Good Practices

**Best Practices/Rules to be followed in Building a Predictive Model Using SAS**

[Venkata Bhaskar Patibandla – XXXX]

# 

# Best Practices in Predictive Modelling using SAS

Venkata Bhaskar Patibandla, [Venkatabhaskar22@gmail.com](mailto:Venkatabhaskar22@gmail.com)

## Introduction

The use of predictive analytics has enabled organizations to more accurately predict their business outcomes, improve business performance, and increase profitability. Model management is not a one-time activity but an essential business process. Models must be well developed and validated to explain that they are working as expected and address the business needs.

Outcomes analysis is also necessary to ensure that the scores derived from applying the model to new data are accurate and to verify that model performance.

This GPC paper tells us SAS-based strategies for effectively managing SAS predictive models in a production environment and tools to perform the same.

## Abbreviations

|  |  |
| --- | --- |
| **Abbreviation** | **Expansion** |
| AI | Artificial Intelligence |
| ANLC | Analytics Life Cycle |
| BI | Business Intelligence |
| CI Matrix | Classification Matrix |
| CV | Coefficient of Variation |
| EDA | Exploratory Data Analysis |
| E-Miner | Enterprise Miner |
| ML | Machine Learning |
| MSE | Mean Square Error |
| PCA | Principal Component analysis or variable reduction |
| P-Value | Probability Values |
| R | R Programming Tool |
| RMSE | Root Mean Square Error |
| SAS | Statistical Analysis on System |
| SDLC | Software Development life Cycle |
| VIF | Variance influence Factor |

## Problem Faced

Knowing an Exact model and accurate model for the business problem is painful thing. In many cases we will do data monitoring to find abnormalities and to make better decision making. The current environment predictive analytics is the Major branch of science where all the organisations will follow. This is looking at data in a way that helps make predictions about what might happen in the future.

Normally People can go through the entire process of preparing the analytical base tables, developing models, and deploying the accurate model in less than two months and whereas some projects might take 10 months or longer to deploy a accurate model with some models never being deployed.

Delays in deploying a model obviously result in lost opportunities and might even result in a model that no longer provides useful predictions due to changes in market and economic conditions. Model deployment for many organizations simply takes too long.

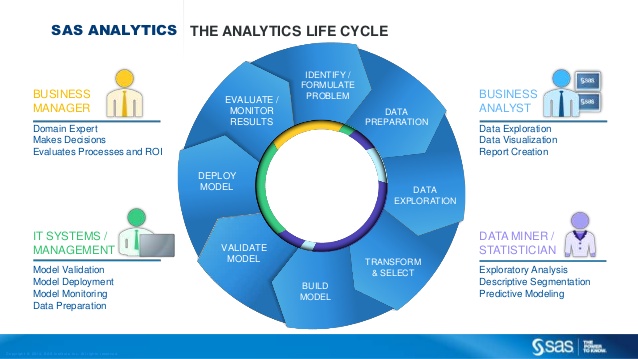
Model deployment is delayed due to technical integration issues. For example, lack a common integrated framework for comparing build models with best model. Models are also often developed by more than one user using a host of different algorithms and with different test data with different methods. There should be sophisticated tool to support all the different algothims in a one central framework to choose the best accuracy of the model, where in which SAS / or any other tools are limited before SAS enterprise Miner. A data mining tool which is very effective and integrated framework. We will discuss this in this paper.

## Solution Approach/ Remedial Action

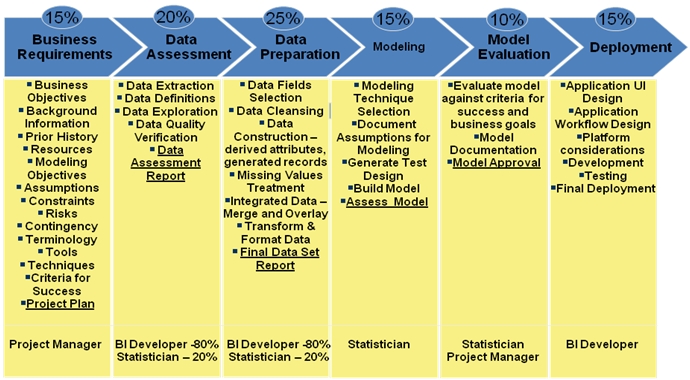
**A Guide or / Process for Building an Effective Model Using SAS.**

**Modelling Development Life Cycle: -**

Like SDLC, in Predictive modelling, Modelling life cycle exists to build better and accurate predictive models. It is advised to follow by developers when we build the model, the life cycle process illustrated as follows.



