<https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/>

**Introduction**

Linear and Logistic regressions are usually the first algorithms people learn in predictive modeling. Due to their popularity, a lot of analysts even end up thinking that they are the only form of regressions. The ones who are slightly more involved think that they are the most important amongst all forms of regression analysis.

The truth is that there are innumerable forms of regressions, which can be performed. Each form has its own importance and a specific condition where they are best suited to apply. In this article, I have explained the most commonly used 7 forms of regressions in a simple manner. Through this article, I also hope that people develop an idea of the breadth of regressions, instead of just applying linear / logistic regression to every problem they come across and hoping that they would just fit!

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

Regression analysis is an important tool for modelling and analyzing data. Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.

**Why do we use Regression Analysis?**

Let’s say, you want to estimate growth in sales of a company based on current economic conditions. You have the recent company data which indicates that the growth in sales is around two and a half times the growth in the economy. Using this insight, we can predict future sales of the company based on current & past information.

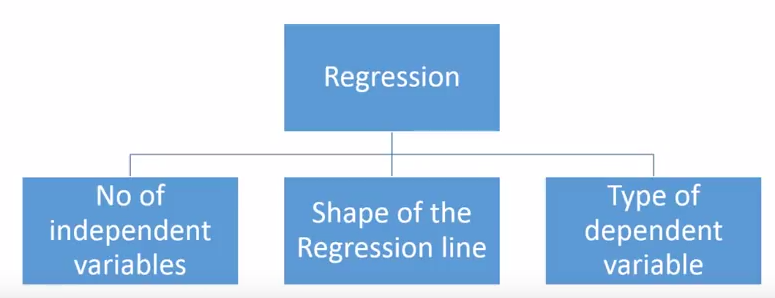
There are multiple benefits of using regression analysis. They are as follows:

* It indicates the **significant relationships** between dependent variable and independent variable.
* It indicates the **strength of impact** of multiple independent variables on a dependent variable.

Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

**How many types of regression techniques do we have?**

There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line). We’ll discuss them in detail in the following sections.



For the creative ones, you can even cook up new regressions, if you feel the need to use a combination of the parameters above, which people haven’t used before. But before you start that, let us understand the most commonly used regressions:

## 1. Linear Regression

It is one of the most widely known modeling technique. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the dependent variable is continuous, independent variable(s) can be [continuous or discrete](https://en.wikipedia.org/wiki/Continuous_and_discrete_variables), and nature of regression line is linear.

Linear Regression establishes a relationship between **dependent variable (Y)** and one or more **independent variables (X)** using a **best fit straight line** (also known as regression line).

It is represented by an equation **Y=a+b\*X + e**, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

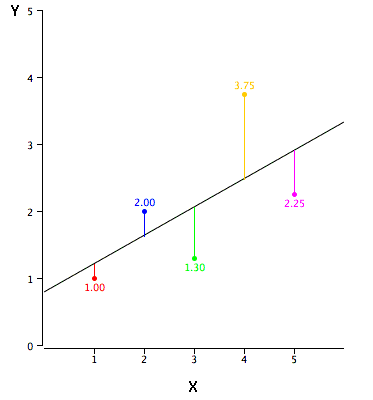
[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Linear_Regression1.png)

The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable.  Now, the question is “How do we obtain best fit line?”.

#### How to obtain best fit line (Value of a and b)?

This task can be easily accomplished by Least Square Method. It is the most common method used for fitting a regression line. It calculates the best-fit line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. Because the deviations are first squared, when added, there is no cancelling out between positive and negative values.

[Least_Square](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Least_Square.png)

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/reg_error.gif)

We can evaluate the model performance using the metric **R-square**. To know more details about these metrics, you can read: Model Performance metrics [Part 1](https://www.analyticsvidhya.com/blog/2015/01/model-performance-metrics-classification/), [Part 2](https://www.analyticsvidhya.com/blog/2015/01/model-perform-part-2/) .

#### Important Points:

* There must be **linear relationship** between independent and dependent variables
* Multiple regression suffers from **multicollinearity, autocorrelation, heteroskedasticity**.
* Linear Regression is very sensitive to **Outliers**. It can terribly affect the regression line and eventually the forecasted values.
* Multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable
* In case of multiple independent variables, we can go with **forward selection**, **backward elimination** and **step wise approach** for selection of most significant independent variables.

## 2. Logistic Regression

Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can represented by following equation.

