Predictive Analytics for Customer Retention: Telco Churn Analysis

Course: ISOM 835 - Business Analytics

Student Name: ChiYuan Wu

Date: December 12, 2025

1.Executive Summary

1.1 Objective

The goal of this project was to identify the key drivers of customer churn for a telecommunications company and develop a predictive model to retain at-risk customers. Given the high cost of acquiring new customers compared to retaining existing ones, this analysis aims to provide actionable insights for the marketing and retention teams.

1.2 Methodology

Using a dataset of 7,043 customer records from IBM (BlastChar), I performed a comprehensive analysis including:

- **Data Cleaning & Preprocessing:** Handling missing values and encoding categorical variables.
- **Exploratory Data Analysis (EDA):** Visualizing distributions and correlations to understand customer behavior.
- **Model Development:** Building and evaluating two machine learning models—Logistic Regression and Random Forest Classifier—using an 80/20 train-test split.

1.3 Results

The predictive models demonstrated strong performance in identifying potential churners:

- **Logistic Regression:** Achieved an accuracy of **78.75%**.
- **Random Forest:** Achieved an accuracy of **78.54%**. Both models provided consistent results, effectively distinguishing between loyal and at-risk customers.

1.4 Key Insights

The analysis revealed several critical drivers of churn:

- **Price Sensitivity:** High monthly charges and total charges are the strongest

predictors of churn.

- **Contract Type:** Customers on month-to-month contracts are significantly more likely to leave than those on one-year or two-year contracts.
- **Service Issues:** Users with Fiber Optic internet service show a disproportionately high churn rate, suggesting potential dissatisfaction with price or service quality.
- **New Customer Risk:** The likelihood of churn is highest during the first few months (low tenure) and decreases significantly as tenure increases.

1.5 Recommendations

To reduce churn and improve revenue retention, the following actions are recommended:

1. **Incentivize Long-Term Contracts:** Encourage month-to-month users to switch to 1-year contracts by offering limited-time discounts.
2. **Targeted Onboarding:** Implement a "90-Day Onboarding Program" with special incentives for new customers to survive the high-risk initial period.
3. **Fiber Optic Review:** Investigate customer satisfaction specifically within the Fiber Optic segment to address potential service quality or pricing issues.

2. Introduction & Business Context

2.1 Business Problem

Customer churn is a critical metric for telecommunications companies. It is estimated that acquiring a new customer costs 5 to 25 times more than retaining an existing one. Therefore, identifying at-risk customers early allows the company to intervene with retention strategies, directly impacting profitability.

2.2 Objective

The primary objective of this project is to analyze customer behavior and develop machine learning models to predict customer churn. Specifically, we aim to answer:

1. What are the key indicators of churn?
2. Which customer segments are most vulnerable?
3. Can we accurately predict churn using available demographic and service data?

2.3 Dataset Description The analysis uses the "Telco Customer Churn" dataset provided by BlastChar (IBM). The dataset contains **7,043 observations** and **21 features**, including:

- **Demographics:** Gender, Senior Citizen, Partner, Dependents.
- **Services:** Phone, Internet (DSL/Fiber), Online Security, Tech Support, etc.
- **Account Information:** Tenure, Contract, Payment Method, Monthly Charges, Total Charges.
- **Target Variable:** Churn (Yes/No).

