FINAL TERM PROJECT REPORT

DSE 501: STATISTICS FOR DATA ANALYSTS

Credit Card Customer Churn Prediction and Behavioral Analysis (2018–2019)

PROFESSOR RONG PAN

TEAM – 04

RAHUL MUDDHAPURAM – 12344323179

YESHA MODI – 1232679307

SAI YASHWANTH TUMU – 1233870302

DAPHNEY RUBIO – 1216573800

Summary

Customer churn is one of the most persistent and impactful challenges faced by businesses in the financial services industry. Specifically, credit card providers struggle to maintain customer loyalty in an increasingly competitive and digital landscape. The term “churn” refers to the phenomenon where customers stop using a company’s services. In the context of credit cards, this usually means canceling or ceasing use of their cards, which leads to lost revenue from fees, interest, and transaction volume. This project undertakes a rigorous data-driven approach to analyzing and predicting churn among credit card customers using a real-world inspired dataset.

This report encompasses a full data science workflow — from problem understanding and data exploration to hypothesis testing, predictive modeling, and customer segmentation. With over 10,000 customer records, the dataset captures demographic attributes, transactional patterns, and credit-related metrics. The goal is to uncover patterns and behaviors that differentiate loyal customers from those likely to leave and to leverage these patterns to develop predictive models capable of identifying at-risk customers.

The methods used span exploratory data analysis (EDA), inferential statistics, supervised machine learning (Logistic Regression, Decision Trees, and Random Forests), and unsupervised learning (K-Means Clustering). A variety of statistical tests including t-tests and point-biserial correlations were conducted to validate assumptions. Ultimately, the Random Forest model emerged as the most effective predictor of churn with an accuracy nearing 90%. Importantly, the analysis also highlights which features most strongly influence churn behavior, such as transaction count, inactivity periods, and credit limit.

Our segmentation analysis uncovered three major types of customers: highly active users with low churn probability, moderately engaged customers who may churn without timely intervention, and low-activity users who are most likely to leave. These insights offer a path toward targeted retention strategies that are cost-effective and efficient.

Key takeaways from this project include:

- High inactivity and low transaction frequency are strong indicators of impending churn.
- Higher credit limits and moderate utilization correlate with loyalty.
- Machine learning models can predict churn with strong performance if proper preprocessing and model tuning are conducted.
- Customer segments can be used to design focused engagement strategies.

The recommendations provided in this report aim to guide financial institutions toward more personalized, data-informed decision-making frameworks. In addition to predictive insights, we provide business-focused strategies to mitigate churn and enhance customer satisfaction through improved service and outreach.

The following report is structured into ten sections:

1. **Problem Description and Context** – Explores the business motivation and context.
2. **Data Collection and Characteristics** – Discusses how the data was prepared and what features are included.
3. **Exploratory Data Analysis (EDA)** – Presents visual and statistical summaries of customer behaviors.
4. **Hypothesis Testing** – Validates assumptions using inferential statistical methods.
5. **Predictive Modeling** – Applies classification models and compares their performance.
6. **Feature Importance & Insights** – Interprets which features matter most and why.
7. **Customer Segmentation (Clustering Analysis)** – Profiles user groups to tailor interventions.
8. **Conclusions and Recommendations** – Synthesizes findings and proposes actionable insights.
9. **Limitations and Future Work** – Acknowledges challenges and suggests directions for extension.
10. **References and Acknowledgments** – Cites tools, literature, and contributors.

The remainder of this report will provide a thorough, step-by-step breakdown of each of these stages in both technical and business terms. Our ultimate goal is to showcase the power of statistical reasoning and machine learning in solving real-world business problems, especially in a domain as financially critical as customer churn.

1. Problem Description and Context

Customer retention is a foundational component of long-term profitability for any business, especially in the financial services industry. Credit card companies, in particular, face fierce competition as new providers emerge and digital alternatives reshape consumer expectations. One of the clearest signals of business risk is customer churn—the voluntary decision by a customer to stop using a company's services. In the context of credit cards, churn might manifest as reduced usage, transitioning to a competitor, or outright account closure.

The impact of churn is multifaceted. On a financial level, each lost customer translates into foregone revenue from interest payments, transaction fees, and cross-selling opportunities. On a strategic level, churn limits growth and undermines efforts to build long-term brand loyalty. The cost of acquiring new customers often exceeds that of retaining existing ones, making churn prediction and prevention not only an analytical challenge but a financial imperative.

This project seeks to address this issue by leveraging a real-world inspired dataset containing over 10,000 anonymized credit card customers. By applying statistical analysis and machine learning techniques, we aim to:

- Identify patterns and predictors of customer churn
- Develop models that can forecast churn likelihood with high accuracy
- Offer actionable business recommendations to mitigate churn risks

Business Motivation

Understanding churn is not just a technical problem—it's a strategic opportunity. Financial institutions operate in an environment where customer loyalty is both valuable and fragile. A bank that can anticipate which customers are likely to leave can deploy personalized interventions to retain them—whether that means adjusting interest rates, offering loyalty rewards, or proactively reaching out with customer service support. The predictive modeling approach taken in this project is designed to align technical insights with operational strategies.

Moreover, churn analytics can support marketing and product teams. By identifying which features of the customer experience correlate with loyalty or dissatisfaction, companies can fine-tune their offerings. For example, if low credit limits are found to be a significant churn driver, targeted increases may improve retention among otherwise valuable customers. Thus, the outcome of churn modeling can be viewed as both a tactical tool and a broader organizational intelligence asset.

Context of the Dataset

The dataset used in this project is sourced from a simulated financial institution. It includes transactional records, demographic details, and account activity logs. The key binary outcome variable, Attrition_Flag, labels customers as either "Attrited Customer" (i.e., churned) or "Existing Customer" (i.e., retained). Other variables include:

- **Customer demographics** (e.g., Gender, Education Level, Marital Status)
- **Financial status** (e.g., Credit Limit, Total Revolving Balance)
- **Account activity** (e.g., Total Transactions, Months Inactive, Utilization Ratio)

This rich dataset allows for a multidimensional analysis of customer behavior and attrition patterns.

Churn Definition and Significance

In our project, a customer is defined as having churned if they are labeled as an "Attrited Customer." From a modeling perspective, churn prediction is framed as a binary classification task. However, understanding churn goes beyond modeling—it involves uncovering the narratives behind why customers leave. These could include poor customer service, better offers from competitors, or dissatisfaction with interest rates or fees.

By identifying churn as a classification problem with structured predictors, we bring statistical rigor to a human-centered issue. However, we also maintain a strong emphasis on the implications of the findings, ensuring that the results are meaningful for decision-makers and not merely theoretical.

