Homework 2

Tiffany Le

Introduction

Companies, especially in the streaming services industry, actively pursue strategies to boost revenue by offering services aligned with media trends and consumer profiles and enhancing customer experience and retention. Customer churn refers to customers leaving or opting out of the streaming service, which can result in the company's revenue loss. This data analysis aims to assist a streaming service company in predicting whether customers will "churn" from the service. To achieve this goal, a dataset, provided by the Streaming Service's company, on their customers provides key independent variables to build two different predictive models and one recommendation system. These key independent variables are the customers':

- **gender** Self-disclosed gender identity (Male, Female, Nonbinary, or other).
- **age** age in years
- **income** Self-reported annual income in thousands
- **monthssubbed** Months subscribed to the service
- **plan** Months subscribed to, `P` for premium, `A` for ad-free, `B` for basic (with adds)
- **meanhourswatched** Mean hours of content watched per month
- **competitorsub** Whether or not the customer is subscribed to your competitor's streaming service, `0` for no, `1` for yes
- **topgenre** Most common genre of content the user watches, includes many categories.
- **secondgenre** Second most common genre of content the user watches, includes many categories
- **numprofiles** Number of user profiles associated with the account
- **cancelled** Whether or not the user has canceled the service in the past, `0` for no, `1` for yes

- **downgraded** Whether or not the user has downgraded the service at some point in the past (Premium is the highest plan, then Ad Free, then Basic), `0` for no, `1` for yes
- **bundle** Whether or not the user purchased their plan as a "bundle" with another service, `0` for no, `1` for yes
- **kids** Whether or not the user has an active Kids profile on their account, 0 for no, 1 for yes
- **longestsession** The length of the longest watch session from the user, in minutes

Introduction cont.

These predictive models can be incredibly valuable and helpful if successful. By accurately predicting and identifying if a customer is at risk of churn, the company can prevent the risk by implementing personalized content recommendations or marketing strategies. Furthermore, these models can not only assist with a higher rate of customer retention and satisfaction, but they can also lead customers to share the benefits of the service with their friends, family, and peers, leading to company growth and its customer population.

Methods

The data analysis of the business churn dataset was independently completed by two different predictive models: the Logistic Regression Model and the Gradient Boosting Tree Model. After completing the two models, a recommendation system was completed to assist in churning mitigation and prevention.

Data Preparation

Before fitting the independent variable data to the models, the data was prepared through multiple steps. First, the "null" or empty values were dropped from the dataset and its indices were reset. Afterward, the data was split into two lists: `x` (independent variables) and `y` (whether or not the customer has "churned"). Following this, Train-Test-Model Validation with an 90/10 split was utilized and outputted the four splots: X_train, X_test, y_train, and y_test. Regarding the split, 90% of the data was split into the training dataset that will be used to build the model. The remaining 20% will be used as the testing dataset to assess the model's performance on unfamiliar data. After splitting the dataset 90/10, two lists were created, one to contain only independent variables defined as continuous or numerical variables and the other to contain independent variables defined as categorical variables. Following this step, the continuous variables' z-score was computed, and the categorical variables (gender,

plan, competitorsub, topgenre, secondgenre, cancelled, downgraded, bundle, kids) were applied with One Hot Encoding.

Logistic Regression Model

After preparing the data, an empty logistic regression model was created. Subsequently, the logistic regression pipeline was constructed using the dataset's z-score and the empty logistic regression model. The model was then trained using X_train and y_train. Finally, the logistic regression model was utilized to make two predictions on whether a customer will churn: one by inputting X_train and the other by inputting X_test into the model.

Gradient Boosting Tree Model

The model creation process was also repeated for the gradient boosting tree model, but instead of creating an empty logistic regression model, an empty gradient boosting tree model was created.

Assessing the Performance of Each Model

After completing the four predictions, the two model's performance is conducted by calculating the training and testing sets' Accuracy, Precision, Recall, F1 Score, and ROC AUC.

