**Predicting the Success Rate of Indian Movies**

**OVERVIEW:**

Bollywood is one of the biggest movie industries in the Asian continent. It produces around 150-200 movies every year. It has generated 3 billion USD revenue in the year 2011. It is also estimated that, it might enhance its revenue to 4.5 billion USD by the year 2016. But the question is, whether it is going to be easy task for it to reach the target.

Yashraj movies are one of the biggest production houses in India which produces 30 films every year. But not all movies they had produced were successful. We would like to use the data driven approach to estimate their success/failure rates. The approach can be used to help Yashraj to produce more successful movies and increase their revenue.

**Below is a one of those examples through which we are motivated:**

• Amazon.com have used data mining approach to predict the type of show that would be liked by viewers

• They had 14 pilot episodes posted on their website

• They monitored the number of viewers watched, comments and patterns on their site

• They also captured the rating for those pilot episodes, how much it was viewed and how many shared it with others.

• Based on the prediction they released their first TV show “Alpha house” which received 7.5 rating on IMDB.

**Data mining problem:**

The project begins with the unsupervised data model and proceeds with the generation of different data model techniques.

Proposed Data Mining Method/Model – **OneR Classification and J48 Tree**

The some of the data mining problem that we might encounter are as follows:

Selection of attributes those are relevant to get the target variable

Selection of the number of years to predict target variable is also difficult

**Target variable:**

• Using the data sets available we are trying to predict what kind of movie would be successful

• This will be an unsupervised approach as we are using classification model to predict the target variable

**Features useful:**

The useful features can be such as:

1. Year and Month of the movie release.

2. Actor

3. Director

4. Genre

**Business value:**

The prediction would help the production house to gain more profits and avoid the possible errors.

**DATA:**

**Data Instances/Attributes:**

We collected the data from the IMDB database which is an open source and will be sufficient to predict the success rate of the future productions from Yashraj productions using WEKA tool. All the data in the source is in text format and should be converted to executable format. The data attributes include the following:

• Number of Votes

• Genre

• Running time

• Languages

• Actors

• Directors

**Preliminary Results:**

No results have been predicted yet. However we have few data mining models in mind that  would help us get the target variable. We are still in the process of cleansing the data for the  project.

**Issues:**

**Finding the data**

We  chose  Bollywood  movies  for  our  project,  however  we  couldn’t  find  any  data  set  on  Bollywood movies for free. We had to download the data set from movielens which host data  for Hollywood movies. The data set have lot of information from which we had to choose the  relevant attributes for our project.

**Converting the data**

The data sets we found on movielens.com were not weka friendly. They were not in .arrf or .csv format. They are in a format that we cannot open even with notepad or WordPad. The data  sets we have on movie lens are web compatible, so we are trying to copy each instance of data into excel sheet. However copying 500 instances of around 10 attributes is not an easy task. We are still trying to find software that could convert these files to Weka friendly format.

**Preparation of data**

The next step is going to be, preparing the data to load into the weka. Not all data sets included in the file are going to be useful and hence we might remove some of the unwanted data to  avoid possible confusions.

We are going to make use of the classification method and see which all the movies that fall into different categories and what are the attributes they are associated with. This might help us in identifying the most important attribute in determining the class­ The Target Variable.

By this checkpoint, we have an update on our project analysis by data modeling.

**Data Modelling**

Choosing appropriate model for data mining is not an easy task. Many times the data we chose might not give the results we intend to. Data models in a large picture can be divided into two categories

1. Supervised model
2. Unsupervised model

Supervised data model: We use this model when we know our target variable, and when we know the categories of the data. Training data is provided in supervised model.

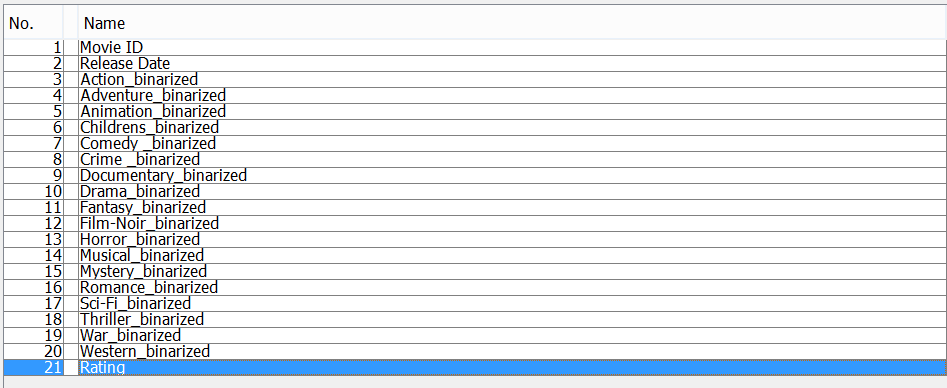
Unsupervised model: This model is used when the target variable is not known or the categories of the data are unknown. Training data is not provided in this model. This is often used to categorize or cluster the target variables.

