imdb

June 25, 2018

0.1 Capstone Project

Predicting the metascore of a movie based on the given features.

* Sai Kiran Rebba

June 25th,2018.

0.2 I.Definition:

Project Overview:

- Now a days everyone is watching movies and give rating or review for that whether it is good or bad.With this information I have an idea to predict the score of the particular movie based on the rating given by the audience and voting.
- Hence an analysis is made by Preetish panda, which was posted in Linkedin. Source :https://www.linkedin.com/pulse/analyzing-imdb-movie-dataset-preetish-panda
- By the sheer exploration of the data it is possible to develop certain score based on the related charactersitics or features which can be used to give information about that movie whether it is good or bad.
- Hence, in this project I developed a model that can predict the metascore by using rating of a particular movie and votes polled to that movie and revenue i.e howmuch money invested on that movie and genres i.e comedy,drama,horror,adventures,sci_fi,romance etc and actors

etc.

Problem Statement:

- The goal is to develop a model that can predict the 'Metascore ',of a movie based on the important features in the data set.
- The tasks involved are :-
- Download and preprocess the imdb data.
- Train a BenchMark Model and record it's performance.

- Then three supervised learning models were trained using the training data and a comparision is done based on the performance metric and decided which among the three is the best model.
- The best model thus selected is optimized using GridSearchCV technique.
- The Optimized model is then compared with the Benchmark model and deciding which is the best for the given data.
- Then the best model is validated aganist unseen data and documenting the intuition.
- The final model can be applied to determine the metascore of a particular movie with the givrn features.

Metric:

- The current problem is a Regression task, since it takes certain features as inputs and attempts to find a score that helps an individual to get an idea about particular movie.
- Hence, Coefficient Of Determination is considered as the performance metric that can be applied to compare the performances of the scores obtained from the BenchMark and the Optimal Model considered.
- The CoEfficient Of Determination(Rˆ2) is the key output of the Regression Analysis.It can be defined as the proportion of the varaince in the dependent variable that is predictable from the independent variable.

It's values ranges from 0 to 1, and the results are given intuition by, if:-

The value of R^2 -> 0 , indicates that the model is a worst fit to the given data.
The value of R^2 -> 1 , indicates that the model is the best fit to the given data.
The value of R^2 in between 0 and 1 -> indicates that the respective variability exhibited by th The formula for CoEfficient Of Determination(R^2) is given by :- fed29779d54adeccdec58f0894870c6

From the above formula SSreg is called Sum of Squares of Residuals, also called the Residual Sum

And the SStot is called the Total Number of Squares. Source :-https://en.wikipedia.org/wiki/Tota

0.3 Loading the required data -> Data Acquisition:

- Data acquisition is the initial step in machine learning.Here the data is acquired from kaggle.com.
- Link to the dataset: https://www.kaggle.com/nielspace/fork-of-imdb-data-set/data

0.4 Data Preprocessing:

- Data preprocessing means remove the unnnecessary data fro our dataset.
- Cleaning the data like the rows which are empty are being removed from the datasset,if have been persist in the dataset then they abnormal results.

- We can observe that the features, 'give us with the quantitative information about each data point.

- The target varaible, 'Chance of Admit', will be the variable we intend to predict. These are stored in the varaibles features and scores respectively

- Here I think movie title ,Description and its rank are not play a key role in our prediction .So, I remove those features from out data.

0.5 Data Exploration:

- Data Exploration is a crucial step in the process of Machine Learning.It helps us to understand the patterns and available features in a data set from which we can determine the sort of actions that we can perform for further analysis.

- Data Exploration gives an intuition that a cursory investigation of the data-set is necessary for familiarizing ourself with the data through an explorative process and is a fundamental practice to help you better understand and justify your results.
- Since, the main goal of this project is construct a working model that has the capability of predicting the 'Metascore', we will need to separate the dataset into features and the target variable.

