Title Page

Assignment 3: Ensemble II Models using SAS Enterprise Miner

Melissa Hunfalvay

Email: Melissa.Hunfalvay@gmail.com

Data 640 9040

Spring 2022

Professor Steve Knode

Date: February 23th, 2022

Introduction

The dataset chosen was “car lemon dataset.csv” (Figure 1). The data included various features of cars (such as color and make) and whether the cars were determined to be good or bad purchases. The dataset included 34 variables and 72,983 rows or observations. The type of variables included numeric (n = 15) and character (n = 19) variables. Of the numeric variables, 11 were interval, the other two were binary.

The purpose of the analysis was to identify the bad vehicle purchases, from good vehicle purchases. Ensemble models are “models of models” based on the concept that large numbers of models are more accurate than any one model (Surowiecki, 2005).

“Is Bad Buy” was the target binary variable (“0” = not a bad buy, “1” = bad buy). The target variable was unevenly proportioned and skewed heavily in favor of not bad purchases (n = 64,007; 87.70%). Bad purchases totaled 8,976 of the 72,983 observations (12.30%). The imbalance percentage is 75.4%.

Data Cleaning and Preparation

Data cleaning and preparation tasks are outlined in Table 1 for the various heterogeneous models created. This includes Support Vector Machines (SVM), different forms of decision trees, and regression, neural networks, and Naïve Bayesian. The varying models require different forms of cleaning and preparation prior to generating results. Decision trees require the least amount of data preparation compared with the other models examined thus far.

Table 1: Summary of data steps for each model type

Cleaning Task	Support Vector Machines	Decision Tree Models	Regression, Neural Networks, Naïve Bayes
Accept SAS rejected variables?	Yes	Yes	Yes
Reject ID variables?	Yes	Yes	Yes
Impute missing values?	Yes, default methods used	No, not needed for decision tree models (Lindoff & Berry, 2011)	Yes, default methods used
Adjust outliers?	Yes, 3 standard deviations from the mean	No, not needed for decision tree models (Hastie, Tibshirani, Friedman, 2009)	Yes, 3 standard deviations from the mean
Transform skewness?	Yes, Warrenty Cost Log 10 transformation	No, not needed for decision tree models (Lindoff & Berry, 2011)	Yes, Warrenty Cost Log 10 transformation
Correlated variables?	Yes, remove vehicle year	Yes, remove vehicle year	Yes, remove vehicle year
Correlation matrix?	No	Yes, as ensemble methods require independence of input (Rokach, 2012).	Yes, in order to strengthen the unquiness of each input variable for weighting purposes (Sarma, 2013).
Transform Interval variables?	Yes, change interval variables maximum normal	No, not needed for decision tree models (Lindoff & Berry, 2011)	Yes, change interval variables maximum normal
Transform Class variables?	Yes, change class variables to dummy indicators	No, not needed for decision tree models (Lindoff & Berry, 2011)	Yes, change class variables to dummy indicators

Predictive Model Development

Using the Ensemble Node four different input model variations were developed (SAS Enterprise Miner (n.d.). For example, “The Champions” model used the prior champion RBF-Model 9, as input to the Ensemble Node (see Table 2). If new algorithms were inserted, various parameters were experimented with to find an ideal model input. For instance, in the regression, the input parameters of regression type (logistic) and model selection (stepwise, forward etc.) were iterated over to find the lowest misclassification rate prior to running the Ensemble node.

Table 2: Ensemble Node “Models of Models”

Model Name	Ensemble Model Concept	Algorithms Included	Algorithm Parameters
New Classifier	This model included classification algorthims not used so far in this course in order to determine how they stacked up against previous models.	Decision Tree (simple)	Custom significance level 0.1
		Regression	Type: Logistic. Model selection: stepwise
		Neural Networks	Default settings
		Niave Bayes via HP BN Classifer	Custom significance level 0.1
The Champions	This model used prior "champion" model along with the New Classifer model as inputs into the ensemble node to see if together they formed an better outcome.	Added RBF Champion model	Cutoff threshold 0.24 (custom – best fit). RBF Degree: 3
Rocking the Boat	This model mixes up new models with a poor performing decision tree (bagging) model based on the premise that the ensemble model can be more accurate than the individual models only if the indivdual models disagree with one another (Knode, 2022).	Regression	Type: Logistic. Model selection: stepwise
		Neural Networks	Default settings
		Niave Bayes via HP BN Classifer	Custom significance level 0.1
		Bagging (Bag5)	Custom significance level 0.1
The Individuals	Includes each of the individual model results using the model comparison node to include past "champions" and "best in class" to determine if any of these models alone beat the Ensemble Node results.	Logistical Regression	As decribed above
		Neural Networks	As decribed above
		Niave Bayes	As decribed above
		SVM RBF (model 9)	As decribed above
		Gradient boost (Gradient7)	1000 trees, 0.5 cutoff, tree depth 5
		Bagging (Bag5)	As decribed above
		Boosting (Boost3)	20 trees, 0.5 cut off, 0.13 splitting criterion, max tree depth 6, Splitting rule: nominal variables = GINI index
		Decision Tree (simple)	As decribed above
		Random Forest (Forest6)	200 trees, 0.5 cut off criterion, 0.05 sig level, tree depth max 50

Ensemble Node parameters were then varied to include average, maximum and voting methods for the heterogenous new model (see Table 3, Knode 2022). Finally, the three Ensemble Node models and each individual model were compared to one another to determine the overall "Champion" model.

Table 3: Examination method for each Ensemble Model

Method of Examination	Method Explained
Average Method	Averages the posterior probabilties for the class variable
Maximum Method	Takes the maximum of the posterior probabilities for the class variable
Voting -Average	Averages the posterior probabilities for class varaibles from the models which are in the majority group
Voting - Proportion	The posterior probabilities are calculated as a ratio of the number of models in the majority group and the total number of models included

Selection criteria for each model was set to Validation: Misclassification rate. Data was partitioned 70% training, 30% validation for each model.

Imbalanced Target

The binary target variable "*Is Bad Buy*" was imbalanced. The type of model being created determined if an adjustment was needed for the imbalanced target variable. For instance, decision trees type models showed no change in results (from assignment 2; Abbott, 2014) when using the Cutoff Node. In contrast, SVM models require an adjustment for this imbalance (Linoff & Berry, 2011). Therefore, where appropriate, two methods were used to adjust for the imbalanced target variable. A cutoff criterion of 0.13, which was reflective of the proportion of rare events in the dataset, was implemented using the Cutoff Node. The cost function was also calculated to determine overall cost of each model. Proportions of cost for false negatives was set to $1 and false positives to $6.90. The True results cost was $0 (see Figure 2).

