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**:

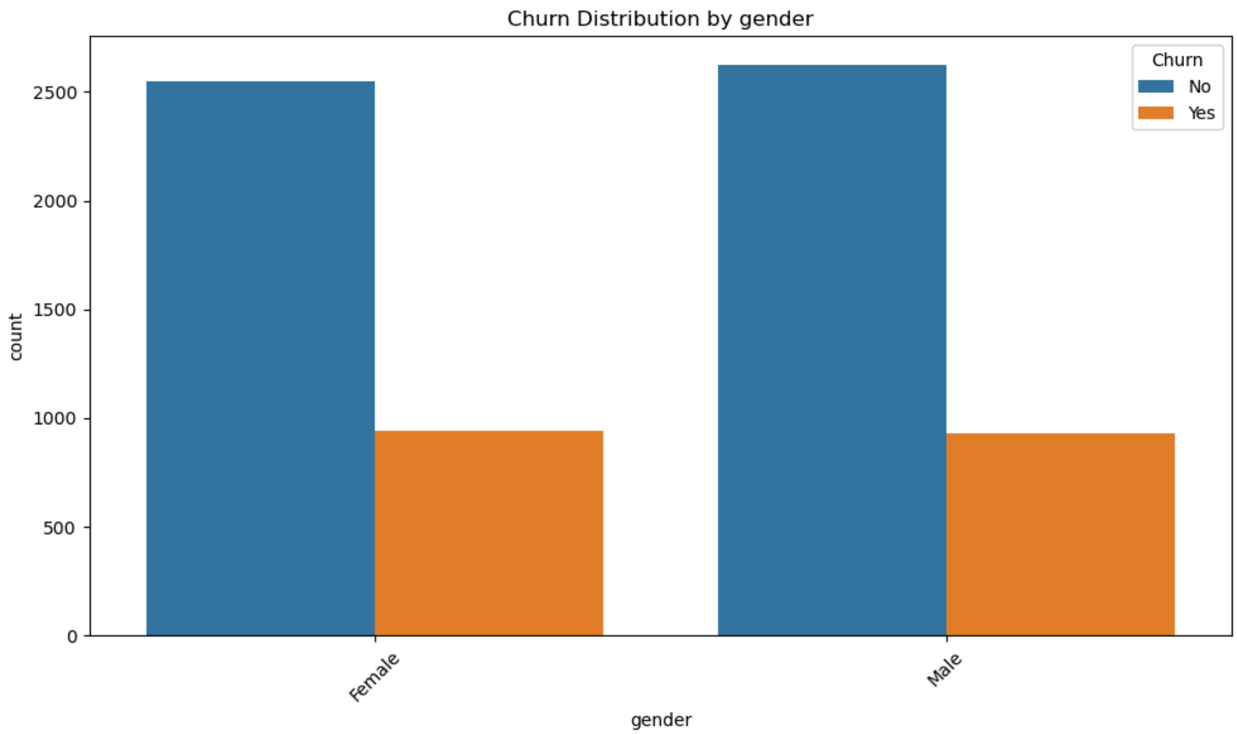
    - Some categorical features were grouped to reduce sparsity.

    - `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)

Univariate Analysis



Churn Distribution by Gender

Chart Title: Churn Distribution by Gender
X-axis: Gender
Y-axis: Number of Customers

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 ($p > 0.05$).

