Comprehensive Machine Learning Project Documentation — TV Series Renewal Prediction

1. Project Overview & Objectives

This project addresses a critical business question for television production studios and streaming platforms: predicting whether a TV series will be renewed for another season. By developing a robust predictive model, this initiative aims to provide a data-driven framework for strategic content and investment decisions. The project was executed with strict adherence to the Bar-Ilan University (BIU) Machine Learning methodology, ensuring a structured, transparent, and reproducible lifecycle from initial data preparation through final model evaluation.

The analysis is based on a comprehensive TMDB TV series dataset. The initial dataset contained 168,639 records, each described by 29 distinct features, providing a rich foundation for predictive modeling.

The core characteristics of the dataset are summarized below:

- **Dataset:** TMDB TV Series Data
- **Initial Records:** 168,639
- **Initial Features:** 29
- **Target Variable:** in_production (Boolean), representing whether a series is ongoing (1) or has ended (0).

This document details each phase of the project, beginning with the foundational step of Data Preparation.

2. Data Preparation

The objective of the Data Preparation phase was to transform the raw, unprocessed dataset into a clean, standardized, and structurally consistent format suitable for analysis and modeling. This foundational step is essential for ensuring the quality, reliability, and integrity of all subsequent project phases.

The preparation process followed a sequential and methodical workflow:

1. **Data Loading and Inspection:** The TMDB dataset was loaded into memory, and its initial structure was inspected to verify the number of rows and columns, column data types, and non-null value counts.
2. **Data Type Conversion:** To ensure proper handling by analytical libraries, columns containing textual or categorical information, initially typed as object, were explicitly converted to the string data type. Similarly, columns representing dates (first_air_date, last_air_date) were converted to the datetime format.
3. **Categorical Feature Reduction:** To reduce noise and simplify features with high cardinality, a strategy of grouping rare categories was implemented. A frequency threshold of 1.0% was established; any category appearing in less than 1% of the non-null records within a feature

was mapped to a consolidated "Other" category. This transformation was applied to seven high-cardinality columns, including original_language and origin_country.

4. **Basic Text Cleaning:** Free-text fields, such as overview and tagline, underwent a basic cleaning process. This included converting all text to lowercase and removing special characters to standardize the content for potential text-based feature engineering in later stages.

The key decisions made during this phase are documented in the table below.

Key Data Preparation Decisions

Area	Decision	Rationale
Missing Values	Kept as NaN	To be handled systematically in the subsequent Data Cleansing stage.
Category Reduction	Threshold = 1%	Balances feature simplification with the retention of meaningful information.
Skipped Columns	genres, production_companies	Deferred due to high complexity and multi-value text format.
Duplicate id	Documented, not removed	The non-uniqueness of the id column was noted for handling in the Data Cleansing stage.

As a result of this phase, the dataset was successfully standardized. Data types were validated, and the complexity of categorical features was effectively managed. The prepared dataset provided a solid and reliable foundation for the next stage of the project: Exploratory Data Analysis.

3. Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) phase was conducted to uncover underlying patterns, identify anomalies and data quality issues, and form hypotheses about the dataset's structure and predictive potential. This critical investigation provides the necessary context to guide subsequent data cleansing and feature engineering efforts.

The analysis yielded several key findings:

- **Data Quality Issues:** A detailed review of numeric variables revealed significant right-skew in distributions for features like popularity and vote_count, indicating the presence of a long tail of highly popular shows. The analysis also confirmed the existence of extreme outliers and high levels of missing data, particularly in text-based attributes such as overview and tagline.

- **Categorical Distributions:** An examination of categorical features highlighted significant class imbalances. The dataset is heavily dominated by "Scripted" shows, which account for 76.7% of the type category. Similarly, the status feature is concentrated in two primary classes: "Ended" (55.3%) and "Returning Series" (40.0%).

- **Correlations:** The relationships between numeric variables were assessed using a Pearson correlation matrix. The analysis found that most pairwise correlations are weak, which effectively mitigates concerns about multicollinearity among the predictive features.

The correlation matrix for key numeric features is presented below.

Feature	number_of_ seasons	number_of_ episodes	vote_count	vote_average	popularity	episode_run_ time
number_of_ seasons	1.00	0.42	0.07	0.16	0.19	0.06
number_of_ episodes	0.42	1.00	0.04	0.09	0.34	0.03
vote_count	0.07	0.04	1.00	0.11	0.22	0.02
vote_average	0.16	0.09	0.11	1.00	0.13	0.16
popularity	0.19	0.34	0.22	0.13	1.00	0.02
episode_run_ time	0.06	0.03	0.02	0.16	0.02	1.00

Numeric Feature Correlation Matrix (Pearson)

Interpreting the correlation matrix, the strongest positive relationship was observed between number_of_seasons and number_of_episodes (r=0.42). This finding is logically consistent, as series with more seasons naturally tend to have a higher total number of episodes. The next strongest correlation was observed between popularity and number_of_episodes (r=0.34), which suggests a logical relationship where shows with more episodes tend to accumulate higher popularity scores.

