Phase 3 - MetaData

Anuradha Nitin Bhave
ab5890@rit.edu

Himanshi Chetwani
hc9165@rit.edu

Omkar Vaidya
ov5232@rit.edu

ABSTRACT

This document provides an account of the various process that were a part of this project, from data extraction to analysis of the results obtained. This project works with movie data sets, and implements data mining techniques to identify relations between the success of the movie and its features.

1. INTRODUCTION

Movie industry is one of the biggest industries in the world and generates a huge amount of revenue. For every movie released, its success depends upon various factors such as cast, reviews, genre, language, etc. The Internet Movie Database also known as IMDB[1] is an enormous online database and contains information about different movies, television programs, etc. and includes information related to cast, production crew, rating, reviews, etc. The project implemented consists of movie datasets from IMDB and MovieLens[2] which together amounts to almost 0.9 million records described using 11 attributes. In order to determine whether a movie will be a success or not it is important to understand its features and different factors that affect it. The main aim is to find patterns, trends, and relationship between attributes and making a prediction about whether the movie will be a success or not based on the insight collected.

Section 2 says about the domain of the project. Section 3 explains about the choice of the data sets. Section 4 describes the data sets in detail. Section 5 explains about the attributes selected. Section 6 says about the motivation behind the project. Section 7 deals with data handling. Section 8 explains the design of the project. Section 9 talks about the data mining tasks selected. Section 10 contains visualization of data. Section 11 explains the implementation of the data mining tasks. Section 12 contains the results of the implementation. Section 13 talks about how the CRISP methodology was followed in the project. Section 14 says about the lessons learned during the project. Section 15 tells about the project is doing when its complete now. Section 15 talks about the future work that can be done.

2. DATA SET DOMAIN - TOPIC OF THE PROJECT

The domain of the data sets is entertainment industry particularly feature films.

3. CHOICE OF DATA SETS

The choice of data sets was largely influenced by how much metadata about these movies is present in the data. The first set was the IMDb data set [1], which is a subset of IMDb data, and is freely accessible. This data set is refreshed daily, and it holds records of various movies, tv shows, documentaries and short films.It was filtered to only include information about movies and short films. The second set was the MovieLens data set [2], which contained records of movies, along with their tags, and user ratings.

4. DATA SET DESCRIPTION

4.1 Description of IMDb Data set

The IMDB data set [1] is, in itself, a collection of data sets which describe movies, short films, documentaries, and TV shows through their run time, cast and crew, average rating and number of votes received on IMDb. It is freely accessible. This data set is refreshed daily, and it holds records of various movies, tv shows, documentaries and short films.

4.2 Description of MovieLens data set

The MovieLens Data set [2] contains records which describe a movie, the ratings it receives and the tags associated with that movie. Tags are given by users, and are often single lines that describe the movie via its genre, actor, or director.

5. CHOICE OF ATTRIBUTES

After charting out the project requirements, and evaluating attributes at hand,the following attributes were chosen to be a part of the final merged data set.

- ID : A unique identifier for every record in the data set. This is a numeric monotonic attribute.
- Title : A string representing the name of the movie used when the movie was released.
- Genre : A string array representing up to three genres of the movie.
- Language : A string representing the language of the title
- Region : A string representing the region of origin of the title.
- Title Type : The type/format of the title

- numVotes : Numeric attribute, represents the number of votes a movie receives on IMDb[1].
- averageRating : Numeric attribute, average of averageRating which represents the average rating of the movie on IMDb and the rating of a movie on IMDb.
- Runtime : Numeric attribute, represents the total running time of the movie in minutes.
- isAdult : Boolean attribute to represent PG rating of the title, 0 if not an adult title, 1 if adult title
- isOriginalTitle : Boolean attribute to represent if a title is an original or not, 0 for a title that is not original and 1 for a title that is original.

6. PROJECT MOTIVATION

In the present scenario, movies are a great form of entertainment. There are some movies which tend to be superhit on the box office while some do not turn out to be that great. There are various factors that affect the success of the movie and are based on various features such as the cast of the movie, genre, reviews, ratings, etc. So the main aim of the project is to develop and understand the insights about these features of the movie.

