TMDB Box Office Prediction

Guo Yi, Tang

The University of Adelaide, a1756700@student.adelaide.edu

In this project, we are performing predictions for TDMB box office revenue in the given test file by using different machine learning algorithms. The dataset was collected from Kaggle. We will perform exploratory data analysis at the beginning stage to gain some insight and patterns from the data. Since we know which variable is useful and highly correlated with the target variable which is revenue, we will proceed to the data cleaning and pre-processing stage. We will focus on filling up missing values and drop the columns that are believed to be not highly correlated with the target variable. Hence, we will create a new dataset with a clear information of the movies' release year, date, and month. Before we do the modelling, we need to split the dataset into training and test dataset. Linear, Random Forest, Lasso and Ridge Regression model are used to predict the revenue generated from the given data set name "test" on Kaggle. Nonetheless, we will compare each models' accuracy with RMSE and R2 score. There are also graphs plotted during each stage to provide a better visualization on the results. This report is focus on predicting the revenue from given test set with the best model we used, which is Random Forest Regression model. Final submission is submitted on Kaggle.

Additional Keywords and Phrases: Machine Learning, Linear, Ridge, Lasso, Random Forest

ACM Reference Format:

Guo Yi, Tang. 2021. TMDB Box Office Prediction

1 INTRODUCTION

Movies are noticeably one of the most influential sectors in today's world. Part of the reason is because the positive impacts exceed the negatives that could possibly change a person's attitude. Thus, the numbers of audience have increased significantly over the years. This led to the film studios and the related stakeholders' incentives to generate a maximum revenue from the movies. One of the useful ways is implement a prediction method that can provide a predicted revenue on that movie based on the numerical and categorical information. The main aim for this project is to predict the TMDB overall worldwide box office revenue on given "train.csv" with 4398 movies. The dataset was taken from Kaggle to perform training and testing process. In the dataset, there are 10 parameters that could probably affect the revenue generated from 7398 movies. Thus, we will split the testing and training set from the given file named "train.csv" into 600 and 2400 movies. The goal is to promote a model with the highest accuracy in predicting the revenue generated.

2 METHODOLOGY

The next subsections provide information on the approach for data cleaning, preprocessing and the model used for this paper.

2.1 Data Cleaning and Pre-processing

Since our target variable in this dataset is the revenue generated from TDMB movies, we plotted the histogram for the Normal Distribution Revenue and the Log Distribution Revenue to observe its skewness.

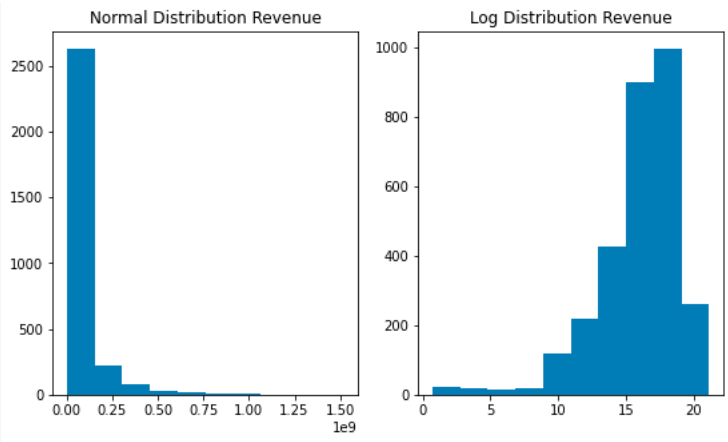


Figure 1: Histogram for Revenue (normal and log)

The plan is to use the log transform value to perform the prediction and the exponential method from numpy of the log for test.csv prediction for submission at the end. Firstly, there are some missing values and unreadable input, hence we need to clean and handle them. The cleaning process is performed separately on train.csv and test.csv considered that we only use train.csv for most of the time. In this project, we handle missing values by replacing them with 0. On the other hand, we chose to drop the non-numerical variables for training and testing purposes. One of the features release dates was being handled by separating them into three different columns, which are release year, release month and release date.

