12/10/17

Ensemble Models

Wesley Clark

iamwesleyclark@gmail.com

(301) 254-9395

Introduction

The analytics team has been tasked by a bank to develop models that identify clients utilizing the subscription deposit feature. SAS Enterprise Miner (SAS, 2015) was used to develop several ensemble models to analyze the *subscribed_deposit* variable retrieved from the Bank Marketing data base (UMUC 2017). SEMMA(Sample, Explore, Modify, Model, Analyze) methodology was employed using SAS Enterprise Miner software. The aim of this analysis was to identify clients that use subscribe deposit feature. To accomplish this four different groups of models were produced. Group 1: Bagging, Boosting, HP Forest and Gradient Boosting, Group 2: Logistic Regressions(LR), Group 3: Support Vector Machines(SVM), Group 4: Neural Network(NN), Group 5: Ensemble Models(EM). An overview of each group is provided in Figure 2. These models were then analyzed and compared. This report details the development, analysis and subsequent results of each model and group.

Sample

A direct marketing campaign was conducted by a bank. In this campaign, a client was contacted in order to determine if the subscription deposit feature was being utilized. The dataset originally contained 4521 unique observations across 17 variables. Each observation reflected a unique person. Of the 17 variables, 7 were interval, 6 were nominal and 4 were binary. A detailed description of variables is presented in Figure 1.

Data Sampling: Very few of the observations used subscription deposit feature. To adjust for this, the data was appropriately sampled at a 90% sample size and a penalty function was

employed. The penalty function was weighted to 0.85, 0.15 for Decision 1 and 2. This was done because 88.548% of people did not use this feature. A penalty function increased the likelihood of a true positive, and weighted the clients that are using subscription deposit.

Category	Variable	Description
Interval Variables	**Last_contact_day** **Age** **Last_contact_duration_sec** **Number_of_contacts** **Previous_contacts** **Days_passed** **Avg__credit_balance**	Day of the month last attempted contact Age of client Duration of phone call How many contacts Count of previous contacts Number of days prior to last contact Current balance in clients primary account
Nominal Variables	**Contact_type** **Education** **Job** **Last_contact_month** **Marital_Status** **Outcome_previous_campaign**	Style of contact(cellular, telephone) Education received(basic47, university) Employment(admin, retired) Month(jan, feb) Personal Realtionship(Married, Single) Last call description(failure, success)
Binary Variables	**Has_credit_in_default** **Has_housing_loan** **Has_personal_loan** ***Subscribed_deposit***	Default description(yes, no) Housing loan existing(yes, no) Personal loan existing(yes, no) Using subscribe deposit(yes, no)

(Figure 1: Variable Description. Interval Variables, Nominal Variables and Binary Variables)

Data Partitioning: The data was partitioned for some models, and not for others. The ensemble models did not contain any direct partitioning, however used models that had already been partitioned. The SVM and LR models did employ data partitioning. Initially, partitioning was implemented at 70% training, 20% validation and 10% test sets. Evidence of overfitting was observed in the SVM models. The misclassification rate for the training data was significantly lower than validation and test sets. To adjust for this, the data partition was changed to 50%

training, 30% validation and 20% test for SVM models alone. Sensitivity among the validation set was improved as a result. The LR models were left at 70/20/10 partitioning.

Model Type	Number of Models	Group
Bagging, Boost, HP Forest, Gradient Boosting	10	1
Logistic Regression	6	2
Suppor Vector Machine	4	3
Neural Network	1	4
Ensemble Model	4	5

