NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

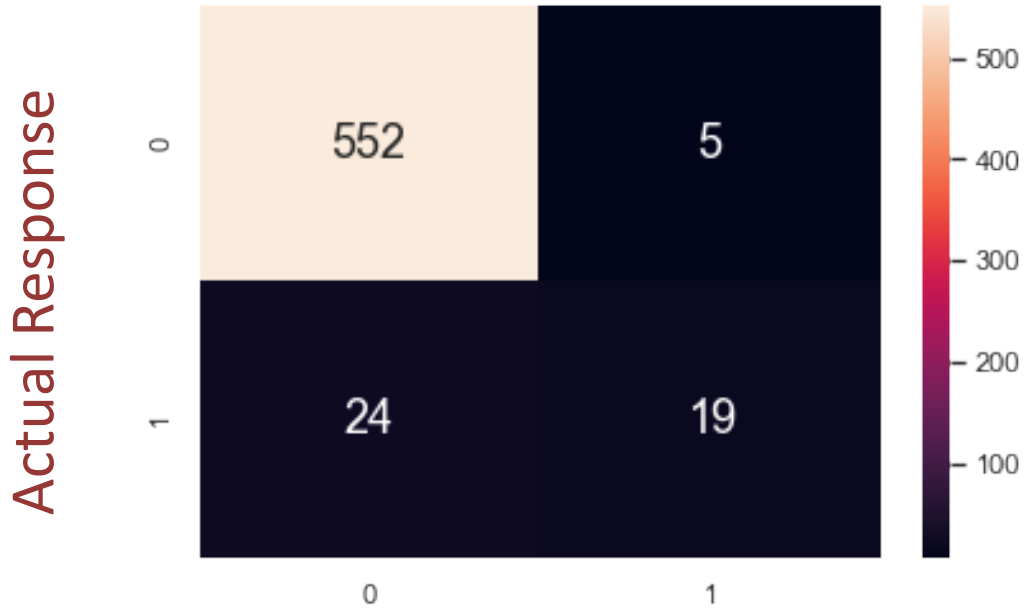
Review Session 6

Classification, Cross validation, Accuracy and Bias Variance

Dr Smitha K G

Admin Announcements

1. LAMS DS deadline: 3rd March 11.59 pm (hard deadline)
2. DS Theory Quiz in Recess Week : **8 March, Friday**. Slots : 12:30 pm – 2:00 pm and 2:30 pm to 4:00 pm. Lab allocations and FAQs posted.
3. Mini-Project details posted on NTULearn. Finalize on the team and the dataset and start working towards the same. Keep your TA updated. Deadline for the same is 1st March



Data Science

Binary Classification

Goodness of Fit of the Model

TP : True predicted as True 19
TN : False predicted as False 552
FN : True predicted as False 24
FP : False predicted as True 5

$$accuracy = \frac{552 + 19}{552 + 19 + 5 + 24}$$

"Positive" : 1 Legendary
"Negative" : 0 Non-Legendary

$$tpr = \frac{19}{19 + 24}, \quad fnr = \frac{24}{24 + 19}$$

$$fpr = \frac{5}{5 + 552}, \quad tnr = \frac{552}{552 + 5}$$

Confusion Matrix : https://en.wikipedia.org/wiki/Confusion_matrix

Actual **N**	**TN**	**FP**
Actual **P**	**FN**	**TP**
	Predicted **N**	Predicted **P**

25	**0**
0	**75**

0	**25**
0	**75**

When will you be happy?

