Telecom Customer Churn Prediction — End-to-End ML & BI Project

Project Overview

Customer churn remains one of the most critical challenges for telecom service providers. Retaining an existing subscriber is significantly more cost-effective than acquiring a new one. This project focuses on identifying customers who are likely to churn by combining **business intelligence (Power BI)** and **machine learning (Random Forest)** to help organizations proactively act on churn risks.

The goal:

1. Provide business decision-makers with clear insights on why customers churn.
2. Build a predictive model that identifies customers at risk, enabling timely retention actions.

Dataset Summary

The dataset consists of customer-level telecom information, including demographics, service subscriptions, billing patterns, and churn status.

Size: **6007 rows, 32 features**

Key data categories include:

- Personal info (Age, Gender, Marital status, State)
- Service subscriptions (Internet type, Security add-ons, Streaming services)
- Contract/billing (Payment method, Contract term, Paperless billing)
- Usage & revenue (Monthly charge, Extra data charges, Long-distance charges, Total revenue)
- Target variable: Customer_Status (Stayed / Churned)

Business Understanding

Telecom companies lose revenue when customers cancel services. To reduce churn, they need to predict:

- ✓ *Which customers are likely to leave?*
- ✓ *What patterns contribute to churn?*
- ✓ *Which customer segments should retention campaigns focus on?*

This project supports those outcomes by:

- Conducting exploratory analysis to understand churn drivers.
- Developing a machine learning model to flag customers at risk.

Exploratory Data Analysis (EDA)

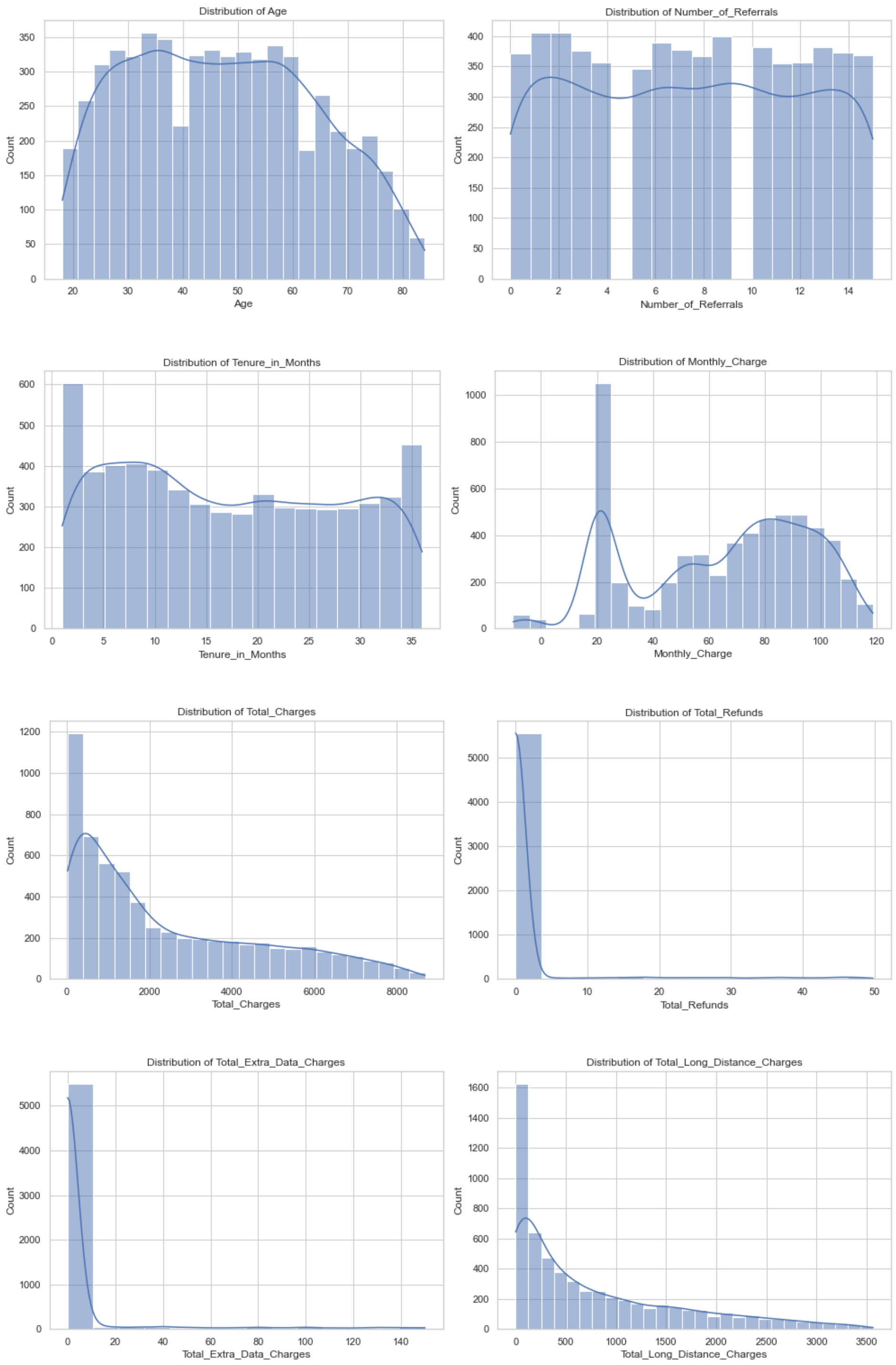
EDA helps ensure business and ML value by:

- Understanding characteristics & distribution of variables (univariate analysis)
- Identifying relationships between churn and predictor variables (bivariate analysis)
- Detecting skewness, imbalance, and outliers affecting modelling.
- Finding features with strong influence on churn behaviour.

Although the project already includes a Power BI dashboard, EDA remains important because:

- Power BI is visual storytelling for business users.
- EDA is statistical validation for ML modelling.
- EDA ensures the model is trained on well-understood, properly prepared data.

Univariate Analysis:



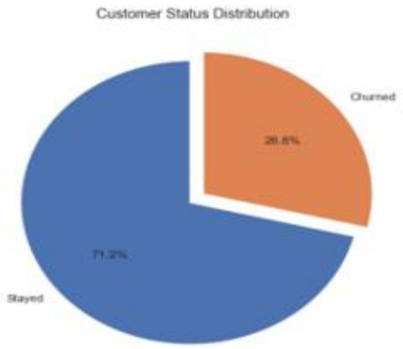
Key Observations

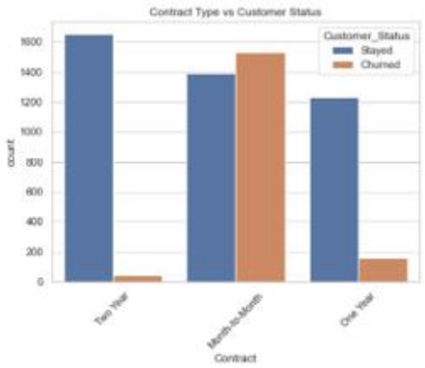
- **Age and Tenure** are well spread: diverse customer base with no extreme clustering.
- **Billing-related features (Monthly/Total Charges)** are right-skewed: few customers generate the highest revenue.
- **Usage-based features (Extra Data & Long-Distance Charges)** show high skewness: only a minority of users consume heavy add-ons.
- **Low referral count overall**: **Number of Referrals** shows most customers are single-referral users, suggesting limited peer-driven acquisitions i.e. Limited organic acquisition.
- **Total refunds are concentrated at the lower end**, suggesting that refund requests are rare, but the small number of high-refund cases may hint toward dissatisfaction among specific segments.

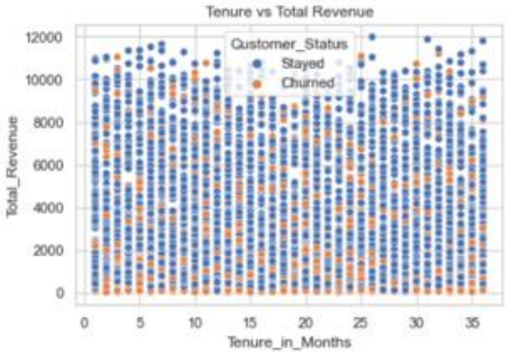
About Outliers:

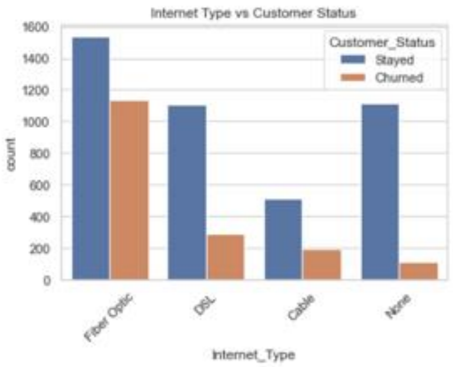
Most variables — particularly Total Charges, Monthly Charges, Add-on Usage, and Refunds — exhibit outliers due to a very small group of heavy-spending or refund-claim customers. These outliers are not errors but genuine behavioral variations. Removing them would erase meaningful customer segments such as *high-value users* or *high-risk refund seekers*.

