Project Report

By: Vaidehi Purohit (SKS/A2/C83458)

Title

Customer Churn Analysis and Prediction Using Machine Learning

1. Introduction

Customer retention is one of the most critical challenges faced by subscription-based businesses such as telecom, banking, SaaS platforms, and streaming services. Acquiring new customers is significantly more expensive than retaining existing ones. Therefore, predicting which customers are likely to leave (churn) and understanding the reasons behind churn is extremely valuable for businesses.

This project focuses on analyzing customer churn data using data science techniques. The project combines data preprocessing, exploratory analysis, customer segmentation, machine learning modeling, and business interpretation to provide both technical and practical insights.

2. Problem Statement

The main objectives of this project are:

- To understand customer behavior and identify key factors contributing to churn
- To analyze patterns and trends using exploratory data analysis
- To segment customers into meaningful groups
- To build machine learning models that predict customer churn
- To provide actionable business recommendations based on the analysis

The core question addressed is:
"Can we predict which customers are likely to churn and why?"

3. Dataset Overview

The dataset used in this project contains customer-level data from a telecom company. It includes information related to:

- Customer demographics (e.g., gender, senior citizen status, dependents)
- Account information (tenure, contract type, payment method)
- Services subscribed (internet, phone service, streaming services, security features)
- Billing information (monthly charges, total charges)
- Target variable: Churn (Yes/No)

The dataset contains over 7,000 customer records and more than 20 features, making it suitable for both analysis and machine learning modeling.

4. Data Cleaning and Preprocessing

Before performing any analysis, the dataset was carefully cleaned and prepared.

The following preprocessing steps were performed:

- Converted the TotalCharges column into numeric format, as it was originally stored as text
- Handled missing values in TotalCharges using median imputation
- Encoded the target variable Churn into binary format (0 = No, 1 = Yes)
- Removed the customerID column since it is an identifier and does not provide predictive value
- Applied one-hot encoding to categorical variables so that machine learning models could process them
- Ensured that the dataset contained no invalid or inconsistent values before modeling

These steps ensured that the dataset became reliable, consistent, and suitable for both statistical analysis and machine learning.

5. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand customer behavior and identify patterns related to churn.

Key Findings from EDA:

- The overall churn rate was found to be approximately **26%**, meaning around one in four customers leaves the service
- Customers with **short tenure (new customers)** showed significantly higher churn compared to long-term customers
- Customers with **higher monthly charges** were more likely to churn, suggesting price sensitivity
- Customers on **month-to-month contracts** churned far more frequently than those on 1-year or 2-year contracts
- Long-term contract customers demonstrated strong loyalty and lower churn rates

EDA helped transform raw data into meaningful business insights and guided further steps such as segmentation and modeling.

6. Customer Segmentation

To better understand different customer behaviors, segmentation was performed using two important factors:

Segmentation by Tenure:

- 0–12 months
- 12–24 months
- 24–48 months
- 48+ months

Results showed that customers in the first 12 months had the highest churn risk.

Segmentation by Monthly Charges:

- Low charges
- Medium charges
- High charges

Customers in the high monthly charge group showed a higher tendency to churn, indicating dissatisfaction with pricing or perceived value.

Customer segmentation helps businesses target retention strategies more effectively rather than treating all customers the same.

7. Machine Learning Model Building

Two machine learning models were built and compared:

1. **Logistic Regression**
2. **Random Forest Classifier**

The dataset was split into training and testing sets. Features were used to predict whether a customer would churn.

Random Forest performed better than Logistic Regression, indicating that churn depends on complex interactions between multiple factors rather than simple linear relationships.

This modeling step demonstrates practical application of machine learning in solving real-world business problems.

8. Model Evaluation

Models were evaluated using multiple performance metrics:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix

Special attention was given to **recall for churned customers**, because correctly identifying customers who are likely to leave is more important than simply achieving high accuracy.

This evaluation ensures that the model is not only mathematically accurate but also practically useful.

