, lead to sustainable development through responsible spending and long-term customer relationships, where financial institutions are at the forefront to harvest benefits from engagement, minimize risk, and improve service of customer needs in a competitive and data-driven market.

4. Introduction:

This is an investigation into customer retention in a bank's credit card user base, following an extensive number of factors that may impact whether the customers will stay or go. Analyzing thousands of customer profiles, this study tries to trace patterns regarding their demographic, credit-use, behavioral, and financial parameters. Information on age, gender, and income category helps understand which segments of customers demonstrate higher loyalty or are more prone to attrition. Credit usage patterns, such as transaction amount, frequency, and utilization rate, give a view into the level of customers' engagement with their cards, pointing to satisfaction or disengagement. The behavioral indicators in this model include inactivity in accounts or contact with customer support that may suggest dissatisfaction or low engagement— warning signals for attrition. Finally, financial drivers, including credit limits and open-to-buy balances, indicate how the bank addresses credit offerings with regard to the usage needs of the customer, which might have an effect on loyalty. Identification of these categories allows this study to provide actionable insights for the bank to improve their engagement strategies. These findings underpin the proactive customer retention efforts for high-value customers to reduce churn, enhance customer satisfaction, secure revenue streams, and strengthen the competitive edge of the bank in the evolving financial landscape.

4.1 About Dataset:

The dataset, titled “Bank Churners” was sourced from a bank’s internal database and contains anonymized data on over 10,000 customers. Each record represents a unique customer and includes information about their age, gender, marital status, education level, credit card category, transaction behavior, and account activity levels. This data was pre-processed to ensure anonymity, and categorical variables were encoded where necessary to facilitate analysis.

The primary challenges in this dataset include managing categorical variables with multiple levels, addressing class imbalances (as there are fewer attrited customers compared to retained ones), and handling "Unknown" entries or missing values in critical columns like Education_Level and Marital_Status. These challenges necessitate careful preprocessing to ensure accurate analysis and reliable insights into the factors influencing customer retention and attrition.

4.2 Null Values

A preliminary analysis of the dataset indicates the absence of explicit null values in the provided data frame. However, the dataset does contain instances of "Unknown" in categorical columns like Marital_Status and Education_Level, which may represent missing or unreported data. These "Unknown" values pose challenges, as they might bias the analysis if not addressed appropriately. Handling these instances may involve assigning them to a separate category or imputing values based on similar records, depending on the extent and impact of the missing data on model accuracy.

QUESTIONS TO ADDRESS:

- Determine the correlation
- Identify and remove the outliers
- Identify the best model

5.PRE-PROCESSING:

Outliers are data points that deviate significantly from the overall pattern of the dataset. Their presence can greatly impact data analysis, as they may skew statistical summaries and affect the performance of machine learning models. Identifying and addressing outliers is essential, as they may represent data errors or highlight unique patterns that require special attention.

In this dataset, key variables like Total_Revolving_Bal, Avg_Open_To_Buy, Total_Amt_Chng_Q4_Q1, Total_Trans_Amt, Total_Trans_Ct, Total_Ct_Chng_Q4_Q1, and Avg_Utilization_Ratio were analyzed for potential outliers. There were no NULL Values in the dataset.

Several outlier detection methods were employed to ensure comprehensive identification and handling:

IQR Method: The interquartile range method defines outliers as points that lie beyond a specified distance from the quartiles (e.g., 1.5 times the IQR). This approach is commonly used for skewed data as it is less sensitive to extreme values than the Z-score method.

