Module 2: Technique Practice

ALY6040 – Data Mining, Northeastern University

Professor Justin Grosz

11/19/23

Submitted by: Sudhamshu Vidyananda

Introduction

Netflix, a leader in the streaming service industry, constantly seeks to enhance viewer engagement and satisfaction. Our project revolves around a comprehensive dataset from Netflix, providing insights into viewer behaviors and preferences. The primary objective is to understand and predict factors influencing episode completion by Netflix users.

Business Problem

The central business question is: *"What features contribute to a viewer completing an episode, and can we predict whether an episode will be completed?"* Addressing this question is crucial for Netflix as it directly impacts viewer retention, content strategy, and overall user experience. By predicting episode completion, Netflix can tailor its content recommendations, improve viewer satisfaction, and ultimately drive subscription renewals and growth.

Approach and Methodology

The approach to follow to solve the business problem are Data Cleaning, Exploratory Data Analysis, Predictive Modelling, Model Comparison and Analysis and Interpretation and Recommendation.

Data Cleanup

In the realm of data analytics, particularly for a data-driven company like Netflix, the integrity and quality of data are paramount. The process of data cleaning is a critical step in preparing for robust and accurate analysis. It involves refining the dataset to ensure it accurately represents the real-world scenarios it is supposed to model. The first step was to find out the basic statistical features of the dataset and through this process, we could find out that Netflix dataset contained 6 numerical and 4 categorical values. **Season, Episode, Time Watched, Gender, Completed** and **Time of Day** had null/missing values in them. Missing values in

Season column were replaced with **mode** of Season by grouping each Show. This approach ensures that missing values are filled logically based on the viewing patterns of each specific show. By grouping by 'Show', the method ensures that the imputation is tailored and relevant. It acknowledges that different shows have different popular seasons. Missing values in **Episode** column were replaced with **mode** of Episode by grouping each Show and Season together. By grouping 'Show' and 'Season', the imputation is made contextually relevant. Different shows and seasons may have varying numbers of episodes, so this method ensures that the imputed values are sensible within the specific context of each show and season. Missing values in **Time Watched** column were replaced with **median** of Time Watched by grouping each Show with its Season and Episode. This approach acknowledges that different shows and episodes may inherently have different typical viewing times. Using the median (as opposed to the mean) is a robust approach for handling outliers. The median is less affected by extremely long or short viewing times, providing a more representative value for typical viewing duration. Missing values in **Gender** column were again replaced with **mode** of Gender by grouping each Show with its Season and Episode. By using the mode gender for specific show-episode combinations, the method minimizes bias. It acknowledges that different shows or episodes might have varying gender distributions. This ensures a more contextual and accurate imputation, reflecting the actual viewer demographics of each episode and this method avoids the introduction of arbitrary data, which can skew analysis results. Missing values in **Completed** column were replaced with **mode** of Completed by grouping each Show with its Season and Episode and by taking the average of Time Watched per Show, Season and Episode. If the average duration is >= 45 minutes, we logically assume that most viewers likely watched the entire episode. By calculating the average time watched per episode, the imputation is tailored to the viewing patterns specific

to each show and episode. This ensures a more accurate and contextually relevant imputation compared to using a global average or a static value. Missing values in **Time of Day** were replaced by **mode** of this column by grouping Day and Show with its Season and Episode. Viewing habits can vary significantly across different shows and even between different seasons or episodes of the same show. Imputing Time of Day based on Day, Show, Season, and Episode is a contextually relevant method. It recognizes that viewing times can be show-specific and may vary across different episodes and days. This method respects the inherent patterns in the dataset, avoiding arbitrary or uninformed imputation, which could lead to biased analysis. Dataset was checked to see if it contained any **special characters** which might create confusion for the stakeholders and found none. Dataset was also checked for **duplicated records** and found none. Duplicated records will disrupt the data modelling process and will not produce accurate results and needs to be removed from the dataset. Date, Season, Episode, Time Watched, Completed and Time of Day variables had **incorrect datatype** tagged to them. 'Date' conversion facilitates trend analysis over time, crucial for understanding viewer habits and seasonal patterns. Integer conversions Clarifies that 'Season', 'Episode', 'Time Watched', 'Completed', and 'Time of Day' are not just numerical but are discrete units, important for accurate counting, categorization, and interpretation in the context of shows. Dataset is thoroughly checked for outliers. Identifying **outliers** is crucial for data cleaning and preprocessing. It ensures that the dataset is ready for analysis and that the results are not skewed by extreme values. The outliers are not removed from the dataset because we will end up losing important information as the show preference and viewing duration depends on individual to individual.