The insights gained from this EDA phase provided a clear, evidence-based roadmap for the targeted interventions required in the subsequent Data Cleansing stage.

4. Data Cleansing

The Data Cleansing stage was strategically designed to methodically address the data quality issues identified during Exploratory Data Analysis. Its purpose was to resolve duplicates, manage missing values, mitigate the impact of outliers, and correct distributional skew, thereby creating a robust and high-quality dataset suitable for reliable model training.

A sequence of targeted cleansing operations was performed:

1. **Duplicate Removal:** Duplicate records were removed based on the id column to ensure that each unique TV series was represented only once in the dataset. This action resulted in the removal of 3,934 records, reducing the dataset size to 164,705 rows.

2. **Handling Placeholder Zeros:** Zero values within the vote_average and episode_run_time columns were identified as placeholders for unknown or missing data rather than valid measurements. These zeros were replaced with NaN (Not a Number) to ensure they would be correctly handled during imputation. To preserve the original information, binary indicator columns (_was_zero) were created to flag records that originally contained these placeholders.

3. **Missing Value Imputation:** A systematic strategy was employed to handle missing values. Numeric columns were imputed using the median, a measure robust to the influence of outliers. Categorical columns were filled with the string "Unknown" to explicitly represent missing information. Additionally, for columns with over 30% missing data, binary flags (_was_missing) were created to capture the absence of information as a potentially predictive signal.

4. **Outlier Capping:** To mitigate the influence of extreme values on the model, a capping strategy was applied. This involved using a combination of the Interquartile Range (IQR) method and percentile-based capping (retaining values between the 0.5 and 99.5 percentiles).

5. **Skewness Correction:** Right-skewed numeric features that had been capped (popularity_capped, vote_count_capped, etc.) were transformed using a log1p function. This transformation helps to normalize their distributions, making them more suitable for linear models and improving overall model stability.

The impact of these transformations on key numeric features is summarized below.

Data Transformation Impact Summary

Feature	Transformation	Skewness (Before → After)	Max Value (Before → After)
popularity_capped	log1p	1.19 → 0.82	5.23 (capped)
vote_count_capped	log1p	9.03 → 2.65	513.48 (capped)
number_of_episodes_capped	log1p	5.225 → 0.362	422.48 (capped)
episode_run_time	Capping	N/A	6032.00 → 180.00

These cleansing operations significantly improved the overall quality and integrity of the data. The resulting dataset, characterized by mitigated outliers and more normalized distributions, provided a solid foundation for the feature engineering and selection phase.

5. Feature Engineering & Feature Selection

This phase had a dual objective: first, to create new, informative features from the existing data to enhance predictive power (Feature Engineering), and second, to systematically identify and select the most relevant subset of features for the final model (Feature Selection).

Feature Engineering

To capture additional predictive signals, several new features were engineered from the cleaned dataset. These additions were designed to provide insights that were not explicitly present in the original data:

- **Text-Derived Features:** To quantify the amount of descriptive text available for each series, overview_len (character count) and overview_words (word count) were created as simple numeric proxies.
- **Date-Derived Features:** Temporal aspects of a series's lifecycle were captured by engineering first_air_year, series_age (calculated as the number of years since the first air date), and active_days (the duration between the first and last air dates).
- **Missingness Indicators:** To explicitly model the absence of data as a potential feature, additional binary flags such as first_air_date_was_missing were created.

Feature Selection

A methodical, multi-step process was employed to reduce the feature space from 108 candidate features (post-encoding) to a final, optimized set of 30.

1. **Train-Test Split:** The dataset was partitioned into an 80% training set and a 20% testing set. Stratification was used to ensure that the proportion of the target variable (in_production) was maintained in both splits.
2. **Encoding:** All categorical features (identified by the _grp suffix) were transformed using one-hot encoding to convert them into a numerical format suitable for modeling.
3. **Initial Filtering:** A VarianceThreshold was applied to automatically remove any zero-variance features that provide no predictive information.
4. **Univariate Ranking:** An ANOVA F-test (SelectKBest) was conducted to score and rank each feature based on its individual statistical relationship with the target variable. This provided a quantitative measure of each feature's predictive potential.
5. **Correlation Pruning:** To eliminate redundancy, highly correlated numeric-like features were identified using a threshold of |r| >= 0.95. For each redundant pair, one feature was removed based on transparent, deterministic tie-breaking rules (e.g., dropping first_air_year in favor of the more direct series_age).
6. **Final Selection:** The final set of 30 features was constructed by selecting the top-ranked features from the ANOVA F-test results, after accounting for the features that were removed during the correlation pruning step.

