Final Project Executive Summary

Tej Rai

University of Wisconsin La-Crosse

DS 740: Data Mining

August 7th, 2024

Final Project Executive Summary

When you subscribe to a service what makes you leave? Customer churn is critical for many businesses, particularly in subscription-based industries. Understanding and predicting customer churn can help companies develop strategies to retain customers and improve their bottom line. This analysis investigates the key factors contributing to customer churn for a telecom company using two machine learning models: Artificial Neural Network (ANN) and XGBoost. The goal is to identify the most influential churn predictors and assess these models' performance in predicting churn.

The findings from this project could help customer retention teams in telecommunication companies develop targeted strategies for reducing churn. The models will help identify high-risk customers and the factors contributing to their likelihood of leaving. It can also help marketing and strategy teams design personalized marketing campaigns based on customer segmentation. The model's insights can help understand customer preferences and tailor offerings to keep them engaged.

Data Description and Preparation

First, we utilized the Telco Customer Churn data set (via Kaggle), which consists of 7,043 actual observations and 21 variables, including customer demographics, account information, and service usage details. Key steps in data cleaning included:

1. Handling missing values: Rows with missing values in the ‘TotalCharges’ column were removed.
2. Handling unimportant variables: Columns like ‘customerID’ were removed.
3. Encoding Categorical variables: One-hot encoding was applied to categorical variables to prepare the data for machine learning models.

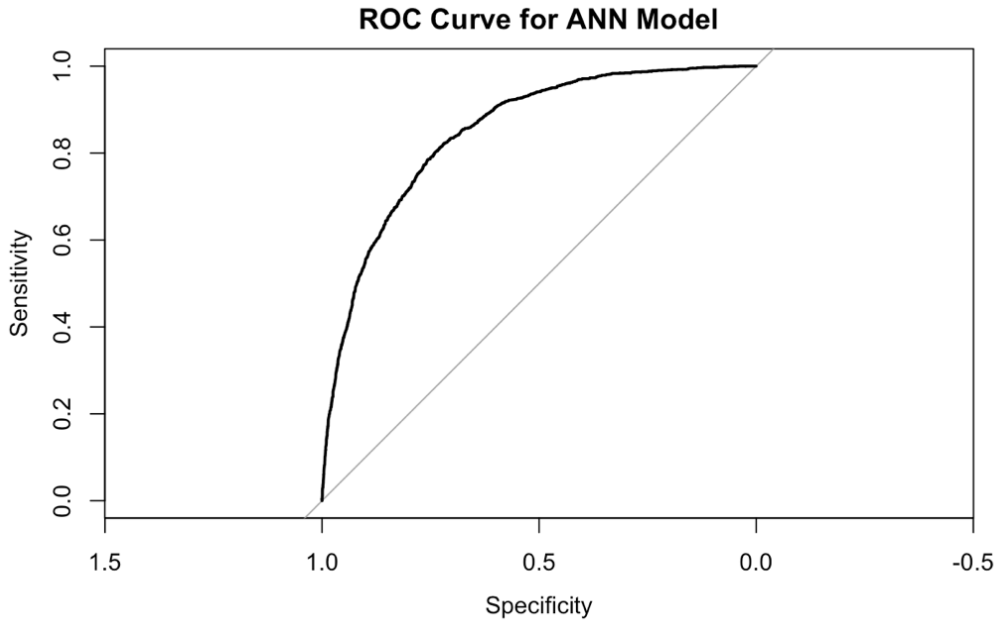
Model Selection and Tuning

After preparing the data, we implemented two supervised learning techniques: ANNs and XGBoost, because they were capable of handling complex non-linear data patterns and work well with non-quantitative binary response variables (in this case Churn as ‘Yes’ or ‘No’). Through cross-validation both models were tuned, and the optimal patterns were found (ANN: size =1, decay = 0.3) (XGBoost: nrounds = 25, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.8, min_child_weight = 1, subsample = 1). Note, in attempt to save time, the models’ tuning parameters were optimized for size, decay, nrounds, and max_depth respectively.

We performed an outer layer of validation to help select the best model also. The resulting confusion matrix demonstrated both models performed similarly well. The XGBoost outputted an accuracy = 80.66%, sensitivity = 90.63%, and specificity = 53.13%, while the ANN outputted an accuracy = 80.53%, sensitivity = 89.72%, and specificity = 55.16%. While the results were very close, the ANN was the chosen model because it had a 2.03% greater specificity, meaning it had a reduced rate of false positives. A false positive in this context occurs when the model predicts that a customer will churn (Yes), but the customer does not churn (No). Since we want customer retention and marketing teams to have accurate and effective targeted campaigns, we don’t want unnecessary retention efforts directed at customers who don’t intend to leave. The ANN final best model’s confusion matrix is visualized below:

	Predicted Outcome		
Actual Outcome		Will not Churn	Will Churn
	Did not Churn	4613 (True Negatives)	550 (False Positives)
	Did Churn	810 (False Negatives)	1059 (True Positives)

We also checked the final model's performance by generating a ROC curve and evaluating its AUC. The ROC curve is visualized below:

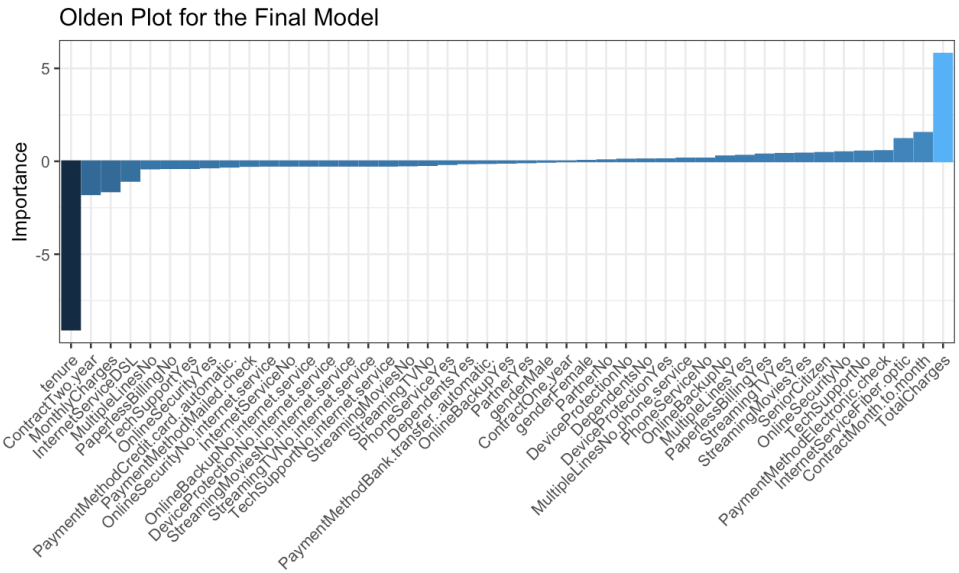


The model had an AUC value of 0.8488, meaning the model provides a reliable basis for making decisions about customer retention efforts. The model can correctly identify a significant

proportion of customers who will churn, which means the model allows the teams to target at-risk customers effectively.

Results and Interpretations

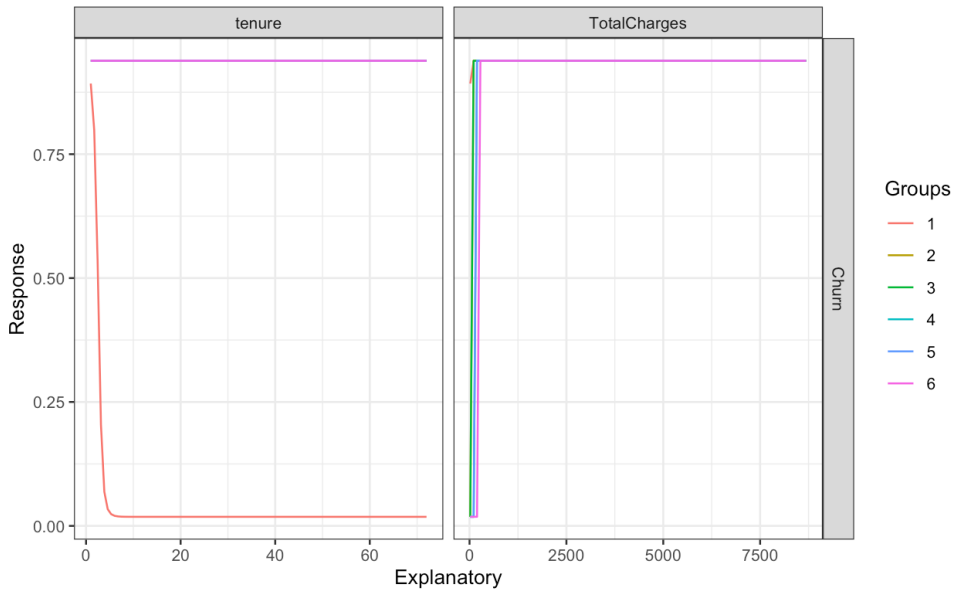
Furthermore, after determining which model was best and whether it was adept or not, we evaluated the important variables in predicting churn and the relationships between churn and the variables. To find the most important variables we created an olden plot to visualize the magnitude of importance a variable had on churn. The olden plot is displayed below.



From the olden plot, we can see the two variables with the greatest magnitude and importance were ‘tenure’ and ‘TotalCharges’. This makes sense because customers who have been with the service for a longer period (higher tenure) are less likely to churn as long-term customers tend to be more loyal and satisfied. In terms of business action, it would make sense to continue to nurture long-term relationships through reward programs to maintain loyalty or consistent engagement, but significant effort does not need to be targeted towards greater tenured customers. On the flip side, it makes sense that customers with higher total charges are more

likely to churn because they may perceive the service as too costly and decide to leave. It would make sense to target these high-charge customers with special offers or loyalty programs.

In addition to the olden plot, we created a lek profile to better visualize the relationship between the important variables and churn. The lek profile is visualized below:



A lek profile helps interpret the model by showing how different levels of each predictor when other variables are held constant affect the predicted outcome and in this case the churn's sensitivity to change in the predictor. As we said before, we can see that as tenure increases there is a decline in the probability of churn and as total charges increase there is a sharp increase in churn probability. Therefore, it would be very helpful for a customer retention team to implement some sort of loyalty or rewards program or give special offer towards customers who have high total charges in order retain them.

Final Thoughts & Recommendations

While our model did show some success and offer insights towards what variables impacted customer churn, there are still improvements that can be done. When the dataset is examined closely, it can be noticed that there is about a 26.54% churn rate in the data set. This

signifies that the data is highly imbalanced which results in a class being underrepresented. Various resampling techniques could be applied to balance the data set to enhance the model training and evaluation, ensuring fair consideration of both churn and non-churn cases.

In addition, the models selected were run on a large data set, and tuning multiple parameters with a typical laptop was difficult due to the time constraint of the project. Tuning parameters on more computing power with more time would be helpful to improve model performance.