Exploratory Data Analysis

EDA is done to answer the hypothesis questions and have a visual interpretation of it.

Hypothesis 1: Viewers are more likely to complete watching episodes during certain times of the day.

This hypothesis stems from the idea that viewer engagement may fluctuate throughout the day due to various factors like work schedules, daily routines, and prime time programming. To support this hypothesis, we can visualize the completion rate across different times of the day using the dataset provided. We'll create a bar chart that shows the average completion rate for each time category. **In Figure 1 (see Appendix)**, the first-time segment (0) has a slightly higher completion rate (approximately 26.61%) compared to the second time segment (1) with a completion rate of about 25.43%. This means that users prefer to watch the shows in the morning rather than at night. Although the difference is not very large, it does suggest that the time of day may have an impact on the likelihood of episode completion. Even a slight variation in completion rates can be significant at scale for a platform like Netflix. The time segment with the higher completion rate might be considered a more engaging period, potentially influencing when Netflix might release new episodes or promote certain content. These insights could lead to more targeted marketing strategies and content scheduling that align with viewers' preferences and habits, aiming to maximize viewer engagement and retention.

Hypothesis 2: "Viewing preferences for specific shows differ significantly between male and female viewers."

This hypothesis suggests that there are distinct patterns in show preferences based on gender, which could inform content recommendations and marketing strategies tailored to different demographics. A grouped bar chart showing the count or proportion of completed episodes across top shows, separated by gender, could support this hypothesis. It would visually compare male and female preferences for various shows. **In Figure 2 (see Appendix)**, Stranger Things

show have higher viewing rate among both females and males compared to other shows. The differences in bar heights across shows indicate that gender may indeed play a role in show preferences and completion rates. The graph supports this hypothesis by visually demonstrating differences in the number of completed episodes between genders for each show. The variation in completed episode counts suggests that Netflix could tailor its recommendations and promotional content based on gender to increase viewer engagement. For shows with a more pronounced gender preference, targeted marketing strategies could be developed to attract viewers from the less represented gender, potentially increasing the audience base for those shows.

Correlation Matrix

The correlation matrix heatmap provided in **Figure 3 (see Appendix),** shows the relationship between various variables in the Netflix dataset. The variables that show a stronger red color in the 'Completed' row/column might be positively correlated with episode completion. Conversely, variables that show a stronger blue color in the 'Completed' row/column might be negatively correlated with episode completion. Variables with white or lighter-colored cells in the 'Completed' row/column do not have a strong linear relationship with episode completion. 'Time Watched' has a correlation coefficient of **0.69** with 'Completed' which indicates that it is a strong predictor and can be used in a logistic regression model to predict episode completion. On the other hand, 'Time of Day' has a correlation coefficient of **-0.013** with 'Completed' which indicates that it may not be a significant predictor in the model.

Data Analysis

Data modeling plays a crucial role in various fields by providing insights into complex datasets, facilitating informed decision-making, and identifying trends and patterns. It helps

businesses optimize their operations, tailor marketing strategies, and predict customer behavior. We will be looking at 3 approaches in modeling i.e., Logistic Regression, Logistic Regression with PCA variables and Decision Tree Analysis. Using all these modeling techniques, *we will predict whether an episode will be completed or not*

