TELECOM CHURN PREDICTION PROJECT DATA REPORT

This project aims to build a binary classifier to predict customer churn for SyriaTel, helping the company identify customers likely to churn. It involves data preprocessing, analysis, model selection, and evaluation to achieve accurate predictions.

BUSINESS OVERVIEW

Customer churn refers to customers ceasing business with a company, impacting revenue and profitability in competitive markets. For SyriaTel, reducing churn is crucial to retaining customers and ensuring sustainable growth.

Churn detection identifies patterns in customer behavior, such as service complaints or reduced usage, that indicate a likelihood of leaving. It involves analyzing customer data over time, including interactions and usage patterns, as well as feedback. Advanced churn prediction models use historical data and machine learning to forecast at-risk customers.

Detecting churn is challenging due to the complexity of customer behavior, influenced by various factors making root causes hard to pinpoint. Telecom datasets are often imbalance, with fewer churn cases, which significantly impact model accuracy. Evolving customer behaviors and changing churn factors further complicate consistent predictions. External influences like market conditions also add additional layers of complexity to the churn detection process.

The stakeholders in this project include SyriaTel management, who'll leverage insights to reduce churn, and marketing and customer service teams, who will target high-risk customers with retention strategies. Data science and IT teams responsible for implementing, maintaining, and updating the churn prediction model are also stakeholder in this project.

PROPOSED SOLUTIONS

Exploratory Data Analysis (EDA): Analyze customer data to uncover trends and identify key factors contributing to churn.

Model Selection and Training: Train and evaluate machine learning models, including decision trees and logistic regression.

Model Evaluation: Assess the model's performance using metrics like accuracy, precision, recall, and F1-score, focusing on correctly predicting high-risk churn customers.

PROJECTED CONCLUSION

This project aims to help SyriaTel reduce churn by providing a predictive model for predicting customers who might churn soon. The model aims to enable data-driven decision-making, allowing the company to target high-risk customers and enhance overall satisfaction.

PROBLEM STATEMENT

SyriaTel faces significant challenges with customer churn, which threatens profitability and market stability. To address this, the company seeks a machine learning solution to predict churn, identify contributing factors, and provide actionable insights for retention strategies to reduce revenue loss and enhance customer loyalty.

OBJECTIVES

Predict Customer Churn: Develop a classifier to categorize customers as "churn" or "non-churn," using historical data and improving performance metrics like precision, recall, and F1-score.

Identify Key Drivers of Churn: Analyze model output to uncover factors strongly linked to churn, providing actionable insights for prioritizing high-risk customers.

Guide Retention Strategies: Support SyriaTel with data-driven strategies, optimizing resource allocation for high-risk, high-value customers.

METRICS OF SUCCESS

To evaluate the logistic regression and decision tree models for customer churn prediction, we used metrics that show accuracy and generalization to unseen data. The key metrics include;

- Precision - for resource optimization.
- Recall - for identifying at-risk customers.
- F1-score -for balancing precision and recall.
- AUC-ROC -for robustness in distinguishing churners.

With these metrics, we suggested the best-performing model among the models built for deployment, ensuring effective prediction on future customer data.

DATA UNDERSTANDING

The dataset for this project, titled "Churn in Telecoms" from Kaggle, includes customer demographics, usage statistics, and service details to predict churn. Kaggle is a reputable platform offering diverse, real-world data, making this dataset ideal for modeling customer churn in the telecom industry. You can access the dataset directly via this link.

UNDERSTANDING THE ROWS AND COLUMNS.

The dataset contains 3,333 rows and 21 columns, providing different information on SyriaTel customers, including whether they’ve churned or not.

Column Breakdown:

1. **State**: The state where the customer is located (categorical).
2. **Account Length**: The duration of the customer’s account in days (numerical).

