IDARE®

AI Success Metrices

Module 10 ML Success and Performance Assessment

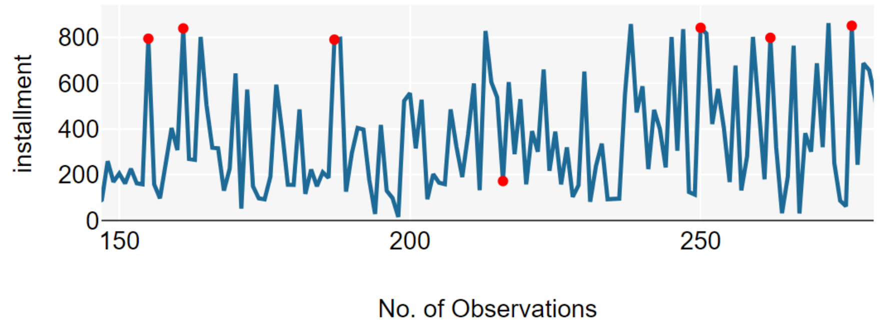
IDARE®

ML Success Tactics
Minimizing Data Uncertainty

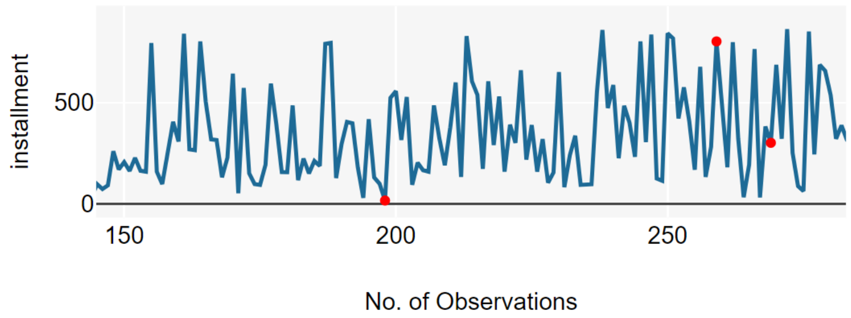
Data Quality

Types of Trouble you may face in Data

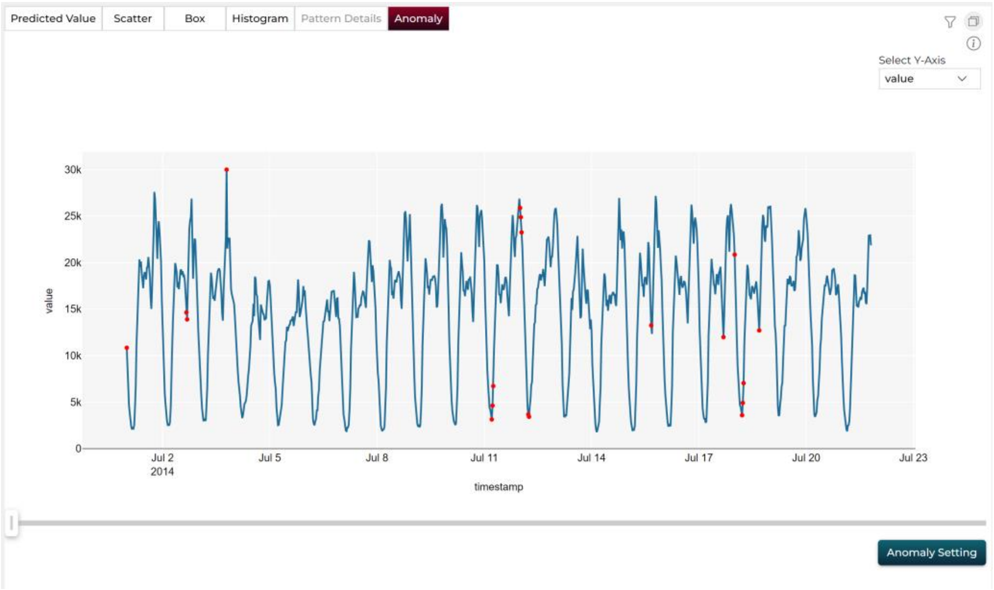
- **Missing Value**, Null Value, Blank Value, inf value, NA value
  - Removing is the easy option, however in many cases those values means major cause of the target.
  - Replace with either 0 or avg or other statistical parameters will be wise
- **Anomalous Data or Outlier**
  - Sudden picks and valleys in the data
  - Use anomaly detector to detect or isolate
  - Talk to domain expert or use your knowledge to understand
    - That anomaly is the part of a process or means something. Twik the anomaly detector to isolate right anomalies
    - Or simply data error
  - If those anomalies mean something, categories them based on their recognized category, if not remove them
  - Removing anomalies for a variable will lead to removing other variables from that point so be careful

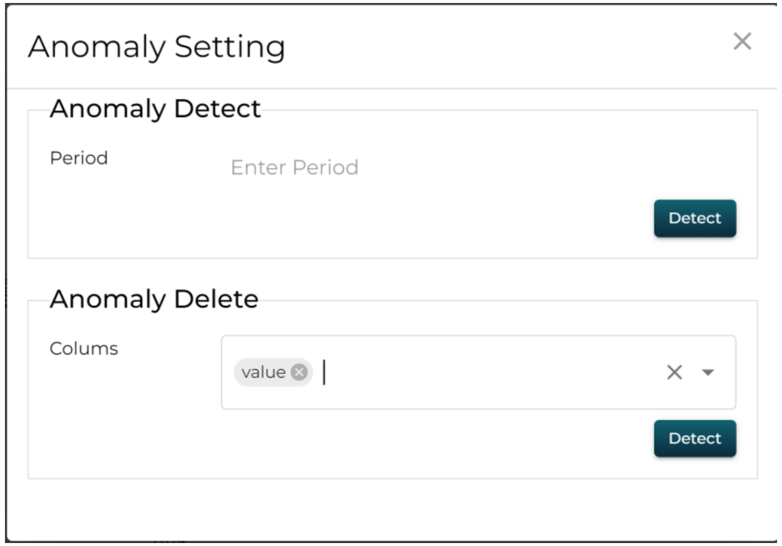


Twiking Anomaly parameter shows lesser anomalies



Anomaly Detection Example





ML Success Tactics
Variable Selection (Feature Engineering)

