**Project Summary: B2B Customer Churn Prediction Using LightGBM**

**Executive Summary**

This project used a state-of-the-art machine learning model to predict customer churn for a B2B telecom provider (Telco), leveraging the LightGBM gradient boosting framework in Python. By analyzing a dataset of over 7,000 business customer records, including contract details, payment methods, and usage metrics, the model achieved impressive results: an accuracy of 84.40% and an AUC-ROC score of 0.93.

For each customer, the model assigns a churn probability, enabling precise identification of high-risk B2B customers to reduce revenue loss and improve retention rates. Visualizations, such as the correlation heatmap, ROC curve (AUC = 0.93), precision-recall curve (AUC = 0.93), churn distribution bar chart, and distribution of churn probabilities, highlight the model’s exceptional discriminative power and actionable insights. Key business findings reveal that month-to-month contracts, high monthly charges, and shorter tenures are primary drivers of churn, with approximately 20.6% of customers (1,484 out of 7,043) exhibiting churn probabilities above 0.7, signaling a critical segment for retention efforts. These insights drive targeted retention strategies, offering incentives for long-term contracts, optimizing pricing, and enhancing service offerings and loyalty programs. Ultimately these retention efforts should deliver significant ROI by minimizing churn-related costs, strengthening customer loyalty (thus tenure), and enhancing competitive positioning in the B2B telecom market.

**Project Overview**

I set out to create a smart predictive model using LightGBM to spot B2B customers at risk of leaving a telecom provider, so they could roll out targeted retention programs to keep them on board and boost profitability. Losing customers to churn hits hard, it can cost millions in revenue each year, making it crucial to nail accurate churn predictions and get those retention programs going as soon as possible. One big win from cutting churn is the cost savings: it’s way cheaper to keep existing customers than to win new ones, customer acquisition costs (CAC), like marketing, sales, and onboarding, usually outpace customer retention costs (CRC) by 5 to 25 times. Plus, keeping customers happy doesn’t just save money; it builds a loyal base of advocates who’ll spread the word about the company and bring in referrals. Let your customers do the marketing for you!

**Methodology**

* **Data Preprocessing and Engineering**: The Telco Customer Churn dataset, containing 7,043 customer records with 21 features (e.g., contract type, tenure, monthly charges), was preprocessed to handle missing values and encode categorical variables. New features, such as ChargePerMonthRatio (ChargePerMonthRatio=TotalCharges/tenure+1) and TenureContractInteraction (TenureContractInteraction=tenure×Contract), were engineered to capture nuanced relationships driving churn. Of course, there is the option of engineering additional features to help improve the results.
* **Class Balancing**: SMOTE (Synthetic Minority Oversampling Technique) was applied to address the imbalanced distribution of churn (minority class) vs. no-churn (majority class), ensuring the model effectively learned from both outcomes. The churn distribution bar chart (found below) highlighted this imbalance, justifying the approach.
* **Model Development**: LightGBM, a high-performance gradient boosting framework, was selected for its speed and accuracy on tabular data. Hyperparameters were optimized using Optuna, a Bayesian optimization library, to maximize the AUC-ROC score, achieving early stopping at iteration 91 with a validation AUC of 0.93. Optuna helped to automatically test and find the best settings (or “hyperparameters”) for LightGBM, such as learning rate, number of trees, and depth of decision rules. Instead of manually guessing or testing hundreds of combinations, Optuna used smart algorithms to explore options efficiently, saving time and effort.
* **Evaluation**: The model was evaluated on a held-out test set, yielding an accuracy of 84.40% and an AUC-ROC of 0.93, demonstrated by the ROC curve (AUC = 0.93), precision-recall curve (AUC = 0.93), and churn probability distribution, showcasing robust predictive power for identifying churn risk.
* **Churn Probability by Customer**: For each customer, the model predicts a churn probability (0 to 1), enabling detailed risk assessment. High-risk customers (probability > 0.7) were identified, with 1,484 customers (20.6%) meeting this criterion, as shown in the distribution of churn probabilities in the table found below. Results were saved as a CSV file for CRM integration and retention targeting.
* **Visualization and Insights**: Feature importance, correlation matrices, churn distribution, ROC curves, precision-recall curves, and a histogram of churn probabilities were generated using Matplotlib and Seaborn, providing actionable insights into churn drivers and model performance.

**Key Results**

* **Performance Metrics**: The model achieved an accuracy of 84.40% and an AUC-ROC of 0.93, which is quite good, enabling precise identification of at-risk customers.

