Predictive Analytics for Customer Churn

(Applied machine learning Project)

Valerie Garcia (vlg170000@utdallas.edu)

Patrick Gervadis Ninan (pgn230000@utdallas.edu)

Albin Poulose (axp220239@utdallas.edu)

Srivani Kakumani (sxk230164@utdallas.edu)

Ashiq Mohammed Al Ameen (axa220171@utdallas.edu)

Executive Summary

In this project we attempt to analyze customer data to predict whether or not the customer to this video streaming service will churn. We use a logistic regression and decision tree model for their enhanced interpretability because our final goal is to provide suggestions as to how the company can decrease churn. Since we care most about churning customers vs loyal customers we increased the sensitivity of our models by utilizing oversampling in our training data. Ultimately we found that younger accounts with lower engagement and higher monthly charges are those that are most likely to churn.

Introduction

This project aims to predict customer churn in a subscription-based service using anonymized data. The model seeks to identify customers at risk of canceling their subscriptions by analyzing subscription types, payment methods, and customer interactions. Companies are motivated to retain revenue, enhance customer satisfaction, and ensure business sustainability. Additionally, predicting churn provides a competitive advantage and optimizes resource allocation. Ultimately, proactive churn management enables targeted retention strategies and fosters stronger customer relationships.

What is Customer Churn?

Customer churn refers to the phenomenon where customers discontinue their relationship or subscription with a company or service provider. In this case we will analyze whether or not a customer churned from a video streaming service. Churn is an important metric for businesses as it directly impacts revenue, growth, and customer retention.

Dataset Description

We utilized the Kaggle Dataset Predictive Analytics for Customer Churn: Dataset (kaggle.com). This dataset contains 24,3787 anonymized customer records regarding whether or not a customer churned. It also includes a "test" dataset that contains all the features but no target "churn" column and so was not used for this project. The data is clean with no missing values or duplicate records.

Target

Churn: 1/0 corresponding to yes/no, whether the customer churned

Numeric Features

AccountAge: Age of the customer's subscription account (in months)

MonthlyCharges: Monthly subscription charges

TotalCharges: Total charges incurred by the customer

ViewingHoursPerWeek: Average number of viewing hours per week

SupportTicketsPerMonth: Number of customer support tickets raised per month

AverageViewingDuration: Average duration of each viewing session

ContentDownloadsPerMonth: Number of content downloads per month

UserRating: Customer satisfaction rating (1 to 5)

Categorical Features

SubscriptionType: Type of subscription plan chosen by the customer (e.g., Basic, Premium, Deluxe)

PaymentMethod: Method used for payment (e.g., Credit Card, Electronic Check, PayPal)

PaperlessBilling: Whether the customer uses paperless billing (Yes/No)

ContentType: Type of content accessed by the customer (e.g., Movies, TV Shows, Documentaries)

MultiDeviceAccess: Whether the customer has access on multiple devices (Yes/No)

DeviceRegistered: Device registered by the customer (e.g., Smartphone, Smart TV, Laptop)

GenrePreference: Genre preference of the customer (e.g., Action, Drama, Comedy)

Gender: Gender of the customer (Male/Female)

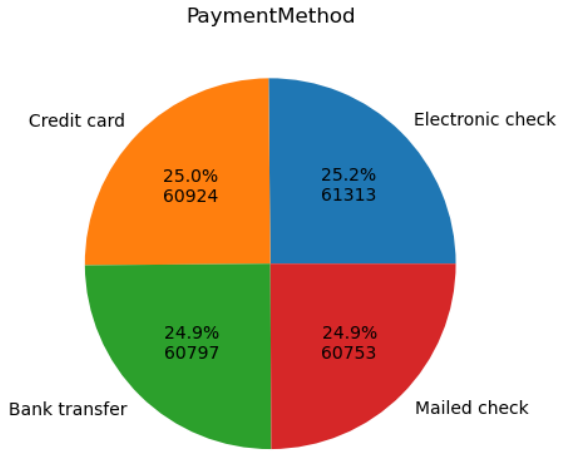
ParentalControl: Whether parental control is enabled (Yes/No)

