Application of Decision Tree to Predict Gross Income of a Movie

Aditya Mandhare
Computer Engineering Department, Mumbai University
Mumbai, India

ABSTRACT
Decision Trees are powerful and popular tools for prediction. These trees are especially useful where in addition to the accuracy of the prediction, the reason for the prediction is also important. This paper presents an application where these decision trees would be used to predict the gross income of a bollywood movie. The paper identifies several factors influencing a movie's income and applies decision tree learning to predict the gross income.

General Terms
Classification technique, machine learning, prediction algorithm, Bollywood, movies, gross income.

Keywords
Decision Tree, Entropy, Information Gain, Splitting Factor, attribute-value pairs, Data Mining.

1. INTRODUCTION
Starting in 1913, Bollywood which is popularly known as Hindi Cinema Industry is a large part of Indian film industry. It is based in Mumbai, Maharashtra, India. In 2013 this industry contributed INR 500000 crore (1 crore = 10 million) to the Indian economy and which is about 0.5% of the GDP in 2013 [8]. Today the success of a movie is decided by the money it makes. Several factors make a film successful. If a movie gets certain combination of these factors right, then it can expect its gross income to be high, and may suffer losses if it gets them wrong. In this paper a survey of various factors influencing a movie's gross income has been made. The decision tree learning method is worked upon this available set of records. With the help of these records, a decision tree has been constructed. This tree will be useful in predicting the gross income of any Bollywood film. The paper uses 'Indian National Rupee (INR)' in crore (1 crore = 10 million) for expressing income and investment of movies.

The rest of this paper is organized as follows:

- Section 2 covers Decision Tree Learning
- Section 3 covers Determining attribute value pairs for movies
- Section 4 covers Determining the Test Attributes
- Section 5 covers Construction of Decision Tree and classifying Test Set
- Section 6 covers Decision Tree in Rule Form
- Section 7 covers Conclusion
- Section 8 covers Acknowledgement
- Section 9 covers References

2. DECISION TREE LEARNING
This is a method which is commonly used in Data Mining. It is supervised classification learning. This technique is best suited to problems where instances are represented by attribute-value pairs. Thus every instance has a fixed set of attributes, and each attribute can take a value from a set of values. In this method a tree model is created. Each of the interior node corresponds to one of the input attribute. For each possible value of the attribute the tree branches out to its children node. Finally the leaf node represents a value of the target attribute. Thus given a set of values of the input attribute, we can find a path from root node to leaf node and can reach one of the values of the target attribute.

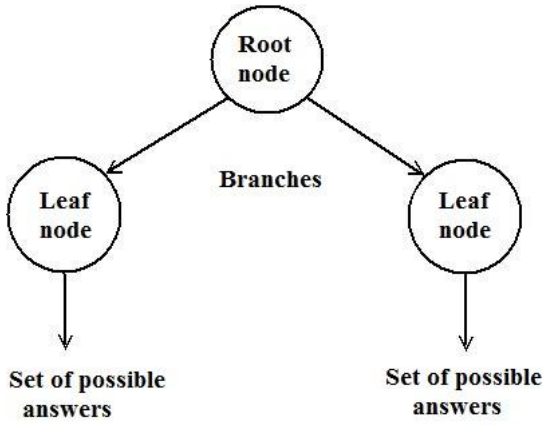

Fig 1: A sample decision tree

The tree is constructed in a top down recursive divide and conquer manner. In the beginning, all the training examples are at the root. Attributes are categorical and examples are partitioned recursively based on selected attributes. The test attributes are selected on the basis of statistical measure. The partitioning process is stopped when, either all the samples belong to the same class, that is, the same value of the target attribute, or there are no remaining attributes for further partitioning, or there are no samples left.

