ISYE 6740 – Fall 2022
Final Project

Team Member Names: Kimberly Pierce

Project Title: Predicting the Longevity of Television Shows

Problem Statement

How we consume media has evolved exponentially over the years and with that, a change in public perception of television shows. In the early 2000's there were new crime dramas every season, but now fantasy stories like Game of Thrones prevail viewership. Although there are obvious trends in genres of shows, we want to see what kinds of factors influence longevity across multiple years. In this project, we are going to explore what, if any, factors influence the longevity of a television show and determine if a prediction is viable through regression analysis.

Data Source

When it comes to anything involving television shows or movies, the first place to find the most accurate and intensive data is IMDb. The website provides datasets that are refreshed daily and are contained in gzipped, TSV formatted files. There are seven distinct datasets found at datasets.imdbws.com, and a massive amount of data to filter through, but only four of the seven are most useful for the problem at hand, with a select number of columns from each:

- title.basics.tsv.gz contains information about each title and has nine columns, with seven being used, including:
  - tconst: a unique identifier for each title
  - titleType: the format of the title
    - filtered to only include 'tvSeries' and 'tvMiniSeries'
  - isAdult: a Boolean column, 0: non-adult, 1: adult title
  - startYear: series start year
  - endYear: series end year
  - runtimeMinutes: primary run time in minutes
  - genres: up to three genres associated with the title
    - this was split into two columns, with the third column deleted due to lack of sufficient data across the data points
- title.principals.tsv.gz contains the principal cast and crew and has six columns, with only two being used, including:
  - tconst: a unique identifier for each title (this is the same identifier as title.basics above)
  - category: category of job a person was in

        - this was changed into twelve columns of counts of primary job categories per title in order to see if the amount of people in certain job categories influence longevity
- title.episode.tsv.gz contains television episode information and has four columns, with three being used, including:
    - parentTconst: a unique identifier of the parent television series (this is the same identifier as 'tconst' from the two previous datasets)
    - seasonNumber: season number of each episode
        - this was changed to the maximum season number per television series
    - episodeNumber: episode number per season for its television series
        - this was changed to the average of the maximum number of episodes per season per series
- title.ratings.tsv.gz contains the IMDb ratings per title and has three columns, with only two columns being used, including:
    - tconst: a unique identifier for each title (the same identifier for the others above)
    - averageRating: weighted average of all the individual user ratings

Each of the descriptions can also be further explained from imdb.com/interfaces.

From the initial merging of the four datasets using the column "tconst", further manipulation of the data needed to be performed. With the "title.basics.tsv.gz" dataset, a new column was created to determine how many years each show had run since there were several datapoints that did not have seasons, such as news programs and documentaries. Furthermore, the ongoing television shows were changed to have an endYear of 2023 and the "genres" categorical variable columns were transformed into a set of 25 dummy variable columns in order to evaluate the regression models more easily.

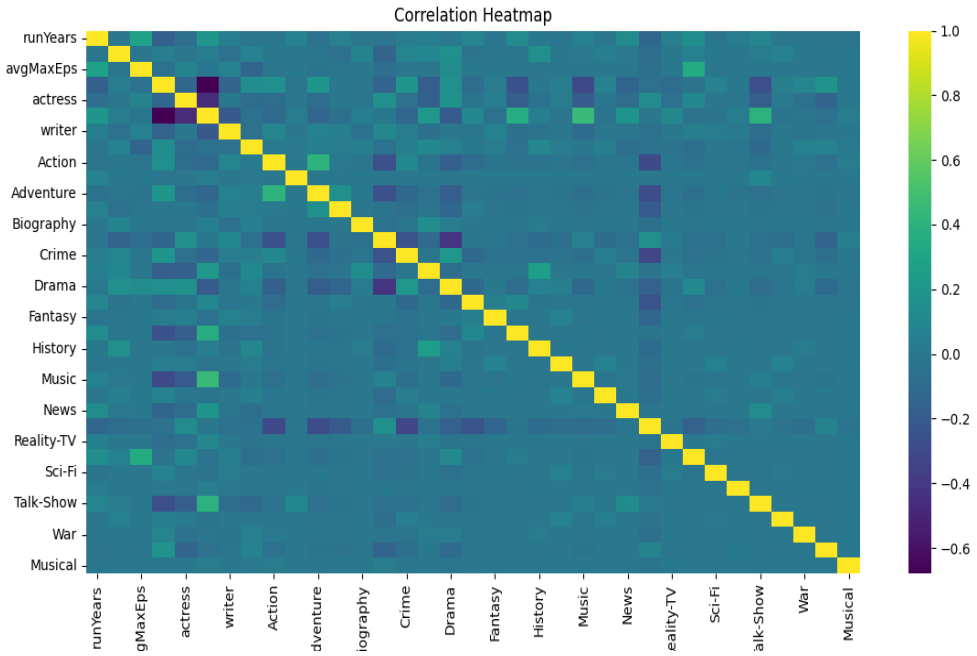
Finally, with the "title.basics.tsv.gz" and "title.principals.tsv.gz" datasets, the "isAdult" column and several of the job category columns were deleted. This is due to the columns only containing zero values once the dataset merges were completed, resulting in a dataset with 4,285 rows and 35 columns. Although this is quite a few datapoints fewer than the original datasets, there are still significant amounts of data to perform regression analysis.