2.2 Data Modelling

2.2.1 Linear Regression Model

Linear regression model is known as one of the simplest supervised machine learning algorithms that a data scientist uses for predicting modelling. The model we used was ordinary least squares (OLS). It is a linear least squares method to estimate the unknown parameters in linear regression model that would find the “best fit line”. This means the smaller the sum of squared errors, the higher the prediction accuracy. The errors is the difference between the predicted value of y subtract by the actual value of y.

$$\hat{y} = \beta 0 + \beta 1 x$$

Figure 2: Linear regression (OLS)

In our case, y is the revenue, and it is assumed to have a linear relationship with all the variables which are x. Based on the given data from the input variables, β can be estimated and it is selected to minimize the sum of squared errors. Thus, find the best fits line in the training set and predict the revenue based on the line.

2.2.2 Lasso Regression Model

Lasso regression is a part of the linear regression that implement shrinkage. This means that the data values are shrunk close to a central point such as mean. It provides a simple, scarce models as it can use with more than one parameter. The reason why this model was chosen is because it is suitable for models showing high levels of multicollinearity as we could use it to automatically select or eliminate some parameters. Lasso Regression's mechanism is like OLS.

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Figure 3: Lasso Regression Model formula

Tuning parameter (λ) is the symbol for the amount of shrinkage. When λ is 0 that means no parameter is eliminated which happen to be equal to linear regression. However, as λ increases, coefficients will be set to 0 and be eliminated eventually, thus bias increases. On the other hand, as λ decreases, the variance will increase.

2.2.3 Ridge Regression Model

Ridge Regression is similar with Lasso regression and OLS. However, the disadvantage is it will reduce the magnitude of coefficients.

For OLS:

$$\underline{\hat{\mathbf{B}}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underline{\mathbf{Y}}$$

Figure 4: Linear Regression Model formula (beta)

For Ridge:

$$\underline{\tilde{\mathbf{B}}} = \left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}'\underline{\mathbf{Y}}$$

Figure 5: Ridge Regression Model formula (beta)

K is known as ridge parameter which take the identity matrix to the cross-product matrix and form a new matrix (X`X + kI). However, choosing a value for k is not easy, hence it is not commonly used as much as OLS.

2.2.4 Random Forest Regression Model

Random Forest Regression model is a supervised learning algorithm that combines predictions from multiple machine learning algorithms to generate a higher accuracy prediction.

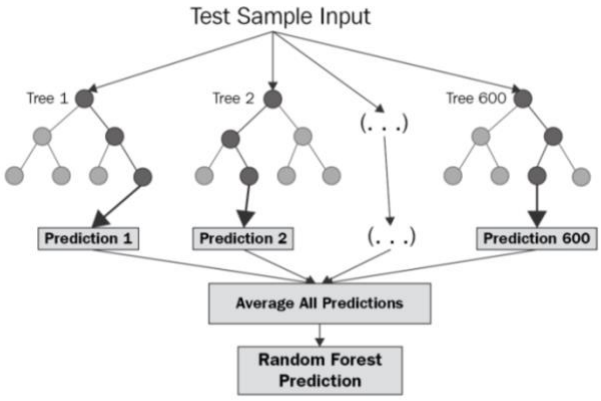


Figure 6: Random Forest Regression Model

The way it operates is by formulating multiple decision trees during training process and generate the mean of the classes as the prediction. It begins with selecting a random k data point from training set, build a decision tree and choose the number of trees we want to build. It is important to choose the number of trees for the model to prevent error such as overfitting. We have used this random forest for our last prediction model.

3 IMPLEMENTATION

The next subsections provide information on the steps and results derived from exploratory data analysis, preprocessing and the model used for this project.

3.1 Exploratory Data Analysis (EDA)

In this step, we perform the Exploratory Data Analysis (EDA) to gain insights from the data. This is because it will help us to explore the patterns in the dataset. Firstly, we look at the shape of train and test dataset. We observed that there is no revenue column in test dataset, which is expected as we will only use it at the later stage. Hence, we printed out the head of the train dataset and plot a scatter graph on revenue and budget. To do so, we need to import libraries such as pandas, numpy, matplotlib, seaborn and wordcloud. From figure 7, we can see it does not show a clear trend between budget and revenue, but it does indicate that high budget usually come with higher revenue.