A graph of a graph

AI-generated content may be incorrect.

* **Feature Importance**: As highlighted in the feature importance chart below, contract type (“Contract”) stands out as the top driver of customer churn, making it the most critical factor in predicting whether B2B customers will leave. The next most important churn predictors are MonthlyCharges, ChargePerMonthRatio, and TotalCharges, indicating that pricing and cost-related factors significantly influence churn risk. This analysis reveals that customers on month-to-month contracts are far more likely to churn, while those with longer tenures demonstrate stronger loyalty. These insights drive our retention strategy, targeting these high-risk customers to switch to long-term contracts, helping to save millions in revenue and boost customer loyalty.

A graph with numbers and text

AI-generated content may be incorrect.

* **Correlations**: As shown in the correlation heatmap below, we uncovered some important connections that help explain why B2B customers might churn. There’s a strong positive link between monthly charges and total charges, those bright red squares mean higher monthly bills often lead to higher overall costs over time (which makes sense), signaling potential pricing pain points for customers. On the flip side, we found a clear negative connection between tenure and contract type, shown by the blue squares, which indicates that longer-tenured customers are more likely to have month-to-month or one-year contracts rather than two-year contracts. This suggests they may prefer flexibility over long-term commitment, potentially increasing their churn risk unless we intervene with retention strategies like loyalty discounts or long-term contract incentives. These insights guide pricing and retention strategies, helping retain high-value customers by offering long-term contract incentives and reviewing pricing for those facing high charges. There are, of course, other correlations which may help in crafting retention programs.

A graph with a red line

AI-generated content may be incorrect.

* **Churn Distribution**: The bar plot of churn distribution highlights the initial class imbalance, with approximately 5,000 customers not churning (Churn = 0) and 2,000 churning (Churn = 1), reinforcing the business need for predictive intervention as well as SMOTE to balance the classes.

A graph with a number of blue squares

AI-generated content may be incorrect.

* **Churn Probability Distribution**: A histogram of churn probabilities showed that approximately 21.0% of B2B customers (1,484 out of 7,043) have a churn probability above 0.7, identifying a critical segment for targeted retention efforts. The distribution also revealed a concentration of low-risk customers (probabilities < 0.3), allowing efficient resource allocation. As an example, we wouldn’t want to provide pricing discounts to low-risk customers as that would be throwing money away. Also, keep in mind, I have defined ‘high-risk customers’ as having a churn probability of 0.7 but that number can be increased or decreased as-needed.

A graph of a number of churn probabilities

AI-generated content may be incorrect.

Retention Programs

**Business Implications**

The insights and predictive capabilities of this model drive significant business results for Telco. Let’s take a look:

* **Reduced Churn and Revenue Loss**: By assigning churn probabilities to each customer and identifying 1,484 high-risk customers (probability > 0.7), Telco can implement targeted retention campaigns, such as offering discounts for long-term contracts or personalized service upgrades. The churn probability distribution and churn distribution bar chart highlight that focusing on these high-risk customers could protect a significant amount of annual revenue.
* **Improved Customer Loyalty**: Highlighting tenure as a protective factor against churn enables the development of loyalty programs or incentives for long-term clients. The goal is to encourage stronger B2B relationships, increasing customer lifetime value (CLV).
* **Optimized Pricing Strategies**: The correlation between high monthly charges and increased churn, suggests opportunities to review pricing, perhaps offering tiered plans or discounted bundled services to retain price-sensitive businesses while maintaining profitability.
* **Competitive Advantage**: Leveraging AI to predict and mitigate churn, as validated by the ROC curve’s high AUC, churn probability distribution, and realistic probability assignments (as found in the CSV file), positions Telco as a data-driven leader in the industry, enhancing customer satisfaction and market share while minimizing customer acquisition costs.

**Technologies Used**

Python, LightGBM, Pandas, Scikit-learn, Optuna, imbalanced-learn (SMOTE), Matplotlib, Seaborn

**What is LightGBM?**

LightGBM (short for "Light Gradient Boosting Machine") is a powerful, fast, intuitive and efficient machine learning tool used to make accurate predictions from data, such as predicting which B2B customers are likely to churn. Think of it as a smart algorithm that learns from patterns in data to identify trends, in this case, who might churn and why. I prefer it over XGBoost in this instance since LightGBM has straightforward parameters, is faster given the large dataset and is more memory efficient.