Customer Churn Analysis

The analysis follows a disciplined data science lifecycle: data acquisition, data quality management, exploratory analysis, statistical validation, feature engineering and selection, and dimensionality reduction using Principal Component Analysis (PCA). The purpose of this report is not only to explain *what* was done, but *why* each decision was made and *how* it supports business decision-making.

---

1. Business & Analytical Problem Definition

1.1 Business Context

In the telecommunications sector, customer churn poses a major financial risk due to high customer acquisition costs and relatively low switching barriers. Understanding churn behavior is essential for improving customer lifetime value, optimizing pricing strategies, and designing effective retention programs.

The dataset used in this project represents real-world telecom customers with varying contract structures, billing methods, service bundles, and tenure durations.

1.2 Problem Statement

The organization seeks to answer the following core question:

> **Which customer characteristics and behaviors most strongly influence churn, and are these effects statistically significant?**

1.3 Analytical Objectives

This project aims to: - Identify key churn drivers using **data-driven and statistically validated methods** - Quantify relationships between churn and customer attributes - Engineer business-meaningful features that improve explanatory power - Reduce feature dimensionality while preserving interpretability - Provide insights that can directly inform retention strategies

1.4 Target Variable Definition

- **Churn (Binary Target Variable)**
  - `1` → Customer has churned
  - `0` → Customer has been retained

---

2. Related Work & Literature Review

Previous studies in customer churn analytics consistently highlight the importance of tenure, pricing sensitivity, and contractual commitment. Research in telecom churn modeling emphasizes the value of combining: - Exploratory Data Analysis (EDA) - Statistical hypothesis testing - Feature engineering - Dimensionality reduction

Rather than relying exclusively on predictive models, this project adopts a **transparent and explainable analytical approach**, aligning with best practices in modern data science and business analytics.

---

3. Dataset Overview

3.1 Data Source

- **Dataset:** Telco Customer Churn Dataset
- **Nature:** Real-world, customer-level transactional and subscription data

3.2 Feature Composition

- **Numerical Features:** tenure, MonthlyCharges, TotalCharges
- **Categorical Features:** Contract, PaymentMethod, service subscriptions, billing preferences

Each observation represents a unique customer and their associated service attributes at the time of analysis.

---

4. End-to-End Analytical Workflow

Workflow Steps:

1. Data Loading and Initial Inspection

2. Data Wrangling and Quality Assurance

3. Exploratory Data Analysis (EDA)

4. Feature Engineering

5. Feature Selection

6. Probability Analysis & Hypothesis Testing

7. Dimensionality Reduction using PCA

8. Insight Consolidation and Business Interpretation

5. Data Wrangling & Data Quality Management

Data wrangling was treated as a critical analytical phase to ensure accuracy, consistency, and downstream reliability.

5.1 Data Loading and Structural Inspection

- Dataset imported using `pandas.read_csv()`
- Structural checks performed using `.head()`, `.tail()`, `.info()`, and `.shape()`
- Purpose: verify schema integrity, detect anomalies, and understand data scale

5.2 Data Type Standardization

- `TotalCharges` initially stored as string due to formatting issues
- Converted to numeric using `pd.to_numeric(errors='coerce')`
- Invalid entries deliberately converted to missing values for controlled handling

5.3 Missing Value Treatment (Business Logic)

- Missing values detected primarily in `TotalCharges`
- Rather than statistical imputation, a domain-based rule was applied:

  TotalCharges = tenure × MonthlyCharges

- This preserves financial consistency and reflects real billing behavior

5.4 Duplicate Detection and Removal

- Duplicate rows identified using `.duplicated()`
- Duplicates removed to avoid analytical distortion

5.5 Categorical Data Standardization

- All categorical values were normalized by:
  - Trimming whitespace
  - Converting text to lowercase

This step prevents artificial category inflation during encoding.

5.6 Outlier Detection and Treatment

- Outliers detected using the **Interquartile Range (IQR)** method
- Outliers were **capped**, not removed:
  - Preserves extreme but valid customer behavior
  - Stabilizes scaling and statistical tests

5.7 Final Dataset Preparation

- Removed `customerID` (identifier only)

- Converted `SeniorCitizen` to categorical
- Encoded `Churn` as binary
- Reset index to ensure clean structure

5.8 Data Quality Validation

- No remaining missing values
- No duplicate records
- Clean dataset exported as `cleaned_telco_churn.csv`

---

6. Exploratory Data Analysis (EDA)

EDA was conducted to uncover behavioral patterns, detect churn risk signals, and guide feature engineering.

