**IMDB RATINGS PREDICTION**

END TERM REPORT

BY:

ANIRUDDHA DAS

SECTION: K21ML

ROLL NO.: A52

TEJAS THUKRAL

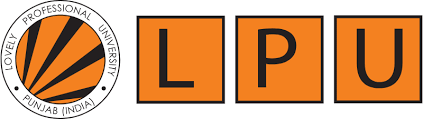
SECTION: K21ML

ROLL NO.:

UJJWAL SHIVHARE

SECTION: K21ML

ROLL NO.:



**Department of Machine Learning**

**School of Computer Science Engineering**

**Lovely Professional University, Jalandhar**

November-2022

**STUDENT DECLARATION**

This is to declare that this report has been written by us. No part of the report is

copied from other sources. All information included from other sources have been

duly acknowledged. I am aware that if any part of the report is found to be copied, We

will take full responsibility for it.

Signature:

Name:

Roll No.:

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| S.No. | Title | Page No. |
| 1. | Cover Page | 1 |
| 2. | Student Declaration | 2 |
| 3. | Table of Contents | 3 |
| 4. | Bonafide Certificate | 4 |
| 5. | Objective | 5 |
| 6. | History | 5-7 |
| 7. | Project Description | 7-9 |
| 8. | Working | 9-11 |
| 9. | Conclusion | 11 |

**BONAFIDE CERTIFICATE**

Certified that this project report “IMDB RATINGS PREDICTION

” is the bonafide work of “Tejas Thukral”, “Aniruddha Das”, “Ujjwal Shivhare” who all carried out the project work under my supervision.

Signature of the Supervisor:

Name: Dr. Dhanpratap Singh

ID:25706

Department: Computer Science Engineering

Objective : IMDB BOX OFFICE PREDICTION , as the name implies is a project about estimating ratings of movies through various features that the movie has, the purpose of the project is to show the correlation between ratings and other elements of the movie such as gross income, genre and runtime. Thus, the model that is created, can be used to predict success of the movie.

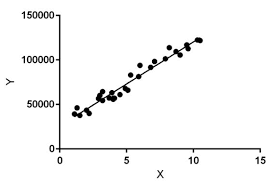
History : Since the last decades, movie infotainment sites such as rotten tomatoes and IMDB have become really popular, they provide a platform for movie-goers to critic a movie and also give them information about various features about the movies itself such as director, actors, gross revenue among others.

It seems there is a visible correlation between ratings of the movie and public perception, gross revenue generated, runtime at theatres and cinema halls, genre and famous name who worked in the movie or presented it/ made it possible. Machine learning models can be used to predict ratings of the movies by describing various features about them and vice versa.

How is it done? : There are various machine learning techniques that can be used to train a model, be it supervised or unsupervised or reinforcement learning models and techniques, for this project, we will use two ML-based regression algorithms to predict IMDB ratings to a certain degree of accuracy then we will compare which one better for this kind of prediction

What is regression? : Before we delve into the inner workings of the regression algorithms, we must understand what regression is, Regression is a statistical analysis technique often used in economics to predict various facets of economy, it is used to statistically analyse the relation of one dependant variable with an independent variable. It takes past input and as it is a supervised machine learning algorithm, it should be labelled data, it then makes probabilistic determination between given two or more feature sets to reach an outcome, there are various kinds of regression algorithms that are popular and in-use, they are:

1. Linear Regression : Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. These algorithms are relatively very simple and are considered one of the fundamental algorithms for beginners to dabble into.



