MIS 6324.
Group 8

ANALYSIS OF SOCCER METRICS FOR PREDICTING PLAYER PERFORMANCE

GROUP 8

- ABHINAV SINGH TILWAR
- ADITYA AROLKAR
- GIRIDHARAN RAJAGOPALAN
- RANJIT NAIR
- YASIN FAZWANI

Table of Contents

EXECUTIVE SUMMARY

A performance indicator is a selection, or combination, of action variables that aims to define some aspects of a performance in a given sport and, these performance indicators, should relate to successful performance or outcome.

Utilizing this definition it seems necessary to define “action variables”, which we have then categorized into physical, technical, tactical and other parameters and will discuss them individually. The sport has developed on such a huge scale that we are in a situation to evaluate players on their tender age to determine how good the basic attributes can impact on turning the player to professional. An estimate of what basic trait a player possesses is going to help the clubs/national teams by leaps and bounds. The investment on youth players is constantly monitored with the updated stats and help them on decisions with return of investment.

Manager and Scout these days are facing a tough time in analyzing player’s performance and take right decisions for better performance of players. Clubs or Managers needs to analyze data from various perspectives, find patterns, and see the "big picture." With a "big picture view" of individual player performance, managers are better able to identify potential problems and take the appropriate prescriptive or preventative actions to ensure each player is performing to the best of their abilities. Thus effective analysis can have a large impact on player’s professional success.

Business Intelligence can assist clubs or Manager to dramatically predict performance of players. Intelligence provides the knowledge on what to base decisions and select appropriate targets for pattern recognition. While the use of intelligence supports in recognizing patterns of players’ professional progress, it also provides clubs or Manager with the ability to effectively manage resources, budget, and meet their responsibility for player’s professional excellence. Such analysis can be enhanced with the help of BI Predictive Analytics Techniques.

Among the various analysis techniques practiced by Clubs or Managers, three main approaches for the identification of player’s successful performance are:

- Trend Analysis - Analyzing player’s preferences and performance which proves to be necessary offerings in order to boost player’s results, retention and satisfaction.
- Tracking Player Performance
- Monitoring performance indicators

By combining qualitative forms of assessment with quantitative analysis derived from Business Intelligence, clubs or Managers can:

- Retain players on basis of their training performance stats and potential.
- Track progression towards defined goals and take action where necessary to ensure desired outcomes are achieved.

PROJECT BACKGROUND

In this project, we are analyzing player's performance on the basis of predictive analysis in Business Intelligence. We are using a second hand data. The existing master dataset is tagged historically from the sites mentioned below:

http://www.paulvi.nl/fifa/

http://www.ultimatedb.nl/

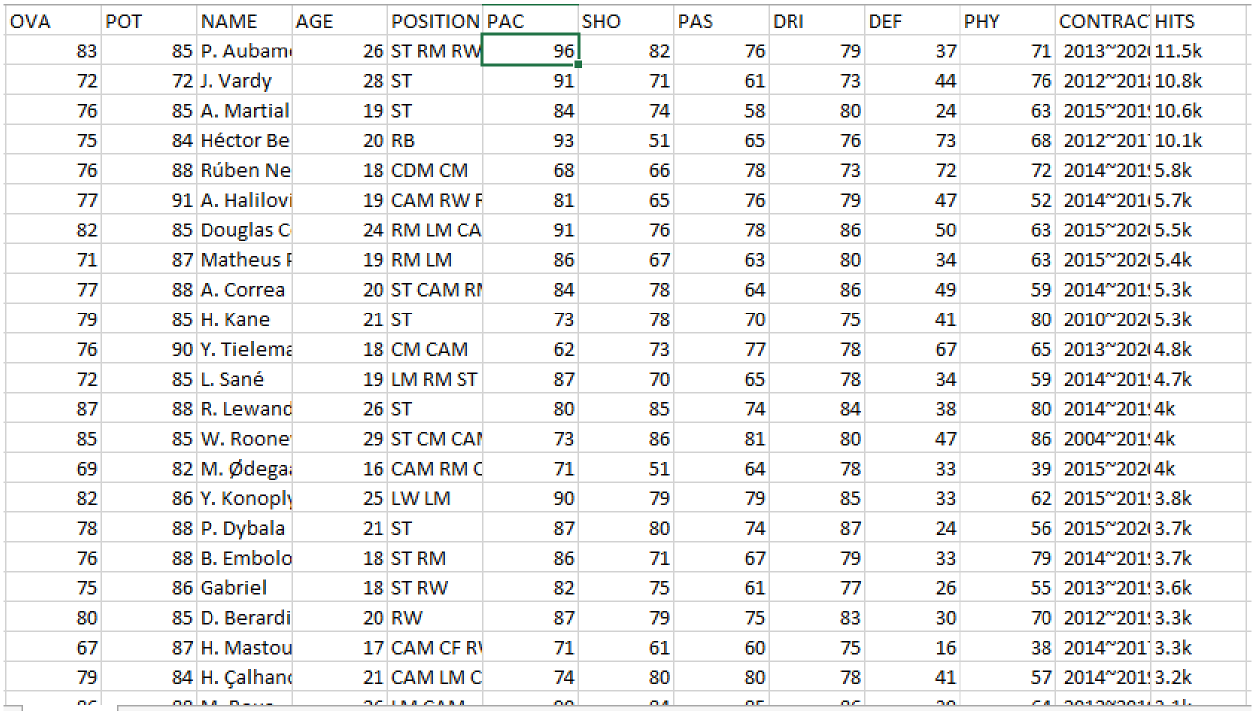
http://sofifa.com/

The main objective of this project is to identify and analyze a model which would help the clubs/manager/scout in predicting the player's performance. We will develop a model that will help us to identify those attributes which are most important and helpful while predicting the overall potential, performance. Thus the result will help the clubs/managers/scouts to address poor performance well ahead in a more efficient and effective way so that better performance and better results can be obtained in time.

The attributes for FIFAPlayerPerformance.xlsx dataset are:

1. ***Name***: This attribute describes the name of the player. It is nominal.[Rejected]
2. ***Age***: This attribute describes the age of the player. It is a numerical value and the datatype is interval.
3. ***Position***: This attribute describes the position in which a particular player plays. It is nominal and contains values like "CF- Center Forward", "CB- Center Backward"
4. ***PAC***: This attribute describes the pace of a player. It is an interval type value. It helps in determining the pace with which player shoots the ball and plays on field.
5. ***SHO***: This attribute describes the shooting rate of a particular player. It is an interval and contains numeric values.
6. ***DRI***: This attribute basically describes the rate with which a particular player dribbles with the ball. It is an interval and contains numeric values.
7. ***PAS***: This attribute describes the passing rate of a particular player. It contains numerical values and hence is an interval.

