Predictive Modeling to Control Customer Churn Rates

Members:
Meng-Wei (Vivian) Wu
Kexin (Shera) Huang
Zonghai (Liam) Wan

Table of Contents

Executive Summary

The COVID-19 pandemic had a significant impact on the global economy, causing a recession and large-scale layoffs started from top ranking technology firms. It has influenced companies to strive to maintain current profit levels or increase business revenue to survive. For B2C businesses, this strongly involves predicting customer churn rate and understanding the parameters that may affect customers' exit. We decided to predict the customer churn rate of the telecommunication industry because of its relative stability compared to tech firms.

In this report, we are going to apply exploratory data analysis to identify patterns and trends that indicate a higher likelihood of churn. It will be used in developing four predictive models.The Random Forest model showed that elderly, couples and male consumers have higher retention rate. Important features identified for marketing campaigns would address specific concerns of at-risk customers. By reducing churn, the company can increase customer satisfaction and loyalty, leading to higher retention rates and increased revenue.

Background and Introduction

Global economy has been facing regression since Covid-19. In 2023, we've seen large-scale lay-offs across major technology conglomerates like Twitter, Meta, Amazon, etc. The recent bank-run of Silicon Valley Bank brought people back to 2008 when Lehman Brothers' bankruptcy. Last weekend, UBS bought Credit Suisse to avoid a bank crisis. Under such bad economic circumstances, companies' top priority is to "survive." In other words, to "maintain or increase the revenue." For most of the "B to C" business, it's essential to maintain existing customers while enrolling new customers. Therefore, the prediction of customer churn rate and understanding of parameters that might affect customers to exit is essential for generating revenue in the end.

In this project, we picked the telecommunication industry to be the targeting industry for customer churn rate prediction for the following reasons. The Telecommunication (Telco) industry is the bandwidth carrier for technology but it is relatively stable compared to the tech industry. Those base stations need time to build and are not easily replaced in a short-term period. However, in this post-pandemic era, customers' behaviors “leap-frogged 5 - 10 years ahead” according to McKinsey. In order to balance out the leap-forwarding customer behaviors, it’s eager to understand customers’ needs and attraction.

Problem-solving Strategies and Model Alignment

The telecommunications industry traditionally attempts to solve the problem of customer churn through reactive measures, such as offering incentives or benefits to existing customers who are at risk of churning. These incentives may include discounts, promotions or free services, which can help to retain customers in the short term. However, this approach may not be sustainable in the long term, as it can be costly and may not address the underlying reasons for churn.

A more proactive approach to reducing churn involves analyzing customer data to identify patterns and trends that indicate a higher likelihood of churn. The analysis can be used to develop targeted marketing campaigns that address the specific concerns of at-risk customers. For example, if customers are more likely to churn due to high charges for internet and phone services, the company could offer a lower-priced bundle that includes these services. Additionally, the company could invest in improving its customer service to address any concerns that customers may have. For example, if customers are not satisfied with the services, the company should improve the quality of the services they provide and strengthen the training of their customer service representatives. Last but not least, the company should anticipate the customer lifetime based on customers behavior patterns and develop effective strategies and solutions to extend the length of customers lifetime in the company.

The proactive approach aligns with the business model of the telecommunications industry, which is focused on providing high-quality services to customers. By analyzing customer data and identifying ways to improve the customer experience, the company can increase customer satisfaction and loyalty, which can lead to higher retention rates and increased revenue. Additionally, by reducing churn, the company can reduce its customer acquisition costs, which can further increase profitability. Overall, the proactive approach to reducing churn is a more sustainable and effective strategy than reactive measures, and it aligns with the long-term goals of the telecommunications industry.