In addition to this nature of duties of model deployment stated as, Here the role may change depend on the individual organisation requirement and decision involved.



\*\*\* A data scientist/Or a Statistician/ Data modeller will work on the below phases, it may be all or few areas as required.

1. Data Assessment
2. Data Preparation
3. Modelling
4. Model Evaluation /Validation

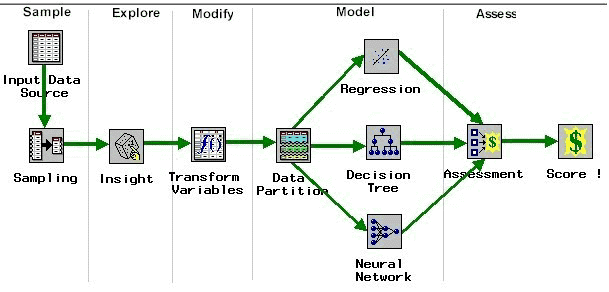
The Main advantage using SAS Enterprise Miner is nothing but the above said all the process is integrated and it will follow a **SEMMA** Approach. (Sample / Explore/ Modify/ Model/ Assess) in which this principle the total model building approach will fall.

This tool is very flexible on data mining which include not only predictive modelling but also integrated machine learning algorithms to compare model accuracy and also a great Visualation tool to project the data insights. Thus we recommend SAS enterprise has great advantage than any other tool.

**Overview of SEMMA:**

Enterprise Miner nodes are arranged into the following categories according the SAS process for data mining: SEMMA.

1. Sample — Identify input data sets (identify input data; define source and target variables, define categorical or ordinal to target variables to identify the model based, sample from a larger data set; partition data set into training, validation, and test data sets).
2. Explore — Explore data sets statistically and graphically (plot the data, obtain descriptive statistics, identify important variables, perform association analysis).
3. Modify — prepare the data for analysis (create additional variables or transform existing variables for analysis, identify outliers, replace missing values, modify the way in which variables are used for the analysis, introduce dummy variables, breakdown of variables using PCA or factor analysis).
4. Model — ﬁt a predictive model (model a target variable by using a regression model, a decision tree, a neural network, or a user-deﬁned model).
5. Assess — compare competing predictive models (build charts that plot the percentage of respondents, percentage of respondents captured, lift, and proﬁt).

**** **SEMMA Approach Process flow:**

## SAMPLE:

1. In Sample phase, the input dataset is taken and will define the response and Explanatory variables. As per the requirement we will define the classification of the response variable (i.e. Categorical / Nominal / ordinal etc.) The original sample is divided into training and validation datasets based on the volume of data. Generally, 50 -50 or 60-40% is suggested. This will enable to reduce the time to fit the model.

\*\* The best practice in sampling method is nothing but to follow the the seed values used to generate the random numbers for the samples. So that the sampling will generated on random and will repeated numbers of times of process for no of samples. This will be taken care automatically in Sampling Node in SAS Enterprise Miner.

1. The next Step in sample method is Sample Partitioning. To partition the sample into training, validation, and test data sets, follow the Sampling node with a Data Partition node. Given percentage of split in sampling node in Data partition node area.
   1. **Training Data set:** Training data set is nothing but the using this data sample we will build the model. Based on the requirement we will fit classification models or other algorithms. The Best practice used for Training data set use as much of data in a population sample, this gives the accurate prediction (as we know the sample Size N is more the sample error will be less).

\*\* The accuracy of the Training dataset is nothing but the Model accuracy obtained when you build the model on Training dataset is close while testing accuracy on the testing data. In that case the model is best fit for the data.

