**Business Case:** A Insurance company wants to develop a model that accurately classify the risk associated with the person & then fix the premium according to the risk category.

In the advanced & one click shopping world, apart from the model prediction, number of variable for the model plays a crucial role. A customer wants to get the quote for his insurance in less amount of time & with as much as less information asked by the company.

* **SEMMA Approach**

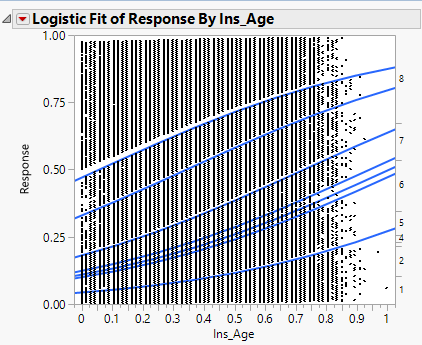
* **Sample**
  + Make validation column and Stratified splitting has to be done. In 40,40,20 ratio (tentative ratio)

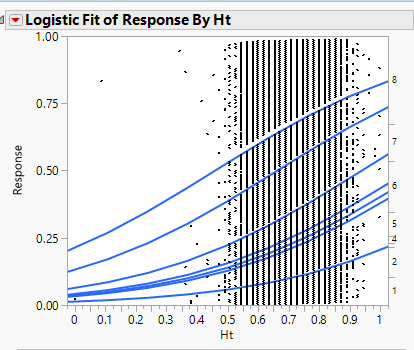
* **Explore - Data Exploring**
  + Operations done in JMP in the train file.
    - Changed the variable type as provided in the instructions
      * Type of dummy variables (Medical\_Keyword) is being considered as categorical since there are only 0s and 1s.
    - Created a new column - calculate sum of all Medical\_Keyword columns

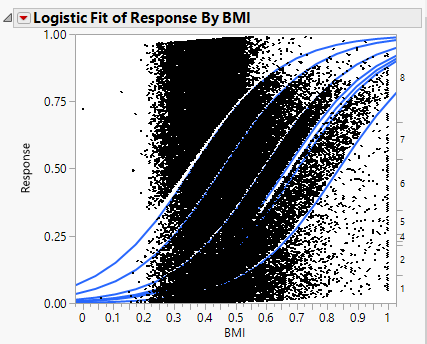
* Objective of this project is to classify the risk from 1-8.
  + 1 being highly subjected to risk and 8 being least subjected to risk
    - Above is the inference based on the distrib plot and graph builder plot

* A look at the response variable shows that there is no specific distribution being followed while estimating the risk.

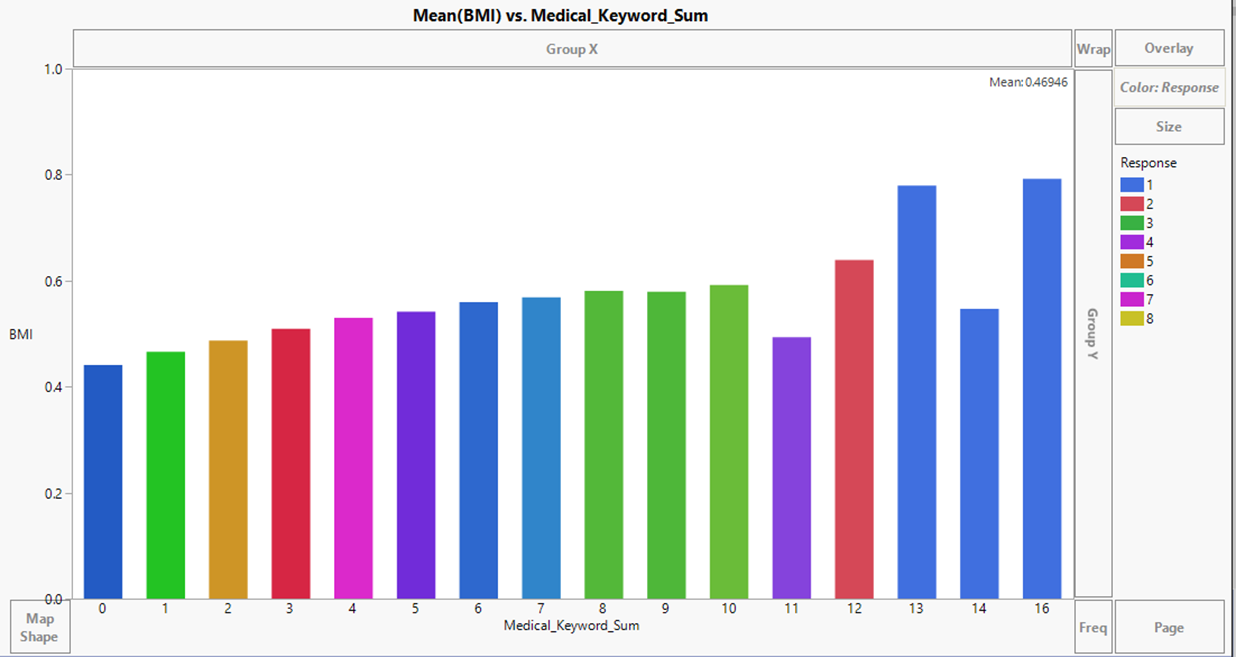
A lot more explorations can be done… just try to analyze the data and create a story… we can discuss our stories.