Accuracy Measures and Results

The purpose of this assessment was to identify rare (positive or "1") events, which are bad buys. Therefore, models were assessed using the following order of importance to achieve this goal. First, true positives, second, sensitivity, third, precision, fourth, F1 score, fifth, accuracy (see Tables 4 to 8; Figures 3 to 6).

Table 4: New Classifier Model Results from each Ensemble Node Evaluation Criteria

Model Characteristics		*Evaluation Criteria*																*Notes*	*Conclusion*
Method Used	**Dataset**	**FN #**	**FN %**	**TN#**	**TN %**	**FP #**	**FP %**	**TP #**	**TP%**	**Sensitivity %**	**Precision %**	**F1 Score**	**Accuracy %**	**Cost $**	**Specificity %**	**Misclassification Rate**	**Lift**		
Average	Training	240	76.43	2213	98.84	26	1.16	74	23.57	23.57	74.00	35.75	89.58	$1,682	98.84	10.00%	5.71	Low TP values, not identifying enough bad buys. Sensitivity low. Accuracy rate good. Precision good.	Reject
Average	Validation	107	79.26	951	98.96	10	1.04	28	20.74	20.74	73.68	32.37	89.32	$748	98.96	11.00%	4.58	No overfitting.	
Maximum	Training	232	73.89	2206	98.53	33	1.47	82	26.11	26.11	71.30	38.23	89.62	$1,634	98.53	10.00%	5.55	A more balanced model compared with the average method. Still not identifying enough bad buys.	Accept: Best in class for "New Classifier"
Maximum	Validation	104	77.04	994	98.32	17	1.68	31	22.96	22.96	64.58	33.88	89.44	$735	98.32	11.00%	4.59	A slightly overfit model w/ 4% different in sensitivity from training to validate, though accuracy remained the same.	
Voting Average	Training	240	76.43	2213	98.84	26	1.16	74	23.57	23.57	74.00	35.75	89.58	$1,682	98.84	10.00%	5.57	Voting average same as average and voting proportion results	Reject
Voting Average	Validation	107	79.26	951	98.96	10	1.04	28	20.74	20.74	73.68	32.37	89.32	$748	98.96	11.00%	4.89	No overfitting.	
Voting Proportion	Training	240	76.43	2213	98.84	26	1.16	74	23.57	23.57	74.00	35.75	89.58	$1,682	98.84	10.00%	5.29	Proportion showed no differences with the average voting method.	Reject
Voting Proportion	Validation	107	79.26	951	98.96	10	1.04	28	20.74	20.74	73.68	32.37	89.32	$748	98.96	11.00%	4.68	No overfitting.	

Table 5: The Champions Model Results from Each Ensemble Node Evaluation Criteria

Model Characteristics		*Evaluation Criteria*																*Notes*	*Conclusion*
Method Used	**Dataset**	**FN #**	**FN %**	**TN#**	**TN %**	**FP #**	**FP %**	**TP #**	**TP%**	**Sensitivity %**	**Precision %**	**F1 Score**	**Accuracy %**	**Cost $**	**Specificity %**	**Misclassification Rate**	**Lift**	*Overall, adding the RBF champion model did not improve the sensitivity or TP rate much from the "New Classifier" model.*	
Average	Training	240	76.43	2213	98.84	26	1.16	74	23.57	23.57	74.00	35.75	89.58	$1,682	98.84	10.00%	5.71	Low TP values, not identifying enough bad buys. Sensitivity low. Accuracy rate good. Precision good.	Reject
Average	Validation	107	79.26	950	98.86	11	1.14	28	20.74	20.74	71.79	32.18	89.23	$749	98.86	11.00%	4.57	No overfitting.	
Maximum	Training	232	73.89	2206	98.53	33	1.47	82	26.11	26.11	71.30	38.23	89.62	$1,634	98.53	10.00%	5.55	Best TP capture and sensitivity.	Accept: Best in class for "The Champion"
Maximum	Validation	104	77.04	944	98.23	17	1.77	31	22.96	22.96	64.58	33.88	88.96	$735	98.23	11.00%	4.59	No overfitting.	
Voting Average	Training	241	76.75	2215	98.93	24	1.07	73	23.25	23.25	75.26	35.52	89.62	$1,687	98.93	10.00%	5.59	Voting average same as average and voting proportion results	Reject
Voting Average	Validation	107	79.26	951	98.96	10	1.04	28	20.74	20.74	73.68	32.37	89.32	$748	98.96	11.00%	4.43	No overfitting.	
Voting Proportion	Training	241	76.75	2215	98.93	24	1.07	73	23.25	23.25	75.26	35.52	89.62	$1,687	98.93	10.00%	5.28	Proportion showed no differences, except in lift values, with the average voting method.	Reject
Voting Proportion	Validation	107	79.26	951	98.96	10	1.04	28	20.74	20.74	73.68	32.37	89.32	$748	98.96	11.00%	4.67	No overfitting.	

Table 6: Rocking the Boat Model Results from Each Ensemble Node Evaluation Criteria

Model Characteristics		*Evaluation Criteria*																*Notes*	*Conclusion*
Method Used	**Dataset**	**FN #**	**FN %**	**TN#**	**TN %**	**FP #**	**FP %**	**TP #**	**TP%**	**Sensitivity %**	**Precision %**	**F1 Score**	**Accuracy %**	**Cost $**	**Specificity %**	**Misclassification Rate**	**Lift**	*As the prior two models didn't create much change, this model adds a poor performing decision tree (bagging) model based on the premise that the ensemble model needed individual models with more disagreement (Knode, 2022).*	
Average	Training	240	76.43	2213	98.84	26	1.16	74	23.57	23.57	74.00	35.75	89.58	$1,682	98.84	10.00%	5.40	Did not identify enough TP. Highest cost.	Reject
Average	Validation	107	79.26	950	98.86	11	1.14	28	20.74	20.74	71.79	32.18	89.23	$749	98.86	11.00%	4.58	No overfitting.	
Maximum	Training	227	72.29	2179	97.32	60	2.68	87	27.71	27.71	59.18	37.74	88.76	$1,626	97.32	11.00%	5.09	This model identified the highest number of true positives and highest sensitivity. Precision moderate. Accuracy good.	Overall "Champion" Ensemble Node model
Maximum	Validation	103	76.30	939	97.71	22	2.29	32	23.7	23.70	59.26	33.86	88.59	$733	97.71	11.00%	4.73	No overfitting.	
Voting Average	Training	238	75.80	2213	98.84	26	1.16	76	24.2	24.20	74.51	36.54	89.66	$1,668	98.84	10.00%	5.58	Did not identify enough TP. Highest lift.	Reject
Voting Average	Validation	107	79.26	950	98.86	11	1.14	28	20.74	20.74	71.79	32.18	89.23	$749	98.86	11.00%	4.58	No overfitting.	
Voting Proportion	Training	238	75.80	2213	98.84	26	1.16	76	24.2	24.20	74.51	36.54	89.66	$1,668	98.84	10.00%	5.23	Did not identify enough TP.	Reject
Voting Proportion	Validation	107	79.26	950	98.86	11	1.14	28	20.74	20.74	71.79	32.18	89.23	$749	98.86	11.00%	4.74	No overfitting.	