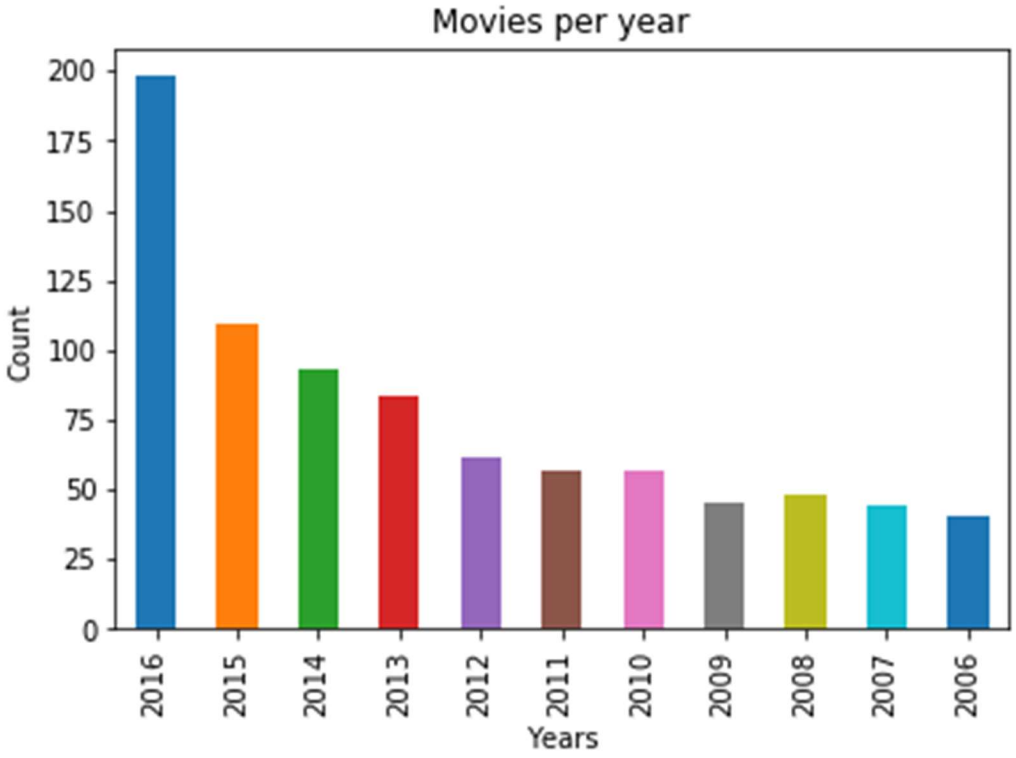
Feature set Exploration:

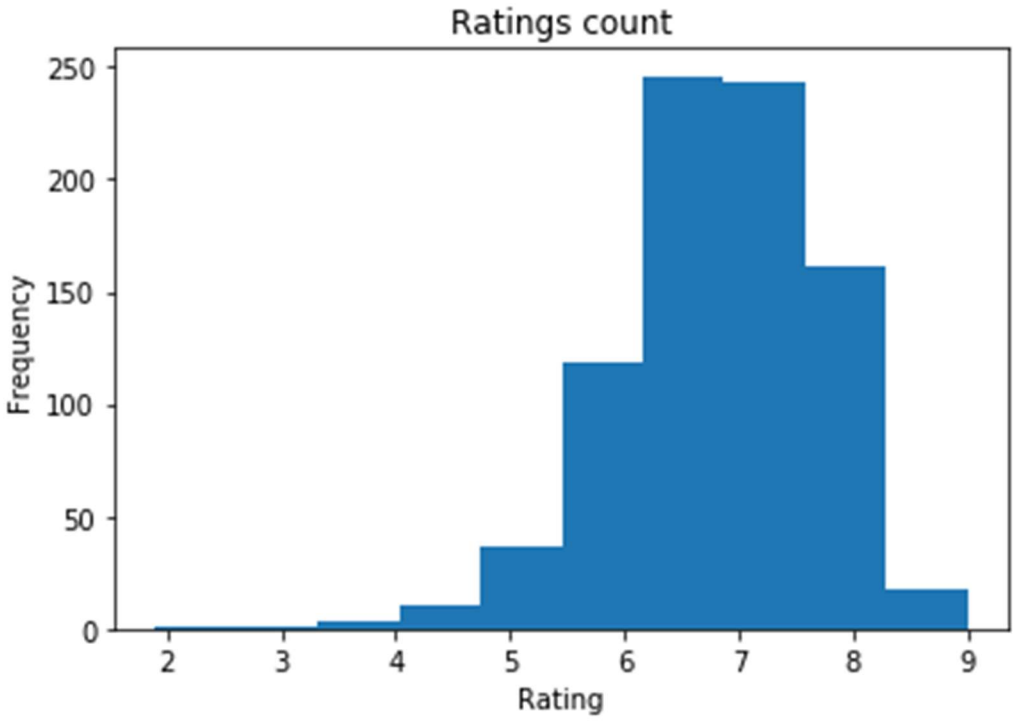
- Genre -> It tells us about the type of the movie.for example Action,Adventure,Horror,Thriller,Fantasy,Sci_fi etc.
- Director -> Director of that movie.
- Actors -> Actors which are acted in that movie.
- Year -> In which year the movie was released.
- Runtimeinminutes -> Duration of the movie in minutes like 120 minutes,180 minutes etc.
- Rating -> Rating of that movie ranges from 0 to 10 in float values.
- Votes -> Number of votes polled for that movies.
- Revenueinmillions -> Budget of that movie in million.

