Comparing Supervised Learning Models to Predict Netflix Customer Churn

Team 5B: Boram Gaudet, Chris Gu, Jason Jia, Srita Kothuri

BANA 273, Section B

December 5, 2025

Executive Summary

The goal of this project was to apply machine learning to predict Netflix user churn based on a publicly available Netflix dataset. This process included the analysis of various descriptive variables using three different machine learning models: logistic regression, decision tree, and random forest. Across all three models, features related to engagement emerged as the most influential in predicting customer churn: average watch time per day, watch hours per month, and last login. The decision tree algorithm had the highest recall metric out of the three models tested, and so we selected it as the final model.

Introduction

For most companies, customers make up the foundation of the firm's daily operations. Therefore, it is essential for a company whose business is driven by subscription plans to obtain a clear understanding of what factors drive churn. Investing in this kind of market research is important for online streaming platforms such as Netflix, because customer churn can have a large financial impact. Predicting churn enables the firm to create more targeted and effective marketing interventions, redirect expenses from acquiring new customers to lower expense endeavors like retaining customers, and ultimately optimize overall spend on marketing and advertising.

To understand the key factors in predicting customer churn, our group acquired a publicly available data set titled, *Netflix Customer Churn*, from the database website, Kaggle. This synthetic data set contains 5,000 unique customer values and additionally provided information across 14 different variables:

- Customer_id: unique identification number for each customer
- Age
- Gender: male, female, other
- Subscription_type: basic, standard, premium
- Watch_hours: hours of watched content on Netflix per month
- Last_login_days: number of days since last log in
- Region: South America, Europe, Asia, Africa, North America, Other
- Device: tablet, laptop, mobile, TV, desktop
- Monthly_fee: range from $8.99 - $18.00
- Churned: Churn (1), no churn (0)
- Payment_method
- Number_of_profiles: range from 1-5
- Avg_watch_time_per_day: average watch time per day in hours
- Favorite_genre: horror, sci fi, romance, drama, documentary

To gain an understanding of which of these factors are most influential in predicting customer churn, our team turned to supervised machine learning models. Specifically, we employed three types of classification algorithms—logistic regression, decision tree, and random

forest—and compared the performance of each. Classification models assign data to predefined categories (classes) by learning patterns from labeled training data. Then, they use these learned characteristics to predict the correct class for new, unseen test data. Our group decided to run three different classification algorithms to not only gain a more accurate prediction status, but to also explore differences amongst the models. For example: Does each model differ in which attribute they rank the most influential for predicting churn? How does each model differ in evaluation metrics?

Through our analysis, we aimed to answer two main questions:

1) Of the 14 customer attributes, which was most influential in predicting if a customer churns or not?
2) Are there differences between each of the classification models—either in which factor they ranked as most influential, or their evaluation metrics?

To determine which model performs best in predicting user churn, we used the F1 score as the primary metric. The F1 score combines precision and recall into a single value, so it tells us not only how well the model identifies genuine churn users (recall) but also how reasonably well it identifies potential churn users (precision). In this case, Netflix is indeed concerned with finding potential churn users, but it also doesn't want to waste retention resources on too many loyal users who will never churn. Therefore, focusing on the F1 score led us to choose models that strike a practical balance: they reliably identify churn users while controlling the number of unnecessary "potential churn" alerts, which aligns perfectly with the business goal of effectively retaining users.

Exploratory Data Analysis

To get a better understanding of the data, our group conducted preliminary data analysis prior to preprocessing the data. This included visualizing key variables, as seen in the figures below, and using Python functions to assess the average and standard deviation of the continuous variables in our data set.