8. ***DEF***: This attribute describes the defense of a particular player i.e. how well he can defend the ball. It is an interval and contains numerical values.
9. ***PHY***: This attribute describes the overall physique of a particular player while playing in a soccer match. It is an interval and contains quantifiable data.
10. ***POT***: This attribute describes the potential of a particular player to perform in a match. It contains numerical values and hence it is an interval
11. ***OVA***: This attribute describes the Overall Ability of a player to perform in a particular soccer match.[Output Target]
12. ***Hits***: This attribute describes the no of hits that a player has had till now in his career.[Rejected]
13. ***Contract***: This attribute basically describes the contract that a particular player has with a particular club. It is nominal.[Rejected]



OVA	POT	NAME	AGE	POSITION	PAC	SHO	PAS	DRI	DEF	PHY	CONTRAC	HITS
83	85	P. Aubam	26	ST RM RW	96	82	76	79	37	71	2013~202	11.5k
72	72	J. Vardy	28	ST	91	71	61	73	44	76	2012~201	10.8k
76	85	A. Martial	19	ST	84	74	58	80	24	63	2015~201	10.6k
75	84	Héctor Be	20	RB	93	51	65	76	73	68	2012~201	10.1k
76	88	Rúben Ne	18	CDM CM	68	66	78	73	72	72	2014~201	5.8k
77	91	A. Halilovi	19	CAM RW F	81	65	76	79	47	52	2014~201	5.7k
82	85	Douglas C	24	RM LM CA	91	76	78	86	50	63	2015~202	5.5k
71	87	Matheus F	19	RM LM	86	67	63	80	34	63	2015~202	5.4k
77	88	A. Correa	20	ST CAM RI	84	78	64	86	49	59	2014~201	5.3k
79	85	H. Kane	21	ST	73	78	70	75	41	80	2010~202	5.3k
76	90	Y. Tielema	18	CM CAM	62	73	77	78	67	65	2013~202	4.8k
72	85	L. Sané	19	LM RM ST	87	70	65	78	34	59	2014~201	4.7k
87	88	R. Lewand	26	ST	80	85	74	84	38	80	2014~201	4k
85	85	W. Roone	29	ST CM CAI	73	86	81	80	47	86	2004~201	4k
69	82	M. Ødega	16	CAM RM C	71	51	64	78	33	39	2015~202	4k
82	86	Y. Konoply	25	LW LM	90	79	79	85	33	62	2015~201	3.8k
78	88	P. Dybala	21	ST	87	80	74	87	24	56	2015~202	3.7k
76	88	B. Embolo	18	ST RM	86	71	67	79	33	79	2014~201	3.7k
75	86	Gabriel	18	ST RW	82	75	61	77	26	55	2013~201	3.6k
80	85	D. Berardi	20	RW	87	79	75	83	30	70	2012~201	3.3k
67	87	H. Mastou	17	CAM CF R	71	61	60	75	16	38	2014~201	3.3k
79	84	H. Çalhan	21	CAM LM C	74	80	80	78	41	57	2014~201	3.2k

ANALYSIS APPROACH

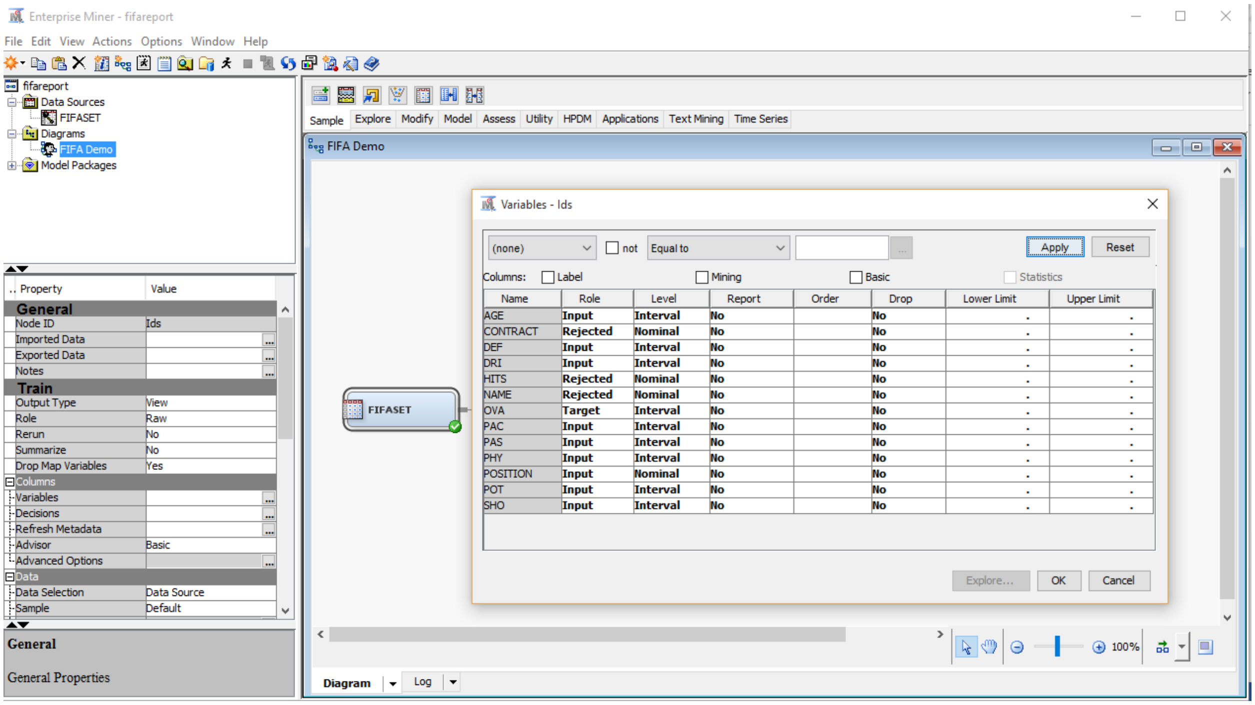
The Player's performance can be predicted using various Predictive Analytics techniques namely Classification, Decision Tree, Linear Regression Model, Logistic Regression, Neural Networks.

For our project we have used Decision Tree, Linear Regression, and Neural Networks methodologies. Later on we did a model comparison on these methodologies to see which model would be better for performing predictive analytics. We have used SAS Enterprise Miner 9.4 tool.

The results for Predictive Analysis of Player's Performance Data are interpreted into the following steps:

Importing Data.

Step 1: The excel file is imported as a data source into SAS Enterprise Miner.


