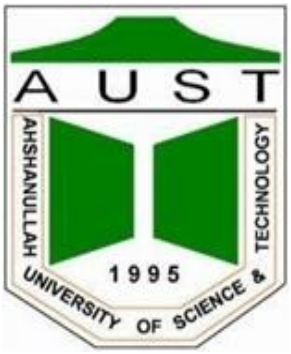

Ahsanullah University of Science and Technology

Department of Computer Science & Engineering

Project Report

On

Google Play Store Apps Rating Prediction

Course No.	CSE 4108
Course Name	Artificial Intelligence Lab

Submitted To:

Md. Siam Ansary	Tonmoy Hossain
Department of CSE, AUST	Department of CSE, AUST

Submitted By:

Name	Tahiya Ahmed Chowdhury
ID No.	17.02.04.048
Session	Fall – 2020
Section	A (A2)
Date of Submission:	September 09, 2021

Introduction:

The Google Play Store is every android user's go to application installing platform. Apps of multi-dimensional categories, varying sizes and android requirements are published. Our aim is to develop two such models which will predict the rating of apps based on category, size, total no of installs and also rating count. This will help software companies, startups and developers to get an idea of the user preferences in the marketplace.

Brief Description of Dataset:

From the main dataset Google-Playstore, 550 rows were selected and 5 columns were used for training the model. The columns category, size, total no of installs and rating count was allocated as the features and rating as the targeted.

Description of the Models:

1. Multiple Linear Regression Model:

Multiple linear regression model simply follows the methods of linear regression except that the targeted data depends on more than one feature data. That is, the number of independent variable here is more than one while dependent remains the same.
For the creation of our model, we followed the steps:

I. The necessary libraries such as numpy, pandas, matplotlib, sklearn,etc. were imported.
II. Then after importing the dataset Google-Playstore.csv, the data was then preprocessed by extracting 550 rows and 5 columns and performing necessary operations such as removing columns with null values, rearranging their orders and changing their datatypes,etc with the help of pandas library modules.
III. The feature data columns category, size, total no of installs and rating count were allocated in matrix X and target data rating in y.
IV. Using SimpleImputer module from sklearn the missing datas were adjusted.
V. The categorical feature i.e the first column was encoded using One Hot Encoder.
VI. 25% of the dataset was assigned for test while rest used for training the supervised model.
VII. Finally using the Linear Regression module from sklearn library, the model was trained.
VIII. Lastly the test results were predicted and evaluated by determining r2 score, RMSE, MSE and MAE.

2. Polynomial Regression Model:

Multiple linear regression model simply follows the methods of linear regression except that the targeted data depends on more than one feature data. That is, the number of independent variable here is more than one while dependent remains the same.
For the creation of our model, we followed the steps:

I. The necessary libraries such as numpy, pandas, matplotlib, sklearn,etc. and the same dataset was imported.
II. Data was preprocessed exactly like the above mentioned model for maintaining consistency.
III. The feature data columns category, size, total no of installs and rating count were allocated in matrix X and target data rating in y.
IV. Using SimpleImputer module from sklearn the missing data were adjusted.
V. The categorical feature i.e the first column was encoded using One Hot Encoder.
VI. 25% of the dataset was assigned for test while rest used for training the supervised model.
VII. Finally using the Ploynomial Regression module from sklearn library, the model was trained. The order used was second.
VIII. Lastly the test results were predicted and evaluated by determining r2 score, RMSE, MSE and MAE.

Comparison of the Performance of Models:

Model	R2_score	MSE	RMSE	MAE
Multiple Linear Regression	-0.314	0.814	0.663	0.638
Polynomial Regression	0.242	0.240	0.490	0.338

Here it may be noted that for multiple linear regression the r score is negative which proves that the model is a very poor fit to the data. Also the mean square error, root mean square and mean absolute error is higher than that of polynomial regression.

But in polynomial regression model, the r score is 0.242 which is a weak score which implies that this model is a weak fit to the data as well. But not as much as the other one. However the mean square error, root mean square and mean absolute error values are acceptable.

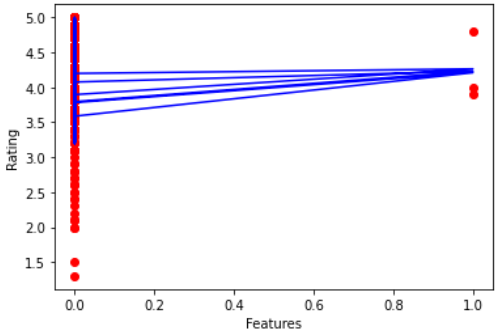

Fig1 : Multiple Linear Regression Model

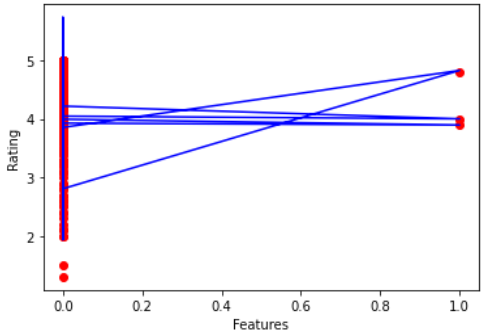

Fig2 : Polynomial Regression Model

For both the models, the red line represents the actual data and the blue one represents the

predicted outcome. As the value of r score is not satisfactory thus in the figures above neither the linear regression model have a linear graph nor the polynomial regression model have a curve which proves that data is not fitting properly. However for multiple regression model the blue line is almost consistent with the red one which means that predictions made by the polynomial models is not that bad which was reflected in the RMSE,MSE,MAE scores as well.

Conclusion:

After analyzing the performance of our models we can come to the conclusion that although both models are not the perfect fit for our data, yet performance wise polynomial regression model is working better among them for this dataset.