Movie Production Analysis & Prediction

Project
Presentation Slides

ID: 320210207
Name: Yousef Ibrahim Gomaa Mahmoud

Introduction

- **Problem Introduction:**
  - In this notebook, we explore various aspects of movie production and promotion, aiming to uncover patterns and insights that can contribute to the success of a film.
  - The task is to analyze this dataset and build a predictive model that can help stakeholders make informed decisions about movie production and marketing strategies.

Dataset

- **Data Description:**
  - **The goal is to perform classification and regression on "Collection." The dataset is found to have key parameters, such as:**

	Marketing expense	Production expense	Multiplex coverage	Budget	Movie_length	Lead_ Actor_Rating	Lead_Actress_rating	Director_rating	Producer_rating
0	20.1264	59.62	0.462	36524.125	138.7	7.825	8.095	7.910	7.995
1	20.5462	69.14	0.531	35668.655	152.4	7.505	7.650	7.440	7.470
2	20.5458	69.14	0.531	39912.675	134.6	7.485	7.570	7.495	7.515
3	20.6474	59.36	0.542	38873.890	119.3	6.895	7.035	6.920	7.020
4	21.3810	59.36	0.542	39701.585	127.7	6.920	7.070	6.815	7.070

Critic_rating	Trailer_views	3D_available	Time_taken	Twitter_hastags	Genre	Avg_age_actors	Num_multiplex	Collection
7.94	527367	YES	109.60	223.840	Thriller	23	494	48000
7.44	494055	NO	146.64	243.456	Drama	42	462	43200
7.44	547051	NO	147.88	2022.400	Comedy	38	458	69400
8.26	516279	YES	185.36	225.344	Drama	45	472	66800
8.26	531448	NO	176.48	225.792	Drama	55	395	72400

Dataset

Feature	Datatype
Marketing expense	float64
Production expense	float64
Multiplex coverage	float64
Budget	float64
Movie_length	float64
Lead_ Actor_Rating	float64
Lead_Actress_rating	float64
Director_rating	float64
Producer_rating	float64
Critic_rating	float64
Trailer_views	int64
3D_available	object
Time_taken	float64
Twitter_hastags	float64
Genre	object
Avg_age_actors	int64
Num_multiplex	int64
Collection	int64

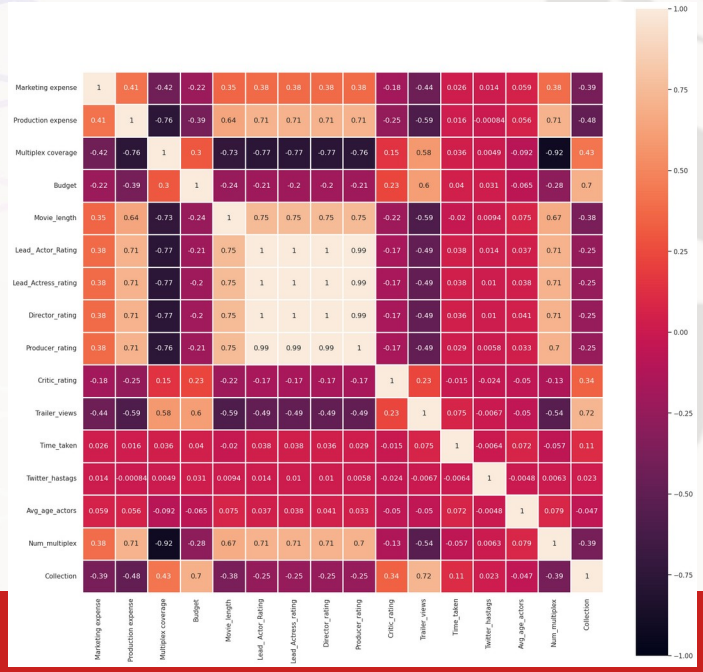
Table 1. *Dataset Datatypes*

Preprocessing

- **Data Preprocessing:**
  - The data is filtered for missing values and/or duplicates, approaches such as dropping or filling with mean/mode/median are viable options.
  - Conclusion
    - Data types are correctly casted.
    - Missing values found in the data given, which are in turn replaced with mean values since it is float64 type.
    - No duplicated records.

Preprocessing

- **Data Preprocessing:**
  - A correlation matrix is done to evaluate the dependencies and correlation between the features.