2.1 Information Gain
Consider a set of samples "S". Suppose there two classes in which samples are to be classified, "P" and "N". Let there be "p" elements of class "P", and "n" elements of class "N". The amount of information, needed to decide if an arbitrary example in S belongs to P or N is defined as, [1][2][3]

$$\mathrm{I(p,n)} = -\frac{\mathrm{p}}{p+n} log_2 \frac{\mathrm{p}}{p+n} - \frac{\mathrm{n}}{p+n} log_2 \frac{\mathrm{n}}{p+n}$$

Let us assume that using an attribute A, a set S will be partitioned into sets {S1, S2, S3....Sv}. Si contains pi examples of P and ni examples of N, the entropy or the expected information needed to classify objects in subtrees Si

is,

$$E(\mathrm{A}) = \sum_{i=1}^{v} \frac{p_i+n_i}{p+n} I(p_i, n_i) \qquad [1][2]$$

Gain(A) is the expected reduction in entropy caused by knowing the value of attribute A.

$$Gain(\mathrm{A}) = I(p_i, n_i) - E(A) \qquad [1][2]$$

After computing information gain for each attribute, the one with highest information gain is chosen as the test attribute for set S. The above process is continued till the stopping conditions mentioned previously are reached. The input sample to be classified is of the form, (x, Y) = (x1, x2, x3, ..., xk, Y). The dependent variable, Y, is the target attribute. In the problem we try to obtain the value of this attribute. The vector x is composed of the input variables, x1, x2, x3 etc., that are used for that task. [1][2][3]

2.2 Advantages of Decision Tree [3]

2.2.1 Ease of Understanding

Decision trees are able to generate easy to understand rules. Thus humans can easily perceive these rules. These rules can also be easily used with database languages like SQL.

2.2.2 Required Computation

The decision tree learning method is not computation intensive. Thus by using decision tree method complex problems can be solved with minimum computations.

2.2.4 Error Handling

The decision trees are robust to errors. These trees can work efficiently even if some data is missing. Thus given a sample set of data, it is possible to construct a tree even if the values of some attributes of a record are missing.

3. DETERMINING ATTRIBUTE VALUE PAIRS FOR MOVIES

In this paper we will predict the gross income of a film by applying this decision tree learning. For this purpose we would be needing attribute-value pairs. There are several factors or attributes which the paper has identified which can influence the success of a bollywood movie. Some of these attribute-value pairs are as follows,

3.1 Critic Rating

This is a rating given to a film by the film critics. This rating can play a crucial role in helping the film gain popularity. It is the critic's view about the film, the acting and the story. The critics rate the film. It is usually on a 0-5 scale. The different values are, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5.

3.2 Film CBFC Category

This paper checks if the film is having an 'A'(adult) certificate or not. This is provided by Central Board of Film Certification (CBFC). It tells about the age group of the audience which is fit to watch the movie. Yes = Y (‘A’ certificate) No = N (Does Not have ‘A’ certificate)

3.3 Presence of a Superstar

This is an important deciding factor as far as the bollywood movies are concerned. There is a lot of fan following for each superstar which can act as crowd puller. This paper considers two values. Yes = Y No = N

3.4 Number of Songs

Songs have an important role to play in Indian cinemas and many films are remembered solely for their songs. Songs released before the movie and can be a good crowd puller.

3.5 Investment in movies

This tells how much money has been put in to make the movie. This paper has collected data about the expenditure on each film. The budget of a film has been expressed in crores (1 crore = 10 million).

3.6 Opening day collection

This is an excellent measure of how much a film would earn. It measures the response of the audience on the day of release. This amount of opening day collection is expressed in crores.

The above mentioned factors would be used to predict the value of the leaf node, that is, the film's gross income. It can have values in one of the following ranges; 0-75, >75 in crore (1 crore = 10 million).

