Predicting Movie Success Report

Presented by: Team 15 – SC

Ola Adel Badawy

Salma Essa Fouad

Mohamed Yossry Mohamed

Mohammed Abdallah ElSaid

Farah Mohamed Abdelaziz

- **Data Cleaning**

  - ### Dropping Unnecessary Features

    ['homepage', 'id', 'original_title', 'overview', 'tagline','release_date' , 'movie_id']

    We decided to drop them, because they are not necessary for answering the questions and no clear relation between them and the targeted feature.

  - ### Dropping Duplicate Features

    ['original_language', 'status']

    More than 99% of movies has status 'released', so the model can't be trained of unreleased movies efficiently where the ratio between the number of features and number of training rows is not consistent, also more than 95% of data their language is 'En', so the model can't be trained of a movies released with another language.

  - ### Dropping Missing Rows

    As runtime feature is critical for deciding the targeted feature and no way to predict it, we removed all training samples that doesn't has runtime, also we removed all training samples that doesn't has vote average because the model can't be trained with unlabeled data.

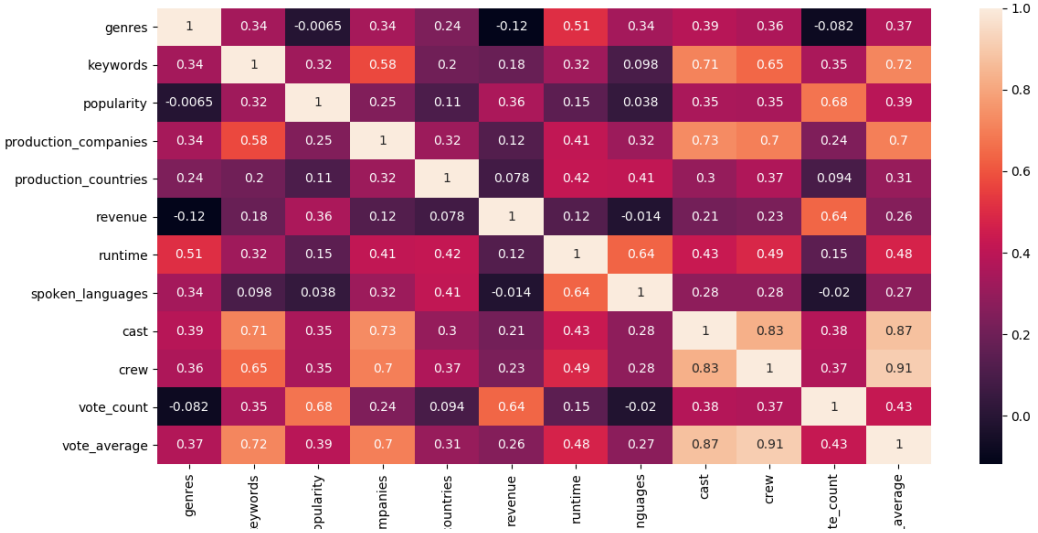
  - ### Categorical Feature Mapping

    We convert the categorical feature into numerical by mapping each unique category to the mean vote averages corresponding to it.

Finally, after data cleaning we includes the following features

['budget','runtime','revenue','genres','keywords','production_companies','production_countries','spoken_languages','cast','crew','popularity','vote_average','vote_count']

- **Features Correlation Matrix**



- **Improvements Techniques**

  Feature scaling (min-max normalization)

- **Training and Testing Sets**

  After dropping the missing rows we get 3793 samples. data splitted into training, testing in 80 – 20 manner.

- # Regression Techniques

  - ## Linear Regression

    It basically gives us an equation, where we have our features as independent variables, on which our target variable is dependent upon.

    $$Y = \theta_1 X_1 + \theta_2 X_2 + \dots \theta_n X_n$$

    MSE = 0.0009762154930126408

    Score = 0.9058087007794992

  - ## Polynomial Regression

    Polynomial regression is another form of regression in which the maximum power of the independent variable is more than 1. In this regression technique, the best fit line is not a straight line instead it is in the form of a curve.

