Homework 2

Ryan Jewik

Introduction

My model is predicting whether or not a customer to a streaming subscription service is at risk of cancelling their subscription. It uses predictors gender, age, income, months subscribed, the users' plan, the mean hours watched, if they are subscribed to competitors, their favorite two genres, the number of profiles they have, if they have cancelled or downgraded before, if they have kids or a bundle, and their longest viewing session. If we are able to understand which customers are at a high risk of cancelling, we can focus on those customers more (like giving better recommendations) in order to retain their subscription, and we can also look at how each variable effects their cancellation. Understanding the impact of each variable can tell a company what's more important for subscription retention. For example if kids show a strong coorelation with subscription retention, you might want to add more kids shows.

Methods

We began with testing two types of models. The first being a logistic regression model and the second being a gradient boosting tree.

To prepare our models all that was necessary was creating dummy variables for our categorical variables, plan, topgenre, secondgenre, and gender. Afterwards we performed a train test split and predicted on our models.

Our plan is to choose one of these models to use to get a subset of users who are high risk, then recommend them content they might like using a k-nearest-neighbors model.

We measured the performance of each model using accuracy, precision, recall, and ROC AUC. Both models performed almost identically with the gradient boosting tree performing neglibigly better. With those results we proceeded with the gradient boosting tree.

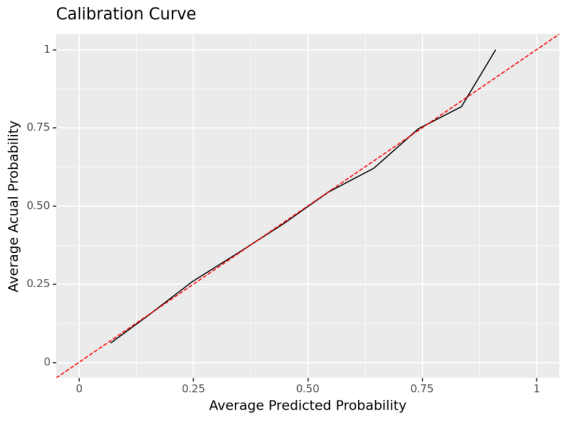
We then used that model and tested it on a new data set, and took the 200 most at-risk users to test a recommendation system on them. Using those users' data, we fit a knearest neighbors model, to group together users with similar characteristics. The characteristics we measured

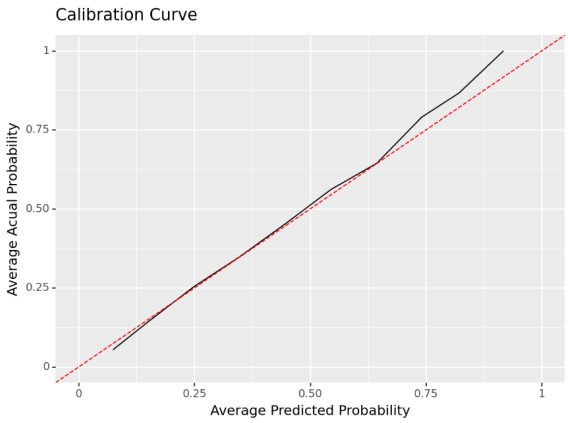
in this model included income, age, and the average hours watched. We grouped users into 10 groups, and added those groups as a column in our data frame.

Results

The two models performed almost identically, which made the choice difficult. A gradient boosting tree runs much slower as it continuously builds upon itself and the collection of other trees its tests on. This results in strong predictions particularly with lots of data. This is the reason why I chose the gradient boosting tree over the logistic regression.

Both of our models were calibrated very well. Both models would be used very similarly. As we had shown with the recommendations, we can target users who are at a higher risk of cancelling their subscription, or as an alternative we can analyze the predictors to see what to focus on. An example of that would be if you see that lots of women are at a high risk of cancelling, you might want to cater more shows or recommendations for women to hopefully retain their subscription.





Discussion/Reflection

A few sentences about what you learned from performing these analyses, and at least one suggestion for what you'd add or do differently if you were to perform this analysis again in the future.

This analysis displays the relationship between a customer and their subscription retention. A subscription service might adjust their marketing and/or their recommendations based on which customers are likely going to cancel their subscription and their attributes. I was able to find that both the logistic regression and gradient boosting tree performed really well in this prediction task, but I still think there was room for improvement. While the gradient boosting

tree performed well, it was very slow likely due to the large number of attributes and number of users. It might be worth looking more closely at the coefficients of each of the indepdent variables and possibly removing any that aren't very impactful on the prediction. This might reduce our runtime to something more favorable.