Table 7: The Individual Models Results from Each Ensemble Node Evaluation Criteria

Model Characteristics		*Evaluation Criteria*													*Notes*	*Conclusion*
Method Used	**Dataset**	**FN #**	**FN %**	**TN#**	**TN %**	**FP #**	**FP %**	**TP #**	**TP%**	**Sensitivity %**	**Precision %**	**F1 Score**	**Accuracy %**	**Specificity %**		
Logistical Regression	Training	97	3.80	1220	47.79	1019	39.91	217	8.50	69.11	17.56	28.00	56.29	54.49	Good identification of TP (8.5 out of 12.3%). Sensitivity, accuracy, specificity moderate. Precision poor.	Reject
	Validation	49	4.47	619	56.48	342	13.40	86	7.85	63.70	20.09	30.55	64.32	64.41	No overfitting	
Neural Networks	Training	81	3.17	1482	58.05	757	29.65	233	9.13	74.20	23.54	35.74	67.18	66.19	Good identification of TP (9.13 out of 12.3%). Sensitivity, accuracy, specificity moderate to good. Precision poor.	Consider for Champion
	Validation	43	3.92	509	46.44	452	17.70	92	8.39	68.15	16.91	27.10	54.84	52.97	Overfitting (see accuracy measures)	
Niave Bayes	Training	84	3.29	1017	39.84	1222	47.87	230	9.01	73.25	15.84	26.05	48.84	45.42	Good identification of TP (9.01 out of 12.3%). Sensitivity good. Precision poor. Accuracy and specificity poor to moderate.	Reject
	Validation	37	3.38	437	39.87	524	20.52	98	8.94	72.59	15.76	25.89	48.81	45.47	No overfitting	
SVM RBF (model 9)	Training	**4**	**0.17**	584	22.88	1655	64.83	**309**	**12.12**	**98.62**	15.75	27.16	35.00	26.09	Good identification of TP. False negatives low. High sensitivity.Trade-off with accuracy and FP.	Consider for Champion
	Validation	**9**	**0.82**	220	20.08	720	65.66	**147**	**13.44**	**94.25**	16.99	28.79	33.52	23.42	No evidence here of overfitting, but see ROC	
Gradient boost (Gradient7)	Training	235	9.19	2224	87.13	15	0.57	79	3.1	25.22	84.47	38.85	90.24	99.35	Poor identification of TP, poor sensitivity. High specificity. Tradeoff	Reject
	Validation	101	9.26	955	87.10	6	0.59	33	3.05	24.78	83.79	38.24	90.15	99.33	No overfitting	
Bagging (Bag5)	Training	237	9.28	2208	86.49	31	1.21	77	3.02	24.52	71.30	36.49	89.50	98.62	Poor identification of TP, poor sensitivity. High specificity. Tradeoff. Good precision	Reject
	Validation	108	9.85	951	86.77	10	0.91	27	2.46	20.00	72.97	31.40	89.23	98.96	No overfitting	
Boosting (Boost3)	Training	130	5.11	1178	46.16	1061	41.55	184	7.19	58.46	14.75	23.56	53.34	52.63	All values moderate. Nothing good or great.	Reject
	Validation	58	5.26	500	45.61	461	42.09	77	7.05	57.27	14.35	22.95	52.65	52.01	No overfitting	
Decision Tree (simple)	Training	247	9.67	2233	87.47	6	0.24	67	2.62	21.34	91.78	34.63	90.09	99.73	High precision, specificity, accuracy. Low sensitivity and TP identification	Reject
	Validation	110	10.04	956	87.23	5	0.46	25	2.28	18.52	83.33	30.30	89.51	99.48	No overfitting	
Random Forest (Forest6)	Training	240	9.41	2222	87.04	17	0.67	74	2.88	23.43	81.13	36.36	89.92	99.24	High precision, specificity, accuracy. Low sensitivity and TP identification	Reject
	Validation	102	9.34	955	87.09	7	0.61	32	2.96	24.07	82.91	37.30	90.05	99.30	No overfitting	

Models were assessed for their generalizability by comparing training and validation datasets on the following metrics; first, misclassification rate, second, cumulative lift, third, cost. Where available, ROC index was also used.

Table 8: Model Generalizability Characteristics

Model Characteristics		**Training**				**Validation**				**Notes**
Model Name	Method	Misclassification Rate (%)	Cumulative Lift	ROC Index	Cost ($)	Misclassification Rate	Cumulative Lift	ROC Index	Cost ($)	
New Classifier	Maximum	10%	5.55	N/A	$1,634	11%	4.59	N/A	$735	High lift
The Champions	Maximum	10%	5.55	N/A	$1,634	11%	4.59	N/A	$735	High lift
Rocking the Boat	Maximum	10%	5.09	N/A	$1,626	11%	4.73	N/A	$733	Hight lift valid
Logistical Regression	N/A	10%	3.59	0.71	$1,688	11%	2.80	0.68	$680	No overfitting
Neural Network	N/A	10%	3.78	0.78	$1,316	11%	2.95	0.68	$749	Overfitting
Naïve Bayes	N/A	11%	3.01	0.66	$1,802	11%	2.71	0.64	$779	Expensive. Not great ROC
SVM RBF (model 9)	N/A	0.09	6.67	0.92	1685.00	0.11	3.67	0.69	780.00	Overfitting
Gradient boost (Gradient7)	N/A	10%	2.59	0.81	$1,633	10%	1.97	0.77	$707	Not great lift
Bagging (Bag5)	N/A	10%	3.01	0.62	$1,666	11%	2.62	0.60	$755	No good ROC
Boosting (Boost3)	N/A	47%	3.82	0.74	$1,961	47%	3.82	0.74	$859	Very poor accuracy
Decision Tree (simple)	N/A	9%	2.83	0.61	$1,710	10%	2.60	0.60	$764	Not good ROC
Random Forest (Forest6)	N/A	10%	3.88	0.77	$1,675	10%	3.74	0.76	$713	OK

Conclusion and Takeaways

The goal of this analysis was to determine the bad buys (true positives) within the car dataset. Of the Ensemble Nodes the winning model was “Rocking the Boat” using the maximum method (Figure 7: high lift rate in validation). However, this model was not adequately determining true positives (Figure 8). Five of the individual models achieved higher true positive and sensitivity scores compared with the Ensemble Node winner.