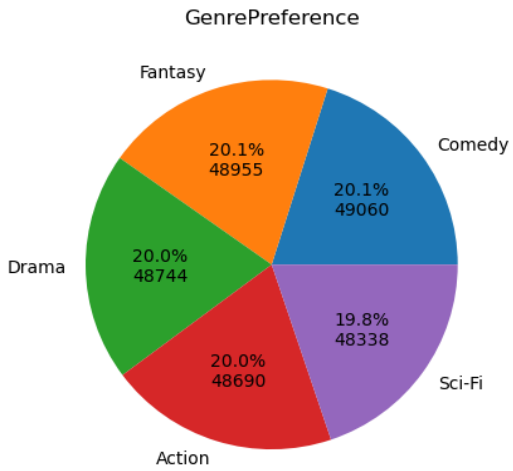
SubtitlesEnabled: Whether subtitles are enabled (Yes/No)

Data Exploration

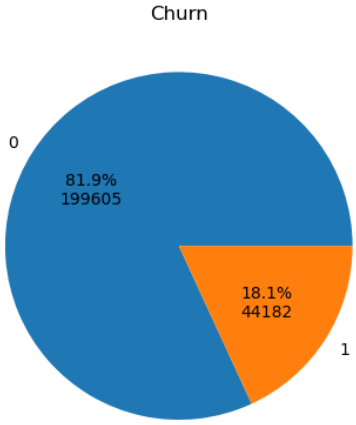
We did an initial analysis of the data to get an idea of any underlying trends.

First we analyzed the categorical features: SubscriptionType, PaymentMethod, PaperlessBilling, ContentType, MultiDeviceAccess, DeviceRegistered, GenrePreference, Gender, ParentalControl, and SubtitlesEnabled. All the categorical features were equally distributed, som examples below:

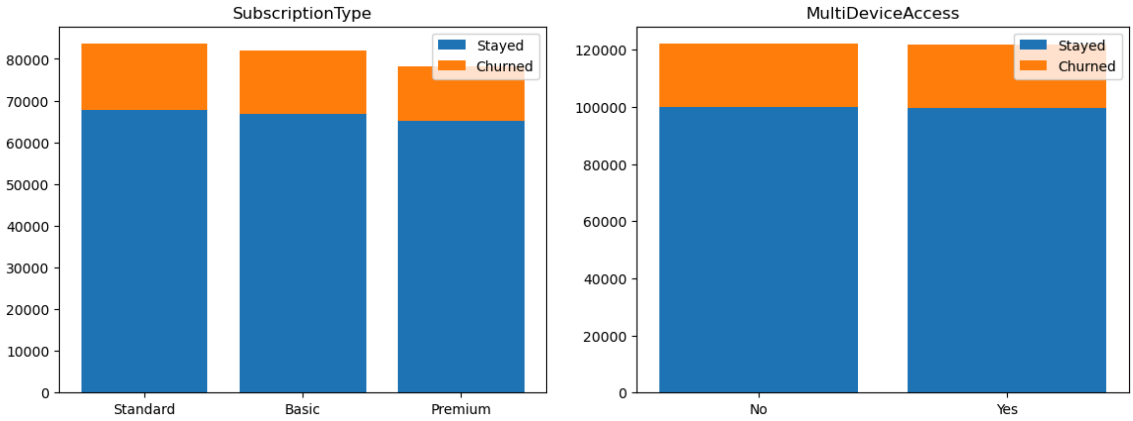




The exception was our target variable “Churn” which demonstrated that 18% of customers churned.