Right use of Correlation & Causation in variable selection in training

Variable Selection Defines success of an AI Solution

Most Important Process in Solution Creation is to determine the right predictor variables or features

- Variable that doesn't have any affect on the target
- that cause the target
- The relationship between the selected variables i.e. correlation coefficient
- Check Feature Importance after each analysis
- Create science driven variables
- Perform extensive parametric study
- Try to stick to One Algorithm during Variable selection

- Start your analysis unselecting these variables

- Try keeping these variables under any circumstances unless there are compelling reasons
- If p-value < 0.05 try keep the variable

If Pearson correlations coefficient
- high w.r.t target try keeping these variables
- high w.r.t other predictor variable, try not to use one of the two

Recommended Practice for being successful in variable selection process

Key practices

- Brainstorm to understand the problem and study the target
- Utilize your Domain expertise
- Gather domain knowledge
    - Do extensive literature survey
    - Talk to domain experts
- Use critical and analytical skills to determine

Correlation and Causation

Right use of Correlation & Causation in variable selection in training

Correlation & Causation for Variable Selection

- **Causation:**
  - Variable that directly comes from domain expertized are used or not
  - Variables that are low p-values considered or not
  - High P-values are avoided or not unless domain experts recommends that
  - IF p-values are 0 check t-value, high t-value suggest high significance with the target

- **Correlation:**
  - If variables are highly correlated or high coefficient with respect to TARGET should be used
  - If variables within themselves are highly correlated should be avoided unless domain knowledge suggested or parametric study suggested

Variable Statistics | Validation Result | Variable Importance | Compare KPI

Variables	Data Type	Missing Values Count	p value	t value	Pearson Correlation
Lever_Pos	float	0	0.4	0.84	0
Ship_Speed	int	0	0	11.33	0
GT_Shft_torq	float	0	0	106.99	0
GT_RPM	float	0	0	61.11	0
Gas_Genrtr_RPM	float	0	0	100.25	0.01
Strbrd_Proplr_Trq	float	0	0	-139.25	0
Port_Proplr_Trq	float	0	0	-139.25	0
HP_Trbin_exit_temp	float	0	0	-96.02	-0.04
GT_Comprsr_inlet_air_Temp	int	0	0.04	2.11	
GT_Comprsr_outlet_air_Temp	float	0	0	16.55	-0.02
HP_Trbin_exit_press	float	0	0	24.18	0
GT_Comprsr_inlet_air_Press	float	0	0.04	2.11	
GT_Comprsr_outlet_air_Press	float	0	0	-177.8	-0.02
HP_Trbin_exahst_gas_press	float	0	0	12.86	0.01
Trbin_Injecton_Cntrl	float	0	0	-36.89	-0.02
Fuel_flow	float	0	0	82.69	-0.02
GT_Trbin_dcay_coeff	float	0			1

Cross Validation for Variable Selection

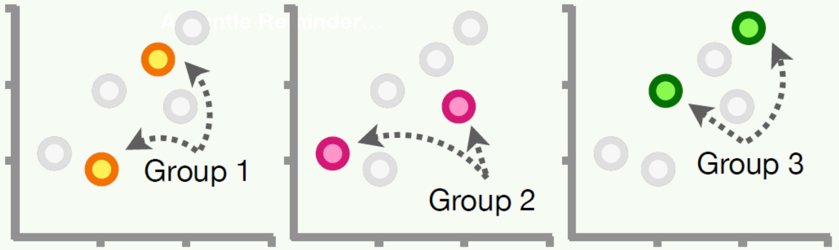
Right use of cross validation in training

Cross Validation: Details Part 3

IDARE®

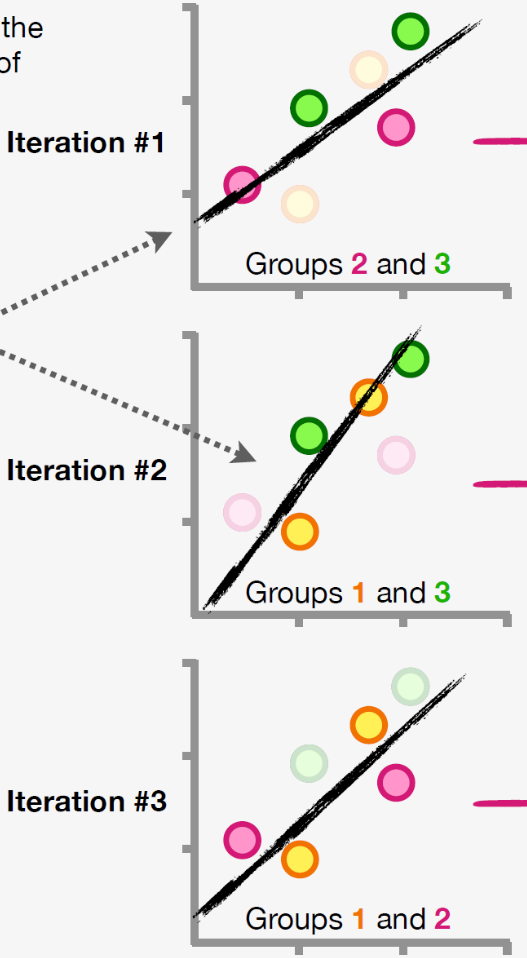
8 Because we have **3** groups of data points, we'll do **3** iterations, which ensures that each group is used for Testing. The number of iterations are also called **Folds**, so this is called **3-Fold Cross Validation**.