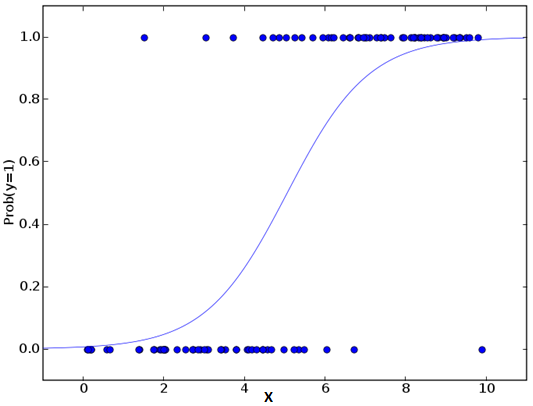
odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence

ln(odds) = ln(p/(1-p))

logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk

Above, p is the probability of presence of the characteristic of interest. A question that you should ask here is “why have we used log in the equation?”.

Since we are working here with a binomial distribution (dependent variable), we need to choose a link function which is best suited for this distribution. And, it is [**logit**](https://en.wikipedia.org/wiki/Logistic_function) function. In the equation above, the parameters are chosen to maximize the likelihood of observing the sample values rather than minimizing the sum of squared errors (like in ordinary regression).

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Logistic_Regression.png)

#### Important Points:

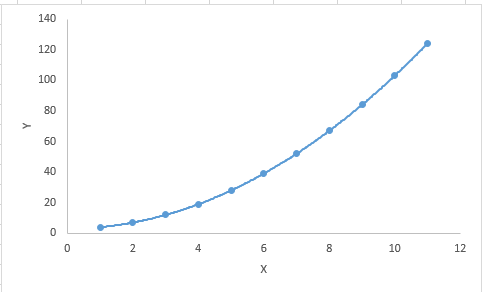
* It is widely used for **classification problems**
* Logistic regression doesn’t require linear relationship between dependent and independent variables.  It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio
* To avoid over fitting and under fitting, we should include all significant variables. A good approach to ensure this practice is to use a step wise method to estimate the logistic regression
* It requires **large sample sizes** because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square
* The independent variables should not be correlated with each other i.e. **no multi collinearity**.  However, we have the options to include interaction effects of categorical variables in the analysis and in the model.
* If the values of dependent variable is ordinal, then it is called as **Ordinal logistic regression**
* If dependent variable is multi class then it is known as **Multinomial Logistic regression**.

## 3. Polynomial Regression

A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation:

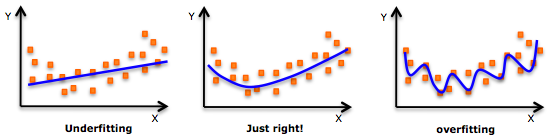
y=a+b\*x^2

In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Polynomial.png)

#### Important Points:

* While there might be a temptation to fit a higher degree polynomial to get lower error, this can result in over-fitting. Always plot the relationships to see the fit and focus on making sure that the curve fits the nature of the problem. Here is an example of how plotting can help:

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/02/underfitting-overfitting.png)

* Especially look out for curve towards the ends and see whether those shapes and trends make sense. Higher polynomials can end up producing wierd results on extrapolation.

## 4. Stepwise Regression

This form of regression is used when we deal with multiple independent variables. In this technique, the selection of independent variables is done with the help of an automatic process, which involves no human intervention.

This feat is achieved by observing statistical values like R-square, t-stats and AIC metric to discern significant variables. Stepwise regression basically fits the regression model by adding/dropping co-variates one at a time based on a specified criterion. Some of the most commonly used Stepwise regression methods are listed below:

* Standard stepwise regression does two things. It adds and removes predictors as needed for each step.
* Forward selection starts with most significant predictor in the model and adds variable for each step.
* Backward elimination starts with all predictors in the model and removes the least significant variable for each step.

The aim of this modeling technique is to maximize the prediction power with minimum number of predictor variables. It is one of the method to handle [higher dimensionality](https://www.analyticsvidhya.com/blog/2015/07/dimension-reduction-methods/) of data set.

## 5. Ridge Regression

Ridge Regression is a technique used when the data suffers from multicollinearity ( independent variables are highly correlated). In multicollinearity, even though the least squares estimates (OLS) are unbiased, their variances are large which deviates the observed value far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

Above, we saw the equation for linear regression. Remember? It can be represented as:

y=a+ b\*x

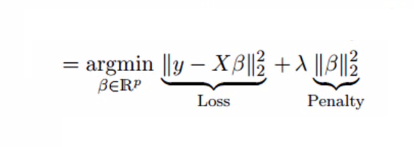
This equation also has an error term. The complete equation becomes:

y=a+b\*x+e (error term),  [error term is the value needed to correct for a prediction error between the observed and predicted value]

=> y=a+y= a+ b1x1+ b2x2+....+e, for multiple independent variables.