3. **Area Code**: The area code associated with the customer's phone number (categorical).
4. **Phone Number**: The customer's phone number (categorical, though not useful for analysis).
5. **International Plan**: Whether the customer has an international calling plan (binary: yes/no).
6. **Voice Mail Plan**: Whether the customer has a voicemail plan (binary: yes/no).
7. **Number Vmail Messages**: Number of voicemail messages the customer has (numerical).
8. **Total Day Minutes**: Total minutes used by the customer during the day (numerical).
9. **Total Day Calls**: Total number of calls made by the customer during the day (numerical).
10. **Total Day Charge**: Total cost of the customer's day-time calls (numerical).
11. **Total Eve Minutes**: Total minutes used by the customer during the evening (numerical).
12. **Total Eve Calls**: Total number of calls made by the customer during the evening (numerical).
13. **Total Eve Charge**: Total cost of the customer's evening calls (numerical).
14. **Total Night Minutes**: Total minutes used by the customer during the night (numerical).
15. **Total Night Calls**: Total number of calls made by the customer during the night (numerical).
16. **Total Night Charge**: Total cost of the customer's night-time calls (numerical).
17. **Total Intl Minutes**: Total international minutes used by the customer (numerical).
18. **Total Intl Calls**: Total number of international calls made by the customer (numerical).
19. **Total Intl Charge**: Total cost of international calls made by the customer (numerical).
20. **Customer Service Calls**: Number of calls the customer has made to customer service (numerical).
21. **Churn**: Whether the customer has churned (binary: 0 for staying, 1 for churned).

This dataset doesn't have any missing values or duplicates.

The **churn** column is particularly important as it is the target variable for the classification models, indicating whether a customer has churned. Features such as call minutes, charges, and customer service interactions offer rich data to analyze patterns that may correlate with churn behavior, providing actionable insights for customer retention strategies.

DATA ANALYSIS

Data analysis was performed on the dataset and the following are some of the findings:

Distribution of the target class: There's a substantial imbalance in our dataset. 85.5% of the rows belong to the "False" class while 14.5% of rows belong to the "True" class. This also shows that 15% of customers at SyriaTel have churned.

Area Code Analysis: From the visualization of this analysis, it was observed that area code 415 had the highest number of SyriaTel customers, followed by 510 and 408. Changed the data type of area code to object as its an identifier. This distribution, however, is most likely influenced by the population size of these areas. The organization might want to further investigate the number of users in these areas as well as the population to streamline their services to suit sizes of people they are serving.

Churn by State: Analyzed churn by state, finding that WY, VA, and AL had the highest retention, while TX, NJ, and MD had the most churn. These trends we also visualized.

Top 20 States with Highest Churn: Identified TX, NJ, MD, MI, and NY as the states with the highest churn, which may, however, be influenced by population size. Visualized this data with a bar plot.

Top 20 States with Highest Customer Service Calls: Found that WV, NY, OR, MN, and VT had the highest number of customer service calls, with many high-churn states also ranking here. Also investigated the link between churn and service calls but still kept in mind that these high numbers could be due to the population size of these areas.

SUMMARY FINDINGS IN EDA AND DATA ANALYSIS

- **Customer retention duration for churned customers**: On average, it took approximately 102 days (3.4 months) for a churned customer to stop using SyriaTel's services.
- **Call patterns**: The total number of local calls made across different times of the day is relatively consistent, whereas the number of international calls is significantly lower.
- **Service plan**: The majority of SyriaTel customers do not subscribe to international or voicemail plans.
- **Churn by area code**: Area code 415 reports the highest number of churns, but it is also the area where SyriaTel services are most utilized.
- **State-level trends**: States such as TX, NJ, NY, MD, CT, OR, and MI exhibit both the highest number of churns and the highest volume of customer service calls.
- **Class imbalance**: The dataset is imbalanced, with 85% of the target class labeled as "False" (not churned) and only 15% labeled as "True" (churned).

DECISION AFTER DATA ANALYSIS

From the analysis carried out, these are the conclusions about the data that was used to build the classifier to help predict churn among SyriaTel customers.

1. The phone number column is an identifier. It doesn't provide any useful predictive value about churn. So it was be dropped.
2. All the other columns went through feature selection and a test for multicollinearity to decide which ones are kept for modeling.
3. Encoding was performed for the state, area code, churn, international plan, and voice mail plan columns as they are categorical variables.
4. A test for multicollinearity(vif) was be performed to further decide columns to drop for an effective model.