4. DETERMINING TEST ATTRIBUTES

The paper has collected data of 25 recent bollywood films [4][5][6][7]. The values of each attribute have been collected. Thus by applying a top-down recursive divide-and-conquer technique, a tree can be constructed. Traversing this tree for any input record can help predict the input film's gross income. In Fig.2, for each movie, attributes like, critic’s rating, CBFC adult certificate, presence of a superstar, number of songs, investment, first day collection and their values have been provided. Thus we have attribute value pairs of each record. Out of these attributes, ‘adult’, ‘star presence’ are binary ie. They have values either Yes = Y or No = N. But other attributes like ‘rating’, ‘songs’, ‘invest’, ‘first day collection’ and the class gross income are continuous. We first convert them into categorical form. Here we decide upon the range in which the values of these attributes lie. Attribute ‘rating’ can be categorized as ‘< 3’ and ‘>= 3’. Attribute ‘songs’ can be categorized as ‘<= 6’ and ‘> 6’. Attribute ‘invest’ can be categorized as ‘0-50’ and ‘>50’. Attribute ‘first day collection’ can be categorized as ‘<= 10’ and ‘> 10’. The class ‘gross income’ can be categorized as income ‘0-75’ or income ’> 75’ (crore). In Fig.3, the table shows attributes which are either binary of categorical. Now the decision tree learning can be applied on this table and a tree can be built.

id	name	rating	adult	star_presence	songs	invest	first_day_collection	gross_income
1	Jai Ho	2.5	N	Y	9	75	18.8	183
2	Singham Returns	3	N	Y	5	115	32.09	156.77
3	Entertainment	3	N	Y	8	80	11	72.02
4	Main Tera Hero	3	N	N	6	42.5	6.6	80
5	Gunday	3.5	N	Y	8	72	15	150
6	Hasee To Phasee	3.5	N	N	6	25	4	62
7	Heropanti	3	N	N	6	25	6.63	65
8	Humshakals	2.5	N	Y	6	75	5.1	100
9	Queen	4	N	N	8	12.5	2	98.2
10	Gulaab Gang	3	N	Y	7	12	2.41	14
11	Shaadi Ke Side Effects	3.5	N	Y	7	50	5.7	70
12	Bewakoofiyaan	2	N	Y	7	22	2	25
13	Darr @ the Mall	2	N	N	3	8	1	6
14	Mardaani	3.5	Y	Y	1	15	3.46	35
15	Jannat 2	3	Y	Y	7	28	8.5	84
16	Gori Tere Pyaar Mein	2.5	N	Y	8	30	3.25	12.65
17	Zila Gaziyabad	1.5	Y	Y	5	36	5	16
18	Himmatwala	2.5	N	Y	5	50	12.14	70
19	Student Of the Year	3.5	N	N	7	45	7.48	71
20	Race 2	3	N	Y	5	60	19.45	162
21	Chennai Express	3.5	N	Y	8	75	29.25	395
22	Pizza	2.5	N	N	5	1.5	0.75	4.35
23	Dhoom 3	4	N	Y	6	125	30.9	530
24	Bhaag Milkha Bhaag	4	N	Y	6	30	8.5	164
25	Shuddha Desi Romance	3.5	N	N	9	25	6.75	55

Fig 2: Initial table with attribute-value pairs for 25 movies (Training Set)

