Data 640: Predictive Modeling Fall 2023

Assignment # 3– Ensemble Models 2

Simon A Hochmuth

Email: shochmuth3@student.umgc.edu

Professor: Ed Herranz

Introduction

Churn is the percentage of accounts that cancel or choose not to renew a subscription, by doing so companies lose revenue due to the lost payments. For this reason, there is a vested interest for companies to correctly identify customers likely to leave and attempt to retain their subscription (ProductPlan 2021). The goal of this analysis is to utilize the data within the churn dataset to build and find the best predictive model that predicts whether a customer is likely to leave a company. For this analysis a neural network, decision tree boost, logistic regression, and ensemble model were used to build 11 predictive models within the SAS Enterprise Miner application. After the models are built, the 11 will be compared to see which had the best performance in predicting churn. After finding the best model within this analysis, these will be compared against previous assessments of ensemble and SVM models to figure out which model did the best across all sets of analysis conducted.

Data Set Description

As discussed, the dataset chosen had the goal of conducting predictive analysis to predict whether a customer is likely to leave the telephone company. The churn dataset contained 3333 observations and had 21 total variables, including the target variable of churn. There are no missing values in all the data. The original dataset provided by UMGC, shown in figure 1, has 21 variables consisting of customer personal data like phone number, plan type, state, and area code and individual plan metadata like number of voicemails, number of international calls, time of calls, number of extra charges for calls, and the total minutes. Overall, the variables had normalized data with the highest skew being 1.32 in the variable total international calls. For that reason, it was decided not to transform the data to make it more normalized. The target variable

churn is a binary variable that states whether someone left or stayed with the telecommunication company. Where naturally the company cares to learn who is more likely to leave because that is a lost revenue source. The target variable was highly skewed, with 14.5% of the data being TRUE or customers that left, and the other 85.5% being FALSE as shown in figure 2 below. Since the target variable is skewed, some work will need to be done to prepare it to be used by the model.

Data Cleansing and Preparation

As discussed earlier, the data in this model had normalized distributions with no missing data, meaning there was not a need to cleanse the data for the model. The first step was simply to pull the data into the SAS Enterprise Miner application, and to accurately set the roles and levels for each variable. Specifically, for churn the role was set to “target” and “binary” was set for the level. This is required for predictive models to work correctly. After the data was explored, the next step was to partition the data to prepare it for the models. It was decided to partition 30% of the data into the training data set, and 70% into the validation data set. After the data was partitioned, the next step was to move towards building the Ensemble models. It was decided not to transform the target variable at this point, because a cutoff node would be used to deal with the imbalance in the data after the models were created.

Predictive Models Developed

After the three previous analyses were completed, a total of 25 models were created to predict churn utilizing various Ensemble and Support Vector Machine (SVM) techniques. Table 1 below shows the 25 models that were developed for this analysis with their parameters and optimal cutoff values chosen. Where figures 13 to 15 show the model diagrams used for each of

these assignments and figures 3 to 12 can be reviewed to see the parameters utilized for each model. During the most recent analysis in assignment 3, Neural Network, Logistic Regression, and Decision Tree Boosting models were first created. Then, these three models were used as inputs for eight new Ensemble models. Meaning a total of 11 models were created during this analysis and will be compared against the other 14 created during assignments 1 and 2. The 8 new ensemble models contained two for each of the maximum, voting, proportion, and average ensemble approaches. Out of the 11 new models, the optimal cutoff values were chosen by reviewing the values in a cutoff table like the one shown in figure 16. A prioritization was placed on true positive accuracy, because as discussed previously a company is more interested in knowing these values. However, overall accuracy and goodness values were also considered when choosing cutoff values to maximize overall model accuracy. The final parameters chosen are in table 1 along with the optimal cutoff values, model type, default parameters, relative figures, and changed parameters.

As previously discussed, the target variable churn was skewed, with 14.5% of the 3333 observations being positive, and the other 85.5% being negative. For that reason, a cutoff node was implemented on each model to address the imbalance in the target data. Since the optimization method and parameters differed between each model, the cutoffs were chosen on a case by case basis. The process of choosing the cutoff values included creating an initial node and setting the value equal to 14.5%, to match the current imbalance in the target data. It was decided that any optimal cutoff chosen would not go lower than this value, to ensure all target true positives are represented within the data. The goal was to choose a cutoff that would maximize true positives and true negatives without lowering model accuracy by a large amount. The goal of maximizing true positives and true negatives is called the goodness rating. The

goodness rating is the sum of the true positive and negative rates, where a model that is 100% accurate in true positive values and 100% accurate in true negative values will have a 100% + 100% = 200% goodness rating. Naturally no model will be 100% accurate for both rates and a 200% goodness rate, but the goal is to pick one that is the closest to achieving it. Where anything above 150% goodness is significant. There was an emphasis put on a higher true positive rate because realistically companies are more interested in knowing whether a person will leave the company for retention targeting. The final SAS diagrams can be seen in figure 13 to 15 below and the optimal cutoffs chosen are listed in table 1 below. The cutoff node tables, like figure 16 below, were analyzed to provide the optimal cutoff values that are listed in table 1 below. Where the goodness scores, misclassification rates, and precision were all evaluated to determine the final optimal cutoff values depicted in the table.

