**Term Project Proposal**

**1) Shixin Li -** [**sl3368@nyu.edu**](mailto:sl3368@nyu.edu)

**2) Li Hao Liu -** [**lhl302@nyu.edu**](mailto:lhl302@nyu.edu)

**3) Kashyap Uppuluri -** [**ku328@nyu.edu**](mailto:ku328@nyu.edu)

**4) Rohit Wankhede -** [**rnw245@nyu.edu**](mailto:rnw245@nyu.edu)

**Business Problem**

* What is the exact business problem?
  + By using the IMDB dataset consisting of information about movies and the prior user ratings for these movies we plan to classify the user base into segments and help devise marketing strategies to target specific segments and ensure a maximum turnout.
  + IMDB offers audiences information about movies. However, we provide producers and directors a way to improve their movies or knowing how their movies will perform in the market.
  + In short, what we want to do is to predict a roughly rating of a new movie based on the IMDB dataset.
* How exactly would it add business value?
  + Help the producers make a lot of money by identifying the segments that might give the highest ratings and target them to ensure maximum turnout yielding high revenue.
  + Before making the movies, producers can know what kind of traits will attract more audiences to watch the movies. After the movie is done, producers can use this model to predict the box office and ratings of the movies.
* What is the use scenario? In other words, in what business context would your solution be deployed?
  + It can be applied in both pre-production and post-production phases to fine-tune features like what actors appeal, length of the movie and more to determine an ideal set of features that gives a high rating.

**Modeling Problem**

* What precisely is the data-mining problem? Be as articulate and clear as possible. Remember that business problems might be general, but data mining problems need to be extremely precise.
  + One of the initial data mining task is to find the set of features which correlate with the user rating of a movie. i.e Feature Selection. On the basis of the features identified a prediction model with the best accuracy has to be devised.
* Is it supervised or unsupervised?
  + Supervised. Target: User Rating.
* What is a data instance or record?
  + A data instance in rating.list

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| New | Distribution | Votes | Rank | Title |
|  | 0000000125 | 1057644 | 9.2 | The Godfather (1972) |
| Note: In this list, movies have been rated on a scale of 1 to 10, 10 being good and 1 being bad. For each movie, the total number of votes, the average rating, and the vote distribution are shown. New movies are indicated by a "\*" before their entry. The vote distribution uses a single character to represent the percentage of votes for each ranking. The following characters codes can appear:  "." no votes cast "3" 30-39% of the votes "7" 70-79% of the votes  "0" 1-9% of the votes "4" 40-49% of the votes "8" 80-89% of the votes  "1" 10-19% of the votes "5" 50-59% of the votes "9" 90-99% of the votes  "2" 20-29% of the votes "6" 60-69% of the votes "\*" 100% of the votes | | | | |

* A data instance in genre.list

|  |  |
| --- | --- |
| Title | Genre |
| The Godfather (1972) | Crime |
| The Godfather (1972) | Drama |
| Genres are the common categories in which we can classify movies, there aren't that many of them. The name and number of Genres are subject to change over time. | |

* What might be the target variable?
  + User rating of prospective movie watchers (IMDB users who rate).
* What features would be useful?
  + Movie description, Length of the movie, genre and more…

**Data Details**

* Give a short description of the data you are planning to use.
  + We are going to use the dataset from the IMDb website. Such as: movie ratings, actors, actresses, genres, directors and more...
* Have you obtained data?
  + Yes, from the IMDb database. It provides Plain Text Data Files for us to analyze.
  + http://www.imdb.com/interfaces
* What is the general size of the data (records, columns, bytes, etc.)
  + Depend on the list we have to use.
  + For example, the ratings list is 43.4MB and contains top 250 movies, bottom 10 movies and more than 660,000 records of movies rating report.
  + Genres list is 76.5MB, it contains descriptions and more than 2,000,000 records of the genres of movies.
* Is the data set already structured with features or will there be feature engineering involved?
  + The data is structured to an extent, but is scattered across different datasets. One of the feature engineering tasks is to collate all relevant features into a single dataset.
  + From the IMDb website, each plain text data file contains a specific feature. For example, genres.list provides us a list of each movie’s genre.