**Data Mining Project Analysis and Report**

Spring 2017

**Team Members**

Sapna Bohra

Sarchina Kumari

Pinky Sindhu

Poonam Dhingra

[**Why IMDB Dataset?**](#_kpdm55rqtypv) **3**

[**Problem Statement**](#_4fngwmcf0umq) **3**

[**Data Preprocessing**](#_p7w6635kfmcl) **3**

[**Describing and Exploring the data**](#_w6ywule0ov42) **4**

[**Data Mining Models**](#_sc20cf77kwqy) **5**

[Correlation](#_kmyjhafst3g7) 5

[Frequency Pattern](#_mrbqxknps76e) 6

[Classification](#_5r88ipe5oeyn) 7

[Predict a basic model for the gross earnings of the movie](#_y4js2vh2bxea) 7

[Predict the profit model of a movie by grouping the Director name](#_5m07de13nryb) 9

[Predict the average IMDb score by grouping the movies based on director](#_j8s0zg9re8y3) 10

[Predict the status of a movie based on popularity of the directors and actors](#_e5l6zgmvp5ye) 12

[Predict the Gross of the movie by the most correlated genre](#_fsidyo471825) 13

[Predict the IMDb score of the movie by the most correlated genres](#_sv2cg52et1r6) 14

[Regression](#_u2ozi1mu01ca) 15

[Clustering](#_ox195c9gu7ib) 16

[Comparison of various classification models](#_p2ad1sg5iyal) 18

[Web Mining of the IMDb reviews](#_vqrac1uv3dtp) 19

[Compare the reviews of successful movies by popular and unpopular director](#_d2n0mf1uzcbh) 21

[Comparing reviews of successful movies in two different genres.](#_j5iixc3nu932) 22

[**Visualizations**](#_i3tqfxt0cmoy) **24**

# 

## Why IMDB Dataset?

The reason behind our choosing IMDB dataset was that the dataset contained attributes suitable for many data mining techniques, such as classification, clustering, associations, correlations, text mining and web mining. The data contained several outliers as well that helped us in understanding and implementing data cleaning.

## Problem Statement

**How can we tell the greatness of a movie before it is released?**

Our objective revolves around analyzing the IMDb dataset based on actors, genres, directors, plot keywords, Facebook likes and reviews to find out interesting patterns, associations and correlations using various data mining techniques to generate a predictive model for the movies yet to be released.

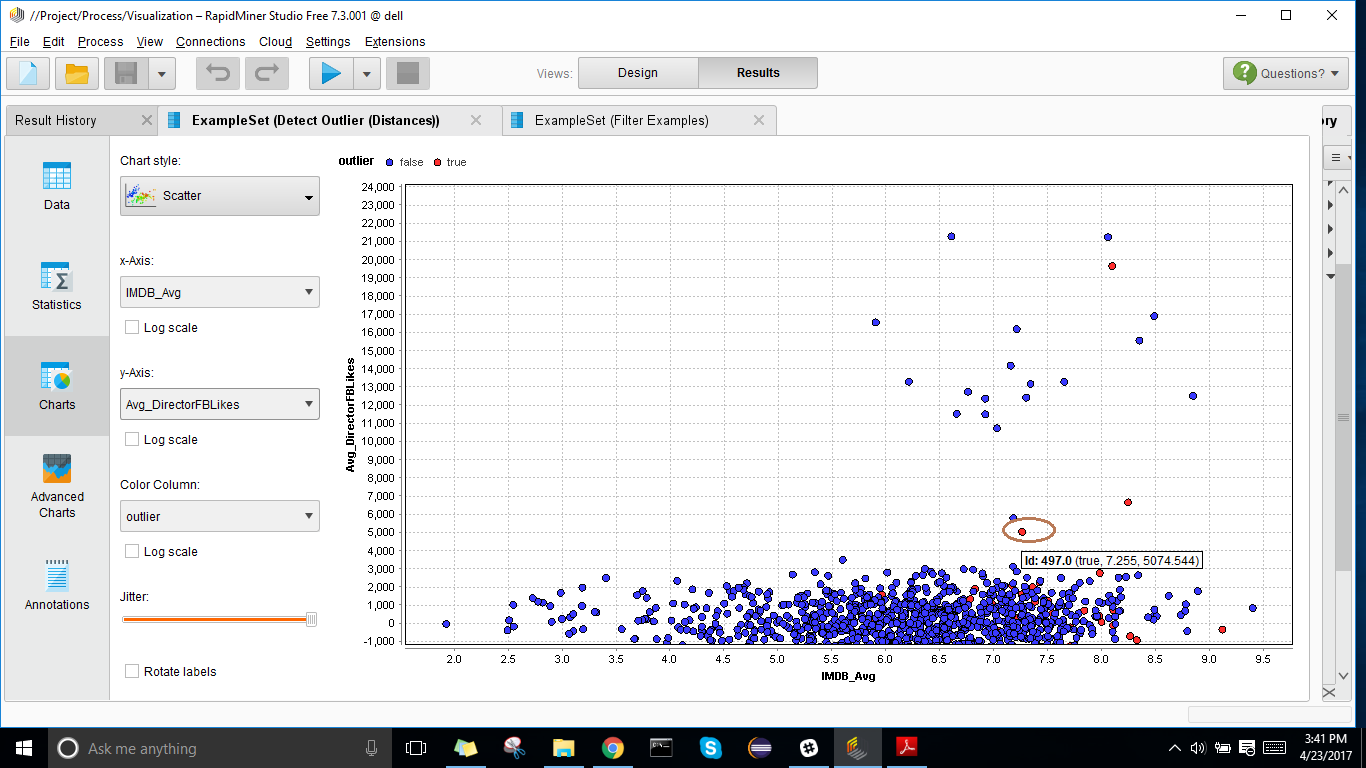
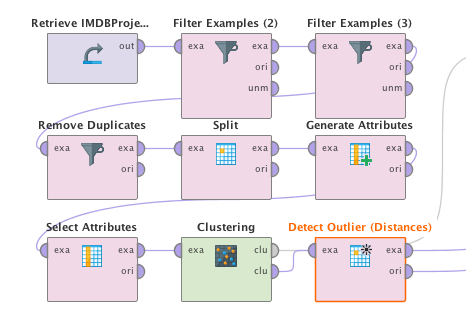
## Data Preprocessing

The dataset contained many missing values, noises and duplicate attributes that would have impacted our analysis through following techniques.

1. **Eliminating Missing Values:** Eliminated the tuples having relevant/classifying attributes as missing values.
2. **Removing Irrelevant Attributes :** The attributes irrelevant for analysis is removed from the datasets. Similarly duplicate attributes conveying the same meaning are also eliminated.

No of Attributes Removed : 10

1. **Outlier Analysis:** We detect several outliers in our datasets. Since the IMDb dataset is popular in US and the popularity have been increased over the recent years, the outliers detected were with the movies released in other countries and released prior to the year 2000. We have detected the outliers using clustering and seen below is the process and our results.



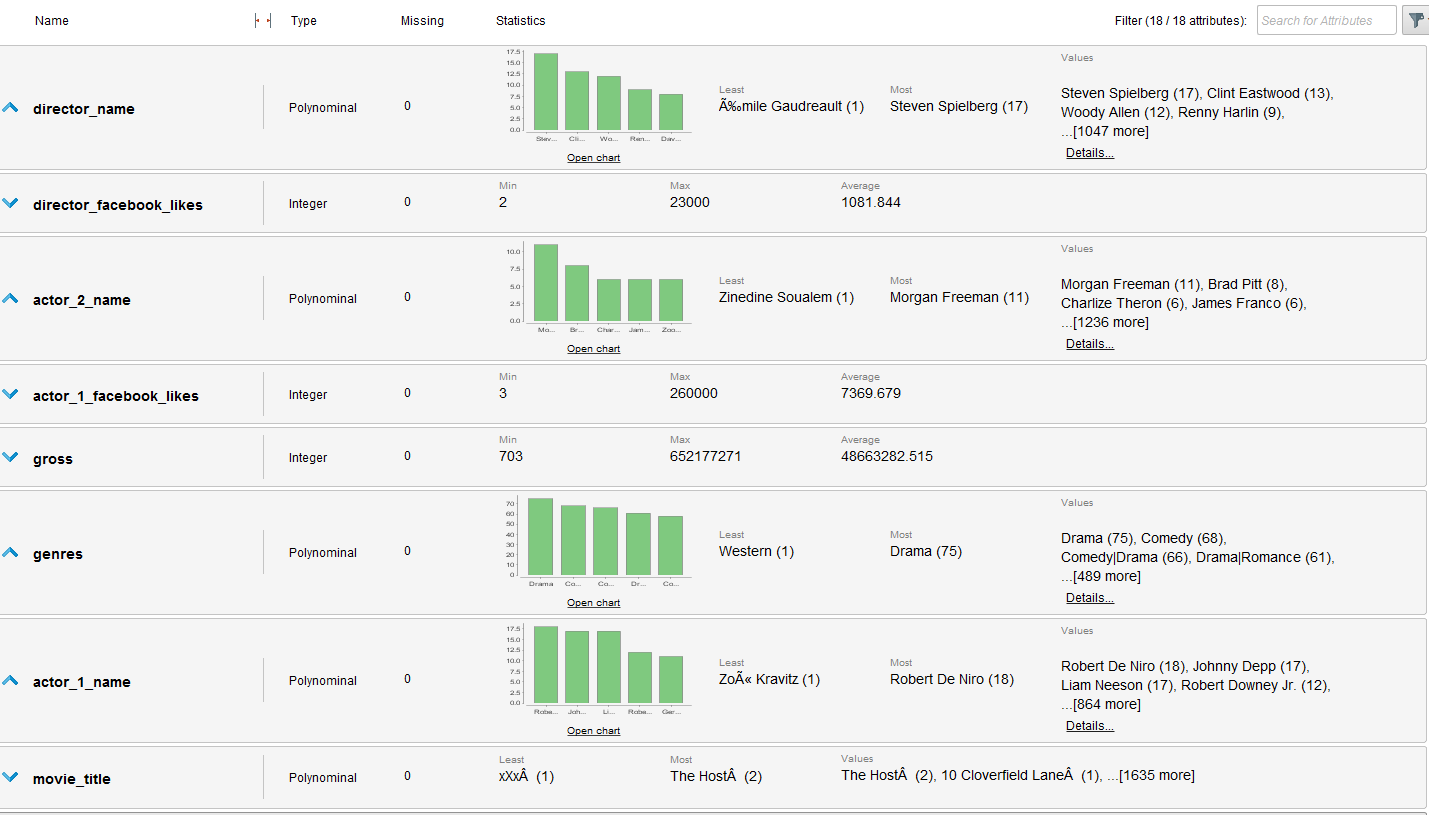
1. **Discrepancy Detection:** There were values as zeros recorded in fields like director facebook likes or user reviews which obviously is a discrepancy, hence we interpreted these type of data as an error and removed those tuples from dataset.
2. **Attribute Subset Selection:** Each cell corresponding to attributes ‘Genre’ and ‘Plot Keywords’ in the datasets contained multiple types of data separated by ‘|’ symbols.. Hence in rapidminer using the “Split” operator, we generated multiple values of each genre/plot keywords in movie. These genres were aggregated (count) based on director for creating classification models.
3. **Getting Rid of Noisy Data:** Since our dataset have many numerical values, it was important for us to normalize those values by discretize them in the rapidminer for effective and meaningful analysis. The discretisation of the numerical values was done into equal frequencies.