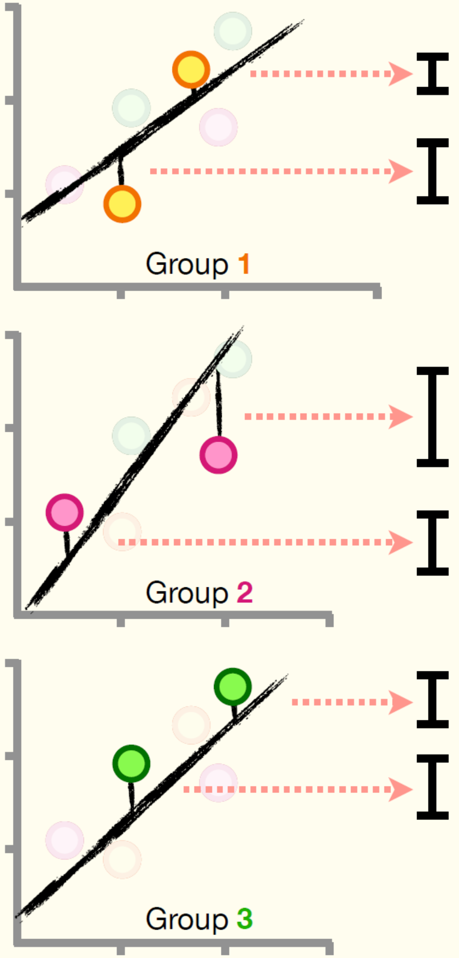
Gentle Reminder: These are the original **3** groups.



9 So, these are the **3** iterations of Training…

NOTE: Because each iteration uses a different combination of data for Training, each iteration results in a slightly different fitted **line**.





10 …and these are the **3** iterations of Testing.

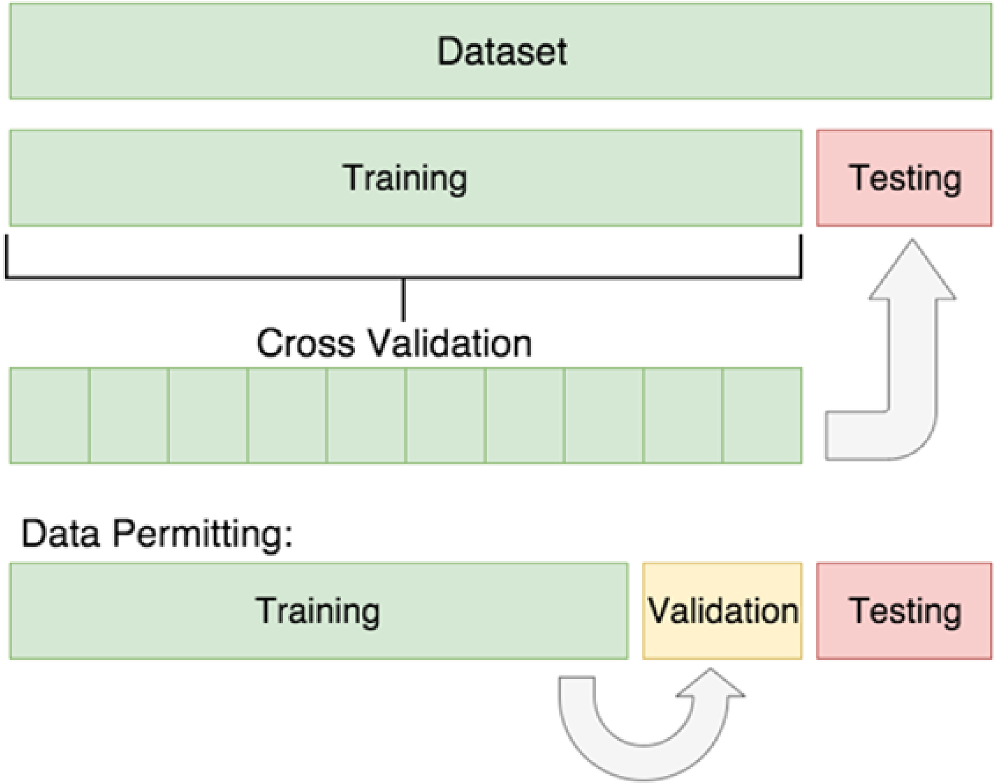
A different fitted line combined with using different data for Testing results in each iteration giving us different prediction errors.

We can average these errors to get a general sense of how well this model will perform with future data…

…or we can compare these errors to errors made by another method.