7. DATA HANDLING

7.1 Data Collection

The data was collected from the two sources mentioned in the Choice of Data Sets Section. [1] is a collection of data sets in itself, and all sets pertaining to the descriptive information associated with a movie such as its run time minutes, language, region,as well as its ratings were considered for this project. [2] contains movie records that describe a movie by its title, genre, rating, and the tag that users give it. A tag is a line, and is often an opinion of the user about the movie, or the names of the cast and crew associated with the movie. Thus, the tag is not uniform in what it describes, it just associates the movie with a line that the user defines. [1] and [2] were merged to obtain the final data set.

7.2 Data Merging

[1] and [2] were combined to obtain a final data set consisting of 0.9 million records. This was done using the merge() of the integrated development environment RStudio, which provides join operations for data frames. First, all data sets of [1] were merged on the basis of a unique identifier. This data frame was later merged with [2] on the basis of the movie title to obtain the final records. Merging on the basis of title ensured that movies with records present in both data sets were not repeated in the final data set. For movies that were present in both the data sets, their average rating was computed as a mean of the two ratings available.

7.3 Data Cleaning

The primary step of data cleaning was to ensure that our final data only consisted of movie records. Records present in [1] were filtered on the basis of the type of their title, using filter() present in the dplyr package of R. Missing values in the run time minutes attribute were replaced by 0. Missing values in the average rating column were replaced by the mean of the column.

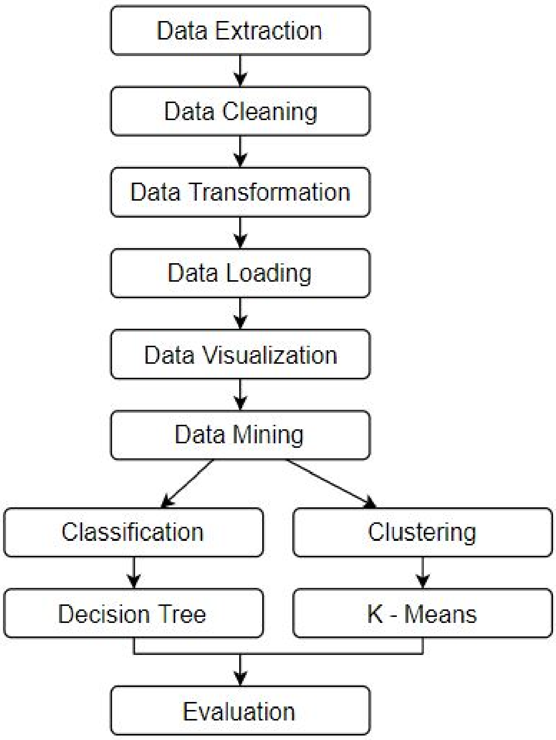


Figure 1: Project Design

7.4 Data Loading

Once the data was cleaned, it was converted to a tab separated file, which could be loaded into MySQL. A relational database was chosen given the consistent structure of the final data set. It also helped ease the process of querying the data to obtain insights.

8. PROJECT DESIGN

The project implementation is divided into two phases, the data management phase and the data mining phase. The different tasks performed in data management are data cleaning, data processing, data cleaning and data loading. After the data is collected, cleaned and loaded the data mining task can be performed. The data mining phase begins with the visualization of the data and exploring it. It is important to understand the distribution of attribute values and types so as to determine the data mining technique to be implemented. The next step is to implement the data mining algorithm. The data mining technique used in the project are Classification and Clustering. After the task is implemented the model is evaluated to obtain the accuracy of the implementation. The design of the project can be seen as follows:

9. CHOICE OF DATA MINING ACTIVITIES

Data Mining is used in order to discover large patterns in huge data sets. There are two types of data mining techniques - supervised and unsupervised. Two data mining activities were chosen for the current project -

