**Capstone Project**

**Predicted Netflix Revenue**

**By**

**Christopher Waters**

**Date: 10/08/2024**

Table of Contents

[Problem statement 3](#_Toc173493655)

[Industry/ domain 3](#_Toc173493656)

[Stakeholders 4](#_Toc173493657)

[Business question 4](#_Toc173493658)

[Data question 5](#_Toc173493659)

[Data 5](#_Toc173493660)

[Data science process 5](#_Toc173493661)

[Data analysis 5](#_Toc173493662)

[Modelling 9](#_Toc173493663)

[Outcomes 14](#_Toc173493664)

[Implementation 14](#_Toc173493665)

[Business answer 14](#_Toc173493666)

[Data answer 14](#_Toc173493667)

[Response to stakeholders 15](#_Toc173493668)

[End-to-end solution 15](#_Toc173493669)

[References 16](#_Toc173493670)

Problem statement

Netflix is the most-subscribed video on demand streaming media service with over 260 million paid memberships in more than 190 countries as of July 2024. However, with the emergence of successful alternative VoD streaming services such as Disney+ and HBO Max, as well as a recent decrease in streaming industry market share, Netflix could experience a stagnation or decrease in revenue. In 2022, Netflix experienced the shortest increase in new subscribers since 2012 (9 million).

A graph of a number of subscribers

Description automatically generated

Fig1: Number of Netflix Subscribers (2001-2023)

Industry/ domain

In recent years, Netflix's total annual expenses have been in the range of tens of billions of dollars. In 2021, Netflix’s total operating expenses were approximately $22 billion. This included content spending, technology and development, marketing, and administrative costs. In Q1 2023, Netflix maintained its position as the leading streaming platform in the US with a market share of 44.21%, down 5.56% from its market share of 49.72% in Q1 2022.

A graph showing different colored lines

Description automatically generated

Fig2: VoD Streaming Services Market Share (March 2020 – March 2023

Stakeholders

Institutional Investors

* Vanguard Group owns 36.6 million shares of Netflix, representing a stake of 8.3% of outstanding shares.
* BlackRock (BLK 2.35%) owns over 31 million shares, representing an equity stake of 6.9% of outstanding shares.
* FMR, LLC: Fidelity Management & Research Company LLC owns 21 million shares of Netflix, representing a 5.1% stake.

Individual Investors

* Reed Hastings Owns 5,426,708 shares, representing a stake of 1.25%. He has served as executive chairman of Netflix since 2023 after 25 years as CEO.
* Ted Sarandos Owns 673,889 shares, representing a stake of fewer than 1% of shares. Sarandos has served as co-CEO of Netflix since July 2020.

Business question

Based on the total revenue average Netflix users contributed throughout their subscription, what will the predicted revenue average be?

Data question

What difference would an increase in the percentage of ‘Premium’ subscribers have on the total revenue average?

Data

Data was obtained from Kaggle.com.

Download Link: <https://www.kaggle.com/datasets/arnavsmayan/netflix-userbase-dataset>

Data science process

Data analysis

The initial dataset comprised of one table with 10 columns and 2500 rows (2423 after clearing outliers). Each row represents a unique user. Data included object/categorical datatypes that described each user’s ‘Gender’ (Male, Female), ‘Subscription Type’ (Basic, Standard, or Premium), the type of ‘Device’ they used (Laptop, Smartphone, Smart TV, or Tablet), and the ‘Country’ in which they are located (Australia, Brazil, France, Germany, Italy, Mexico, New Zealand, Spain, UK, USA). Numeric datatypes included the revenue generated from each user’s monthly subscription (‘Monthly Revenue’) and ‘Age’ (26-51). The date each user joined (‘Join Date’) and the date of their last payment (‘Last Payment Date’) were also featured. No missing values were identified. An alternative to the ‘Monthly\_Revenue’ feature, entitled ‘Subscription\_Revenue’, was created due to the original feature containing six values (10-15) that were found in in all three subscription types. Cost of each subscription type were selected by determining the USD equivalent of what each country paid for each subscription type then finding the average of what all countries would pay in USD. Separate numeric columns for date, month and year were created for both the ‘Join\_Date’ and ‘Last\_Payment\_Date’ columns. Additional columns included the number of days since last payment (‘Days\_Since\_Last\_Pay’), difference in days from when each user joined to their last payment day (‘Difference\_Days’), the total number of payments per user (‘Total\_Payments’), their payment frequency (‘Pay\_Frequency’), the number of months since joining (Months Since Join), the total amount of revenue from each user (‘Total\_Revenue’) and their average monthly revenue (‘Ave\_Monthly\_Rev’). ‘Total\_Revenue’ was the chosen target variable. An end date (07/15/2023) was setup to reflect the last payment date of the dataset. All categorical features were converted to numeric (int) datatypes prior to feature selection.