DATA PREPARATION & FEATURE SELECTION FOR MODELING

SPLITTING

The dataset was split first before any further preprocessing and feature selection. This is to prevent data leakage and leave the testing set untouched to simulate real-world data that the model hasn't 'seen'. **80% of the dataset was used for training the algorithm and 20% for testing the algorithm's performance.**

The **'stratify'** parameter was utilized to ensure that the distribution of the target class in the train and test sets remain consistent with the original dataset, since our dataset is highly imbalanced.

ENCODING

The categorical columns were one-hot-encoded since these machine learning algorithms 'understand' numerical values only. During encoding fitting was done only on the training set, but transformation was done on both the training and test sets.

The target column contains Boolean values so encoding was straightforward here, utilizing the pandas. map() method to assign True (churn) =1 and False (no churn) =0.

MULTICOLLINEARITY

After encoding scaling was done on the dataset and in this section, VIF was calculated on the scaled training data. The calculations were performed only on the training dataset. The Variance Inflation Factor (VIF) measures multicollinearity among independent variables, with each VIF value representing how much the variance of a feature is inflated due to correlation with other features.

For this project, both VIF and model regularization (during modeling) were combined to deal with multicollinearity. Used this VIF analysis to drop a few features with the highest multicollinearity to simplify the data. Then, during modeling, regularization (Lasso or Ridge) was utilized to further fine-tune feature importance and handle any remaining multicollinearity.

CLASS IMBALANCE & FEATURE SELECTION

For addressing the severe class imbalance and also feature selection, Logistic Regression's built-in feature selection strategy (L1/L2) were implemented to refine features and improve model performance. To address class imbalance, the class_weight='balanced' parameter was applied to the model to adjust class weights inversely proportional to their frequencies.

MODELING

Classification models were built and evaluated iteratively to predict customer churn, starting with a simple logistic regression model as the baseline. The logistic regression model was then refined, and later a decision tree model was introduced due to its interpretability and flexibility. Both models were evaluated using precision, recall, F1-score, and AUC-ROC to assess their performance. The goal was to identify the best-performing model among the models built that generalizes well to unseen data and provides actionable insights for addressing churn.

INTERPREATING THE METRICS OF THE BASELINE MODEL

The model's precision for predicting churn (class 1) was 0.27, indicating only 27% of churn predictions are correct, while precision for non-churn (class 0) was 0.92, showing high accuracy for non-churn predictions.

Recall for churn was 0.61, meaning it identifies 61% of churn instances, and recall for non-churn was 0.72. The F1-score for churn was 0.38, indicating poor balance between precision and recall for churn, while non-churn had a strong F1-score of 0.81.

The overall accuracy was 71%, but this is skewed due to the imbalanced dataset, and the AUC-ROC score was 0.694, suggesting moderate model performance in distinguishing churn. Optimization strategies including hyperparameter tuning, alternative models like decision trees, and exploring class balancing techniques such as oversampling or SMOTE were considered.

OPTIMIZATION 1 DISCUSSION

Accuracy: Decreased from 0.71 (baseline) to 0.69 (optimized).

AUC-ROC: Dropped from 0.694 to 0.687, indicating slightly reduced performance in distinguishing churn.

F1-score for Class 1 (Churn): Decreased from 0.38 to 0.37, showing reduced balance between precision and recall for churn.

This decline in these metric scores could be attributed to overfitting which may have occurred during hyperparameter tuning or SMOTE oversampling technique introducing synthetic samples that misalign with data.

To try and optimize this model further, broadening the range of the hyperparameter grid C values, trying alternative resampling techniques like ADASYN were implemented.

OPTIMIZATION 2 DISCUSSION

Recall for the minority class improved (from 61% with SMOTE to 65%), indicating the model detected slightly more churn cases. However, **precision** for the minority class dropped significantly (from 27% to 23%), meaning that many of the predicted churn cases were false positives. Due to this tradeoff, we got a lower **F1-score (34%)** for the minority class compared to SMOTE (37%), since F1-score balances precision and recall. Overall **accuracy**, dropped (from 70% to 64%), suggesting a higher number of incorrect predictions.