...

	age	watch_hours	last_login_days	monthly_fee	churned	number_of_profiles	avg_watch_time_per_day
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000
mean	43.847400	11.649450	30.089800	13.683400	0.503000	3.024400	0.874800
std	15.501128	12.014654	17.536078	3.692062	0.500041	1.415841	2.619824
min	18.000000	0.010000	0.000000	8.990000	0.000000	1.000000	0.000000
25%	30.000000	3.337500	15.000000	8.990000	0.000000	2.000000	0.110000
50%	44.000000	8.000000	30.000000	13.990000	1.000000	3.000000	0.290000
75%	58.000000	16.030000	45.000000	17.990000	1.000000	4.000000	0.720000
max	70.000000	110.400000	60.000000	17.990000	1.000000	5.000000	98.420000

Figure 1. Descriptive Statistics of all numerical variables in Netflix Customer Churn dataset.

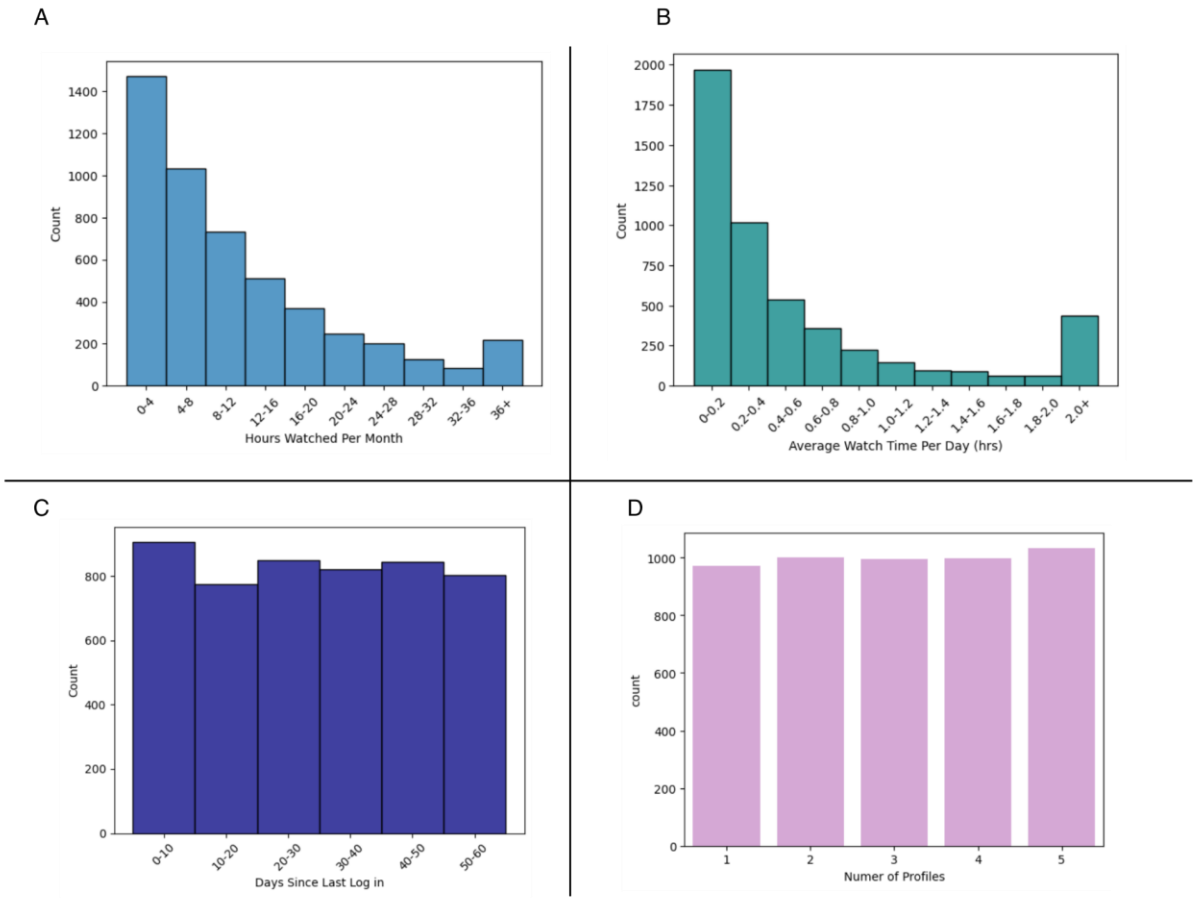


Figure 2. (A) distribution of hours watched per month. (B) average watch time per day in hours. (C) Number of days since last log in. (D) Distribution of number of profiles.