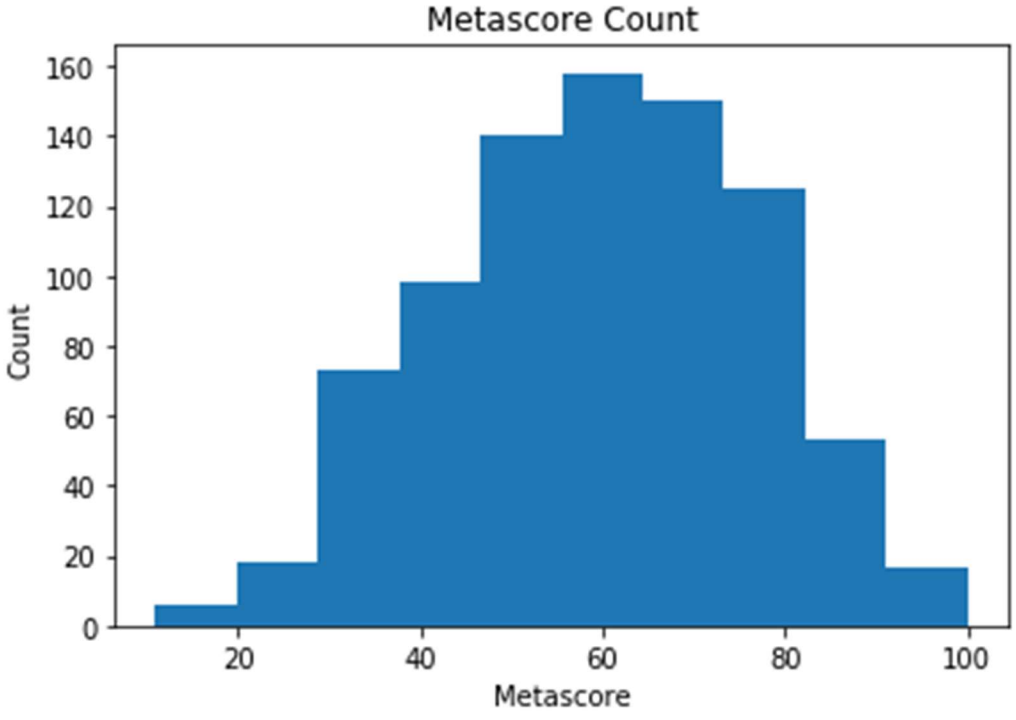
Exploratory Visualisation:

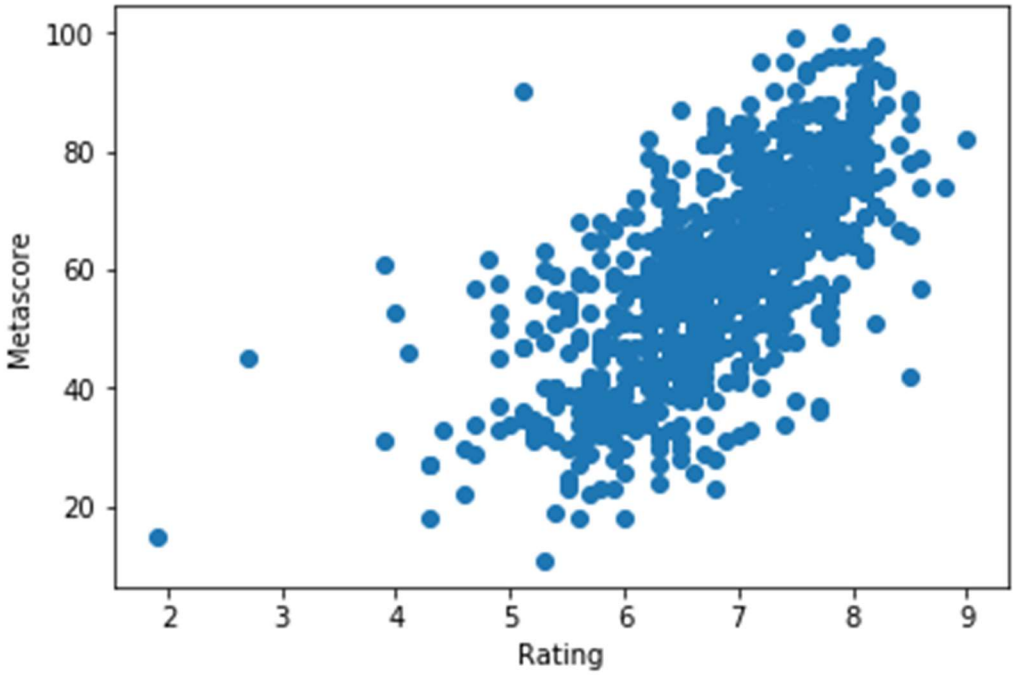
- Exploratory Visualization can be defined as an approach for analyzing data sets to summarize their important characteristics, often by the application of visual methods.

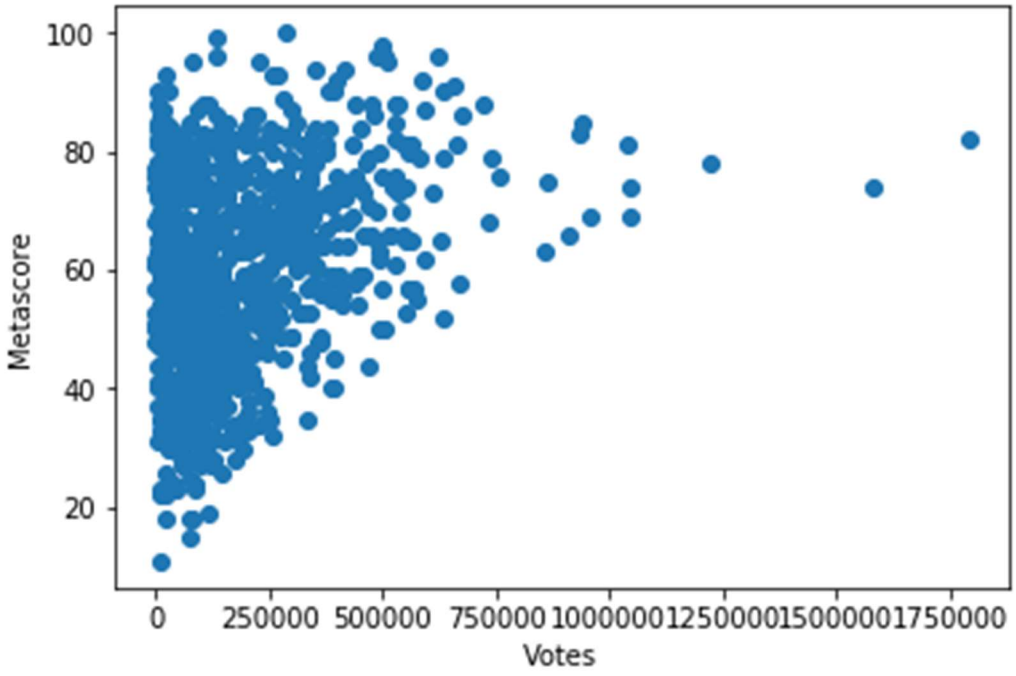
- The Primary theme of Exploratory Visualization is for observing what the data can give us an intuition far beyond the conventional modeling or hypothesis testing tasks.