## Describing and Exploring the data

The dataset was discovered from Kaggle and included 5000 tuples with 28 attributes. The dataset covers movies released over the past 100 years. After the data preprocessing, we selected the below 17 attributes for our analysis. The goal of our project is classification by consider the IMDb score/facebook likes/gross earnings as the label attributes which determine the likeability of the movie

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Data type** | **Attribute** | **Data type** |
| Director Name | Polynomial | Number of voted users | Integer |
| Director Facebook Likes | Integer | IMDB Review Link | Poylnomial |
| Actor 2 Name | Polynomial | Number of user for reviews | Integer |
| Actor 1 Facebook Likes | Integer | Budget | Integer |
| Gross | Integer | Title Year | Poylnomial |
| Genres | Poylnomial | Actor 2 Facebook likes | Integer |
| Actor 1 Name | Poylnomial | IMDB Score | Integer |
| Movie title | Poylnomial | Movie Facebook likes | Integer |
| Plot Keywords | Polynomial |  |  |

The initial statistics and the analysis of the cleaned data on Rapidminer is as below



1. Director Steven Spielberg is the most common director in the dataset, counts to be 17 times, followed by Clint Eastwood (13) and Woody Ellen (12)
2. Morgan Freeman seems to be most common actor followed by Brad Pitt
3. Drama is most popular genre being used in the dataset and comedy seems to be second and it would be no surprise to see that the combination of both is the most popular combination.
4. Robert Di Niro is most popular actor here followed by Jonny Depp.

## Data Mining Models

### Correlation

Correlation analysis have been evaluated on the dataset to identify the strength of a relationship between the attributes as we have selected nearly 20 attributes for the analysis.

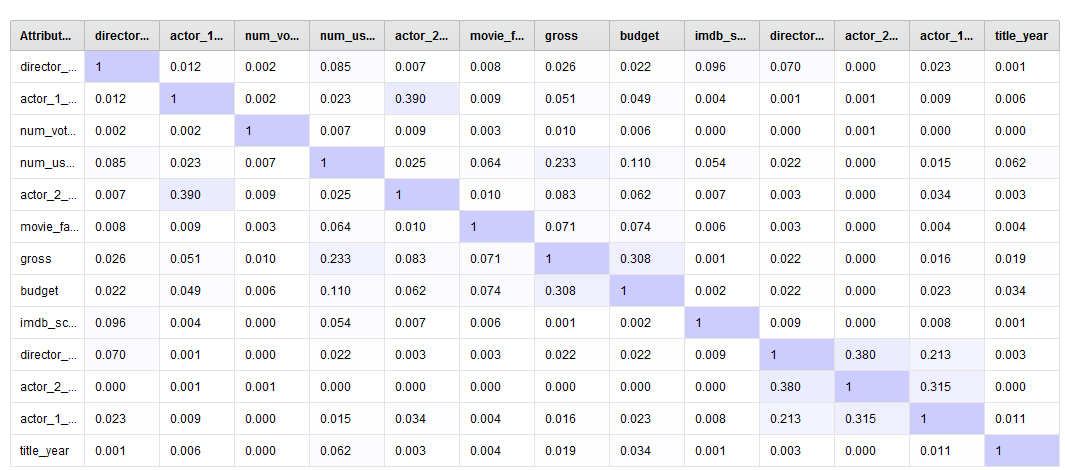
**Analysis**

1. The strongest correlation could be seen between director and actor 2 of the movie, which means that there are high chances for actor 2 and director to be seen together in upcoming movies.
2. We found out a strong relationship between actor 1 and actor 2 so we can say that there is likeability of both appearing in future movies.
3. Similarly there are certain genres of the movie which are highly correlated to each other. The combination of highly correlated genres are family and adventure, family and animation, action and thriller, crime and thriller.

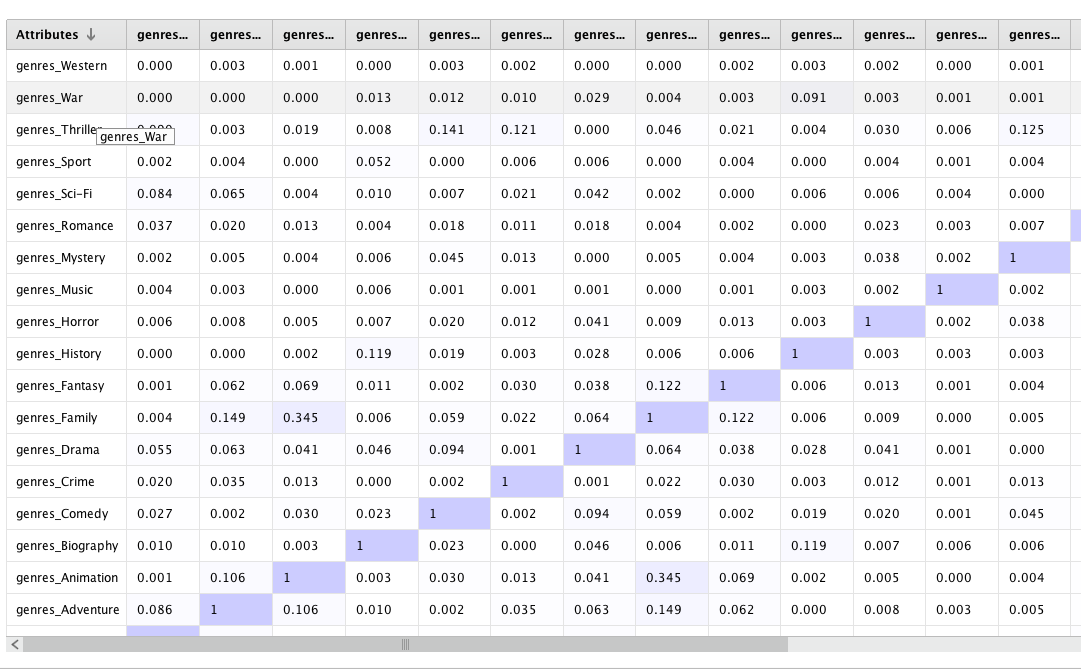
**Process**

**Result**

Correlation of Attributes



Correlation of Genres



### Frequency Pattern

The frequency pattern analysis is conducted based on the plot keywords of the movies. This is analysed by grouping the plot keywords against each genre. We took the plot keywords of movie of genre ‘action’ and ‘drama’. We divided the keywords of entire data into several text files(nearly 30 files) and applied text mining on these words.

**Analysis**

1. The plot keywords associations for the genre Drama, based on Frequency Pattern with minimum support of 0.7

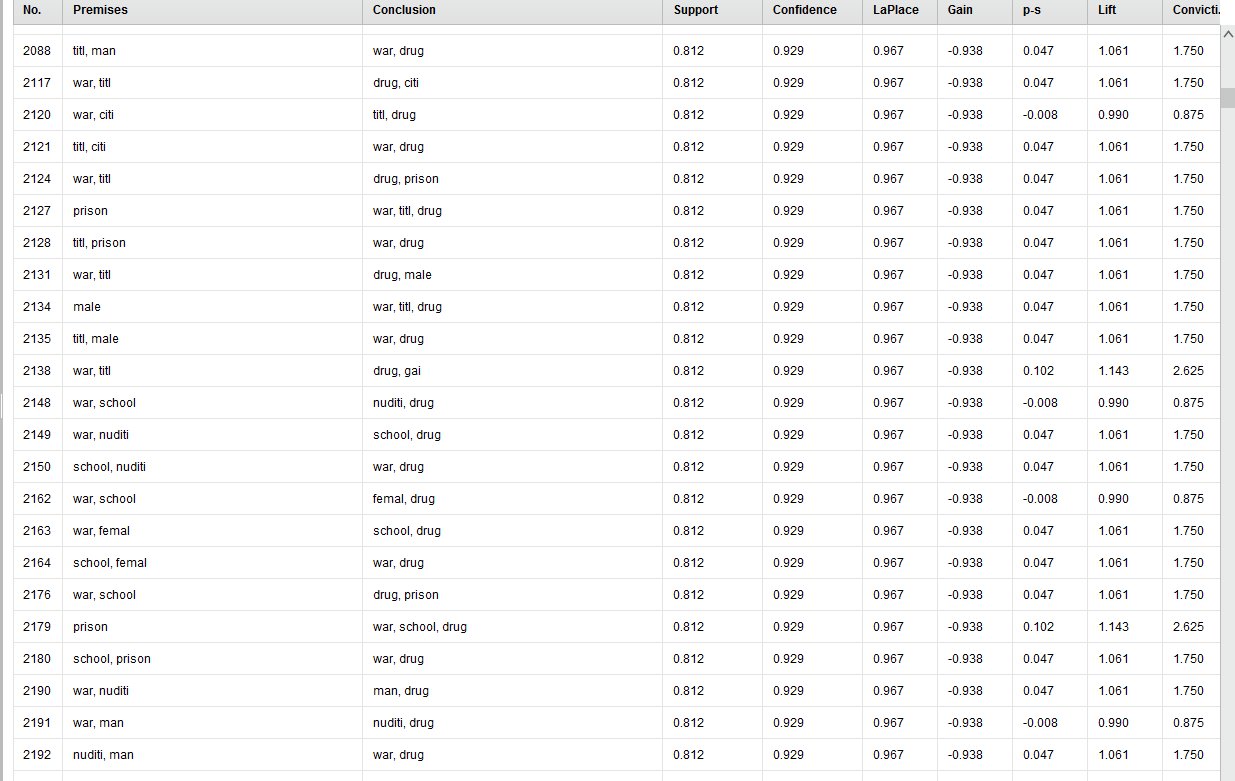
* {Female, Drug}
* {Drug, Citi, York}
* {War, Drug, Citi}
* {Drug, Citi, Prison}