    Quadratic regression, or regression with second order polynomial, is given by the following equation: $Y = \Theta_1 + \Theta_2 X + \Theta_3 X^2$

    MSE = 0.0008851991846504766

    Score = 0.91459051626619

  - ## Ridge Regression

    Ridge regression is an extension for linear regression. It's basically a regularized linear regression model.

    cost function for ridge regression:

    $$\min\left(\left|\left|Y - X(\theta)\right|\right|_2^2 + \lambda||\theta||_2^2\right)$$

    λ given here, is actually denoted by alpha parameter in the ridge function. So by changing the values of alpha, we are basically controlling the penalty term. Higher the values of alpha, bigger is the penalty and therefore the magnitude of coefficients are reduced.

    MSE = 0.0009762440165542012

    Score = 0.9058059486520671

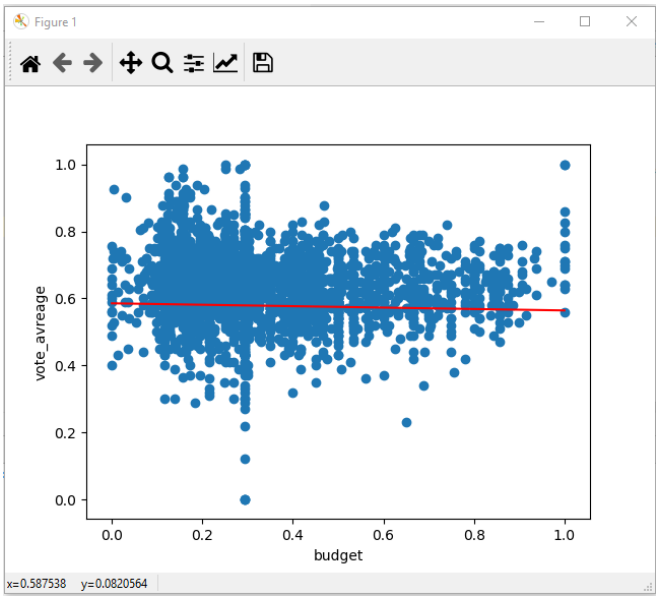
- ## Lasso Regression

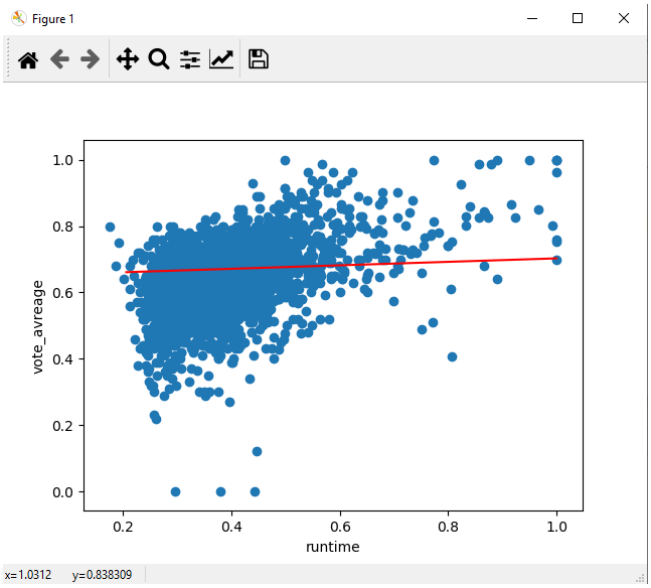
  Mathematics behind lasso regression is quiet similar to that of ridge only difference being instead of adding squares of theta, we will add absolute value of Θ.