Model Development and Analysis

The data we used for analysis is from Kaggle - Telco Customer Churn[1] and the meanings for each feature are listed below:

Feature	Meaning
customerID	Customer ID
gender	Whether the customer is male or a female (1,0)
SeniorCitizen	Whether the customer is a senior citizen or not (1,0)
Partner	Whether the customer has a partner or not (Yes, No)
Dependents	Whether the customer has dependents or not (Yes, No)
tenure	Number of months the customer has stayed with the company
PhoneService	Whether the customer has a phone service or not (Yes, No)
MultipleLines	Whether the customer has multiple lines or nor (Yes, No, No phone service)
InternetService	Customer's internet service provide (DSL, Fiber optic, No)
OnlineSecurity	Whether the customer has online security or not (Yes, No, No internet service)
OnlineBackup	Whether the customer has online backup or not (Yes, No, No internet service)
DeviceProtection	Whether the customer has device protection or not (Yes, No, No internet service)
TechSupport	Whether the customer has tech support or not (Yes, No, No internet service)
StreamingTV	Whether the customer has streaming TV or not (Yes, No, No internet service)
StreamingMovies	Whether the customer has streaming movies or not (Yes, No, No internet service)

[1] Telco Customer Churn
https://www.kaggle.com/datasets/blastchar/telco-customer-churn?datasetId=13996&sortBy=voteCount

Contract	The contract term of the customer (Month-to-month, One year, Two years)
PaperlessBilling	Whether the customer has paperless billing or not (Yes, No)
PaymentMethod	The customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
MonthlyCharges	The amount charged to the customer monthly
TotalCharges	The total amount charged to the customer
Churn	Whether the customer churned or not (Yes, No)

Data Cleaning

To prepare our data for exploratory data analysis and later predictive modeling, we took several steps. First, we checked the data type of each feature and made any necessary conversions to ensure they were in appropriate types. Next, we removed any rows of data that contained null values to ensure the integrity of our data. While we have not yet transformed the categorical data into numeric values, we plan to do so when developing our predictive models. For now, we believe leaving the categorical data as-is will make our exploratory data analysis more straightforward and understandable, allowing us to better identify patterns and relationships within the data.

Exploratory Data Analysis (EDA)

We first plotted the target variable - Churn to understand its distribution and found that the data is imbalanced with 73% non-churned customers and 27% churned customers. In order to make predictions more accurate, We will conduct several techniques which will be discussed later to make it balanced.

After analyzing the variables in our dataset, we categorized them into two types: those with less than six unique categories as categorical and those with more as numeric. Specifically, when examining the customer information, we discovered that gender does not significantly impact churn rates, as both male and female customers exhibit similar distributions of churned and non-churned customers. While the number of SeniorCitizens in our dataset is low, their churn rates are relatively high. Additionally, customers with a partner and those without dependents have lower churn rates than their counterparts.

Turning to subscription services, most customers have subscribed to PhoneService and exhibit no strong preference for changing their telecommunication provider. However, customers who do not have OnlineSecurity, OnlineBackup, DeviceProtection, or TechSupport are less likely to switch providers. Specifically, the proportion of churned customers who subscribed to these services is much lower than that of non-churned customers. Moreover, StreamingTV and StreamingMovies display identical distribution patterns, with both services showing churn rates of around 30-40%.

Further analysis revealed that customers with Month-to-month Contracts are more likely to churn compared to those with One-year or Two-year contracts. Furthermore, approximately 50% of customers who pay by electronic checks switch their service providers, but this is not the case for those who pay by mailed checks, bank transfers, or credit cards. In addition to the categorical variables, our examination of numeric variables revealed that customers with tenures of less than 20, monthly charges between 60 and 120, and total charges of less than 2000 are more likely to churn.

Overall, the exploratory data analysis provides valuable insights into the factors that drive customer churn rates, and these findings can help on feature selections and model development.

Model Development

Before building a model, it is important to preprocess the data to ensure its suitability for analysis. This involves converting categorical features into numerical values and scaling numeric variables to reduce the impact of extreme values, which can improve the generality of the data. Additionally, it is crucial to select important categorical and numeric features based on their respective chi-squared and f-values, which can help prevent overfitting. To address imbalanced data, we can employ the SMOTE package, which can help balance the data and improve the accuracy of the model.

To predict customer churn, we have selected four models to compare their performance: Logistic regression, Decision tree, Random forest, and XGBoost. By comparing the results of these models, we can identify the one that provides the most accurate predictions.

