IMDB Film Prediction with Cross-validation Technique

Shivansh Jagga[1], Akhil Ranjan[2], Prof. Siva Shanmugan G[3]
[1, 2, 3] Department of Computer Science and Technology
[1, 2, 3] Vellore Institute Of Technology, Vellore

Abstract- *The Film Industry has been growing to colossal proportions and these days production of any film requires a considerable investment. According to PwC, the revenue for the industry will grow from 88.3 billion U.S. dollars in 2015 to 104.6 billion in 2019. With such a turnover, the hollywood film industry considerably contributes to the economy of U.S as well. With a lot riding on a movie's success and the highly unpredictable nature of returns, it is imperative to have a system which can predict the success of a film. Thus, a brief hiatus for the same is prudent for any person making a film.*

Keywords*-* Naive Bayes, KNN Algorithm, K-fold Cross Validation, Genetic Algorithm

I. INTRODUCTION

The Film Industry is a multi-billion dollar industry, and generates around $10 billion revenue annually. However, the situation of the industry is not pretty at the moment-- around 80% of the industry's income in the last 10 years has come from less than 10% of the films released. The remaining, 78% on the other hand have lost money.

However, the uncertain behaviour of whether a movie will be successful does makes the movie industry in US market a risky attempt. Because of the huge difficulty and dubiety in the film industry's features, the movie market is one of the riskiest projects for an investor to take in today's world.

Even apart from all this, according to recent research, we can find out whether a film is going to be successful or not and it can be predicted by studying the features of the movie.
Credits of a movie are of the most importance. In this project, the work is to study a new movie's rating, which includes the finances and the viewers' ratings, which is based on the data available on IMDB. Machine learning algorithms including Naïve Bayes theorum and KNN algorithm. Additionally, Genetic Algorithm is also used to optimise the framework.

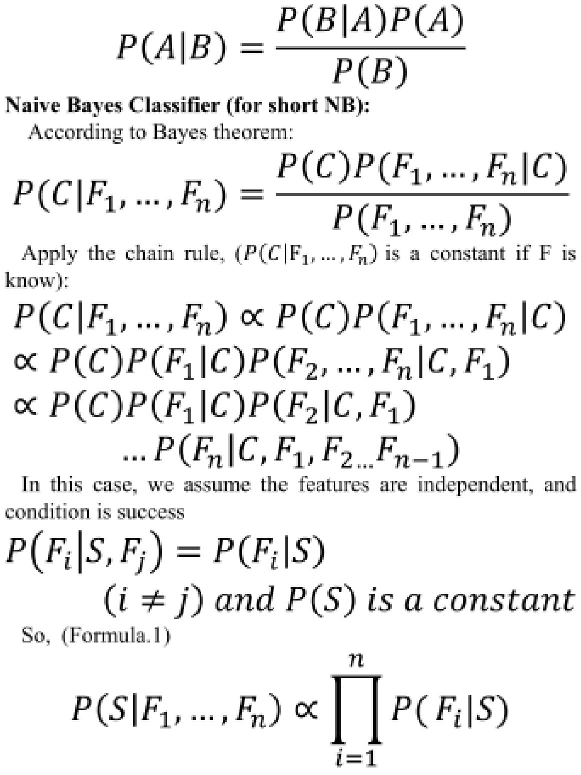
In the 2nd part, theory to support NB classifier is given. The 3rd part is about the data collection. In the 4th and 5th part, the prediction model and the validation of the model is done. The 6th part is about Genetic Algorithm and its working.

II. PRINCIPLE THEORY AND METHOD

A. Naive Bayes Theorum

Bayesian Theorem:

"Mathematically, Bayes' theorem gives the relationship between the probabilities of A and B, P(A) and P(B), and the conditional probabilities of P(A|B) and P(B|A). In its most common form, it is: "



$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Naive Bayes Classifier (for short NB):

According to Bayes theorem:

$$P(C|F_1, \ldots, F_n) = \frac{P(C)P(F_1, \ldots, F_n|C)}{P(F_1, \ldots, F_n)}$$

Apply the chain rule, ($P(C|F_1, \ldots, F_n)$ is a constant if F is know):

$$P(C|F_1, \ldots, F_n) \propto P(C)P(F_1, \ldots, F_n|C)$$
$$\propto P(C)P(F_1|C)P(F_2, \ldots, F_n|C, F_1)$$
$$\propto P(C)P(F_1|C)P(F_2|C, F_1)$$
$$\ldots P(F_n|C, F_1, F_{2\ldots}F_{n-1})$$

In this case, we assume the features are independent, and condition is success

$$P\left(F_i\middle|S, F_j\right) = P(F_i|S)$$
$$(i \neq j)\ and\ P(S)\ is\ a\ constant$$

So, (Formula.1)

$$P(S|F_1, \ldots, F_n) \propto \prod_{i=1}^{n} P(\,F_i|S)$$

B. Maximum Likelihood Estimation

In statistics, maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model.