**Fig 1: A graph showing Linear Regression’s ‘Best Fit’ approach**

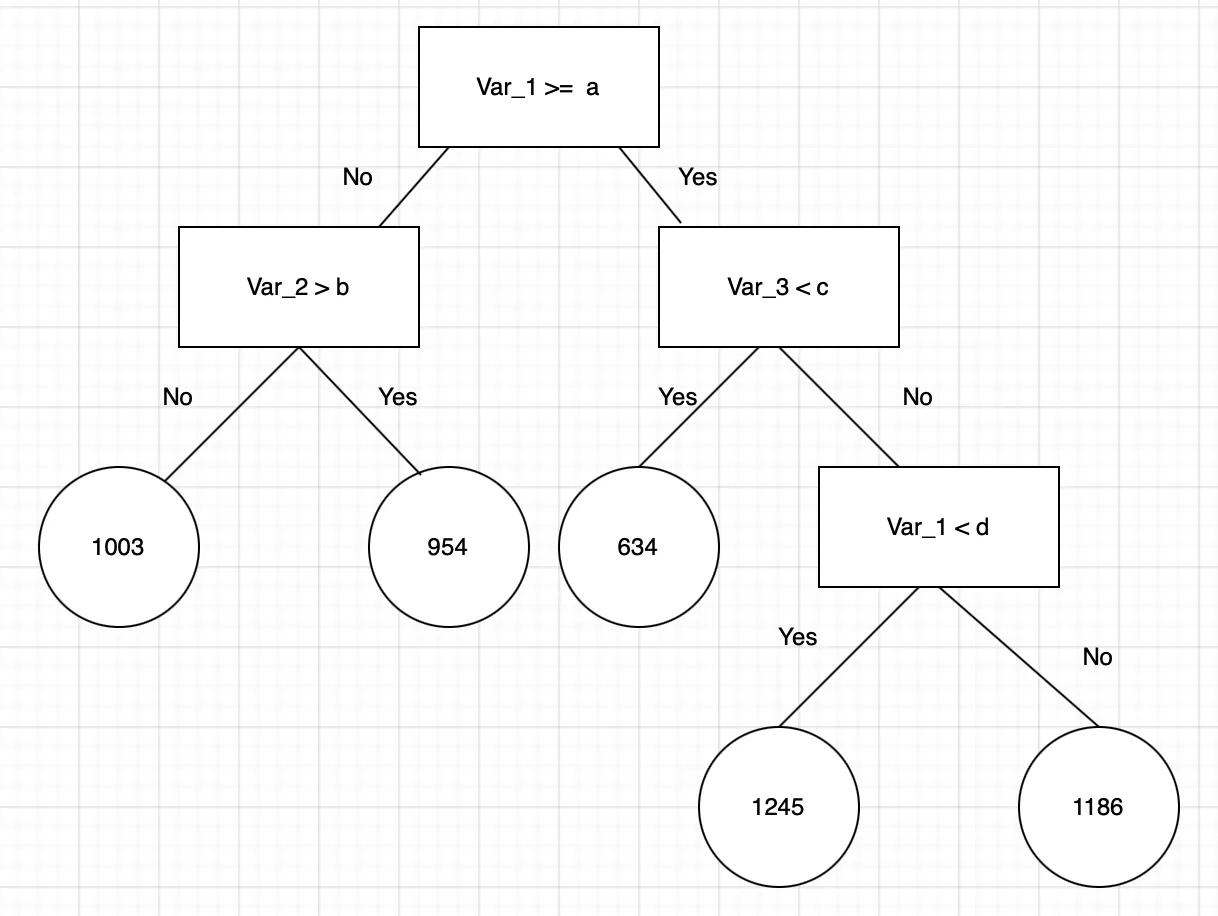
This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

We will use scikit-sklearn module to make the model in python, which contains the above-mentioned technique.

1. Random Forest Regression : Random Forest is a popularly-used machine learning algorithm developed by Leo Breiman and Adele Cutler, which combines the outcomes of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

How does it work? :

**Decision Trees** are used for both regression and classification problems. As the name implies, they appear like trees where each node can have multiple children, each split is the result of some loss function(Usually mean squared error). An example of a decision tree is below:



**Fig 1: Example of a decision tree**

**Ensemble Learning** is the way of using multiple decision tress at the time of making the model which are trained with same data, basically averaging the results of this ensemble to make better predictions. Our requirement, for ensemble learning is that the errors of each model are independent and different from tree to tree.

**Bootstrapping** is the process of randomly sampling subsets of a dataset over a given number of iterations and a given number of variables. These results are then averaged together to obtain a more powerful result. Bootstrapping is an example of an applied ensemble model.

We will use the sklearn module for training our random forest regression model, specifically the RandomForestRegressor function. The RandomForestRegressor documentation shows many different parameters we can select for our model. Some of the important parameters are highlighted below:

**n\_estimators** — the number of decision trees you will be running in the model

**criterion** — this variable allows you to select the criterion (loss function) used to determine model outcomes. We can select from loss functions such as mean squared error (MSE) and mean absolute error (MAE). The default value is MSE.

**max\_depth** — this sets the maximum possible depth of each tree

**max\_features** — the maximum number of features the model will consider when determining a split

**bootstrap** — the default value for this is True, meaning the model follows bootstrapping principles (defined earlier)

**max\_samples** — This parameter assumes bootstrapping is set to True, if not, this parameter doesn’t apply. In the case of True, this value sets the largest size of each sample for each tree.

Other important parameters are **min\_samples\_split**, **min\_samples\_leaf**, n\_jobs, and others.

**Project Description** : Now that we know about the regression techniques we are going to use in this project, we will now make the preparations for the model itself.

* Libraries : Matplotlib : We will generate plots, heatmaps which show correlation, scatterplots, etc., to make our project more appealing and easier to understand.