Preprocessing

- **Data Preprocessing:**
  - Afterwards, we use Z-Test and ANOVA (Analysis of Variance), which are statistical tests used in different scenarios, to make inferences about population parameters or to compare means across different groups.
  - Results:
    - Z-test: Failed to Reject the Null Hypothesis
    - ANOVA: Reject the Null Hypothesis

Machine Learning

- **Machine Learning Techniques:**
  - Some machine learning techniques are evaluated.
  - K-NN (All kinds)
  - LDA
  - Naive Bayes
  - Decision Tree
  - Neural Networks

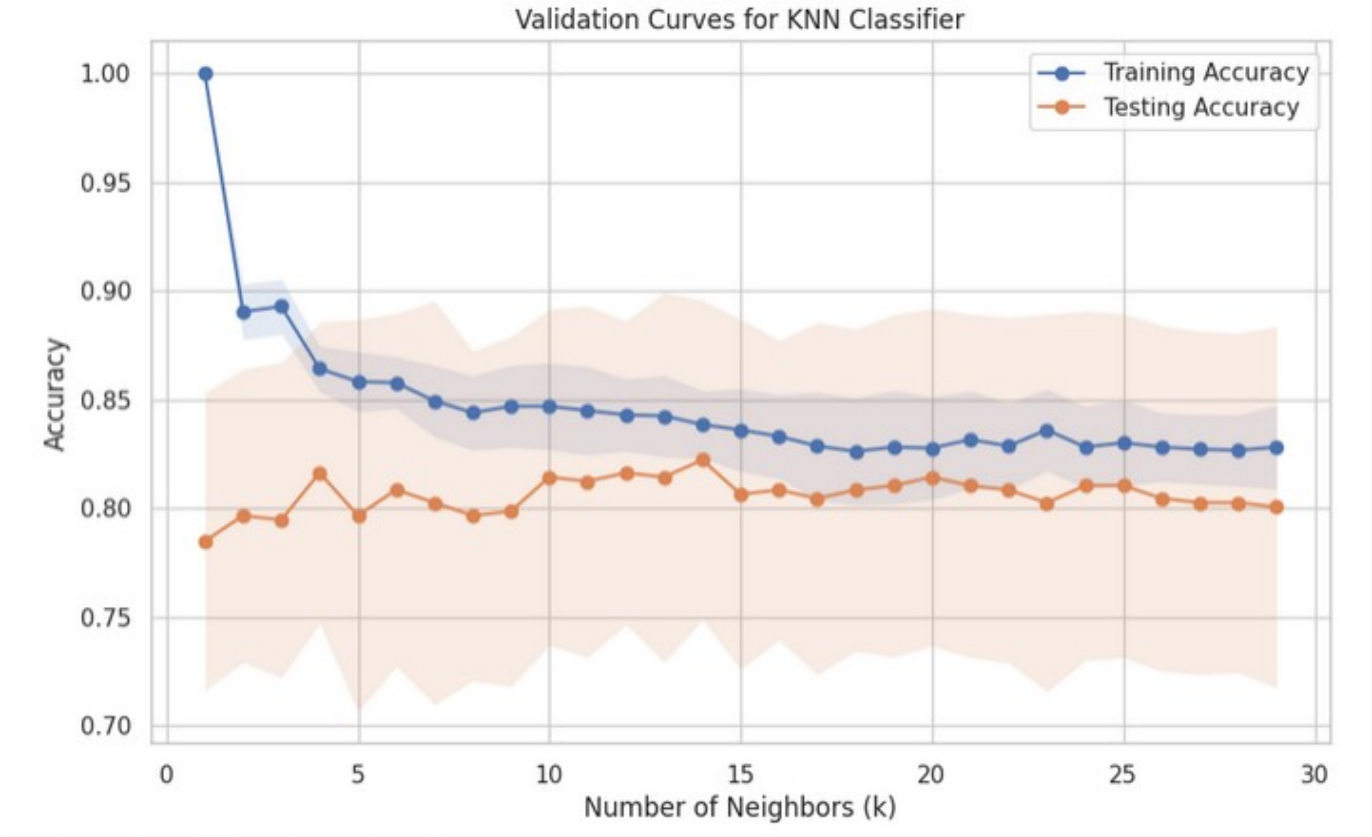
E-JUST

Machine Learning

- **K-NN:**
  - **Results: (Accuracies & Mean Absolute Error are shown here, other evaluation metrics in the documentation) (K=2 & K=7 respectively)**
    - **K-NN with PCA: 79.41176470588235%**
    - **K-NN with LDA: 85.29411764705883%**
    - **K-NN with SVD: 77.45098039215686%**
    - **K-NN Regressor (different distances)**
      - **Euclidean Distance: 6898.0392156862745**
      - **Manhattan Distance: 7085.994397759105**
      - **Cosine Distance: 7223.529411764704**