In a linear equation, prediction errors can be decomposed into two sub components. First is due to the **biased** and second is due to the **variance**. Prediction error can occur due to any one of these two or both components. Here, we’ll discuss about the error caused due to variance.

Ridge regression solves the multicollinearity problem through [shrinkage parameter](https://en.wikipedia.org/wiki/Shrinkage_estimator) λ (lambda). Look at the equation below.

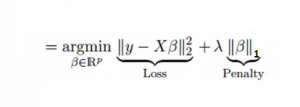
[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Ridge2.png)

In this equation, we have two components. First one is least square term and other one is lambda of the summation of β2 (beta- square) where β is the coefficient. This is added to least square term in order to shrink the parameter to have a very low variance.

#### Important Points:

* The assumptions of this regression is same as least squared regression except normality is not to be assumed
* It shrinks the value of coefficients but doesn’t reaches zero, which suggests no feature selection feature
* This is a regularization method and uses [l2 regularization](https://en.wikipedia.org/wiki/Regularization_(mathematics)).

## 6. Lasso Regression

Similar to Ridge Regression, Lasso (Least Absolute Shrinkage and Selection Operator) also penalizes the absolute size of the regression coefficients. In addition, it is capable of reducing the variability and improving the accuracy of linear regression models.  Look at the equation below: [](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Lasso.png)Lasso regression differs from ridge regression in a way that it uses absolute values in the penalty function, instead of squares. This leads to penalizing (or equivalently constraining the sum of the absolute values of the estimates) values which causes some of the parameter estimates to turn out exactly zero. Larger the penalty applied, further the estimates get shrunk towards absolute zero. This results to variable selection out of given n variables.

#### Important Points:

* The assumptions of this regression is same as least squared regression except normality is not to be assumed
* It shrinks coefficients to zero (exactly zero), which certainly helps in feature selection
* This is a regularization method and uses [l1 regularization](https://en.wikipedia.org/wiki/Regularization_(mathematics))
* If group of predictors are highly correlated, lasso picks only one of them and shrinks the others to zero

## 7. ElasticNet Regression

ElasticNet is hybrid of Lasso and Ridge Regression techniques. It is trained with L1 and L2 prior as regularizer. Elastic-net is useful when there are multiple features which are correlated. Lasso is likely to pick one of these at random, while elastic-net is likely to pick both.

[Elastic_Net](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Elastic_Net.png)

A practical advantage of trading-off between Lasso and Ridge is that, it allows Elastic-Net to inherit some of Ridge’s stability under rotation.

#### Important Points:

* It encourages group effect in case of highly correlated variables
* There are no limitations on the number of selected variables
* It can suffer with double shrinkage

Beyond these 7 most commonly used regression techniques, you can also look at other models like [Bayesian](https://en.wikipedia.org/wiki/Bayesian_linear_regression), [Ecological](https://en.wikipedia.org/wiki/Ecological_regression) and [Robust regression](https://en.wikipedia.org/wiki/Robust_regression).

## How to select the right regression model?

Life is usually simple, when you know only one or two techniques. One of the training institutes I know of tells their students – if the outcome is continuous – apply linear regression. If it is binary – use logistic regression! However, higher the number of options available at our disposal, more difficult it becomes to choose the right one. A similar case happens with regression models.

Within multiple types of regression models, it is important to choose the best suited technique based on type of independent and dependent variables, dimensionality in the data and other essential characteristics of the data. Below are the key factors that you should practice to select the right regression model:

1. Data exploration is an inevitable part of building predictive model. It should be you first step before selecting the right model like identify the relationship and impact of variables
2. To compare the goodness of fit for different models, we can analyse different metrics like statistical significance of parameters, R-square, Adjusted r-square, AIC, BIC and error term. Another one is the [Mallow’s Cp](http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/regression-and-correlation/goodness-of-fit-statistics/what-is-mallows-cp/) criterion. This essentially checks for possible bias in your model, by comparing the model with all possible submodels (or a careful selection of them).
3. Cross-validation is the best way to evaluate models used for prediction. Here you divide your data set into two group (train and validate). A simple mean squared difference between the observed and predicted values give you a measure for the prediction accuracy.
4. If your data set has multiple confounding variables, you should not choose automatic model selection method because you do not want to put these in a model at the same time.
5. It’ll also depend on your objective. It can occur that a less powerful model is easy to implement as compared to a highly statistically significant model.
6. Regression regularization methods(Lasso, Ridge and ElasticNet) works well in case of high dimensionality and multicollinearity among the variables in the data set.