The decision for the champion model came down to balancing the “tradeoffs” of the model and understanding the goal of the analysis for the car sales company. The Naïve Bayesian model (Figure 9; SAS Enterprise Miner, n.d.) was the most “balanced” model with good identification of true positives and sensitivity (Figure 10). However, cost was high, and this model had a low ROC index. The neural networks model was similarly adequate but was overfitting results (Figure 11). Logistic regression had slightly lower numbers across the important categories. Therefore, the overall champion model remains the SVM (model 9) as it has high true positive identification, low false negatives, and high sensitivity. This model is not perfect however, as tradeoffs occur in precision, false positives, and accuracy. A caveat in using this model is that it is overfitting and therefore, should be used with caution on a large scale.

A limitation of the analysis is that there are other combinations of models, parameters and algorithms that could be explored. For instance, as the greatest change occurred in the Ensemble Node with the addition of a poor performing model, it would be interesting to have an equal number of poor, and well performing models, to “rock the boat” further causing the individual models to disagree further. In summation, I would still recommend the car sales company use SVM model 9 to identify the bad car sales, helping to improve inventory which will lead to happier customers.

References

Abbott, D. (2014). *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst.* Indianapolis, IN: Wiley Publishing.

Adjorlolo, S. (2018). Diagnostic accuracy, sensitivity, and specificity of executive function tests in moderate traumatic brain injury in Ghana. *Assessment*, *25*, 498–512. DOI: 10.1177/1073191116646445

Lindoff, G.S., & Berry, M.J.A. (2011). *Data Mining Techniques.* Wiley. Indianapolis, IN.

Knode, S. (2022). Ensemble Model 2: Bagging, Boosting & Random Forests. Retrieved February 17th, 2022, from: https://learn.umgc.edu/d2l/le/content/627222/viewContent/25080886/View

SAS Enterprise Miner (n.d.). *HP Bayesian Network Classifier Node.* Cary, NC: SAS Institute Inc. Retrieved February 17th, 2022, from: https://documentation.sas.com/doc/en/emref/15.1/n1twcnhiink4kyn1tu7za95ixh3h.htm

SAS Enterprise Miner (n.d.). *Ensemble Node.* Cary, NC: SAS Institute Inc. Retrieved February 17th, 2022, from: https://documentation.sas.com/doc/en/emref/15.1/n1upixvafylkyon1nfbokdnx3hpu.htm

Sarma, K.S. (2013). Chapter 7: Comparison and Combination of Different Models. In *Predictive Modeling with SAS Enterprise Miner*. Cary, NC: SAS Institute Inc.

Surowiecki, J. (2005). *The Wisdom of Crowds.* Anchor Books. New York, NY.

Appendix

Name	Label	Role	Level	Number of Levels	Percent Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness
Transmission	Automatic, manual	Input	Nominal	3	0.00137	.	.	.	.	.
TopThreeAmericanName	Manufacturers	Input	Nominal	5	0	.	.	.	.	.
VehicleAge	Years	Input	Nominal	10	0	.	.	.	.	.
WarrantyCost	Zip code where bought	Input	Interval	.	0	462	7498	1276.581	598.8468	2.070831
VNZIP1	Color	Input	Interval	.	0	2764	99224	58043.06	26151.64	-0.10353
Color	Size category e.g. SUV	Input	Nominal	17	0	.	.	.	.	.
Size	Manufacturer country	Input	Nominal	13	0	.	.	.	.	.
Nationality	Demand status	Input	Nominal	5	0	.	.	.	.	.
PRIMEUNIT	Alloy, covers	Input	Nominal	3	0	.	.	.	.	.
WheelType	At acquisition	Input	Nominal	4	0	.	.	.	.	.
VehBCost	Auction market price	Input	Interval	.	0	1	45469	6730.934	1767.846	0.715931
VehOdo	Online purchase	Input	Interval	.	0	4825	115717	71500	14578.91	-0.45315
MMRAcquisitionAuctionAveragePric	Retail market price	Input	Interval	.	0.024663	0	35722	6128.909	2461.993	0.463641
IsOnlineSale	Auction provider	Input	Binary	2	0	.	.	.	.	.
MMRAcquisitionRetailAveragePrice	Guarantee	Input	Interval	.	0.024663	0	39080	8497.034	3156.285	0.209214
Auction	Auction clean price	Input	Nominal	3	0	.	.	.	.	.
AUCGUART	Retail clean price	Input	Nominal	3	0	.	.	.	.	.
MMRAcquisitionAuctionCleanPrice	Auction clean price	Rejected	Interval	.	0.024663	0	36859	7373.636	2722.492	0.466501
MMRAcquisitonRetailCleanPrice	Unique buyer ID	Rejected	Interval	.	0.024663	0	41482	9850.928	3385.79	0.1763
MMRCurrentAuctionCleanPrice	Wheel Type ID	Rejected	Nominal	21	0	.	.	.	.	.
BYRNO	Make of car	Rejected	Interval	.	0	835	99761	26345.84	25717.35	2.129225
WheelTypeID	Model of car	Rejected	Nominal	5	0	.	.	.	.	.
Make	Retail average price	Rejected	Nominal	21	0	.	.	.	.	.
VehYear	Retail clean price	Rejected	Nominal	10	0	.	.	.	.	.
MMRCurrentRetailAveragePrice	Auction average price	Rejected	Nominal	21	0	.	.	.	.	.
MMRCurrentRetailCleanPrice	Car ID	Rejected	Nominal	21	0	.	.	.	.	.
MMRCurrentAuctionAveragePrice	Car submodel	Rejected	Nominal	21	0	.	.	.	.	.
Model	State car purchased	Rejected	Nominal	21	0	.	.	.	.	.
RefId	Car trim level	Rejected	Interval	.	0	1	73014	36511.43	21077.24	-.000203
SubModel	Bay avoidable purchase	Rejected	Nominal	21	0	.	.	.	.	.
VNST		Rejected	Nominal	21	0	.	.	.	.	.
Trim		Rejected	Nominal	20	4.367863	.	.	.	.	.
IsBadBuy		Target	Binary	2	0	.	.	.	.	.
PurchDate		Time ID	Interval	.	0	.	.	.	.	.