Enterprise Miner - fifareport
File Edit View Actions Options Window Help
fifareport
Data Sources
FIFASET
Diagrams
FIFA Demo
Model Packages
Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series
FIFA Demo
FIFASET
Variables - Ids
Apply
Reset
Explore...
OK
Cancel
Diagram
Log
100%

Property	Value
General	
Node ID	Ids
Imported Data	...
Exported Data	...
Notes	...
Train	
Output Type	View
Role	Raw
Rerun	No
Summarize	No
Drop Map Variables	Yes
Columns	
Variables	...
Decisions	...
Refresh Metadata	...
Advisor	Basic
Advanced Options	...
Data	
Data Selection	Data Source
Sample	Default

General

General Properties

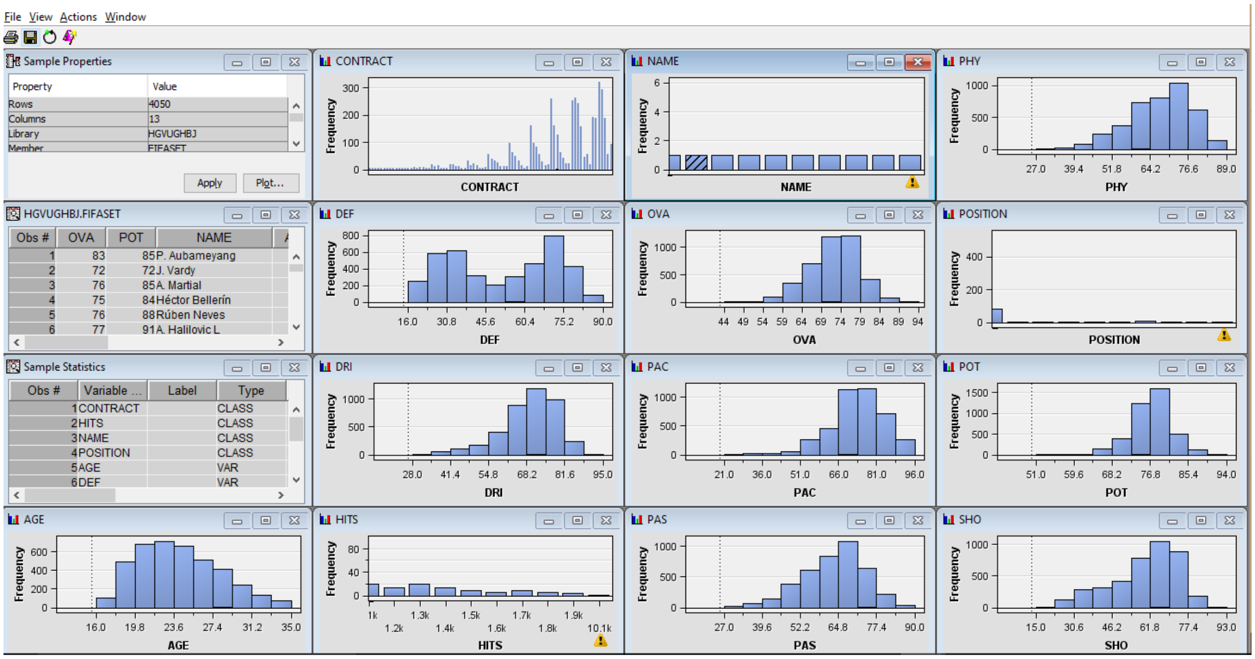
(none) □ not Equal to

Columns: □ Label □ Mining □ Basic □ Statistics

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
AGE	Input	Interval	No		No	.	.
CONTRACT	Rejected	Nominal	No		No	.	.
DEF	Input	Interval	No		No	.	.
DRI	Input	Interval	No		No	.	.
HITS	Rejected	Nominal	No		No	.	.
NAME	Rejected	Nominal	No		No	.	.
OVA	Target	Interval	No		No	.	.
PAC	Input	Interval	No		No	.	.
PAS	Input	Interval	No		No	.	.
PHY	Input	Interval	No		No	.	.
POSITION	Input	Nominal	No		No	.	.
POT	Input	Interval	No		No	.	.
SHO	Input	Interval	No		No	.	.

Checking Missing Values

Step 2: As our dataset didn't contain any null values so there is no need to add the Data Replacement Node.



Sample Properties

Property	Value
Rows	4050
Columns	13
Library	HGVUGHBJ
Member	FIFASET

HGVUGHBJ.FIFASET

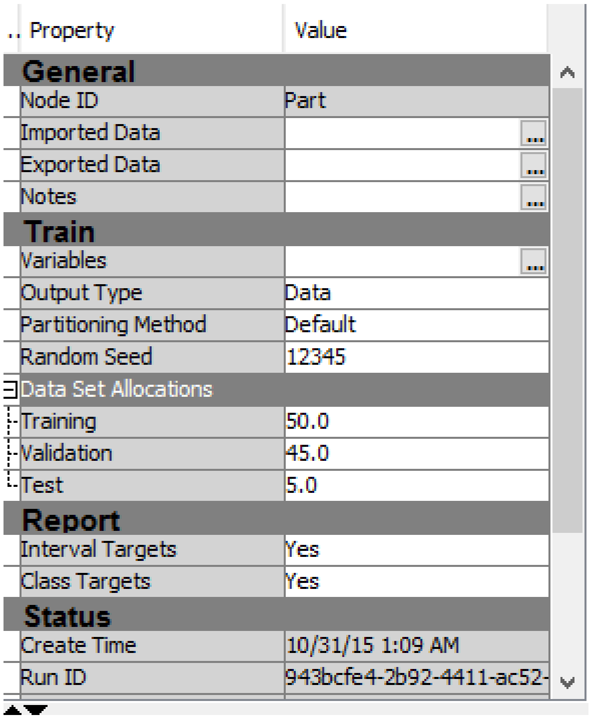
Obs #	OVA	POT	NAME
1	83	85	P. Aubameyang
2	72	72	J. Vardy
3	76	85	A. Martial
4	75	84	Héctor Bellerín
5	76	88	Rúben Neves
6	77	91	A. Halilovic L

Sample Statistics

Obs #	Variable ...	Label	Type
1	CONTRACT		CLASS
2	HITS		CLASS
3	NAME		CLASS
4	POSITION		CLASS
5	AGE		VAR
6	DEF		VAR

