**Final Project Milestone Report**

**Introduction:**

This project aims to measure how reviews and ratings of movies released in theatres relate to their sales at the box office. So, we are using the dataset which consists of data for 45,000 movies released on or before July 2017. It is a compilation of data collected from The Movie Database (TMDb) and the GroupLens website and includes figures for cast, crew, plot keywords, budget, revenue, posters, release dates, languages, and more. The dataset is complemented with ratings from users of the GroupLens website.

This dataset is divided into the following six files:

* Movie metadata: Contains information on the 45,000 movies, including budget, revenue, release dates, and languages.
* Movie plot keywords: Keywords from plots of all the movies.
* Movie credits: Cast and crew information for all the movies.
* Reference keys: The unique identifiers for all the movies in the TMDb and IMDb movie databases.
* Reference keys subset: The unique identifiers for a subset of the movies (9,000) in the TMDb and IMDb movie databases.
* Movie ratings: 100,000 ratings for a subset of the movies (9,000) from 700 users.

We believe this dataset is correct because GroupLens has retrieved the data from MovieLens, which is a movie recommendation service. This data was created by selected users of MovieLens and each selected user has rated atleast 20 movies.

We are taking the US Unemployment Dataset as a contextual dataset. This dataset, which was last updated one year ago, contains a collection of unemployment rates in the US, dating from January 2010 until 2020, grouped by categories (education level, race, and gender). This dataset contains two files one indicating the unemployment rate of the US and the other showing unemployment rate of the US in state wise. We trust this data is correct as it was put together with U-3 rate data collected from the U.S. Bureau of Labour Statistics (BLS) and is to be combined with the overall unemployment data that can be retrieved from the BLS website at<https://data.bls.gov/cgi-bin/surveymost?bls>. This is a trusted source of data as BLS is a government agency that collects data, perform analysis, and provide results to US public and other government agencies.

Below are the columns in the file unemployment rate of the US:

* Year: Represents year in which an unemployment rate is calculated.
* Month: Represents month in which an unemployment rate is calculated.
* Primary school: Unemployment rate based on education qualifications of adults.
* Date: Merging of month and year columns.
* High school: Unemployment rate based on education qualifications of adults.
* Associates Degree: Unemployment rate based on education qualifications of adults.
* Professional Degree: Unemployment rate based on education qualifications of adults.
* White: Unemployment rate based on race of adults.
* Black: Unemployment rate based on race of adults.
* Asian: Unemployment rate based on race of adults.
* Hispanic: Unemployment rate based on race of adults.
* Men: Unemployment rate based on gender of adults.
* Women: Unemployment rate based on gender of adults.

Below are the columns in the file unemployment rate of the US in state wise:

* State: Represents states in United States
* Month: Represents Month of the year
* Unemployment: Unemployment rate in a particular state in the specified month.
* Date: Date when unemployment rate was recorded.

**Data Adequacy:**

In the real-world processing and analysis of the data can be a difficult task and to know whether our data is adequate or if we can trust the conclusions, we need to be sure about accuracy. The most important thing is to get deep and familiar with the data and dig into any mistakes.

To ensure the results are adequate, we will be performing following steps:

1. **Separate data from analysis and make analysis repeatable.**

It is the best practice to separate the data and to analyze the process. This also makes it easy to analyze different data and get the results.

1. **If possible, check your data against another source.**

To make sure the data is adequate we must compare it with multiple sources.

1. **Get down and dirty with the data.**

Play with the data and do some exploratory data analysis. Ask questions about the data and run some checks on it.

1. **Unit testing of code.**

Unit testing is a great way to check for problems at each level in the model. It is a good idea, but it can complicate if the process is too long

1. **Document your process**.

This step can be monotonous, but it will help in the future. Documentation communicates to others to check your logic and validate it.

1. **Get feedback from others**

Review from others is a very important task as other people’s insights can be useful in improving the model.

The Movies Dataset has the potential to determine how unemployment rate affects the box office revenues, the changes in revenues with respect to year, how movie plot affects the box office collection, how factors like unemployment, movie plot, budget affects the box office collections. Apart from this, the dataset consists metadata of more than 45000 movies where the data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages which can answer every question listed below and hence the provided dataset is adequate.

