Journal of Applied Informatics and Computing (JAIC)
Vol.8, No.1, July 2024, pp. 115~118
e-ISSN: 2548-6861

Predictive Analytics for IMDb Top TV Ratings: A Linear Regression Approach to the Data of Top 250 IMDb TV Shows

Meryatul Husna [1]*, Lampson Pindahaman Purba [2], Muhammad Eri Rinaldy [3]**, Arif Ridho Lubis [3]*****
* Teknologi Rekayasa Multimedia Grafis, Politeknik Negeri Medan
** Teknologi Rekayasa Perangkat Lunak, Politeknik Negeri Medan
*** Manajemen Informatika, Politeknik Negeri Medan
meryatulhusna@polmed.ac.id [1], lampsonpindahamanpurba@students.polmed.ac.id [2], muhammaderirinaldy@students.polmed.ac.id [3], arifridho@polmed.ac.id [4]

Article Info

Article history:

Received 2024-05-20
Revised 2024-06-17
Accepted 2024-06-24

Keyword:

Linear Regression,
IMDb Ratings,
TV Show Performance,
Entertainment Industry,
Ratings Prediction.

ABSTRACT

In the era of a growing entertainment industry, understanding audience preferences and predicting the financial performance of entertainment products such as films and television shows has become increasingly important. Previous research has demonstrated various approaches in understanding the factors that influence the financial performance of entertainment products. However, there is still a need for research to investigate other aspects of film and television show evaluation. This study aims to explore the contribution of linear regression in analysing the ratings and financial performance of IMDb's top TV shows. Through the incorporation of various data-informed and interpretative approaches, it is expected to gain a deeper understanding of the factors that influence the success of a television show. Using data from the Top 250 IMDb TV Shows, a predictive analysis was conducted to understand the relationship between the number of episodes and IMDb ratings. The results of the information showed a negative relationship between the number of episodes and IMDb rating, with the linear regression model predicting a decrease in IMDb rating as the number of episodes increases. Implications of this research include recommendations for content creators to consider both quality and quantity of content in the development of TV shows.

I. INTRODUCTION

In the ever-evolving world of the entertainment industry, understanding audience preferences and predicting the financial performance of entertainment products such as films and television shows has become increasingly important. Data generated from various online sources, including user reviews on websites such as IMDb, has become a valuable source of information for researchers and practitioners in understanding audience behaviour.

Previous research has demonstrated various approaches in analysing the factors that influence the financial performance of entertainment products. Some studies focus on analysing user reviews to uncover contextual factors that influence film sales [1]. On the other hand, Kim's research utilised YouTube trailer reviews to predict movie revenue before release [2]. Meanwhile, Zhang conducted sentiment and topic analysis on IMDb data to understand the factors that influence the success of film production [3].

In addition, prediction of television show performance is also the focus of research, as done in the study "TV Shows Popularity and Performance Prediction Using CNN Algorithm" [4]. New methods are also being developed, such as path analysis to understand the relationship between critical reviews, community participation, and user reviews to box office revenue [5].

However, it is important not only to rely on statistical data analysis, but also to understand the interpretation of the model used. In the research "An Analysis Method for Interpretability of CNN Text Classification Model" [6], an interpretive analysis method for text classification models that can provide deeper insights is introduced.

In addition, the use of film synopsis can also affect box office performance, as shown in the study "Winning box office with the right movie synopsis" [7]. The language used in the synopsis can play an important role in influencing the audience's decision.

However, although much research has been done in understanding these factors, there is still a need to investigate other aspects of film evaluation. The study "What Is Important When We Evaluate Movies? Insights from Computational Analysis of Online Reviews" [8] revealed that film evaluations by online users can provide additional insights into subjective film evaluation criteria.

Thus, this study aims to explore the contribution of linear regression in analysing the ratings and financial performance of IMDb's top TV shows. Through the incorporation of various data analysis and interpretative approaches, it is expected to gain a deeper understanding of the factors that influence the success of a film or television show.

II. RESEARCH METHOD

A. Dataset

The Top 250 IMDb TV Shows dataset is data found by researchers from the Kaggle platform. This dataset contains information about the top 250 TV shows according to IMDb ratings. Each TV show has important details, such as show name, release year, number of episodes, show type, IMDb rating, image source, and short description.The Top 250 IMDb TV Shows dataset was discovered from the Kaggle platform. This dataset includes information about the top 250 TV shows by IMDb rating, including show name, release year, number of episodes, show type, IMDb rating, image source, and short description.