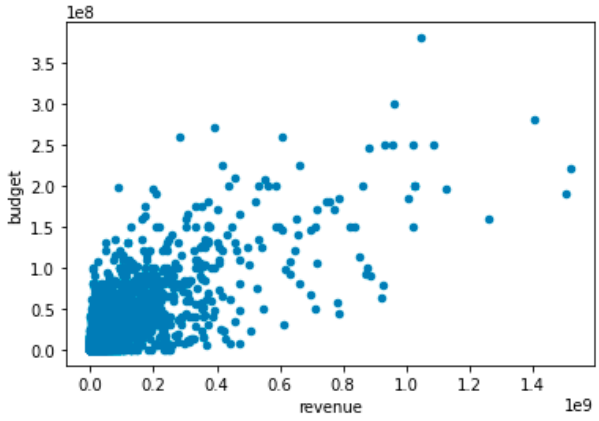


Figure 7: scatter plot for budget vs revenue

We explored the popularity of the movies and plotted a barplot for top 15 most popular movies and it shows that “Wonder Woman” and “Beauty and the Beast” are ranked first and second for most popular movies.

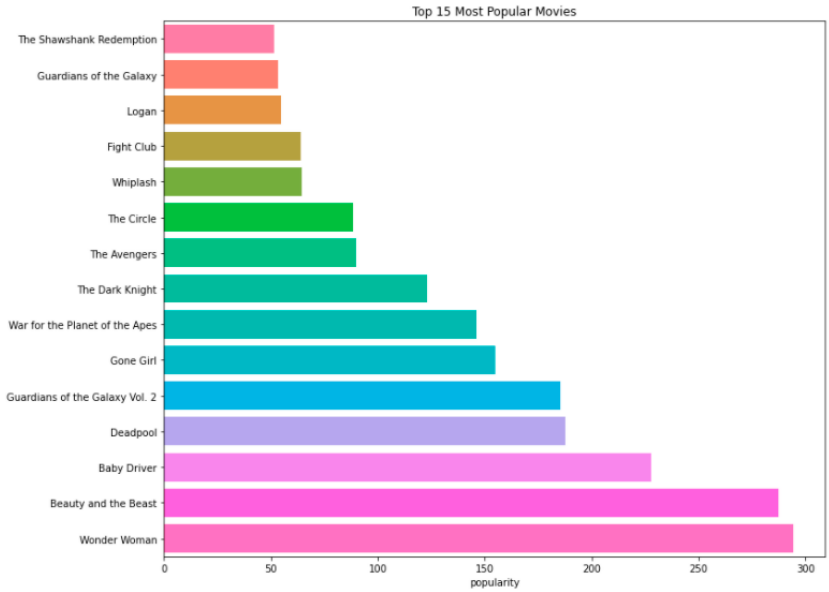


Figure 8: top 15 most popular movies

When we looked at top 15 high budget movies, we noticed that “Pirates on the Caribbean: On Stranger Tides” spent more than $350million to produce.

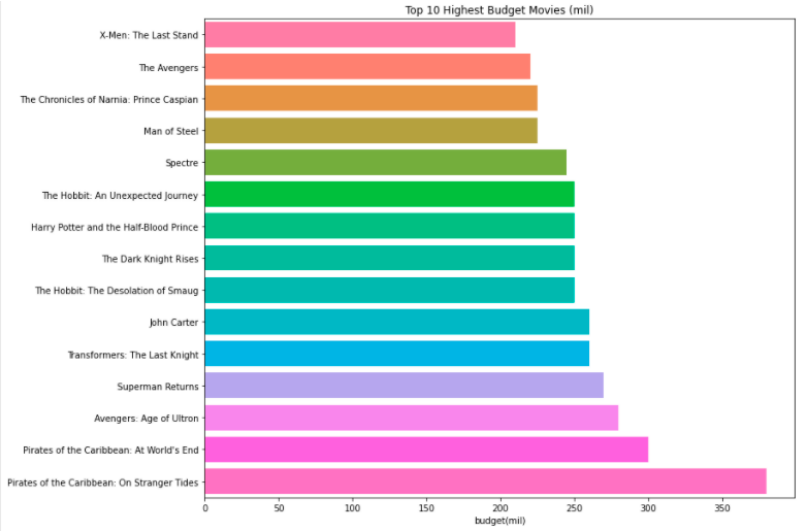
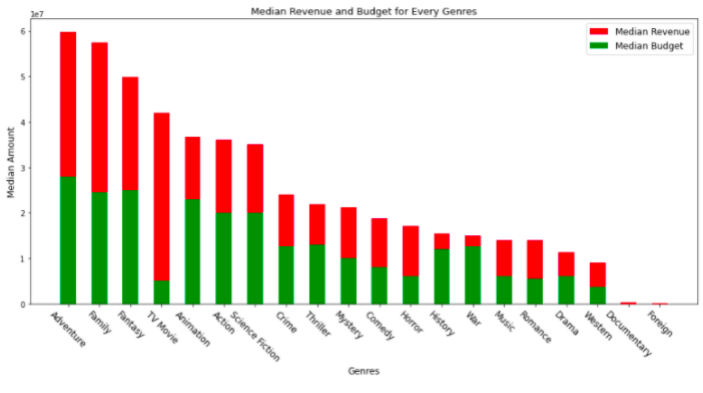
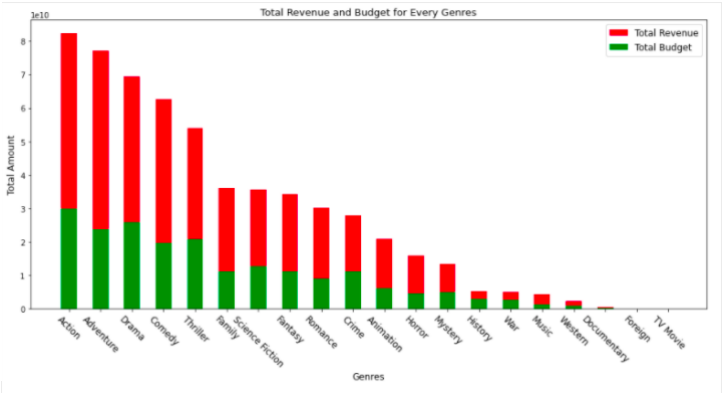


