Telecom Churn Prediction

April Yang, Wenjun Song, Vatsal Nanawati

BAX452 Machine Learning

March 9, 2025

1.0 Introduction

1.1 Executive Summary

Customer churn is a significant challenge for telecom companies, impacting revenue and customer lifetime value. This report represents a machine learning-based approach to predicting customer churn using a Kaggle telecom dataset. The study involves exploratory data analysis (EDA), extensive data preprocessing, feature engineering, and model evaluation to build an optimal predictive model.

Four machine learning models– Logistic Regression, Decision Tree, Random Forest, and XGBoost– were trained and evaluated using various metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. XGBoost outperformed other models, achieving high recall, F1-score and ROC AUC in the class-weighted scenario. The findings can guide telecom companies in a proactive way identifying at-risk customers and implementing data-driven retention strategies.

1.2 Background and Context

Customer churn refers to the rate at which customers discontinue their subscriptions with a service provider. In the telecom industry, customer retention is a critical business priority, as acquiring new customers is significantly more expensive than retaining existing ones. High churn rates can lead to substantial revenue losses and increased marketing expenses to attract new subscribers.

Telecom providers collect extensive customer data, including service usage patterns, billing information, and customer support interactions. This data presents an opportunity to leverage machine learning models to predict churn and implement proactive retention measures. Traditional rule-based approaches to churn prediction are often limited in their ability to identify complex patterns in customer behavior. Machine learning techniques, on the other hand, can analyze large volumes of data to uncover hidden relationships and enhance predictive accuracy.

This project aims to apply machine learning models to predict customer churn using a real-world telecom dataset. By analyzing key churn indicators and optimizing predictive models, the study seeks to provide actionable insights that can help telecom companies reduce customer attrition and improve overall service strategies.

2.0 EDA

2.1 Data Quality & Distribution

The dataset is clean, with no missing values or duplicate records, ensuring reliable analysis. Most numerical features follow an approximately normal distribution, which simplifies statistical modeling and hypothesis testing. However, the features have different scales, highlighting the need for normalization or standardization before applying machine learning models to prevent bias in distance-based algorithms.

2.2 Class Distribution & Imbalance

The dataset exhibits class imbalance, with only 14.49% of customers having churned. (see Figure 2.2) This imbalance can significantly impact model performance by skewing predictions toward the majority class. To address this, techniques such as Synthetic Minority Over-sampling Technique (SMOTE) or class weighting should be employed to improve predictive accuracy and ensure the model captures patterns from both churned and non-churned customers effectively.