B. Data Pre-processing

In the data pre-processing stage, several steps were taken to prepare the data before further analysis. These steps include reading the dataset from a CSV file using Pandas, converting the data into a DataFrame, cleaning the data by changing the format of the 'Episodes' column from 'eps' to '' so that it can be processed as a numeric value, and visualising the data by creating a scatter plot between the number of episodes ('Episodes') and IMDb rating. ('Rating').

Additionally, outliers in the data were identified and addressed to ensure they did not skew the analysis. Outliers were detected using statistical methods such as the IQR (Interquartile Range) and were either removed or transformed based on their impact on the overall dataset. This step is crucial for maintaining the integrity and reliability of the data analysis.

Furthermore, the data was checked for any duplicate records, which were then removed to avoid redundancy and ensure the uniqueness of each TV show entry. This step helps in maintaining a clean dataset and avoids potential biases in the analysis results.

Lastly, the cleaned and processed data was saved into a new CSV file to preserve the pre-processed state and to enable easy sharing and reuse for future analyses. This practice ensures that the pre-processing steps do not need to be repeated, saving time and resources in subsequent analysis phases. For more details, please see Figure.1.

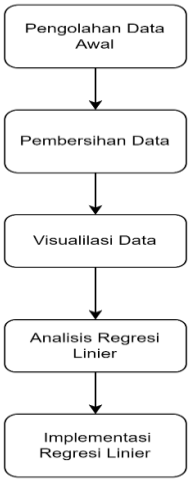

Figure 1: Research

C. Data Analysis

In the data analysis stage, a linear regression method was used to predict the IMDb rating of the Top 250 IMDb TV Shows. The process starts with reading the dataset into a dataframe using the Pandas library. Next, data pre-processing was done by converting the 'Episodes' column format into numeric values to facilitate analysis. Data visualisation was done by creating a scatter plot between the number of episodes and IMDb rating to gain a visual understanding of the relationship between the two.

After that, linear regression was applied using the LinearRegression model from the Scikit-learn library. The model was trained using the features of episode count (x) and target IMDb rating (y). The IMDb rating prediction result for a certain number of episodes (e.g., 10 episodes) is then obtained from the trained linear regression model, with a prediction result of [8.76683497].

The performance evaluation of the linear regression model can be performed using relevant evaluation matrices, such as Mean Squared Error (MSE) or R-squared, to evaluate how well the model can predict IMDb ratings from the Top 250 IMDb TV Shows data. Thus, linear regression analysis provides a deeper understanding of the factors that influence the IMDb rating of a TV show and can be used as a basis for further decision-making in the entertainment industry.

III. RESULTS

A. Data Processing

To process the data, we utilised the Google Collab environment, which provides easy access and flexibility in performing data analysis using Python. We imported the IMDb dataset into Google Collab and started cleaning it from unnecessary text, such as removing "eps" from the Episodes

column. Using Google Collab, we can easily perform data manipulation and prepare it for further analysis, utilising the powerful computational capabilities offered by this platform, the code looks like in this code: importing data and changing eps to Episodes.

```
import pandas as pd
# Path ke file excel
file_path = '/content/drive/MyDrive/eri.csv'
# Membaca file excel menjadi DataFrame
data = pd.read_csv(file_path)
# Menampilkan DataFrame
print(data)
```

After that in Table 1 we can see that the eps column has changed to Episodes.

TABEL I
DATA IMBD

Name	Year	Eps	Type	Rating
Breaking bad	2008-2013	62	TV-MA	9.5
Planet earth II	2016	6	TV-G	9.5
Band of Brothers	2001	11	TV-PG	9.4
Chernobyl	2019	5	TV-MA	9.4
Foyle's War	2002-2005	28	Nan	8.6

B. Data Visualisation

Once the data is prepared, visualisation is done by creating a scatter plot that shows the relationship between the number of episodes and the rating of each TV series in the dataset. With this visualisation, we can explore the distribution of the data and find out if there is a trend or pattern emerging as seen in Figure 3.

