IMDB Box Office Prediction Report

Course: COMP_SCI_7209 Big Data Analysis and Project

Student Name: Kengtian Lu

Student ID: 1840363

Abstract

The aim of this project is to predict movie box office by implementing advanced predictive models to study the correlation between relevant features and box office of movies on IMDB. This report will cover the following 5 areas:
1 Introduction
2 Methodology
3 Experiments
4 Results and findings
5 Conclusions

Introduction

The project is a Kaggle competition for estimating the total global box office receipts of more than 7000 movies listed on iMDB. Cast, crew, narrative keywords, budget, poster, release date, language, production firm, and nation are just a few of the data elements offered.
Exploratory data analysis conducted at the beginning of the project identified several asymmetrically distributed attributes, some of which were category and numerical, as well as some missing values. Several data pre-processing stages were carried out in order to get the data ready for machine learning modelling. Several models were trained during the training and testing phase, the results were compared, and an ideal model was chosen.
The explanation of the model validation is a significant project restriction. According to the guidelines, a "small training set" was created using 80% of the training sample and a "small validation set" using the remaining 20%, which then predicted the box office's final value.

Data Sets:

The training set	train.csv
The test set	test.csv
Interpretation of feature data	Data description.txt

Methodology

Overall, this project was completed by using an iterative workflow method rather than a linear process, since this allowed us to test out various models and techniques until an acceptable degree of accuracy was discovered. For instance, it was discovered often throughout the project that feature data could not be better investigated for analysis, and each time this occurred, it was necessary to return to the data pre-processing procedure to assess what should be done.
Even if the workflow is iterative, when the workflow is broken down into its individual steps, each task may still be thought of as a linear process. The specifics of how each activity will be carried out are then detailed.

1) Exploratory data analysis

Exploratory analysis is about having a thorough understanding of data before it is processed. The purposes of exploratory analysis are as follows.
Quickly describe a data set: number of rows/columns, type of data, data preview.
Analysis of dirty and poor data: analysis of missing data, abnormal data, invalid data types and incorrect values.
Visualisation of data distribution: bar charts, histograms, box plots, etc.
Understanding correlations (relationships) between variables: a basis for subsequent analysis.

2) Data pre-processing

Data quality analysis is an important part of the data preparation process in data mining. It is a prerequisite for data pre-processing and the basis for the validity and accuracy of data mining analysis conclusions. Only credible data can ensure reliable data mining. If the data is not of high quality, the results of data mining will be affected. General data pre-processing includes outlier detection, noisy data detection, processing of missing values and processing of duplicate data.

3) Research models

It is time to create a mapping function that can predict the target variable through the independent variables once the data is clean and ready to be processed. Numerous models were investigated in order to discover this function in order to comprehend the underlying assumptions, constraints, and limitations of each model. The pursuing models were looked into:

3.1 Random Forest

Through the concept of integrated learning, whose fundamental building block is the decision tree, which is really a subset of machine learning known as Ensemble Learning, Random Forest is an algorithm that combines many trees. Since each decision tree is a classifier by default (assumed to be a classification issue), N trees will each produce N classification results for a single input sample. The Bagging concept is used in its most basic form when the random forest combines all categorisation votes and assigns the

category with the highest votes as the final output.
Random forests are a very adaptable and useful technique with a number of properties:
Superior accuracy compared to all existing algorithms;
On huge data sets, it can operate effectively;
Handling input samples with high-dimensional attributes without dimensionality reduction being necessary;
The capacity to assess the significance of each characteristic in relation to the categorisation problem;
During the generation process, a neutral estimate of the internal generation error may be produced;
For the default value problem, positive outcomes are also attained.

3.2 XGBoost Regression
Extreme Gradient Boosting is a version of the boosting process that concentrates on lowering bias, or the model's inaccuracy. To prevent overfitting, it consequently employs a number of rather basic base learners. To lessen the bias of the model, the main concept is to continuously creating new trees, each of which is based on the difference between the previous tree and the goal value.

3.3 Ridge Regression
A regularised regression method is called ridge regression. It is used to improve estimates by adding bias to existing estimates. Additionally, it resolves the multicollinearity issue since ridge regression results in erroneous estimate.

3.4 Lasso Regression
By creating a first-order penalty function, lasso regression creates an improved model. By ultimately locating certain indicators (variables) with coefficients of zero, the explanatory power is great. Lasso regression is effective at handling data with multicollinearity and, like ridge regression, is a biassed estimator. In contrast to ridge regression, which has a minimal possibility of predicting coefficients equal to zero, making it difficult to filter variables.