Ideal TPR = 1, FPR = 0

Bad? TPR = 1, FPR = 1

Bad? TPR = 0, FPR = 0

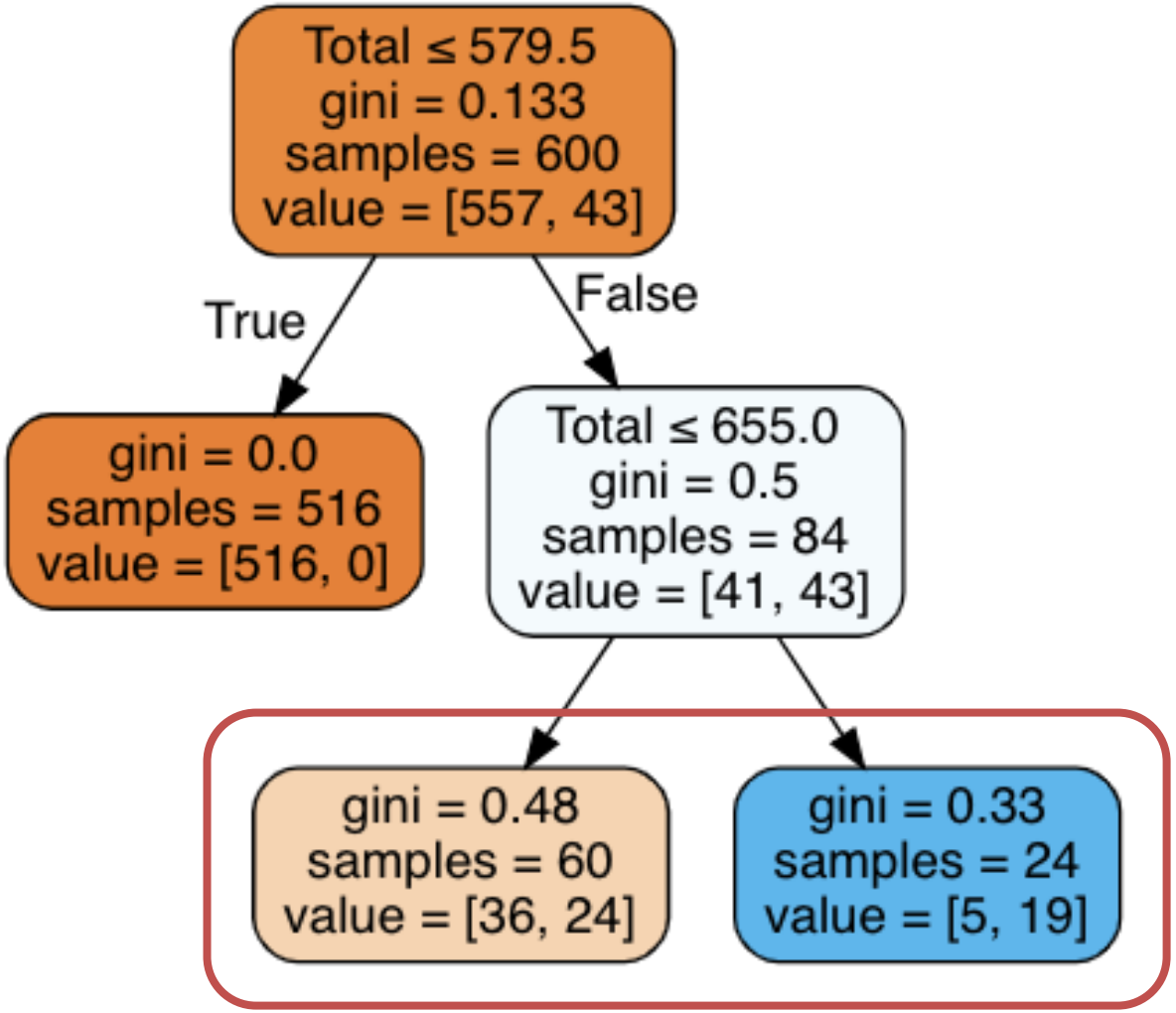
Trash TPR = 0, FPR = 1

25	**0**
75	**0**

0	**25**
75	**0**

Balancing classes to achieve the desired TPR and FPR is a tricky thing to do. ☺

F1 Score for balanced FPR and FNR : https://en.wikipedia.org/wiki/F-score



Data Science

Binary Classification

How does a Tree "decide" classes?

The tree doesn't! **You** decide it on your own by choosing **Decision Threshold**.

If Proportion > T, you call it Positive, and else, you call it Negative class.

Default Threshold for Trees = 0.5

Experiment with the Decision Threshold!

Use your tree to find Leaf Nodes.