Name	Model Type	Default parameters values	Notable parameter changes from default	Figure of Parameters used	Optimal Cutoff Value	Assignment
RBF1	SVM RBF	Figure 9	N/A	1	0.2	#1 SVM
Poly2	SVM 2nd Polynomial	Figure 9	N/A	2	0.37	#1 SVM
Poly3	SVM 3rd Polynomial	Figure 9	N/A	3	0.44	#1 SVM
Linear	SVM Linear	Figure 9	N/A	N/A	0.14	#1 SVM
Sigmoid 1	SVM Sigmoid	Figure 9	N/A	+/-1	0.27	#1 SVM
Sigmoid 2	SVM Sigmoid	Figure 9	N/A	+/-2	0.14	#1 SVM
Decision Tree Bag1	Bagging	Figure 3	No changes	Figure 3	0.17	#2 Ensemble
Decision Tree Bag2	Bagging	Figure 3	Nominal/Ordinal Criterion = Gini, Interval criterion = Variance, category size = 50, max depth = 10 , 0.15 significant level	Figure 4	0.17	#2 Ensemble
Decision Tree Boost1	Boosting	Figure 3	No changes	Figure 3	0.18	#2 Ensemble
Decision Tree Boost2	Boosting	Figure 3	Nominal/Ordinal Criterion = Gini,	Figure 4	0.2	#2 Ensemble
HP Forest1	Random Forest	Figure 5	No changes	Figure 5	0.15	#2 Ensemble
HP Forest2	Random Forest	Figure 5	Proportion of Obs in each sample = 0.15, Significance level = 0.05, max categories = 30, minimum category size = 5	Figure 6	0.15	#2 Ensemble
Gradient Boost 1	Gradient Boost	Figure 7	No changes	Figure 7	0.15	#2 Ensemble
Gradient Boost 2	Gradient Boost	Figure 7	Max Depth = 2, max n iterations = 75, train proportion = 15	Figure 8	0.17	#2 Ensemble
Neural Network Basic	Neural Net	Figure 10	No changes	Figure 10	0.16	#3 Ensemble
Decision Tree Boost Basic	Decision Tree Boost	Figure 11	No changes	Figure 11	0.45	#3 Ensemble
Logistic Regression basic	Logistic Regression	Figure 12	No changes	Figure 12	0.17	#3 Ensemble
Ensemble avg2	Ensemble Avg	Figure 10/11	Average w/ Neural Net & DT boost	Figure 10/11	0.3	#3 Ensemble
Ensemble vote2	Ensemble Vote	Figure 10/11	Vote w/ Neural Net & DT boost	Figure 10/11	0.35	#3 Ensemble
Ensemble max2	Ensemble Max	Figure 10/11	Max w/ Neural Net & DT boost	Figure 10/11	0.47	#3 Ensemble
Ensemble proportion2	Ensemble Proportion	Figure 10/11	Proportion w/ Neural Net & DT boost	Figure 10/11	0.4	#3 Ensemble
Ensemble avg1	Ensemble Avg	Figure 11/12	Average w/ Log Regression & DT boost	Figure 11/12	0.3	#3 Ensemble
Ensemble vote1	Ensemble Vote	Figure 11/12	Vote w/ Log Regression & DT boost	Figure 11/12	0.37	#3 Ensemble
Ensemble max1	Ensemble Max	Figure 11/12	Max w/ Log Regression & DT boost	Figure 11/12	0.47	#3 Ensemble
Ensemble proportion1	Ensemble Proportion	Figure 11/12	Proportion w/ Log Regression & DT boost	Figure 11/12	0.4	#3 Ensemble

Table 1: Shows the names of the 25 models used in Assignment 1 to 3 SVM & Ensemble analysis, as well as the training parameters, and optimal cutoff values chosen. Referenced figures can be found in the appendix.

Results

Now that the models were created with optimal cutoff values, the last step was to analyze the results to compare which had the best performance in predicting churn across all three analyses completed. Since companies care more about customers that leave, a higher focus was put on true positives when choosing the cutoffs shown in table 1 above. Choosing cutoff values was to ensure the models aren't overfit towards the non-churn target events which make up

85.5% of the observations. This means there is a chance a model guessed that most values of churn = FALSE and would get very close to this accuracy. Since guessing everything false would end up giving a model a relatively good score of 85.5%, the goal was to utilize the goodness value to evaluate each of these models to ensure that both true positive and true negative values were considered. Figures 17 and 18 in the appendix show that all the models were created across all the analyses, while figure 19 below chose the top 11 performing models. As discussed previously the sensitivity rate or true positive rate would be prioritized when comparing each model since people who churn are the focus of the company.

