|  |  |
| --- | --- |
| Text, logo  Description automatically generated  **movie popularity prediction**  Dr /merna | Abstract  What can we say about the success of a movie before it is released? Are there certain companies (Pixar?) that have found a consistent formula? Given that major films costing over $100 million to produce can still flop, this question is more important than ever to the industry. Can we predict which films will be highly rated, whether or not they are a commercial success?  Team CS\_55  [Course title] |

**Team members:-**

* Mariem karam sayed ali
  + - 20201700815
* Maryam abdelnabi Mohamed hassen
  + - 20201700813
* Shoruk ragab abd-elazem saleh
  + - 20201700398
* Zeyad Khaled abd-elazem sayed
  + - 20201700304
* Marwa thabet geat atallah
  + - 20201700798

**Milestone 1 “Preprocessing and regression techniques”**

**First: preprocessing**

* **Fill in missing values**

Imputing missing data is very important in the Feature Engineering process. There are many techniques to impute the missing values in such an accurate manner.

* Numeric Data is imputed using the mean strategy.
* Categorical Data is imputed using the most-frequented value.
* **Drop noisy data**

Starting the Pre-processing by neglecting the unnecessary columns either which don't affect the target value (such as case index) or contain too many missing values.

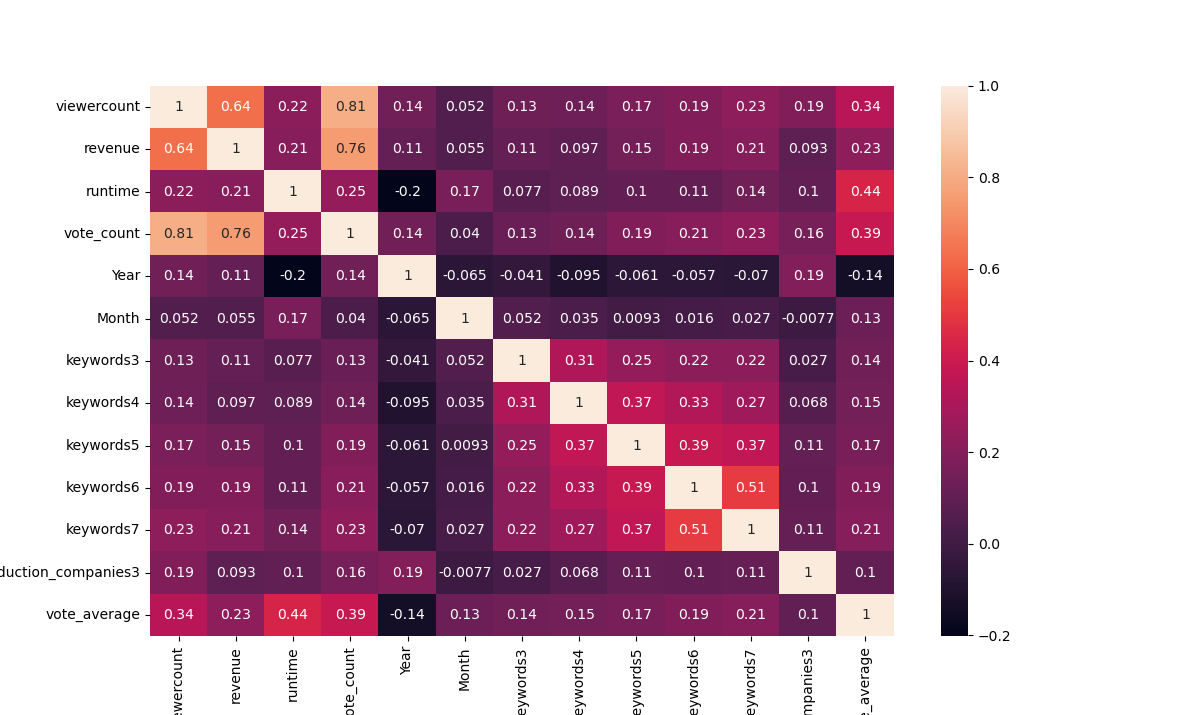
-Dropping ***ID*** column because it has unique data.