9. Business Recommendations

Based on analysis and model insights, the following recommendations were proposed:

1. **Improve onboarding for new customers**
   Since new customers churn the most, better onboarding, support, and engagement during the first few months can improve retention.
2. **Offer flexible pricing plans**
   High monthly charges increase churn risk. Discounts, customized plans, and loyalty rewards could reduce this risk.
3. **Promote long-term contracts**
   Customers on long-term contracts are more loyal. Incentives should be provided to encourage contract upgrades.
4. **Provide value-added services**
   Customers without services like online security or tech support tend to churn more. Offering trials or bundles could increase engagement.
5. **Use predictive model for proactive retention**
   The churn prediction model can be used to identify high-risk customers early and target them with personalized offers.

10. Limitations of the Project

While the project is effective, it has certain limitations:

- Only historical structured data was used
- No real-time behavioral or interaction data was available
- Class imbalance was not deeply handled
- External factors (competition, customer satisfaction surveys, etc.) were not included

Acknowledging limitations shows critical thinking and professionalism.

11. Future Improvements

If given more time and resources, the project can be enhanced by:

- Using advanced algorithms such as XGBoost or Gradient Boosting
- Applying SMOTE to handle class imbalance
- Performing feature importance analysis more deeply
- Building a real-time web dashboard for predictions
- Deploying the model as an API or web application
- Integrating customer support interaction data

12. Conclusion

This project demonstrates the practical use of data science and machine learning to solve a real business problem. Through structured analysis, customer segmentation, and predictive modeling, meaningful insights were generated about customer churn behavior.

The project not only highlights technical skills such as data preprocessing, EDA, and machine learning, but also emphasizes business thinking by connecting analysis results to real-world retention strategies.

Overall, this project serves as a strong foundation for applying data science techniques to real-world decision-making.

