Paper 3356-2015

Improving the performance of two stage model using Association node of SAS® Enterprise Miner 12.3™

Girish Shirodkar, Ankita Chaudhari, and Dr. Goutam Chakraborty,

Oklahoma State University

ABSTRACT

Over the years, few published studies have discussed ways to improve the performance of two-stage predictive models. This study is an attempt to demonstrate how we can leverage the Association node in SAS® Enterprise Miner™ to improve the classification accuracy of a two-stage model. We find that creating dummy variables to capture association rules with lifts higher than1.6 and using those as potential input variables improve fit statistics of both the first and the second stage models of the two stage model. For example, for the first stage model, the validation ASE improved to 0.228 from 0.238, cumulative lift to 1.56 from 1.40. For the second stage model, the misclassification rate improved to 0.240 from 0.243 and the final prediction error to 0.17 from 0.18.

INTRODUCTION

Two stage model is frequently used in marketing and other business applications where modeling takes into account the conditional dependence that exists between two target variables. For example, in marketing applications of a promotional offer, we can model whether a customer is going to take the offer or not followed by how much the customer is willing to spend based on the acceptance of the offer. While two stage models have been in use for a while, few studies have suggested how to improve performance of such models. This study, which is based on ten year (1999-2008) data of 130 US hospitals and integrated delivery networks, is an attempt to employ two stage modeling methodology and improve its performance by integrating it with the results from the Association node of SAS® Enterprise Miner 12.3™.

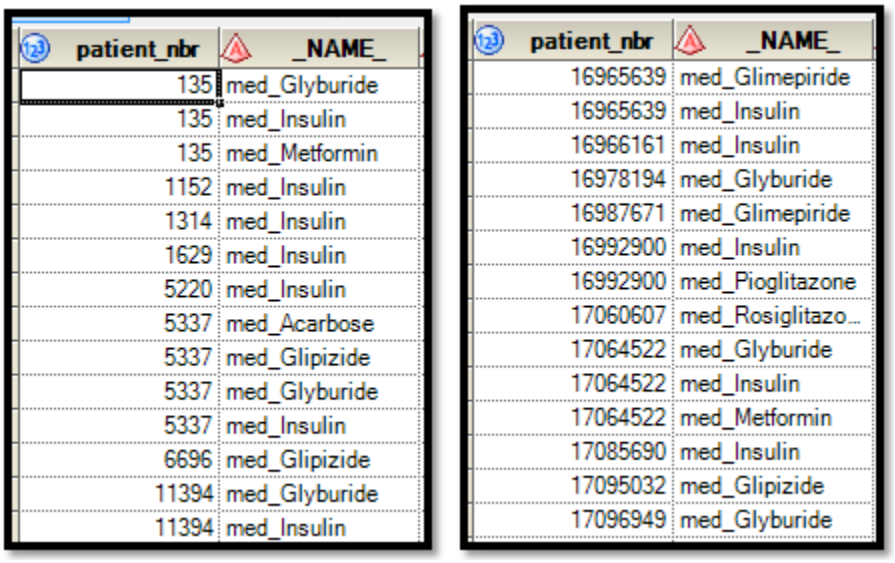
The context of this study involves diabetic patients and whether or not they are readmitted to a hospital for treatment within a certain period of time. The first stage target variable indicates if a diabetic patient will be readmitted in near future and second stage target variable indicates if the patient will be readmitted within thirty days or not.

DATA PREPARATION

The data, which is based on the records of numerous US hospitals and integrated delivery networks, initially consisted of 99,947 observations and 38 variables and provided by UCI machine library. It is a multivariate

dataset consisting of variables related to patient admission records, lab results, prescribed medications and patient readmission records. Data exploration via the StatExplore node of SAS® Enterprise Miner 12.3, identified that dataset was free from the missing values but three of the variables related to inpatient and outpatient visits were right skewed. These three variables were log-transformed using the Transform node in SAS® Enterprise Miner 12.3.