Calibration

A visual comparison metric was also assessed, which is the calibration for the testing sets of both models. The calibration score of a model helps us improve the model's prediction probability and reliability. Additionally, calibration is essential when the probability estimate of a customer belonging to the "churn" or "not churn" is very important to the audience.

Recommendation System

A recommendation system was utilized to predict the "high-risk customers"' (customer's with the highest predicted probability of churning) top ten movies and films for the streaming service to suggest content for the customers individually.

Data Preparation

Two new datasets were utilized to accomplish this recommendation system: a new customer dataset and a favorite film dataset. The New Customer dataset is data on customers not trained in the logistic model. The Favorite Films dataset is data on the films the customers enjoy and deem their favorite. The data was prepared through multiple steps before fitting the independent variables into the model. First, the "null" or empty values were dropped from both datasets, and their indices were reset.

Identifying High-Risk Customers

To identify the high-risk customers, the data from the "New Customers" dataset was used to predict the high-risk customers utilizing the trained Logistic Regression model. Afterwards, a new column was added to the "New Customers" dataset to store the predicted probabilities (rounded to two decimal places.) Afterward, the top 200 high-risk customers were selected from the dataset based on the highest predicted "churning" probability and assigned to a new dataset called "High-Risk Customer."

Nearest Neighbors Model

A list was created to contain the independent variables that would be utilized to calculate the nearest neighbors: age, income, and meanhourswatched. Following this step, the z-score was computed which standardizes the continuous variables. After preparing the data, an empty Nearest Neighbors model was created with `n_neighbors = 10` to find the ten (10) nearest neighbors of a data point (customer). Subsequently, the Nearest Neighbor pipeline was constructed using the dataset's z-score and the empty nearest neighbor model, and the pipeline was fitted with the data from the independent variables' columns in the "Favorite Films" dataset.

Addition of the Recommended Content

Afterwards, the `named_steps` attribute was utilized to extract the neighbors and distances, and the values are assigned to the `distances` and `neighbors` variables. Following this, the nearest neighbors' indices are populated to the column "neighbors" to the "High Risk Customer" dataset. Finally, the modified dataset is outputted to a .csv file named "myfile.csv."

Application of the Recommended Content

After obtaining the .csv file, I utilized a third-party application called "Shiny App" to predict for movies each customer might like based on what the most similar users in the dataset like.

Results

After conducting the performance analysis on the two models, both of the models' scores (Accuracy, Precision, Recall, F1, and ROC AUC) were very similar. The performance metrics for each model's training and testing seats are displayed in the table below. The calibration curve metrics for each model's testing set are displayed below the table.

	Accuracy	Precision	Recall	F1	ROC AUC
Logistic: Train	~0.741	~0.606	~0.281	~0.384	~0.737
Logistic: Test	~0.739	~0.587	~0.274	~0.373	~0.726
GBT: Train	~0.744	~0.620	~0.274	~0.380	~0.742
GBT: Test	~0.738	~0.587	~0.26	~0.361	~0.724

