

|  |
| --- |
| Data Scientist  2018 |
| Image result for tempus |
| September 18  Tempus  Authored by: Ratnam Dubey |

Contents

[Overview and Motivation 3](#_Toc524964522)

[Classification Problem 3](#_Toc524964523)

[Problem Solve Approach 4](#_Toc524964524)

[Data Observation and Cleaning 5](#_Toc524964525)

[Exploratory Data Analysis 5](#_Toc524964526)

[Data Cleanup 6](#_Toc524964527)

[Response Variable Analysis 6](#_Toc524964528)

[Challenges with standard Machine learning techniques 6](#_Toc524964529)

[Approach to handling Imbalanced Datasets 7](#_Toc524964530)

[Data Level approach: Resampling Techniques 7](#_Toc524964531)

[Train and Test Data Split 7](#_Toc524964532)

[Synthetic Minority Over-sampling Technique 7](#_Toc524964533)

[PCA Analysis 8](#_Toc524964534)

[Machine learning 10](#_Toc524964535)

[Supervised Learning 10](#_Toc524964536)

[Model Selection 10](#_Toc524964537)

[Tuning Hyperparameter 11](#_Toc524964538)

[Model Testing. 14](#_Toc524964539)

[Confusion Matrix 14](#_Toc524964540)

[Area Under the ROC curve (AUC – ROC) 15](#_Toc524964541)

[References. 16](#_Toc524964542)

# ****Overview and Motivation****

The goals of our project is to develop accurate classifiers for large-scale data analysis of a dataset. The Task Given by the Tempus is to develop the Machine learning pipeline which includes from cleaning of the Data to classifying the correct predictors.

Data Overview

Data Consist of the Various parameters given in the Binary format and the response variable also provided in the Binary format.

In total of 16563 columns-predictors and 1 response variable that needed to be predicted with help of 16563 predictors. Also, there are total 530 Observations provided in the dataset.

All the predictors are in the Binary format i.e. (0 or 1) and the response variable is also in the (0 or 1)

Which means it’s a classification class problem which needed to be solved with the help of 16563 predictors.

## Classification Problem

In [machine learning](https://en.wikipedia.org/wiki/Machine_learning) and [statistics](https://en.wikipedia.org/wiki/Statistics), classification is the problem of identifying to which of a set of [categories](https://en.wikipedia.org/wiki/Categorical_data) (sub-populations) a new [observation](https://en.wikipedia.org/wiki/Observation) belongs, on the basis of a [training set](https://en.wikipedia.org/wiki/Training_set) of data containing observations (or instances) whose category membership is known. Examples are assigning a given email to the ["spam" or "non-spam"](https://en.wikipedia.org/wiki/Spam_filtering) class, and assigning a diagnosis to a given patient based on observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.). Classification is an example of [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition).

In the terminology of machine learning, classification is considered an instance of [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning), i.e. learning where a training set of correctly identified observations is available. The corresponding [unsupervised](https://en.wikipedia.org/wiki/Unsupervised_learning) procedure is known as [clustering](https://en.wikipedia.org/wiki/Cluster_analysis), and involves grouping data into categories based on some measure of inherent similarity or [distance](https://en.wikipedia.org/wiki/Distance). - Wikipedia

# ****Problem Solve Approach****

The tempus data science problem needed to be solved in the sorted manner to avoid the confusion as to accomplish this its been decided to solve the problem in Five phases

As to accomplish this below methodology has been adopted for the analysis and successful prediction of the data.

Phase 1 – Consist of Data Observation and Cleaning of the Data

Phase 2 –PCA Analysis

Phase 3 – Machine Learning approach Iterative (till satisfactory result).

Phase 4 – Model Testing.

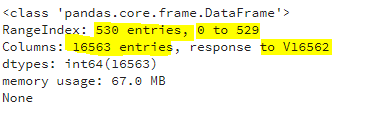
Phase 5 – Model Deployment

# Data Observation and Cleaning

## Exploratory Data Analysis

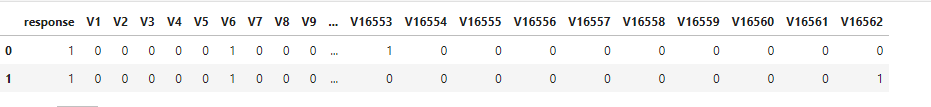
Exploration of the Dataset is important part of the Machine learning problems.

