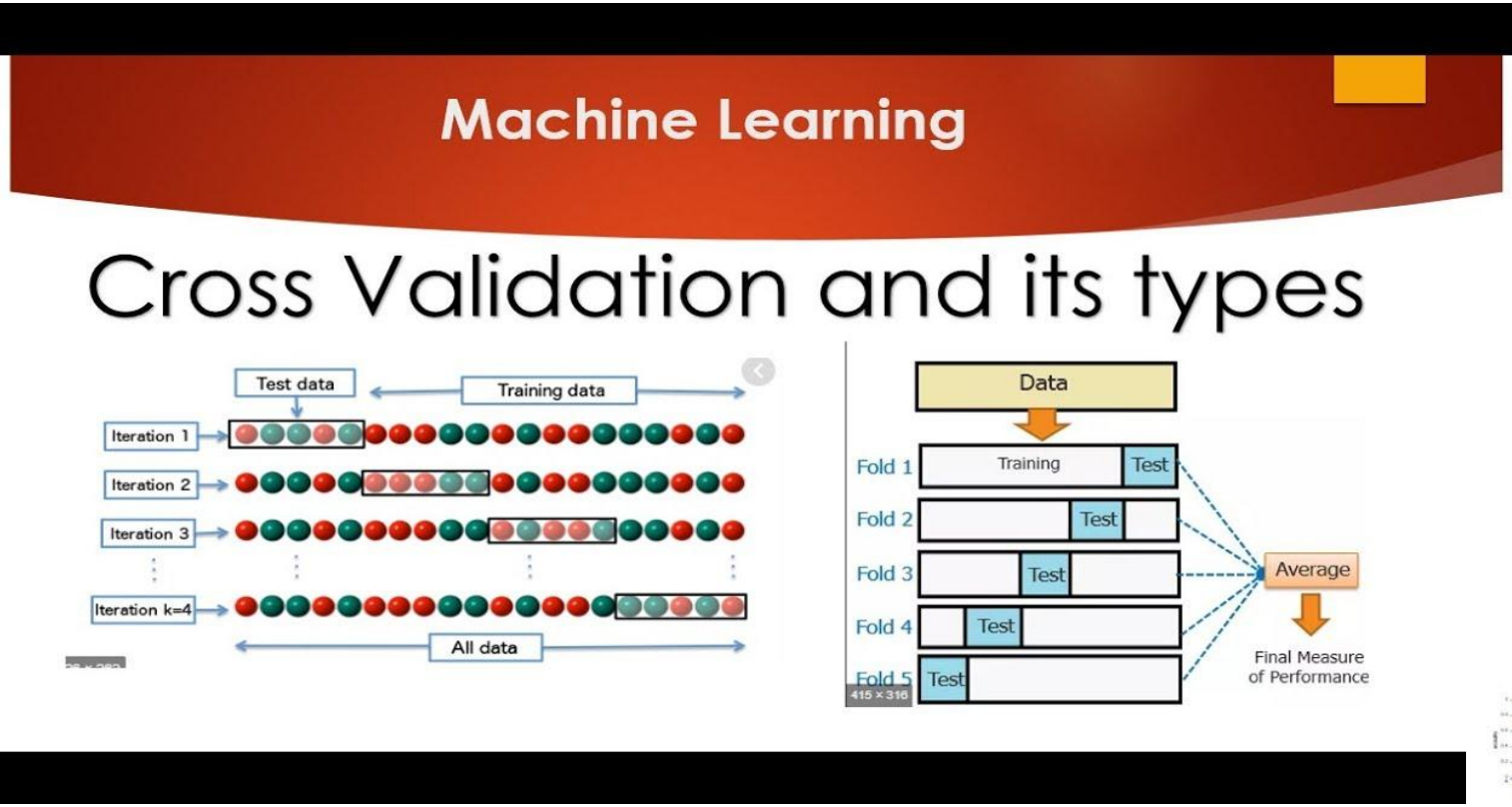
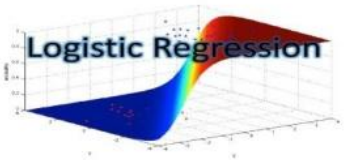
Cross Validation

INTERNSHIPSTUDIO

- **Cross-validation** is a statistical **technique** which involves partitioning the data into subsets, training the data on a subset and use the other subset to evaluate the model's performance.
- To reduce variability we perform multiple rounds of **cross-validation** with different subsets from the same data.

