Paper 3361-2015

The More Trees, the Better! Scaling Up Performance Using Random Forest in SAS® Enterprise Miner™

Narmada Deve Panneerselvam, Spears School of Business, Oklahoma State University, Stillwater, OK 74078.

ABSTRACT

Random Forest (RF) is a trademark term for an ensemble approach of Decision Trees. RF was introduced by Leo Breiman in 2001.This paper demonstrates this simple yet powerful classification algorithm by building an income-level prediction system. Data extracted from the 1994 Census Bureau database was used for this study. The data set comprises information about 14 key attributes for 45,222 individuals. Using SAS® Enterprise Miner™ 13.1, models such as random forest, decision tree, neural network, gradient boosting, and logistic regression were built to classify the income level(>50K or <50k) of the population. The results showed that, the random forest model was the best model for this data, based on the misclassification rate criteria. The RF model predicts the income-level group of the individuals with an accuracy of 86.23%, with the predictors capturing specific characteristic patterns. This demonstration using SAS® can lead to useful insights into RF for solving classification problems.

INTRODUCTION

The objective of this study is to demonstrate the high performance of Random Forest under different experimental settings using SAS Enterprise Miner ™. The goal is to build a best-fit model to predict the income level of people (>50 K or < 50 K). The best model is chosen by compare and contrast of misclassification rates from different classification models.

RANDOM FOREST – LITERATURE REVIEW

In reference with the literature, Random Forest is a combination of Random Space Method and Randomized Node Optimization. The Random Space Method also known as bagging is an attractive choice in classification problems.

Bagging has been shown to provide impressive improvement in performance, by creating an ensemble of hundreds or thousands of trees with a bootstrap sample in a single procedure [5]. The significance of RF is that each tree has access to different subspace of feature sets. This random selection of features to split each node results in favorable error rate.

OUT-OF-BAG DATA

The data that is not in the bootstrap sample used is the Out-of-Bag (OOB) data. OOB data is used as test set for the tree grown on the bootstrap sample. OOB data is also used to find the unbiased estimate of error called Out-of-Bag error rate and the variable importance. Out-of-Bag error rate is the percentage of time the Random Forest predictors are correct. [Figure 1]

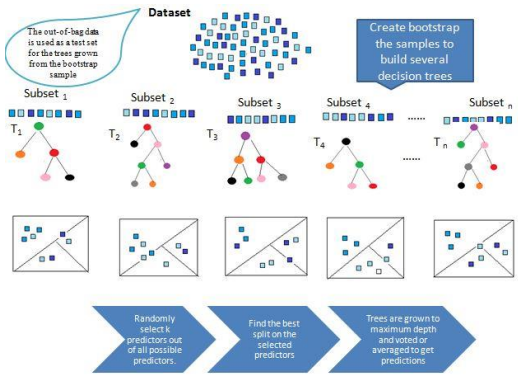

Figure 1. An overview of RF

Random Forest's ensemble approach trains several weak learners in parallel and combine their predicting probabilities Generally weak learners show low bias and high variance, fortunately averaging all weak learners result in low bias and low variance. The bootstrap sample randomly selects m - number of variables in each split to reduce correlation.

DATA

The data used for this study was extracted by Barry Becker from the 1994 census bureau database [6].The dataset includes 45222 instances recorded with fourteen key attributes, that where considered essential for decision making .The goal is to build a best-fit model to predict the income level of people (>50 K or < 50 K).The key attributes and the target of this study is listed [Figure 2]

