Machine Learning Project Overview: Predicting TV Show Popularity Score

Research Question
What is the predicted popularity rank of a TV show given its attributes?

Prior Knowledge
TV show popularity is driven by content quality, celebrity influence, and genre appeal. Marketing, social media engagement, and platform accessibility further shape audience reach. Demographics and societal trends also play key roles in determining a show's success.

Goals

- Using a supervised machine learning model to predict TV show popularity rankings based on well-cleaned, selected features from over 150K records of the Full TMDB TV Shows dataset.
- Achieving a low error score on evaluation metrics to reflect model performance.
- The ultimate goal is to help production companies uncover the 'recipe' for a successful TV show, leading to increased revenue.

Key factors
Selected time period (2018-2023), preprocessing techniques, model improvement and evaluation strategies.

Project Stages

1. Data Preparation

- **Dataset Scope:** Filtering records (first air date: 2018-2023, ~50K records).
- **Feature Selection:** Dropping columns with >55% nulls, duplicates, URLs.
- **Categorical Handling:** Reducing to top 5/10 categories; others mapped to "Other."
- **Text Isolation:** Separating text features into df_text.

2. Exploratory Data Analysis (EDA)

- **Visualizations:** AutoViz, box plots, missing value metrics (msno).
- **Feature Analysis:** Numeric, categorical, Boolean, target variable (log transformation check).
- **Outlier Detection:** IQR, correlation & distribution changes.
- **Normality Check:** Most features are skewed.
- **Handling Missing Data**:
    - **Categorical:** KNN & random min-max imputations for genres, networks, language, origin country.
    - **Numerical:** MICE imputation.
    - **Dates:** Filling last_air_date based on status.

3. Feature Engineering & Selection:

- **Text:** Word Cloud (top 5 common words).
- **Dates:** Extracting Year & Month.
- **Encoding:** One-Hot Encoding & Label Encoding.
- **Feature Selection:** Lasso, Ridge, Gradient Boost, Random Forest—final 24 features out of 60.

4. Model Selection & Evaluation

- **Data Split:** Train+Valid (80%), Test (20%).
- **Models:** Linear Regression, Decision Tree, Random Forest - **the chosen one**, AdaBoost, GBM, SVM, XGBoost (Linear Regression provided better results on RMSE and MSE while SVM on MAE).
- **Metrics:** MSE, MAE, RMSE, RMSLE (decision was based on RMSLE).

5. Model Fine-Tuning

- **Hyperparameters:** n_estimators, max_features, max_depth, min_samples_split, min_samples_leaf, bootstrap.
- **Optimization:** Grid-search & cross-validation – **demonstrated lower performance than the base model** - stayed with the base model.