March 2025

ML Model Prediction – TMDB

PROJECT BY -EREZ LEVY

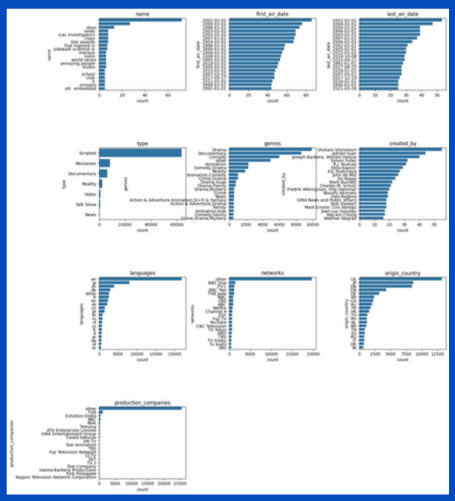
1. **Data Collection & Preperation:**

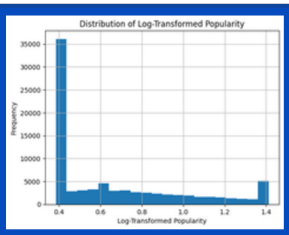
Reduce Large Categories:Dropping unnecesery Dataset columns
Cleaning common words & Punctuation marks from the Dataset tmdb
Converting Dataset to Lower case words
Identifying & Replacing rare features in columns into 'other'

count	mean	std	min	25%	50%	75%	max	
id	168639.0	111307.074704	76451.662352	1.0	45936.5	97734.000	196923.5000	251213.000
number_of_seasons	168639.0	1.548497	2.942872	0.0	1.0	1.000	1.0000	240.000
number_of_episodes	168639.0	24.465082	134.799622	0.0	1.0	6.000	20.0000	20839.000
vote_count	168639.0	13.305054	190.809059	0.0	0.0	0.000	1.0000	21857.000
vote_average	168639.0	2.333843	3.454334	0.0	0.0	0.000	6.0000	10.000
popularity	168639.0	5.882644	42.023216	0.0	0.6	0.857	2.4315	3707.008
episode_run_time	168639.0	22.603348	47.950427	0.0	0.0	0.000	42.0000	6032.000

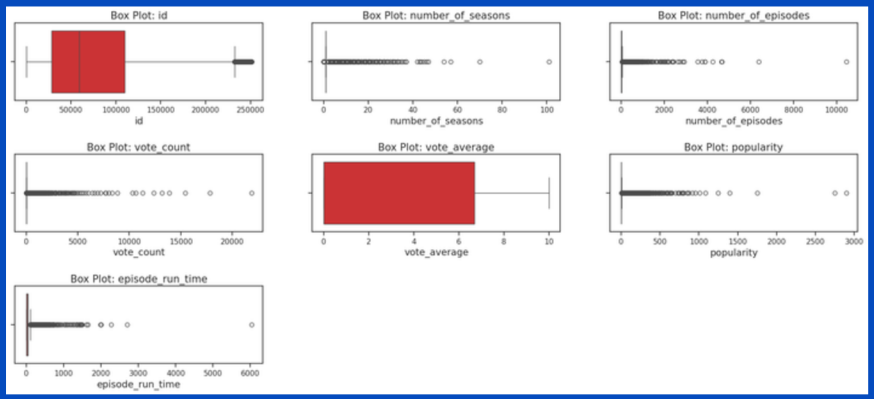
2. Exploratory Data Analysis (EDA)

Instant Reports:





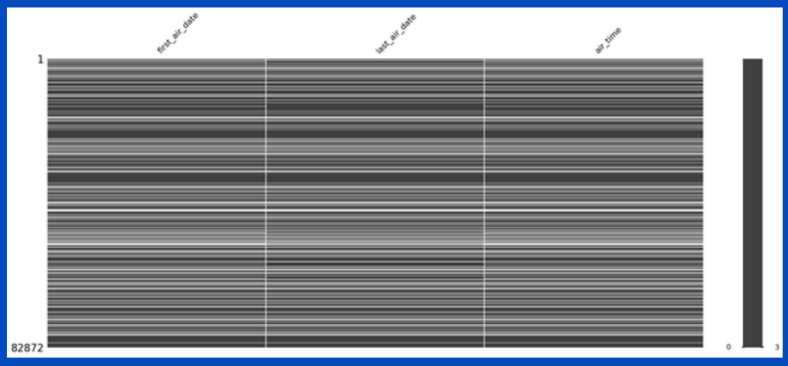
Creating data set of continuese data-Numeric Columns