* + BMI vs Medical illness
  + Generally, the higher the BMI greater is the risk
  + But if BMI is too low & person does not have any medical history is also a high risk as they may be may not be able to resist any virus or germs.



* **Modify - Data Modification** 
  + Missing Values (Refer to 'Missing values' sheet in excel for more analysis)
    - Most of the values are missing for Family\_Hist *(Can some inference be taken?)*
      * *For example, Some people would not like to disclose their personal details*
      * *They may think that these details are irrelevant to their current situation*
      * *They may fear for more premium if they have a significant family health issues*
      * *They may not aware of family history*
    - A number of values (around 90%) are missing for certain Medical History variables and all the missing data is for discrete variables only *(Can these be ignored? Can we infer something from this?)*
      * *For example, many people might not measure their BP, sugar levels for long time and hence might not be filled.*
      * *As the data is 90% missing, there could be some system error as well.*
      * *The variable may be the no of years people have quit smoking/drugs?*
      * *People may have avoided the fields fearing rejection of insurance or high premium rates and it may also be an optional field*
    - Not many values are missing for Employment info *(Can these be imputed?)*
      * *One of the possible reason could be the variable may be salary or working hours which the user might not want to disclose.*
      * *May be certain people would not like to disclose if they are in unemployment period in past X years.*

*(More such inferences would help)*

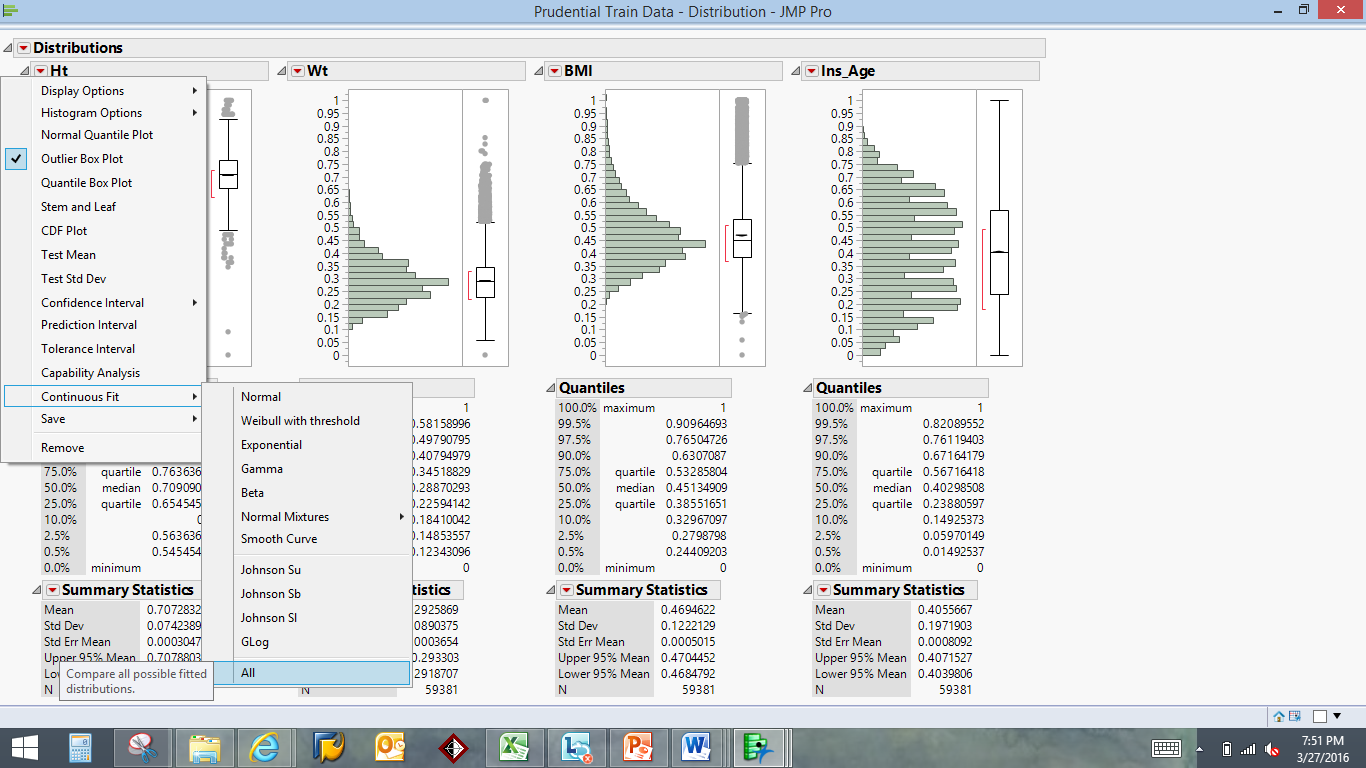
*People avoid mentioning the job details since their work involves serious health risk like working in hazardous conditions, unsanitary environment, exposure to radioactive substances, lead. The question may be related to the employment travel to undeveloped countries where there are more chances of getting a disease. Either people don’t want to disclose the information or they are not sure if their job involves future travel to these countries.*