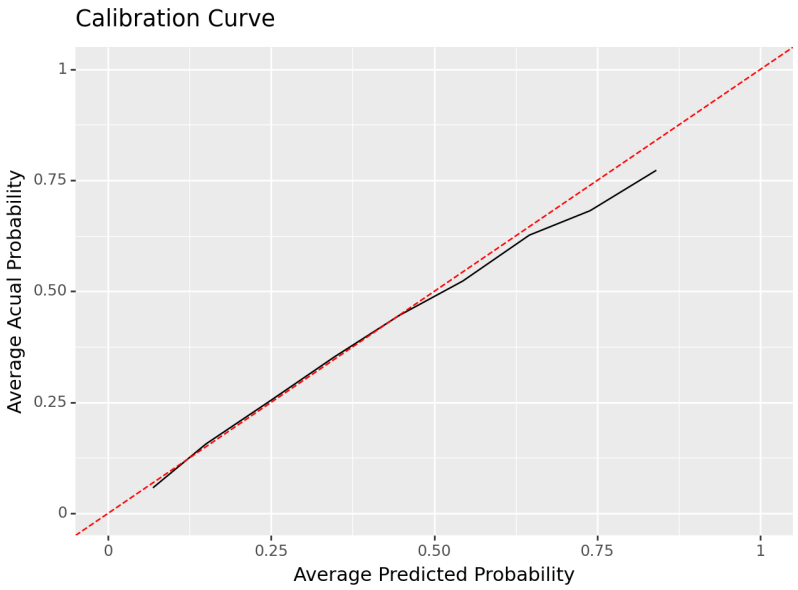

Figure 1: A graph showing the Logistic Regression's calibration curve (solid line) against the perfectly calibrated curve (dotted curve)

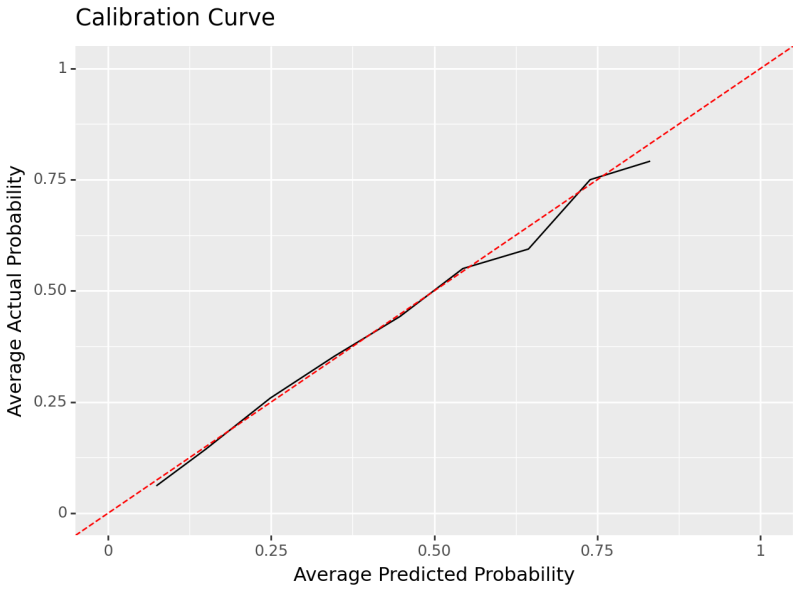

Figure 2: A graph showing the Gradient Boosting Tree's calibration curve (solid line) against the perfectly calibrated curve (dotted curve)

Model Comparison

The models' performances were tested through five (5) different metrics: Accuracy, Precision, Recall, F1 Score, and ROC AUC. The Accuracy of a model measures how much of the model was correctly predicted (TRUE Positive and TRUE Negative). The Precision measures the model's ability to make correct optimistic predictions. The Recall measures the model's ability to identify positive cases from all actual positive cases correctly. The F1 Score is a combination or mean of precision and recall that measures the model's ability to balance precision and recall. The ROCAUC is the area under the ROC curve, which measures the model's ability to distinguish between positive and negative cases. Finally, a visual calibration metric was utilized and tested on the models' test sets.

Fit of the Model

While there is a difference between the logistic and gradient boosting tree regression training and testing sets' Accuracy, Precision, Recall, F1 Score, and ROC AUC, the difference is not significant enough to show signs that either model is overfitting the data. With that in mind, the difference between the training and testing sets for the logistic model is less than the gradient boosting tree model.

Interpretation of Accuracy

The logistic regression model achieved a 0.3% lower accuracy on the training data and 0.1% higher accuracy on the testing data than the gradient boosting tree model. The logistic regression model correctly classified whether "a customer churned or not" for approximately 74.1% of the training data. The logistic regression model correctly classified approximately 73.9% of the samples on the testing data. The gradient boosting tree regression model correctly classified whether "a customer churned or not" for approximately 74.4% of the training data. The gradient boosting tree regression model correctly classified approximately 73.8% of the samples on the testing data.

