

Using SAS Enterprise Miner to predict breast cancer at early stage

Gibson O. Ikoro., Queen Mary University of London; B. de la Iglesia, University of East Anglia

ABSTRACT

Breast cancer is the leading cause of cancer-related deaths among women worldwide, and its early detection can reduce mortality rate. Using a data set containing information about breast screening, we constructed a model that can provide early indication of a patient's tendency to develop breast cancer. This data set has information about breast screening from patients who were believed to be at risk of developing breast cancer. The most important aspect of this analysis is that we excluded all patients with symptoms commonly associated with breast cancer, while keeping patients with symptoms that are less likely or unknown to be associated with breast cancer as input predictors. The hope was that a model could be developed that would identify women at high risk of developing breast cancer. This group could then be subjected to more intense screening with a view to earlier detection of the cancer and thus improved outcomes. The target variable is a binary variable with two values, 1 (indicating a type of cancer is present) and 0 (indicating a type of cancer is not present). SAS® Enterprise Miner™ 12.1 was used to perform data validation and data cleansing, to identify potentially related predictors, and to build models that can be used to predict at an early stage the likelihood of patients developing breast cancer. We compared two models: the first model was built with an interactive node and a cluster node and the second was built without an interactive node and a cluster node. Classification performance was compared using a receiver operating characteristic (ROC) curve and average squares error. Interestingly, we found significantly improved model performance by using only variables that have a lesser or unknown association with breast cancer. The result shows that the logistic model with an interactive node and a cluster node has better performance with a lower average squared error (0.059614) than the model without an interactive node and a cluster node.

INTRODUCTION

All women are at risk of developing breast cancer during their life time (Jemal, Siegel et al. 2008).

The threat of breast cancer begins at puberty and rises gradually until the peri-menopausal years when it increases dramatically (Delen, Walker et al. 2005). Around the age of 75, the risk levels off (Hooper, Madhavan et al. 2010). Nevertheless, it is proved that the possibility of developing breast cancer is related to many different factors (Botha, Bray et al. 2003). Therefore, women with certain risk factors are more likely to be affected than others, age being one of the most important ones. The probability of a woman developing breast cancer in her lifetime is approximately 1 in 8 (Botha, Bray et al. 2003). According to the National Cancer Institute Surveillance, Epidemiology, and End Results Program, SEER (2000 - 2002), a woman's lifetime risk of developing breast cancer in the UK is as shown in Figure 1

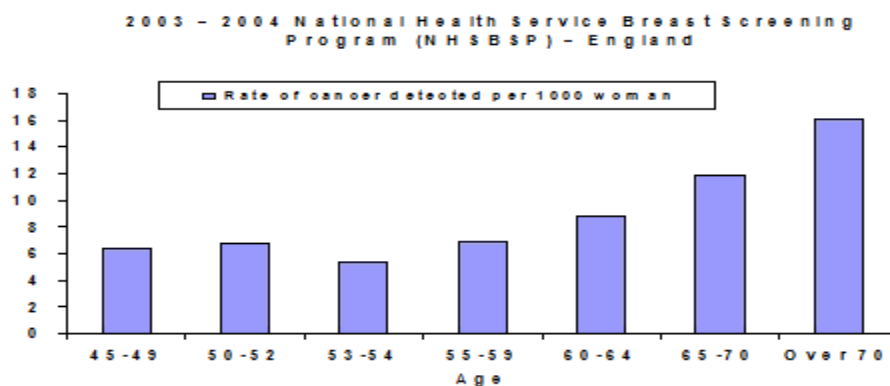


Figure 1. A woman's lifetime risk

Breast cancer is a major health problem that represents a significant worry for many women and their physicians (de la Iglesia, Potter et al. 2011). Currently recognised risk factors that associated with breast cancers development do not allow us to divide the population into a high risk group who need screening and a low risk group who do not (Hooper, Madhavan et al. 2010). This could be as a result of lack of an effective algorithm. The research questions for this study are as follows:

- **Can we identify which women should be screened?** Extracting information from a large database may help to identify a high risk group and allow a more effective and efficient screening programme (de la Iglesia, Potter et al. 2011).
- **Can we predict the likelihood of asymptomatic patients developing breast cancer?** This may involve excluding the entire variables which are one way or another strongly associated with breast cancer while keeping the variables that are less or unknown to have association with breast cancer as input predictors. Some of the Variables that are clearly associated with breast cancer include: Clinical findings on right breast, Clinical examination information concerning patient's right breast, Mammography examination information concerning patient's breast etc. Strictly speaking these variables identify patients who already have breast cancer at the time of examination (Hooper, Madhavan et al. 2010). So we are looking for predictor variables in patients who did not originally have breast cancer but who developed it subsequently.