4) Training and Testing
In order to run the model and partition the dataset in order to forecast the target variables, training is required. 80% of the training samples were used to create a "small training set," while the remaining 20% were used to create a "small validation set." On the basis of the accuracy results from the four models, the best model was then selected to be run on the test set to provide estimates.

5) Evaluation
An ideal algorithm was ultimately discovered. The root mean square error, or RMSE, was the acceptable measure that was selected by all methods.

Experiment

1) Implementation description

Then, more implementation information is provided for each workflow step.

Exploratory data analysis in progress

Importing the dataset and making observations on the dataset to roughly view the dispersion of features

Visualise the feature dispersion

Transformation of the target data, Revenue, using logarithm transformation

Descriptive operations on data

Distributing target variables

Exploring numerical and textual galactic features

Find outliers, missing values

Plotting against univariate variables

Find relevant features

```
train.info()
```

started 02:44:42 2022-10-17, finished in 19ms

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 23 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   id                     3000 non-null   int64
 1   belongs_to_collection  604 non-null    object
 2   budget                 3000 non-null   int64
 3   genres                 2993 non-null   object
 4   homepage               946 non-null    object
 5   imdb_id                3000 non-null   object
 6   original_language      3000 non-null   object
 7   original_title         3000 non-null   object
 8   overview               2992 non-null   object
```

Figure 1 Overview of data

