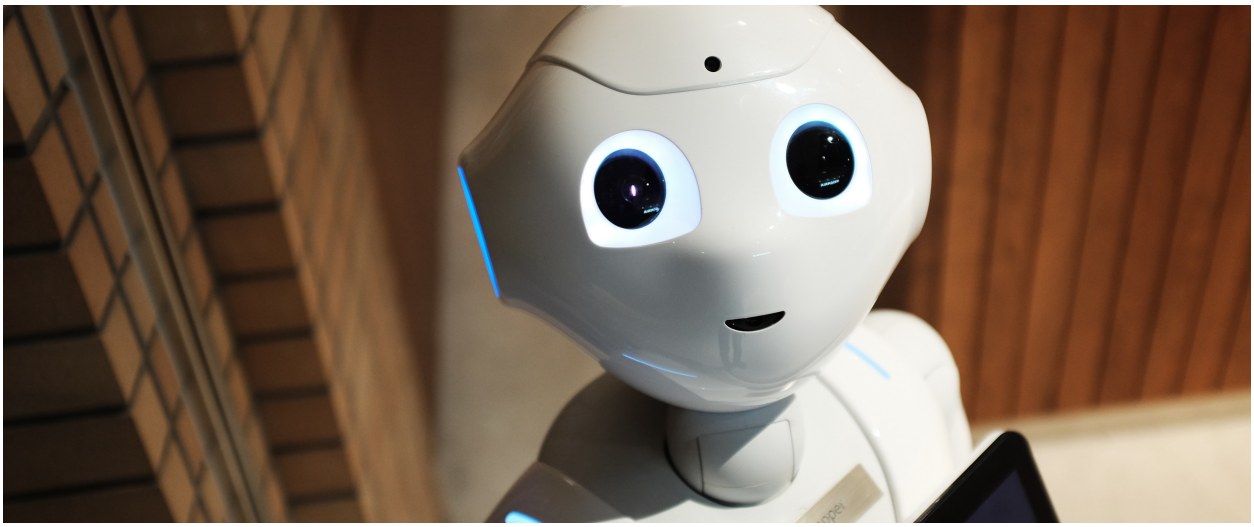


MDL - Assignment 1

- Team Name : 404-TeamNotFound
- Team Number : 6
- Team Members :
  1. `Aditya Malhotra` : 2020101052
  2. `Karmanjyot Singh` : 2020101062

Task 1 : Linear Regression

> Write a brief about what function the method `LinearRegression().fit()` performs.

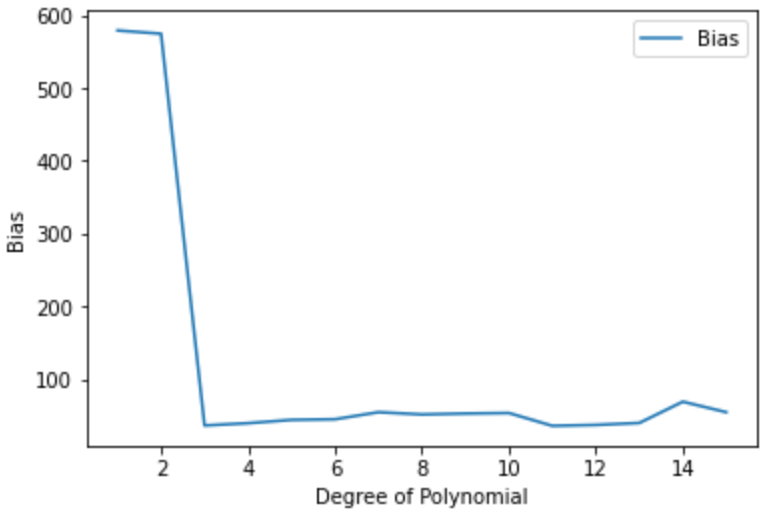
- given the data set , **x_train** and **y_train** , then we can generate the best fit model using this library function as :
  - `LinearRegression(fit_intercept=False).fit(x_train , y_train )`
  - **Note** : `x_train` is the generated polynomial feature for the actual `x` data set value , which is generated using the `PolynomialFeatures()` method of sklearn library

- The function `LinearRegression().fit()` returns a machine learning model such that fits the given data set (which is passed into the function as parameter) while minimizing the sum of the squares of the predicted value and the actual value for the given data set , and thus generating the **best-fit** linear model for the given data set.
- It generates a line y = mx + c (initializes the coefficient) such that the generated line , best-fits the given data set .