5.1 BOXPLOT ANALYSIS:

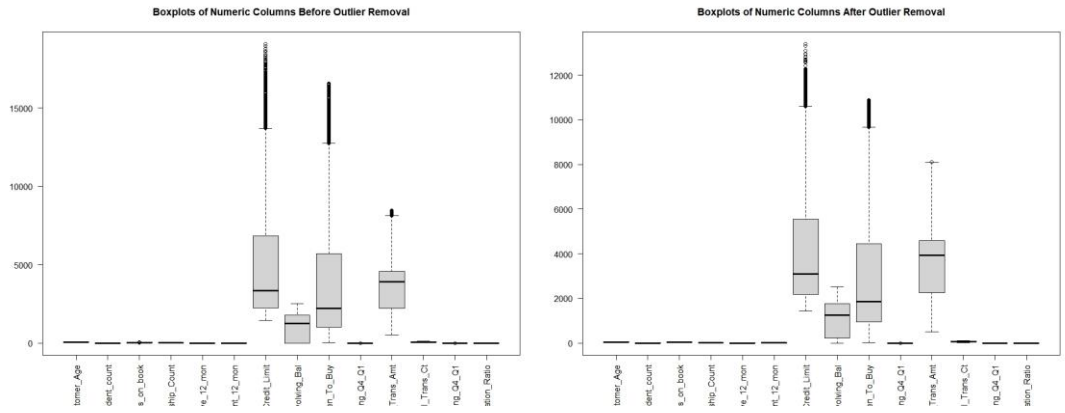
Boxplot analysis was implemented for the Bank Churners dataset for detection of the outliers.
A boxplot, also known as a box-and-whisker plot, is a graphical representation that summarizes the distribution of a dataset. In the boxplot visualization, certain variables such as Customer_Age, Months_on_book, Months_Inactive_12_mon, Contacts_Count_12_mon, Credit_Limit,

Avg_Open_To_Buy, Total_Amt_Chng_Q4_Q1, Total_Trans_Amt, Total_Trans_Ct, Total_Ct_Chng_Q4_Q1, Naive_Bayes_Classifier_Attrition_Flag, Card_Category, Contacts_Count_12_mon, Dependent_count, and Education_Level exhibit significant outliers. The presence of these outliers is crucial to identify, particularly in the context of regression analysis, as outliers can heavily influence the regression model. They can skew the estimated relationships between independent and dependent variables, distorting the coefficients, which may lead to biased or inaccurate predictions.

On the other hand, variables such as CLIENTNUM, Dependent_count, Total_Relationship_Count, Total_Revolving_Bal, and Avg_Utilization_Ratio show no significant outliers, suggesting that these variables follow a more consistent distribution. Their lack of extreme values can make them more reliable for modeling, as they are less likely to distort regression results.

5.2 Dealing with Outliers:

To find and eliminate the dataset's outliers, we applied the Interquartile Rule.



Outliers have been removed from the dataset

6. Correlation Check:

A correlation plot in R is a method of visualization of the correlation matrix used to summarize different relationships among several variables. The correlation matrix tabulates the correlation coefficients between pairs of variables. It depicts graphically the intensity and the direction of such correlations, allowing the identification of patterns and dependences among Variables For instance, using the corrplot package in R, one has the possibility of producing a correlation plot with color intensities and numerical annotations to emphasize the correlation coefficients of the matrix. This kind of plot provides a way of visualizing groups of variables that have similar patterns of correlations, creating knowledge around the relationships between variables.

In our dataset, the correlation plot shows all variables presenting a correlation coefficient close to 0, thus showing no significant dependencies between them. Nonetheless, some striking exceptions do exist.