Figure 9: top 15 high budget movies

The variable ‘genres’ was used for visualization where we used pandas series function to hold information in genres column and create a new dataframe. The dataframe consists of the total movies, total revenue, total budget, median budget, and median revenue. Figure 10 shows us which genre generated more revenue over the years. Based on the graph for total revenue and budget for every genre, we can see action, adventure and drama are the top 3 genres that generated the most revenue and high revenue does come with high budget too. Thus, we noticed median revenue and budget graph indicates that TV movie had generate a high median revenue, however, there was only one movie made in this genre. Therefore, it should not be considered.





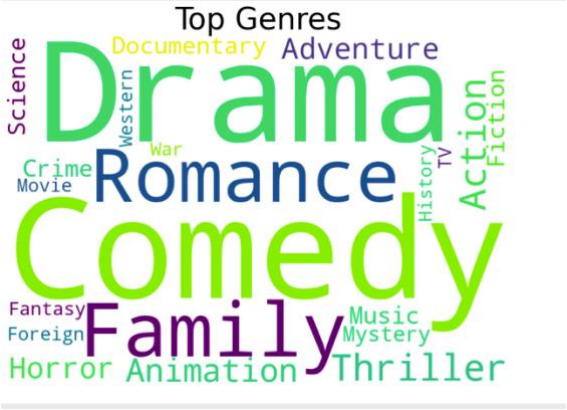


Figure 10: Top genres, budget, revenue and median budget and revenue

In terms of the relationship between revenue and runtime, figure 13 shows us that most of the movie runtime falls within 140 minutes.

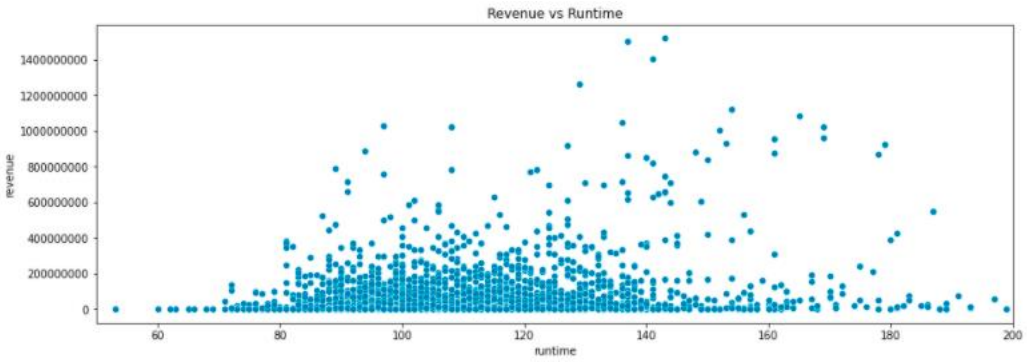


Figure 11: revenue vs runtime

From the graphs below, we can see 2017 has the highest total revenue in one year. The revenue for movie has increased over the years with the highest growth in 2017. This indicates that movie will generate high revenue when the release month is on midyear which are May, June, and July, as well as end of the year in November and December. It seems like movie did better when it was released on 11th, 16th, 30th of the month. On the other hand, movie tends to perform worse when it was released at the beginning of the month. 31st should not be taken into consideration because it does not happen in every month.

