Predicting box-office success of Movies

Team No: 33

- Ankush Khandelwal (201302077)
- Gorang Maniar (201364092)
- Ishaan Arora (201301011)

Mentors

Irshad Ahmad Bhat

Sayyad Nayyaroddeen

Aim

This project aims at predicting the box office success of movies by determining their Gross Production. Pattern recognition techniques used to predict the same are Multilayer perceptron(MLP) , Logistic Regression and Support Vector Machines(SVM). The predicted values were then compared to the actual values and efficiency of each model was calculated and comparison of the three models are shown.

Reference Paper

Predicting box-office success of motion pictures with neural networks by Ramesh Sharda and Durus Delen :

https://www.researchgate.net/publication/222530390_Predicting_box-office_success_of_motion_pictures_with_neural_networks

Dataset

- Movies released during 2000-2010.
- Data extracted for the movies from different sources like
  - www.omdbapi.com
  - www.rottentomatoes.com
  - www.metacritic.com

Dataset

- Following are the features of the dataset :

Type	Features
Nominal	Actors, Director, Writer, Production-House, Genre
Numeric al	Budget, IMDB Rating, No of Rating, IMDB Votes, Metascore, Tomato Meter, Tomato User Rating, Tomato Reviews, Tomato Fresh, Tomato Rotten.

Pre-Processing

- For nominal features like Actors, Directors, etc. a numerical value was assigned and normalised.

- For numerical features were extracted, normalised and used.

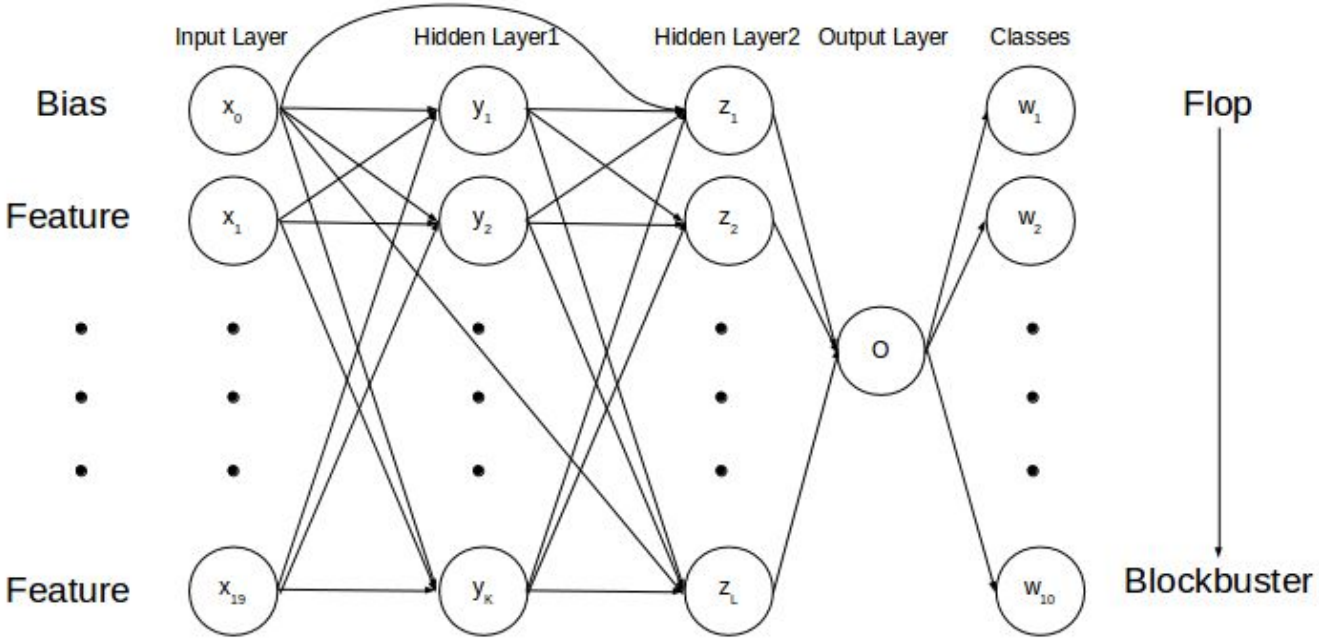
- For each nominal feature, the value assigned to it was the normalised equivalent of its average over income to expenditure ratio for all the movies where it appeared.