Interpretation of Precision

The logistic regression model achieved a 0.14% lower precision on the training data and the same accuracy on the testing data compared to the gradient boosting tree model. Approximately 60.6% of the "customer churn" (1 or positive) predictions made by the logistic model on the testing data were correct. On the training data, approximately 58.7% of the "customer churn" predictions made by the logistic model were correct. Approximately 62% of the "customer churn" (1 or positive) predictions made by the gradient boosting tree model on the testing data were correct. On the training data, approximately 58.7% of the "customer churn" predictions made by the gradient boosting tree model were correct.

Interpretation of Recall

The logistic model achieved a 0.07% higher recall on the training data and 1.4% on the testing data than the gradient boosting tree model. The logistic model correctly identified approximately 28.1% of the actual "customer churn" (1 or positive) for the training data. The logistic model correctly identified approximately 27.4% of the actual "customer churn" (1 or positive) for the testing data. The gradient boosting tree model correctly identified approximately 27.4% of the actual "customer churn" (1 or positive) for the training data. The gradient boosting tree model correctly identified approximately 26% of the actual "customer churn" (1 or positive) for the testing data.

Interpretation of the F1 Score

The logistic model achieved a 0.4% lower F1 score on the training data and 1.2% higher on the testing data than the gradient boosting tree model. The logistic model and gradient boosting tree balance making accurate optimistic predictions (precision) and capturing the proportion of the actual positive cases (recall). The logistic model's training data had an F1 Score of approximately 38.4%, and its testing data had a score of approximately 37.3%. The gradient

boosting tree model's training data had an F1 Score of approximately 38%, and its testing data had a score of approximately 36.1%.

Interpretation of ROC AUC - Area Under the Receiver Operating Characteristic Curve

The logistic model achieved an ROC AUC of 0.68% lower on the training data and 0.28% than the gradient-boosting tree model. The logistic and gradient boosting tree models' ROC AUC scores indicate that both models had good discrimination toward distinguishing between positive and negative cases. The logistic model had approximately 0.737 ROC AUC score for its training data and 0.726 ROC AUC score for its testing data. The gradient boosting tree model had approximately 0.742 ROC AUC score for its training data and 0.724 ROC AUC score for its testing data. In Figures #3 and #4 (Logistic) and Figures #5 and #6 (Gradient Boosting below are the ROC curves for both the logistic and gradient boosting tree models.

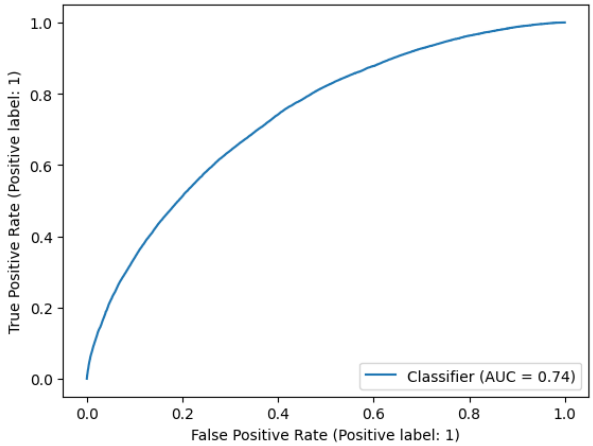

Figure 3: A graph showing the ROC Curve for the Logistic Regression's model on the Training data.