```
train.describe()
```

started 02:44:42 2022-10-17, finished in 22ms

	id	budget	popularity	runtime	revenue
count	3000.000000	3.000000e+03	3000.000000	2998.000000	3.000000e+03
mean	1500.500000	2.253133e+07	8.463274	107.856571	6.672585e+07
std	866.169729	3.702609e+07	12.104000	22.086434	1.375323e+08
min	1.000000	0.000000e+00	0.000001	0.000000	1.000000e+00
25%	750.750000	0.000000e+00	4.018053	94.000000	2.379808e+06
50%	1500.500000	8.000000e+06	7.374861	104.000000	1.680707e+07
75%	2250.250000	2.900000e+07	10.890983	118.000000	6.891920e+07
max	3000.000000	3.800000e+08	294.337037	338.000000	1.519558e+09

Figure 2 Data Describe

Data pre-processing

Locate missing and void values

compensating for missing values

Use the average value to fill in the blanks for data-based attributes.

Normalizing abnormal asymmetric feature distributions

Graphing the outcomes

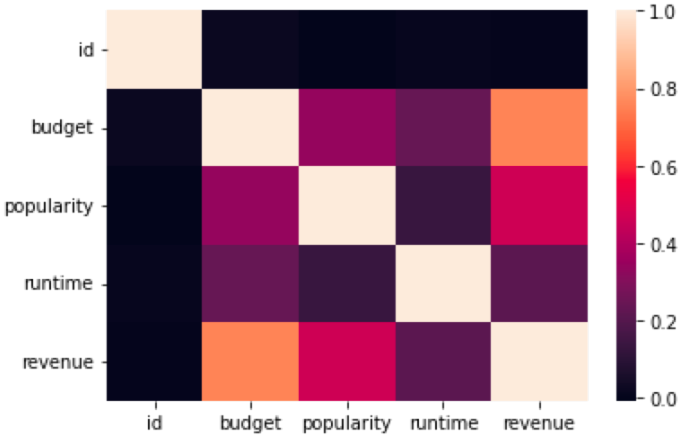

Figure 3 Correlation of key features and objectives

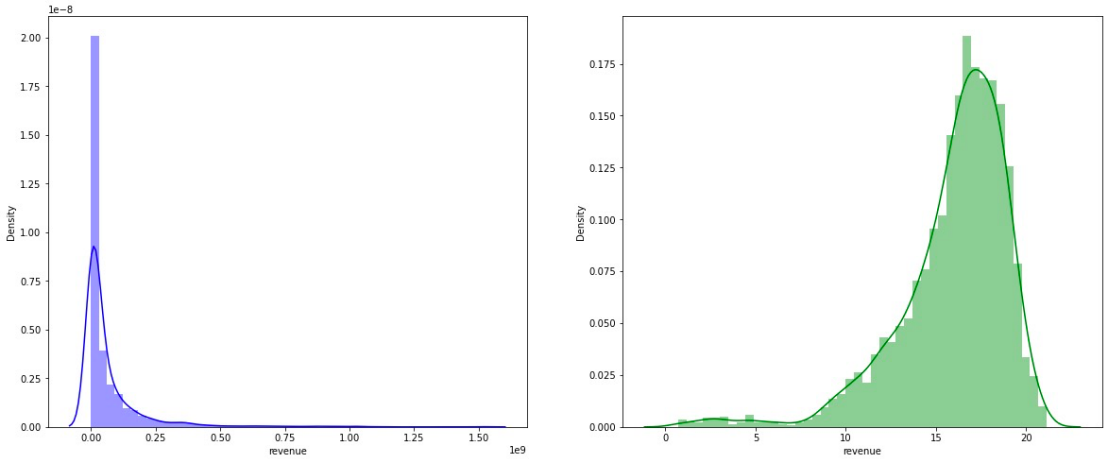

Figure 4 Logarithm Transformation

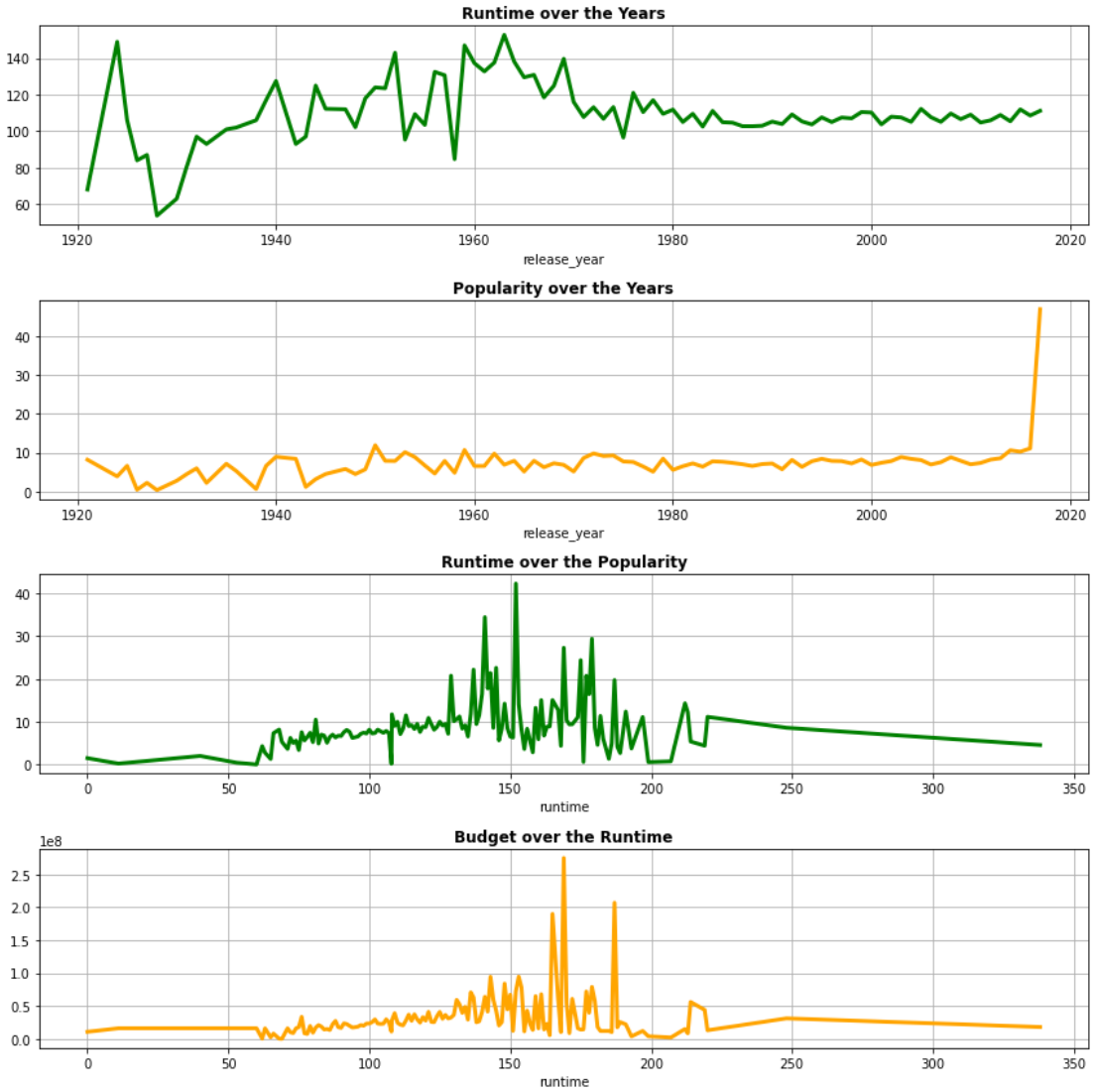

Figure 6 The relevance of key features in addition to the target

Feature Engineering

This part is complex. The project carried too many features that needed to be classified and refined, so the process was modified repeatedly

Converting Json format to nominal format, then nominal format to numerical format

Research Models

This is a challenging step and requires identifying the assumptions and limitations of each model.

Splitting the training and test sets

Define the evaluation function

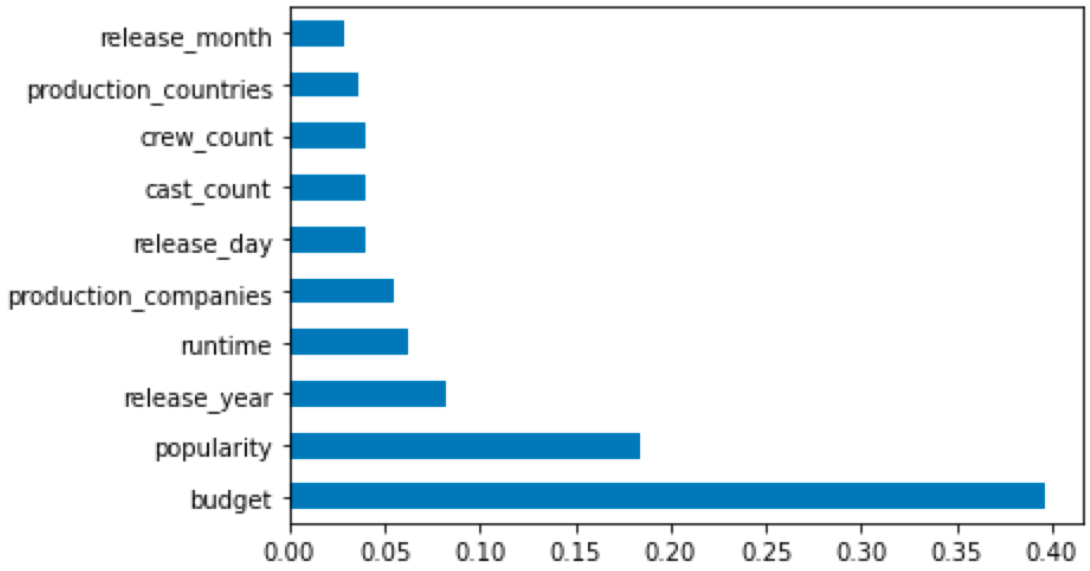

Figure 5 Feature Importance onRandom forest

Evaluation

Evaluation using RMSE

All models have RMSE values between 2 and 3, with the random forest model having the smallest value in the period, so the random forest model can be considered optimal.