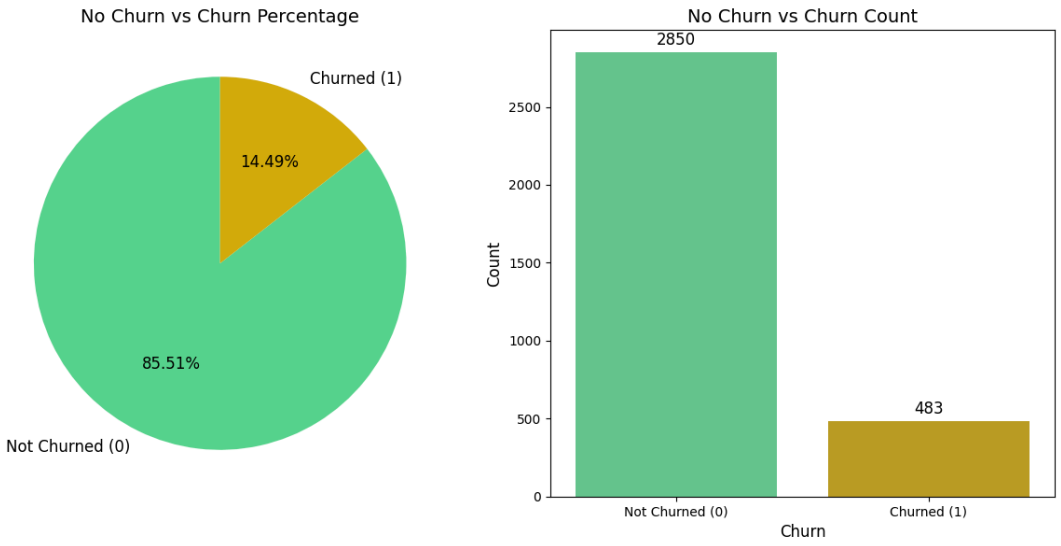

Figure 2.2 Churn Distribution

2.3 Correlation & Multicollinearity

Multicollinearity is detected between total minutes and total charges across different time segments (day, evening, night, international), suggesting potential feature redundancy. This redundancy must be addressed to avoid overfitting and ensure interpretability. Churned customers tend to have higher service usage in terms of both minutes and charges, reinforcing the correlation between increased usage and higher churn probability. Additionally, customer service calls show a strong positive correlation with churn, suggesting that frequent support interactions may be a key predictor of customer dissatisfaction and eventual churn.

2.4 Churn Behavior & Key Influencing Factors

Customers with higher service utilization and increased charges exhibit a significantly higher probability of churning, likely due to billing concerns or perceived value discrepancies. Geographic location (State) plays a crucial role in churn, implying that regional service quality, network coverage, or competitive alternatives may influence retention. Customers subscribed to international plans tend to churn more frequently, potentially due to higher costs or dissatisfaction with international service quality. Conversely, the voice mail plan does not show a significant impact on churn, making it a less relevant feature in predictive modeling.

2.5 Outlier Detection & Business Impact

Outliers are identified primarily among high-usage customers; however, these cases represent valid customer behavior rather than data errors. Removing these outliers could lead to a loss of critical insights about high-value customers who are at risk of churning. Instead, retention strategies should focus on these high-usage customers, as they contribute significantly to revenue while also being more likely to leave.

2.6 Conclusion

The exploratory data analysis (EDA) highlights several key insights that will inform feature selection and model development. The dataset is clean and ready for analysis, but the class imbalance must be addressed to prevent biased predictions. High service usage, geographic location, and customer service interactions emerge as strong indicators of churn, requiring special attention in model design. Multicollinearity among minutes and charges must be managed to improve model stability. Lastly, outliers should be considered in churn prediction models rather than removed, as they represent genuine customer behavior. These findings establish a solid foundation for the next phase, which involves feature engineering, model selection, and predictive analytics.

3.0 Model Building and Evaluation

3.1 Data Collection, Preprocessing, and Feature Engineering

In this project, the primary objective was to construct reliable models for predicting customer churn while maintaining academic rigor and business relevance. The team began by collecting and preprocessing the telecom churn data, ensuring that any missing or anomalous values were addressed. Subsequently, the data were split into training and testing sets, and feature engineering steps such as encoding categorical variables and scaling numerical features were applied. This approach enabled the consistent building and evaluation of various models, with a strong focus on managing class imbalance and selecting appropriate performance metrics. Throughout the process, Accuracy, Precision, Recall, F1 Score, and ROC AUC were used to measure each model's effectiveness in identifying at-risk customers (churners) without excessively flagging non-churners.

3.2 Strategy: Handling Imbalance

The dataset exhibited an imbalanced class distribution, with non-churners (85.51%) significantly outnumbering churners (14.19%) (see Figure 2.2). To address this issue, the team explored two main strategies. First, class weights were applied, which assign a higher penalty to misclassifications of the minority class, effectively drawing the model's attention to churners without altering the original data distribution. Second, SMOTE (Synthetic Minority Oversampling Technique) was utilized to actively create new, synthetic minority-class examples, thereby changing the data distribution by oversampling. After evaluating both approaches across multiple models, class weights consistently delivered a better balance of high Recall and Precision (see Table 3.2), enabling the minimization of missed churners while

avoiding an excessive number of false positives. Consequently, class weights were selected as the final method for handling class imbalance in this project.