Pattern Recognition Techniques

- **Multilayer Neural Network**
  - Number of input features = 19 (Actor, Director, Genre, Production house etc.)
  - Number of Hidden Layers = 2 (optimal)
  - Number of Hidden units in hidden layer = 10 (optimal)
  - Number of output neuron = 1
    - Further Classification to **10 classes**
  - Activation function = Sigmoid $F(x) = 1/(1+e^{-x})$

Class No.	1	2	3	4	5	6	7	8	9	10
Range (in Million $)	<1 (Flop)	>1 <10	>10 <20	>20 <40	>40 <65	>65 <100	>100 <140	>140 <180	>180 <250	>250 (Blockbuster)

Neural Network Architecture



K = 10

L = 10

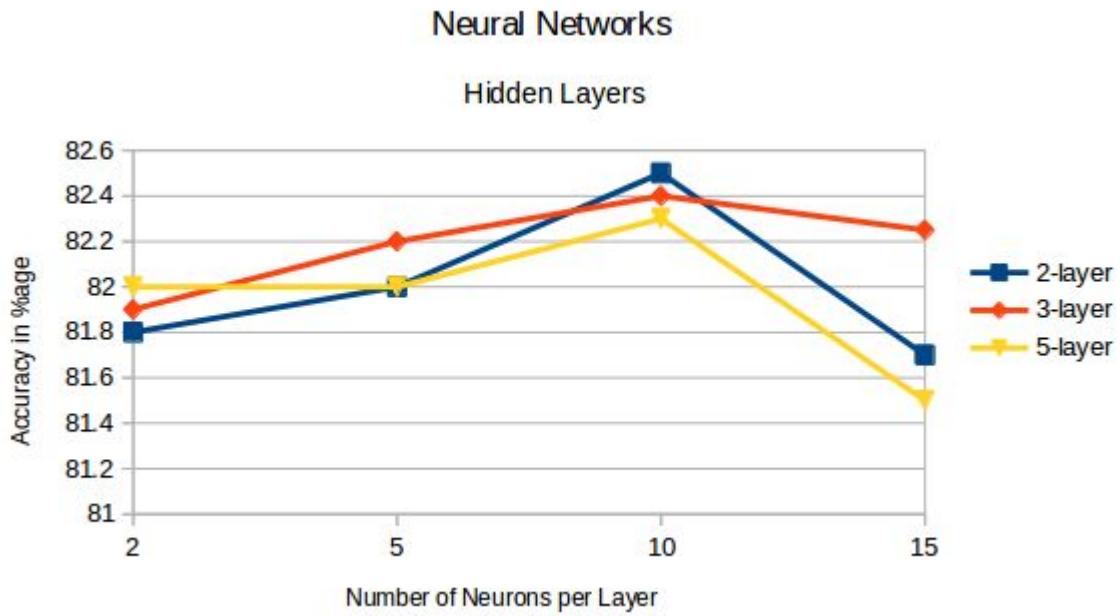
Specifications

- Size of Data Vector **1161 X 20** (1161 - Number of Movies and 20-Features including bias)
- Initially weights given (-0.25 to 0.25) which is determined by

$$-1/\sqrt{(Nh)} < d < +1/\sqrt{(Nh)}$$

- Optimal number of Nh is determined by calculating efficiency on variable number of Hidden Layer Neurons (Between Number of input neurons and output neurons)
- Best case : #Nh=2 & #Neurons in Hidden layer = 10

Finding Optimal Number of Hidden Layers



Process

- Weights Initialization.
- a feature vector for all movies in training data is feeded (feed forward) into the network and error for each is calculated by backpropagation.
- For eg: In 2- hidden layer perceptron
  Star_wars =

.05	1	0.7	0.05	0.08	0.05	0.05	0.05	0.55	0.03	0.36	0.18	0.40	0.52	0.10	0.66	0.40	0.03	0.01	0.05