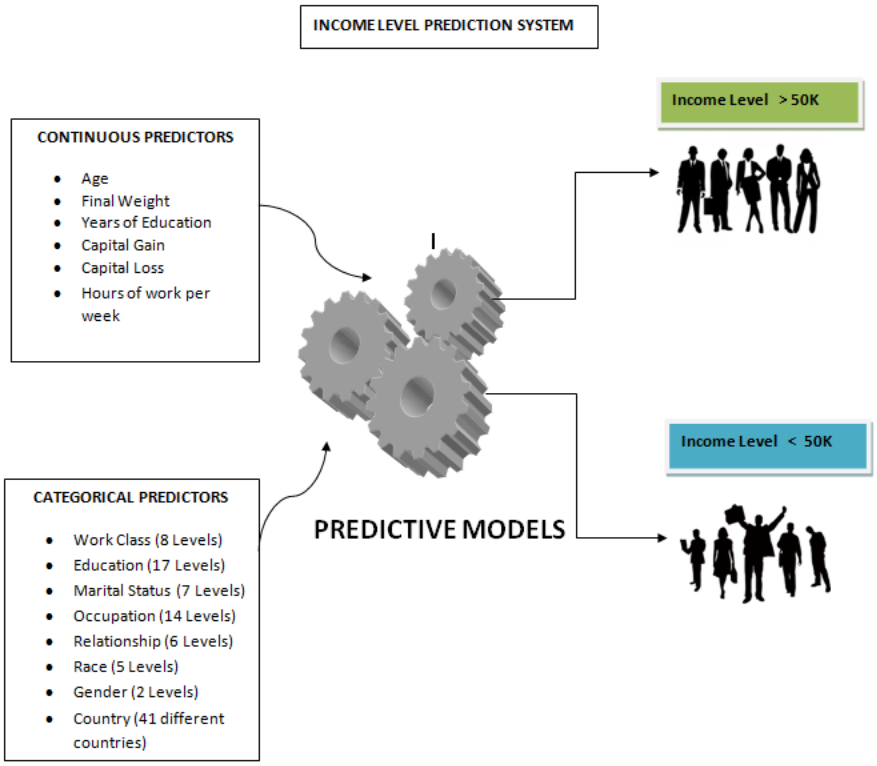

Figure 2. Income Level Prediction System

DESCRIPTIVE STATISTICS

Before model building, preliminary investigation is done to understand the data. Association between categorical predictors and target is identified using PROC FREQ. Variables Education, Occupation, Relationship, Gender; Race has association with the target because the distribution of target in each level changes, as the level of the predictor variables change. Further, to confirm the significance of this association Chi Square test was performed.

Variable	ChiSquare	p-value	Cramer's V
Education	4429.6533	<0.0001	0.3688
Occupation	4031.9743	<0.0001	0.3519
Relationship	6699.0769	<0.0001	0.4536
Gender	1518.8868	<0.0001	0.216
Race	330.9204	<0.0001	0.1008
Country	317.2304	<0.001	0.0987

Table 1. Results of Chi Square Test from SAS

From the results of the Chi Square test, we can conclude there is evidence to prove significant association between the categorical predictors and the target (Income Level) at 5% level. The predictors: Relationship, Education and Occupation seem to have stronger association based on the Cramer's V value [Table 1].

DATA PREPARATION

The variable summary is given in Table 2. The data was reasonably clean with no missing values and outliers, with exception of few continuous predictors showing skewness and kurtosis [Table 3]. To correct the shape, transformation is done for these variables using SAS Enterprise Miner ™.

Variable Summary

Role	Measurement Level	Frequency Count
INPUT	BINARY	1
INPUT	INTERVAL	5
INPUT	NOMINAL	8
TARGET	BINARY	1

Table 2. Variable Summary result from SAS Enterprise Miner ™

Interval Variable Summary Statistics

Variable	Label	Missing	N	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
age		0	32561	17	90	38.58	13.64	0.5587	-0.166
capital_gain		0	32561	0	99999	1077.65	7385.29	11.9538	154.799
capital_loss		0	32561	0	4356	87.30	402.96	4.5946	20.377
fnlwgt		0	32561	12285	1484705	189778.37	105549.98	1.4470	6.219
h_per_w		0	32561	1	99	40.44	12.35	0.2276	2.917

Table3. Interval Variable Summary Statistics from SAS Enterprise Miner ™

The continuous variables are explored using PROC UNIVARIATE. Three variables fnlwgt, capital_gain, capital_loss seems to have right skewed distribution.[Figure 3]