* Insurance\_Hist5 is missing for 40% of the records *(What should be done on this?)*
  + *Any reasons?*

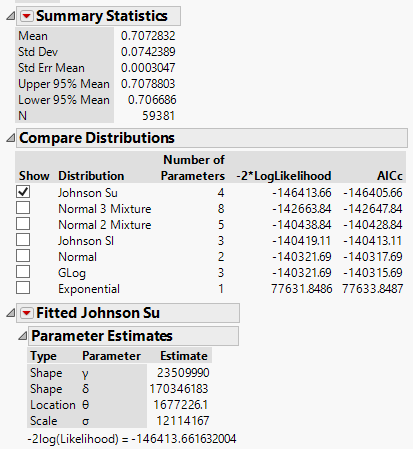
*People may not be willing to disclose the years of no insurance coverage or insurance cancellations due to non-payment*

* Almost all the missing variables are numerical (continuous and discrete) variables … Why is this so??
  + People usually tend to answer more when provided with some options rather than putting some values
  + Medical\_History\_1, Medical\_History\_10, Medical\_History\_15, Medical\_History\_24, Medical\_History\_32 – same as 2nd missing value

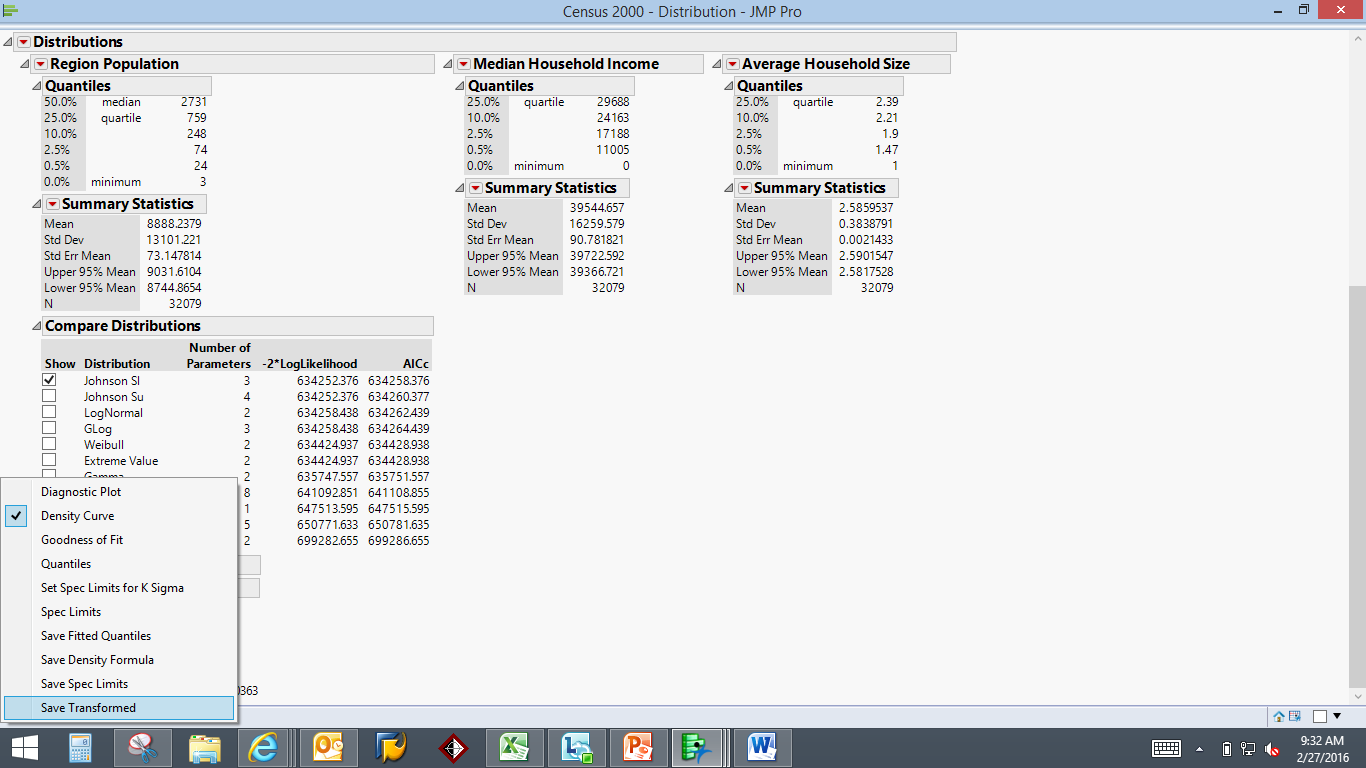
* Outliers
  + Need to identify separately for continuous and categorical variables.
  + Refer to 'Outlier' sheet in excel for detailed analysis
  + Though the number of 0s/1s distribution is highly skeptical for Dummy variables but they cannot be considered as outliers assuming that only a few have diseases.
  + *Another person redoing the outlier analysis would be perfect* **Jessica’s notes and screenshots below:**
