Predictive Analysis of Customer Churn

COMS W4995 Project - Group 7

Dhvanil Nanshah
dn2572

Ta-Wei Huang
th3061

Lu Liu
ll3721

Yaqing Wang
yw3955

Leah Uzzan
lu2166

I. Introduction

Customer retention is essential for the success and longevity of any business. "Churn" refers to the percentage of customers who leave a service provider or supplier within a certain period. Industries such as the credit card sector focus mainly on reducing churn and identifying customers at risk of leaving. High churn rates can severely impact revenue, especially for businesses with subscription-based models.

JB Link Teleco, a company in California offering phone and internet services, has experienced significant customer churn despite its early rapid growth and investments in expanding its market presence and infrastructure. Addressing this high churn rate is crucial for the future stability and growth of JB Link Teleco. This report performs a whole data processing pipeline and suggests data-driven strategies with various machine-learning models to decrease customer churn.

II. Exploratory Data Analysis

This dataset is particularly notable for its imbalance, with a smaller number of churn cases (1,869) compared to non-churn cases (5,174). The primary objective in analyzing this dataset is to predict the minority class, which is the customers likely to churn. To utilize all data, we used a stratified split to maintain the original ratios of the Churn and Not Churn classes in the training and test data.

In handling missing data, we observed that three features, Customer Satisfaction, Churn Reason, and Churn Category, have a high missing data proportion (over 70%). Therefore, we decided to exclude these features from our analysis. Additionally, the information potentially offered by Churn Reason and Churn Categories was deemed redundant, as their insights are already reflected in the patterns of other features.

The factors affecting telecom customer churn include less likelihood of referring friends, higher churn rates for specific offers like provider contracts, and more churn among phone service subscribers. Higher monthly charges, shorter service tenures, and senior customer status also contribute to churn. Long tenure correlates with tolerance for higher charges. Customers with higher long-distance charges, possibly due to value or lack of competition, are less likely to churn.

III. Methodology

A. Data Preprocessing

1) **Missing Data:** As described in the previous section, we excluded these attributes (Customer Satisfaction, Churn Reason, and Churn Category) from further analysis.
2) **Data Duplication Verification:** Furthermore, we confirmed the absence of duplicate records by matching the count of unique Customer IDs to the total number of database entries, ensuring data integrity.
3) **Data Standardization:** Numerical features with high variance can impact the effectiveness of models. To mitigate this, we implemented this standardization with StandardScaler and transformed features to have a mean of zero and a standard deviation of one, which avoids bias towards features with larger scales.

B. Feature Engineering

1) **Zipcode to County Mapping** In our feature engineering process, we mapped zip codes to counties to improve geospatial data representation. The high number of unique values for cities and zip codes was reduced by creating a county column, decreasing unique geospatial data points from thousands to 57 counties.
2) **Preliminary Feature Selection:** We eliminated features with a substantial proportion of missing values. Furthermore, we removed Longitude, Latitude, City, and Zip Code from our dataset since the County feature already represents geographic information. Additionally, Customer ID was excluded from the model training data as it does not contribute any meaningful information for modeling purposes.
3) **Categorical Feature Encoding** We applied one-hot encoding to convert categorical data into integer data. For the County feature, which comprises 57 unique categories, we opted for target encoding instead of one-hot encoding to prevent the issue of high dimensionality, which often accompanies one-hot encoding with features having a large number of categories. Additionally, our analysis revealed that none of the features in our dataset were found to require or benefit from ordinal encoding.

C. Baseline Models Overview

1) ***Initial Model Selection****:* An assortment of baseline models (see Figure 1) was created to help in the selection process of which machine learning models to invest the greatest resources in developing. We began our model selection processing utilizing all of the standard ML binary classification models available through Sci-kit-Learn, including logistic regression, decision trees, AdaBoost, gradient boosting classifier,