Boxplots of features below

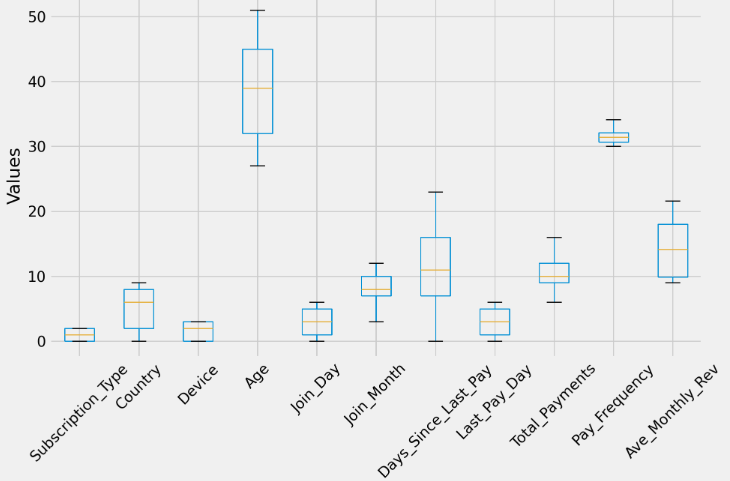


Fig3: Boxplots of ‘Subscription\_Type’, ‘Country’, ‘Device’, ‘Age’, ‘Join\_Day’, ‘Join\_Month’,

‘Days\_Since\_Last\_Pay’, ‘Last\_Pay\_Day’, ‘Total\_Payments, ‘Pay\_Frequency’ and ‘Ave\_Monthly\_Rev’

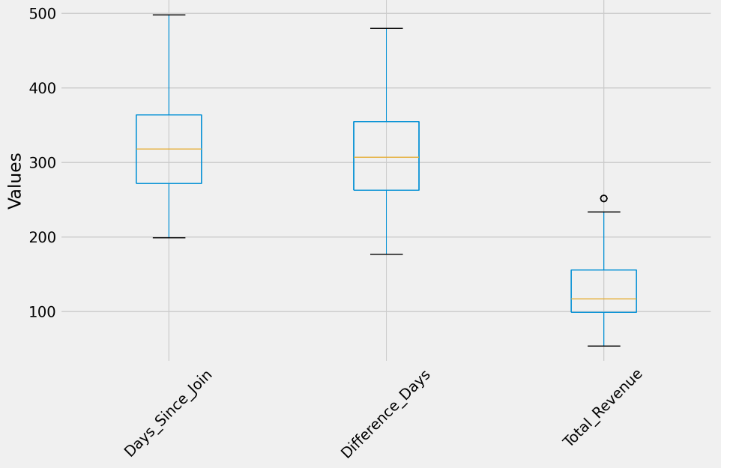


Fig4: Boxplots of ‘Days\_Since\_Join’, ‘Difference\_Days’ and ‘Total\_Revenue’

Pie charts below show the percentage of users by ‘Subscription\_Type’ and the ‘Total\_Revenue’ each type contributed.

A pie chart with text

Description automatically generated

Fig5: Percentage of Users by ‘Subscription\_Type’

A pie chart with different colored circles

Description automatically generated

Fig6: Percentage of ‘Total\_Revenue’ by ‘Subscription\_Type

A Pearson correlation heatmap was plotted to see how each feature correlated with the target variable and each other.

