**Data Preparation/Feature Engineering**

**1. Overview**

Data preparation and feature engineering are crucial steps for any machine learning project. They involve cleaning and transforming raw data into a format that can be easily understood and processed by machine learning algorithms. Most scientists agree that these steps make up 80 percent of the work involved in a data science project.

**2. Data Collection**

The dataset has 500 instances and 9 attributes. It comes from the website Kaggle. The dataset consists of attributes like Serial No, GRE Score, TOEFL Score, University Rating, SOP (Statement of Purpose), LOR (Letter of Recommendation), CGPA (Undergraduate Cumulative GPA), Research, and Chance of Admit. The data format is in CSV with the size 16KB. This dataset is chosen because in our reference, it uses almost the same thing and the attributes are related to the performance of students who want to apply to a university

**3. Data Cleaning**

The steps taken to clean the raw data are as follows:

- Identify missing values

- check the duplicate values

- check whether the outliers or not in our dataset

- Identify highly correlated variables, variables with nearly no variance

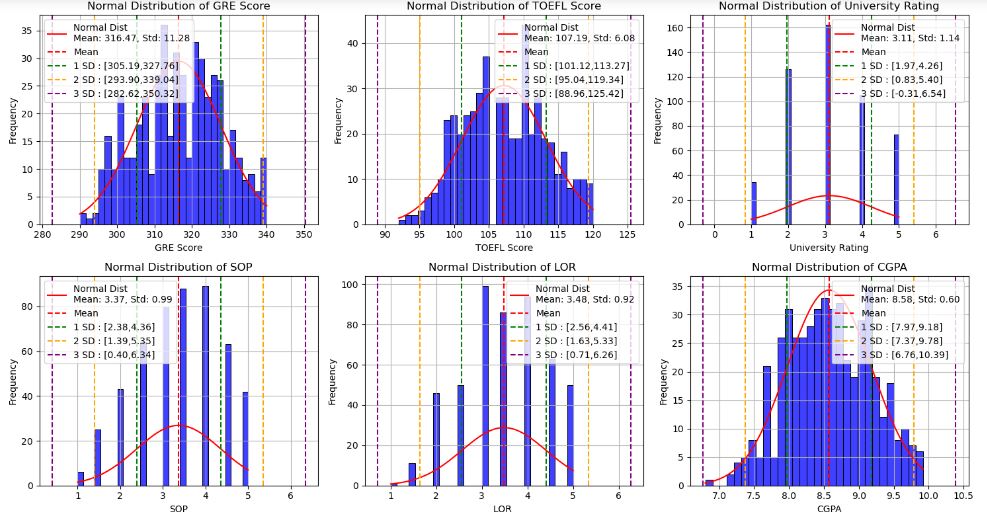
**4. Exploratory Data Analysis (EDA)**

Summarize the exploratory data analysis performed on the dataset. Include visualizations and

key insights gained during this phase.

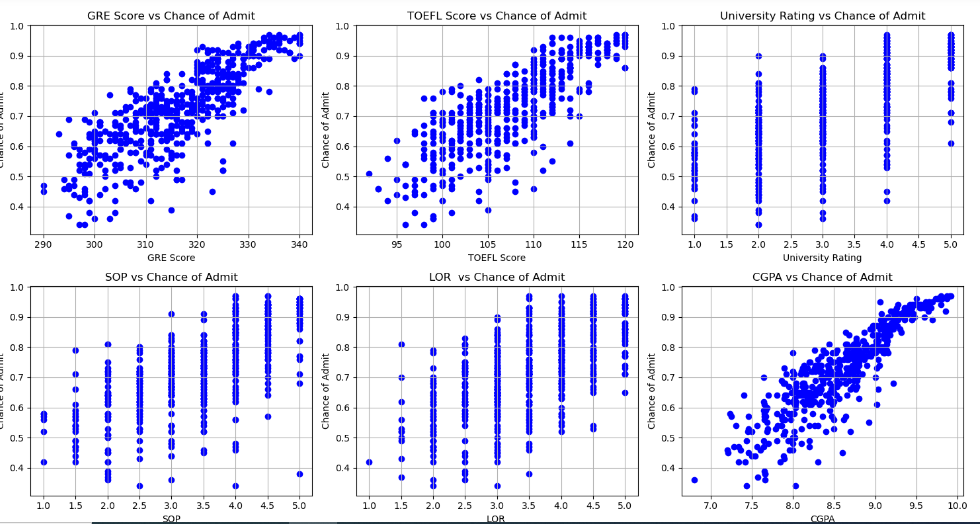
A comprehensive Exploratory Data Analysis was conducted on the dataset to gain insights into its underlying structure, variable distributions, and the relationships among the features.

