Data Science : it is a process of using data to find the solution(Data collection, Data Cleansing, Data exploration, feature engineering, model building, evaluation, deployment)

AI : any technique that enables computer to mimic human

ML : technique which learns from the examples or past data that too without explicitly programmed

1. supervised(labeled data) 2. un-supervised(no-labeled data- clustering) 3. re-enforcement learning(reward based, feedback based learning)

DL : subset of ML, it train itself to perform a task using neural network. works well with un-structured data(audeo,video,images)

Errors :

Bias : gap between actual and predicted value

small sample size does not have enough variation of data

exit pol result based on one city only

high error with training and testing both

underfitting

solution : get-more-training-data, increase number of parameters, increase complexity of model, increase training time until cost function is minimised

irreducible errors : can not be reduce

variance : tells howmuch scattered the predicted value from the actual value

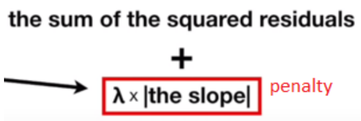
model perform well with training set but not with testing set

overfitting

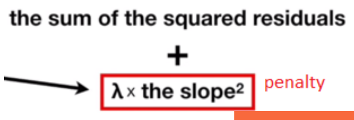
solution : adding-more-data, remove-some-features, Regularization , Cross-Validation, Ensembleling, early-stop, dropouts, reduce-hidden layer

adding some penalty to the model by reducing the co-efficient so that errors can be reduced

1. L1 Regularization - Lasso : exclude useless variables and less errors



2. L2 Regularization - redge : reduce more errors



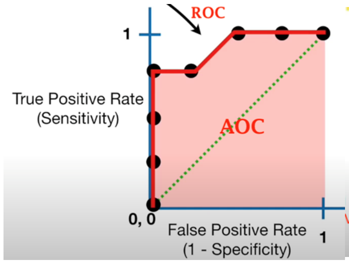
Regression Assumtion:

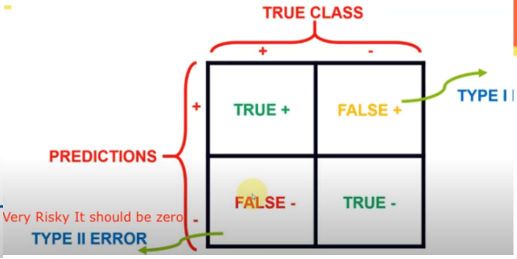
1. there should not be multi co-linearity
2. linear relation should be there between independent and dependent variable
3. Homoscadasticity should be there
4. No Outlier should be there
5. Data should be normally distributed

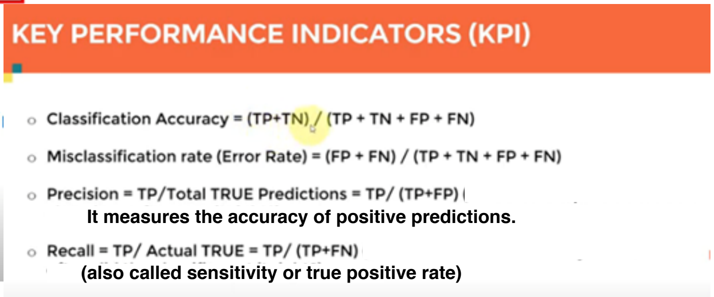
Evaluation Matrix

- **Regression**
  - Mean absolute Error
  - Mean Squared Error
- **classification**
  - Confusion Matrix

AOC ROC Curve(used with Binary classification problem)

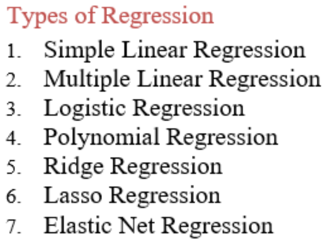






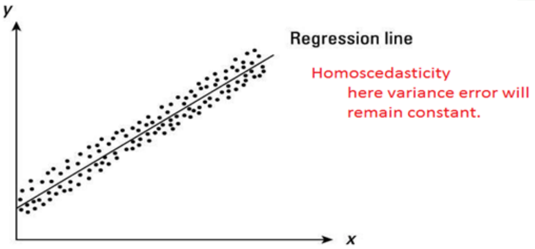
KEY PERFORMANCE INDICATORS (KPI)