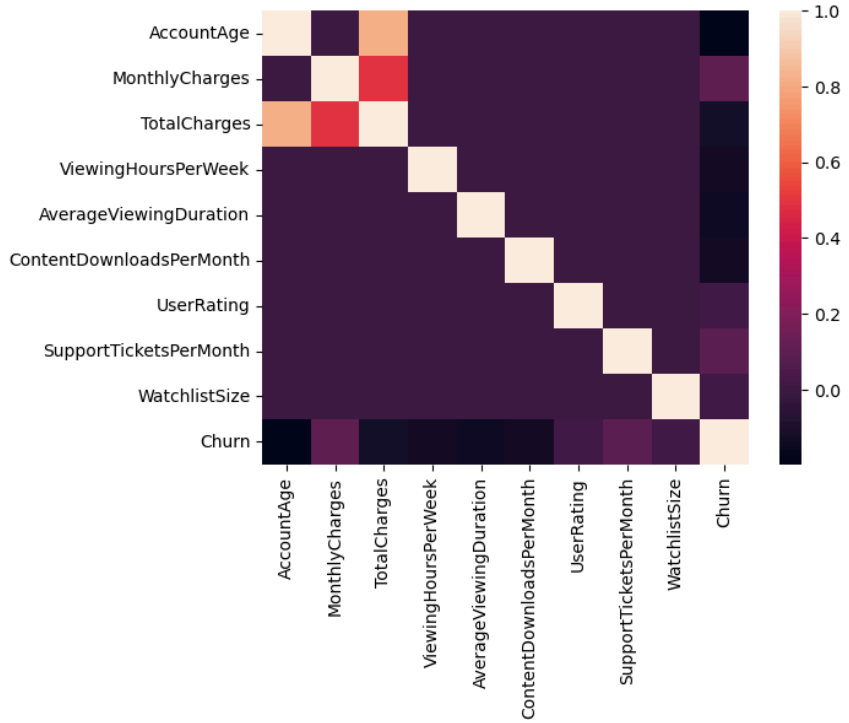


We analyzed whether any categorical feature corresponded with customer churn, with none standing out. Most showed that churn was evenly distributed amongst categories.

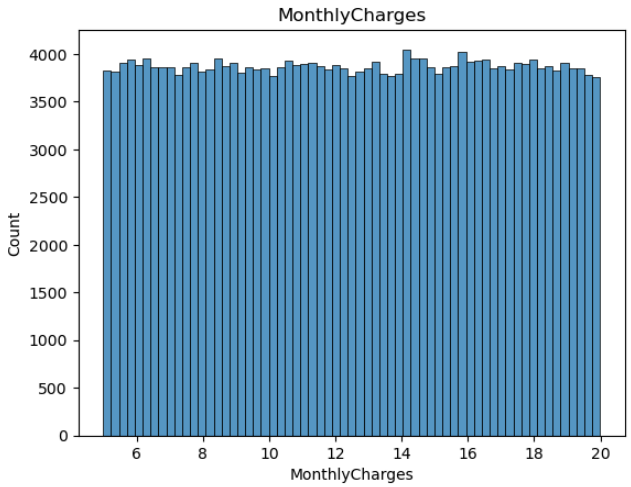


We then analyzed our numerical features: AccountAge, MonthlyCharges, TotalCharges, ViewingHoursPerWeek, AverageViewingDuration, ContentDownloadsPerMonth, UserRating, SupportTicketsPerMonth, and WatchlistSize. An initial correlation heatmap revealed that most features are uncorrelated. AccountAge and MonthlyCharges correlated positively with TotalCharges which makes sense because an older account has more time to accrue charges and an account with a higher amount of MonthlyCharges is more likely to have a higher TotalCharge amount. Furthermore AccountAge, TotalCharges, ViewingHoursPerWeek, AverageViewingDuration, and ContentDownloadsPerMonth were

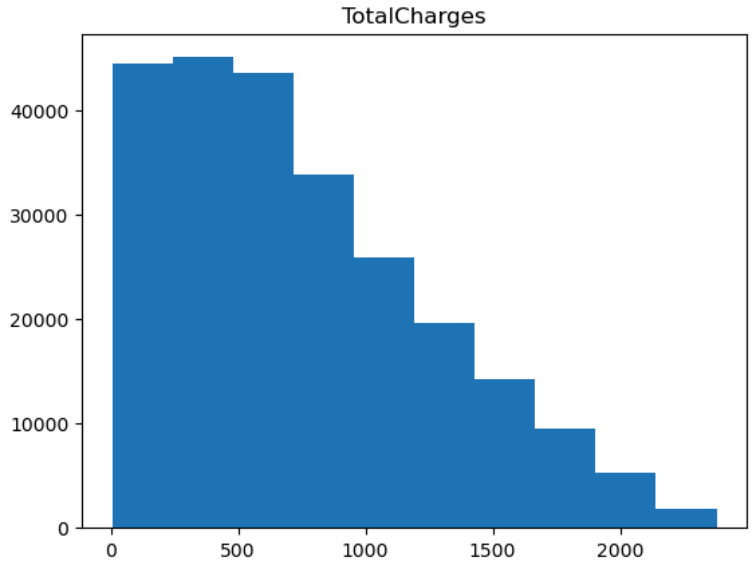
negatively correlated with Churn.