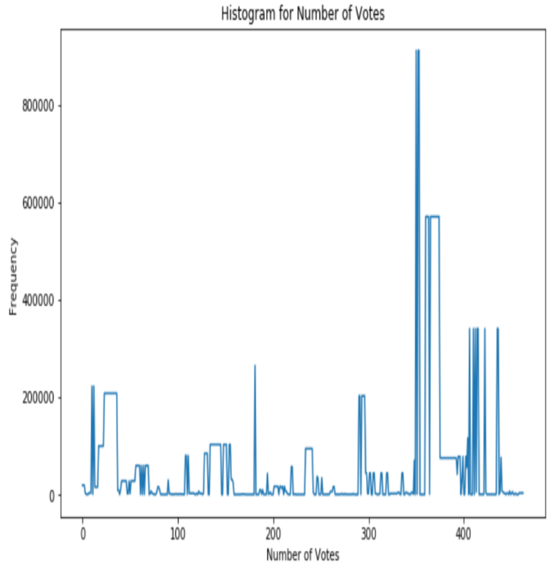


Figure 2: Histogram of number of votes

- Classification - Classification is a type of supervised data mining technique. It is used to identify which class a new data point belongs to. This is done by training the model from a set training data set. The reason it falls under a supervised machine learning technique is that it uses a class variable to map the new data point to a class.

- Clustering
  Clustering is an unsupervised data mining technique. It is mainly a way to group data into logical sets, to identify if any underlying similarities between these patterns can be studied. This is done by randomly choosing cluster centers and then calculating the distance of every point from the cluster centers to determine which cluster the point belongs to.Since points that are similar are grouped together, a point is assigned to that cluster that it is closest from. This technique is unsupervised because it does not have a target variable to adhere to, instead it just groups most similar objects together.

10. DATA VISUALIZATION

The attributes that were most important for the classification and clustering processes were visualized to understand their range, and correlations.

11. PROJECT IMPLEMENTATION - CHOICE OF APPROPRIATE TECHNIQUE

In the given project as mentioned two data mining techniques have been used - Classification and Clustering. These two techniques have previously been explained in brief.

- Classification - Classification is a supervised machine learning technique. Classification helps us identify why certain items are alike. Here a target variable is required for the specified data set. In order to have a

