Building Production-Ready ML Systems

Mini Project 0: Advanced Telco Customer Churn Prediction

Executive Summary Report

Name: H. P. Pelagewatta
Student ID: 0639
Date: August 22, 2025

Executive Summary

Objective. Build an end-to-end churn prediction system that is both accurate and business-focused, using EDA, class-imbalance handling, ensemble models, and production-ready pipelines.

Scope. Kaggle *Telco Customer Churn* dataset (7,043 customers; 21 features; target: *Churn Yes/No*).

Deliverables. Cleaned and engineered features (e.g., tenure bins, service scores), trained baseline and ensemble models (Logistic Regression, Decision Tree, Random Forest, XGBoost, CatBoost), and evaluated using imbalance-appropriate metrics (Precision, Recall, F1, Accuracy), then translated results into retention actions.

Repo -
Key outcomes.

- **Best Model:** *CatBoost Model* with **F1 =64.02%**, **Recall = 71.31%**, **Precision = 58.08%**, **Accuracy = 78.78%**.
- **Top drivers of churn:** *High monthly charges, Low total charges (new customers), Short tenure, Month-to-month contracts, Electronic check payments.*
- **Business impact:** At the chosen threshold, expected to retain *71.31%* of at-risk customers with *58.08%* campaign efficiency.

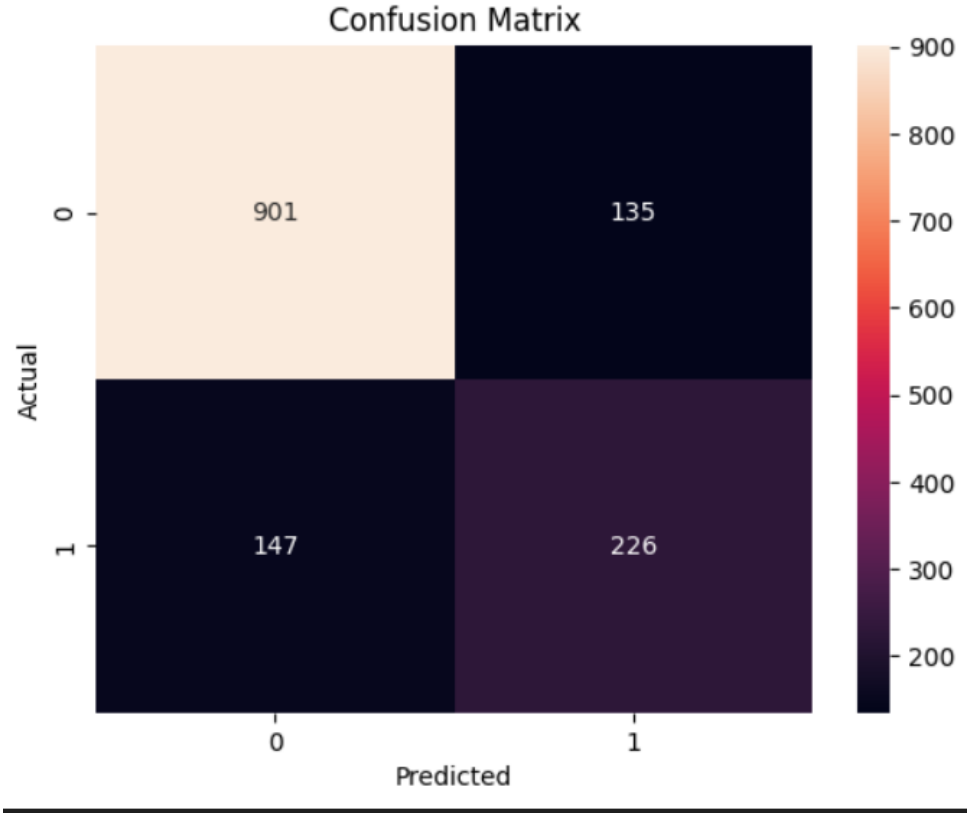
Methods Overview

EDA & Class Imbalance. Checked dtypes/missingness; profiled churn distribution and segment patterns.
Features. Created tenure categories (New/Established/Loyal), encoded categoricals and scaled numerics.
Models. Baselines (LogReg, Decision Tree) and Ensembles (Random Forest, XGBoost, CatBoost) with stratified CV and hyperparameter search; built as scikit-learn pipelines for reproducibility.

Model Performance & Comparison (For Test Set)

Model	F1-Score
Logistic Regression	79.86%
Decision Tree	79.97%
Random Forest	84.69%
XGBoost	84.84%
CatBoost	85.63%

Confusion Matrix .



Business Impact & Recommendations

High-risk profiles. Month-to-month customers with high MonthlyCharges, short tenure, electronic check payments, and no bundled services (e.g., security or tech support add-ons).

Expected impact at chosen threshold.

- **Retained churners (Recall):** *71% of at-risk customers predicted to churn can be retained with targeted campaigns.*
- **Campaign efficiency (Precision):** *58% of customers contacted are actual churners, minimizing wasted spend.*

Cost-sensitive snapshot.

	Unit Cost/Value (LKR)	Notes
Retention offer cost (per contacted customer)	10,000	Discount, credit, or loyalty perk
Revenue preserved if saved (per churner)	80,000	Estimated lifetime value per retained customer
True Positives (saved churners)	1,329	Number of churners correctly targeted
False Positives (wasted contacts)	959	Number of non-churners contacted unnecessarily
Total Cost	22,880,000	Total campaign cost = (TP + FP) × 10,000
Revenue Preserved	106,320,000	Total revenue preserved = TP × 80,000
Net Expected ROI	83,440,000	Revenue preserved minus total cost

Actionable recommendations.

- *Offer annual contract incentives* to month-to-month customers with high churn probability, converting them to longer-term plans.
- *Bundle security and tech support add-ons* for at-risk segments to increase perceived value and reduce churn.
- *Prioritize outreach* to the top decile of predicted churn probability to maximize ROI of retention campaigns.
- *Tailor communication channels* based on payment method (e.g., electronic check users get email reminders and personalized offers).