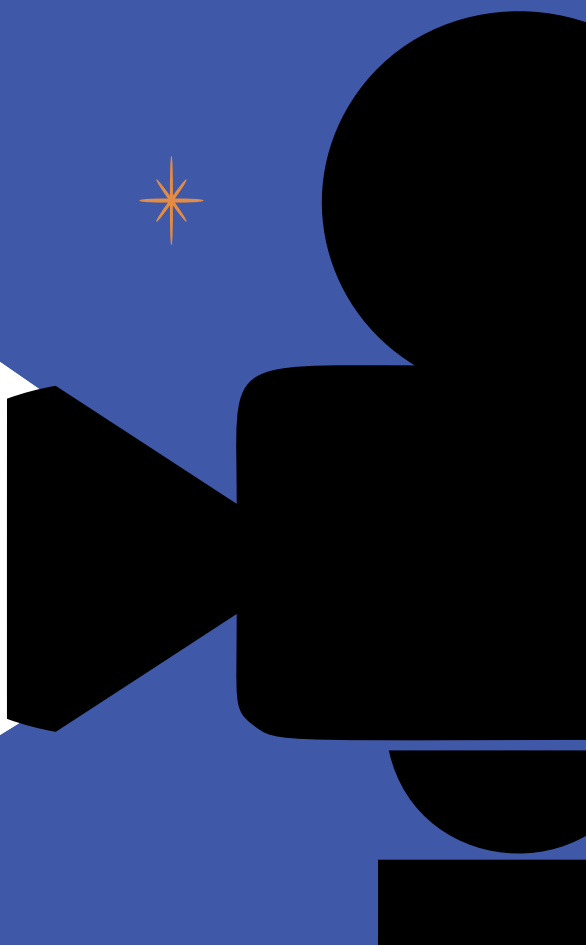
CCAI312

VOX CINEMA

PREDICT GROSS THE OF A MOVIES

Khadeja Y. Njaai 2005528
Ritaj M. Almutairi 2006532
Razan I. Alkhamisi 2006008

Prepared To:
Dr. Safa Alsafari



Content

Introduction

This report is made for our pattern recognition course project. In this report we will discuss a few subjects like the problem description of our project, the methods used, and results of each method. and conclusion.

1.Problem Description

Vox cinema finds it hard to know whether a movie will increase its profit or not, so will the tickets be sold out? so, the CEO suggested predicting if the film will succeed or not by relying on IMBD data.

IMBD is a website that takes all the individual votes cast by the registered users and uses them to calculate a single rating of a movie.

Vox cinema will use the IMDB data to know if the number of votes will increase the gross or not.

2.Data Understanding and Cleaning

In the stage of understanding the data, we divided the process into the:

- **Data Description**
- **Data Monitoring**
- **Data pre-processing (data preparation and cleaning)**
    1. drop the unnecessary and unwanted columns
    2. convert the data type
- **Data analysis and transformation:**
    1. data exploration
    2. pattern identification
    3. engineering new features

Data Description & Data After Cleaning:

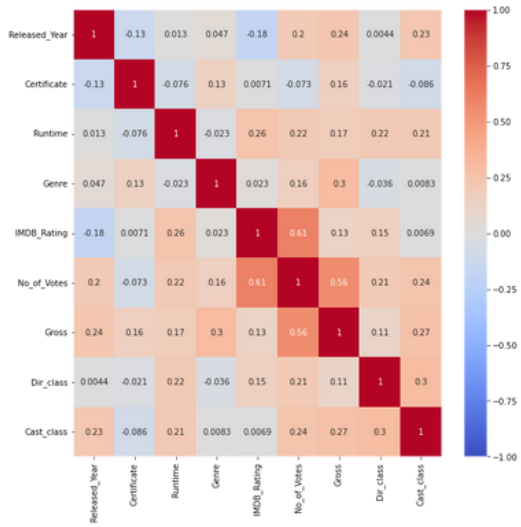
Below is a brief description of the used data set, and clarify changes to data after the cleaning phase

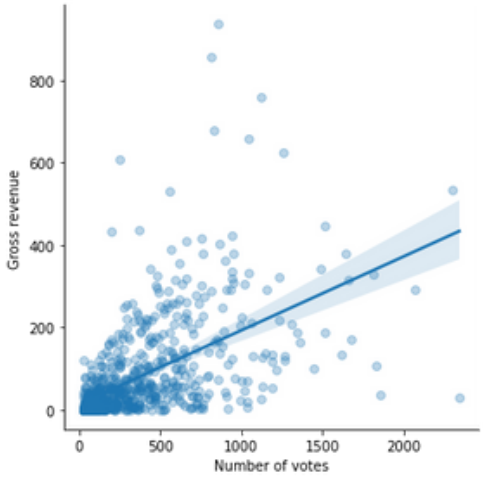
Original data information: 16 Columns X 1000 Rows

Data information after cleaning: 10 Columns X 750 Rows

Name of the data	Description of the data	Data type	Changes
Poster Link	Link of the poster that imdb using	object unique values	DROP
Series_Title	Name of the movie	object unique values	DROP
Released_Year	Year at which that movie released	object	convert to Integer
Certificate	Certificate earned by that movie	object	change to integer
Runtime	Total runtime of the movie	object	change to integer
Genre	Genre of the movie	object	change to integer
IMDB_Rating	Rating of the movie at IMDB site	float64	
Overview	Movie summary	object	DROP
Meta_score	Score earned by the movie	float64	DROP
Star1	Name of the Stars	object	Marge to cast_class
Star2	Name of the Stars	object	Marge to cast_class
Star3	Name of the Stars	object	Marge to cast_class
Star4	Name of the Stars	object	Marge to cast_class
No_of_Votes	Total number of votes	int64	
Gross	Money earned by that movie	object	change to Integer
Director	Name of the Director	object	change to Integer Dir_class

Find the Correlation between the Features:





this graph shows a high correlation between Gross and No_of_votes, and based on this correlation we decide to use Gross as Label and No_of_Voltes as a Feature.