[https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/  
https://machinelearningmastery.com/spot-check-regression-machine-learning-algorithms-python-scikit-learn/  
https://www.kaggle.com/amar09/regression-algorithms-using-scikit-learn](https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/)

https://github.com/mattnedrich/GradientDescentExample

**7 Important Model Evaluation Error Metrics Everyone should know**

[**Tavish Srivastava**](https://www.analyticsvidhya.com/blog/author/tavish1/)**, February 19, 2016**

[](https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+NLP101+2018_T1/about?utm_source=AVTopbanner)

## Introduction

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.

Many ingenuous analyst, don’t even check model accuracy. Once they are finished building a model, they hurriedly map predicted values on unseen data. This is an incorrect approach.

Simply, building a predictive model is not your 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.

In our industry, we consider different kinds of metrics to evaluate our models. The choice of metric completely depends on the type of model and the implementation plan of the model. After you are finished building your model, these 7 metrics will help you in evaluating your model accuracy. Considering the rising popularity of cross – validation, I’ve also mentioned its principles in this article.



## Table of Contents

1. Confusion Matrix
2. Gain and Lift Chart
3. Kolmogorov Smirnov Chart
4. AUC – ROC
5. Gini Coefficient
6. Concordant – Discordant Ratio
7. Root Mean Squared Error
8. Cross Validation (Not a metric though!)

## Warming up: Types of Predictive models

When we talk about predictive models, we are talking either about a regression model (continuous output) or a classification model (nominal or binary output). The evaluation metrics used in each of these models are different.

In classification problems, we use two types of algorithms (dependent on the kind of output it creates):

1. **Class output** : Algorithms like SVM and KNN create a class output. For instance, in a binary classification problem, the outputs will be either 0 or 1. However, today we have algorithms which can convert these class outputs to probability. But these algorithms are not well accepted by the statistics community.
2. **Probability output** : Algorithms like Logistic Regression, Random Forest, Gradient Boosting, Adaboost etc. give probability outputs. Converting probability outputs to class output is just a matter of creating a threshold probability.

In regression problems, we do not have such inconsistencies in output. The output is always continuous in nature and requires no further treatment.

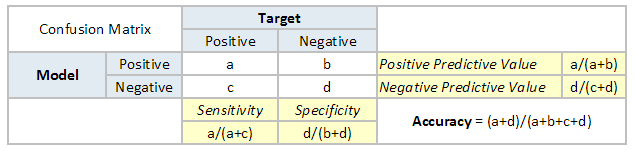
**Illustrative Example**

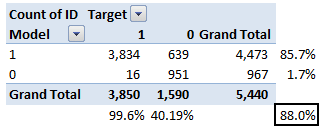
For classification model evaluation metric discussion, I have used my predictions for the problem BCI challenge on Kaggle ([link](https://www.kaggle.com/c/inria-bci-challenge)) . The solution of the problem is irrelevant for the discussion, however the final predictions on the training set has been used for this article. The predictions made for this problem were probability outputs which have been converted to class outputs assuming a threshold of 0.5 .

## 1. 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)

[](https://www.analyticsvidhya.com/blog/wp-content/uploads/2015/01/Confusion_matrix1.png)

The accuracy for the problem in hand comes out to be 88%.  As you can see from the above two tables, the Positive predictive Value is high, but negative predictive value is quite low. Same holds for Senstivity and Specificity. This is primarily driven by the threshold value we have chosen. If we decrease our threshold value, the two pairs of starkly different numbers will come closer.

In general we are concerned with one of the above defined metric. For instance, in a pharmaceutical company, they will be more concerned with minimal wrong positive diagnosis. Hence, they will be more concerned about high Specificity. On the other hand an attrition model will be more concerned with Senstivity.Confusion matrix are generally used only with class output models.

## 2. Gain and Lift charts

Gain and Lift chart are mainly concerned to check the rank ordering of the probabilities. Here are the steps to build a Lift/Gain chart:

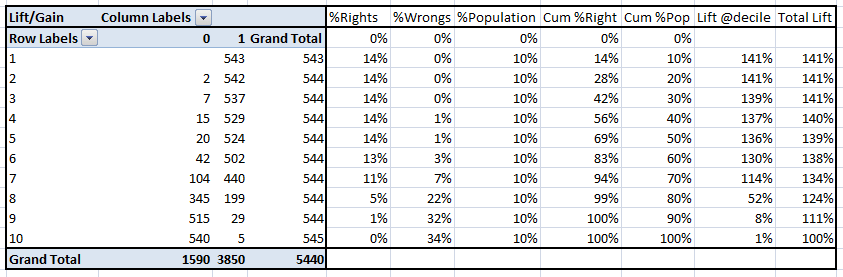
Step 1 : Calculate probability for each observation

Step 2 : Rank these probabilities in decreasing order.

