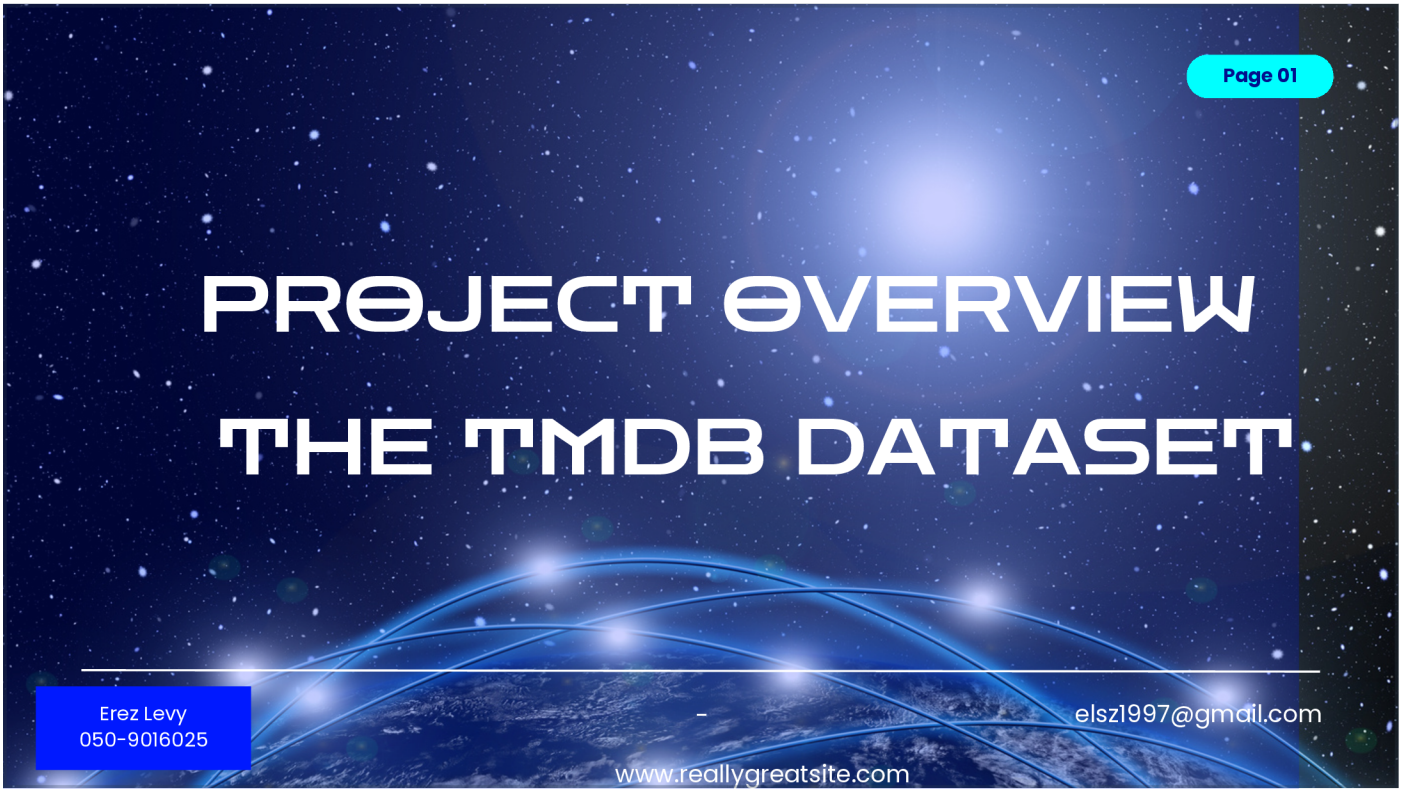

Page 01
PROJECT OVERVIEW
THE TMDB DATASET
Erez Levy
050-9016025
-
elsz1997@gmail.com
www.reallygreatsite.com

GUIDING QUESTIONS:

1. What am I trying to find out?

I aim to build a predictive model to determine the success of a TV show based on features such as vote count, vote average, and popularity.

CASE STUDIES

2. What do I already know?

The TMDB (The Movie Database) dataset provides extensive data on 150,000 TV shows, including ratings, genres, production details, and popularity metrics.

GUIDING QUESTIONS:

3. What is I aiming to achieve?

Success is defined by the creation of an accurate, well-validated machine learning model that can predict a TV show's success based on selected features.

GUIDING QUESTIONS:

4. What factors affect My results?

Data quality and completeness Feature selection (e.g., number of seasons, episodes, production company, language) Model selection and hyperparameter tuning , Handling of missing data and outliers Balancing the dataset if needed.