- Check the normal distribution of each column



Scatterplot:

This is especially useful for visualizing interactions between variables, including linear and non-linear relationships.



**5. Feature Engineering**

For the feature engineering, we used scaling and normalizing the data to transform the features to a similar scale. We also utilized the PolynomialFeatures function from the scikit-learn library to create polynomial features.

**6. Data Transformation**

The data scaling is used to plot the normal distribution curve.

After normalization, we noticed that the difference is almost insignificant, and the metric R² score for the decision tree has decreased.

**Model Exploration**

**1. Model Selection**

Multiple linear regression is a statistical technique utilized to predict the value of a dependent variable by employing multiple independent variables. The purpose of MLR is to construct a model that represents the linear correlation between the independent variables x and dependent variable y, which will then be examined.

Y = β0 + β1x1 + β2x2 + β3x3 + β4x4 + . . . + βnxn

The purpose of linear regression is to predict the output variable Y based on the input variable(s) X. Given a new value of X, the linear regression model can compute a predicted value of Y by applying the learned coefficients β0, β1, β2, …βn.

Linear regression uses a method called **Ordinary Least Squares (OLS)** to estimate the coefficients β0, β1, β2, …βn ​, etc. The goal is to minimize the difference between the actual values of Y and the predicted values by minimizing the **sum of squared residuals (errors).**

**Cost function (MSE)**

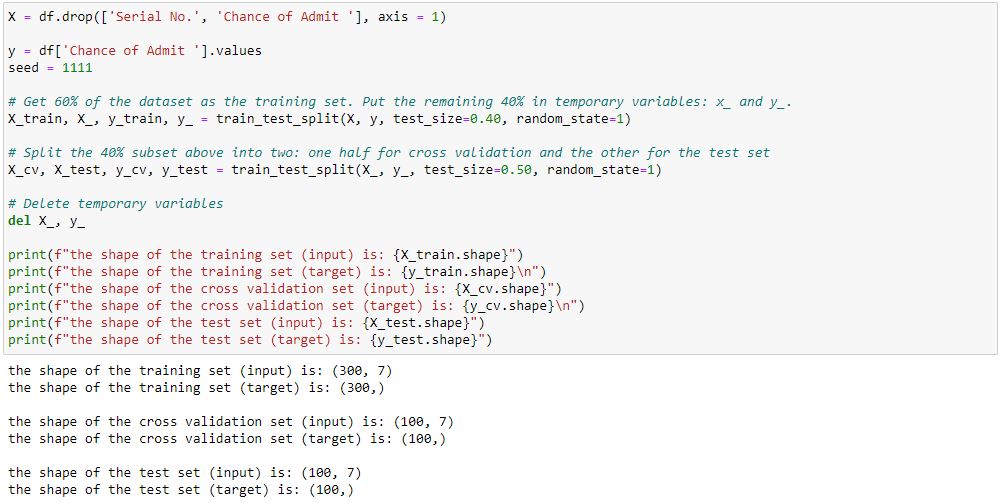
 ​ is the actual value of the dependent variable for the i-th observation.