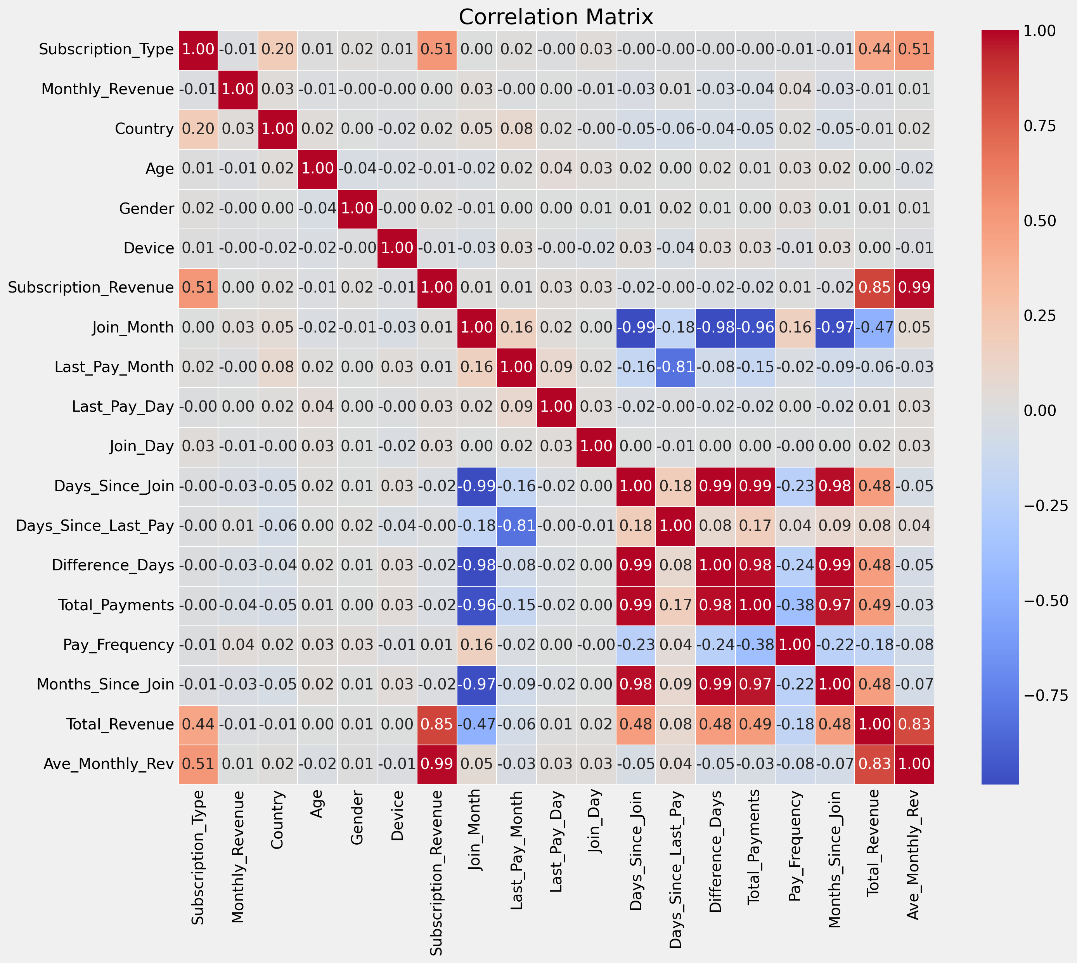


Fig7: Correlation Matrix of Features

Lasso regression was run to determine importance of each feature that showed a moderate-to-strong correlation with target.