	Optimal Cutoff	Sensitivity Rate (%)	Specificity Rate (%)	Overall Classification Accuracy (	Overall Precision rate (%)	Goodness Value (%) (Sensitivity + Specificit	Model Differences	Assignment
Ensemble avg2	30%	86.30%	93.10%	92.11%	97.55%	179.40%	Avg Ensemble Log Reg & Dec Tree Boost	#3 Ensemble 2
Ensemble vote2	35%	86.30%	92.99%	92.01%	97.50%	179.29%	Vote Ensemble Log Reg & Dec Tree Boost	#3 Ensemble 2
Decision Tree Boost 1	18%	84.90%	92.80%	91.00%	80.90%	177.70%	ProbF, ProbChisq , Entropy, Max category size = 5, max depth 6 , 0.2 significance level	# 2 Ensemble 1
Neural Network Basic	16%	83.56%	93.69%	92.20%	97.10%	177.25%	default NN model	#3 Ensemble 2
Ensemble proportion1	40%	82.20%	94.97%	93.11%	96.90%	177.17%	Proportion Ensemble w/ Neural Net & Dec Tree Boost	#3 Ensemble 2
Ensemble proportion2	40%	82.19%	94.97%	93.11%	96.90%	177.16%	Proportion Ensemble Log Reg & Dec Tree Boost	#3 Ensemble 2
Decision Tree Bagging 1	17%	84.90%	90.90%	90.00%	97.30%	175.80%	ProbF, ProbChisq , Entropy, Max category size = 5, max depth 6 , 0.2 significance level	# 2 Ensemble 1
Ensemble max2	47%	85.61%	89.95%	89.32%	97.34%	175.56%	Max Ensemble Log Reg & Dec Tree Boost	#3 Ensemble 2
Gradient Boost 1	15%	80.80%	94.60%	92.60%	96.65%	175.40%	Max Depth = 4, max n iterations = 50, train proportion = 40	# 2 Ensemble 1
Decision Tree Bagging 2	17%	75.40%	98.90%	95.50%	94.00%	174.30%	Variance, Gini Criterion, Max category size = 50, max depth 10 , 0.15 significance level	# 2 Ensemble 1
Ensemble vote1	37%	80.13%	92.99%	91.10%	96.48%	173.12%	Vote Ensemble w/ Neural Net & Dec Tree Boost	#3 Ensemble 2

Figure 19: Top models from figures 17 and 18 were chosen in a more condensed table based on goodness ratings.

When looking at the goodness rates across the models, it is shown that the clear winner was the Average Ensemble method with Logistic Regression and Decision Tree Boost input nodes performed the best, followed very closely by the same ensemble model but with the voting

method. The best model had 179.4% goodness rating, with 86.3% accuracy in the true positives, and 92.1% overall accuracy. The 179% goodness rating shows this model has amazing performance in respect to correctly predicting true positives and negatives, while the 92.1% overall accuracy shows an increase from 85.5%. Remember, if a model predicted 100% of the values to be FALSE, it would automatically get a score of 85.5% accuracy, and this model showed increase accuracy from this baseline. Overall, all the top models by goodness used the ensemble technique, while the SVM did not produce any model in the top 11. On an important note, the top SVM model produced a 163% goodness rating, which is 16% lower than the best performing ensemble model but is still a relatively good score. While the goodness level was very high at 179% for the top model, it is important to note that all models had higher accuracy for true negatives than true positives. This is expected because as previously discussed, there are more negative values leading to a bias in the model. Figure 19 above shows the top 11 models had over 80% accuracy in true positives. These results were very favorable, since the baseline of guessing all positive values is 14.5%, an accuracy of 80% or higher is a significant increase from what was originally provided. This proved that there was a significant improvement on both true positive and negative predictions, which shows the models provide significant insight into the prediction of churn.

Conclusions and Takeaways

The results of the 25 models showed that the Average Ensemble method with Logistic Regression and Decision Tree Boost input nodes and parameters described in table 1 above was the best model for the purposes of this analysis. The model had a cutoff of 30%, a goodness rating of 179.4%, with 86.3% accuracy in predicting people that would churn, and an overall accuracy of 92.11%. Other models did well also, with the next 10 models being within 6% of the

goodness rate of the top model. Other cutoff values were not assessed, because the highest performing value was individually chosen for each model. Therefore, the Average Ensemble model described proved to add the most value to the telecommunication companies' analysis of customers leaving the company and is the one recommended for the company to utilize. Naturally, the company would want to have close to 100% accuracy in true positives which is why it is suggested to expand upon the model. While the model is being expanded, it is also a recommendation to start to use this to target customers who are predicted to be true positives to build retention. Even with 86.3% true positive accuracy, there is a high likelihood of success of reaching someone who is planning on leaving to build their satisfaction. Future work can also include further diving into the results of both the SVM and Ensemble models to figure out why some had higher true positives rates, while others had higher true negative rates. Also understanding if there is overlap between true positive values, and if some models caught true positives that the champion model missed. With the top 11 models having goodness ratings within 6% of the champion model, each of these prove to show significant insight into predicting churn. However, some limitations of each of these models are the lack of insight into how each prediction was made which will make it difficult to decipher how these differences arose. Another limitation is because the cutoff value was used due to the bias within the target value, so there is no telling if the models missed some fundamental patterns due to the implemented cutoff.