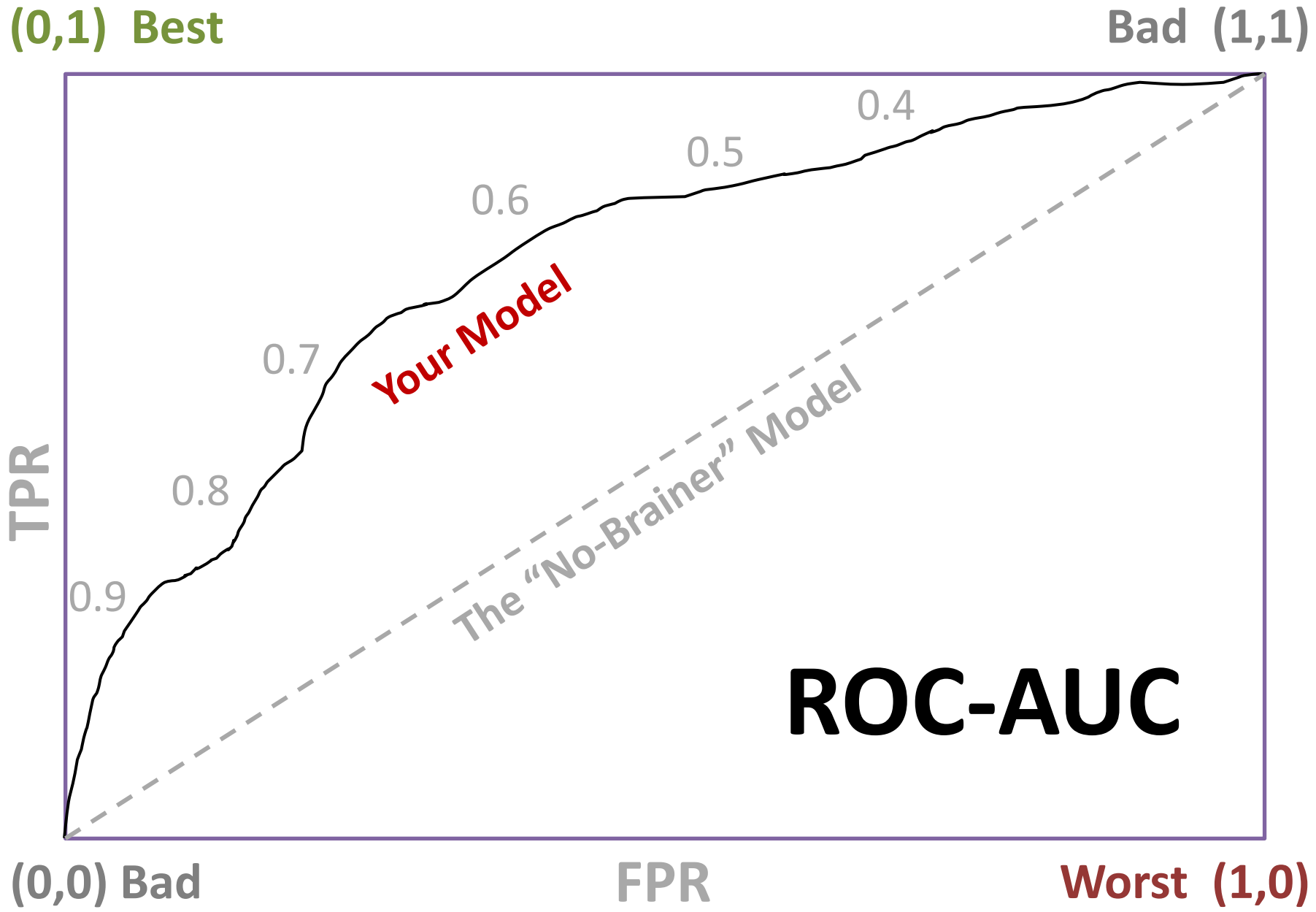
Vary your decision threshold T in steps from 0 to 1 and note the TPR and FPR.