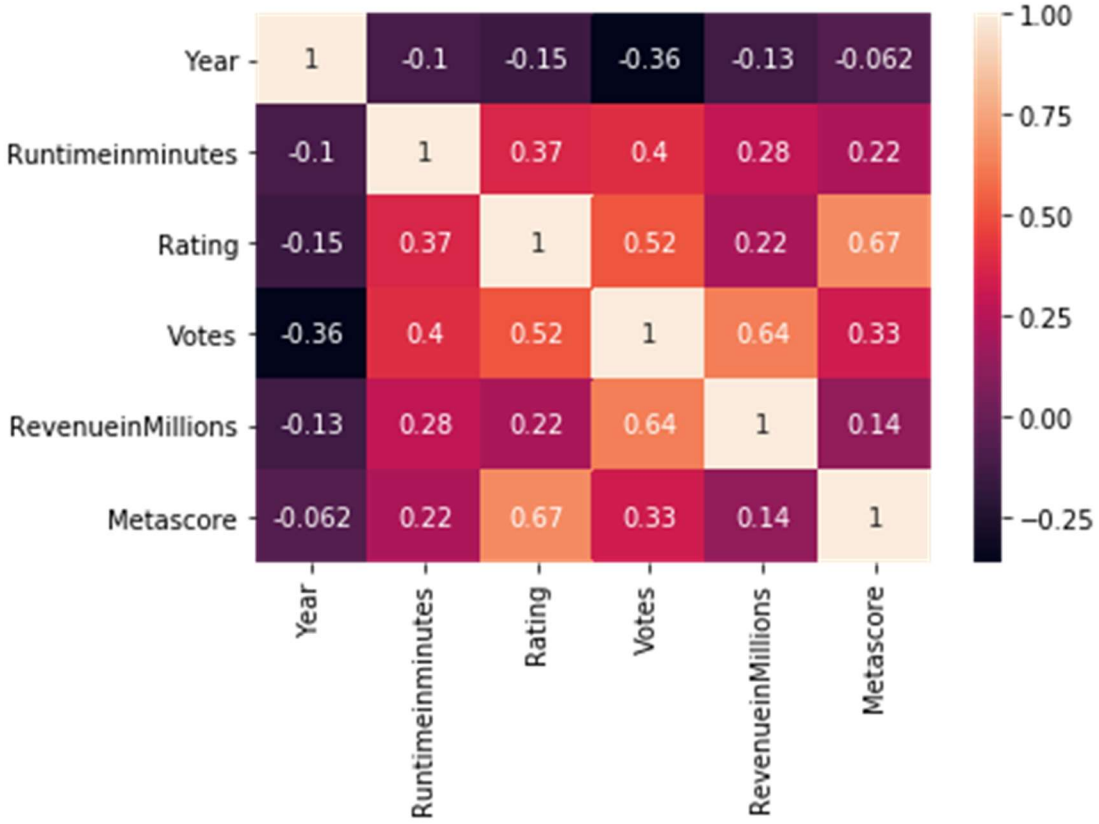

Ratings count
Frequency
Rating
0
50
100
150
200
250
2
3
4
5
6
7
8
9


Metascore Count
Count
Metascore
0
20
40
60
80
100
120
140
160
20
40
60
80
100


Metascore
100
80
60
40
20
2
3
4
5
6
7
8
9
Rating


Metascore
100
80
60
40
20
0
250000
500000
750000
1000000
1250000
1500000
1750000
Votes

Heat map which shows us the correlation between features.