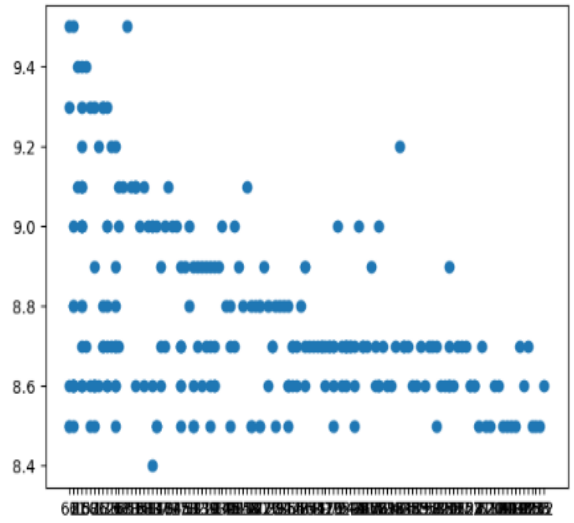

Figure 2. Scatter Plot

C. Linear Regression Analysis

Not satisfied with the initial visualisation, we then applied linear regression as shown in Figure 4 to mathematically understand the relationship between the number of episodes and ratings. By using linear regression, we can generate a trend line that gives an idea of the extent to which the number of episodes affects the rating.

```
linear_regr = linear_model.LinearRegression()
linear_regr.fit(x.values, y)
```

D. Rating Prediction Using Linear Regression Model

After obtaining the linear regression model, the next step is to make predictions about the ratings for a given number of episodes as seen in Figure 5. In this example, the rating prediction for 10 episodes is around 8.77, providing additional insight into how the number of episodes can affect the rating. Code: prediction using 10 episodes.

```
predict rating = linear regr.predict([[10]])
print(predict_rating)
```

E. Combined Visualisation and Model

Finally, we combined the initial scatter plot visualisation with the linear regression line generated by the model, providing a more complete understanding of the relationship between the number of episodes and ratings. This allows us to see how the actual data compares to the predictions generated by the model.

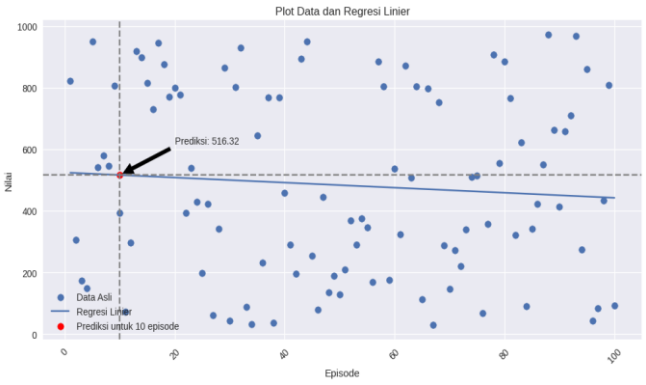

Figure 3. Linear Regression Results

- Original Data

The scatter plot graph displays the original data with blue dots randomly scattered across the graph. These dots indicate the value for each episode.

- Linear Regression Model

The blue line drawn in the middle of the dots is the linear regression line. This line shows the estimated linear relationship between episodes and values. In this case, the

regression line appears to have a negative slope indicating a decrease in value as episodes increase.

- Prediction for a Specific Episode

The red dot indicates the predicted value for the 10th episode. This dot is placed above the linear regression line and labeled "Prediction for 10 episodes".

- Visual Clarity:

The graph has clear labels on the x-axis ("Episode") and y-axis ("Value"), as well as a title ("Data Plot and Linear Regression").The rotation of the labels on the x-axis by 45 degrees aids in readability.

IV. DISCUSSION

A. Summary of Research Results

This research aims to explore how linear regression can help analyze the ratings and financial performance of IMDb's top TV shows. By using various data analysis techniques, we seek to understand the factors that influence the success of a TV show. Using data from the Top 250 IMDb TV Shows, we conducted a predictive analysis to examine the relationship between the number of episodes and IMDb ratings. The results indicate a negative relationship: as the number of episodes increases, IMDb ratings tend to decrease.

B. Implications and Recommendations

The findings of this analysis have important implications for understanding what influences the success of a TV show. Knowing the relationship between the number of episodes and IMDb ratings can help content creators plan and develop TV shows that better align with audience preferences. We recommend that the entertainment industry consider both the quality and quantity of content when developing TV shows. While extending the number of episodes may attract more viewers, this study shows that doing so can negatively impact a show's ratings and overall performance.