Task 2 : Calculating Bias and Variance

Bias

- It is defined as the difference between the average prediction of our model and the correct value which we are trying to predict
- It is a measure of how much the value predicted by our model varies from the actual value , a **High Bias** describes huge variation of the predicted values from the actual value of the data set.
- The following Plot describes the variation of the bias with the degree of the polynomial functions



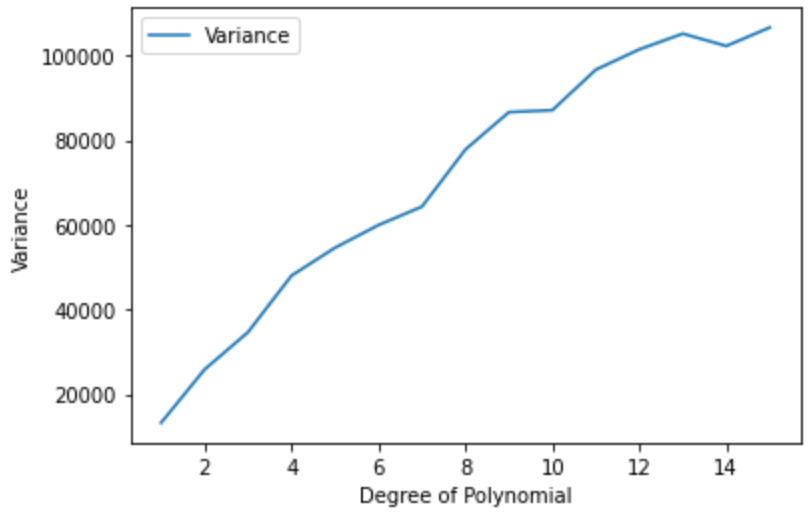
Plot : Bias vs Degree of the polynomial

- From the plot , we observe that there is a sharp decrease in the calculated bias value from degree=2 to degree = 3 , for the given data set , meaning thereby the

lower degree polynomials (of degree < 3) were **under fitting** the given data set and thus leading to a unusually **high bias** for them , and thus the predicted values of the model varied greatly from the actual value of the test training set

Variance

- Variance is used to measure the spread of data . A higher value of variance means , a high variation of the predicted value around the actual value of the dataset.
- Which is justified as , due to the increasing degree , and hence greater complex polynomial terms , and thus greater variation.



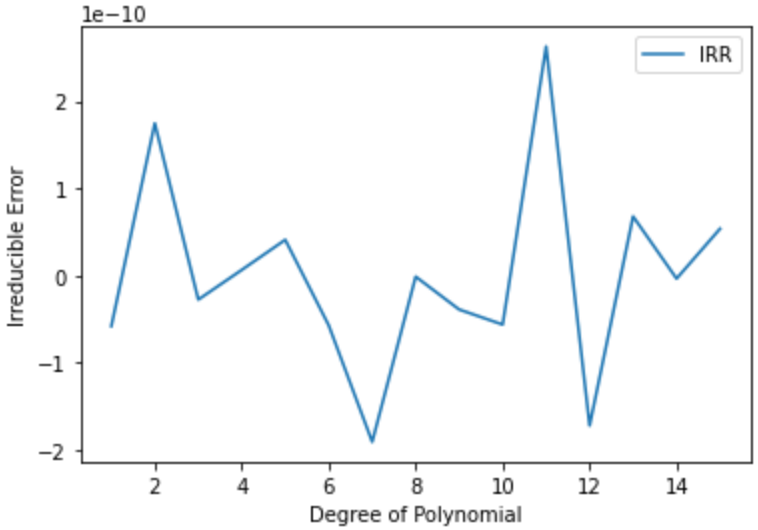
Plot : Variance vs Degree of polynomial

degree	bias	variance
1	578.834	13188.5
2	574.428	25798.8
3	36.3969	34620.7
4	39.5212	47992.5
5	44.0603	54580.9
6	44.9234	59945
7	54.641	64305
8	51.5572	77804.3
9	52.7812	86644.9
10	53.506	87112.9
11	35.9098	96692.2
12	37.1573	101484
13	39.8405	105196
14	69.0515	102285
15	54.6107	106659

Data : Bias and Variance of the trained models

Task 3 : Calculating Irreducible Error

- Irrespective of how good the model is trained , there always exist some error which can't be removed , which is attributed to the noise in the data (i.e. the external parameters apart from the existing parameters in the data set that determine the output , and thus cause errors).