This dataset also contained medication prescription history which shows if a particular medicine was prescribed to a patient or not. In order to perform the association analysis for the prescribed medications, the data was first split into two partitions, namely the data for the patients who were not readmitted and the patients who are readmitted. Both the datasets were then transposed using PROC TRANSPOSE and the resulting datasets kept only two columns: Patient number and prescribed medications name as shown below.

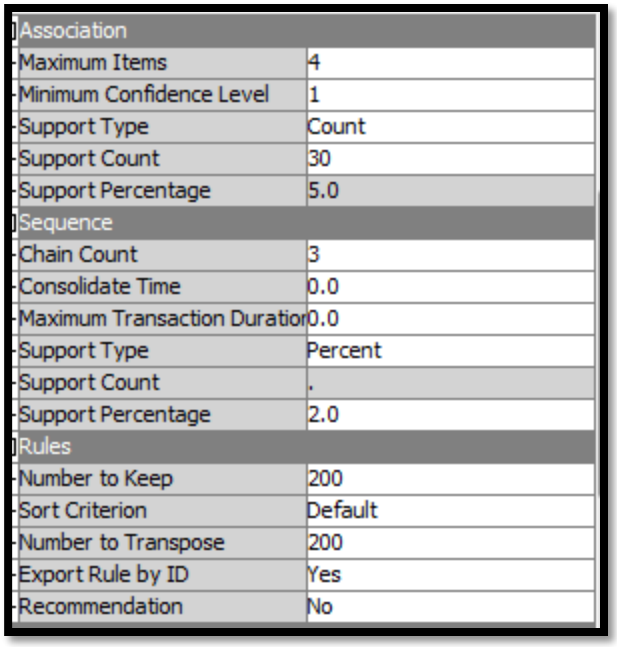


patient_nbr	_NAME_
135	med_Glyburide
135	med_Insulin
135	med_Metformin
1152	med_Insulin
1314	med_Insulin
1629	med_Insulin
5220	med_Insulin
5337	med_Acarbose
5337	med_Glipizide
5337	med_Glyburide
5337	med_Insulin
6696	med_Glipizide
11394	med_Glyburide
11394	med_Insulin

patient_nbr	_NAME_
16965639	med_Glimepiride
16965639	med_Insulin
16966161	med_Insulin
16978194	med_Glyburide
16987671	med_Glimepiride
16992900	med_Insulin
16992900	med_Pioglitazone
17060607	med_Rosiglitazo...
17064522	med_Glyburide
17064522	med_Insulin
17064522	med_Metformin
17085690	med_Insulin
17095032	med_Glipizide
17096949	med_Glyburide

Figure 1: Association analysis data samples for two datasets

The resulting transposed datasets (readmitted and non-readmitted patients medication data) were then imported in SAS® Enterprise miner 12.3 as “transaction” datasets. Association rules were then generated using the “Association” node. Below are the primary settings of Association node used for generating the association rules for prescribed medications.



Association	
Maximum Items	4
Minimum Confidence Level	1
Support Type	Count
Support Count	30
Support Percentage	5.0
Sequence	
Chain Count	3
Consolidate Time	0.0
Maximum Transaction Duration	0.0
Support Type	Percent
Support Count	.
Support Percentage	2.0
Rules	
Number to Keep	200
Sort Criterion	Default
Number to Transpose	200
Export Rule by ID	Yes
Recommendation	No

Figure 2: Association node properties panel settings used

After the association rules generation for readmitted and non-readmitted patients, one major challenge was to determine the basis for choosing the rules to be used in the predictive modeling. It is always good to know the proportion of the data where the given rule is true or how often the right side of the rule is true. But in order to use the association rules in predictive modeling, it was imperative to know how much better the rule for prediction is than just the random guess. After considering all of the above metrics, "Lift" value of the association analysis rules was used as a basis for rule selection. We chose a cutoff value of lift as 1.6 and all the rules above that cut-off value were selected from the association analysis of both the datasets (readmitted and non-readmitted patients). Following tables summarize the selected rules that were used as a dummy variable as input in predictive modeling.