From above itemsets we can say that whenever there is a Drama movie based on drug, there are chances that the plot keywords would contain War, Female and New York city and prison.

1. The plot keywords associations for the genre Action, based on Frequency Pattern with minimum support of 0.7

* Rescue
* Soldier
* Terrorist
* Murder
* Spy

**Result**



### Classification

#### **Predict a basic model for the gross earnings of the movie**

**Classification Type :** Decision Tree, Naive Bayes

**Attributes Selected :** Director, Actor, Movie, IMDb, Budget, **Gross as Label,** Genre, Plot Keywords

**Analysis :** Gross Earnings of the movie depends mainly upon the likeability of the director and actor, facebook likes of the movie and also on the IMDb score of the movie.

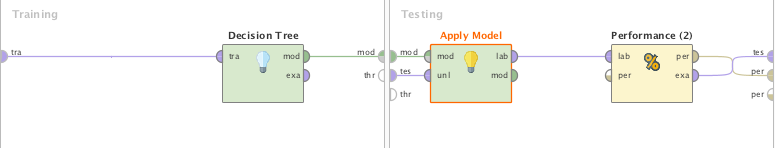
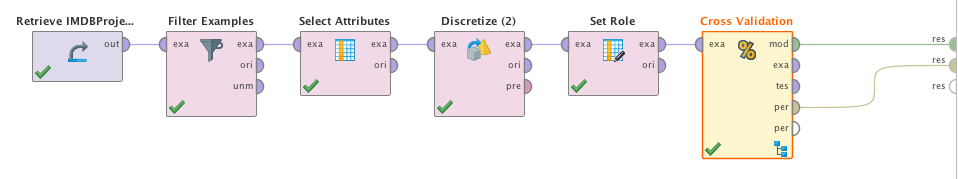
**Decision Tree :** The classification model chosen for this model is Decision Tree. Using the decision tree as the classification, we generated the graphical representation of the earnings of the movie based on the attributes like likeability of the actor/director, budget of the movie, IMDb score of the movie, the genre and the plot of the movie. The decision tree starts with a single root (the facebook likes of the director) which then branches off into a number of solutions.The criterion used for the classification is **gain ratio**, since it is usually a good measure for deciding the [relevance](https://en.wikipedia.org/wiki/Relevance) of an attribute.

**Naive Bayes :** We have ran the same model with Naive Bayes classification assuming strong independence among attributes. Though the prediction percentage was good, it gives poor recall percentage for the classification.

**Operators**: The classification model is generated by discretizing the numerical values of the dataset (facebook like, IMDb scores etc) on the processed and cleaned dataset. The relevant attributes are selected using the ‘Select Attributes’ operator. Then we applied Cross Validationto our predictive models by partitioning the dataset randomly into **10** equal size subsamples.

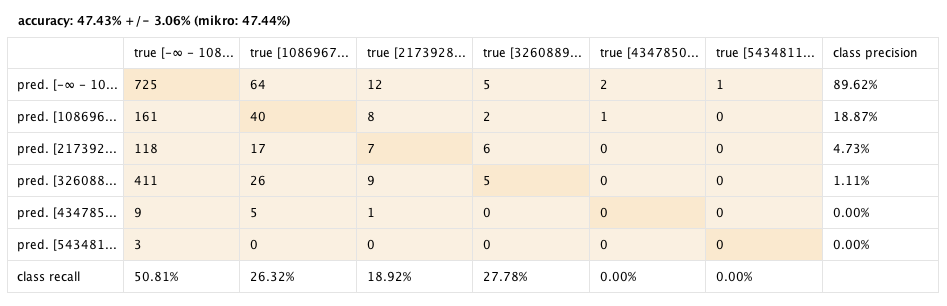
**Conclusion:**The visual representation of the decision tree was easy to interpret our basic earnings predictive model. Thelimitation of our decision making is that we can only select from the known alternatives.

**Process :**



**Result**

Confusion Matrix



Decision Tree

#### **Predict the profit model of a movie by grouping the Director name**Screen Shot 2017-04-22 at 1.15.55 PM.png

**Classification Type:** Naive Bayes

**Attributes Selected**: Count of all Genres, Average number of voted user, Average IMDB score, Average director Facebook likes, Average Gross, Average budget, Average Movie Facebook likes, **Director name** as ID, **Earnings (Profit & Loss)** as Label,Average number of users for reviews.

**Analysis:**

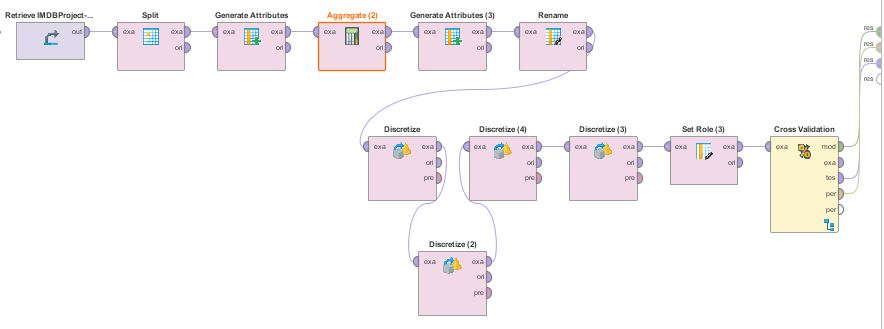
**Precision for Profit Versus Loss**

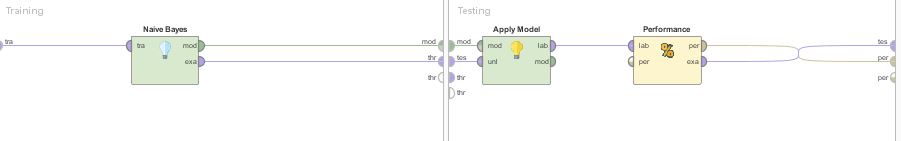
Our data predicted 251 movies to be the profit and 178 movies were actually profit which is a good percentage to predict the precision. However the model is not so good for predicting the precision of loss as out of 800 movies that we predicted loss only 474 were actually loss.

**Recall for Profit Versus Loss**

However, the model is good for recalling loss of the movie as out of 543 movies that were actually loss, 474 were predicted to be loss by model whereas only 178 movies were predicted to be a profit out of 504 movies that were actually a profit.

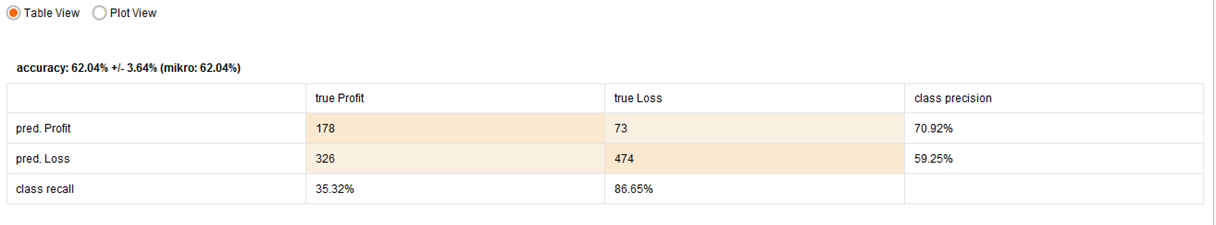
**Process**:





**Results:**

Confusion Matrix



**Conclusion:**

The model is good for predicting the loss of movies.

#### **Predict the average IMDb score by grouping the movies based on director**

**Classification Type**: Naïve Bayes

**Attributes used:** Count of all Genre, Average number of voted user, Average IMDB score, Average director Facebook likes, Average Gross, Average budget, Average Movie Facebook likes, Average number of users for reviews.

**Process**

1. Calculated average of all the numeric attributes by using Aggregate operator.
2. Discretize all the numeric values by using Discretization by Binning operator. And we have discretized IMDB Average by user specification into 4 classes: Good, Very Good, Average and Excellent.
3. Aggregated by the director name.
4. Generated new attributes by splitting all the Genre which helps us to predict the IMDB score by taking all categories of movies into consideration like Action, Fantasy, Romance and many more
5. Generated new attribute named earning which gives Profit and Loss of the movies.
6. Renamed all the attributes.
7. Labelled IMDB Average and using Director Name as ID.
8. Used Cross Validation to check the performance of the model i.e Naïve Bayes.

