Customer Churn Analysis & Retention Strategy for a Subscription Business (OTT/SaaS)

Business Analytics Case Study Report

Note: This report is structured in two parts:

- **Part A (Sections 1-6)**: Technical analysis and findings directly from data analysis
- **Part B (Sections 7-10)**: Strategic business framework and professional presentation materials

1. Executive Summary

This case study examines customer churn patterns for a California-based telecommunications company serving 7,043 subscribers. Through comprehensive data analysis using Python (Pandas, NumPy, Matplotlib, Seaborn), a churn prediction framework was developed that identified 2,387 high-risk customers representing 40.4% of monthly revenue ($184,271).

Key Findings from Analysis:

- Overall churn rate: 26.5%
- High-risk customers demonstrate 54.9% churn probability
- Month-to-month contracts show 42.7% churn vs. 2.8% for two-year contracts
- New customers (< 12 months tenure) exhibit the highest churn vulnerability
- Customers without technical support are 2.7x more likely to churn

Analytical Output: A 4-factor risk scoring model successfully segments customers by churn probability, enabling data-driven prioritization of retention efforts.

PART A: TECHNICAL ANALYSIS & FINDINGS

This section contains the actual data analysis performed, methods used, and empirical findings.

2. Introduction

Project Overview

Project Name: Customer Churn Analysis & Retention Strategy for Subscription Business

Dataset: IBM Telco Customer Churn dataset containing 7,043 customer records with 33 variables including demographics, service usage, contract details, billing information, and churn outcomes from Q3 operations in California.

Analytical Goal: Identify patterns in customer churn behavior and develop a predictive risk classification system using exploratory data analysis techniques.

Analytical Objectives

The analysis aimed to:

1. Understand which customer segments experience higher churn rates
2. Identify key variables associated with customer churn

3. Quantify churn rates across different customer characteristics
4. Develop a risk classification methodology based on observable patterns
5. Calculate revenue exposure from customers exhibiting high-risk characteristics

3. Situation

Dataset Overview

Source: IBM's Telco Customer Churn dataset - a publicly available dataset used in industry to model real-world subscription churn scenarios.

Scope: Analysis of a fictional telco company that provided home phone and Internet services to 7,043 customers in California during Q3.

Data Structure: 7,043 observations with 33 variables including:

- **Customer Demographics:** Gender, Senior Citizen status, Partner, Dependents
- **Location Data:** Country, State, City, Zip Code, Latitude, Longitude
- **Account Information:** Customer ID, Tenure Months
- **Services:** Phone Service, Multiple Lines, Internet Service (DSL/Fiber Optic/Cable), Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies
- **Billing:** Contract type, Paperless Billing, Payment Method, Monthly Charges, Total Charges
- **Churn Indicators:** Churn Label (Yes/No), Churn Value (1/0), Churn Score (0-100), CLTV, Churn Reason

Initial Data Assessment

Upon loading the dataset, initial exploration revealed:

- All 33 columns present with varying data types (6 int64, 3 float64, 24 object)
- Dataset size: 1.8+ MB in memory
- Churn Value column serves as the primary target variable (binary: 0 or 1)

4. Task

Analysis Objectives

Primary Goal: Perform exploratory data analysis to identify patterns and factors associated with customer churn, then develop a scoring mechanism to classify customers by risk level.

Specific Deliverables:

1. Clean and prepare the dataset for analysis
2. Identify key variables correlated with churn through segmentation analysis
3. Quantify churn rates across different customer segments
4. Build a composite risk scoring model based on empirical findings

5. Calculate revenue concentration in high-risk customer segments
6. Visualize patterns for clear communication of findings

Technical Requirements

- Handle missing values and data type inconsistencies
- Perform segmentation analysis across multiple dimensions (tenure, contract type, services, pricing)
- Create visualizations to illustrate churn patterns
- Develop a simple, interpretable scoring methodology
- Calculate aggregate financial metrics

5. Action

Methodology & Analysis Steps

Step 1: Data Loading and Initial Exploration

Tools Used: Python (Pandas, NumPy, Matplotlib, Seaborn) in Jupyter Notebook

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

data = pd.read_csv('Telco_customer_churn.csv')

df = pd.DataFrame(data)
```

Initial Data Profiling:

- Used .info() to examine data types and non-null counts
- Used .describe() to generate statistical summary of numeric variables
- Used .head() and .tail() to inspect sample records
- Examined dataset structure: 7,043 rows × 33 columns

Key Findings from Profiling:

- Count column: All values = 1.0 (constant)
- Zip Code range: 90001 to 96161
- Tenure Months: Mean = 32.37, Range = 0 to 72 months
- Monthly Charges: Mean = $64.76, Range = $18.25 to $118.75
- Churn Value: Binary (0 or 1)

Step 2: Data Quality Assessment and Cleaning

Missing Value Analysis: Used .isna().sum() to identify missing data:

- **Churn Reason:** 5,174 missing values (73.5% of records)
  - *Interpretation:* Expected, as only churned customers would have a reason
- **Total Charges:** Initially stored as object type, preventing numerical analysis

Data Type Issue Resolution:

