Milestone 2 Report: Feature Engineering, Selection, and Modeling

1. Project Objective

The goal of this project is to develop an interactive data science tool (dashboard) for analyzing movie datasets (IMDb, Rotten Tomatoes, Netflix). This tool aims to facilitate data-driven decision-making by allowing users to explore relationships, trends, and patterns, ultimately understanding key factors influencing movie success, audience engagement, and critical reception.

2. Data Used

The analysis utilizes the following datasets:

- **Rotten Tomatoes Movies Dataset:** Contains metadata about movies, including ratings, genres, runtime, release dates, and production details. (Primary dataset for modeling in this milestone).
- **Rotten Tomatoes Reviews Dataset:** Includes audience and critic reviews.
- **Netflix Movies Dataset:** Information about movies available on Netflix.
- **IMDb Movies Dataset:** Details on movies, user ratings, director, and box office earnings.
- **TMDB Credits Dataset:** Information on cast and crew for movies.

These datasets were sourced from public repositories like Kaggle. Preprocessing steps from Milestone 1 (handling missing values, duplicates, outliers, scaling, encoding) were applied, and the cleaned_rotten_movies.csv file served as the input for this milestone. After loading and feature engineering, the dataset used for modeling contained 8077 samples.

3. Technology Stack

- **Programming Language:** Python
- **Data Manipulation & Analysis:** Pandas, NumPy
- **Machine Learning & Statistics:** Scikit-learn (sklearn), SciPy
- **Data Visualization:** Matplotlib, Seaborn
- **Development Environment:** Jupyter Notebook
- **(Future) Dashboarding:** Plotly Dash or Streamlit

4. Project Timeline

- **Milestone 1: Data Collection, Preprocessing, EDA:** January 17, 2025 - February 20, 2025 (5 weeks) - *Completed*
- **Milestone 2: Feature Engineering, Selection, Modeling:** February 21, 2025 - March 26, 2025 (5 weeks) - *Current*
- **Milestone 3: Tool Development and Finalization:** March 27, 2025 - April 30, 2025 (5 weeks)

5. Exploratory Data Analysis (EDA) Report Summary (from Milestone 1)

Key insights derived from the initial EDA include:

- **Distributions:** Numerical features like ratings (`tomatometer_rating`, `audience_rating`) showed varied distributions. Review counts (`tomatometer_count`, `audience_count`) were often right-skewed. Movie `runtime` generally followed a normal distribution.
- **Relationships:** Pairplots revealed potential correlations, such as a weak positive trend between critic (`tomatometer_rating`) and audience (`audience_rating`) scores. Many relationships appeared non-linear.
- **Outliers:** Box plots highlighted significant outliers, especially in review counts (`tomatometer_count`, `audience_count`) and `runtime`, indicating movies with exceptionally high review volumes or unusual lengths.
- **Missing Data:** Addressed during preprocessing using imputation (median for numerical, mode for categorical) or dropping rows/columns based on the extent of missingness.
- **Categorical Data:** Features like `genres`, `content_rating`, and `production_company` were analyzed for frequency and distribution.

6. Feature Engineering

New features were created from the preprocessed `rotten_tomatoes_movies` dataset to enhance model predictive power:

- **Time-Based Features:**
  - `release_year`: Extracted from `original_release_date`.
  - `release_month`: Extracted from `original_release_date`.
  - `release_dayofweek`: Extracted from `original_release_date`.
  - `movie_age`: Calculated as the difference between the current year (2025) and `release_year`.

- **Ratio/Interaction Features:**
    - `audience_tomatometer_ratio`: Ratio of `audience_count` to `tomatometer_count` (plus a small epsilon to avoid division by zero).
    - `runtime_rating_interaction`: Product of `runtime` and `tomatometer_rating`.
- **Categorical/Text Features:**
    - `num_genres`: Calculated by splitting the `genres` string and counting the elements.
    - `production_company_encoded`: High-cardinality `production_company` feature was label encoded after filling missing values with 'Unknown'.
    - Categorical features like `content_rating` were one-hot encoded if present and not already handled.
- **Target Variable Encoding:** The target variable `tomatometer_status` (with original values 'Rotten', 'Fresh', 'Certified-Fresh') was label encoded into numerical representation (e.g., 0, 1, 2) for modeling. The mapping was: `{'Certified-Fresh': 0, 'Fresh': 1, 'Rotten': 2}`.