The final set of 30 selected features is summarized by category in the following table.

Final 30 Features by Category

Feature Category	Count	Example Features
Missingness Flag	11	episode_run_time_was_missing, poster_path_was_missing
Categorical (One-Hot)	13	original_language_grp_en, genres_grp_Reality
Numeric (Prepared)	3	popularity_log, vote_count_log, number_of_seasons_capped
Date-Derived	1	series_age
Text Length	1	overview_words
Numeric (Plain)	1	number_of_seasons

The resulting set of 30 high-quality, de-correlated features was now ready for the model training and evaluation stage.

6. Modeling & Evaluation

The primary goal of the modeling stage was to systematically compare the performance of several machine learning algorithms to identify the most effective model for predicting TV series renewal. A rigorous evaluation protocol was employed, utilizing 3-fold cross-validation for hyperparameter tuning on the training set and a final evaluation on an untouched holdout test set to assess generalization performance.

Four candidate models were selected to represent a diverse range of algorithmic approaches:

- **Logistic Regression:** A reliable and interpretable linear model that serves as a strong baseline.
- **Linear SVM:** A margin-based linear classifier, chosen for its efficiency and effectiveness in finding optimal separating hyperplanes.
- **Random Forest:** A non-linear ensemble model capable of capturing complex feature interactions and robust to overfitting.
- **K-Nearest Neighbors (KNN):** A distance-based, non-parametric model used to test for the presence of cluster structures within the feature space.

The performance of each model was evaluated using cross-validation, with ROC-AUC as the primary selection metric.

3-Fold Cross-Validation Results

Model	CV ROC-AUC	CV F1-Score	Best Parameters
Random Forest	0.9249	0.8120	{'max_depth': 15, 'min_samples_split': 5, 'n_estimators': 200}
KNN	0.9056	0.7901	{'n_neighbors': 15, 'p': 1, 'weights': 'uniform'}
Logistic Regression	0.8509	0.7222	{'C': 1, 'solver': 'lbfgs'}
Linear SVM	0.8500	0.7196	{'C': 2}

Based on the cross-validation results, the **Random Forest** classifier was selected as the champion model. It demonstrated superior performance on both the primary metric (ROC-AUC) and the secondary metric (F1-Score), indicating its strong ability to both discriminate between classes and maintain a good balance of precision and recall. The champion model was then evaluated on the holdout test set to confirm its performance on unseen data.

Final Model Evaluation on Holdout Test Set (Random Forest)

Metric	Score
ROC-AUC	0.9251
F1-Score	0.8112
Precision	0.8090
Recall	0.8135
Accuracy	0.8441

To verify that the model generalizes well and is not overfit, its performance on the cross-validation set was compared to its performance on the test set. The CV ROC-AUC was 0.9249, while the Test ROC-AUC was 0.9251. The negligible difference of +0.0002 confirms that the model is robust and its performance on the training data is consistent with its performance on new, unseen data.

The selected Random Forest model demonstrates high predictive accuracy and robustness, making it a reliable tool for this prediction task.

7. Conclusions & Future Work

This project successfully designed, implemented, and validated a high-performing machine learning model for predicting the renewal of TV series. Following a systematic methodology, the process transformed a raw dataset into a clean, feature-rich input for a Random Forest classifier that achieved excellent predictive accuracy on unseen data.

Summary of Insights

- **Technical Insight:** The project underscored the criticality of a systematic data preparation and feature engineering pipeline. Features derived to represent data missingness (e.g., episode_run_time_was_missing) and temporality (series_age) proved to be highly predictive, as confirmed by their high ANOVA F-scores during feature selection.
- **Performance Insight:** The final Random Forest model achieved a high ROC-AUC score of 0.925 on the holdout test set. This result indicates a very strong capability to distinguish between series that will be renewed and those that will be discontinued.
- **Business Insight:** The model's success demonstrates that renewal prediction can be effectively data-driven. The analysis revealed that key predictive factors include a series's age, the completeness of its metadata, and its primary language and country of origin. These insights can be used to inform content strategy, acquisition priorities, and production investment decisions.