Methodology

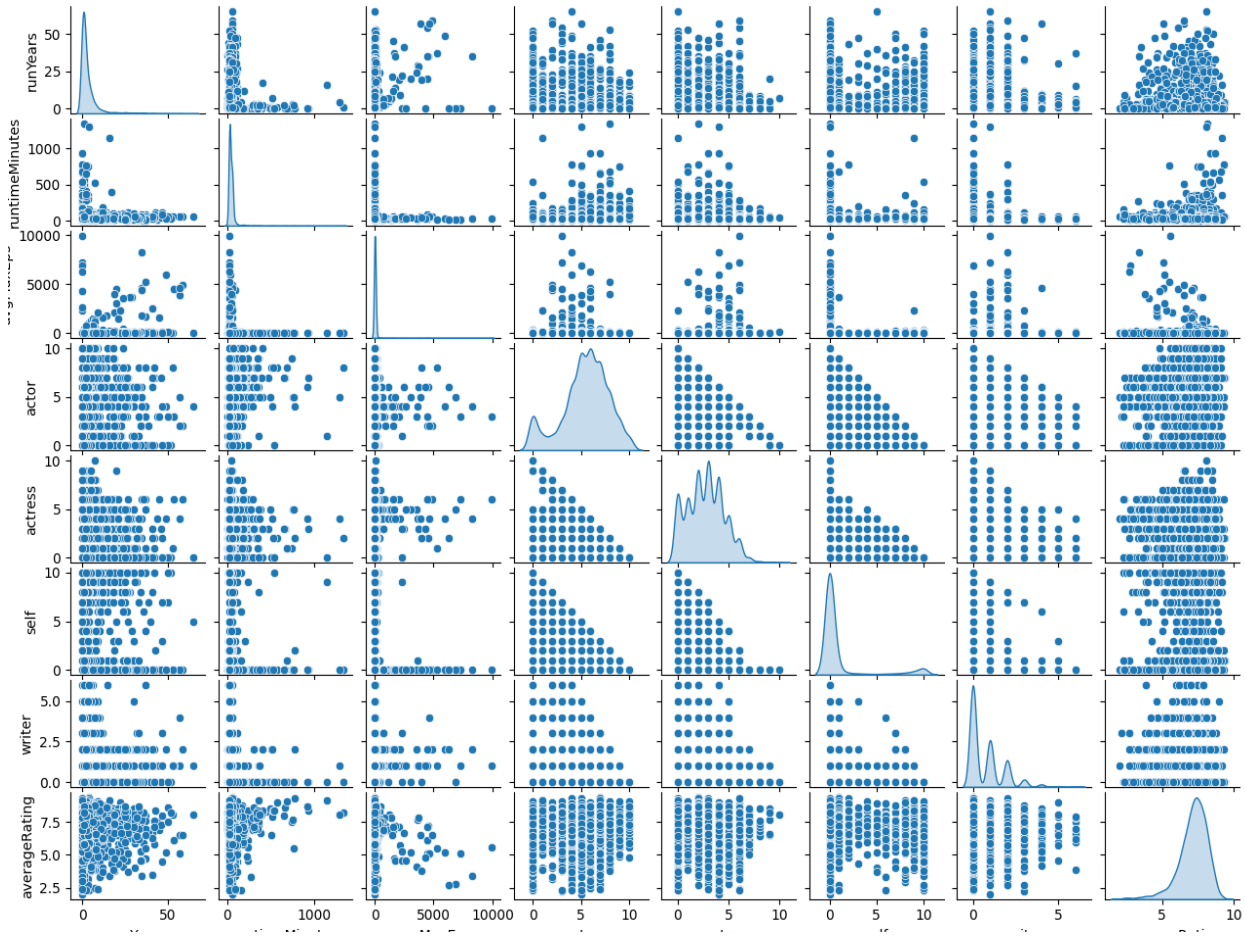
For this problem, we trialed several regression analyses to determine the predictability of long-term television shows. This was performed in three stages: Determining Correlation and Directionality, Determining Regression Model Type, and Model Estimation.