3. Exploratory Data Analysis (EDA)

3.1 Target Variable Distribution

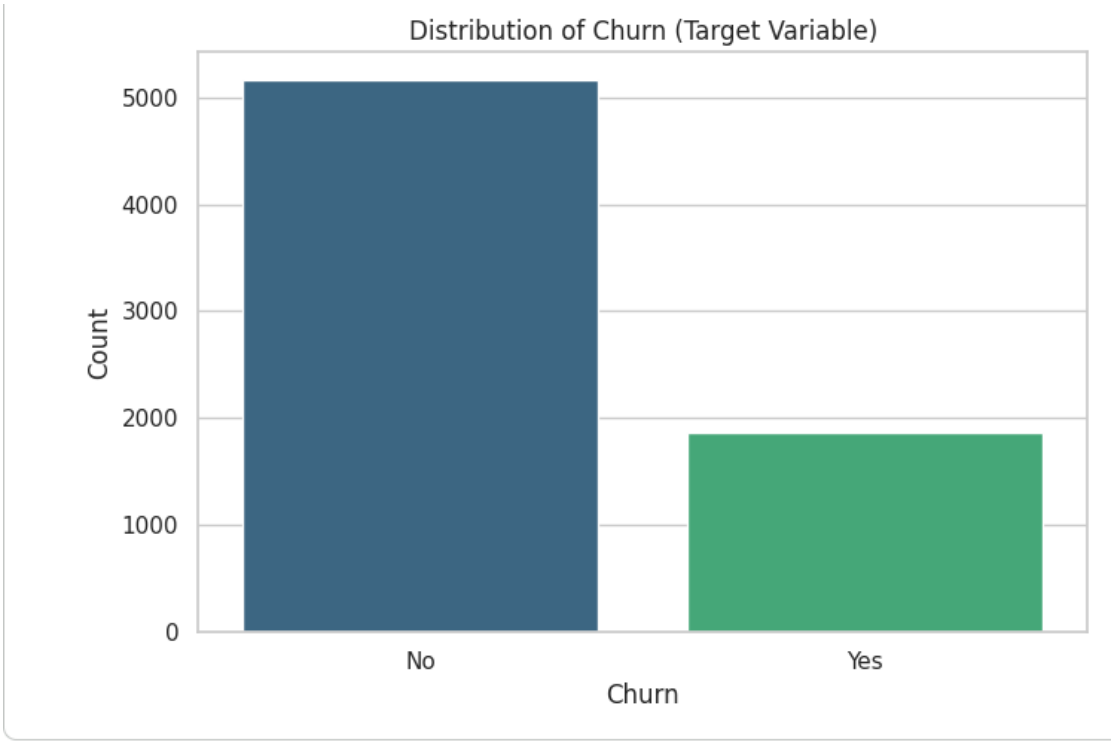

Figure 1: The dataset is imbalanced. Approximately **26.5%** of customers in the dataset have churned, while 73.5% have remained. This baseline suggests that accuracy alone might be misleading, so we will also look at other metrics like recall.