References:

ProductPlan. (2021, August 12). *Churn*. Product Roadmap Software | ProductPlan. https://www.productplan.com/glossary/churn/

UMGC. (2023). *Churn at a Cellular Phone Company-mod3 Excel Spreadsheet*. Data-640 UMGC. https://learn.umgc.edu/d2l/le/content/920742/viewContent/31268436/View

Appendix

Name	ler	Drop	Lower Limit	Upper Limit	Number of Levels	Percent Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
account_length		No	.	.	.	0	1	243	101.0648	39.82211	0.096606	-0.10784
area_code		No	.	.	3	0	.	.	.	.	.	.
churn		No	.	.	2	0	.	.	.	.	.	.
customer_service_calls		No	.	.	10	0	.	.	.	.	.	.
international_plan		No	.	.	2	0	.	.	.	.	.	.
number_vmail_messages		No	.	.	.	0	0	51	8.09901	13.68837	1.264824	-0.05113
phone_number		No	.	.	21	0	.	.	.	.	.	.
state		No	.	.	21	0	.	.	.	.	.	.
total_day_calls		No	.	.	.	0	0	165	100.4356	20.06908	-0.11179	0.243182
total_day_charge		No	.	.	.	0	0	59.64	30.56231	9.259435	-0.02908	-0.01981
total_day_minutes		No	.	.	.	0	0	350.8	179.7751	54.46739	-0.02908	-0.01994
total_eve_calls		No	.	.	.	0	0	170	100.1143	19.92263	-0.05556	0.206156
total_eve_charge		No	.	.	.	0	0	30.91	17.08354	4.310668	-0.02386	0.025487
total_eve_minutes		No	.	.	.	0	0	363.7	200.9803	50.71384	-0.02388	0.02563
total_intl_calls		No	.	.	.	0	0	20	4.479448	2.461214	1.321478	3.083589
total_intl_charge		No	.	.	.	0	0	5.4	2.764581	0.753773	-0.24529	0.60961
total_intl_minutes		No	.	.	.	0	0	20	10.23729	2.79184	-0.24514	0.609185
total_night_calls		No	.	.	.	0	33	175	100.1077	19.56861	0.0325	-0.07202
total_night_charge		No	.	.	.	0	1.04	17.77	9.039325	2.275873	0.008886	0.085663
total_night_minutes		No	.	.	.	0	23.2	395	200.872	50.57385	0.008921	0.085816
voice_mail_plan		No	.	.	2	0	.	.	.	.	.	.

Figure 1: Shows churn data set with all relevant variables.

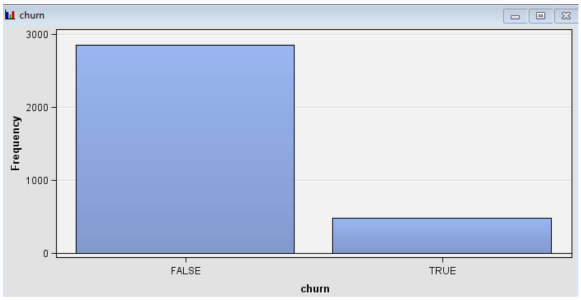
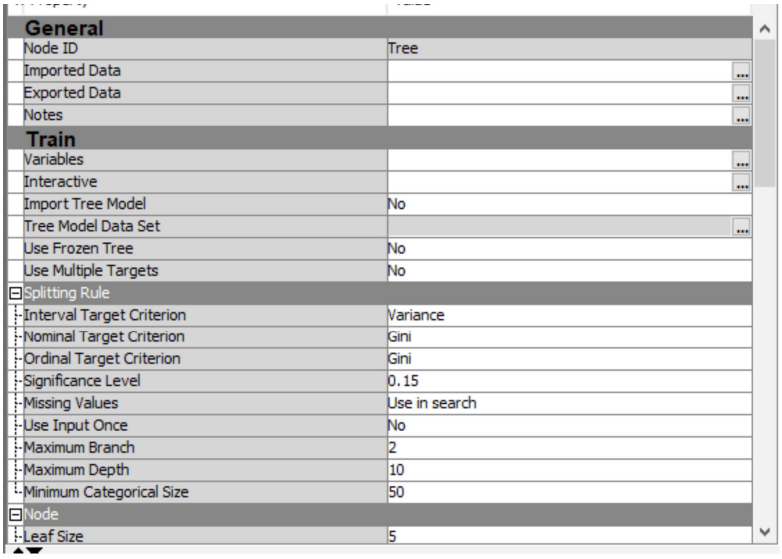


Figure 2: Bar chart of target variable Churn with 483 out of 3333 (14.5%) equal to True

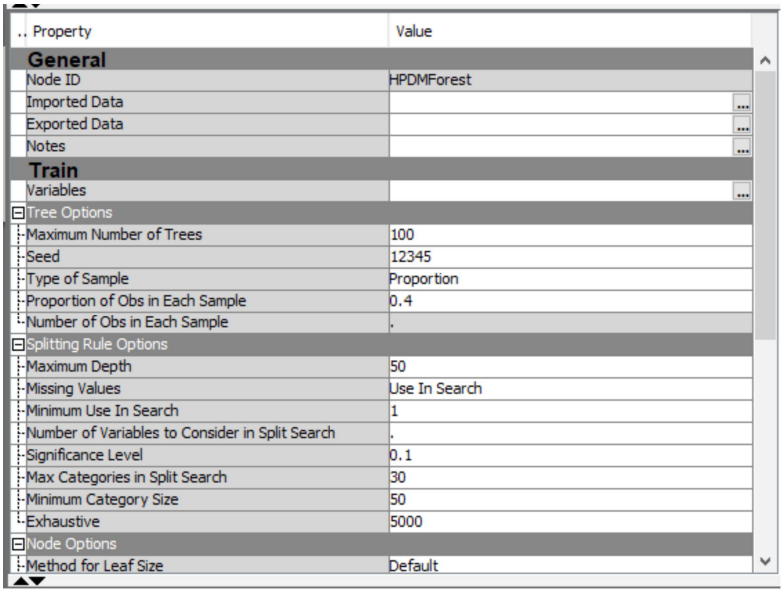
General	
Node ID	Tree3
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Interactive	
Import Tree Model	No
Tree Model Data Set	
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5

Figure 3: Default training parameters used for decision tree (bag1) and decision tree (boost 1) models.