  MSE = 0.010369757119433706

  Score = -0.000538203574537155

- **regression line plots**





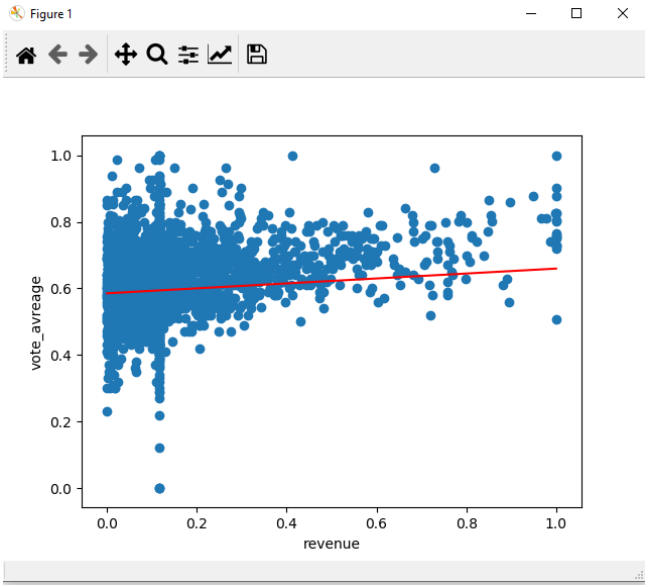


Figure 1
vote_avreage
revenue
0.0
0.2
0.4
0.6
0.8
1.0

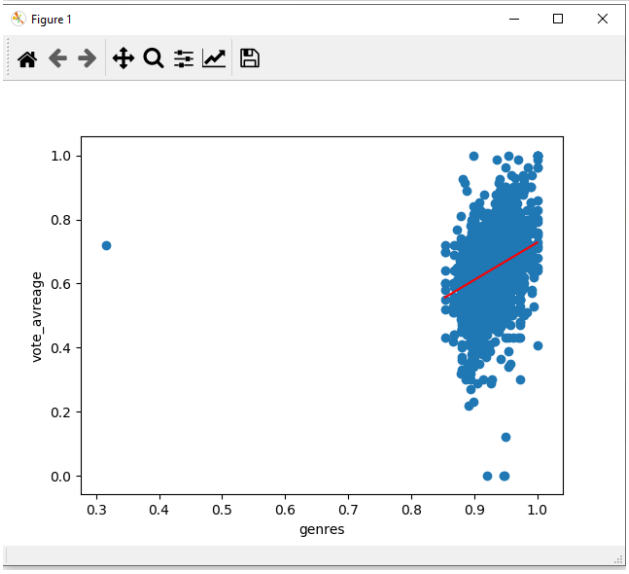


Figure 1
vote_avreage
genres
0.0
0.2
0.4
0.6
0.8
1.0
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

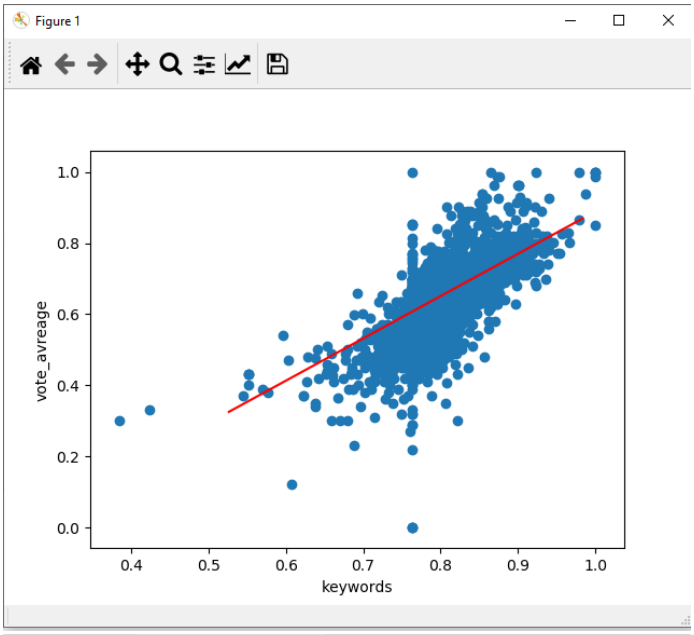


Figure 1
vote_avreage
keywords
0.0
0.2
0.4
0.6
0.8
1.0
0.4
0.5
0.6
0.7
0.8
0.9
1.0