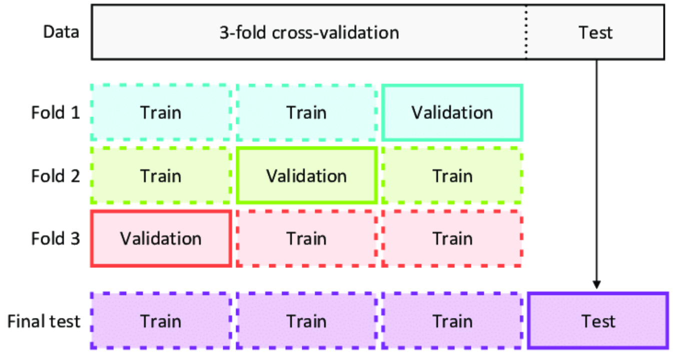
Cross Validation with Data Split

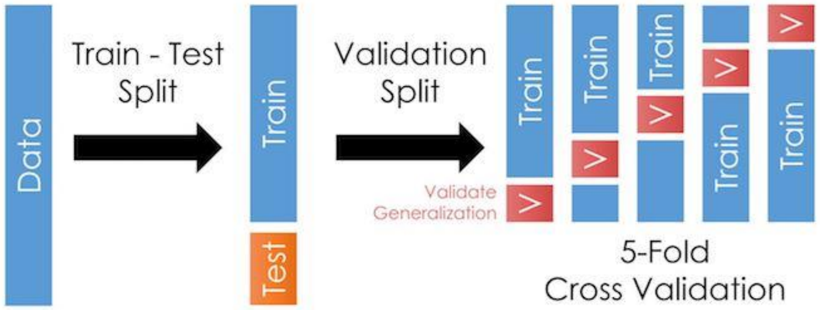
Train Data: Data Sample ML will Learn From
Test Data: Unseen data or Out of Sample data potentially you will see when in production
Validation Data: Part of Train data kept unseen for cross validation



Kinds

- 3-fold
- 5-fold
- 10-fold

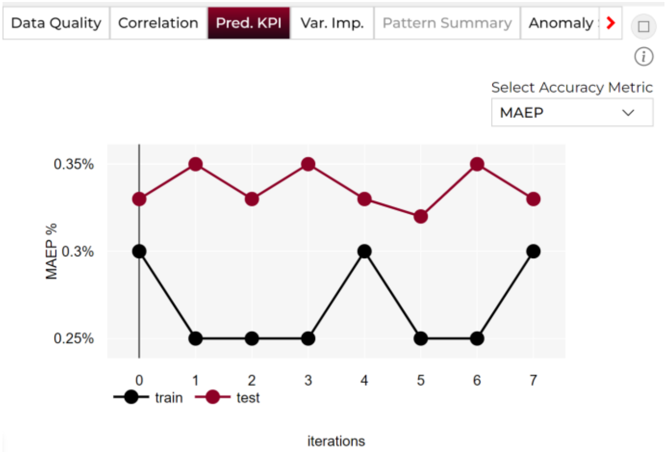




Error Check with Cross-Validation

Error to Look at

- Train Error (Bias)
- Cross Validation Error
- Test Error (Variance)
- Variation of Error between folds



Variable Statistics	Validation Result	Variable Importance	Compare KPI			

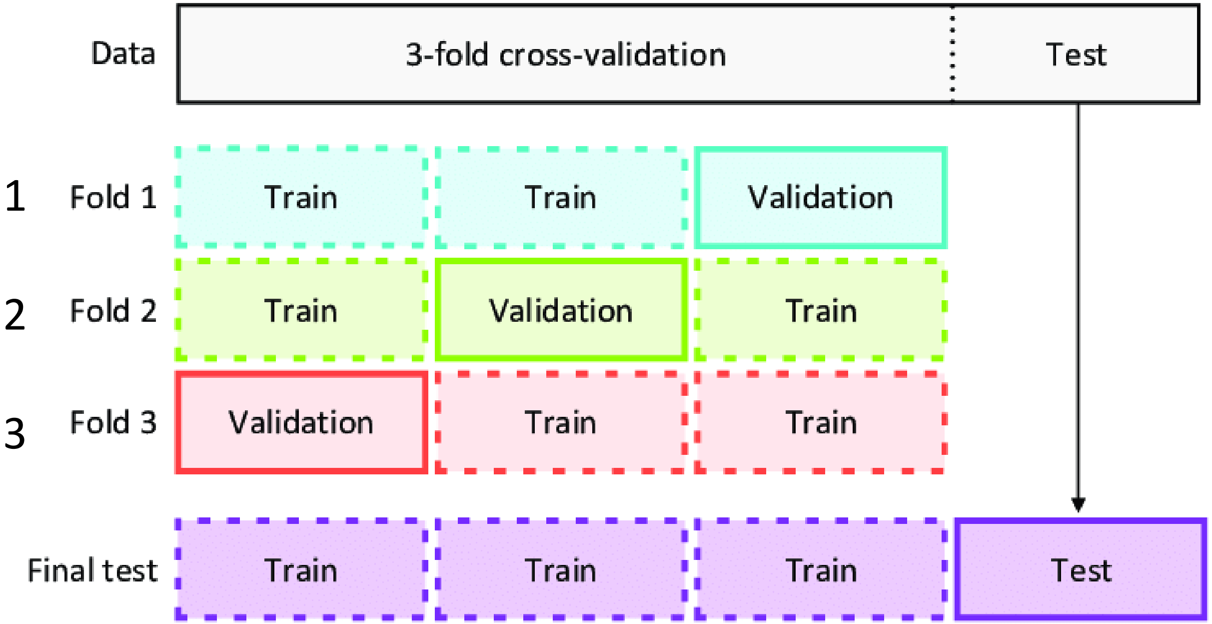
ML Algorithm	Mean CV MAEP	Test MAEP	Split_1 CV MAEP	Split_2 CV MAEP	Split_3 CV MAEP	Stand Dev of CV MAEP
Random Forest	0.2	0.32	0.28	0.08	0.24	-0.09
XGB	0.23	0.35	0.28	0.11	0.28	-0.08
Linear regression	0.21	0.33	0.22	0.16	0.26	-0.04