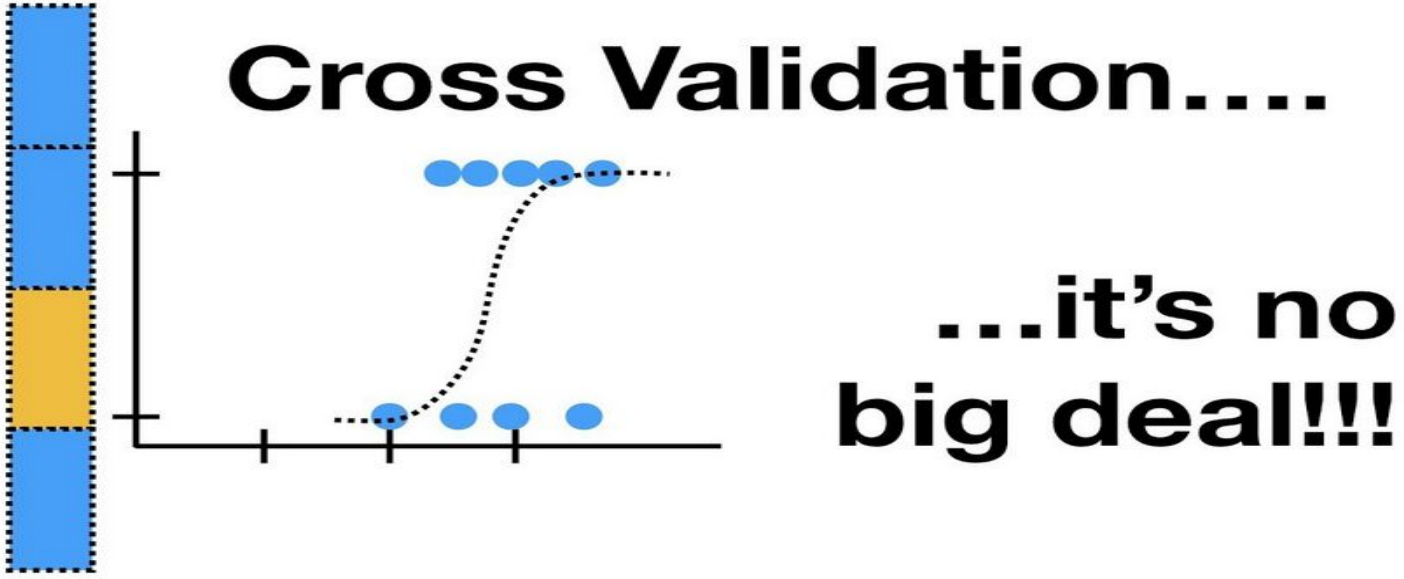


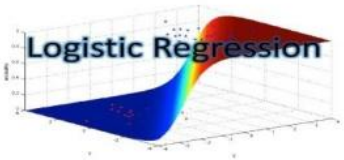


K-fold Cross-Validation

INTERNSHIPSTUDIO

- In k-fold cross-validation, the data is divided into k folds. The model is trained on k-1 folds with one fold held back for testing.
- This process gets repeated to ensure each fold of the dataset gets the chance to be the held back set.
- Once the process is completed, we can summarize the evaluation metric using the mean or/and the standard deviation.