Model 1: Logistic Regression

Before running the analysis, the categorical variables need to be converted to numerical to perform the analysis using label encoding. Here in **Figure 4(a) (see Appendix)**, '**Completed'** variable is taken as the target variable and '**User Id', 'Date', 'Day', 'Show', 'Season', 'Episode', 'Time Watched', 'Gender'** and **'Time of Day'** are taken as independent variables. As per the results **Figure 4(b) (see Appendix)**, we see that **Show, Season** and **Time Watche**d have p-value < 0.05. Hence, we can say that these are the significant variables to predict the completion of an episode. Among these variables, **Show** is having higher co-efficient values when compared to Season and Time Watched which indicates that Completion rate is highly dependent on the show. The specific show watched is a predictor of whether an episode is completed. **Season** is having the negative and least co-efficient value compared to Show and Time Watched which indicates that completion rate is negatively dependent on Season. As the season number increases, the likelihood of completing an episode decreases. **Time Watched** is having a positive correlation indicating that as the time watched increases, so does the likelihood of completing an episode. The 2 main factors that signify the completion of an episode is Show and Time duration. The model's Pseudo R-squared is **0.4908**, which is a measure of the model's goodness of fit. It indicates that approximately 49% of the variability in episode completion is explained by the model. Variables like User ID, Date, Day, Episode, Gender, and Time of Day do not seem to provide strong predictive power in this model. The accuracy of the test dataset is

81.2% and training dataset is 82.06%. The confusion matrix **Figure 4(d) (see Appendix)** shows that 394 of them were predicted as completed and 94 of them were predicted as not completed. The model incorrectly predicted episode as completed 53 times when the actual result was not completed. The model incorrectly predicted episode as not completed 60 times when the actual result was completed. The top right corner represents the false positives — instances where the model predicted the episode would be completed (1s) but it was not (0s). This is a type of error we may want to minimize because it indicates a prediction of completion when the episode was not completed. To avoid this, we can adjust the decision threshold. Logistic regression outputs a probability score between 0 and 1, and by default, a threshold of 0.5 is used to classify predictions. If we increase this threshold, we will get fewer false positives but potentially more false negatives which is comparatively better than having more false positives.

Model 2:Logistic Regression with PCA

PCA is a dimensionality reduction technique that transforms the original variables into a new set of variables (principal components), which are uncorrelated and ordered so that the first few retain most of the variation present in the original variables. Here in **Figure 5(a) (see Appendix)**, we see that, we have taken 2 PCA component to check the accuracy of the model. The choice of taking only two Principal Component Analysis (PCA) components for analysis is typically driven by the goal of dimensionality reduction while retaining as much of the variation in the data as possible. x1 is having p-value >0.05 and hence it is not a significant variable. x2 is having p-value < 0.05 and a positive correlation with completed target variable. The accuracy of training set is 84.31% whereas the accuracy of test set is 85.7% which is the percentage of correct predictions on new, unseen data. The fact that the second principal component (x2) is significant **Figure 5(b) (see Appendix)**, suggests that it captures some underlying patterns or

features in the data that are influential in predicting whether an episode is completed. The insignificance of the first principal component (x1) indicates it may not be capturing relevant information for the prediction task or that the variance it explains is not related to the outcome. The high significance of x2 (p-value practically 0) also suggests that this component is a strong predictor in the model. Since PCA components are combinations of the original features, x2 likely represents a specific combination of the original features that has a strong relationship with the likelihood of episode completion. The accuracy of this model is higher than model 1 which suggests that the model 2 with PCA may have a better generalization on unseen data, given the higher testing set accuracy. By reducing dimensionality, PCA may be helping the logistic regression model focus on the most informative signals from the dataset. The second model's accuracy does not drop much from the training to the testing set, which suggests that it is consistent and likely not overfitting, an indication of good model performance. The confusion matrix **Figure 5(c) (see Appendix)** shows that shows that 411 of them were predicted as completed and 104 of them were predicted as not completed. The model incorrectly predicted episode as completed 36 times when the actual result was not completed. The model incorrectly predicted episode as not completed 50 times when the actual result was completed. The false positive is much lesser in this model when compared to model 1 which shows that this is a better model, and it has correctly predicted true positive, true negative more than first model and false negative is also comparatively less indicating an overall good model.