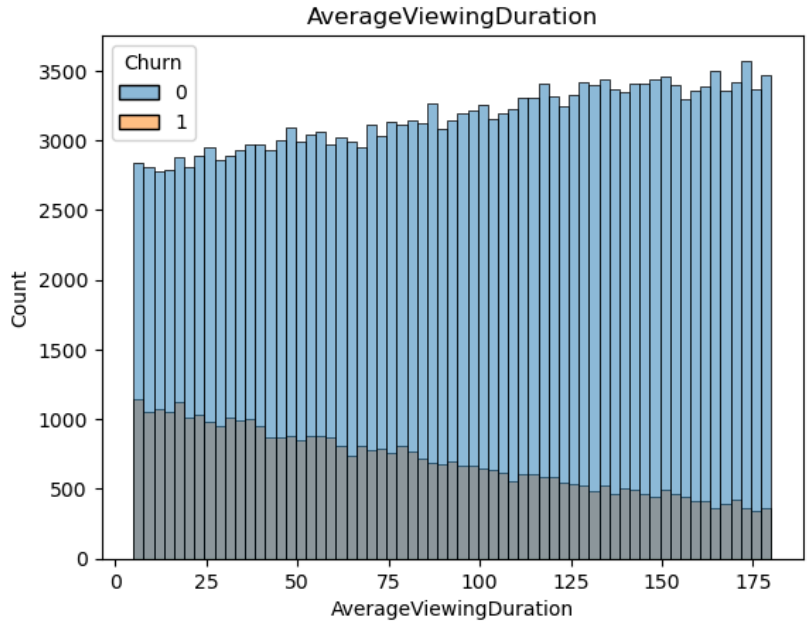
Much like the categorical features, our numerical features were uniformly distributed across their ranges:



With the exception of TotalCharges which is right skewed, with a skew of 0.69. We attempted a log transformation to normalize the distribution but the resulting distribution had a skew of -1.27, so we abandoned the attempt.



We created histograms for the numerical features that layered the histograms for the churned versus unchurned customers. This revealed promising negative correlations for AccountAge, TotalCharges, ViewingHoursPerWeek, AverageViewingDuration, and ContentDownloadsPerMonth. This confirms what we saw in the correlation matrix. We also saw a positive correlation with MonthlyCharges and SupportTicketsPerMonth.



Modeling

We chose to analyze customer churn using logistic regression and a decision tree. We chose their models for their interpretability as the final goal is not just classification but to inform the company on how to decrease customer churn. We divided our dataset into 80% training data and 20% test data. We one-hot encoded our nonbinary categorical variables, dropping one column for logistic regression to avoid the dummy variable trap.

Logistic Regression

Our initial logistic regression model used a min-max scaler on numerical features so that they were in the range 0-1, the same as our categorical features. This was done to increase the interpretability of the resulting coefficients. Here are the top coefficients of the resulting model sorted by absolute value. Coefficients below a coefficient of .3 were not included for length and because they were not deemed relevant enough.

Feature Name	**Coefficient**
AccountAge	-2.049127
AverageViewingDuration	-1.525997
ViewingHoursPerWeek	-1.314889
ContentDownloadsPerMonth	-1.310734
MonthlyCharges	1.014004
SupportTicketsPerMonth	0.784656

From model coefficients we can interpret that the most important feature is AccountAge which indicates that younger accounts are more likely to churn. AverageViewingDuration, ViewingHoursPerWeek and ContentDownloadsPerMonth were next important, all indicating that lower engagement indicates a customer is more likely to churn. Higher MonthlyCharges and SupportTicketsPerMonth also increased a customer's likelihood to churn.