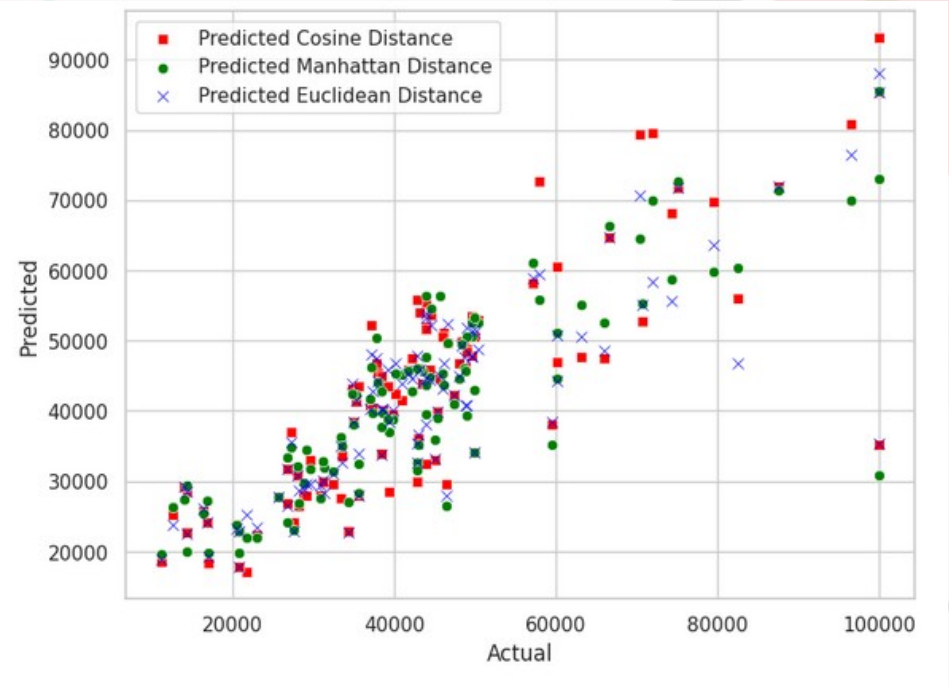
Machine Learning

- **K-NN:**
  - Different K values
  - Highest yield at 2
  - Stagnates at 0.55



Machine Learning

- **K-NN Regressor:**



Machine Learning

- **Naive Bayes:**
  - **Results:** 0.7549019607843137

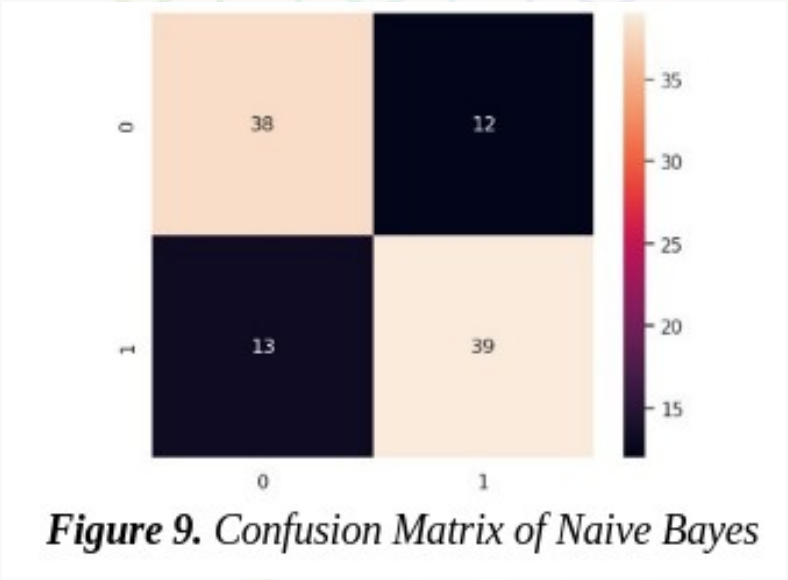


Figure 9. *Confusion Matrix of Naive Bayes*

- **Decision Tree:**
  - **Results:** 0.7745098039215687

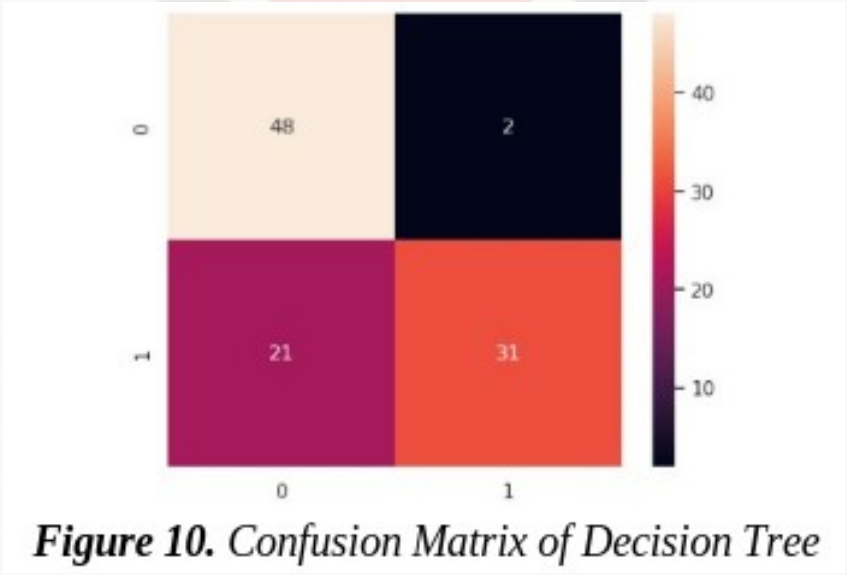


Figure 10. *Confusion Matrix of Decision Tree*

Machine Learning

- **Decision Tree:**
  - **Depth 7**

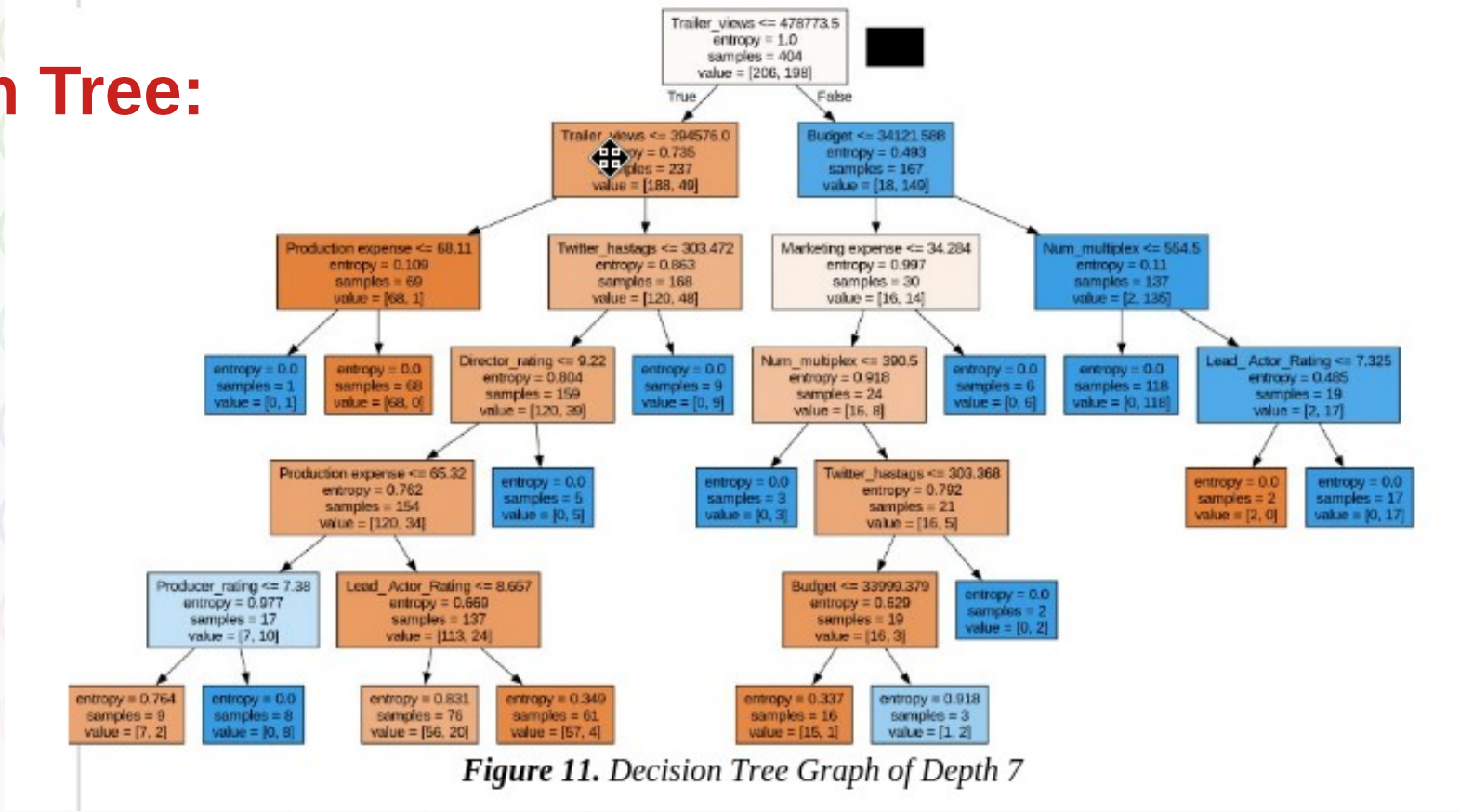


Figure 11. *Decision Tree Graph of Depth 7*

Machine Learning

- **Linear Discriminant Analysis:**

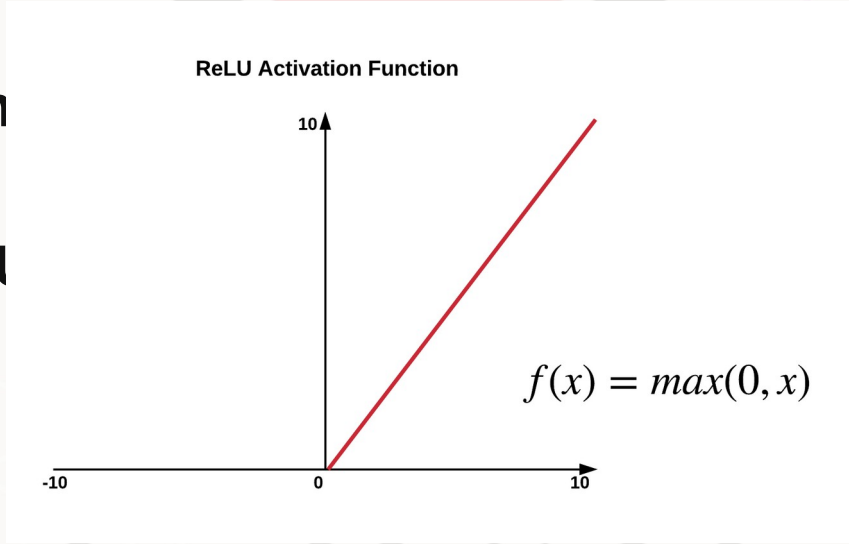
Linear Discriminant Analysis Classifier:

```
# Linear Discriminant Analysis (as a classifier)
lda = LinearDiscriminantAnalysis()