*(**Figure 2: Group Overview** including Number of Models and Model Type)*

Explore

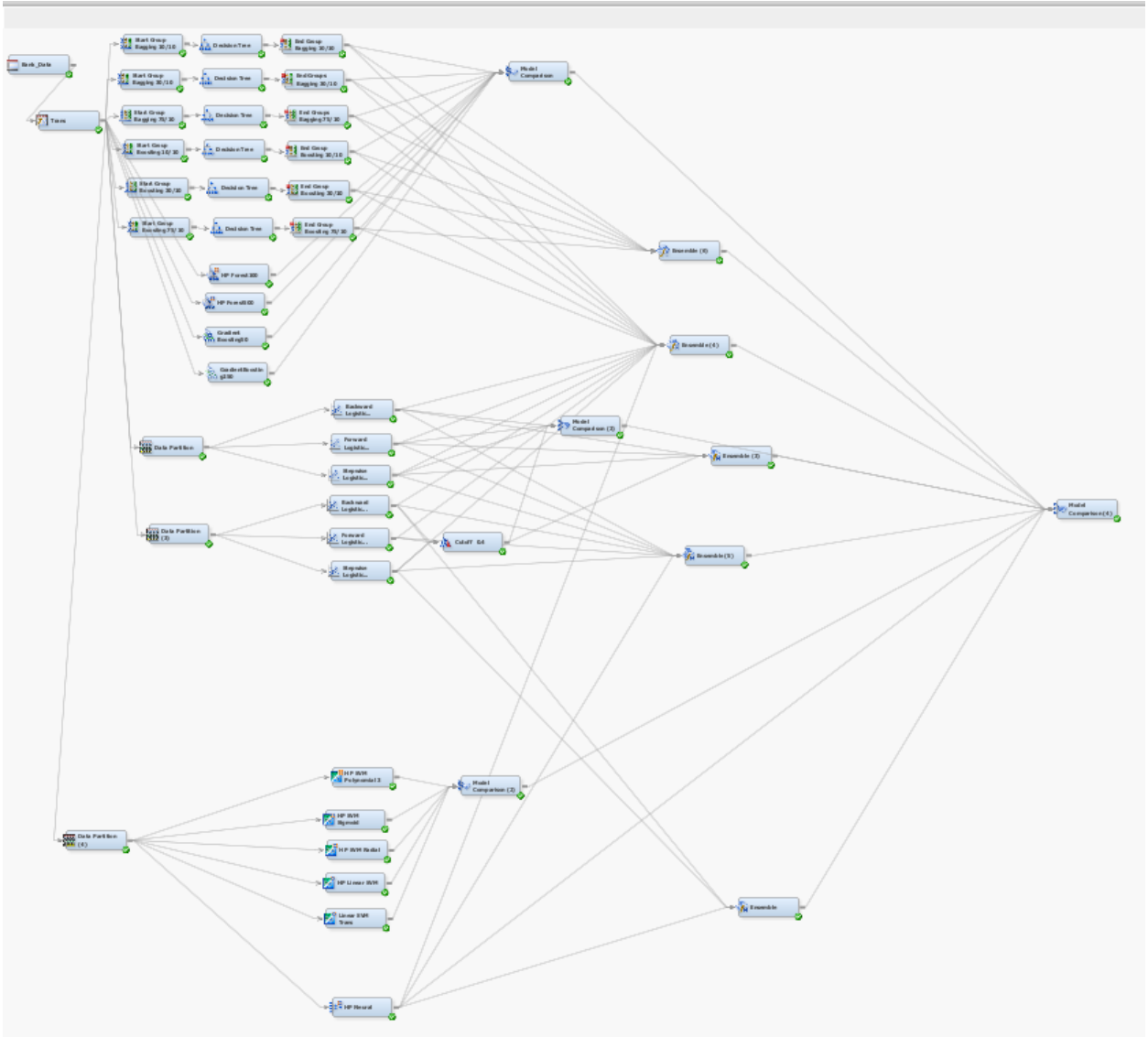
In order to understand what variables were in need of change, the data was examined. Two variables were considered for elimination, days_passed and last_contact_day. There was no year value present for either of these variables. Lacking this information made predictive insight less likely. Although a year would have been preferred, ultimately there was no evidence that predictive value was absent. It was hypothesized that seasonal correlations could be present. For this reason, both days_passed and last_contact_day were therefore included in the analysis.

Missing Data: There was no missing data present in our dataset.

Data Transformation: While the dataset did not contain any missing values, there was some skewness. To adjust for this some transformations were performed.

Modify

Of the 17 variables, 3 contained skewness that was high enough to warrant transformation. Number_of_contacts, Previous_contacts and Avg__credit_balance all contained somewhat skewed entires with skewness values greater than 5. These three variables were transformed and



the skewness was reduced to less than 1. The rest of the variables did not require any transformation.

Model

(Figure 3: Model representation with nodes. Enlarged view of each group available in the appendix)

In total 25 models were constructed. 10 models utilized Bagging, Boosting and HPForest. Of these, 3 models were bagging models with index counts of 10, 30 and 75. 3 models were Boosting models, which also contained index counts of 10, 30 and 75. 2 models were gradient boosting models and contained either 50 or 250 iterations. Additionally 2 HP Forest models contained iteration counts of 100 and 500 respectively were made. These models comprised Group 1. Group 2 was made up of 6 Logistic regressions. 3 of these implemented a cutoff set to 0.4. Group 3 contained 5 SVM's. 1 HP SVM using a Sigmoid kernel, 1 HP SVM utilizing the Polynomial Kernel with Polynomial Degree set to 3, 1 HP SVM with Radial Basis Function kernel and 1 HP Linear SVM set to Interior Point. Group 4 consisted of 1 Neural Network with model selection criterion set to misclassification. The makeup of each model is listed in Figure 4.

