Movie Popularity Prediction

Milestone 1 Report

Team ID	CS_6

Name	ID
محمد محروس محمد احمد عبدالرحمن	20201700731
صالح عادل صالح محمد	20201700415
شهاب مصطفى فهمى على	20201700405
عبدالرحمن حسني محمد كامل	20201700434
علي محمد علي شارب	20201701108
عمر ايمن حسن غباشي	20201701112

The Report

1) Preprocessing Techniques

1- We dropped Columns based on Certain reasons for Certain Columns

These columns were (homepage, id, status).

How: We got the number of nulls from the homepage column and found it more than half of the data.

We got the most frequent value from the id column and found that there is no repeated value.

We got the most frequent value from the status and found that its number is about all the data.

2- Drop any row that any value on it is null.

3- Split the data by 70 for the train and 30 for the test.

4- (budget):

1- we apply feature scaling on it.

2- We replace the zeros with the column's median.

5- (genres):

1-we got a list containing ids of dictionary and implemented algorithm on it.

2-an algorithm is going to relate the genres' rows with each corresponding runtime value using the ID of each genre to get a numeric value instead of each row of the genres.

6-(keywords):

1- we extracted a list of values of the names from the column.

2- feature encoding.

3- feature scaling.

4- replacing zeros with median.

6- (original language): we applied feature encoding.

6- (Original title):

1- using RE, we removed the stop words, applied the stemming on it.

2- feature encoding.

3- Feature scaling.

7- (overview):

1- using RE, we removed the stop words, applied the stemming on it.

2- feature encoding.

3- Feature scaling.

8- (viewer count):

1- Feature scaling.

2- replaced the outliers with the median.

9- (production companies):

1-we got a list containing ids of dictionary and implemented algorithm on it.

2-an algorithm is going to relate the product company' rows with each corresponding runtime value using the ID of each product company to get a numeric value instead of each row of the product company.

10- (production countries):

1-we got list containing iso_3166_1's of dictionary and implement algorithm on it.

2-an algorithm is going to relate the production countries' rows with each corresponding runtime value using the iso_3166_1 of each production countries to get a numeric value instead of each row of the production countries.

11- (release date): algorithm to calculate the difference (in years) between the current local date and the release date.

12- (revenue):

1- We apply feature scaling on it.
2- We replace the zeros with the column's median.

13- (runtime)

We apply feature scaling on it and use it on dictionaries columns.

14- (spoking language)

1- we extracted a list of values of the iso_639_1 from the column.
2- feature encoding.
3- feature scaling.
4- replacing zeros with median.

15- (tagline)

1- using RE, we removed the stop words, applied the stemming on it.
2- feature encoding.

3- Feature scaling.

16- (title)

1- using RE, we removed the stop words, applied the stemming on it.
2- feature encoding.
3- Feature scaling.

17- (vote count)

1- Feature scaling.
2- replaced the outliers with the median.

18- The Preprocessing is done on the training data and the test data but the different between them that the train data is doing fit and transform in feature encoder and feature scaling while the test data is doing transform only.

2) Feature Selection (Perform Analysis)

- We apply Correlation on dataset after the preprocessing techniques.

1- Apply concatenation on training data and testing data.
3- Apply correlation that the correlation of target column > 0.1.
4- We got on the top features from the correlation to train data and test data.

3) Regression techniques

1- Apply Linear regression.

a. Fit train data and target train data.
b. Predict test data.

c. Calculate the mean square error on this data.
d. MSE = 0.5127449030301584.
e. Calculate the r2_score = 33.49379138051339.

2- Apply Polynomial regression.

a. Apply polynomial regression.
b. Fit and transform on the train data.
c. Fit the transform features to the linear regression.
d. Fit the polynomial data and target train data.
e. Predict on test data.
f. Calculate the mean square error = 0.4688330884532879.
g. Calculate the r2_score = 39.18942732706475.

3- Apply Ridge regression.

a. Fit train data and target train data.
b. Predict test data.
c. Calculate the mean square error on this data.
d. MSE = 0.513919416135518.
e. Calculate the r2_score = 33.3414497128542.

4) Differences Between Each Model

1 When we applied polynomial regression the mean square error and the accuracy became better from linear regression.

2 When we applied ridge regression the mean square error and the accuracy became almost like linear regression.

5) Features on our regressions

1- (Use) => Based on correlation of the features used, their correlation is greater than 0.1 (genres, viewercount, production companies , revenue, runtime , vote count).