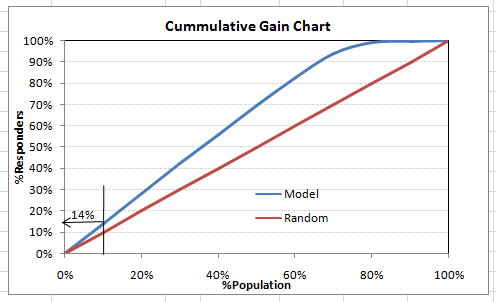
Step 3 : Build deciles with each group having almost 10% of the observations.

Step 4 : Calculate the response rate at each deciles for Good (Responders) ,Bad (Non-responders) and total.

You will get following table from which you need to plot Gain/Lift charts:

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

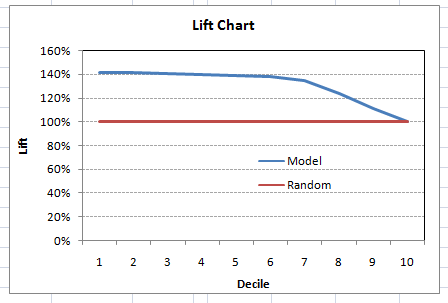
This is a very informative table. Cumulative Gain chart is the graph between Cumulative %Right and Cummulative %Population. For the case in hand here is the graph :

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

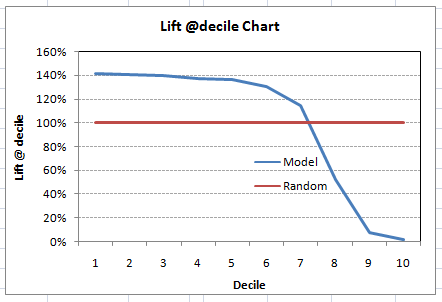
This graph tells you how well is your model segregating responders from non-responders. For example, the first decile however has 10% of the population, has 14% of responders. This means we have a 140% lift at first decile.

What is the maximum lift we could have reached in first decile? From the first table of this article, we know that the total number of responders are 3850. Also the first decile will contains 543 observations. Hence, the maximum lift at first decile could have been 543/3850 ~ 14.1%. Hence, we are quite close to perfection with this model.

Let’s now plot the lift curve. Lift curve is the plot between total lift and %population. Note that for a random model, this always stays flat at 100%. Here is the plot for the case in hand :

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

You can also plot decile wise lift with decile number :

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

What does this graph tell you? It tells you that our model does well till the 7th decile. Post which every decile will be skewed towards non-responders. Any model with lift @ decile above 100% till minimum 3rd decile and maximum 7th decile is a good model. Else you might consider over sampling first.

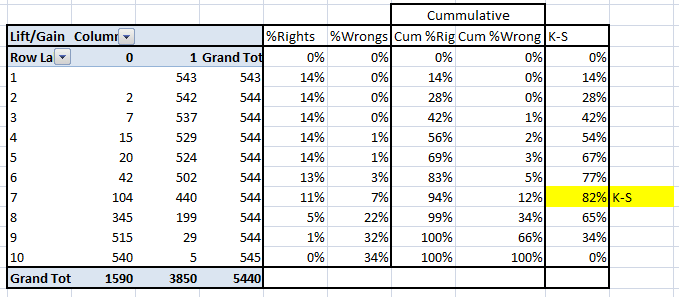
Lift / Gain charts are widely used in campaign targeting problems. This tells us till which decile can we target customers for an specific campaign. Also, it tells you how much response do you expect from the new target base.

## 3. Kolomogorov Smirnov chart

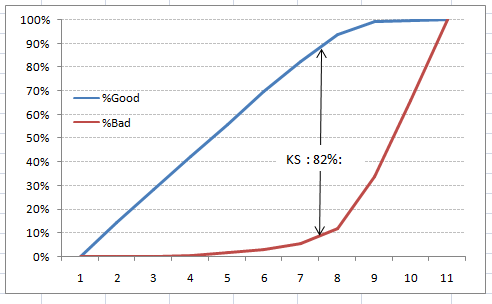
K-S or Kolmogorov-Smirnov chart measures performance of classification models. More accurately, K-S is a measure of the degree of separation between the positive and negative distributions. The K-S is 100, if the scores partition the population into two separate groups in which one group contains all the positives and the other all the negatives.

On the other hand, If the model cannot differentiate between positives and negatives, then it is as if the model selects cases randomly from the population. The K-S would be 0. In most classification models the K-S will fall between 0 and 100, and that the higher the value the better the model is at separating the positive from negative cases.