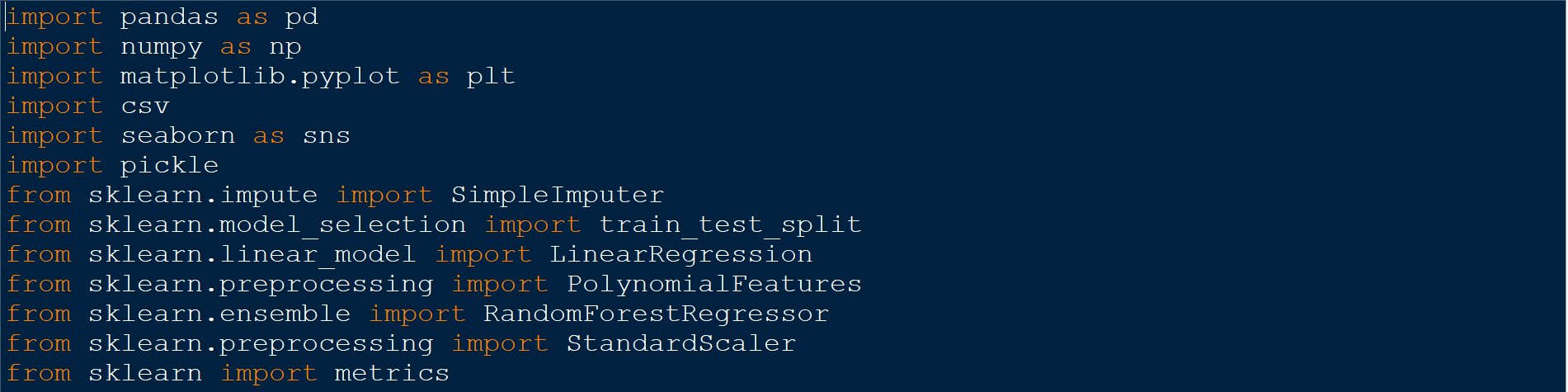
Pandas : We will use the dataframes and trainer model function to make our data train-able.

Numpy : It has several helpful functions and libraries such as numpy array etc to greatly simplify calculations.

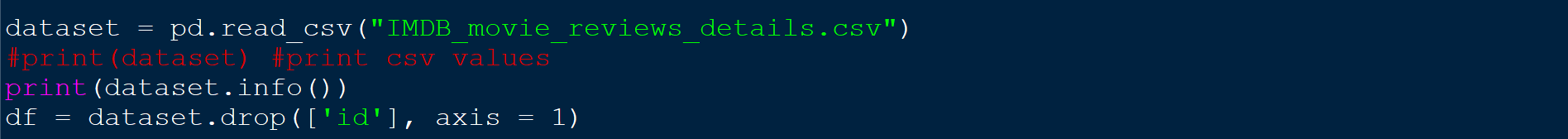
Seaborn : An extension and enhancement of plotting and graphing capabilities of matplotlib.

* Working :

