Learn to Build a Polynomial Regression Model from Scratch

Overview

Polynomial regression is a form of regression analysis in which the relationship between the dependent variable and independent variable is mapped using n^{th} degree polynomial. If we apply linear regression to the linear dataset, then it gives good results. On the other hand, if we apply linear regression to non-linear data then the results are drastic. So, for such cases, we need polynomial regression which will capture the non-linear relationship in the data. Hence, in polynomial regression, the original features are converted into polynomial features of the required degree and modeled using the linear model.

In this project, we will be building a polynomial regression model to predict the points scored by the sports team. The project will give in-depth intuition about the business problem, exploratory data analysis, data preprocessing, and model building. The project will also cover a detailed explanation of various regression metrics.

Aim

- To predict points scored by the sports team using polynomial regression

Data Description

The dataset contains information about the points scored by sports teams based on various attributes.
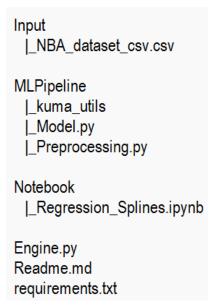
Tech Stack

➔ Language: Python
➔ Libraries: pandas, numpy, scipy, matplotlib, seaborn, sklearn, statsmodel

Approach

- Data Preprocessing
  - Outlier removal
  - Imputing null values
  - Onehot encoding

- Model Building
  - Linear regression model building

- ○ Polynomial regression model building

- ● Model Evaluation
  - ○ Evaluation of model on test data
  - ○ Discussion on various regression matrix-like R-squared, AIC,AICC, F-statistics

Modular Code Overview

```
Input
  |_NBA_dataset_csv.csv

MLPipeline
  |_kuma_utils
  |_Model.py
  |_Preprocessing.py

Notebook
  |_Regression_Splines.ipynb

Engine.py
Readme.md
requirements.txt
```

Once you unzip the polynomial_regression.zip file, you can find the following folders within it.

1. Input
2. ML_Pipeline
3. Notebook
4. Engine.py
5. Readme.md
6. requirements.txt

1. The Input folder contains the data that we have for analysis. In our case, it is the NBA dataset

2. The Notebook folder contains the jupyter notebook file of the project

3. The ML_pipeline is a folder that contains all the functions put into different python files, which are appropriately named. These python functions are then called inside the Engine.py file

4. The requirements.txt file has all the required libraries with respective versions. Kindly install the file by using the command **pip install -r requirements.txt**

5. **All the instructions for running the code are present in Readme.md file**

Takeaways
1. What is a distribution plot?
2. What is a boxplot?
3. What is a violin plot?
4. How to detect outliers?
5. How to treat outliers?
6. What is pandas imputer?
7. What is iterative imputer?
8. What is a KNN imputer?
9. What is an LGBM imputer?
10. Univariate analysis
11. Chatterjee correlation
12. What is ANOVA?
13. Implementation of ANOVA
14. Data preprocessing
15. What is AIC?
16. What is likelihood?