- Cross Validation Avg. Error
- Standard Deviation

Error Fold 1

Error Fold 2

Error Fold 3



AI Success Metrices

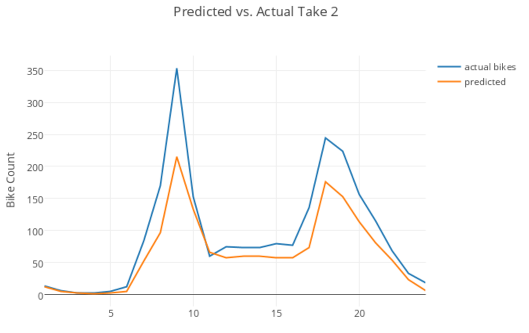
Module 10 ML Success and Performance Assessment

Performance Assessment

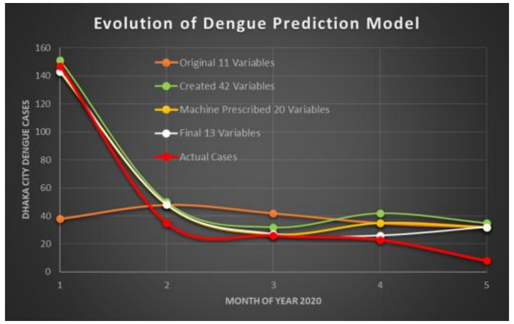
Fundamentals

Performance Check: Regression

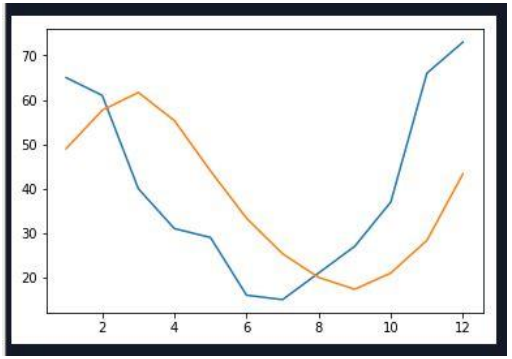
- The zeroth law: Compare actual versus predicted line chart for Test data
- Check whether predicted line captures the pattern of the actual or now
- If pattern doesn't matches, major work will be needed in variable selection



Captures the pattern
But error will be high.
BIAS is ok though
variances are high



All prediction captures
the pattern except one

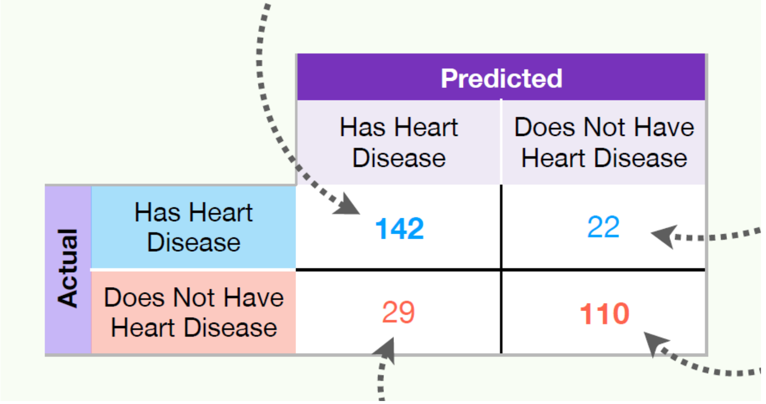


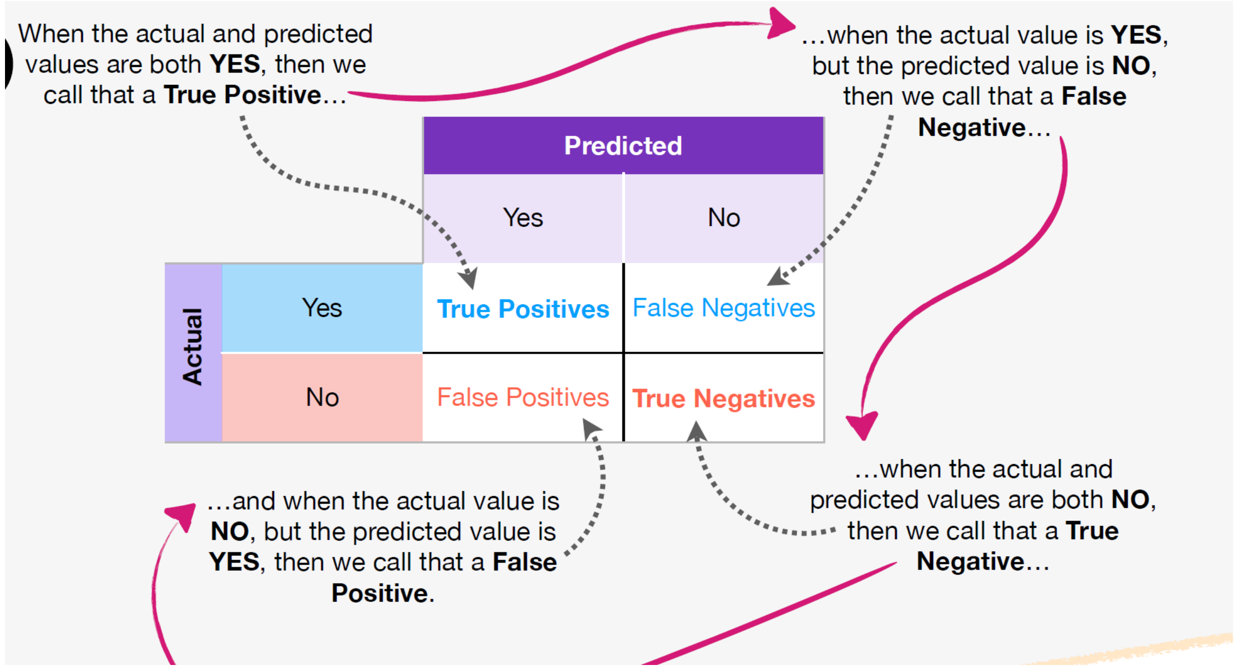
Captures the pattern,
the pattern shifted