  + To reduce the effect of an outlier, go to Analyze 🡪 Distribution, when in distribution, go to “Continuous Fit.” Since we do not know which *type* we want, select “All” and it will decide for you :



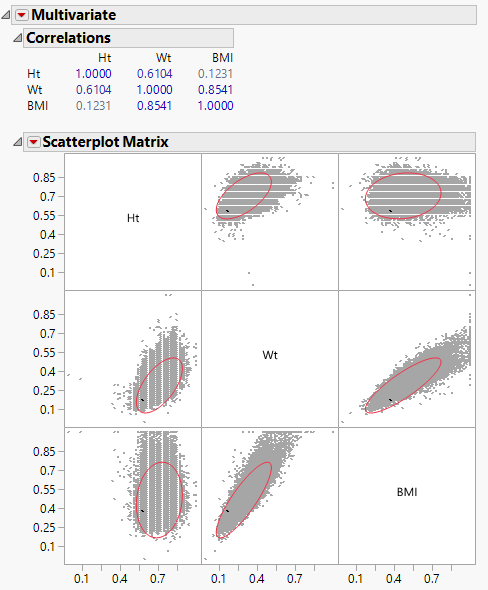
This will create a new column that normalizes the data and will reduce the weight any outlier has on the data:

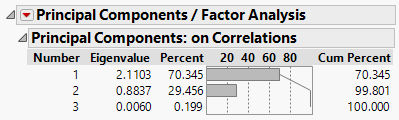


To save this in the dataset, click on the red arrow

🡨 Will create new column in your data set

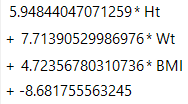
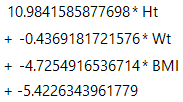
* Run PCA on Ht,Wt,BMI
  + - In JMP, choose Analyze, Multivariate Methods, Multivariate, then red arrow 🡪 Principal Components 🡪 on Correlations





We would choose to extract the first two components because although the first component is the only group >1 Eigenvalue, it only accounts for 70.3% of the total variance. As would be expected, the principle components are highly correlated between BMI/Weight and Height/Weight.