Assess

There was no single model that performed the best across the board. For this model to be useful to the bank, it must first off predict who will use the subscribe deposit feature, as well as achieve a level of success in predicting who will not use it. If we wanted to build a model that had 100% success rate at predicting who would use the subscribe deposit without placing any weight on who would not, we could instead just predict that everyone use the subscribe deposit. Such a prediction would be completely useless to the bank. One application of this is marketing

Model	False Negative	Ture Negative	False Positive	True Positive	Total Incorrect Classifications	Percentage YES YES	Lift	Misclassification Rate	ROC Index
Bagging 10/10	409	3566	37	57	446		4.2194	10.9609%	0.866
Bagging 30/10	413	3568	35	53	448	1.3025	3.8712	11.0101%	0.892
Bagging 75/10	398	3552	51	68	449	1.671%	4.0422	11.0347%	0.898
Boosting 10/10	103	2357	1246	363	1349	8.921%	6.1510	33.1531%	0.98
Boosting 30/10	214	2984	619	252	833	6.193%	8.3878	20.472%	0.998%
Boosting 75/10	27	1502	2101	439	2128	10.789%	8.7320	52.230%	0.99
Gradient Boosting50	436	3587	16	30	—	0.737	3.7153	11.1084%	0.882
Gradient Boosting250	348	3526	67	118	—	2.900%	4.12930	10.0991%	0.913
HP Forest100	410	3589	14	56	424	1.302%	4.5501	10.4203%	0.928
HP Forest500	410	3588	15	56	425	1.376%	4.3444	10.4448%	0.928
SVM Polynomial 3	53	667	54	40	107	4.913%	2.7754	13.145	0.78
HP SVM Sigmoid	78	652	69	15	147	2.417%	0.08539	13.319	0.586
HP SVM Radial	84	712	9	9	93	4.312%	1.707842	11.4251%	0.772
HP Linear	81	709	12	12	101	2.212%	5.1235	11.4251%	0.871
Backward LR	66	703	18	27	94	3.523%	4.2311	10.314%	0.895
Forward LR	66	703	18	27	94	$3.523	4.2311	10.314%	0.895
Stepwise LR	66	703	18	27	94	3.523%	4.2311	10.314%	0.895
Backward LR Cutoff	99	1045	36	75	135	$2.991	4.319	11.393%	0.893
EM LR+NN	62	702	19	31	81	3.8099	—	10.00%	—
EM Bag+Boost+ LR+NN	301	3560	43	165	344	4.051%	—	8.00%	—
EM LR	63	701	21	29	83	3.5233	—	11.00%	—
EM Bag+Boost	308	3582	21	158	329	3.8831	—	8.00%	-
EM Bag+Boost+LR	63	706	15	30	407	3.1212	—	11.00%	—
NN	247	3530	73	219	320	5.382%	—	9.0%	—
EM Bag+Boost+LR+NN	62	707	14	31	76	3.2013%	-	10.89%	—

the subscribe deposit feature. We are most interested in marketing it to people that will likely

use it, but we are also interested in saving money by not spending marketing dollars on people who will not use subscribe deposit.

(Figure 4: Model Comparison. Each model compared among several categories)

The models were all compared across False Negative, True Negative, False Positive, True Positive, Total Incorrect Classifications(misclassifications), Lift, Misclassification Rate and ROC Index. A view of Lift and ROC Curve compared can be seen in in Figure 5 and 6 in the appendix. Lift can be said to be the measure of how much better a model is compared to having no model at all (LC, 2017). Although the SVM models had very good lift values, the misclassification rate and percent YES YES were mostly off the mark. The Boosting 75/10 model was excellent at predicting clients that would use the subscription deposit feature. It achieved by far the highest percent, at 10.871. Unfortunately this model also achieved the highest misclassification rate at 52%. Because the misclassification rate was so inordinately high, this model could not be considered a winner. A view at Figure 4 is quick to show that each group performed differently. A tradeoff between sensitivity and specificity (PSU 2013) will lead to the best model. A key takeaway from this investigation was that of all the groups, the EM's performed especially among all facets. They were well rounded. Of the all the groups, the single best model was the model that incorporated Group 1, 2 and 4: Ensemble Model Bag+Boost+LR+NN. This model performed remarkably well and achieved a very low misclassification rate of 8%. Additionally it achieved a YES YES value of 4.051%. This is indicative of a reasonably high predictive value of clients that will use the subscribe deposit feature. A view of the classification chart for this model is available in figure 7 of the appendix.

This model is the recommended model to use going forward. One way to improve this model could be to add more types of analysis, such as a decision tree, to the champion EM and potentially increase performance even more.