IDARE®

Performance Check Classification

- The zeroth law: check confusion matrix
- check true positive, false positive, true negative and false negative
- Reduce false positive or false negative based on the problem

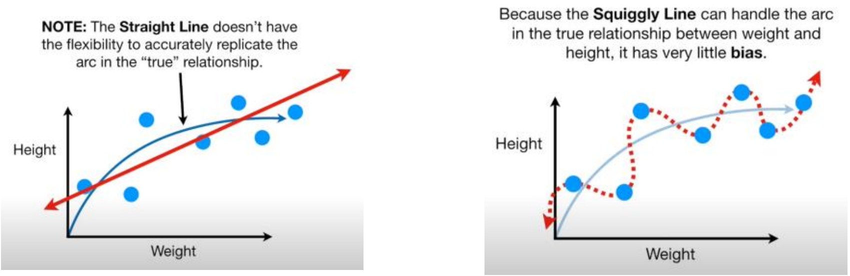




Understanding Model Stability or Prediction Consistency

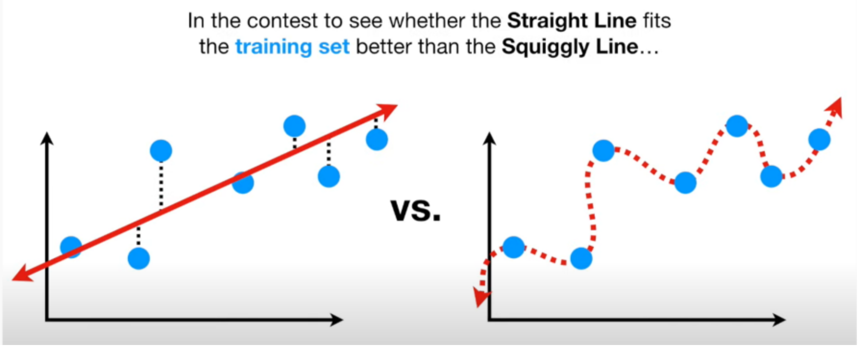
Model Stability or Consistence

Remember Bias and Variance?

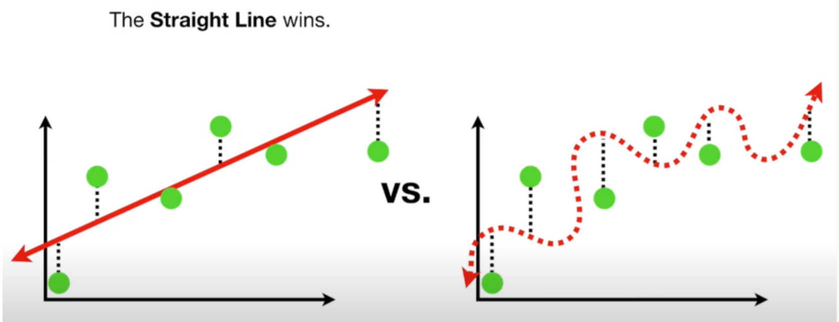


Training data set trains with 2 algorithm

4. Bias: How well the algorithm learns the true behavior or how complicated your model is
   - *Low bias: over complicated model or too many variable used*
   - *High bias: over simplified model or too less variable used*
   - Sum of Squared Error for each predicted points with training data set
5. Variance: Measures the differences between actual and predictions.
   - Sum of Squared Error for each predicted points with test data set

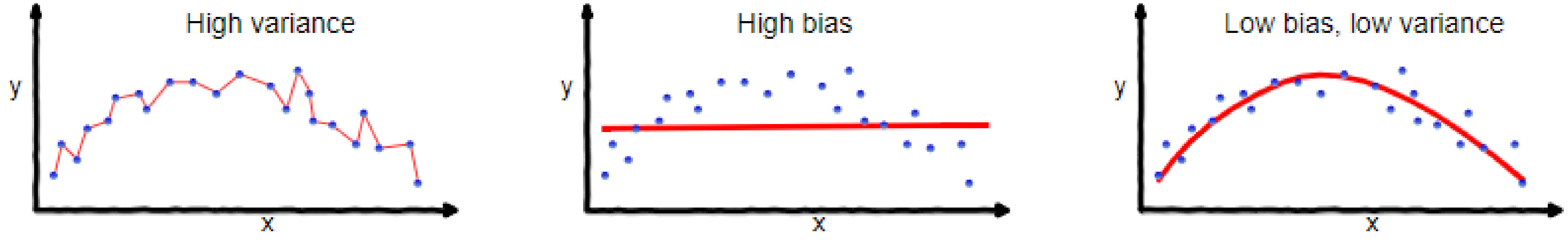


Error for Training Data sets



Error for Test Data sets

Overfitting Underfitting



overfitting underfitting Good balance

Overfitting: Training error low ,testing error high→ Model Low Bias high variance

- Extra unrelated variables cause reduce bias and cause more error later, leads to instable and inconsistent result