7. Feature Selection

The goal was to select the most relevant numerical features for predicting the `tomatometer_status`.

- **Methods Used:**
    - **ANOVA F-test:** `SelectKBest` with `f_classif` scoring was used to evaluate the relationship between each numerical feature and the categorical target. The top 10 features identified were: `['runtime', 'tomatometer_rating', 'tomatometer_count', 'audience_rating', 'tomatometer_top_critics_count', 'tomatometer_fresh_critics_count', 'tomatometer_rotten_critics_count', 'release_dayofweek', 'movie_age', 'runtime_rating_interaction']`.
    - **Random Forest Importance:** A Random Forest Classifier was trained on all numerical features, and feature importances (mean decrease in impurity) were extracted.
- **Selected Features:** Based on the Random Forest Importance scores, the following top 10 features were selected for modeling:
  ['tomatometer_rating', 'runtime_rating_interaction', 'tomatometer_fresh_critics_count', 'tomatometer_count', 'tomatometer_rotten_critics_count', 'audience_rating', 'tomatometer_top_critics_count', 'runtime', 'movie_age', 'audience_count']
-

- **Justification:** Random Forest Importance was chosen as the primary method because it captures non-linear relationships and feature interactions. The selected features represent a mix of original ratings, counts, engineered interactions, critic counts, time-based information, and runtime deemed most influential by the model. Features with low importance were excluded.

8. Data Modeling

- **Objective:** Train and evaluate classification models to predict the `tomatometer_status` (Rotten, Fresh, Certified-Fresh).
- **Data Splitting:** The dataset (containing 8077 samples and 10 selected features) was split into:
    1. Training Set (60% - 4845 samples)
    2. Validation Set (20% - 1616 samples)
    3. Test Set (20% - 1616 samples) Stratification based on the target variable (`y_selected`) was used during splitting.
- **Feature Scaling:** `StandardScaler` was applied to the training data and used to transform the validation and test sets.
- **Models Trained:**
    1. Logistic Regression (with `multi_class='ovr'`)
    2. Decision Tree Classifier
    3. Random Forest Classifier (`n_estimators=100`)

8.1. Model Performance on Validation Data

Models were trained on the training set and evaluated on the validation set. Performance metrics (weighted averages for precision, recall, F1; ROC AUC using weighted One-vs-Rest) are summarized below:Export to Sheets

8.2. Model Comparison and Analysis

- Based on the validation set performance, the **Random Forest Classifier** demonstrated the highest performance across all metrics, achieving near-perfect scores with an Accuracy of 0.9845 and an ROC AUC of 0.9993.
- The Decision Tree also performed exceptionally well (Accuracy 0.9821, ROC AUC 0.9871), suggesting the features selected are highly predictive and the decision boundaries might be relatively clear.
- Logistic Regression provided strong baseline performance (Accuracy 0.9264, ROC AUC 0.9859) but was outperformed by the tree-based ensemble methods.
- The Random Forest model was selected as the best model for final evaluation on the test set due to its superior performance on the validation data.

8.3. Best Model Performance on Test Data

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Random Forest	0.9845	0.9846	0.9845	0.9845	0.9993
Decision Tree	0.9821	0.9821	0.9821	0.9820	0.9871
Logistic Regression	0.9264	0.9250	0.9264	0.9251	0.9859