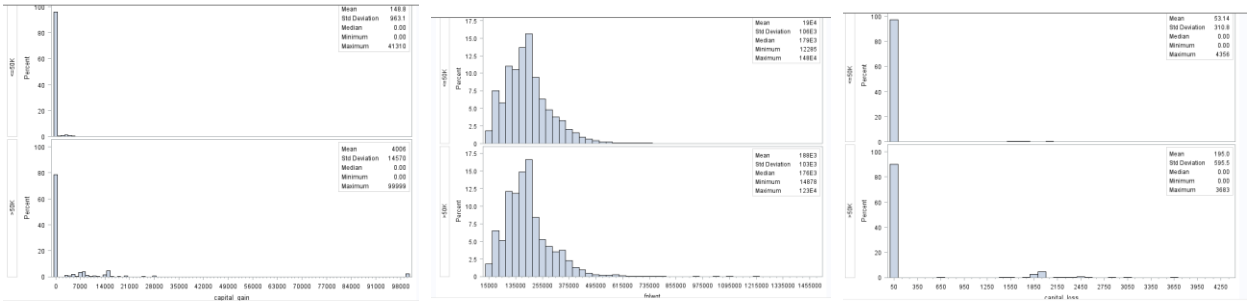

Figure 3. SAS Result for PROC UNIVARIATE

This skewness could have impact in the accuracy and performance of Income Level Prediction System. So to improve the model performance these independent variables should be transformed to be more symmetrical.

DATA TRANSFORMATION

The three variables fnwgt, capital_loss and capital_gain were transformed using Log and Square Root transformations respectively in SAS Enterprise Miner ™ [Figure 4]

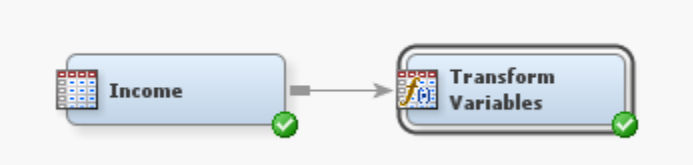

Figure 4. Data Transformation in SAS Enterprise Miner ™

Transformations Statistics

Source	Method	Variable Name	Formula	Standard Deviation	Skewness ▼	Kurtosis
Input	Original	capital_gain		7139.662	12.27204	164.457
Input	Original	capital_loss		402.3355	4.622779	20.7793
Output	Computed	LOG_capital_lo...	log(capital_los...	1.580807	4.320244	16.68782
Output	Computed	LOG_capital_g...	log(capital_gai...	2.44967	3.096213	7.790507
Input	Original	fnlwgt		104952.5	1.382399	5.736962
Output	Computed	LOG_fnlwgt	log(fnlwgt + 1)	0.630352	-0.84326	0.809473

Table4. Transformation Statistics from SAS Enterprise Miner ™

The values of the skewness and kurtosis reduced considerably, and this could significantly improve the performance of the system [Table 4].

PREDICTICE MODELING

Data partition is done prior to model building, 70% of the data is used for training and 30% for validation. Decision Tree, Neural Network, Logistic Regression, Gradient Boosting and Random Forest models were built for the income level prediction data with default settings. Alongside three other random forest models were built with different experimental settings. The most optimal settings were selected to build the best

model. [Figure 5]. Various trial and error settings were experimented on the data in order to reach the optimal settings.