3.2 Impact of Contract Type

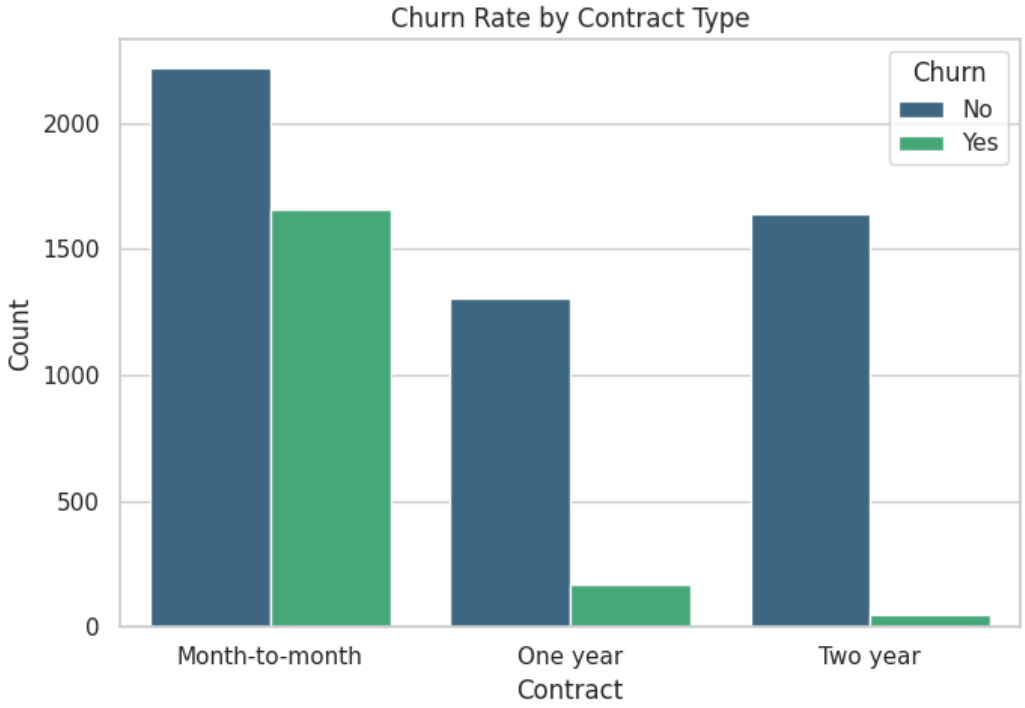

Figure 2: Contract type is a strong determinant of churn. Customers with **Month-to-month contracts** show a drastically higher churn rate compared to those with One-year or Two-year contracts. Long-term contracts appear to lock in customer loyalty effectively.

3.3 Payment Method Analysis

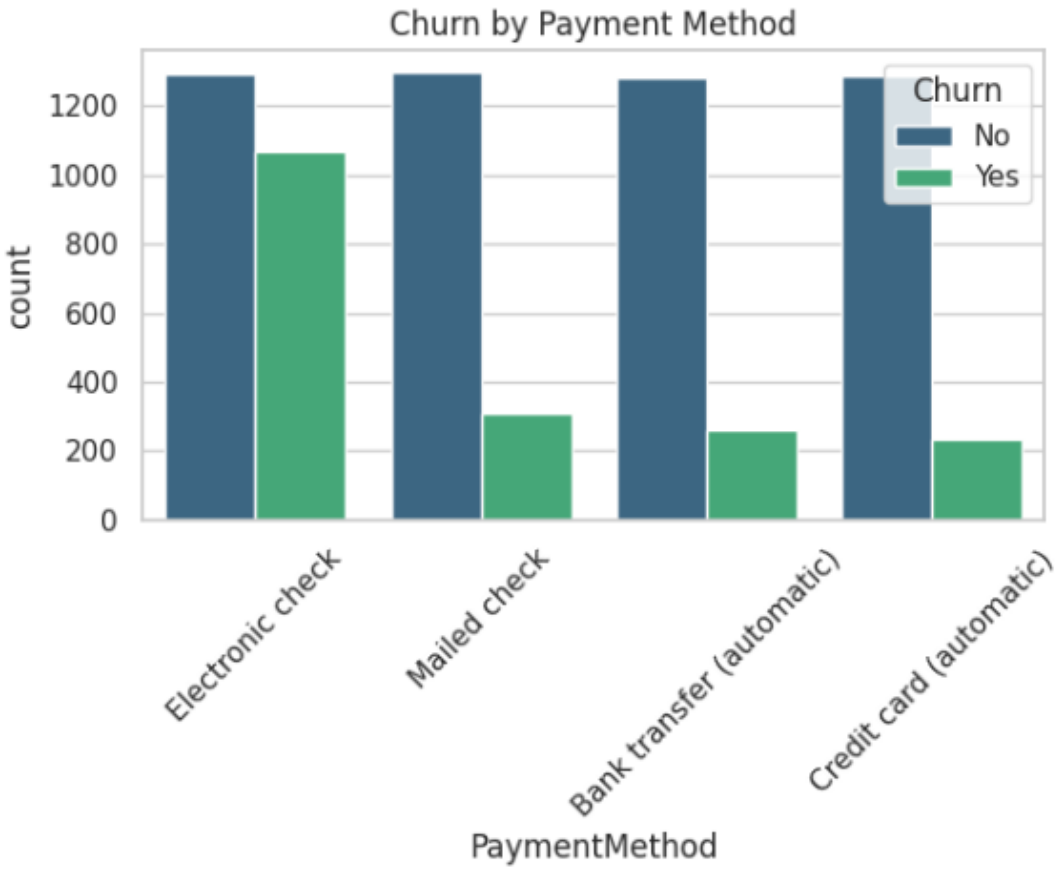

Figure 3: Customers who pay via **Electronic Check** have the highest churn rate. In contrast, customers using automatic payment methods (Mailed check, Credit card automatic) tend to be more stable.