References

Lift Charts. (n.d.). Retrieved November 16, 2017, from https://www3.nd.edu/~busiforc/handouts/DataMining/Lift%20Charts.html

Receiver Operating Characteristic Curve (ROC). Retrieved November 19, 2017, from https://onlinecourses.science.psu.edu/stat504/node/163

SAS Institute Inc. Getting Started with SAS Enterprise Miner 14.1 Cary, NC: SAS Institute Inc. Available from https://www.sas.com/en_id/software/studio.htm

University of Maryland University College (2017) bank marketing campaign.csv [Data file] Retrieved from https://learn.umuc.edu/d2l/le/content/249259/Home

Appendix

Figure 1: **Variable Description**. Interval Variables, Nominal Variables and Binary Variables

Category	Variable	Description
Interval Variables	**Last_contact_day** **Age** **Last_contact_duration_sec** **Number_of_contacts** **Previous_contacts** **Days_passed** **Avg__credit_balance**	Day of the month last attempted contact Age of client Duration of phone call How many contacts Count of previous contacts Number of days prior to last contact Current balance in clients primary account
Nominal Variables	**Contact_type** **Education** **Job** **Last_contact_month** **Marital_Status** **Outcome_previous_campaign**	Style of contact(cellular, telephone) Education received(basic47, university) Employment(admin, retired) Month(jan, feb) Personal Realtionship(Married, Single) Last call description(failure, success)
Binary Variables	**Has_credit_in_default** **Has_housing_loan** **Has_personal_loan** ***Subscribed_deposit***	Default description(yes, no) Housing loan existing(yes, no) Personal loan existing(yes, no) Using subscribe deposit(yes, no)

Figure 2: Group Overview: A description of each group, including Number of Models and Model Type

Model Type	Number of Models	Group
Bagging, Boost, HP Forest, Gradient Boosting	10	1
Logistic Regression	6	2
Suppor Vector Machine	4	3
Neural Network	1	4
Ensemble Model	4	5

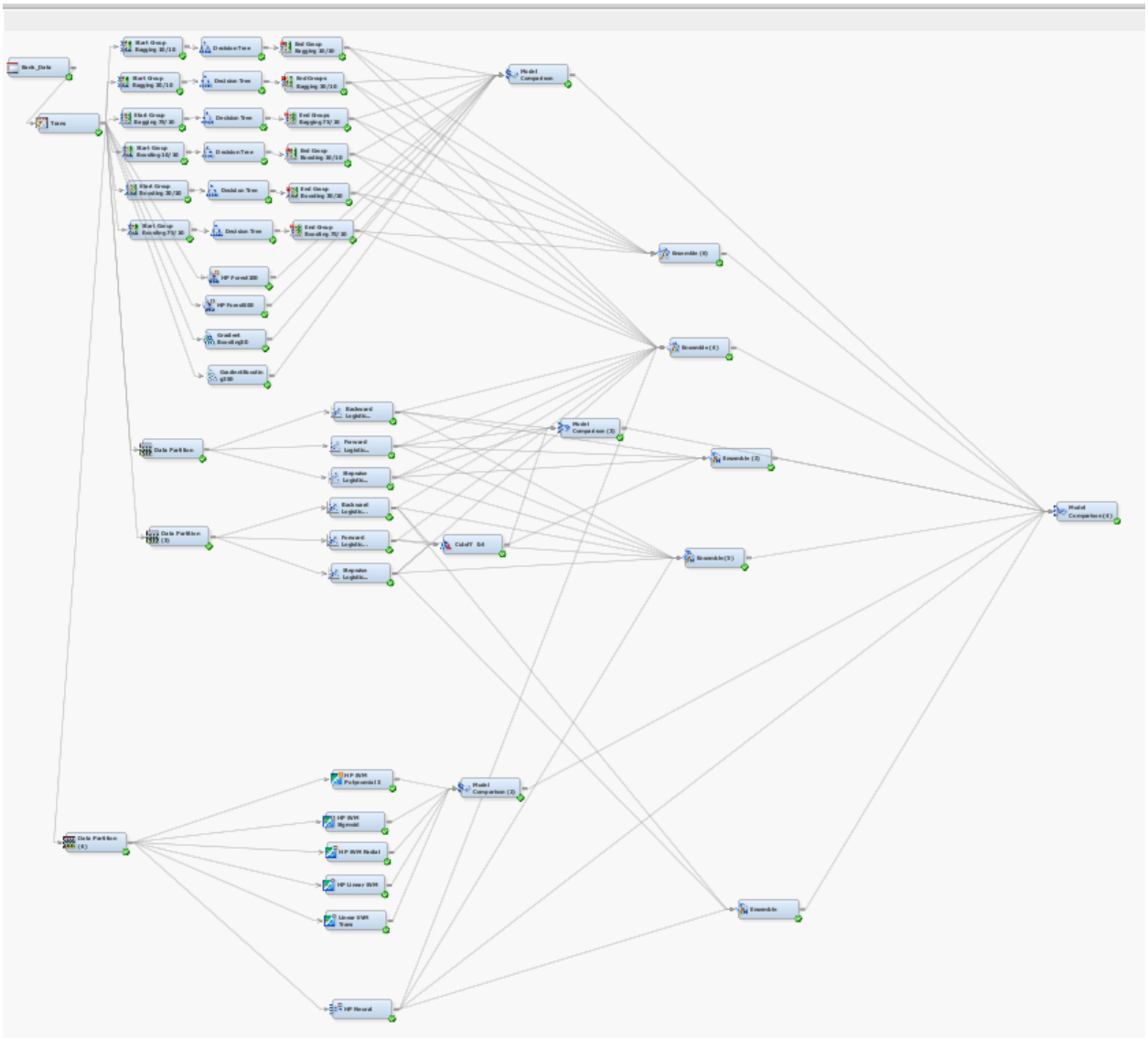
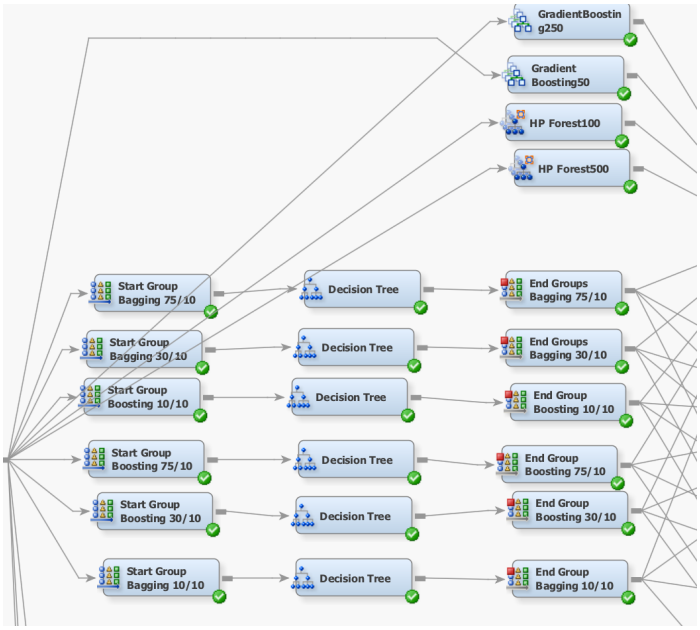