* 1. **Validation or Testing datsets:** Once you're finished training, then you run against your testing set and verify that the accuracy is sufficient

## EXPLORE:

The main important Step in Building a Model is exploring the data (EDA – Exploratory data analysis). In this phase user will understand the Raw data and getting insights and data behaviour of the sample by [analyzing](https://en.wikipedia.org/wiki/Data_analysis) [data sets](https://en.wikipedia.org/wiki/Data_set) to summarize their main characteristics and with visual methods and possibly formulate hypotheses that could lead to new data collection and experiments.

This Hypothesis Testing is very important to lead/assume crucial decisions like Normality required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed

Some of the Best practices/techniques used in EDA are

1. Variance factor.
2. Missing data analysis.
3. Box plot for understanding outliers and their removal
4. Scatter plot for understanding the cluster / or Residuals
5. Multi collinearity between the variables to avoid the influence of one over other.
6. Central Tendency and Mean deviation techniques.

\*\* Multi collinearity and Variable Influence Factor (VIF) are the two major issues in EDA where the independent variables will have highest correlation among each other it cannot able to participate in Model and it will lead poor results of the accuracy. So we need remove if any exists.

1. Use Chi –Square Stats in Explore Node for doing Exploratory variable anlaysis. When Outputs the come Check for missing values, Large standard Deviations, Measure Skewness and normality. Use default method to replace missing values with Mean values.

## MODFIY:

Mainly in our Statistical models our data contains categorical or nominal or ordinal data. Sometimes we see we have so many groups with no definite relation in between them. The solution is to use dummy variables - variables with only two values, zero and one

For example, imagine that you wanted to predict depression from year in school: freshman, sophomore, junior, or senior. Obviously, "year in school" has more than two levels.

What you need to do is to recode "year in school" into a set of dummy variables, each of which has two levels. The first step in this process is to decide the number of dummy variables. it's simply k-1, where k is the number of levels of the original variable.

You could also create dummy variables for all levels in the original variable, and simply drop one from each analysis.

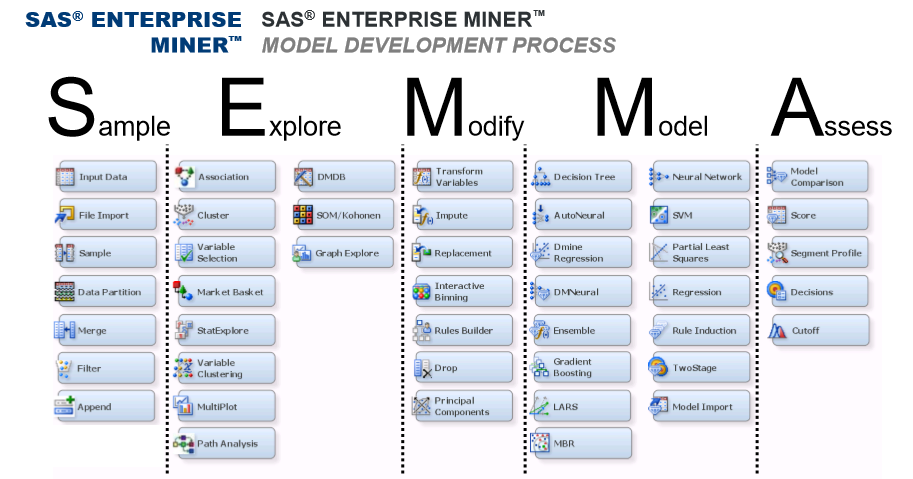
In this instance, we would need to create 4-1=3 dummy variables. In order to create these variables, we are going to take 3 of the levels of "year of school", and create a variable corresponding to each level, which will have the value of yes or no (i.e., 1 or 0). In this instance, we can create a variable called "sophomore," "junior," and "senior." Each instance of "year of school" would then be recoded into a value for "sophomore," "junior," and "senior." If a person were a junior, then "sophomore" would be equal to 0, "junior" would be equal to 1, and "senior" would be equal to 0.

The use of dummy variables will introduce accurate results interpretation in the model. Use that dummy variables in the regression model instead of original independent variables. This is one of the good practice and main step in building a model in the modify STEP.

