Overall information about data:

Total no of rows :4824

Total no of Columns: 28

Highest imdb_score: 9.2

Lowest imdb_score:1.6

- Analyzing the data using histogram and heatmap, with the correlated features defined
- X='num_critic_for_reviews','duration','num_voted_users','num_user_for_reviews', 'movie_facebook_likes','director_facebook_likes'
- y='imdb_score' (prediction)
- training data = 80%
- testing data = 20%
- #Linear regression model is the best compared to Decision Tree, SVM, and KNN, i.e.,
- #Mean squared error using linear regression 0.6764002653438028
- #Mean absolute error using linear regression 0.6577251825742132

Introduction:

The goal of this analysis was to predict movie ratings (IMDb scores) using various features(28) related to movies and their directors. The dataset used for this analysis contains information about several movies, including attributes such as the number of critic reviews, movie duration, number of votes from users, number of user reviews, movie-related Facebook likes, and director-related Facebook likes. The analysis aimed to explore the relationships between these features and movie ratings, and to develop a predictive model using linear regression.

Data Analysis:

The dataset was initially subjected to exploratory data analysis to understand the distribution of variables and their relationships. Two primary visualizations were used: Boxplots, histograms and heatmaps.

1. Histogram Analysis:

Histograms were used to visualize the distribution of each numerical feature. This helped to identify any potential outliers or trends in the data.

2. Heatmap Analysis:

A heatmap was used to visualize the correlations between features. Correlation values were calculated and displayed using a color gradient. This allowed us to identify which features had a stronger linear relationship with the target variable, IMDb scores.

Selected Features:

The following features were selected for use in the prediction model:

- `num_critic_for_reviews`: The number of critic reviews for the movie.

- `duration`: The duration of the movie in minutes.

- `num_voted_users`: The number of users who voted for the movie.

- `num_user_for_reviews`: The number of user reviews for the movie.

- `movie_facebook_likes`: The number of Facebook likes for the movie's page.

- `director_facebook_likes`: The number of Facebook likes for the movie's director.

Model Selection:

After preparing the data, a linear regression model was chosen as the predictive model. Several other models were considered, including Decision Tree, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). The decision to choose linear regression was based on its comparative performance metrics.

Mean Squared Error (MSE): 0.6764

Mean Absolute Error (MAE): 0.6577

These metrics indicate the quality of the predictive model. Lower values of MSE and MAE suggest better predictive performance.

Conclusion:

In this analysis, explored the relationships between various movie-related features and IMDb scores. The selected features were used to train a linear regression model for predicting movie ratings. Among the tested models (Decision Tree, SVM, KNN, and linear regression), the linear regression model demonstrated the best performance. The model's predictive accuracy was evaluated using MSE and MAE, resulting in values of 0.6764 and 0.6577, respectively. This indicates a reasonable predictive capability for the selected features