Pre-report for Movie Classification Dataset

Strategic Decision Making with PowerBi

—

Prepared By:

Sameer Kumar Samal
2023JULB01137

Table of Contents

Problem Statement

To analyze the factors influencing the box office collection of movies and develop a predictive model for future movie success.

Key Questions:

1. What are the most significant factors affecting movie collections (revenue)?
2. How do marketing and production expenses correlate with revenue generation?
3. Does 3D availability or the presence of high-rated lead actors impact the movie's performance?
4. How can this data be used to optimize resource allocation for maximum profitability in movie production and promotion?

Data Requirements

- **Key Metrics:** Marketing Expense, Production Expense, Budget, Revenue, Ratings (Lead Actor, Actress, Director, Producer, Critic).
- **Qualitative Attributes:** Genre, 3D Availability, Technical Oscar Status.
- **Missing Values**: Handle nulls, e.g., in Time taken.
- **Outliers**: Identify and address in numerical columns.
- **Consistency**: Standardize categorical variables (e.g., 3D_available as "YES”/NO") and validate realistic numerical values.
- **ROI**: (Collection−Budget)/Budget (Collection - Budget) / Budget (Collection−Budget)/Budget.
- **Marketing Efficiency**: Collection/Marketing Expense Collection / Marketing
- Expense Collection/Marketing Expense.
- **Production Efficiency**: Collection/Production Expense Collection / Production Expense Collection/Production Expense.
- **Overall Rating**: Weighted average of actor, actress, director, producer, and critic ratings.
- **Engagement Metrics**: Correlation of trailer views, Twitter hashtags, and multiplex coverage with collections.
- Genre-wise collections.
- Impact of 3D availability and technical Oscars on revenue.
- Budget category trends (low, medium, high).
- **Independent Variables**: Marketing Expense, Production Expense, Multiplex Coverage, Ratings, Twitter Hashtags.
- **Dependent Variable**: Collection (Box Office Revenue).

Data Collection

- **Expenses & Budget:** Internal reports, trade publications, entertainment databases (e.g., IMDb, Box Office Mojo).
- **Revenue:** Box office tracking platforms (e.g., Box Office Mojo, The Numbers).
- **Engagement:** YouTube analytics, Twitter API, or social media aggregators.
- **Ratings:** Professional platforms (e.g., IMDb, Rotten Tomatoes, Metacritic).
- **Genre:** Official movie classifications.
- **3D Availability:** Production announcements or theater specifications.

- **Technical Oscars:** Award databases (e.g., Academy Awards lists).
- **Multiplex Coverage:** Distribution data from theater chains or distributors.
- **Time Taken:** Press releases, interviews, or tracking tools.
- **Actor Age:** Calculated using birthdates from IMDb or Wikipedia.

Data Validation

- **Column Names:** Ensure clarity and consistency (e.g., no extra spaces).
- **Data Types:** Verify correct types (e.g., numerical as float, categorical as string).
- **Missing Values:** Handle with imputation (mean/median for numerical, mode for categorical) or remove if >30%.
- **Empty Strings:** Replace place holders like "N/A" or "Unknown".
- **Numerical:** Validate ranges (e.g., Ratings: 0-10, Expenses/Collections: no negatives).
- **Categorical:** Ensure valid values (e.g., 3D_available: "YES"/"NO").
- **Consistency:** Collections ≥ Budget; Time taken is plausible.
- Remove duplicate rows and ensure unique identifiers.
- **Summary Stats:** Identify anomalies using mean, median, and standard deviation.
- **Outliers:** Detect using IQR or Z-scores.
- Validate logical links (e.g., Budget ≥ Marketing + Production Expense).
- Confirm positive correlations (e.g., Collection with Marketing Expense).
- Ensure proper encoding for categorical variables (e.g., one-hot encoding).
- **Histograms/Boxplots:** Detect outliers in numerical data.
- **Scatter/Bar Charts:** Validate relationships and frequency of categorical data.

Data Cleaning

- Replace missing numerical values with the mean/median and categorical values with the mode.
- Identify and remove duplicate rows.
- Remove unrealistic values (e.g., negative numbers) or replace outliers with statistical measures.
- Ensure numerical columns are floats/ints and categorical columns are properly classified.
- Remove spaces, standardize with underscores.
- Standardize categorical values for consistency.
- Add new metrics like ROI or marketing efficiency.
- Ensure no missing values, verify consistency, and validate derived metrics.

Tools

- **Excel/Google Sheets:** Quick cleaning, pivot tables, basic charts.
- **Python (Pandas, NumPy):** Programmatic cleaning and manipulation.
- **Power BI:** Interactive dashboards (slicers, bar charts, scatter plots).
- **Tableau:** User-friendly interactive visuals (e.g., ROI insights).
- **Python (Matplotlib, Seaborn, Plotly):** Detailed and interactive visualizations.

- **PowerPoint:** Structured presentations with charts and insights.
- **Google Data Studio:** Real-time shareable dashboards.
- **Excel Dashboards:** Dynamic reports using pivot charts and slicers.
- **SPSS/R:** Advanced statistical analysis and modelling.
- **SQL:** Querying and processing large datasets.
- **Notion:** Document insights and share with teams.
- **Google Docs/Slides:** Collaborative data summaries and presentations.

Dashboard

- **Showcase key metrics:** revenue trends, ROI, expenses vs. collections, genre impact.
- **Target audience:** Management or analysts.
- Load the dataset into Power BI.
- Create Visuals
  - **KPIs:** Total revenue, average ROI.
  - **Bar Chart:** Revenue by genre.
  - **Scatter Plot:** Expenses vs. revenue.
  - **Pie Chart:** 3D vs. non-3D movies.
  - **Line Chart:** Collection trends.
- Filter by genre, 3D availability, expense ranges.
- Use logical grouping and clear titles, labels, and legends.
- Ensure consistent and clean layout.

Storytelling

- Understand Your Audience.
  - **Executives:** Focus on high-level ROI and profitability.
  - **Marketing Team:** Emphasize marketing impact on revenue.
  - **Production Team:** Highlight genre success and ratings.
- Present Key Insights
  - **Revenue by Genre:** Bar chart—Drama and action dominate; niche genres offer higher ROI.
  - **Marketing vs. Revenue:** Scatter plot—Diminishing returns beyond a budget threshold.
  - **3D vs. Non-3D:** Pie chart—3D movies generate 35% of revenue but cost more.
  - **Ratings and Collections:** Heatmap—Critic ratings correlate more strongly with revenue.
  - **Production Duration:** Line chart—Shorter cycles (<2 years) yield better collections.
- Address Challenges
  - Does marketing spend efficient?
  - Are niche genres underutilized despite high ROI?
- Recommendations
  - Boost marketing budgets for action and drama.
  - Leverage critic reviews to increase audience trust.
  - Focus on shorter production cycles for timely releases.
- Conclude with Vision
  - Data-driven strategies align production and marketing for optimized returns.
- Make It Visual
  - Use annotated graphs, infographics, and trend highlights to engage the audience.