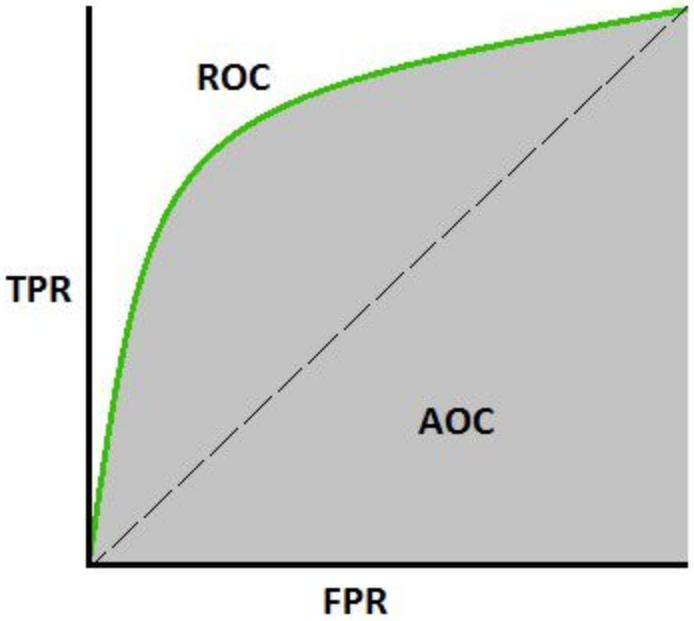


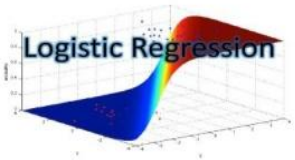
ROC Curve

AUC - ROC curve is a performance measurement for classification problem at various thresholds settings.

- ROC (receiver operating characteristic curve) is a probability curve and AUC represents degree or measure of separability.
- It tells how much model is capable of distinguishing between classes.
- Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.

The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.

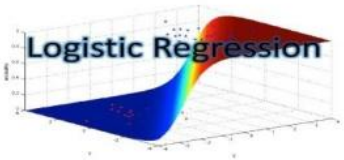




Definition of the Terms

- Positive (P) : Observation is positive (for example: is an apple).
- Negative (N) : Observation is not positive (for example: is not an apple).
- True Positive (TP) : Observation is positive, and is predicted to be positive.
- False Negative (FN) : Observation is positive, but is predicted negative.
- True Negative (TN) : Observation is negative, and is predicted to be negative.
- False Positive (FP) : Observation is negative, but is predicted positive.

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



ROC Curve

INTERNSHIPSTUDIO

Defining terms used in AUC and ROC Curve.

TPR (True Positive Rate) / Recall /Sensitivity

$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

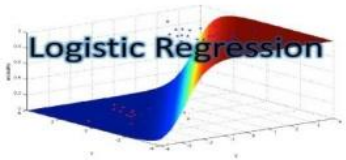
Specificity

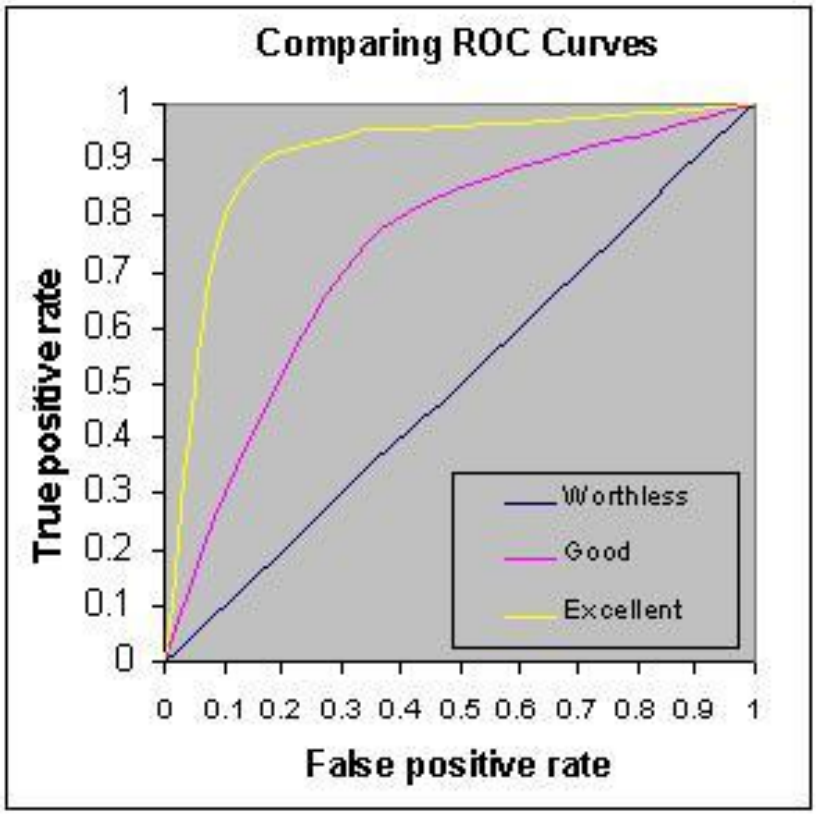
$$\text{Specificity} = \frac{TN}{TN + FP}$$

FPR

$$\text{FPR} = 1 - \text{Specificity}$$

$$= \frac{FP}{TN + FP}$$

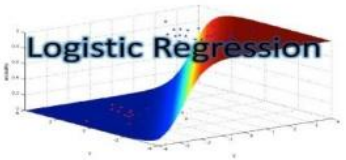




Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm.