The exploratory data analysis shows that the distribution among many of the key variables are largely evenly distributed. The days since last log in, number of profiles, gender, and age have almost equal distribution among customers. However, attributes that varied the most were those that involved watch time. The highest average watch time per day and per month were between 0-0.2 hours and 0-4 hours respectively.

Data Analysis

Logistic Regression

The first analytical model we ran our data through was Logistic Regression. The logistic regression model is a type of classification tool that predicts binary outcomes. We chose to apply this tool first because it can handle both numerical and categorical independent variables, allowing us to predict the log odds that an event will occur. First, we ran our regression model without any preprocessing steps, meaning that we did not scale any of our numerical data. However, we created dummy variables for our categorical variables, as this was necessary for our model to run properly.

After using a StandardScaler to standardize the values of all our numerical variables, we ran a logistic regression model and used K-Fold cross-validation to help boost the robustness of our model. K-Fold cross-validation is a resampling method that uses a prefixed number of folds (k) to resample and shuffle through the data in an iterative fashion. This ensures that overfitting does not occur as the samples involved in the training of the data are consistently changed throughout.

Our group also compared the feature importance before and after conducting the preprocessing steps and K-Fold cross-validation. The results of the confusion matrix, accuracy readout, and feature importance varied with and without preprocessing the data. The figures below show the differences.