T = 0 : Everyone P

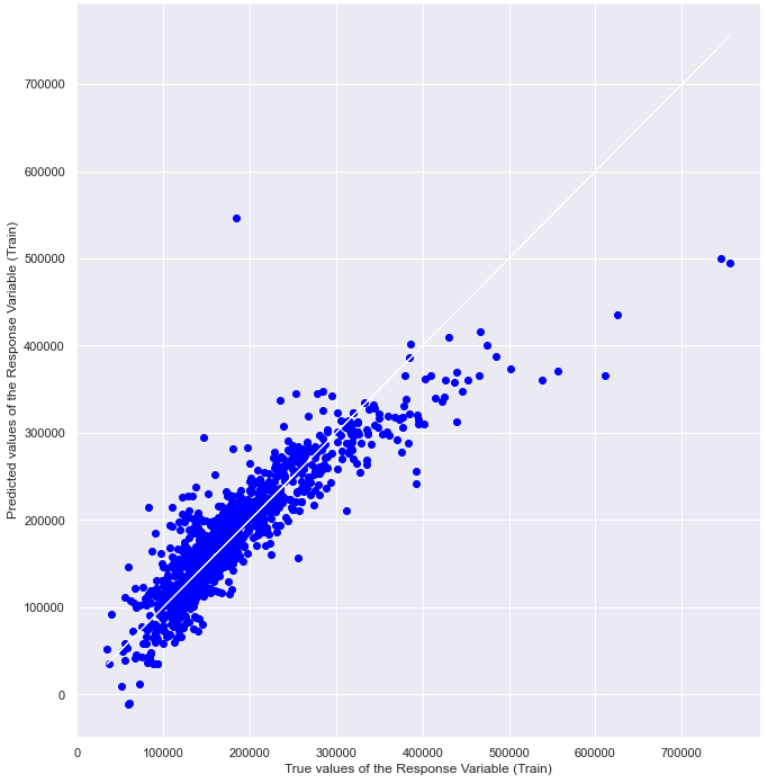
TPR = 1, FPR = 1

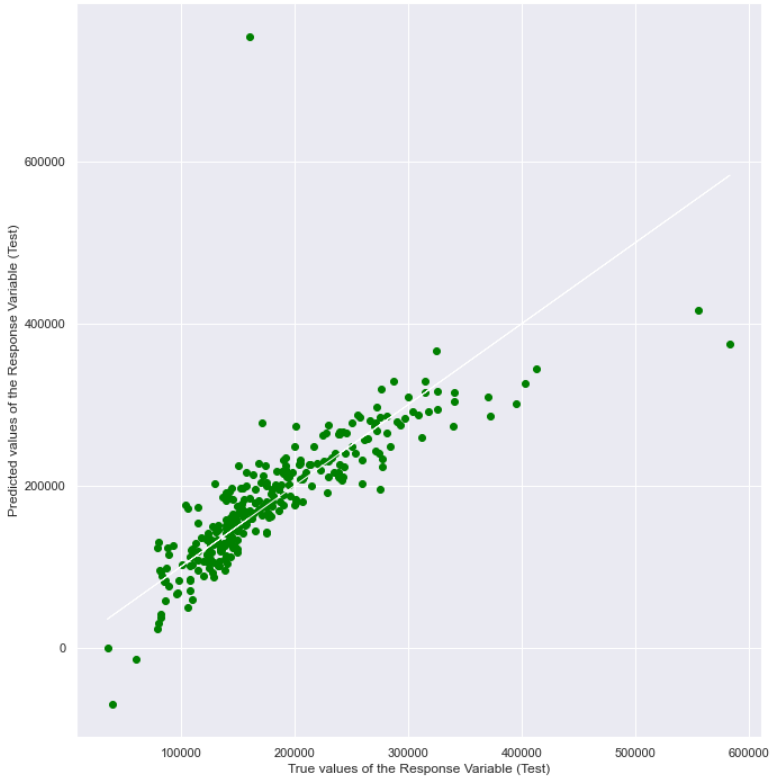
T = 1 : Everyone N

TPR = 0, FPR = 0



MULTI-VARIATE LINEAR REGRESSION





Explained Variance (R^2) on Train Set : 0.744
Mean Squared Error (MSE) on Train Set : 1484.39 *10^6
Mean Squared Error (MSE) on Test Set : 1795.3 * 10^6

Features taken: 'GrLivArea','TotalBsmtSF', 'GarageArea','OverallQual'

K FOLD CROSSVALIDATION

cross_val_score ()

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_validate.html#sklearn.model_selection.cross_validate
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html#sklearn.model_selection.KFold

How can we select the features that I need for regression fit?

https://towardsdatascience.com/5-feature-selection-method-from-scikit-learn-you-should-know-ed4d116e4172

https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html#sklearn.feature_selection.SelectKBest