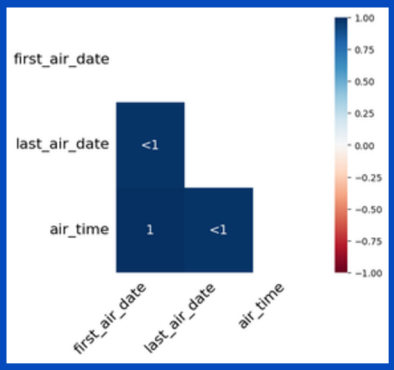
3. Data Cleansing

IQR columns & cap Outliers
Creating Dummies - tmdb_eda 'adult' column
Convert 'first_air_date' & 'last_air_date' to datetime objects & adding Column 'air_time'
Missing Values & Imputation Methods using MICE and KNN
Evaluate dropping the null values (up tp 5%) from the tmdb_eda Data Set

Plotting the missingness (nullity) matrix



Missingness Correlation HeatMap



4. Feature Eng. & Features Selection

Analyze the relationship between categorical and numerical features using box plots or violin plots.

Visualize correlation matrices using heatmaps for a clearer view of feature relationships.

Normalization/Standardization: Bringing Features to a Common Scale

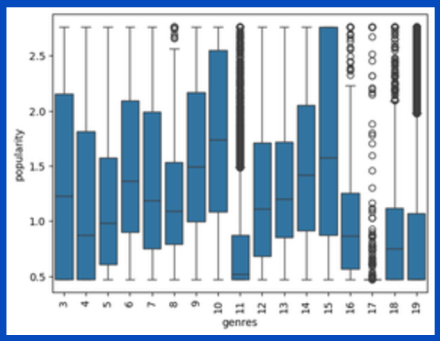
Sekuness-how much a distribution leans to one side or the other.

Feature Selection: Choosing the Right Ingredients for the Model

One-Hot Encoding

Spearman's Rank Correlation: Instead of Pearson's correlation (which is sensitive to outliers), use Spearman's rank correlation.

Categorical vs. Numerical:



Heat map:

Selected variables - recommended by 3 or more models

#	Column	Non-Null Count	Dtype
0	id	82872 non-null	float64
1	number_of_episodes	82872 non-null	float64
2	vote_average	82872 non-null	float64
3	genres	82872 non-null	int64
4	origin_country	82872 non-null	int64
5	episode_run_time	82872 non-null	float64
6	popularity_log	82872 non-null	float64
7	air_time	82872 non-null	float64
8	log_popularity	82872 non-null	float64
9	weighted_vote_average	82872 non-null	float64
10	popularity_score	82872 non-null	float64
11	popularity	82872 non-null	float64

dtypes: float64(10), int64(2)

5. Imbalance Techniques & Model Selection & Fine Tuning & Model Evaluation

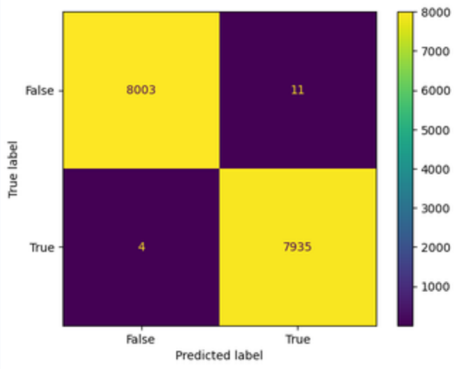
Applying imbalance Techniecs for the tmdb_model Dataset

Classification Report

Dataset Imbalansing Techniques-ROS,RUS,SMOTE,SMOTETomek

Model Selection: Choose appropriate machine learning models based on the problem at hand (e.g., regression, classification, clustering).

Confusion Metrix



Best Model Selection

model	MSE	RMSE	MAE	RMSLE	
1	Decision Tree	4.349159e-26	2.085464e-13	1.425803e-13	8.648281e-13
2	RandomForest	4.708293e-10	2.169860e-05	6.777158e-06	1.579591e-05
6	XGB	4.280811e-06	2.069012e-03	1.072934e-03	1.918067e-03
4	GBM	2.103887e-05	4.586815e-03	2.428622e-03	4.697174e-03
0	Linear Regression	2.570268e-03	5.069781e-02	3.726432e-02	7.461799e-02
5	SVR	2.383856e-03	4.882475e-02	4.134523e-02	1.340399e-01
3	ADABoost	3.800955e-03	6.165189e-02	5.480208e-02	1.891613e-01

Best Model (DecisionTreeRegressor)

supervised machine learning algorithm used for regression tasks, meaning it predicts a continuous target variable. It works by creating a tree-like structure where each internal node represents a decision based on a feature, and each leaf node represents a predicted value.

Fine Tuning - best model parameters :

Best Parameters for DecisionTree Regressor :

Regressor_max_depth : 10
Regressor_min_samples_leaf: 1
Reggressor_min_samples_split 2

Mean Squared Error : 4.967849782870243e-07

R-Squared : 0.9999994503652786