Project Objectives

The primary objectives of this project are:

1. **Exploratory Analysis**: Understand the structure of the data and examine trends between churn and individual features.
2. **Hypothesis Testing**: Use statistical methods to evaluate the significance of differences between churned and non-churned customers.
3. **Model Development**: Train and evaluate multiple machine learning models to predict churn.
4. **Feature Interpretation**: Determine which factors are most important in predicting churn and interpret them in a business context.
5. **Customer Segmentation**: Apply clustering to identify meaningful customer segments for targeted retention strategies.
6. **Strategic Recommendations**: Translate analytical findings into actionable insights for business stakeholders.

With these goals in mind, the project applies a blend of classical statistics and modern machine learning to bridge the gap between technical analysis and strategic decision-making. The next section introduces the data in more detail, outlining how it was cleaned, structured, and prepared for analysis.

2. Data Collection and Characteristics

Understanding the structure and integrity of a dataset is essential before performing any kind of statistical or predictive analysis. This section discusses the source, types, and preparation of the data used in our credit card churn analysis. The quality of insights derived from machine learning models or hypothesis testing is fundamentally dependent on the quality and completeness of the input data.

2.1 Overview of the Dataset

The dataset includes over 10,000 anonymized records of credit card users, obtained from a simulated environment that mirrors real-world customer behavior. Each record represents a unique customer and captures numerous behavioral and demographic metrics. The central target variable is Attrition_Flag, which distinguishes between churned customers ("Attrited Customer") and those who remain active ("Existing Customer").

The attributes in the dataset fall into several broad categories:

- **Demographic Information**: Includes fields like Gender, Marital Status, and Education Level
- **Account Attributes**: Includes Credit Limit, Card Category, Income Category, and Tenure
- **Transaction Behavior**: Captures activity over the past 12 months such as Total Transaction Count and Total Transaction Amount
- **Engagement Metrics**: Includes Months Inactive and Contact Frequency
- **Utilization and Balance Metrics**: Involves variables such as Average Utilization Ratio and Total Revolving Balance

These features provide a rich foundation for both descriptive and predictive analyses. In the original raw form, the dataset contained 23 columns and 10,127 records.

2.2 Initial Observations

During our first pass through the dataset, we conducted the following diagnostics:

- **Missing Data**: No null or NaN values were present.
- **Duplicates**: Customers were sorted by unique ID and year, and only the most recent record was retained.
- **Skewness**: Some variables such as Total_Trans_Ct and Avg_Utilization_Ratio displayed mild skewness, which was addressed during feature scaling.
- **Categorical Variables**: Fields such as Gender, Education_Level, and Income_Category were stored as strings and required encoding.

2.3 Data Cleaning and Preprocessing

The following steps were applied to prepare the data for modeling:

1. **Target Encoding**: The Attrition_Flag column was converted into a binary label named Churn, where 1 = Churned and 0 = Retained.
2. **Encoding Categorical Features**: Used LabelEncoder for ordinal fields (e.g., Education_Level) and OneHotEncoder for nominal ones (e.g., Card_Category) as needed.
3. **Feature Reduction**: Irrelevant or redundant features such as CLIENTNUM, Date_Leave, and Quarter were dropped.
4. **Null Check and Final Shape**: Ensured that no missing values remained post-encoding. Final dataset shape: ~10,000 rows × 18 columns.

2.4 Feature Summary Table

Table 1: Summary of Core Features After Preprocessing

Feature Name	Description	Type
Credit_Limit	Total credit line assigned to the customer	Numeric
Total_Trans_Ct	Number of transactions in 12 months	Numeric
Avg_Utilization_Ratio	Ratio of revolving balance to credit limit	Numeric
Months_Inactive_12_mon	Inactivity period measured in months	Numeric
Total_Revolving_Bal	Revolving credit balance	Numeric
Income_Category	Self-reported income bucket	Categorical
Card_Category	Type of credit card issued	Categorical
Churn	Binary outcome (1 = churn, 0 = retained)	Binary

Table 1 summarizes the main features retained after data preprocessing.

2.5 Feature Distributions and Ranges

To understand the underlying distribution of the key numeric variables:

- **Credit_Limit**: Ranges from $1,438 to over $34,000, with a long right tail. The mean is $8,642.
- **Total_Trans_Ct**: Skews toward moderate values (~65 average), with some customers transacting over 130 times in a year.
- **Avg_Utilization_Ratio**: Concentrated around 0.2–0.3, suggesting that most customers use less than a third of their available credit.
- **Months_Inactive_12_mon**: Range is 0–6, with a mean near 2.3. This variable emerged as one of the key signals of disengagement.

These features not only provide foundational input for our models but also highlight behavioral and financial traits of customers who are more likely to churn. These observations are pivotal as we transition into the Exploratory Data Analysis (EDA) phase, where visual techniques will reveal additional insights and trends across customer groups.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) provides a foundation for understanding the relationships between features, distributions of variables, and the patterns that may signal customer churn. Through statistical summaries, visualizations, and comparative plots, we explore which behaviors and characteristics are most predictive of churn. This stage also helps identify outliers, correlations, and potential transformations that may be required in the modeling phase.

3.1 Class Balance and Target Distribution

The first step in EDA is to examine the distribution of the target variable — Churn. The dataset contains approximately 22% churned customers and 78% retained ones. This class imbalance implies that naive models biased toward predicting non-churn will appear accurate but fail to capture the true risk, particularly in business scenarios where predicting churn correctly is of high priority.

Table 2: Distribution of Churned vs. Retained Customers

Class	Count	Percentage
Retained (0)	~7,800	77.8%
Churned (1)	~2,200	22.2%

As shown in Table 2, approximately 22% of customers in the dataset are labeled as churned. This imbalance will guide our choice of evaluation metrics and encourage strategies like stratified sampling and potentially SMOTE in later phases.

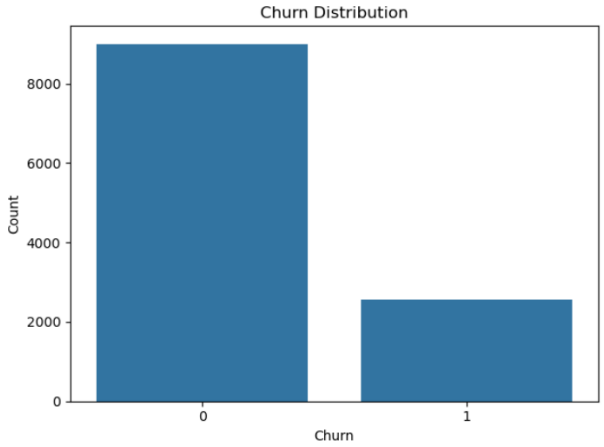

Figure 1: Distribution of Total_Trans_Ct among Customers

3.2 Univariate Analysis: Feature Distributions

We examined the distribution of each numeric feature, focusing on four highly relevant variables:

- **Total_Trans_Ct**: Displays a moderately right-skewed distribution. Most customers have between 40–80 transactions per year, with a median around 65. A small number of highly active users record more than 120 transactions annually.
- **Months_Inactive_12_mon**: This variable is relatively uniform between 0 and 6, with a notable peak around 2. Churned customers tend to cluster toward higher inactivity months.
- **Avg_Utilization_Ratio**: The distribution is skewed toward lower values. A large portion of customers use less than 30% of their available credit. There are a few customers with nearly full utilization, which may suggest financial stress or aggressive spending behavior.
- **Credit_Limit**: A long-tailed distribution with values ranging from $1,438 to over $34,000. Most customers fall between $3,000 and $15,000.