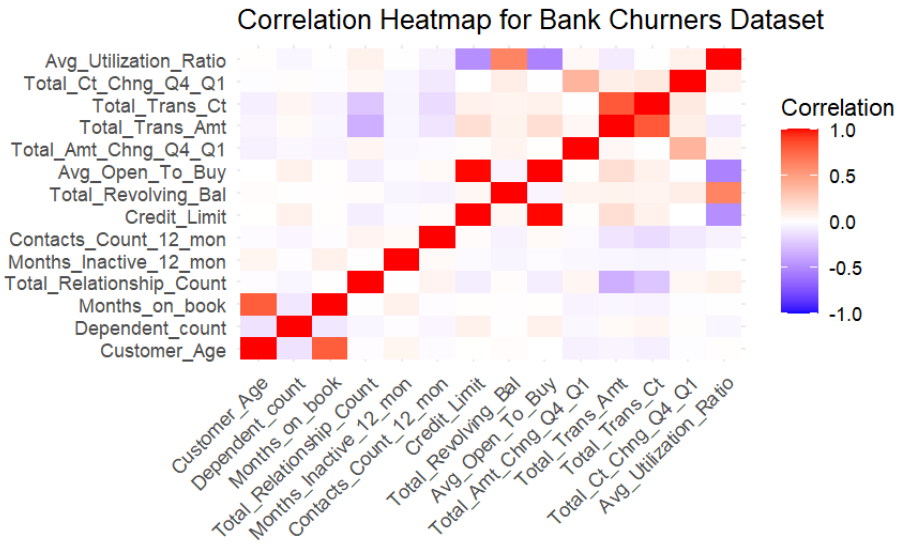


Figure 1. Correlation Plot

Understanding these correlations is crucial for regression analysis, helping identify potential multicollinearity issues and guiding the selection of influential variables. This preliminary analysis lays the foundation for a more effective and insightful regression model, ultimately enhancing the reliability and interpretability of the dataset.

7. Regression Models

Initially, we built a linear regression model using all of the variables to explain the *Total_Trans _Amt.* In t his model we included all of the predictor variables in the dataset. However, After reviewing the summary , we have observed that not all of the variables are statistically significant and for most of the variables th e p- values are higher than 0.05. The p-values above the conventional threshold of 0.05 indicates that som of the variables are not contributing effectively to explain the variance in the model. Having thse variable s in the model decreases the predictive power of the model.

Residual standard error: 835.1 on 5152 degrees of freedom
Multiple R-squared: 0.7194, Adjusted R-squared: 0.7177
F-statistic: 426.1 on 31 and 5152 DF, p-value: < 2.2e-16

Stepwise AIC

To address the redundant variable issue, we performed stepwise selection method based on the Akaike Information Criterion(AIC). This step ensures that we retain most of the significant variables and removing those variables that do not explain the variance in the data. We performed both forward AIC and backward AIC and got the same AIC value and same variable suggestion. The Stepwise AIC step iteratively evaluates the models by adding or by dropping variables and finds out the model that has least AIC score. In simple words we could say this step removed all of the unnecessary complexity in the data.
Call:
lm(formula = Total_Trans_Amt ~ Attrition_Flag + Customer_Age +

```
Dependent_count + Marital_Status + Card_Category + Total_Relationship_Count +
Contacts_Count_12_mon + Credit_Limit + Total_Revolving_Bal +
Total_Amt_Chng_Q4_Q1 + Total_Trans_Ct + Total_Ct_Chng_Q4_Q1,
data = train_data)
```

The revised model showed similar R-squared value but was more parsimonious and efficient. This helped us into getting more insights on the data and made the data more suited for further analysis. This model explains 71.88% of the variance in the data with an adjusted R-squared of 71.79% indicating a good fit to the data. The F-statistic (825.4, $p < 2.2e-16$) confirms that the model is statistically significant overall.

Analysis of various models.

We the tried to develop a comprehensive understanding of the relationships between predictors and *Total_Trans_Amt* and for this we analyzed various models. First we analysed a first order model and then we increased complexity and analysed interaction models, second order models and finally logarithmic models. The analysis of which and the summary is discussed below.

1. First-Order Model

   Residual standard error: 834.8 on 5167 degrees of freedom
   Multiple R-squared: 0.7188, Adjusted R-squared: 0.7179
   F-statistic: 825.4 on 16 and 5167 DF, p-value: < 2.2e-16

2. First-Order Model with Interactions

   - **Min Residual**: -3096.4, slightly increased in magnitude, indicating the largest underprediction in this model.
   - **1st Quartile (1Q)**: -472.3, suggesting a reduction in residual spread compared to the first-order model.
   - **Median**: -47.7, closer to zero, indicating a more balanced residual distribution.
   - **3rd Quartile (3Q)**: 399.1, similar to the first-order model.
   - **Max Residual**: 4440.9, lower than in the first-order model, showing reduced overprediction extremes.