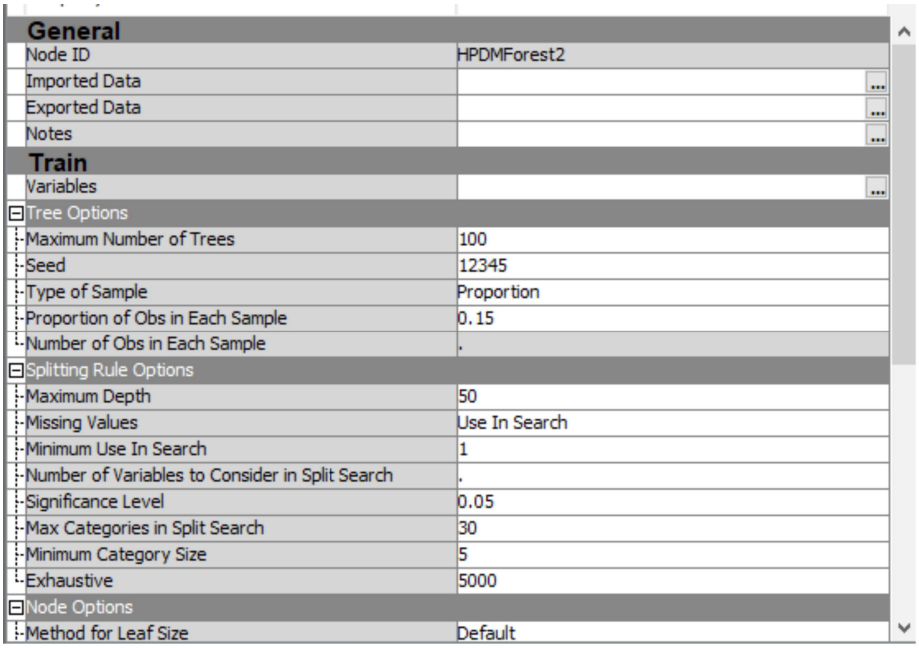
General	
Node ID	Tree
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Interactive	
Import Tree Model	No
Tree Model Data Set	
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	Variance
Nominal Target Criterion	Gini
Ordinal Target Criterion	Gini
Significance Level	0.15
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	10
Minimum Categorical Size	50
Node	
Leaf Size	5

Figure 4: Changed parameters used for decision tree (bag2) and decision tree (boost 2) models.



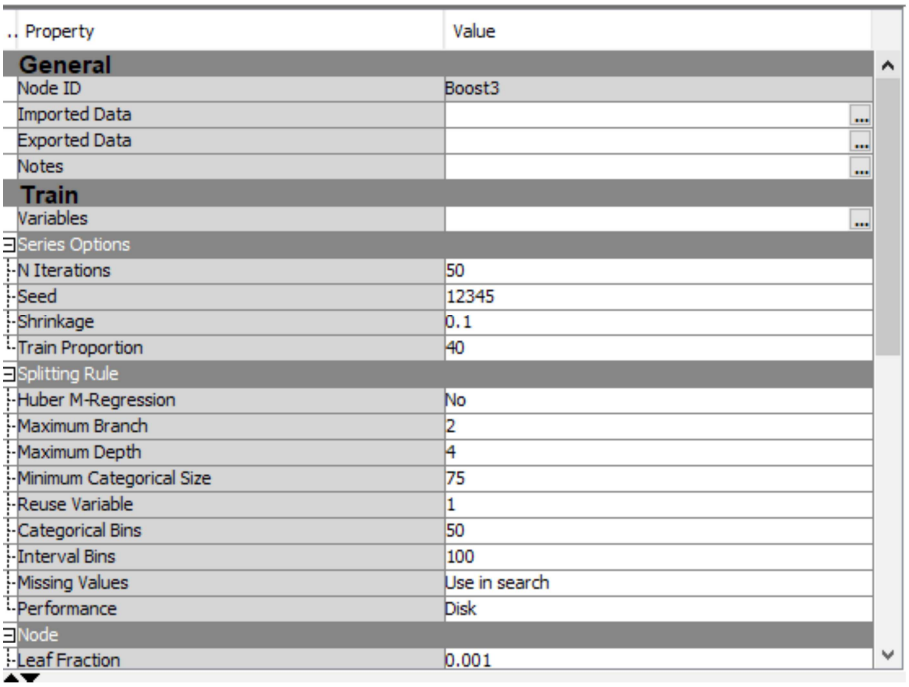
Property	Value
General	
Node ID	HPDMForest
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Tree Options	
Maximum Number of Trees	100
Seed	12345
Type of Sample	Proportion
Proportion of Obs in Each Sample	0.4
Number of Obs in Each Sample	.
Splitting Rule Options	
Maximum Depth	50
Missing Values	Use In Search
Minimum Use In Search	1
Number of Variables to Consider in Split Search	.
Significance Level	0.1
Max Categories in Split Search	30
Minimum Category Size	50
Exhaustive	5000
Node Options	
Method for Leaf Size	Default

Figure 5: Training parameters used for HP Forest Model 1.