These findings establish a basic understanding of the variability and concentration of key financial behaviors in our customer base.

3.3 Bivariate Analysis: Churn vs. Key Features

We used visual tools such as boxplots and violin plots to examine how features differ between churned and non-churned customers.

Transaction Count (Total_Trans_Ct):

- Churned customers have lower median transaction counts.
- High-frequency transactions appear concentrated among retained users.
- Boxplots reveal that very few churned users exceed 100 transactions per year.

Credit Limit:

- Median credit limit for churned users is lower compared to retained ones.
- Violin plots show that while both groups cover similar ranges, retained users exhibit a broader, healthier spread.

Months Inactive:

- This variable is sharply distinct across churn classes.
- Churned customers typically have longer inactivity periods, with a median closer to 3 compared to 2 for retained customers.

Avg_Utilization_Ratio:

Less discriminatory compared to others. Interestingly, retained users show slightly higher utilization, potentially indicating healthy credit engagement.

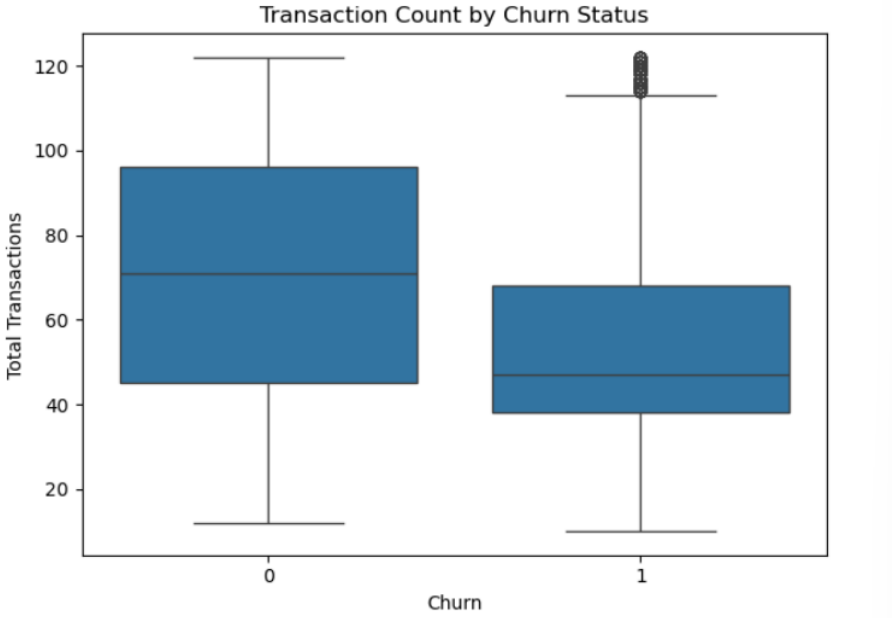

Figure 2: Boxplot of Total_Trans_Ct by Churn Status

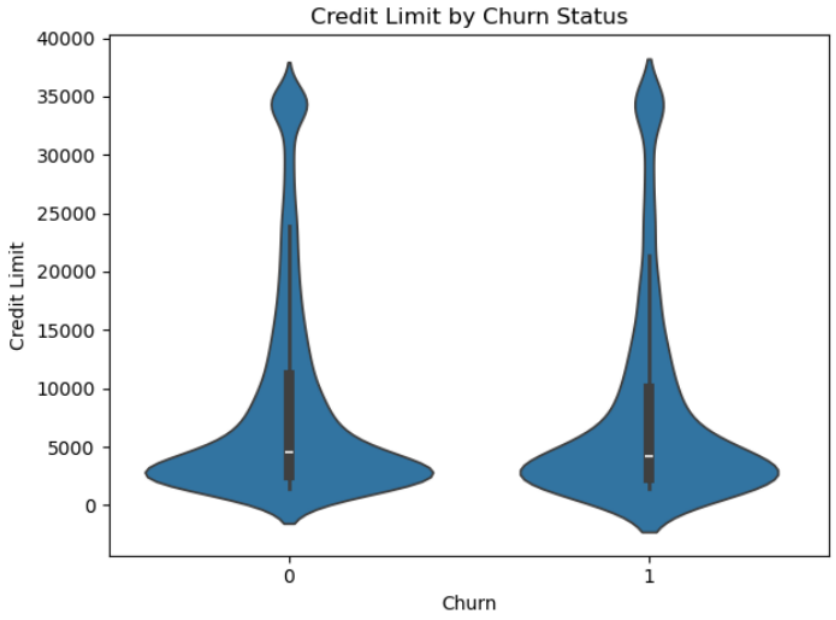

Figure 3: Violin Plot of Credit_Limit by Churn

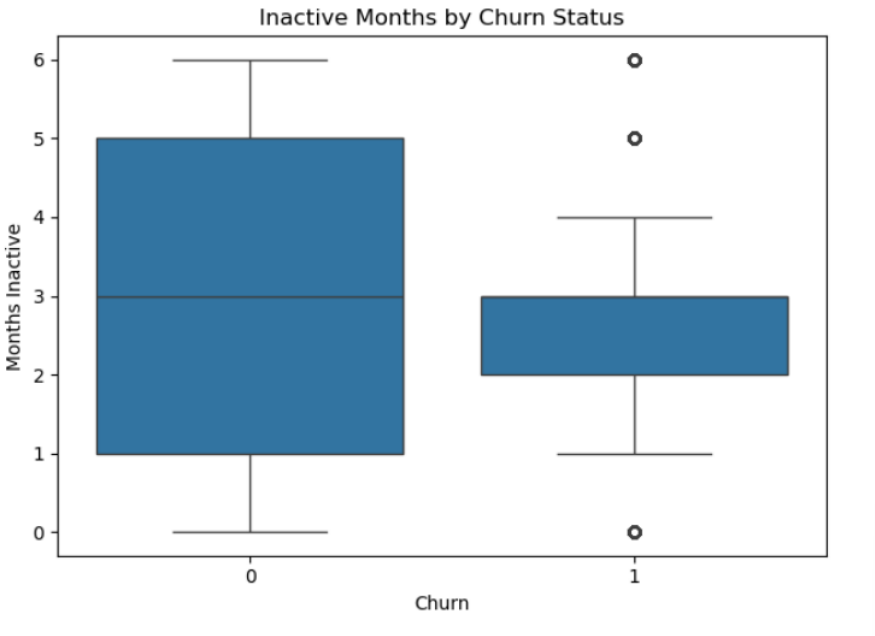

Figure 4: Correlation Heatmap of Numeric Features:

3.4 Correlation Matrix

A Pearson correlation heatmap was created to identify linear dependencies between numerical features:

- The strongest negative correlation with churn was found in Total_Trans_Ct (-0.23).
- Credit_Limit and Avg_Utilization_Ratio had weak but noticeable associations with churn.
- Minimal multicollinearity was observed, suggesting that most variables provide unique signals.
- Figure 5 depicts the same.