Table 1: Selected association rules and dummy variables for readmitted patients

Lift	Association Rule	Dummy Variable Name
1.85	med_Rosiglitazone ➔ med_Metformin & med_Glimeperide	Readmission_Association_A
1.83	med_Acarbose ➔ med_Glipizide	Readmission_Association_B
1.82	med_Metformin ➔ med_Rosiglitazone & med_Insulin & med_Glyburide	Readmission_Association_C

1.71	med_Metformin & med_Insulin ➔ med_Rosiglitazone & med_Glimepiride	Readmission_Association_D
1.67	med_Rosiglitazone & med_Glyburide ➔ med_Metformin	Readmission_Association_E
1.66	med_Metformin ➔ med_Rosiglitazone & med_Glimepiride	Readmission_Association_F
1.61	med_Pioglitazone & med_Metformin ➔ med_Insulin & med_Glyburide	Readmission_Association_G
1.61	med_Pioglitazone & med_Insulin & med_Glyburide ➔ med_Metformin	Readmission_Association_H

Table 2: Selected association rules and dummy variables for non-readmitted patients

Lift	Association Rule	Dummy Variable Name
1.98	med_Acarbose ➔ med_Glyburide	NO_Readmission_Association_A
1.82	med_Rosiglitazone & med_Insulin ➔ med_Metformin & med_Glimepiride	NO_Readmission_Association_B
1.79	med_Rosiglitazone & med_Glyburide ➔ med_Metformin	NO_Readmission_Association_C
1.76	med_Glyburide & med_Glipizide ➔ med_Metformin	NO_Readmission_Association_D
1.74	med_Pioglitazone & med_Insulin ➔ med_Metformin & med_Glimepiride	NO_Readmission_Association_E
1.74	med_Pioglitazone & med_Metformin ➔ med_Insulin & med_Glyburide	NO_Readmission_Association_F
1.72	med_Rosiglitazone ➔ med_Metformin & med_Insulin & med_Glimepiride	NO_Readmission_Association_G

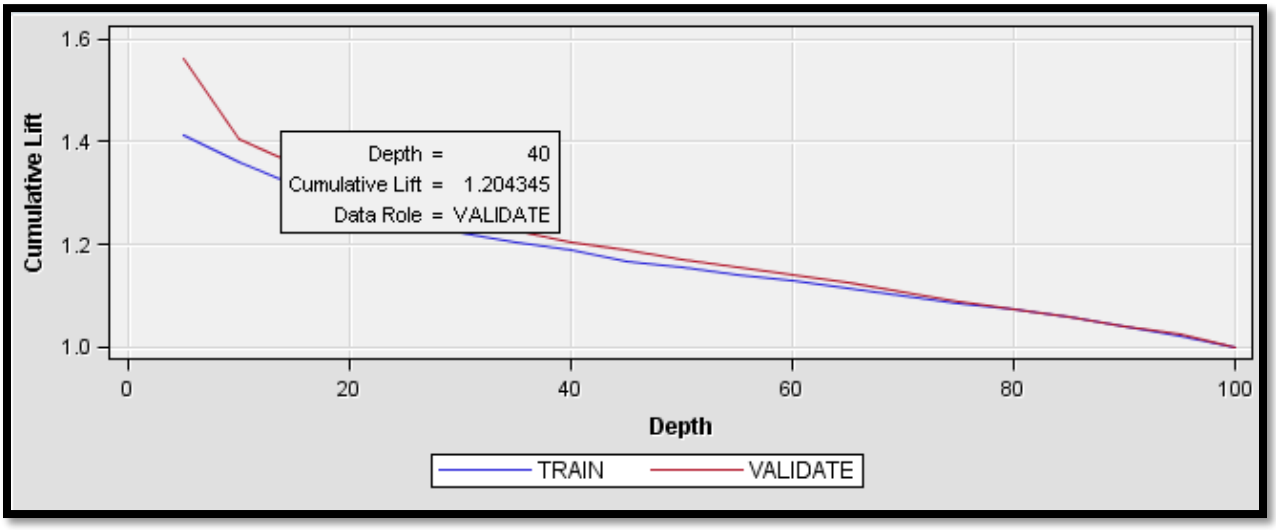
1.69	med_Nateginide ➔ med_Pioglitazone	NO_Readmission_Association_H

Whenever any of the above mentioned association rule was true for given patient record in the main data, corresponding dummy variable was assigned the value of '1' or '0' otherwise. Along with the above mentioned dummy variables, two other variables; namely "Readmission_Association_Flag" and "No_Readmission_Association_Flag", were created if any of the rules generated by the association analysis was true for given patient record, irrespective of the lift value.