First, identify Predictor (Input) and Target (output) variables. Next, identify the data type and category of the variables.



Consists of 16563 Predictor and 1 Target variable with having 530 observations.

Data Overview



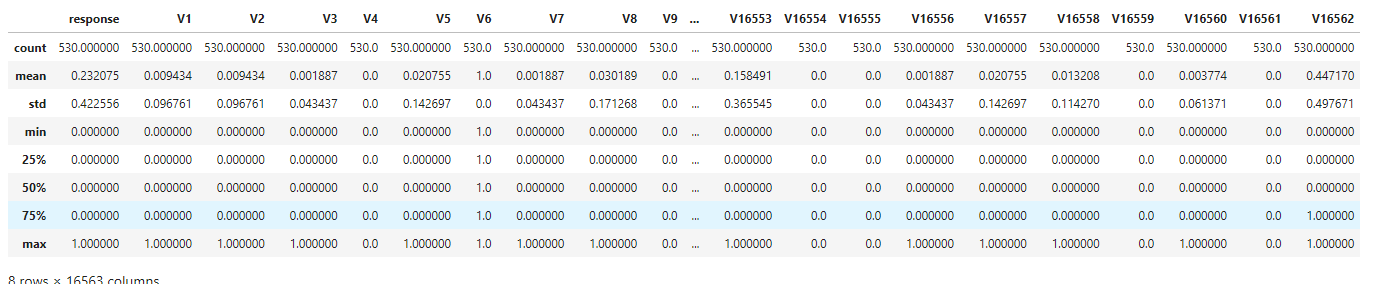
Shape of the Data



Number of Dimensions in the data



Data Describe



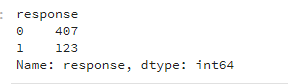
## Data Cleanup

In Exploration of the Data it is been found that out of 16563 predictors some of them doesn’t show any variance in the data. Which states that it consists of single value in the 530 observations.

For Example, V4 has 530 observations and out of 530 observations none of the observations have different values i.e. the value is completely 0 or completely 1 for all the observations. Removal of such variables are necessary to save the computational power. Although this can also be accomplished with PCA, but the computational space can be saved with manual removal of such observations.

## Response Variable Analysis

Exploring the response variable and the distribution of the response variable.



Response Variable is Uneven which is the problem why it is a problem?

This is a scenario where the number of observations belonging to one class is significantly lower than those belonging to the other classes. This problem is predominant in scenarios where anomaly detection is crucial like electricity pilferage, fraudulent transactions in banks, identification of rare diseases, etc. In this situation, the predictive model developed using conventional machine learning algorithms could be biased and inaccurate.

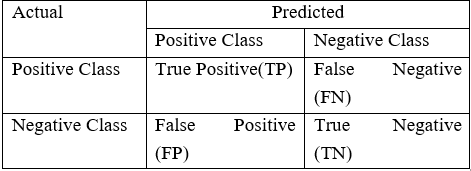
This happens because Machine Learning Algorithms are usually designed to improve accuracy by reducing the error. Thus, they do not consider the class distribution / proportion or balance of classes.

## Challenges with standard Machine learning techniques

The conventional model evaluation methods do not accurately measure model performance when faced with imbalanced datasets.

Standard classifier algorithms like Decision Tree and Logistic Regression have a bias towards classes which have number of instances. They tend to only predict the majority class data. The features of the minority class are treated as noise and are often ignored. Thus, there is a high probability of misclassification of the minority class as compared to the majority class.

Evaluation of a classification algorithm performance is measured by the Confusion Matrix which contains information about the actual and the predicted class.

  
Accuracy of a model = (TP+TN) / (TP+FN+FP+TN)

However, while working in an imbalanced domain accuracy is not an appropriate measure to evaluate model performance. For e.g.: A classifier which achieves an accuracy of 98 % with an event rate of 2 % is not accurate, if it classifies all instances as the majority class. And eliminates the 2 % minority class observations as noise.

## Approach to handling Imbalanced Datasets

### Data Level approach: Resampling Techniques

Dealing with imbalanced datasets entails strategies such as improving classification algorithms or balancing classes in the training data (data preprocessing) before providing the data as input to the machine learning algorithm. The later technique is preferred as it has wider application.