DATASET

This paper presents an analysis of a real database containing information from patients who underwent breast screening. Data are examined of those patients who were believed to be at risk of developing specific types of breast cancer. The information used was dated from 9 January 1997 to 27 June 2002 with 13078 records overall. **Table 1** is a sample of the original dataset.

Name	Min	Max	Mean	Std Dev	Unique	Missing	Type
HysAge	0	4716	42.202	104.717	--	14997	NUM.DISC
HRTCurrent	--	--	--	--	7	9068	CATEGORICAL
MenopauseAge	0	5217	49.881	110.49	--	14790	NUM.DISC
MenarcheAge	0	1621	10.971	21.815	--	1287	NUM.DISC
PregAgeFirst	0	333	22.591	10.258	--	5395	NUM.DISC
PregNo	0	42	2.35	1.954	--	2875	NUM.DISC
ChildNo	0	50	1.951	1.566	--	2827	NUM.DISC
ContraPillCurrent	--	--	--	--	6	6151	CATEGORICAL
ClinicalFindR	--	--	--	--	27	1484	CATEGORICAL
ClinicalFindL	--	--	--	--	28	1613	CATEGORICAL
ExClinicalR	--	--	--	--	6	1569	CATEGORICAL
ExMammoR	--	--	--	--	7	8944	CATEGORICAL
ExUSR	--	--	--	--	8	12567	CATEGORICAL
ExMRIR	--	--	--	--	4	17011	CATEGORICAL
ExOtherR	--	--	--	--	3	17014	CATEGORICAL
ExClinicalL	--	--	--	--	6	1595	CATEGORICAL
ExMammoL	--	--	--	--	8	8916	CATEGORICAL
ExUSL	--	--	--	--	8	12171	CATEGORICAL
ExMRIL	--	--	--	--	4	17013	CATEGORICAL
ExOtherL	--	--	--	--	5	17011	CATEGORICAL
CytFNAR	--	--	--	--	7	15017	CATEGORICAL
CytNipDischR	--	--	--	--	7	16852	CATEGORICAL
CytHistBiopR	--	--	--	--	7	16841	CATEGORICAL
CytOtherR	--	--	--	--	5	17004	CATEGORICAL
CytFNAL	--	--	--	--	7	14961	CATEGORICAL
CytNipDischL	--	--	--	--	7	16815	CATEGORICAL
CytHistBiopL	--	--	--	--	7	16832	CATEGORICAL
CytOtherL	--	--	--	--	4	17002	CATEGORICAL
TripleConcR	--	--	--	--	5	621	CATEGORICAL
TripleConcL	--	--	--	--	5	602	CATEGORICAL
DiagnosisR	--	--	--	--	48	569	CATEGORICAL
DiagnosisL	--	--	--	--	48	335	CATEGORICAL
BCNPresentAtDiag	0	4	--	--	5	10823	CATEGORICAL
BCNSeeBCN	1	3	--	--	3	2238	CATEGORICAL
TreatDiscussed	0	1	--	--	2	0	CATEGORICAL
WrittenInfo	0	1	--	--	2	0	CATEGORICAL
GPContacted	0	1	--	--	2	0	CATEGORICAL
RTrauma	0	1	--	--	2	0	CATEGORICAL
LTrauma	0	1	--	--	2	0	CATEGORICAL
LogDeleted	0	1	--	--	2	0	CATEGORICAL

Table 1. Description of Variable and Frequency Distribution

- Number of instances (records): 13078
- Number of fields (attributes): 154
- Target attribute – Diagnosis (= YES/NO).
 - Diagnosis = Yes (7.1%)
 - Diagnosis = No (92.9%)

DATA PREPARATION

To start, we need to prepare the data. This process involves purging some variables, missing value imputation, remodelling, transformation and investigating if there is spurious correlation between input and the target.