id	name	rating	adult	star_presence	songs	invest	first_day_collection	gross_income
1	Jai Ho	< 3	N	Y	> 6	> 50	> 10	> 75
2	Singham Returns	>= 3	N	Y	<= 6	> 50	> 10	> 75
3	Entertainment	>= 3	N	Y	> 6	> 50	> 10	0 - 75
4	Main Tera Hero	>= 3	N	N	<= 6	0-50	<= 10	> 75
5	Gunday	>= 3	N	Y	> 6	> 50	> 10	> 75
6	Hasee To Phasee	>= 3	N	N	<= 6	0-50	<= 10	0 - 75
7	Heropanti	>= 3	N	N	<= 6	0-50	<= 10	0 - 75
8	Humshakals	< 3	N	Y	<= 6	> 50	<= 10	> 75
9	Queen	>= 3	N	N	> 6	0-50	<= 10	> 75
10	Gulaab Gang	>= 3	N	Y	> 6	0-50	<= 10	0 - 75
11	Shaadi Ke Side Effects	>= 3	N	Y	> 6	0-50	<= 10	0 - 75
12	Bewakoofiyaan	< 3	N	Y	> 6	0-50	<= 10	0 - 75
13	Darr @ the Mall	< 3	N	N	<= 6	0-50	<= 10	0 - 75
14	Mardaani	>= 3	Y	Y	<= 6	0-50	<= 10	0 - 75
15	Jannat 2	>= 3	Y	Y	> 6	0-50	<= 10	> 75
16	Gori Tere Pyaar Mein	< 3	N	Y	> 6	0-50	<= 10	0 - 75
17	Zila Gaziyabad	< 3	Y	Y	<= 6	0-50	<= 10	0 - 75
18	Himmatwala	< 3	N	Y	<= 6	0-50	> 10	0 - 75
19	Student Of the Year	>= 3	N	N	> 6	0-50	<= 10	0 - 75
20	Race 2	>= 3	N	Y	<= 6	> 50	> 10	> 75
21	Chennai Express	>= 3	N	Y	> 6	> 50	> 10	> 75
22	Pizza	< 3	N	N	<= 6	0-50	<= 10	0 - 75
23	Dhoom 3	>= 3	N	Y	<= 6	> 50	> 10	> 75
24	Bhaag Milkha Bhaag	>= 3	N	Y	<= 6	0-50	<= 10	> 75
25	Shuddha Desi Romance	>= 3	N	N	> 6	0-50	<= 10	0 - 75

Fig 3: Table with attribute-value pairs properly categorized for 25 movies (Training Set)

4.1 Level 1

Calculating the Gain for all attributes, it has been found that the Gain of 'invest' is highest. Therefore, invest becomes the root node of the tree. The tables in fig.4 and fig.5 are its children nodes. From the table in fig.5 it can be inferred that all films with budget or investment more than 50 crore would earn more than 75 crore. But we still need to decide for the movies with investment in the range 0-50 in fig.4.

4.2 Level 2

Comparing the Gain values of attributes in the table shown in fig.4, 'rating' has maximum gain. Therefore it is the next test attribute. The tables in fig.6 and fig.7 are its two children nodes. The fig.6 clearly shows that movies with budget in the range 0-50 crore and critics rating less than 3 have gross income in the range 0-75. But we still need to decide for those films whose rating is more than or equal to 3.

4.3 Level 3

Now, comparing the Gain values of attributes of the table in fig.7 we get that attribute 'adult' has the highest gain. Thus we get two tables. Table in fig.8 shows movies having CBFC 'A' certificate and table in fig.9 shows movies which do not have 'A' certificate.

4.4 Level 4

Now we calculate Gains of attribute of tables in fig.8 and fig.9. We get that gain of attribute 'songs' is highest in both the tables. Thus 'songs' is the new test attribute in both the cases. Table in fig.10 shows that movies with investment in the range '0-50', rating '>=3', adult= 'Y' and songs'<=6' have gross income in the range '0-75'. Table in fig.11 shows that movies with investment in the range '0-50', rating '>=3', adult= 'Y' and songs'>6' have gross income in the range '>75'. Tables in fig.12 and fig.13 show that movies with investment in the range '0-50', rating '>=3', adult= 'N' and songs'<=6' and songs '>6' respectively. The table in fig.12 shows that, it is required to split it further, in order to reach to a conclusion. However, table in fig.13 shows that all films with investment in the range '0-50', rating '>=3', adult= 'N' and songs'>6' have gross income in the range '0-75'.