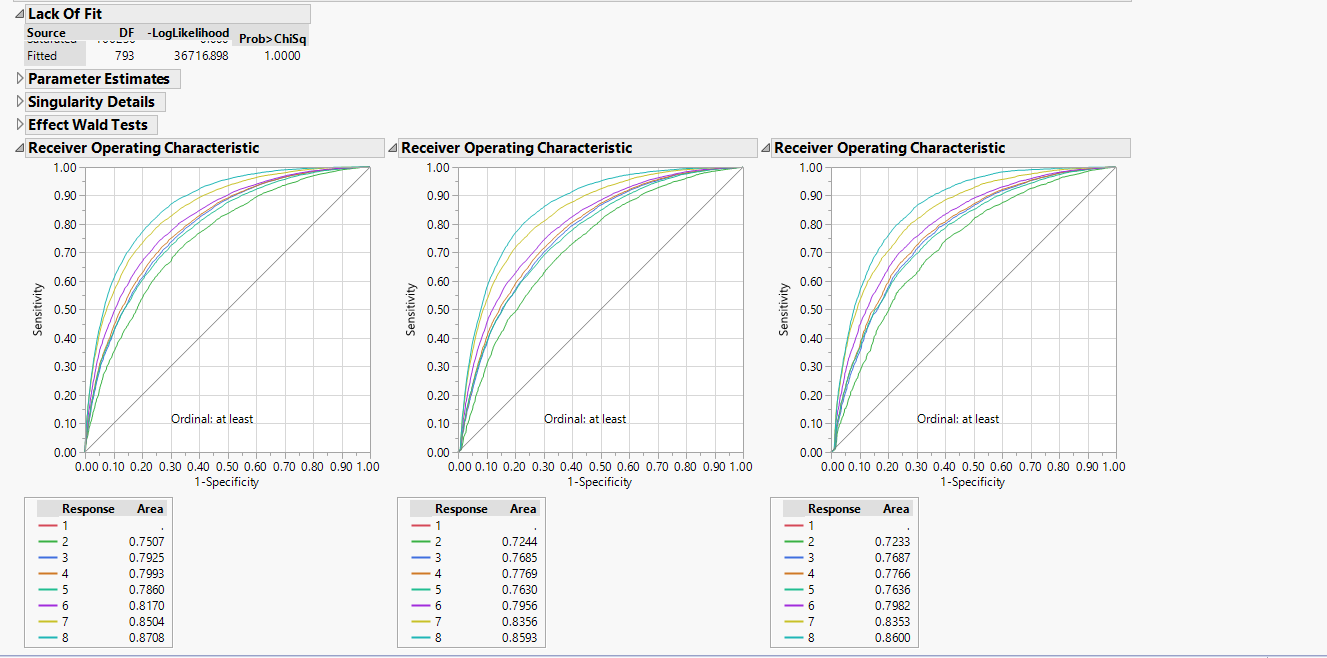
The formulas for the first and second components are as follows:

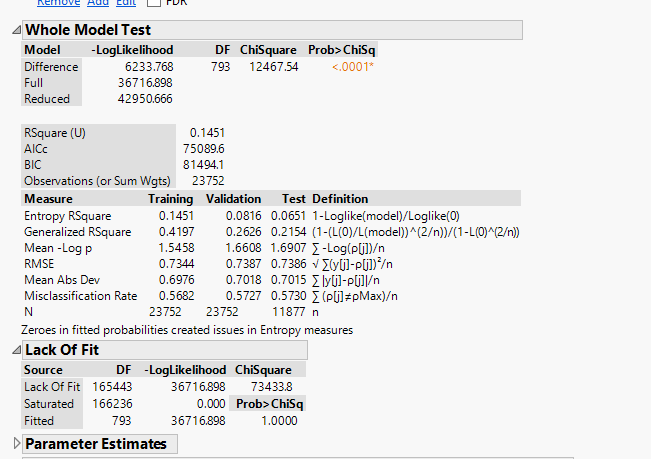
 

* **Model - Which Model has to be chosen**
  + This is a multi-label classification problem.
  + I have checked it is working fine with 'Ordinal Logistic' and 'Decision Tree' Algorithm. *We may have to check which other algorithms could also give better results. Neural Network*
  + *Compare the models built on different parameters such as ROC, AUC, R-Sq, etc.*
  + Find out which is the better model