This model had an accuracy of .83, a precision of .57, and a sensitivity of .12 on the holdout test set. In order to increase the sensitivity of our model we oversampled the training data so that it has a 50/50 ratio of churned to not-churned records. This resulted in a new model that had an accuracy of .68, a precision of .32, and a sensitivity of .70. Coefficients for the top features are listed below.

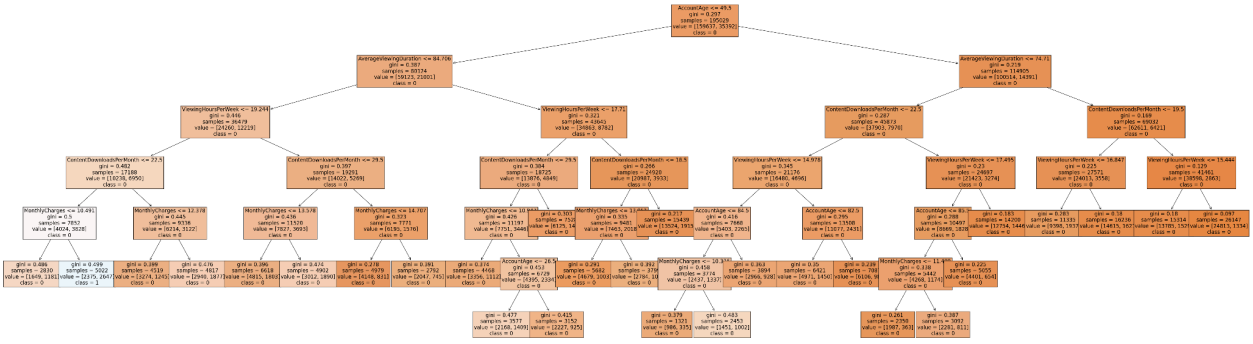
Feature Name	**Coefficient**
AccountAge	-2.034687
AverageViewingDuration	-1.537041
ViewingHoursPerWeek	-1.321582
ContentDownloadsPerMonth	-1.320304
MonthlyCharges	1.049503
SupportTicketsPerMonth	0.795614

The most important features of the model remain remarkably consistent even with oversampling which strengthens our conviction in which features are most important.

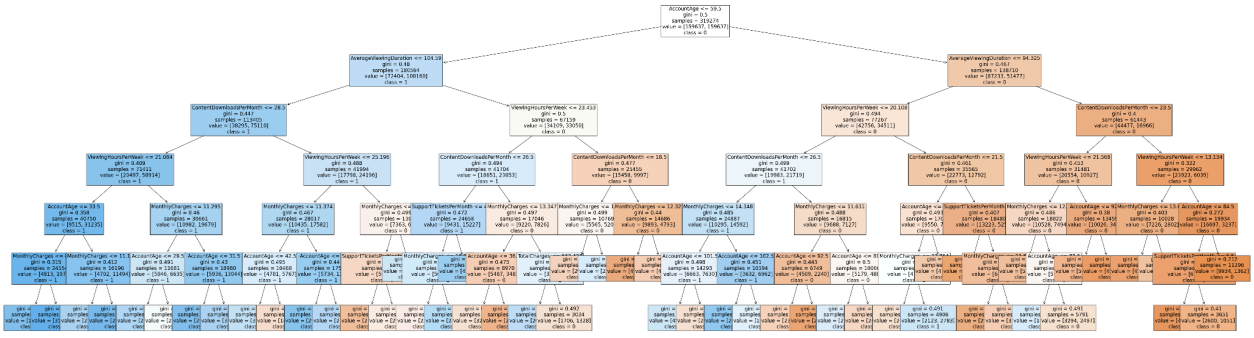
Decision Tree

We created a decision tree using the CART algorithm with gini impurity. We used a grid search to find the best hyperparameters for early stopping. The resulting tree had an accuracy of .82 on the test data. The precision was .53 and sensitivity was .07. This is a disappointing result as this model doesn't perform much better than a baseline model that classifies every customer as not-churned. In fact we can see that every leaf node except one classifies the input as not-churned. The one churn classified node has the rule: AccountAge <= 49.5, AverageViewingDuration <= 84.706, ViewingHoursPerWeek <= 19.244, ContentDownloadsPerMonth <= 22, MonthlyCharges > 10.491. This indicates to us that young accounts with low engagement (viewing duration, viewing hours per week, and content downloads per month)

and high monthly charges are the most likely to churn.