Figure 3: Model Overview: A map of all the models made. Enlarged view of each group is provided below.

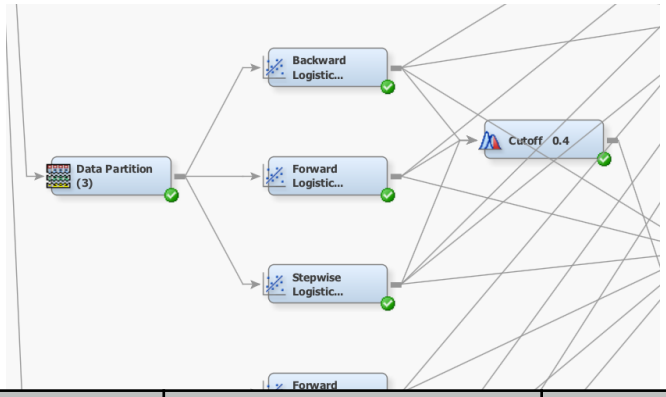
Group 1: Bagging Boosting, HP Forest and Gradient Boosting



Model Name	Type	Index Count	Minimum Group Size	Maximum Number of Trees	Iterations
Bagging 10/10	**Bagging**	10	10	N/A	N/A
Bagging 30/10	**Bagging**	30	10	N/A	N/A
Bagging 75/10	**Bagging**	75	10	N/A	N/A
Boosting 10/10	**Boosting**	10	10	N/A	N/A
Boosting 30/10	**Boosting**	30	10	N/A	N/A

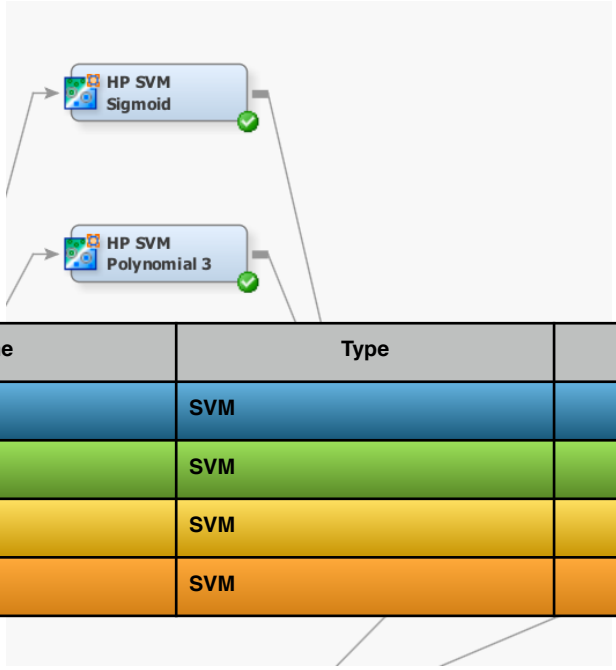
Model Name	Type	Index Count	Minimum Group Size	Maximum Number of Trees	Iterations
Boosting 75/10	**Boosting**	75	10	N/A	N/A
Gradient Boosting50	**Boosting**	N/A	N/A	N/A	50
Gradient Boosting250	**Boosting**	N/A	N/A	N/A	250
HP Forest100	**HP Forest**	N/A	N/A	100	N/A
HP Forest500	**HP Forest**	N/A	N/A	500	N/A