A. DERIVING TARGET VARIABLE

Breast cancer is an overarching term used to describe the different types of cancer that affect the breast. In collaboration with domain experts we identified seven types of breast cancer that affect people, namely, axillary cancer (AXCA), invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC), ductal carcinoma in situ (DCIS), malignant axillary lymphadenopathy (MAL), Paget's disease (PAG) and Breast cancer (BC). DiagnosisR is the patient's diagnosis for right breast while DiagnosisL is the patient's diagnosis for left breast.

In remodelling the target variable, we combined the variable DiagnosisR and DiagnosisL in the dataset to form the variable "BCPresent". Where the variable "DiagnosisR" refers to patient's diagnosis for right breast and the variable "DiagnosisL" is the patient's diagnosis for left breast. We do this by assigning value "1" to "BCPresent" field when a type of cancer is present and value "0" when it is not. The result is shown in Figure 3 and SAS code is displayed Figure 2.

```
data BCPresentdata;
  set Cancerdata;
  If DiagnosisL in ('AXCA','IDC','ILC','DCIS','MAL','PAG','BC') |
    DiagnosisR in ('AXCA','IDC','ILC','DCIS','MAL','PAG','BC')
  then BCPresent = 1;
  else BCPresent = 0; run;
```

Figure 2. SAS sample code

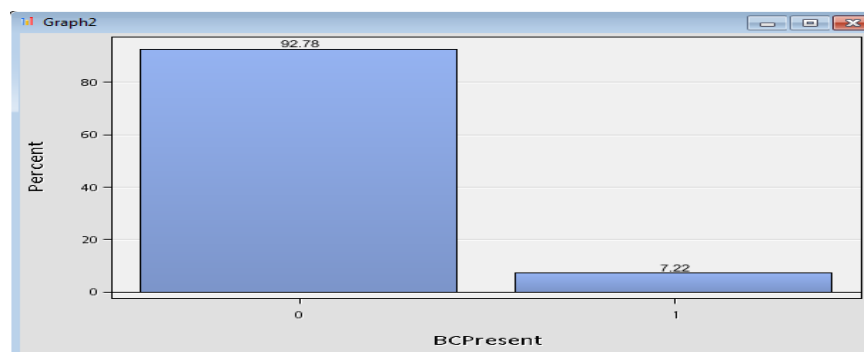


Figure 3. Breast Cancer: Target

B. MISSING VALUE IMPUTATION, TRANSFORMATION AND PARTITION

Using Multiple Imputation technique, missing values were replaced in the data sets. Multiple Imputation provides a useful strategy for dealing with data sets with missing values (Carpenter and Kenward 2013). Instead of filling in a single value for each missing value, multiple imputation procedure replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute (Rubin 1996). The missing cutoff was set to 50% (i.e the maximum percent of missing allowed for a variable to be imputed. Variables whose percentage of missing exceeds this cutoff are ignored). The number of class (categorical) variable is 109 and interval variable is 6. The imputation technique for class input variables was discriminant function method while regression method was used to impute the continuous variable. Using Partition node, the data was split into 50% for training, 30% for validation and 20% for testing. Result is shown in Table 2.

Partition Summary		
Type	Data Set	Number of Observations
DATA	EMWS3.Ids3_DATA	248482
TRAIN	EMWS3.Part_TRAIN	124240
VALIDATE	EMWS3.Part_VALIDATE	74545
TEST	EMWS3.Part_TEST	49697

Table 2: Results from the Partition node

Now, we turn to data transformation. Transforming highly skewed numerical inputs improves the model performance. Skewness is a measure of symmetry in a distribution (Beck and Shultz 1986). The skewness for a normal distribution is zero which implies symmetry. The Transform variable node can make a variety of transformation of interval-scaled variables. Each of the continuous variables was examined using StatExplore node. Few variables have too large single categories (which mean the values of these variables did not vary much) and will contribute nothing to the target variable, hence they were dropped.