13. Evaluation Metrics

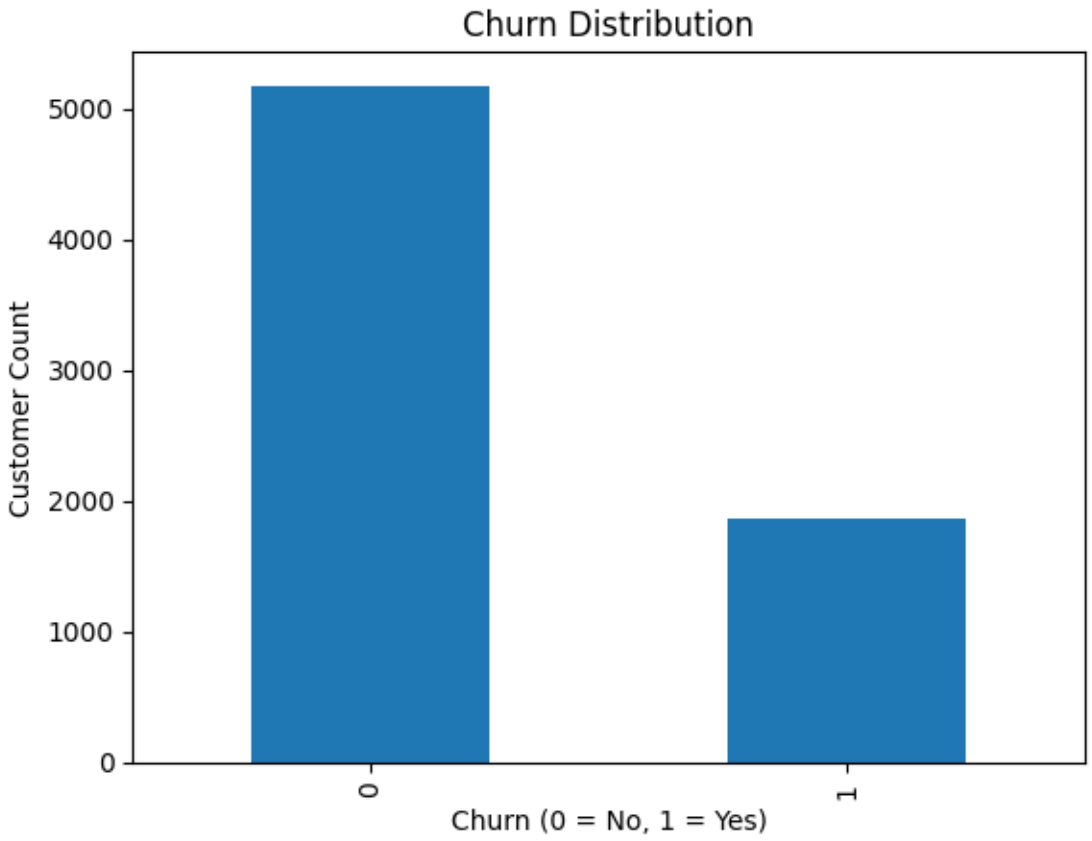

Figure 1: Churn Distribution

Image description:
A bar chart showing the number of customers who did not churn (Churn = 0) versus those who churned (Churn = 1).

This chart illustrates the overall distribution of customer churn. Approximately 26% of customers have churned, indicating a significant retention challenge and highlighting the importance of churn prediction for the business.

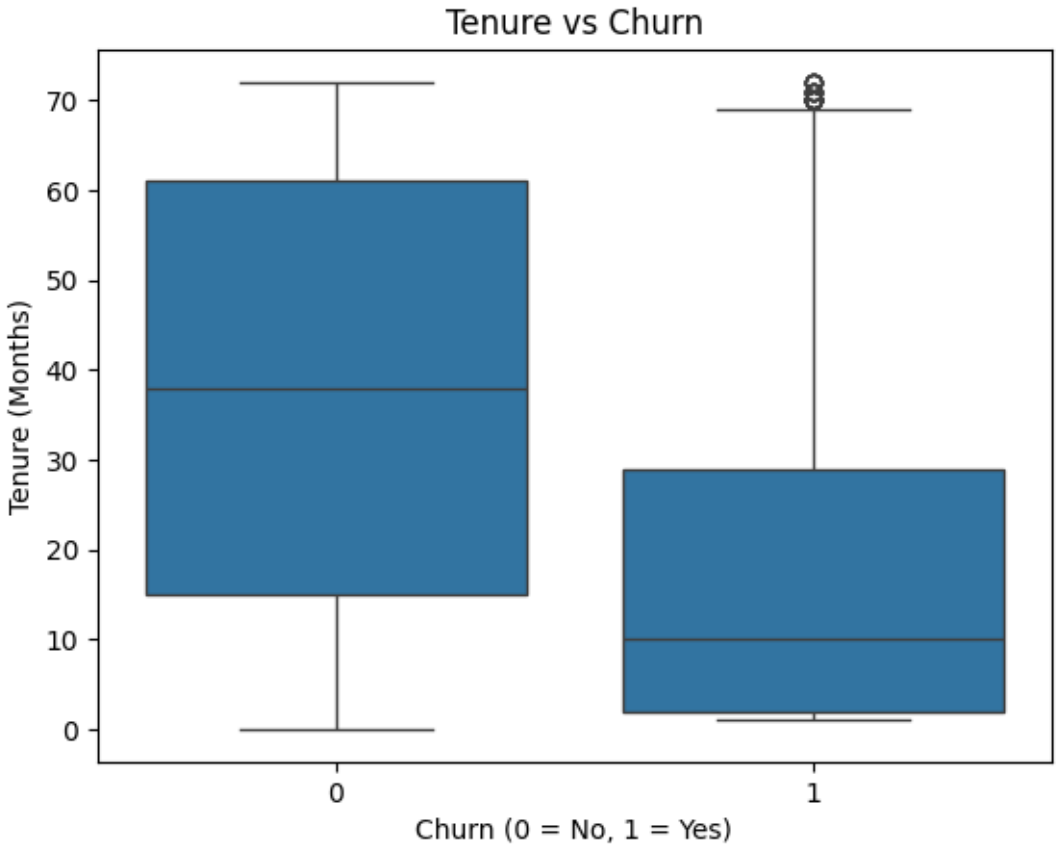

Figure 2: Tenure vs Churn (Boxplot)

Image description:
A boxplot comparing customer tenure (in months) for churned and non-churned customers.