Group 2: Logistic Regressions



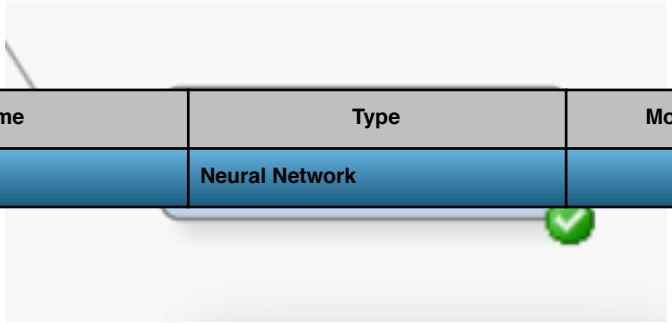
Model Name	Regression Type	Cutoff
Backward Logistic Regression 1	**Backward**	None
Forward Logistic Regression 1	**Forward**	None
Stepwise Logistic Regression 1	**Stepwise**	None
Backward Logistic Regression 2	**Backward**	0.3
Forward Logistic Regression 2	**Forward**	0.3
Stepwise Logistic Regression 2	**Stepwise**	0.3

Group 3: Support Vector Machines



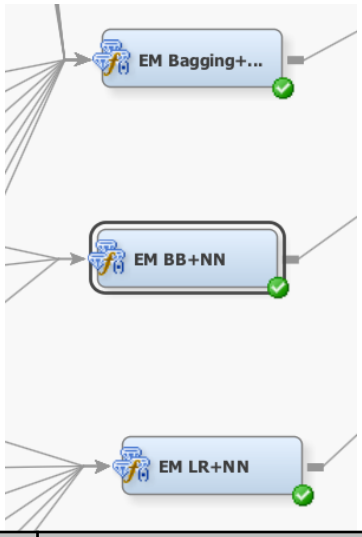
Model Name	Type	Kernel
HP SVM Sigmoid	**SVM**	Sigmoid
HP SVM Polynomial 3	**SVM**	Polynomial
HP SVM Radial	**SVM**	Radial
HP Linear SVM	**SVM**	None

Group 4: Neural Network



Model Name	Type	Model Selection Criterion
Neural Network	**Neural Network**	Misclassification

Group 5: Ensemble Models



Model Name	Type	Makeup
EM Bagging+Boosting+LR+NN	**Ensemble Model**	Bagging Boosting Logistic Regression Neural Network
EM BB + NN	**Ensemble Model**	Bagging Boosting Neural Network
EM LR + NN	**Ensemble Model**	Logistic Regression Neural Network
EM LR + BB	**Ensemble Model**	Logistic Regression Bagging and Boosting

Model	False Negative	Ture Negative	False Positive	True Positive	Total Incorrect Classifications	Percentage YES YES	Lift	Misclassification Rate	ROC Index
Bagging 10/10	409	3566	37	57	446		4.2194	10.9609%	0.866
Bagging 30/10	413	3568	35	53	448	1.3025	3.8712	11.0101%	0.892
Bagging 75/10	398	3552	51	68	449	1.671%	4.0422	11.0347%	0.898
Boosting 10/10	103	2357	1246	363	1349	8.921%	6.1510	33.1531%	0.98
Boosting 30/10	214	2984	619	252	833	6.193%	8.3878	20.472%	0.998%
Boosting 75/10	27	1502	2101	439	2128	10.789%	8.7320	52.230%	0.99
Gradient Boosting50	436	3587	16	30	—	0.737	3.7153	11.1084%	0.882
Gradient Boosting250	348	3526	67	118	—	2.900%	4.12930	10.0991%	0.913
HP Forest100	410	3589	14	56	424	1.302%	4.5501	10.4203%	0.928
HP Forest500	410	3588	15	56	425	1.376%	4.3444	10.4448%	0.928
SVM Polynomial 3	53	667	54	40	107	4.913%	2.7754	13.145	0.78
HP SVM Sigmoid	78	652	69	15	147	2.417%	0.08539	13.319	0.586
HP SVM Radial	84	712	9	9	93	4.312%	1.707842	11.4251%	0.772
HP Linear	81	709	12	12	101	2.212%	5.1235	11.4251%	0.871
Backward LR	66	703	18	27	94	3.523%	4.2311	10.314%	0.895
Forward LR	66	703	18	27	94	$3.523	4.2311	10.314%	0.895
Stepwise LR	66	703	18	27	94	3.523%	4.2311	10.314%	0.895
Backward LR Cutoff	99	1045	36	75	135	$2.991	4.319	11.393%	0.893
EM LR+NN	62	702	19	31	81	3.8099	—	10.00%	—
EM Bag+Boost+ LR+NN	301	3560	43	165	344	4.051%	—	8.00%	—
EM LR	63	701	21	29	83	3.5233	—	11.00%	—
EM Bag+Boost	308	3582	21	158	329	3.8831	—	8.00%	-
EM Bag+Boost+LR	63	706	15	30	407	3.1212	—	11.00%	—
NN	247	3530	73	219	320	5.382%	—	9.0%	—
EM Bag+Boost+LR+NN	62	707	14	31	76	3.2013%	-	10.89%	—