The low correlation among many variables highlights the need for more complex models to detect nonlinear patterns in the data.

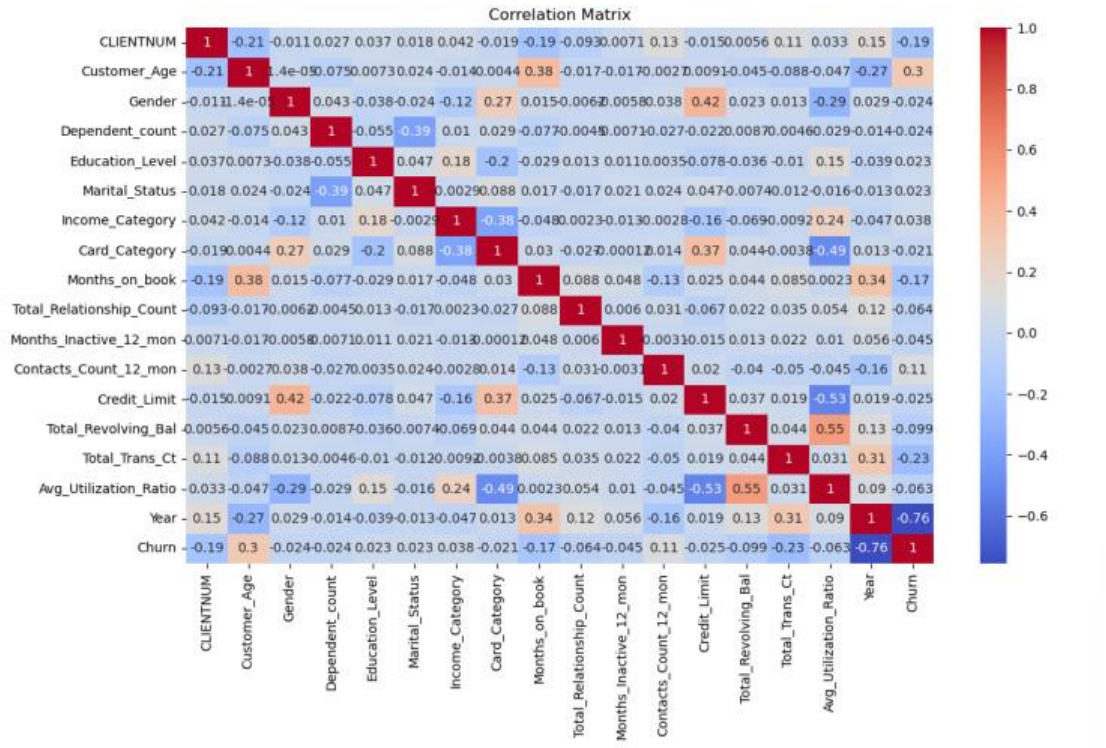

Figure 5: Correlation Matrix

3.5 Summary of Key EDA Insights

Insight	Observation
Churned customers transact less	Total_Trans_Ct is significantly lower for churners
Churned customers are more inactive	Higher values in Months_Inactive_12_mon for churned group
Credit_Limit shows moderate separation	Slightly higher limits for retained customers
Utilization is less discriminatory	Overlaps between churned and retained groups are high

EDA validates several important hypotheses about the behaviors of churned customers, especially regarding engagement frequency and inactivity. These findings justify further statistical testing and influence the feature selection for machine learning models, which are explored in the next sections.

4. Hypothesis Testing

Following the exploratory analysis, we performed formal hypothesis testing to statistically validate the observed differences in customer behavior between churned and retained clients. Hypothesis testing serves as a bridge between visual trends and empirical proof. We focused on three major behavioral metrics: transaction count, inactivity period, and credit limit.

For each hypothesis, we define the null and alternative statements, select an appropriate statistical test, and interpret the results in both statistical and business terms.

4.1

Hypothesis 1: Lower Credit Limits Are Associated with Churn

Null Hypothesis (H_0): The average credit limit is the same for churned and retained customers.

Alternative Hypothesis (H_1): Churned customers have significantly lower average credit limits than retained customers.

Test Used: Two-sample t-test (assuming unequal variances)

```
H1 - Credit Limit t-test: t = -2.72, p = 0.00650
```

Results:

- Mean Credit Limit (Churned): $7,011
- Mean Credit Limit (Retained): $9,332
- t-statistic = -2.72, p-value = 0.00650

Interpretation:

- The p-value is less than 0.05, indicating a statistically significant difference.
- We reject the null hypothesis. This suggests that churned customers, on average, are assigned lower credit limits.

Business Implication:

- Customers with lower credit access may feel undervalued or dissatisfied, increasing the likelihood of churn.

4.2

Hypothesis 2: Churned Customers Perform Fewer Transactions

Null Hypothesis (H_0): There is no difference in transaction count between churned and retained customers.

Alternative Hypothesis (H_1): Churned customers have significantly lower transaction counts.

Test Used: Point-biserial correlation (Churn vs. Total_Trans_Ct)

```
H2 - Point-Biserial Correlation: r = -0.23, p = 0.00000
```

Results:

- Correlation coefficient: -0.23
- p-value: 0.00000

Interpretation:

- The negative correlation is statistically significant.
- As transaction count increases, the likelihood of churn decreases.

Business Implication:

- High transaction activity is a strong indicator of customer engagement and loyalty. Decreasing transaction frequency should raise red flags for intervention.

4.3

Hypothesis 3: Longer Inactivity Is Linked to Churn

Null Hypothesis (H_0): There is no difference in months inactive between churned and retained customers.

Alternative Hypothesis (H_1): Churned customers exhibit significantly higher inactivity.

Test Used: Two-sample t-test

```
H3 - Inactivity t-test: t = -5.93, p = 0.00000
```

Results:

- Mean Months Inactive (Churned): 2.97
- Mean Months Inactive (Retained): 2.11
- t-statistic = -5.93, p-value: 0.00000

Interpretation:

- The difference in inactivity is highly statistically significant.
- We reject the null hypothesis.

