Netflix Churn Classification with Machine Learning Models

Brendan George Lauterborn
Department of Computer Science
Towson University
Towson, United States
blaute3@students.towson.edu

***Abstract*—Accurate prediction of Netflix customer churn is essential for business success and customer retention. This paper aims to predict whether a Netflix customer will churn or not. Multiple binary classification techniques were tested to determine the best model.**

I. Introduction

Correctly predicting and analyzing customer churn is essential to a business' health and success. Losing customers only hurts Netflix and developing ways to reduce this can save a business. This paper shows research on predicting customer churn using the Netflix Customer Churn dataset published on the Kaggle machine learning repository.

Multiple regression techniques are applied and evaluated to determine customer churn. These techniques include Logistic Regression, Support Vector Machine, and Random Forest. The performance of the model is tested using accuracy, F1, and ROC.

II. Literature Review

This section discusses the literature review aimed at predicting customer churn. We highlight limitations or unexplored areas of their research methods.

Mohan and Jadhav performed a study on OTT platform churn that provided a benchmark for our research. The dataset they used consisted of 317 respondents of a questionnaire. They trained and tested four models that predicted customer churn. These models that they used were Decision Tree, Random Forest, Ada Boost, and Gradient Boost. 80% of the data was used for training and 20% was used for testing their models. The models that performed best in their study were the Random Forest and Gradient Boost models, having an F1 score of .833 and .825 respectively.

The main limitation of this study was the dataset used. 76.02% of the users have multiple streaming accounts. The number one platform was Netflix, where 46.69% respondents use this platform. The dataset is also a survey, rather than actual Netflix data. This allows for potential bias in response due to self reporting. The attributes of their dataset do not capture important usage of the streaming service, such as watch time, last login date, and completion. The attributes in their survey mostly focus on quality of content and demographics. The churn attribute was also not a factor of if the customer left, but rather if they plan on leaving.

The second study that we looked at focused on Netflix's retention strategy. Singh analyzed how predictive analytics and subscription based services can mitigate churn. The dataset for this study was built using sentiment surveys, user reviews, and customer feedback. Singh built a logistic regression model that used the independent variables subscription tier, monthly usage hours, perceived content quality, satisfaction with pricing, age, and customer tenure. In this model Singh determined that the most statistically significant attribute was price satisfaction. Having a higher satisfaction resulted in a lower churn. Also, subscription tier and tenure were deciding factors in predicting customer churn.

Random Forest was the next binary classification model that was built. In this model, the top prediction features mentioned were pricing satisfaction score, viewing frequency, tenure, plan type, and age. This model obtained an accuracy of .84. Singh highlights that users who are younger are more sensitive to price changes in the subscription.

Once again, one limitation we observe is how the dataset is obtained. This approach allows for people who are more vocal to speak up and participate. Thus, allowing for a self-selection bias. Users with a strong positive or negative view are more likely to participate. Therefore, the dataset may not accurately represent the user base of Netflix. However, the dataset focuses more on sentiment analysis and pricing satisfaction. Available: https://www.ijprems.com

III. Problem Definition

This study aims to predict whether Netflix member will churn based on the attributes that make up the data set. This problem is formulated as a binary classification problem. Churned is represented as a 1 and not-churned is represented as a 0. This problem is extremely relevant because it has clear meaning towards a business. From a business standpoint, Netflix wants to maximize their profits by retaining as many customers as they can. Therefore, identifying customers who churn allows Netflix to improve their product in certain ways to increase their revenue.

The dataset contains 5,000 records of customer data. Each customer record contains customer_id, age, gender, subscription_type, watch_hours, last_login_days, region, device, monthly_fee, payment_method, number_of_profiles, avg_watch_time_per_day, favorite_genre, and the target variable churned.

IV. Approach Description

We begin our experiment by performing data preprocessing and encoding. From there, we used Logistic Regression, Support Vector Machines, and Random Forest to predict Netflix customers churn.

A. Data Preprocessing and Encoding

We first begin by dropping the attribute customer_id as it is not relevant in our study. The dataset contains a few categorical attributes that need to be encoded. These attributes are gender, subscription_type, region, device, payment_method, and favorite_genre.

We used ordinal encoding for subscription_type as each tier offers more benefits to the customer and is ordered. The tiers were basic, standard, and premium. We encoded gender, region, device, payment_method, favorite_genre using one-hot encoding. The remaining attributes, age, watch_hours, last_login_days, monthly_fee, number_of_profiles, and avg_watch_time_per_day are kept as is since they are numerical values already. Each model is trained using 70% of the data and tested with the remaining 30%.