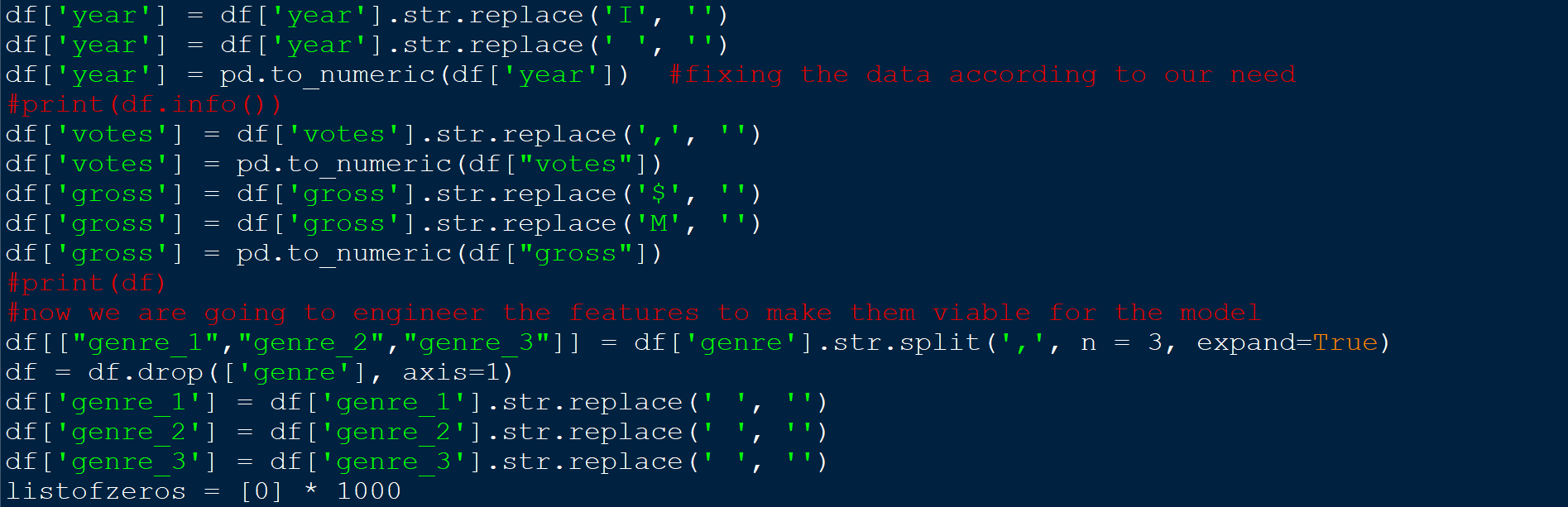
Step 1: Import all essential libraries



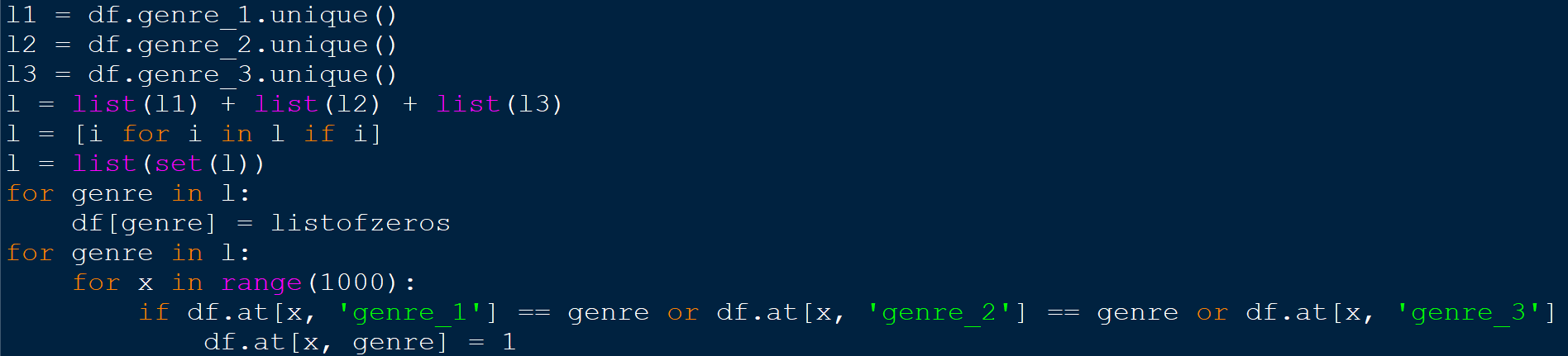
Step 2: Initialize the dataset



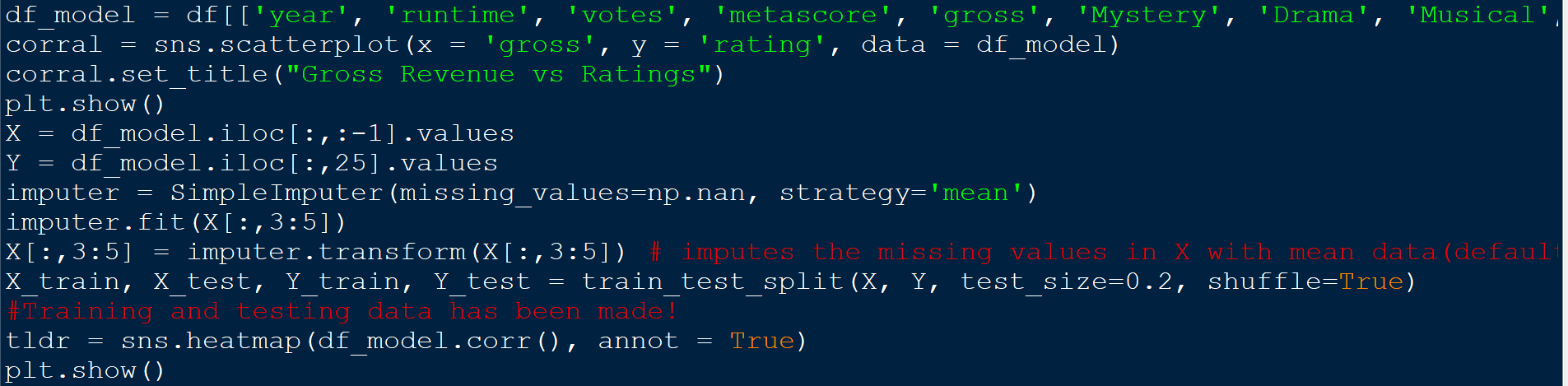
Step 3: Clean and refine the dataset