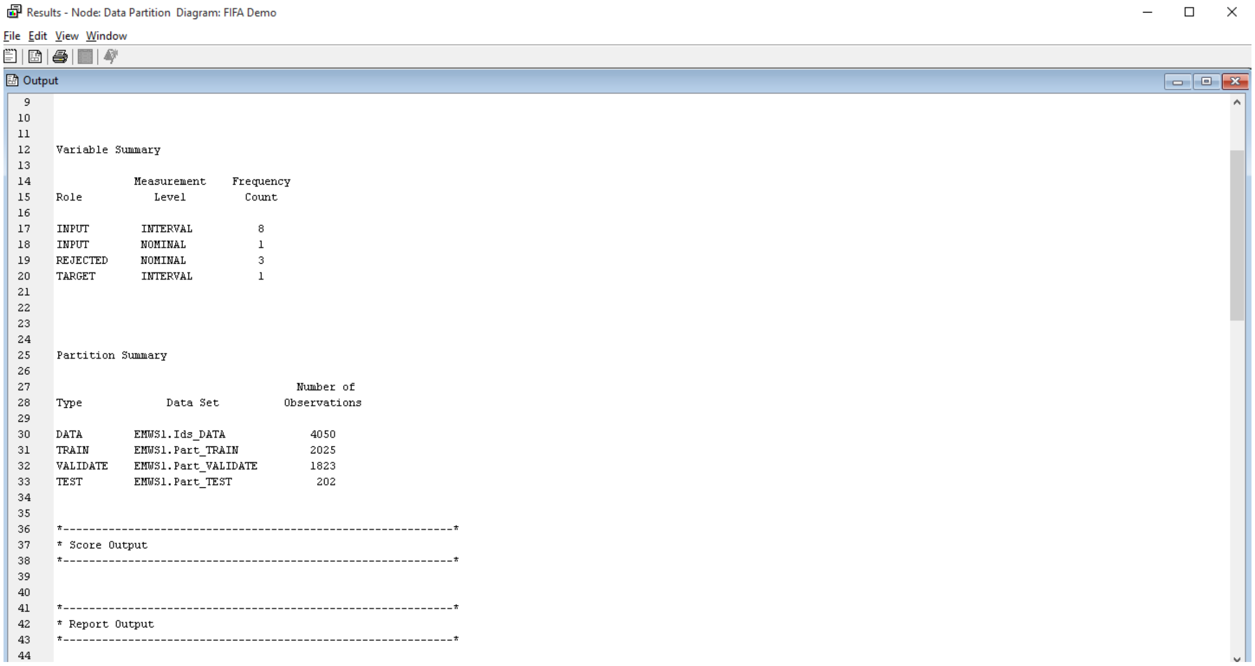
Data Partition

Step 3: Next Step is to add a **Data Partition** Node. For our project we have selected 50 for Training data, 45 for Validation Data and 5 for test data.



Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	50.0
Validation	45.0
Test	5.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	10/31/15 1:09 AM
Run ID	943bcfe4-2b92-4411-ac52-

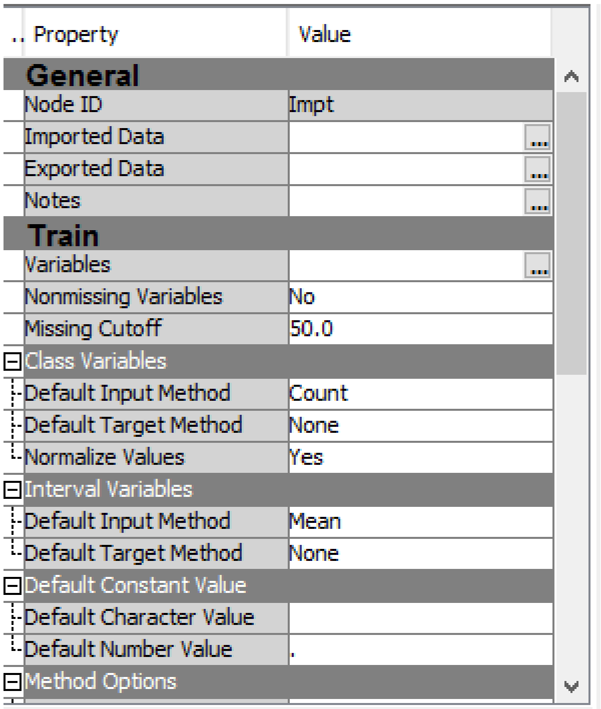
The below screenshot shows the result that we obtain after running the Data Partition Node. This result basically describes the distribution of target variable in training and validation dataset.



```



Variable Summary

            Measurement    Frequency
Role           Level         Count

INPUT        INTERVAL          8
INPUT        NOMINAL           1
REJECTED     NOMINAL           3
TARGET       INTERVAL          1




Partition Summary

                                     Number of
Type             Data Set          Observations

DATA        EMWS1.Ids_DATA            4050
TRAIN       EMWS1.Part_TRAIN          2025
VALIDATE    EMWS1.Part_VALIDATE       1823
TEST        EMWS1.Part_TEST            202


*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*


*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*

```

Adding Impute Node

Step 4: Now we perform the linear regression analysis. But before that, we add an impute node to our diagram so that if there are any missing values present they get replaced.



.. Property	Value
General	
Node ID	Impt
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Nonmissing Variables	No
Missing Cutoff	50.0
⊟ Class Variables	
Default Input Method	Count
Default Target Method	None
Normalize Values	Yes
⊟ Interval Variables	
Default Input Method	Mean
Default Target Method	None
⊟ Default Constant Value	
Default Character Value	
Default Number Value	.
⊟ Method Options	

As we can see in the above screenshot in the Properties Panel of Impute Node, in the class variables section we have given default input method as “Count” so that if any missing values are found in the categorical inputs they get replaced by the most frequent category. Similarly if any missing values are found in the interval variables they get replaced by the mean of non-missing values.

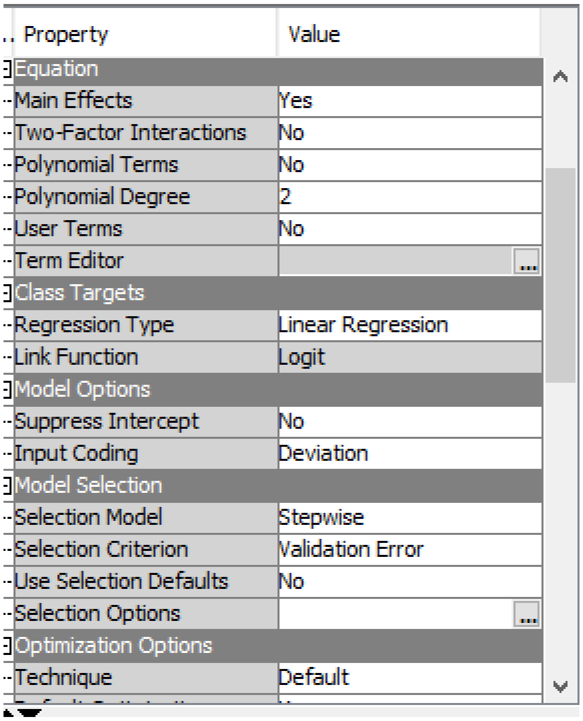
Linear Regression Analysis

Step 5: In this step we perform **Linear Regression Analysis** by adding a Regression Node to our diagram.