This boxplot shows that customers who churn generally have much lower tenure compared to retained customers. This suggests that new customers are at the highest risk of leaving and require stronger onboarding and engagement strategies.

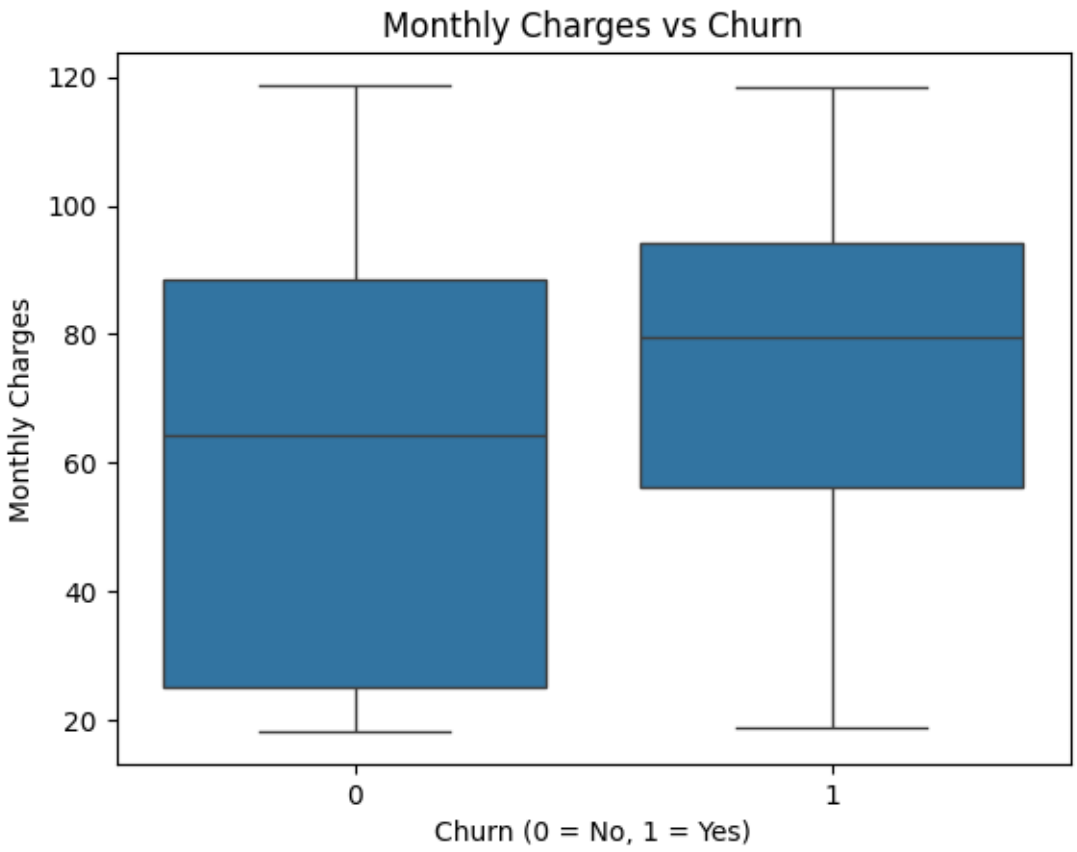

Figure 3: Monthly Charges vs Churn (Boxplot)

Image description:
A boxplot comparing monthly charges for churned versus non-churned customers.

The plot indicates that customers who churn tend to have higher monthly charges than those who stay. This suggests that pricing sensitivity plays an important role in customer dissatisfaction and churn behavior.