The main objective of balancing classes is to either increasing the frequency of the minority class or decreasing the frequency of the majority class. This is done to obtain approximately the same number of instances for both the classes.

### Train and Test Data Split

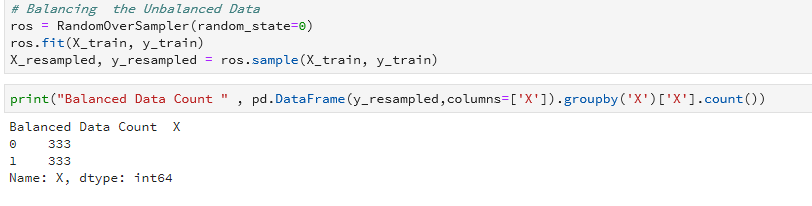
Split arrays or matrices into random train and test subsets.

* Size of training data is 80% of the Original dataset
* Size of the test data is 20% of the Original dataset

It is needed to be done before oversampling of the data takes place so that biased output problem can be avoided.

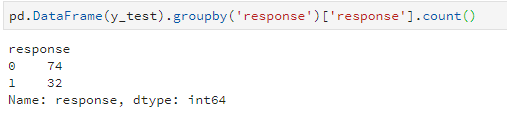
### Synthetic Minority Over-sampling Technique

Over-Sampling increases the number of instances in the minority class by randomly replicating them to present a higher representation of the minority class in the sample. This technique is followed to avoid overfitting which occurs when exact replicas of minority instances are added to the main dataset. A subset of data is taken from the minority class as an example and then new synthetic similar instances are created. These synthetic instances are then added to the original dataset. The new dataset is used as a sample to train the classification models.



Balanced Dataset Count is 333 for response variable 1 and 0

Test Data size is 106 Observations (Non-Balanced Dataset – Test Data)



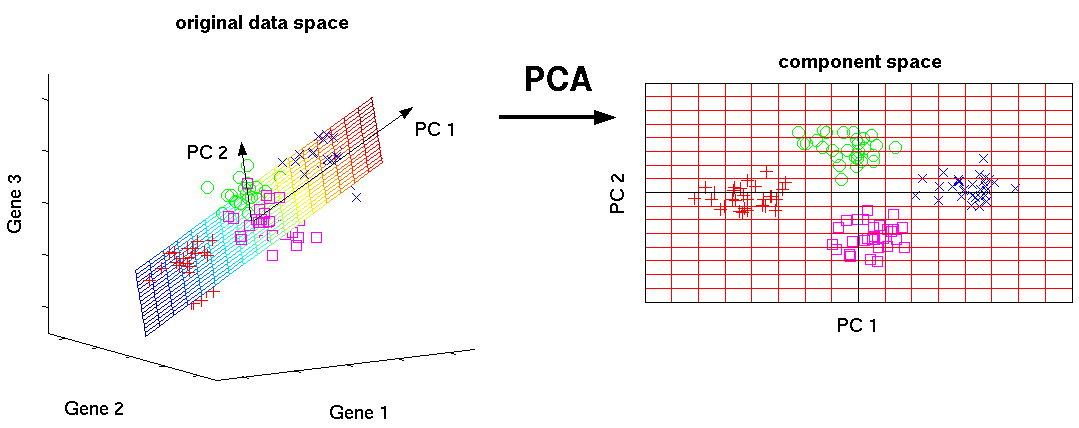
# PCA Analysis

In simple words, principal component analysis is a method of extracting important variables (in form of components) from a large set of variables available in a data set. It extracts low dimensional set of features from a high dimensional data set with a motive to capture as much information as possible. With fewer variables, visualization also becomes much more meaningful. PCA is more useful when dealing with 3 or higher dimensional data. It is always performed on a symmetric correlation or covariance matrix. This means the matrix should be numeric and have standardized data.

