ST635 Final Project Report

Vinit Late & Sahil Bora

Abstract

This project paper goes deep into analyzing customer churn in a telecommunications dataset. Customer churn is an essential metric for subscription-based businesses, representing the percent of customers stopping service usage over a certain period. With the help of advanced statistical modeling and EDA, the findings from this study highlight main predictors of churn, evaluating customer segmentation and proposing actionable ways of retaining customers. Key findings show month-to-month contracts and fiber optic internet are significant contributors to churn, while tenure is beneficial to retaining customers. This study applies PCA and K-Means Clustering to uncover latent structure and find different customer clusters with targeted business strategies.

Introduction

Customer retention is a cornerstone of sustainability for subscription-based businesses; hence, churn directly impacts revenue and growth. The drivers of customer churn are essential to comprehend in developing effective retention strategies. The current study makes use of a robust telecommunications dataset including 7,043 observations and 21 variables on customer demographics, service usage, account information, and churn status.

The project objectives are two fold:

1. To identify and quantify the key drivers of customer churn using statistical and machine learning techniques.
2. To segment customers into actionable groups for precision marketing and retention efforts.

Through the application of Multiple Linear Regression (MLR), Logistic Regression, PCA, and K-Means Clustering, this analysis provides a comprehensive framework to address churn challenges. The methodologies ensure robust interpretability and actionable insights, positioning this study as a valuable contribution to customer retention analytics.

Methodology

Data Preparation

To ensure data quality and consistency, the following preprocessing steps were performed:

- Rows with missing TotalCharges values (11 observations) were removed to avoid bias.
- Categorical variables were standardized. Responses indicating "No internet service" were consolidated into "No" for clarity.
- Numerical variables (tenure, MonthlyCharges, and TotalCharges) were normalized to account for scale differences, facilitating compatibility with Principal Component Analysis (PCA).

These steps ensured that the dataset was suitable for statistical modeling applications.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) uncovered critical trends and relationships in the data:

- Churn Distribution: 26.6% of customers churned, while 73.4% retained their services.

- Tenure: Bimodal distribution revealed clusters of short- and long-tenured customers.
- Monthly Charges: Most customers were concentrated below $75, indicating a skew toward lower spending.
- Internet Service: Fiber optic users exhibited disproportionately higher churn rates.
- Contract Type: Month-to-month contracts had significantly higher churn rates compared to annual or biennial contracts.

EDA provided an empirical basis for model selection and hypothesis generation, enabling a focused approach to understanding churn.

Statistical Models

Multiple Linear Regression (MLR)

Business Context: The goal was to understand the factors driving MonthlyCharges to optimize pricing strategies and identify high-revenue customers MLR revealed that tenure, InternetService type, and contract type significantly influence charges, with an Adjusted R-squared of 87.2%, ensuring strong predictive accuracy. This provides actionable insights for pricing adjustments and targeting high-value customers.

- Adjusted R-squared = 87.2%, indicating strong predictive accuracy.
- Residual diagnostics confirmed normality and homoscedasticity.
- VIF values were all below 5, confirming minimal multicollinearity.

Logistic Regression

Business Context: We aimed to predict customer churn probability, enabling targeted retention strategies for at-risk customers. This model highlighted tenure, contract type, and InternetService as key churn predictors. Month-to-month contracts and Fiber optic services were strongly associated with higher churn, emphasizing the need for loyalty programs and service improvements.

- AUC = 0.841, indicating excellent model performance.
- Odds ratios revealed significant predictors, such as:
  - Tenure: Longer tenure decreases churn likelihood.
  - InternetService: Fiber optic users are significantly more likely to churn.
  - Contract: Month-to-month contracts significantly increase churn likelihood.

Principal Component Analysis (PCA)

Business Context: PCA was used to simplify the dataset by identifying the most impactful variables driving customer behavior while preserving key information.PCA reduced dimensionality to two principal components, explaining 95% of the variance. Tenure and TotalCharges emerged as the primary contributors, enabling better customer segmentation and visualization for decision-making.

- PC1 explained 72.7% of variance.
- Cumulative variance of 95% was achieved by the second component.

K-Means Clustering

The focus was on grouping customers into actionable segments for tailored marketing and retention strategies.

K-Means segmented customers into three clusters:

- Cluster 1 (High Spenders): Long tenure and high charges.
- Cluster 2 (Discount Seekers): Long tenure and low charges.
- Cluster 3 (New Customers): Short tenure and moderate charges.

Clusters were validated by intra-cluster variance and visualization.

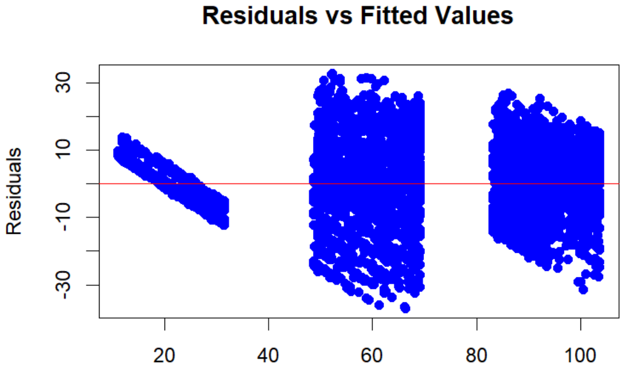
Results

Multiple Linear Regression

The MLR model highlighted:

- InternetService: Fiber optic users incurred $34.5 higher charges compared to DSL users.
- Contract Type: Annual contracts marginally increased monthly charges.

