BU425 Final Report

Cole Hanniwell, Ryan Kieswetter, Einstein Oyewole, Jared Um

BUSINESS PROBLEM

Over the past decade, we have seen significant growth in the telecommunications industry with several new players infiltrating the global marketplace. While this has aligned with societal trends regarding technological growth, it has created a fiercely competitive atmosphere and made it more difficult for service providers to retain customers. Regarding our proposed business problem, we wanted to address the issue of customer churn for service providers within the US telecommunications industry. Both customer retention and acquisition are key organizational processes that businesses in all industries must consider. The success of a business in a competitive market, such as those of the telecommunications industry, is often dependent on their ability to meet or exceed customer expectations.

Given the numerous substitutes for customers to evaluate, it is clear that there is high buyer power within the telecommunications industry. With several large players such as AT&T, Verizon, and more, the risk of customer churn is high given the competitive nature and numerous alternative providers. Our goal is to accurately predict customers which are likely to churn and better address uncertainties. The intended outcomes of the following analysis include the identification of which features or factors cause customer churn, as well as create a model that will be able to accurately predict this churn at 80% accuracy. Through these results, we aim to provide functionality that will be able to give telecommunication providers more insight on their customers.

MOTIVATION

The main motivation behind our analysis is to enable telecommunications providers to proactively engage with the high risk of churn customers and reduce their overall customer churn. This will allow service providers to highlight which features are impacting customer churn, ultimately leading to the development of strategies to mitigate this issue. It can also be assumed that the identification of these features will also help to promote customer acquisition as these telecommunication providers will have a better understanding of consumer preferences and behaviour.

With increased efficiency surrounding customer acquisition and retention, service providers will also be able to increase their Customer Lifetime Value to Customer Acquisition ratio (LTV:CAC ratio), indicating higher profitability for the company. The LTV:CAC ratio measures the lifetime value of a customer against the cost of acquiring them. This quantitative measure is important specifically to the telecommunications industry because of the competitive environment and the difficulty associated with acquiring customers. For these telecommunications providers, both customer acquisition as well as retention are key processes that need to be considered in order to be profitable.

DATASET

The dataset we used for the analysis is sourced from Kaggle.com. It contains 4250 samples of customer accounts from a US based telecom provider. Each sample consists of 19 independent variables along with a boolean response variable. Of the 19 independent variables 2 are string variables, 2 are boolean variables and the remaining 15 variables are integers. The Dataset contains no null or missing values. The data can be segmented into customer level data and call level data. The customer level data consists of the state a customer is located in (string), area code of the registered number (string), account length (int), international calling plan (boolean) and a voicemail plan (boolean). The call level data consists of the number of voice-mail messages (int), The total number, minutes and cost of day calls (int), the total number, minute and cost of evening calls (int), the total number, minute and cost of night calls (int), the total number, minute and cost of international calls (int), and finally the number of customer service calls (int).

The main concern we initially had with the dataset was regarding call level data as it was unclear whether the data was cumulative (i.e each additional call is added to total call minutes) or if the data is averaged by a time sequence (i.e total call minutes divided by length of account). To determine what option was correct we binned the account lengths in 10 month intervals and took the average of each interval based on total minutes of day calls. If the data was cumulative we would see a consistent increase as account length increases. If the data was averaged already we would see a relatively flat total minutes of day calls across each account length. As shown in exhibit A, the call level data has already been averaged and as such there was no need to manipulate call level data to account for account length.

ANALYSIS

The dataset selected has a significant problem with the minority class of the response variable. The response variable has a large unbalance between its two classes having only 14% of the observations where the customer churns and the remaining 86% where the customer does not churn. The dataset only contains 4250 observations, so having only 600 observations where the customer churns may cause a challenge for our models to learn their patterns. Using synthetic minority oversampling technique (SMOTE) will minimize the effect of both issues. SMOTE is an algorithm that creates synthetic observations of the minority class (customer will churn) in the dataset. SMOTE works by selecting a random minority observation in the dataset and selecting its k (used k = 5) nearest neighbors in the feature space. SMOTE will then select one of these k nearest neighbors at random and generates a random data point that is along the line connecting the two neighbors. This data point is then added to the dataset as a synthetic observation for the minority class. The synthetic observation is realistic as it is near other minority observations in the feature space. The outcome of using SMOTE on the original dataset resulted in a new dataset with 7304 observations and a balanced class distribution. With a balanced dataset, models can now be trained easier as they can better recognize patterns in the minority observations with a balanced dataset.