## MODEL

This is the crucial part of the Phase where in this stage Sample data is free with missing values, outliers and free from Multiple collinearity, and variable reduction completed etc.,

The different stages in which various techniques used in SAS enterprise miner were stated below.

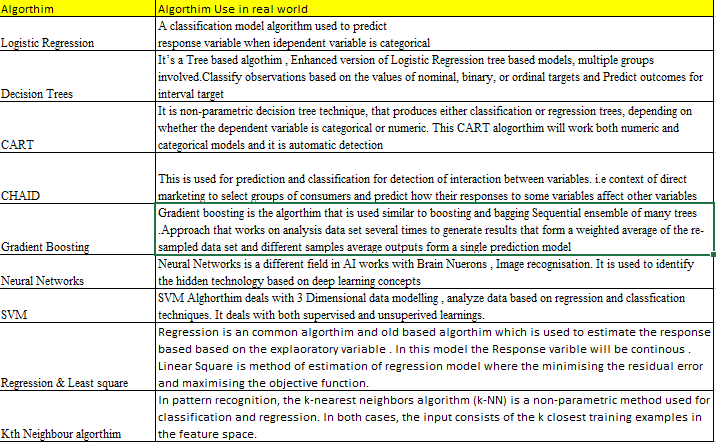


Ideally in modelling we got supervised and unsupervised models. They classified as

1. Classification and Prediction (Supervised) -- Model knows target variable
2. Clustering (unsupervised) -- Model don’t know Target variable

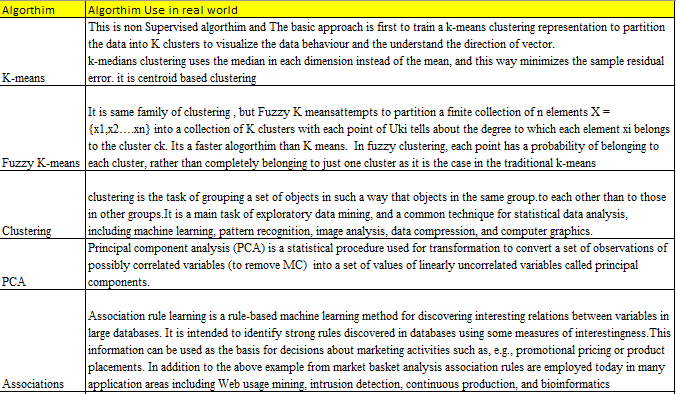
**Supervised Algorithms (Machine learning – Algorithms):**

In a brief, Machine learning techniques or Predictive Model algorithms used in SAS enterprise miner are as follows. These are Important or widely used Supervised algorithms used in current market



**Un-Supervised Algorithms (Machine learning – Algorithms):**

In a brief, Machine learning techniques or Predictive Model algorithms used in SAS enterprise miner are as follows. These are Important or widely used Un -Supervised algorithms used in current market



Once the Appropriate model is created, the model outputs statistics such as Model scores, Model Accuracies parameters such as MSE, P values, ROC curve, uplift and cross lift charts, RMSE, VIF, Classification matrix and other Model related statistics).

## ASSESS:

In SAS enterprise miner, Once the SAS model is built on training dataset with Gini Statistics it will compared with many different models.

\*\* For example: On current Sample if we build a model based on logistic regression, depending on the nature of the data and business requirement the model will be compared with different machine learning algorithms such as random forest, decision tree, naive Bayes or transformation of the response variable etc.

The above mentioned step is a best practice to compare the model accuracy with different models and draw conclusion to fit the right model for data on training dataset. SAS enterprise miner will do this automatically in ASSESS Step. The Assessment Node will take of the comparison of the different model and evaluating the model accuracy. Alternatively, once the model building completed the same model will have validated against testing dataset sample and perform the accuracy with different random SEED sample with the technique called Boosting and Bagging.

