PRINCIPLES OF DATA MANAGEMENT

UNIVERSITY of HOUSTON
DIVISION OF RESEARCH

Step-by-Step Guide

- **How to create a model?**

  ➢ **Prepare the Data**

  - Data Preprocessing and Required Data Analysis

  - Feature Selection and Feature Engineering
    - Select Independent Features
      1. Correlation Analysis
      2. Univariate Selection
      3. Recursive feature elimination

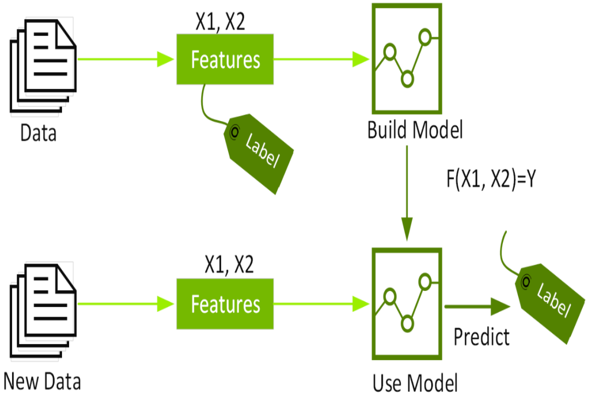

Fig-1
Image source: NVIDIA

Step-by-Step Guide

- **How to Train the Model?**
  - Select an appropriate model or algorithm for your problem
    - Regression Model : Continuous Data
      - **ORDINARY LEAST SQUARES REGRESSION (OLS)**

$Y = \beta_0 + \Sigma_{j=1..p} \beta_j X_j + \varepsilon$

Y: Target variable ; β0 : intercept ; e: random error[0-σ²]

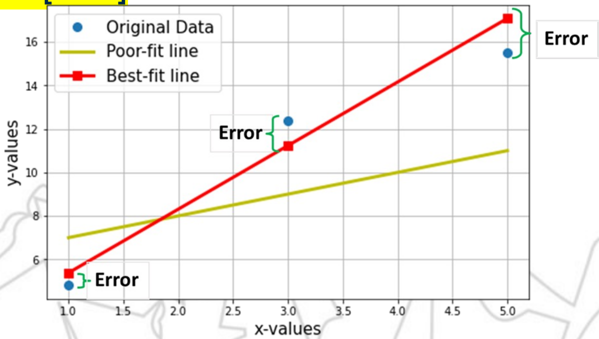

Fig -2
Image source: Analytics Vidhya

Step-by-Step Guide

- **XGBOOST(Extreme Gradient Boosting): Regressor**
  - **Boosting**:
    - Ensembles are constructed from decision tree models
    - Trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models.
  - **Gradient boosting:**
    - Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm.

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma).$$

L : Loss Function

Step-by-Step Guide

- Fit the model to the training data
- How to evaluate the model quality?

Regression Model Evaluation

- **Mean Squared Error (MSE):**
  - MSE is calculated by taking the average of the squared differences between the predicted and actual values

    MSE = (1/n) * Σ(y_actual - y_pred)^2
  - **Interpretation of MSE**
    - Higher MSE values -> larger errors between the predicted and actual values
    - MSE is sensitive to outliers

Step-by-Step Guide

- **What is R-squared?**
  - R-squared (R^2) is a statistical measure that represents the proportion of the variance in the dependent variable that can be explained by the independent variables in a regression model.

$$R^2 = 1 - (SSR / SST)$$

SSR: Sum of Squared differences b/w predicted and actual values

SST: Sum of Squared differences b/w the actual and mean of dependent features

- **Interpretation of R-Squared**
  - **Higher R^2 -> Model is a good fit**
  - **Lower R^2 -> Higher Residuals/errors**
  - **Values must be between 0 and 1**