MDL Assignment 1

Team Name: Alpha Come

***Team No*: 9**

***Team Members*:**

- Ayush Agrawal (2020101025)
- Shreyansh Agrawal (2020101013)

Task 1: Linear Regression

`LinearRegression().fit()` is a function from the scikit-learn library that is used to calculate the coefficients for predicting the model for a data set using linear regression. Given a dataset of x and y, it tries to find a and b such that the sum of the square of distances of predicted value of y and real value of y is minimum.

It tries to fit the function

$$y = ax + b$$

to the given dataset. The function implements the Ordinary Least Squares method to estimate the unknown parameters in the linear regression model. The least-squares solution is computed using the singular value decomposition of matrices.

This function can be used with `PolynomialFeatures().fit_transform()` to get the coefficients for a n-degree polynomial of the form -

$$y = a_n x^n + a_{n-1} x^{n-1} + a_{n-2} x^{n-2} + \cdots + a_1 x + a_0$$

Task 2: Calculating Bias and Variance

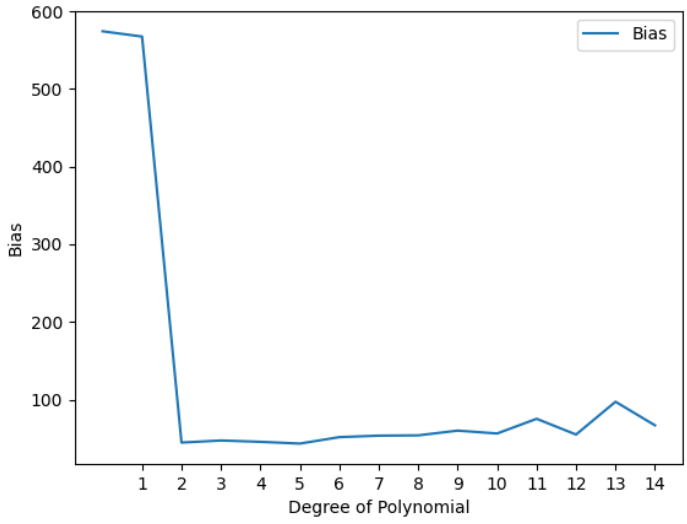
Degree	Bias	Variance
1	574.003	33228.8
2	567.017	50617.3
3	47.6863	74968.6
4	54.7216	97726.5

Degree	Bias	Variance
5	53.8527	103737
6	51.2185	115234
7	52.9566	135544
8	52.0966	153593
9	51.0936	179925
10	54.378	191473
11	57.653	215707
12	71.509	222969
13	56.8478	219251
14	91.0967	232571
15	68.3832	227581

Bias:

Bias is a measure of accuracy of predictions made by the model. If a model has high bias then it means prediction made by the model are inaccurate. Mathematically, it is defined as the mean of difference between correct value and predicted value.

$$Bias = E[\hat{f}(x) - f(x)]$$



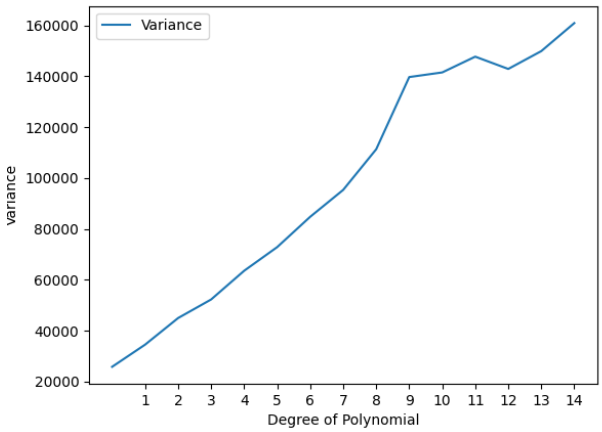
In our models, we observe that there is a sharp decrease in bias from degree 2 to 3 and thereafter, it remains almost constant with a slight increase. The high bias of degree 1 and 2 polynomials can be attributed to the phenomenon of underfitting. The degree 1 and 2 polynomials are too simple which do not consider the variations in data well. They make predictions which are based on assumptions which do not fit the given data.

The degree 3 polynomial has the lowest bias which means it suits the given data best. It captures the dependence of dependent variable(Area of House) on independent variable(Cost of House) very well. It fits the data best and generalizes well to data not yet seen by the model before. The bias then increases gradually because the higher degree polynomials are too restrictive for the data. This is known as overfitting where the model extracts too much information from training data but fails to correctly predict the test data values.

Variance:

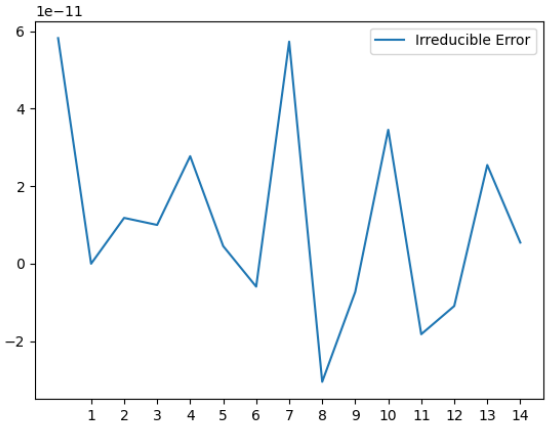
Variance is a measure of precision of the predictions made. In other words, it measures the spread of the prediction values. It has no relation with the actual values but how close or far are the predictions relative to themselves. If a model has high variance, the model captures the noise from input data. Mathematically, it is defined as:

$$Variance = E[\hat{f}^2] - E[\hat{f}]^2$$



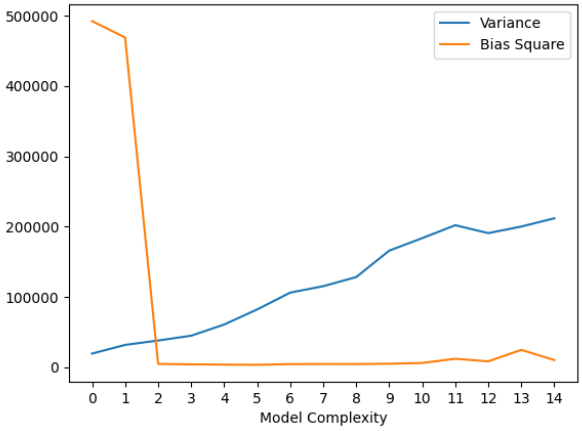
In our models, we observe that the variance of models increases as the degree of polynomials increase. The higher complexity of the models attempts to fit the noise that occurs in the training data. The degree 1 and 2 polynomials have relatively low variance due to smaller number of features. As the number of features increase, the predicted values become more scattered which explains their high variance values.

Task 3: Irreducible Error



As can be clearly seen by the graph of irreducible error vs. model complexity, there is no clear pattern between the two, which is pretty intutive considering that the irreducible error is random noise and is independent of the model or regression technique used. The irreducible error is close to zero implying that the given data has very little noise.

Task 4: *Bias^2 vs Variance*



As is clearly evident from the graph, the bias^2 is very high in the begining beacuse of underfitting, but variance is low. Whereas, with increase in the degree, the variance increases and bias decreases as the model is now overfitting. The sweet spot is at degree 3 as the total error is least for the model with degree 3. While the bias of for further degrees until degree 10 stays low, the variance steadily increases; hence the total error increases too. The optimal model for this type of data is a cubic function hence, we can say that the dependent variable is order of degree 3 of independent variable.