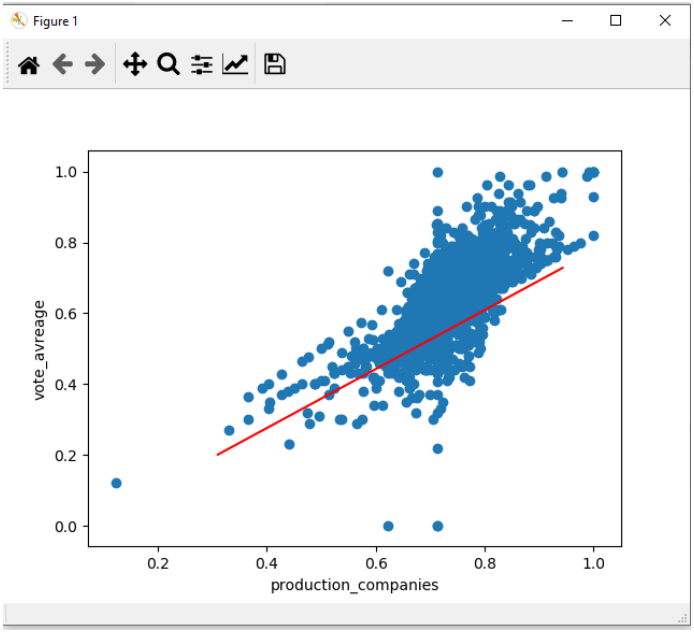


Figure 1
vote_avreage
production_companies
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.4
0.6
0.8
1.0

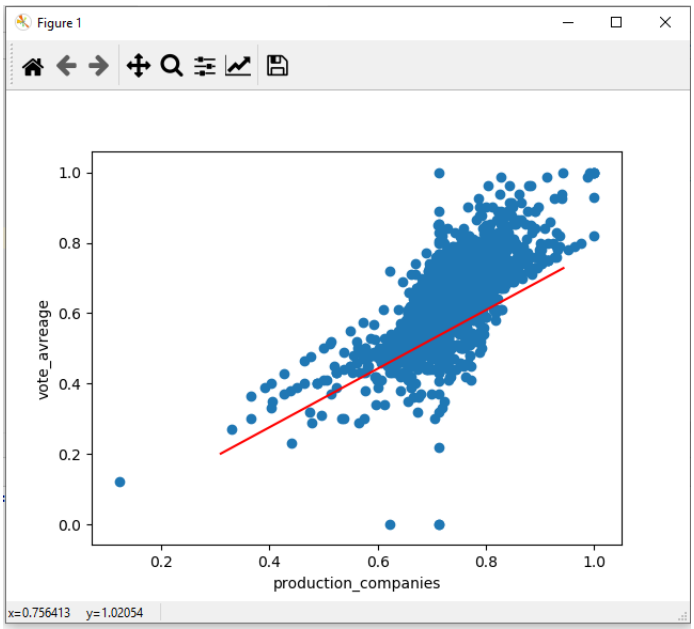


Figure 1
vote_avreage
production_companies
1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
x=0.756413 y=1.02054

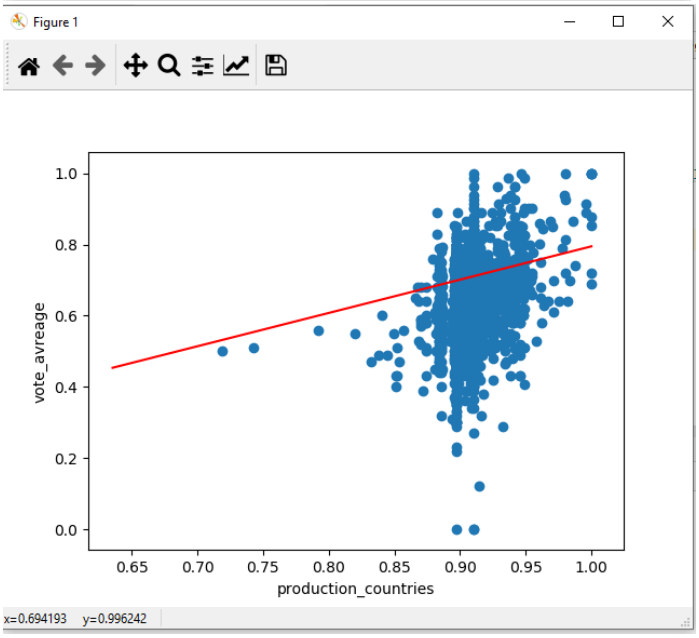


Figure 1
vote_avreage
production_countries
1.0
0.8
0.6
0.4
0.2
0.0
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
x=0.694193 y=0.996242