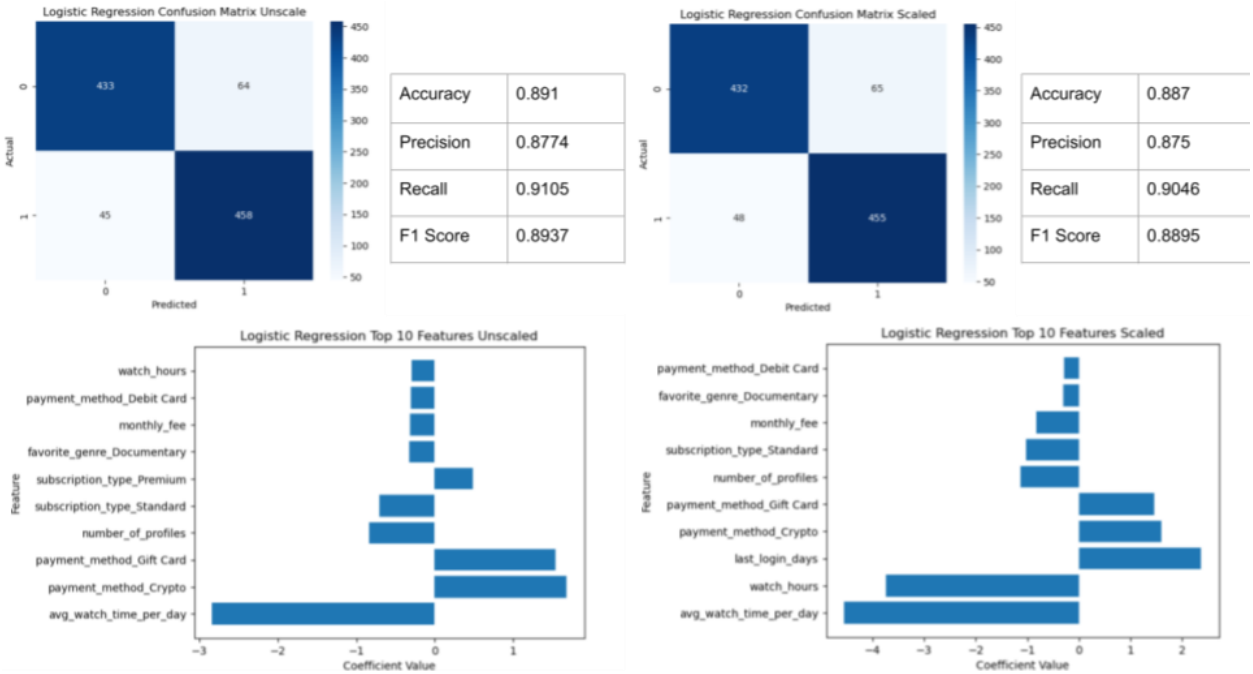


Figure 3. Confusion Matrix, accuracy metrics, and feature importance for preprocessed logistic regression (A) and cleaned logistic regression (B).

Running the data set through the logistic regression model showed us that the top predictor in Netflix customer churn is average watch time per day, meaning customers who

watch more content each day are significantly less likely to churn. In both the raw and preprocessed data, a 1-hour increase in average watch time per day decreases the odds of churn by 92% and 99%, respectively. However, preprocessing the data does change the remaining top 9 features, with watch hours and last log in days becoming more influential in the revised data set. Additionally, the F1 metric on the raw data was 89.4% but slightly declined to 89.8% after adjusting the variables.

Decision Tree

Following our logistic regression model analysis, our team determined that the decision tree algorithm would also be useful in predicting churn. Decision tree models split data into branches based on true/false conditions, creating a series of decision rules that classify users into one of two classes (churn or not churn). Decision trees are useful because they can capture non-linear relationships that are often present between predictor and target variables, and the tree-like-diagram makes the model highly interpretable as opposed to a logistic regression model, whose pathways are not made clear.

Similar to the logistic regression model, our group compared the confusion matrix, evaluation metrics, and top 10 influential features before and after cross-validation. We used hyperparameteric tuning via GridSearchCV to find the optimal combination of hyperparameters for our decision tree. This process outlines the appropriate depth (layers) of our decision tree to ensure we're not overfitting or underfitting the training data. The reason our group chose GridSearchCV out of the many other hyperparametric tuning methods is due to the model's robust capabilities. Although it is computationally expensive, this method of hyperparametric tuning runs through every possible hyperparameter possible, ensuring that the most optimal value is reached. As part of this process, we used 5-fold cross-validation to ensure that our data is still valid, without overloading our computer.

Similar to the logistic regression model, the decision tree algorithm found average time watched per day to be the most influential factor, contributing roughly 52% of the total predictive power of the model. The F1 metric for this model, both before and after validation, is higher than that of the logistic regression model, proving that the decision tree is a better suited algorithm to predict churn.

It is also important to note that compared to the previous model, the decision tree compiles a different list of the remaining top 9 influential features. Both models found that the number of profiles are the top three most influential factors. However, the variables other than average watch time per day are differ between each of the models.