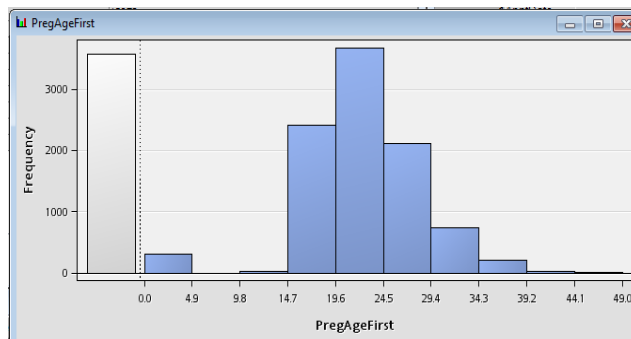


Figure 4. PreAgeFirst before transformation

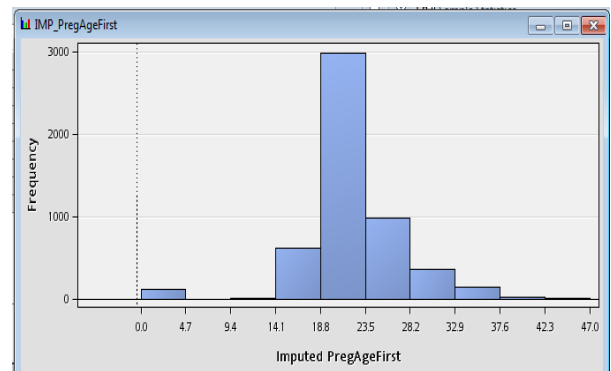


Figure 5. PreAgeFirst after transformation

Some variables were highly skewed, for example the distribution of the age at first full-term pregnancy in Figure 4 looks highly skewed to the left. It lacks the bell-shaped curve, suggesting that the distribution is not normal. We solve this problem by transforming the variable using Maximum Normal Transformation. The result of our transformation is show in Figure 5. Now the variable looks like a bell-shape curve, indicating that it is drawn from a normal distribution. Similarly, the same process was applied in the remaining continuous attributes to investigate for missing and extreme values.

C. INTERACTIVE GROUP NODE AND CLUSTERS NODE

In this section we show how interactive group node can be used to improve the output results from Cluster Node and subsequently increase the model performance overall. Interactive Group node perform variable grouping, which is a binning transformation performed on the input variables. The Output Variables window displays each variable's Gini Statistic and information value (IV). IV is used to evaluate a characteristic's overall predictive power. Variable is rejected if its IV is less than 0.10. The result and flow diagram are displayed in Table 3 and

Figure 6 respectively.

Obs	Variable	Gini Statistic	Information Value	Level for Interactive
1	PainL	15.937	0.273	BINARY
2	PregAgeFirst	12.040	0.052	INTERVAL
3	PainR	11.687	0.184	BINARY
4	ChildNo	7.989	0.035	NOMINAL
5	PregNo	7.349	0.034	NOMINAL
6	Side	5.813	0.052	NOMINAL
7	BPastHistMilL	5.767	0.015	BINARY
8	PainNonCyclicalL	4.701	0.081	BINARY
9	BPastHistMilR	4.651	0.010	BINARY
10	SmokingPerDay	1.929	0.006	INTERVAL
11	PainCyclicalL	0.000	0.000	BINARY
12	PainCyclicalR	0.000	0.000	BINARY
13	PainNonCyclicalR	0.000	0.000	BINARY
14	PainResolvedL	0.000	0.000	BINARY
15	PainResolvedR	0.000	0.000	BINARY

Table 3. Gini Statistic and information value from Interactive group

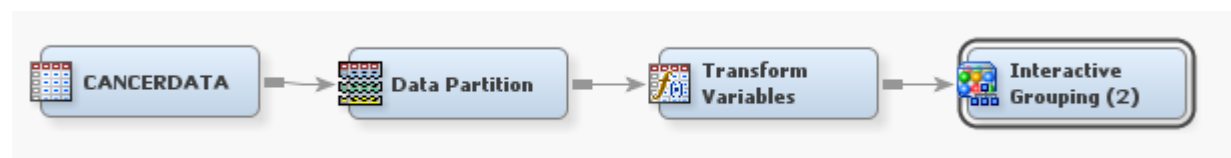


Figure 6. Flow Diagram including Interactive Grouping node

Figure 7 and Figure 8 Cluster results show the difference between clustering without variable grouping and clustering after variable grouping respectively. The first one did not use the Interactive Grouping Node before Cluster Node while the second used the Interactive Grouping Node before Cluster Node. As we can see, the cluster result in Figure 8 is better and the clusters appear to be well distributed.

