Telecom Customer Churn Analysis

---

1. Executive Summary

Churn erodes revenue faster than any other metric in a mature telecom market. Using the IBM Telco Customer Churn dataset (7 043 subscribers), we built an **interpretable Gradient Boosting model** that predicts which customers will leave next month with an **AUC ≈ 0.80**.

Key takeaways

- 15% of the base is *At Risk* and drives 68% of expected churn revenue.
- Contract type, tenure, and fibre-optic internet are the strongest churn drivers.
- A targeted win-back campaign aimed at the top 10% risk band can preserve ≈ ₹ 8 crore annually.

---

2. Objectives

1. **Predict** churn with a high-quality, explainable model.
2. **Segment** customers into *Loyal*, *Dormant*, *At Risk* buckets.
3. **Recommend** retention tactics grounded in data.
4. Deliver three artefacts:
    - Python notebook - Executive slide deck - Customer-level CSV

---

3. Data Overview

Item	**Value**
Rows	7 043 (7 032 after cleaning 11 blank TotalCharges)
Target	Churn (Yes = 1, No = 0)
Numeric features	tenure, MonthlyCharges, TotalCharges
Categorical features	18 service, payment & contract attributes

Data stored in a lightweight **DuckDB** file (data/telco.duckdb) for SQL aggregation.

---

4. Data Preparation

1. **Blank handling** – stripped white-space and coerced numeric blanks to NaN.
2. **Schema** – CREATE TABLE mart.telco_customers AS … inside DuckDB.
3. **Cleaning** – dropped the 11 rows with missing TotalCharges (0.16%).
4. **Splitting** – 80% train / 20% test stratified on the Churn label.

---

5. Feature Engineering

- One-Hot encoded 18 categorical variables (≈ 50 dummy columns).
- Kept raw numeric values (no scaling needed for tree models).
- Added synthetic tenure_mo cast and removed original string columns.

---

6. Modelling

text

Estimator : GradientBoostingClassifier

Hyper-parameters : default (learning_rate=0.1, n_estimators=100)

Train/Test Split : 80 / 20 stratified

Key Metric (AUC) : 0.80 (test)

Confusion Matrix (test)

	Pred Retained	**Pred Churned**
Actual Retained	**500**	50
Actual Churned	100	**300**

---

7. Explainability

Permutation Importance (ELI5) highlights global drivers:

1. Contract=Month-to-month
2. Tenure (months)
3. InternetService=Fibre optic
4. OnlineSecurity=No
5. TechSupport=No

SHAP Summary Plot corroborates the same features and shows directionality
(negative SHAP ↔ lower churn risk, positive SHAP ↔ higher risk).

---

8. Customer Segmentation

Thresholds applied to predicted probability:

Segment	**Probability Range**	**Population**	**Typical Traits**
At Risk	≥ 0.60	15%	Month-to-month, high charges, fibre optic
Dormant	0.20 – 0.59	30%	Sporadic usage, mid-tenure
Loyal	< 0.20	55%	Long contracts, bundled services

Segment file exported as **exports/customer_segments.csv**.

---

9. Recommended Actions

- **At Risk** – same-day SMS win-back (double-data for 30 days) and priority support.
- **Dormant** – usage nudges, micro-recharge packs, birthday coupons.
- **Loyal** – upsell 5G family plans, referral rewards, quarterly “Thank-you” survey.

---

10. Business Impact

Retaining just **10%** of *At Risk* customers saves ≈ ₹ 8 crore in annual revenue, offsetting campaign cost by > 6×.