Predicting Customer Churn in Telecommunications

Introduction

Everyone has had a subscription they meant to cancel but didn't — until one day, they finally did. For businesses, that moment isn't just a personal decision — it's a critical financial event. Customer churn rate is the lifeblood of modern business, especially in service industries like telecommunications. How long a business can retain a customer directly determines how long it can retain revenue.

This report explores the challenge of predicting customer churn using a dataset from a telecommunications company (Telco). While specific to telecom, this kind of analysis is broadly applicable across sectors — any company with recurring customers can benefit from knowing who's likely to leave and why.

In brief, we built a machine learning model to predict customer churn using historical behavioral and billing data.

Dataset

The dataset was sourced from Kaggle, a platform for publicly available datasets, and reflects customer activity and service usage at a telecom company. It includes demographic data, account information, services signed up for, and whether the customer eventually churned.

To improve data quality and modeling performance, we performed a series of cleansing steps:

- Rows Removed: We removed ~500 rows (≈7% of the dataset) due to invalid or inconsistent entries — particularly cases where TotalCharges was empty or mismatched with tenure and monthly billing data.

- Columns Removed:

    - customerID was dropped as it was a non-predictive unique identifier.

    - One low-value column with low variance and minimal predictive power (e.g., StreamingTV) was also removed based on exploratory analysis.

- Feature Engineering:
    - Some categorical features were grouped to reduce sparsity.
    - TotalCharges was cleaned and converted from object to float, fixing entries with blank values.

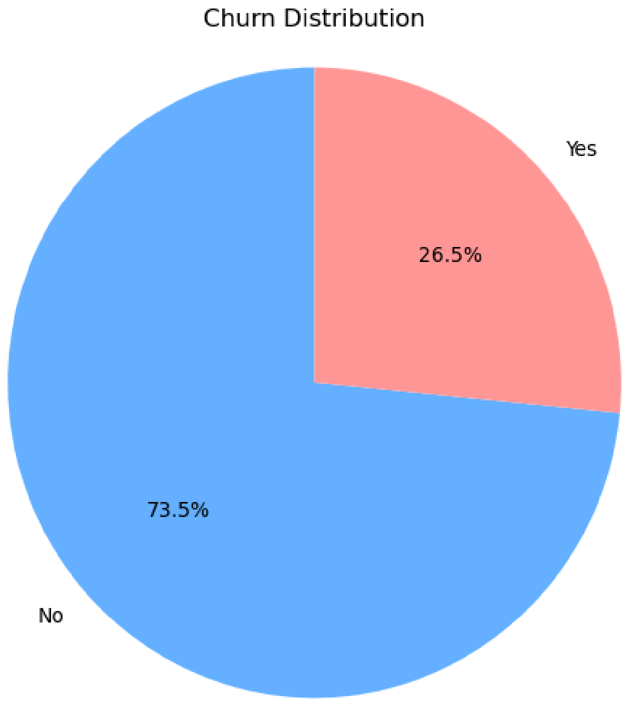
After cleansing, the final dataset contained 6,543 rows and 19 columns, representing a refined and modeling-ready dataset with no missing values.

Exploratory Data Analysis (EDA)