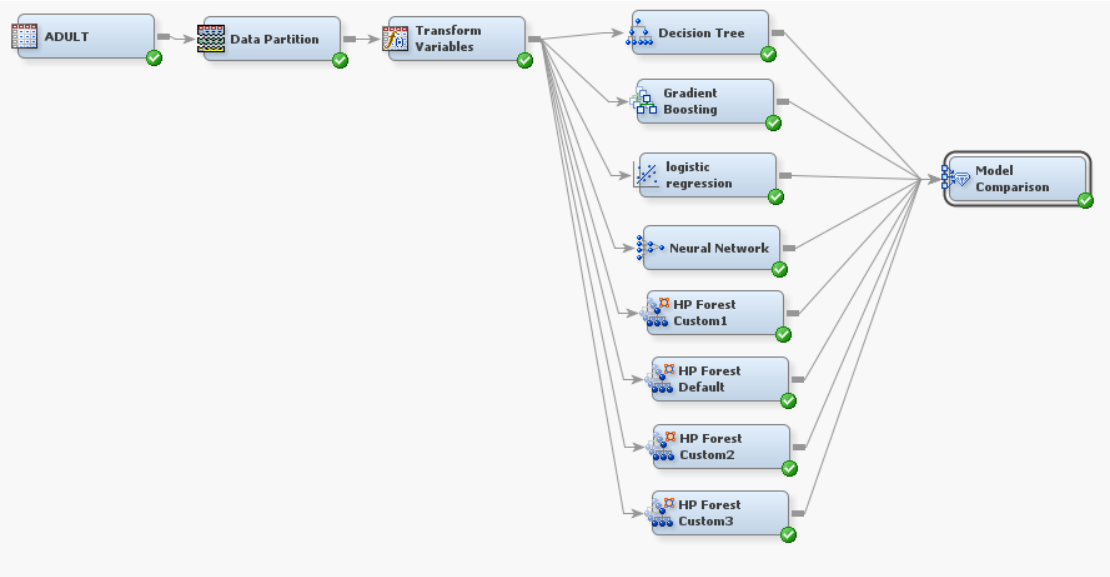

Figure 5. Models built for this study in SAS Enterprise Miner ™

GOODNESS-OF-FIT

The Model Comparison node in SAS Enterprise Miner ™ is more convenient to do the honest assessment and select the best performing model.

- In reference to Figure 6, HP Random Forest with default settings outperforms other traditional models. Further fine tuning is done to identify the model with high performance.
- Three different customized RF model settings and their corresponding misclassification rate is given below [Table 5].
- Lower the misclassification rate, the model performs better. Customized Random forest model with 75 trees, split depth 50 and number of variables for each split node being 9, excels all other models built in this study [Figure 6].

Model	No of Trees	Variables in each split node	Split node depth	Training	Validation
HP Random Forest Custom 3	75	7	50	0.131186	0.139318
HP Random Forest Custom 2	50	9	50	0.131011	0.137988
HP Random Forest Custom 1	75	9	50	0.130704	0.137783

Table5. Table with RF custom model settings

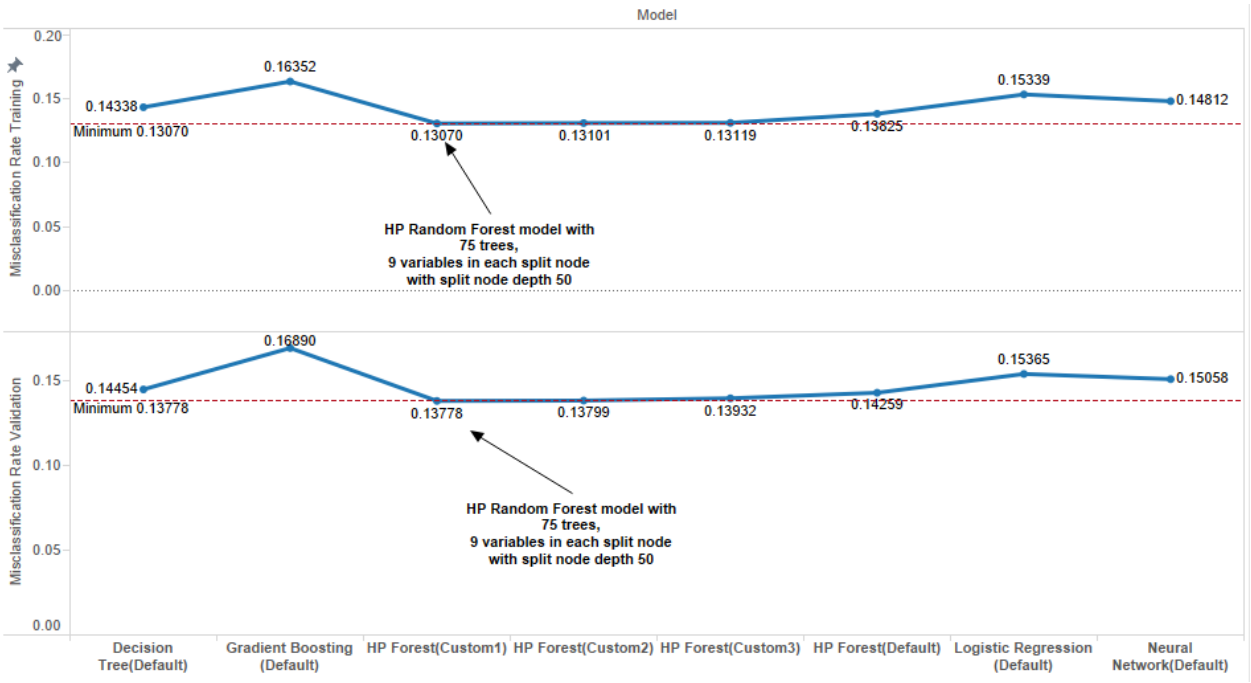

Figure 6. Plot on Misclassification Rate across models

Though random forest with default settings outstrips the performance of other models, the random forest model with customized setting scores high compared to all models, with misclassification rate being the selection criteria. HP Random Forest Custom 1 is the best-fit model for the data used in this study. The default [Table 6] and the best model [Table 7] settings are given.