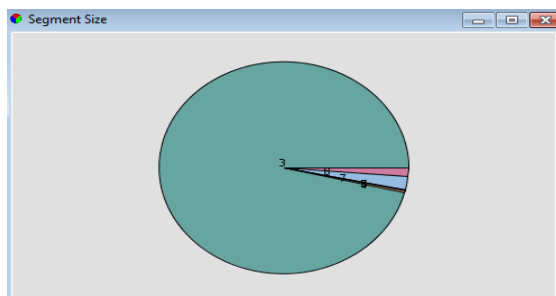


Figure 7. Without Interactive Group Node

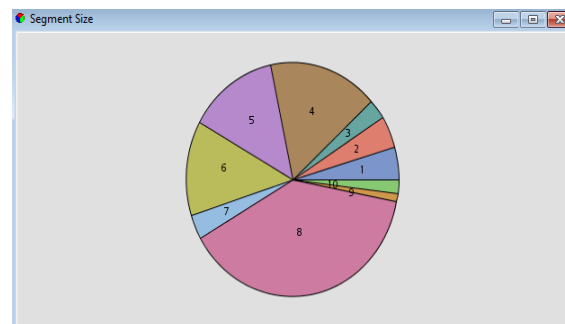


Figure 8. With Interactive Group Node

Moreover, the effect of using Interactive Grouping Node before Cluster Node can also be seen in their Chi-Square plot in Figure 9 and Figure 10. The Chi-Square plot highlights inputs that are associated with the target. Unlike Figure 9, the cluster segments (which we used as an input variable) in Figure 10 has the best chi-square among all the variables, indicating a strong association with the target variable.

