Capstone Project 2: Milestone Report 2

1. The submission demonstrates understanding of how to describe a **problem statement**.

Problem statement: Why it's a useful question to answer and for whom (source this from your proposal)

With so many new options available for customers to watch and stream their favorite movies and TV shows it is key for companies to be able to stand out by providing the pertinent content for all their customers. Being able to accurately recommend new movies and TV shows for individual users keeps the user from venturing to another provider. Being able to predict what combination of movie or TV show features, such as cast, director, genre will give the company insight to what additional content would be best to add to their library.

2. The submission demonstrates an understanding of how to describe a dataset and detail how it was **collected**, **cleaned**, and **wrangled**.

I have acquired datasets from Rotten Tomatoes, IMDb, Amazon, Netflix, and MovieLens. These datasets are vast enough to cover the scope of this project. They also allow me to break up the project into two possible routes; one focusing on a content level prediction and the other to focus on a user level prediction.

The data can be separated to accommodate being applied to either the content or user level prediction. For the content based recommendation system I will utilize the datasets from Rotten Tomatoes, IMDb, Amazon, and Netflix. These datasets are more geared towards the content. They provide information regarding movies available on different platforms, their genres, release dates, directors, cast, and ratings. The ratings given are not associated with individual users, so this doesn't allow for creating a prediction model for recommending a movie or TV show to an individual user. What it does allow us to do is create models to predict how the content would be rated based on it’s features. This helps to solve the problem of determining what currently existing content should be added to the platform. It can also provide a type of blueprint to develop new content around by predicting the combination of features that produce the best ratings.

The IMDb dataset allows for a more complex recommendation system. The dataset contains information on users, broken down by gender, age group, and location (us or non-us). From this dataset we can develop a recommendation model on the demographic level to predict how a certain demographic would rate a movie or TV show that they have not yet watched. This helps to solve relatively the same problem as above, but on the individual level. It allows a company like Netflix to determine what movies or TV shows to recommend that an individual user from a certain demographic should watch next. In doing so creating higher customer satisfaction and increasing customer retention.

3. The submission demonstrates successful application of **exploratory data analysis** (visualization and inferential statistics), for example histograms, scatter plots and hypothesis testing appropriately. And successful application of **machine learning.**

In my exploratory data analysis I used the plot_tree function on a smaller group of the data set, that of the male only demographic, in order to visualize the decision tree breakdown. Next, to keep my decision tree plot from being too large, I set the num_trees equal to 1. This showed the decision tree breakdown from ‘males_allages_avg_vote’ to ‘males_allages_votes’ and then to ‘males_18age_avg_vote.’ This breakdown shows how the decisions are automatically mapped in order of importance. To confirm this I then created a bar chart displaying the feature importance using;
plt.bar(range(len(model_0.feature_importances_)), model_0.feature_importances_)

This shows us the level of importance based on the features index. It has index 0 as having the highest level of importance with a value around 0.8 or 80%. This makes sense considering that feature is ‘males_allages_avg_vote.’ Now this can be considered an overfitting issue, since we are creating a model to predict the average vote based on a set of features using the average vote of all males doesn’t give us much insight into what features are good at predicting the average vote. From here we can make decisions on if we should omit these types of features in order to create a more useful model.

4. The submission demonstrates successful application of **data storytelling** techniques to articulate hypotheses and inferences, appropriate to their target audience.

Based on the findings from our decision tree plot and feature importances it’s clear that some of the features in our datasets need to be removed in order to tell a meaningful story. Once removing the features that the model was overly reliant on we were able to find more meaningful data out of the feature importances bar charts. For example in the model using only the male demographic features, after removing the ‘‘males_allages_avg_vote’ feature we found that the age groups of 30-45 and 45+ gave us the highest level of importance in predicting the average vote of a certain film.

From here we can flip the model to take in these values as our dependent variable to discover what features of the film have the highest level of importance to these particular demographics. This then would be a good indicator of what film features would correspond to higher average vote scores. Having this information allows us to know what types of films should be recommended to these specific demographics.

So, to tell this story, I took all of my object column features and created a model with ‘males_45age_avg_vote’ as my dependent variable to find the most important features associated with this demographic. To save time, I only used 1000 rows of my total dataset, but I was still able to get a respectable root mean squared error of 0.86073 so I could be confident in my results. After running the feature importance function on my model, then cleaning the data so that I could find the features that were associated with the column numbers, I was able to find the top 5 features for predicting the ‘males_45age_avg_vote.’ These five are Animation, Germany, Twentieth Century Fox, Universal Pictures, and Sam Katzman Production.