Predicting Movie Revenue
using Movie Metadata

Quazi Ishtiaque Mahmud (2012331051)
Asif Mohaimen (2012331054)

Course Teacher
Ayesha Tasnim
Assistant Professor

Machine Learning
Dept. of Computer Science & Engineering
Shahjalal University of Science & Technology
August 2017

1 Dataset

We have used the famous IMDB 5000 Movie Dataset provided by kaggle. This dataset contains 5000+ movie metadata scraped from IMDB. It contains 28 variables for 5043 movies, spanning across 100 years in 66 countries. There are 2399 unique director names, and thousands of actors/actresses.

The variables are listed below:

Movie Title, Color, Number of Critics for Reviews, Movie's Facebook Likes, Duration, Director's Name, Director's Facebook Likes, Actor 3's Name, Actor 3's Facebook Likes, Actor 2's Name, Actor 2's Facebook Likes, Actor 1's Name, Actor 1's Facebook Likes, Gross, Genre, Number of voted users, Cast's total Facebook likes, Number of faces on poster, Plot Keywords, Movie's IMDB Link, Number of user for reviews, Language, Country, Content Rating, Budget, Title Year, IMDB Score, Aspect Ratio.

We actually didn't use these many features to train and predict using our models. These features have continuous values so they are all normalized so that, they have values within range 0 to 1. The features we have used to train and test are listed below:

Director's Facebook Likes, Number of Critics for Reviews, Actor 3's Facebook Likes, Actor 2's Facebook Likes, Actor 1's Facebook Likes, Movie's Facebook Likes, Duration, Number of voted users, Number of user for reviews, Budget, Year, IMDB Score, Total Facebook Likes.

Then the Genres are kept in the following categories and they are considered as binary features meaning if a movie has the following genre Action, Adventure, Mystery the feature value will be like this, [1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0]: Action, Adventure, Fantasy, Sci-fi, Thriller, Comedy, Family, Horror, War, Animation, Western, Romance, Musical, Documentary, Drama, History, Biography, Mystery, Crime.

The content rating is also considered as binary feature. But it only refers to a certain category. So the feature will be like this for Parental Guidance Suggested, [0, 1, 0, 0, 0]. The categories of content ratings are listed below:

General Audiences, Parental Guidance Suggested, Parents Strongly Cautioned, Restricted, Adults Only.

For the features Language & Country, we have kept 1 if the language is English and 0 for all other languages & for USA, 1 and 0 for others respectively.

2 Related works

There were some previous works performed by some famous researchers in the past. We have tried to study some of their works and these are the result findings listed below:

1. **Stanford Project**, 2015
   - Accuracy - 57.9% (Bingo) using SVM Linear Kernel
   - Accuracy - 79.0% (One Class Away) using Naive Bias Classifier
2. **Stanford Project**, 2013
   - Accuracy - 52.0% (Bingo) using Logistic Regression
3. **Stanford Project**, 2011
   - Accuracy - 29.6% (Bingo) using Modified K-means clustering
   - Accuracy - 57.6% (One Class Away) using Modified K-means clustering
4. **Sharda et al.**, Springer, 2006
   - Accuracy - 36.9% (Bingo) using Artificial Neural Network
   - Accuracy - 75.2% (One Class Away) using Artificial Neural Network

We actually couldn't find their datasets for our training or testing purpose. We only found two researches that are performed on this Kaggle IMDB 5000 Movie Dataset. They are listed below:

1. **Markovic Vasilije**, 2016

    - Accuracy - 48.0% (Bingo) using Random Forest Model

2. **Alexei Novikov**, 2016

    - Accuracy - 47.0% (Bingo) using Random Forest Model

3 Our findings

3.1 Defining Class Label

The revenue generated by movie is a continuous value. But to classify we need a fixed number of labels. So to determine the label of a class, we take the $log10$ of that revenue generated by the movie to be our class label. So our class labels were 0 to 8. So we have 9 classes here.

3.2 Architecture of Different Models

All the models were implemented by using the Python SkLearn Machine Learning library.