MODEL BUILDING

Once the data was prepared and ready for analysis, it was partitioned into Training and Validation datasets with a ratio of 70:30. Next, we used two stage sequential modeling wherein the first stage predicted if the diabetic patient was readmitted and the second stage predicted whether the readmission happened before thirty days.

For the first stage of predictive modeling, various models such as Decision Trees with ProbChisq, Entropy and Gini as selection criterions, Forward Logistic Regression, Backward Logistic Regression, Stepwise Logistic Regression, Rule Induction and MBR were built based on the previously selected variables. Using Model Comparison Node in SAS® Enterprise Miner 12.3, competing models were diagnosed and compared with each other. The backward logistic regression model outperformed competing models for the first stage. After inclusion of dummy variables from association analysis, many fit indices improved such as the validation ASE to 0.228 from previous 0.238, cumulative lift of 1.56 versus 1.40.



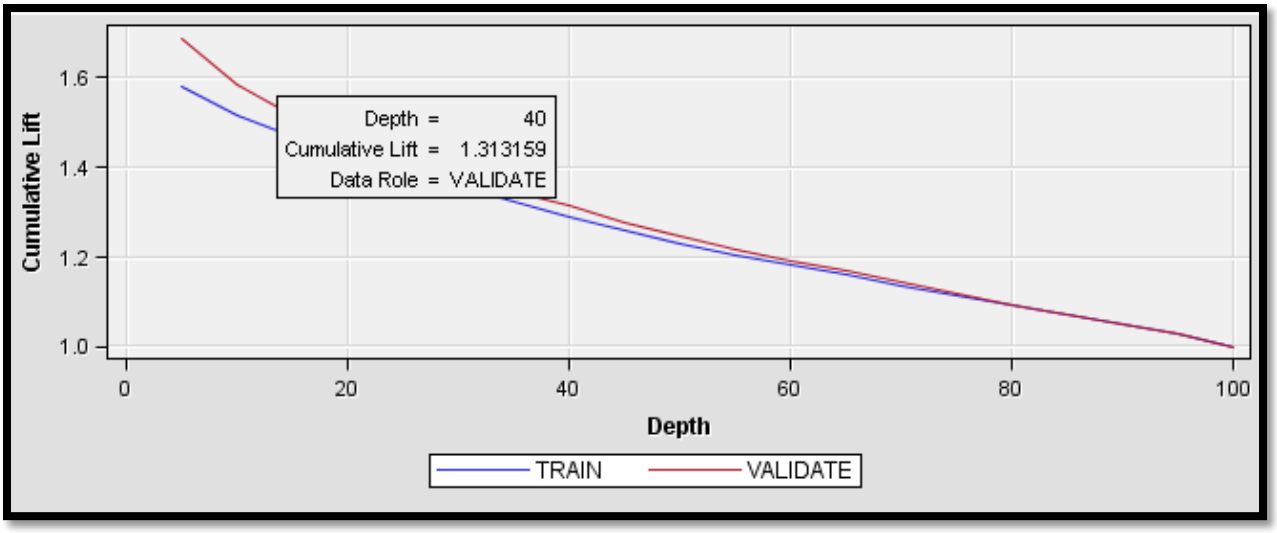


Figure 3: Cumulative lift before and after inclusion of dummy variables from association analysis for the first stage of modeling