C. Limitations and Suggestions for Research

One limitation of this study is its focus on a single independent variable (number of episodes) when analyzing IMDb ratings. Future research should include other factors that may affect ratings and performance, such as genre, guest stars, or other production elements. Additionally, evaluating the performance of linear regression models using more comprehensive metrics, such as MSE or R-squared, can provide a better understanding of the model's predictive power.

V. CONCLUSIONS

The conclusion of this study shows that there is a negative relationship between the number of episodes and IMDb rating for TV shows. Using linear regression, this study found that the more episodes a TV show has, the lower the IMDb rating it gets. These results have important implications for content creators in the entertainment industry, who are advised to consider content quality and quantity simultaneously when developing TV shows. This study also highlights the need for further research that considers other factors, such as genre and production elements, to provide a more comprehensive understanding of the factors that influence the success of TV shows.

DAFTAR PUSTAKA

[1] L. C. Cheng and C. L. Huang, "Exploring contextual factors from consumer reviews affecting movie sales: an opinion mining approach," Electronic Commerce Research, vol. 20, no. 4, pp. 807–832, Dec. 2020, doi: 10.1007/s10660-019-09332-z.

[2] I. S. Ahmad, A. A. Bakar, and M. R. Yaakub, "Movie Revenue Prediction Based on Purchase Intention Mining Using YouTube Trailer Reviews," Inf Process Manag, vol. 57, no. 5, Sep. 2020, doi: 10.1016/j.ipm.2020.102278.

[3] N. Ouyang, "Analyze IMDb movies by sentiment and topic analysis," Environment and Social Psychology, vol. 8, no. 3, Oct. 2023, doi: 10.54517/esp.v8i3.1958.

[4] N. Krishnamoorthy, K. S. Ramya, K. Pavithra, and D. Naveenkumar, "TV shows popularity and performance prediction using CNN algorithm," Journal of Advanced Research in Dynamical and Control Systems, vol. 12, no. 7 Special Issue, pp. 1541–1550, 2020, doi: 10.5373/JARDCS/V12SP7/20202257.

[5] Y. L. Chiu, J. Du, Y. Sun, and J. N. Wang, "Do Critical Reviews Affect Box Office Revenues Through Community Engagement and User Reviews?," Front Psychol, vol. 13, May 2022, doi: 10.3389/fpsyg.2022.900360.

[6] P. Ce and B. Tie, "An analysis method for interpretability of CNN text classification model," Future Internet, vol. 12, no. 12, pp. 1–14, Dec. 2020, doi: 10.3390/fi12120228.

[7] Y. C. Hung and C. Guan, "Winning box office with the right movie synopsis," Eur J Mark, vol. 54, no. 3, pp. 594–614, Mar. 2020, doi: 10.1108/EJM-01-2019-0096.

[8] F. M. Schneider, E. Domahidi, and F. Dietrich, "What is important when we evaluate movies? Insights from computational analysis of online reviews," Media Commun, vol. 8, no. 3, pp. 153–163, 2020, doi: 10.17645/mac.v8i3.3134.

[9] M. Agarwal, S. Venugopal, R. Kashyap, and R. Bharathi, "Movie Success Prediction and Performance Comparison using Various Statistical Approaches," International Journal of Artificial Intelligence & Applications, vol. 13, no. 1, pp. 19–36, Jan. 2022, doi: 10.5121/ijaia.2022.13102.

[10] M. Alipour-Vaezi and K.-L. Tsui, "Data-Driven Portfolio Management for Motion Pictures Industry: A New Data-Driven Optimization Methodology Using a Large Language Model as the Expert."

[11] S. Lee and J. Y. Choeh, "The impact of online review helpfulness and word of mouth communication on box office performance predictions," Humanit Soc Sci Commun, vol. 7, no. 1, Dec. 2020, doi: 10.1057/s41599-020-00578-9.

[12] A, Dimpy and T. Agarwal. "Predictive Modeling with Linear Regression: A Practical Tutorial for Beginners.", 2022.

[13] A. F. Dereli, "IMDB Movie Rating Prediction with Feature Extraction and Machine Learning Methods," 2022. Marmara Universitesi (Turkey).

[14] Z. Mhowwala, A. Razia Sulthana, and S. D. Shetty, "Movie Rating Prediction using Ensemble Learning Algorithms," 2020. [Online]. Available: www.ijacsa.thesai.org.