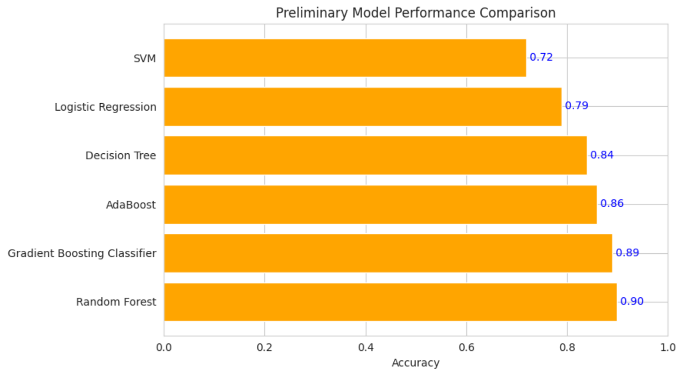

Fig. 1. Preliminary Models

support vector machines, and Random Forests. These baseline models were each created using default hyperparameters and trained on three different sets of training data. These sets include training data utilizing stratified random sampling (the default for sklearn's train_test_split), undersampling of the majority class (non-churn), and oversampling using SMOTE of the minority class (churn). The classification reports were taken for the best outcome using whichever training set for each model and compared. The primary metrics of interest were accuracy, precision, and recall. The precision and recall result hierarchy of the models, for both minority and majority classes, were in agreement with the accuracy hierarchy displayed in Figure 1. On this basis, the Random Forest and Gradient Boosting Classifier models were chosen for further development. [Discuss if space]

*2) **Hyperparameter Tuning**:* Gradient Boosting Classification, AdaBoost and Random Forest models had extensive hyperparameter tuning done utilizing both grid search and random search with k-fold-cross-validation. Approximately 2500 models of Random Forest were fitted in tuning the number of trees, number of features per tree, max depth, minimum sample split, minimum samples per leaf, and Bootstrapping (True or False). Gradient Boosting Classifier had around 1500 models fit in tuning the number of estimators, the learning rate, max depth, the minimum sample split, the minimum samples per leaf, and the sub-sample value. Both models used 5-fold cross-validation. AdaBoost had around 360 models fit in tuning its parameters. The key hyperparameters tuned for AdaBoost were the base estimator (decision tree), max depth of the base estimator, and the number of estimators (trees), and cross-validation was used. The random search values for each hyperparameter were bound utilizing industry standards (such as 0.01-0.3 for the learning rate of gradient boosting classifiers) or based on intuition paired with trial and error. Once optimal hyperparameters were found by the random search, they were then fed into a grid search where a small range of values around those optimal values would be tested using 5-fold cross-validation. The exception is AdaBoost which did not have the time to have a grid search done following the random search. The best parameters resulting from the final searches are used to train our optimized machine-learning models.

*3) **Decision Boundary Tuning**:* The final stage of tuning involved taking our trained, optimized machine learning models and adjusting the decision boundary in order to find an optimal balance between recall and precision for the minority and majority classes without compromising too heavily on the overall accuracy of the model. The decision boundary was explored by viewing the classification reports from the test data for decision boundaries between 0.3 and 0.6 for both models as deviating too far to either end would compromise the overall performance of the model notably.

IV. Results

TABLE I
MODEL RESULT COMPARISON
0: NOT CHURN; 1: CHURN

	Adaboost	Gradient Boosting	Random Forest
Accuracy	0.93	0.93	0.91
Precision (Majority)	0.95	0.96	0.93
Recall (Majority)	0.96	0.94	0.93
Precision (Minority)	0.88	0.83	0.81
Recall (Minority)	0.85	0.89	0.80
AUC	0.91	0.89	0.86

A. Best Models Evaluation

As seen in table I, all three final models boast a similar final accuracy, with AdaBoost and Gradient Boosting performing the best overall with a 0.93 accuracy. In addition, both AdaBoost and Gradient Boosting dominate Random Forest across all metrics shown for both the minority and majority classes where relevant. Gradient Boosting performed 0.01 better than AdaBoost in majority group precision and 0.04 better for minority group precision but slightly worse for all other metrics. In particular, the minority group is of interest, and although the Gradient Boosting has a .04 advantage in minority precision, the AdaBoost performs .05 better in minority recall. Since we aim to delineate strong identifying factors for churn among consumers and not just predict churn among consumers, it is critical to give attention to the minority class recall and precision. A higher minority class recall means that we are doing a better job predicting a large portion of the totality of churns. A higher precision means that we are doing a better job of ensuring that those we predict to be churns are actually churns or reducing false positives, in other words. From a business standpoint, a lower precision comes with the issue that you are able to less correctly identify potential churners and may end up over-allocating resources to marketing to the wrong groups of potential customers, along with the risk of not offering deals and incentives to current customers that are churn risks. Conversely, a low recall comes with the risk of you not identifying enough, even if accurately, of potential churners, thus running similar risks in marketing as above, along with added risks such as making financial decisions, such as those about growth and provided services that may be detrimental to the health of the company. This means that both recall and precision carry similar importance, and although Gradient Boosting and AdaBoost each have a strength in one of those minority group metrics, the overall metrics give a slight edge to AdaBoost. This is reinforced by AdaBoost's better performance on the roc-auc score (.91) versus Gradient Boost (0.89). Paired with the fact that AdaBoost underwent significantly less rigorous hyperparameter results in its favor, suggesting both more potential and marginally better current results.