0.6 ALGORITHMS AND TECHNIQUES

By observing the problem, it is quite evident that it is a 'Regression' Problem.

It is important to understand the intuition behind the consideration of specific model, since it has to generate an optimal possibility of results that can improve the model's performance on a sample of new data.

Hence, taking the performance of the model into consideration, I chose three Supervised Machine Learning Algorithms that can be better compatible for the data being available.

They are :-

Ensemble Methods - i) Random Forests, ii)GradientBoostingRegressor iii)Lasso

Ensemble Methods - Random Forests :- Random forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes in case of the classification tasks or mean prediction for the regression tasks of the individual trees.

The Random forests have a stroke of brilliance when a performance optimization happens to enhance precision of the model, or vice versa.Tuning down the fraction of features that is considered at any given node can let you easily work on datasets with thousands of features.

Since Random Forests perform well on almost every machine learning problem and they also show less overfit behavior when compared to Decision Trees.

GradientBoosstingRegressor: Gradient Boosting Regresor is a ensemble algorithm for both classification and regression.GBTs build trees one at a time, where each new tree helps to correct errors made by previously trained tree. With each tree added, the model becomes even more expressive. There are typically three parameters - number of trees, depth of trees and learning rate, and the each tree built is generally shallow.

Although it may seem GBDTs are better than random forests, GBDTs are prone to overfitting, however there are strategies to overcome same and build more generalized trees using a combination of parameters like learning rate (shrinkage) and depth of tree. Generally the two parameters are kept on the lower side to allow for slow learning and better generalization.

Lasso: It is one of the regresion technique.It uses regularisation.LASSO stands for Least Absolute Shrinkage and Selection Operator. I know it doesn't give much of an idea but there are 2 key words here – 'absolute' and 'selection'. https://www.analyticsvidhya.com/blog/2016/01/completetutorial-ridge-lasso-regression-python/#four

0.7 Data Preprocessing:

Feature Tranformation :-

- Feature transformation is an important concept in the Data Preprocessing phase which involves the tranformation of relevent numerical data into the range of 0 to 1 by using min_max_scaler which is available in sklearn.preprocessing.
- We also need to transform the categorial data into numerical data by using get_dummies().

Grid Search CV:- Grid Search technique is an approach for tuning the parameters that are used to build and evaluate a model based on the individual combination of parameters of an algorithm specified in a grid.

Grid Search rigorously checks for the combination of hyper-parameters in order to find the best model.

Finally it's easy to figure out the combination that has high cross validation accuracy with respect to the parameters considered that eventually contribute to the optimization of the learning algorithm

By the next approach, I will chose the best of these of three models, that can be an optimized model for the data set.

I will then use the performance metric (r2_score) and compare the three potential based on their scores, the model which has the best r2_score will be eventually considered for further analysis.

Eventually I'll optimize the selected model by 'GridSearchCV' and evaluate the model by comparing the final r2_score of the optimized model and the benchmark model.

BENCHMARK MODEL

DEFINITION :- A Bench Mark Model can be defined as a standard model that already shows a better performance on a given data.The factors on which our results or the solution is tested, are mostly going to be the amount of training/testing data, and then we compare your solution with that of the benchmarked solution obviously based on a performance metric(here r2_score).

The main theme here is to understand which model works delivers the best solution than their existing solution.So, it can be achieved by sheer analysis, implementing standard algorithms and observance and coming to the conclusion that the model shows good solutions or results than the benchmark model's solution.

Since, the problem is a 'Regression' task , I'm implementing a 'Linear Regression' model as my BenchMark Model.

IMPLEMENTATION In the further section of the project, I'll intuitively select the best out of the three models that I considered for the current problem by using the performance metric(r2_score), based on the results generated, I'll decide best of the three models, which is optimal for the given problem.

INITIAL MODEL EVALUATION :- In this section, I'll clearly show the coding implementation of the three supervised learning models

Import the necessary libraries and intialize the models and store them in respective variables.
And finally comparing the r2_scores of the three learning models and decide which one is the best.