For SVM linear kernel we have set C=2 and gamma=2. For SVM polynomial kernel we set C=2 and gamma=1.

For MLP we have used 1 hidden layer with 20 neurons. For activation function we used 'relu', the rectified linear unit function, which returns f(x) = max(0, x).

For Logistic Regression we have used 1e-5 as our value for c which is the inverse of regularization strength.

3.3 Results

We have tested several models from the python SKlearn library on this dataset. We have used 70% data for training purpose and 30% data for testing purpose. As our dataset was short and to compare with other researchers we have used 5-fold cross validation. Our findings are shown in figure 3.1 for bingo accuracies in the labeled models. Logistic Regression and MLP gave better bingo accuracies. So we have calculated one class away accuracies for them in figure 3.2. Our comparison with authors of this specific dataset is shown in figure 3.3

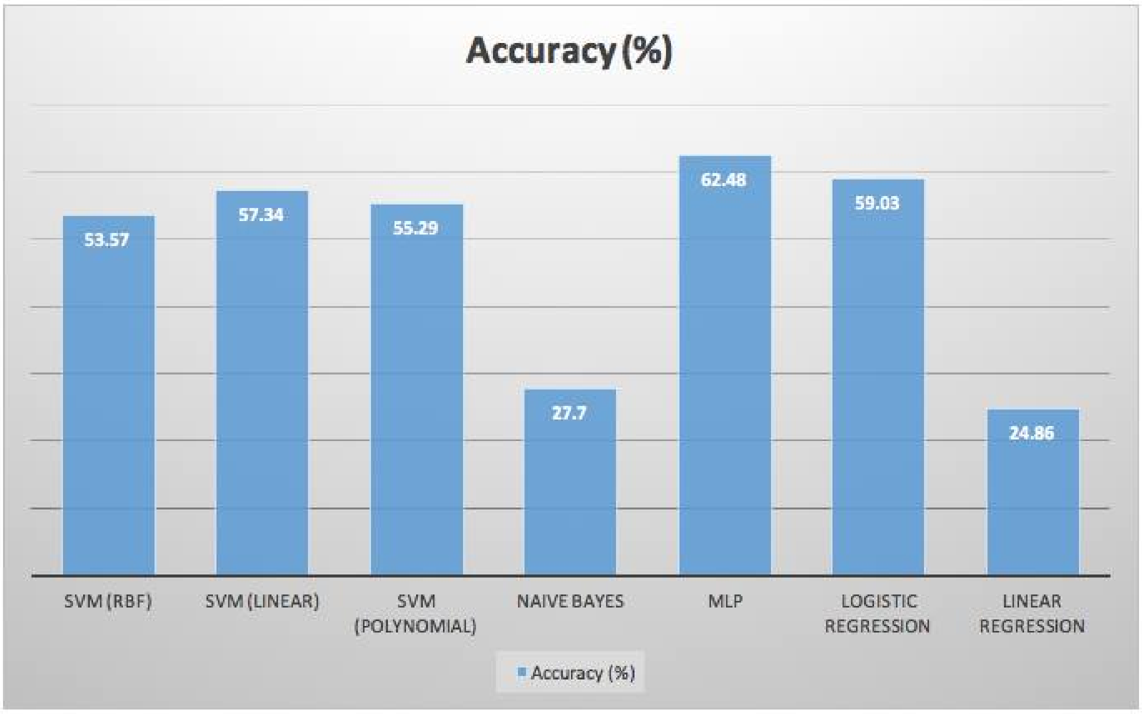

Figure 3.1: Bingo Accuracy for Different Models

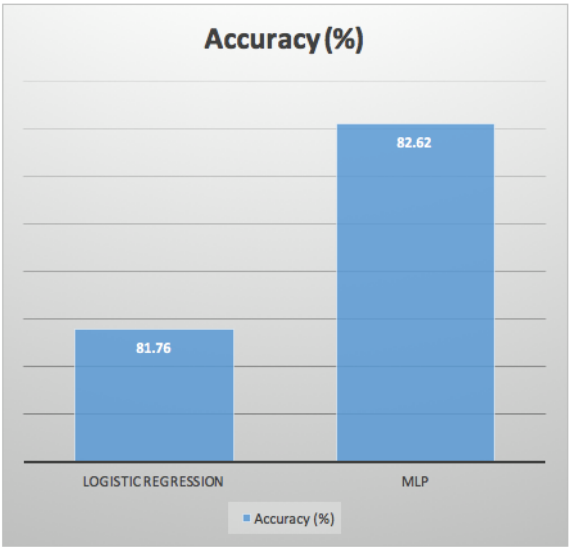

Figure 3.2: One Class Away Accuracy for Different Models

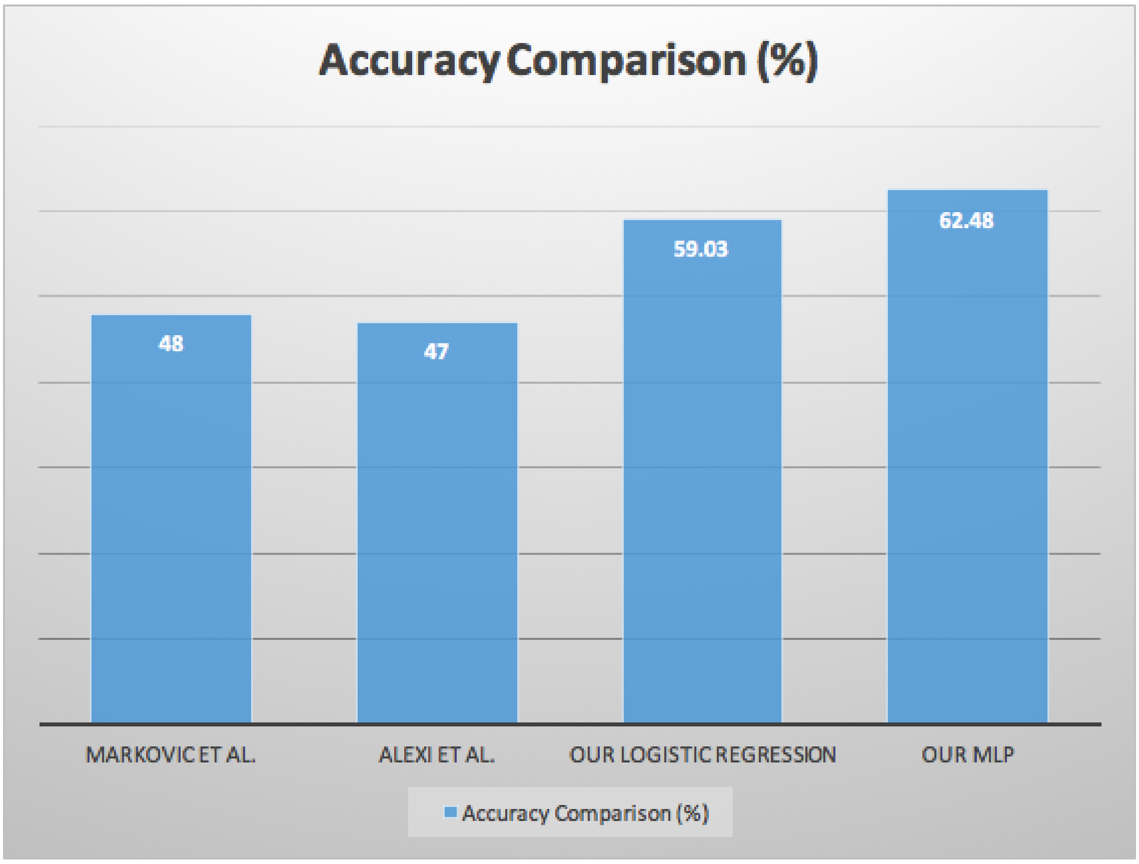

Figure 3.3: Result Comparison with other Authors