Plot : irreducible error vs degree of polynomial

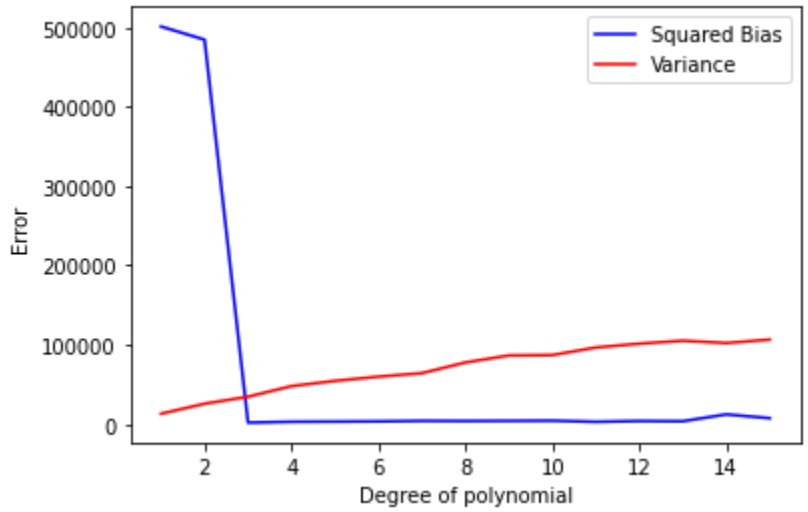
degree	Mean Sq Error	irreducible error
1	514365	-5.82077e-11
2	510444	1.74623e-10
3	36581.3	-2.77396e-11
4	51076.8	6.36646e-12
5	57909.4	4.09273e-11
6	63523.6	-5.72982e-11
7	68474.3	-1.90994e-10
8	81844.6	-1.36424e-12
9	90818.1	-3.91083e-11
10	91539.3	-5.63887e-11
11	99745.2	2.62844e-10
12	105507	-1.72349e-10
13	108925	6.77574e-11
14	114638	-3.63798e-12
15	114222	5.36602e-11

Plot : Data , MSE and Irreducible error for given data set

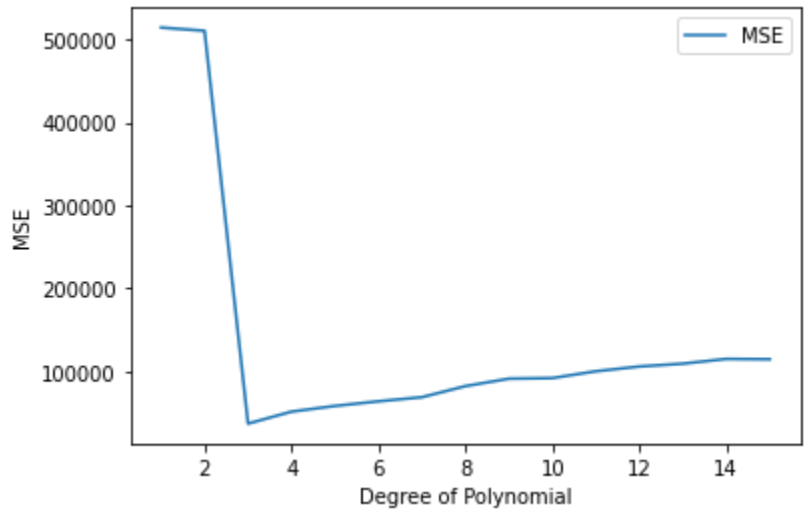
Task 4 : Plotting Bias^2 and Variance

- As we observe , the value of variance and bias^2 decreases steeply from degree = 1,2 to degree = 3 , which could be explained as the , linear and quadratic polynomials **under fits** the given data and thus high value for bias and variance .
- We could observe that , the bias and variance are minimized (global minima) for degree 3 (cubic) polynomial , which thus is the **best fit** for the curve , which is also justified as for **degree = 3** , the value of MSE (Mean Square Error) is the least and hence , it is the best fit polynomial class function for our given data set . Thus best fit model will be of the type ax^3 + bx^2 + cx + d , whose coefficients could be determined using the training data set

