Francois Bailly and Walker Kirk

MIS545

August 2nd, 2021

Project Description

1. **Problem Statement**

With thousands of movies coming out every year, it is becoming increasingly difficult for producers, writers, and directors to make educated decisions when it comes to the types of movies they need to work on, or for actors to decide whether or not to participate in a movie.

The recent COVID-19 crisis has resulted in lower movie theater attendance and with the proliferation of streaming services, it is pivotal that filmmakers make pictures that appeal to the public’s preferences to drive higher audiences. We aim to develop an analysis that will be able to find correlation between different attributes within a wide array of movies.

No.1: Can users be categorized based on their ratings?

No.2: Are tags a better predictor of movie ratings than the movie genre?

No.3: Can we predict the rating of a future movie using certain attributes and the number of voting?

1. **Dataset description**

The data being used is from MovieLens, which is available from the MovieLens.org website (<https://grouplens.org/datasets/movielens/>). The version of the dataset that we selected is the smallest one, which contains 100,000 ratings and 3,600 tag applications applied to 9,000 movies by 600 users. The entire MovieLens dataset consists of 4 relational files. Each subset is contained in a comma-separated text file (csv). 3 subsets are selected for this study: movies, ratings, and tags.

Table 1. Lists the dataset composition.

| MoveLiens Subset | Field Name | Data Type | Data Format | Description |
| --- | --- | --- | --- | --- |
| movies | movieId | Independent | numeric | Movie unique identifier. |
| title | Independent | string | Title of the movie. |
| genres | Independent | Pipe delimited string array | List of genres related to a given movie. |
| ratings | userId | Independent | numeric | User unique identifier. |
| movieId | Independent | numeric | Movie unique identifier. |
| rating | Independent | numeric | Rating of the movie (1-5) |
| timestamp | Independent | date | Date at which the rating was given. |
| tags | userId | Independent | numeric | User unique identifier. |
| movieId | Independent | numeric | Movie unique identifier. |
| tag | Independent | string | Tag associated with a given movie. |
| timestamp | Independent | date | Date at which a given user provided a given tag for a movie. |

1. **Data mining techniques used**

For this project, we chose to use two predictive methods; Naïve Bayes, and Support Vector Machine. Naïve Bayes is particularly well suited for this case study, as it performs well in cases of categorical inputs, and is widely used for predictions and forecasting. Support Vector Machine on the other hand, is useful for classification, regression, and outlier detection. This will be useful for us to answer whether or not users can be categorized based on their ratings.

In addition, we will also use cluster analysis as a descriptive method, so that we can group similar users together. In our case, the dependent variable will be *rating*, as this is the variable that we are trying to predict. *Tag* and *genre will be the* independent variables that help us to predict *rating*.