General	
Node ID	HPDMForest2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Tree Options	
Maximum Number of Trees	100
Seed	12345
Type of Sample	Proportion
Proportion of Obs in Each Sample	0.15
Number of Obs in Each Sample	.
Splitting Rule Options	
Maximum Depth	50
Missing Values	Use In Search
Minimum Use In Search	1
Number of Variables to Consider in Split Search	.
Significance Level	0.05
Max Categories in Split Search	30
Minimum Category Size	5
Exhaustive	5000
Node Options	
Method for Leaf Size	Default

Figure 6: Training parameters used for HP Forest Model 2.



Property	Value
General	
Node ID	Boost3
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Series Options	
N Iterations	50
Seed	12345
Shrinkage	0.1
Train Proportion	40
Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	4
Minimum Categorical Size	75
Reuse Variable	1
Categorical Bins	50
Interval Bins	100
Missing Values	Use in search
Performance	Disk
Node	
Leaf Fraction	0.001

Figure 7: Training parameters used for Gradient Boost Model 1.

Property	Value
General	
Node ID	Boost4
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Series Options	
N Iterations	75
Seed	12345
Shrinkage	0.1
Train Proportion	15
Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	2
Minimum Categorical Size	75
Reuse Variable	1
Categorical Bins	50
Interval Bins	100
Missing Values	Use in search
Performance	Disk
Node	
Leaf Fraction	0.001

Figure 8: Training parameters used for Gradient Boost Model 2.

Train	
Variables	
Maximum Iterations	25
Use Missing as Level	No
Tolerance	1.0E-6
Penalty	1.0

Figure 9: Default training parameters used for each of the SVM models.

Property	Value
General	
Node ID	Neural
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Continue Training	No
Network	
Optimization	
Initialization Seed	12345
Model Selection Criterion	Profit/Loss
Suppress Output	No
Score	
Hidden Units	No
Residuals	Yes
Standardization	No

Figure 10: Neural Network parameters used for ensemble models in Assignment 3

Property	Value
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.

Figure 11: Decision tree boosting parameters used for ensemble models in Assignment 3

Property	Value
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	None
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
Convergence Criteria	
Uses Defaults	Yes
Options	...
Output Options	
Confidence Limits	No
Save Covariance	No

Figure 12: Log Regression parameters used for ensemble models in Assignment 3

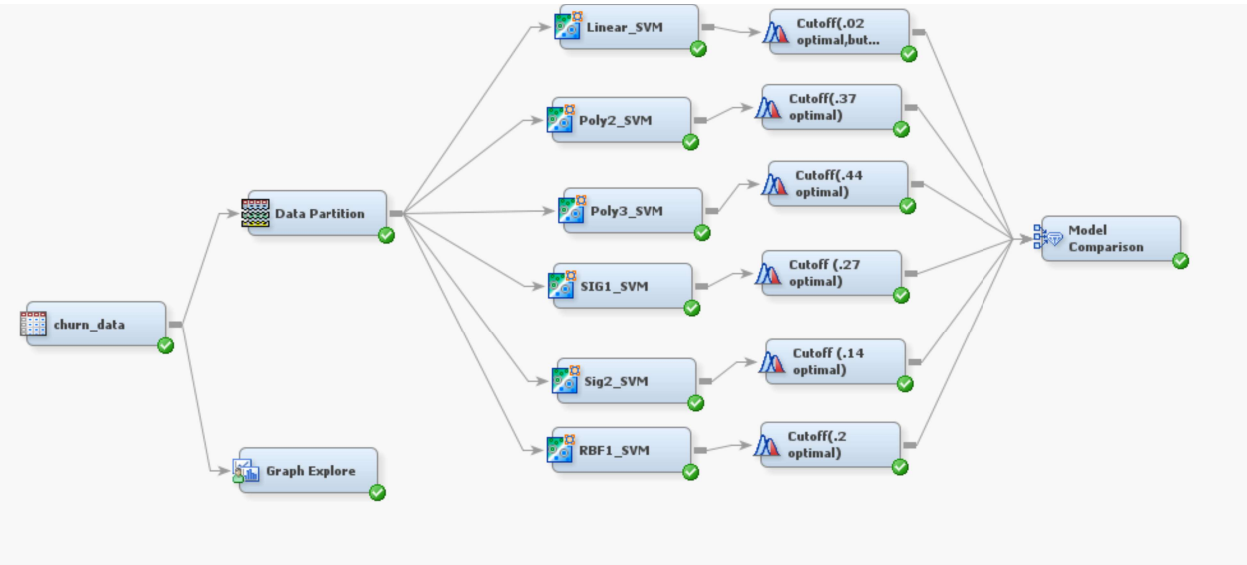


Figure 13: Full diagram of all 6 SVM models from Assignment 1, showing the original data, partition, models, and cutoff nodes.