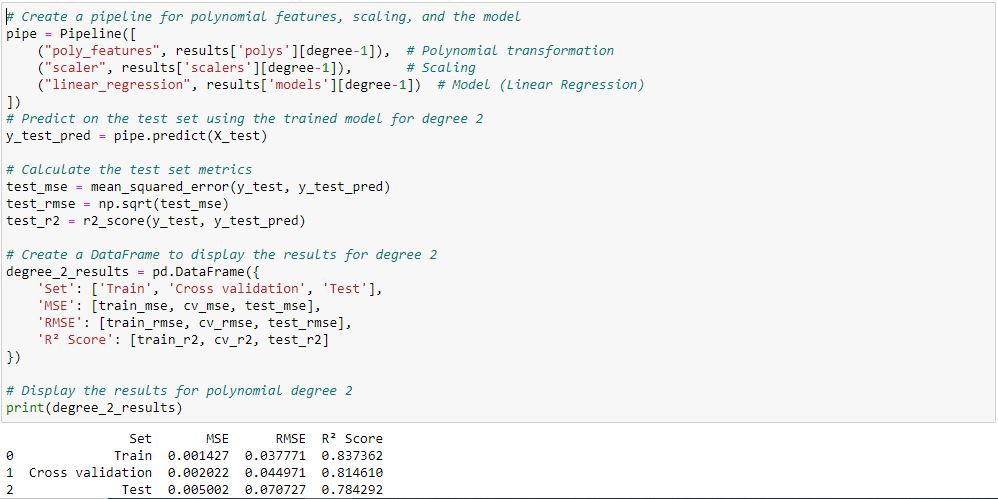
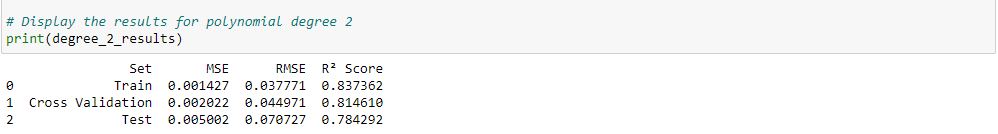
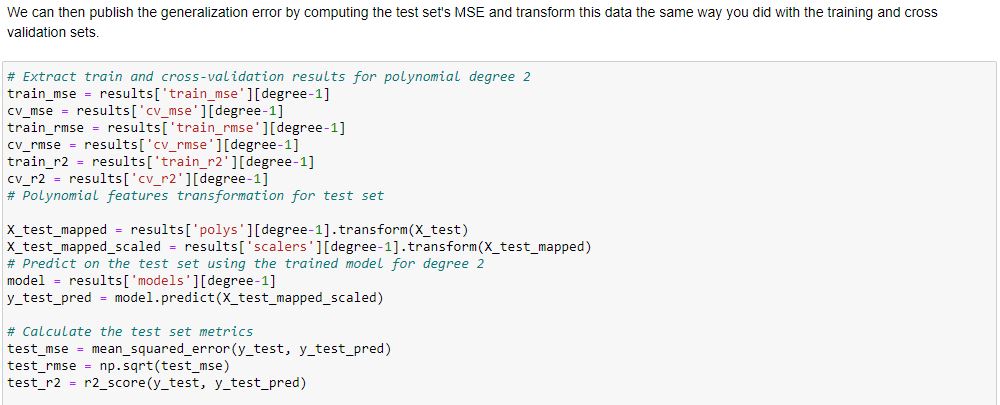
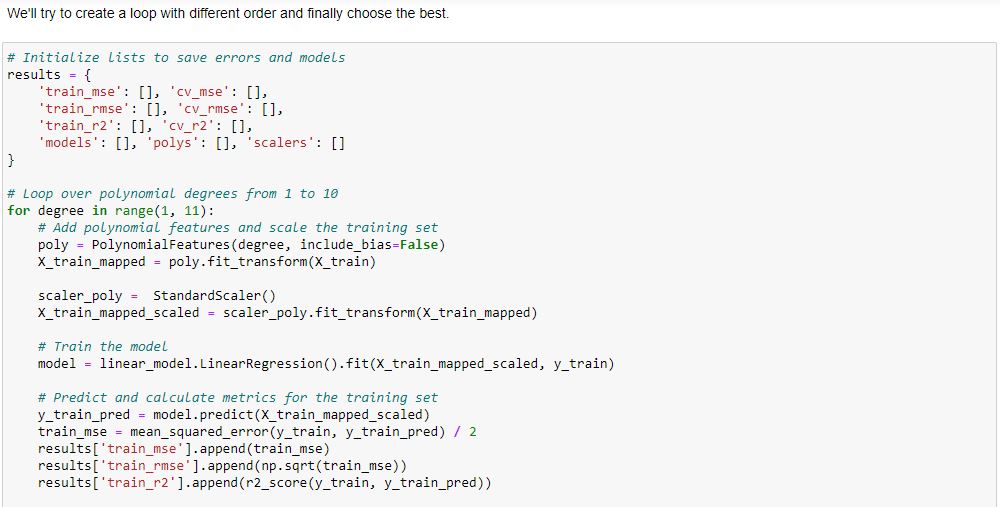
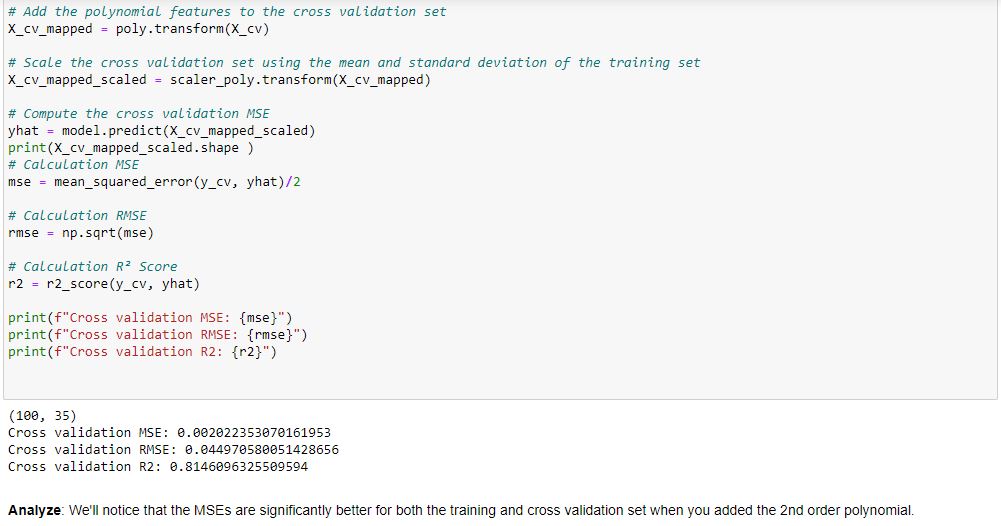