4.5 Level 5

Again calculating the Gain of all the attributes of table in fig.12, it can be inferred that, the new test attribute is 'star_presence'. The table in fig.14 shows that the movies with presence of a star actor earn more than 75 crore. The table in fig.15 shows that the movies without presence of a star actor earn in the range '0- 75' crore.

id	name	rating	adult	star_presence	songs	invest	first_day_collection	gross_income
4	Main Tera Hero	>= 3	N	N	<= 6	0-50	<= 10	> 75
6	Hasee To Phasee	>= 3	N	N	<= 6	0-50	<= 10	0 - 75
7	Heropanti	>= 3	N	N	<= 6	0-50	<= 10	0 - 75
9	Queen	>= 3	N	N	> 6	0-50	<= 10	> 75
10	Gulaab Gang	>= 3	N	Y	> 6	0-50	<= 10	0 - 75
11	Shaadi Ke Side Effects	>= 3	N	Y	> 6	0-50	<= 10	0 - 75
12	Bewakoofiyaan	< 3	N	Y	> 6	0-50	<= 10	0 - 75
13	Darr @ the Mall	< 3	N	N	<= 6	0-50	<= 10	0 - 75
14	Mardaani	>= 3	Y	Y	<= 6	0-50	<= 10	0 - 75
15	Jannat 2	>= 3	Y	Y	> 6	0-50	<= 10	> 75
16	Gori Tere Pyaar Mein	< 3	N	Y	> 6	0-50	<= 10	0 - 75
17	Zila Gaziyabad	< 3	Y	Y	<= 6	0-50	<= 10	0 - 75
18	Himmatwala	< 3	N	Y	<= 6	0-50	> 10	0 - 75
19	Student Of the Year	>= 3	N	N	> 6	0-50	<= 10	0 - 75
22	Pizza	< 3	N	N	<= 6	0-50	<= 10	0 - 75
24	Bhaag Milkha Bhaag	>= 3	N	Y	<= 6	0-50	<= 10	> 75
25	Shuddha Desi Romance	>= 3	N	N	> 6	0-50	<= 10	0 - 75

Fig 4: Table of movies with invest in range 0-50 (Level 1)

id	name	rating	adult	star_presence	songs	invest	first_day_collection	gross_income
1	Jai Ho	< 3	N	Y	> 6	> 50	> 10	> 75
2	Singham Returns	>= 3	N	Y	<= 6	> 50	> 10	> 75
5	Gunday	>= 3	N	Y	> 6	> 50	> 10	> 75
8	Humshakals	< 3	N	Y	<= 6	> 50	<= 10	> 75
20	Race 2	>= 3	N	Y	<= 6	> 50	> 10	> 75
21	Chennai Express	>= 3	N	Y	> 6	> 50	> 10	> 75
23	Dhoom 3	>= 3	N	Y	<= 6	> 50	> 10	> 75

Fig 5: Table of movies with invest greater than 50 (Level 1)

id	name	rating	adult	star_presence	songs	invest	first_day_collection	gross_income
12	Bewakoofiyaan	< 3	N	Y	> 6	0-50	<= 10	0 - 75
13	Darr @ the Mall	< 3	N	N	<= 6	0-50	<= 10	0 - 75
16	Gori Tere Pyaar Mein	< 3	N	Y	> 6	0-50	<= 10	0 - 75
17	Zila Gaziyabad	< 3	Y	Y	<= 6	0-50	<= 10	0 - 75
18	Himmatwala	< 3	N	Y	<= 6	0-50	> 10	0 - 75
22	Pizza	< 3	N	N	<= 6	0-50	<= 10	0 - 75

Fig 6: Table of movies with rating less than 3 (Level 2)

id	name	rating	adult	star_presence	songs	invest	first_day_collection	gross_income
4	Main Tera Hero	>= 3	N	N	<= 6	0-50	<= 10	> 75
6	Hasee To Phasee	>= 3	N	N	<= 6	0-50	<= 10	0 - 75
7	Heropanti	>= 3	N	N	<= 6	0-50	<= 10	0 - 75
9	Queen	>= 3	N	N	> 6	0-50	<= 10	> 75
10	Gulaab Gang	>= 3	N	Y	> 6	0-50	<= 10	0 - 75
11	Shaadi Ke Side Effects	>= 3	N	Y	> 6	0-50	<= 10	0 - 75
14	Mardaani	>= 3	Y	Y	<= 6	0-50	<= 10	0 - 75
15	Jannat 2	>= 3	Y	Y	> 6	0-50	<= 10	> 75
19	Student Of the Year	>= 3	N	N	> 6	0-50	<= 10	0 - 75
24	Bhaag Milkha Bhaag	>= 3	N	Y	<= 6	0-50	<= 10	> 75
25	Shuddha Desi Romance	>= 3	N	N	> 6	0-50	<= 10	0 - 75