Model 3: Decision Tree

A Decision Tree is a non-parametric supervised learning method used for classification and regression tasks. Its importance lies in its simplicity and interpretability. For Decision Tree

analysis, ‘Completed’ is taken as target variable and ‘**User Id’, ’Date’, ‘Day’, ’Show’, ’Season’, ’Episode’, ’Time Watched’, ‘Gender’** and **‘Time of Day’** are taken as independent variables. By setting the depth of decision tree to **4 Figure 6(a), 6(b) (see Appendix)**, **Time Watched, Show** and **Day** show as significant variables with Time Watched having the highest importance suggesting it to be the main driver in predicting whether an episode is completed. The model shows an extremely high level of accuracy for both training set (98.6%) and test set (98.5%) which suggests excellent model performance. However, such high accuracy can sometimes be a sign of overfitting, especially if the decision tree is deep and complex, capturing noise in the training data that does not generalize well to unseen data. In this case, though, the high-test accuracy suggests the model generalizes well because of no over-fitting. The Decision Tree model has a much higher accuracy than both logistic regression models. While the first logistic regression model had an accuracy of 81.19%, and the second with PCA had around 85%, the Decision Tree outperforms them on this metric. The classification report provides precision, recall, and F1-score for both classes. The model has high scores across all metrics for both classes, indicating a balanced performance on predicting both completed and not completed episodes. The confusion matrix **Figure 6(c) (see Appendix)** shows that shows that 439 of them were predicted as completed and 153 of them were predicted as not completed. The model incorrectly predicted episode as completed 8 times when the actual result was not completed. The model incorrectly predicted episode as not completed 1 time when the actual result was completed. The false positive is much lesser in this model when compared to model 1 and model 2 which shows that this is a best model, and it has correctly predicted true positive, true negative more than first and second model and false negative is also comparatively less indicating an overall good model.

Interpretation and Recommendation

Interpreting the results across all three models (two logistic regression models and one decision tree model) for Netflix's context suggests that **Time Watched** is the most significant predictor whether an episode will be completed. The decision tree model highlighted this variable's importance and achieved high accuracy in classification. The presence of Show as a significant variable in the decision tree model also implies that certain shows have a higher likelihood of being watched completely. These insights and focusing on these variables can guide Netflix in enhancing user engagement and content strategy. Below is the comparison of Accuracy, Precision, False Positive and False Negative of all the 3 models:

Table: Model Comparison

Model Name	Accuracy	Precision	False Positive	False Negative
Logistic Regression	81.19%	0.64	53	60
Logistic Regression(PCA Variables)	85.7%	0.74	36	50
Decision Tree	98.5%	0.95	8	1

Based on these interpretations, here are two recommendations for Netflix:

Content Personalization and Recommendation Improvements:

Given the significance of **Time Watched** variable, Netflix should refine its recommendation algorithms to prioritize the engagement time metric. This could involve promoting content that keeps viewers engaged for longer periods, suggesting that such content is more likely to be completed and enjoyed. Analyzing which specific shows are completed more often can provide

Netflix with insights into viewer preferences, guiding content creation and acquisition strategies towards genres or show types with higher completion rates.

User Engagement and Retention Strategies:

Netflix could develop features that encourage longer viewing sessions, such as seamless auto-play functionality for episodes, better content curation, or interactive features that engage viewers. Analyzing viewing patterns related to the **Time of Day** and **Day** - although these were not significant in the logistic regression models, the decision tree did not rule them out, could lead to personalized viewing suggestions tailored to when users are most likely to start and complete watching an episode. It's worth noting that while the decision tree model showed very high accuracy, Netflix should ensure these results are not due to overfitting and validate the model's performance with additional data. Moreover, in taking actions based on these recommendations, Netflix should consider the diverse preferences of its global audience and the risk of creating echo chambers where viewers are only recommended a narrow range of content types.