Figure 4: Model Comparison. Each model compared

Figure 5: Lift Comparison among highest scoring models

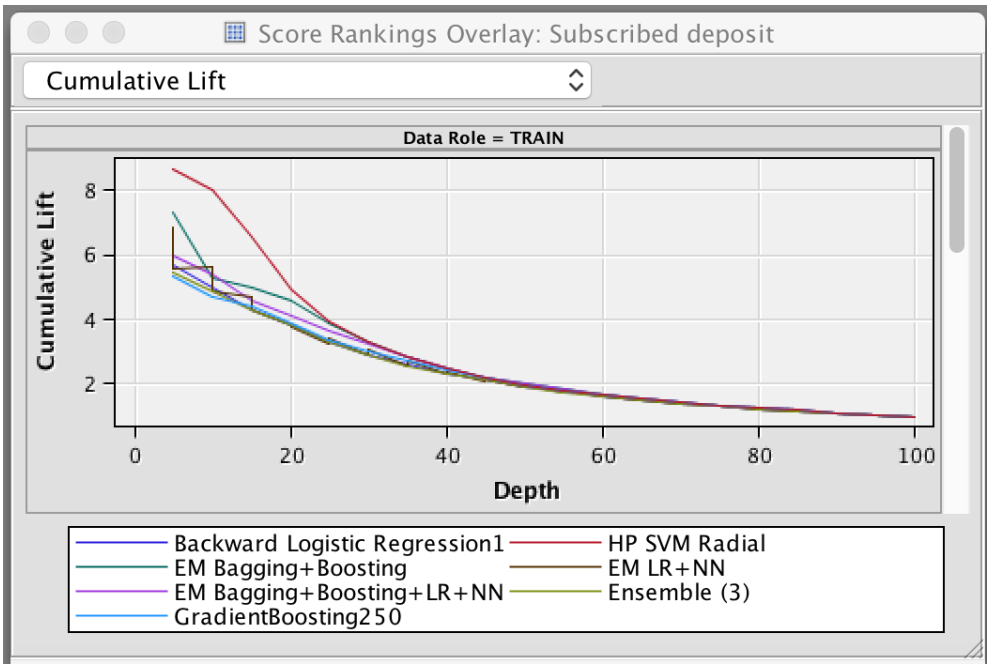
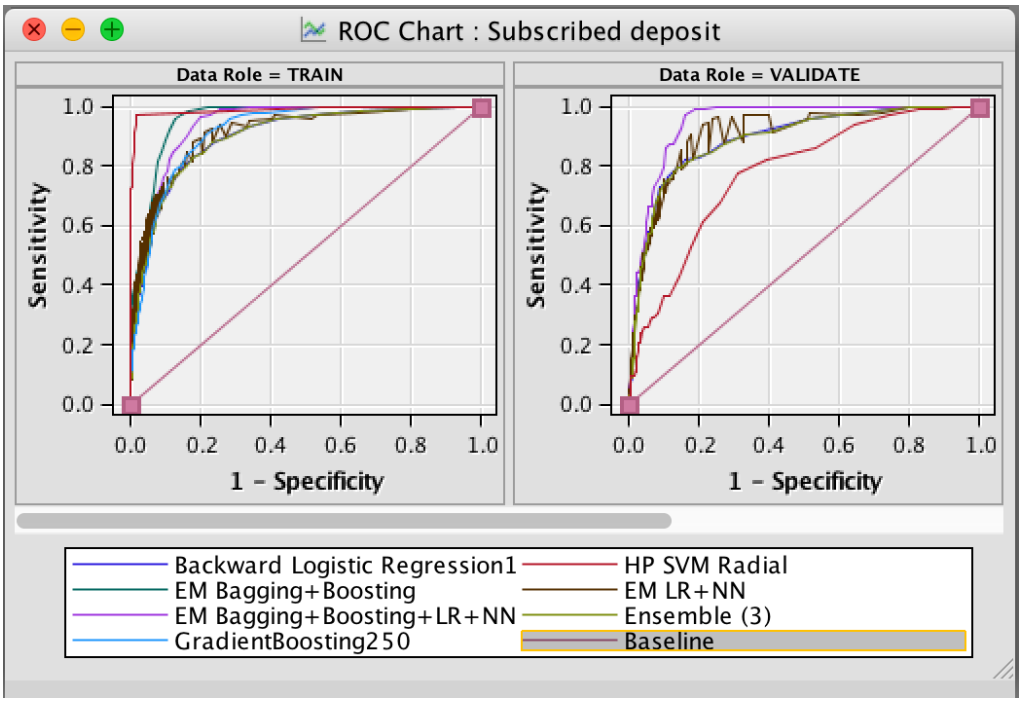