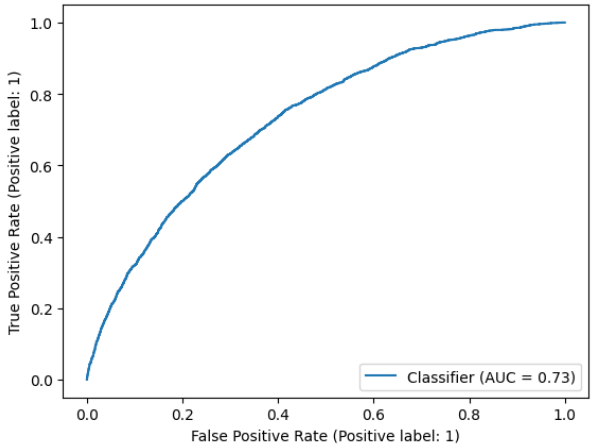

Figure 4: A graph showing the ROC Curve for the Logistic Regression's model on the Testing data.

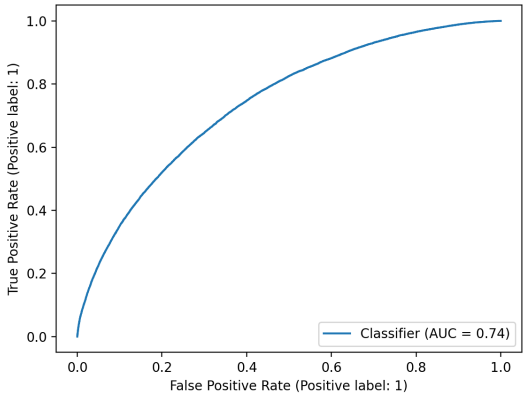

Figure 5: A graph showing the ROC Curve for the Gradient Boosting Tree's model on the Training data.

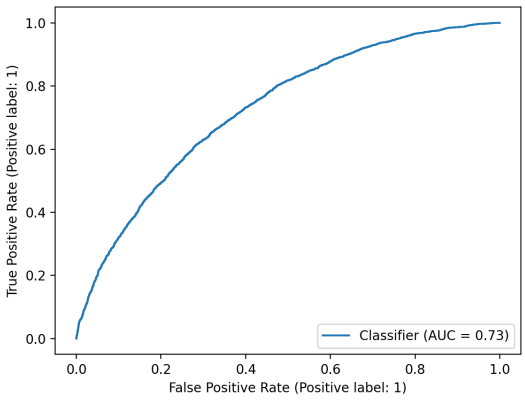

Figure 6: A graph showing the ROC Curve for the Gradient Boosting Tree's model on the Testing data.

Interpretation of Calibration

In regards to the calibration for each model, the calibration for each model was graphed with the x-axis being "Average Predicted Probability" and the y-axis being "Average Actual Probability." The closer the model's graphed curve is to the perfect calibrated model curve (dotted curve/line), the better calibrated the model is. A perfectly calibrated model is visualized as a y = x curve/line. From the graphs (Figure #1 [Logistic] and Figure #2 [GBT]), we can see that Logistic's Calibration Curve is consistently closer to the perfect calibration curve compared to the Gradient Boosting Tree's Calibration Curve.

Model Trust and Its Caveats

The statistical comparison (Accuracy, Precision, Recall, F1, ROC AUC) between the two models reveals that the Logistic Regression and Gradient Boosting Tree models performed similarly predicting whether a customer will "churn." That said, the Logistic Regression model was much better calibrated than the Gradient Boosting Tree. At first glance, both models' calibration was similar. Still, at about 0.55 Average Predicted Probability, the Gradient Boosting Tree calibration curve began to deviate more from the perfect calibration curve. Regarding this data analysis, calibration is a crucial metric for measuring the models' performances and ensuring that the predicted probabilities align with the actual probability. Calibration is a metric for determining the reliability of the predicted probabilities (churn), making decisions, evaluating the model's reliability, and enhancing the performance of ensemble methods (e.g., Random Forests, Gradient Boosting Tree, etc.).

Another point that was considered in choosing the model, the level of the time/space complexit and memory usage costs of the gradient boosting tree model outweighed the benefits of the model. As mentioned, the percentage difference between the two models was minor, but the model complexity and training time difference are much more significant. The reasoning is that the tree model is an ensemble method that combines multiple decision trees to make predictions, and each tree is iteratively constructed and involves many branches and nodes.