Fig 7: Table of movies with rating greater than or equal to 3 (Level 2)

id	name	rating	adult	star_presence	songs	invest	first_day_collection	gross_income
14	Mardaani	>= 3	Y	Y	<= 6	0-50	<= 10	0 - 75
15	Jannat 2	>= 3	Y	Y	> 6	0-50	<= 10	> 75

Fig 8: Table of movies with rating >= 3 and adult = Y (Level 3)

id	name	rating	adult	star_presence	songs	invest	first_day_collection	gross_income
4	Main Tera Hero	>= 3	N	N	<= 6	0-50	<= 10	> 75
6	Hasee To Phasee	>= 3	N	N	<= 6	0-50	<= 10	0 - 75
7	Heropanti	>= 3	N	N	<= 6	0-50	<= 10	0 - 75
9	Queen	>= 3	N	N	> 6	0-50	<= 10	> 75
10	Gulaab Gang	>= 3	N	Y	> 6	0-50	<= 10	0 - 75
11	Shaadi Ke Side Effects	>= 3	N	Y	> 6	0-50	<= 10	0 - 75
19	Student Of the Year	>= 3	N	N	> 6	0-50	<= 10	0 - 75
24	Bhaag Milkha Bhaag	>= 3	N	Y	<= 6	0-50	<= 10	> 75
25	Shuddha Desi Romance	>= 3	N	N	> 6	0-50	<= 10	0 - 75

Fig 9: Table of movies with rating >= 3 and adult = N (Level 3)

id	name	rating	adult	star_presence	songs	invest	first_day_collection	gross_income
14	Mardaani	>= 3	Y	Y	<= 6	0-50	<= 10	0 - 75

Fig 10: Table of movies with rating >= 3 and adult = Y and songs less than equal to 6 (Level 4)

id	name	rating	adult	star_presence	songs	invest	first_day_collection	gross_income
15	Jannat 2	>= 3	Y	Y	> 6	0-50	<= 10	> 75

Fig 11: Table of movies with rating >= 3 and adult = Y and songs more than 6 (Level 4)

id	name	rating	adult	star_presence	songs	invest	first_day_collection	gross_income
6	Hasee To Phasee	>= 3	N	N	<= 6	0-50	<= 10	0 - 75
7	Heropanti	>= 3	N	N	<= 6	0-50	<= 10	0 - 75
24	Bhaag Milkha Bhaag	>= 3	N	Y	<= 6	0-50	<= 10	> 75

Fig 12: Table of movies with rating >= 3 and adult = N and songs less than or equal to 6 (Level 4)

id	name	rating	adult	star_presence	songs	invest	first_day_collection	gross_income
10	Gulaab Gang	>= 3	N	Y	> 6	0-50	<= 10	0 - 75
11	Shaadi Ke Side Effects	>= 3	N	Y	> 6	0-50	<= 10	0 - 75
19	Student Of the Year	>= 3	N	N	> 6	0-50	<= 10	0 - 75
25	Shuddha Desi Romance	>= 3	N	N	> 6	0-50	<= 10	0 - 75

Fig 13: Table of movies with rating >= 3 and adult = N and songs more than 6 (Level 4)

id	name	rating	adult	star_presence	songs	invest	first_day_collection	gross_income
24	Bhaag Milkha Bhaag	>= 3	N	Y	<= 6	0-50	<= 10	> 75

Fig 14: Table of movies with rating >= 3 and adult = N and songs less than or equal to 6 and star_presence =Y (Level 5)

id	name	rating	adult	star_presence	songs	invest	first_day_collection	gross_income
6	Hasee To Phasee	>= 3	N	N	<= 6	0-50	<= 10	0 - 75
7	Heropanti	>= 3	N	N	<= 6	0-50	<= 10	0 - 75