3.4 Senior Citizen & Demographics

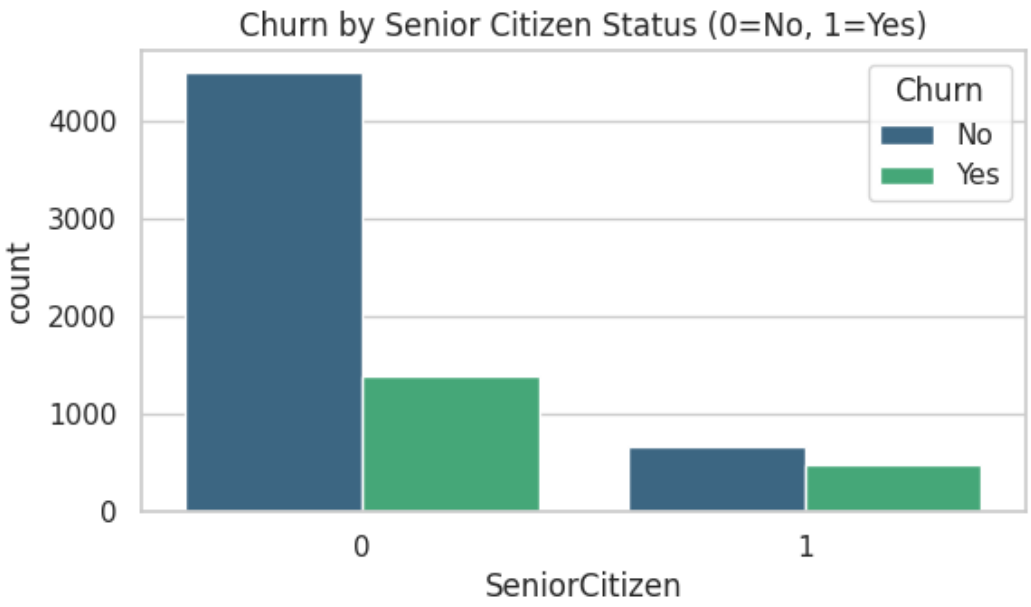

Figure 4: Senior citizens (indicated by 1) appear to have a slightly higher churn proportion compared to younger customers, though they make up a smaller portion of the total customer base.