Figure 1: Variables after cleaning

Enter weight values for the decisions

Level	DECISION1	DECISION2
1 ...	0.0	6.9
0 ...	1.0	0.0

Figure 2: Cost Function Values

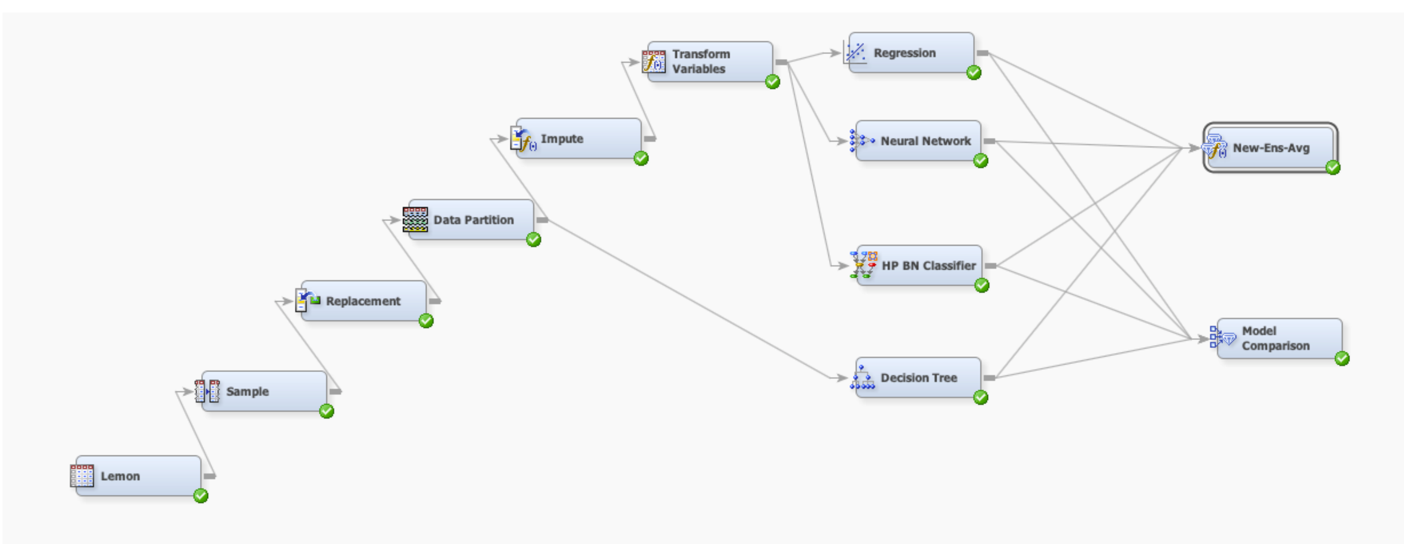

Figure 3: “New Classifier” model diagram using the average method for the Ensemble Node

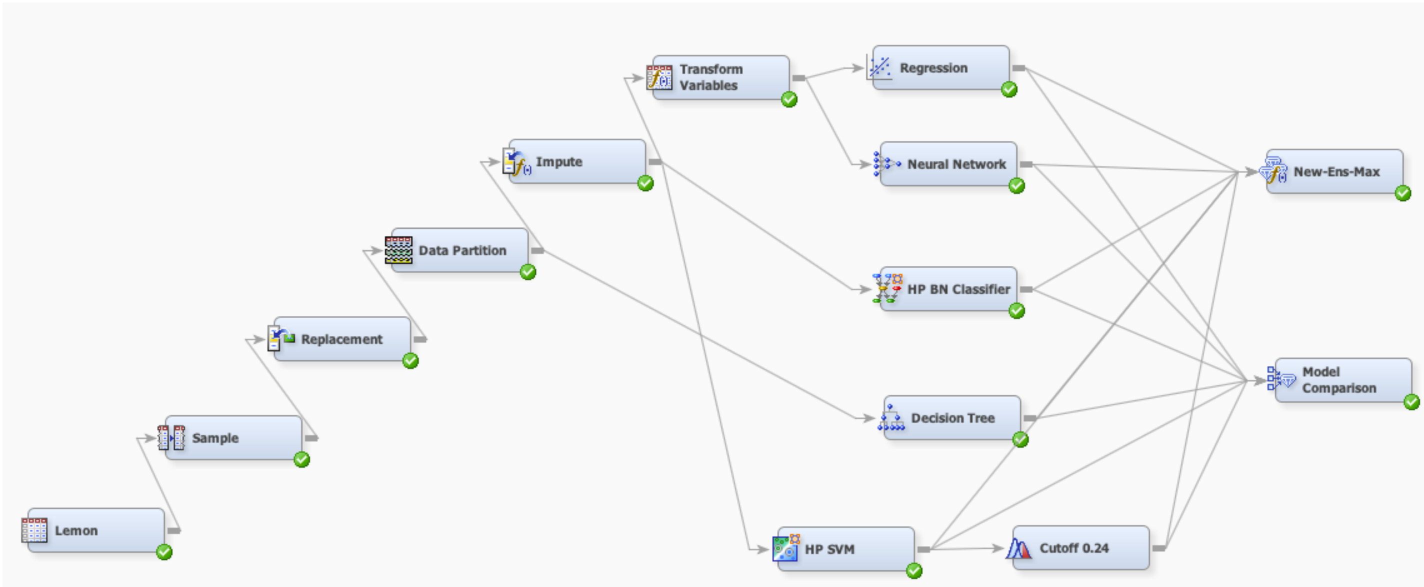

Figure 4: “The Champions” model diagram using the maximum method for the Ensemble Node

