A Comparison of Model Building via Rapid Predictive Modeler (RPM) vs. Full Suite of Enterprise Miner

Srikar Rayabaram, MSIS and SAS & OSU Data Mining Certificate Program, Oklahoma State University, Stillwater, OK, 74078
Faculty Advisor: Dr. Goutam Chakraborty

Introduction

Today most large organizations use analytics for better decision making. Even with the widespread availability of point-and-click interface for advanced predicting modeling and analytics software such as SAS® Enterprise Miner, building good predictive models still requires analysts to pre-process and manipulate data as well as to configure and compare many different models. The job of pre-processing, configuring and comparing requires a person with deep statistical or data modeling knowledge. Large businesses can afford experienced statisticians or data scientists who can handle the tasks mentioned above. However, this is not the case with most of the SMEs (small and medium size businesses).

SAS ® has recently released a component called "Rapid Predictive Modeler (or, RPM)". The RPM can be run from SAS® Enterprise Guide™. SAS® contends that using RPM makes it very easy for a person with minimal training in the area of statistics or data modeling to quickly develop a *reasonable* predictive model. In this research, we test this contention by a controlled experiment where novices (graduate students with no training in statistics or Enterprise Miner) and experts (students with statistics classes and exposure to Enterprise Miner) are given identical instructions and the same data and are asked to build the same model first using RPM and then using the full power and functionalities of Enterprise Miner. Time spent in building the model as well as participant's perceptions and satisfaction with the process of building models are explored in this poster.

The Experiment process

The following steps were executed as a part of this experiment.

1) The user is asked to start off this experiment by answering the part of the survey which mainly deals with the user's background with respect to data mining and statistics.

2) The user is then asked to develop a predictive model with the Rapid Predictive modeler module in SAS® Enterprise Guide™ 5.1 (32 bit version) based on a credit related data provided. This dataset was used to create a risk model and has 30 variables with 999 observations.

3) Next, the user is asked to develop a predictive model similar to the model developed in the earlier step but with the SAS® Enterprise Miner™ 7.1 using the same dataset.

4) The user is then asked to complete a battery of questions that ask them about perceptions of and satisfaction with the process of building models.

The Survey

The survey is divided into two parts. The first part mainly deals with the user background. Based on the responses to these questions, the user is classified as an "Expert" user or "Novice" user. If a user has any data mining/statistics experience and worked on SAS tools, then the user is classified as an expert user,. Otherwise the user is classified as a "novice" user. After collecting the user background information, he/she is asked to develop the predictive model using RPM and SAS Enterprise Miner based on the steps mentioned in the document provided.

Part B of the questionnaire mainly deals with the user feedback on the whole experiment process as well as questions on how he/she felt about building the predictive model using RPM and building the predictive model using SAS Enterprise Miner in terms of speed of building the models, ease of use and understanding of results.

Survey Results

A total of 44 people responded to the survey. Based on the survey data, 22 of the responders were classified as "Expert" users and the remaining 22 were classified as "Novice" users. All the captured responses were saved as a SAS table.
This table has following columns:

1) Category – This variable tells us whether the user is an "expert" user or a "novice" user.

2) Speed – This column deals with the user's response on what model was built quickly.

3) Ease - This column deals with the user's response on what model was easy to build.

4) Understanding – This column deals with the user's response on what model is better in terms of understanding the results.

5) RPM Execution Time – Amount of time (in minutes) a user spent in building the model using RPM.

6) EM Execution Time – Amount of time (in minutes) a user spent in building the model using SAS® Enterprise Miner™.

	Category	Speed	Ease	Understanding	RPM Execution Time	EM Execution Time
1	Novice	RPM	RPM	EM	3	88
2	Novice	RPM	RPM	EM	4	100
3	Novice	RPM	RPM	EM	3	50
4	Novice	RPM	RPM	EM	3	101
5	Expert	RPM	RPM	EM	5	33
6	Expert	RPM	RPM	EM	6	36
7	Novice	RPM	RPM	EM	8	27

Figure: Snapshot of survey results

An "expert" user on an average is taking nearly 7.54 minutes to build the predictive model using the Rapid Predictive Modeler component. The average time to build the model using SAS® Enterprise Miner™ by an expert user is 32.5 minutes.

A "novice" user is taking nearly 7.04 minutes to build the predictive model using the Rapid Predictive Modeler component. The average time to build the model using SAS® Enterprise Miner™ by a novice user is 74.3 minutes.

Summary Statistics

Results

The MEANS Procedure

Category	N Obs	Variable	Mean	Std Dev	Minimum	Maximum	N
Expert	22	RPM Execution Time	7.5454545	1.0568269	5.0000000	9.0000000	22
		EM Execution Time	32.5000000	4.4373952	25.0000000	39.0000000	22
Novice	22	RPM Execution Time	7.0454545	2.6453422	3.0000000	11.0000000	22
		EM Execution Time	74.3181818	18.7474096	27.0000000	101.0000000	22

Also, a two-sample t- test was performed on RPM execution time and Enterprise Miner execution time for all the users.