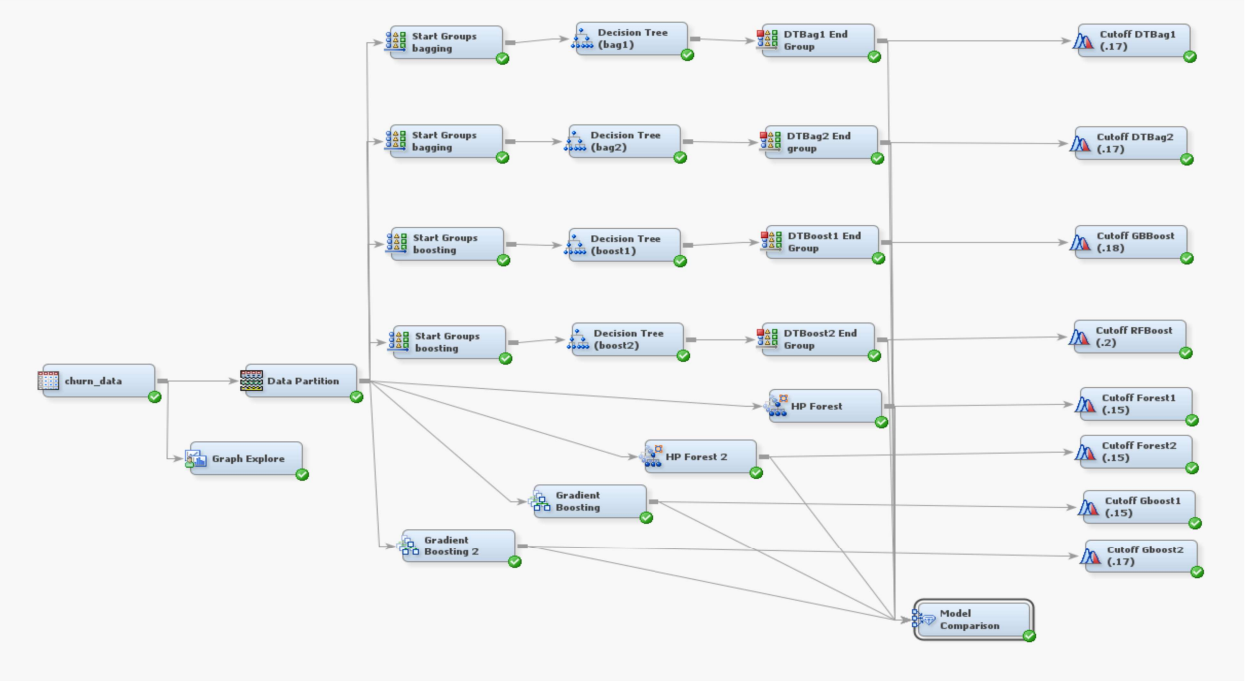


Figure 14: Full diagram of all 8 Ensemble models from Assignment 2, showing the original data, partition, models, and cutoff nodes.

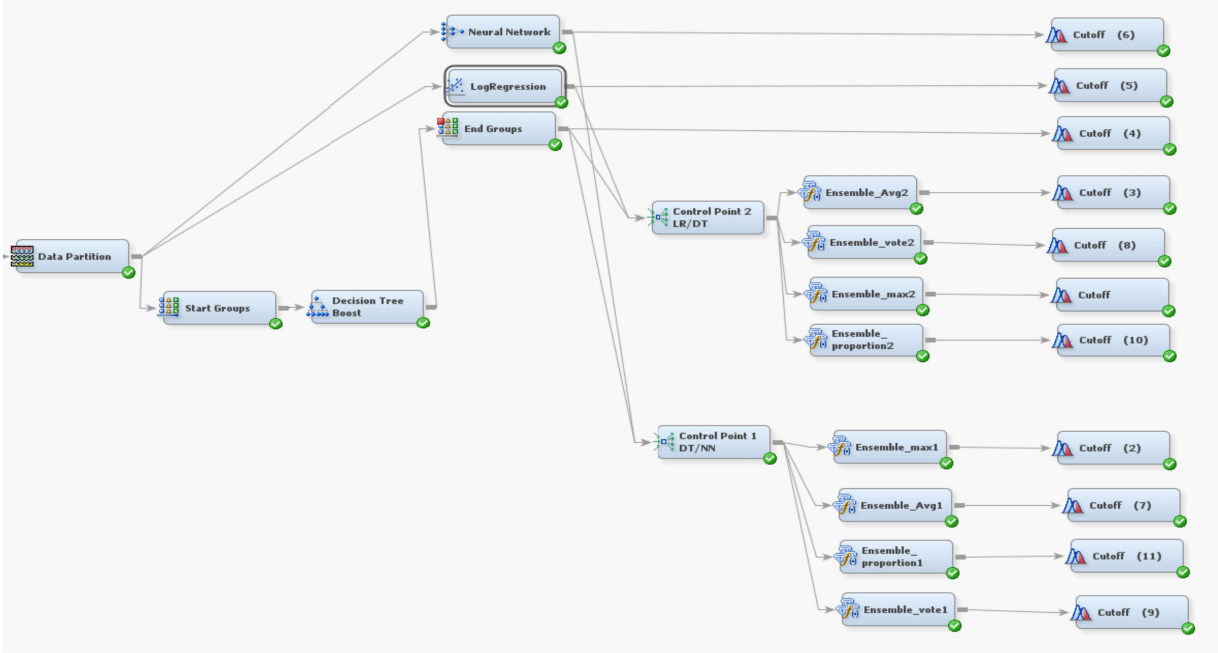


Figure 15: Full diagram of all 11 Ensemble models from Assignment 3, showing the original data, partition, models, and cutoff nodes.