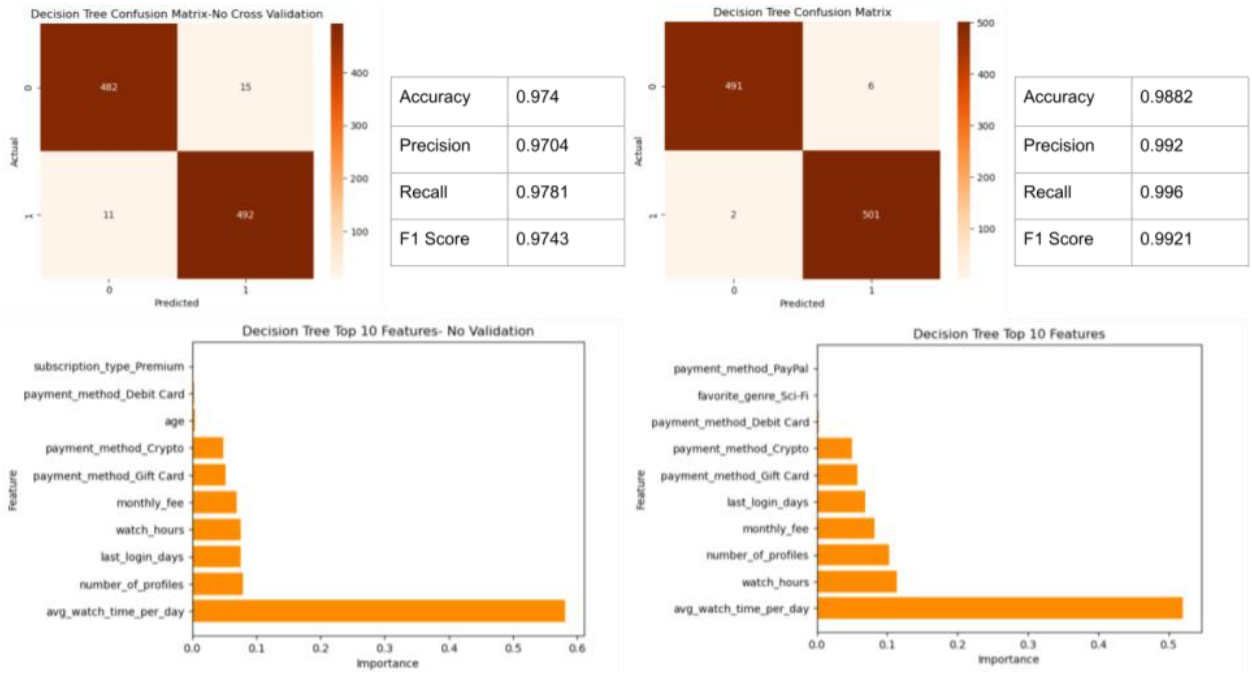


Figure 4. Confusion matrix, precision metrics, and feature importance in decision tree before hyperparametric tuning (A) and after (B).

Random Forest

The last model our group used to address the business problem was the random forest algorithm. Random forest seemed to be an obvious choice, as it builds on the structure of decision trees, but offers greater robustness. As an ensemble method, it builds many decision trees and averages their predictions, which helps minimize overfitting and captures more complex interactions between features than a single tree can. Additionally, random forest handles large datasets well, making it well suited for our analysis.

Instead of using GridSearchCV as a validation method for random forest, our group used RandomSearchCV. This method is very similar to GridSearchCV, but instead of parsing through every possible value, RandomSearchCV decreases the computational complexity by testing random combinations of values. Similar to the decision tree model, we used a 5-fold cross-validation strategy. Figure 5 shows the differences between values before and after conducting parametric tuning.

Consistent with all our findings so far, validation increased the recall, precision, and accuracy measurements of the model. Specifically, F1 for the decision tree model is the highest out of all three models we have assessed so far (0.992). Additionally, average time watched per day continues to be the most influential feature, contributing approximately 53.3% of the model's predictive power—slightly higher than the 52% observed in the decision tree model. This increase may be due to random forest's ensemble approach, which aggregates multiple trees and can better capture the full range of patterns associated with this feature.