Property	Value
Tree Options	
Maximum Number of Trees	50
Seed	12345
Type of Sample	Proportion
Proportion of obs in each sa	0.6
Number obs in each sample	.
Splitting Rule Options	
Maximum Depth	50
Missing Values	Use In Search
Minimum Use In Search	1
Number vars to consider in	.
Significance Level	0.05
Max Categories in Split Sea	30
Minimum Category Size	5
Exhaustive	5000
Node Options	
Method for Leaf Size	Default
Smallest percentage of obs	0.001
Smallest number of obs in n	5
Split Size	.
Score	
Variable Selection	Yes

Table 6. Default RF setting in SAS EM

Property	Value
Tree Options	
Maximum Number of Trees	75
Seed	12345
Type of Sample	Proportion
Proportion of obs in each sa	0.6
Number obs in each sample	.
Splitting Rule Options	
Maximum Depth	50
Missing Values	Use In Search
Minimum Use In Search	1
Number vars to consider in	8
Significance Level	0.05
Max Categories in Split Sea	30
Minimum Category Size	5
Exhaustive	5000
Node Options	
Method for Leaf Size	Default
Smallest percentage of obs	0.001
Smallest number of obs in n	5
Split Size	9
Score	
Variable Selection	Yes
Status	
Create Time	2/1/15 9:47 PM

Table7. Optimal RF setting for this data in SAS EM

RESULT DISCUSSION

Based on the misclassification rate iteration plot [Figure 7], after 75 trees the misclassification rate stays steady around 0.137. The default of 50 trees for Random Forest is not the optimal experimental setting that suits the data used in our study.