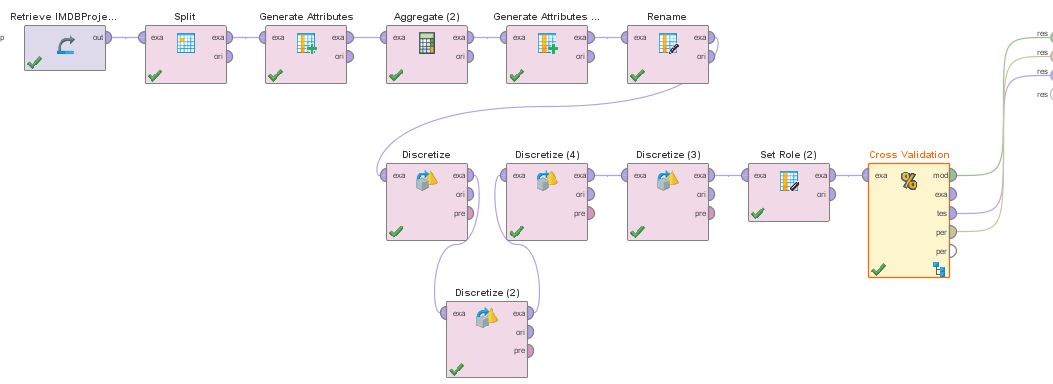


Fig: Process for predicting IMDB Average

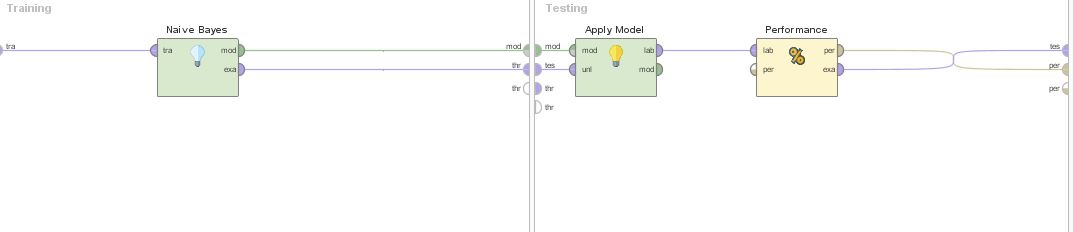


Fig: Subprocess for Cross Validation

**Analysis**: Based on our prior analysis, we performed correlation matrix and found that most of the above attributes are independent of each other. So, Naïve Bayes is the best fit for the above process. Also it outperforms as this data has both nominal and numeric attributes.

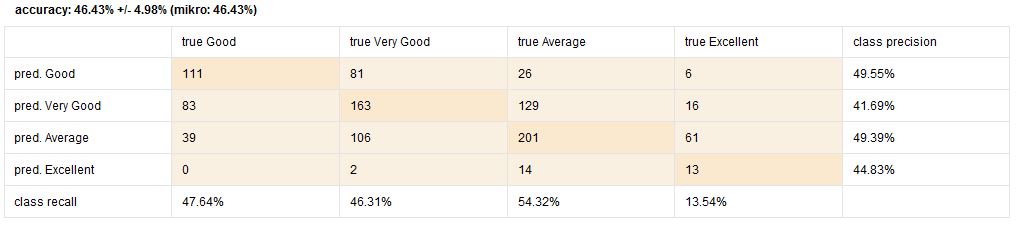
**Class Recall for Average IMDB:**

Our Dataset has total 233 Good movies, but model predicted 111 movies to be Good  
out of 233. For Very Good, out of 352 model predict 163. For Average, our model has best recall of 54.32% (201 out of 370) and for Excellent recall is poor (only 13%).

**Class Precision for Average IMDB:**

The model has best class precision for Good movies and least for Very Good movies. For Good, it predicts 111 movies to be Good from 224 actual Good movies and for Very Good it predicts only 163 out of 391.

**Result:**



**Conclusion:** Overall accuracy of our model is 46.43% with 0.621 +/- 0.21 of absolute error and 0.647 +/- 0.019 of root mean squared error. The model is best for predicting Average movies. We also tried using other Models like k-nn and decision tree but the both have very low accuracy as compared to Naïve Bayes.

#### **Predict the status of a movie based on popularity of the directors and actors**

**Classification Type**: Decision Tree

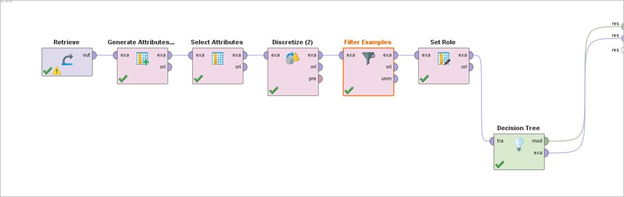
**Attributes selected** : Actor Facebook likes, Director Facebook likes, Movie Facebook likes, Movie status as labeled, Director name as Id.

**Analysis:** Movie status that is if movie is Flop, Good, very good, excellent depends upon the Actor’s and Director’s popularity.

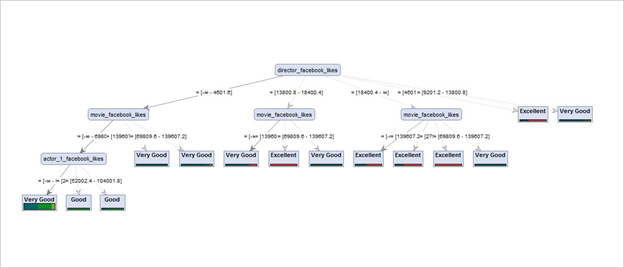
**Decision Tree :** The classification model chosen is Decision Tree. Its predicting the Move status based on the popularity of the actor’s and director’s. We have discretized Actor Facebook likes, Director Facebook likes, Movie Facebook likes.

And Movie status is generated from the IMDB score by keeping a range of the score like movies which have IMDB score of 8-10 are Excellent movies.

**Process:**



**Results:**



**Conclusion :** The decision tree starts with the director Facebook likes and further split into multiple branches. Second level get splits from Movie Facebook likes and then Actor Facebook likes. As we have discretized the values, it’s taking few ranges and based on that it’s predicting the movie status.

#### **Predict the Gross of the movie by the most correlated genre**

**Classification Type :** Decision Tree

**Attributes Selected :** Director, Actor, Movie, IMDb, Budget, **Gross as Label,** Genre, budget, facebook likes

**Generated Attributes** : action-thriller, adventure-family, anime-family, crime-thriller, mystery-thriller

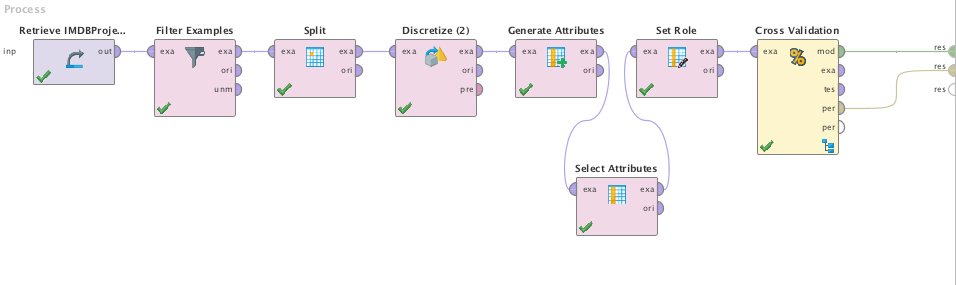
**Analysis :** Gross Earnings of the movie depends mainly upon the highly correlated genre combination anime-family, followed by the actor and movie faceboook likes.

**Decision Tree :** The classification model chosen for this model is Decision Tree. Using the decision tree as the classification, we generated the graphical representation of the gross of the movie based on the attributes like likeability of the actor/director, budget of the movie, IMDb score of the movie, the correlated genres. Being ‘anime-family’ as the highly correlated genre, the decision tree root begins from ‘anime-family’

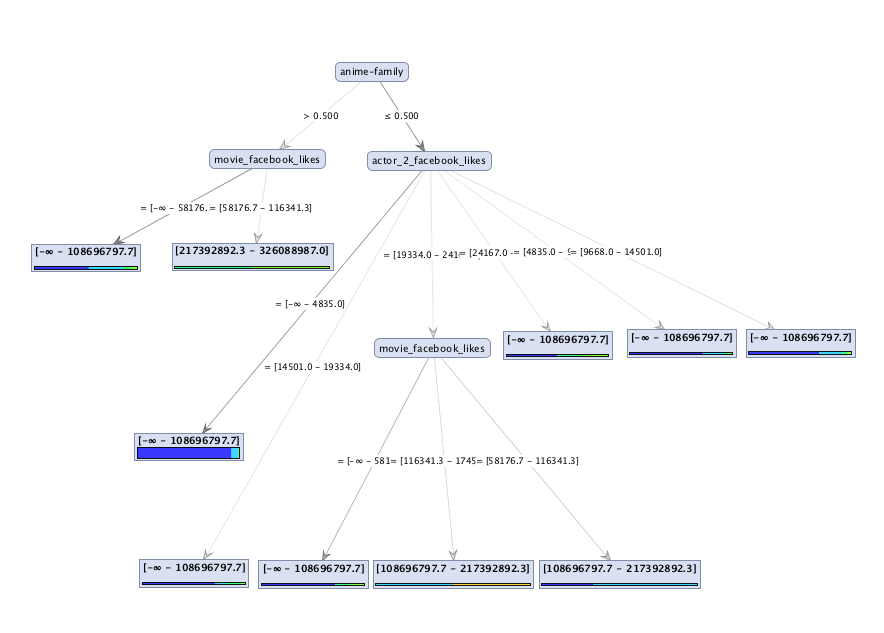
**Operators**: The classification model is generated by discretizing the numerical values of the dataset (facebook like, IMDb scores etc) on the processed and cleaned dataset. The relevant attributes are selected using the ‘Select Attributes’ operator. Then we applied Cross Validationto our predictive models by partitioning the dataset randomly into **10** equal size subsamples. The ‘Generate Attributes’ operator is used to create the set of new attributes which are strongly correlated(based on our correlation analysis).