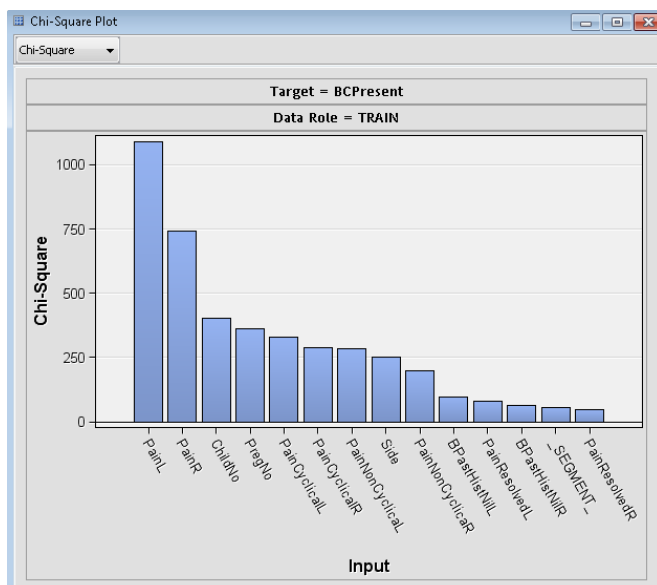


Figure 9. Chi-Square: without Interactive Group Node

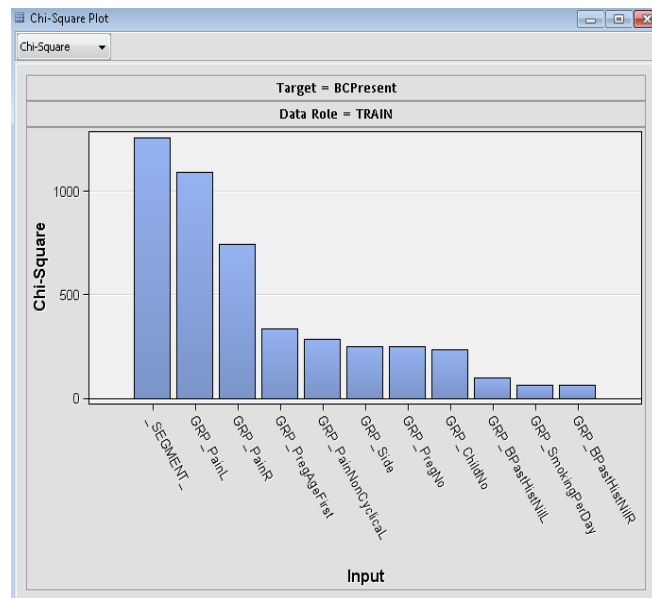


Figure 10. Chi-Square with Interactive Group Node

MODEL BUILDING

The model was developed using SAS® Enterprise Miner™ 12.1. The process flow diagram is shown in Figure 11. In the model, the dataset node (Cancerdata node) was added to the diagram workspace as the data source of the study. Then the Data Partition node was connected to Cancerdata node in order to split the original dataset into training and validation dataset. Thereafter, Transformation Node was connected to the Data Partition to transform continuous variables, followed by two regression models: In the first model, Transformation Node is followed by Interactive Group Node which performs binning transformation on the input variables in order to improve cluster result and variable predictive power. Thereafter, Cluster node was connected to the Interactive group node which performs the actual clustering. The cluster segments with other variables were used as an input to improve prediction. Logistic Regression nodes were connected to the Cluster node. Similarly, the second model consists the entire node as in the first mode except the Interactive group node and cluster node. The second model is made up of Dataset node, Data Partition Node, Transformation Node and Logistics Node. Then the performances for both models were assessed and compared by connecting them towards Models Comparison node.

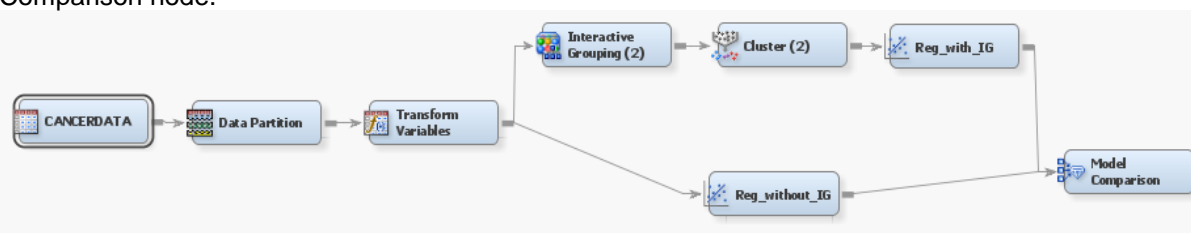
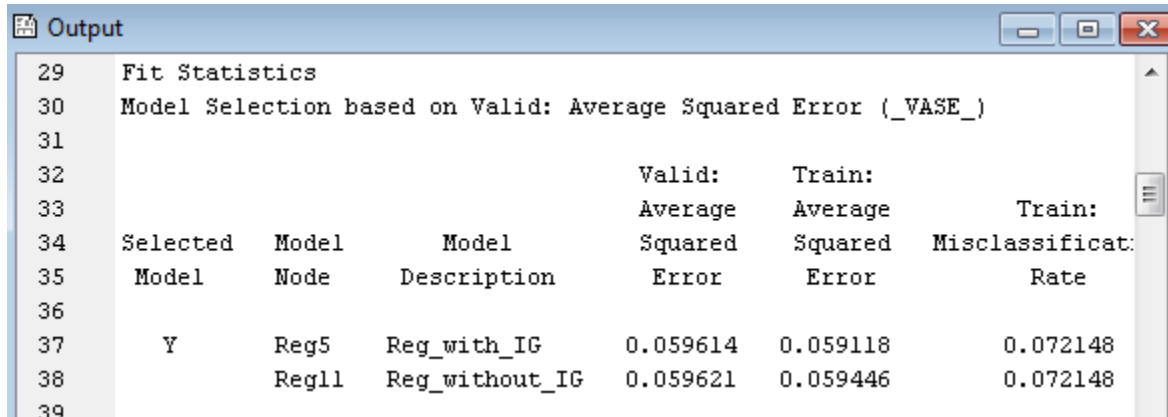


Figure 11. SAS Enterprise Miner Model Flow Diagram

RESULTS

The models were compared based on validation average squared error. Small value of average squared error gives a better model. Table 4 shows the model fit statistics - average squared error for both training and validation data partitions.



Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error	Train: Misclassification Rate
Y	Reg5	Reg_with_IG	0.059614	0.059118	0.072148
	Reg11	Reg_without_IG	0.059621	0.059446	0.072148

Table 4. Validation: Average Square Error of Models Compared.

It clearly shows that regression model with an Interactive Node and Cluster Nodes (Reg with IG) has the lowest averaged squared error with values (0.059614). Hence, the best model.

Figure 12. shows the Receiver Operating Characteristics (ROC) curves for both training, validation and testing data.

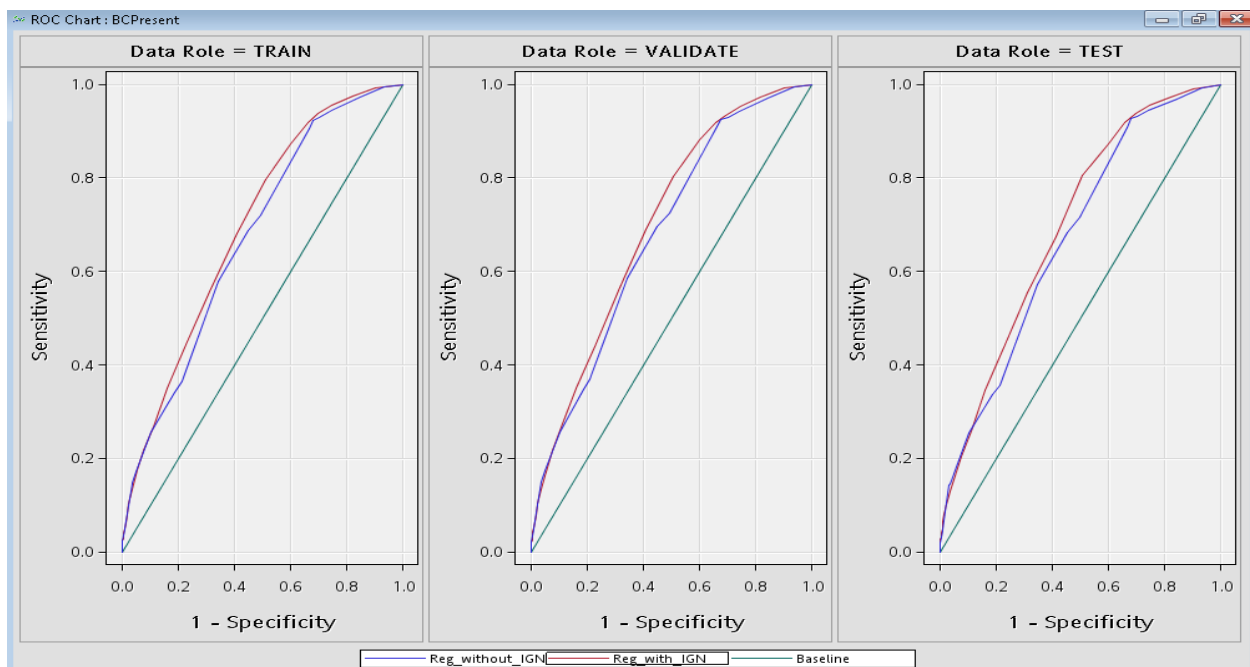


Figure 12. ROC Chart

ROC curves are used to evaluate and compare the performance of diagnostic tests; they can also be used to evaluate model fit. ROC curve is a plot of the proportion of true positives (events predicted to be

events) versus the proportion of false positives (non-events predicted to be events) . The ROC curve of the regression model built with an Interactive Group Node and Cluster Node (Reg with IG (the red line)) has the highest area under the curve, indicating a superior performance than the model built without an Interactive Group Node and Cluster Node.