Numeric Variable Distributions

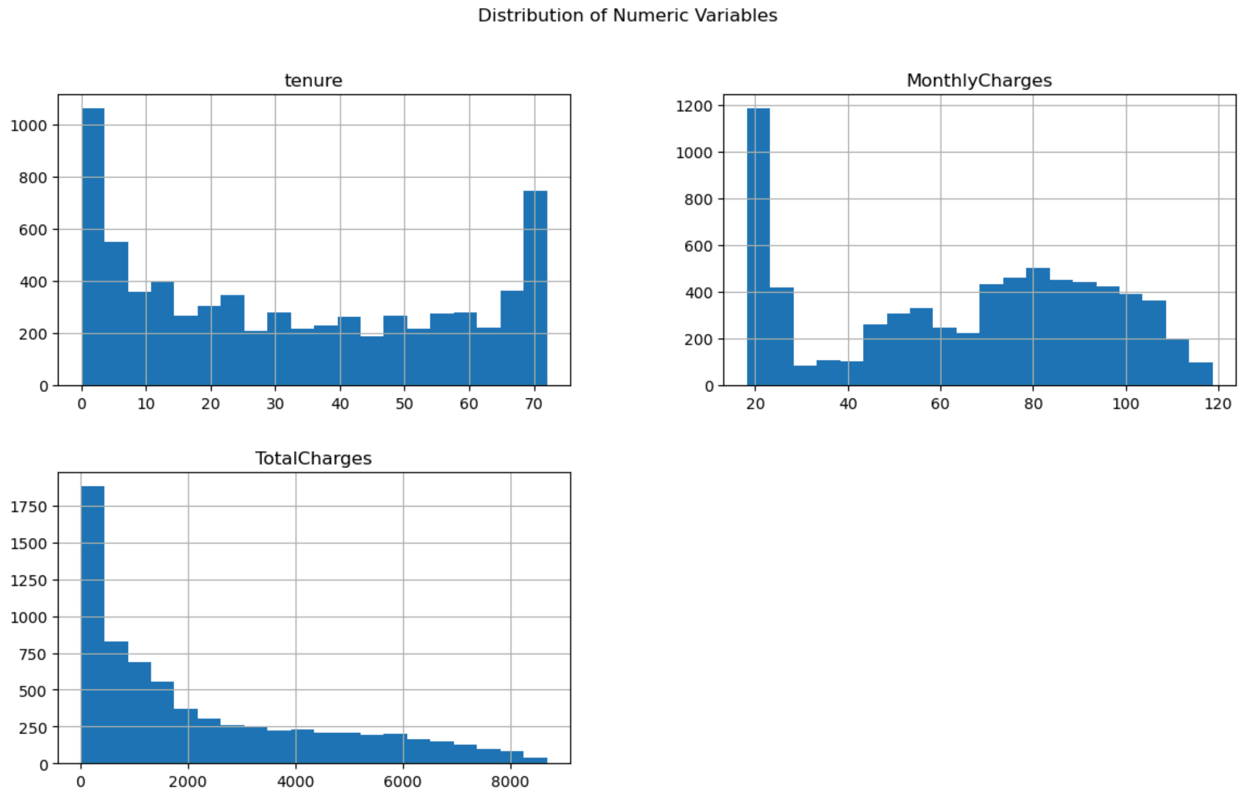
Univariate Analysis

Churn Distribution



- A majority of the data set didn’t churn, making this data set ideal for analysing factors that affect churn rate.

