Department of Computer Engineering

Bu-Ali Sina University

Probability and Statistics for Engineers Course

Implementation of Linear Regression in python

By:

Zahra Sharafian

Reyhaneh sadat Khosravi

Academic Supervisor:

Professor Sara No-doust

Fall 2023

Table of Contents

1.Introduction

Linear regression is a fundamental statistical technique used to model the relationship between a dependent variable and one or more independent variables.
It is a widely applied method in various fields, including economics, finance, biology, and education. The goal of linear regression is to establish a linear equation that best represents the underlying pattern in the data, allowing for the prediction of the dependent variable based on the values of the independent variables.

In this project, we implement linear regression in Python using the graduate-admissions dataset. The dataset contains features such as GRE (Graduate Record Examination) scores, TOEFL (Test of English as a Foreign Language) scores, university rankings, and other factors that are believed to influence graduate admissions outcomes.

Our goal is to build a linear regression model to understand the relationship between specific features and the likelihood of admission to graduate programs. By analyzing and interpreting the coefficients obtained from the model, we aim to identify the factors that play a significant role in influencing the admission process.
In this project we focus on exploring the importance of each feature such as CGPA (Cumulative Grade Point Average) in graduate admissions. We aim to understand how CGPA, as a measure of academic performance, plays a role in the decision-making process for admission to graduate programs.

2.Linear Regression

Linear regression is a statistical method that is used to predict a continuous dependent variable (target variable) based on one or more independent variables (predictor variables). This technique assumes a linear relationship between the dependent and independent variables, which implies that the dependent variable changes proportionally with changes in the independent variables. In other words, linear regression is used to determine the extent to which one or more variables can predict the value of the dependent variable.

Linear regression also has limitations, such as assuming a linear relationship between variables, but it remains a valuable and widely used tool in the machine learning landscape. Its simplicity and efficiency make it an essential algorithm for various regression tasks.

It is a fundamental tool in statistics and machine learning and has various applications in fields such as economics, finance, biology, and engineering.

3.Linear Regression Model Representation

The representation of linear regression is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). both the input values (x) and the output value are numeric.
in a simple regression problem (a single x and a single y), the form of the model would be:

$$y = B0 + B1*x$$

More explanations of the formula are given in the implementation section.

4.Dataset

In this project, we have utilized a dataset related to university admissions based on CGPA, TOEFL scores, motivation letter acceptance rates, and other admission requirements. For ease of use with the desired data, we have changed the names of some columns (such as the "Chance of admit" to “Chance” column) to have a more convenient analysis in the code. On the other hand, for better understanding and accuracy in the visuals and results, we have normalized the acceptance probability between 0 and 1, instead of using a scale of 0 to 100.

A display of data in dataset and their distribution

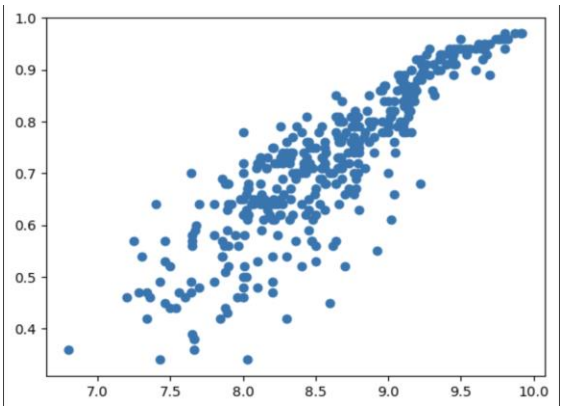

Figure 1: Dataset -> CGPA & Chance

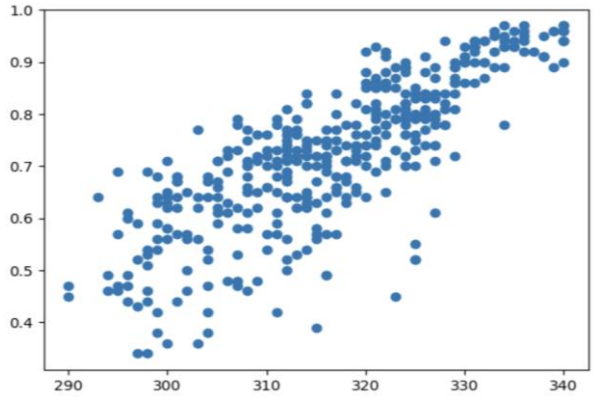

Figure 2: Dataset -> GRE_Score & Chance

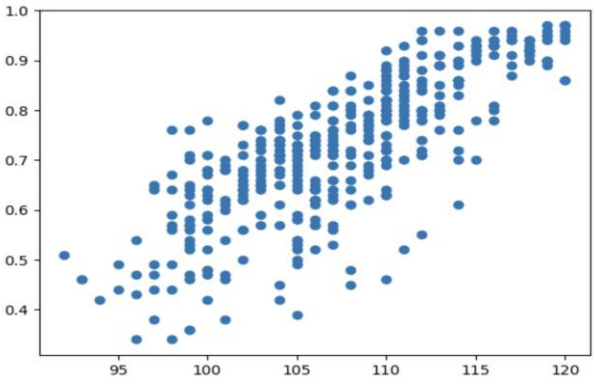

Figure 3: Dataset -> TOEFL Score & Chance

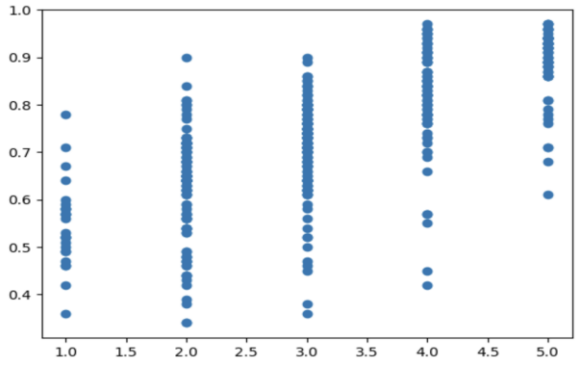

Figure 4: Dataset -> University Rating & Chance

5.Implementation

In summary, the code implements simple linear regression, calculates the slope and intercept, makes predictions and plots the regression line.