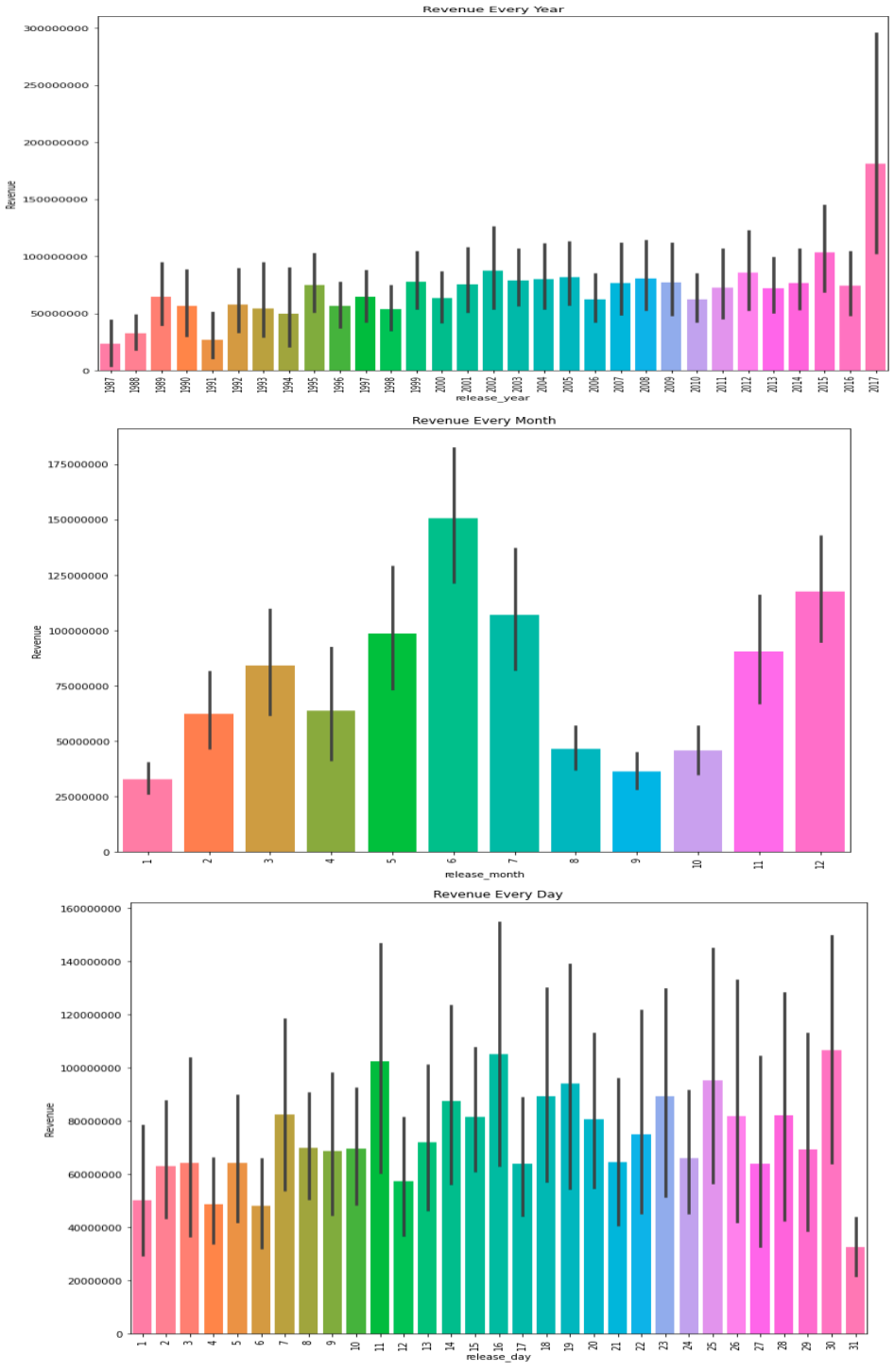


Figure 12: Revenue every year, month, and day

3.2 Data Pre-Processing

Pre-processing is a process which we handle noisy or missing values in the datasets. There are approximately 23 columns in the dataset, and it is necessary to perform cleaning and filling those null values before we perform the modelling. In our case, we will fill all the null values with zero. Dates also being handles by separating them into different columns to gain a clearer visualization. Moreover, there are columns that are not needed for modelling purposes such as belongs to collection, homepage, overview, and others. Thus, we will drop those columns and create a new dataset name “train”. This is supported by the correlation heatmap that shows the variables that are highly correlated with the revenue.

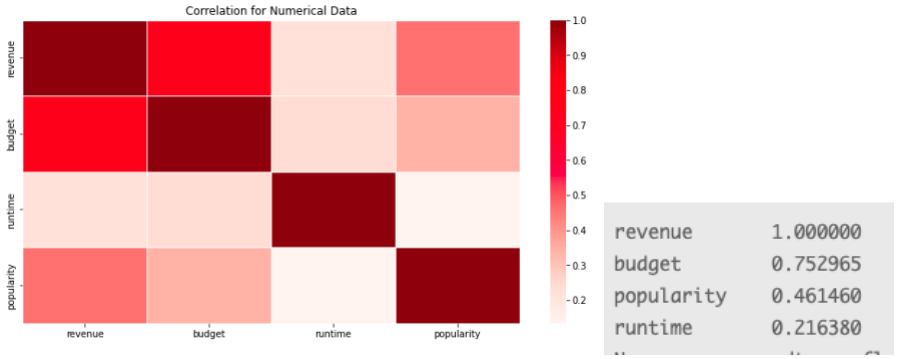


Figure 13: correlation for numerical data

3.3 Data Modeling

In this step, we split train dataset into 70% for training and 30% for testing. We proceed with the application of different regression models to make comparison on the accuracy. First we applied it with linear regression model and calculated the RMSE. In this case, the parameter is set to be default, unlike Ridge and Lasso that requires different levels of regularization which results a penalty. Hence, we also applied Ridge and Lasso regression model as it will automate the variable selections and eliminations. Lastly, we perform a training in random forest regression model and look at the RMSE. Lastly, we pick the random forest regression model as it has the lowest RMSE to perform a prediction on test set given on Kaggle and make the final submission on Kaggle competition.

4 EVALUATION

Root Mean Square Error (RMSE) is used to evaluate the accuracy of the models for this project. This means only the one with the lowest RMSE will be selected. Besides, R2 score is used to evaluate the models’ efficiency as well. The following table shows the RMSE and R2 score for each model.

Table 1: Evaluation Metrics

Regression Model	RMSE	R2 test score	R2 train score
Linear	2.800770479620727	0.2856665580072234	0.2548825929611215
Random forest	2.3429321143989013	0.8153719260024755	0.4785783264200548
Lasso	2.862331748660421	0.23955045175507317	0.22176706876261842
Ridge	2.7983341172935283	0.25617836953711515	0.28347078795332326