-Eliminating those which have more than **50%** of total data missing or showing no values. For this dataset, we are Dropping ***home-page****.* Their Missing values percentages were very high, **62%**. Which they cover almost the entire column, eliminating these features won't cause any significant loss of values or even affects the prediction process. The right evaluation was also considered as it is important to be aware of dealing with such missing values because there may be a chance of losing important data.

* **Feature selection**

It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.

By using the Pearson Coefficient of Correlation we are getting the top 90% correlation features with each other .



* **label encoding**

**movies-regression-dataset** have categorical features and list of dictionaries

* **handle list of Dictionary**

first convert list of dictionary to list using function(Loctolist())then fit the Xtrain to extract categorical columns from list using function (fit\_List())and apply changes on Xtrain and Xtest using function (transform\_List())

* **handle categorical data**

using label encoding change categorical data to numeric to can use it to fit model

* **feature scaling**

Feature scaling is a method used to normalize the range of independent variables or features of data. It makes the flow of ***gradient descent*** smooth and helps algorithms quickly reach the minima of the cost function.

We used **Standardization** which is a very effective technique that re-scales a feature value so that it has distribution with *0 mean* value and *variance* equals *1*.

* **Date separating**

We have to split the date time stamp into few features like Year, Month and Day to easy model fitting.

* **Outlier handling**

Outliers in data can spoil and deceive the training process of machine learning models, resulting in less accurate models and eventually bad performance.

To impute the outliers, we can use a variety of imputation values, ensuring that no data is lost.  
As impute values, we can choose between the**mean, median, mode, and boundary values.**

**Second: regression algorithms used.**

1. Ridge regression

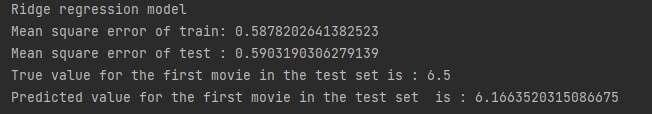
2. Multiple regression

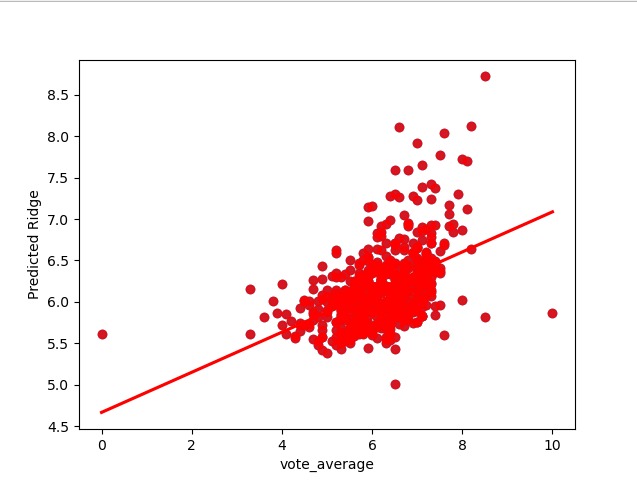
3. Support Vector regression (SVR)

