Comparing Model Performances Using JMP Pro 10: Predicting Whether a Customer Would Subscribe to a Term Deposit When Contacted for a Bank Direct Marketing Campaign

Krishna Reddy Gade, Dinesh Yadav Gaddam
Management Information Systems, Oklahoma State University, Stillwater, OK 74075

Introduction

"Who is the potential customer?" - has always been the million dollar question for a marketing manager. Companies invest considerable amounts of budget in promoting their products but the real challenge has always been to achieve a good Return On Marketing investment (ROMI). The answers to this question has led to the research of an approach which improves the ability of a company to pick a pool of selected customers and direct their promotions to them. This approach is called directed marketing.

Through this poster we try to demonstrate the capabilities of Model Comparison algorithm provided by JMP Pro 10 and come up with the best model which would predict whether a customer would subscribe to a term deposit or not when contacted through the directed marketing approach. The data used is real world data obtained from UCI learning repository. The data was collected from a Portuguese marketing campaign relating to bank term deposit subscription. The dataset contains 45211 observations, 16 Input variables and a target variable and contains no missing values. No multi-collinearity was observed among input variables and important variables for model building were identified using the chi-square test.

Dividing the dataset into training and validation datasets

A sample dataset is drawn at random comprising of 4521 observations which is about 10% of the original dataset. JMP Pro 10 provides a much easier way of generating a validation column. From JMP, choose Cols -> New Column, Initialize Data -> Random -> Random Indicator. Random Indicator column will default to values 0 (Training), 1 (Validation) and 2 (Test). For all general purposes we chose 60% for training and 40% for validation. For neural network we decided to use the KFold validation criterion which divides the original data into K subsets. In turn, each of the K sets is used to validate the model fit on the rest of the data, fitting a total of K models. The model giving the best validation statistic is chosen as the final model.

Developing the models and comparing them using the model comparison algorithm

Each time a model is built it is evaluated for its performance and then the model (prediction formula) is saved to the data table as a new column. Once all the models are built Model Comparison is used to study the performance of these models. We chose to study/compare the performances of five models which are 1. Neural Network with one hidden layer, three activity functions and KFold validation method (number of folds=5), 2. Decision tree with settings set for the best split, 3. Bootstrap Forest, 4. Boosted Tree and 5. Stepwise Logistic Regression model. Models are compared based on Goodness Of Fit and the different methods that are used to measure how good the model fits the data are R-Squared value, RMSE, ROC and Lift curve. R-Squared provides a measure of how well observed outcomes are replicated by the model, as the proportion of total variation of outcomes explained by the model. RMSE (Root mean squared error) is an estimate of the standard deviation of the random component in the data. Area under the curve is a measure of accuracy and is measured by the area under the ROC curve. Lift curve: gives the ratio of the proportion of cases with the primary outcome in the selected fraction to the proportion of cases with the primary outcome overall.

Results

Measures of Fit for Subscribed?

Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Neural		0.3913	0.4777	0.2172	0.2600	0.1351	0.0961	4494
Partition		0.3677	0.4525	0.2257	0.2640	0.1382	0.1008	4494
Bootstrap Forest		0.4868	0.5753	0.1832	0.2372	0.1251	0.0770	4494
Boosted Tree		0.3465	0.4295	0.2333	0.2628	0.1523	0.1019	4494
Fit Nominal Logistic		0.3086	0.3875	0.2468	0.2698	0.1445	0.0977	4494

Figure 1 Measures of fit statistics from Model Comparison

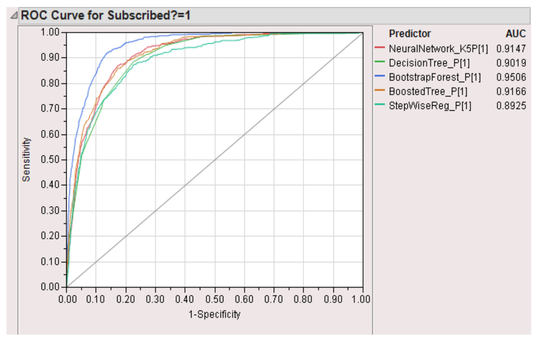

Figure 2 ROC curve

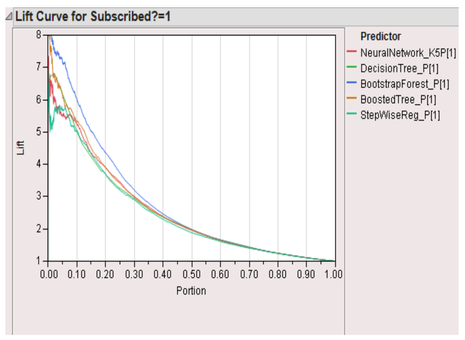

Figure 3 Lift curve

Comparing model performances

R-squared: It is observed that the Bootstrap Forest fairs better than the rest of the models explaining about 57.5% of the total variance. Root mean squared error (RMSE):It is observed that Bootstrap Forest fairs better than the rest of the models as it has the lowest RMSE value of 0.2372.Misclassification rate: It is observed that the Bootstrap Forest is the model which has the lowest misclassification rate of 0.077 when compared to the other models. AUC: It is observed that the bootstrap forest fairs better than the other models with an AUC value of 0.95 meaning that the neural network model has an accuracy rate of 95.06%. Lift curve: From the lift curve statistics it could be said that if bootstrap model is used to prepare a list of customers who would be contacted and of those list if top 5% are contacted then there is a 6 times better chance that the customer subscribes to the product as compared to the case when the customers were to be picked up at random.

An overall comparison of the models- ranking of the model based on R-Squared, RMSE, Misclassification, Lift at 5% and AUC values is given below.

	R Squared	RMSE	Misclassification	Lift at 5%	AUC
Bootstrap Forest	1	1	1	1	1
Boosted Tree	4	3	5	2	3
Decision Tree	3	4	4	3	4
Neural Network	2	2	2	5	2
Nominal logistic	5	5	3	4	5

Table 1 Ranking matrix

Conclusion

Bootstrap Forest, Decision Tree, Neural Network, Boosted Tree and Nominal Logistic Regression models are built to address the business problem. Bootstrap Forest emerged as a clear winner among all the models that were built. Model comparison algorithm is used to study/compare R-squared, RMSE, Misclassification Rate, Lift and the Area Under Curve statistics. Through this paper we tried to demonstrate the various capabilities of JMP Pro 10 such as Model validation using train and validate methodology, Model building with Bootstrap Forest-a Random-Forest technique, Gradient-Boosted Tree, Neural network with three activation functions, Nominal Logistic and a Decision Tree and finally the Model comparison algorithm for comparing fits across multiple fit predictions.

References

- http://www.jmp.com/support/help/Partition_Method.shtml#1034541
- http://blogs.sas.com/content/jmp/2012/05/16/making-better-predictive-models-quickly-with-jmp/
- http://www.academia.edu/3114404/Using_data_mining_for_bank_direct_marketing_an_application_of_the_CRISP-DM_methodology.
- http://www.aaai.org/Papers/KDD/1998/KDD98-011.pdf.
- http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#.

Acknowledgement

We would like to thank Dr. Goutam Chakraborty Professor (Marketing) and founder of SAS and OSU Data Mining Certificate Program - Oklahoma State University for his generous support throughout this project.