For the case in hand, following is the table :

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

We can also plot the %Cumulative Good and Bad to see the maximum separation. Following is a sample plot :

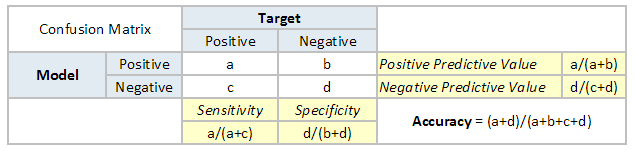
[](https://www.analyticsvidhya.com/blog/wp-content/uploads/2015/01/KS_plot.png)

The metrics covered till here are mostly used in classification problems. Till here, we learnt about confusion matrix, lift and gain chart and kolmogorov-smirnov chart. Let’s proceed and learn few more important metrics.

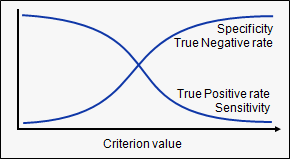
## 4. Area Under the ROC curve (AUC – ROC)

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.

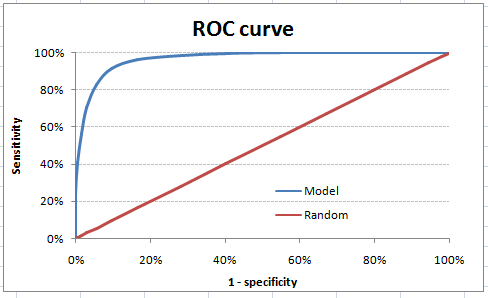
Let’s first try to understand what is ROC (Receiver operating characteristic) curve. If we look at the confusion matrix below, we observe that for a probabilistic model, we get different value for each metric.

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

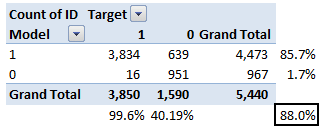
Hence, for each sensitivity, we get a different specificity.The two vary as follows:

[](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 case in hand.

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

Let’s take an example of threshold = 0.5 (refer to confusion matrix). Here is the confusion matrix :

[](https://www.analyticsvidhya.com/blog/wp-content/uploads/2015/01/Confusion_matrix2.png)

As you can see, the sensitivity at this threshold is 99.6% and the (1-specificity) is ~60%. This coordinate becomes on point in our ROC curve. To bring this curve down to a single number, we find the area under this curve (AUC).

Note that the area of entire square is 1\*1 = 1. Hence AUC itself is the ratio under the curve and the total area. For the case in hand, we get AUC ROC as 96.4%. Following are a few thumb rules:

* .90-1 = excellent (A)
* .80-.90 = good (B)
* .70-.80 = fair (C)
* .60-.70 = poor (D)
* .50-.60 = fail (F)

We see that we fall under the excellent band for the current model. But this might simply be over-fitting. In such cases it becomes very important to to in-time and out-of-time validations.

**Points to Remember:**

1. For a model which gives class as output, will be represented as a single point in ROC plot.

2. Such models cannot be compared with each other as the judgement needs to be taken on a single metric and not using multiple metrics. For instance, model with parameters (0.2,0.8) and model with parameter (0.8,0.2) can be coming out of the same model, hence these metrics should not be directly compared.

3. In case of probabilistic model, we were fortunate enough to get a single number which was AUC-ROC. But still, we need to look at the entire curve to make conclusive decisions. It is also possible that one model performs better in some region and other performs better in other.

### **Advantages of using ROC**

Why should you use ROC and not metrics like lift curve?

Lift is dependent on total response rate of the population. Hence, if the response rate of the population changes, the same model will give a different lift chart. A solution to this concern can be true lift chart (finding the ratio of lift and perfect model lift at each decile). But such ratio rarely makes sense for the business.

ROC curve on the other hand is almost independent of the response rate. This is because it has the two axis coming out from columnar calculations of confusion matrix. The numerator and denominator of both x and y axis will change on similar scale in case of response rate shift.

## 5. Gini Coefficient

Gini coefficient is sometimes used in classification problems. Gini coefficient can be straigh away derived from the AUC ROC number. Gini is nothing but ratio between area between the ROC curve and the diagnol line & the area of the above triangle. Following is the formulae used :

Gini = 2\*AUC – 1

Gini above 60% is a good model. For the case in hand we get Gini as 92.7%.

## 6. Concordant – Discordant ratio

This is again one of the most important metric for any classification predictions problem. To understand this let’s assume we have 3 students who have some likelihood to pass this year. Following are our predictions :

A – 0.9

B – 0.5

C – 0.3

Now picture this. if we were to fetch pairs of two from these three student, how many pairs will we have? We will have 3 pairs : AB , BC, CA. Now, after the year ends we saw that A and C passed this year while B failed. No, we choose all the pairs where we will find one responder and other non-responder. How many such pairs do we have?