```
df['Total Charges'] = pd.to_numeric(df['Total Charges'], errors='coerce')
```

- Converted 'Total Charges' from object to numeric
- Revealed 11 additional null values after conversion

Missing Value Treatment:

```
df['Total Charges'].fillna(df['Total Charges'].median(), inplace=True)
```

- Imputed 11 missing Total Charges values with median ($1397.47)
- Used median instead of mean to avoid skewing from outliers
- Final null check confirmed: Total Charges = 0 nulls

Data Quality Validation:

- Verified all numeric columns properly formatted
- Confirmed no remaining nulls in analysis columns
- Dataset ready for exploratory analysis

Step 3: Exploratory Data Analysis - Tenure

Analysis Method:

```
df.groupby('Tenure Months')['Churn Value'].mean()
```

- Grouped customers by tenure months (0-72)
- Calculated mean churn rate for each tenure group
- Created line plot visualization

Findings:

Tenure Period	Churn Rate	Observation
0 months	0.0%	New signups (no churn yet)
1 month	61.99%	Highest churn point

Tenure Period	Churn Rate	Observation
2 months	51.68%	Still very high
3-4 months	~47%	Elevated churn continues
12+ months	Declining	Inverse relationship

Visualization: Created line plot showing churn rate declining as tenure increases, with steep drop-off after first year.

Key Insight: Customers with lower tenure show significantly higher churn rates, indicating that the risk of churn is highest during the initial months of subscription.

Step 4: Exploratory Data Analysis - Contract Type

Analysis Method:

df.groupby('Contract')['Churn Value'].mean().sort_values(ascending=False)

Findings:

Contract Type	Churn Rate	Magnitude
Month-to-month	42.71%	Baseline
One year	11.27%	3.8x lower
Two year	2.83%	15x lower

Visualization: Created bar plot comparing churn rates across contract types.

Key Insight: Month-to-month customers churn significantly more than customers on long-term contracts, suggesting low commitment and price sensitivity.

Step 5: Exploratory Data Analysis - Pricing

Analysis Method:

- Used box plot to compare Monthly Charges distribution between churned (Churn Value = 1) and retained (Churn Value = 0) customers
- Examined median, quartiles, and outliers for each group

Findings:

- Churned customers show higher median monthly charges
- Distribution of churned customers skews toward premium pricing tiers
- Price differential most pronounced in combination with other risk factors

Visualization: Box plot showing Monthly Charges on y-axis, Churn Value (0/1) on x-axis.

Key Insight: Customers with higher monthly charges are more likely to churn, especially when combined with short tenure and month-to-month contracts.

Step 6: Exploratory Data Analysis - Technical Support

Analysis Method:

```
df.groupby('Tech Support')['Churn Value'].mean().sort_values(ascending=False)
```

Findings:

Tech Support Status	Churn Rate	Relative Risk
No	41.64%	Baseline
Yes	15.17%	2.7x lower
No internet service	7.41%	5.6x lower

Key Insight: Customers without technical support exhibit higher churn, indicating that service quality and support availability are critical retention drivers.

Step 7: Risk Scoring Model Development

Model Design: Based on the four strongest churn predictors identified in exploratory analysis, created a composite risk score:

```
df['ChurnRiskScore'] = 0

df['ChurnRiskScore'] += (df['Tenure Months'] < 12)        # +1 point

df['ChurnRiskScore'] += (df['Contract'] == 'Month-to-month') # +1 point

df['ChurnRiskScore'] += (df['Monthly Charges'] > df['Monthly Charges'].median()) # +1 point

df['ChurnRiskScore'] += (df['Tech Support'] == 'No')       # +1 point
```

Risk Classification Logic:

```
def risk_category(score):

   if score >= 3:

      return 'High Risk'

   elif score == 2:

      return 'Medium Risk'

   else:

      return 'Low Risk'


df['ChurnRiskCategory'] = df['ChurnRiskScore'].apply(risk_category)
```

Scoring Rationale:

- Each factor contributes equally (1 point)
- Simple additive model for interpretability
- Threshold of 3+ points defines high risk (presence of 3-4 risk factors)
- Score range: 0-4 points

Step 8: Model Validation

Validation Method:

df.groupby('ChurnRiskCategory')['Churn Value'].mean()

Results:

Risk Category	Actual Churn Rate	Validation
High Risk	54.88%	2.1x overall rate
Medium Risk	22.34%	Close to overall rate
Low Risk	5.91%	4.5x lower than overall

Model Performance:

- Clear separation between risk categories
- High-risk group shows dramatically elevated churn
- Low-risk group shows substantially reduced churn
- Model successfully identifies distinct behavioral segments

Step 9: Financial Impact Calculation

Revenue at Risk Analysis:

revenue_risk = df[df['ChurnRiskCategory'] == 'High Risk']['Monthly Charges'].sum()

total_revenue = df['Monthly Charges'].sum()

revenue_risk_percentage = revenue_risk / total_revenue

Findings:

- **High-risk customer count:** 2,387 customers (33.9% of base)
- **Revenue at risk:** $184,270.90 monthly
- **Percentage of total revenue:** 40.40%
- **Annualized risk:** $2,211,250.80

Additional Metrics Calculated:

```
overall_churn_rate = df['Churn Value'].mean()  # 26.54%

avg_monthly_charge = df['Monthly Charges'].mean()  # $64.76
```

Step 10: Visualization of Revenue Distribution

Final Visualization:

```
df.groupby('ChurnRiskCategory')['Monthly Charges'].sum().plot(kind='bar')

plt.title('Revenue Distribution by Churn Risk Category')

plt.ylabel('Monthly Revenue')
```

Visual Insight: Bar chart showing revenue concentration across Low, Medium, and High risk segments, highlighting the disproportionate revenue exposure in the high-risk category.

6. Results

Quantitative Findings

Overall Churn Metrics

- **Total customer base:** 7,043 customers
- **Overall churn rate:** 26.54%
- **Average monthly charge:** $64.76
- **Total monthly revenue:** $456,119.35

Risk Model Performance

Risk Category	Customer Count	% of Base	Actual Churn Rate	Model Accuracy
High Risk	2,387	33.9%	54.88%	Strong predictor (2.1x baseline)
Medium Risk	2,164	30.7%	22.34%	Moderate (near baseline)
Low Risk	2,492	35.4%	5.91%	Strong predictor (0.22x baseline)

Model Validation: The risk scoring model successfully separates customer populations with dramatically different churn probabilities:

- High-risk customers are 9.3x more likely to churn than low-risk customers
- Clear gradation across all three risk categories
- Model demonstrates strong predictive validity using only four simple variables

Churn Drivers Quantified

1. Contract Type Impact:

Contract	Churn Rate	Relative to Month-to-Month
Month-to-month	42.71%	Baseline
One year	11.27%	73.6% reduction
Two year	2.83%	93.4% reduction

Key Finding: Two-year contracts reduce churn by 15-fold compared to month-to-month.

2. Tenure Impact:

Tenure Range	Churn Rate	Pattern
0-1 months	61.99%	Critical vulnerability period
2-6 months	47-52%	High risk continues
12+ months	<30%	Stabilization
24+ months	<20%	Strong retention

Key Finding: First-year customers are at highest risk; retention improves dramatically after 12 months.

3. Technical Support Impact:

Tech Support	Churn Rate	Risk Multiplier
No	41.64%	2.7x vs. Yes
Yes	15.17%	Baseline
No internet service	7.41%	0.5x vs. Yes

Key Finding: Customers without technical support are 2.7 times more likely to churn.

4. Pricing Impact:

- Churned customers show higher median monthly charges
- Price sensitivity amplified when combined with other risk factors
- Customers paying above-median charges contribute to elevated risk scores

Financial Impact Analysis

Revenue Exposure by Risk Category

Risk Category	Monthly Revenue	% of Total	Annual Exposure
High Risk	$184,270.90	40.40%	$2,211,250.80
Medium Risk	$148,437.55	32.54%	$1,781,250.60
Low Risk	$123,410.90	27.06%	$1,480,930.80

Critical Finding: Despite representing only 34% of the customer base, high-risk customers account for 40% of monthly revenue.

Expected Revenue Loss Calculation:

- High-risk customers: $184,271 × 54.88% churn rate = **$101,119 monthly expected loss**
- Annualized: **$1,213,428 in revenue at immediate risk**

Pattern Summary

The analysis identified four interconnected churn patterns:

Pattern 1: The "New Customer Cliff" Customer churn spikes dramatically in months 1-4, then gradually stabilizes. This suggests issues with onboarding, initial expectations, or early service experience.

Pattern 2: The "Commitment Paradox" Customers on month-to-month contracts churn at 15x the rate of two-year contract customers. Lack of commitment barrier allows easy exit, while commitment itself signals satisfaction.

Pattern 3: The "Price-Value Gap" Higher-paying customers churn more frequently, particularly those without supplementary services like technical support. This indicates perceived value does not match price paid.

Pattern 4: The "Support Safety Net" Technical support acts as a powerful retention mechanism, reducing churn by 63.6%. Support availability may signal service quality and provide problem resolution before churn occurs.

Visual Evidence

All findings were supported by clear visualizations created during analysis:

1. **Line plot:** Tenure vs. Churn Rate showing inverse relationship
2. **Bar chart:** Contract Type vs. Churn Rate highlighting 15x difference
3. **Box plot:** Monthly Charges comparison between churned/retained customers
4. **Bar chart:** Revenue Distribution across Risk Categories showing concentration