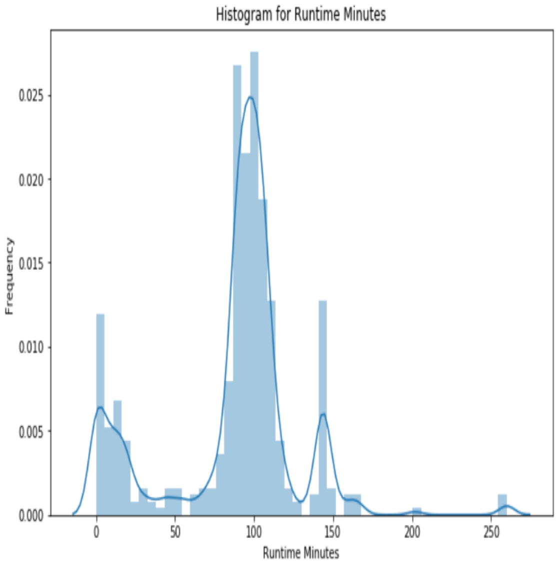


Figure 3: Histogram of runtime minutes

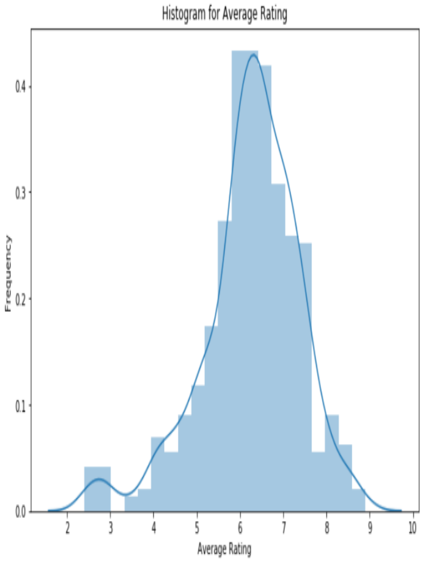


Figure 4: Histogram of average rating

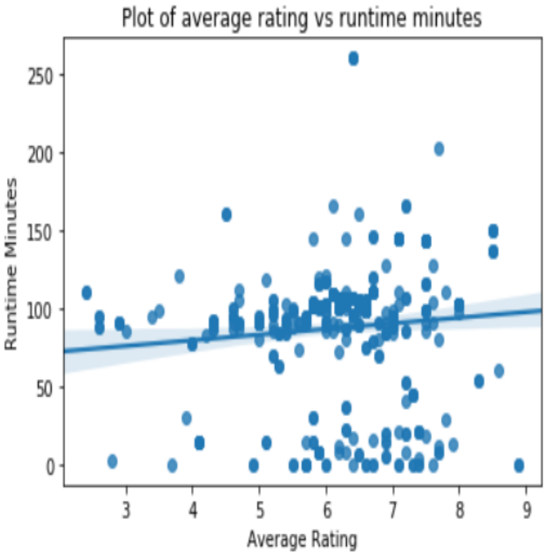


Figure 5: Plot of average rating vs run time minutes

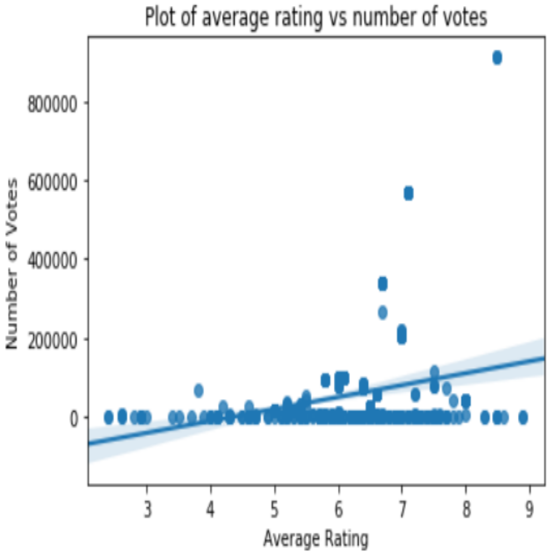


Figure 6: Plot of average rating vs number of votes

target variable a feature called opinion is introduced. The feature variable opinion was created based on an existing variable Avg_Rating. There were two values assigned to it - Hit or Flop. If the Avg_Rating is greater than or equal to five, the movie is considered a hit and a flop otherwise.

There are different types of classifiers that are available such as - Decision Trees, Naive Bayes, KNN etc. This project uses Decision Trees as the classification technique. In order to implement classification using Decision Trees, it was empirical that target variable, Opinion must be a nominal attribute.

The purpose of this classification was to predict which movies are more likely to be a hit and which ones are more likely to be a flop. The classification was divided into two steps - training or the learning step and test or the prediction step. The model finally predicts if the movie is a it or a flop.

Decision trees was used to perform the classification which use an if..then structure. The reason decision tree was chosen is because the structural visual representation makes decision making of predicting the class the new data point belongs to. The first step that went into creating a decision tree involved partitioning the instances. The choice of attributes that would be partitioned first was based on a purity measure. Each of the branches made from the attributes become the rules of the decision tree, leading to the final decision. There are several selection measures available which help in the splitting of attributes such as Information Gain, Gini Index, Gain Ratio. The project implements both Gini Index and Information Gain(Entropy).

The code is implemented in python and uses the libraries - pandas, sklearn, numpy, IPython, pydotplus. 10 fold cross validation is performed on the code. This causes the data set to be split into 10 subsets. An average overall estimate is provided in the end and the results are averaged out. This even prevents any form of overfitting of training set. The accuracy of the model is also very high in both cases (gini and entropy) estimating to 93%.

- Clustering
  The aim of clustering was to better understand the data, and to derive patterns that could relate the success of a movie to its characteristics such as average rating, run time minutes, region and language.
  Keeping this aim in mind, the type of clustering chosen was clustering using partitioning methods, since a top down approach would give more information, than other forms of clustering such as hierarchical and density based clustering.
  The algorithm chosen to perform clustering was k-means. K-means is a clustering algorithm that clusters data into k number of groups. K is a parameter given by the user, and should at least be two. The reason why this algorithm was chosen over other available clustering algorithms was because it requires the least amount of computation, which was a significant factor to be considered for a data set of 0.9 million records.
  In order to identify the optimal number of clusters, an error curve of k values between 5 to 20 was plotted. Kmeans was run on the data set for varying values of