For our model we know the target variable. We are trying to get the overall rating of a movie based on existing parameters like Actors rating, Director rating, Movie Genre, Release date.

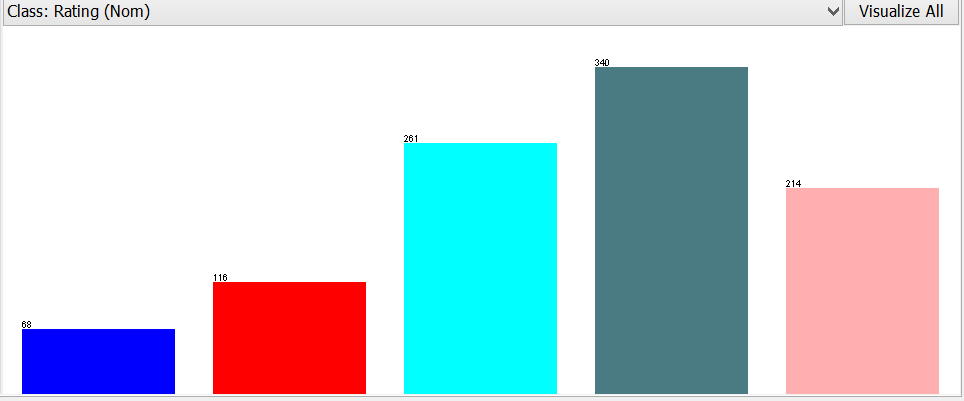
We have used J48 Classification and Naïve bayes models in our project. However initially when we built the data model with the attributes we had the accuracy of the result was very poor.

**Model #1**

Below is the list of data attributes that were used to build our model which resulted in more inaccurate results than accurate results.



**Attributes used for model 1**



**Overall Rating of movies when converted to nominal from numeric**

All the attributes when loaded into Weka were in numeric format. However to support/build our model we had to convert the attributes from numeric to nominal and binary. Binary converts the data into two instances 0 and 1 i.e. low and high. Whereas nominal values are stored as numbers too however these numbers act as indexes into an array of possible attribute values which is very efficient.

The above graph represents the nominal values for overall rating. Rating is between 1 and 5 so we can see five differently colored columns and the values on top of those bars are the number of movies that fall into that rank.