The dependent variable is “whether customer churn or not.” Therefore, logistic regression model is the first model that came to our mind for binary outcome prediction. In the Logistic regression model, we first define a function for metrics set-up to evaluate the logistic regression model we performed, which are “accuracy,” “precision,” “recall score,” , “F1 score,” and “F2 score.” We later introduced lasso regression to choose features for our final model. We created a logistic regression model object (lasso_log) and set the regularization penalty to L1 (Lasso) by setting the penalty parameter to “ L1.” We then employed a logistic regression model and predicted the outcome using the model. In order to improve our logistic regression model interpretability, the weights of variables are also calculated. The plot (Appendix 1) shows that: positive weights indicate that the variable is positively associated with the outcome variable, while negative weights indicate a negative association; the larger the weight, the stronger the association between the variable and the outcome. We can see from the plot that “total charges” has the highest positive association with outcome variable “churn” while the longer the “tenure”, the less likely the customers are likely to “churn.” The overall logistic regression model accuracy is 0.79.

For the decision tree model, we used *Decision Tree Classifier* to create an instance of the DecisionTreeClassifier class, which represents a decision tree model. In our python code, the tree is built recursively, splitting the data into smaller subsets at each internal node, based on a set of decision rules that maximize the information gain. The overall decision tree function is only 0.76, even less than logistic regression model accuracy. Moreover, since the decision tree is only one single tree and may not be comparable to the random forest model employed by bootstrapping, we later conducted a random forest model.

In the random forest model, we employed the *Random Forest Classifier* function from the scikit-learn library. This function allowed us to specify the ideal number of trees to grow, which we set at five hundred, and select entropy as the splitting criterion. Entropy is calculated at each node split where the lower the value indicates higher in purity of that node. Our goal is to minimize entropy in our forest growing process.

For our XGBoost model, we used the *XGBClassifier* function from the XGBoost library. This model operates similarly to the Random Forest model by building a model with multiple decision trees. Instead of using ‘bagging’ technique, Gradient Boosting Decision Tree improves the weakness of previous trees in sequence. In this model, we have set the number of tree improvements to be one hundred, and the model yielded a similar result to the Random Forest model.

In evaluating the prediction results of the trained models, the Random Forest model yielded the highest accuracy among the four techniques utilized in the analysis (*Appendix 2*). The overall accuracy of predicting customer churn rate using out-of-sample validation is 0.846, indicating that the Random Forest model correctly predicted 84.6% of the test data. In further analysis, we noted that whether a client is elderly, has a partner, client’s gender, or whether client has device protection play a significant role in the churn prediction process.

Recommendations and Business Insights

After analyzing the data, we have identified some key recommendations for telecommunications companies looking to reduce their churn rate using predictive models.

1. We suggest that companies should focus on targeting specific customer groups that are more likely to churn, such as SeniorCitizens, customers with a Partner, and Male customers. These three features are found to significantly impact the churn rate, and targeting these groups could help to mitigate the problem.
2. SeniorCitizens are a valuable customer segment as they are willing to pay higher monthly charges than other customers. Therefore, the company should design services that cater to their needs to ensure their satisfaction and retention.

3. To strengthen the customer base and improve customer loyalty, we recommend creating affordable and appropriate service plans specifically for customers with less than six months of tenure. This will help to establish a strong foundation of customers and encourage loyalty.
4. We recommend that the company consider phasing out electronic checks as a PaymentMethod since it has a high churn rate. Instead, the focus should be on bank transfers or credit cards, which are more secure and reliable payment methods. By adopting these recommendations, telecommunications companies can reduce their churn rate and improve customer satisfaction and retention.

