Dependent variable — Independent variable

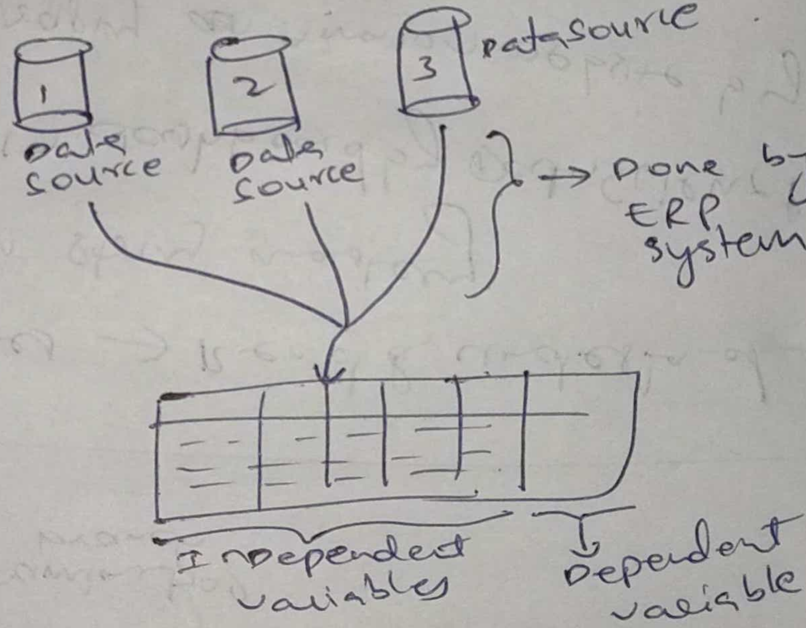
Association relationship

Linear Regression [supervised]
(Predictive Analytics)

Steps in building LR-model

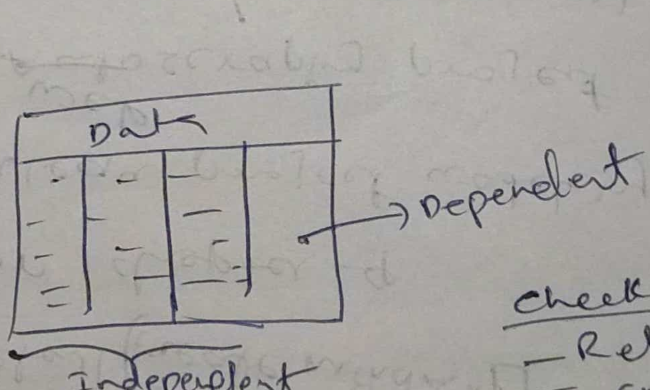
several iteration needed

①



Done by ERP system

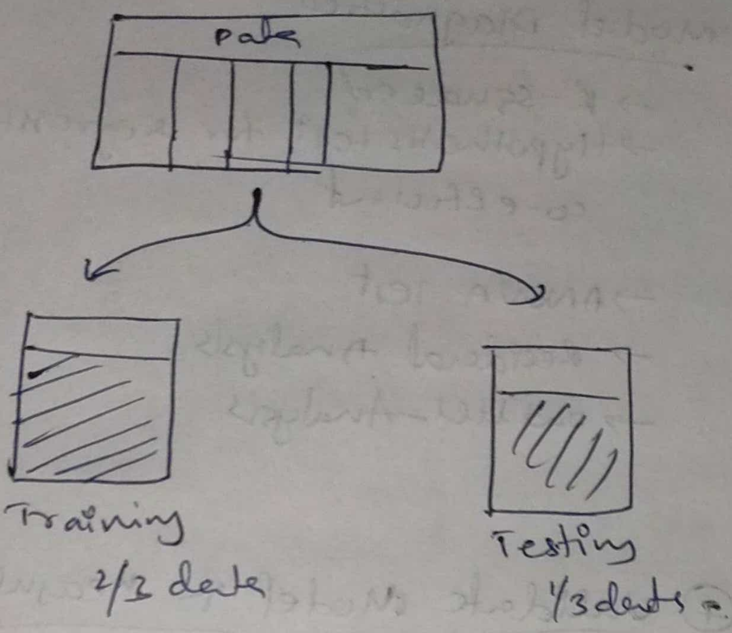
②



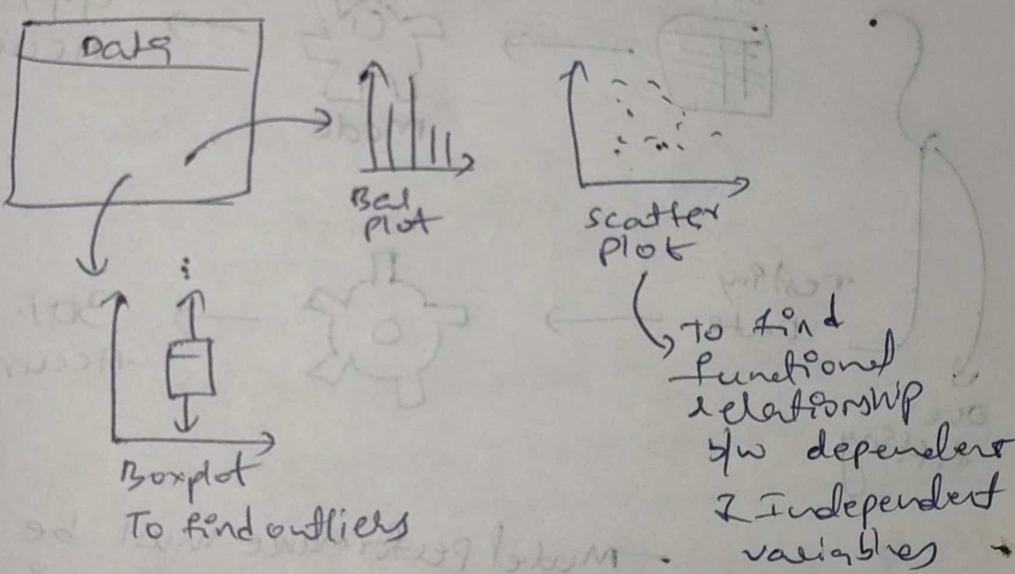
Check for
- Reliability
- completeness
- Accuracy
- Missing data
- Outliers