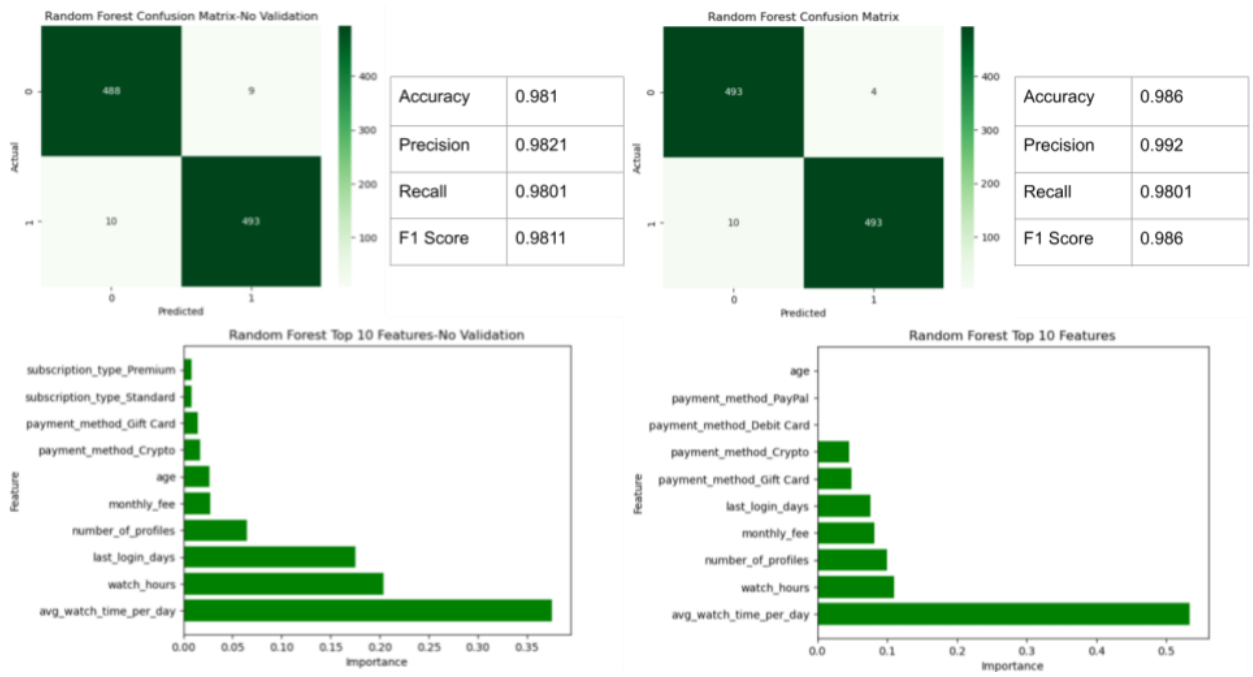


Figure 5. Confusion matrix, precision metrics, and feature importance in decision tree before hyperparametric tuning (A) and after (B).

Discussion

Conclusions

Predicting customer churn for Netflix by running data through three different classification models gives rise to three main conclusions.

Firstly, average daily time spent watching content on Netflix emerged as the strongest predictor of churn. Individuals who watch more each day have significantly lower odds of cancelling their subscriptions. This was echoed across all three models. Days since last login and monthly watch hours and number of profiles are closely followed in terms of influence. While each model predicted varying levels of influence for these four attributes, they remained the top three predictors in all models that we assessed.

Secondly, each classification model differed in its accuracy, precision, and recall metrics. The random forest model and decision tree model performed similarly, both achieving high scores in the 0.95s across all three metrics, while logistic regression fared the worst. Since the tuned decision tree achieved the highest F1 score among the three models, we ultimately selected it as our primary model. Its F1 score suggests that it strikes a good balance between correctly flagging likely churners and avoiding too many false alarms. In practice, the tree is flexible enough to learn nonlinear patterns and simple interactions in the data, while the random forest's heavier averaging appears to smooth out some of these signals and slightly lowers its F1. Logistic regression, on the other hand, is limited by its linear structure, so it cannot fully capture the more complex relationships in this dataset and therefore performs worse overall.

Lastly, our analysis also shed light on the importance of preprocessing data and cross-validation. In the analysis of the regression model, our group compared predictions before both

preprocessing and validation with the post-processing and validated scores. All three accuracy parameters increased in value. This was similar to what was seen in the decision tree and random forest models before and after validation with hyperparameter tuning.

Overall, our analysis highlights how each machine learning model differs in setup, capability, and strength, revealing the advantages and limitations they bring to forecasting outcomes on our dataset.