GUIDING QUESTIONS:

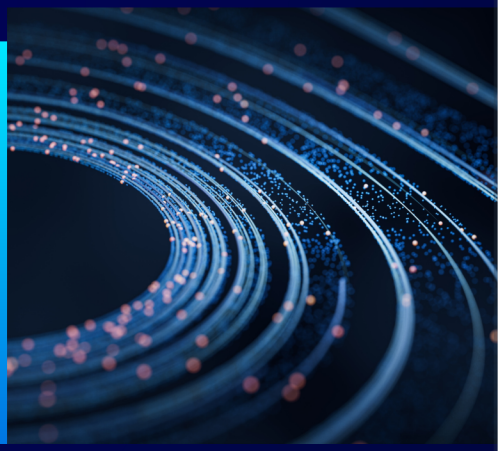
5. Is there something new I can use?

Advanced machine learning models such as Gradient Boosting and Random Forest for better predictions.
Feature engineering techniques to improve model accuracy.
Sentiment analysis from show descriptions to add another predictive elements.

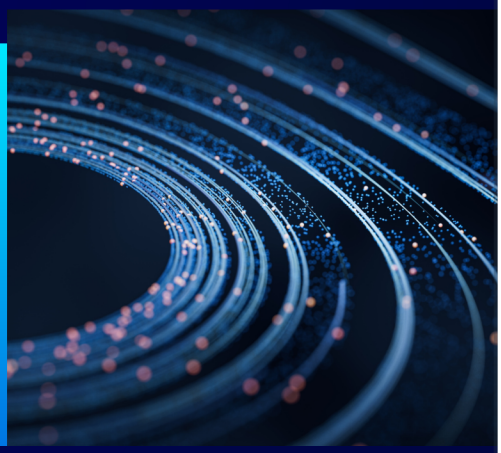
PROJECT STAGES OVERVIEW

1. Data Preparation:

Clean and preprocess TMDB dataset. Merge relevant tables and remove unnecessary columns & rows , reduce large categories standardize numerical features and handle missing values.



PROJECT STAGES OVERVIEW



2. Exploratory Data (EDA):

Visualize trends in TV show popularity, ratings and genres creating istance reports for understanding and analyzing correlations between show features and success metrics.

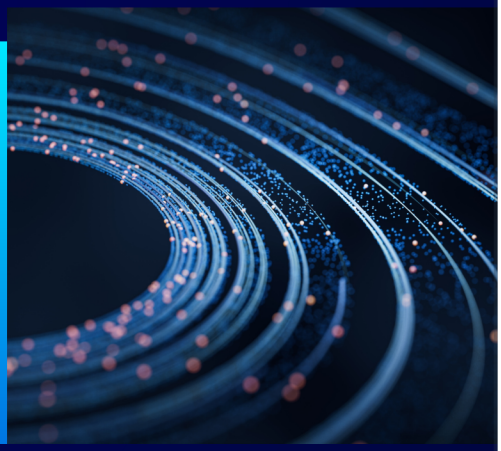
PROJECT STAGES OVERVIEW

3. Data Cleansing :

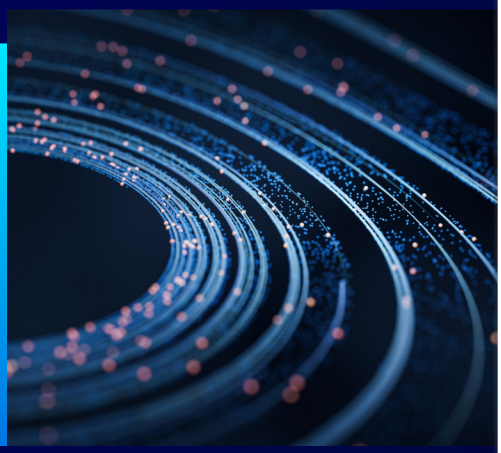
Identify and Detect and handle outliers which might skew the data analysis and model performance.

Address missing data through imputation methods (e.g., mean, median, mode).

Reduce highly imbalanced categories.



PROJECT STAGES OVERVIEW



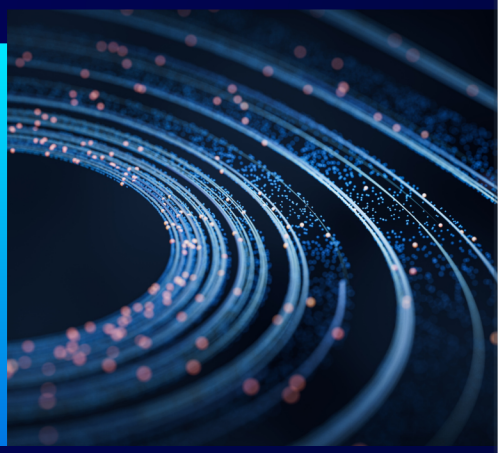
5. Feature Engineering:

Extract additional features such as episode duration and scale numerical features to ensure they have similar ranges, which helps certain algorithms perform better.

Feature selection choosing the most valuable features that predict the target value by running penalty models creator influence.

Using Principal Component Analysis (PCA) if needed.