C. KNN Algorithm

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).

The weakness of KNN: "It treats variables the same when calculating distance even though in almost any real problem, some of variables have a greater impact on the final price than others." It is computationally intensive.

D. K-Fold Cross Validation

K-Fold Cross-validation is a statistical method to evaluate how accurate a predictive model will be. In K-Fold Cross-Validation the collected samples are divided by K roughly equal parts. One of them is chosen as the testing set and the rest are training set. After fitting the model with parameters to the training set, we can compute its error in the testing set.

The cross-validation process can be repeated k times and the average result can be used to produce estimation. "The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once."

"The disadvantage of this method is that the training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation. Data Collection And Feature Generation.

E. Genetic Algorithm

GAs are adaptive heuristic search algorithms constructed on the ideas of natural selection and genetics. Genetic Algorithms exploit historical data to set the path of the search into a region of better performance within the search space.

It repeatedly modifies a population of individual solutions. At each iteration, the algorithm randomly selects a set of individuals from the population and uses them as parents to produce the children for the next generation. Before generating successive generation, the cost function for the entire population is calculated to get a ranked list of solutions. The whole process will iterate for many times until some condition is met.

III. DATA COLLECTION

The data is derived by IMDB and collected via IMDB's API. A python script to get the film's information from 1975 to now (as the people's taste changes). In the business feature, there are unique gross for different countries. However, it was seen that there are not enough films that have worldwide gross. Therefore, the US market was concentrated on.

Because, the film can only be searched by title and ID, we plan to get the information by random generated Id on the IMDB website. We choose directors, writers, actors, US' gross and budget as our input features.

The users' ratings are also taken to indicate the success. The plotting of user's ratings indicates that they accord with normal distribution. (Fig.1) In terms of finance, we take the ratio ($Gross/Budget$ of movie) as the feature. However, its scale (from 0 to 7194) is quite different from the users' ratings' (from 0 to 10). Moreover, its distribution is unable to be calculated. So, we need to scale it. To make it accords with normal distribution, we apply logarithm to the ratio. Then we change it on scale 0 to 10. We take it as the final financial rating. The movie's score is the average value of the users' rating and the scaled ratio. Finally, we take the average of user's rating and financial rating of a film as the film's score.

IV. PREDICTION MODEL

A. Naïve Equations

Apply maximum-likelihood estimates to Naïve Bayes model, according to the Theorem 1 in "The Naive Bayes Model, Maximum-Likelihood Estimation" written by Michael Collins and the EM Algorithm. q(S) is interpreted as the probability of being successful. We get (n= number of samples):

$$q(S|f_i) = \frac{\sum_{t=1}^{n} q(F^t|S)\ \{exist\ j, f_i = F_j^t\}}{\sum_{t=1}^{n} count(\ exist\ j, f_i = F_j^t)}$$

Per*son ' score* = (*the sum of the movies ' scores, which have the person*) / (*the number of the movies*)

Instinctively, some kinds of people are more important than others. We divide people into five parts – directors, writers, main actors, secondary actors and the rest actors. We give a weight to each person. The formula changes to:

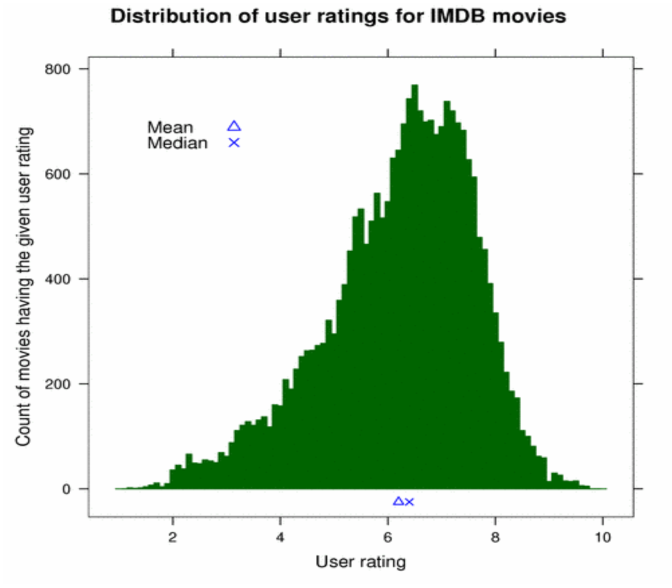
$$P_i = \frac{\sum_{P_i \ in \ movie \ j} M_j * W_j}{\sum_{P_i \ in \ movie \ j} W_j} (Formula \ 1)$$

According to formula 1, we get people's scores. After plotting them and regression, we find it related to the normal distribution function where, μ=5.9, σ= 0.764323:

$$P \sim N(\mu, \sigma)(Formula \ 2)$$

For people not in training set, we use formula 2 to get their scores. Similarly after getting the people's scores, we get the films scores.