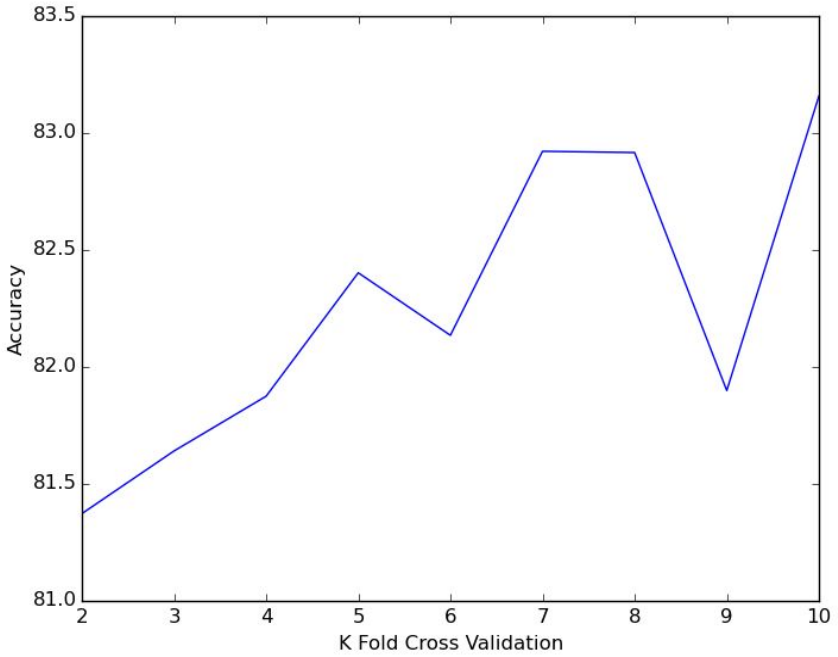
Process

- Now error from each feature vector is calculated and weights are adjusted till error (||expected - actual||) gets minimised.
- Ran for 25000 epochs.
- K-fold Cross validation is applied on the dataset t(k=2 to 10)

K-Fold Cross validation

(#Nh=2 & #Neurons in hidden layer = 10)

K- Fold	Accuracy (in %)
2	81.37
3	81.63
4	81.88
5	82.41
6	82.09
7	82.89
8	82.90
9	81.88
10	83.15



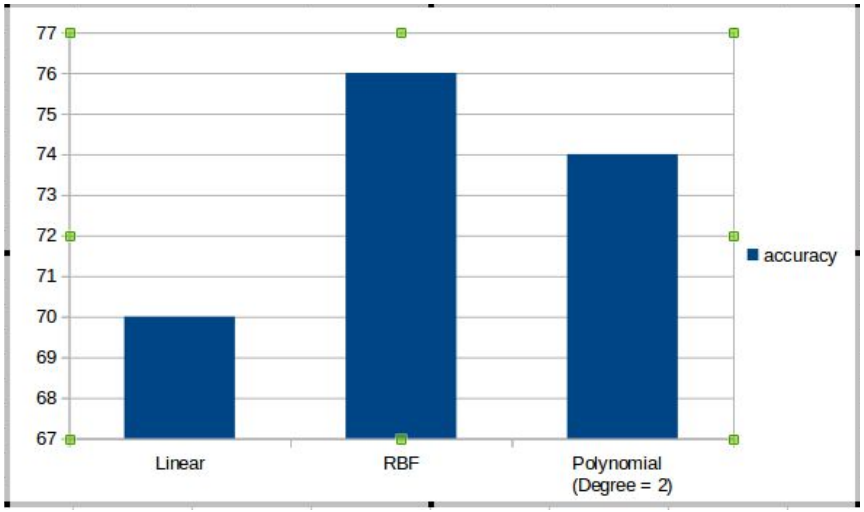
Support Vector Machine (SVM)

- Linear, Radial Basis Functions(rbf) and polynomial kernel funtions of degree 2 are used.
- For RBF Kernel , Manual Grid search is carried out.

Gamma -> C	1	0.1	0.01	0.001
1	50	42	41	40
10	50	57	47	41
100	70	68	72	***76***
1000	57	53	57	54