Results & Findings

First Experiment

13 characteristics with a high correlation to ‘Revenue’ are selected and evaluated in the model, yielding the following results.

Regression	RMSE	Runtime
Random Forest	**2.091**	**1.03s**
XGBoost	**2.2268**	**186ms**
Ridge	**2.5240**	**9ms**
Lasso	**2.5356**	**9ms**

Second Experiment

6 of the 13 features with weaker relativities are removed for model evaluation and the results are as follows

Regression	RMSE	Runtime
Random Forest	/	/
XGBoost	**2.3982**	**177ms**
Ridge	**2.5658**	**8ms**
Lasso	**2.5725**	**9ms**

Result

The diversity of features can be seen to be beneficial to the assessment results. And random forest regression is the best of the four models, while the random forest took more than 1s, the others were all within 200ms..

Conclusions

There is a case to be made for the importance of data pre-processing given the analyses and lessons learned from this project. For instance, employing logarithmic transformations to address distribution asymmetries can dramatically boost speed. Additionally, eliminating outliers will result in improved outcomes. Another crucial duty is classifying characteristics according to their kind.

It is clear from the experimental part that the data are often not linear in nature. Therefore, linear regression will typically not be effective. Polynomial regression also has issues with overfitting issues. We must choose the degree parameters carefully if we wish to apply this approach.

In conclusion, random forest regression is the best option out of XGB regression, ridge regression, Lasso regression, and random forest regression.

Reference

En.wikipedia.org 2022, *Regression analysis - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/Regression_analysis> [Accessed 16 October 2022].