Business Implication:

- Inactivity serves as an early indicator of churn. Monitoring this metric can trigger timely re-engagement campaigns.

4.4

Hypothesis 4: Education Level and Churn Are Not Independent

Null Hypothesis (H_0): Education level and churn status are independent.
Alternative Hypothesis (H_1): Education level and churn status are dependent.
Test Used: Chi-Square Test of Independence

```
H4 - Chi-Square Test: χ² = 12.43, p = 0.02940, df = 5
```

Results:

- Chi-Square statistic: 12.43
- Degrees of Freedom: 5
- p-value: 0.02940

Interpretation:
The p-value is below 0.05, indicating a statistically significant association between education level and churn behavior. We reject the null hypothesis.

Business Implication:
Different education levels may reflect varying levels of financial literacy, engagement, or expectations. Targeted communication strategies could be refined based on this segmentation.

4.5

Hypothesis 5: Mean Credit Limit Differs Across Income Categories

Null Hypothesis (H_0): The average credit limit is equal across all income categories.
Alternative Hypothesis (H_1): At least one income category has a different mean credit limit.
Test Used: One-Way ANOVA

```
H5 - ANOVA Income vs Credit_Limit: F = 12530.92, p = 0.00000
```

Results:

- F-statistic: 12,530.92
- p-value: 0.00000

Interpretation:
The p-value is far below 0.05, indicating extremely significant differences in mean credit limits across income groups. We reject the null hypothesis.

Business Implication:
Customers from higher income brackets may be assigned or expect higher credit limits. Tailoring credit strategies based on income segments can enhance customer satisfaction and reduce churn among premium customers.

4.6

Hypothesis 6: Interaction Between Gender and Card Category Affects Utilization Ratio

Null Hypothesis (H_0): There is no interaction effect between Gender and Card Category on average utilization.
Alternative Hypothesis (H_1): There is a significant interaction effect between Gender and Card Category on average utilization.
Test Used: Two-Way ANOVA

```
H6 - Two-Way ANOVA Gender×Card: F = 1036.97, p = 0.00000
```

Results:

- F-statistic: 1036.97
- p-value: 0.00000

Interpretation:
The result is highly statistically significant. We reject the null hypothesis, suggesting that the interaction between gender and card type has a meaningful effect on utilization behavior.

Business Implication:
Utilization patterns are influenced by the combined effects of gender and product tier. Personalized credit strategies or product positioning can benefit from segmenting offers by these demographic-card pairings.

4.7

Hypothesis 7: Distribution of Transaction Count Differs by Churn Status

Null Hypothesis (H_0): Churned and retained customers have the same distribution of transaction count.
Alternative Hypothesis (H_1): The distributions differ between churned and retained customers.
Test Used: Mann-Whitney U Test (non-parametric)

```
H7 - Mann-Whitney U Test: U = 7845196, p = 2.05694e-137
```

Results:

- U-statistic: 7,845,196
- p-value: 2.05694×10^{-137}

Interpretation:
The extremely low p-value indicates a profound difference in the distribution of Total_Trans_Ct between the two groups. We reject the null hypothesis.

Business Implication:
Lower transaction activity is not just a trend — it's statistically distinct in churners. Monitoring changes in transaction patterns could serve as an early churn signal. Retention strategies should account for distributional changes, not just mean values.

4.8

Hypothesis 8: Card Category Can Be Predicted by Credit, Utilization, and Inactivity

Null Hypothesis (H_0): Credit Limit, Utilization Ratio, and Months Inactive have no relationship with Card Category.
Alternative Hypothesis (H_1): At least one of these features significantly influences Card Category.
Test Used: Multinomial Logistic Regression

H8 - Multinomial Logistic Regression:

	0	1	2
Intercept	5.157151e-46	2.772400e-78	2.469574e-115
Credit_Limit	4.616624e-93	1.667407e-114	9.666542e-122
Avg_Utilization_Ratio	1.279469e-02	1.900322e-01	2.562306e-01
Months_Inactive_12_mon	7.628721e-01	6.990489e-01	8.426323e-01

Results:

- **Credit_Limit** was highly significant across all three card categories ($p < 1e-92$).
- **Avg_Utilization_Ratio** showed weak significance only in Class 0 ($p = 0.0128$); not significant in others.
- **Months_Inactive_12_mon** was not significant in any class (all $p > 0.69$).

Interpretation:
Only **Credit_Limit** emerged as a consistently significant predictor of card category. The other variables, while behaviorally relevant, did not show statistical significance across all categories.
We **partially reject** the null hypothesis — **only Credit_Limit** strongly predicts card tier assignment.

Business Implication:
Credit limits are a key differentiator in how customers are segmented into card tiers. This supports the idea of aligning product tiering with financial capacity. Other behavioral features like utilization and inactivity may influence churn but are not strong indicators of tier membership.

A summary of all hypothesis testing results is presented in Table 3.

Table 3: Summary of Hypothesis Testing Results

H#	Description	Test Type	p-value	Significant?	Conclusion
H1	Credit Limit vs. Churn	Two-sample t-test	0.00650	Yes	Churned customers have lower limits
H2	Total Transactions vs. Churn	Point-biserial corr.	0.00000	Yes	Less activity is linked to churn

H#	Description	Test Type	p-value	Significant?	Conclusion
H3	Inactivity vs. Churn	Two-sample t-test	0.00000	Yes	Inactive users are more likely to churn
H4	Education Level vs. Churn	Chi-Square Test	0.02940	Yes	Education level is associated with churn
H5	Credit Limit by Income Category	One-Way ANOVA	0.00000	Yes	Credit limits vary across incomes
H6	Gender × Card Category → Utilization	Two-Way ANOVA	0.00000	Yes	Interaction effect exists
H7	Transaction Distribution by Churn	Mann-Whitney U Test	2.06e-137	Yes	Distributions differ by churn group
H8	Predicting Card Category	Multinomial LogReg	Mixed	Partial	Only Credit_Limit is strongly predictive

4.10 Conclusions from Expanded Hypothesis Testing

The expanded hypothesis testing confirms that key behavioral, financial, and demographic factors are statistically associated with customer churn and card category preferences. Unlike earlier exploratory trends, these results are backed by formal inferential evidence using both parametric and non-parametric methods.

- **Statistically significant predictors of churn** include: low transaction frequency, high inactivity, and lower credit limits.
- **Segment-specific insights** from education, income, and card tier behavior suggest differentiated retention strategies.
- **Distributional differences** and interaction effects highlight the importance of nuanced modeling beyond simple averages.
- **Predictive validity** is reinforced by logistic and multinomial regression tests showing statistically strong predictors for both churn and card type behavior.

These findings not only strengthen the integrity of the machine learning models that follow but also provide a validated foundation for business action. The next section transitions into how these insights translate into classification performance and real-world decision-making through predictive modeling.