Conclusion

The data analysis, which included logistic regression and decision tree models, provides valuable insights into the factors influencing episode completion on Netflix. The robustness of the decision tree model, coupled with the interpretability of the logistic regression models, offers a comprehensive understanding of viewer behavior. Time Watched emerged as the most influential variable predicting episode completion. This indicates that engagement time is a strong indicator of content stickiness and viewer satisfaction. The logistic regression model with PCA suggested that certain combinations of variables captured in principal components can also predict completion rates effectively. The importance of the Show variable in the decision tree

model implies that specific content characteristics contribute to viewer engagement. Netflix should emphasize metrics that correlate with long viewing times, such as Time Watched, to enhance the recommendation engine. Leverage insights from the Show variable to inform content development and acquisition strategies, focusing on properties that drive complete views. Future work should build on these findings, incorporating additional data sources and modeling techniques to refine predictions and contribute to strategic decisions aimed at increasing viewer engagement. The goal for Netflix is to create a highly personalized, engaging viewer experience that not only recommends content that viewers are likely to start but also content that they are likely to watch through to the end.

Appendix

Figure 1: Bar plot for Completion Rate

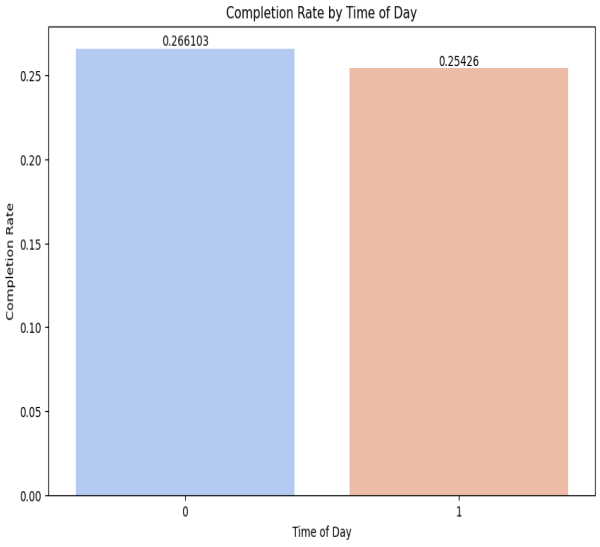
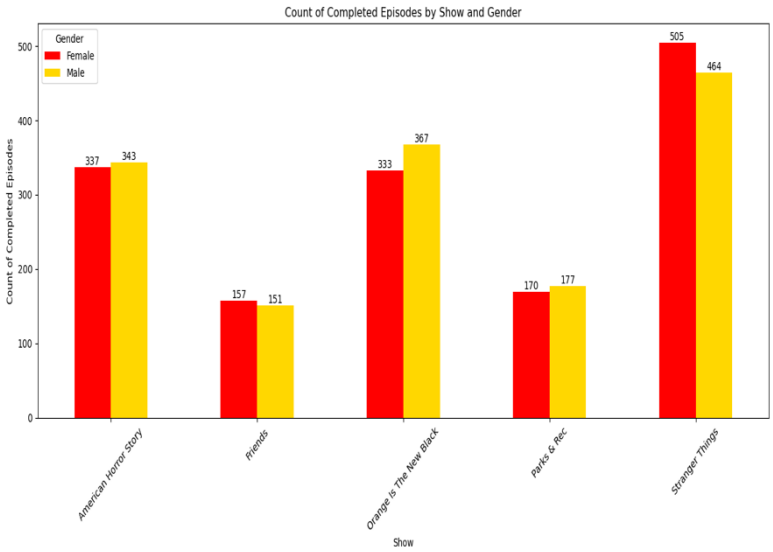