SAS enterprise miner performs different visualization and ROC curve and said the best model with minimum MSE, RMSE and lower VIF and higher confusion matrix accuracy etc., based on the business requirement. Thus the all the model statistics with documented and produced

## Benefits

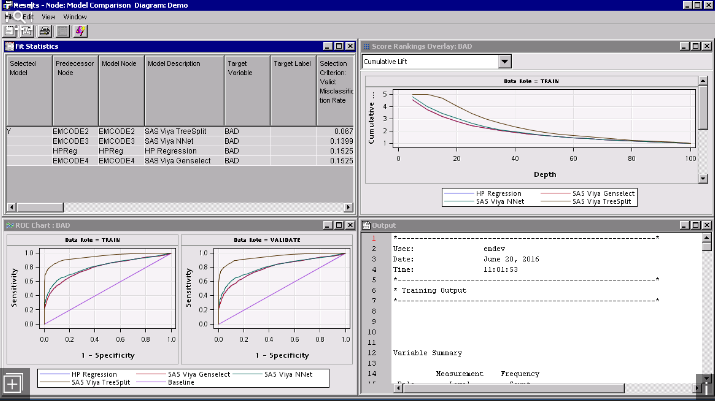
## Quantitative Benefits

In this paper, its talked that best Practices to be followed while building the model building in SAS enterprise miner. this will enable the model to

1. Going from raw data to accurate, business-driven data mining models enabling to collaborate more efficiently.
2. Build more models faster and can automate model deployment and scoring
3. M**odel comparisons, reporting and management**

Examples of model outputs and charts:

Some of the model outputs and evaluation techniques in predictive modelling are shown below.



## Qualitative Benefits

## Learning/ Improvements

This Current GPC paper said above the Best practices or rules/Techniques to be followed explicitly when we building a predictive model using SAS. But in It is recommended that we will see different tools as well like R and Python, WEKA which are also emerging value added tools due to free ware nature. The Same practices can be done and will be extended to that tools in future Paper editions.

## Applicability to Other Projects

The predictive modelling which stated above best practices can be applied in any domain and based on the requirement. But the particular set of few practices is common to all the predictive models Such as outlier treatment, EDA analysis and PCA etc.

This can easily adopt in other Analytical Projects and also with the help of SAS enterprise we can connect/integrate with any legacy database and Pull the extracts to perform the actions. SAS has flexibility to connect to Big data frameworks like Hadoop, MangoDB through SAS Access engine Providers

Comparison of Tools:

* If for beginners entering in analytics industry, it is recommended to learn SAS as your first language. It is easy to learn and holds highest job market share.
* If you are someone who has already spent time in industry, you should try and diversify your expertise be learning a new tool.
* For experts in industry, people should know at least 2 of these. That would add a lot of flexibility for future and open up new opportunities, R / Python is more useful.

The current demand trend in the Market were below as per the research statistics.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Market Research  Survey stats | R | SAS | Python |
| 1 | Availability/Cost | 5 | 3.5 | 5 |
| 2 | Ease of learning | 2.5 | 4.5 | 3.5 |
| 3 | Data Handling Capabilities | 4 | 5 | 5 |
| 4 | Graphical Capabilities | 4.5 | 3.5 | 4.5 |
| 5 | Advancement in tools | 3.5 | 4.5 | 4.5 |
| 6 | Job Scenario | 4 | 4.5 | 4.5 |
| 7 | Customer Service and Community | 4 | 5 | 4 |
| 8 | Deep Learning Support | 4 | 4 | 4 |
|  | Total | 31.5 | 34.5 | 35 |

## References

The GPC paper developed based on the SAS Skill set and data science Technical Expertise and also few references from [www.sas.com](http://www.sas.com)

<https://support.sas.com/resources/papers/proceedings12/337-2012.pdf>

<https://www.kdnuggets.com/2016/11/predictive-science-vs-data-science.html>

[](https://www.google.co.in/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=0ahUKEwjF_f3Y4tHXAhUkT48KHUiCBN0QjRwIBw&url=http://edouard-legoupil.github.io/humanitaRian-data-science/slides/&psig=AOvVaw110Nn552P3UxlQruWpScZN&ust=1511425910574427)

## Appreciations received for the Good Practice

Please refer the XXXX Blogs in emerging Fields Digital and BAS in Data Science Space.