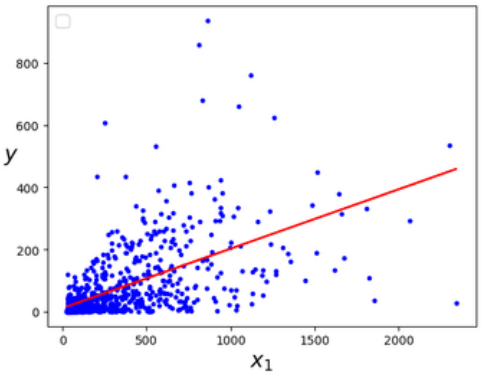
3. Data Modeling

first, we divided the data as 80% as training data, and 20% as testing data,Then we searched and tried several types of models that help us predict the Gross, and give us the best possible result.

3.1 Linear Regression

Predicting the gross of the movies, using linear regression.
We calculated the training error and the test error and got the result shown in the table below.
The gap between the train and test error indicates that this linear regression is underfitting.

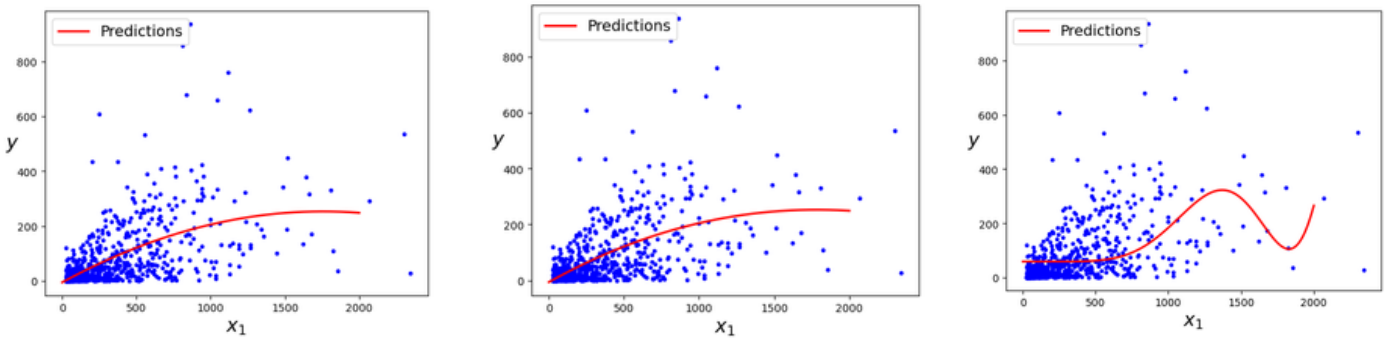
	Train Error	Test Error
Mean absolute error	57.85	53.45
Root mean square error	95.51	88.72
Mean sum of squares (MSE)	9121.30	7870.76



3.2 Polynomial

After finding an underfitting in the linear regression we decide to try polynomial to solve this issue, and experimented with several different degrees (Second, Third, tenth)

	Second Polynomial Degree	Third Degree Polynomial	10th Degree of Polynomial
Trian Error Root mean square error	92.31	92.31	102.17
Test Error Root mean square error	177.55	177.18	170.54



Based on the results, the third degree is the best degree we have reached in this experiment

3.3 KNN(K-Nearest Neighbors) :

We counted the errors in the training and testing and we found that

In the training
the Mean absolute error is equal 49.88 and the root mean square error equals 6842.55 and also we calculate the Mean sum of squares (MSE) and we found that it's equal 6842.55 and finally the training score equals 0.442

In the testing
Mean absolute error equals 64.35 and root mean square error equal 10959.06 and Mean sum of squares (MSE) equal 10959.06
and finally the testing score: 0.27

However, we found that the testing error is greater than the training error so the model going to be under fit model

3.4 Lasso Regression

Lasso regression is a type of linear regression that uses shrinkage. The Lasso is used in underfitting cases.

In order to try different models, we tried using Lasso Regression. The results were interesting.
In the training phase,
The mean absolute error is 55.34, while the root mean error is 91.14 and the Mean sum of squares (MSE) equals 8305.69.

On the other hand, in the testing phase, the Mean absolute error is 64.23, and the root mean square error equals 105.49 while the MSE is11127.84.

Conclusion

After seeing the results of the training errors and the testing errors of all the previous methods. We concluded that the Lasso Regression is the best fit compared to the polynomial features.
This assumption is made based on a comparison of the gap between the training model and the testing model.

	Polynomial	KNN	Lasso Regression
Train Error	92.31	6842.55	91.14
Test Error	177.18	10959.06	105.49

References

https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows

chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://aa.ssdi.di.fct.unl.pt/files/AA-03_notes.pdf