Variable: RPM Execution Time

Category	N	Mean	Std Dev	Std Err	Minimum	Maximum
Expert	22	7.5455	1.0568	0.2253	5.0000	9.0000
Novice	22	7.0455	2.6453	0.5640	3.0000	11.0000
Diff (1-2)		0.5000	2.0143	0.6073		

Category	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Expert		7.5455	7.0769	8.0140	1.0568	0.8131	1.5103
Novice		7.0455	5.8726	8.2183	2.6453	2.0352	3.7804
Diff (1-2)	Pooled	0.5000	-0.7256	1.7256	2.0143	1.6609	2.5602
Diff (1-2)	Satterthwaite	0.5000	-0.7450	1.7450			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	42	0.82	0.4150
Satterthwaite	Unequal	27.537	0.82	0.4174

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	21	21	6.27	< .0001

Variable: EM Execution Time

Category	N	Mean	Std Dev	Std Err	Minimum	Maximum
Expert	22	32.5000	4.4374	0.9461	25.0000	39.0000
Novice	22	74.3182	18.7474	3.9970	27.0000	101.0
Diff (1-2)		-41.8182	13.6227	4.1074		

Category	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Expert		32.5000	30.5326	34.4674	4.4374	3.4139	6.3413
Novice		74.3182	66.0060	82.6303	18.7474	14.4233	26.7913
Diff (1-2)	Pooled	-41.8182	-50.1072	-33.5291	13.6227	11.2325	17.3146
Diff (1-2)	Satterthwaite	-41.8182	-50.3080	-33.3283			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	42	-10.18	< .0001
Satterthwaite	Unequal	23.346	-10.18	< .0001

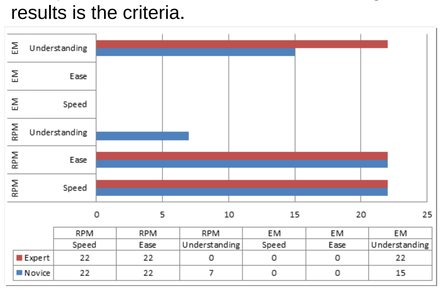
Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	21	21	17.85	< .0001

For RPM execution times, the "p-value" in the t test is > 0.05 and based on this number we cannot reject the null hypothesis which says that the average RPM execution time values for "expert" users and "novice" users are not different.

For EM execution times, the "p-value" in the t test is < 0.05 and based on this number we can reject the null hypothesis which says that the mean EM execution time values for "expert" users and "novice" users are the same.

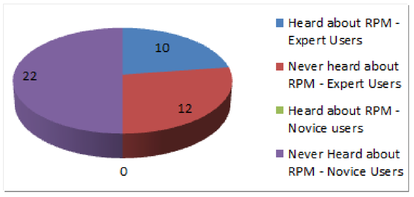
Based on these results, it can be inferred that there is a significant difference in the average execution time between the "expert" group and "novice" group when SAS® Enterprise Miner™ is used to build the model and there is no significant difference in the average execution time between the two groups when Rapid Predictive Modeler is used to build the model.

Also, from the survey data it can be seen that both the groups, i.e. experts and novices prefer RPM if speed of building the model or ease of building is the criteria. All the users in the "expert" category believe that the model built using SAS® Enterprise Miner™ is better if understanding the results is the criteria.



	RPM Speed	RPM Ease	RPM Understanding	EM Speed	EM Ease	EM Understanding
Expert	22	22	0	0	0	22
Novice	22	22	7	0	0	15

From the survey results, it can be seen that very few people knew about RPM before taking part in this experiment. Only 27% of the respondents knew about RPM before taking part in this experiment and all of them were "expert" users.



Conclusion

Based on the survey data we can conclude that most of the people find building the model using Rapid Predictive Modeler component easy and fast. The majority of the people prefer the model built using SAS® Enterprise Miner™ if understanding the model is the criteria. Also, it is evident that novices take more time compared to the "expert" users, in building the model using SAS® Enterprise Miner™ as most of them are have never used it before. However, most of the "expert" users are used to building models using SAS® Enterprise Miner™.

Reference

SAS® Rapid Predictive Modeler Product Brief.

Acknowledgments

I would like to thank Mr. Karthik Nakkeeran from SAS Institute Inc. and Professor Dr. Goutam Chakraborty for their generous support.

Contact Information

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

Srikar Rayabaram,
Email: rayabar@okstate.edu
Srikar Rayabaram is Master's student in Management Information Systems at Oklahoma State University. He is a BASE SAS® 9 and a certified SAS® predictive modeler using Enterprise Miner 6. He has also received his SAS® and OSU Data Mining Certificate in December 2012.

Dr. Goutam Chakraborty,
Email: goutam.chakraborty@okstate.edu
Goutam Chakraborty, Oklahoma State University, Stillwater OK,
Dr. Goutam Chakraborty is a 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®.