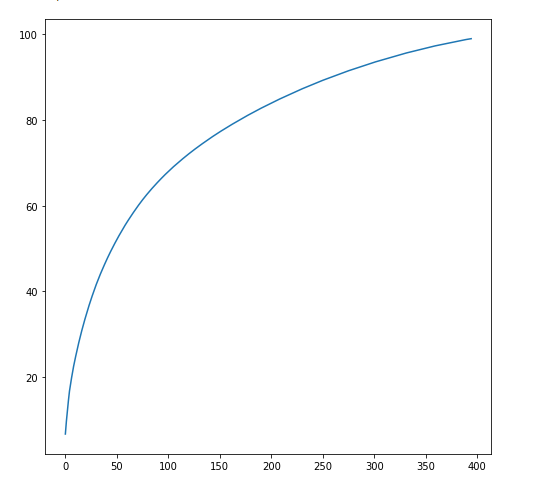
Suppose we have a data set of dimensions 300 (*n*) × 50 (*p*). *n* represents the number of observations and *p* represents number of predictors. Since we have a large p = 50, there can be p(p-1)/2 scatter plots i.e. more than 1000 plots possible to analyze the variable relationship. Wouldn’t it be a tedious job to perform exploratory analysis on this data?

In this case, it would be a lucid approach to select a subset of *p* *(p << 50)* predictor which captures as much information. Followed by plotting the observation in the resultant low dimensional space.

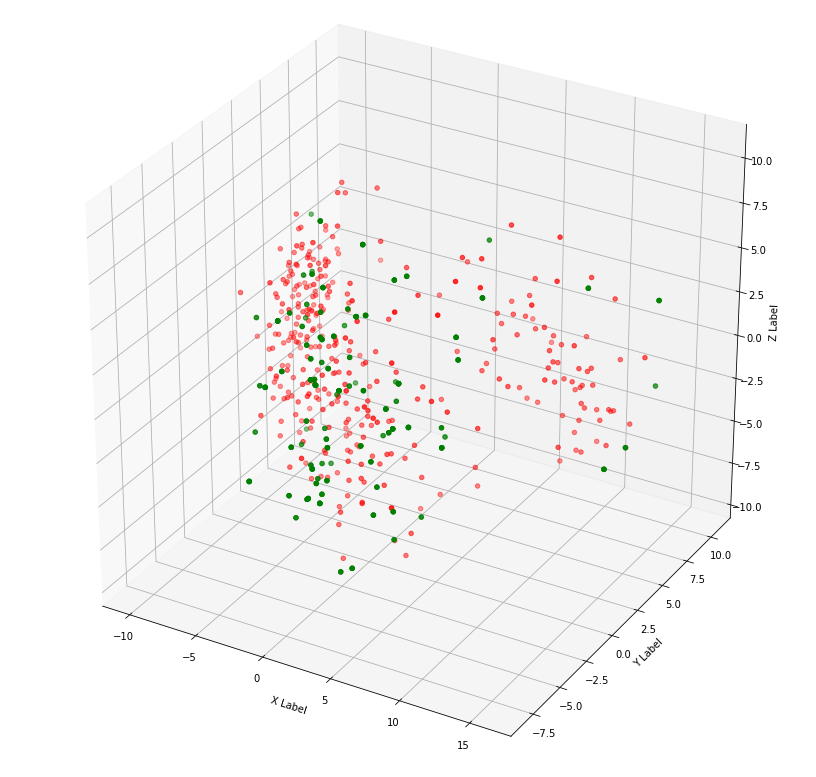
The image below shows the transformation of a high dimensional data (3 dimension) to low dimensional data (2 dimension) using PCA.



Scale of the Data is the important before applying the PCA but in the provided dataset the scalarization of the data is already in place, so we can skip the scalarization step in the process. Capturing the variance



As we have observed that the cumulative variance can be captured with very little features. Using the observed result in transformation of the data. But before that checking the data into the 3d plane for observation purpose only.



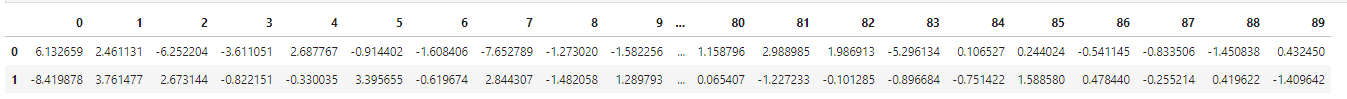
Green dots represent the 1 and 0 represents in the Red Color

It can be observed that the it is hard to classify and segregate the data in the 3 dimensions we need to take the data in more than three dimensions so that data can be classified properly.