This model performance dropped compared to the first optimization and therefore FURTHER OPTIMIZATION STRATEGIES were considered like exploring a decision tree model which can often handle imbalanced data better, since class imbalance was the biggest issue with this dataset.

DECISION TREE BASELINE DISCUSSION

With the decision tree these were the metric scores;

- **Accuracy**: Logistic Regression (64%) vs. Decision Tree (78%) — decision tree improved accuracy by better classifying the majority class.
- **Recall (Minority Class)**: Logistic Regression (65%) vs. Decision Tree (32%) — decision tree struggled with the minority class.
- **Precision (Minority Class)**: Logistic Regression (23%) vs. Decision Tree (27%) — slight improvement for decision tree due to fewer false positives.
- **AUC-ROC**: Logistic Regression (0.686) vs. Decision Tree (0.588) — logistic regression performed better.

One of the advantages of the decision tree here was the higher accuracy and better precision for the minority class, but struggles with recall. However, due to decision tree biases towards majority class, we got poor minority-class recall and lower AUC-ROC.

Again, further optimization strategies were considered since the model's performance was poor. These are the optimization strategies;

1. Apply class weights to the decision trees to address the class imbalance.
2. Optimize hyperparameters like max_depth, min_samples_split.

DECISION TREE OPTIMIZATION DISCUSSION

The optimized decision tree improved accuracy from 78% to 87% by tuning hyperparameters like max_depth and min_samples_split, but recall for class 1 declined from 32% to 27%, possibly due to a focus on overall accuracy. Precision for class 1 increased significantly from 27% to 59% by using class weights to reduce false positives. The **AUC-ROC improved from 0.59 to 0.70, indicating better ability to distinguish between classes**.

MODELS COMPARISON

In this project, multiple models were built to predict customer churn, starting with a baseline logistic regression model that achieved 64% accuracy and an AUC-ROC of 0.6861, with strong precision for the negative class but poor precision for the positive class. To address class imbalance,the logistic regression model was optimized using SMOTE, which improved accuracy to 70% and recall for the positive class but reduced precision

Then, a baseline decision tree was built, which achieved 78% accuracy but struggled with recall (32%) and AUC-ROC (0.5878). However, after hyperparameter tuning, the decision tree's performance improved to 87% accuracy and 0.7024 AUC-ROC, balancing precision (59%) and recall (27%) for the positive class, and performing well for the negative class. **Despite better handling of class imbalance, the decision tree still faced challenges in dealing with the imbalance.**

FINAL MODEL SELECTION

The final model selected was the last optimized decision tree, with **87% accuracy and 0.7024 AUC-ROC**, chosen for its balance between accuracy and interpretability. It decently handled non-linear patterns by recursively splitting the dataset based on significant features.

CONCLUSION

This analysis identified key predictors of churn, which include **international plan**(whether a customer has an international plan or not), **customer service calls**(the number of calls the clients has made to customer service calls, and **account length**(how long the customer has been with the company). Logistic regression initially outperformed in accuracy and AUC-ROC, but an optimized decision tree, enhanced through hyperparameter tuning and feature refinement, ultimately delivered a more balanced performance. Efforts to address class imbalance using SMOTE improved recall for the minority class, while ADASYN had limited impact. The optimized decision tree was selected as the final model, though further tuning is recommended to improve recall and ensure robust churn predictions.

RECOMMENDATIONS

Enhance customer service: Improve the efficiency and quality of customer service to address unresolved issues and reduce churn.

Offer tailored plans for International plan users: Design more competitive or flexible international plans to address dissatisfaction among these customers.

Identify customer likely to churn early: Use the model's predictions to identify customers at risk of churning and deploy personalized retention strategies.

Improve data collection: Enhance data collection practices and collecting more representative data for better modeling in the future.

NEXT STEPS

Best Next Steps for the Project:

1. Further improve and optimize the models.
2. Deploy the optimized decision tree model to a production environment for real-time access and decision-making.
3. Continuously monitor model performance by tracking key metrics like accuracy, precision, recall, and AUC-ROC to detect data drift and performance degradation.
4. Collect more data on customer behaviors, demographics, usage patterns, and service interactions to improve the model's generalization and insights.