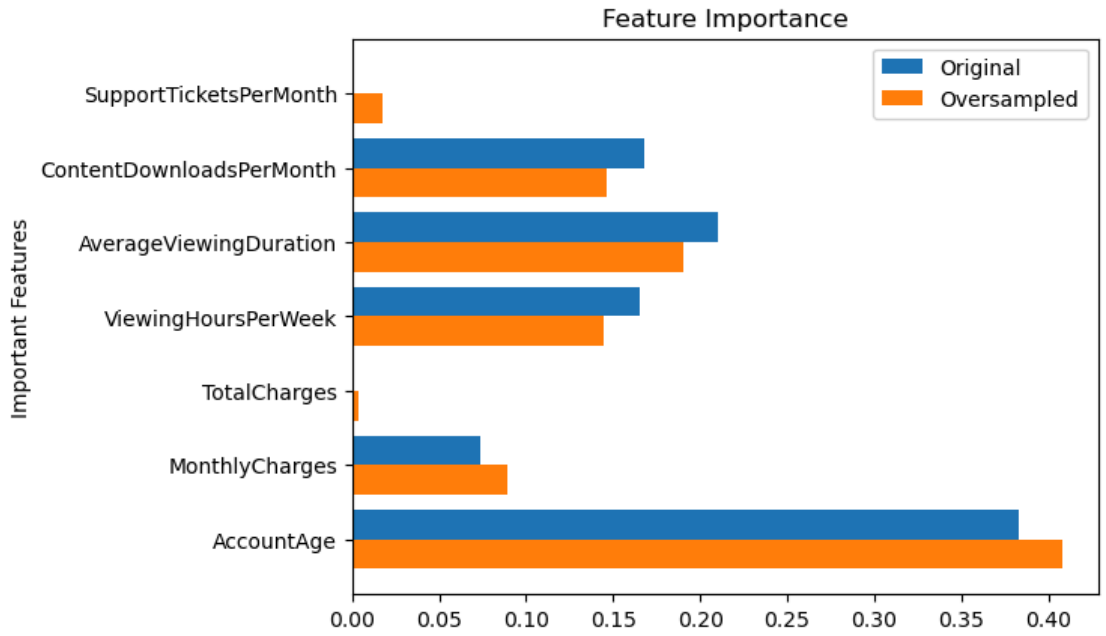


In order to increase the sensitivity of our model we oversampled the training data so that it has a 50/50 ratio of churned to not-churned records. We repeated the process of building a CART tree, using a grid search to again determine best early stopping parameters. The resulting tree had an accuracy of .64, precision of .29, and sensitivity of .69. Though the tree is less accurate we prefer it over the previous tree because of its higher sensitivity, which indicates how well it predicts churned cases, the class we are most concerned with.



When we look at the feature importance of both the original and over-sampled trees we see similar trends. AccountAge is the most important feature indicating new accounts are most likely to churn. Then AverageViewingDuration, ContentDownloadsPerMonth, ViewingHoursPerWeek, and Monthly Charges are the most important features. The oversampled model also includes SupportTicketsPerMonth and