Bivariate Analysis:







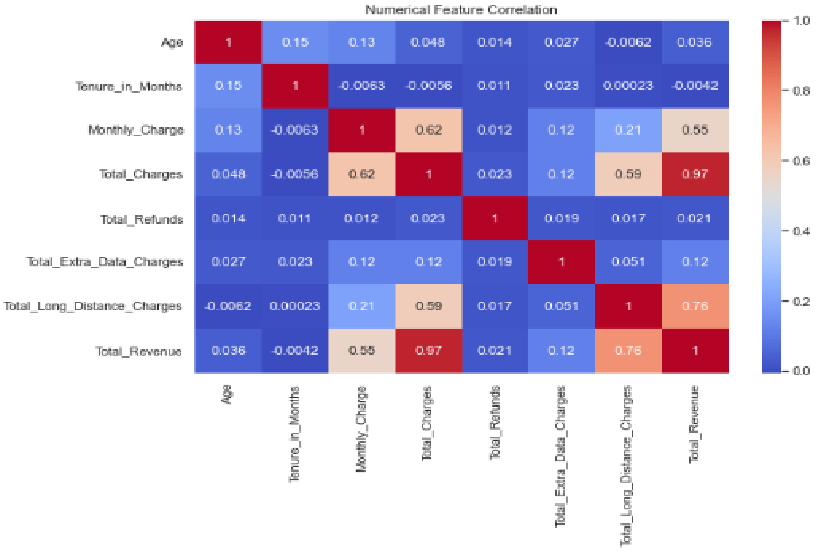


c

Key Observations: Customer Behaviour & Drivers of Churn

- Overall churn rate stands at approx. 29%, indicating that nearly one-third of the customer base is at risk, a material challenge for long-term revenue stability.
- Contract type shows strong influence on churn: customers on month-to-month contracts churn the most, while two-year contract customers show the highest retention — highlighting the value of long-term plans in customer stickiness.
- Tenure and revenue move together: customers who have stayed longer tend to contribute significantly higher total revenue, while churned customers are disproportionately concentrated in the early months of service, signalling early-stage dissatisfaction.
- Internet service type impacts churn: Fiber Optic users show higher churn than DSL or Cable users, hinting at experience or pricing issues specific to that segment; customers with no extra service show mostly retention, suggesting wireless-only segments are not at churn risk.

Correlation Insights



Numerical correlation matrix revealed:

- Total Revenue ≈ Total Charges (high correlation as expected)
- Monthly Charge moderately correlates with churn-associated features
  - Customers who pay high monthly fees usually generate higher revenue, especially when paired with long-distance usage.
  - But if high monthly fees are not matched with loyalty/tenure, they may become high-risk churn customers due to pricing sensitivity.
  - Monthly_Charge does NOT correlate strongly with tenure (≈ −0.006), meaning customers paying higher charges are not necessarily long-term customers.

- Refund-related fields show low correlations individually but are important churn signals behaviourally.
    - Refunds often represent service failure, billing disputes, or dissatisfaction.
    - Even small refund incidents can become triggers for customer frustration.

Machine Learning: Churn Prediction Model

Why Random Forest?

Random Forest was selected because:

- It handles categorical + numerical variables well
- It is robust to outliers and non-linear relationships
- It reduces the risk of overfitting through bagging and ensemble learning

Modeling Steps

1. Dropped non-predictive attributes (Customer_ID, Churn reason)
2. Label-encoding used for categorical features
3. Target encoding:
    - Stayed = **0**
    - Churned = **1**
4. Data split: 80% training, 20% testing
5. Model training using **RandomForestClassifier**
6. Model evaluation using:
    - Confusion matrix
    - Classification report (precision, recall, F1-score)
7. Created predictions on new joiners' dataset, saving only customers predicted to churn

Business Value of the ML Model

The churn model enables:

- Early warnings for high-risk customers
- Customer-specific retention campaigns
- Revenue loss minimization through proactive engagement
- Sales & CRM prioritization based on churn risk probability.

Automation & Production Thinking

To make churn prediction operational:

Component	Function
Power BI Dashboard	Business insights + trends + churn persona profiling
Python/Random Forest	Generates churn probability for each customer
SQL Jobs / Stored Procedures	Automate customer data refresh
Scheduled Model Execution	Weekly churn score generation and export
Automated Alerts	Notify CRM/Retention team with high-risk customer list

This architecture transforms the solution from **analysis** to a **data-driven decision system**.

Key Takeaways

- Churn is not random — it correlates with short tenure, high charges, month-to-month contracts, and low add-on engagement.
- EDA provides the statistical foundation behind the insights shown in Power BI.
- Random Forest delivers strong predictive performance and is interpretable for business users.
- The project enables analytical reporting **and operational value** through automated churn alerts

Possible Future Enhancements

Improvement Area	Idea
Model	Use XGBoost / LightGBM for probability-based scoring
Customer Segmentation	Cluster churners by behavior for personalized offers
Financial Optimization	Retention campaign cost vs revenue preservation analysis
Real-Time Scoring	Deploy through an API for CRM applications

Final Statement

This project blends **data analysis, business intelligence, and machine learning** to solve a highly impactful telecom challenge. It not only identifies **why** customers churn but also **who is likely to churn next**, empowering telecom companies to protect recurring revenue and strengthen loyalty.