We have two pairs AB and BC. Now for each of the 2 pairs, the concordant pair is where the probability of responder was higher than non-responder. Whereas discordant pair is where the vice-versa holds true. In case both the probabilities were equal, we say its a tie. Let’s see what happens in our case :

AB  – Concordant

BC – Discordant

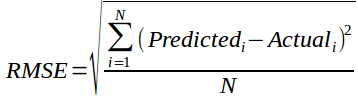
Hence, we have 50% of concordant cases in this example. Concordant ratio of more than 60% is considered to be a good model. This metric generally is not used when deciding how many customer to target etc. It is primarily used to access the model’s predictive power. For decisions like how many to target are again taken by KS / Lift charts.

## 7. Root Mean Squared Error (RMSE)

RMSE is the most popular evaluation metric used in regression problems. It follows an assumption that error are unbiased and follow a normal distribution. Here are the key points to consider on RMSE:

1. The power of ‘square root’  empowers this metric to show large number deviations.
2. The ‘squared’ nature of this metric helps to deliver more robust results which prevents cancelling the positive and negative error values. In other words, this metric aptly displays the plausible magnitude of error term.
3. It avoids the use of absolute error values which is highly undesirable in mathematical calculations.
4. When we have more samples, reconstructing the error distribution using RMSE is considered to be more reliable.
5. RMSE is highly affected by outlier values. Hence, make sure you’ve removed outliers from your data set prior to using this metric.
6. As compared to mean absolute error, RMSE gives higher weightage and punishes large errors.

RMSE metric is given by:



where, N is Total Number of Observations.

Beyond these 7 metrics, there is another method to check the model performance. These 7 methods are statistically prominent in data science. But, with arrival of machine learning, we are now blessed with more robut methods of model selection. **Yes! I’m talking about Cross Validation**.

Though, cross validation isn’t a really a evaluation metric which is used openly to communicate model accuracy. But, the result of cross validation provides good enough intuitive result to generalize the performance of a model.

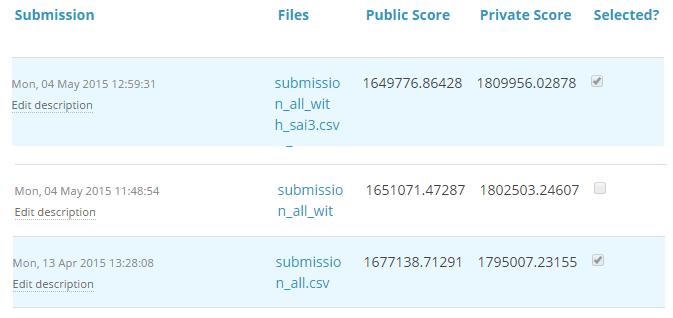
Let’s now understand cross validation in detail.

## 8. Cross Validation

Let’s first understand the importance of cross validation. Due to busy schedules, these days I don’t get much time to participate in data science competitions. Long time back, I participated in TFI Competition on Kaggle. Without delving into my competition performance, I would like to show you the dissimilarity between my public and private leaderboard score.

### Here is an example of scoring on Kaggle!

For TFI competition, following were three of my solution and scores (Lesser the better) :

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/05/kagglescores.png)

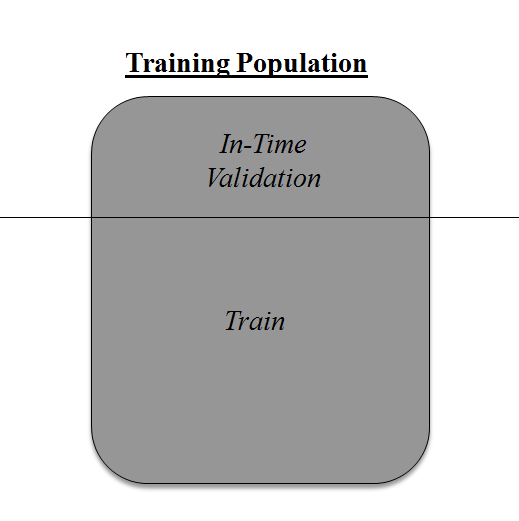
You will notice that the third entry which has the worst Public score turned to be the best model on Private ranking. There were more than 20 models above the “submission\_all.csv”, but I still chose “submission\_all.csv” as my final entry (which really worked out well). What caused this phenomenon ? The dissimilarity in my public and private leaderboard is caused by over-fitting.

Over-fitting is nothing but when you model become highly complex that it starts capturing noise also. This ‘noise’ adds no value to model, but only inaccuracy.

