Project Report: Movie Success Prediction and Sentiment Study

This report integrates findings and methodologies from projects focusing on movie box office revenue prediction and movie sentiment analysis. The objective aligns with your stated project goals: **predicting movie success using IMDb/Kaggle data** and **analyzing the sentiment expressed in viewer reviews** and **movie titles**. The analysis uses various datasets sourced from IMDb.

Project Areas

The combined project covers two primary areas:

1. **Movie Box Office Revenue Prediction**: Building a regression model to predict worldwide gross revenue.
2. **Movie Sentiment Analysis**: Examining sentiment in movie titles across genres and classifying sentiment in user-written reviews.

Dataset

The projects utilize datasets derived from IMDb.

- **Revenue Prediction Dataset**: A CSV file (box_office_data.csv) containing top movies by worldwide gross. Key columns include **Rank**, **Release Group** (title), **$Worldwide**, **$Domestic**, **$Foreign** (revenues), **Year**, **Genres**(comma-separated), **Rating** (MPAA), **Vote_Count** (IMDb), **Original_Language**, and **Production_Countries.** The source is described as the top 1000 or more movies by worldwide gross.
- **Title Sentiment Dataset**: Uses the IMDb box office dataset, including title and genre metadata. Key fields are **Movie Title** and **Comma-separated Genres**.
- **Review Sentiment Dataset**: IMDb movie reviews, specifically 50,000 labeled examples split evenly between positive and negative sentiment. Each row contains a review text and its corresponding label (pos or neg).

Preprocessing & Methodology

Different preprocessing steps and methodologies are applied depending on the analysis area:

- **Revenue Prediction Methodology**:

    1. **Missing Data Handling**: Rows missing the target ($Worldwide) are dropped. Other missing values are dropped or imputed.
    2. **Feature/Target Split**: The target is $Worldwide, and features include Year, Genres, Rating, Vote_Count, Original_Language, and Production_Countries.
    3. **Pipeline Construction**: A scikit-learn pipeline is used, wrapping a RandomForestRegressor. This includes a **Numerical Transformer** (StandardScaler on Year and Vote_Count) and a **Categorical**

Transformer(OneHotEncoder on Genres, Rating, Original_Language, Production_Countries).

4. **Train/Test Split**: Data is split 80% for training and 20% for testing with a fixed random_state=42.
5. **Training & Prediction**: The pipeline is fitted on the training data and predictions are made on the test set.

- **Title Sentiment Methodology**:

  1. **Data Cleaning**: Entries missing title or genre are removed.
  2. **Genre Explosion**: Multi-genre entries are split so each row represents one title-genre pair.
  3. **Title Sentiment Scoring**: A polarity score is computed for each title on a –1 to +1 scale.
  4. **Aggregation**: The average title polarity is calculated for each genre.

- **Review Sentiment Methodology**:

  1. **Preprocessing**: Includes lowercase conversion, removal of HTML tags, punctuation, and extraneous whitespace. Stopword removal or lemmatization is optional as VADER is robust to unprocessed text.
  2. **Loading Data**: The dataset is read into a pandas DataFrame.
  3. **Sentiment Scoring**: VADER (Valence Aware Dictionary and sEntiment Reasoner), a lexicon-based analyzer, is used to compute four scores per review: pos, neu, neg, and compound. The SentimentIntensityAnalyzer() from nltk.sentiment.vader is instantiated.
  4. **Label Assignment**: A simple rule is used: if the compound score is ≥ 0, the predicted label is 'pos'; if it is < 0, the predicted label is 'neg'.

Visualizations

Visualizations are used to interpret results in both prediction and sentiment analysis:

- **Revenue Prediction Visualizations**:

  5. **Actual vs. Predicted Scatter Plot**: Shows test-set actual vs. predicted revenue, highlighting model fit and outliers. Includes a 45° dashed line for comparison.
  6. **Residuals Distribution Histogram**: Displays the distribution of prediction errors (residuals), showing model bias (mean) and variance (width).
  7. **Feature Importance Bar Chart**: Ranks the top 10 most important features from the Random Forest model, indicating which inputs strongly drive predictions (e.g., Vote_Count, Year, Genre indicators).

- **Title Sentiment Visualizations**:

  1. **Genre-wise Average Sentiment Bar Chart**: Bars sorted by mean polarity, showing which genres use more positive or negative language in titles.
  2. **Title Sentiment Distribution Box Plot**: Boxplots per genre showing median, interquartile range, and outliers of title polarity scores.