Techniques to handle
1) Data Imputation
2) New variables deriving
3) Handle categorical data by encoding

③



④



⑤

OLS $\longrightarrow$ To find Regression parameters

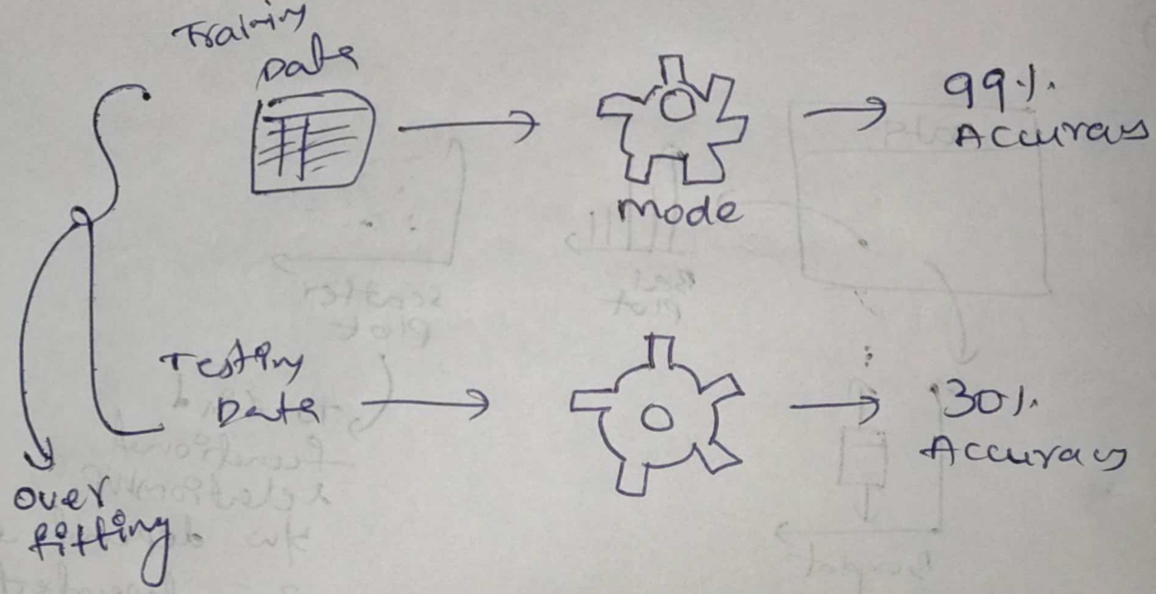
$$\boxed{y = \underbrace{\beta_0}_{\text{const}} + \beta_1 \underbrace{X}_{\text{Grade in 10th}} + \epsilon}$$

You will get β_0 & β_1

6 Model Diagnostics

- → R-squared
- → Hypothesis test for regression co-efficient
- → ANOVA Test
- → Residual Analysis
- → outlier Analysis

7 Validate Model & Measure Accuracy



- Model performance must be consistent on both training & testing dataset

- cross validate the model using multiple training and test datasets

8 Deploy the model

- According to the Business rules

$$\boxed{Y = \beta_0 + \beta_1 X + \varepsilon}$$

Y = Dependent variable

(β_0, β_1) = Regression parameters

ε = Random Error

X = Independent variable

For n observations,

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$$

where $i = 1, 2, 3 \dots n$

$$\boxed{\varepsilon_i = Y_i - \beta_0 + \beta_1 X_i}$$

β_0 & β_1 can be estimated by minimizing sum of squared errors (SSE)

$$SSE = \boxed{\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X)}$$

values of β_0 & β_1 are taken by partial derivatives of SSE

$$\hat{\beta}_1 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}$$

Above method is called as OLS method.

It yields BLUE (Best Lineal Unbiased Estimates)

Myths

→ Residuals follow Normal Distribution

→ (ϵ_i) is constant for various independent variables

→ ϵ_i & x are correlated

→ functional relationship b/w outcome variable & feature is correctly defined.

Properties

① Mean(y_i) = $\hat{\beta}_0 + \hat{\beta}_1 x$

② y_i follows Normal Distribution with mean $\hat{\beta}_0 + \hat{\beta}_1 x$ & variance VAR ϵ_i

$Y = \beta_0 + \beta_1 X + \varepsilon$

constant (1)

OLS API estimates (β_1) only to estimate β_0 a constant term meed to be added as new feature.

$Y_i = \beta_0 + \beta_1 x + \varepsilon$

OLS gives (β_0 & β_1)

$Y_i = 30587.285652 + (3560.587 * 62.1)$

For every 1% increase in grade '10' salary increases by 3560.587

Rahman