In the following section, I will discuss how you can know if a solution is an over-fit or not before we actually know the test results.

### The concept : Cross Validation

Cross Validation is one of the most important concepts in any type of data modelling. It simply says, try to leave a sample on which you do not train the model and test the model on this sample before finalizing the model.

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/05/validation.png)

Above diagram shows how to validate model with in-time sample. We simply divide the population into 2 samples, and build model on one sample. Rest of the population is used for in-time validation.

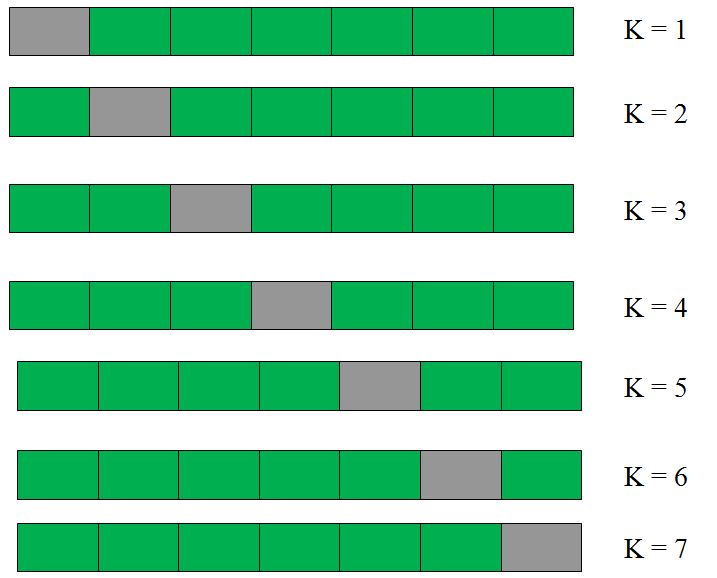
Could there be a negative side of the above approach?

I believe, a negative side of this approach is that we loose a good amount of data from training the model. Hence, the model is very high bias. And this won’t give best estimate for the coefficients. So what’s the next best option?

What if, we make a 50:50 split of training population and the train on first 50 and validate on rest 50. Then, we train on the other 50, test on first 50. This way we train the model on the entire population, however on 50% in one go. This reduces bias because of sample selection to some extent but gives a smaller sample to train the model on. This approach is known as 2-fold cross validation.

### k-fold Cross validation

Let’s extrapolate the last example to k-fold from 2-fold cross validation. Now, we will try to visualize how does a k-fold validation work.

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/05/kfolds.png)

This is a 7-fold cross validation.

Here’s what goes on behind the scene : we divide the entire population into 7 equal samples. Now we train models on 6 samples (Green boxes) and validate on 1 sample (grey box). Then, at the second iteration we train the model with a different sample held as validation. In 7 iterations, we have basically built model on each sample and held each of them as validation. This is a way to reduce the selection bias and reduce the variance in prediction power. Once we have all the 7 models, we take average of the error terms to find which of the models is best.

### How does this help to find best (non over-fit) model?

k-fold cross validation is widely used to check whether a model is an overfit or not. If the performance metrics at each of the k times modelling are close to each other and the mean of metric is highest. In a Kaggle competition, you might rely more on the cross validation score and not on the Kaggle public score. This way you will be sure that the Public score is not just by chance.

### How do we implement k-fold with any model?

Coding k-fold in R and Python are very similar. Here is how you code a k-fold in Python :

from sklearn import cross\_validation model = RandomForestClassifier(n\_estimators=100)

#Simple K-Fold cross validation. 5 folds.

#(Note: in older scikit-learn versions the "n\_folds" argument is named "k".)

cv = cross\_validation.KFold(len(train), n\_folds=5, indices=False)

results = []

# "model" can be replaced by your model object # "Error\_function" can be replaced by the error function of your analysis

for traincv, testcv in cv:

            probas = model.fit(train[traincv], target[traincv]).predict\_proba(train[testcv])

            results.append( Error\_function )

print out the mean of the cross-validated results

print "Results: " + str( np.array(results).mean() )

### But how do we choose k?

This is the tricky part. We have a trade off to choose k.

For a small k, we have a higher selection bias but low variance in the performances.

For a large k, we have a small selection bias but high variance in the performances.

Think of extreme cases :

k = 2  : We have only 2 samples similar to our 50-50 example. Here we build model only on 50% of the population each time. But as the validation is a significant population, the variance of validation performance is minimal.

k = number of observations (n) :  This is also known as “Leave one out”. We have n samples and modelling repeated n number of times leaving only one observation out for cross validation. Hence, the selection bias is minimal but the variance of validation performance is very large.

Generally a value of k = 10 is recommended for most purpose.