3. Second-Order Model

   Residual standard error: 819 on 5158 degrees of freedom
   Multiple R-squared: 0.7298, Adjusted R-squared: 0.7285
   F-statistic: 557.3 on 25 and 5158 DF, p-value: < 2.2e-16

4. Complete Second-Order Model

   Residual standard error: 715.6 on 4904 degrees of freedom
   Multiple R-squared: 0.8039, Adjusted R-squared: 0.7927
   F-statistic: 72.06 on 279 and 4904 DF, p-value: < 2.2e-16

Best Model

We then performed logarithmic models for all of these models and we got best results for logarithmic first order interaction model

The residual summary indicates that the log transformation has improved the residual distribution:

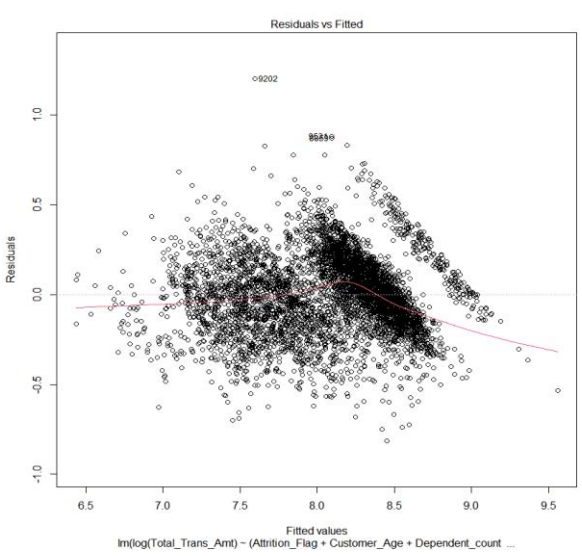
- **Min Residual**: -0.8401, representing the largest underprediction on the log scale.
- **1st Quartile (1Q)**: -0.1406, showing the lower quartile of residuals.
- **Median**: 0, indicating a more symmetric residual distribution around zero.
- **3rd Quartile (3Q)**: 0.1422, reflecting a similar spread to the 1st quartile.
- **Max Residual**: 1.1641, the largest overprediction, showing that residuals are more evenly distributed on the log scale.

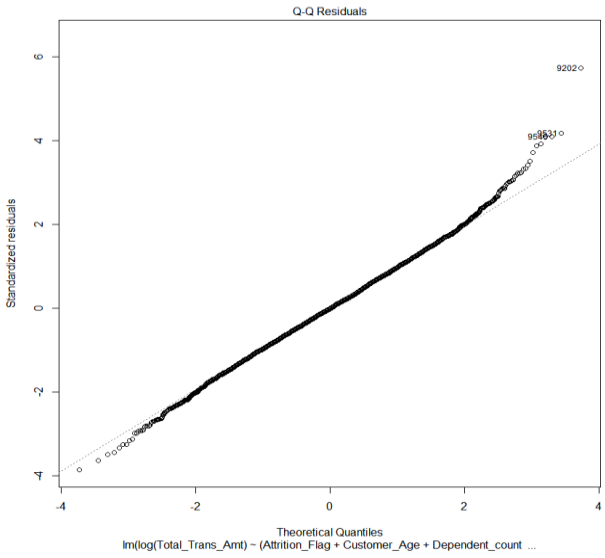
The fit statistics indicate strong explanatory power:

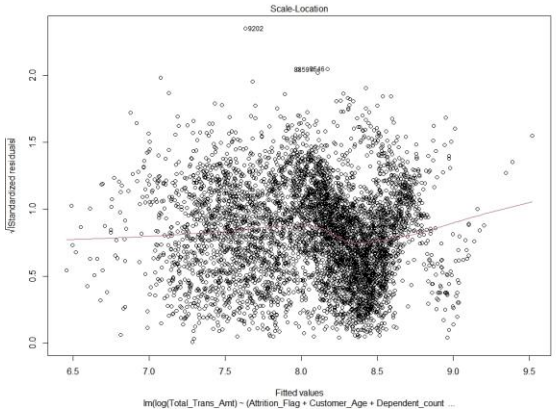
- **Residual Standard Error**: 0.2147, showing a low level of unexplained variability on the log-transformed scale.
- **Multiple R-squared**: 0.8189, indicating that 81.89% of the variability in log(Total_Trans_Amt) is explained by this model.
- **Adjusted R-squared**: 0.8147, slightly lower but still high, demonstrating that the model balances complexity with strong predictive power.

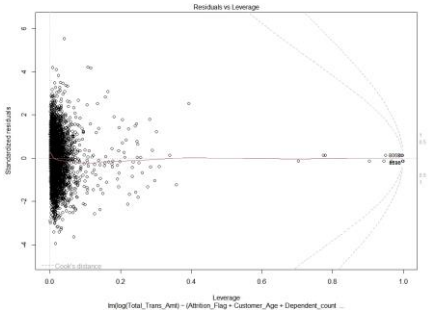
F-statistic: 197.5 on 116 and 5067 degrees of freedom, with a p-value < 2.2e-16, confirming the model's overall significance.

8. Residual Analysis





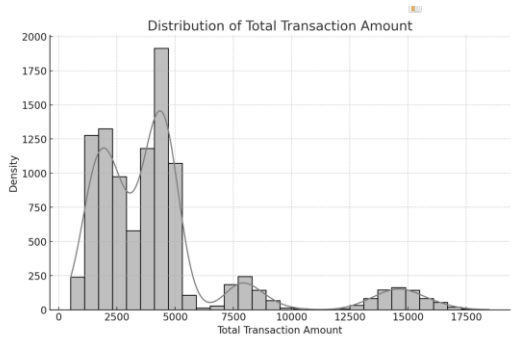




9.5 Conclusion:

The Q-Q plot indicates that while the assumption of normality is reasonable for the central portion of the residuals, the deviations in the tails warrant further investigation. To enhance the reliability of statistical inference, addressing the anomalies identified in the distribution tails is crucial. This may involve modifying the model, applying transformations, or using alternative regression approaches to ensure more accurate and robust predictions**.**

Exploratory Data Analysis (EDA) :



Exploratory Data Analysis (EDA):

The EDA for Total_Trans_Amt began with visualizing its distribution through histograms and density plots. The initial inspection showed a right-skewed distribution, indicating that most customers have lower transaction amounts, with a few having significantly higher amounts. This right skew was confirmed by a skewness value of 2.04.

To address this skewness and approximate normality, a logarithmic transformation was applied to the Total_Trans_Amt variable. This transformation compressed the range of higher values, resulting in a

more symmetric distribution, with a noticeable reduction in skewness, thereby facilitating improved data interpretability and preparation for model training.

Conclusion:

- In conclusion, after exploring various regression models, we found that the Logarithmic-First Order model with interaction effectively captures the demographic impact on credit card usage. The model's Residual Standard Error of 0.2134 indicates a low average prediction error.

- Additionally, a Multiple R-squared of 0.8188 and an Adjusted R-squared of 0.8146 suggest that the model explains approximately 81% of the variance in credit card usage, even when adjusted for the number of predictors.

- The F-statistic of 195.7, along with a highly significant p-value (less than 2.2e-16), further supports the model's overall reliability and relevance. These metrics indicate that this model accurately highlights how demographic factors influence credit card usage, making it a valuable tool for understanding these dynamics.