CHOOSING THE BEST MODEL: Since, it is obvious that the model which has best high r2_score when compared to the other models can be termed as the best optimal model for the current problem, as the fact that if :-

r2_score is 0 -> it indicates that the model is a worst fit to the given data. r2_score is 1 -> it indicates that the model is the best fit to the given data. r2_score in between 0 and 1 -> indicates that the respective variability exhibited by the target varaible. The values can be tabulated as follows :-

0.14 Chossing the best Model

The major analysis that can be obtained from the above tabulated data is that the 'Gradient Boosting Regressor Model' is the best model among the three, since it exhibits a high score of 0.59i.e, the target variable accounted for about 59% of the variance.

And Random Forest Regressor shows a score of '0.53' which indicates that 53% is accounted for the target variable..

Lasso shows a decent score of '0.54' which indicates that 54% is accounted for the target variable.

REFINEMENT In this section of the project, the model('Gradient Boosting Regressor Model') is optimized by the application of 'GridSearchCV' technique for fine tuning the parameters for the final model thus choosen and later calculating the performance metric(r2_score) of the optimized model.

MODEL TUNING Here, I will find the implementaion of GridSearchCV by intially importing the libraries sklearn.grid_search.GridSearchCV and sklearn.metrics.make_scorer.

1. Initialize the regressor('Gradient Boosting Regressor Model') and store it in the varaible 'reg'.
2. Creating a dictionary of parameters, in the variable parameters.
3. Using make_scorer to create a r2_score scoring object. -> score
4. Perform Grid Search on the Regressor lgr_grid using the 'scorer', and store it in grid_obj.
5. Fit the Grid Search Object to the training data (X_train, y_train) and store it in the grid_fit.

0.15 FINAL MODEL EVALUATION

In this part of the project I'll demonstrate the comparison of the performances between the BenchMark Model('Linear Regressor') and the Opmital Model('Gradient Boosting Regressor') based on their performance metrics(r2_score) in a tabular form.

0.16 MODEL VALIDATION

In this part of the project, I'll demonstrate the performance of the Best Model for the given regression task i.e., The Optimal Model aganist unseen data.

Task - Predicting The Metascore of a given movie

Reckon that the data of the three students are as follows :-

0.17 Justification:

Here I take the unseen data from test set because they are also unseen by the user.Initially we divide the dataset into train and test sets.So,I think this is fine. By observing the results above we come to conclusion that:

Finals results are approximately equal to the actual reulsts.But it mainly depends on the rating,votes,Revenueinmillions.Already we visualised the heatmap which shows us the correlation between metascore and these features.

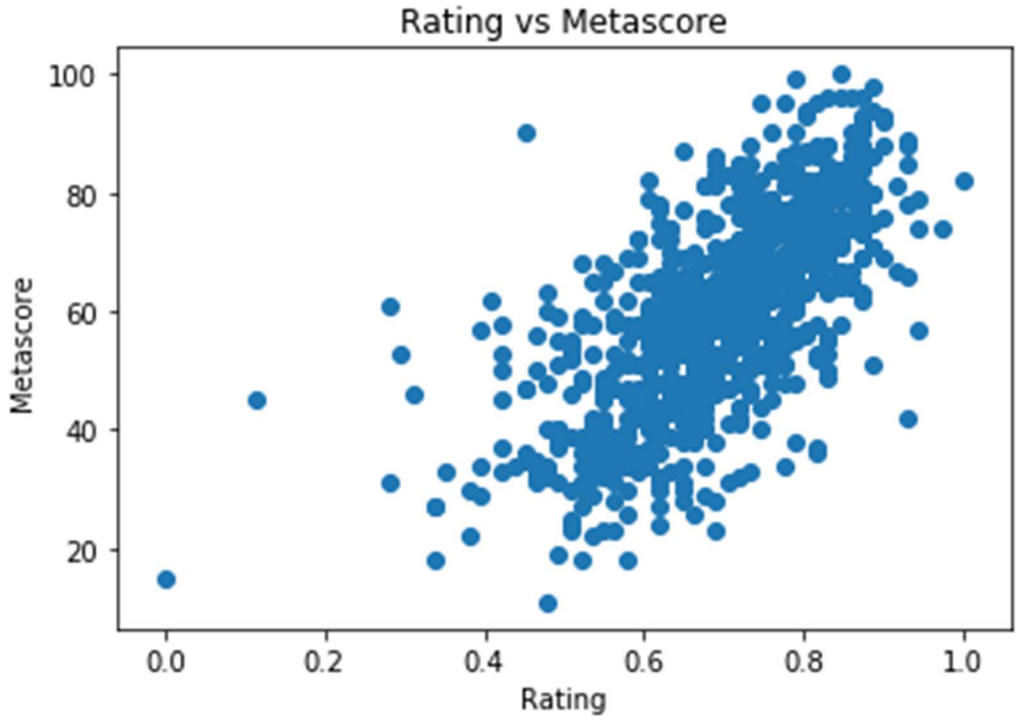
i)Metascore is 56.77.Here runtime is 0.504 and rating is 0.67 but votes ate 0.05 which is very low.So,the metascore is also low.