After iterative approach it has been found that the best segregation of the data is observed in the 5-dimension

space.

Transformation of the data in the 90-D space



# Machine learning

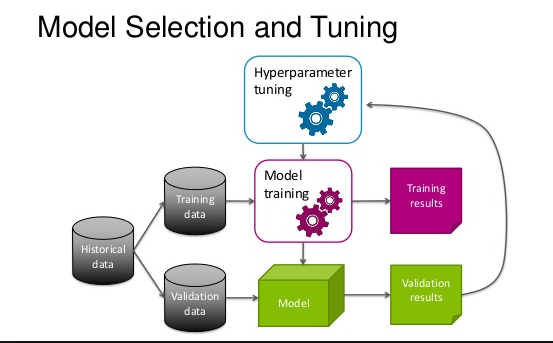
Machine learning is a field of [computer science](https://en.wikipedia.org/wiki/Computer_science) that uses statistical techniques to give [computer systems](https://en.wikipedia.org/wiki/Computer_systems) the ability to "learn" (e.g., progressively improve performance on a specific task) with [data](https://en.wikipedia.org/wiki/Data), without being explicitly programmed

## Supervised Learning

 This algorithm consists of a target / outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using these set of variables, we generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data. Examples of Supervised Learning: Regression, [Decision Tree](https://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/), [Random Forest](https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/), KNN, Logistic Regression etc.

## Model Selection

Model selection is the task of selecting a [statistical model](https://en.wikipedia.org/wiki/Statistical_model) from a set of candidate models, given data. In the simplest cases, a pre-existing set of data is considered. However, the task can also involve the [design of experiments](https://en.wikipedia.org/wiki/Design_of_experiments) such that the [data collected](https://en.wikipedia.org/wiki/Data_collection) is well-suited to the problem of model selection. Given candidate models of similar predictive or explanatory power, the simplest model is most likely to be the best choice



Model performance is evaluated on the AUC\_ROC score:

As to evaluate this process certain models were used for analysis and then selection of the best performing model on the balanced dataset is chosen.

Algorithms evaluated in the selection process are

KNN – K nearest neighbor

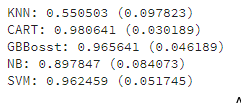
CART – Decision trees

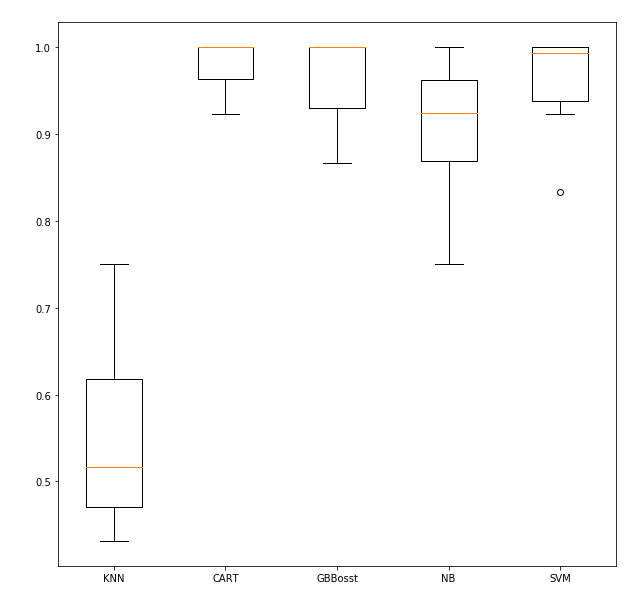
GBBoost – Gradient Boosting

NB – Naïve Bias

SVM – Support Vector Machine

Results in the Model Selection and Performance are given below: -





As it has been observed that the SVM has provided the best results on the Observed dataset so SVM has been selected as the lead model in the selection process.

## Tuning Hyperparameter

Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot).