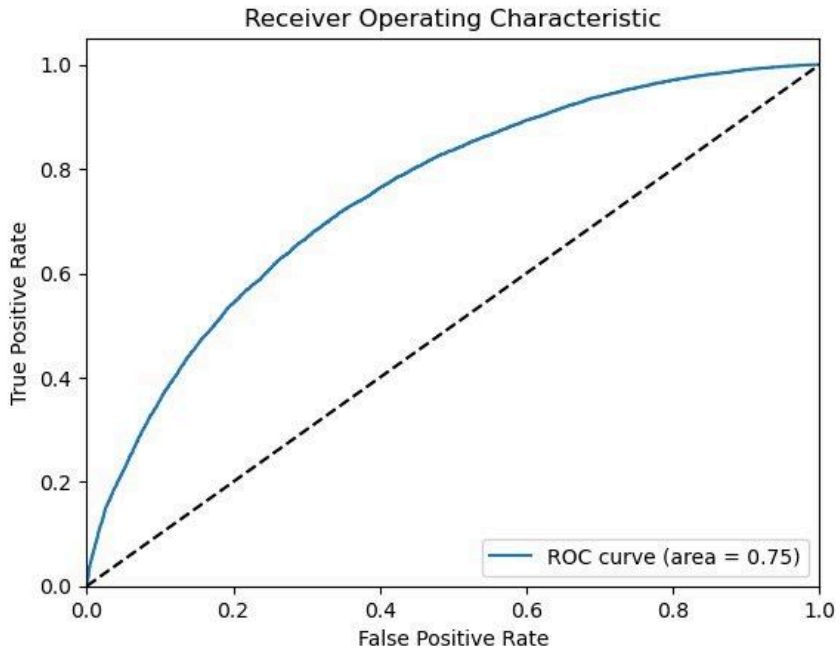
TotalCharges as split criteria, both with positive correlation.



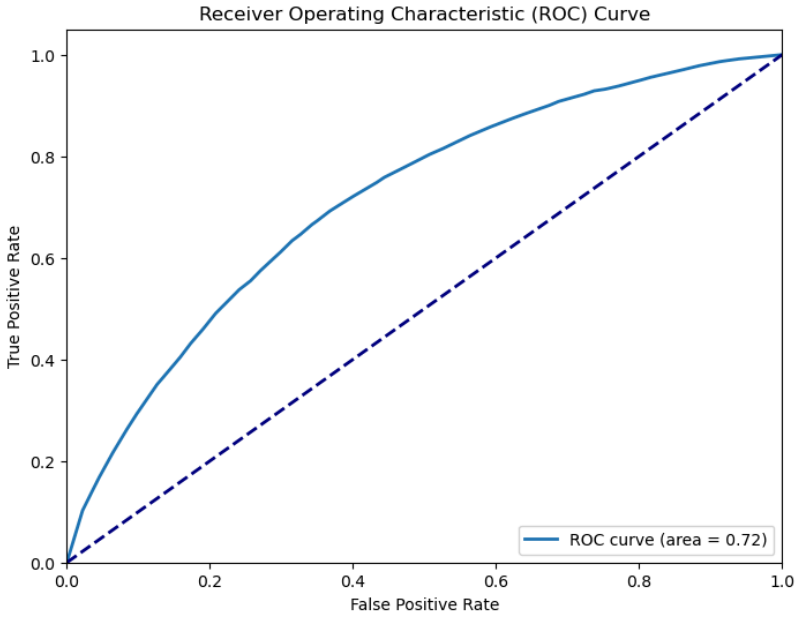
Results and Discussion

We prefer the oversampled models over the models built on original data because we want to prioritize sensitivity to churned cases over just accuracy. Of the oversampled models we prefer the logistic regression model both for its better accuracy/sensitivity and improved interpretability. The oversampled logistic model had an accuracy of .68, a precision of .32, and a sensitivity of .70. Versus the oversampled decision tree with an accuracy of .64, precision of .29, and sensitivity of .69.

This is the ROC Curve for the oversampled logistic regression model.



This is the ROC Curve for the oversampled decision tree model.



The logistic regression model has a higher area under the curve (AUC) at .75 to the decision tree .72. This strengthens our conclusion that the logistic model has greater predictive value.

Conclusion

When we analyze all our models we can see that the customers most likely to churn are newer members (low AccountAge) with low amounts of engagement (AverageViewingDuration, ContentDownloadsPerMonth, and ViewingHoursPerWeek), and high monthly charges. This leads us to the conclusion that the way to reduce customer churn is to target newer subscribers with features that increase engagement and provide discounts for monthly charges. In this analysis we focused on building models with high interpretability: logistic regression and decision trees. In future analyses we might focus on building more accurate but less interpretable models using neural networks, support vector machines, or ensemble models. Hopefully these models will be more accurate and can be used regardless of their interpretability to predict whether a customer is likely to churn and should be targeted with anti-churn measures.