The dataset is collected from TMDB and GroupLens. The Movie Details, Credits and Keywords have been collected from the TMDB Open API. This dataset uses TMDb API but it is not embraced or certified by TMDb. Likewise, their API also provides admittance to information on numerous movies, actors and actresses, crew members, and TV shows. The Movie Links and Ratings have been obtained from the Official GroupLens website.

The data is a representative sample of the population. It accurately represents and reflects the characteristics of the population.

**Will this data enable you to develop models that are useful beyond the data used in this project?**

A model is most useful when integrated into the real-life context of organizational operations and delivers real business value. This purpose should be the end goal of a data science project.

With this end in mind, the project will build a reusable data model, which is devised to be independent of the specific datasets used for its conception.  This model will focus on providing business value by predicting the answer to one particular and relevant business question: *Can box office revenue be anticipated using movie ratings by potential moviegoers?*

Not being constrained by any specific data set makes this predictive model valuable to movie studios beyond the confines of the project scope. For instance, movie studios can deploy this model as part of an automated data analysis pipeline and repeatedly exploit it to estimate box office revenue, even as the underlying data changes.

We also contemplate derivative applications of this predictive model. For example, with an enriched data set, this model can correlate box office revenue with specific movie characteristics such as duration, genres, plot peculiarities, time periods, etc.

**Limitations of the data:**

The used datasets can contribute to understanding the relationship between a movie’s general perception by movie lovers and its box office results. However, we must still recognize several limitations in the data, which include the following:

* **Sample demographics**  
  Without demographic information to categorize the movie raters, it is impossible to determine how well the dataset represents the general moviegoer population.

* **Response bias**  
  Raters provided online movie ratings mostly on an unassisted and voluntary basis. Thus, certain biased groups (franchise devotees, fans of specific actors, enthusiasts of competing franchises, etc.) may have been more likely than other groups to provide a spontaneous rating.

* **Population demographics**  
  For best results, the demographics of the movie raters should reflect the demographics of the moviegoers. Unfortunately, the included data sets do not provide insights into the composition of the moviegoer population as their demographics are not available.

* **Collection environment**  
  Registered users of the MovieLens website provided the movie ratings. The online channel, coupled with the registration requirement, significantly restricted the number and the type of people who contributed.

* **Box office results influence.**  
  With movie ratings provided concurrently with the publishing of box office figures, the result we are trying to predict may have influenced some of the ratings. Some people may have rated more positively the movies doing well in the box office and vice-versa.

**Questions on the dataset:**

* The main question is to find out the relationship between unemployment rate vs the Box office revenue. How does unemployment affect the box office revenue?
* How has the Box office revenue improved or decreased in the available data with respect to the year?
* One of the primary tasks is to investigate how the unemployment rate affects the movie rating, by considering the movies of the same time frame.
* Does movie plot affect the box office revenue?
* The next question is how a movie plot and the rating of the movie affects the unemployment rate. Will the employees be inspired by an Inspirational movie and pursue their passion by leaving their jobs or staying on their current path?
* How the level of unemployment affects the rating of the movies released in the same time frame.
* How the cast of the movie with respect to movie plot affects the unemployment rate
* How the budget of the movies affects the unemployment rate within the same time frame.
* We can investigate if there is a relation between number of votes and unemployment rate.

**Model assessment:**

We will assess the models using the cross-validation method in which it involves partitioning the original observation dataset into a training set, used to train the model, and an independent set used to evaluate the analysis. We will use the data provided in the movie dataset to check whether the rating and reviews did match the box office revenue numbers or not. The data is adequate to test the performance of the model, we have a good amount of data to subset into various parts to create different sets for testing/ evaluating the model.

Limitations of testing the model:

* Linear relationships are only considered.
* There may be variables (like timeline of movie) other than revenue and reviews which are not studied, yet do influence the response variable
* Changes to the model might result in set of tests altogether.

**Team Members:**

Anusha Dasari

Bhavi Shah

Mauro Jasse

Sai Ruchith Reddy Ginnavaram

Vinjal Doshi