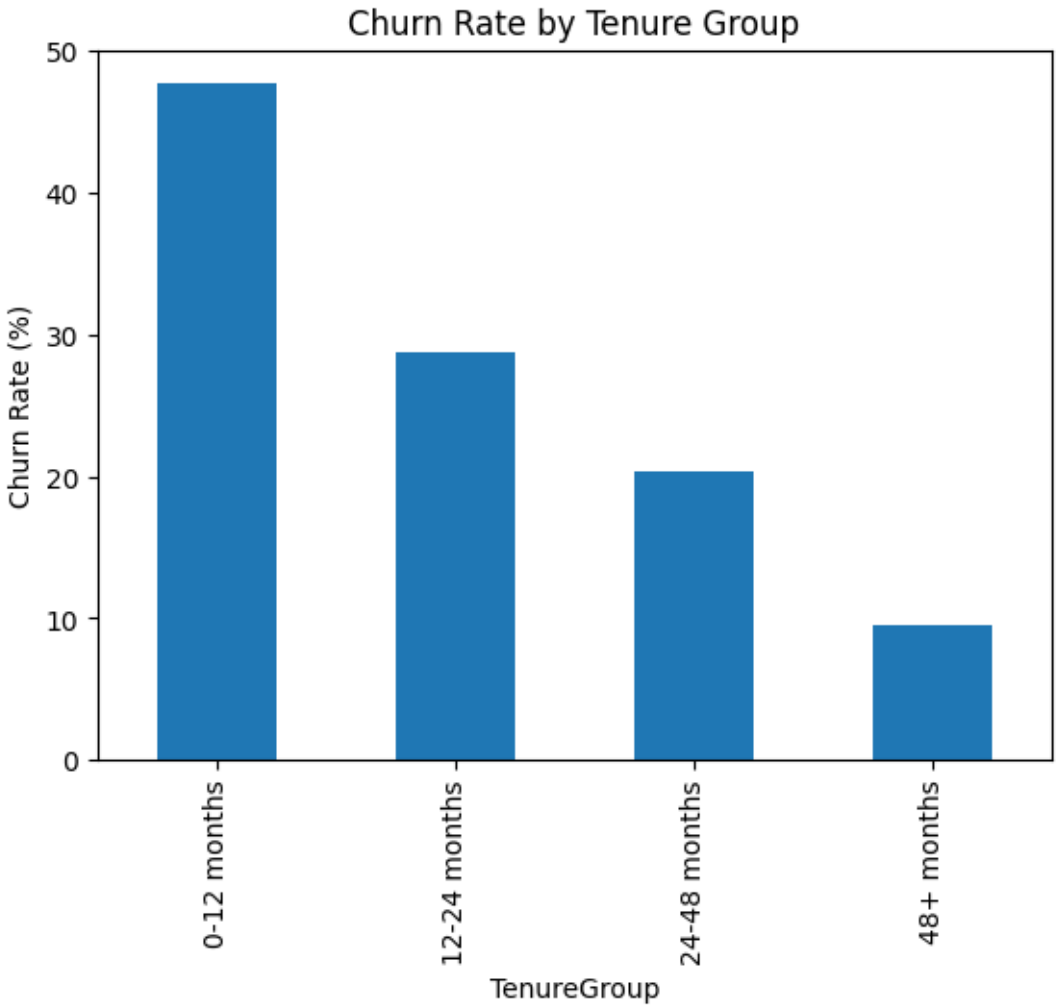

Figure 4: Churn by Tenure Group (Segmentation)

Image description:
A bar chart showing churn rates across tenure segments (0–12 months, 12–24 months, 24–48 months, 48+ months).

The segmentation plot shows that customers within the first 12 months have the highest churn rate, while customers with more than 48 months of tenure have the lowest churn. This confirms that retention efforts should focus strongly on new customers.