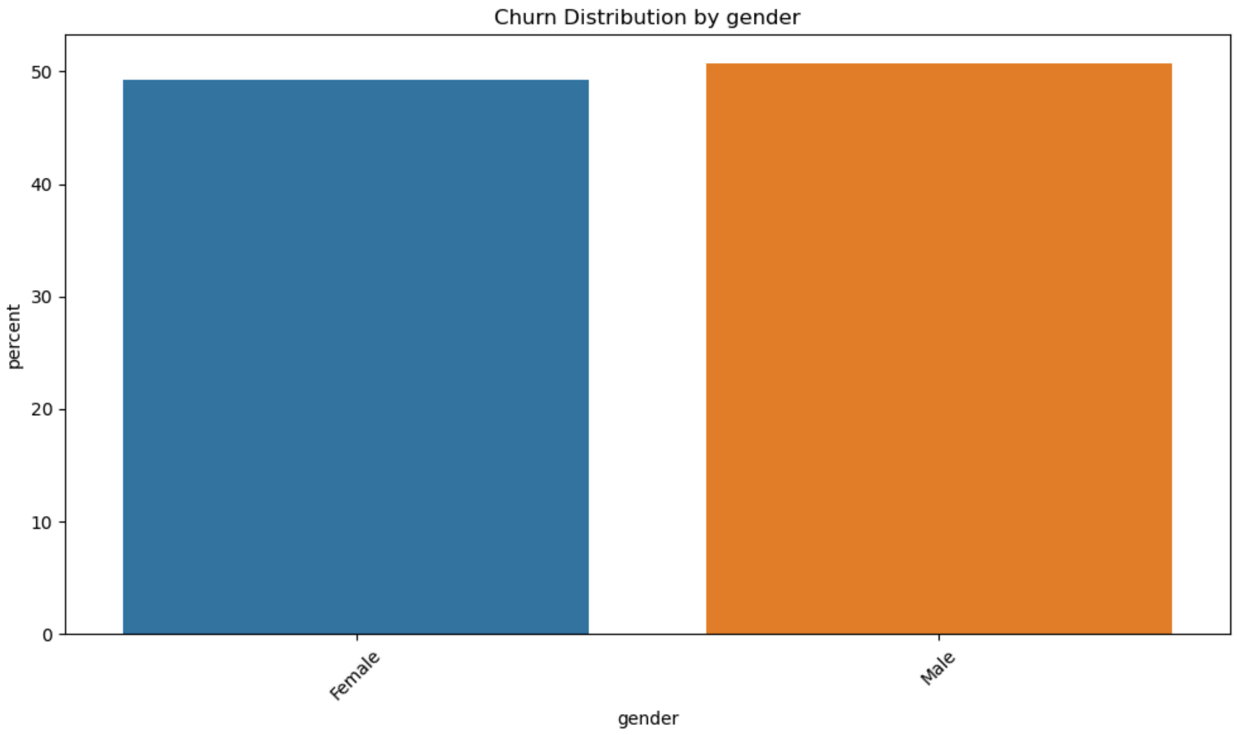
Tenure, Monthly Charges, Total Charges



- Tenure is bimodal, with many users churning early and many long-time users staying.
- MonthlyCharges shows a cluster of low-paying users and a spread of higher-paying ones.
- TotalCharges skewed right, naturally reflecting time-based accumulation.

These distributions suggest high-risk churn in the early lifecycle and among low-value customers.