The first action performed on the new SMOTE dataset was using AutoML method Tree-Based Pipeline Optimization Tool (TPOT) to find the best tree-based method for the dataset. TPOT works by selecting and processing the most influential variables from the dataset and creating different decision tree models with these variables then using cross validation to pick the best tree. The TPOT algorithm generated a pipeline with Extra-Trees model and its tuned parameters which were generated using Hyper Parameter Optimization (HPO). The Extra-Trees Classifier generates random decision trees using all the variables in the dataset and a sample of the observations and averages the results to find the best trees. The Extra-Trees model with the tuned hyper parameters from TPOT had a test accuracy of 97.3% and a precision of 98.7%. Then to gain insights on the model, used the SHAP package to find the Shapely values for the model. Shapely values are a concept used to explain complex ML models by measuring the impact of each individual variable on the model's prediction. The important variables in the extra-trees model based on the shapely values.

The next action performed on the dataset was using Support Vector Machine (SVM) modeling with a polynomial kernel. SVM was used to determine if there was a clear margin of separation between the classes. The polynomial kernel was selected due to the data set being complex and the ability to have a very flexible boundary. The resulting SVM model had an accuracy of 71.7% and a precision of 86.7%. Based on the results of the model it can be concluded that the data is not separable. Once again, the shapely values were used to determine the important variables for the SVM models predictions. Four of the variables had the most impact to the model unlike the extra trees model which incorporated more variables for each prediction. This may indicate why the extra-trees model performed significantly better than the SVM model.

From the result of the model being non-separable and non-linear it violates the assumptions of logistics regression. This prevented the need to test out the logistic regression model and rather try a new model that had no assumptions and was able to handle complex datasets.

Therefore, the next action performed on the dataset was using Neural Networks. Neural networks make an assumption about the data but the dataset does not violate these assumptions. Neural networks allow for the exploration of the data shape by changing the number of layers in the network and the neurons in each layer. The final network included many layers, including batch normalization layers and dropout layers to avoid overfitting and help backwards propagation. The model was trained using the Adam optimizer and included an early callback to avoid overfitting. The model resulted in a validation accuracy of 95% and a precision of 94.4%.

The next analysis led the way back to decision trees. XGBoost is another ensemble decision trees model that was implemented. It stands for Extreme Gradient boosting and it builds decision trees sequentially. The trees try to predict and correct the errors from the previous decision tree. This algorithm gave some really good results with 95% accuracy and 97% recall. Another boosting algorithm that was used in the analysis was Catboost. Catboost is a boosting algorithm that was specifically designed for categorical variables; highlighted in the name, Categorical boosting. Catboost performed really well on the dataset, it 's performance was on par with that of the Extra Trees classifier. Catboost had an accuracy of 97% and a precision score of 98%.

Measuring the precision of each model allows for minimizing the number or false negatives returned from the selected final model. The precision measures the number of true positives returned compared to the total number of true positives and false negatives returned. Subtracting the precision from 100% measures the number of false negatives returned from the model.

RESULTS

The analysis provided extremely promising results greatly exceeding the expected accuracy rating of 80% in the project proposal. Of the various different models (Extra Trees, SVM, Neural Nets, XGBoost, CATBoost) utilized in the analysis, the extra trees classifier with TPOT tuned hyper parameters demonstrated both the highest accuracy and the highest precision. This model produced an accuracy rating of 97.3% and a precision of 98.7% which can be seen in the ROC in exhibit B. This is similar to the results of CAT boost however due to the Extra Trees classifier being a bagging algorithm it is a better fit for our business problem as bagging algorithms avoid overfitting and are generally less complex. The precision rating of 98.7% indicates that the model is extremely efficient at determining true positives from false positives. Additionally, looking at the confusion matrix seen in exhibit C shows the true negative rate to be 98.5% well within the goals of the analysis to limit the number of false negatives.

OUTCOMES

As stated in our business problem, a subgoal of the analysis was identifying which features cause customers to churn within the US telecommunications industry. For our respective dataset, there were 19 established input features that were evaluated in order to determine their significance and casual relationships. Of these independent variables, the features averaging the most impact on the model output were the if the provider offered an international plan and the number of calls made to customer service, respectively. Following these, total minutes of day calls, total charge of day calls, and the customer having a voicemail plan were the next most significant features (see Exhibit D for the average impact each input feature has). Given the following results and the sufficient accuracy of our model, it is clear that these are the most causal features of customer churn. With relation to real-world applications, those in the US telecommunications industry should strongly consider incorporating these features that have causal significance into their product/service offerings in order to retain customers and avoid costs associated with churn.

RECOMMENDATION

With the conclusion of the analysis there are several recommendations the team has surrounding both future analyses which are likely to provide valuable insights for the organization and the implications of the now trained prediction model.

In regards to the prediction model, our recommendation is to integrate the model into the organization's data pipeline to allow for real time updates on customers likely to churn. This will allow the organization to proactively identify and engage with the customers prior to leaving the organization.