ii)Metascore is 41.29.Here runtime is 0.355 and rating is 0.577.So it is giving the low value.

iii)Metascore is 79.69.Here rating is 0.9 which is nearly equal to 1 and votes also nearly half of the max value i.re 0.4 .So,it is giving the highest value.

0.18 Free Form Visualisation:

- The most important feature for this dataset is 'Rating'.So,I plot a scatterplot between RRating and Metascore.
- The correlation between rating and etascore in 0.67.
- If the Rating increases then Metascore also increases.So,it plays a key role in this dataset.



0.19 Reflection:

- The process used for this project can be summarized using the following steps:-
- A Common Problem on Graduate Admissions is found in “Kaggle” and the data set related to it is acquired from the Public Domain.
- The Dataset was downloaded and loaded for the current project and necessary statistics were calculated.
- The Data Set is Explored with scatter-plots to illustrate the correlation of the input features with respect to the target variable.
- The Data Set is preprocessed using the technique “Feature Selection” and irrelevant attributes are removed from the dataset.
- Feature transfromation is vary important for this project.In this project first of all we have to convert the categorial data into numerical data by using get_dummies.
- A function is designed that calculates the performance metric r2_score and returns that score.
- The data is splitted into 75% training and 25% testing data.
- A BenchMark was created for the Regressor i.e, here Linear Regressor acts as a BenchMark Model and is trained using the training data.
- Then three supervised learning models were trained using the training data and a comparision is done based on the performance metric and decided which among the three is the best model.
- The best model thus selected is optimized by the application of ‘GridSearchCV’.
- The optimized model is then compared with the BenchMark Model and came to the conclusion that the Optimzed Model shows a better performance than the BenchMark model.
- Finally the Optimized Model which is selected as the best model for the input data is validated aganist unseen data and the performance and intuition was documented clearly. I personally found that the features which are important for the data is very difficult.Because it’s giving low r2_score values when we run with some important features.Again if we run with some other important then it increased.

But finally it is giving low r2_score value which i think that it is suitable for this dataset.

0.20 IMPROVEMENT

- Potentially the ‘Data Preprocessing’ phase is the crucial part of any Machine Learning Problem, Since during this phase we can potentially identify the flaws in the data set that could actually mess the results and performance of the model thus considered.
- Hence, removing irrelevant data during this phase can predominantly increase the model’s performance and can benefit in generalized results.

- Since, this is a regression task, the Wrapper Method implementation i.e., RFE may not be the best one.
- There is also room for trying Embedded method such as LASSO and Elastic Net and Ridge Regression that don't need the external implementation of the feature selection techniques , since these methods have embedded feature selection and regularization built in. It's worth a trial, since there is always scope for improving the model performance aganist the given dataset.
- Further more, some ensemble methods such as 'XGBOOST' should also take care of larger data dimensions and these methods can themselves be used for feature selection.