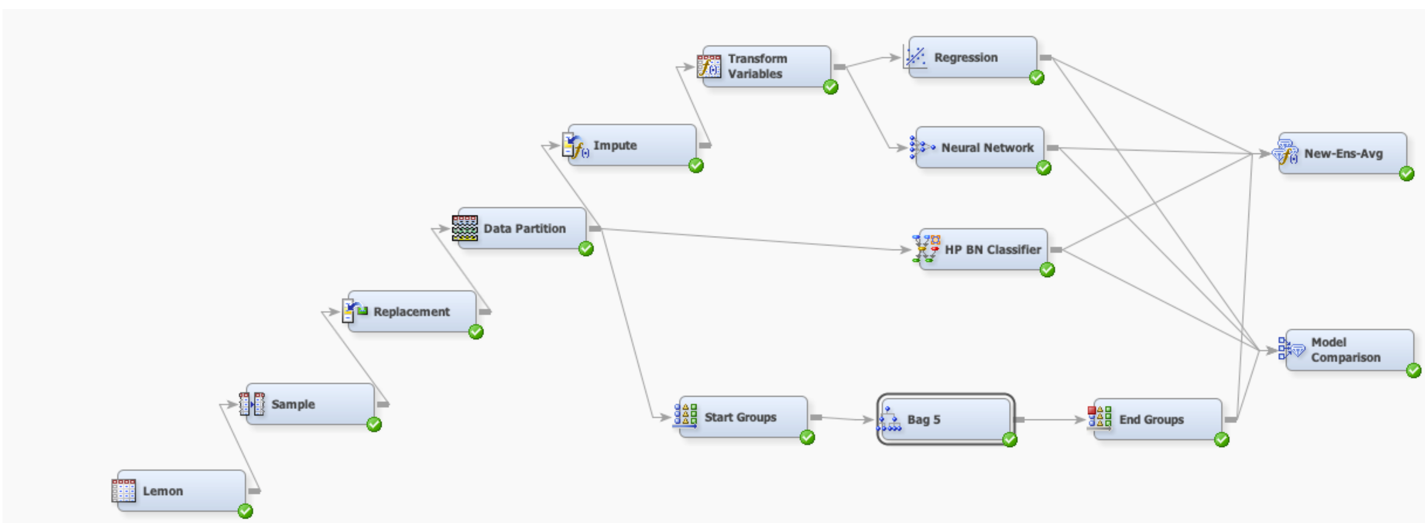

Figure 5: “Rocking the Boat” model diagram using the average method for the Ensemble Node

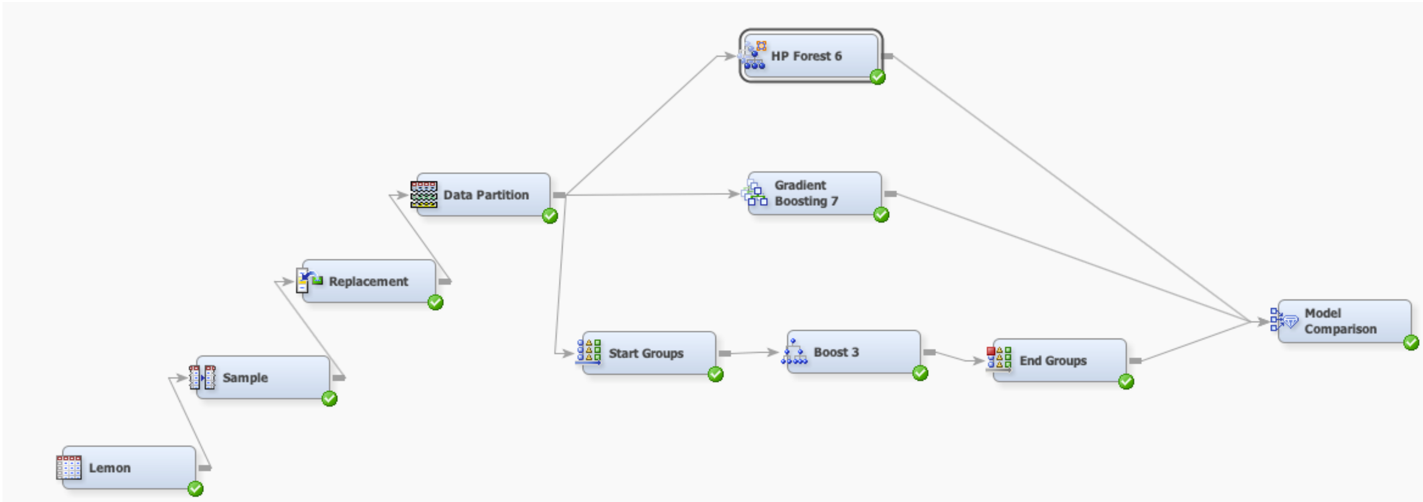

Figure 6: The Individuals model diagram, incorporating any past "best in class" models not yet used for comparative purposes

