GB 656 Final Project Report: Churn Prediction Model

By Team 7: Akshay Saraf, Grant Cook, Huzaifa Ikram & Ori Nozar

Introduction

In this project, we aim to build a classification model that accurately predicts the churn of an individual based on distinctive features and variable information available to us. The model was built using the training dataset provided by the challenge to be evaluated on test data with churn outcomes unavailable to us. Churn is a common business question that many businesses desire to have an effective model to predict customer or employee loyalty or churn based on the data they have access to. Having this model helps businesses allocate resources more efficiently to reduce spending and improve customer satisfaction and customer retention.

Business Framing

Customer churn is a critical issue for telecommunication companies, as it directly impacts revenue and profitability. In a highly competitive market, retaining existing customers is often more cost-effective than acquiring new ones. Churn not only results in the loss of recurring revenue but also incurs additional costs related to marketing and promotional efforts aimed at attracting new customers. Moreover, high churn rates can damage a company's reputation, leading to a negative perception among potential customers. Therefore, understanding and predicting customer churn is essential for telecommunication companies to maintain a stable customer base and ensure long-term business success.

Problem Statement

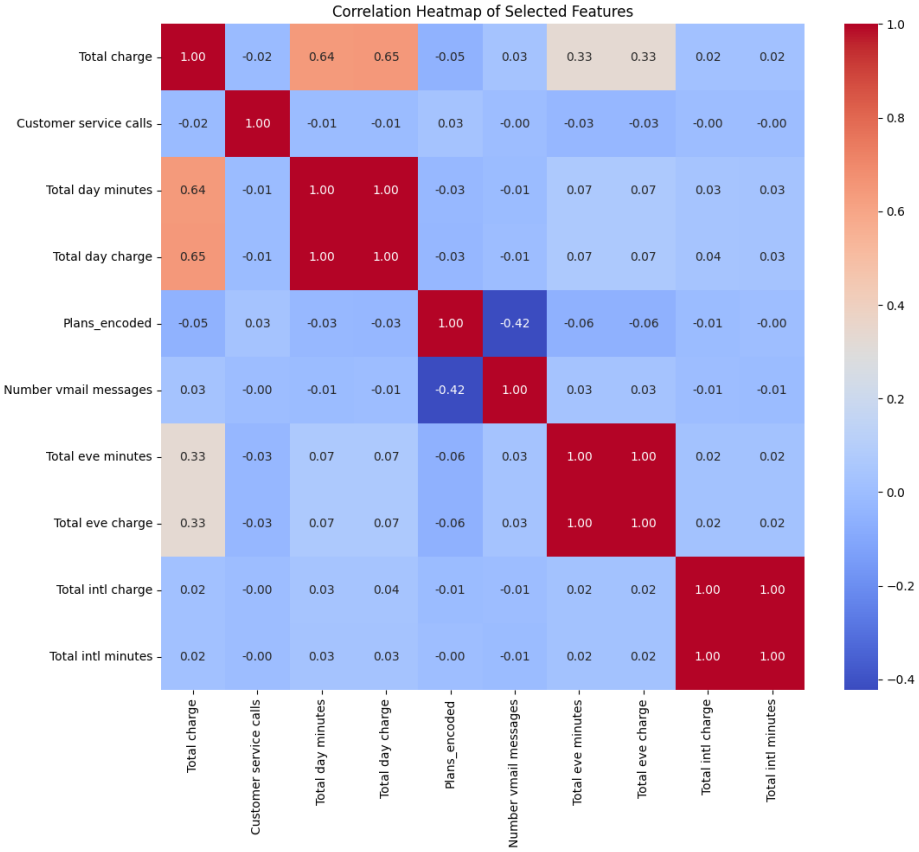
The use case for the churn prediction model is to identify customers who are likely to leave the service in the future. By leveraging the data on customer demographics, usage patterns, and interactions with customer service, the model can provide valuable insights into the factors contributing to churn. The value of this model lies in its ability to enable proactive measures, such as targeted retention campaigns, personalized offers, and improved customer service, to prevent churn. By accurately predicting churn, the company can allocate resources more efficiently, enhance customer satisfaction, and ultimately reduce churn rates. This predictive capability will be instrumental in helping the company retain its customer base, increase customer lifetime value, and achieve a competitive edge in the market.

Data Cleaning

The data we gathered had a mix of objects, strings, and integers. We had to convert all of the data into integers or dummy variables in order to run our classification models. We converted sex and plan into dummy variables. In addition, we converted the null values in the plan column into a value called “no plan”. We also dropped variables that would not provide much insight on a customer's probability to churn like their phone number or state. To better understand how each attribute we collect data on relate to one another we created a correlation matrix. After looking at

the correlation matrix, it was revealed that total eve minutes and total eve charges are very similar variables.