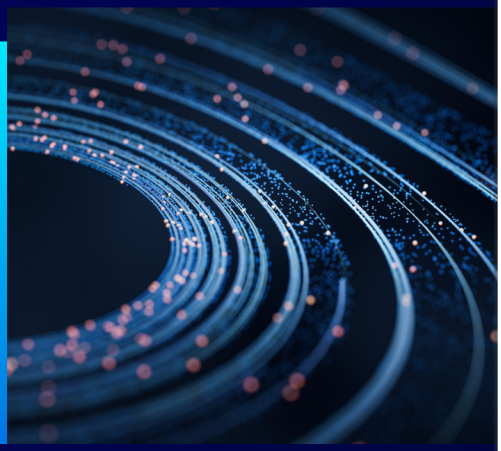
PROJECT STAGES OVERVIEW



4. One-Hot Encoding:

- Convert categorical data (e.g., genres, languages, production companies) into numerical format creating binary columns for each category in a categorical feature.

PROJECT STAGES OVERVIEW

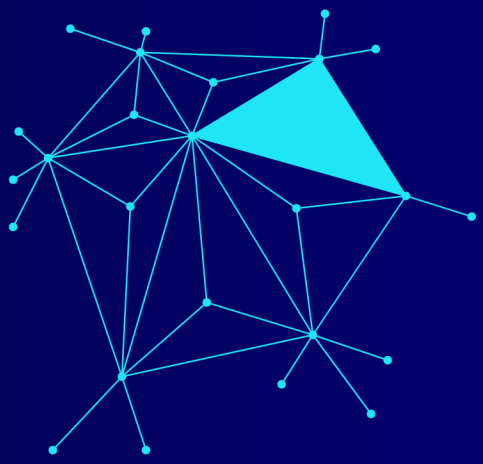


6. Imbalanced Data:

Refer to datasets where the distribution of classes is not equal.
This imbalance can pose a challenge for training machine learning models because the model might become biased towards the majority class and perform poorly on the minority class.
Useing oversampling or undersampling techniques to certain success categories are under presented.

PROJECT STAGES OVERVIEW

7. Model Selection and Fine-Tuning



Compare regression models:
Linear Regression (baseline model) Random Forest Regressor (captures non-linear relationships) Gradient Boosting Regressor (high-performance model).
Train and evaluate models using R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).
Perform cross-validation and hyperparameter tuning to optimize performance

PROJECT SUMMARY:

Decision Tree -The best choice considerations:

- **Regression problem:** The problem involves predicting 'popularity', which is a continuous variable, making it a regression task. Decision Trees (specifically, DecisionTreeRegressor) can handle regression problems effectively.
- **Feature importance:** Decision Trees can help identify which features are most important in predicting TV show popularity, providing valuable insights for business decisions.
- **Data characteristics:** The dataset has a mix of numerical and categorical features, and I suspect there are non-linear relationships between features and popularity, Decision Trees can handle these scenarios.

HOW WILL WE DEPLOY THE MACHINE LEARNING?

The final model will be deployed as a web API or integrated into an analytical dashboard for stakeholders.

Possible deployment through cloud platforms such as AWS or Google Cloud for real-time predictions.

PROJECT INSIGHTS OF MACHINE LEARNING

Page 16

As technology advances, data analysis will become even more integral.

AI, IoT, and predictive analytics are paving the way for real-time decision-making and deeper insights.

AI based Data Sience stays ahead of the curve to offer cutting-edge solutions.

WHO WILL USE AND BENEFIT FROM THE MACHINE LEARNING TMDB MODEL ?

Streaming platforms and content producers:
Gain insights into factors driving TV show success.
Marketers and advertisers: Improve targeted campaigns based on expected popularity trends.
Viewers and recommendation systems: Enhance personalized recommendations based on predicted success.

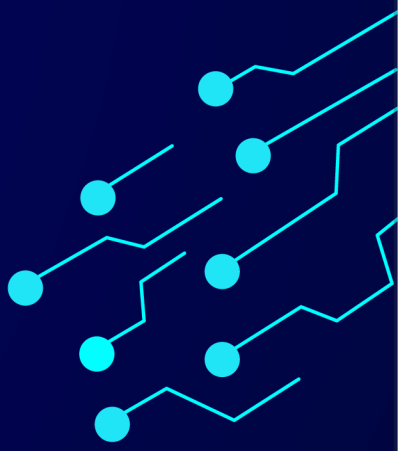
KEY BENEFITS OF ML DATA ANALYSIS

Decision-Making:
Gain actionable insights to inform strategies.

Improved Efficiency:
Streamline processes and identify bottlenecks.

Risk Mitigation:
Predict and prevent potential issues with data-driven foresight.

Competitive Edge:
Stay ahead of trends and market shifts.



THANK YOU

This project aims to provide actionable insights and predictive capabilities that can assist various stakeholders in making data-driven decisions regarding TV show production, marketing, and content recommendations.

Erez Levy ML Project BIU-DS18 - elsz1997@gmail.com