Figure 2: Episode completion by Gender



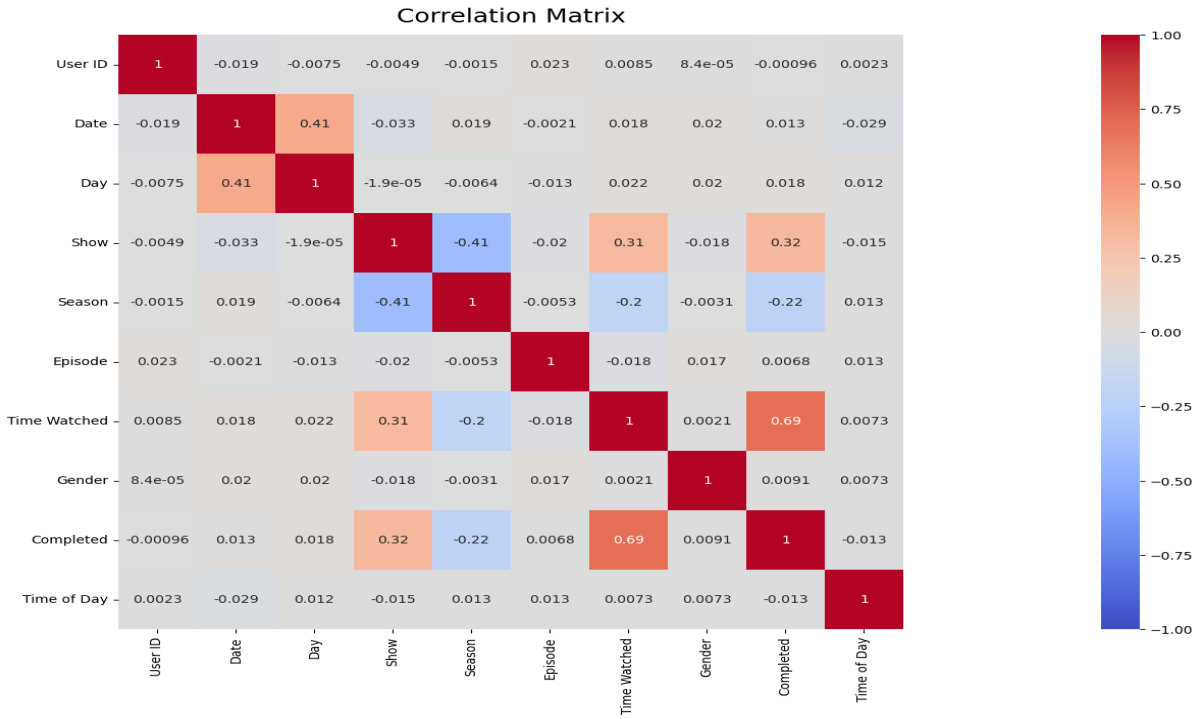


Figure 3: Correlation Matrix

```
Optimization terminated successfully.
         Current function value: 0.292510
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                Completed   No. Observations:                 2403
Model:                          Logit   Df Residuals:                     2393
Method:                           MLE   Df Model:                            9
Date:                Sun, 19 Nov 2023   Pseudo R-squ.:                  0.4908
Time:                        19:27:40   Log-Likelihood:                -702.90
converged:                       True   LL-Null:                       -1380.4
Covariance Type:            nonrobust   LLR p-value:                4.029e-286
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -5.2541      0.376    -13.979      0.000      -5.991      -4.517
User ID    -8.775e-07    1.9e-06     -0.463      0.644   -4.59e-06    2.84e-06
Date           0.0343      0.067      0.509      0.611      -0.098       0.167
Day            0.0124      0.066      0.188      0.851      -0.117       0.142
Show           0.2510      0.058      4.350      0.000       0.138       0.364
Season        -0.1746      0.054     -3.226      0.001      -0.281      -0.068
Episode        0.0168      0.024      0.695      0.487      -0.031       0.064
Time Watched   0.1323      0.006     23.020      0.000       0.121       0.144
Gender         0.1810      0.138      1.310      0.190      -0.090       0.452
Time of Day   -0.1938      0.138     -1.405      0.160      -0.464       0.076
==============================================================================
```

Figure 4(a) Logistic Regression Results

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	-5.254071	0.375850	-13.979159	2.089423e-44	-5.990724	-4.517418
Show	0.250979	0.057696	4.350012	1.361302e-05	0.137896	0.364061
Season	-0.174578	0.054124	-3.225513	1.257471e-03	-0.280659	-0.068497
Time Watched	0.132316	0.005748	23.019627	2.964827e-117	0.121050	0.143582