2- (Discard) (budget, keywords , original language , original title , overview , production countries , release date , spoken language , tagline , title).

6) Sizes of training data and testing data

1- The data is divided into 70% train and 30% test and the shuffle is true when the divide.

2- The number of columns for the train and test data is 16 columns before correlation.

3- The number of rows for the train data is 1859 rows.

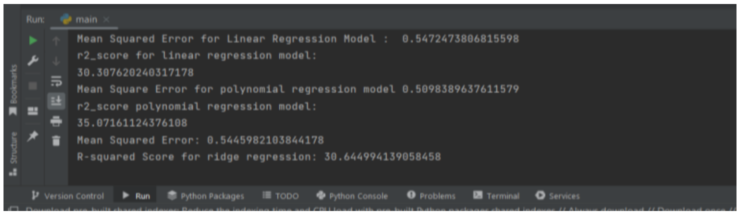
4- The number of rows for the test data is 797 rows.

7) Techniques improve the results.

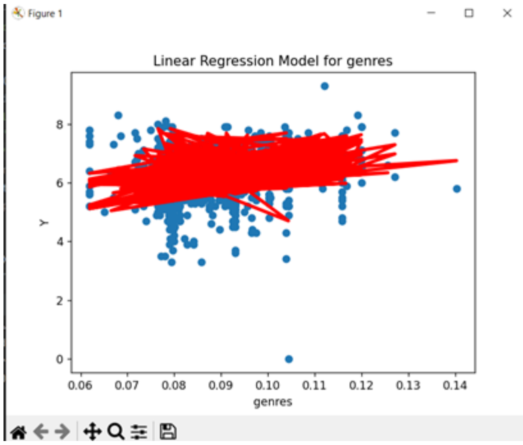
We use polynomial regression and it improves the result and improve the mean square error from 0.5127449030301584 to 0.4688330884532879 and improve the r2_score from 33.49379138051339 to 39.18942732706475.