With any classification model, it is important to look at a heatmap of what features are correlated with each other before choosing which features to include. This helps to reduce multicollinearity in a model, as well as overfitting the model to the training data and misleading information about the most important features. We found this to be an issue in our initial model, because we had two features that had a correlation equal to 1.00, which would affect our model if one of the features was not removed. This can be seen in our heatmap below, where "Total eve charge" intersects with "Total eve minutes" and "Total day charge" shows a dark red tint with a 1.00 correlation. To improve the performance and integrity of our model, we removed one of these variables and ran the model again to reduce the probabilities of the issues mentioned above arising in our analysis. We did this with all high correlating features, removing one from the model. With the cleaned data, we will now be able to better predict the probability of a specific customer churning based on their attributes.



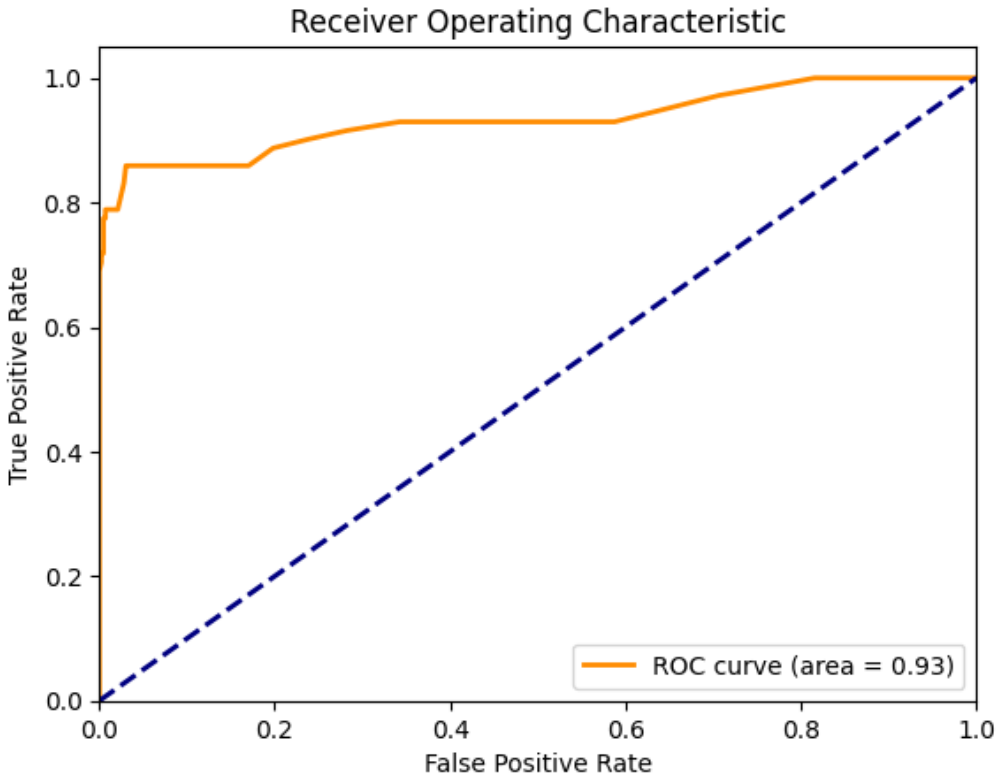
Methods

We decided to use a classification model, specifically a random forest classifier, to predict customer churn. We chose the random forest classifier because it is a decision tree classifier which we thought was best equipped for prediction in this scenario. Using the training set we were able to tune the model's hyper parameters to try and achieve the high AUC and F1 score.

After creating and tuning the model, we uploaded new unseen testing data and had the model predict whether each customer would churn and the specific probability of each customer churning. Both models had very similar results with the same 5 most important features and very similar AUC and F1 scores. With our model's output, we can address the overall concerns our customers have with our service and target individual customers that are highly likely to churn soon. This can allow us to be more efficient when investing in company improvements and marketing to customers.