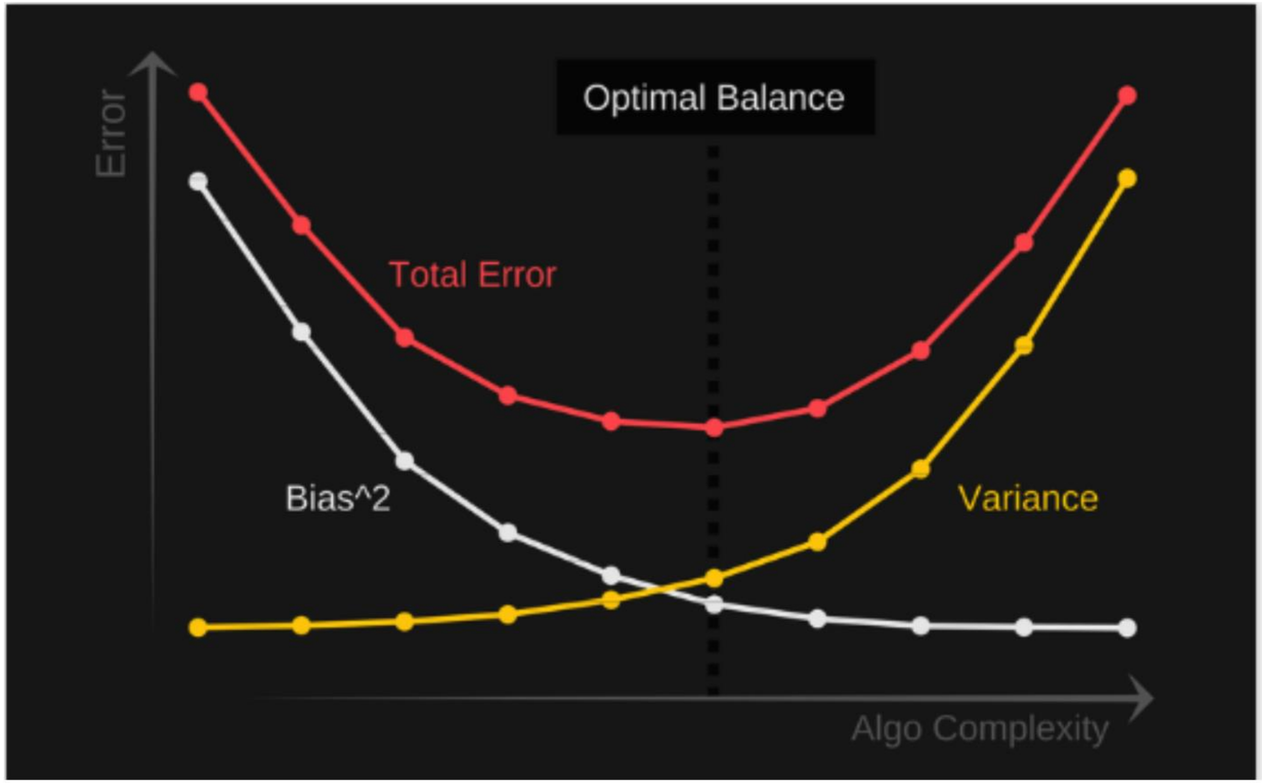
Underfitting: Training error high, testing error high→ Model high Bias, high variance

- Missing important variable increase bias and also cause more error later leads to highly instable and inconsistent prediction

Good Balance: Training error low, testing error low→ Optimal Bias, low variance

- Good variable selection and science driven AI reduces chances of overfitting or underfitting

Total Error = Bias^2 + Variance + Irreducible Error



Total Error Scale on the order of Bias squared, which in most cases is substantially big, downplays testing error. Best way to understand a stable model is to do a eye check.

- Minimum difference between training and testing error
- Both error are low
- Standard deviations of cross validation data sets are low

Error Metrices

IDARE®

Error metrices for Regression

Residual = Observed - Predicted

SSR = Sum of Squared Residuals

$$\textbf{SSR} = \sum_{i=1}^{n} (\text{Observed}_i - \text{Predicted}_i)^2$$

$$\textbf{Mean Squared Error (MSE)} = \frac{\textbf{SSR}}{n}$$

...where ***n*** is the sample size

$$R^2 = \frac{\text{SSR(mean)} - \text{SSR(fitted line)}}{\text{SSR(mean)}}$$

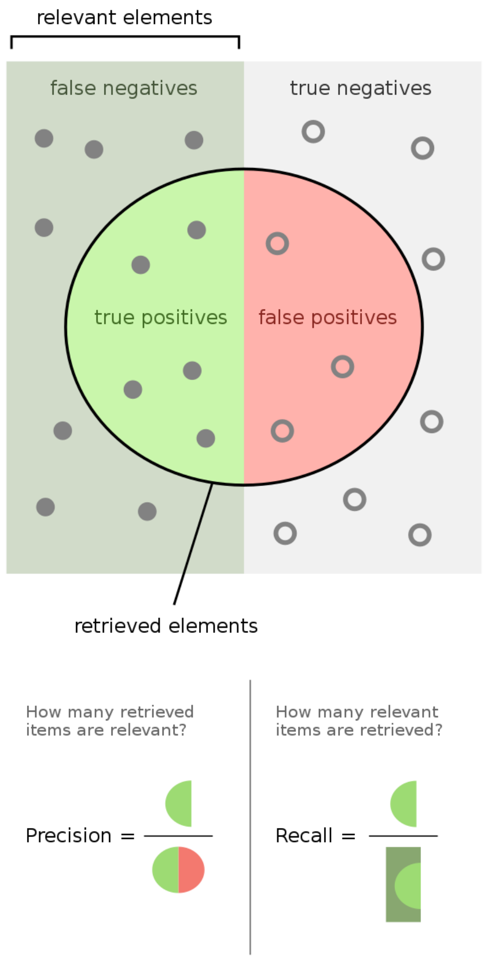
$$RMSE = \sqrt{MSE}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n} \left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