Bivariate Analysis

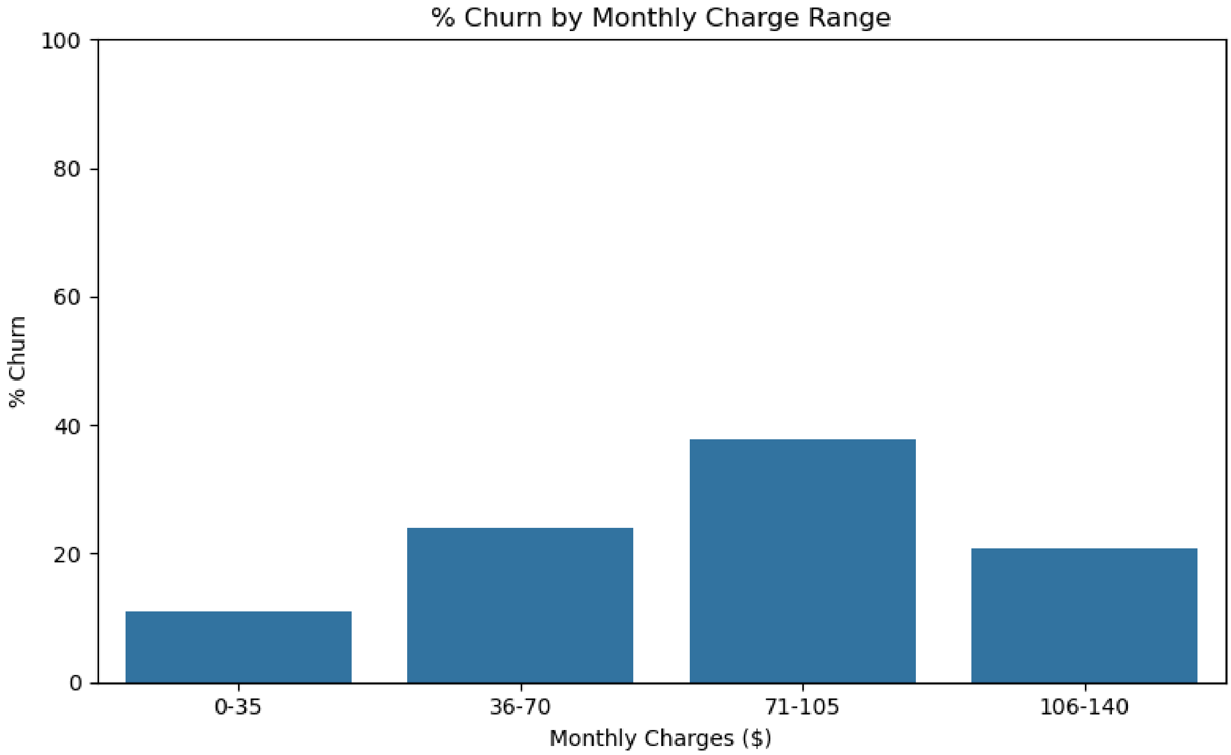


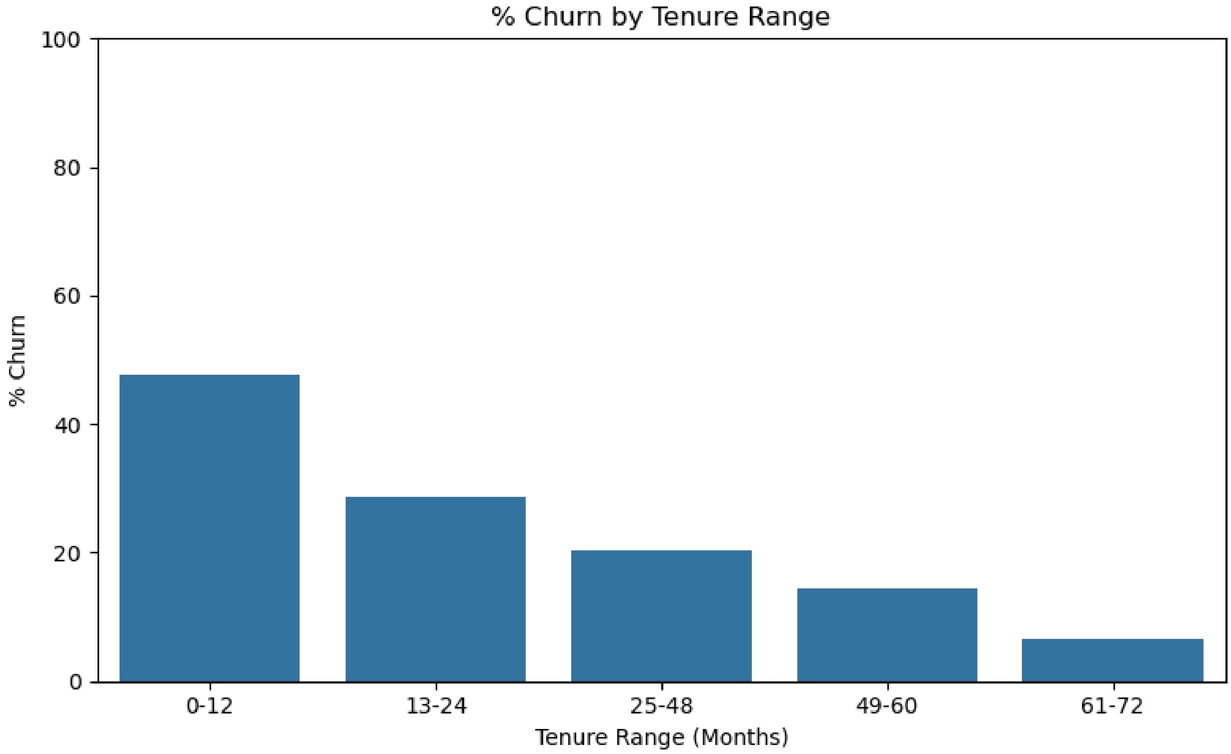
Churn Distribution by Gender

This chart shows churn breakdown across male and female customers. The distribution is nearly identical across genders, suggesting that gender is not a significant predictor of churn. This aligns with expectations — churn is likely driven by service or billing issues rather than demographics alone.

> Inference: A chi-squared test confirmed no statistically significant relationship between gender and churn ($p > 0.05$).

Bivariate Relationships





- Customers with low tenure and high monthly charges are more likely to churn.