That said, I trust the results of the logistic regression model more than that of the gradient-boosting tree's. Still, I would advise the streaming service company to take caution with completely being dependent on the data. The reasoning behind this suggestion is that the streaming service economy is a sensitive economy that is prone to substantial fluctuations based on different variables such as the entertainment offered per service, the cost, the time of the year (summer break, school session, winter break), ratings of the films, seasonal films, and the type of people that utilize the services (e.g., college students, parents, kids, etc.). To combat these issues, I highly recommend the streaming service continue conducting surveys and legally collecting data from their customers to constantly update the model with new data and potentially capture significant data trends.

With the suggestions in mind, I also suggest the CEO of the Streaming Services company to utilize the logistic model to identify the customers who have a high risk of churning based on their consumer profiles. While the model doesn't have more background on the customers besides the information provided (e.g. college student, medical history, etc.), this model is able to gather a generalize view on the churning rates for the customers. That being said, a recommendation system was built in this project as well, and this system was utilized for identifying and suggesting movies that the top 200 high risk customers (that weren't a part of the training or testing dataset) would enjoy based on their profiles. In Figure #7 are the top 10 movie and/or film recommendations for each of the two "high-risk" customers from the new customer ("high-risk") dataset. I would also suggest the CEO to strategically use the movie suggestions generated to identify what films that would need to be kept or added on to the streaming service. The main reason for consumer to subscribe to or churn from a streaming service is for the content that the service provides, so having a list of high-risk consumers' potential favorite films can be crucial towards understanding the consumer demand.

Recommendations

recs	index	gender	age	income	monthssubbed	plan	meanhourswatched	competitorsub	numprofiles	cancelled	downgraded	bundle	kids	longestsession	topgenre	secondgenre	pred	neighbors
Monique, Lip I,Bona,Ai to nikushimi no kanata e,In girum imus nocte et consumimur igni,Chases of Pimple Street, The,Kaidan Oiwa no borei,C'est le vent,Wired 03:36,Daniel's Spark,Goal Club	42	woman	26	57.69	6	B	22.3	0	3	1	0	0	0	122.39	Comedy	Thriller	0.86	[209 317 416 328 268 363 307 111 445 376]
Bona,Space Available,100 Years at the Movies,Dolor de pagar la renta, El,Throwing Down,Orekhovyy khleb,Chases of Pimple Street, The,Goal Club,Clase z tropical,Happy	351	woman	20	52.05	1	P	9.36	1	2	1	0	0	1	113.91	Thriller	ScienceFiction	0.85	[381 307 342 328 445 8 363 370 194 197]

Figure 7: A table showing the top ten movie suggestions for two "High-Risk" customers based on their consumer profile.

Discussion/Reflection

By performing these analyses, we can see the impacts different variables can have on predicting whether the streaming service's customers will churn. Furthermore, running these analyses visualizes the importance of fitting the data to the correct model to reduce the potential errors in prediction and the time and space complexity of a model. We could see this by comparing the logistic regression and gradient boosting tree's training and testing performance metrics: Accuracy, Recall, Precision, F1, ROC AUC, and Calibration. After selecting the logistic model as the model of choice, we were able to utilize the model to predict some movies that high-risk customers would enjoy. This data analysis identified the problem and proposed a solution to the problem.

If this analysis were repeated, I would add a max depth for the decision trees in the gradient boosting ensemble. Not only does it reduce the computational costs of the Gradient Boosting Tree model utilized in this analysis, but it can also lead to minimizing any signs of minimal overfitting and increase the model's performance. Additionally, another item that I would like to change in the future is to add media trends and global news (e.g. Oscar wins, controversies with the actors/actresses of a film, COVID-19 pandemic, etc.) to the analysis, which can be essential factors in predicting whether a customer will churn or not.