First we import necessary libraries (pandas for data manipulation, numpy for numerical operations, and matplotlib for plotting).

Next, the code selects the independent variable (feature) like 'CGPA', and the dependent variable (target) 'chance', from the DataFrame.

Then Calculates the mean of the independent variable (CGPA) and the dependent variable (chance).

Next we calculate the slope (B1) and intercept (B0) using the formulas mentioned earlier. These formulas are based on the least squares method, aiming to minimize the sum of squared differences between the observed and predicted values.

```
# Select features and target variable
X = data['CGPA']
y = data['Chance']

# Calculate the mean of X and y
mean_X = np.mean(X)
mean_y = np.mean(y)

# Calculate the slope (beta1) and intercept (beta0)
numerator = np.sum((X - mean_X) * (y - mean_y))
denominator = np.sum((X - mean_X) ** 2)
beta1 = numerator / denominator
beta0 = mean_y - beta1 * mean_X

# Make predictions
y_pred = beta0 + beta1 * X
```

Figure 5: Algorithm of implementation Linear Regression in Python

5.Result

After implementing linear regression, based on the main function that has been implemented, we calculate the final result of the impact of the CGPA factor on the acceptance rate percentage(Chance of admit) in universities and visualize it on the data within the dataset.

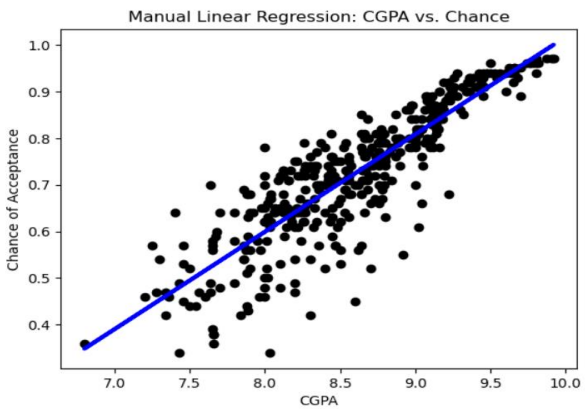

Figure 6:result -> CGPA & Chance

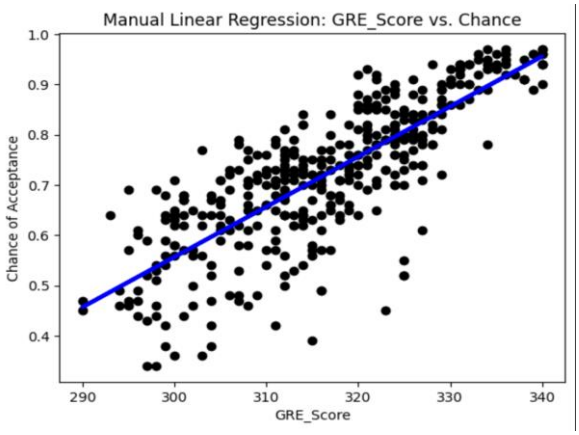

Figure 7: result -> GRE Score & Chance

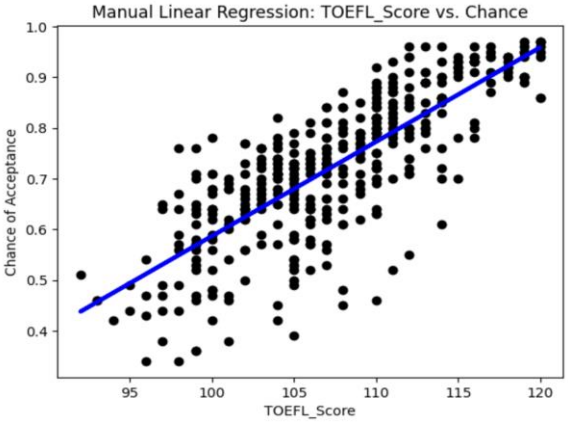

Figure 8:result -> TOEFL Score & Chance

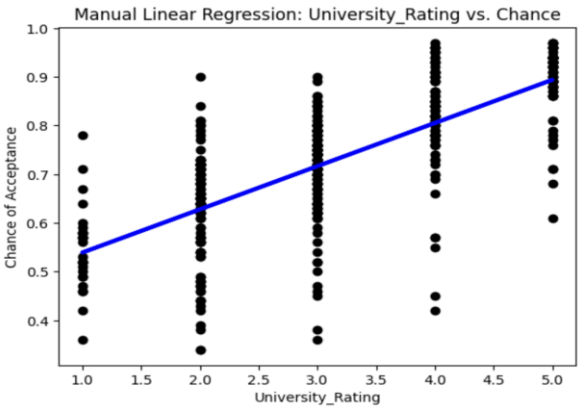

Figure 9: result -> University Rating & Chance

6.published in

6.1.Github

Github Repository

7.Resources

7.1.Helpful links

linear regression for machine learning
Linear-regression-python-implementation
Sklearn source for inspiring implementation of main function

7.2.Searching appropriate datasets for implementing linear regression

Datasets - Linear Regression

7.3.Dataset

Graduate-admissions