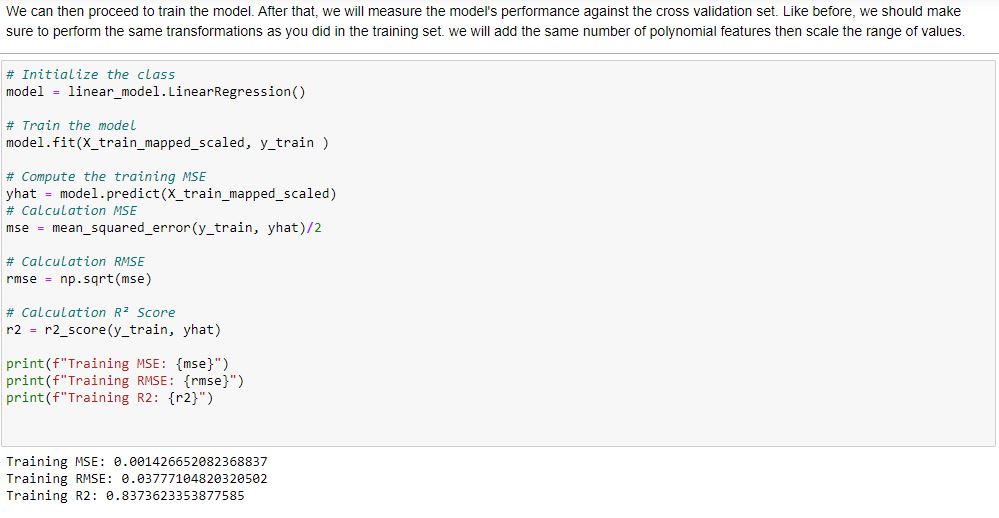
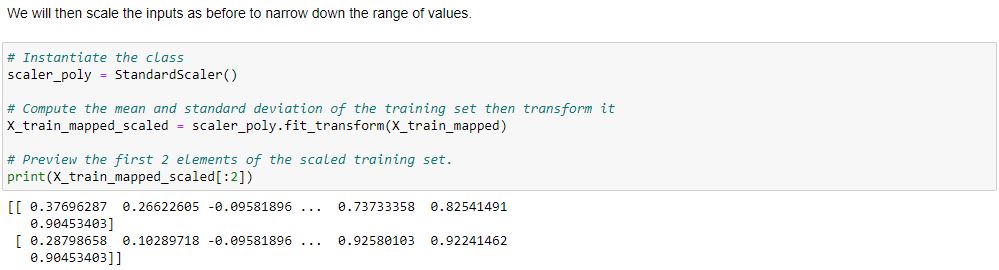
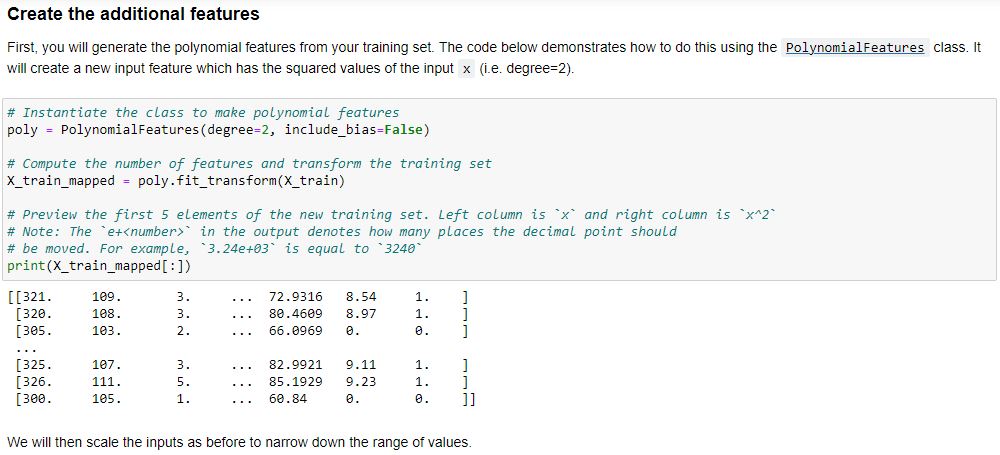
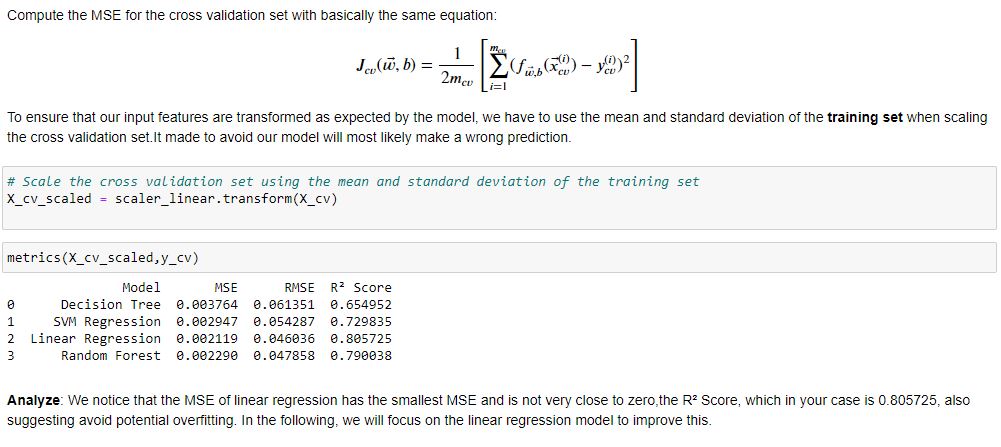