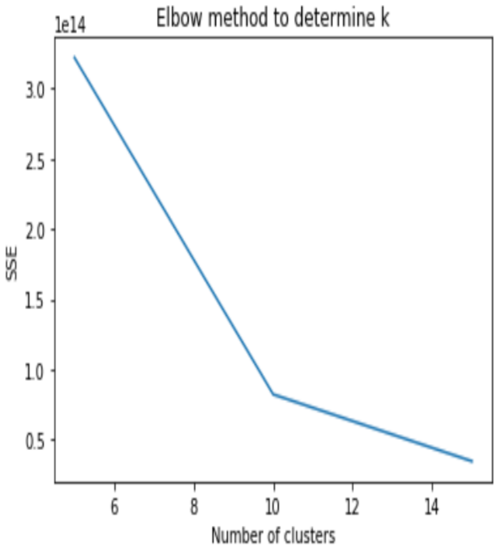


Figure 7: Error curve to determine optimal number of clusters

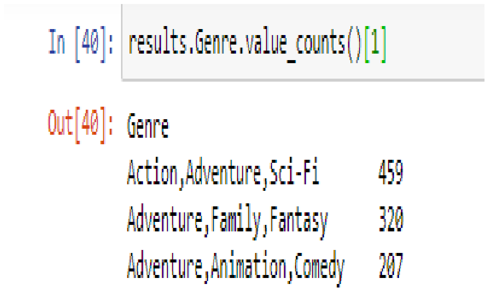
k and the sum of squared errors, which is a measure of variation in a cluster. It was observed that the rate of change in the error, or the knee was at 10. Hence, the number of clusters was determined to be 10. Figure 7 is a plot of number of clusters against their sum of squared error values. The y-axis is scaled, and the top left corner represents the scaling factor of le14.
After the value of k was decided, the kmeans algorithm was implemented on the data set. The python package sklearn was used to implement kmeans. The distance metric chosen was euclidean. The resultant clusters were evaluated using the Davies-Bouldin index and Calinski-Harabasz index.

12. RESULTS OF DATA MINING

The two data mining techniques resulted in the below decision tree and clusters.

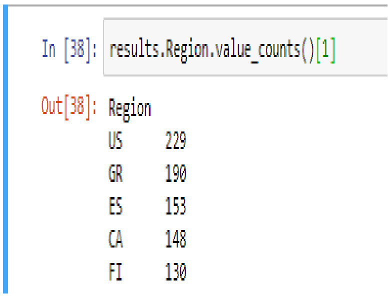
- Results of Classification using Decision Tree. The first step to the result involved finding out the accuracy of the decision trees. Both gini and entropy method were chosen to identify the best split. Both have given an accuracy of approximately 93 %. The final decision tree is visualized.The resultant image is extremely huge and is thus not attached in the report. It can be found as output.png i the final submissions folder.
  Based on the criteria provided - gini/entropy, the best attribute is chosen is RunTimeMins, the samples that are assigned to it, there are two classes 0 or 1, the leaf node gives the final identification of the class. At each level a new attribute is chosen based on the purity.

- Results of Clustering
  10 clusters were obtained by running the kmeans algorithm. The original data was grouped according to clusters to derive information. This was done using the groupby function available in pandas DataFrame package. A quick look at the properties of Cluster 1 shown in the figures 8, 9 and 10 conveys that the cluster mostly consists of movies whose genre is action, adventure and comedy, that are mostly produced in the United States of America, and whose average rating is in the range of 6.5 - 7.5.



```
In [40]: results.Genre.value_counts()[1]

Out[40]: Genre
         Action,Adventure,Sci-Fi       459
         Adventure,Family,Fantasy      320
         Adventure,Animation,Comedy    207
```