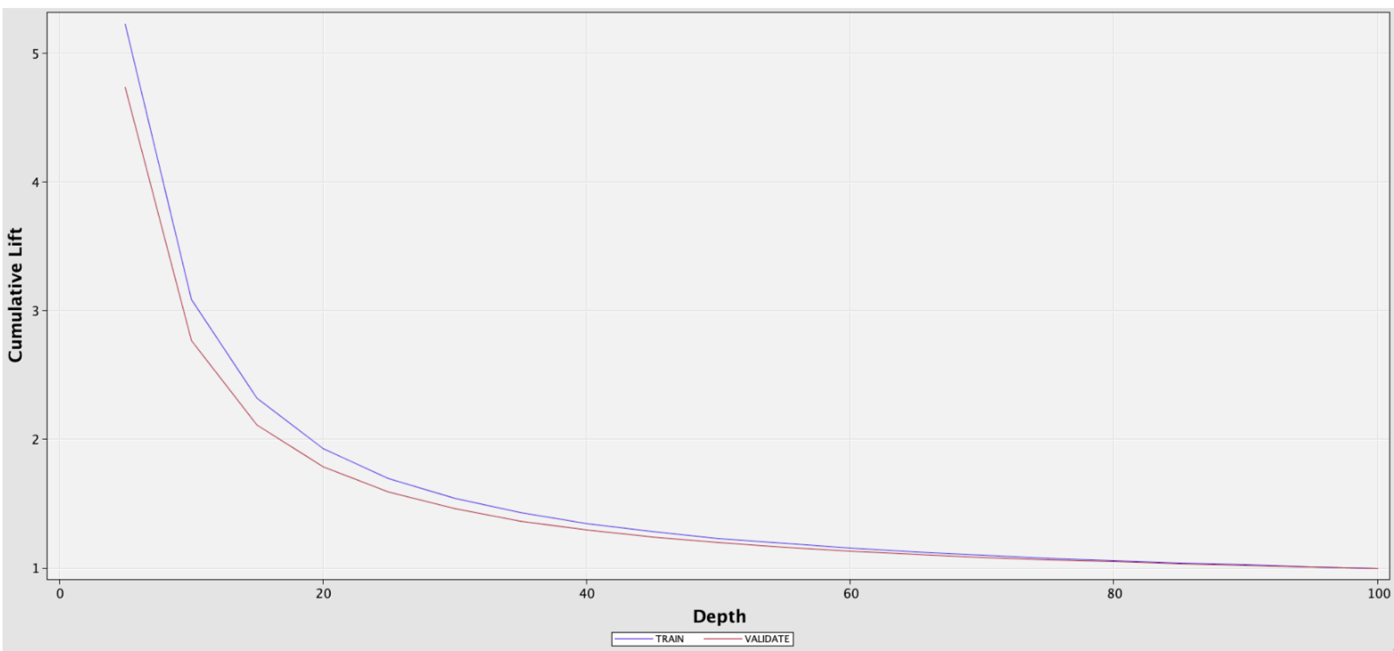

Figure 7: Cumulative lift for the Rocking the Boat model

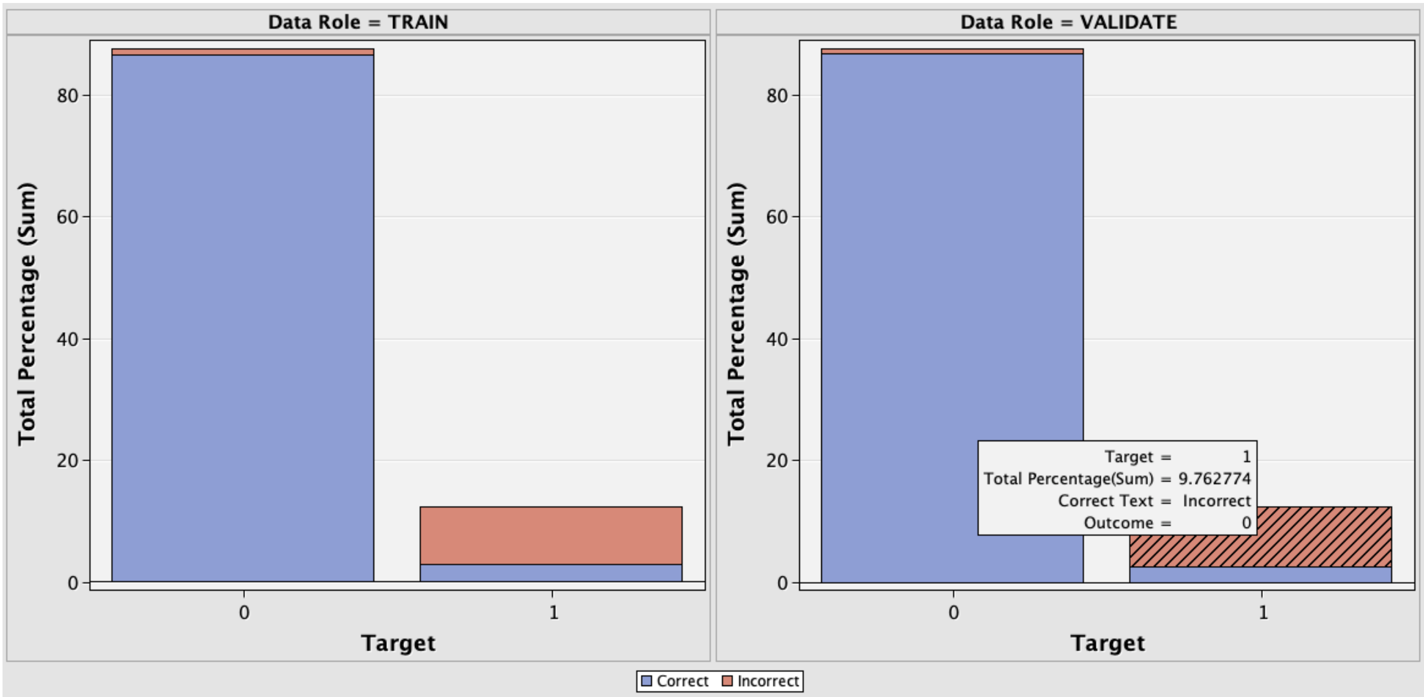

Figure 8: Misclassification Rates for Rocking the Boat model

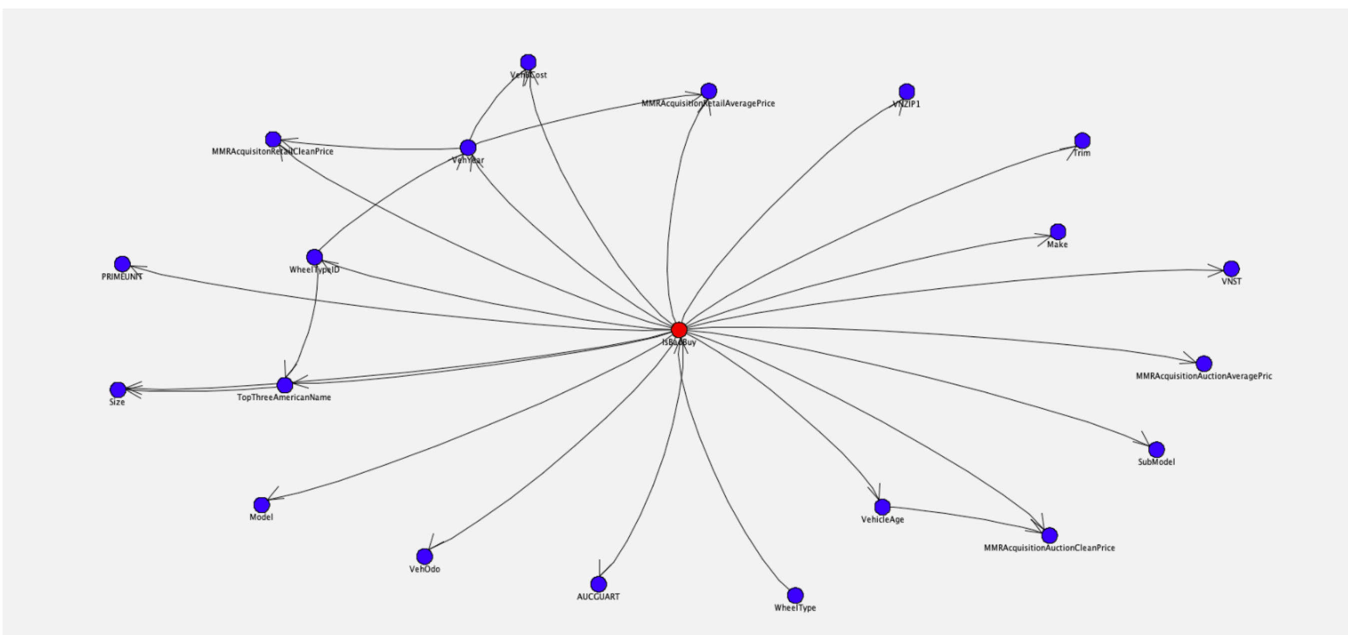

Figure 9: Naïve Bayesian Network (prior to variable cleaning)