Regression Analysis: Regression Analysis is a statistical process for estimating the relationships among variables. It estimates the conditional expectation of the dependent variable given the independent variables.

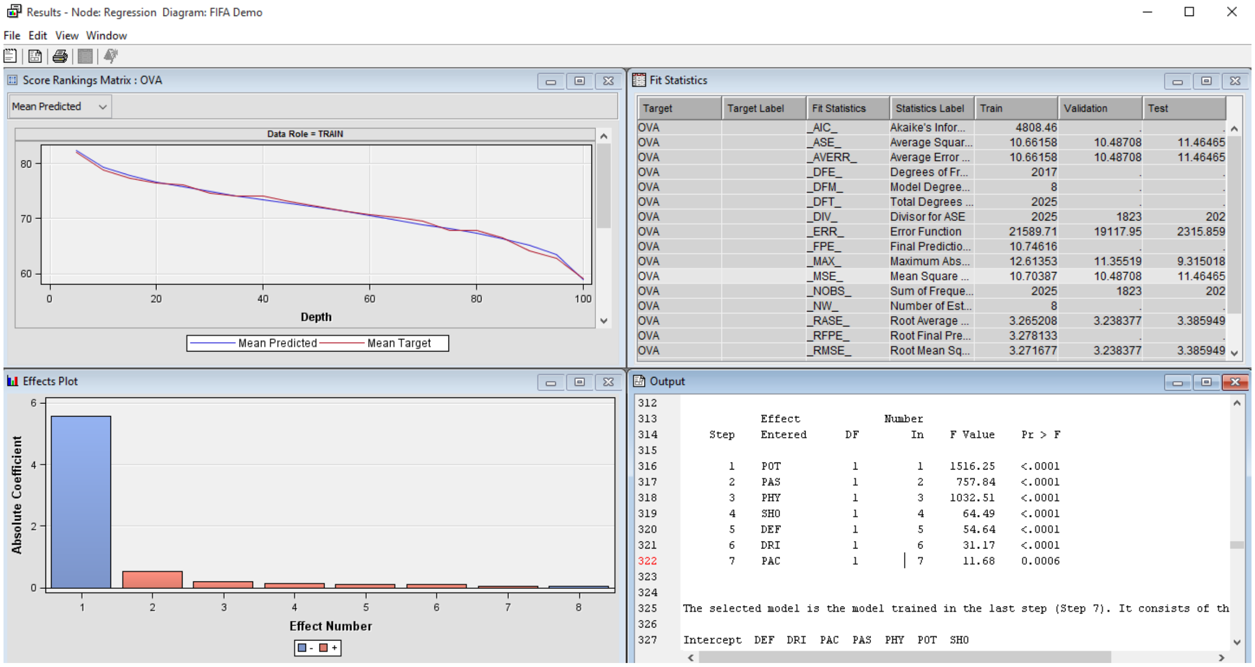
For our project the dependent variable was **OVA** attribute (target) i.e. Overall Ability of a player and other attributes turn out to be independent variables. Basically after performing regression analysis we come to know that which attributes have impact on my target variable i.e. OVA.

We have performed Linear Regression as our target variable is an interval variable. We have used Stepwise Selection Model for performing Linear Regression.



Property	Value
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Linear Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Error
Use Selection Defaults	No
Selection Options	...
Optimization Options	
Technique	Default

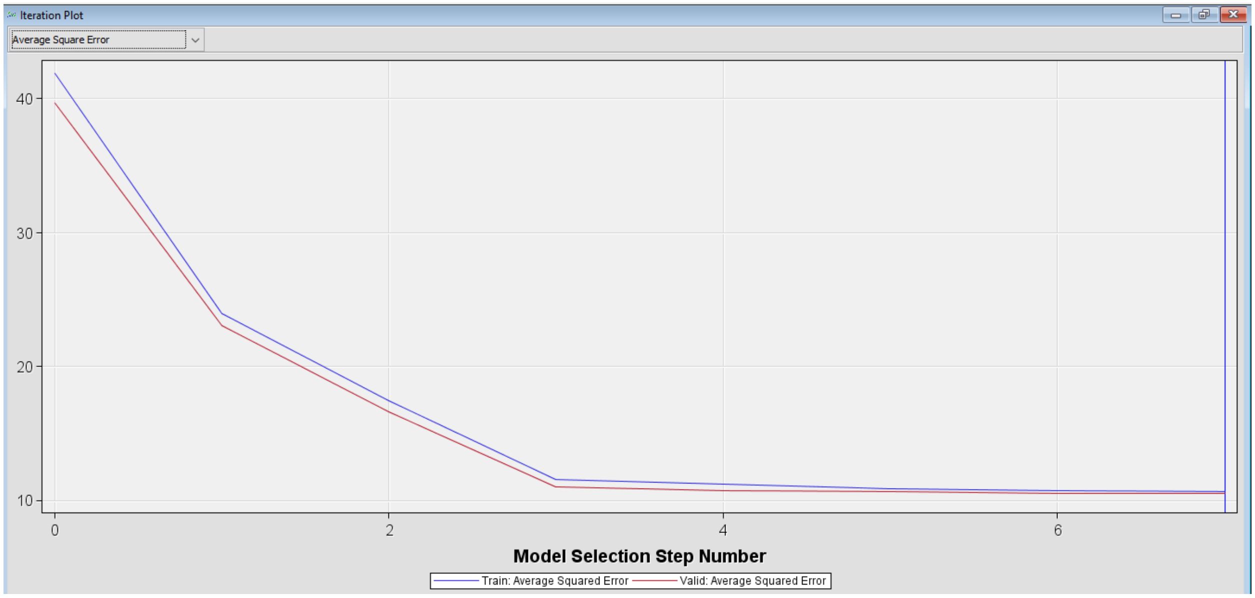
By using Stepwise Selection Model we are able to restrict our inputs as this model considers only those input variables fit for the model which have got a low P-value.