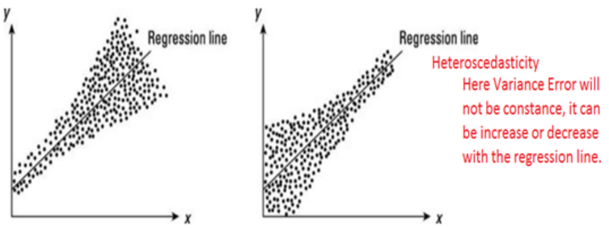
- Classification Accuracy = (TP+TN) / (TP + TN + FP + FN)
- Misclassification rate (Error Rate) = (FP + FN) / (TP + TN + FP + FN)
- Precision = TP/Total TRUE Predictions = TP/ (TP+FP)
  It measures the accuracy of positive predictions.
- Recall = TP/ Actual TRUE = TP/ (TP+FN)
  (also called sensitivity or true positive rate)

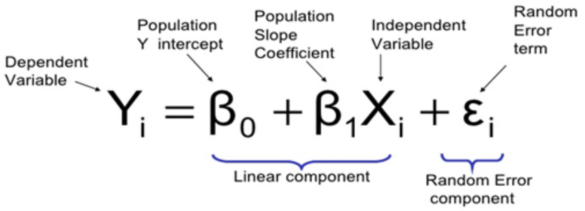


Types of Regression

1. Simple Linear Regression
2. Multiple Linear Regression
3. Logistic Regression
4. Polynomial Regression
5. Ridge Regression
6. Lasso Regression
7. Elastic Net Regression



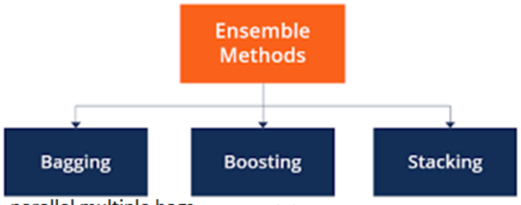




$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

List of classification algorithms

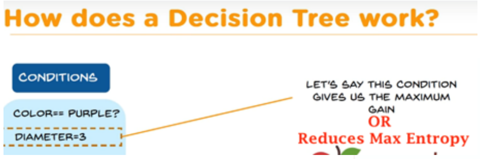
1. Linear Classifiers: Logistic Regression(Single Class Classification), Naive Bayes Classifier(Multiclass classification)
2. KNN(K Nearest Neighbors)
3. Support Vector Machines
4. Decision Trees
5. Random Forest

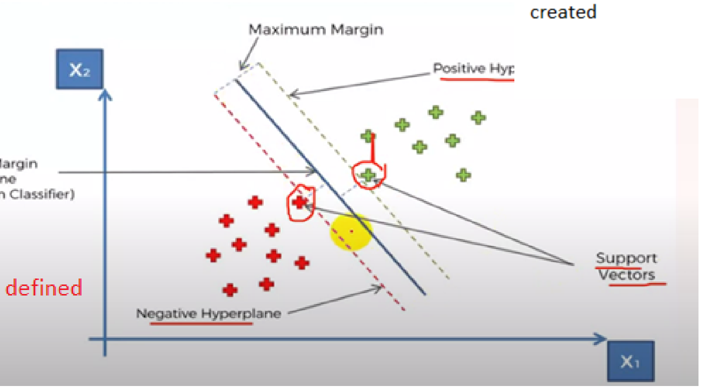


parallel multiple bags created

sequencial multiple weak learner generated one strong model

multiple algo created





Imbalanced-DataSets

1% fraud only then most of the ml model fails to detect that

1. Under Sampling Majority Class
2. Over Sampling Minority Class(take duplicate data and make it the count same as bigger class)
3. Over Sampling Minority Class using SMOTE(it uses k-nearest to generate new data)
4. use Ensemble Method : It divide majority class into multiple batches as minority and training will happens with 1 batch of majority with minority class.
5. Focal Loss : will penalize the majority sample during loss calculation and give more weightage to minority class

The F1 score is the harmonic mean of precision and recall. It balances both precision and recall and is useful when the classes are imbalanced. The F1 score ranges between 0 and 1, where 1 represents perfect precision and recall, and 0 indicates poor performance.