The Movie Database(TMDb) Analysis and Comparison of Predictive Models

A Preprocessing, Analytical and Modeling Case Study using Supervised ML Models

Jack Cox

Introduction: Motivation and Aim

The movie industry has grown immensely over the past few decades generating approximately $10 billion of revenue for the stakeholders annually. A movie's gross revenue prediction is a very important problem in the film industry because it determines all the financial decisions made by producers and investors. Furthermore, a prediction system to assess the success of new movies with the help of the predicted revenue can help the movie producers and directors take proper decisions when making the movie in order to increase the chance of profitability and success.

The goal of this project is to analyze the data and compare models to check the best one through evaluating metrics based on public data for movies extracted from a popular online movie database called The Movie Database (TMDb).

Dataset

The movie dataset used for this project is a database of 5000 movies catalogued by The Movie Database (TMDb). The information available about each movie is its budget, revenue, rating, genre, release date, etc. The dataset is unclean and will take some preprocessing to be ready for evaluation.

Part 1: Data Preprocessing

- Read Dataset
- Define target variables
- Remove null, unwanted, infinite values
- Feature Engineering
    - Encode genre column
    - Create Year column
- Feature Selection
- Feature Transformation

Part 2: Exploratory Data Analysis

- Statistical Analysis
- Correlation Analysis
- Regression Target Analysis
- Year over Year Trend
- Trend Analysis of vote_average
- Genre feature analysis
- General trend analysis

Blueprint of Dashboard

Title

Dropdown

Details from selected movies

Search box