Fig 15: Table of movies with rating >= 3 and adult = N and songs less than or equal to 6 and star_presence =N (Level 5)

5. CONTRUCTION OF DECISION TREE AND CLASSIFYING TEST SET

The final decision tree based on the calculations made on the collected data is given below. With the help of this tree, previously unseen samples can be easily classified. Thus we can now predict the 'Gross income' of other films in future. The Fig.16 shows two test cases which are required to be classified. Now we can use the decision tree to predict the gross income of these movies. Fig. 17 shows the Decision tree which has been prepared using the training set. It can be observed that the tree considers attributes like, invest, rating, songs, adult, star_presence to classify the sample set.

id	name	rating	adult	star_presence	songs	invest	first_day_collection	gross_income
32	Yeh Jawaani Hai Deewani	>= 3	N	Y	> 6	0-50	> 10	?
33	Heroine	< 3	N	Y	<= 6	0-50	<= 10	?

Fig 16: Test Set for the Decision tree

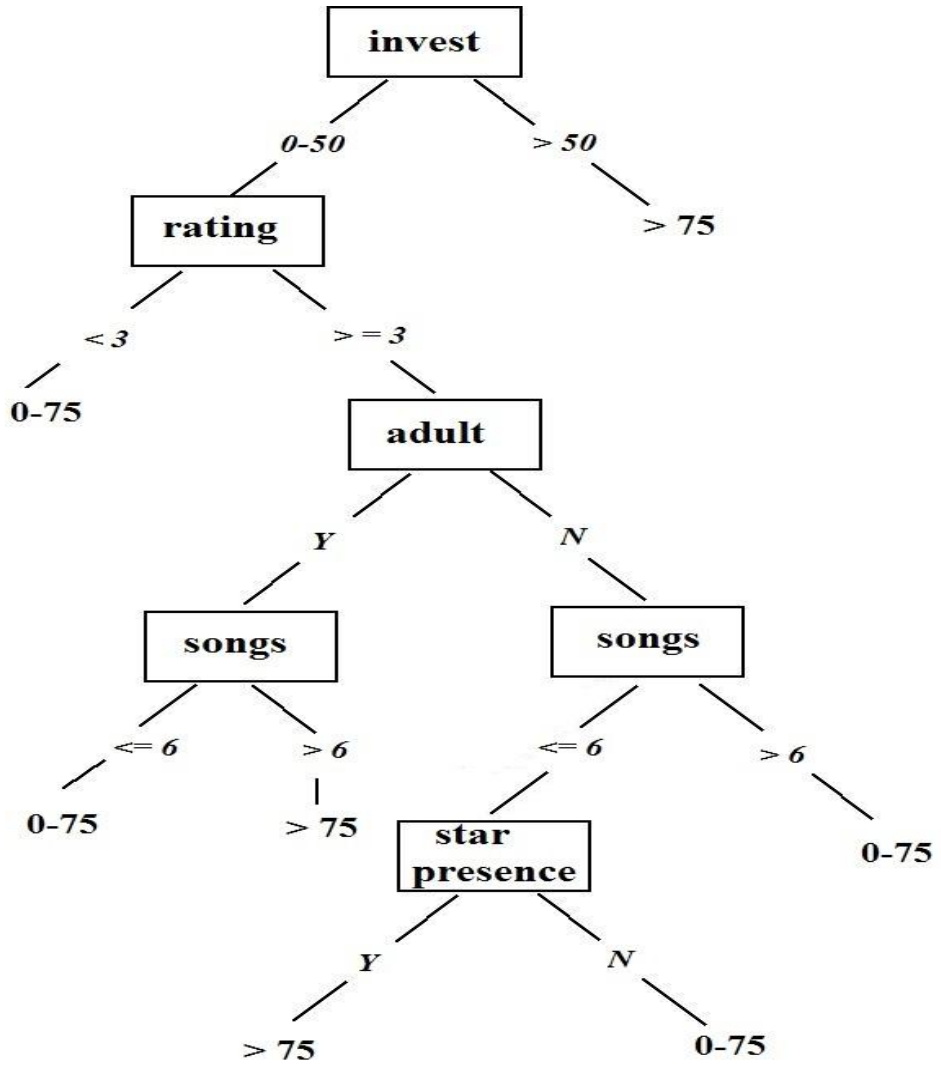

Fig 17: Final Decision Tree Predicting Gross Income of a Movie

id	name	rating	adult	star_presence	songs	invest	first_day_collection	gross_income
32	Yeh Jawaani Hai Deewani	>= 3	N	Y	> 6	0-50	> 10	> 75
33	Heroine	< 3	N	Y	<= 6	0-50	<= 10	0 - 75

Fig 18: Test Set Classified using Decision Tree

By using the Decision Tree, the two test cases can be easily classified. The first test case of movie, 'Yeh Jawaani Hai Deewani' has an investment in the range 0-50 crore(1 crore = 10 million). It has rating greater than 3, it does not have a CBFC 'A' certificate, it has songs less than 6 and casted a super star in lead role. Therefore according to the decision tree, it should have gross income more than 75 crore, as shown in fig. 18. In reality, this movie has earned more than 750 million [4]. The second test case of movie 'Heroine' has investment in the range of 0-50 crore (1 crore= 10 million), but has critic rating less than 3. Therefore according to the decision tree, it should have gross income in the range 0-75 crore. In reality, movie heroine has earned 440 million (44 crore) [4][5]. Thus both the test cases have been correctly classified by this decision tree as shown in fig.18.

6. DECISION TREE IN RULE FORM