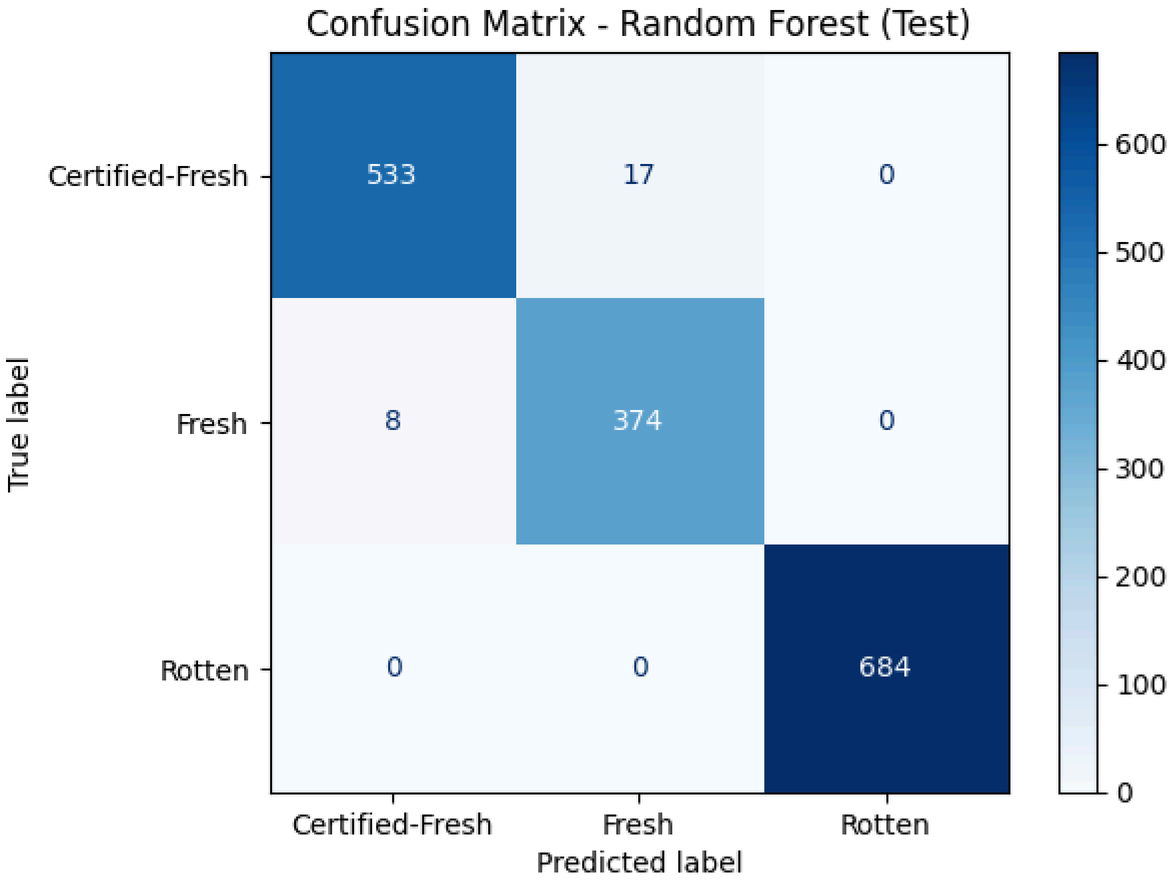
The selected best model (Random Forest Classifier) was evaluated on the held-out test set to estimate its generalization performance on unseen data.

- **Test Performance Metrics:**
  - Accuracy: 0.9845
  - Precision (weighted): 0.9847
  - Recall (weighted): 0.9845
  - F1-Score (weighted): 0.9846
  - ROC AUC (weighted OvR): 0.9987

Classification Report (Validation)

	Precision	**Recall**	**F1-Score**	**Support**
Certified-Fresh	**0.90**	**0.92**	**0.91**	**550**
Fresh	**0.88**	**0.80**	**0.84**	**382**
Rotten	**0.97**	**1.00**	**0.98**	**684**

- **Confusion Matrix (Test Set):**



- **Analysis:** The performance of the Random Forest model on the test set was virtually identical to its performance on the validation set (Accuracy 0.9845, ROC AUC 0.9987), indicating excellent generalization to unseen data. The classification report shows very high precision and recall across all three classes ('Certified-Fresh', 'Fresh', 'Rotten'), demonstrating the model's strong ability to accurately predict the tomatometer status based on the selected features. The near-perfect scores suggest the chosen features are highly informative for this classification task.