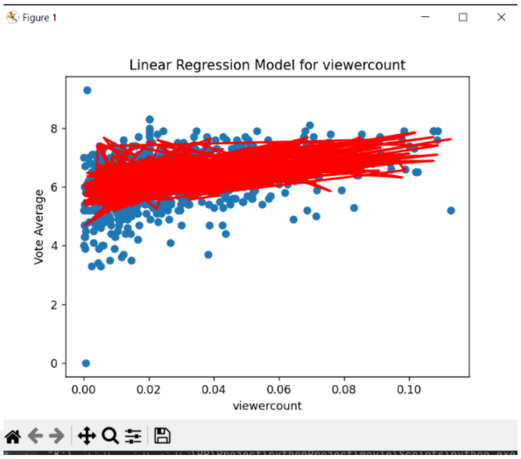
8) Screenshots

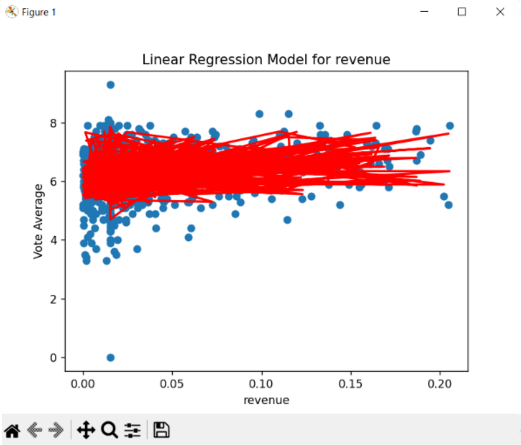
The Resultants of regression

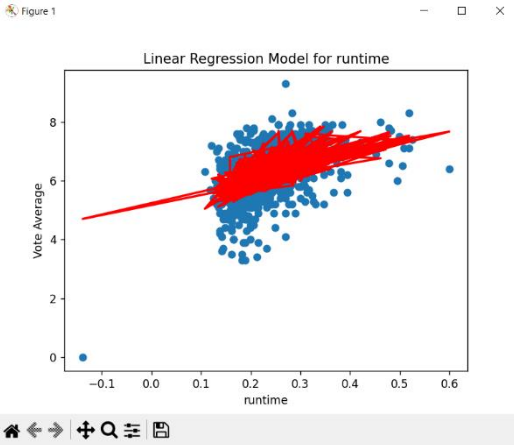


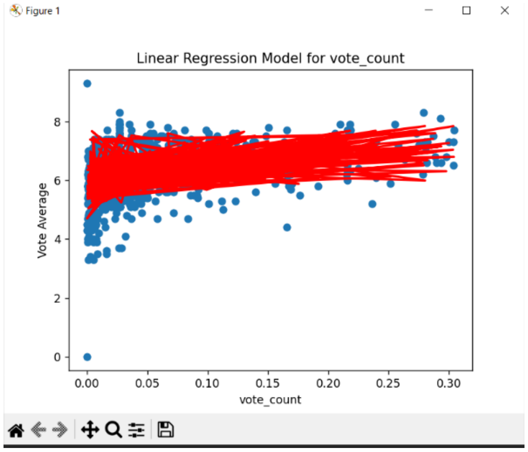
```
Mean Squared Error for Linear Regression Model :  0.5472473806815598
r2_score for linear regression model:
30.307620240317178
Mean Square Error for polynomial regression model 0.5098389637611579
r2_score polynomial regression model:
35.07161124376108
Mean Squared Error: 0.5445982103844178
R-squared Score for ridge regression: 30.644994139058458
```


Linear Regression Model for genres


Linear Regression Model for viewercount


Linear Regression Model for revenue


Linear Regression Model for runtime


Linear Regression Model for vote_count

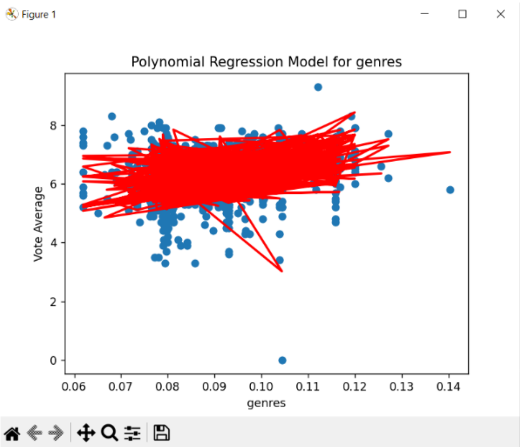


Figure 1
Polynomial Regression Model for genres
Vote Average
0
2
4
6
8
0.06
0.07
0.08
0.09
0.10
0.11
0.12
0.13
0.14
genres

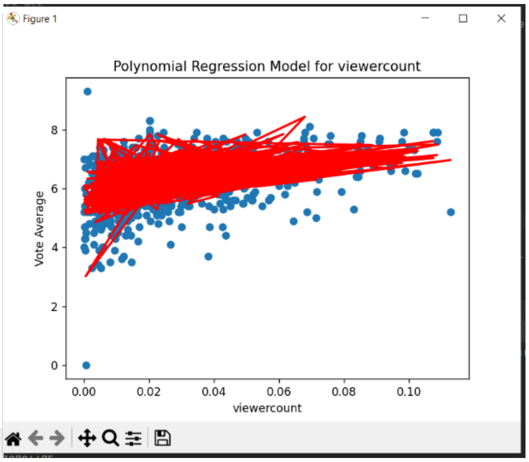


Figure 1
Polynomial Regression Model for viewercount
Vote Average
0
2
4
6
8
0.00
0.02
0.04
0.06
0.08
0.10
viewercount

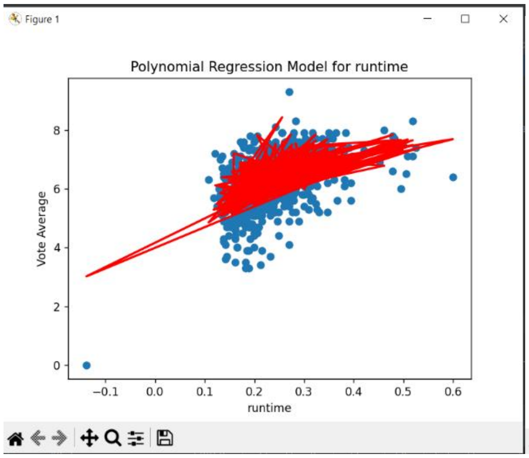


Figure 1
Polynomial Regression Model for runtime
Vote Average
0
2
4
6
8
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
runtime