Although the recall for the minority class in AdaBoost is 4% lower than that in Gradient Boosting, AdaBoost shows a 2% higher recall for the majority class. This aspect is significant in a business context, particularly for identifying customers who are unlikely to churn. Retaining these customers and potentially offering promotions to maintain their loyalty can be a strategic advantage. To determine our final model choice, we compared the AUC curves, as they could provide the aggregate performance measure across all possible classification thresholds. Ultimately, AdaBoost was selected as the final model due to its higher AUC score (.91), indicating its overall effectiveness in classifying both churn and non-churn customers.

B. Feature Importance

TABLE II
TOP 5 FEATURE IMPORTANCE COMPARISON

Rank	Adaboost	Gradient Boosting	Random Forest
1	Number of Referrals	Total Customer Svc Requests	Total Customer Svc Requests
2	Streaming TV	Monthly Change	Monthly Change
3	Total Customer Svc Requests	Product/Service Issues Reported	Number of Referrals
4	Tenure in Months	Contract_Two Year	Contract_Two Year
5	Avg Monthly GB Download	Number of Referrals	Tenure in Months

The table II shows the overview of the most influential features across three distinct predictive models. The key observations include the overlap in feature importance across the models. For instance, 'Total Customer Service Requests' is a top feature for both Gradient Boosting and Random Forest, indicating that frequent customer service requests might be a sign of dissatisfaction with the service provided. It suggests that improving customer service and addressing issues effectively could potentially reduce churn. Conversely, some features like 'Streaming TV' (Adaboost) and 'Product/Service Issues Reported' (Gradient Boosting) are unique to specific models, suggesting model-specific sensitivities to certain data aspects.

Figure 2shows how mixing up a feature affects the model's performance, highlighting which ones are crucial. Total Customer SVC Requests stands out as the top feature for RF and GBM but is not ranked high for AdaBoost in table I. This is in line with AdaBoost's less extensive tuning and suggests the models might be in unanimous agreement if AdaBoost is further tuned. It's a hint for the business to watch customer service interactions closely, as they're clearly linked to outcomes like churn. Investigating which services lead to more requests could be insightful. Similarly, AdaBoost ranks Monthly Charge and Contract_Two_Year differently than other models, which perfectly aligns in ranking with figure 1, further speaking to AdaBoost's potential to widen the gap between it and the remaining models. Despite less fine-tuning, AdaBoost's strong performance hints that AdaBoost is successful due to other factors such as its distribution of weight across multiple features and its better selection of mid and low-tier features with respect to importance.

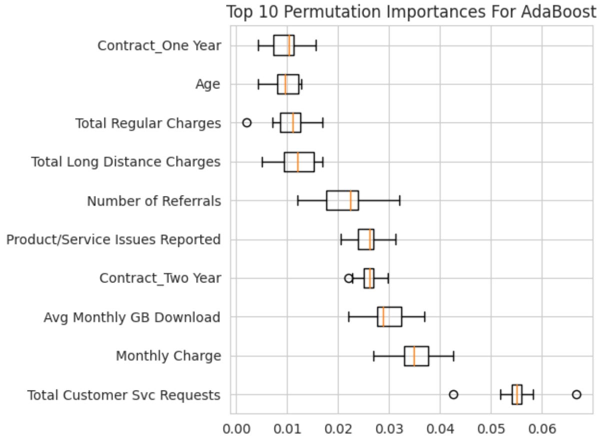

Fig. 2. Feature Importance

V. CONCLUSION

This project effectively addressed the challenge of customer churn at JB Link Teleco, employing data-driven strategies and machine learning models to predict churn likelihood. The application of models like AdaBoost, Gradient Boosting, and Random Forest revealed that the Gradient Boosting model, in particular, stood out with its high accuracy (0.93) and impressive recall for churned customers (0.89). Key features influencing churn included 'Number of Referrals, 'Monthly Charge,' and 'Total Customer Service Requests.'

The insights garnered from this analysis offer actionable strategies for JB Link Teleco, such as refining customer service approaches and enhancing referral programs. While the models demonstrated robust predictive capabilities, future enhancements could include incorporating additional data sources and continuously updating the models to adapt to evolving customer behaviors.

In essence, this project not only provides JB Link Teleco with a strategic tool for reducing customer churn but also sets a precedent for adopting similar analytics-driven approaches in other areas of its business operations. For future improvement, we can combine more domain knowledge to perform better in the feature engineering part and utilize deep learning models to generate better predictions.