Figure 4(b) Significant Variables Logistic Regression

```
Training Set Accuracy: 82.06408655846859%
Testing Set Accuracy: 81.19800332778702%
 ---------------------------------------------------------------------------
Classification resport - Training Set
              precision    recall  f1-score   support

           0       0.87      0.89      0.88      1775
           1       0.67      0.63      0.65       628

    accuracy                           0.82      2403
   macro avg       0.77      0.76      0.76      2403
weighted avg       0.82      0.82      0.82      2403

Classification resport - Test Set
              precision    recall  f1-score   support

           0       0.87      0.88      0.87       447
           1       0.64      0.61      0.62       154

    accuracy                           0.81       601
   macro avg       0.75      0.75      0.75       601
weighted avg       0.81      0.81      0.81       601

 ---------------------------------------------------------------------------
```

Figure 4(c) Classification Report and Accuracy Result – Logistic Regression

```
True Positive = 394
False Positive = 53
False Negative = 60
True Negative = 94
```

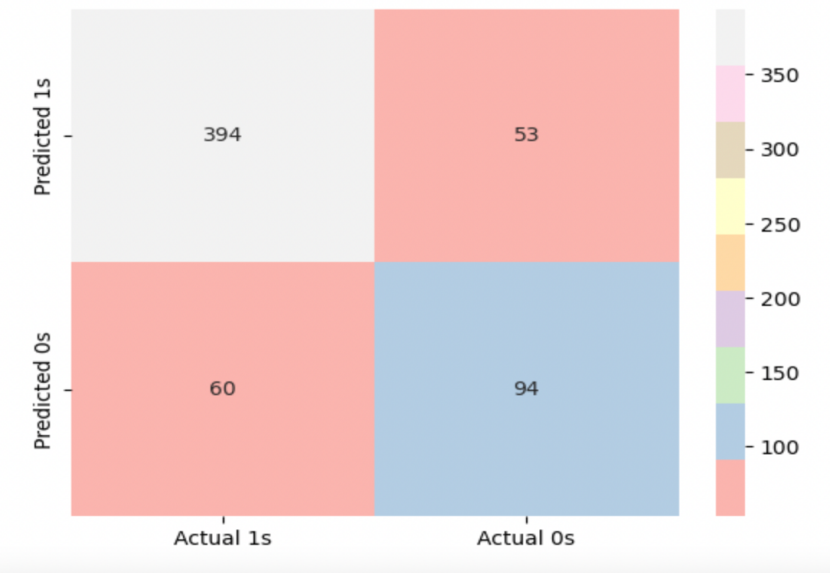


Figure 4(d) Confusion Matrix

```
Optimization terminated successfully.
         Current function value: 0.304367
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                Completed   No. Observations:                 2403
Model:                          Logit   Df Residuals:                     2400
Method:                           MLE   Df Model:                            2
Date:                Sun, 19 Nov 2023   Pseudo R-squ.:                  0.4702
Time:                        21:24:35   Log-Likelihood:                -731.39
converged:                       True   LL-Null:                       -1380.4
Covariance Type:            nonrobust   LLR p-value:                1.364e-282
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.7907      0.080    -22.361      0.000      -1.948      -1.634
x1          2.012e-07   1.85e-06      0.109      0.913   -3.42e-06    3.82e-06
x2             0.1411      0.006     24.659      0.000       0.130       0.152
==============================================================================
Training Set Accuracy: 84.31127756970453%
Testing Set Accuracy: 85.69051580698836%
 ---------------------------------------------------------------------------
Significant variable:
          Coef.  Std.Err.          z          P>|z|    [0.025    0.975]
const -1.790687  0.080079 -22.361453  9.342168e-111 -1.947639 -1.633734
x2     0.141083  0.005721  24.658947  2.950590e-134  0.129869  0.152296
 ---------------------------------------------------------------------------
```