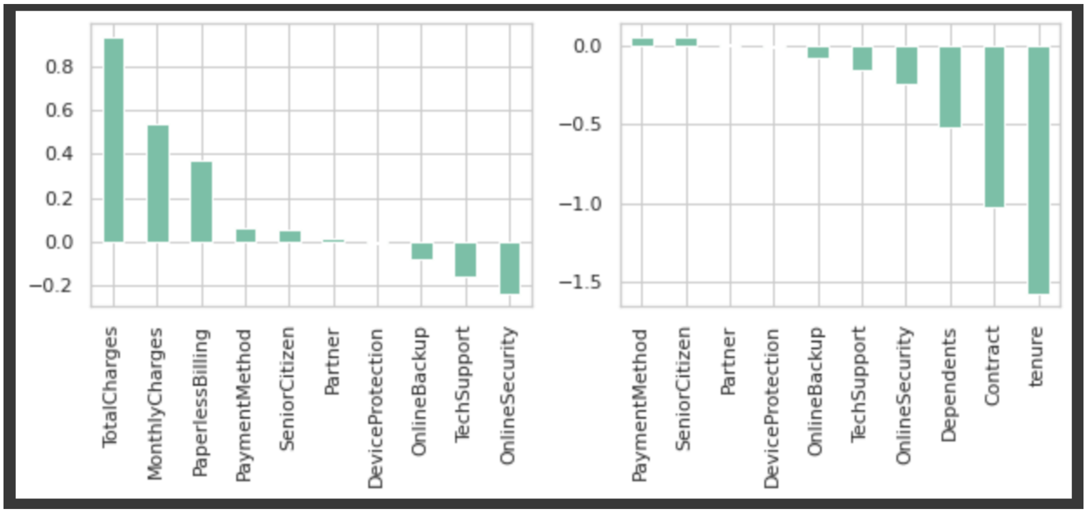
Customer churn rate has been discussed across multiple industries. However, in the telecommunication industry, it is not brought up as often as other industries such as retail, technology, etc. One of the reasons may be that telecom has a relatively higher entry barrier than some other industries - since telecom requires a massive network of base stations. That could be also the reason why previously the telecom conglomerates were not so interested in customer churn rate. However, as mobile phones being the major multifunctional tools in people's daily lives - navigating, communicating, connecting, studying, entertaining, etc. - it's easy to get frustrated when signal is weak and in turn lead to switching the providers. Hence, understanding customer churn rate will not only encourage providers to improve their network and customer relationship, but also by decreasing the customer churn rate will bring more profit and market share to the companies. It's more costly to acquire a new customer than retaining the existing customers. Therefore, It's a win-win situation for both customers and telecom companies. With rapid changing technology - AI-enabled and cloud-based processes - and customers' rapid changing habits, telecom companies will evolve further to be more efficient, effective, and cost-lowering by understanding customers' needs that are derived from customers' churn rate.

Summary and Conclusions

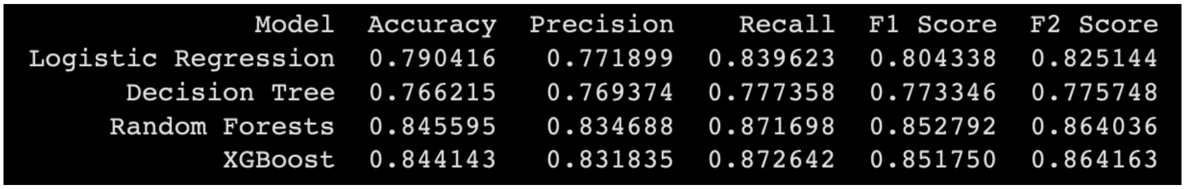
The COVID-19 pandemic has made it critical for businesses to survive and even crucial to focus on increasing profit. Companies need to target specific factors in reducing retention rate. In our case, an effective strategy for the telecommunications industry is to focus on providing high-quality services to customers. Analyzing consumer data to identify potential aspects that could improve customer experience will increase their overall satisfaction and loyalty, resulting in a higher retention rate and possible revenue growth. The dataset used for analysis was cleaned, and exploratory data analysis will help identify patterns and relationships within the data for developing predictive models. Based on our best predicting model, Random Forest 's feature importance rivals that senior citizens, couples and male customers might contain higher retention rate, where we emphasize the importance of consumer satisfaction highlighting the target group.

Appendix

1. Logistic regression model - weights of variables plot



2. Models Results



```
              Model  Accuracy  Precision    Recall  F1 Score  F2 Score
Logistic Regression  0.790416   0.771899  0.839623  0.804338  0.825144
      Decision Tree  0.766215   0.769374  0.777358  0.773346  0.775748
    Random Forests  0.845595   0.834688  0.871698  0.852792  0.864036
            XGBoost  0.844143   0.831835  0.872642  0.851750  0.864163
```