B. Weighted KNN



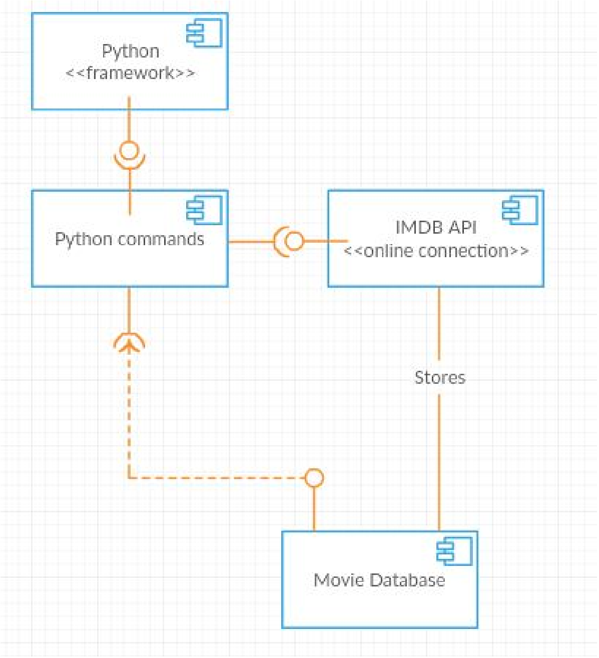
Beside NB, we apply KNN to predict the movies' scores. Each movie has a vector constituted by the five different kinds of people's mean score. The distance between two films' scores is:

$$Dis_{(i,j)} = \sum_{t=1}^{5} (M_t^i - M_t^j)^2$$

To overcome the weakness as stated above, a normal method is to weight the features.

V. TESTING

To test the system, K-Fold Cross-Validation is used and K is 10, because in this method "all observations are used for both training and validation, and each observation is used for validation exactly once." The component diagram is below:



We divide the data into 10 parts. In the K-th time, the K-th part is chosen as the testing set and the rest as the training set. The error is calculated as below. (n=the number of testing set, R is the real score)

$$error = \sum_{i=1}^{n} (M_i - R_i)^2$$

VI. WORKING OF GENETIC ALGORITHM

After we have the validation function and predicting model, we can optimise the weight – w. As, the surface is not steep at the same time not smooth at all and there are a lot of local minimum, GA is chosen to optimise the parameters to avoid local minimum problem.

Roughly, the population size is set to 60, max-iteration times to 110, elite ratio to 0.2 and the mutation probability to 0.2.

Firstly, the domain for each weigh in W. In my experience, "other actors" is the least important kind of people in a movie. Moreover, the fact the weights concern is not the value but the ratio. Therefore, we fix the weight of "other actors" to 0.5 and give the other domains of weight a large range.

So the domain is set as follows:

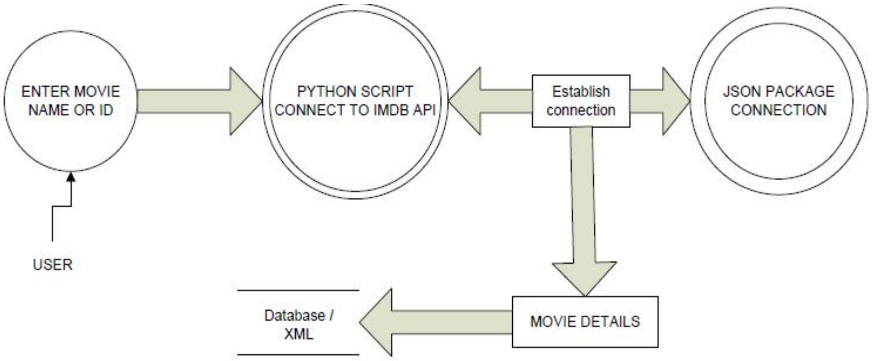
Directors { 3,6 }
Actors { 3, 7 }
Supporting Actors { 0.5 , 2 }
Writers { 0.5 , 1.5 }