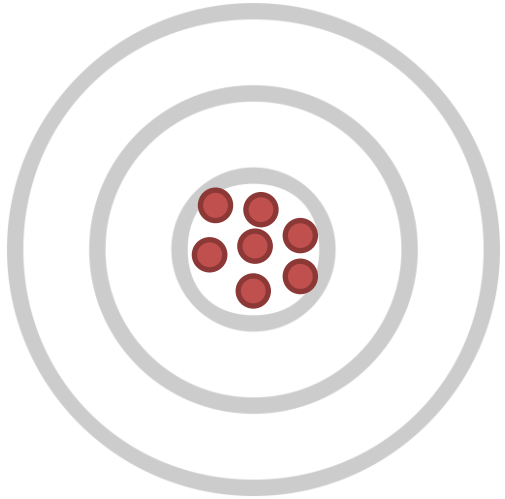
The main concern of Data Science

ACCURACY OF THE MODEL

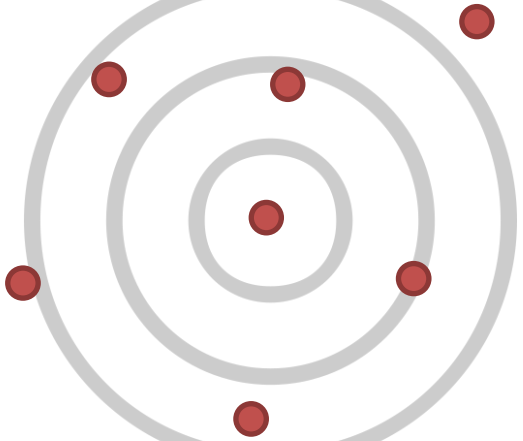

6
5
10.0
2
1

The Dream Model

Low Bias
Low Variance

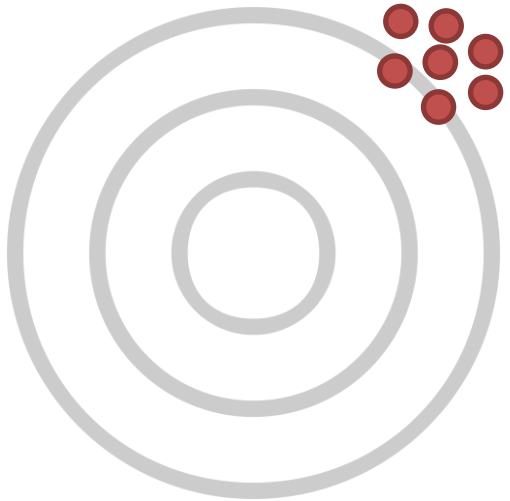


In Practice



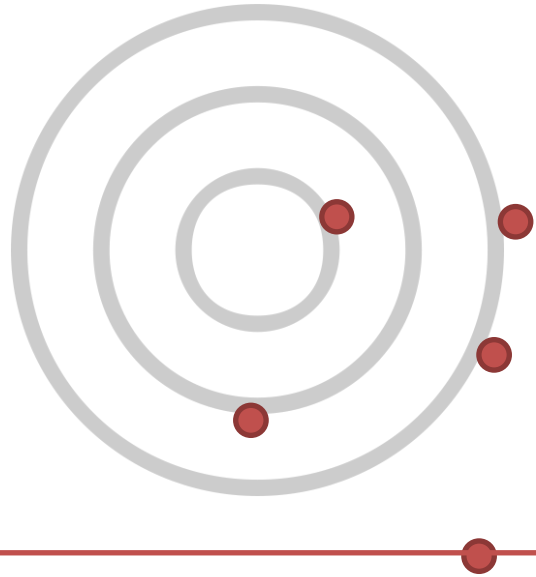