**Conclusion :** The strongly correlated genres have a significance in the gross earnings of the movie.

**Process :**



**Result :**



#### **Predict the IMDb score of the movie by the most correlated genres**

**Classification Type :** Decision Tree

**Attributes Selected :** Director, Actor, Movie, IMDb, Budget, **IMDb as Label,** Genre, budget, facebook likes

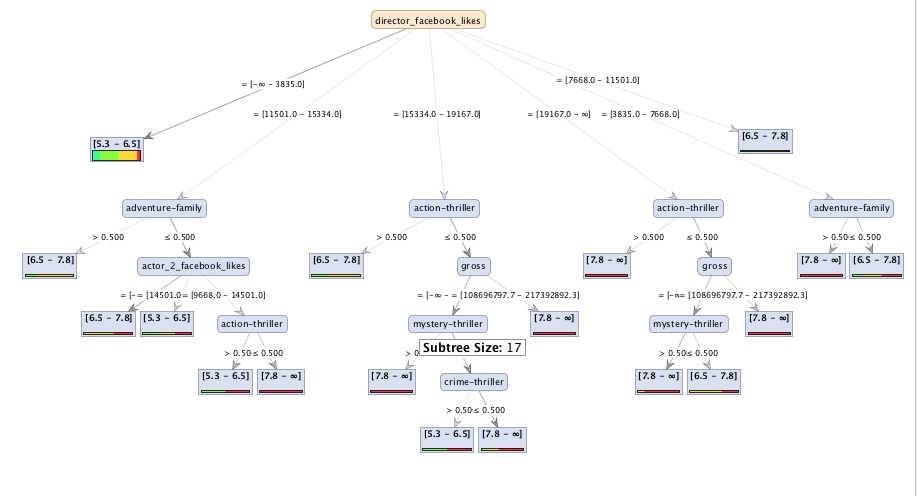
**Generated Attributes** : action-thriller, adventure-family, anime-family, crime-thriller, mystery-thriller

**Analysis :** IMDb score of the movie depends mainly upon the director facebook likes followed by the genre types ‘action-thriller’, ‘adventure-family’, ‘anime-family’, ‘crime-thriller’, ‘mystery-thriller’.

**Decision Tree :** This classification is the extension of the previous model to predict the IMDb score of the movie.

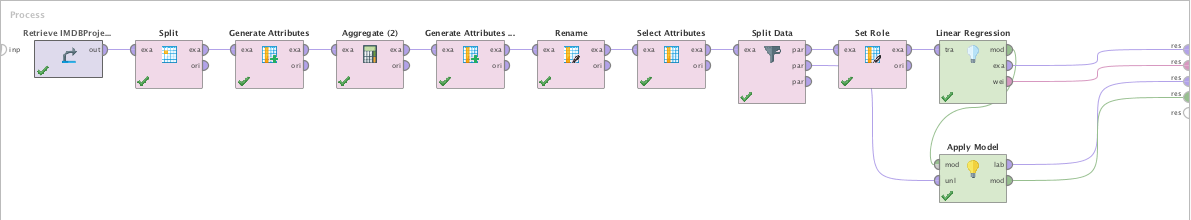
**Conclusion :** The director of the movie detects the likeability of the movie in the IMDb score, followed by the highly correlated popular genres. We can conclude that people who reviews the movie highly consider the director and the genre in which the movie is based upon.

**Result:**



#### **Regression**

**Process**



**Result**



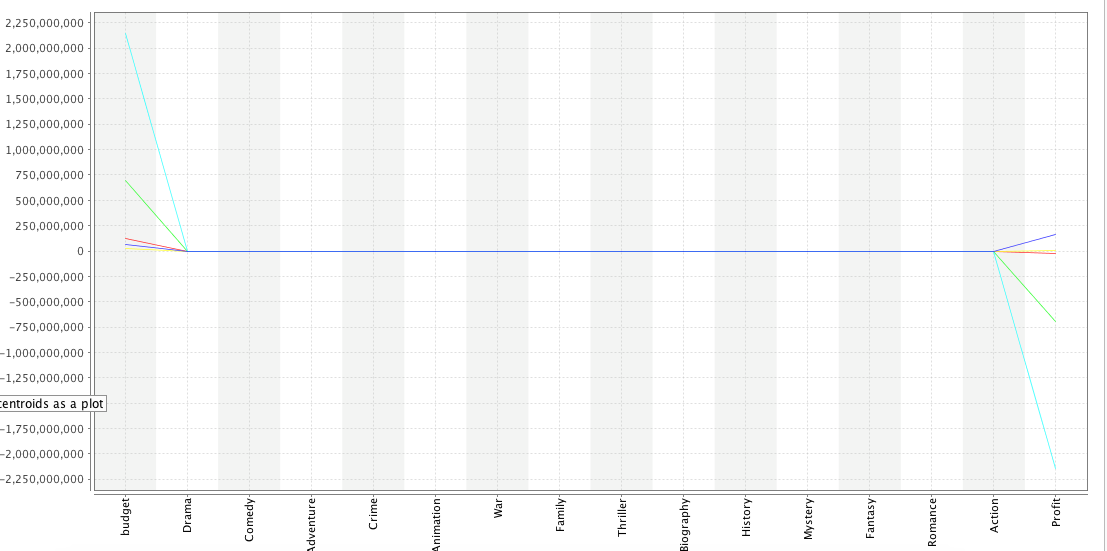
**Conclusion:** The p-value tests the null hypothesis that the coefficient is zero for Num\_Drama, Num\_Adventure, Average Voted User, Num\_Animation and Num\_Thiller. For Average user review has p-value of 0.003. A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. In the result shown above, we can see that the predictor variables of Num\_Adventure, Average Voted User, Num\_Animation and Num\_Thiller are significant because their p-values are 0.000. However, the p-value for Average Gross(1) is greater than the common alpha level of 0.05, which indicates that it is not statistically significant as larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.

#### **Clustering**

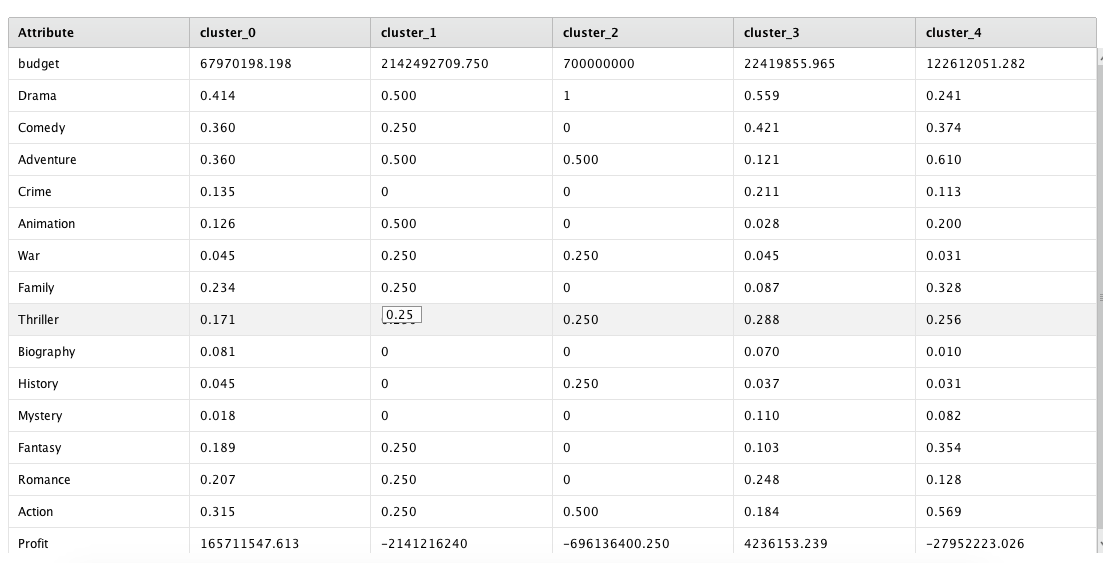
We partitioned a set of data into a set of meaningful sub-classes, called clusters.

**Clustering based on the profit or less earned by the movie**

**Analysis :** When the dataset is clustered into groups after generating the profit and loss of the movie(gross-budget), we can analyze from the centroid table that the cluster with profitable movie contains movies of genres types like drama, action,thriller,crime,fantasy. When movies lacks these genres in the plot of the movie, they incur a loss as seen in the plot below



