Individual project report Data mining

Intro:

Our project used an extensive movies dataset taken from IMDb and contained metadata about movies, actors, and public reviews. We decided to model the ratings of individual movies to see if we could predict how a movie would perform based on meta data such as duration, actors, and budget. We cleaned and transformed our dataset in a preprocessing pipeline, modeled it, and organized the results in a presentation GUI.

We broke up the work for this project into 3 sections for our 3 group members. I was responsible for the preprocessing, Josh was responsible for the modeling, and Adam was responsible for the GUI.

Individual Work:

I worked on the preprocessing functions to clean our dataset. Once we had come to a conclusion about what variables we were going to keep I began work on cleaning the datasets.

This involved 4 steps:

1. Combining and subsetting the data
2. Transforming
3. Imputing
4. Scaling

Combining and subsetting was straightforward, I began with a function that loaded all of our data so that we could begin the process. This function was eventually overwritten by Adam who wrote a function that pulled our data from the internet, however this load function remained as a backup if the internet load did not work.

I then wrote 2 cleaning functions.

Improvements:

* We could do NLP on the description and title of the movies