	Class 1 Predicted	*Class 2 Predicted*
Class 1 Actual	**TP**	**FN**
Class 2 Actual	**FP**	**TN**



Confusion Matrix

Recall:

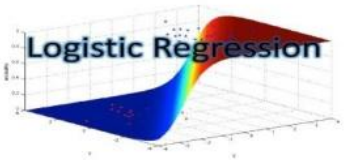
Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (a small number of FN).

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

Precision:

To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High Precision indicates an example labelled as positive is indeed positive (a small number of FP).



Confusion Matrix

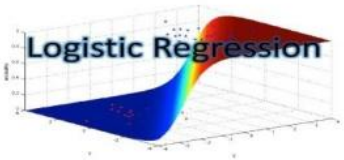
F-measure:

Since we have two measures (Precision and Recall) it helps to have a measurement that represents both of them. We calculate an F-measure which uses Harmonic Mean in place of Arithmetic Mean as it punishes the extreme values more. The F-Measure will always be nearer to the smaller value of Precision or Recall.

$$F\text{-}measure = \frac{2*Recall*Precision}{Recall + Precision}$$

Example to interpret confusion matrix:

n = 165	Predicted: No	Predicted: Yes
Actual: No	50	10
Actual: Yes	5	100



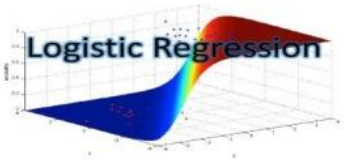
Implementing LR Model

INTERNSHIPSTUDIO

Fitting Logistic Regression

```
[ ] #Fitting Logistic Regresssion to Training Set
    from sklearn.linear_model import LogisticRegression
    classifier = LogisticRegression(random_state =0)
    classifier.fit(x_train,y_train)
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=0, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False)
```



Analyzing Result Parameters

INTERNSHIPSTUDIO

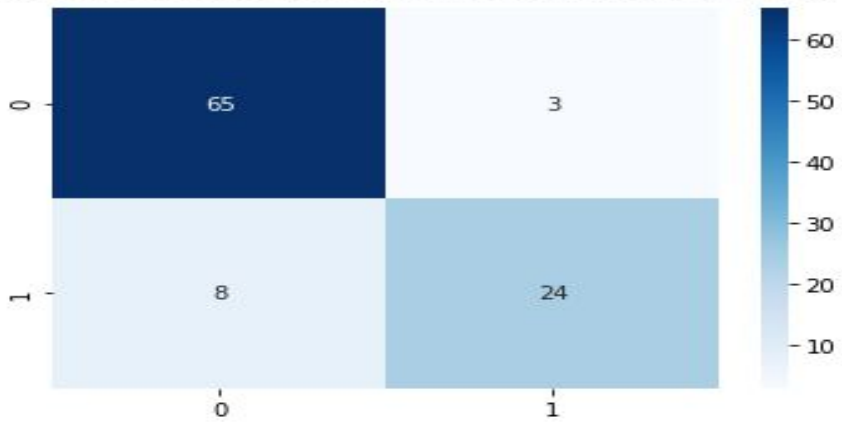
- Predicting the results

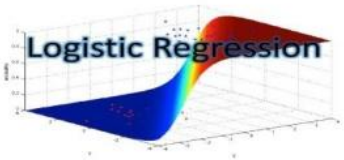
```
[ ]  #Predicting the Test Set Results
     y_pred = classifier.predict(x_test)
     y_pred
```

```
array([0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1,
       0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
       1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1,
       0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1,
       0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1])
```