Figure 8: Three most frequent genres of cluster 1



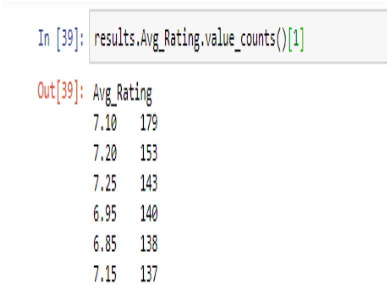
```
In [38]: results.Region.value_counts()[1]

Out[38]: Region
         US    229
         GR    190
         ES    153
         CA    148
         FI    130
```

Figure 9: Three most frequent regions of cluster 1

The cluster quality was evaluated on the basis of Davies Bouldin Index and the Calinski-Harabasz index. The Davies Bouldin Index(DBI) is an internal clustering measure that evaluates cluster scatter, and should be minimized. The DBI obtained for the clustering was 0.494. The Calinksi-Harabasz score should be maximized to indicate good clustering. The score obtained was 6369411.54.

13. ALIGNING TASKS TO CRISP METHODOLOGY

We are ensuring that we follow the CRISP Methodology.



```
In [39]: results.Avg_Rating.value_counts()[1]

Out[39]: Avg_Rating
         7.10    179
         7.20    153
         7.25    143
         6.95    140
         6.85    138
         7.15    137
```

Figure 10: Three most frequent ratings of cluster 1

The steps of the CRISP Methodology and what we have finished up until now are listed below.

- Business Understanding - Completed in Phase 1
- Data Understanding - Completed in Phase 1
- Data Preparation - Completed in Phase 2
- Data Mining - Completed in Phase 3
- Evaluation - Completed in Phase 3
- Deployment - Completed in Phase 3

14. LESSONS LEARNED

Learning is a very crucial part of every project. This project involved a hand in hand learning among all team members. There were multiple things that were learned.

- The project taught us how to formulate the statement and questions.
- The project taught us how to gather data based on requirement from multiple statements.
- The project taught us how to identifying the important attributes based on the client requirements.
- The project taught us how to manage multiple units like data mining and data management.
- The project taught us how to understand the importance of cleaned data and the effect missing values can have on the resultant set.
- The project taught us how to identifying why a particular data mining technique is more suitable than other.
- The project taught us how to implementing the identified mining techniques.
- The project taught us how to train our algorithm.
- The project taught us how to evaluate the algorithm.
- The project taught us how to manage the different phases on a data analysis project using CRISP Methodology.
- The project taught us how to handle big data rather than just implementing smaller sample data sets.
- The project taught us how to document, give demos, report and present the final work.

15. CURRENT STATUS

The final project performs classification and clustering. The classification technique used is Decision Trees which is used to classify and predict if any new movie will be a hit or a flop. The clustering technique used is K Means Clustering which clusters the merged and final data set to identify emergent clusters. This project is now in a completed state.

16. FUTURE WORK - SCOPE OF PROJECT

In its current status, the project presents two data mining tasks that could be used to study patterns that can establish a relation between the success of a movie and its descriptive properties. These tasks can be further developed to develop a predictive analysis system that can determine the performance of a movie before its release. Such a system can be used to analyse whether the movie will be profitable, how well it will perform, and if it turns out to perform poorly, whether it will be able to recover its production costs.

17. ITEMS DELIVERED WITH THIS PHASE

List of items delivered along with this document involve:

- Python code for classification using Decision Trees
- Python code for clustering using K Means Clustering
- Python code for data set visualization
- TXT file of sample subset of data
- TXT file of final data set
- Report for Phase 3
- Final Presentation (. PPT)
- Final Presentation (. PDF)
- README.txt
- Previous phase submission : Metadata_Phase1.zip
- Previous phase submission : Metadata_Phase2.zip

18. CONCLUSION

Movies industries are known for their tremendous revenue generation.Even after a movie is released in theatres, it continues to generate revenues by selling its digital rights to streaming platforms. It is necessary to tap into the data that drives these industries, and to study emerging patterns that link the performance of a movie to its characteristics like the number of votes it receives, its average run time and rating, and so on. Through this project, we have attempted to design two data mining that allow us to correlate the performance of a movie to its aforementioned characteristics.

19. REFERENCES

[1] Imdb data set link.
`https://www.imdb.com/interfaces/`.

[2] Movielens data set link.
`https://grouplens.org/datasets/movielens/`.