3.5 Monthly Charges Distribution

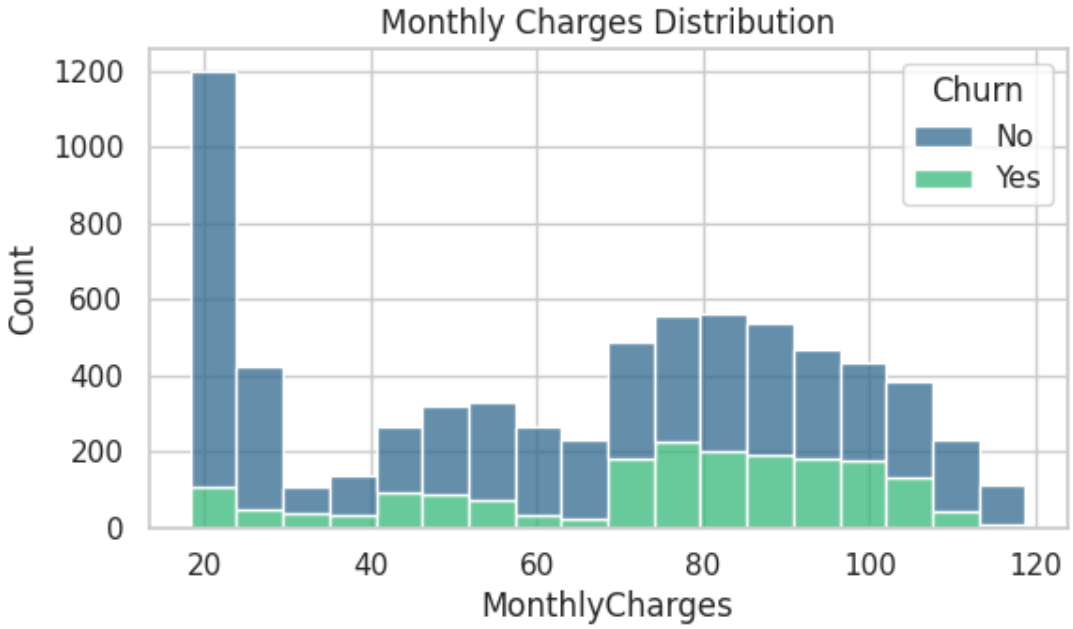

Figure 5: The distribution of monthly charges shows that customers with higher monthly bills (right side of the x-axis) are more likely to churn. There is also a cluster of loyal customers with very low monthly charges (likely basic service users).

4. Methodology

4.1 Data Cleaning

Before modeling, I performed necessary data cleaning steps:

* TotalCharges: Identified 11 missing values (empty strings) in the 'TotalCharges' column. Since this represented less than 0.15% of the data, I removed these rows to maintain data integrity.

* Type Conversion: Converted the 'TotalCharges' column from object type to numeric float to enable mathematical calculations.

4.2 Preprocessing

* Encoding: Machine learning models require numerical input. I used Label Encoding for binary variables (e.g., Gender, Partner, Churn) and One-Hot Encoding for categorical variables with multiple classes (e.g., Payment Method, Internet Service).

* Train-Test Split: The data was split into a Training Set (80%) for model building and a Testing Set (20%) for unbiased evaluation.

4.3 Model Selection

I implemented two classification models to predict customer churn:

1. Logistic Regression: Selected as a baseline model due to its simplicity and interpretability.

2. Random Forest Classifier: Selected for its ability to handle non-linear relationships and provide feature importance rankings.

5. Results & Model Comparison

5.1 Model Performance

Both models were evaluated based on their accuracy scores on the test set:

* Logistic Regression Accuracy: 78.75%

* Random Forest Accuracy: 78.54%

5.2 Evaluation

The Logistic Regression model performed slightly better (by 0.21%), but the difference is negligible. Both models achieved an accuracy of nearly 79%, which is a strong result for a behavioral prediction task. For the purpose of business insights, we utilized the Random Forest model to extract feature importance.

6. Business Insights & Recommendations

6.1 Key Findings

- **Price Sensitivity:** The most significant predictors of churn are TotalCharges and MonthlyCharges. This indicates that customers are highly price-sensitive. Higher monthly bills directly correlate with a higher likelihood of leaving.