Numeric Variable Distributions

Tenure, Monthly Charges, Total Charges

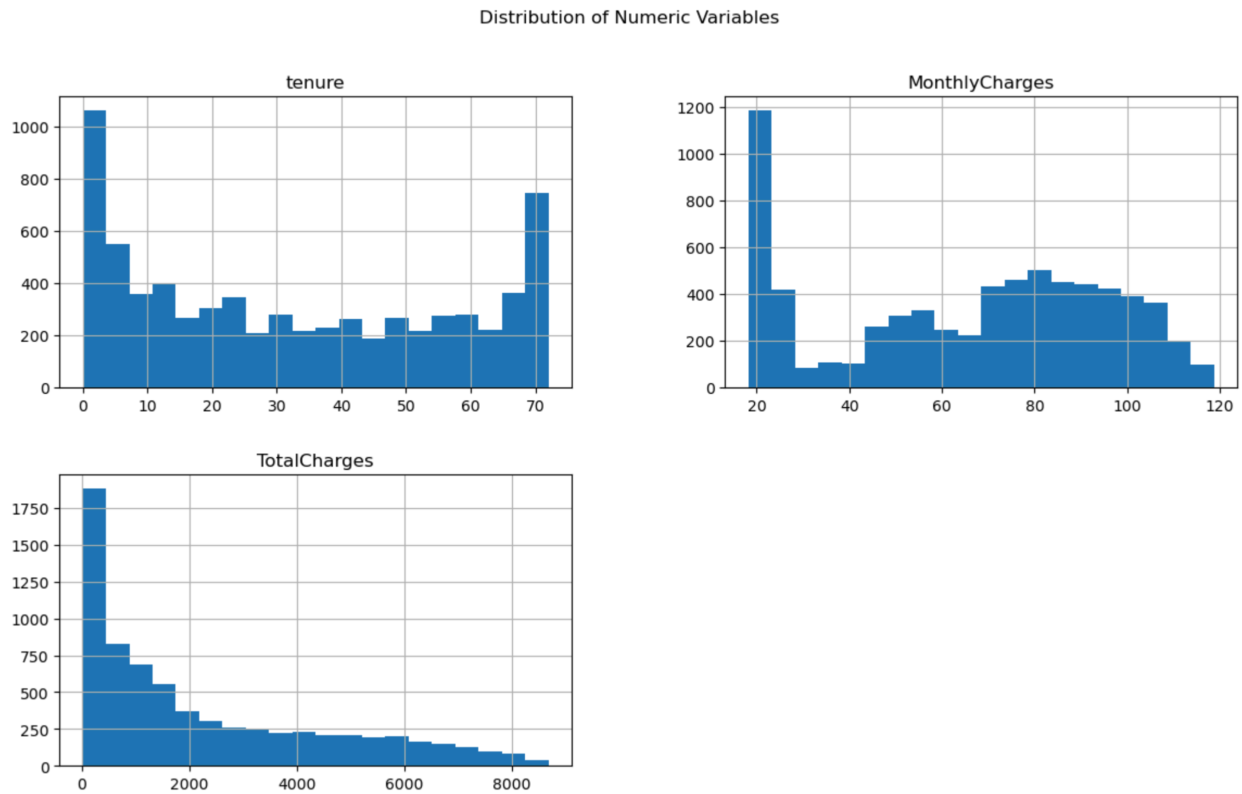


Chart Title: Distribution of Numeric Variables
X-axis: Respective values (e.g., months, dollars)
Y-axis: Frequency

- **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** skews right, naturally reflecting time-based accumulation.