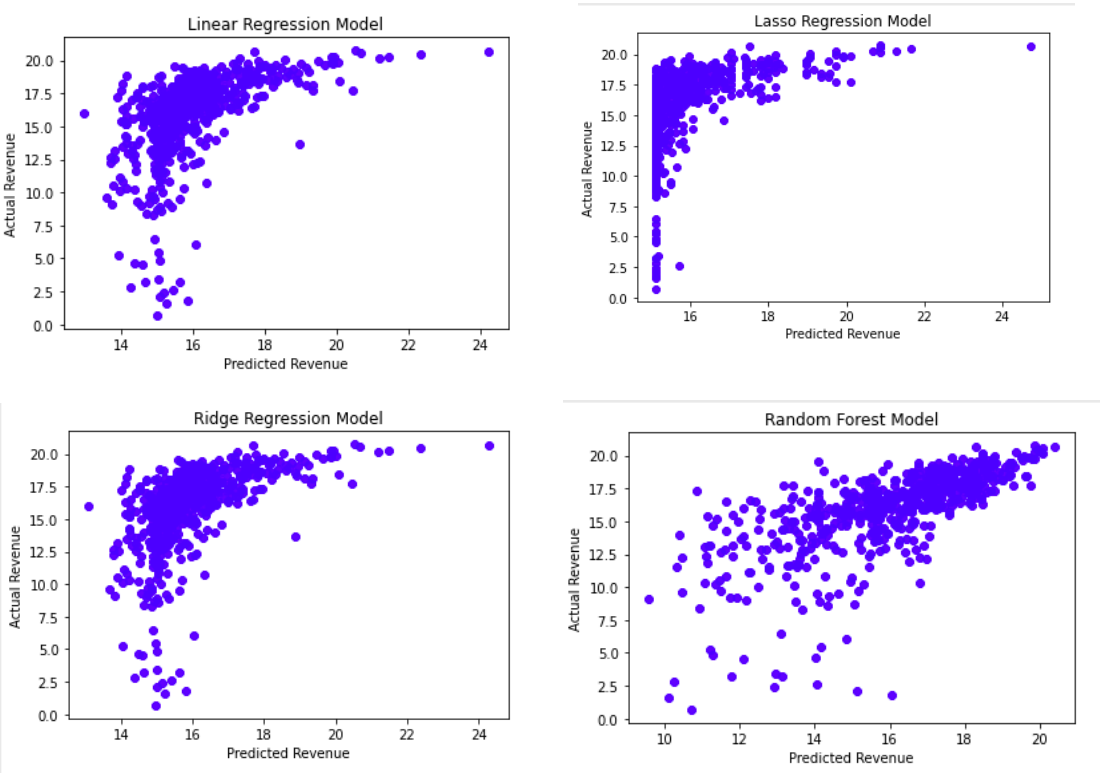


Figure 14: output evaluation

From the above Evaluation output and table, we can observe that random forest regression model has the lowest RMSE and the highest R2 score which makes it the best model among other models. However, the worst model above is Lasso Regression model. There are two graphs to give us some insight on how the actual and predicted points were distributed when we compare revenue with budget and runtime. Lastly, we will use random forest regression model to perform prediction on test.csv file. ID 4590 and 6139 are predicted to have high revenue generated.

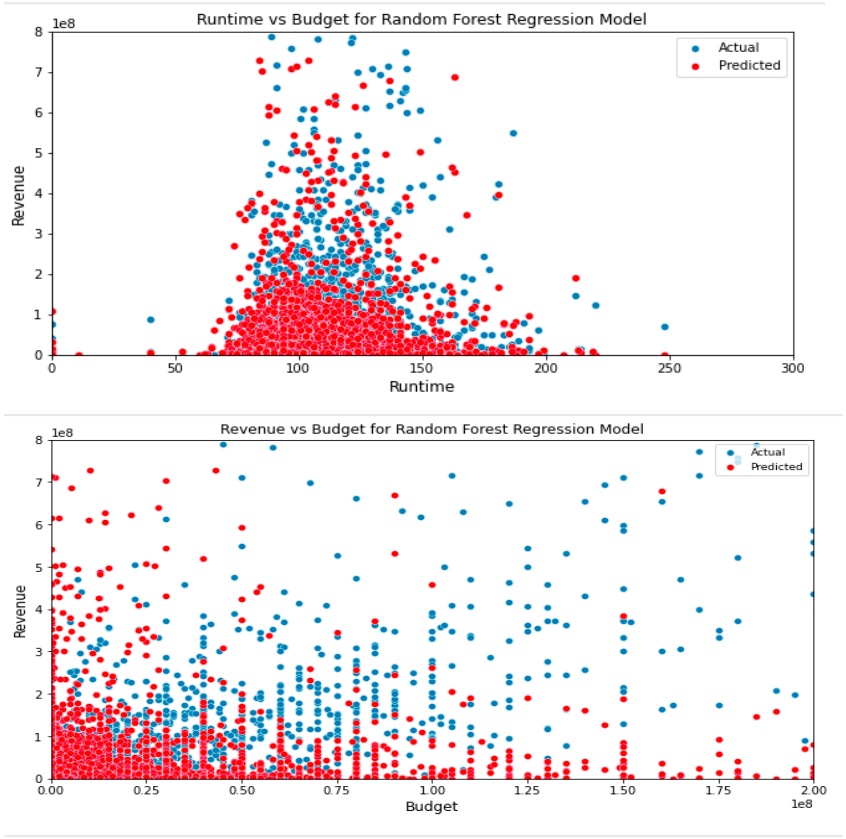


Figure 14: output evaluation