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_1.png)

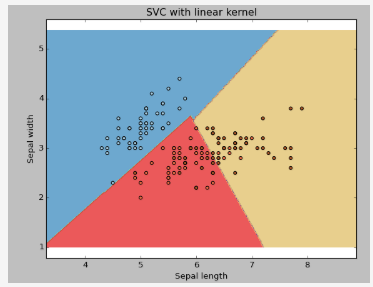
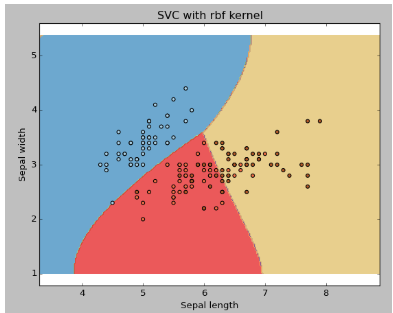
Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line).

Tuning parameters value for machine learning algorithms effectively improves the model performance Below are the list of parameters available with SVM.

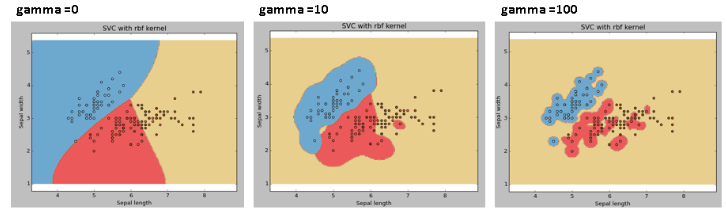
sklearn.svm.SVC(*C=1.0*, *kernel='rbf'*, *degree=3*, *gamma=0.0*, *coef0=0.0*, *shrinking=True*, *probability=False*,*tol=0.001*, *cache\_size=200*, *class\_weight=None*, *verbose=False*, *max\_iter=-1*, *random\_state=None*)

we are optimizing the Parameter with the help of “kernel”, “gamma” and “C”.

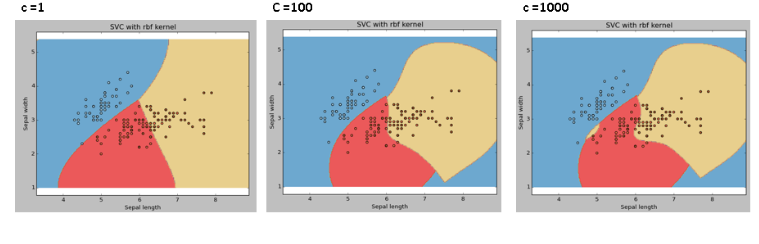
kernel: We have various options available with kernel like, “linear”, “rbf”,”poly” and others (default value is “rbf”).  Here “rbf” and “poly” are useful for non-linear hyper-plane.

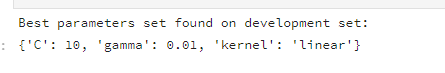
gamma: Kernel coefficient for ‘rbf’, ‘poly’ and ‘sigmoid’. Higher the value of gamma, will try to exact fit the as per training data set i.e. generalization error and cause over-fitting problem.



C: Penalty parameter C of the error term. It also controls the tradeoff between smooth decision boundary and classifying the training points correctly.



With the help of Grid Selection method finding the best parameters for the SVM



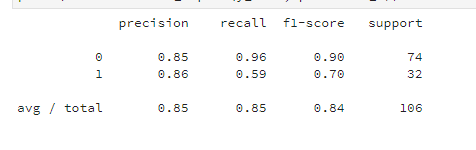
Best parameters for the SVM model are

C – 10

Gamma – 0.01

Kernel – ‘linear

Applying the tuned parameter in the SVM algorithm and checking the performance of the model.



Testing the model on the test data, that been created before Over Sampling of data takes place and scoring of the hyper tuned SVM model on the test data is around 84%

Accuracy is calculated = (TP + TN) / (TP+TN+FN+FP)

# Model Testing.

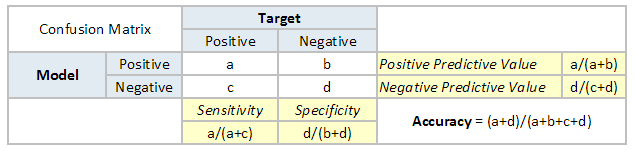
Predictive Modeling works on constructive feedback principle. You build a model. Get feedback from metrics, make improvements and continue until you achieve a desirable accuracy. Evaluation metrics explain the performance of a model. An important aspects of evaluation metrics is their capability to discriminate among model results.