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

Bivariate Relationships

Pairwise Plot of Tenure, MonthlyCharges, TotalCharges vs. Churn

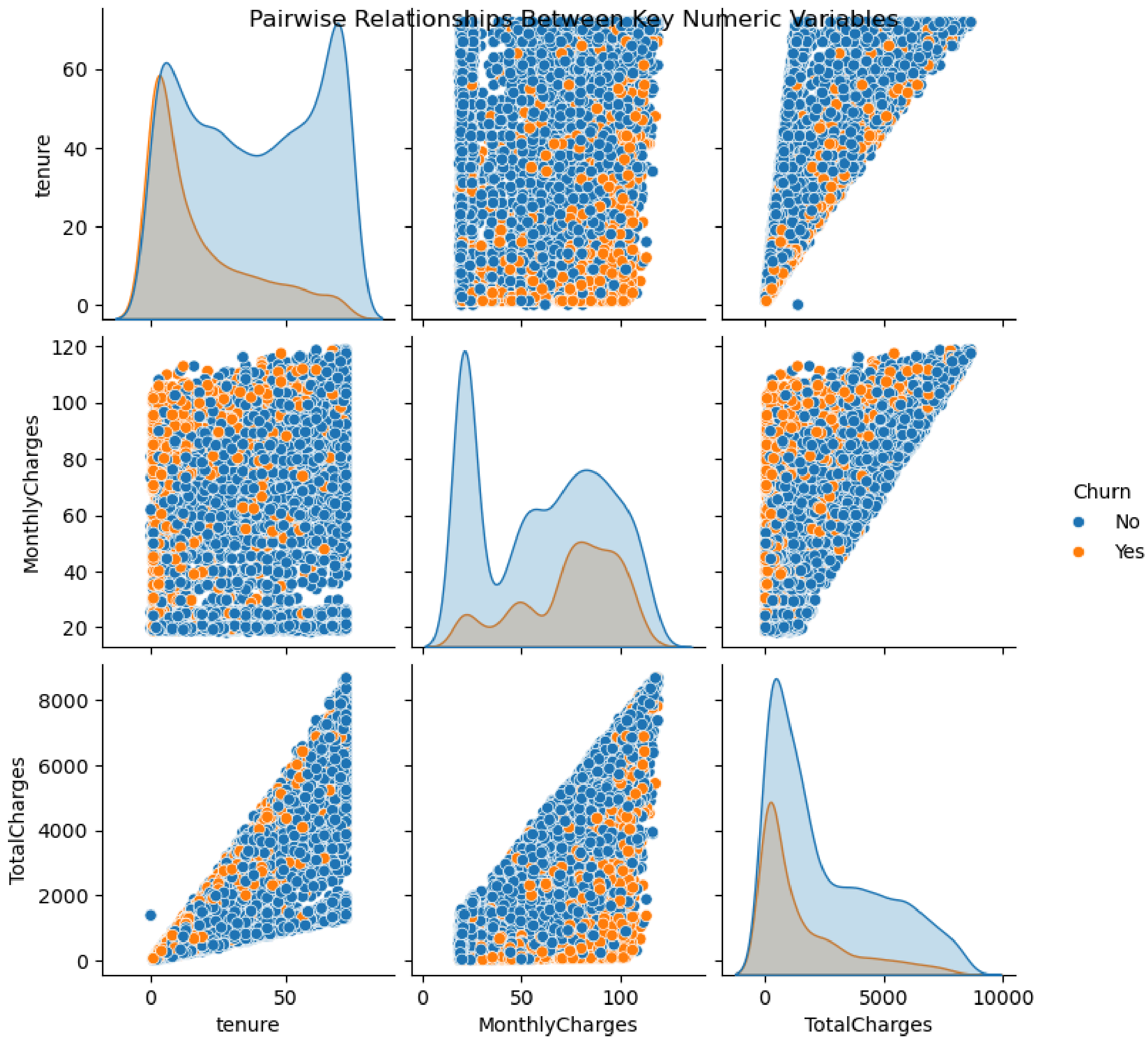


Chart Title: Pairwise Relationships Between Key Numeric Variables
Axes: Feature values (e.g., x: tenure, y: total charges)

This matrix highlights that:

- Customers with **low tenure** and **high monthly charges** are more likely to churn.
- Total charges and tenure are highly correlated (unsurprising, as total = monthly × time).

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** ($\alpha = 0.05$). 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 lower tenure. 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).
- We confirmed this using the **Variance Inflation Factor (VIF > 5)**, indicating multicollinearity.
- To avoid inflated variance in our model coefficients, `TotalCharges` was removed during preprocessing.

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	Optimal Feature Set	Cross-Validated ROC AUC
Logistic Regression	`class_weight=balanced`, `max_iter=1000`	14 features after t-test	0.802
Random Forest Classifier	`n_estimators=100`, `class_weight=balanced`	14 features after t-test	0.832
Support Vector Classifier	`probability=True`, `class_weight=balanced`	14 features after t-test	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

Key Findings:

- **Short tenure, high monthly charges, and month-to-month contracts are statistically significant indicators of churn.**
- **Using t-tests and p-values, we filtered down to the 14 most predictive features.**
- **The final model, a Random Forest classifier, achieved an AUC of 83.2%, indicating strong discriminative ability.**

Business Application:

As mentioned in the introduction, customer churn is directly tied to revenue retention. With this model, Telco (or any subscription-based service) can:

- **Flag at-risk customers early**
- **Offer discounts or loyalty programs to retain high-value individuals**
- **Optimize onboarding experiences for new users**

Next Steps & Assumptions:

Assumptions:

- **We assume that past behavior is predictive of future churn.**
- **No synthetic data was added, and class balance was handled through `class_weight=balanced`.**

Future Work:

- **Deploy this model into a real-time dashboard for live churn risk scoring.**
- **Integrate more behavioral and usage data (e.g., call logs, complaint records).**
- **Run A/B tests to evaluate interventions on customers flagged as high-risk.**