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
OVA		_AIC_	Akaike's Infor...	4808.46	.	.
OVA		_ASE_	Average Squar...	10.66158	10.48708	11.46465
OVA		_AVERR_	Average Error ...	10.66158	10.48708	11.46465
OVA		_DFE_	Degrees of Fr...	2017	.	.
OVA		_DFM_	Model Degree...	8	.	.
OVA		_DFT_	Total Degrees ...	2025	.	.
OVA		_DIV_	Divisor for ASE	2025	1823	202
OVA		_ERR_	Error Function	21589.71	19117.95	2315.859
OVA		_FPE_	Final Predictio...	10.74616	.	.
OVA		_MAX_	Maximum Abs...	12.61353	11.35519	9.315018
OVA		_MSE_	Mean Square ...	10.70387	10.48708	11.46465
OVA		_NOBS_	Sum of Freque...	2025	1823	202
OVA		_NW_	Number of Est...	8	.	.
OVA		_RASE_	Root Average ...	3.265208	3.238377	3.385949
OVA		_RFPE_	Root Final Pre...	3.278133	.	.
OVA		_RMSE_	Root Mean Sq...	3.271677	3.238377	3.385949

```

              Effect                Number
     Step     Entered       DF          In     F Value     Pr > F

        1     POT            1           1     1516.25     <.0001
        2     PAS            1           2      757.84     <.0001
        3     PHY            1           3     1032.51     <.0001
        4     SHO            1           4       64.49     <.0001
        5     DEF            1           5       54.64     <.0001
        6     DRI            1           6       31.17     <.0001
        7     PAC            1           7       11.68     0.0006


The selected model is the model trained in the last step (Step 7). It consists of th

Intercept  DEF  DRI  PAC  PAS  PHY  POT  SHO
```

Looking at the above screenshot we come to know that target variable OVA is dependent on POT, PAC, PAS, PHY, DRI, DEF and SHO attribute of my dataset. Thus the decisive parameter to look for all the players can be read as potential to grow as per this model. As the rest of the values will make up for the position he plays in the game of football.

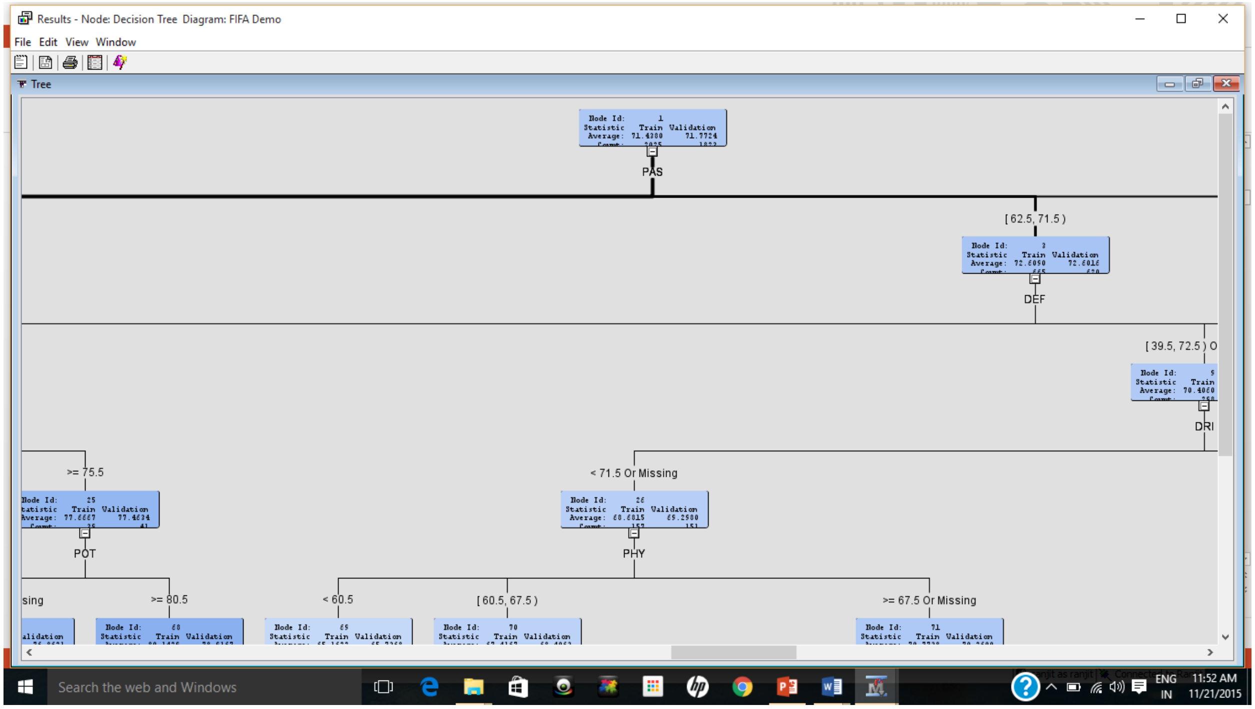


Looking at above screenshot of iteration plot we come to know that the model obtained in step 6 is a good model as it has got the lowest average square error.

Decision Tree Model

Step 6: **Decision Tree Model**.

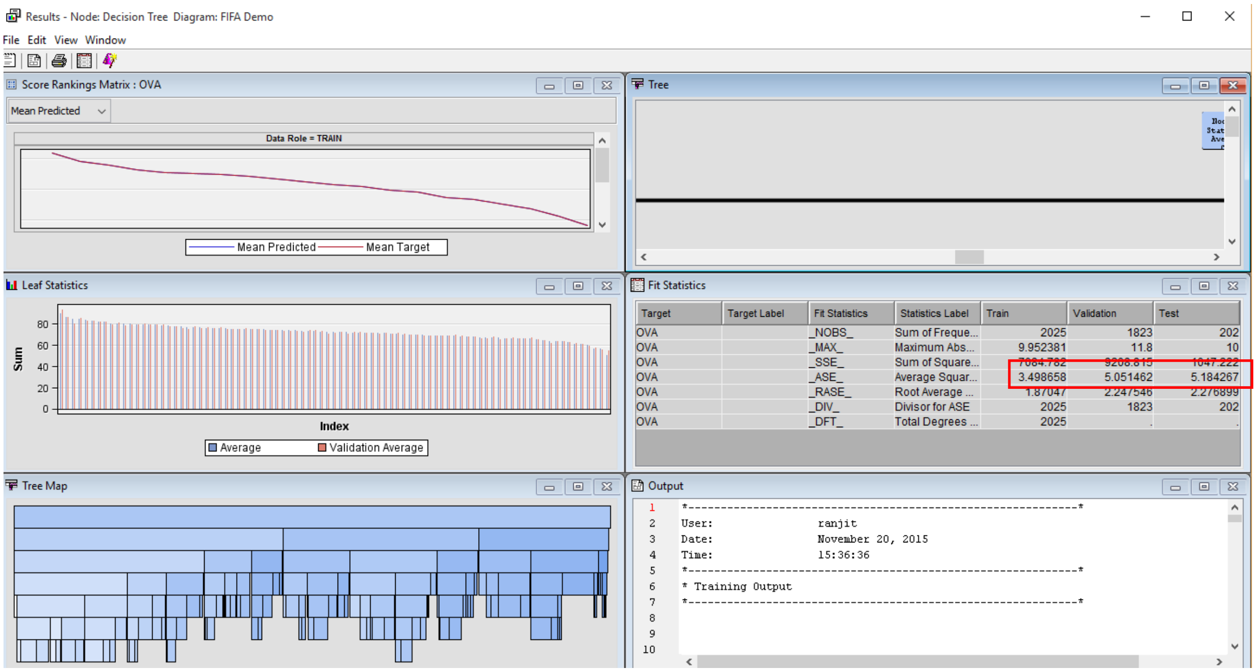
In this step we performed the decision tree analysis by adding a decision tree node to the diagram.