VII. FINAL MODEL

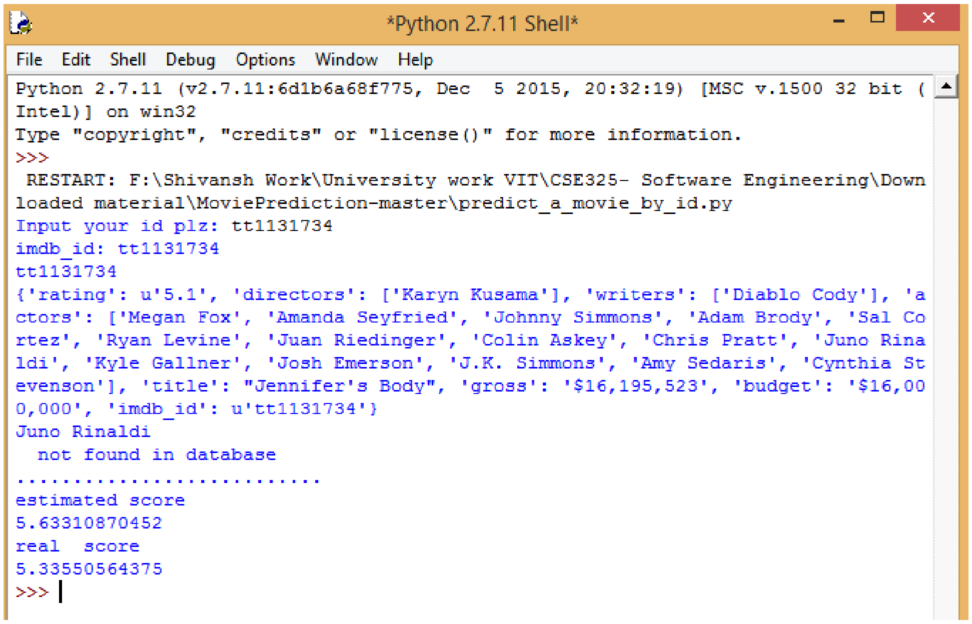
Finally, we have the training data, the prediction system and the optimised weights. Now, we build the final model to predict a movie by combination the GA and NB model.

After the cross validation, the average error based on the whole data (3500 films) is around 0.65, which is acceptable.

A screenshot of the same is given below, along with the data flow model:



Data Flow Diagram



```
Python 2.7.11 (v2.7.11:6d1b6a68f775, Dec  5 2015, 20:32:19) [MSC v.1500 32 bit (
Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> 
 RESTART: F:\Shivansh Work\University work VIT\CSE325- Software Engineering\Down
loaded material\MoviePrediction-master\predict_a_movie_by_id.py
Input your id plz: tt1131734
imdb_id: tt1131734
tt1131734
{'rating': u'5.1', 'directors': ['Karyn Kusama'], 'writers': ['Diablo Cody'], 'a
ctors': ['Megan Fox', 'Amanda Seyfried', 'Johnny Simmons', 'Adam Brody', 'Sal Co
rtez', 'Ryan Levine', 'Juan Riedinger', 'Colin Askey', 'Chris Pratt', 'Juno Rina
ldi', 'Kyle Gallner', 'Josh Emerson', 'J.K. Simmons', 'Amy Sedaris', 'Cynthia St
evenson'], 'title': "Jennifer's Body", 'gross': '$16,195,523', 'budget': '$16,00
0,000', 'imdb_id': u'tt1131734'}
Juno Rinaldi
  not found in database
...........................
estimated score
5.63310870452
real  score
5.33550564375
>>> |
```

VIII. RESULT ANALYSIS AND CONCLUSION

In this project, a big part of the Software Development Life Cycle(SDLC), was the data collection rather than its analysis. We got a model which could be used ideally, although, in reality the collected data has flaws. We must analyze the data and make predictions of the unknown people, maybe using normal distribution or other such techniques, even if it may decrease the accuracy.

Thus, two prediction models are built, on the basis of Genetic Algorithm optimization and the other being KNN model. KNN Algorithm has a lot of computations but does not noticeably increase the accuracy. If we replace GA with something like simulated annealing or ant colony or swarm intelligence optimization, then KNN may be viable, but it doesn’t apply in this case.

The inaccuracy in our model is mainly because of independent features as the combinations of people are important. Example: a new actor could perform well under a good director, or a normal director might make a good film because of a good script. The challenge is considering all these problem.

REFERENCES

[1] Mining gold from the Internet Movie Database, part 1: decoding user ratings By Tom Moertel

[2] http://www.statista.com/statistics/259985/global-filmed-entertainment-revenue/

[3] Tapan Gandhi, Bijay Ketan Panigrahi, Manvir Bhatia, Sneh Anand," Expert model for detection of epileptic activity in EEG signature", Expert Systems with Applications, vol.37 issue 4

[4] Lorenz Cuno, "Genetic algorithms", Availiable:http://www.klopfenstein.net/lorenz.aspx/genetic-algorithms R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[5] Deniz Demir, Olga Kapralova Hongze Lai, Predicting IMDB movie ratings using Google Trends.

[6] https://www.researchgate.net/publication/222530390