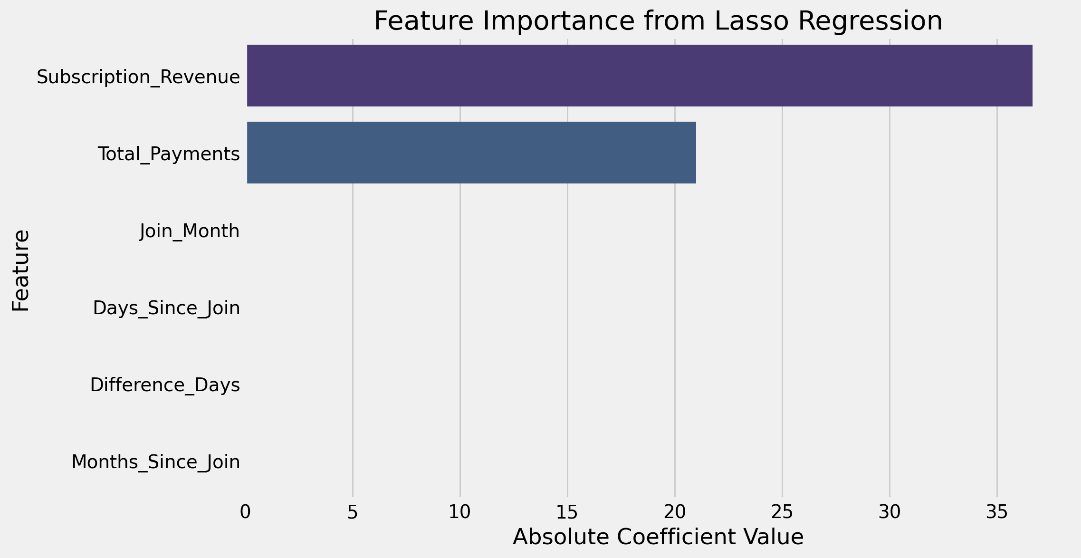


Fig8: Lasso Feature Importance

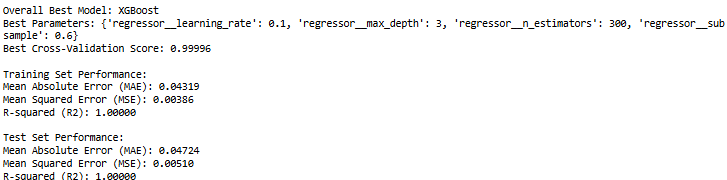
Coefficients showed ‘Subscription\_Revenue’ and ‘Total\_Payments’ were the only features to have any predictive accuracy with scores 36.66 and 21.00, respectively.

Data science process was done in Jupyter notebooks.

Modelling for Business Question

Features were selected using correlation heatmap and feature importance. Each feature that showed a correlation score of 0.1 (or higher) with the target variable was included in a lasso regularisation analysis to determine the importance of each feature on the target based on their absolute coefficient score. ‘Subscription\_Revenue’ and ‘Total\_Payments’ were the only two features that showed high predictive and were selected as the predictors. Seven regression models were used to predict model performance including Decision Tree, Random Forest, K-Nearest Neighbours (KNN), LightGBM, Gradient Boosting, XGBoost, and Support Vector Matrix (SVR). All seven models were fitted through a pipeline and analysed for which one scored the highest Cross-Validation (CV) score across 5 folds. Mean Absolute Error (MAE), Mean Squared Error (MSE) and R-squared (R2) scores were then used to review that model’s performance for accuracy. Histograms below show each model’s performance and the best model’s train and test results.

Business Question Results



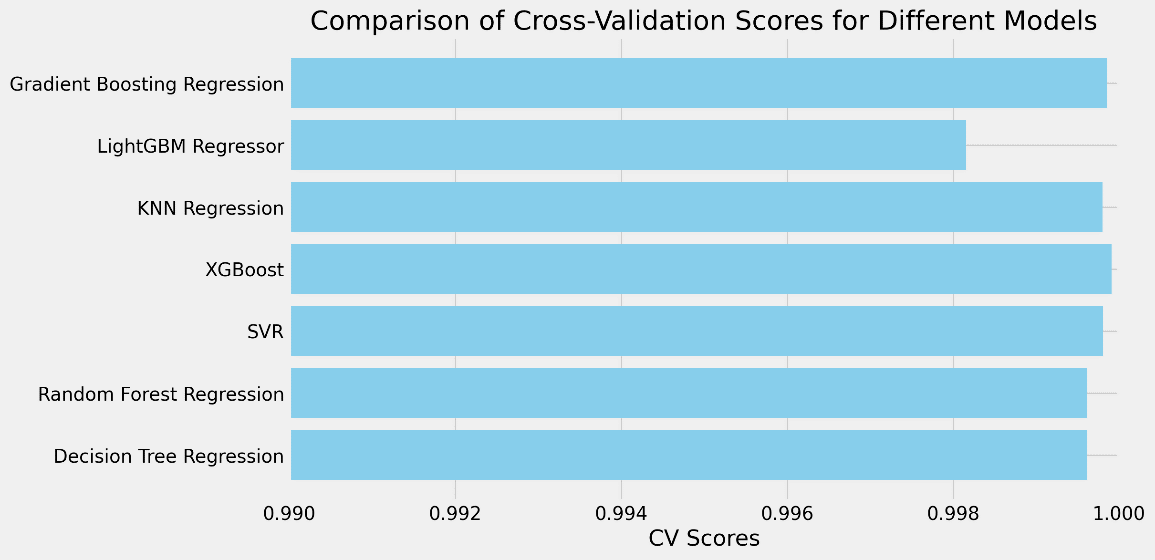


Fig9: Regression Model Cross Validation Scores

A graph of error and error

Description automatically generated with medium confidence

Fig10: MAE, MSE and R2 Scores on Train and Test Data

XGBoost model was shown to outperform the other models with a Cross Validation score of 0.99996 and R2 score of 1.0000. The small difference between the train and test sets for both MAE and MSE indicate minimal overfitting.