This increases the LTV:CAC ratio and overall profitability as customers proactively retained have a much lower cost of acquisition compared to new customers. Additionally, the organization should develop scaled incentives for customers that are likely to churn based on their expected lifetime value to the organization. This will ensure retention incentives are proportioned to the customer value, once again improving the LTV:CAC ratio for the company. The final step for the organization would be to both implement the information into front line customer service associates dashboards and train them on the retention incentive structure. This will help employees consistently apply the retention techniques in their day to day operations as opposed to having a sole retention specialist responsible for reaching out to customers likely to churn.

The second set of recommendations provides forward guidance into areas of interest for future analytical projects. These areas of interest include the factors which had the greatest impact on the likelihood of customer churn (see Exhibit D). The two most deterministic factors we found when utilizing the extra trees classifier were whether the customer had an international plan and the number of customer service calls they have placed.

The insight that if customers have an international plan they are much more likely to churn provides an opportunity to dive deeper into what exactly is offered in both the organizations international call plans and competitors international call plans. By doing so the organization can determine if they are simply having customers with international calling plans churn due to better offers from competitors which would then require a rework of the international calling plan pricing. If this is not the case, then a deeper analysis surrounding the solely international plan customers could be instigated to determine the root cause whether that be poor call connections or some other factors.

The number of customer service calls also has a large effect on customer churn. This is likely because if a customer is continually calling customer service, it means their main problem is unlikely to have been fixed in previous calls. By reducing the number of customer service calls, the organisation can increase profitability due to both reduced call volumes and lower customer churn. The recommendation to facilitate a reduction in service calls involves creating or implementing a dashboard for service agents capable of tracking if a customer's problem has been resolved and what the problem was in the first place. By implementing this system, the organisation can monitor the types of problems customers have and the type of problems agents are incapable of solving. The organisation would then be able to implement initiatives to reduce these calls by proactively solving the problem. If that is not a feasible solution, then they can sufficiently train agents to solve the types of problems customers typically face that lead them to churn.