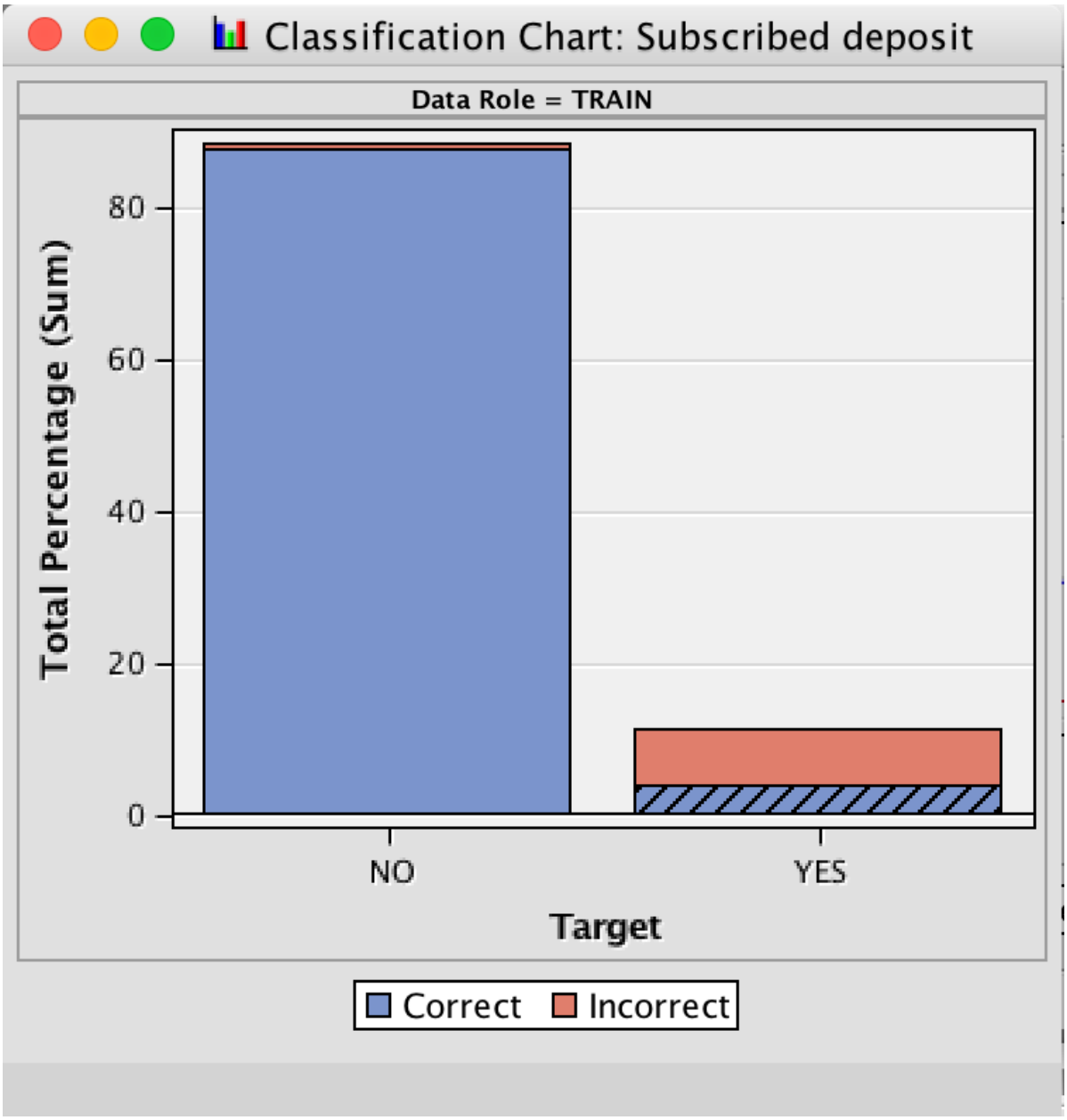
Figure 6: ROC Comparison among highest scoring models

Figure



7:

Classification Chart of Champion EM - Ensemble Model Bag+Boost+LR+NN


Classification Chart: Subscribed deposit
Data Role = TRAIN
Total Percentage (Sum)
80
60
40
20
0
NO
YES
Target
Correct
Incorrect