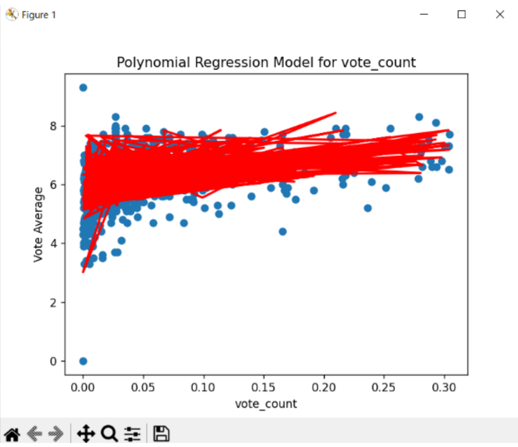


Figure 1
Polynomial Regression Model for vote_count
Vote Average
0
2
4
6
8
0.00
0.05
0.10
0.15
0.20
0.25
0.30
vote_count

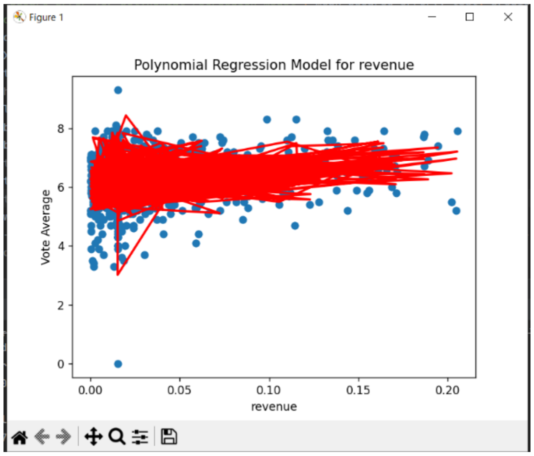


Figure 1
Polynomial Regression Model for revenue
Vote Average
0
2
4
6
8
0.00
0.05
0.10
0.15
0.20
revenue

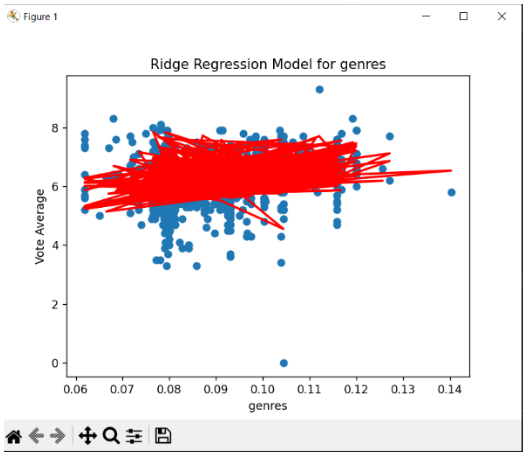


Figure 1
Ridge Regression Model for genres
Vote Average
0
2
4
6
8
0.06
0.07
0.08
0.09
0.10
0.11
0.12
0.13
0.14
genres

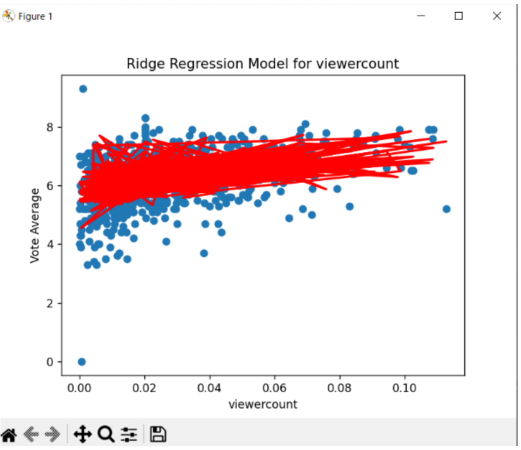


Figure 1
Ridge Regression Model for viewercount
Vote Average
0
2
4
6
8
0.00
0.02
0.04
0.06
0.08
0.10
viewercount

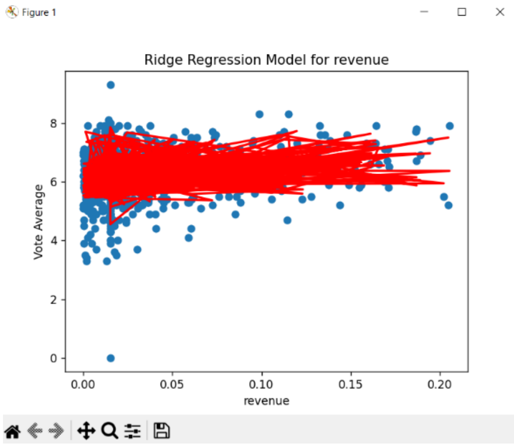


Figure 1
Ridge Regression Model for revenue
Vote Average
0
2
4
6
8
0.00
0.05
0.10
0.15
0.20
revenue

