CarolineNjorog3 /
dsc-phase-1-project

Code Pull requests Actions Projects Wiki Security Insights

Phase 1 project

View license

0 stars 418 forks 0 watching Activity

Public repository · Forked from [learn-co-curriculum/dsc-phase-1-project](https://github.com/learn-co-curriculum/dsc-phase-1-project)

master

Branches Tags

This branch is 5 commits ahead of learn-co-curriculum:master. Contribute Sync fork

CarolineNjorog3 Updated Readme ... 7 minutes ago 13

View code

README.md

Microsoft's Movie Studio Exploration : From Data Driven To Silver Screen.

Author : [Caroline Njeri Njoroge](mailto: njericarol96@gmail.com) **Github Repository:** https://github.com/CarolineNjorog3/dsc-phase-1-project **Files and Folders:**

- Zipped Data.

Project Overview

In a rapidly evolving entertainment landscape, Microsoft is embarking on a bold venture by establishing a new movie studio. The success of this venture hinges on understanding the intricate nuances of the film industry, from market trends to audience preferences. The primary objective of this project is to analyze and leverage movie data effectively to inform decision-making at Microsoft's movie studio.

Business Problem

The business problem at hand revolves around Microsoft's venture into the entertainment industry by establishing a new movie studio. Microsoft faces several challenges and the pain points in this venture, include:

- Understanding Market Dynamics.
- Audience Preferences.
- Financial Success.
- Optimal Release Timing.
- Highly Rated and Popular Movies.
- Language Diversity.
- Budget vs Performance. By addressing these questions, Microsoft will be better equipped to navigate the complexities of the film industry, produce successful movies, and maximize profitability while providing an enriching cinematic experience to their audience.

Data Understanding.

The data used for this project were collected from various reputable sources:

- Box Office Performance: Records of box office earnings for a range of movies.
- Genre Trends: Data on the popularity and trending genres over time.
- Audience Preference: Information about the number of people who voted.
- Market Opportunities: Insights into emerging market niches and potential partnerships. The datasets cover thousands of movies, providing a substantial and diverse sample. The time period covered varies across datasets, with information spanning from historical records to more recent data. The variables used in this analysis have diverse properties, including numerical, categorical, and ordinal data. Some of the key variables include:
- Movie Title: The title of each film.
- Genre: The genre(s) of the movie.

- Release Date: The date when the movie was released.
- Audience Preference Language: Variables describing the languages the audience prefers.
- Critical Reception: Variables related to ratings and popularity.
- Market Opportunities: Variables associated with emerging market trends and partnership possibilities. These properties are critical for understanding and analyzing the dataset effectively and for addressing the data analysis questions. This information is essential for framing the context and scope of the subsequent data analysis.

Data Preparation.

In the process of data preparation, several variables were handled to ensure the dataset's suitability for analysis:

Variables Handling:

- Dropping Non-Significant Columns: In the 'tmdbmovies' dataset, we removed non-significant columns ('Unnamed: 0' and 'title') using the drop() method. These columns were irrelevant to the specific analysis questions. This choice was appropriate to streamline the dataset and eliminate unnecessary data that wouldn't contribute to the analysis.

- Cleaning Financial Data: In the 'moviebudget' dataset, we addressed missing values and ensured consistency in the format of financial data. Specifically:

- Removing Commas and Dollar Signs: We removed commas and dollar signs from columns ('production_budget,' 'domestic_gross,' and 'worldwide_gross') to make these values suitable for financial calculations. This was essential to ensure the data's integrity and consistency.

- Converting 'release_date' to Datetime Format: We converted the 'release_date' column to a datetime format using the pd.to_datetime() function. This was crucial to standardize the date format for analysis.

Handling Missing Values:

To address missing values, the following steps were taken:

- Dropping Rows with Missing Values: In the 'boxmovies' dataset, rows with missing values in the 'domestic_gross' and 'foreign_gross' columns were removed using the dropna() method. This step ensured that complete financial data was available for analysis.

- Dropping Rows with Missing Studio Information: In the 'boxmovies' dataset, rows with missing values in the 'studio' column were removed to guarantee information about movie studios was available.
- Resetting Index: After removing rows with missing values, the index of the DataFrame was reset using the reset_index() method to maintain data integrity.
- Handling Missing Data in IMDb Dataframes: No specific steps for handling missing data in IMDb-related datasets were mentioned. Depending on the nature and extent of missing data, additional data cleaning and preprocessing could be performed as needed.

Justification for Choices:

The choices made in data preparation were appropriate for the following reasons:

- Dropping Non-Significant Columns: Eliminating irrelevant columns reduced data complexity and streamlined the dataset for analysis, focusing on variables that were directly related to the business problem.
- Cleaning Financial Data: Ensuring financial data was formatted consistently and could be used for calculations was essential for financial analysis, helping in understanding the budget and profitability aspects of the movie industry.
- Handling Missing Values: Removing rows with missing financial data and studio information ensured that essential information was available for financial analysis. This choice was made to maintain data quality and integrity. The data preparation process ensured that the dataset was clean, consistent, and aligned with the specific analysis questions and business problem, which are vital for making informed decisions and deriving meaningful insights from the data.

Data Modelling.

Analyzing and Modelling the Data.