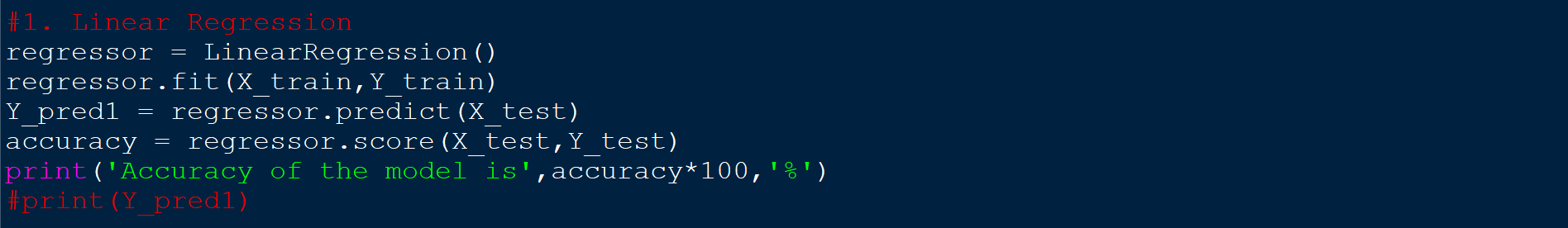
Step 4: Refine the dataset further



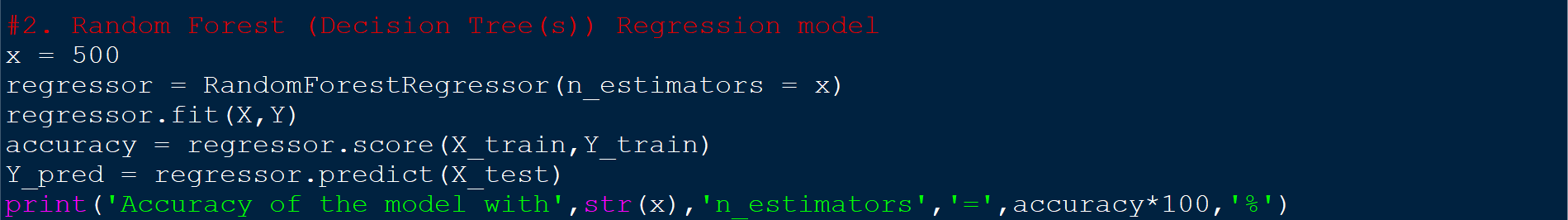
Step 5: Extract useful features and make the dataset out of it



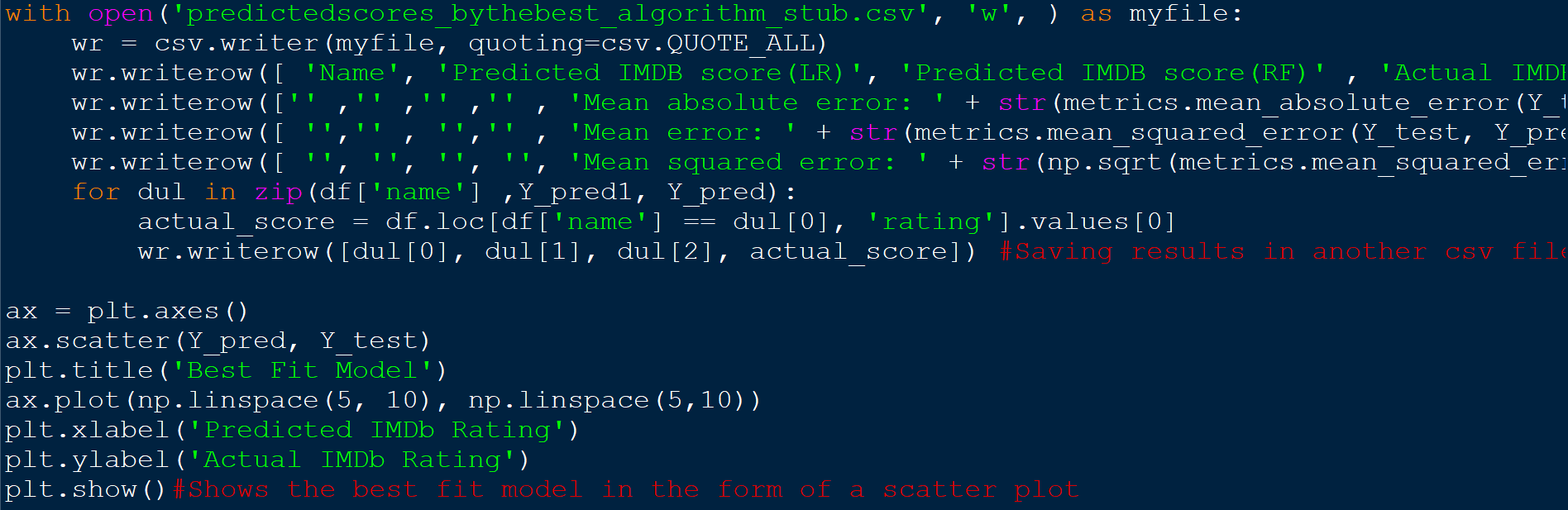
Step 6: Make model using Linear Regression