The decision tree drawn above can also be expressed in the form of rules. These rules can be considered as a ready to use form of the decision tree.

IF invest > 50 THEN gross income >75

IF invest < 50 AND rating < 3 THEN gross income 0-75

IF invest < 50 AND rating >3 AND adult=Y AND songs<=6 THEN gross income 0-75

IF invest < 50 AND rating >3 AND adult=Y AND songs > 6 THEN gross income >75

IF invest < 50 AND rating >3 AND adult=N AND songs<=6 AND star presence =Y THEN gross income > 75

IF invest < 50 AND rating >3 AND adult=N AND songs<=6 AND star presence =N THEN gross income 0-75

IF invest < 50 AND rating >3 AND adult=N AND songs > 6 THEN gross income 0-75

7. CONCLUSION

The paper has analyzed different factors influencing a movie's income. It has applied the decision tree learning method to predict the films Gross Income. It exploits the advantages of decision tree learning, to successfully make required prediction. This prediction can be very useful to many stakeholders, which include producers, distributers and even actors. It can be inferred that decision tree learning is emerging as an important technique in solving the machine learning problems.

8. ACKNOWLEDGMENTS

I would like to acknowledge all the researchers who have done pioneering work in the field of machine learning and artificial intelligence. Also I express sincere thanks to the authors whose papers have been used as reference in this paper. I would specially like to thank all the websites, especially Times Of India, Hindustan Times, bollywoodhungama Websites for providing statistical data about movies needed for the paper.

9. REFERENCES

[1] Rokach, Lior; Maimon, O. (2008). Data mining with decision trees: theory and applications. World Scientific Pub Co Inc. ISBN 978-9812771711

[2] http://www.cs.princeton.edu/courses/archive/spr07/cos424/papers/mitchell-dectrees.pdf

[3] http://staffwww.itn.liu.se/~aidvi/courses/06/dm/lectures/lec3.pdf

[4] http://timesofindia.indiatimes.com/entertainment/hindi/movie-reviews

[5] http://www.hindustantimes.com/entertainment/entertainmentsectionpage-reviews/seclid.aspx

[6] http://www.bollywoodhungama.com/movies/reviews

[7] http://indiatoday.intoday.in/section/67/1/movies.html

[8] http://www.bollywoodlife.com/news-gossip/how-is-the-bollywood-helping-the-indian-economy/