Model Results

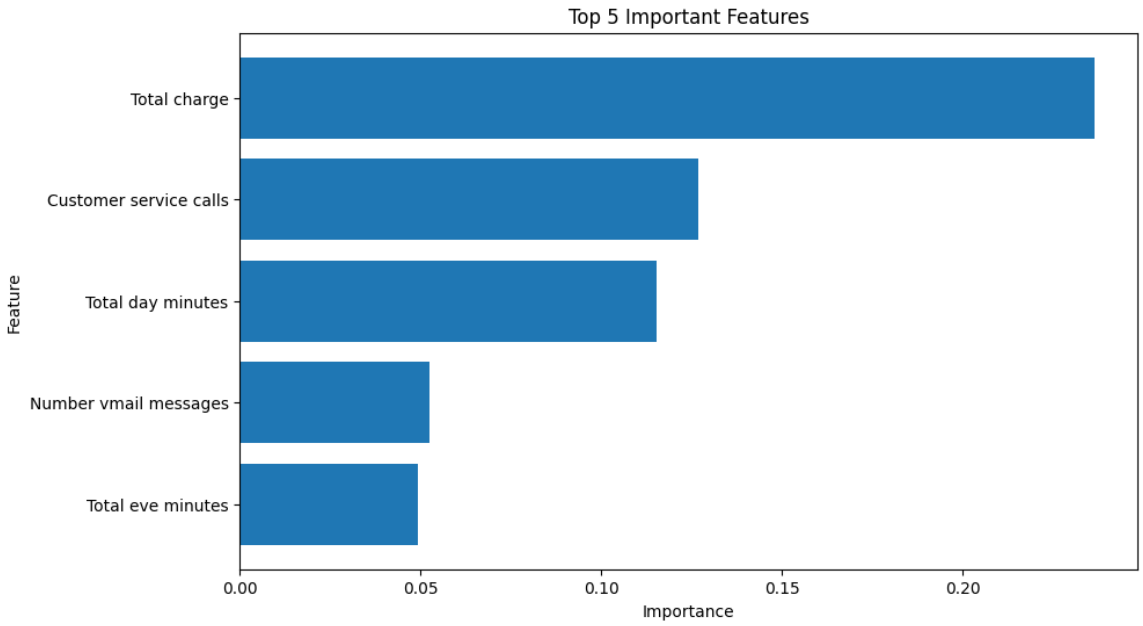
Within our training data, we had two subsets of data: a training and a test set. After training our model using the methods mentioned above, we then tested it on the remaining datapoints in our test set and evaluated the performance. Within our test subset, our model performed well with an **Area Under the Curve (AUC) equal to 0.93**, as seen in the graph below. This outcome metric suggests that the model did very well correctly classifying customers who churned versus those who did not. For reference, complete randomization would have an AUC equal to 0.5 and what is generally considered to be an effective model typically are those with AUC values greater than 0.7. After applying the model to the larger test dataset for the project, the resulting AUC equaled **0.918**, suggesting that the model's performance remains strong when new data is introduced.



More Performance Metrics for the Random Forest Classifier:

Metric	Class 0	Class 1	Accuracy	Macro Avg	Weighted Avg
Precision	0.95	1.00	0.96	0.98	0.96
Recall	1.00	0.69	0.96	0.85	0.96
F1-Score	0.97	0.82	0.96	0.90	0.95
Support	429	71	500	500	500

To support our analysis, we identified the most important features in calculating customer churn probabilities, as seen below. These features are key for driving business decisions to reduce customer churn and improve retention because it highlights important aspects of customer experiences through the scope of our data, allowing them to allocate resources or alter their business strategies based on the areas that impact churn the most.



Recommendations

Based on these features, we can determine three main categories of reasons why customers churn. The cost is the most influential factor, and a solution would be to try and increase automation and optimization for cost savings to pass on to the customer. In addition, since switching services is a hassle for customers, we can offer a price match to incentivize customers to stay with us instead of switching to a competitor. The second reason is related to our customer service and specifically our call system. If we can improve on the experience by offering better training for our employees, automated chat bots, and display resources more transparently we can reduce the time customers spend on calls and boost increase their satisfaction with our service. The third most influential reason customers churn is related to usage. Although we have the least control over their usage, a beneficial initiative would be to include a “pay as you use it”

plan and to run ads informing customers of the value our service provides and why it is important.

Conclusion

In conclusion, our churn prediction model demonstrates significant potential in identifying at-risk customers, enabling proactive retention strategies. By leveraging key features and optimizing model performance, we achieved a high AUC score, indicating robust predictive accuracy. This model empowers businesses to allocate resources efficiently, enhance customer satisfaction and reduce churn rates. Implementing targeted retention campaigns and personalized offers based on our model's insights will drive long-term customer loyalty and business success. Our approach not only addresses current challenges but also sets a foundation for continuous improvement in customer retention strategies.

Kaggle Case Competition Link:

https://www.kaggle.com/competitions/churn-prediction-2024/submissions#