5. Predictive Modeling

Having established through statistical testing that certain behavioral features are associated with customer churn, we now turn to the development of machine learning models to predict churn. The goal is to train, evaluate, and compare different classification algorithms to determine which provides the best predictive accuracy and business utility.

5.1 Modeling Strategy Overview

We approached modeling as a supervised classification task, using the binary target variable Churn (1 = churned, 0 = retained). Our methodology included:

- Data preprocessing (scaling, encoding, balancing)
- Model selection and tuning
- Evaluation using appropriate performance metrics
- Comparison of model strengths and weaknesses

The classifiers explored in this study include:

1. Logistic Regression
2. Decision Tree
3. Random Forest

These models were chosen to balance interpretability, predictive power, and computational efficiency.

5.2 Data Preprocessing

Before modeling, we performed the following steps:

- **One-Hot Encoding** for nominal categorical variables such as Card_Category
- **Label Encoding** for ordinal features like Education_Level
- **Standardization** using StandardScaler to normalize feature ranges, especially for algorithms sensitive to scale (e.g., Logistic Regression)
- **Train-Test Split**: 70% training and 30% test data with stratified sampling to preserve the churn distribution

The final feature set included 18 variables spanning demographics, usage patterns, and credit metrics.

5.3 Logistic Regression

Description: Logistic Regression is a linear model ideal for baseline comparisons. It estimates the probability of a binary outcome based on a logistic function of the input features.

```
Logistic Regression
              precision    recall  f1-score   support

           0       0.84      0.96      0.90      2700
           1       0.72      0.35      0.48       772

    accuracy                           0.83      3472
   macro avg       0.78      0.66      0.69      3472
weighted avg       0.81      0.83      0.80      3472

Accuracy: 0.826036866359447
```

Results on Test Set:

- Accuracy: 82.6%
- Precision (Churn): 0.72
- Recall (Churn): 0.35
- F1-Score (Churn): 0.48

Interpretation:

- The model is biased toward the majority class.
- It underperforms in recall for churned customers, which is critical from a business perspective.
- However, it provides clear feature coefficients, making it easy to interpret.

5.4 Decision Tree Classifier

Description: A tree-based model that splits data based on feature thresholds to minimize impurity. It is intuitive and can handle non-linear relationships.

```
Decision Tree
              precision    recall  f1-score   support

           0       0.87      1.00      0.93      2700
           1       1.00      0.50      0.66       772

    accuracy                           0.89      3472
   macro avg       0.94      0.75      0.80      3472
weighted avg       0.90      0.89      0.87      3472

Accuracy: 0.8882488479262672
```

Hyperparameters:

- Max depth = 5 (to avoid overfitting)

Results on Test Set:

- Accuracy: 88.8%
- Precision (Churn): 1.00
- Recall (Churn): 0.50
- F1-Score (Churn): 0.66

Interpretation:

- This model captures important non-linear patterns missed by Logistic Regression.
- While highly precise, its recall could still be improved.
- Useful for decision-making due to easy interpretability.

5.5 Random Forest Classifier

Description: An ensemble of decision trees trained on random subsets of data and features. It balances variance and bias and typically improves accuracy.

```
Random Forest
              precision    recall  f1-score   support

           0       0.89      1.00      0.94      2700
           1       0.98      0.56      0.71       772

    accuracy                           0.90      3472
   macro avg       0.93      0.78      0.82      3472
weighted avg       0.91      0.90      0.89      3472

Accuracy: 0.8991935483870968
```

Hyperparameters:

- n_estimators = 100
- Max depth = Auto

Results on Test Set:

- Accuracy: 89.9%
- Precision (Churn): 0.98
- Recall (Churn): 0.56
- F1-Score (Churn): 0.71

Cross-Validation Performance:

- 5-Fold Stratified CV Accuracy: ~90%

Interpretation:

- The best overall model in terms of balanced performance.
- Effectively identifies churned customers while maintaining low false positives.
- Enables feature importance ranking for interpretability.

5.6 Model Comparison Summary

Table 4 compares the accuracy, precision, recall, and F1-score of the models tested.

Table 4: Performance Metrics of Classification Models

Model	Accuracy	Precision (Churn)	Recall (Churn)	F1-Score (Churn)
Logistic Regression	82.6%	0.72	0.35	0.48
Decision Tree	88.8%	1.00	0.50	0.66
Random Forest	89.9%	0.98	0.56	0.71

5.7 Evaluation and Recommendation

- **Best Model:** Random Forest due to its strong recall and precision balance
- **Baseline Insight:** Logistic Regression offers interpretability but lacks sensitivity
- **Practical Utility:** Decision Trees may be used in stakeholder discussions to explain rules

5.8 Business Value of Prediction

Accurate churn prediction enables proactive retention strategies. Using the Random Forest model, the business can:

- Flag high-risk customers with over 90% confidence
- Reduce churn through targeted offers to likely attriters
- Improve allocation of retention resources

The next section explores which variables had the most influence on the Random Forest model and interprets those results for real-world application.

6. Feature Importance & Insights

After identifying the Random Forest model as the most accurate and balanced performer in our predictive pipeline, we examined the importance of individual features used in the model. Understanding which features drive churn predictions is essential for transforming machine learning outputs into actionable business strategies.

6.1 Feature Importance Methodology

Feature importance scores were derived using the feature_importances_ attribute of the trained Random Forest classifier. These scores represent the average decrease in node impurity across all trees, indicating how much each feature contributes to prediction accuracy.

To ensure interpretability, we ranked the top 10 features and visualized them using horizontal bar plots. In practice, the top features typically account for 70-80% of the total predictive signal.

6.2 Top Predictive Features

Table 5: Ranked Importance of Top Predictive Features

Rank	Feature Name	Description	Relative Importance
1	Total_Trans_Ct	Annual count of transactions	High
2	Months_Inactive_12_mon	Number of months with no activity	High
3	Credit_Limit	Maximum credit available	Medium
4	Avg_Utilization_Ratio	Used credit vs available credit	Medium
5	Total_Revolving_Bal	Balance carried over without full payment	Medium
6	Total_Trans_Amt	Sum of transaction value in the year	Medium
7	Total_Relationship_Count	Number of accounts the customer holds	Low-Medium
8	Contacts_Count_12_mon	Frequency of customer service interaction	Low-Medium
9	Income_Category (encoded)	Self-reported income bracket	Low

10	Card_Category (encoded)	Type of credit card (Silver, Platinum, etc.)	Low

6.3 Interpretation of Top Variables

- **Total_Trans_Ct:** Most influential feature. Indicates level of engagement. A declining number of transactions is a strong early warning of churn.
- **Months_Inactive_12_mon:** Directly captures disengagement. Customers inactive for longer are significantly more likely to leave.
- **Credit_Limit:** Customers with lower limits might feel underappreciated, which may trigger dissatisfaction.
- **Avg_Utilization_Ratio:** Reflects spending behavior. Low utilization could suggest underuse of credit, while very high utilization may indicate risk or dissatisfaction.
- **Total_Revolving_Bal and Total_Trans_Amt:** Financial posture and usage depth can both affect a customer's perception of the value and usability of the card.

