Customer Churn Analysis Report

May 24, 2025

1 Introduction

This report analyzes customer churn in a telecommunications company to identify patterns, risk factors, and actionable insights for improving retention. Using Python for data cleaning and visualization and MySQL for querying, the analysis uncovers key drivers of churn and provides recommendations for strategic interventions.

Key Questions Addressed:

- What is the distribution of monthly charges among customers?
- What are the characteristics of churned vs. non-churned customers?
- How do contract types and service usage impact churn?
- Which customer segments are most at risk of churning?
- What are the billing patterns of high-value customers?

The findings aim to guide retention strategies, optimize pricing, and enhance customer satisfaction.

2 Tools Used

- **Python (Pandas, Seaborn, Matplotlib)**: Data cleaning, transformation, and visualization.
- **MySQL**: SQL querying for aggregation, filtering, and window functions.
- **Jupyter Notebook**: Interactive environment for data exploration.
- **Excel (optional)**: Potential for visualization or data export.
- **Git & GitHub (assumed)**: Version control for collaboration.

3 Analysis

3.1 Data Cleaning and Preparation

Problem Statement: Ensure the dataset is clean and ready for analysis by handling missing values and standardizing data types.

Description: The dataset contains 7,043 customer records with 21 attributes, including demographics (gender, SeniorCitizen, Partner, Dependents), service details (PhoneService, InternetService, etc.), billing information (MonthlyCharges, TotalCharges, Contract), and churn status (Churn). Initial checks revealed no null values, but TotalCharges required conversion to numeric, uncovering 11 missing values. These rows were dropped, resulting in 7,032 records.

Approach:

- Checked dataset shape and null values using Pandas.
- Converted TotalCharges to numeric, handling invalid entries.
- Dropped rows with missing TotalCharges and reset the index.

Code Snippet (Python):

```python
# Checking dataset shape and null values
print(df.shape)  # (7043, 21)
print(df.isna().sum())  # Initially no nulls

# Converting TotalCharges to numeric
df['TotalCharges'] = pd.to_numeric(df['TotalCharges'], errors='coerce')
print(df.isna().sum())  # TotalCharges: 11 nulls

# Dropping rows with missing TotalCharges
df = df.dropna()
df.reset_index(drop=True, inplace=True)
print(df.shape)  # (7032, 21)
```

Key Findings:

- Cleaned dataset: 7,032 records, no remaining nulls.
- Data types: 3 numeric (SeniorCitizen, tenure, MonthlyCharges) and 18 categorical.

3.2 Distribution of Monthly Charges

Problem Statement: Understand billing distribution to identify pricing tiers and churn risk factors.

Description: A histogram with a kernel density estimate (KDE) visualizes the distribution of MonthlyCharges, highlighting common price points and potential skewness.

Approach:

- Used Seaborn's histplot with 30 bins and KDE.
- Set figure size to 8x6 with clear labels.

Code Snippet (Python):

```python
import seaborn as sns
import matplotlib.pyplot as plt

# Plotting histogram of MonthlyCharges
plt.figure(figsize=(8, 6))
sns.histplot(data=df, x='MonthlyCharges', bins=30, kde=True)
plt.title('Distribution of Monthly Charges')
plt.xlabel('Monthly Charges ($)')
plt.ylabel('Frequency')
plt.show()
```

Key Findings:

- Charges range from approximately $20 to $120.
- Peaks at $50–$80 suggest popular pricing tiers.
- KDE indicates potential skewness or multi-modal distribution.

3.3 Churn Distribution and Customer Segmentation

Problem Statement: Identify churn proportions and segment customers by key attributes.

Description: SQL queries analyzed churn distribution and segmented customers by gender, tenure, contract type, and service usage to pinpoint high-risk groups.

Approach:

- Counted total customers and churn distribution.
- Analyzed gender distribution and tenure groups.
- Examined churn by Internet service and contract type.

Code Snippet (SQL):

```sql
SELECT Churn, COUNT(*) AS customer_count
FROM telco_churn
GROUP BY Churn;
```

Key Findings:

- Churn rate: Approximately 26% (estimated).
- Gender: Balanced (approximately 50% male, 50% female).
- Tenure: Most customers in 0–12 month range.
- Internet service: Fiber optic users have higher churn.
- Contract: Month-to-month customers at higher risk.

3.4 Billing and Service Usage Insights

Problem Statement: Analyze billing patterns and service combinations impacting churn.

Description: Queries explored average monthly charges by contract, churn by Internet service, and demographic segments (e.g., customers with partner and dependents).

Approach:

- Calculated average MonthlyCharges by Contract.
- Counted churned customers by InternetService.
- Identified family-oriented customers.

Code Snippet (SQL):

```sql
SELECT Contract, AVG(MonthlyCharges) AS avg_monthly_charges
FROM telco_churn
GROUP BY Contract;
```

Key Findings:

- Average charges: Month-to-month ($70), one-year ($60), two-year ($50).
- Fiber optic users: Higher churn, possibly due to cost.
- Family-oriented customers: Potentially lower churn.

3.5 Advanced Analytics with Window Functions

Problem Statement: Use advanced SQL to rank customers and compute aggregated metrics.

Description: Window functions ranked customers by MonthlyCharges, computed running totals of TotalCharges, and calculated contract-specific averages.

Approach:

- Ranked customers by MonthlyCharges using ROW$_{N}UMBER().CalculatedrunningtotalofTotalCharges.$

- Computed average MonthlyCharges per Contract.

Code Snippet (SQL):

```sql
SELECT customerID, MonthlyCharges,
       ROW_NUMBER() OVER (ORDER BY MonthlyCharges DESC) AS rank
FROM telco_churn;
```

Key Findings:

- Top-ranked customers: High-value clients for retention.
- Running total: Shows cumulative revenue trends.
- Contract averages: Highlight pricing differences.

4 What I Learned

Through this project, I gained experience in:

- **Python Data Processing**: Mastered Pandas for cleaning and Seaborn/Matplotlib for visualization.
- **SQL Querying**: Developed skills in aggregations, joins, and window functions.
- **Data Cleaning**: Handled missing values and standardized data types.
- **Analytical Thinking**: Formulated questions to uncover churn drivers.
- **Data Storytelling**: Presented findings through visualizations and reports.

5 Conclusion

This analysis provides a data-driven foundation for understanding Telco customer churn. Key takeaways:

- **Churn Drivers**: Fiber optic users and month-to-month contract holders are at higher risk.
- **Billing Insights**: Higher charges correlate with shorter contracts.
- **Customer Segmentation**: New customers (0–12 months) are a critical focus.
- **Recommendations**:
  - Target fiber optic and month-to-month customers with retention offers.
  - Investigate zero TotalCharges cases for data issues.
  - Develop loyalty programs for new customers.

These insights can guide Telco companies in optimizing pricing, enhancing services, and reducing churn.