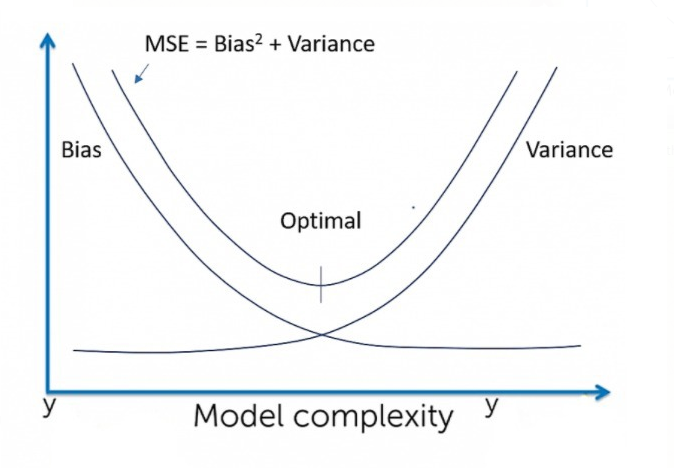
**Measure of Bias and Variance – An Experiment**

**Introduction**

One of the most used matrices for measuring model performance is the predictive errors. The components of any predictive errors are Noise, Bias and Variance. Intent of this article is to measure the bias and variance of a given model and observe the behaviour of bias and variance w.r.t various model such as Linear Regression, Decision Tree, Bagging and Random Forest for a given number of sample sizes.



**Prerequisite**

1. Understanding Bias and Variance
2. Algorithms such as Linear Regression, Decision Tree, Bagging with Decision Tree, Random Forest and Ridge Regression

**Brief of Bias and Variance**

**Bias**: Difference between the prediction of true model and the average models (models build on n number of samples obtained from the population).

True Model: Model build on a population data

Average Model: Average of all the prediction results obtained from various sample obtained from population model.



**Variance**: Difference between the prediction of all the models obtained from the sample with the average model.

where, 

 - Returns the predicted value of x for a population model

 - Returns predicted values of x using a model built on a sample

 - Returns the average of predicted values using n number of models built on n different samples

**Noise**: It is the irreducible error that a model cannot predict.

**Expected behaviour**

|  |  |  |
| --- | --- | --- |
| Algorithm | Bias | Variance |
| Linear Regression | High Bias | Less Variance |
| Decision Tree | Low Bias | High Variance |
| Bagging | Low Bias | High Variance, lesser than Decision tree |
| Random Forest | Low Bias | High Variance, lesser than Decision tree and Bagging |

**Experiment**

Practically it is very difficult and expensive to obtain a population data. Without the knowledge of population data, it is not possible to obtain model bias and variance. So, in order to perform this experiment, we will consider a large dataset to be population. Based on this assumption we will proceed in calculating the bias and variance of various model on this dataset.

**About Data:**

For this example, I am considering a random dataset (dataset is not picked by any criteria).

Dataset : <https://archive.ics.uci.edu/ml/datasets/Physicochemical+Properties+of+Protein+Tertiary+Structure>

This is a data set of Physicochemical Properties of Protein Tertiary Structure. The data set is taken from CASP 5-9. There are 45730 decoys and size varying from 0 to 21 Armstrong.

Attribute : RMSD-Size of the residue.  
F1 - Total surface area.  
F2 - Non polar exposed area.  
F3 - Fractional area of exposed non polar residue.  
F4 - Fractional area of exposed non polar part of residue.  
F5 - Molecular mass weighted exposed area.  
F6 - Average deviation from standard exposed area of residue.  
F7 - Euclidian distance.  
F8 - Secondary structure penalty.  
F9 - Spatial Distribution constraints (N,K Value).

This dataset contains 45730 number of records.

**Population\_Data :**

Superset of all data (Practically it’s not possible, but for the sake of experiment we are considering a large set of data to be our population). In this experiment we consider a data set with 45730 records as population.

**Test\_Data:**

1500 records are extracted from the Population\_Data as Test data.

**Training\_Data:**

Data from the population other than test data are considered as Training\_Data.

**Population\_Model:**

The model built on the Population\_Data.

**Mean\_Model:**

Consider the ‘n’ number of samples being extracted from Training\_Data. We build models on each of these samples. For a given value of x , the mean prediction of these models are considered as predictions of the Mean\_Model for that value of x.

**Model\_Bias:**

Bias for each value of x from test data =

(Prediction of Population\_Model - Prediction of Mean\_Model).

Bias of the model =

Mean (abs (Prediction of Population\_Model - Prediction of Mean\_Model))

**Model\_Variance :**

Variance of the model = Variance (Prediction of Mean\_Model, Prediction of Sample\_Model)

i.e. Difference between the prediction of each models obtained from various samples and the prediction value of the mean model followed by squared and mean of the obtained value. Gives us the info of how much the models from the sample vary from the mean model.

**Steps undertaken for the experiment (Pseudo code):**

1. Considering a data set of 45730 as Population\_Data
2. Extracting Test\_Data of 1500 records from Population\_Data. So, the remaining data is considered to be Training\_Data
3. Build Population\_Model. Collect predictions from the Population\_Model using Test\_Data
4. Build Mean\_Model.

30 random samples have been extracted from Training\_Data. Models are built on each of these samples. The mean predictions using Test\_Data of these models are collected.