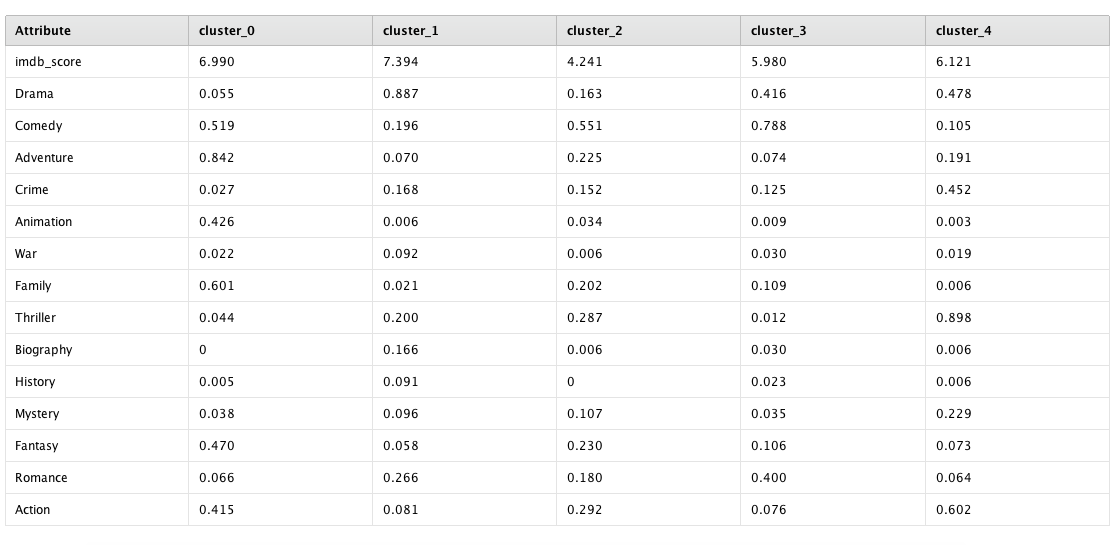
Centroid Table

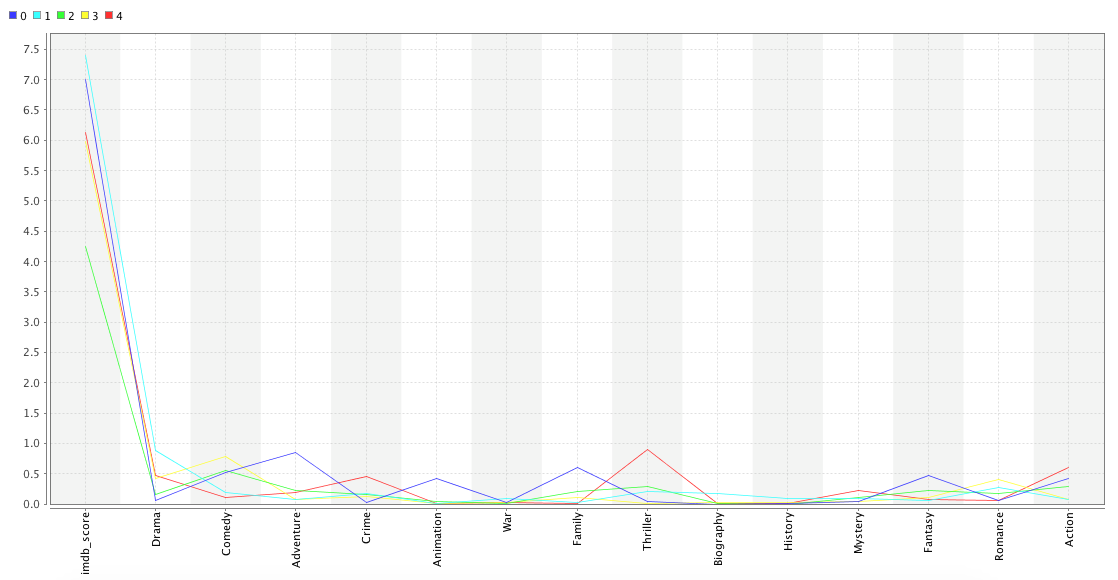


**Clustering based on IMDb**

**Analysis** : When clustering with Genres and IMDb score, it can be detected that the movies of type Genre drama give the highest imdb score. After Drama it is followed by Adventure, family, fantasy and animation as seen in the plot below. Also from the centroid table it can be analysed that the genre biography does not have much impact on the IMDb score.

Centroid Table





#### **Comparison of various classification models**

**ROC Curve**

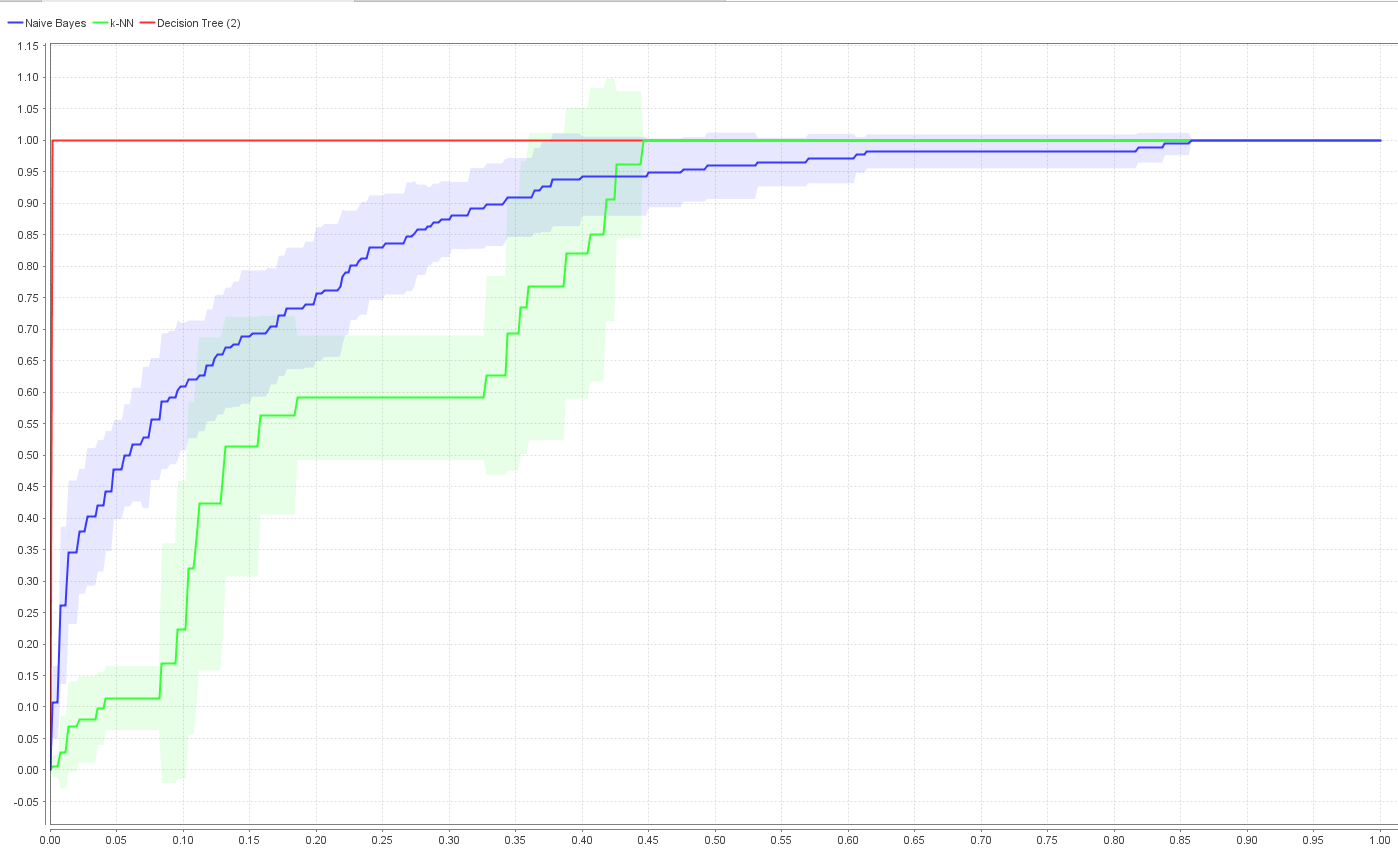
**Parameters Description:**

Number of folds: 10

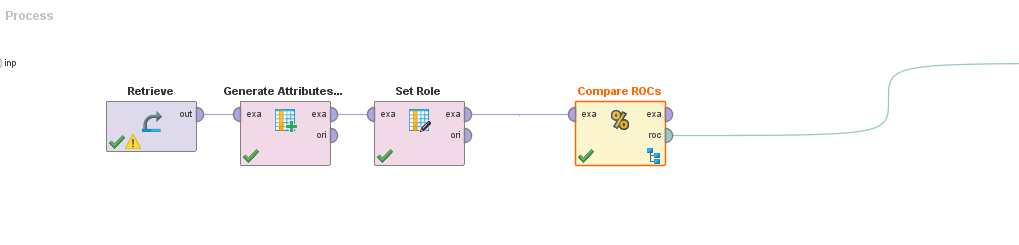
Split ratio: 0.7

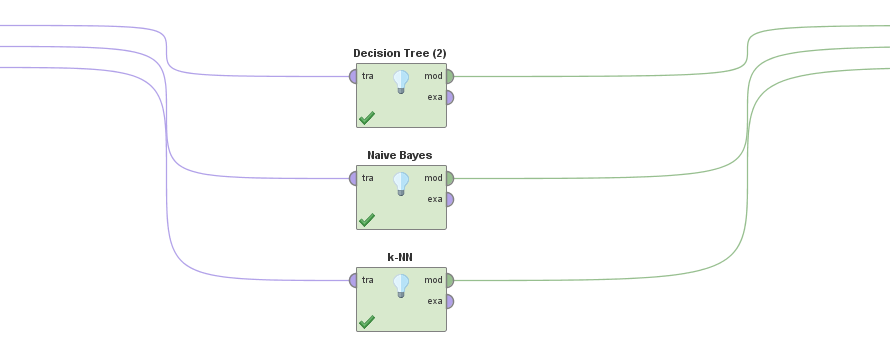
Sampling method: Stratified sampling

**Analysis:** Our analysis show that our data is best for using Decision Tree having probability 1.00 followed by Naive Bayes and K-NN, respectively. Hence our prediction models are based on Decision trees and Naive Bayes.



