## Model Development Phase

| | |
|---|---|
| Date | 19 June 2025 |
| Team ID | SWTID1749705685 |
| Project Title | Movie Box Office Gross Prediction<br>using Machine Learning |
| Maximum Marks | 5 Marks |

### Feature Selection Overview

During the feature selection phase, each input variable was carefully evaluated based on its correlation with the target variable (box office revenue), domain relevance, data availability, and impact on model performance. The selected features were chosen to balance predictive power and model simplicity. Each feature listed below includes a brief description, selection status, and justification to ensure transparency and clarity in the decision-making process.

| Feature | Description | Selected (Yes/No) | Reasoning |
|---|---|---|---|
| Budget | Total production cost of the movie (in millions) | Yes | Strong correlation with revenue; essential for financial success estimation. |
| Genres | Primary genre category (eg., Action, Comedy, Fantasy, etc.) | Yes | Influences audience interest and ticket sales; one-hot or label encoded. |
| Popularity | Popularity score based on TMDB metrics | Yes | Captures public interest; improves model's predictive power. |

| | | | |
|---|---|---|---|
| Runtime | Duration of the movie in minutes | Yes | Affects audience retention and theater scheduling; moderate correlation. |
| Vote Average | Average user rating in TMBD | Yes | Indicates quality perception; positively associated with box office revenue. |
| Vote Count | Total number of votes achieved | Yes | Reflects the reach and engagement; improves generalization. |
| Director | Name of the director, mapped numerically | Yes | Top directors often influence box office outcomes; mapped using top 10. |
| Release Month | Month in which the movie is released (1 - 12) | Yes | Seasonality affects success (summer/holiday releases perform better). |
| Month of the Week (MOW) | Week of the Month when movie was released | Yes | Replaces day-of-week to capture broader timing trends within a month. |
| Homepage | Indicates if the movie has an official website | No | Sparse data; not significantly correlated with revenue in our dataset. |
| Production Company | Studio or distributor behind the film | No | Too many unique categories; leads to high cardinality and overfitting. |
| Spoken Languages | Number of languages the movie is available in | No | Low variance in dataset; does not contribute significantly to predictions. |