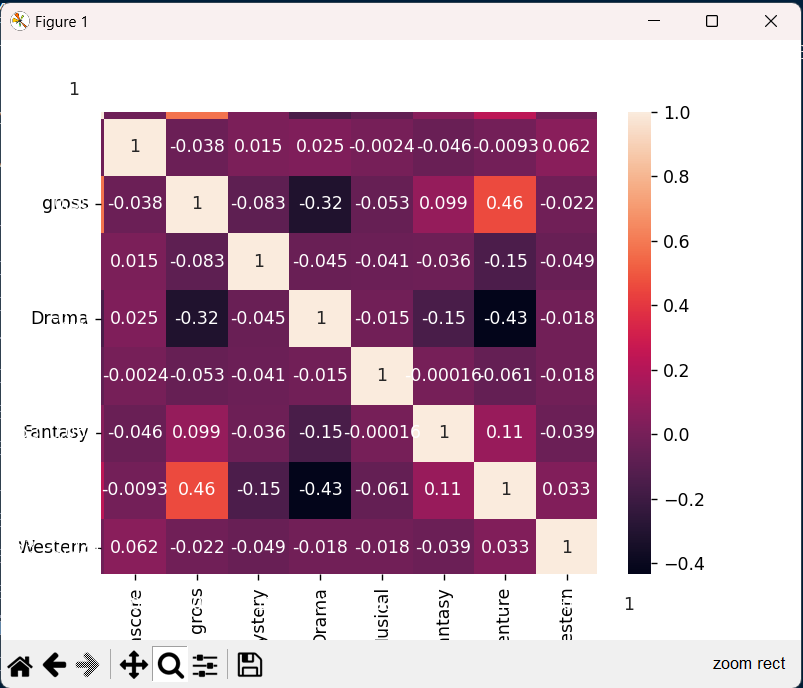
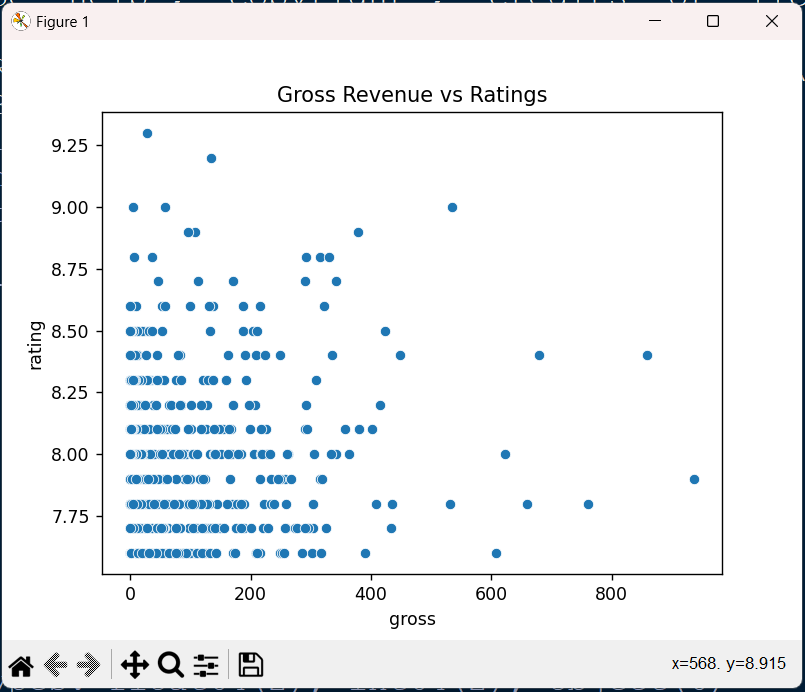