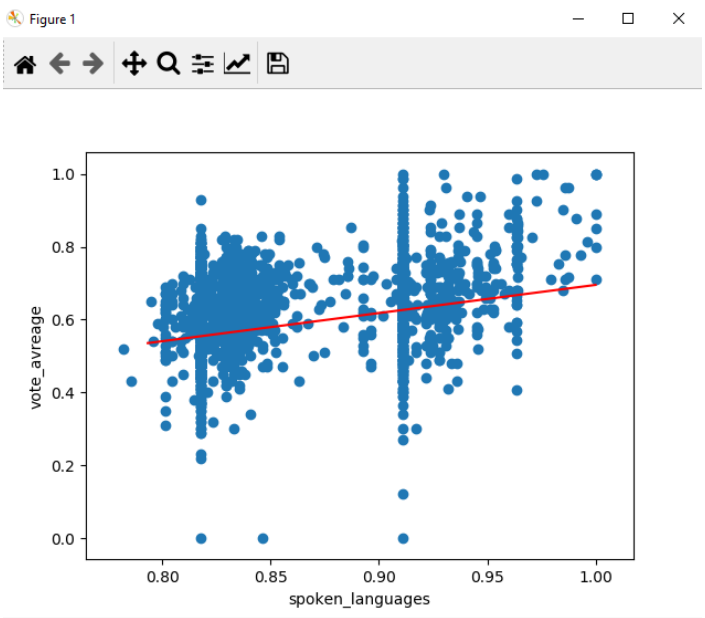


Figure 1
1.0
0.8
0.6
0.4
0.2
0.0
vote_avreage
0.80
0.85
0.90
0.95
1.00
spoken_languages

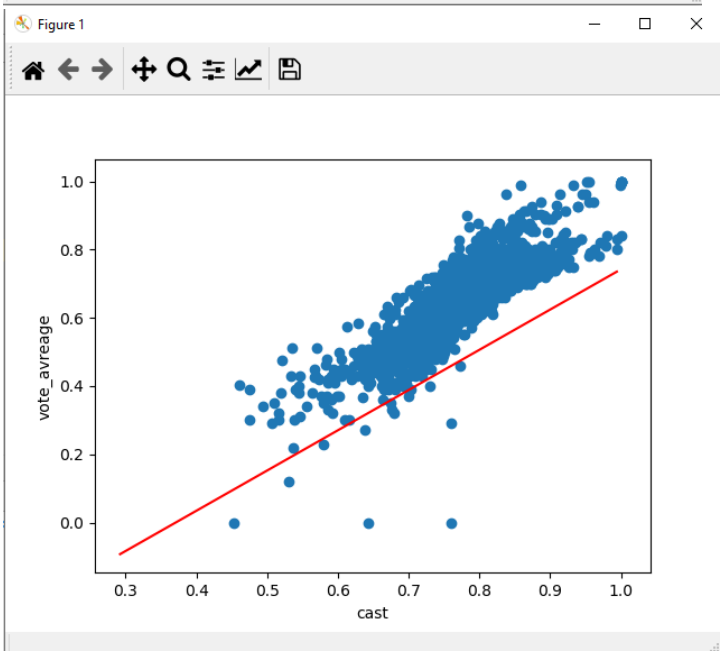


Figure 1
1.0
0.8
0.6
0.4
0.2
0.0
vote_avreage
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
cast