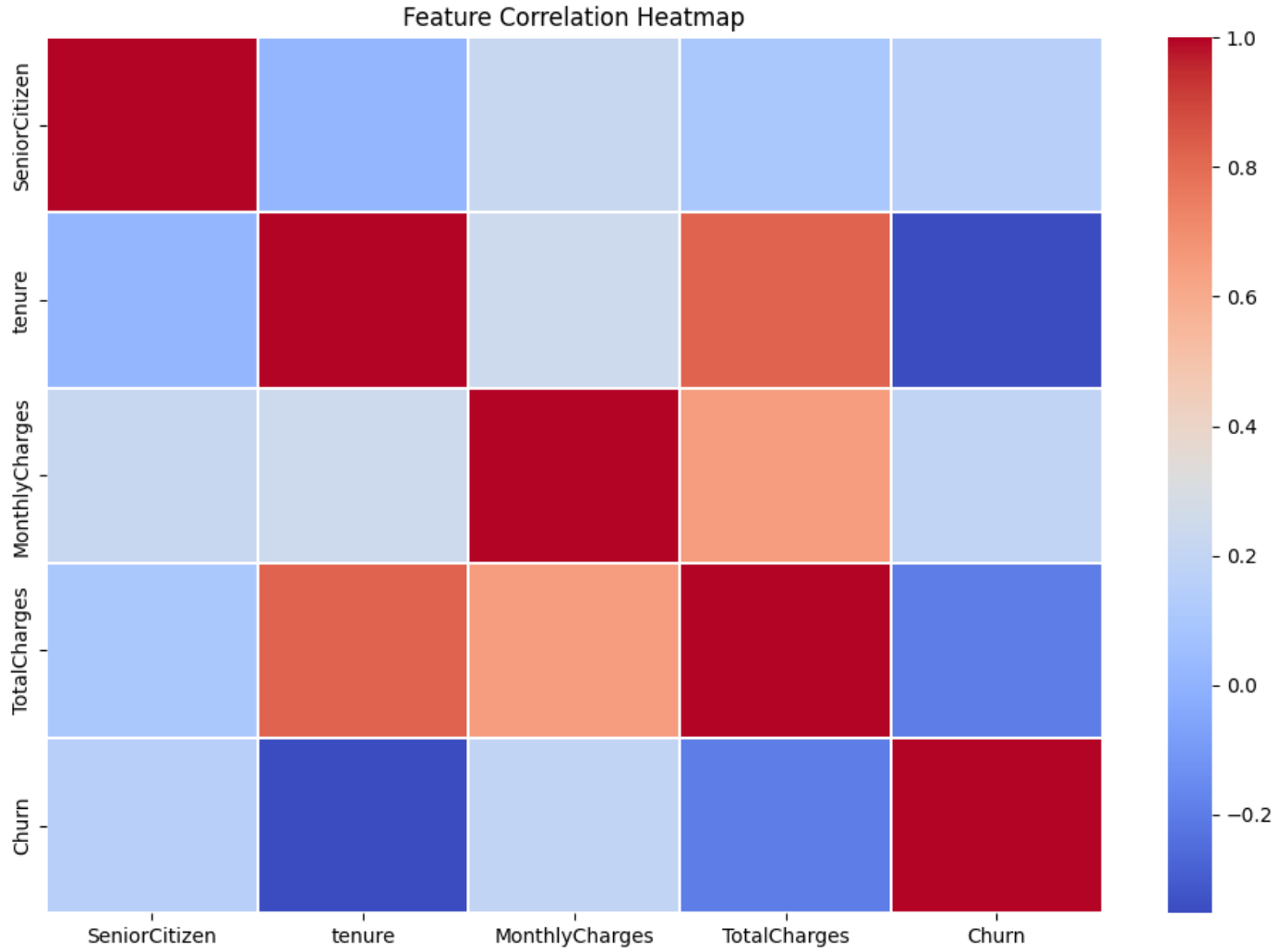

Figure 5: Correlation Heatmap

Image description:
A heatmap showing correlations between numerical features such as tenure, monthly charges, total charges, and churn.

The correlation heatmap shows that features such as tenure, monthly charges, and contract-related variables have noticeable relationships with churn. These correlations help identify which factors are most influential in predicting customer churn.