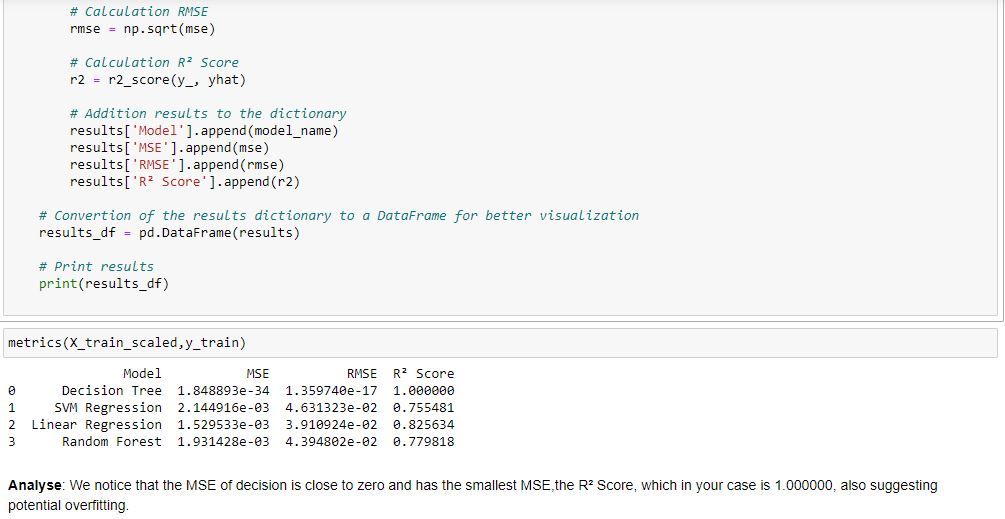
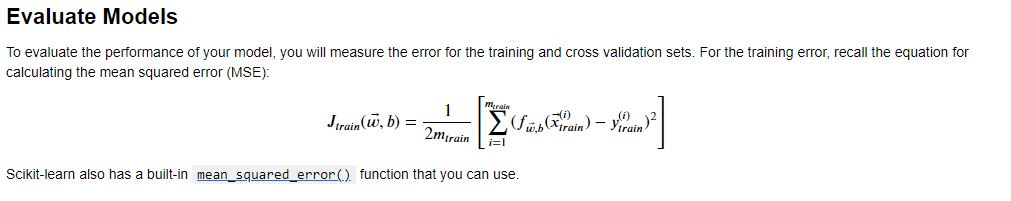
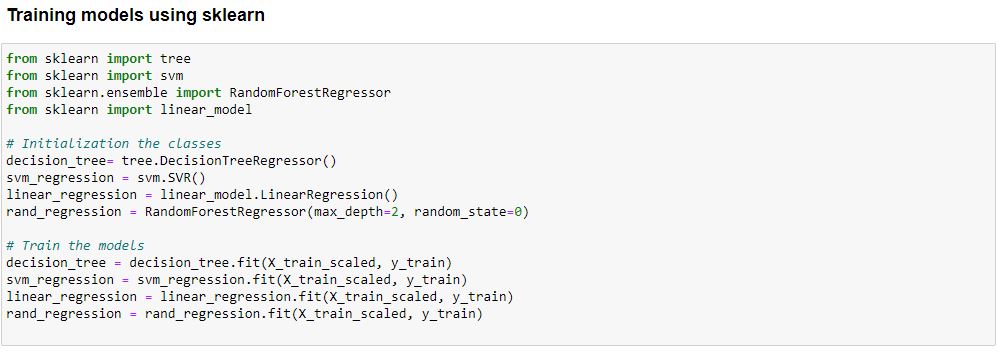
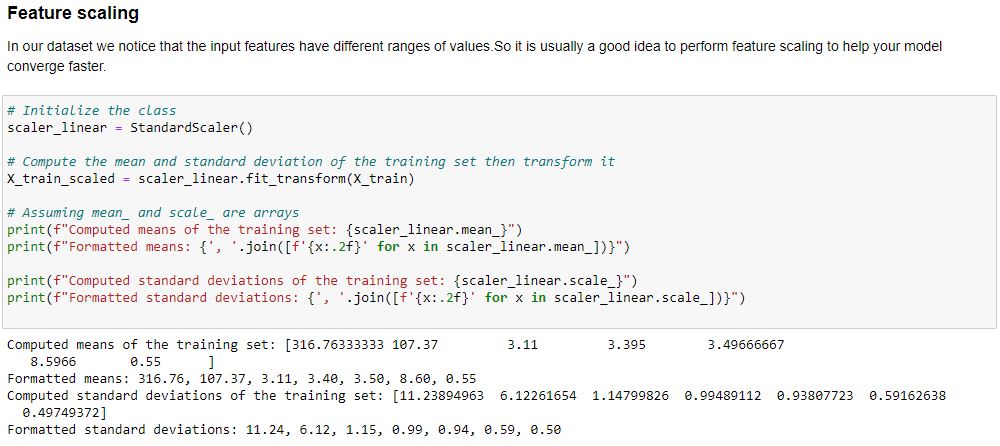
 is the predicted value from the linear model.