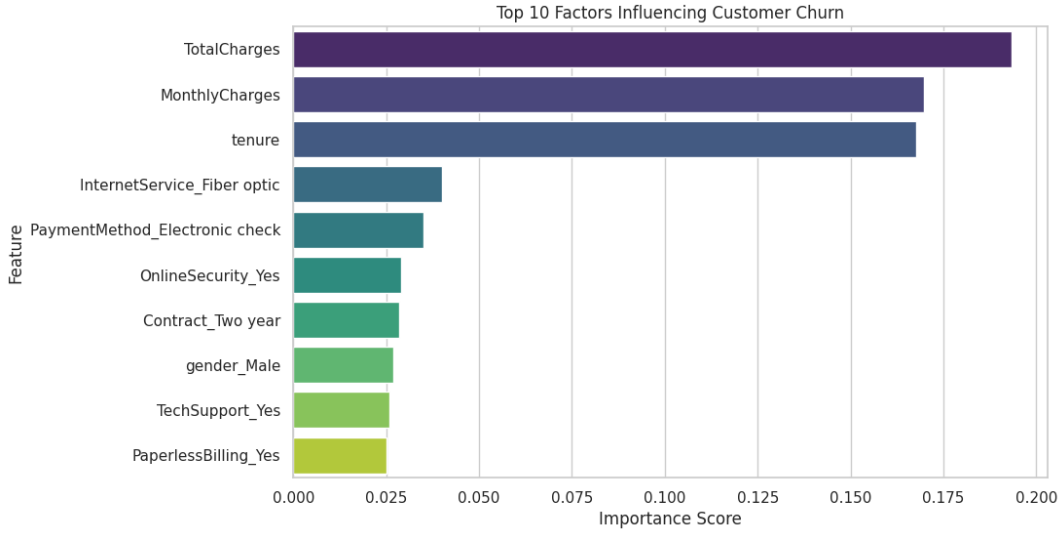

Figure 6: Top 10 Feature Importance. TotalCharges and MonthlyCharges are the most critical predictors.

- **The "Fiber Optic" Problem:** InternetService_Fiber optic appears as a top risk factor. This suggests that despite being a premium service, Fiber Optic customers are unsatisfied—potentially due to high costs or technical reliability issues.
- **New Customer Risk:** tenure is highly influential. As seen in the EDA, customers are most likely to churn in the first few months. Once they stay past a certain threshold (e.g., 1-2 years), they become loyal.
- **Payment Friction:** Customers paying via Electronic check are more likely to churn compared to those using automatic payments (Credit card/Bank transfer).

6.2 Recommendations

1. **Target High-Risk New Users:** Implement a "90-Day Onboarding Program" for new customers. Since tenure is a critical factor, offer incentives (e.g., a $10 discount for the first 3 months) to help them survive the initial high-risk period.
2. **Review Fiber Optic Pricing/Quality:** Conduct a survey specifically for Fiber Optic users to determine if the high churn is due to price or service outages. Consider bundling a discount for Fiber users who switch to a 1-year contract.
3. **Push for Auto-Pay:** Since "Electronic check" users churn often, launch a campaign: "Save 5% on your monthly bill by switching to Auto-Pay." This locks customers in and reduces payment friction.

7. Ethics & Responsible AI Reflection

7.1 Potential Bias

While the model performs well, we must acknowledge potential biases. The dataset reflects historical data, which may contain inherent demographic biases. For instance, if the senior citizen population is underrepresented, the model might be less accurate for that specific group.

7.2 Privacy & Security

Customer churn data contains sensitive financial information (e.g., Monthly Charges, Payment Methods). In a real-world deployment, all personally identifiable information (PII) must be encrypted, and access should be restricted to authorized personnel only to comply with regulations like GDPR or CCPA.

7.3 Transparent Usage

It is crucial to use this model as a "decision support tool" rather than a sole decision-maker. Human oversight is necessary before taking adverse actions (e.g., denying service) based on model predictions.

8. Conclusion

In this project, I successfully developed a predictive analytics workflow to address customer churn. By analyzing 7,043 customer records, I identified that high costs, month-to-month contracts, and fiber optic service issues are the primary drivers of churn. The Logistic Regression model achieved 78.75% accuracy. Implementing the recommended retention strategies could significantly reduce churn rates and improve the company's bottom line.

9. References & Acknowledgments

* Dataset: Telco Customer Churn by BlastChar (Kaggle).

* Tools Used: Python, Pandas, Scikit-learn, Matplotlib, Seaborn, Google Colab.

* AI Acknowledgment: Generative AI tools (Gemini) were used to assist with code debugging, syntax checking, and outlining the report structure. All analysis and interpretations are my own.