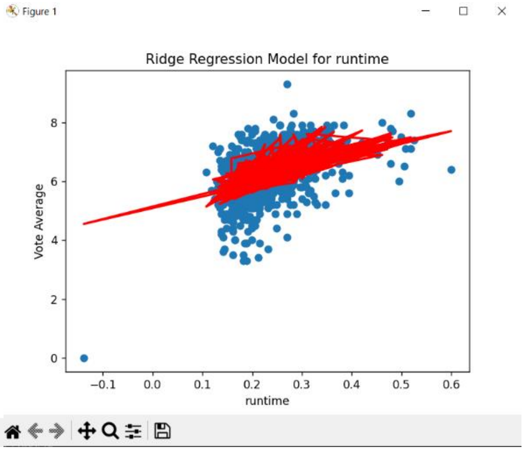


Figure 1
Ridge Regression Model for runtime
Vote Average
0
2
4
6
8
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
runtime

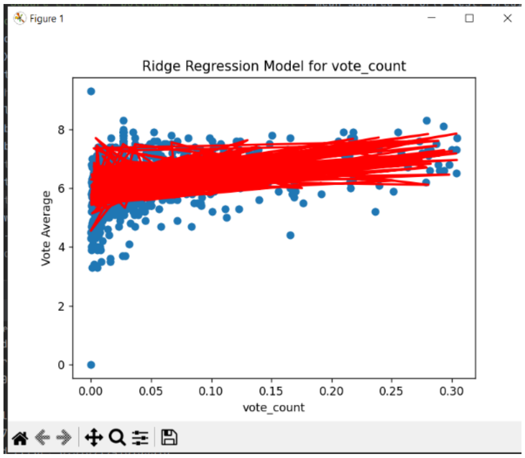
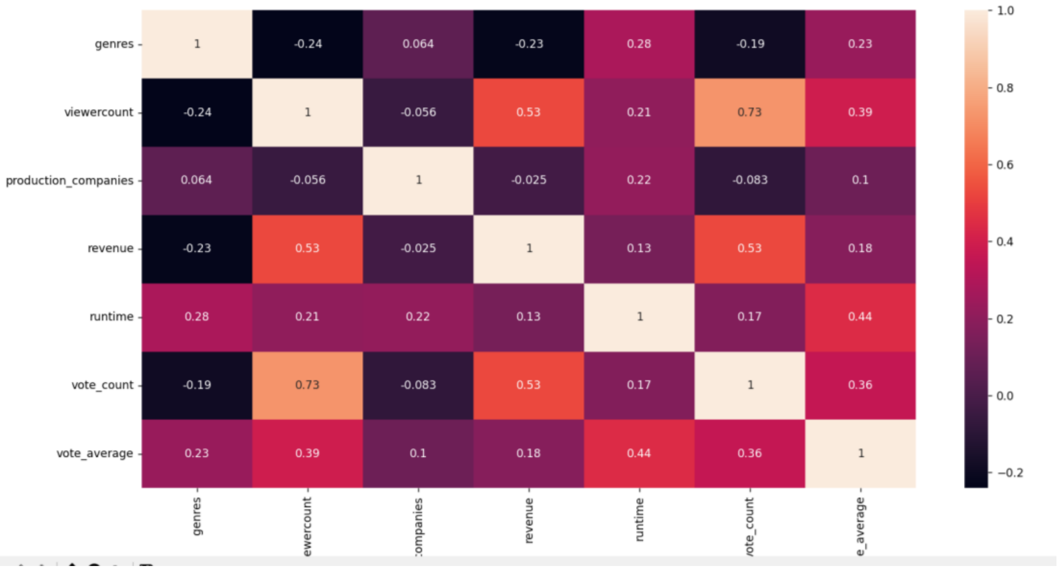


Figure 1
Ridge Regression Model for vote_count
Vote Average
0
2
4
6
8
0.00
0.05
0.10
0.15
0.20
0.25
0.30
vote_count

The Resultants of correlation



9) Conclusion

We used preprocessing techniques on this data and the data were coming bad values before the preprocessing but after the preprocessing the data became better.
We used correlation after the preprocessing techniques, and it extracted the best columns.
We used linear regression and polynomial regression and the mean square error for the polynomial became better than the mean square for the linear regression.