1. Compute Model\_Bias

Bias of the model =

Mean (abs (Prediction of Population\_Model - Prediction of Mean\_Model))

1. Compute Model\_Variance:

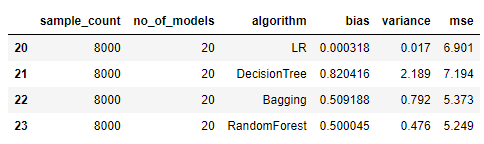
Model\_Variance =

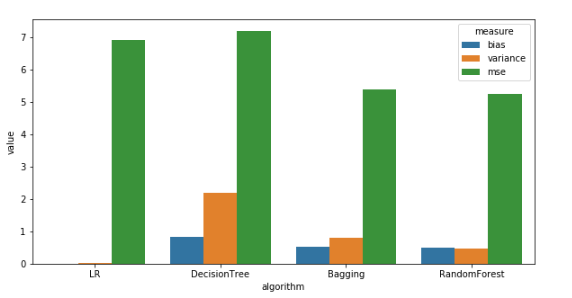
Var (Prediction of Mean\_Model, Prediction of Sample\_Model)

The Model\_Bias and Model\_Variance are being collected for different algorithms such as Linear Regression, Decision Tree, Bagging, Random Forest.

**Results of the Experiment:**

**Bias and Variance for sample size of 8000**

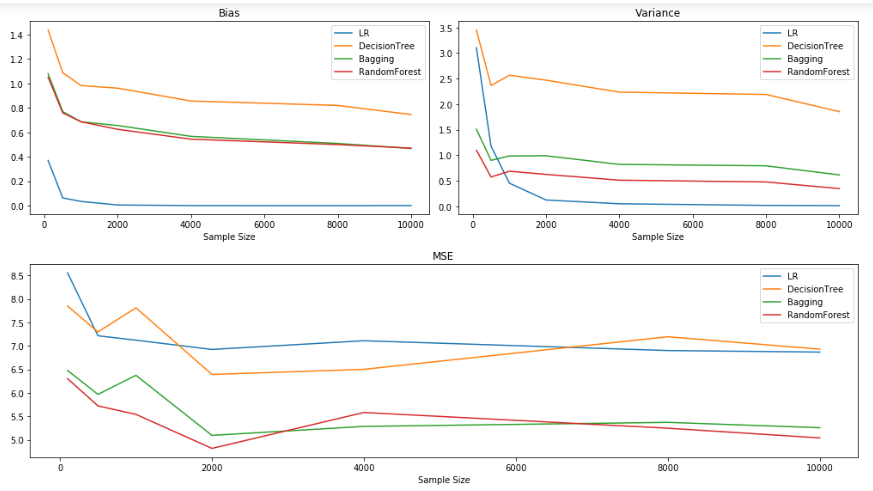




**Observations:** (For sample size of 8000)

* Linear Regression has the least Bias which is not as expected. This hints us that the data is more suited for Linear Regression.
* Variance: Linear Regression < Random Forest < Bagging < Decision Tree, which is as expected.
* Bias: Random Forest < Bagging < Decision Tree, which is also as expected.

**Bias and Variance for sample sizes:[100, 500, 1000, 2000, 4000, 8000, 10000]**

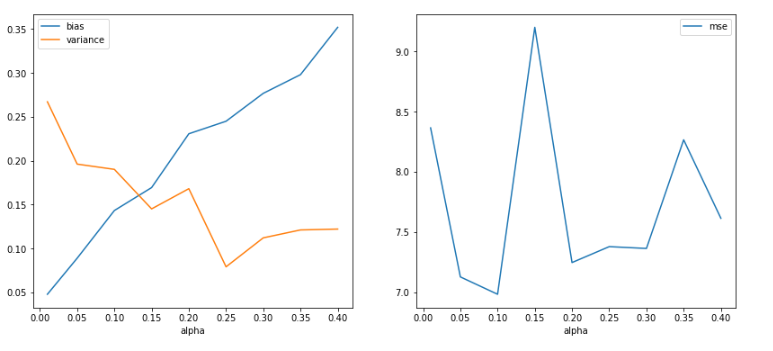


**Observations:**

* It could be observed that the increase in the sample size aids in decrease in Bias and Variance. But often it is quite expensive to obtain data with higher sample size. So, increasing the sample size might not be a viable solution for reducing bias and variance of the model.

**Ridge Regression for a sample of 1000**

* Observing changes in Bias and Variance with various values of alpha for a sample of 1000.
* Visually enables us to view the bias and variance trade of point. In ideal scenario, the alpha value is a tuning parameter i.e. model is trained with various values of alpha and cross validation scores is recorded and the alpha which gives us the best score is selected for the best model.
* As per the figure below, it can be observed that with the increase in the value of alpha, bias value increases and the variance decreases. (This graph is in line with the graph portrayed in book ISLR, Chapter 6 : Linear Model Selection and Regularization, Figure 6.17)



**Conclusions & Observations**

* Computation of Bias and Variance of a model is not feasible in an ideal scenario.
* By conducting an experiment in a controlled environment helped us realize model’s (Linear Regression, Decision Tree, Bagging Decision Tree, Random Forest) behaviour w.r.t various sample sizes of data.
* High sample sizes and good representation of sample over population will help us obtain the required bias variance trade off point.
* Observation of variation in bias and variance for different values of alpha in Ridge regression.

**References**

1. James, G.; Witten, D.; Hastie, T. & Tibshirani, R. (2013), An Introduction to Statistical Learning: with Applications in R, Springer.
2. Markgraf, Bert. "How to Calculate Bias" sciencing.com, <https://sciencing.com/how-to-calculate-bias-13710241.html>. 8 September 2020.
3. Srivastava, P. (2018, September 23). *End your bias about bias and variance!!* Medium. <https://towardsdatascience.com/end-your-bias-about-bias-and-variance-67b16f0eb1e6>
4. Bias Variance Trade off image is taken from https://gadictos.com/bias-variance/