Results from the best model of first stage revealed that factors such as emergency room admissions, female patients and patients with glucose serum level higher than 300 dominated the patient's chance of readmission.

```
Intercept  AlCresult  LOG_number_emergency  LOG_number_inpatient  LOG_number_outpatient
Med_diabetes  Medication_Change  No_Readmission_Association_Flag  Readmission_Association_Flag
 age_group  gender  max_glu_serum  med_Acarbose  med_Chlorpropamide  med_Glimepiride
med_Glipizide  med_Glyburide  med_Insulin  med_Metformin  med_Miglitol  med_Nateglinide
med_Pioglitazone  med_Repaglinide  med_Rosiglitazone  med_glyb_metf  medical_specialty
num_lab_procedures  num_medications  num_procedures  number_diagnoses  race  time_in_hospital


    Likelihood Ratio Test for Global Null Hypothesis: BETA=0

    -2 Log Likelihood          Likelihood
  Intercept     Intercept &          Ratio
       Only      Covariates     Chi-Square       DF     Pr > ChiSq

  84872.263       80701.161      4171.1023       67         <.0001
```

Figure 4: Selected model after inclusion of dummy variables from association analysis

The predicted probability of a patient being readmitted from the first stage of modeling was used as an independent variable for the second stage of predictive modeling. Again, for the second stage of predictive modeling, various models such as Decision Trees with ProbChisq, Entropy and Gini as selection criterions,

Forward Logistic Regression, Backward Logistic Regression, Stepwise Logistic Regression, Rule Induction and MBR were built based on the previously selected variables. Using Model Comparison Node in SAS® Enterprise Miner 12.3, competing models were diagnosed and compared with each other. The backward logistic regression model outperformed competing models for the second stage. After inclusion of dummy variables from association analysis, many fit indices improved such as the misclassification rate to 0.240 from previous 0.243, final prediction error of 0.17 versus 0.18.