Following results are interpreted with the help of above decision tree:

1. The attribute with the highest information gain is "PAS". We can say that the players who have got a passing rate of less than 67.5 are categorized as a defense player and the player who has got a passing rate of more than 67.5 are categorized as players who have the potential to perform well in the positions other than defense. Like the midfielders are the engine of the game who need to connect the dots of the manager's plans. Passing is an attribute will be a plus for the players.
2. As we see from this model on further drilldown on this branch out, say players with less passing rate has Defense as the next deciding factor to judge his overall ability. Similarly, the highest passing range players has shooting an offensive metric as the next deciding factor.
3. The next criteria on which the splitting is done gives further information about contribution towards the overall ability of the player.

Decision Tree Output:



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
OVA		_NOBS_	Sum of Freque...	2025	1823	202
OVA		_MAX_	Maximum Abs...	9.952381	11.8	10
OVA		_SSE_	Sum of Square...	7084.782	9208.815	1047.222
OVA		_ASE_	Average Squar...	3.498658	5.051462	5.184267
OVA		_RASE_	Root Average ...	1.87047	2.247546	2.276899
OVA		_DIV_	Divisor for ASE	2025	1823	202
OVA		_DFT_	Total Degrees ...	2025	.	.

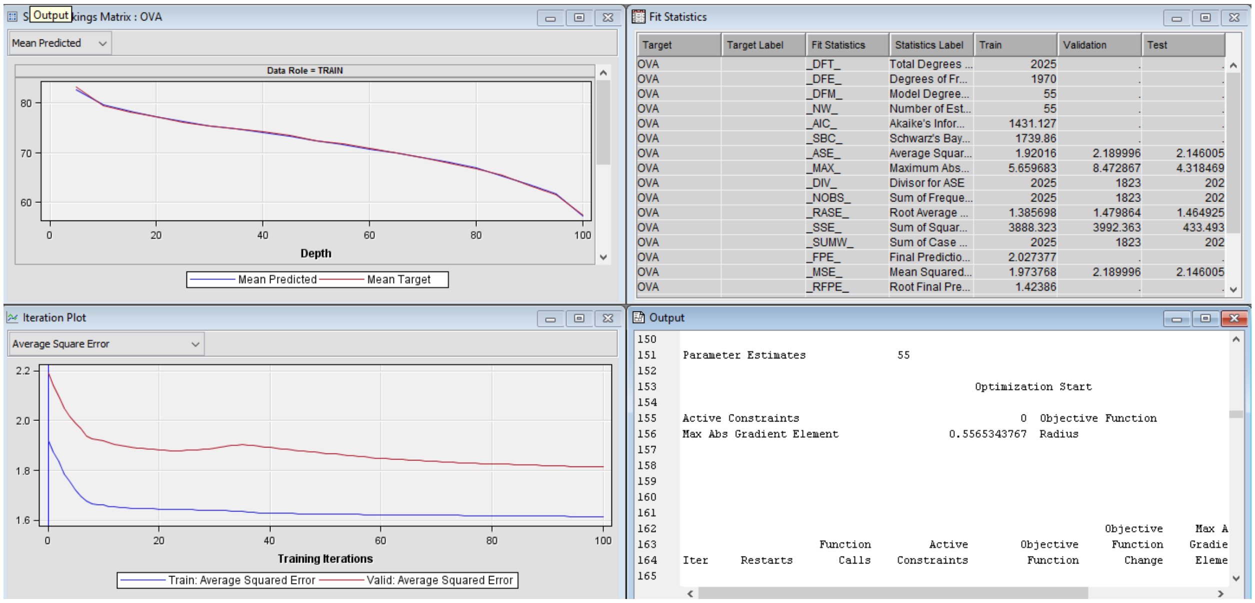
```
*------------------------------------------------------------*
User:                    ranjit
Date:                    November 20, 2015
Time:                    15:36:36
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*
```

The average square error on the prediction ranged from ~5 to 6 on training and validation data set as highlighted above.

Neural Network Model

Step 7: Further we included Neural Network Model for Predictive analysis.

Neural Network: A neural network can be thought of as a regression model that has the ability to approximate virtually any association between the inputs and the target.



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
OVA		_DFT_	Total Degrees ...	2025	.	.
OVA		_DFE_	Degrees of Fr...	1970	.	.
OVA		_DFM_	Model Degree...	55	.	.
OVA		_NW_	Number of Est...	55	.	.
OVA		_AIC_	Akaike's Infor...	1431.127	.	.
OVA		_SBC_	Schwarz's Bay...	1739.86	.	.
OVA		_ASE_	Average Squar...	1.92016	2.189996	2.146005
OVA		_MAX_	Maximum Abs...	5.659683	8.472867	4.318469
OVA		_DIV_	Divisor for ASE	2025	1823	202
OVA		_NOBS_	Sum of Freque...	2025	1823	202
OVA		_RASE_	Root Average ...	1.385698	1.479864	1.464925
OVA		_SSE_	Sum of Squar...	3888.323	3992.363	433.493
OVA		_SUMW_	Sum of Case ...	2025	1823	202
OVA		_FPE_	Final Predictio...	2.027377	.	.
OVA		_MSE_	Mean Squared...	1.973768	2.189996	2.146005
OVA		_RFPE_	Root Final Pre...	1.42386	.	.

```

Parameter Estimates              55


                                           Optimization Start

Active Constraints                             0  Objective Function
Max Abs Gradient Element            0.5565343767  Radius






                                                                       Objective     Max A
                         Function         Active       Objective        Function    Gradie
Iter      Restarts          Calls    Constraints        Function          Change     Eleme