6.1 Univariate Analysis

- Tenure distribution reveals strong early-life churn concentration
- MonthlyCharges show right-skewed distribution with higher churn exposure at upper ranges

6.2 Bivariate & Multivariate Analysis

- Month-to-month contracts exhibit significantly higher churn
- Higher MonthlyCharges combined with short tenure increase churn likelihood
- Payment method and billing preferences show measurable churn differences

6.3 Key EDA Insights

- The first 6–12 months represent the highest churn risk window
- Pricing sensitivity plays a major role in churn decisions
- Contract length is a strong proxy for customer commitment

These findings directly motivated the subsequent feature engineering phase.

---

7. Feature Engineering & Feature Selection

7.1 Feature Engineering Strategy

Features were engineered to represent latent customer behaviors:

- **TenureGroup:** Customer lifecycle segmentation
- **AvgMonthlySpend:** Normalized spending intensity
- **IsLongContract:** Long-term commitment indicator

- **PaperlessAndElectronic:** Digital engagement behavior
- **HasTechSupportOrSecurity:** Dependency on support services
- **ServiceCount:** Service bundle depth

Each engineered feature has a clear business interpretation and modeling purpose.

7.2 Train-Test Split and Leakage Prevention

- Data split performed prior to encoding and scaling
- Stratified sampling ensured churn class balance
- Prevented information leakage

7.3 Preprocessing Pipelines

Two pipelines were implemented to support different analytical needs:

- **StandardScaler Pipeline:** Logistic Regression, RFE, L1 regularization
- **MinMaxScaler Pipeline:** Chi-Square feature selection (non-negative constraint)

Both pipelines include imputation and one-hot encoding.

7.4 Feature Selection Techniques

Filter Methods: - Correlation analysis (numeric features) - Chi-Square test (categorical features)

Wrapper Method: - Recursive Feature Elimination (RFE) with Logistic Regression

Embedded Method: - L1-regularized Logistic Regression

7.5 Final Feature Selection Logic

- Features selected by **multiple independent methods** were prioritized
- This consensus-based approach increases robustness and interpretability

---

8. Probability Analysis & Hypothesis Testing

8.1 Probability Modeling

- MonthlyCharges modeled using a normal distribution
- Probability of charges exceeding defined thresholds calculated

8.2 Statistical Hypothesis Tests

Test	Relationship Tested	Result
t-test	MonthlyCharges vs Churn	Statistically Significant
Chi-Square	PaymentMethod vs Churn	Statistically Significant

Test	Relationship Tested	Result
ANOVA	Contract vs MonthlyCharges	Statistically Significant
Correlation	Tenure vs Churn	Statistically Significant

All null hypotheses were rejected at α = 0.05.

9. Dimensionality Reduction using PCA

9.1 PCA Preparation

- All features encoded and standardized
- PCA applied to full feature matrix

9.2 Explained Variance Analysis

- Majority of variance captured by first few components
- Indicates high feature redundancy

9.3 Principal Component Interpretation

- **PC1:** Usage intensity and financial exposure
- **PC2:** Contractual commitment and billing behavior

9.4 PCA Validation Against EDA

PCA projections confirm EDA conclusions, reinforcing confidence in identified churn drivers.

10. Business Conclusions & Strategic Implications

Key Churn Drivers

- Short tenure (early lifecycle customers)
- High monthly charges
- Month-to-month contracts
- Limited service engagement

Strategic Recommendations

- Strengthen onboarding and early engagement programs
- Promote long-term contracts through incentives
- Introduce pricing personalization for high-risk segments
- Increase service bundling to raise switching costs

11. References

1. Verbeke et al. – Predictive Modeling for Customer Churn
2. IBM Telco Customer Churn Dataset
3. Hastie, Tibshirani, Friedman – The Elements of Statistical Learning
4. Scikit-learn Official Documentation
5. Imani, M., Joudaki, M., Beikmohammadi, A., & Arabnia, H. R. (2025). *Customer churn prediction: A systematic review of recent advances, trends, and challenges in machine learning and deep learning*. *Machine Learning and Knowledge Extraction, 7*(3), 105. https://doi.org/10.3390/make7030105 MDPI
6. Liu, S. (2025). *Literature Review on Customer Churn Prediction in Telecom Industry*. *Theoretical and Natural Science, 132*, 27–32. https://doi.org/10.54254/2753-8818/2025.DL27120 tns.ewapub.com
7. Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). *Customer churn prediction in telecom using machine learning in a big data platform*. *Journal of Big Data, 6*(28). https://doi.org/10.1186/s40537-019-0191-6 SpringerLink