Figure 5(a) Logistic Regression (PCA variables) Results

```
 ---------------------------------------------------------------------------
Classification Report for Training Set:
              precision    recall  f1-score   support

           0       0.88      0.92      0.90      1775
           1       0.73      0.63      0.68       628

    accuracy                           0.84      2403
   macro avg       0.80      0.77      0.79      2403
weighted avg       0.84      0.84      0.84      2403

Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.89      0.92      0.91       447
           1       0.74      0.68      0.71       154

    accuracy                           0.86       601
   macro avg       0.82      0.80      0.81       601
weighted avg       0.85      0.86      0.85       601
```

Figure 5(b) Classification Report – Logistic with PCA

```
True Positive = 411
False Positive = 36
False Negative = 50
True Negative = 104
```

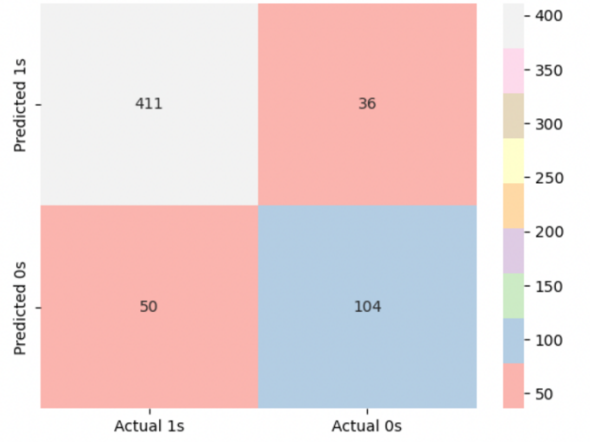


Figure 5(c) Confusion Matrix

```
Training Set Accuracy: 98.59%
Testing Set Accuracy: 98.5%
 ---------------------------------------------------------------------------
Classification Report for Training Set:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      1775
           1       0.96      0.98      0.97       628

    accuracy                           0.99      2403
   macro avg       0.98      0.98      0.98      2403
weighted avg       0.99      0.99      0.99      2403

Classification Report for Test Set:
              precision    recall  f1-score   support

           0       1.00      0.98      0.99       447
           1       0.95      0.99      0.97       154

    accuracy                           0.99       601
   macro avg       0.97      0.99      0.98       601
weighted avg       0.99      0.99      0.99       601
```

Figure 6(a) Decision Tree Classification Report and Accuracy Result

```
Significant Variables:
          Feature  Importance
6  Time Watched    0.689920
3          Show    0.309178
2           Day    0.000901
0       User ID    0.000000
1          Date    0.000000
4        Season    0.000000
5       Episode    0.000000
7        Gender    0.000000
8   Time of Day    0.000000
```

Figure 6(b) Significant Variables – Decision Tree

```
True Positive = 439
False Positive = 8
False Negative = 1
True Negative = 153
```

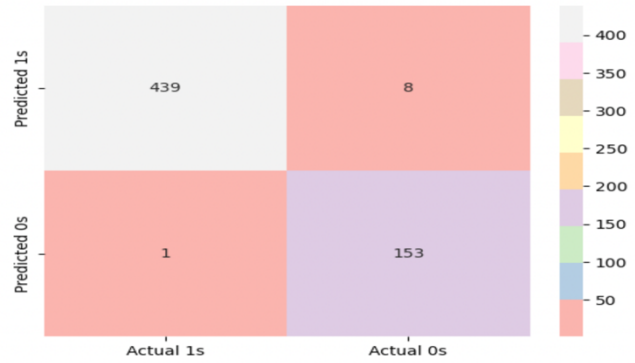


Figure 6 (c) Confusion Matrix

References

Dasu, T., & Johnson, T. (2003). *Exploratory data mining and data cleaning*. John Wiley & Sons.

Dayton, C. M. (1992). Logistic regression analysis. *Stat*, *474*, 574.

Rokach, L., & Maimon, O. (2005). Decision trees. *Data mining and knowledge discovery handbook*, 165-192.

Maćkiewicz, A., & Ratajczak, W. (1993). Principal components analysis (PCA). *Computers & Geosciences*, *19*(3), 303-342.