B. Methodology

The first model that we used was Logistic Regression. This model is a good baseline because it is simple and linear. Given input features as a matrix and the binary dependent variable, Logistic Regression learns coefficients and assigns them to the linear combination of input features. From there, the sigmoid function is applied which squishes the real value of the equation between 0 and 1. This value is the probability that the customer will churn. A threshold is applied such that any probability greater than .50 results in a 1.

The next model that we used was SVM. SVM attempts to find the best hyperplane that separates each class of the data while maximizing the margin between the classes. A larger margin results in better performance on unseen data. This hyperplane is found by using support vectors. SVM can use kernel functions to create non-linear boundaries. SVM is extremely effective on high dimensional data because the model depends on the number of support vectors and not the dimensionality of the data.

The last model that we decided to use was Random Forest. Random Forest builds many decision trees independently. Each tree is trained on a random subset of the data, and therefore gives its own prediction. The predictions are then averaged and the result becomes the final prediction. Due to the building of multiple decision trees, this approach is expected to give a more accurate prediction of Netflix customer churn. This model can also handle categorical data and therefore encoding is not necessary.

V. Experimentation

To evaluate our models, we use accuracy, F1 score, and the Receiving Operator Characteristic(ROC). Accuracy measures the amount of correct predictions to total predictions. The F1 score uses precision and recall. Precision is the amount of true positives over the sum of true positives and false positives. It measures the accuracy of positive predictions. Recall is the amount of true positives over the sum of true positives and false negatives. Essentially, the ratio of true positives to the actual amount of total positives. The F1 score is a good evaluation because it combines the mean of both precision and recall. If one measure is bad, then that greatly impacts the score. The AUC-ROC graph is used to determine how well our model classifies our data. The ROC curve plots true positive rates vs. false positive rates. The AUC measures the area under the ROC curve. More area under the ROC curve indicates a better model.

A. Logistic Regression

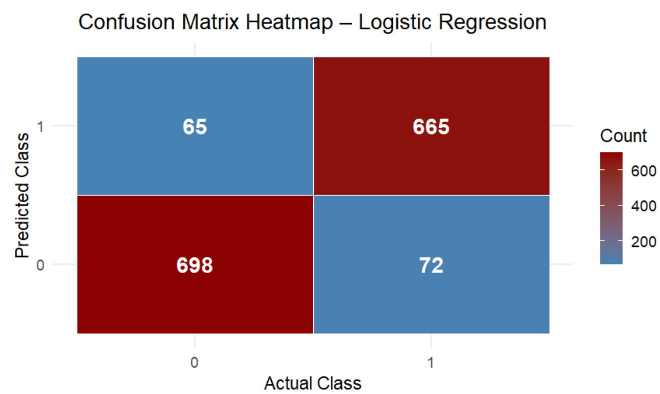

Fig. 1: Logistic Regression Confusion Matrix

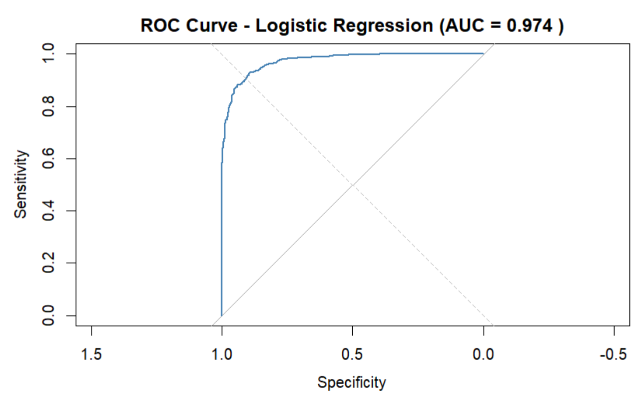

Fig. 2: Logistic Regression ROC

The first experiment that we ran was the logistic regression model. This model produced an accuracy of .909. Meaning that 90.9% of the time, we correctly predicted whether the customer churned. Fig. 1 shows that our model correctly predicted 0 698 times and correctly predicted 1 665 times. Our model resulted in 65 false positives and 72 false negatives. To further evaluate our model, we calculated the F1 score. The logistic regression model had an F1 value of .907. Meaning our model had a good balance between precision and recall as we captured most customers who churn. Our model also had an AUC value of .974 as shown in Fig. 2. This shows that our logistic regression model was able to properly assign churn to the true churn customer. Overall, our model was accurate and effective in predicting customer churn.

From our logistic regression model, we also evaluated the summary statistics. The attributes subsription_type, watch_hours, last_login_days, monthly_fee, number_of_profiles, avg_watch_time_per_day, payment_methodCrypto, and payment_methodGift Card were statistically significant. Each one of these attributes had a p-value less than .01, indicating that they have a strong effect on churn.