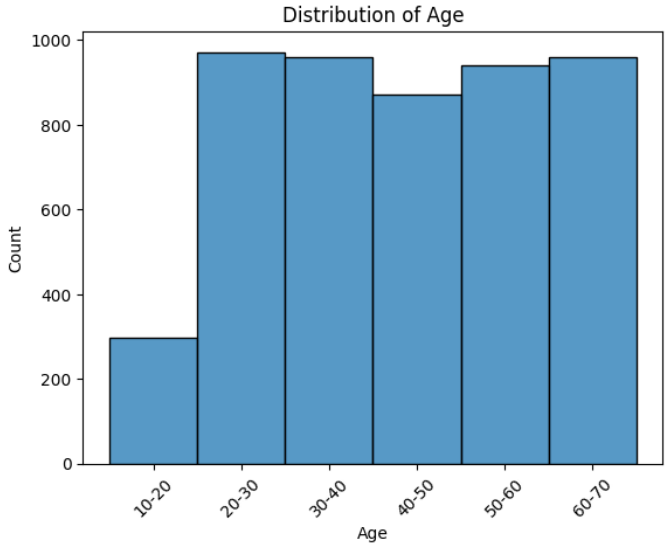
Business Implications

Implementing machine learning models in business contexts could be incredibly powerful for Netflix. The company could deploy the most accurate machine learning model, decision tree in our case, to continuously monitor engagement behavior and flag users that have reduced their average watch time per day or per month. From here, the marketing teams at the company could customize campaigns strategies to target these low-engagement users with personalized content, reminders, or incentives. This could look like discounts on monthly subscription plans, emails regarding new shows hitting the platform, and customer surveys to understand what has caused a change in the customer's behavior. These strategies could not only ensure that Netflix retains their strong customer base, but also prevents them from exerting extra force and money on trying to lure in new customers to keep profits up.