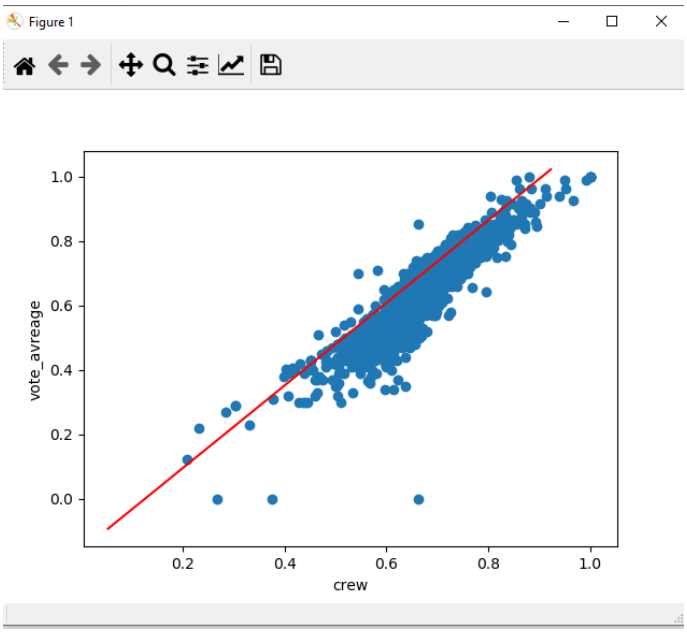


Figure 1
vote_avreage
crew
0.0
0.2
0.4
0.6
0.8
1.0

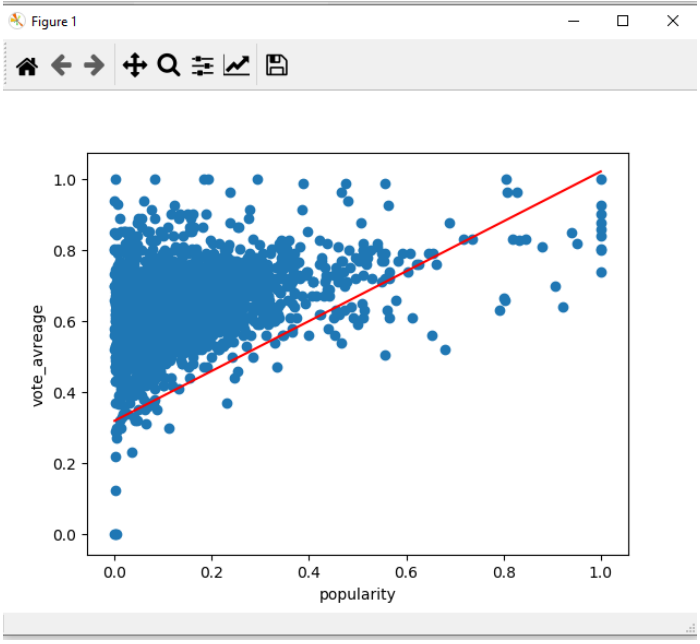
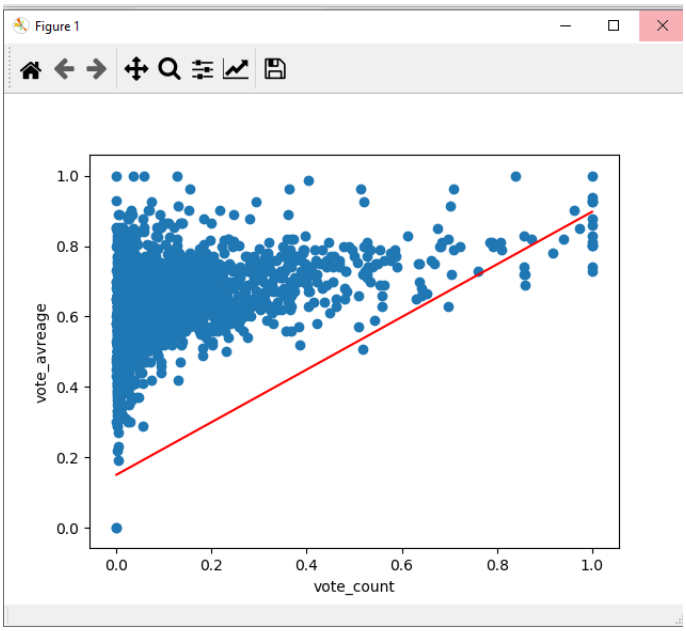


Figure 1
vote_avreage
popularity
0.0
0.2
0.4
0.6
0.8
1.0



- **Conclusion**
  - Dropping unnecessary features helps the model to decrease the MSE where no clear relation between them and the targeted feature.
  - All computed MSE based on different regression methods are approximately equal.
  - Best regression method is the polynomial regression.