Homework 2

Graedon Beeler

Introduction

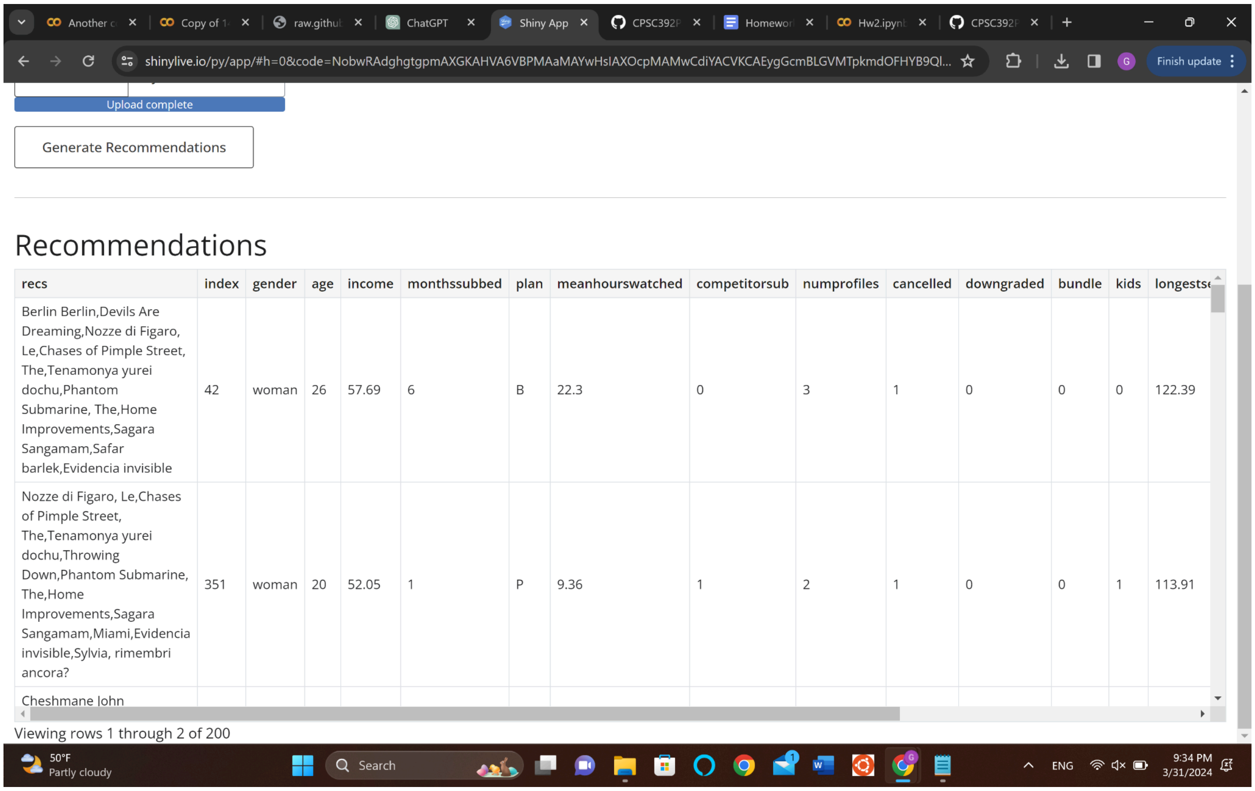
In this project, the goal is to predict the probability of a customer leaving a television streaming service given multiple variables, and then suggest films they might like based on what similar profiles liked, to convince the user to keep their subscription. The variables used to predict the probability of “churning” are gender, age, income, months subscribed, subscription plan, mean hours watched, subscription to a competitor, top genre, secondary genre, number of profiles, past cancellation, downgraded in the past, bundle, has kids, longest session in minutes. If this model is useful, it would be an extremely useful marketing tool to personalize the experience of the streaming platform for each user.

Methods

The models used in this project are Logistic Regression, Gradient Boosting Tree, and K nearest neighbor. Both Logistic Regression and Gradient Boosting Tree models were trained and tested to evaluate which was better at correctly guessing true positives and negatives for the churn value. To process this data, numerical data was normalized using a z-score, and categorical variables were normalized using one-hot encoding. After the models were tested, and the optimal choice was selected, a set of 200 of the most likely customers to churn were paired with customers of similar demographics/preferences using the KKN model.

Results

The Logistic Regression and Gradient Boosting tree compared vary similarly in terms of most performance metrics, with ROC/AUC values both in the 0.73-0.74 range. They were also very accurate in terms of calibration, although the Gradient Boosting Tree seemed slightly more calibrated. However, there seems to be a big discrepancy between time complexity, as the Logistic Regression would complete in around 1-5 seconds, while it would take upwards of 30-40+ seconds for the Gradient Boosting Tree to finish. I would recommend using the Logistic Regression model, because expensive calculations like a Gradient Boosting Tree will slow down your program, and users do not want to wait to see recommendations. For the program that suggests films for users, I suggest taking those films and making a “Picked for you” section at the top of the app with these suggestions.



Recommendations

recs	index	gender	age	income	monthssubbed	plan	meanhourswatched	competitorsub	numprofiles	cancelled	downgraded	bundle	kids	longests
Berlin Berlin,Devils Are Dreaming,Nozze di Figaro, Le,Chases of Pimple Street, The,Tenamonya yurei dochu,Phantom Submarine, The,Home Improvements,Sagara Sangamam,Safar barlek,Evidencia invisible	42	woman	26	57.69	6	B	22.3	0	3	1	0	0	0	122.39
Nozze di Figaro, Le,Chases of Pimple Street, The,Tenamonya yurei dochu,Throwing Down,Phantom Submarine, The,Home Improvements,Sagara Sangamam,Miami,Evidencia invisible,Sylvia, rimembri ancora?	351	woman	20	52.05	1	P	9.36	1	2	1	0	0	1	113.91
Cheshmane John														

Viewing rows 1 through 2 of 200

Discussion/Reflection

From performing these analyses, I learned more about how Logistic Regression and Gradient Boosting trees worked in comparison to each other. I also learned about the K-nearest neighbors model and a practical application for it. If I were to perform this analysis again in the future, I would focus on understanding the big picture of what every piece of code in this project does, as a lot of it makes sense, but some that was reused from in class code I might not have a full grasp on.