- **Review Sentiment Visualizations**:

  1. **Distribution of Compound Scores**: Overlaid histograms for true positive vs. true negative reviews, showing score separation.

2. **Confusion Matrix Heatmap**: A 2×2 grid showing counts of true positives, false positives, true negatives, and false negatives, highlighting error types.
3. **Precision/Recall/F1 Bar Chart**: Grouped bars comparing metrics for pos and neg classes, identifying potential bias towards one class.

(detailed visualizations are mentioned in the individual reports)

Results & Output

- **Revenue Prediction Results**:
    - Mean Squared Error (MSE): **1.886×10^{16}**. The large MSE is noted due to high variance in box office figures.
    - R^2 Score: **0.528**. This indicates the model explains approximately 52.8% of the variance in worldwide gross.
    - Interpretation suggests **moderate predictive power**.
- **Title Sentiment Results**:
    - Average polarity is **highest in Family and Animation** (titles skew positive).
    - Average polarity is **lowest in Horror and Thriller** (titles skew negative).
    - Distribution observations show **Comedy and Action** have wide sentiment ranges. **Adventure and Sci-Fi**cluster around neutral polarity.
- **Review Sentiment Results**:
    - Overall Accuracy: **~85%** (varies slightly with preprocessing).
    - Class-Specific Metrics: Positive class metrics are slightly higher (Precision ≈ 0.84, Recall ≈ 0.88, F1 ≈ 0.86) than negative class metrics (Precision ≈ 0.86, Recall ≈ 0.82, F1 ≈ 0.84).
    - Key Observations: Positive reviews tend to receive higher compound scores (~+0.8) while negative reviews cluster near –0.5. The confusion matrix often shows slightly more false negatives (positive reviews misclassified as negative). Precision/Recall imbalance suggests focusing on reducing false negatives.

Limitations

Several limitations are identified across the projects:

- **Title-Only Focus** (Title Sentiment): Does not consider full review text, plot details, or audience reception.
- **Lexicon Constraints** (Title Sentiment, Review Sentiment): Basic polarity scoring or VADER (a fixed lexicon) may misinterpret proper nouns, nuanced language, context, sarcasm, or domain-specific language.
- **Genre Overlap** (Title Sentiment): Multi-genre films contribute equally to each genre's aggregate, potentially diluting signals.
- **Feature Granularity** (Revenue Prediction): Genres and production countries are one-hot encoded without grouping rare categories.
- **Model Simplicity** (Revenue Prediction): Does not account for marketing budgets, star power, franchise effects, or seasonal release windows.
- **Imbalanced Distribution** (Revenue Prediction): Extreme outliers (blockbusters) can skew error metrics.

- **Threshold Sensitivity** (Review Sentiment): The simple compound ≥ 0 rule is simplistic.
- **Balanced Dataset Assumption** (Review Sentiment): Real-world data may be skewed, affecting model behavior under class imbalance.

Future Work

Areas for future improvement are suggested for each project:

- **Title Sentiment Future Work**:

    4. Keyword Frequency Analysis: Identify common positive/negative terms per genre.
    5. Advanced Sentiment Models: Apply transformer-based analyzers.
    6. Metadata Correlation: Explore links between title sentiment and box office performance or critic scores.
    7. Temporal Trends: Track how title sentiment by genre evolves over decades.
- **Revenue Prediction Future Work**:

    1. Add Metadata: Incorporate director, cast popularity, budget, release month.
    2. Hyperparameter Tuning: Grid search over model parameters (e.g., tree depth, number of estimators).
    3. Alternative Models: Gradient Boosting (XGBoost, LightGBM) or neural networks.
    4. Error Analysis: Analyze large residuals to find systematic biases.
    5. Cross-Validation: Use k-fold CV for more robust performance estimates.
- **Review Sentiment Future Work**:

    1. Threshold Optimization: Use ROC analysis to choose an optimal compound threshold.
    2. Ensemble Methods: Combine VADER with machine learning models (e.g., logistic regression on TF-IDF features).
    3. Deep Learning Approaches: Fine-tune a pre-trained transformer (e.g., BERT).
    4. Error Analysis: Manually inspect misclassified examples.

References

The sources reference specific tools and datasets:

- Scikit-Learn Documentation
- IMDb Datasets
- Hutto, C.J. & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text