Scatterplot and histogram show goodness of fit between Actual and Predicted scores and distribution of residuals, respectively.

A graph of a graph

Description automatically generated with medium confidence

Residuals plot shows a bell curve spanning from -0.2 to 0.2 suggesting that the model has a good fit with errors that are small and symmetrically distributed around zero.

Histogram of Actual and Predicted total revenue averages in test set.

A red and blue bar graph

Description automatically generated

Fig11: Actual and Predicted Total Revenue Averages

No differences found between Actual and Predicted averages.

Modelling for Data Question

Data was altered to simulate the presence of an additional 74 ‘Premium’ subscribers that were ‘Standard’ in the previous dataset (a 10% increase in ‘Premium users). Data was then fitted through the pipeline again and analysed for model performance.

Data Question Results

The following pies charts show the difference upgrading Standard’ (‘13’) subscribers to ‘Premium’ (‘18’) had on Total Revenue.

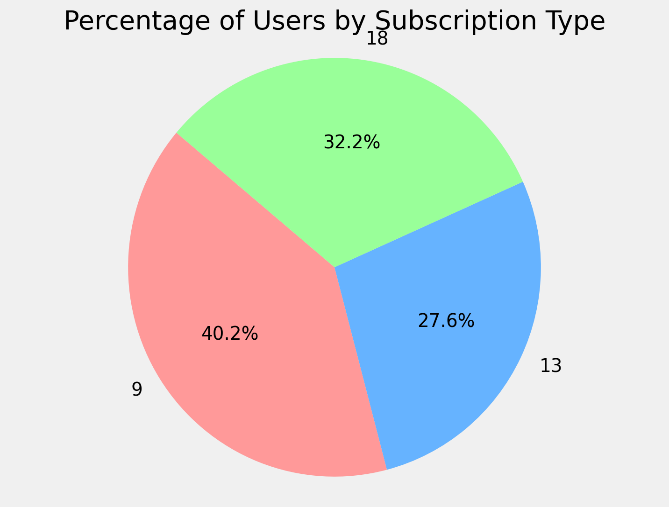


Fig12: Percentage of User by ‘Subscription Type’ (post-upgrade)

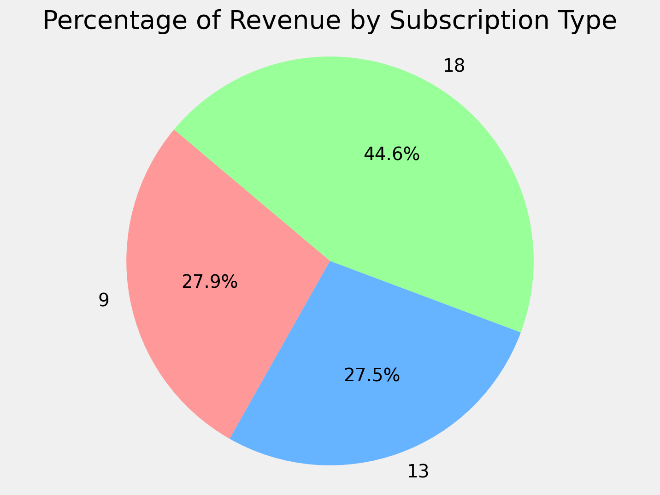


Fig13: Percentage of ‘Total Revenue’ by ‘Subscription Type’ (post-upgrade)

Histograms below display model performance following subscription upgrade.

