**Project Summary: Predicting Customer Churn in Style (with LightGBM)**

**By Brent Janaky, 2025**

**Executive Summary (a.k.a. The TL;DR)**

Let’s face it—losing customers hurts. Like, a punch-in-the-bottom-line kind of hurt. That’s why this project tackled B2B customer churn head-on for a telecom provider (“Telco”) using LightGBM, a machine learning model that’s both lightning-fast and scarily accurate.

Armed with a dataset of over 7,000 customers and features like contract type, billing info, and tenure, the model achieved a crisp **84.4% accuracy** and an impressive **0.93 AUC-ROC** score. Translation? It’s really good at spotting customers who are on their way out the door.

From these predictions, Telco can roll out laser-focused retention efforts like loyalty perks and long-term contract sweeteners—because it’s way cheaper to keep customers happy than to woo new ones with flowers and onboarding costs.

**The Game Plan: Project Overview**

My goal: build a model smart enough to spot which B2B customers are itching to break up with Telco, and why. Customer churn is sneaky and expensive—like trying to fill a leaky bucket with sparkling water.

But here’s the upside: churn prediction = ROI booster. If we know who’s likely to churn, we can intervene early, keep them around, and save millions. Plus, happy customers become brand cheerleaders. And unlike your friends, they *love* talking about their phone plans.

**Methodology (a.k.a. What Actually Happened)**

**Data Wrangling & Feature Magic**

The dataset had 7,043 customer records with goodies like contract type, tenure, and monthly charges. We cleaned it up, encoded the categories, and engineered a few new features like:

* **ChargePerMonthRatio** = TotalCharges ÷ (Tenure + 1)
* **TenureContractInteraction** = Tenure × Contract

Why? Because sometimes the real story is in the math we invent.

**Class Imbalance: The “Churners” Were Outnumbered**

Churners were the minority (only about 28%), so we brought in **SMOTE** (a.k.a. Synthetic Minority Oversampling Technique) to balance the data and give our model a fair shot at recognizing churn.

**Modeling: Enter LightGBM**

LightGBM was chosen because:

* It’s fast
* It’s efficient
* It doesn’t complain about big datasets

Hyperparameters were tuned with **Optuna**, a brilliant optimization library that finds the best settings without asking 500 annoying questions.

**Evaluation**

On the test set, the model crushed it:

* **Accuracy**: 84.4%
* **AUC-ROC**: 0.93

Translation: It doesn’t just guess well—it *knows* who’s out the door and who’s sticking around.

A graph of a graph

AI-generated content may be incorrect.

**Deep Dive: What We Learned**

**Customer Risk: Who’s Packing Their Bags?**

Every customer got a churn probability (between 0 and 1). Anyone scoring above 0.7 was flagged as high-risk of vanishing. That’s **1,484 customers** or 21%—prime candidates for retention campaigns. File this under: *don’t let them ghost you*.

**Top Churn Drivers**

According to feature importance:

* **Contract Type** was the Beyoncé of predictors—front and center.
* **MonthlyCharges**, **ChargePerMonthRatio**, and **TotalCharges** followed close behind.

Customers on month-to-month contracts and those with high charges? Yeah, they’re flight risks. Time to hit them with loyalty rewards and pricing makeovers. We need to sweeten the deal or lose everything.

A graph with numbers and text

AI-generated content may be incorrect.

**Key Correlations (from our Heatmap of Truth)**

* High monthly charges = high total charges. Shocking, I know.
* Longer tenure customers tend to be on flexible contracts. Maybe it’s time to tempt them with a juicy 2-year deal?

A graph with a red line

AI-generated content may be incorrect.

**Churn Distribution & SMOTE Justification**

The churn bar chart showed ~5,000 loyal customers vs. ~2,000 churners. SMOTE stepped in to make sure our model didn’t play favorites.

A graph with a number of blue squares

AI-generated content may be incorrect.

**Churn Probability Distribution**

The histogram revealed:

* ~21% are high-risk (prob > 0.7)
* Many are low-risk (prob < 0.3)

So don’t throw discounts at the loyal crowd—they’re not going anywhere. Save those incentives for the fence-sitters.

A graph of a number of churn probabilities

AI-generated content may be incorrect.

**Business Implications (a.k.a. Why You Should Care)**

**1. Reduce Churn, Save Millions**

Targeting those 1,484 high-risk customers with well-timed interventions = revenue protection.

**2. Build Loyalty Like a Boss**

We now know that longer tenure = less churn. So let’s build programs to keep that clock ticking.

**3. Price Smarter, Not Harder**

If high charges lead to churn, maybe it’s time to tweak the pricing. Tiered plans, bundles, loyalty discounts—the usual bag of tricks, now powered by data.

**4. Flex That Competitive Muscle**

Telco can now brag (politely, of course) about using AI to reduce churn. Customers stay. Revenue grows. Market share expands. Everyone high-fives.

**The Stack**

* **Python**
* **LightGBM**
* **Pandas**
* **Scikit-learn**
* **Optuna**
* **SMOTE**
* **Matplotlib & Seaborn** for those eye-catching visuals

**LightGBM: The Star of the Show**

If LightGBM were a person, it’d be the one who aces every group project. It’s fast, accurate, and surprisingly easy to work with—perfect for structured/tabular data like this. It beat out XGBoost this round due to better speed and memory usage.