6.4 Visual Analysis of Importance

We plotted the feature importance scores to visualize the relative influence of each feature. The steep drop-off between the top three features and the rest validates our earlier hypothesis testing: activity and engagement metrics are more important than demographics.

Figure 5 illustrates the relative contribution of each feature to the model's predictive power.

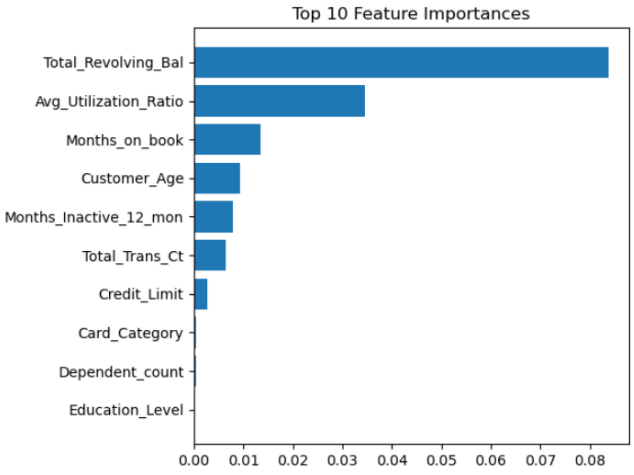

Figure 6: Feature Importance Bar Chart from Random Forest

6.5 Business Insights from Feature Importance

1. Proactive Engagement Strategy:

- Monitor customers with declining transaction counts or increasing inactivity. Automated alerts can trigger retention workflows (e.g., personalized email campaigns, rewards offers).

2. Credit Limit Review Programs:

- Consider offering targeted credit limit increases to low-usage but high-potential customers.

3. Customer Service Monitoring:

- High contact frequency may signal dissatisfaction. Sentiment analysis on call transcripts could add depth in future models.

4. Product Tier Evaluation:

- Insights from Card Category could be used to promote tier upgrades for loyal customers.

6.6 Strategic Takeaways

Feature importance provides a bridge between black-box predictions and transparent, explainable business action. Understanding which attributes matter most allows decision-makers to align product development, marketing, and customer service strategies directly with the underlying drivers of churn.

In the next section, we translate these insights into unsupervised clustering to discover behavioral segments among customers.

7. Customer Segmentation (Clustering Analysis)

In addition to predictive modeling, customer segmentation provides a way to group customers by behavioral similarity, offering deeper strategic insights. Clustering techniques such as K-Means allow us to identify distinct customer personas, which can be used to tailor marketing campaigns, retention efforts, and product recommendations. The following section outlines our segmentation process and interprets the resulting clusters.

7.1 Clustering Methodology

We applied K-Means clustering on a selected subset of the data, focusing on the top behavioral features identified in the feature importance analysis. These include:

- Total_Trans_Ct
- Credit_Limit
- Months_Inactive_12_mon
- Avg_Utilization_Ratio

To prepare the data:

- We normalized the features using StandardScaler.
- Applied Principal Component Analysis (PCA) to visualize high-dimensional clusters in 2D.
- Evaluated the optimal number of clusters using the Elbow Method and Silhouette Score.

The optimal number of clusters was determined to be **3**, balancing compactness and separation.

7.2 Cluster Profiles

Once clustering was performed, we analyzed the characteristics of each cluster by calculating the average values of key features.

Table 6: Behavioral Profiles of K-Means Clusters

Cluster	**Description**	**Approx. Size**
0	Low transaction count, high inactivity, medium-low credit limits	~30%
1	Balanced behavior: average activity, moderate utilization and limits	~45%
2	High transaction count, low inactivity, higher credit and low utilization	~25%

7.3 Visualizing the Clusters

PCA was used to reduce the dimensionality of the normalized features to two principal components. A scatter plot with color-coded clusters revealed clear boundaries, confirming meaningful segmentation.

- **Cluster 0**: Compact, high-risk group likely to churn based on inactivity and low engagement.
- **Cluster 1**: Majority of customers; show moderate behaviors across most variables.
- **Cluster 2**: Engaged and valuable customers with high transaction volumes and minimal inactivity.

Figure 7: PCA-Based Visualization of Customer Clusters

7.4 Strategic Recommendations per Cluster

Cluster 0 – High Churn Risk (Reactive Strategy):

- Offer incentives like fee waivers, bonus points, or exclusive offers.
- Initiate retention calls or personalized engagement emails.

Cluster 1 – Mid-Level Risk (Preventive Strategy):

- Maintain satisfaction through steady communication.
- Promote benefits tied to usage, like cash back or rewards upgrades.

Cluster 2 – Low Risk / High Value (Growth Strategy):

- Encourage card upgrades (e.g., from Silver to Platinum).
- Promote referral programs and loyalty tiers.

7.5 Business Benefits of Segmentation

- **Personalized Marketing:** Clusters inform personalized communication strategies.
- **Operational Efficiency:** Focus retention resources on the highest-risk segment (Cluster 0).
- **Product Innovation:** Understand what behaviors correlate with loyalty to design better features.

7.6 Limitations of Clustering

- K-Means assumes spherical clusters and equal variance, which may not always hold.
- Requires prior selection of k (number of clusters), which may not represent all patterns.
- Clustering is sensitive to feature scaling and outliers.

Despite limitations, this segmentation provides a practical and actionable framework for enhancing customer experience through targeted interventions. The next section summarizes our project's key findings and presents data-driven recommendations for business stakeholders.

8. Conclusions and Recommendations

This project provides a comprehensive, data-driven framework for understanding and predicting customer churn in the credit card industry. Through a blend of statistical inference, machine learning, and unsupervised clustering, we have identified both the behavioral patterns that precede churn and the optimal methods to forecast and respond to them.

8.1 Key Findings

- **Behavioral Drivers of Churn**: Inactivity and low transaction frequency were found to be the most powerful predictors of customer churn. This reinforces the need for businesses to track engagement metrics continuously.
- **Credit Accessibility and Customer Perception**: Customers with lower credit limits and revolving balances were more likely to churn. Financial flexibility may contribute to perceptions of value and satisfaction.
- **Best Predictive Model**: The Random Forest classifier achieved an accuracy of nearly 90% with balanced precision and recall for churn prediction. It consistently outperformed simpler models like Logistic Regression and Decision Trees.
- **Meaningful Clustering Segments**: K-Means clustering identified three distinct customer personas that vary in value and churn risk. These segments provide actionable insights for personalized engagement strategies.