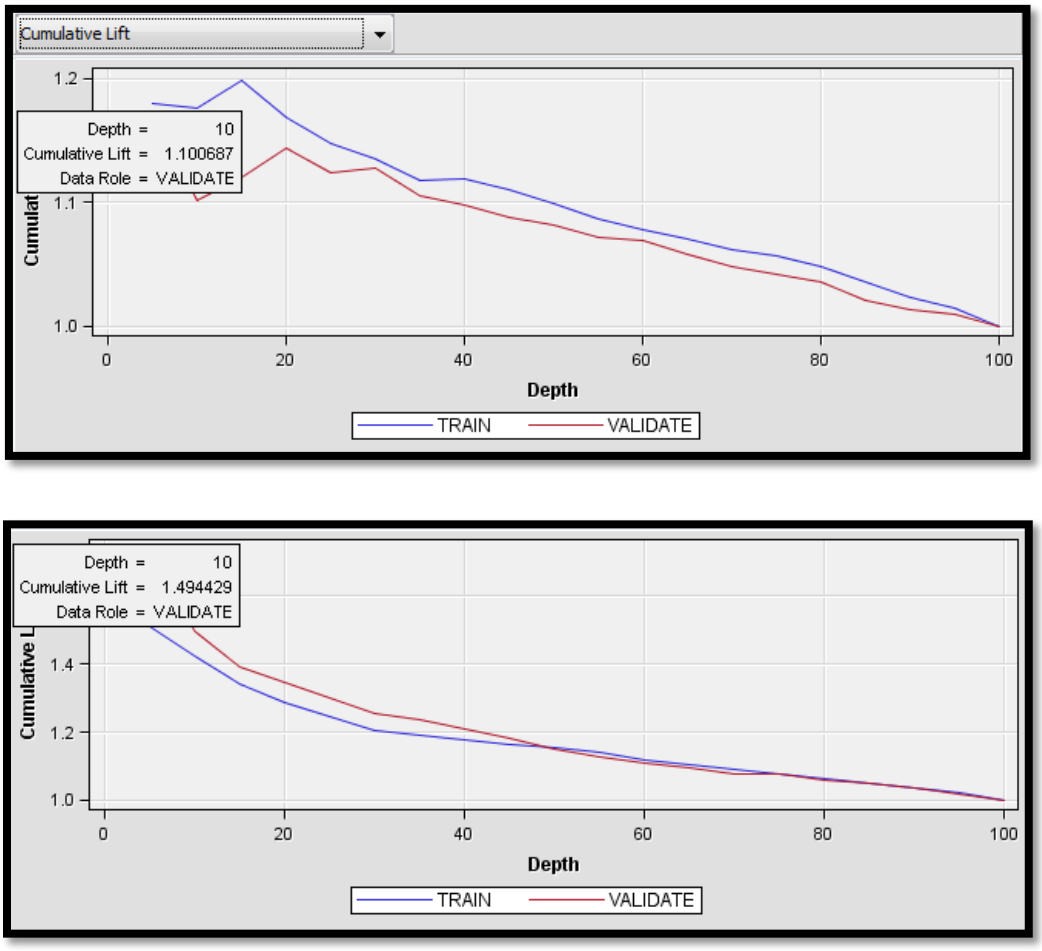


Figure 5: Cumulative Lift before and after inclusion of dummy variables from association analysis for the second stage of modeling

Table 3: Confusion matrix for the second stage of two stage predictive modeling

		Predicted	
		After 30 Days	**Before 30 Days**
Actual	**After 30 Days**	9926 (75.30%)	81 (0.62%)
	Before 30 Days	3013 (23.32%)	96 (0.74%)

CONCLUSION

The purpose of this study was to investigate whether new variables created through association analysis can improve the performance of the two stage predictive modeling. We found that after inclusion of dummy variables created through association analysis, various performance measures of the first stage model of the two stage modeling such as Average squared error, cumulative lift and AIC of the model improve significantly. Also the fit statistics such as misclassification rate as well as final prediction error improved for the second stage model after inclusion of association analysis results.

This study gives some broad ideas about how to leverage results of association analysis node to improve results of two-stage model node in SAS® Enterprise Miner 12.3™. The study revealed that factors such as emergency room admissions, female patients and patients with glucose serum level higher than 300 dominated the patient's chance of readmission while factors such as medication change, 70-80 years old patients, higher inpatient visits and time spent influenced the patient's chance of readmission before thirty days. We hope that our findings provide insightful information to doctors about what types of prescribed medications influence readmission of diabetic patients along with admission information and patient demographics.

REFERENCES

Gelman, Andrew. August 3, 2005. "*Two-Stage Regression and Multilevel Modeling: A Commentary*". Available at http://www.stat.columbia.edu/~gelman/research/published/459.pdf

Angrist, Joshua and Imbens, Guido. "*Two stage least square estimation of average causal effects in models with variable treatment intensity*". Available at http://scholar.harvard.edu/imbens/files/wo-stage_least_squares_estimation_of_average_causal_effects_in_models_with_variable_treatment_intensity.pdf

Brusilovskiy, Eugene and Brusilovskiy, Dmitry. June, 2008. "*Joint regression models for sales analysis using SAS*". Available at http://www.bisolutions.us/web/graphic/BISolutions-JOINT-REGRESSION-MODELS-FOR-SALES-ANALYSIS-USING-SAS.pdf

CONTACT INFORMATION

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

Girish Shirodkar
Oklahoma State University, Stillwater, Oklahoma
Phone: 405-780-5321
E-mail: girish.shirodkar@okstate.edu

Girish Shirodkar is a Master's student in Management Science and Information Systems at Oklahoma State University. He is a Base SAS®, Advanced SAS® and SAS® Certified Business Analyst (Regression and Modeling) and has JMP® Exploration certificate. In May 2014, he received his SAS® and OSU Data Mining Certificate.

Ankita Chaudhari
Oklahoma State University, Stillwater, Oklahoma
Phone: 405-780-5321
E-mail: Ankita.chaudhari@okstate.edu

Ankita Chaudhari is a Master's student in Management Science and Information Systems at Oklahoma State University. She is a Base SAS®, Advanced SAS® and SAS® Certified Business Analyst (Regression and Modeling) and has JMP® Exploration certificate. In May 2014, she received her SAS® and OSU Data Mining Certificate.

Dr. Goutam Chakraborty
Oklahoma State University, Stillwater, Oklahoma
E-mail: goutam.charabotry@okstate.edu

Dr. Goutam Chakraborty is Ralph A. and Peggy A. Brenneman professor of marketing and founder of SAS® and OSU Data Mining Certificate and SAS® and OSU Business Analytics Certificate at Oklahoma State University. He has published in many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired M2007 data mining conference. He has over 25 years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.