**Process:**

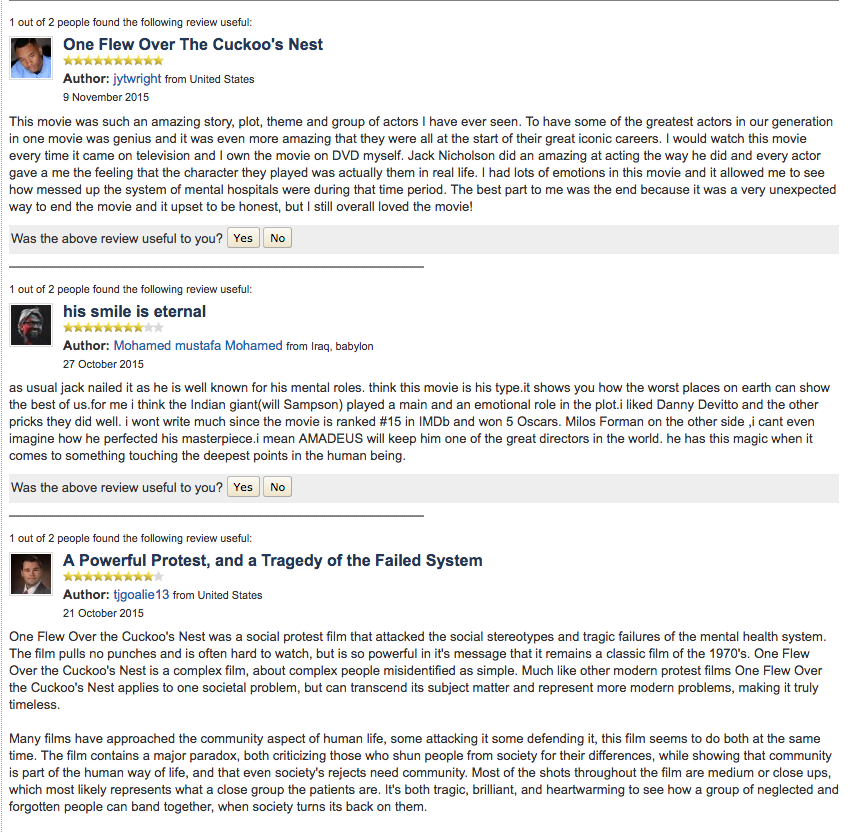




### **Web Mining of the IMDb reviews**

**Purpose :** The IMDb website contains thousands of user reviews for each movie in the dataset. The purpose is to scrap the latest reviews from the IMDb website and analyse some meaningful patterns based on the reviews. The snapshot of the review in the website is as below:

**Sample Review :**

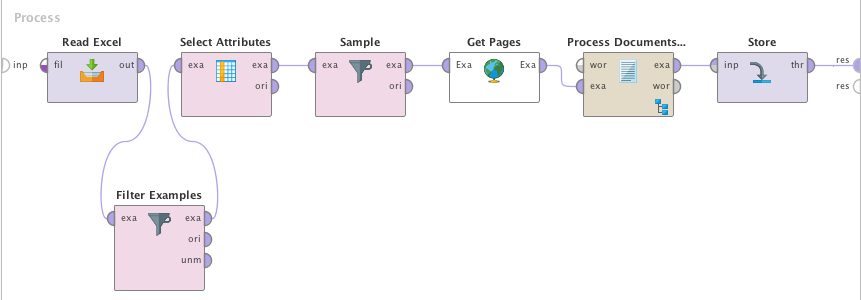


**Procedure :**

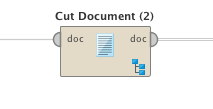
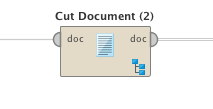
1. Exploring the IMDb reviews from the website.
2. The reviews are sprawled across multiple pages for each movie, and the latest movie reviews will be listed on the first page.
3. Extract the reviews from the website and saved the reviews of each movie in RM
4. Done text mining on saved reviews of selected movie.
5. Create Word List results to Excel file.
6. Use Excel to create word list for Word Cloud
7. Compare the reviews using the cloud.

**Process :**

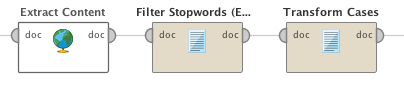
Step 1



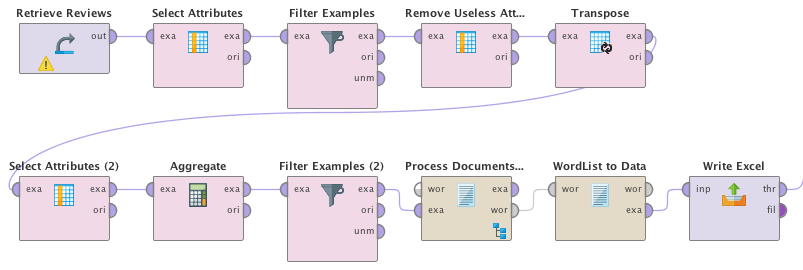
Step 2 & 3

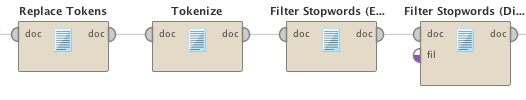
Step 3



Step 4



Step 5



**Operators :**

**Step 1** : After feeding the dataset to the process, ‘Get Pages’ operator fetch the reviews from the website using the the IMDb review link corresponding to each movie is present in the dataset. The Process Documents from Data operator generates word vector from the words from the review link.

**Step 2 &3** : Step 1 is followed by the two levels of sub process using ‘Cut Document’. The operator use regular expressions to scrap reviews related to every user. The second subprocess takes only the review inside the paragraph content of html.

**Step 4** : Using ‘Extract Content’ operator, the content is extracted from the HTML. Finally the extracted content is stored in the repository.

**Step 5** : The repository saves multiple reviews corresponding to each movie. Since every movie contains multiple reviews, several filtering and manipulation is required to place every review corresponding to each movie in multiple rows. The ‘transpose’, ‘aggregate’, ‘filter’ and ‘select’ operators are used in unison to manipulate the reviews into multiple rows. After manipulation of the reviews, the word vector is created.

**Step 6 :** After generating word vector, the vector is tokenized after replacing certain unwanted tokens. ‘Filter Stop Words’ operators are used to remove english stop words and also the stop words from the customised file of stop words. Finally data set is created and is imported into excel.

#### **Compare the reviews of successful movies by popular and unpopular director**

**Details**

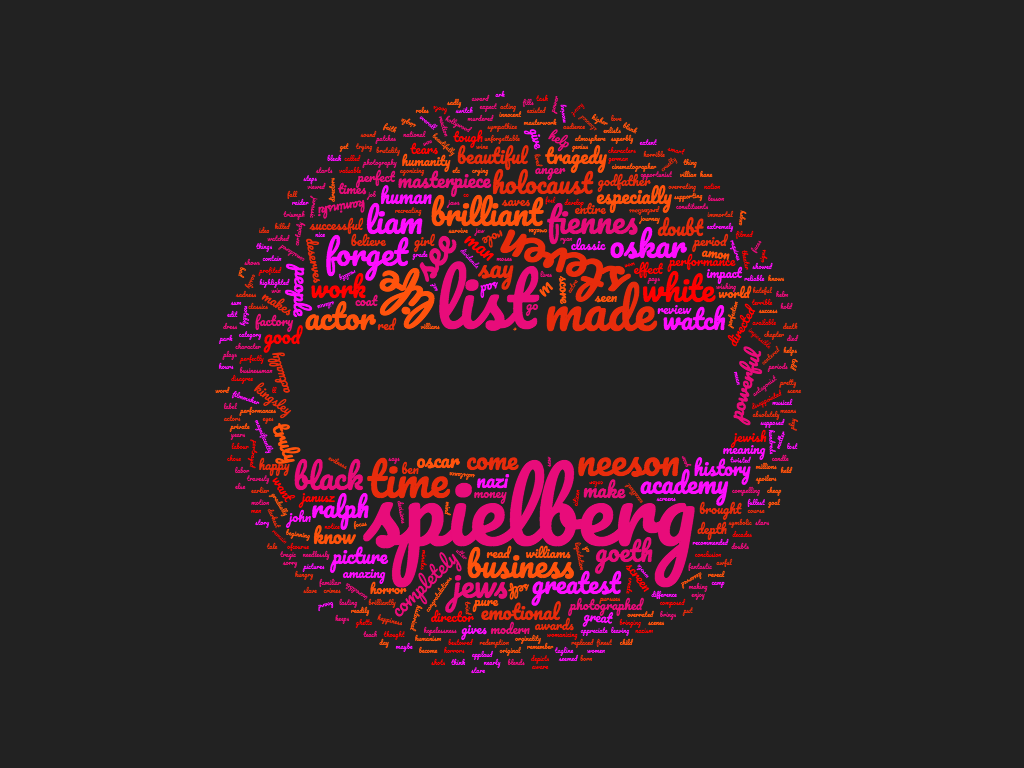
**Movie 1:** Schindler's List

**Facebook Like of the Director** : 14000

**Director Name:** Steven Spielberg

**IMDb Score:** 8.9

***Generated Word Cloud* :**



**Movie 2**: Star Wars: Episode V - The Empire Strikes Back

**Facebook Like of the Director**: 883

**Director Name:**  Irvin Kershner

**IMDb Score:** 8.8

***Generated Word Cloud* :**



**Analysis:** Although both movies have similar rating but it could be seen from reviews that the most emphasized words in movie with famous director revolve around the director itself whereas in the later, the most emphasized keywords are based on the story, and characters. So we can say that the movie with famous director score high IMDb rating because of the director however, latter is famous because of the story.

**Conclusion**: We can hence conclude that the users always comment on the popular actor and director, rather than the plot of the movie when the movie is comprised of popular people.