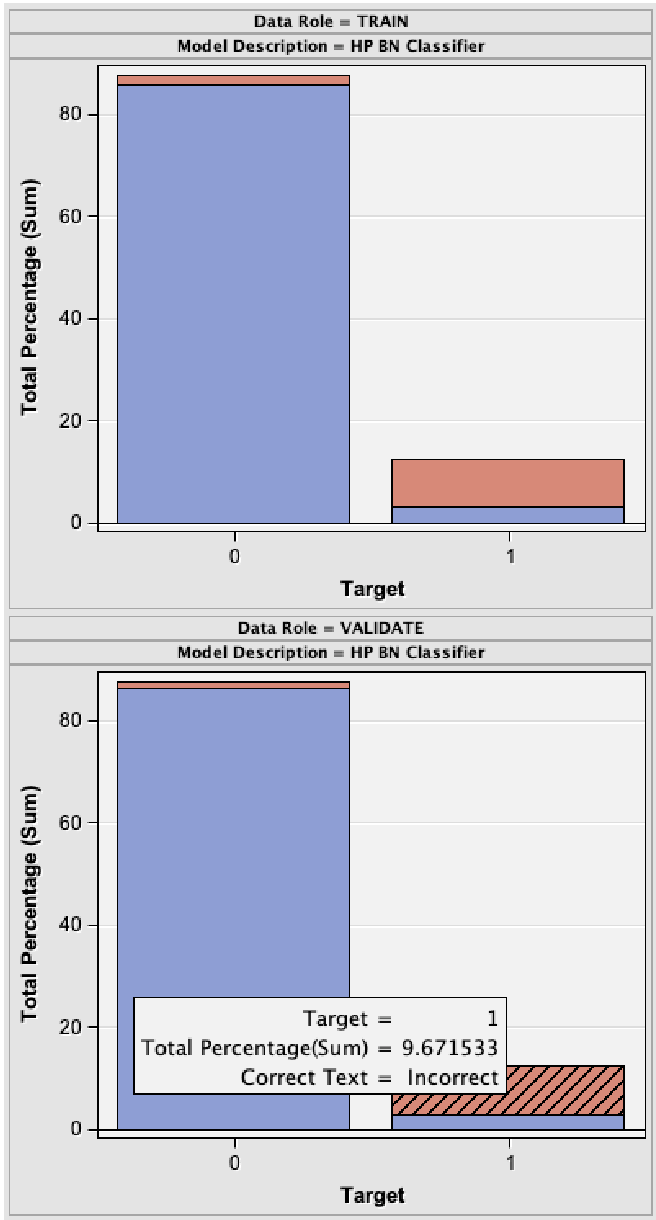

Figure 10: Misclassification rate showing incorrect target variable results in the validation dataset

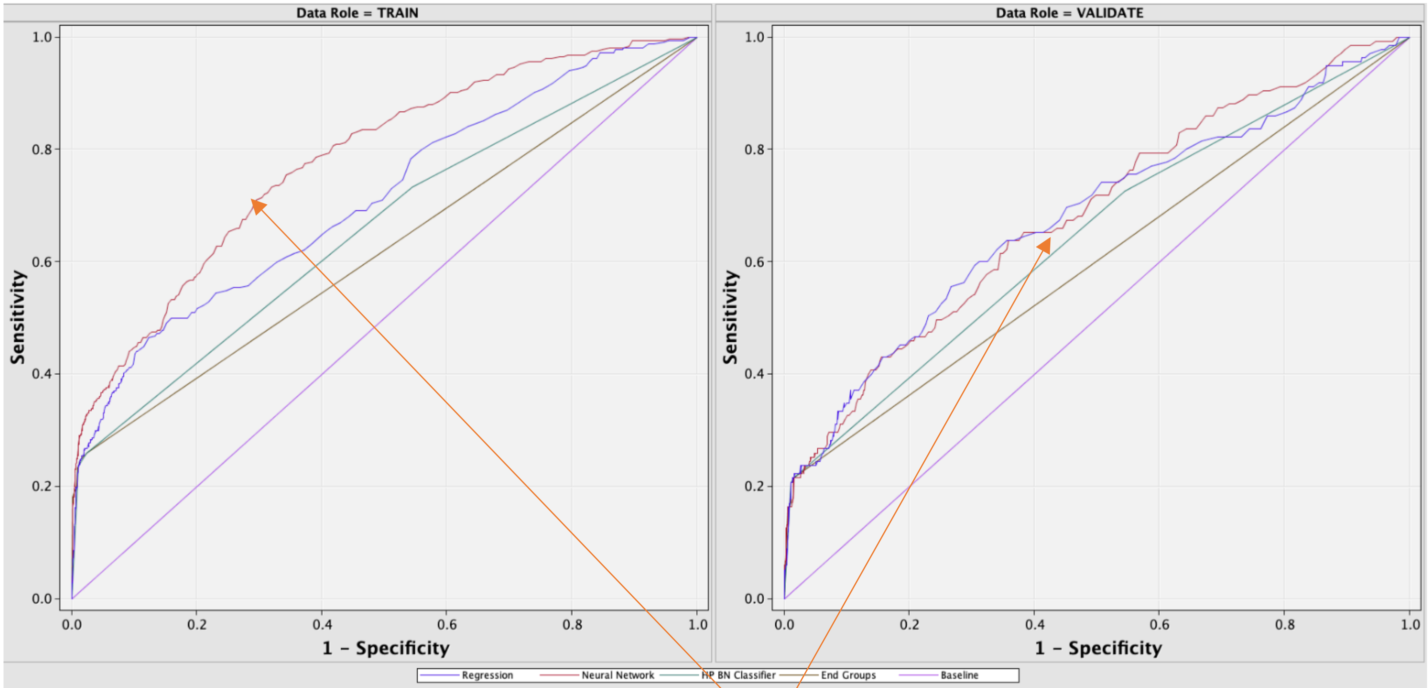

Figure 11: Neural Networks overfitting. ROC Index 0.78 in training and 0.68 in validation datasets