```

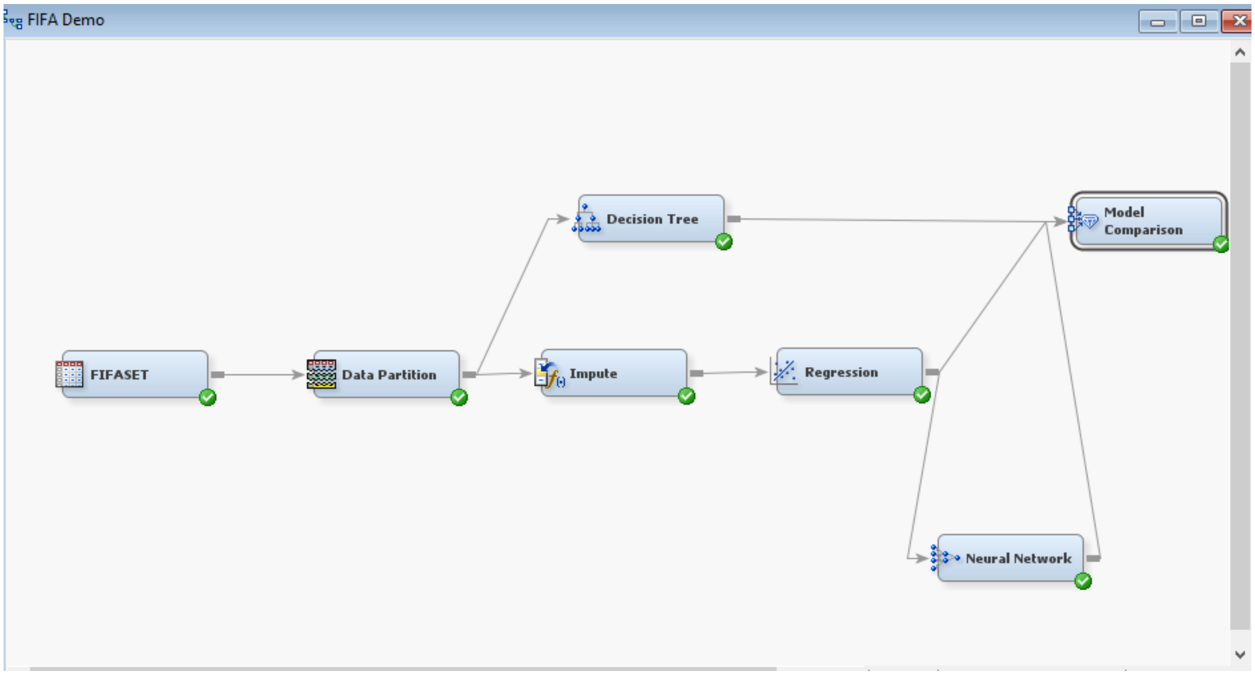
For performing predictive analysis by using neural network we have connected the regression node and the neural network node so that Neural network node contains only those inputs which are important to estimate the target variable and no of estimated weights gets reduced. By using the above approach our neural network model contains an estimated weight of 55.

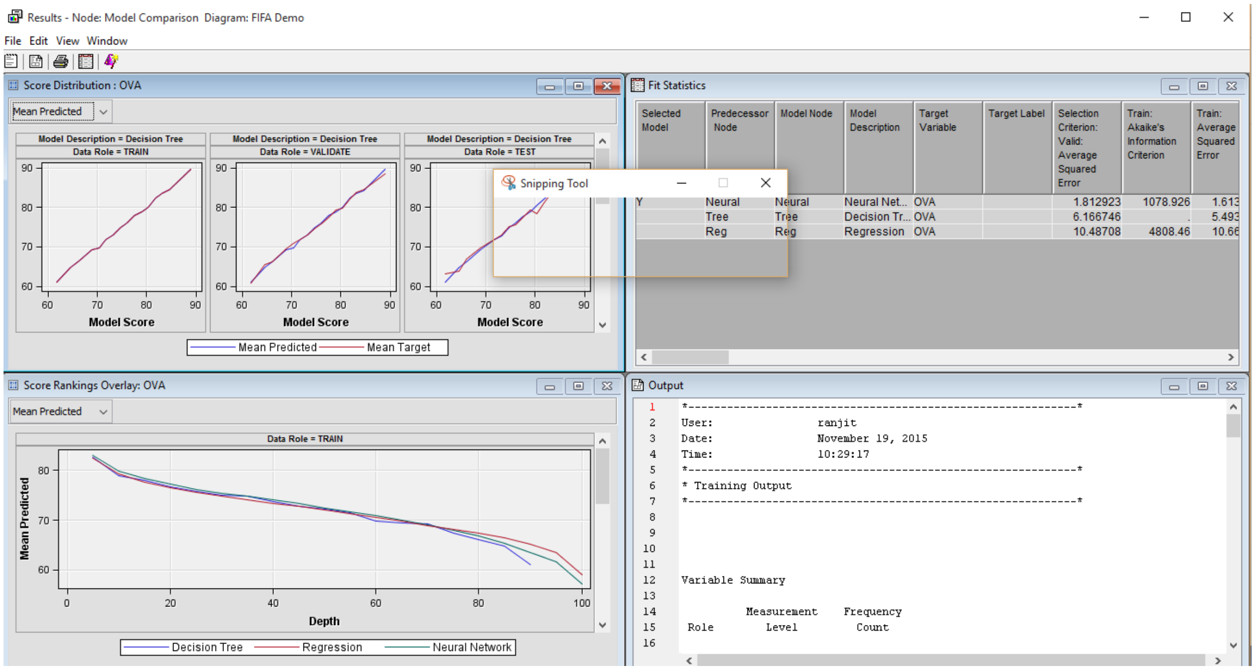
Moreover for performing neural network analysis we have increased the Number of Hidden Units to 6 so that we get a much more optimized result and a better model. After using the above approach the average square error of the neural network model gets reduced to 1.92.

Model Comparison

Step 8: Next step we did was to compare all the above predictive models and compare their results.

For this we added **Model Comparison** node to our diagram and connected all my 3 predictive models to the Model Comparison Node.





Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Akaike's Information Criterion	Train: Average Squared Error
Y	Neural	Neural	Neural Net...	OVA		1.812923	1078.926	1.613
	Tree	Tree	Decision Tr...	OVA		6.166746	.	5.493
	Reg	Reg	Regression	OVA		10.48708	4808.46	10.66

```
*------------------------------------------------------------*
User:                ranjit
Date:                November 19, 2015
Time:                10:29:17
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
 Role        Level         Count

```

After performing the model comparison we come to know that Neural Network is the best model as compared to the other model as valid average square error for neural network model is less compared to the other models.

CONCLUSION

From the analysis, we have set the benchmark to predict player's potential Based on these factors, we provide the following recommendations:

- How much of an impact could basic player's attributes like pace, dribbling, shooting…etc, have on his overall potential.
- Using neural networks try to predict the same output of current professional footballer on to the younger budding footballer.
- On behalf of this analysis Manager or Scout can get a clear picture of a younger player potential and can invest on them.

By implementing these recommendations, Manager can help players to improve their performance. The ultimate goal is:

Identifying players Performance→Invest on identified players with required potential.