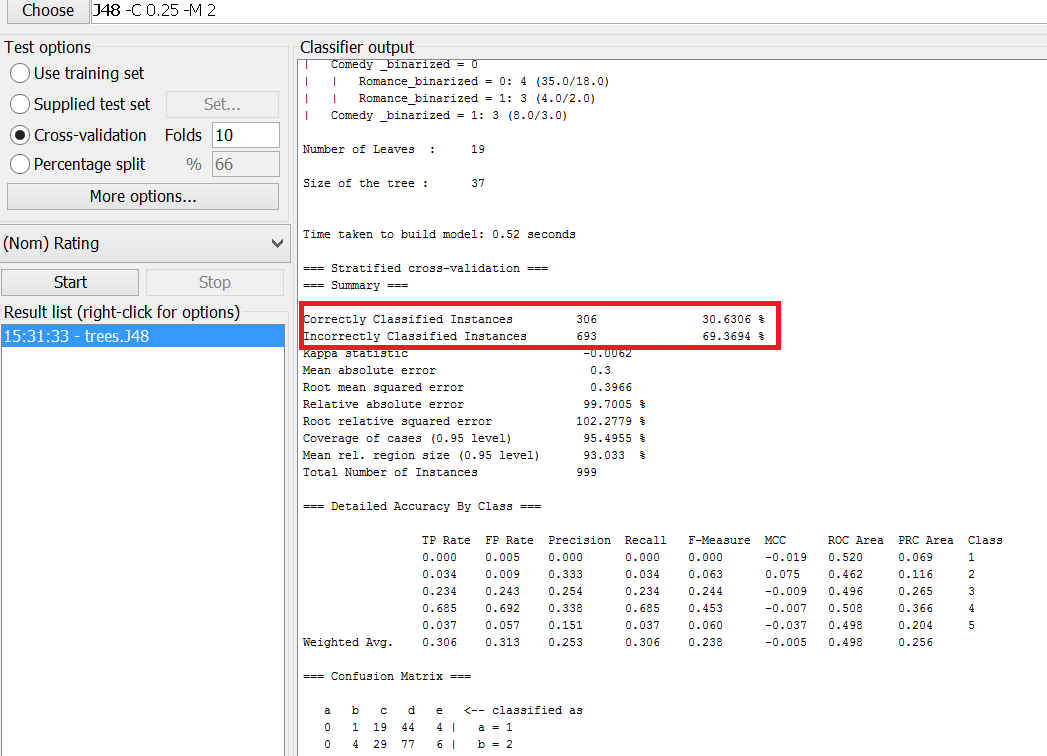
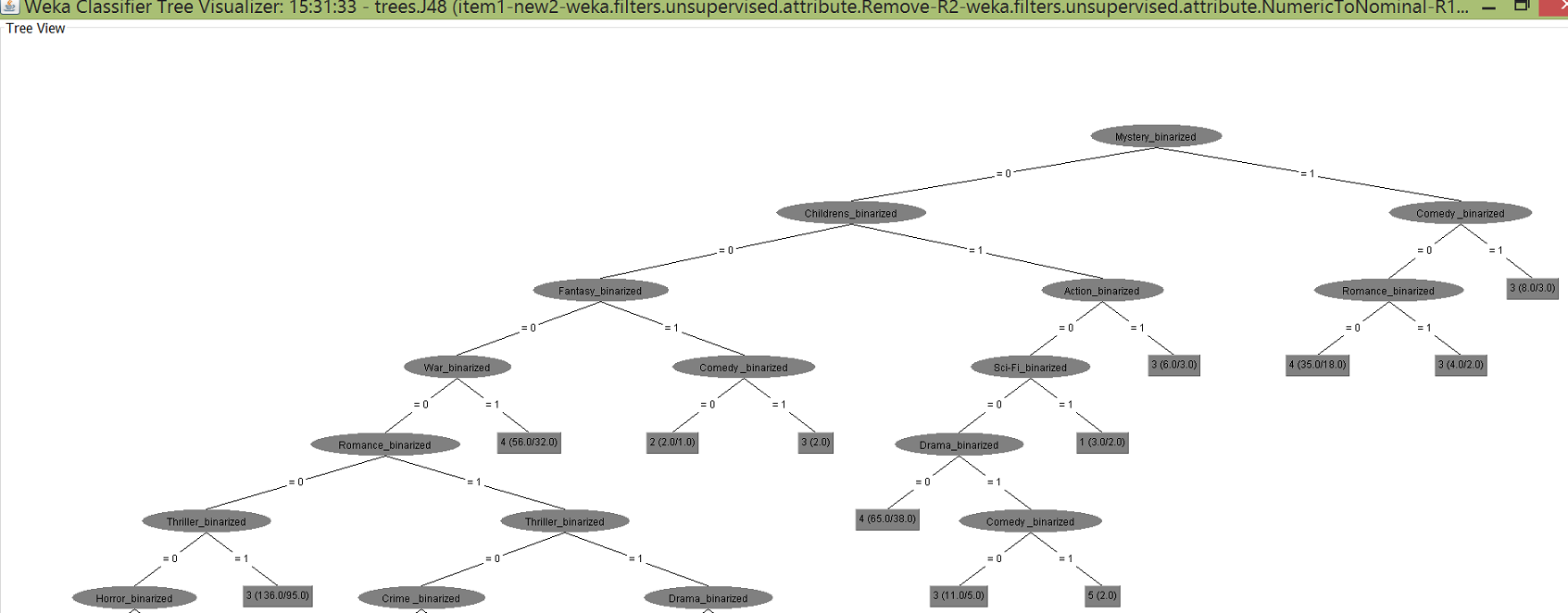


Fig 1.1- **J48 classification result**

**Fig1.2 J48 Result**



As highlighted in fig 1.1 we can see the incorrect instances are higher compared to correct instances. Also if we look at the tree we can see the branches are based on the genre of the movie. The result was not much different from j48 model when we fed the same data to Naïve Bayes model.

We had to go back to the data preparation task and add two attributes manually. This made us reduce the number of data instances from 1000 to 200.

**PROJECT MANAGEMENT**

|  |  |  |
| --- | --- | --- |
| **Team Member** | **Roles and Skills** | **Contributions** |
| Mani Karthik | Project Execution and Management | Will be coordinating with all the team members for overall status of the project. Develop a project plan and manage the deliverables according to the plan. |
| Rajendra | Interpretation and data collection | Collect the data from source and finding the attributes that would perfectly workout for classification |
| Meghana | Analysis | Analyzing the algorithm and providing the possible outcome |
| Srujan | Data conversion | Conversion of data collected into a readable format which will help us executing in WEKA |

**DELIVERABLES AND CHECKPOINTS**

|  |  |  |  |
| --- | --- | --- | --- |
| **Checkpoint Date** | **Expected Deliverable** | **Responsible Team Members** | **Checkpoint Results** |
| 02/15/2016 | Project Submission and overview | Mani Karthik | Project has been submitted and is in progress |
| 03/08/2016 | Data collection, roles and responsibilities | Rajendra and Srujan | Data set has been collected and converted into CSV format |
| 04/10/2016 | Preliminary Analysis | Meghana | Preliminary analysis using the first model which resulted in changes of data attributes as the result was not accurate |
|  |  |  |  |