|  |  |
| --- | --- |
| **Strengths** | **Weaknesses** |
| Linear regression is straightforward to understand and explain, and can be regularized to avoid overfitting. In addition, linear models can be updated easily with new data using [stochastic gradient descent](https://www.quora.com/What-is-an-intuitive-explanation-of-stochastic-gradient-descent). | Linear regression performs poorly when there are non-linear relationships. |

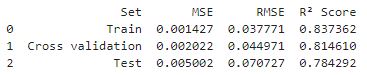
**2. Model Training**

Training model:





**3. Model Evaluation**



 Training **set**:

* The Mean Squared Error (MSE) of 0.001427 and Root Mean Squared Error (RMSE) of 0.037771 indicate minimal error.
* The R² score of 0.837362 suggests that the model explains approximately 83.7% of the variance in the target variable, indicating a robust fit on the training data.

 Cross**-validation set**:

* The MSE of 0.002022 and RMSE of 0.044971 show slightly higher error compared to the training set, which is typical when assessing a model’s generalizability.
* The R² score of 0.814610 suggests that the model maintains strong explanatory power with 81.5% of the variance explained in unseen data.

 Test **set**:

* The MSE of 0.005002 and RMSE of 0.070727 indicate an increase in error, which is expected when testing the model’s performance on completely unseen data.
* The R² score of 0.784292 implies that the model still explains around 78.4% of the variance in the test set, affirming its predictive capability even on new data.

**4. Code Implementation**