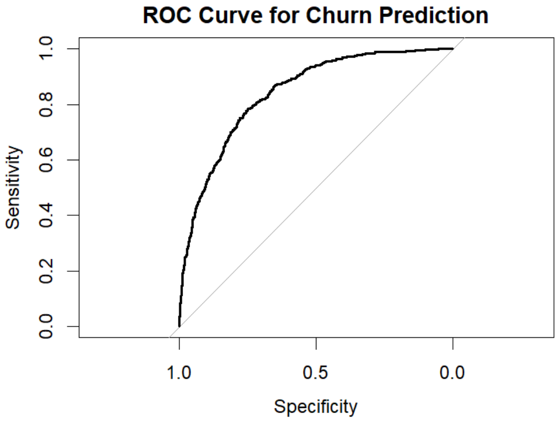
Residual diagnostics confirmed model validity, with minimal violations of linearity or normality assumptions.



Logistic Regression

The logistic model pinpointed key churn predictors:

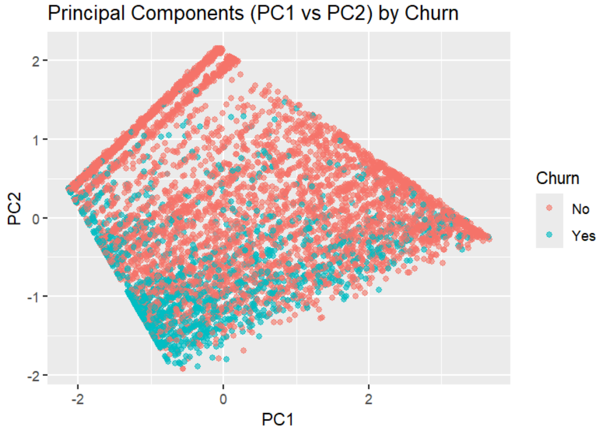
- Tenure: Reduced churn likelihood by 3.2% per additional month.
- InternetService: Fiber optic users showed a significantly higher churn probability.
- Contract Type: Month-to-month contracts had an odds ratio of 2.63, underlining their churn-driving effect.



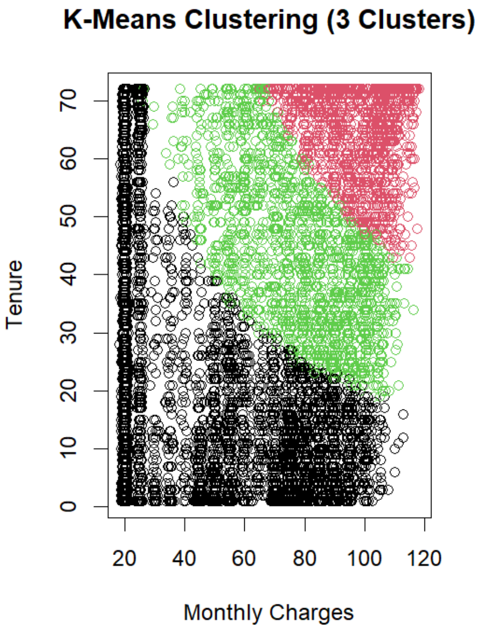
Principal Component Analysis (PCA)

Dimensionality reduction identified latent structures:

- The first principal component correlated strongly with tenure and TotalCharges.
- Variance explained reached 98% by the second component.



N K-Means Clustering



Clustering analysis produced actionable insights:

- Cluster 1 (High Spenders): These customers are characterized by long tenure and high monthly charges, contributing significantly to the company's revenue. However, this group exhibited the highest churn rate at 32.13%, indicating potential dissatisfaction despite their high spending.
- Cluster 2 (Discount Seekers): This segment consists of cost-conscious customers with long tenure and low monthly charges. With a churn rate of only 13.79%, they represent the most stable group, likely due to their preference for economical service options.
- Cluster 3 (New Customers): Comprised of customers with short tenure and moderate charges, this group exhibited a churn rate of 22.24%. Their behavior reflects an early-stage relationship with the service, making them susceptible to churn due to unmet expectations or initial dissatisfaction.

Summary

This project identified significant predictors of churn, including contract type and internet service, and highlighted actionable customer segments. Month-to-month contracts and fiber optic internet emerged as primary churn drivers, while tenure demonstrated a protective effect against churn. The clustering analysis segmented customers into High Spenders, Discount Seekers, and New Customers, offering tailored retention strategies for each group.

Improvements and Future Scope

- **Enhanced Data Collection and Variable Refinements**: Collect additional attributes like satisfaction scores and break down TotalCharges into granular metrics for deeper analysis.
- **Refined Clustering and Model Fine-Tuning**: Evaluate merging clusters and optimize logistic regression models to enhance specificity and sensitivity.
- **Real-Time Analytics and Longitudinal Studies**: Build real-time monitoring systems and study customer transitions over time to improve lifecycle management.
- **Holistic Retention Strategies**: Integrate customer feedback and develop loyalty programs tailored to specific customer segments.

In conclusion, the analyses conducted—spanning regression models, PCA, and clustering—provide actionable insights into customer behavior and churn dynamics. By identifying key drivers of MonthlyCharges and churn, simplifying data dimensions, and segmenting customers into meaningful clusters, the findings empower data-driven strategies for customer retention, engagement, and revenue optimization. These insights lay the foundation for targeted marketing, improved service offerings, and enhanced customer satisfaction, ensuring sustainable business growth.