THE BEST MODEL

The best stepwise regression model is displayed in Table 5

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
GRP_PregAgeFirst	2	3756.0222	<.0001
GRP_Side	1	34.7967	<.0001
WOE_BPastHistNilL	1	60.3934	<.0001
WOE_BPastHistNilR	1	10.7966	0.0010
WOE_ChildNo	1	24.0336	<.0001
WOE_PainL	1	593.8576	<.0001
WOE_PainR	1	458.1053	<.0001
SEGMENT	16	112278.451	<.0001

Table 5. The best model

The best model (Regression model with an Interactive Group Node and Cluster Node) is the model that consists of the following effects: Number of children (ChildNo), No past history for left breast (BPastHistNilL), No past history for right breast (BPastHistNilR), the side of Breast that was screened (Side), Age at first pregnancy (PregAgeFirst), lack of pain on the left breast (PainL) and lack of pain on the right breast (PainR). Interestingly, Table 5 shows that most of the variables such as Number of children, No past history for breast cancer, lack of pain etc which are less likely to be associated with breast cancer are statistically significant at the 5% level.

DISCUSSION AND CONCLUSION

The important aspect of this work is that the analysis excluded all patients with symptoms commonly associated with breast cancer. The input predictors were demographic variables and variables associated with the patients' social, medical and obstetric history. The hope was that a model could be developed that would identify women at high risk of developing breast cancer. This group could then be subjected to more intense screening with a view to earlier detection of the cancer and thus improve outcomes.

Two of the variables in the best fitting model have long been recognised as risk factors for breast cancer (age at first pregnancy and number of children) with late age at first pregnancy and having no children being associated with an increased risk. The other variables are non-discriminatory as the majority of women will have no pain and no symptoms in their breast.

The results suggest that most breast cancers occur in women without a significant history of breast symptoms. As a result, it is not possible to identify a group within the population who should be subjected to more intensive screening than the rest of the population.

Another interesting thing to note is the significant improvement made to the whole model by adding an Interactive Group Node and Cluster Node. Before adding Interactive Group Node, the Cluster Node produced a very poor result. However, incorporating Interactive Group Node increases the performance of the Cluster Node as well as the overall performance of the model.

REFERENCES

- Beck, J. R. and E. K. Shultz (1986). "The Use of Relative Operating Characteristic (Roc) Curves in Test-Performance Evaluation." *Archives of Pathology & Laboratory Medicine* **110**(1): 13-20.
- Botha, J. L., et al. (2003). "Breast cancer incidence and mortality trends in 16 European countries." *Eur J Cancer* **39**(12): 1718-1729.
- Carpenter, J. R. and M. G. Kenward (2013). *Multiple imputation and its application*. Chichester, West Sussex, John Wiley & Sons.
- de la Iglesia, B., et al. (2011). "Performance of the ASSIGN cardiovascular disease risk score on a UK cohort of patients from general practice." *Heart* **97**(6): 491-499.
- Delen, D., et al. (2005). "Predicting breast cancer survivability: a comparison of three data mining methods." *Artif Intell Med* **34**(2): 113-127.
- Hooper, L., et al. (2010). "Effects of isoflavones on breast density in pre- and post-menopausal women: a systematic review and meta-analysis of randomized controlled trials." *Hum Reprod Update* **16**(6): 745-760.
- Jemal, A., et al. (2008). "Cancer statistics, 2008." *CA Cancer J Clin* **58**(2): 71-96.
- Rubin, D. B. (1996). "Multiple imputation after 18+ years." *Journal of the American Statistical Association* **91**(434): 473-489.

ACKNOWLEDGMENTS

The authors thank Samuel Leinster, an Emeritus Professor of Medical Education at University of East Anglia, for his valuable feedback and suggestions during reviewing and proof-reading this paper. We also grateful to SAS Institute Inc. for the funding of this study to be presented in SAS Global Forum 2015.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Gibson Ikoro
Organization: Queen Mary University of London
Address: Mile End Rd, London E1 4NS
E-mail: gibson.ikoro@gmail.com

Name: B. de la Iglesia
Organization: University of East Anglia
Address: Norwich Research Park, Norwich, Norfolk NR4 7TJ
E-mail: bli@cmp.uea.ac.uk

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.