Low Bias
High Variance

In Practice

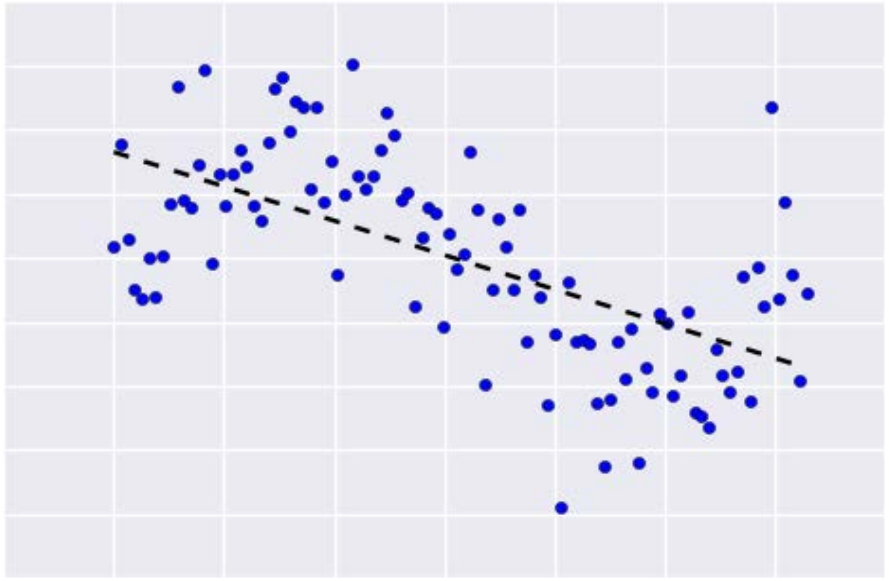


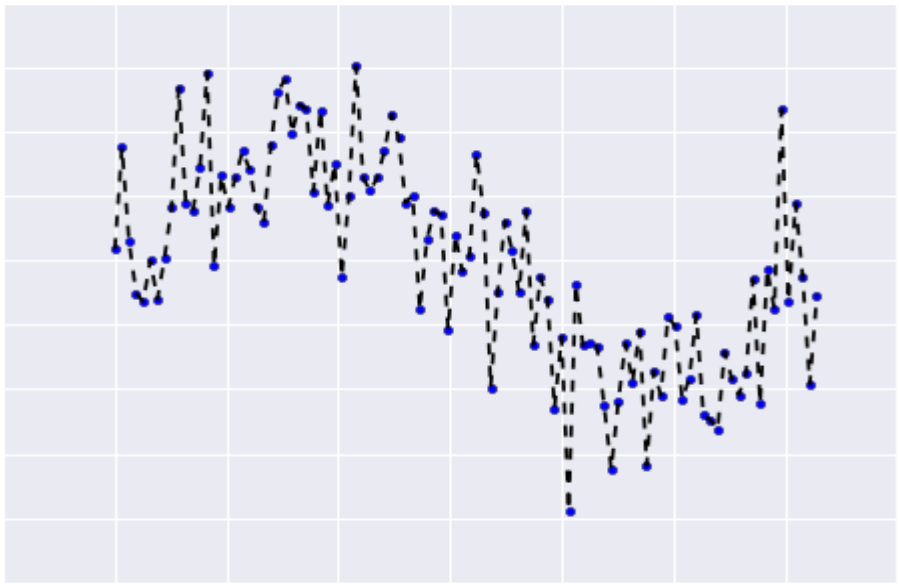
High Bias
Low Variance

Garbage Model



High Bias
High Variance

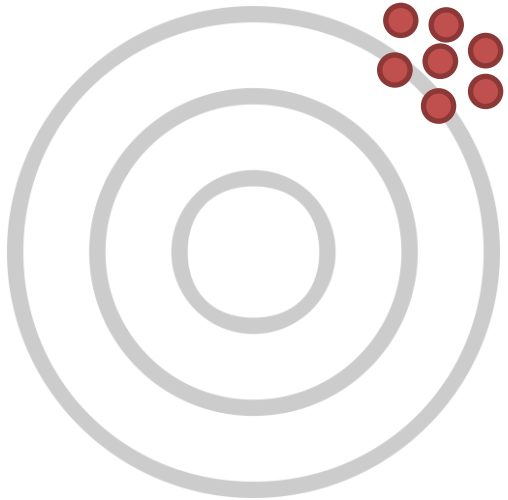


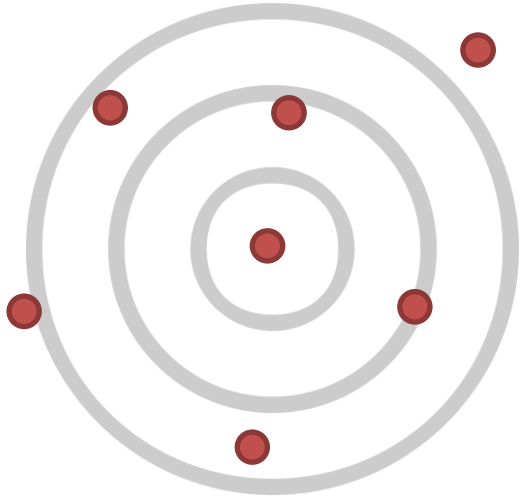