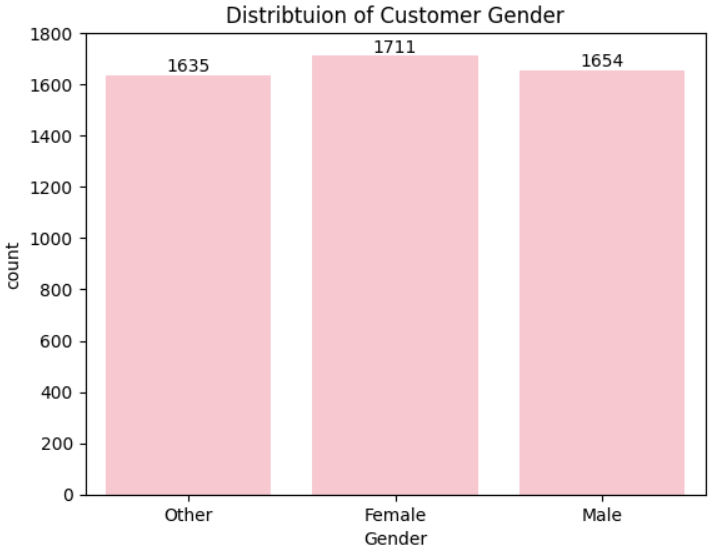
Limitations and Future Directions

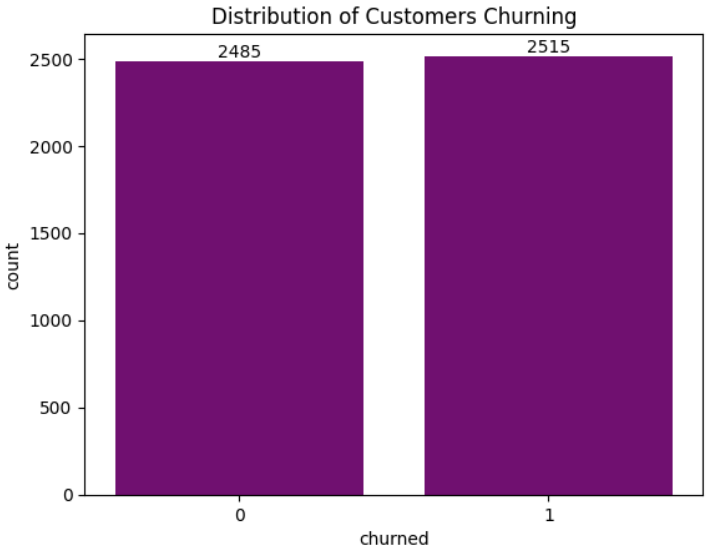
The foremost limitation of our work lies in the fact that we used a synthetic data set procured from an online public database. While the methods we used are valid, the conclusions we have drawn from this analysis do not necessarily represent real-world patterns or genuine Netflix customer behavior. Furthermore, although we evaluated three supervised machine learning models, additional algorithms that may be better suited to our business question were not explored due to our group's limited technical familiarity. Therefore, future work could benefit from testing a broader range of supervised learning methods to more confidently determine which model most effectively predicts customer churn.

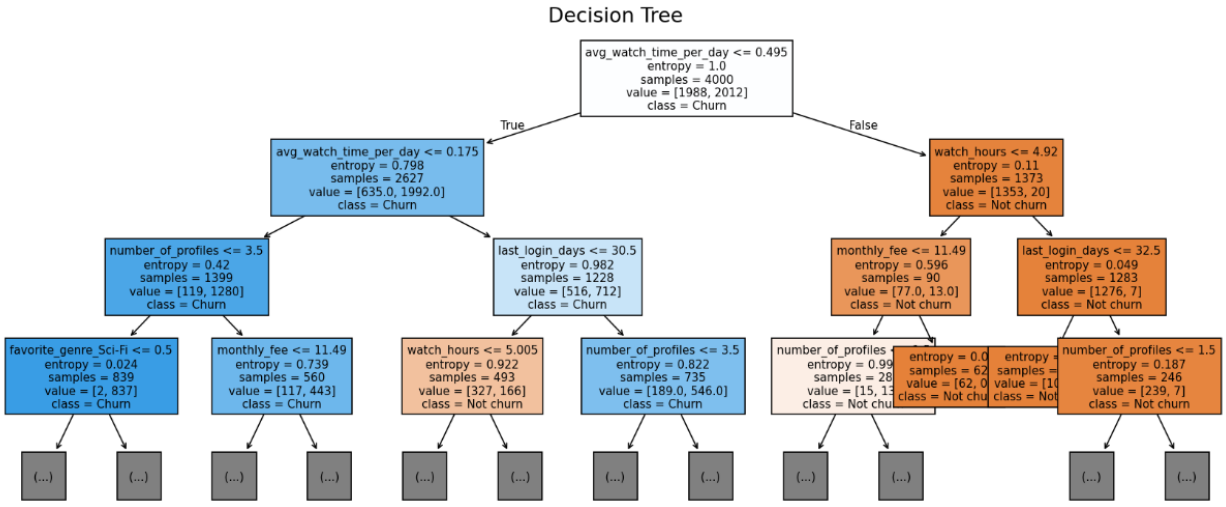
Appendix


Distribution of Age
Count
Age
0
200
400
600
800
1000
10-20
20-30
30-40
40-50
50-60
60-70


Distribution of Customer Region
Region
Count
0
200
400
600
800
1000
803
867
841
765
873
851
Africa
Europe
Asia
Oceania
South America
North America


Distribtuion of Customer Gender
1800
1600
1400
1200
1000
800
600
400
200
0
count
1635
1711
1654
Other
Female
Male
Gender


Distribution of Customers Churning
2500
2000
1500
1000
500
0
count
2485
2515
0
1
churned



Single Decision Tree from Random Forest (Tree #0)