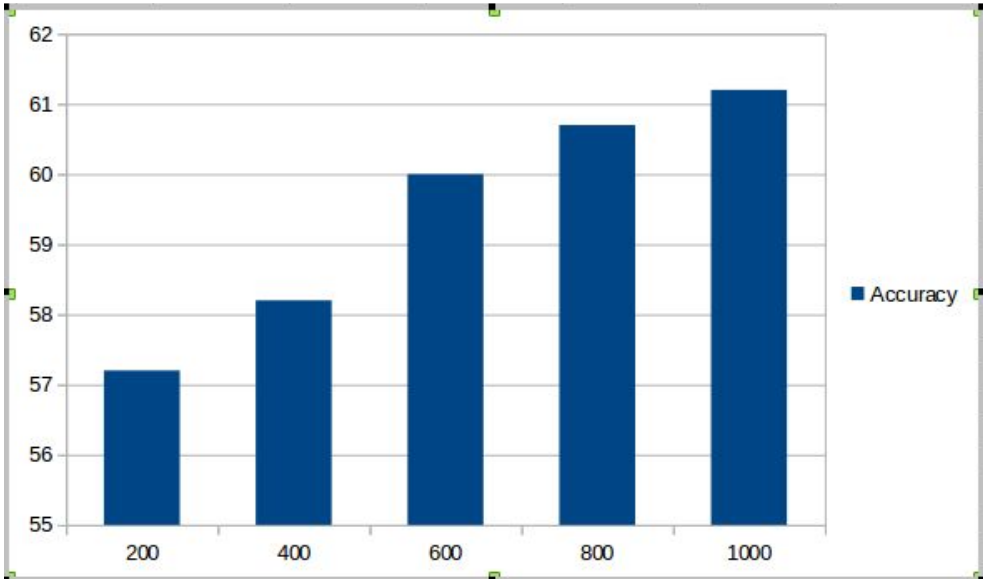
Various SVM Accuracies

Kernel Functions	Accuracy (in %)
Linear	70
RBF	76
Polynomial (degree = 2)	74

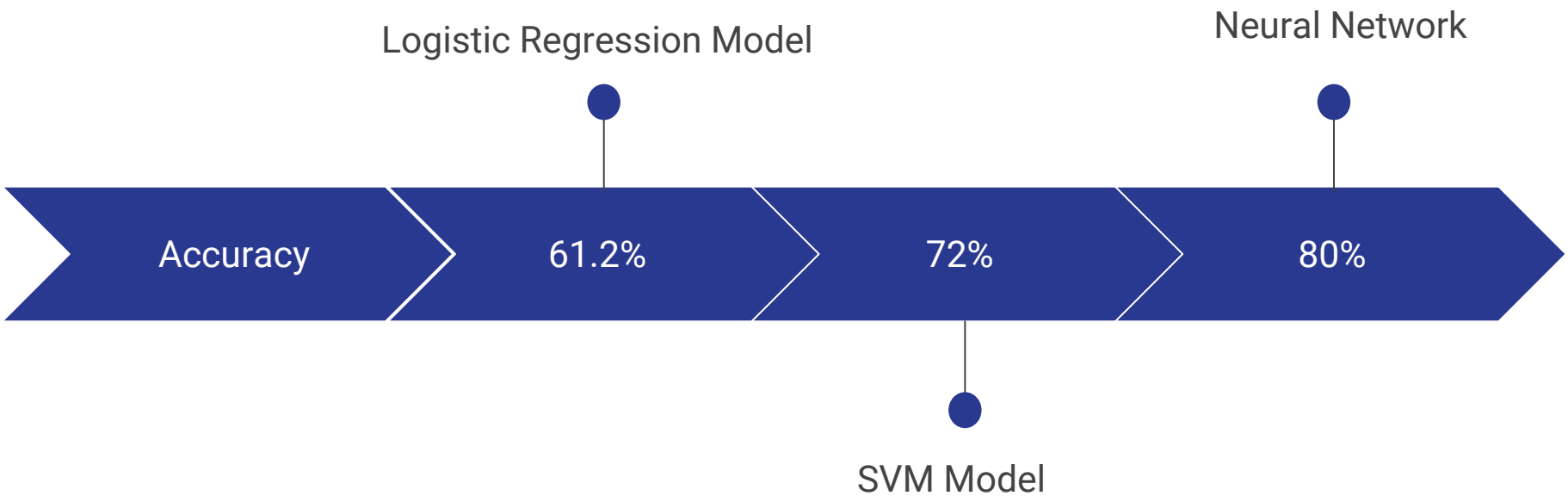


Logistic Regression

Accuracy = 61.2%



Techniques Used



Observations

- Logistic Regression Model showed an increase in accuracy as the number of training samples increased. When number of samples were large, the accuracy observed was 61.2%

- SVM was implemented using 3 different kernel functions, linear, polynomial and radial basis functions . Maximum accuracy was observed in RBF kernel. Average accuracy observed was 72%

Observations

- Neural Networks exhibited the best accuracy among the techniques used. It showed an accuracy of 82% after validating with K-fold Cross Validation.
- Comparing the three techniques using the exact same experimental conditions Neural Networks performed significantly better.

Logistic Regression	SVM	Neural Network
61.2%	72%	80%

Future Scope

- Including Google Search Popularity Index as a feature in our dataset would act as a user based reaction and thus is expected to perform better.

- Sentiment analysis of data from social media sites and trending hashtags could improve the performance.

- Newly developed hybrid techniques using genetic and other intelligent algorithms can also be used as the results.

Thank You