B. SVM

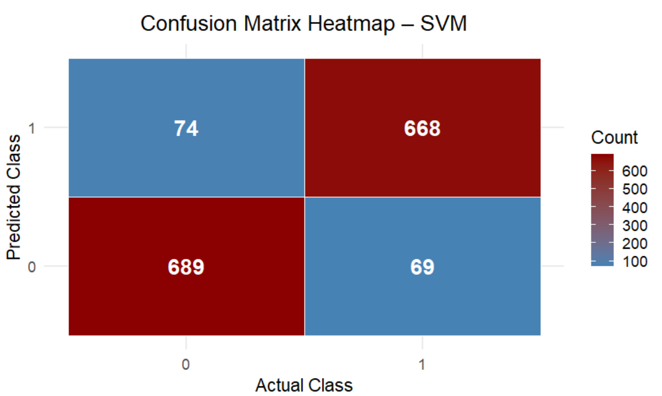

Fig. 3: SVM Confusion Matrix

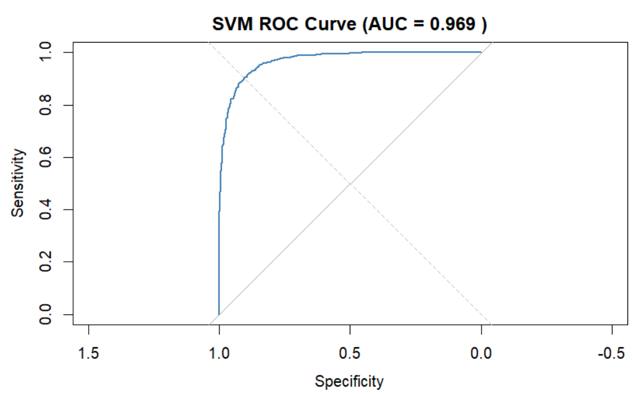

Fig. 4: SVM ROC

The next experiment we ran was the SVM model. This model produced an accuracy of .905. Meaning that 90.5% of the time, we correctly predicted whether the customer churned. Fig. 3 shows that our model correctly predicted 0 689 times and correctly predicted 1 668 times. Our model resulted in 74 false positives and 69 false negatives. Compared to our logistic regression model, we did slightly worse. To further evaluate our model, we calculated the F1 score. The SVM model had an F1 value of .903. Meaning our model had a good balance between precision and recall as we captured most customers who churn. Our model also had an AUC value of .969 as shown in Fig. 4. Again, our SVM model did slightly worse than the logistic regression model in terms of F1 score and AUC. Overall, our SVM model was still accurate and effective in predicting customer churn.

C. Random Forest

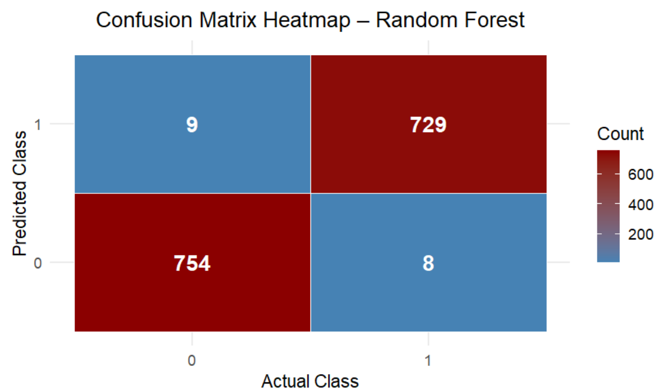

Fig. 5: Random Forest Confusion Matrix

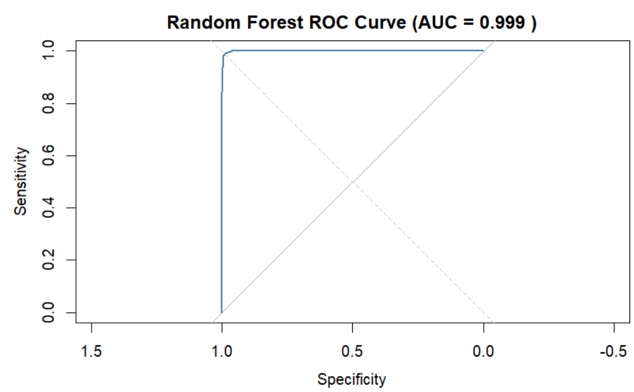

Fig. 6: Random Forest ROC

The last experiment that we ran was the Random Forest model. This model had 500 trees and had an accuracy of .989. This is extremely high and outperformed both the logistic regression and SVM models. Our model only had 9 false positives and 8 false negatives. On top of that, the F1 score was extremely high sitting at .988. The F1 score was significantly higher than the previous two models as well. The area under the ROC curve was .999. Overall, this model raised some suspicions, so we investigated further to ensure that our results were accurate by modeling the random forest feature importance table.