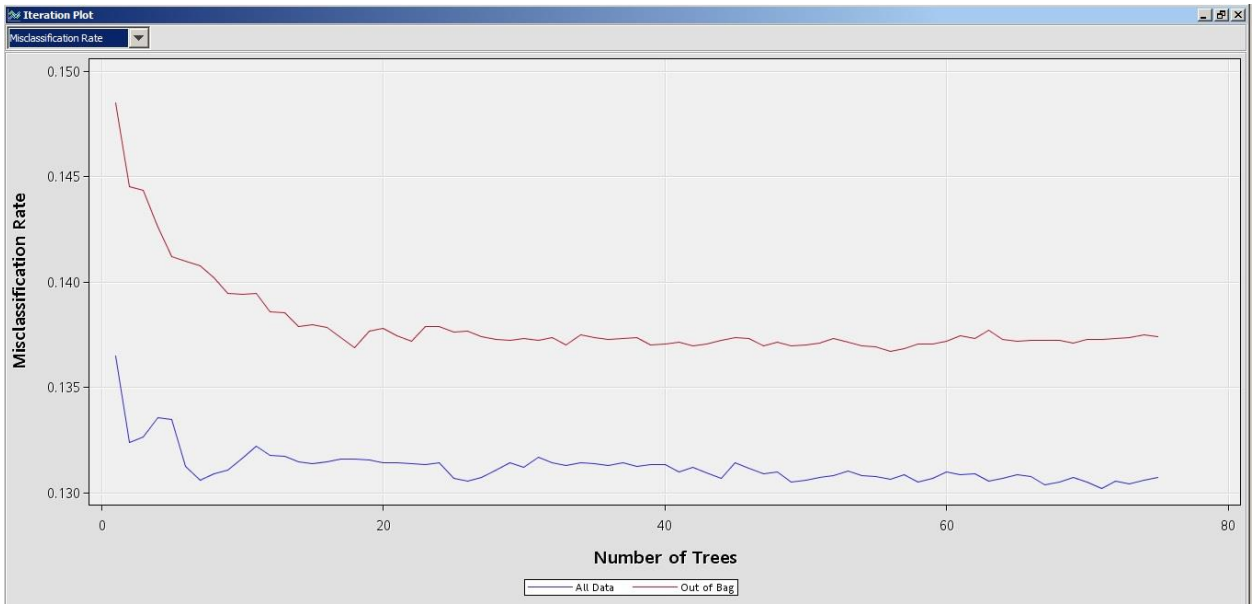

Figure 7. Misclassification Rate Iteration Plot of HP RF custom 1 model

Along with the misclassification error rate criteria, the predictive power of the binary classifier can be evaluated using the ROC chart [Figure 8].The HP Forest Custom 1 model scores high compared to the other models. Area under curve (AUC) in the ROC for the HP Forest Custom 1 model is around 0.9063 (nearly 1). The significance of this AUC is that, closer the value is to 1 the model performs better.

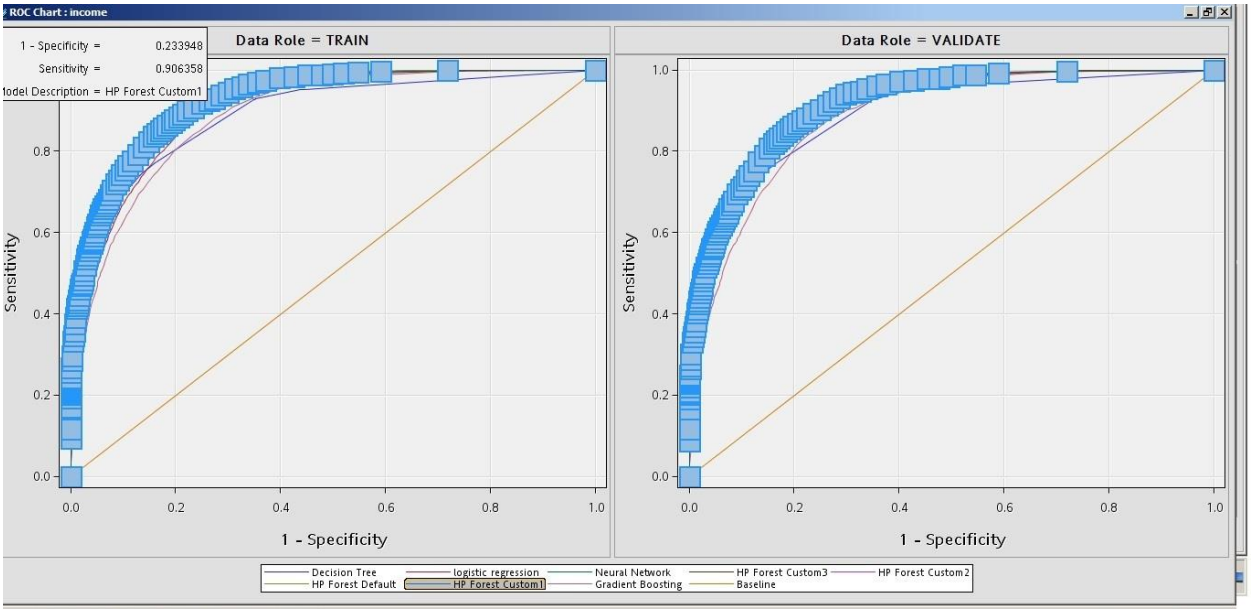

Figure 8. ROC Chart for all model built from SAS Enterprise Miner ™

Out-of-Bag data eliminates the need for separate test data and evaluates the performance of the classifier. Estimates of the variable's importance are based on the margin of cases. [9] Variables with an out-of-bag margin reduction less than or equal to zero are rejected**.**