- As we see further , as the degree of polynomials increases , the generated model starts **over fitting** the training data , and thus performs poorly on the test data (and hence high variance)



Plot : Variance and Squared Bias



Plot : MSE vs Degree of polynomial

Actual Value V/S Predicted Values

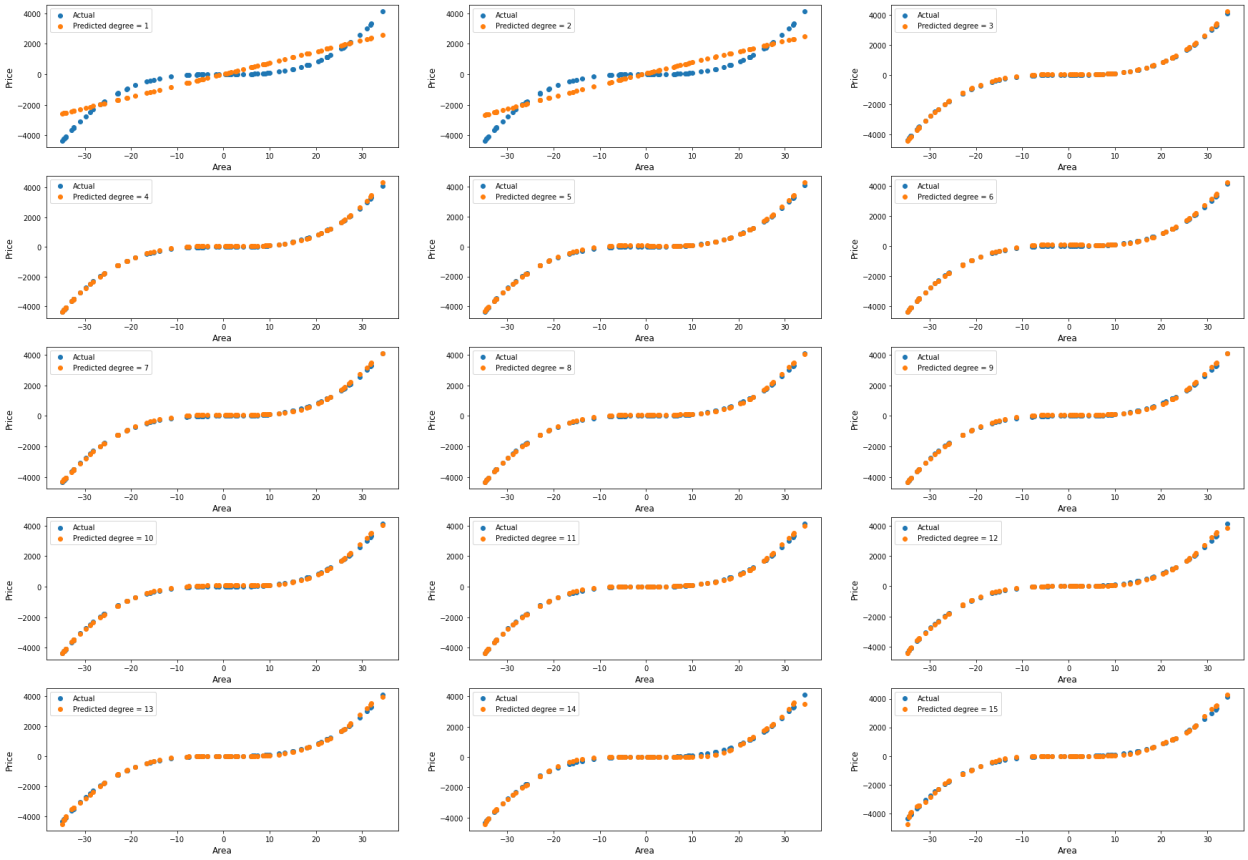


Figure : Actual value vs Predicted Model