Assignment 2

Jonathan Lopez

Introduction

The purpose of this assignment is to analyze the customer data from a streaming service company, to determine whether customers are likely to churn or stop their use of the service. Analysis also included exploring mushroom classification to determine whether mushrooms are poisonous or edible. For the business churn analysis, the aim was to predict customer churn based on customer demographics such as gender and income, subscription history such as months subscribed and plan, and behavioral data. For the mushroom classification analysis, classification was determined using categorical attributes such as mushroom appearance, including but not limited to mushroom cap space, color and odor. For this analysis, the model implementation included Logistic Regression, Gradient Boosting, Categorical Naive Bays, and K-Nearest Neighbors (KNN). These models could impact the streaming services ability to implement customer retention strategies and for the mushroom classification, provide a means of determining whether a mushroom is edible or harmful if ingested.

Methods

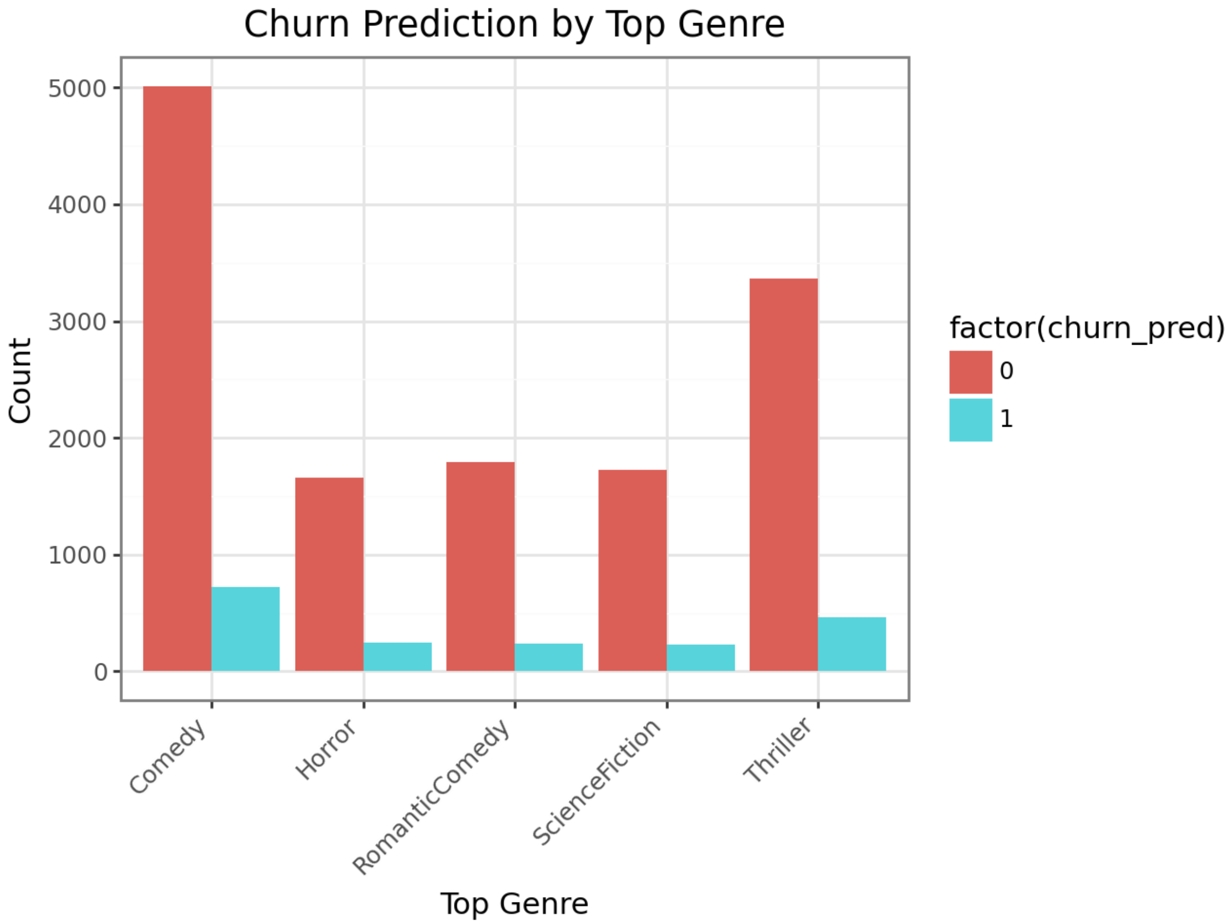
For the business churn model analysis, we built two models including a Logistic Regression and Gradient Boosting model. Data preprocessing for the churn models involved applying z-score scaling to continuous variables and one hot encoding the categorical variables before train-test splitting to a 80-20 ratio. For the mushroom analysis, the models used included Categorical Naïve Bayes, KNN and logistic regression. One-hot encoding was applied to all models for the dataset and evaluation for both data sets utilizing Accuracy, Precision, Recall and ROC AUC. For the mushroom classification, a mushroom detector function was implemented to predict if a mushroom is poisonous or edible.

Results

For the ***Business Churn analysis***, the Gradient Boosting model demonstrated more balanced performance metrics with a slightly higher precision score than the Logistic Regression, with very close train and test accuracy scores. The train/test scores indicate no overfitting, but the low recall for both models suggest moderate underfitting. For both models, the ROC AUC scores were moderate at around .599 - .603, which indicates issues with calibration and the models credibility in estimating churn. Based on the results, the Gradient

Boosting model should be chosen for production due to its slightly higher precision and fewer false positives, which would allow it to be utilized to identify potential churn and implement targeted retention strategies. Gradient boosting is also more advantageous due to its ability to model more complex relationships; however, it would require more computational resources when compared to a logistic model. The bar chart showing churn predictions by top genre showed customers who favor comedy or thriller were least likely to churn, when compared to other genres. These insights would all be useful for the streaming company to implement and target more retention strategies amongst their customers.

Model	Train Accuracy	Test Accuracy	Precision	Recall	ROC AUC
Logistic Regression	.7409	.7416	.6171	.275	.603
Gradient Boosting	.7435	.7414	.6219	.264	.599



For the ***Mushroom Classification***, the KNN and Logistic Regression models achieved the best performance with perfect scores in all categories when compared to the Categorical Naïve Bayes which had slightly lower recall and precision scores. The mushroom detection function utilized all three models and returned correct classifications for the mushroom containing indices tested on. While the function and train test score indicate high accuracy, there are potential concerns regarding overfitting due to the 1.0 for both test and train on KNN and LR models, indicating need for further validation before deploying into commercial use. I would however recommend using the function/application perhaps as an advisory tool to aid in mushroom verification. The primary concern is the risk of misclassification due to false negatives, which would pose serious ethical and health concerns for both the company and consumers.

Model	**Train Accuracy**	**Test Accuracy**	**Precision**	**Recall**	**ROC AUC**
Categorical Naïve Bayes	1.0	.93	.97	.88	.994
KNN	1.0	1.0	1.0	1.0	1.0
Logistic Regression	1.0	1.0	1.0	1.0	1.0

Discussion/Reflection

This analysis highlighted the importance complexity and simplicity when determining the best model for the data set to be evaluated and predicted on. While Logistic regression provided an easier interpretation, it did not capture more complex relationships like the Gradient Boosting model. Future improvements include possibly integrating cross-validation to enhance the model's metrics reliability and lasso regression to reduce the risk of overfitting.