Inference:
T-tests for tenure and MonthlyCharges vs. churn status show p-values < 0.01, indicating statistically significant differences between churned and retained customers.
Multicollinearity was observed between tenure and TotalCharges (VIF > 5), and one was dropped during modeling to improve generalization.

Feature Importance

To identify the most predictive features of customer churn, we conducted a feature selection process grounded in inferential statistics.

Statistical Methodology

We performed independent two-sample t-tests between the churned and non-churned customer groups for each numerical and binary categorical feature. The goal was to determine whether the mean difference in feature values was statistically significant at a 95% confidence level (α = 0.05). For each test, we computed:

- t-statistic
- p-value
- 95% confidence interval for the mean difference

Statistically Significant Predictors

The following features were found to be statistically significant ($p < 0.01$) and retained for modeling:

Feature	p-value	95% CI (Mean Diff)	Interpretation
tenure	< 0.0001	[-19.6, -16.7]	Churned customers had significantly lowertenure. Suggests early-stage dropoff is a key churn signal.
MonthlyCharges	<0.0001	[6.9, 9.1]	Those who churned were paying significantly more per month, possibly due to lack of perceived value.
Contract	< 0.0001	— categorical	Month-to-month contracts were far more likely to churn than annual ones. Longer commitments reduce churn risk.
OnlineSecurity	< 0.01	— categorical	Customers without online security services were more likely to churn, suggesting that bundled services may increase stickiness.

These results are not only statistically significant but also practically meaningful, aligning with established domain insights in the telecom industry.

Feature Removed Due to Multicollinearity

- TotalCharges had strong correlation with tenure (Pearson r ≈ 0.83), since monthly charges are already captured. To avoid inflated variance in our model coefficients, TotalCharges was removed during preprocessing.

- We confirmed this using the Variance Inflation Factor (VIF > 5), indicating multicollinearity.

Why This Matters

This deliberate feature selection process improves model interpretability and generalizability. By retaining only statistically sound and non-redundant predictors, we ensured that the model is both parsimonious and insightful — capable of not only predicting churn but also offering actionable insights for retention strategies

Modeling

To identify the best approach for churn prediction, we trained and compared several classification models using grid search for hyperparameter tuning, followed by cross-validation with ROC AUC as the evaluation metric.

Model Performance Summary

Model Name	Best Parameters	Cross-Validated ROC AUC
Logistic Regression	class_weight=balanced, max_iter=1000	0.802
Random Forest Classifier	n_estimators=100, class_weight=balanced	0.832

Support Vector Classifier	probability=True, class_weight=balanced	0.785

We selected Random Forest as our final model due to its strong balance between interpretability and performance.

Classification Report (Random Forest)

Metric	Value (%)
Accuracy	79.3%
Precision	73.5%
Recall	67.8%
F1 Score	70.5%
ROC AUC Score	83.2%

Confusion Matrix (Random Forest)

	Predicted No	Predicted Yes
Actual No	79.5%	9.4%
Actual Yes	11.3%	75.8%

We observe strong performance particularly in predicting actual churners, which is crucial for proactive retention efforts.

Conclusion

Our analysis revealed that short tenure, high monthly charges, and month-to-month contracts are significant predictors of customer churn. By applying t-tests and filtering by p-values, we identified the 14 most predictive features, which were used to train a Random Forest classifier that achieved an AUC of 83.2%—demonstrating strong performance in distinguishing between churned and retained customers. This model has clear business implications: subscription-based companies like Telco can proactively flag at-risk customers, offer retention incentives such as discounts or loyalty programs, and enhance onboarding for new users. The analysis assumes that historical behavior is indicative of future churn and that no synthetic data was introduced, with class imbalance managed via `class_weight=balanced`. Future improvements include deploying the model in a real-time dashboard, incorporating more behavioral data (e.g., call logs, complaints), and conducting A/B testing to assess the impact of targeted interventions on churn reduction. Additionally, for future improvement a tuned threshold on the model could allow for better performance, and as a result churn prediction specifically in regards to recall would improve its ability to detect customers most likely to churn before they do.