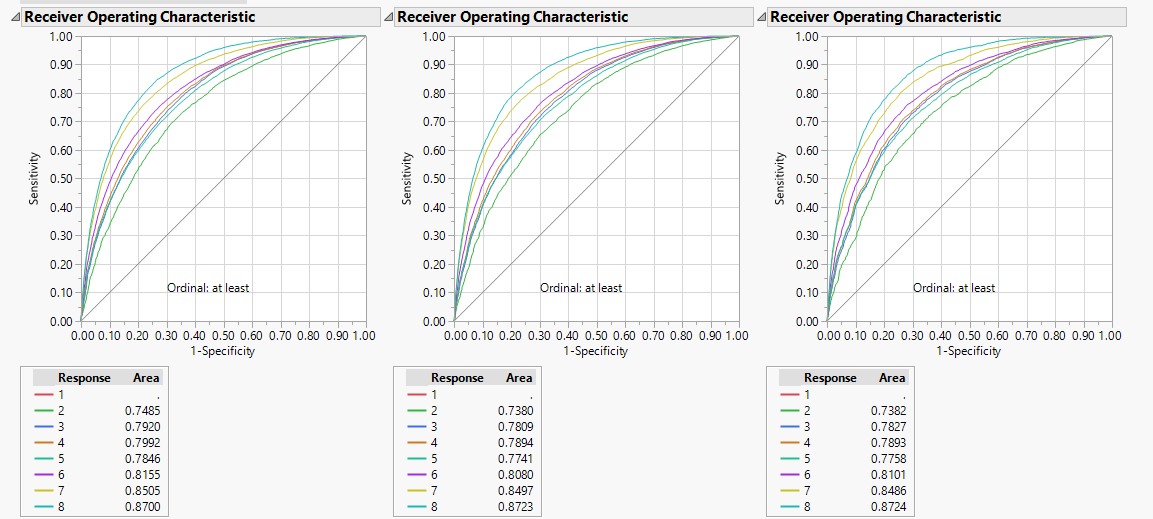
* **Analysis -** 
  + Conclusions that we can draw.
    - *An example could be try providing the user multiple choice mandatory questions rather than value on scale questions.*
      * *This could give us the complete data.*
    - *Further inferences and conclusions.*

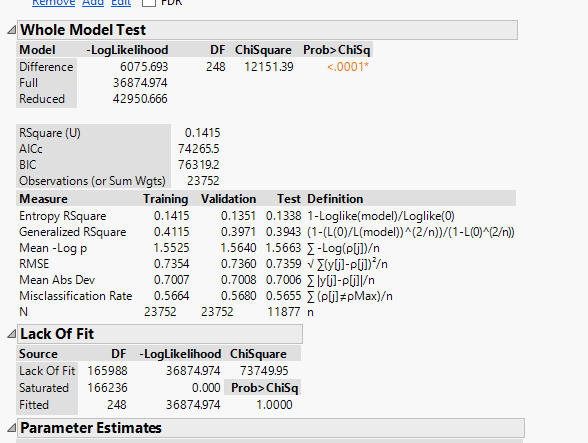
Logistic Regression with Medical SUM



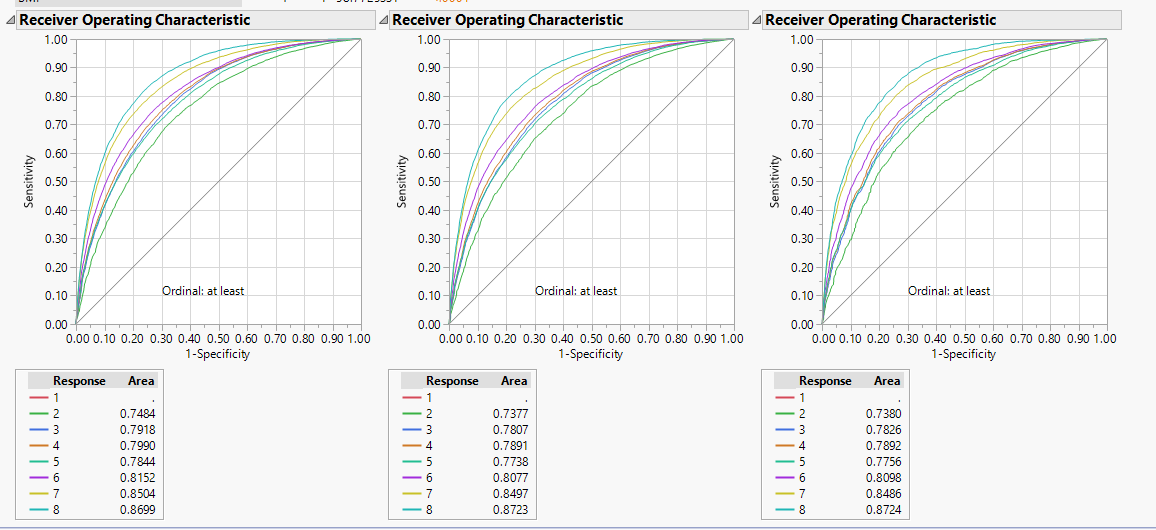


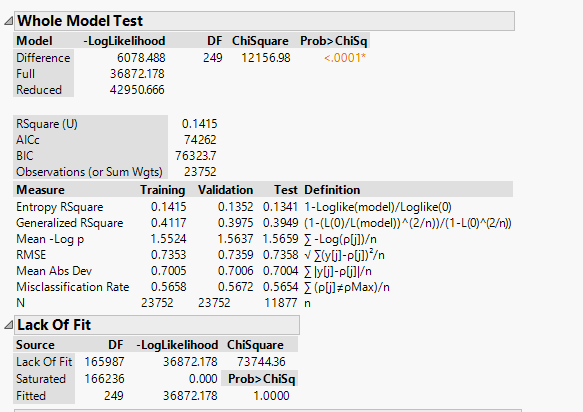
Logistic Regression with Medical Individuals/PCA/history 2 change to continuous:





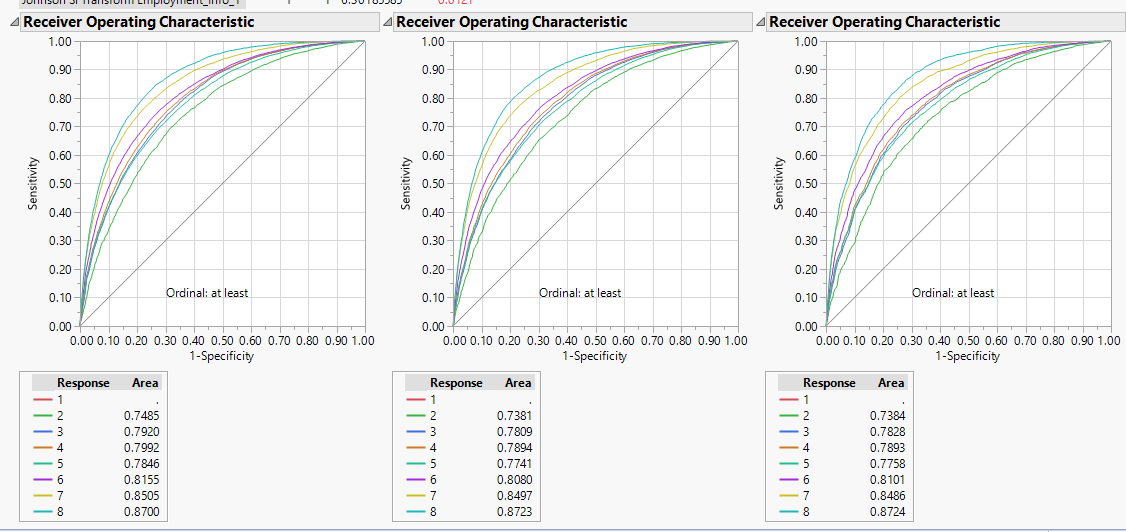
Logistic Regression with Medical Individuals/BMI, HT, WT(without PCA)/history 2 change to continuous:

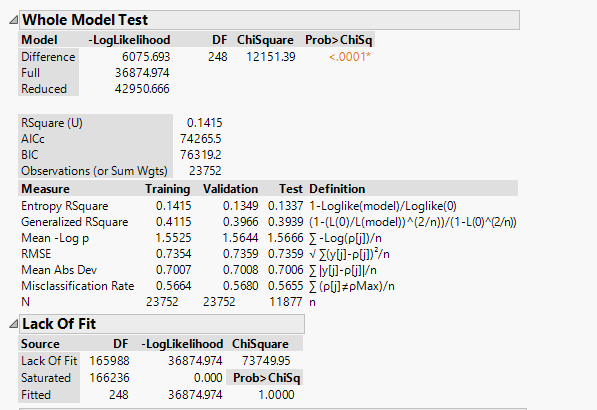




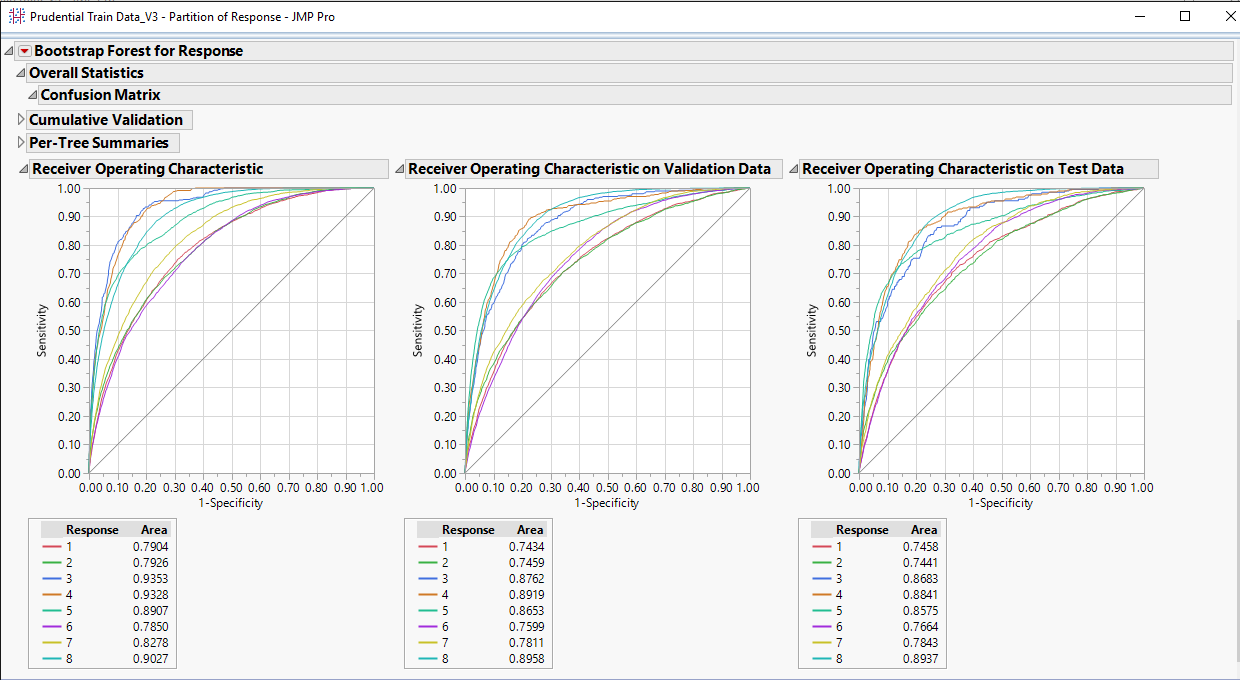
Medical history 2 being defined as categorical is creating the whole issue & variance in r square for the validation & test set.

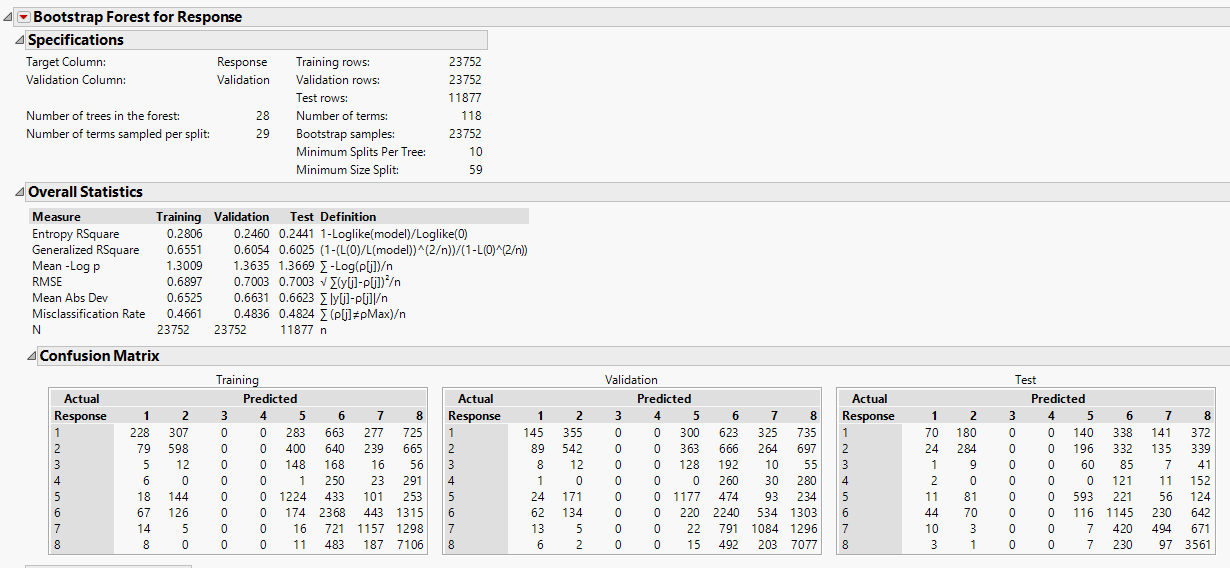
Keywords/princompnents/transformation





Bootstrap Forest: Add column contribution





If you modify the popup comes before bootstrap value, it will give more better results.

Prediction is the probability. ROC, lift, contribution. Compare it with logistic or neural net.