<u>Determining Correlation and Directionality</u>

Once the dataset was finalized, a general heatmap of correlation between the different variables was performed. From the plot below, it shows that there is not a significant amount of correlation between different variables but there are higher instances of correlation between ‘runYears’ and the average maximum episodes and the column ‘self’. Everything else either has a strong negative correlation or a correlation between -0.2 and 0.2.



After viewing the correlation heatmap, we split the data into an 80/20 train-test split using the sklearn package in Python for analysis. This was to ensure the randomized state remained the same throughout all different tests of regression later on. In the proposal, it was suggested to use scatterplots between the higher correlated data to see if any nonlinear transformations could occur for a better fit. For this, we ran a pairplot using Kernel Density Estimation in order to estimate the probability density function based on kernel weights of the different variables and view the scatterplots of a few of the variables. The genres were excluded solely on the fact that they are dummy variables so a scatterplot would not be a good indicator of fit in this instance.



From the KDE plots, it can be shown that a normalization of the data before running it for regression analysis should be performed. Outside of that, any further transformations have the potential to skew the data to no longer be representative of the data. This leads into determining the appropriate regression model for fitting this data.

Determining Regression Model Type

Using 'runYears' as the dependent variable, we ran Logistic, Lasso, Linear, and Ridge regressions in order to determine the best fit for the data. The goal was to determine the correct type of regression through analysis of the data and attempt different types in order to show why certain types of regression work better in this instance.

Starting with Logistic Regression, from the beginning of running this model, it was apparent that it would not be a good fit due to the fact that 'runYears' is not a categorical dependent variable and had to first be transformed using sklearn's LabelEncoder function. After running this, the adjusted R-squared value for the test data was -1.82. This shows that it is an incorrect regression model to use for this data.

From there, we ran a Lasso regression model to see if a sparser model would be a better indicator of predictability considering the amount of sparse data from the genre categories. However, there was no indication of multicollinearity between any of the variables, so Lasso was not a good fit. This is further proved after running a regression model and obtaining an adjusted R-squared value of -0.003. This is better than the Logistic model but is still a rather bad indicator of tv show longevity.

Finally there were linear and ridge regressions to consider. After looking into some of the KDE scatterplots above, a linear model could be suggested for fitting a regression model for this data. For ridge regression as with Lasso above, multicollinearity was not indicated within the data and therefore ridge would not be the best fit for the data. Furthermore, there are not many significant predictor variables so keeping all predictor variables was not necessary in this case. This leads to the choice of regression model being Linear for this project.

Model Estimation

Finally for the methodology, we used the trained model and fit it to our test data to estimate the runYears of television shows implementing sklearn's LinearRegression function. This resulted in an R-squared value of 0.17. While it is not particularly significant, it is at the very least positive, where Logistic and Lasso regression were not. The original model included all 26 genre categories, so the next step was to test the p-values for each and drop any variable that was not statistically significant. From the p-value analysis, it was found that 21 genre categories were not significant in the regression model and were not included in the training set. After evaluation, the model had ten predictor variables and the R-squared value increased to 0.20.

Evaluation

An R-squared value of 0.20 is not a significant result but is an improvement from the other types of regression that were studied in this project. The MSE value is also 0.93, which is quite high for a regression model. For this problem, it was not statistically significant enough to conclude a predictability model for longevity of television shows. The remaining factors of avgMaxEps, job categories of actor, actress, self, and writer, and the genre categories of Adult, Crime, Family, Gameshow, and News were not sufficient in providing a significant predictive model.

Final Results

There were some instances in this project that made the analysis difficult. The dataset itself had over a million datapoints before filtering to television shows and even through merging the four datasets, there was quite a lot of data to sort through for one person. In addition to this, IMDb has a vast range of data that has many missing values, including older television shows that only aired a few

episodes but had no end date in the dataset. The majority of time spent on this project was on cleaning the data, rather than analysis of regression.

In the future, there is potential for a more thorough analysis of this problem. There could be different factors that are better viewed in a Time Series analysis and given more time to accurately look into all 50,000 television shows on the IMDb dataset could show a better result than what was found in this project. For now, there is no conclusive prediction of longevity of television shows using regression analysis.