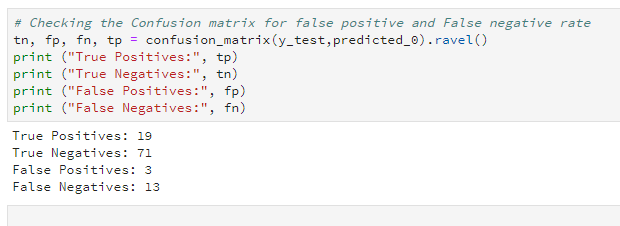
Simply, building a predictive model is not a motive. But, creating and selecting a model which gives high accuracy on out of sample data. Hence, it is crucial to check accuracy of the model prior to computing predicted values.

## Confusion Matrix

A confusion matrix is an N X N matrix, where N is the number of classes being predicted. For the problem in hand, we have N=2, and hence we get a 2 X 2 matrix. Here are a few definitions, you need to remember for a confusion matrix :

* Accuracy : the proportion of the total number of predictions that were correct.
* Positive Predictive Value or Precision : the proportion of positive cases that were correctly identified.
* Negative Predictive Value : the proportion of negative cases that were correctly identified.
* Sensitivity or Recall : the proportion of actual positive cases which are correctly identified.
* Specificity : the proportion of actual negative cases which are correctly identified.

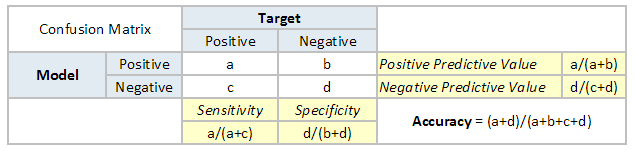
[](https://www.analyticsvidhya.com/blog/wp-content/uploads/2015/01/Confusion_matrix.png)



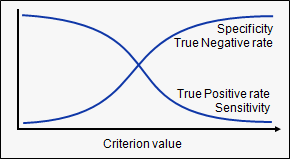
## Area Under the ROC curve (AUC – ROC)

This is again one of the popular metrics used in the industry.  The biggest advantage of using ROC curve is that it is independent of the change in proportion of responders. This statement will get clearer in the following sections.

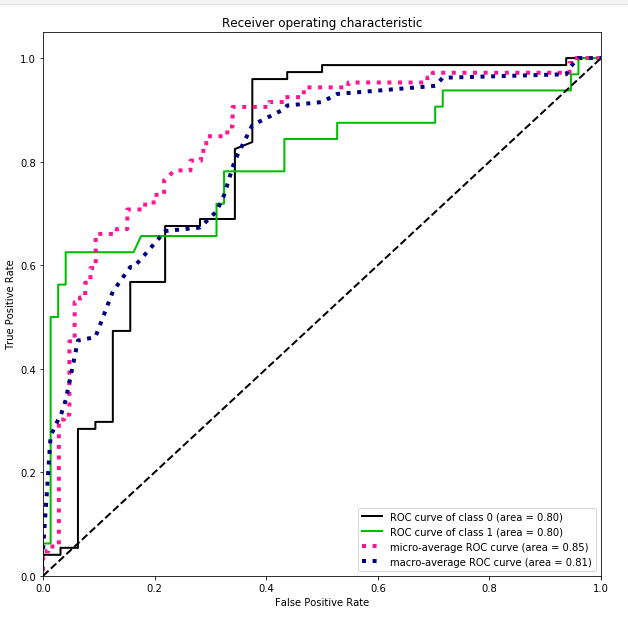
ROC is (Receiver operating characteristic) curve. If we look at the confusion matrix below, we observe that for a probabilistic model, we get different value for each metric.

[](https://www.analyticsvidhya.com/blog/wp-content/uploads/2015/01/Confusion_matrix.png)

Hence, for each sensitivity, we get a different specificity. The two vary as follows:

[](https://www.analyticsvidhya.com/blog/wp-content/uploads/2015/01/curves.png)

The ROC curve is the plot between sensitivity and (1- specificity). (1- specificity) is also known as false positive rate and sensitivity is also known as True Positive rate. Following is the ROC curve for the data.

 AUC & ROC score is around 80% which is considered as the Good model. Deployment of the model can depend on the use-case it can be extracted in any format and can be used for the prediction of the values.

# References.

* Wikipedia for theory content
* Google Images for Images data
* Scikit learn documentations