#### **Comparing reviews of successful movies in two different genres.**

**Details:**

**Movie 1:** One Flew Over the Cuckoo's Nest

**Genre:** Drama

**IMDb Score:** 8.7

**Generated Word Cloud :**



**Movie 2:** Inception

**Genre:** Action & Adventure

**IMDb Score:** 8.8

**Generated Word cloud:**



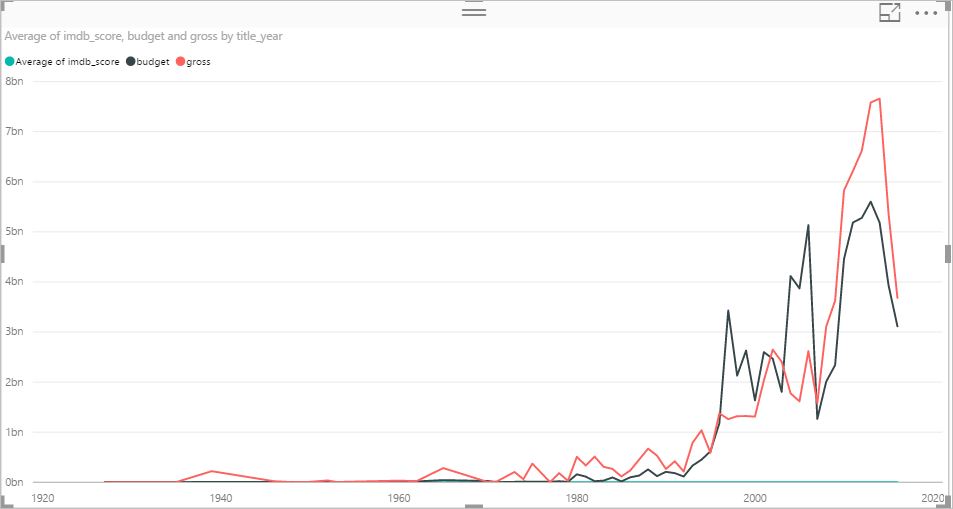
**Analysis:** When it comes to drama, people like to discuss about the plot, story and characters of the movie whereas in action they love to discuss the appearance,look & feel of the movie, the direction and the ideas.

**Conclusion :** We can conclude that when the movie is of type Genre - Drama, the users speak about the plot and characters of the movie rather than the visual effects of the movie. However if the movie if of type Genre - Action, the users speak more about the action, visual effects of the movie rather than the plot of the movie.

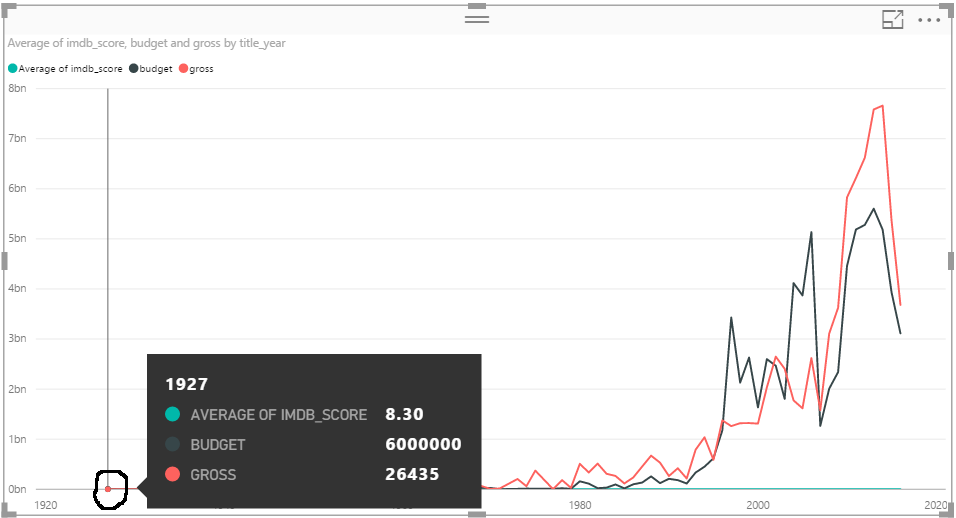
## Visualizations

Here are some data visualizations that we tried using the tool Power BI Desktop. These visualizations give us some insights into our dataset and help us to conclude few things.

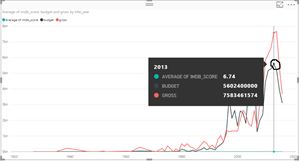
**1.** The role of Gross and Budget in movie’s success?



The above chart is taking data from the year 1927 to 2016 and is helping in predicting the average impact of budget/gross over the period of time. If we hover at the year 1927, budget and gross of the movie is not high, but IMDB score is quite good that is 8.3.



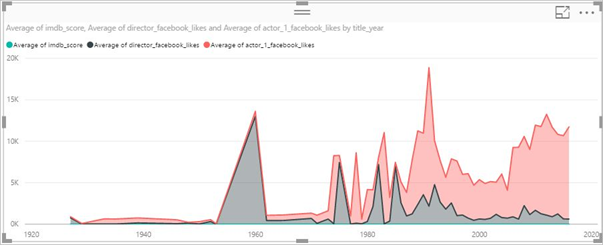
And further, if we analyze the movie with the highest budget and Gross both are giving IMDB score within the range of 6 -7.



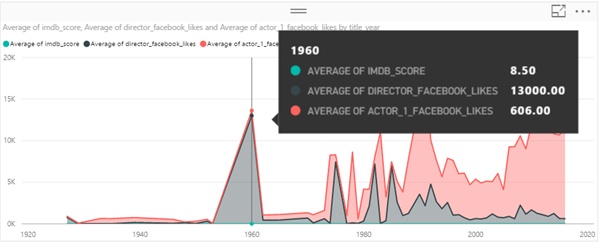


**Conclusion** - After this analysis, we concluded that Budget and Gross has some impact on the movie's success but not that much. Few movies with the small budget have good IMDB scores. If a movie has a high budget and gross that does not necessarily mean that it will be a successful film.

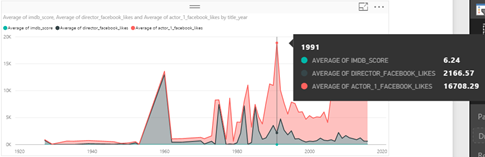
The role of an actor’s and director’s popularity in movie’s success?



Above graph shows that how an actor's and director's fame are making a difference to the movie's success rate. We have analyzed that popularity of a director has good impact on the success rate, but Actor's popularity doesn't guarantee success.

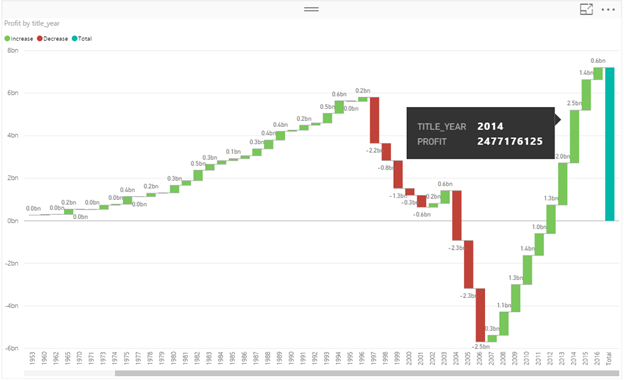


Here we can see that the movies that are made in 1960's have good IMDB score (8.50) and director's and actor's popularity is also very high(average).



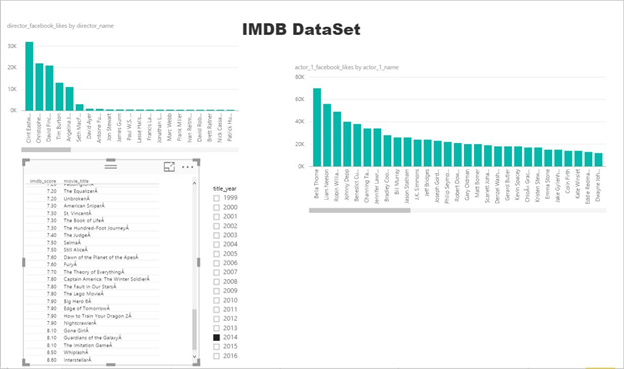


**Conclusion** – In future, it can be easily predicted that the movies whose directors are known and recognized have higher chances of success.

**3.** **Profit and Loss by year and the factors influencing Profit**

By the graph, we can see that from 1997 to 2001, there was a big loss for the movie industry. And again the graph started raising from 2007 to 2016 which shows a notable profit for the film industry. Now let's see what factors played significant role for the benefits of the movie industry. There were total 95 movies during 2014, which made the profit.

The dashboard below shows the top directors and actors who were part of the movie along with movie title along with the IMDB score which is quite good.



4.  **Top profitable Actors and Directors of last 10 years.**

**Top Profitable Actors:**

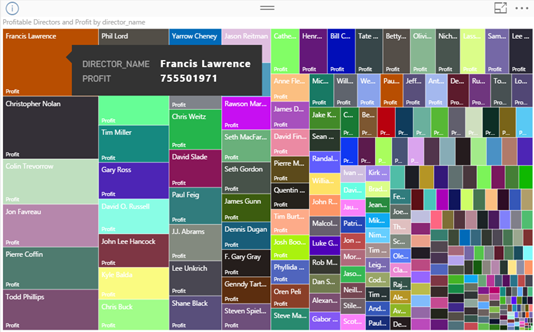
Below is the visualization which gives details about the top profitable actors from the last 10 years like Jennifer Lawrence is the highest profitable actor. Based on that, it can be assumed that the movies with these actors (most profitable) will have a good IMDB score.



**Profitable Directors**

Top Profitable Directors:

Below is the visualization which gives details about the top profitable directors from the last 10 years like Francis Lawrence is the highest profitable director and so on. Based on that, it can be assumed that the movies with these directors (most profitable) will be more successful and have a good IMDB score.



**Conclusion:**

**After analysing all the different models, following are our final conclusions:**

1. Director Name plays significant role in likability of a Movie (Average IMDB, Profit/ Loss).
2. The gross of a movie depends upon the Genre and Actor popularity followed by director.
3. We found strong correlation between certain actors and directors, plot keywords, in certain combination of genres that determines the success of the movie.
4. The movie gross and budget doesn’t play an important role in IMDb score.
5. When a movie is successful, the IMDb users usually comments about the popular actor/director of the movie irrespective of the plot of the movie.
6. Actors popularity doesn’t guarantees a movie success.
7. Top profitable Actors from the last 10 years data are: Jennifer Lawrence, Bradley Cooper, Steve Carell, Robert Pattinson.
8. Top profitable Directors from the last 10 years data are: Francis Lawrence, Christopher Nolan, Colin Trevorrow, Jon Favreau, Pierre Coffin, Todd Phillips.