UNDER-FIT

High Bias

Low Variance

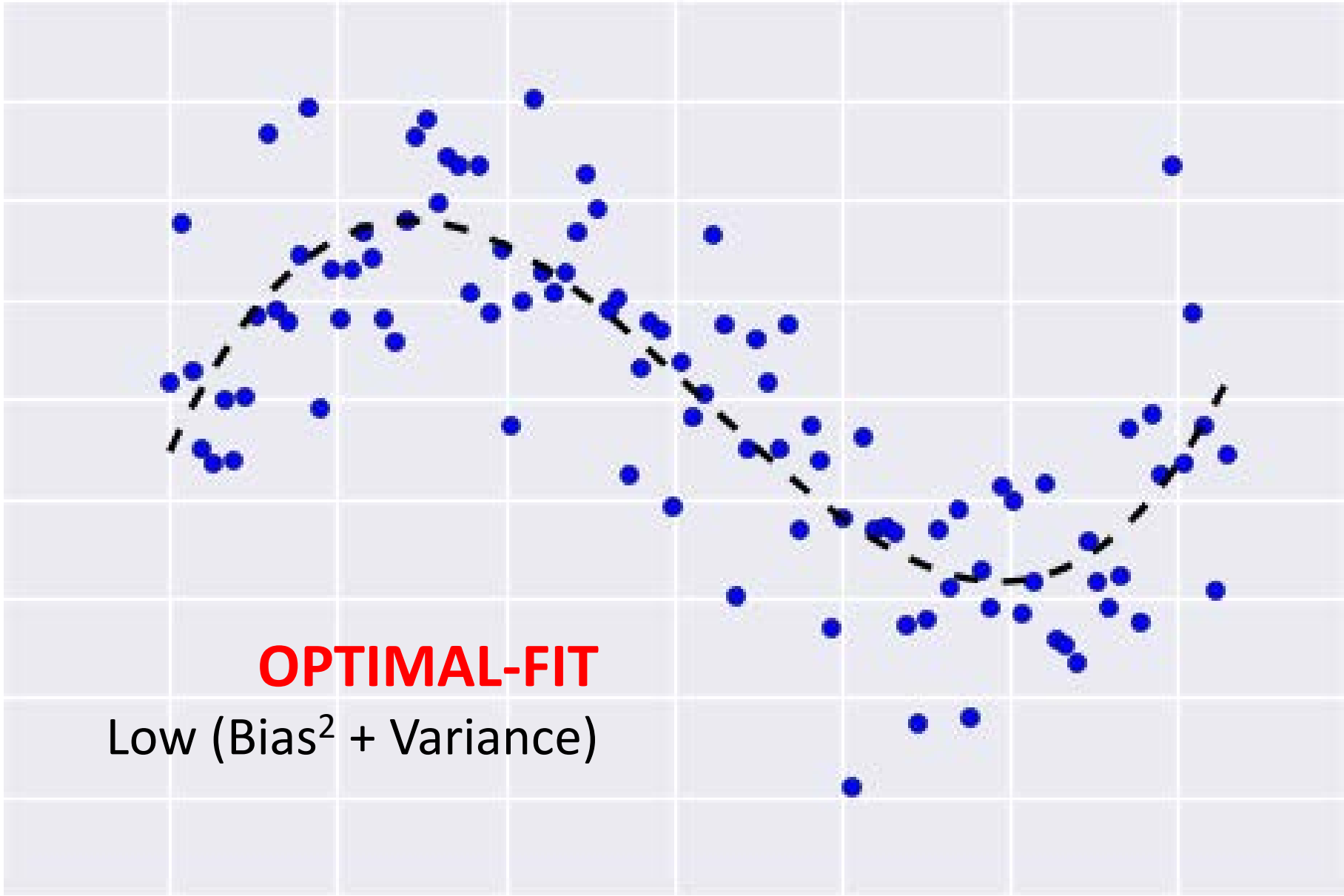




OVER-FIT

Low Bias

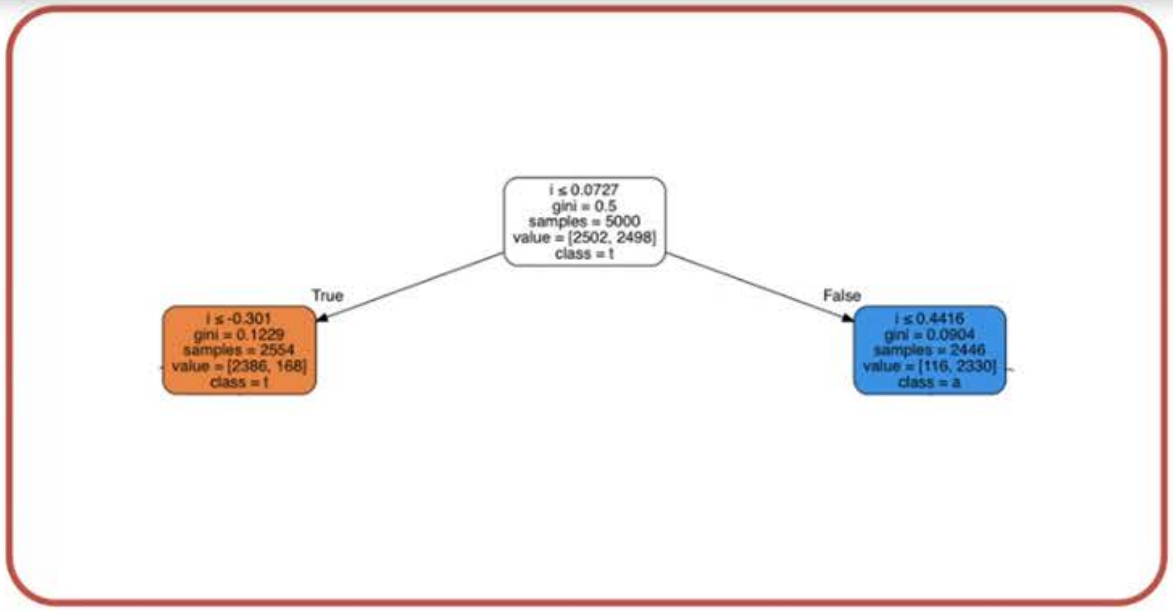
High Variance

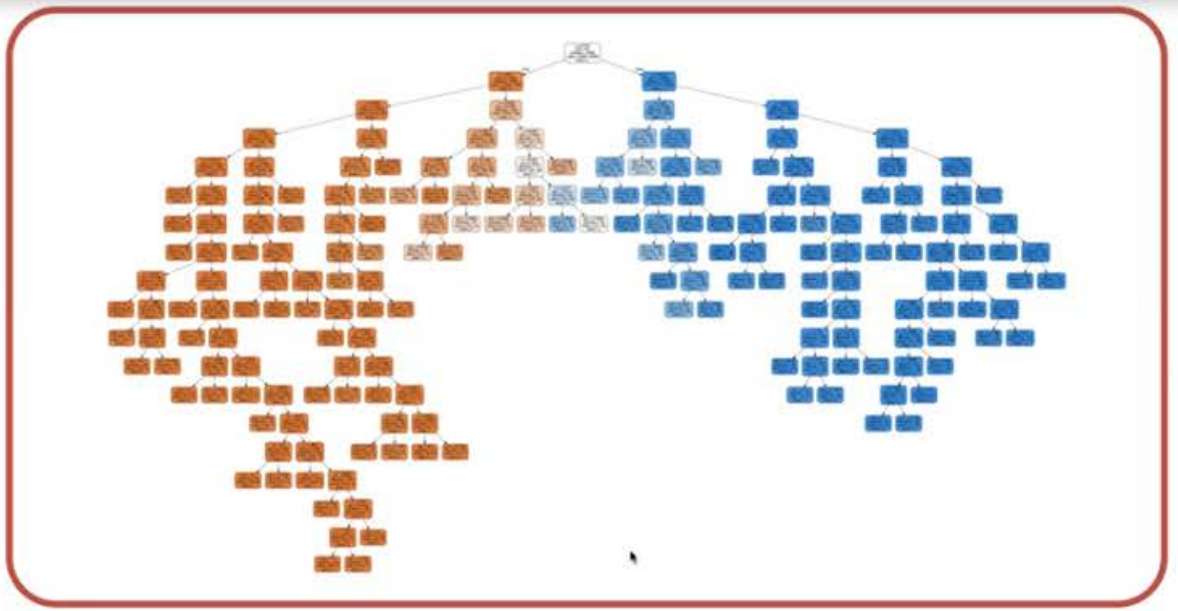


OPTIMAL-FIT

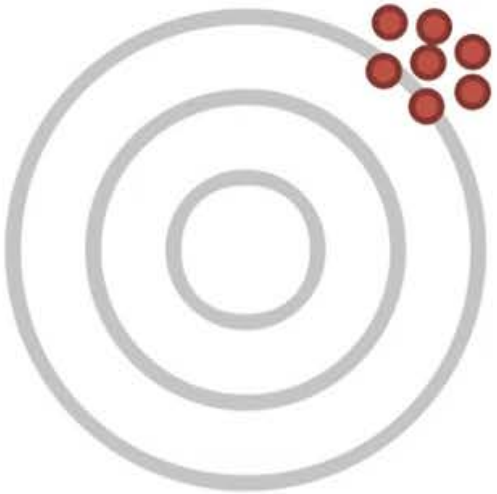
Low (Bias^2 + Variance)