MAEP = Sum of absolute Error / Sum of Actuals

Error metrices for Classification

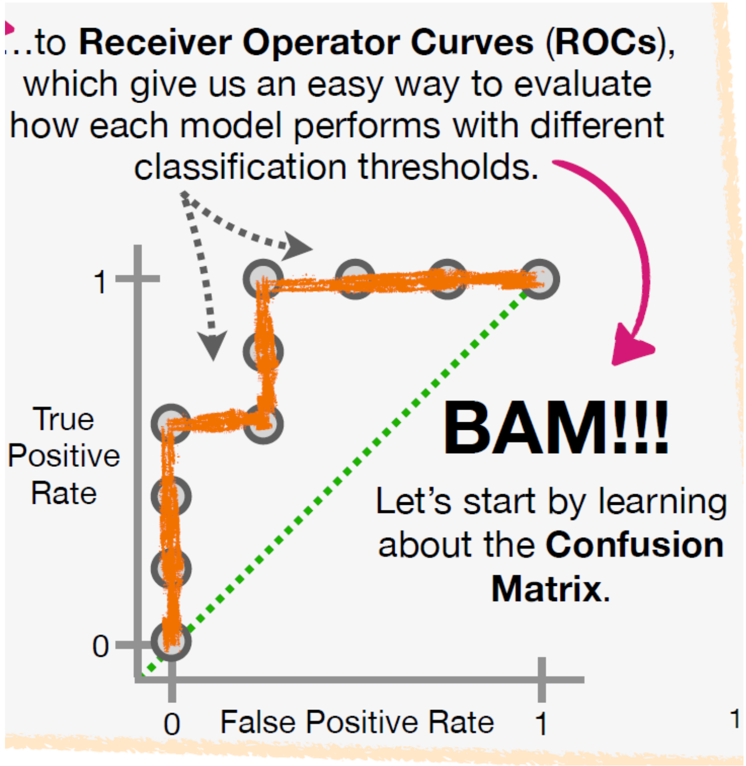


$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$



AI Success Metrices

Module 10 ML Success and Performance Assessment

ML Parameter on Performance (**Hyper Parameter Tuning**)

Hyper Parameter (HP) Tuning for ML Algorithm

Hyper parameters are variables specific to machine learning algorithms that helps the ML learning process.

Hyper parameters has no influence on the performance of the model but affect the speed and quality of the learning process.

The HP tuning process is little expensive, costs computation time as it runs many time to find best models

Following are some key parameters applicable for many ML algorithm

- **No. of Iteration:** how many times the algorithm search for best results
- **No. of Trees or Layers:** no. of Paths and combinations of path to reach the outcome
- **Depth:** How many elements or parameters of a tree or paths to consider
- **Learning Rate:** How small the step of the learning
- **Bootstrap:** Way of data sampling, combinations of rows or variables etc
- **Solver:** Algorithm to tune the hyperparameter

IDARE®

Hyperparameters for different Algorithm

Random Forest (RF)

- **Criterion:** Criterion is a loss function to measure the quality of a split inside a tree.
  - Mean Squared Error and Mean Absolute Error
- **The maximum number of features:** The number of features to consider when looking for the best split. Decreasing the maximum number of features helps control overfitting.
  - All, Square Root, Logarithm: Use the logarithm (base 2) of the total number of features
- **Maximum depth of each tree:** The deeper the tree, the more branches it has and it captures more information about the data.
- **Bootstrap:** It's a sampling technique
- **The number of trees in the forest:** The default value for this parameter is 100, which means that 100 different decision trees will be constructed in the random forest. A higher number of trees give you better performance but makes the training slower.

XG Boost

- **Maximum depth of each tree:** Same as RF
- **The number of trees in the forest:** Same as RF
- **Learning Rate:** Lower learning rate means the model is more robust to overfitting but makes the training slower.

Neural Network

1. **Hidden Layers and Neurons:** Hidden Layers Similar like Trees
2. **Activation Function:** decides whether a neuron's input to the network is important or not in the process of prediction
3. **Solver:** Solver is an algorithm to optimize the weights of the neural network.
Stochastic Gradient Descent & Adam:
4. **Initial Learning Rate:** How small the step of the learning
5. **The number of epochs**: no. of iterations.

IDARE®

Hyperparameter Examples

Random Forest Regressor

Hyperparameters	Values	
Criterion	☑ Mean Squared Error ☐ Mean Absolute Error	Loss of Accuracy
Maximum number of features	☑ All ☐ Square Root ☐ Logarithm	Solver
Bootstrap	◉ True ○ False	Data Sampling
The number of trees in the forest	◉ List ○ Range 100	No. of Iteration
Maximum depth of each tree	◉ List ○ Range None	Depth

XGBoost Regressor

Hyperparameters	Values	
The number of trees	◉ List ○ Range	100 No. of Paths
Maximum depth of each tree	◉ List ○ Range	6 Depth
Learning Rate	◉ List ○ Range	0.3 Learning Rate

Neural Network Regressor

Hyperparameters	Values	
Hidden Layers and Neurons	100	Depth
Activation Function	☑ Rectified Linear Unit ☐ Hyperbolic tan Function	Loss of Accuracy
Solver	☐ Stochastic Gradient Descent ☑ Adam	Solver
Initial Learning Rate	◉ List ○ Range 0.001	Learning Rate
The number of epochs	◉ List ○ Range 200	No. of Iteration

IDARE®

Final Selection of Model and Variables

Decide the best ML models and Variables based on

- Check which ML model's Variable Importance most consistent with the physical understanding of the target
- Select your model by setting 1 error metric.
- The changes of errors based on ML models and different selected variables are very similar for between the error metrices
- Consider minimum difference between training and testing error
- Consider when Both error are the lowest
- Consider when all the cross-validation errors are similar or Standard deviations of cross validation data sets are the lowest
- Use your judgement