VI. Analysis

Our experiment consisted of training and testing three models to predict customer churn. These models included Logistic Regression, SVM, and Random Forest. The Logistic Regression and SVM models had an accuracy of .909 and .905 respectively. They were both very similar in their accuracy in predicting customer churn. However, the Random Forest model had an accuracy of .989. This model was fairly more accurate than the previous two. The Logistic Regression and SVM models both performed well in their F1 scores; having a score of .909 and .903 respectively. The Random Forest model outperformed both models again, having an F1 score of .988.

Despite the Logistic Regression model being simple, it performed really well and better than we expected. It was surprising to see the Logistic Regression model to outperform the SVM. We figured the SVM would be slightly more accurate in determining customer churn due to it being more complex. The Random Forest model also surpassed expectations. One possible explanation for this may be due to the fact that we trained the Random Forest model using the original categorical variables rather than the encoded variables. The dataset must have been too simple as well. Taking a look at Fig. 7. The left panel shows the % increase in mean squared error when each attribute is permuted. The right panel in Fig. 7 shows how much each attribute helps the variance when splitting the data. From this we can see that there are quite a few features that are very important to the model. The features that are most important include number_of_profiles, payment_method, avg_watch_time_per_day, watch_hours, and last_login_days. While avg_watch_time_per_day, watch_hours, and last_login_days dominate the churn prediction. These attributes capture inactivity specifically, which increases the accuracy of our model. Compared to the studies conducted by Mohan, Jadhav, and Singh, our results produced higher predictive results. Suggesting activity to be the most important factor in predicting Netflix customer churn.

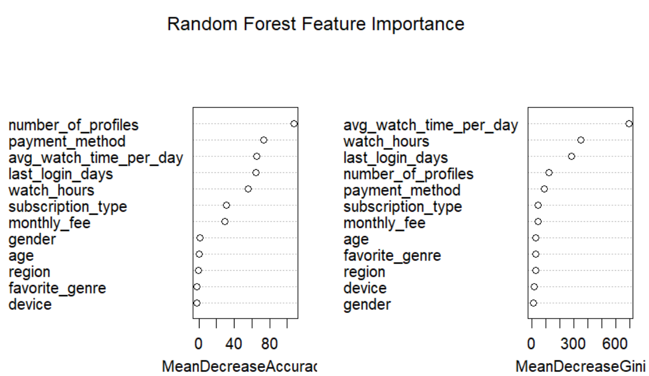

Fig. 7: Feature Importance

The study carried out by Mohan and Jadhav had limiting factors due to the dataset they used. Their dataset was obtained by a questionnaire and focused mainly on demographics, multiple streaming services, and content quality. However, Singh's study included sentiment analysis, pricing satisfaction, and activity to predict churn. This added a factor to the analysis of the data.

The dataset that we used was synthetic and may have been too simple. Our study focused more on behavioral features such as watch time, last login date, number of profiles, and payment method. Therefore, complementing Singh's work more than Mohan and Jadhav.

References

[1] Wadood, A. (2023). *Netflix Customer Churn Dataset*. Available: https://www.kaggle.com/datasets/abdulwadood11220/netflix-customer-churn-dataset

[2] M. Mohan and A. Jadhav. (2022). *Predicting Customer Churn on OTT Platforms: Customers with Subscription of Multiple Service Providers*. *Journal of Information and Organizational Sciences*, vol. 46, no. 2, pp. 433–451. Available: https://doi.org/10.31341/jios.46.2.10

[3] A. Singh. (2025). *Subscription-Based Pricing Models: Impact on Customer Retention – A Case Study of Netflix's Strategy*.

[4] GeeksforGeeks. (2023). *Categorical Data Encoding Techniques in Machine Learning*. Available: https://www.geeksforgeeks.org/machine-learning/categorical-data-encoding-techniques-in-machine-learning/

[5] GeeksforGeeks. (2023). *Encoding Categorical Data in R*. Available: https://www.geeksforgeeks.org/r-language/encoding-categorical-data-in-r/

[6] GeeksforGeeks. (2025). *F1 Score in Machine Learning*. Available: https://www.geeksforgeeks.org/machine-learning/f1-score-in-machine-learning/

[7] GeeksforGeeks. (2025). *AUC – ROC Curve in Machine Learning*. Available: https://www.geeksforgeeks.org/machine-learning/auc-roc-curve/

[8] GeeksforGeeks. (2025). *Understanding Logistic Regression*. Available: https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/

[9] GeeksforGeeks. (2025). *Random Forest Algorithm in Machine Learning*. Available: https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/

[10] GeeksforGeeks. (2025). *Support Vector Machine Algorithm in Machine Learning*. Available: https://www.geeksforgeeks.org/machine-learning/support-vector-machine-algorithm/