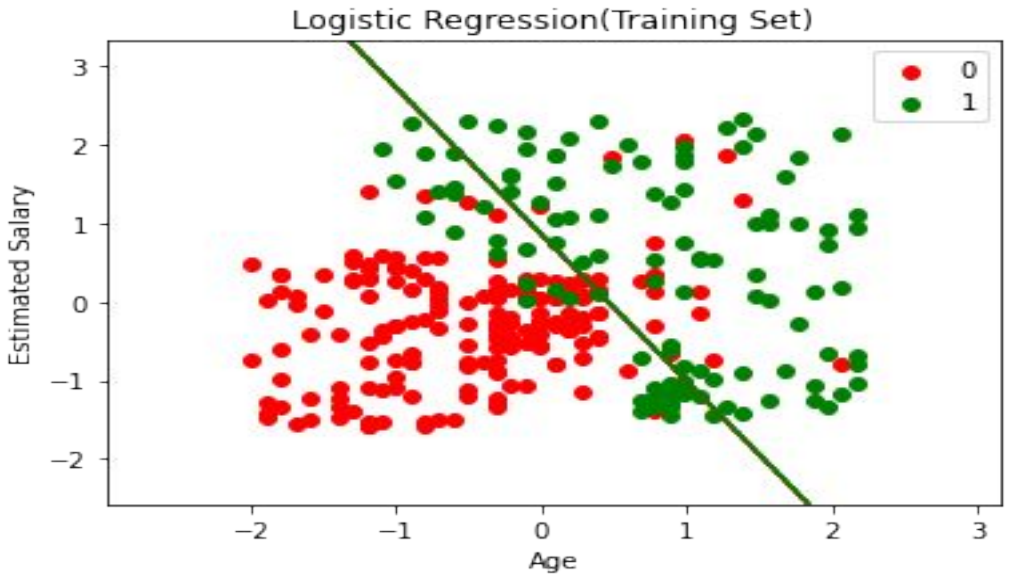
```
[ ]  import seaborn as sns
     sns.heatmap(cm, annot=True, cmap='Blues')
```

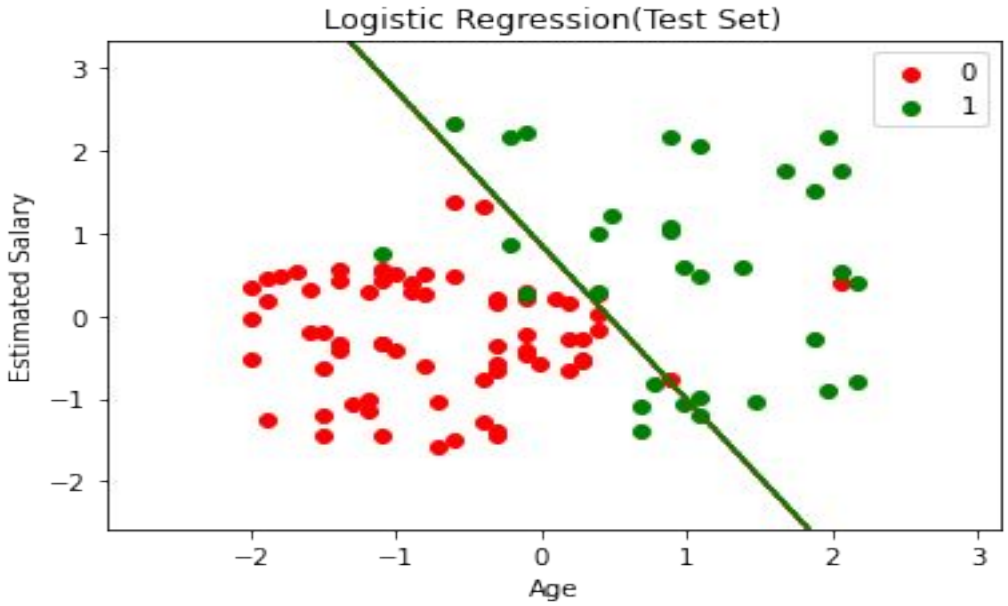
```
/usr/local/lib/python3.6/dist-packages/statsmodels/
  import pandas.util.testing as tm
<matplotlib.axes._subplots.AxesSubplot at 0x7fa988e
```

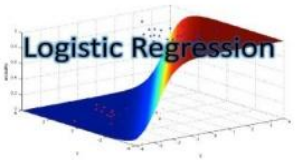




Analyzing Result Parameters







INTERNSHIPSTUDIO

Q.1 What is Cross Validation?

Q.2 What is Confusion Matrix?

Q.3 What is ROC Curve?

Q.4 What is TP ?

Q.5 What is Precision ?

Q.6 What is Recall ?