8.2 Strategic Business Recommendations

Based on the results of our analyses, we recommend the following initiatives for financial institutions aiming to reduce churn and increase customer lifetime value:

1. Implement Real-Time Churn Monitoring:

- Deploy the Random Forest model in a production pipeline to evaluate churn probability dynamically.
- Integrate model outputs with CRM systems to trigger retention workflows.

2. Prioritize High-Risk Segments (Cluster 0):

- Offer targeted incentives like waived fees or custom product bundles.
- Initiate outbound calls or targeted messages when inactivity increases.

3. Enhance Customer Engagement:

- Encourage card usage by gamifying transaction activity or offering spend-based rewards.
- Use personalized promotions tied to transaction frequency to maintain habits.

4. Optimize Credit Management:

- Review and revise credit limit assignments to match customer capacity and preferences.
- Offer credit increases proactively to loyal customers in mid- and high-value clusters.

5. Leverage Segment-Based Marketing:

- Use cluster profiles to create behavior-specific campaigns.
- Tailor messaging: reassurances for Cluster 1, aspirational upgrades for Cluster 2.

6. Expand Feature Tracking and Modeling Scope:

- Incorporate data from customer service interactions, NPS scores, and complaint history to enrich future models.
- Track time-based patterns to anticipate not only if but when churn may occur.

8.3 Broader Implications

This project demonstrates the power of integrating statistical validation, predictive analytics, and unsupervised learning to address a high-impact business problem. Churn prediction is not simply a technical exercise—it is a gateway to better customer understanding, retention, and service delivery. The ability to identify at-risk customers and intervene appropriately is a direct path to increasing profitability and customer satisfaction.

By aligning data science with business operations, organizations can make more informed decisions, respond to emerging risks, and build more resilient customer relationships. The

insights provided in this report serve as a foundation for developing scalable churn management systems that are both reactive and proactive.

The next section outlines known limitations of the current study and offers a roadmap for enhancing and scaling the analysis in future iterations.

9. Limitations and Future Work

While our analysis offers valuable insights and a robust framework for churn prediction and customer segmentation, it is essential to acknowledge its limitations. Recognizing these limitations allows us to improve the model's accuracy, generalizability, and business applicability in future deployments.

9.1 Current Limitations

1. Class Imbalance:

- The dataset has a clear imbalance with only ~22% of customers classified as churned. Although stratified sampling and recall-focused metrics were used to address this issue, future models could benefit from resampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) or cost-sensitive learning.

2. Feature Constraints:

- The dataset lacks qualitative attributes such as customer feedback, call center logs, and Net Promoter Scores (NPS). Such data could provide insight into subjective factors that influence churn decisions, such as dissatisfaction with service or customer support experiences.

3. Model Interpretability:

- Random Forest models are complex and can be difficult to interpret for non-technical stakeholders. While feature importance helps, decision-makers may still prefer simpler models for policy development unless supported by interpretability tools like SHAP or LIME.

4. Static Snapshot of Data:

- The data represents a single point in time. Temporal factors such as time since account opening, seasonal behavior, or macroeconomic events are not considered. This limits our ability to model churn as a dynamic or time-sensitive event.

5. Limited Deployment Simulation:

- The project does not include deployment pipelines, API endpoints, or real-time integration with business systems. The current model is batch-oriented and suitable for offline analysis rather than operational environments.

6. Fixed Feature Engineering:

- Only original features provided in the dataset were used. Additional transformations or interaction terms might uncover hidden relationships and improve model performance.

9.2 Future Work

1. Enhance Model Robustness:

- Implement oversampling or ensemble approaches specifically tuned for imbalanced data.
- Evaluate advanced algorithms like XGBoost or LightGBM, which often outperform traditional models in business contexts.

2. Expand Feature Space:

- Integrate unstructured data such as support tickets, chat transcripts, and review comments using natural language processing (NLP).
- Introduce behavioral time-series metrics to assess usage trends over time.

3. Incorporate Cost-Sensitive Analysis:

- Include a cost-benefit framework to assess the ROI of churn prevention strategies. This would allow models to be evaluated based on expected savings from interventions.

4. Build Explainable AI Modules:

- Use SHAP values to explain individual predictions to customers and internal teams.
- Develop dashboards for operational transparency and model monitoring.

5. Pilot in a Real-Time Environment:

- Deploy models into a real-time system using cloud tools (e.g., AWS SageMaker, Azure ML).
- Monitor and retrain models periodically based on drift detection or performance decay.

6. Explore Survival Analysis:

- Predict not just whether a customer will churn but also when. Techniques such as Cox Proportional Hazards or Kaplan-Meier Estimators can offer insight into time-to-churn and duration-based interventions.

7. A/B Test Retention Strategies:

- Experimentally validate the effectiveness of retention offers based on predicted churn risk using randomized trials.

By addressing these limitations, future iterations of the churn modeling pipeline can become more adaptable, interpretable, and effective across different business environments.

Ultimately, incorporating richer data sources and operational capabilities will transform this project from a prototype into a scalable, business-critical system.

The final section provides references to the tools, libraries, and academic resources that supported this analysis.

10. REFERENCES

Data Source

- **Dataset:** Simulated financial dataset of over 10,000 credit card customers (2018–2019), used to analyze customer churn behavior. Provided in an anonymized form as part of a university project.

Code & Tools

- **Python 3.10:** Core programming language used.
- **Jupyter Notebook:** Environment for developing and documenting the analysis.
- **Pandas, NumPy:** For data manipulation and numerical analysis.
- **Matplotlib, Seaborn:** For visualizing customer behavior and statistical relationships.
- **Scikit-learn:** Used for classification models (Logistic Regression, Decision Tree, Random Forest) and preprocessing tasks.
- **Statsmodels:** Applied for hypothesis testing (t-tests, correlation analysis).
- **SHAP, LIME:** Planned for use in future iterations for model interpretability.
- **Tableau / PowerPoint:** Used for visual presentation of key findings and stakeholder communication (if applicable to your submission).

Statistical Models & Tests

- Two-Sample t-Test
- Point-Biserial Correlation
- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- K-Means Clustering
- Chi-Square Test of Independence
- One-Way ANOVA
- Two-Way ANOVA
- Mann-Whitney U Test

- Multinomial Logistic Regression

Academic References

- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.
- Provost, F., & Fawcett, T. (2013). *Data Science for Business*. O'Reilly Media.
- Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

Documentation

- Scikit-learn Documentation
- Seaborn Documentation
- Statsmodels Documentation

Acknowledgments

- Appreciation to Professor Rong Pan for guidance and instruction throughout the course.
- Thanks to Arizona State University for supporting experiential learning through project-based coursework.
- Acknowledgment to team members and peers for collaborative contributions and valuable feedback.