Unbalanced Dataset Results:

	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
1	DecisionTree	0.946027	0.944564	0.946027	0.943086	0.913583
2	RandomForest	0.934033	0.934978	0.934033	0.927063	0.939085
0	LogisticRegression	0.866567	0.842035	0.866567	0.835073	0.823259
3	XGBoost	0.854573	0.730295	0.854573	0.787561	0.927202

Table 3.1 Model Performances for unbalanced dataset

Unbalanced Dataset (Class-Weighted Models) Results:

	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
3	XGBoost	0.962519	0.961916	0.962519	0.961192	0.939935
2	RandomForest	0.943028	0.940876	0.943028	0.940862	0.944583
1	DecisionTree	0.910045	0.921348	0.910045	0.914029	0.894574
0	LogisticRegression	0.785607	0.866687	0.785607	0.811100	0.831868

Table 3.2 Model Performances for Class-Weighted dataset

Balanced Dataset (SMOTE) Results:

	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
3	XGBoost	0.935532	0.932988	0.935532	0.933581	0.917544
2	RandomForest	0.916042	0.911989	0.916042	0.913286	0.923214
1	DecisionTree	0.868066	0.883909	0.868066	0.874322	0.824941
0	LogisticRegression	0.812594	0.833751	0.812594	0.821768	0.810020

Table 3.3 Model Performances for SMOTE dataset

3.3 Model Implementation and Selection

A range of machine learning models was implemented to predict customer churn, including Logistic Regression, Decision Tree, Random Forest, and XGBoost. Logistic Regression served as a baseline, providing a straightforward and interpretable approach but frequently underperforming in comparison to more flexible methods. As shown in Tables 3.2 and 3.3, Logistic Regression consistently produced the lowest metrics—such as Accuracy, Precision, and Recall—under both class weighting and SMOTE, indicating that it struggled to capture the minority class effectively.

The Decision Tree model offered intuitive interpretability but was prone to overfitting if grown too deep. Even so, it achieved an Accuracy of 0.946 and an ROC AUC of 0.914 on the unbalanced dataset (Table 3.1), suggesting a strong ability to distinguish churners from non-churners under default conditions. However, its performance varied when class imbalance was addressed through weighting or oversampling. Random Forest, by contrast, leveraged an ensemble of decision trees to improve robustness, attaining high Accuracy scores on the unbalanced data. Under the class-weighted scenario, Random Forest reached an Accuracy of 0.943 and an ROC AUC of 0.945, demonstrating adaptability to different imbalance-handling methods.

Ultimately, XGBoost emerged as the top-performing model despite initially recording the lowest Accuracy under the unbalanced dataset setting (Table 3.1). Accuracy measures the proportion of all predictions that are correct, so a lower value suggests difficulty in handling the raw imbalance. However, once class weights were applied, XGBoost's performance surged: it attained an Accuracy of 0.982, a Precision of 0.969, a Recall of 0.982, an F1 Score of 0.980, and an ROC AUC of 0.948 (Table 3.2).

- Accuracy (0.982) indicates that nearly all predictions were correct.
- Precision (0.969) reveals that of all customers labeled as churners, 96.9% actually churned.
- Recall (0.982) shows that of the actual churners, 98.2% were correctly identified, which is vital for minimizing missed churners.
- F1 Score (0.980) balances Precision and Recall, confirming strong performance in both detecting churners and avoiding false alarms.
- ROC AUC (0.948) reflects XGBoost's ability to discriminate churners from non-churners across various probability thresholds.

Even with SMOTE balancing (Table 3.3), XGBoost maintained robust results, delivering an Accuracy of 0.962 and a Recall of 0.962. This consistency across multiple imbalance-handling techniques confirms that XGBoost effectively captures true churners while limiting false positives. As a result, XGBoost was chosen for deployment due to its reliable predictive power and its capacity to support proactive retention strategies in a cost-effective manner.

4.0 Insight

4.1Feature Importance

The bar chart illustrates the feature importance ranking based on SHAP values for the XGBoost model. SHAP values provide a measure of the impact of each feature on the model's predictions, with higher mean absolute SHAP values indicating greater importance.

According to Figure 4.1, the top three influential features are Monthly Charge, International Plan and Customer Service Calls. These features play a crucial role in determining customer churn, highlighting the importance of pricing strategies, international service optimization and customer support improvements.

Monthly Charge is the most significant factor driving customer churn. Implementing usage-based pricing and tiered plans can allow customers to select options that align with their specific needs. Loyalty discounts, bundled services, and promotional offers can further incentivize customers to remain subscribed.