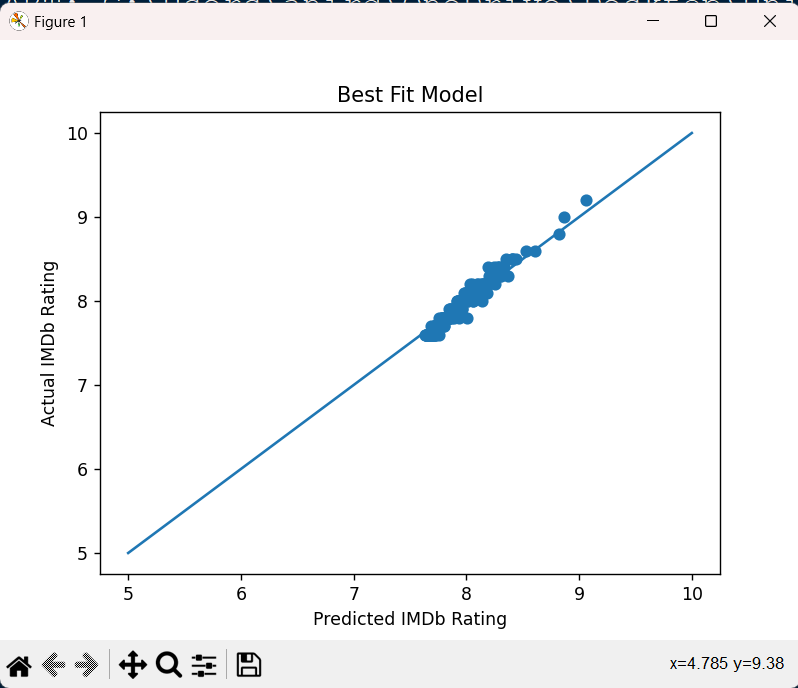
Step 7: Make model using Random Forest Regression



Step 8: Save the predicted outcomes in an csv file and make best fit graph for the model with best accuracy