EXHIBIT

Exhibit A

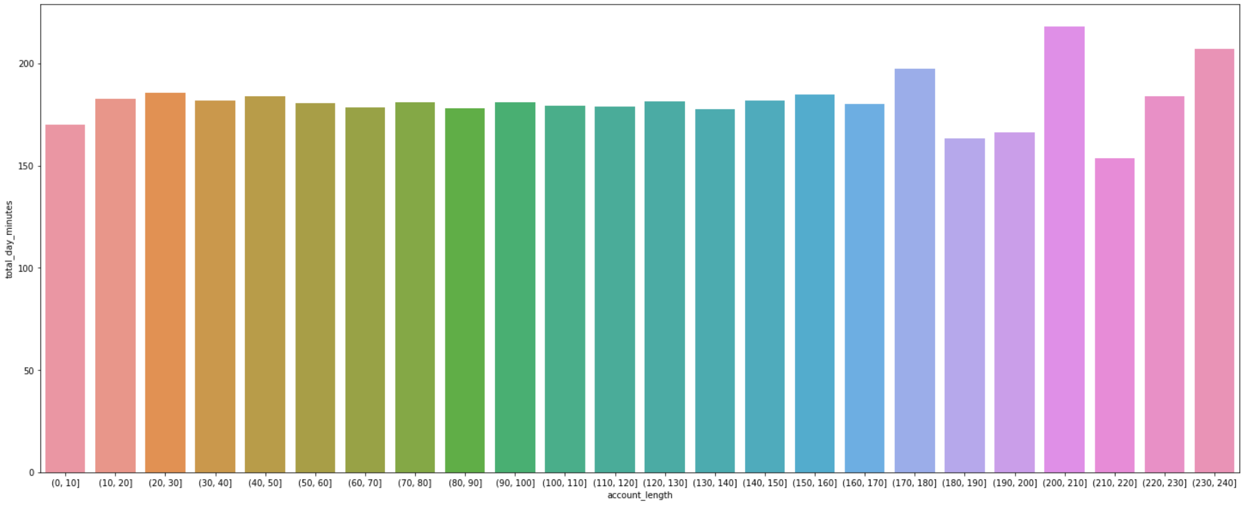


Exhibit B

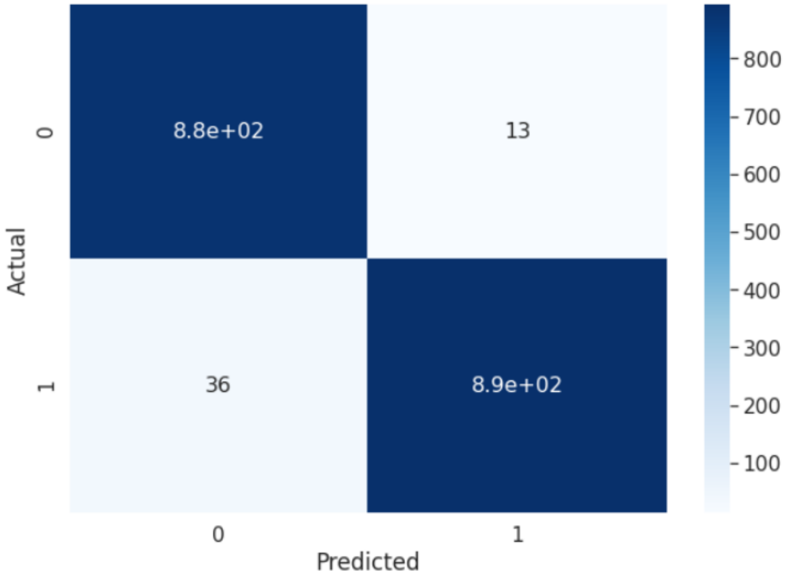


Exhibit C

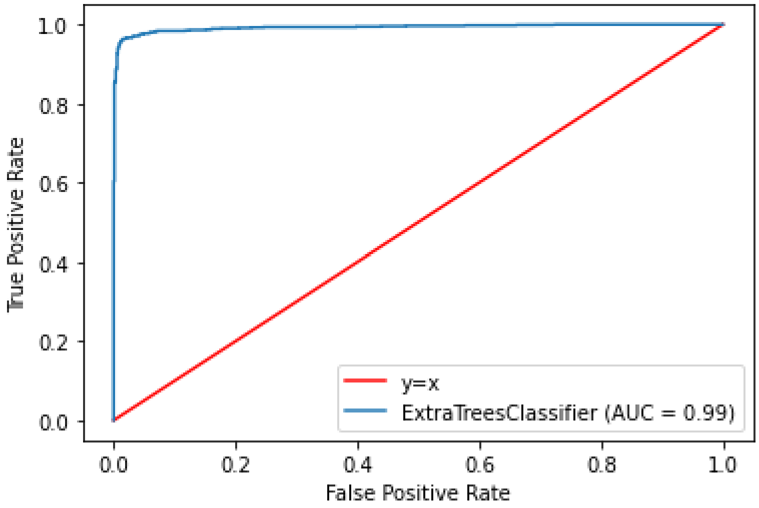
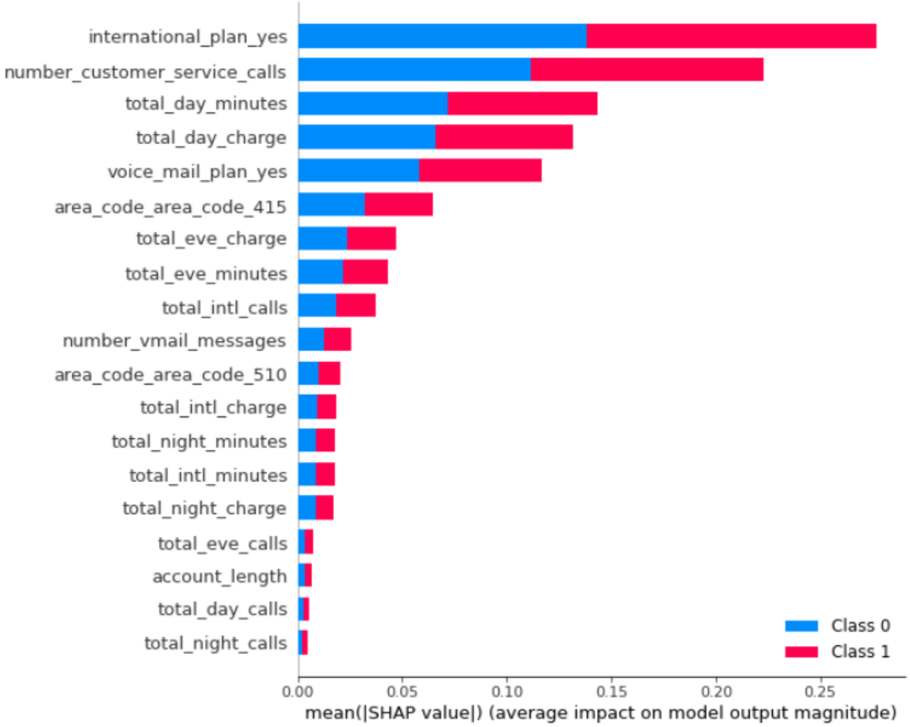


Exhibit D


international_plan_yes
number_customer_service_calls
total_day_minutes
total_day_charge
voice_mail_plan_yes
area_code_area_code_415
total_eve_charge
total_eve_minutes
total_intl_calls
number_vmail_messages
area_code_area_code_510
total_intl_charge
total_night_minutes
total_intl_minutes
total_night_charge
total_eve_calls
account_length
total_day_calls
total_night_calls
Class 0
Class 1
0.00
0.05
0.10
0.15
0.20
0.25
mean(|SHAP value|) (average impact on model output magnitude)