Variable Importance

Variable Name	Number of Splitting Rules	Gini Reduction	Margin Reduction ▼	OOB Gini Reduction	OOB Margin Reduction	Label
relationship	286	0.048217	0.096434	0.03194	0.064208	
LOG_capita...	720	0.031744	0.063488	0.02065	0.041825	Transforme...
marital_stat...	306	0.024156	0.048313	0.01585	0.031828	
education_...	592	0.022477	0.044954	0.01409	0.029308	
occupation	695	0.015389	0.030778	0.00874	0.019225	
education	273	0.013064	0.026129	0.00836	0.017043	
age	976	0.006505	0.013011	0.00253	0.006766	
LOG_capita...	842	0.006317	0.012634	0.00309	0.007330	Transforme...
h_per_w	787	0.004755	0.009511	0.00166	0.004841	
workclass	524	0.002115	0.004229	0.00025	0.001782	
gender	253	0.000802	0.001603	0.00029	0.000798	
LOG_fnlwgt	239	0.000698	0.001395	-0.00035	0.000117	Transforme...
race	169	0.000260	0.000519	-0.00015	0.000057	
country	98	0.000230	0.000461	-0.00008	0.000137	

Table8. Variable Importance result from the HP RF custom 1 model

The variables: relationship, LOG_Capital_Gain, and marital status are the important predictors selected from the HP Random Forest Custom 1 model [Table 8]. Random Forest has an advantage of selecting important variables automatically. The classification table of the HP Random Forest Custom 1 model further quantifies the performance of the best classifier [Table 9]. This income-level prediction model classifies individuals with an accuracy of 86.23%, with the predictors capturing specific characteristic patterns.

Classification Table

Data Role=TRAIN Target Variable=income Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
<=50K	<=50K	88.5593	95.0644	16449	72.1701
>50K	<=50K	11.4407	38.7138	2125	9.3234
<=50K	>50K	20.2466	4.9356	854	3.7469
>50K	>50K	79.7534	61.2862	3364	14.7596

Data Role=VALIDATE Target Variable=income Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
<=50K	<=50K	88.0914	94.6474	7020	71.8600
>50K	<=50K	11.9086	40.3486	949	9.7144
<=50K	>50K	22.0556	5.3526	397	4.0639
>50K	>50K	77.9444	59.6514	1403	14.3618

Table9. Classification Table from the HP RF custom 1 model

CONCLUSION

Different statistical models, RF with default settings and RF with three different experimental settings were built to compare and contrast the high predicting power of Random Forest using SAS Enterprise Miner ™. The More Trees, the Better! Random Forest improves the accuracy of the model without over fitting the data and overcomes the limitations of Decision Trees. RF also handles unbalanced data with great efficiency. Random forest is one of the sophisticated algorithms used to solve regression and classification problem. As per literature, RF can be used for Survival Analysis, which is not in the scope of this demonstration. This could be a prospectus area for further extension of this study.

REFERENCES

[1] Leo Breiman October 2001 Volume 45, Issue 1, pp 5-32. "Random Forests - Machine Learning."

[2]Amit, Yali; Geman, Donald (1997). "'Shape quantization and recognition with randomized trees". Neural Computation."

[3]Ho, Tin Kam (1998). "The Random Subspace Method for Constructing Decision Forests." IEEE Transactions on Pattern Analysis and Machine Intelligence.

[4]Department of Statistics CMU Rebecca C. Steorts "Bagging and Random Forests" March 18 2014. http://www.stat.cmu.edu/~rsteorts/slides/slides_lecture15.pdf

[5]From Wikipedia, the free encyclopedia. "RandomForests" http://en.wikipedia.org/wiki/Random_forest

[6]Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

[7] Miguel Maldonado, Jared Dean, Wendy Czika, and Susan Haller. 2014. "Leveraging Ensemble Models in SAS® Enterprise Miner™." *Proceedings in SAS Global Forum 2014*.

[8] Leo Breiman and Adele Cutler . "Random Forests" https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[9] Geng, Ming, 2006. "A comparison of logistic regression to random forests for exploring differences in risk factors associated with stage at diagnosis between black and white colon cancer patients."

[10]Theodoro Kuolis, April 1 2003. "Random Forest: Presentation Summary"

ACKNOWLEDGMENTS

The dataset used for this study was from Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

I would like to thank Dr. Goutam Chakraborty, for his invaluable support and guidance.

RECOMMENDED READING

- *Getting Started with SAS® Enterprise Miner™ 13.1*

SAS CONTACT INFORMATION

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

Narmada Deve Panneerselvam,
Oklahoma State University, Stillwater, OK, USA
Phone: 405-385-4046
Email: narmada@okstate.edu
www.linkedin.com/pub/narmada-deve-panneerselvam