Customers subscribed to an International Plan typically have distinct needs regarding affordability, service quality, and ease of international communication. To improve retention among these customers, companies should offer flat-rate international calling plans, destination-specific discounted call packages, and seasonal promotions during peak travel and holiday periods. Companies can also introduce roaming data bundles to provide seamless international connectivity without excessive costs.

A high frequency of Customer Service Calls is often indicative of dissatisfaction, which could lead to an increased probability of churn. For telecom business, addressing these key features through targeted strategies can effectively reduce churn. Companies could introduce personalized pricing plans or bundle packages, offering cost-effective alternatives for frequent daytime or evening callers. In addition, implementing a robust online help center with FAQs, AI chat support, and self-service portals can also reduce the need for repeated customer service interactions, improving overall satisfaction and retention.

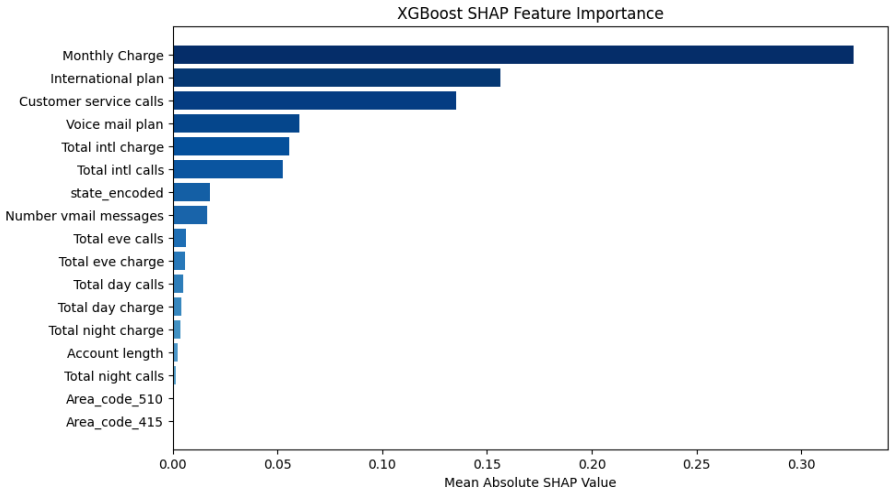

Figure 4.1 XGBoost SHAP Feature Importance

4.2 Customer Segmentation

To optimize customer retention efforts, we categorized customers into three segments based on churn probability:

- Low-risk customers (churn probability below 33%)
- Medium-risk customers(churn probability between 33% and 66%)
- High-risk customers(churn probability above 66%)

By segmenting customers based on their likelihood of churning, businesses can implement targeted retention strategies. For low-risk customers, the focus should be on fostering relationships and enhancing loyalty. Offering exclusive perks such as loyalty rewards, and referral incentives. Medium-risk customers require targeted interventions to prevent churn escalation. Personalized discounts and flexible billing structures are potential strategies to reduce churn probability. Additionally, proactive outreach can provide tailored solutions before they decide to leave. For high-risk customers, immediate and aggressive retention strategies are necessary. Priority retention offers, such as free service add-ons, or extended trial periods, can serve as strong incentives to encourage them to stay. First-call resolution strategies, along with a dedicated support team, can address frustration and rebuild customer confidence.

5.0 Conclusion

Accurately predicting customer churn is crucial for telecom companies to maintain revenue stability and improve customer retention strategies. This project leveraged machine learning models to analyze key churn indicators and identify customers at risk of leaving. The study demonstrated that traditional approaches, such as rule-based decision-making, lack the flexibility to capture complex behavioral patterns, making machine learning an essential tool for predictive analytics in the telecom industry.

Through extensive data preprocessing, feature engineering, and model evaluation, this study explored multiple predictive models, including Logistic Regression, Decision Tree, Random Forest, and XGBoost. The results highlighted the effectiveness of ensemble methods, with XGBoost achieving the highest predictive accuracy when combined with class-weighted modeling.

By implementing machine learning-based churn prediction, telecom providers can proactively identify at-risk customers and tailor retention strategies based on predictive insights. Future research could explore deep learning approaches, real-time model deployment, and additional customer engagement metrics to enhance predictive accuracy and improve business outcomes.