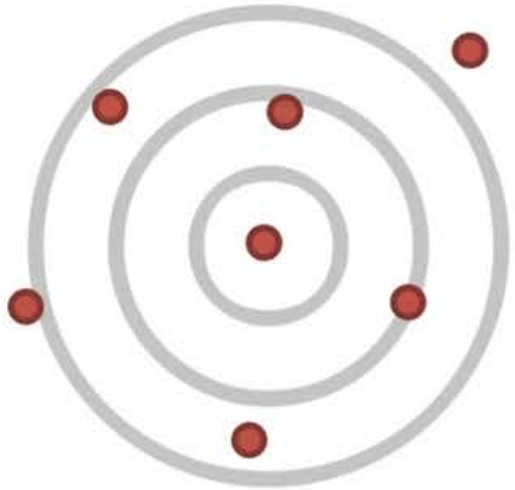
https://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model



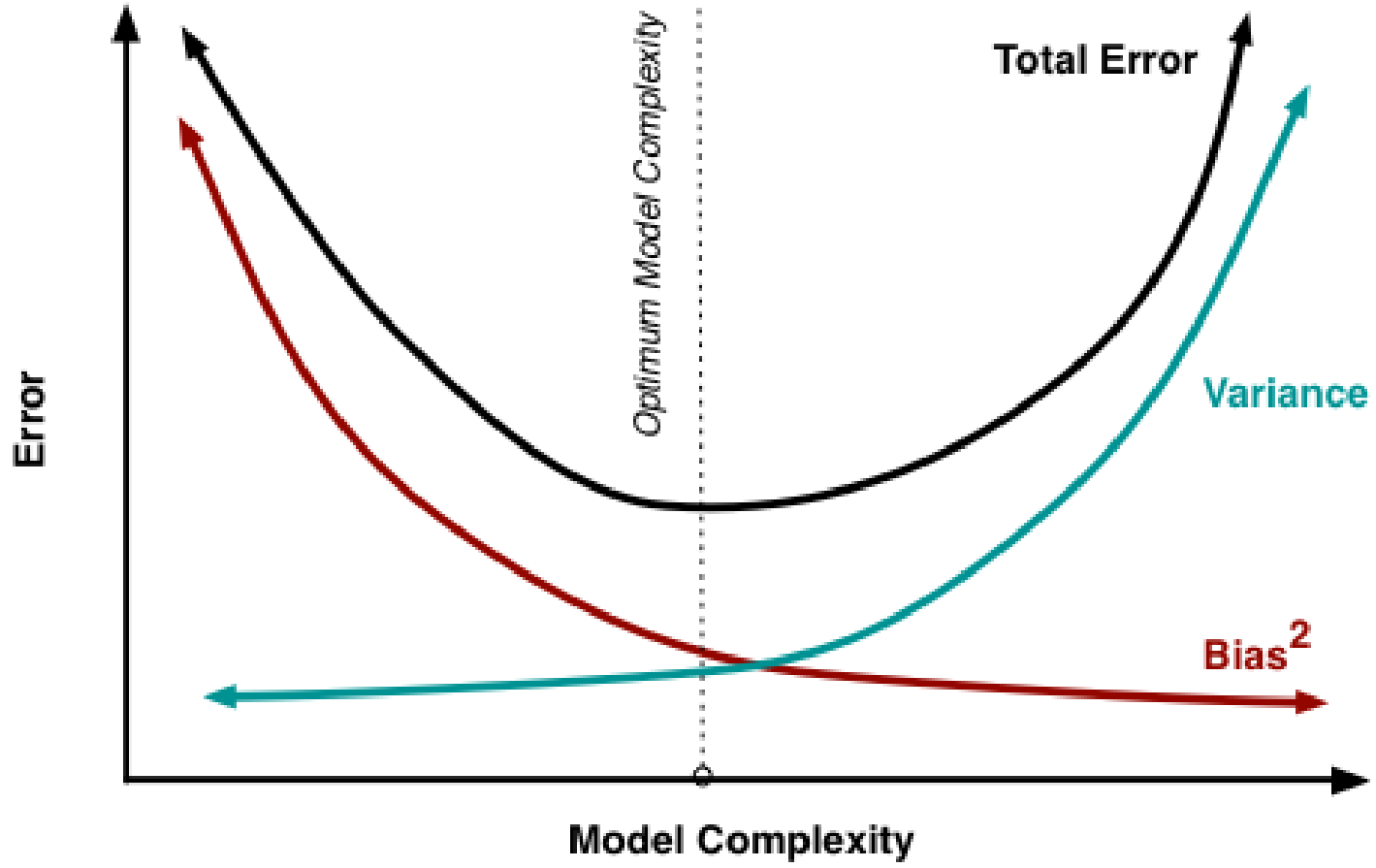


UNDER-FIT
High Bias
Low Variance





OVER-FIT
Low Bias
High Variance



Model Complexity is the Hyper-Parameter to Tune