A white background with black text

Description automatically generated

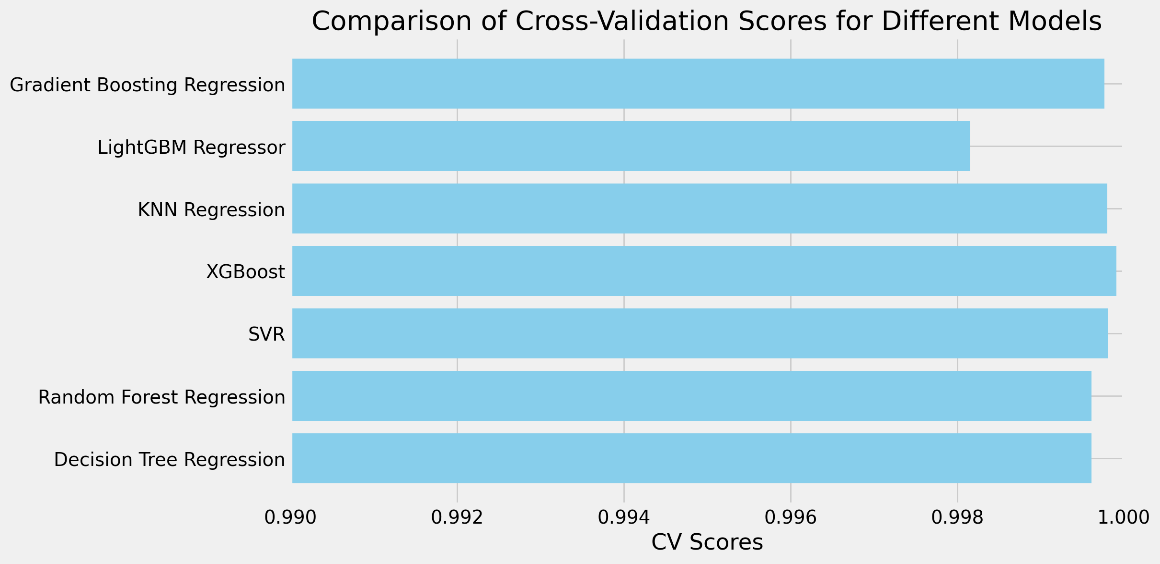


Fig14: Regression Model Cross-Validation Scores (post-upgrade)

A graph of a bar graph

Description automatically generated with medium confidence

Fig15: MAE, MSE and R2 Scores on Train and Test Data

XGBoost model once again produced the highest cross validation score of 0.99991 and R2 score of 1.0000. A decrease in the difference between train and test MSE/MAE scores indicate a reduction in overfitting.

A graph with a red line

Description automatically generated

Residuals plot shows a sustained good fit for the model.

Histogram of change in both Actual and Predicted revenue averages after increasing the number of ‘Premium’ subscribers.

A graph with red and blue rectangles

Description automatically generated

Fig16: Actual and Predicted Total Revenue Averages

Total revenue averages remained equal for both Actual and Predicted scores.

Outcomes

The EDA and feature engineering process was used to identify important features. XGBoost gave the best results on the test set out of all models for both the original and upgraded data. The process confirms the possibility to predict total revenue with high accuracy.

Implementation

Using web frameworks like Flask or FastAPI could be ideal for defining endpoints that will accept new data and process the input data using the saved preprocessing pipeline. Cloud platforms such AWS, Google Cloud, or Azure could be considered for model deployment.

Business answer

The average Predicted revenue was equal to the Actual revenue average (129.48), suggesting a consistent revenue average of $129.48 will be found for every 485 subscribers over a 9-month timeframe.

Data answer

Increasing the number of ‘Standard’ accounts to ‘Premium’ showed a 1.51% increase in both Actual and Predicted averages (130.99).

Response to stakeholders

Model performance demonstrated high accuracy in predicting total revenue average for a specified percentage of users across a specified timeline. Furthermore, analysis showed what difference an increase in subscription upgrades can have on revenue with the same number of users across the same timeframe. Implementing marketing strategies to encourage/incentivise current users to upgrade their subscription accounts is recommended.

End-to-end solution

.

Inclusion of more complex data with a larger sample across an annual timeframe with cancelled/upgraded subscriptions would be advantageous to monitoring how model would continue to perform in future.

.

References

<https://www.kaggle.com/datasets/arnavsmayan/netflix-userbase-dataset>

https://www.geeksforgeeks.org/regression-in-machine-learning/?ref=header\_outind

https://www.w3schools.com/python/python\_ml\_train\_test.asp

https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

https://backlinko.com/netflix-users

https://www.similarweb.com/blog/insights/media-entertainment-news/streaming-q1-2023/

https://www.techopedia.com/who-owns-netflix-stock