**Third: Differences Between Each Model:**

1. **Ridge regression**

A model tuning method used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

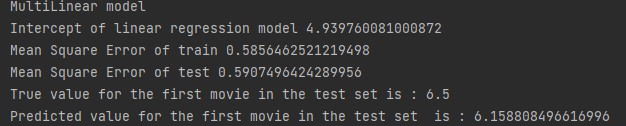
 Error: 0.6

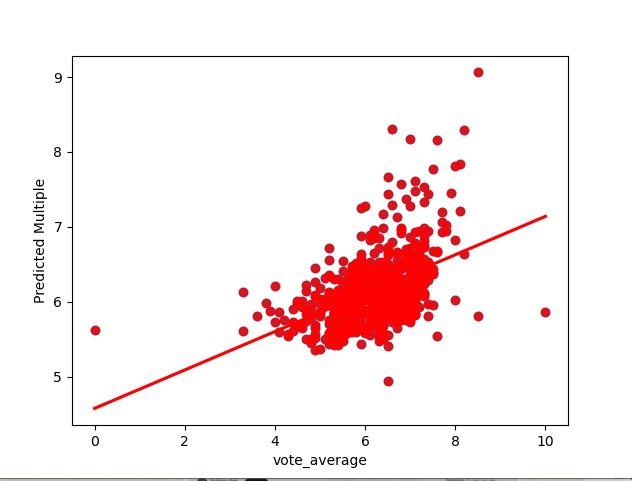
****

**2-Multiple regression**

a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables

Error: 0.6

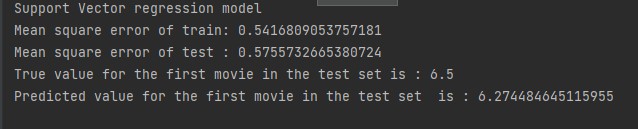


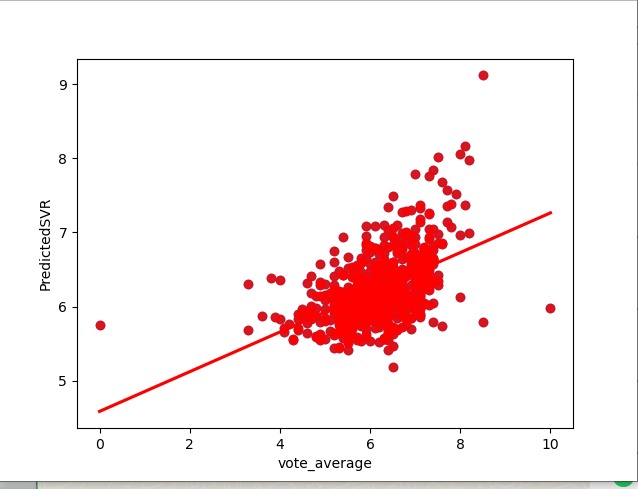
****

**3- Support Vector regression (SVR)**

Use the of idea of SVM with continues dependent variable we get best line fit The data (hyper line) and ignore the error between real data and predicted data within the decision boundary lines

Error: 0.58

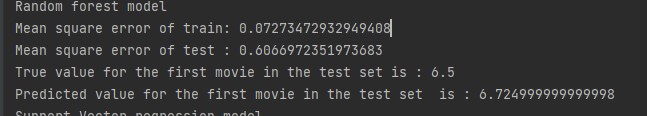
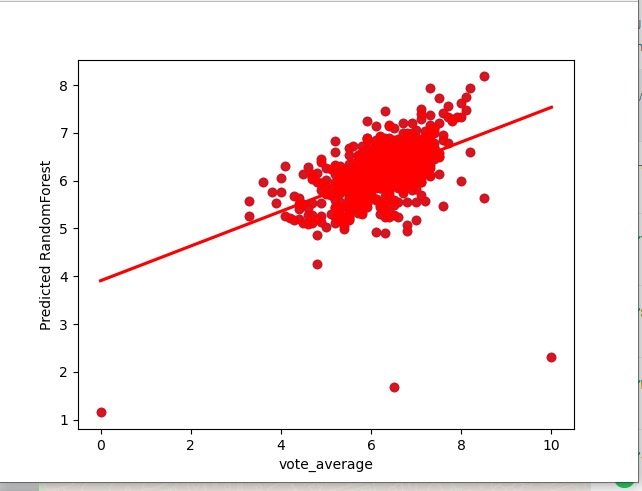




**3-Random Forest Regression**

learning algorithm based on decision tree learners. The estimator fits multiple decision trees on randomly extracted subsets from the dataset and averages their prediction.

Error : 0.61

****

**Fourth:Training & Testing Sizes:**

* **Training set:** is 80% of the total data size.
* **Testing set:** is 20% of the total data size.
* **Validation set:** no validation set was used.

**Fifth : Conclusion:**

In most practices, the best metrics to optimize have been obtained by the Multiple linear regression, Support vector Regression ,Ridge Regression and Random forest regression. But in the movie popularity Prediction, It differs. So it is always related to the dataset and how their features relate with each other, and there is no static way to obtain the best metric.