Technical Report

Project Title: Customer Churn Analysis and Retention Strategy

1. Introduction

Customer churn is a major challenge for subscription-based businesses, as losing existing customers directly impacts revenue and increases customer acquisition costs. Understanding why customers leave is essential for designing effective retention strategies. This project presents an end-to-end business analysis of customer churn using real-world data, combining data preprocessing, exploratory analysis, statistical validation, and predictive modeling to generate actionable business insights.

2. Business Problem and Objectives

The business is facing a growing number of customers discontinuing their services. The objective of this project is to analyze customer behavior, identify key drivers of churn, and propose data-driven recommendations that can help reduce churn and improve customer retention.

Objectives:

- Identify factors influencing customer churn
- Analyze customer behavior patterns
- Build predictive models to identify churn-prone customers
- Provide actionable business recommendations

3. Dataset Description

The dataset used in this project consists of 500 customer records with the following attributes:

- Tenure
- Monthly Charges
- Total Charges
- Contract Type
- Payment Method
- Paperless Billing
- Senior Citizen Status
- Churn (target variable)

The dataset meets the minimum size requirement and is suitable for business analytics and machine learning tasks.

4. Project Architecture and Workflow

The project follows a structured and modular workflow aligned with industry best practices. The analysis is divided into distinct stages: data cleaning, exploratory data analysis, and advanced analysis. Each stage produces outputs that are used by subsequent stages, ensuring reproducibility and clarity.

The raw dataset is preserved separately, and all transformations are applied only to processed data to maintain data integrity.

5. Data Cleaning and Preparation

Data preprocessing was performed using Python and the Pandas library. The raw dataset was loaded, and non-informative identifiers were removed to prevent bias. Missing values in numerical fields were handled using median imputation to reduce the effect of outliers. Categorical variables such as contract type and payment method were converted into numerical format using one-hot encoding to make them suitable for analysis and modeling.

After preprocessing, the cleaned dataset was saved as a separate CSV file to ensure consistency across exploratory analysis and modeling stages. This approach allows the analysis to be reproduced and audited easily.

6. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to understand data distributions and relationships between variables. Visualizations were created using Matplotlib and Seaborn to identify patterns and trends.

Key visual analyses included:

- Distribution of churned and non-churned customers
- Comparison of monthly charges between churned and retained customers
- Tenure distribution segmented by churn status
- Impact of contract type on churn
- Correlation analysis among numerical variables

EDA revealed that customers with shorter tenure and higher monthly charges are more likely to churn. Month-to-month contracts showed a significantly higher churn rate compared to long-term contracts.

7. Statistical Analysis

To validate insights observed during EDA, statistical hypothesis testing was performed. An independent t-test was conducted to determine whether there is a significant difference in

monthly charges between churned and non-churned customers. The test results indicated a statistically significant difference, confirming that pricing plays an important role in customer churn behavior.

8. Predictive Modeling

Predictive analysis was performed to identify customers at high risk of churn. The dataset was split into training and testing sets to evaluate model performance objectively.

Two machine learning models were implemented:

- Logistic Regression
- Decision Tree Classifier

Logistic Regression was chosen for its interpretability and ability to estimate churn probabilities. The Decision Tree model was used to capture non-linear relationships and provide intuitive decision rules. Model performance was evaluated using accuracy and classification metrics such as precision, recall, and F1-score.

Both models successfully identified churn-prone customers, with Logistic Regression providing stable and explainable results, while the Decision Tree offered insights into feature-based decision paths.

9. Code Implementation Overview

The project was implemented using Python with a focus on clarity and maintainability. Pandas was used for data manipulation, NumPy for numerical operations, Matplotlib and Seaborn for visualization, Scikit-learn for machine learning models, and SciPy for statistical testing.

The code was structured to follow the analytical workflow, ensuring that data preprocessing, visualization, and modeling were clearly separated. This modular approach improves readability, debugging, and future extensibility of the project.

10. Results and Insights

The analysis identified tenure, monthly charges, and contract type as the most influential factors contributing to customer churn. Customers with shorter tenure and higher charges on month-to-month contracts were found to be at the highest risk of churn.

Predictive models demonstrated that churn behavior can be effectively estimated using historical customer data, enabling proactive retention strategies.

11. Business Recommendations

Based on the findings, the following recommendations are proposed:

- Implement targeted retention programs for new customers
- Encourage customers to switch from month-to-month to long-term contracts
- Review and optimize pricing strategies for high-risk customer segments
- Integrate churn prediction models into customer relationship management systems

12. Limitations and Future Scope

The analysis is limited to historical customer data and does not include behavioral or interaction-based features. Future work can include customer engagement metrics, real-time churn prediction, and advanced machine learning models to improve performance.

13. Conclusion

This project demonstrates how data-driven analysis and machine learning can be used to address a real business problem. By combining data preprocessing, exploratory analysis, statistical validation, and predictive modeling, meaningful insights were generated that can support strategic decision-making and improve customer retention.