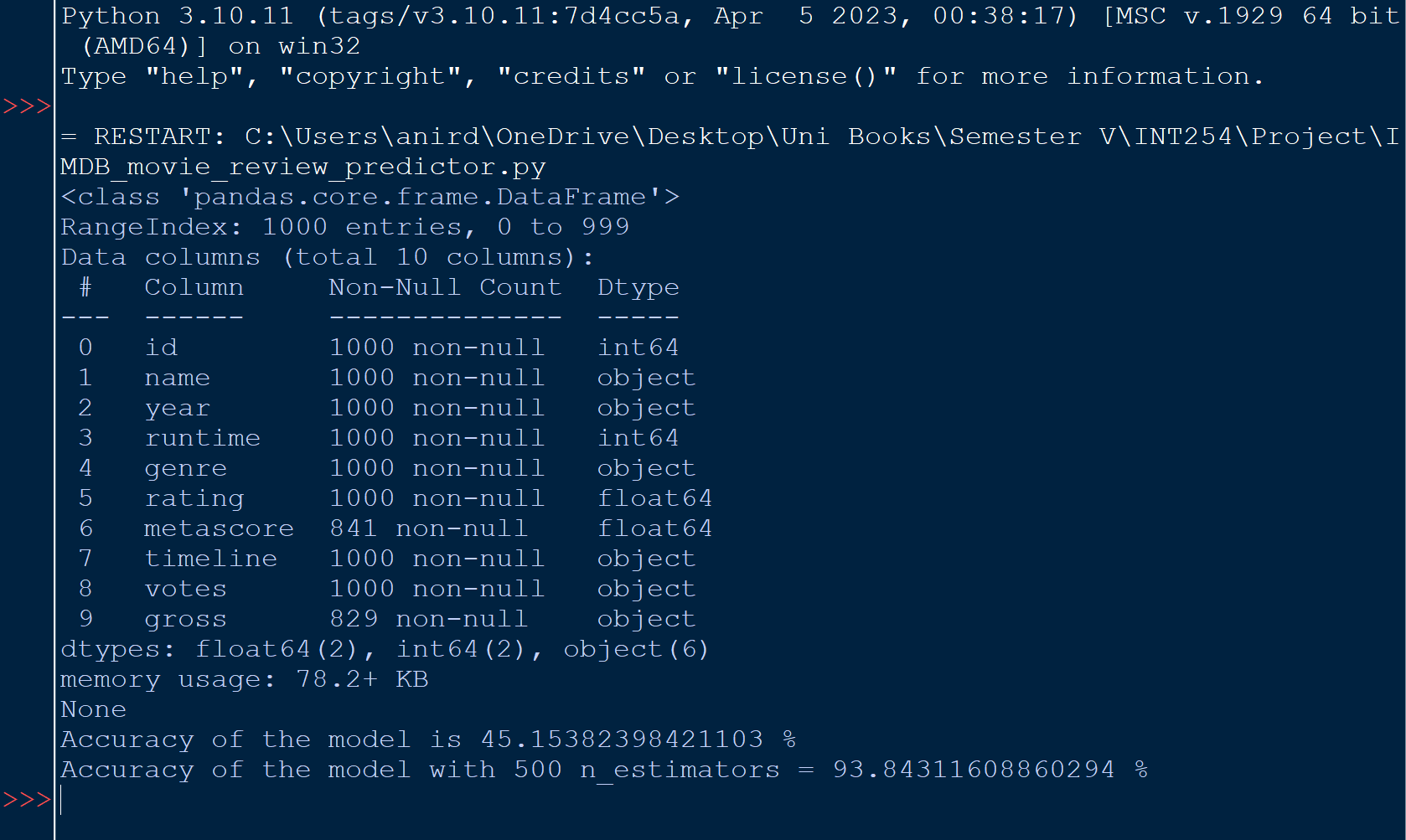




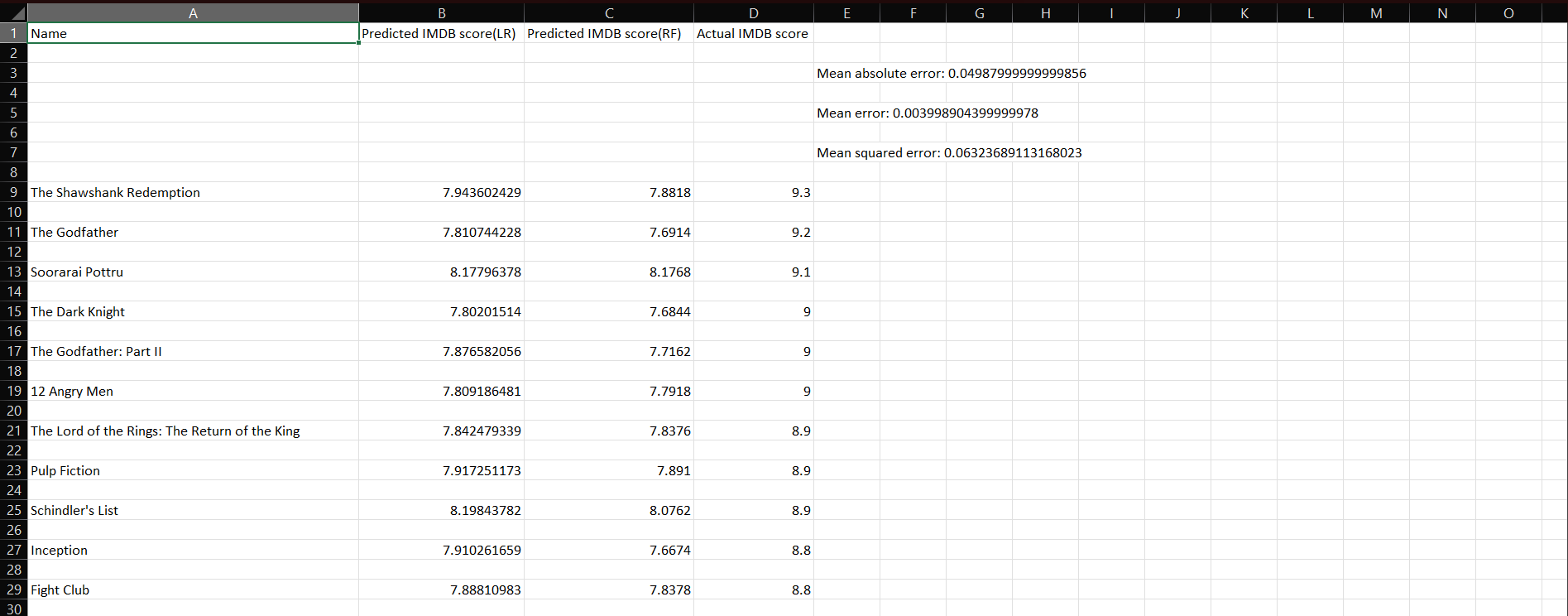
Step 9: (Optional)Save the model for later use, using pickle library



**Conclusion:**

****

**Fig 3**

****

**Fig 4:** CSV file for the outcomes generated by the models

It is evident by this project, that linear regression seems to be less accurate for this kind of dataset which has a ton of features all interdependent and co-dependent on each other, in such cases, Random forest regression shall be used as decision trees are superior in extracting useful features from such kind of labelled data

Overall, this project has been a great learning tool for understanding these two specific regression algorithms, their pros and cons and their strengths in relation to kind of datasets they favour.

…