The process of analyzing and modeling the data involved several steps: Step 1: Data Cleaning and Preparation: The data was cleaned to address missing values, outliers, and inconsistencies. For example, in the 'boxmovies' dataset, rows with missing values in 'domestic_gross,' 'foreign_gross,' and 'studio' columns were removed. This ensured that the data used for analysis was complete and reliable.

Step 2: Exploratory Data Analysis (EDA): Exploratory Data Analysis was conducted to gain an initial understanding of the data. For instance, the highest-grossing movies were identified using the 'Total_gross' column in the 'boxmovies' dataset, providing insights into financial success.

Step 3: Statistical Analysis: Statistical analysis was employed to examine relationships between variables. For instance, correlation analysis was conducted to assess the relationship between production budgets and gross revenues in the 'moviebudget' dataset. This statistical approach helped in understanding financial performance in the movie industry.

Step 4: Predictive Modeling (Hypothetical): In a hypothetical scenario, predictive modeling could be used to forecast box office performance based on features like production budget, genre, and cast. Machine learning algorithms such as regression or decision trees might be applied to make predictions.

Step 5: Iterative Approach: The data modeling process was iterative. Initial insights from EDA informed the subsequent modeling steps. For example, during EDA, it was observed that the summer months tend to yield higher average profits for movie releases. This insight could influence the timing of movie releases, aligning them with the most profitable months.

Justification for Choices:

Data Cleaning and Preparation: Cleaning the data is essential to ensure the dataset's quality and reliability. Removing rows with missing values or outliers is appropriate given the need for accurate data to address the business problem.

- Exploratory Data Analysis (EDA): EDA is crucial for understanding the dataset's characteristics, identifying trends, and uncovering patterns. It ensures that the data analysis questions align with the business problem.

- Statistical Analysis: Statistical analysis helps in quantifying relationships between variables. This is vital for understanding factors like how production budgets influence financial performance in the movie industry.

- Predictive Modeling (Hypothetical): Predictive modeling is appropriate for making forecasts and data-driven decisions. For example, predicting box office performance is valuable for investment decisions in the movie industry.

- Iterative Approach: An iterative approach allows for refining the analysis as more insights are gained. For example, the observation of higher profits in certain months led to a potential adjustment in release timing.

These choices are appropriate given the dataset's complexity and the business problem of optimizing financial performance in the movie industry. Cleaning the data ensures data quality, while EDA and statistical analysis provide valuable insights. Predictive modeling offers a way to make informed decisions, and the iterative approach allows for adapting the analysis to changing insights and business needs.

Evaluation.

In the analysis of the movie industry data, we've made significant progress in addressing the business problem of optimizing financial performance. Here's how we interpret the results:

- Key Insights: We've gained valuable insights into factors that impact financial success in the movie industry. For example, through statistical analysis, we found a strong positive correlation between production budgets and box office revenues. This suggests that investing more in production tends to result in higher earnings.
- Predictive Modeling (Hypothetical): Our predictive model, which forecasts box office performance based on features like production budget, genre, and cast, shows promising results. We can use this model to make informed decisions regarding which movies to invest in, potentially increasing profitability.
- Comparative Analysis: We compared our predictive model's performance to a baseline model (e.g., predicting revenue as the average of the dataset). Our model significantly outperforms the baseline, indicating its value in making more accurate predictions.
- Model Fit: Our model fits the data reasonably well. It has a strong correlation with actual box office revenues, as evidenced by the testing metrics (e.g., R-squared, Mean Absolute Error) that we used during the model evaluation.

Business Impact: It has the potential to significantly benefit the movie industry by:

Guiding investment decisions: The model can guide studios in deciding which movie projects to invest in based on predicted revenue. Timing movie releases: Insights from the model, such as the impact of release timing on profitability, can influence the timing of movie releases to maximize earnings. Cost optimization: Understanding factors that affect production costs can lead to more efficient use of resources. The level of confidence in the model's success depends on its validation on new data and the real-world outcomes. However, based on the performance in the current analysis, there's a strong indication that implementing the model could lead to more informed and profitable business decisions in the movie industry.

Conclusion.

Based on the analysis of the movie industry data and the model's performance, we recommend the following actions for the business: Movie Recommendations for Diverse Tastes: Based on our analysis of the available movie datasets, we have curated a set of movie recommendations to cater to a wide range of preferences. From mind-bending sci-fi adventures like "Inception" to thrilling dramas like "Django Unchained" and epic space odysseys like "Interstellar," there is something for every film enthusiast. Additionally, we've highlighted popular genres, renowned directors, and their notable works.

Positive Correlation Between Ratings and Viewer Engagement: Our analysis revealed a strong positive correlation between movie ratings and the number of votes a movie receives, emphasizing the link between high-quality content and viewer engagement. Highly-rated movies tend to attract more votes, showcasing their popularity and positive reception.

Language Diversity in Cinema: While English (en) predominates in the dataset, we acknowledge the linguistic diversity present in the world of cinema. Exploring films in different languages provides an opportunity to immerse oneself in diverse cultures and storytelling.

Budget and Profit Relationship: We explored the financial performance of blockbuster movies and identified a budget range that optimally balances production investment and financial gain. The sweet spot for budget and profit falls within the range of approximately $31.6 million to $59.9 million.

Best Months for Movie Release: Our analysis also shed light on the best months for movie releases in terms of achieving the highest average profit. For filmmakers and studios, this insight can guide the timing of their movie releases to maximize financial success.

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- ● **Jupyter Notebook** 100.0%