lda.fit(x_train, y_train)
lda_pred = lda.predict(x_test)
# Accuracy
accuracy = print('Accuracy Score: ', format(accuracy_score(y_test, lda_pred)))
# Precision
precision = print('Precision Score: ', format(precision_score(y_test, lda_pred, pos_label='Success')))
# Recall
recall = print('Sensitivity/Recall Score: ', format(recall_score(y_test, lda_pred, pos_label='Success')))
# F1-score
f1score = print('F1-Measure/F1-Score: ', format(f1_score(y_test, lda_pred, pos_label='Success')))
```

```
Accuracy Score:  0.8235294117647058
Precision Score:  0.8863636363636364
Sensitivity/Recall Score:  0.75
F1-Measure/F1-Score:  0.8125000000000001
```

Machine Learning

- **Neural Networks:**
  - 4 Dense Layers:
    - 1024 filters, “RELU” activation function.
    - 512 and then 256 filters, “RELU” activation function.
    - Output layer, “Sigmoid” activation function.
    - Optimizer: ADAM
    - Results: 0.8529411554336548



Conclusion

- **Conclusion:**
  - The study suggests that predicting gross revenue during production is not very accurate. The developed models are imperfect, as they do not consider various variables like plot, social media sentiment, stardom, and awards. The use of more advanced techniques, may improve revenue predictions in the future.
  - **Best model to be used:**

Neural Networks → 0.8529411554336548

Conclusion

Model	Accuracy
K-NN (PCA)	0.7941176470588235
K-NN (LDA)	0.8529411764705883
K-NN (SVD)	0.7745098039215686
K-NN	0.7745098039215687
LDA	0.8235294117647058
NAIVE BAYES	0.7549019607843137
DECISION TREE	0.7843137254901961
NEURAL NETWORKS	0.8529411554336548

Table 6. Table of All Models' Accuracies

Movie Production Analysis & Prediction

Thank you.