Cutoff	Overall Classification Rate	True Positive Rate	True Negative Rate	Overall Precision Rate
0.21	86.12774	62.32877	90.18692	72.67473
0.2	85.52894	63.0137	89.36916	71.83991
0.19	84.93014	65.06849	88.31776	71.19912
0.18	84.93014	68.49315	87.73364	71.50442
0.17	84.43114	69.17808	87.03271	70.97265
0.16	82.83433	71.91781	84.69626	69.56952
0.15	82.23553	74.65753	83.52804	69.33989
0.14	79.54092	74.65753	80.37383	67.12337

Figure 16: table showing overall rates for the 2nd Gradient Boost model. (0.17 optimal)

	Optimal Cutoff	Sensitivity Rate (%)	Specificity Rate (%)	Overall Classification Accuracy (%)	Overall Precision rate (%)	Goodness Value (%) [Sensitivity + Specificity]	Model Differences	Assignment
Decision Tree Bagging 1	17%	84.90%	90.90%	90.00%	97.30%	175.80%	ProbF, ProbChisq , Entropy, Max category size = 5, max depth 6 , 0.2 significance level	# 2 Ensemble 1
Decision Tree Bagging 2	17%	75.40%	98.90%	95.50%	94.00%	174.30%	Variance, Gini Criterion, Max category size = 50, max depth 10 , 0.15 significance level	# 2 Ensemble 1
Decision Tree Boost 1	18%	84.90%	92.80%	91.00%	80.90%	177.70%	ProbF, ProbChisq , Entropy, Max category size = 5, max depth 6 , 0.2 significance level	# 2 Ensemble 1
Decision Tree Boost 2	20%	58.90%	95.90%	90.50%	82.00%	154.80%	Variance, Gini Criterion, Max category size = 50, max depth 10 , 0.15 significance level	# 2 Ensemble 1
HP Forest 1	15%	64.40%	90.53%	86.72%	93.70%	154.93%	Proportion of Obs in each sample = 0.4, Significance level	# 2 Ensemble 1
HP Forest 2	15%	83.56%	86.30%	85.90%	96.80%	169.86%	Proportion of Obs in each sample = 0.15, Significance level = 0.05, max categories = 30, minimum category size = 5	# 2 Ensemble 1
Gradient Boost 1	15%	80.80%	94.60%	92.60%	96.65%	175.40%	Max Depth = 4, max n iterations = 50, train proportion = 40	# 2 Ensemble 1
Gradient Boost 2	17%	78.30%	87.40%	86.05%	95.97%	165.70%	Max Depth = 2, max n iterations = 75, train proportion = 15	# 2 Ensemble 1
Polynomial 2nd Degree	37%	85.00%	78.00%	79.55%	68.50%	163.00%	2nd degree	#1 SVM
Polynomial 3rd Degree	44%	76.00%	75.00%	95.50%	64.50%	151.00%	3rd Degree	#1 SVM
Sigmoid +/- 1	27%	53.00%	51.00%	51.00%	51.00%	104.00%	sig = +/- 1	#1 SVM
Sigmoid +/- 2	14%	47.10%	98.19%	84.00%	64.00%	145.29%	sig = +/- 2	#1 SVM
RBF = 1	20%	88.00%	74.00%	76.00%	67.00%	162.00%	rbf = 1	#1 SVM
Linear	2%	12.90%	98.30%	85.00%	72.00%	111.20%	rbf = 1	#1 SVM

Figure 17: Assignment 1 & 2 model statistical table. The first decision tree model did the best with a 178% goodness value

	Optimal Cutoff	Sensitivity Rate (%)	Specificity Rate (%)	Overall Classification Accuracy (%)	Overall Precision rate (%)	Goodness Value (%) [Sensitivity + Specificity]	Model Differences	Assignment
Neural Network Basic	16%	83.56%	93.69%	92.20%	97.10%	177.25%	default NN model	#3 Ensemble 2
Decision Tree Boost Basic	45%	81.00%	84.90%	84.40%	96.40%	165.90%	default DT with boosting	#3 Ensemble 2
Logistic Regression basic	17%	79.50%	85.40%	84.50%	96.05%	164.90%	default LR model	#3 Ensemble 2
Ensemble avg2	30%	86.30%	93.10%	92.11%	97.55%	179.40%	Avg Ensemble Log Reg & Dec Tree Boost	#3 Ensemble 2
Ensemble vote2	35%	86.30%	92.99%	92.01%	97.50%	179.29%	Vote Ensemble Log Reg & Dec Tree Boost	#3 Ensemble 2
Ensemble max2	47%	85.61%	89.95%	89.32%	97.34%	175.56%	Max Ensemble Log Reg & Dec Tree Boost	#3 Ensemble 2
Ensemble proportion2	40%	82.19%	94.97%	93.11%	96.90%	177.16%	Proportion Ensemble Log Reg & Dec Tree Boost	#3 Ensemble 2
Ensemble avg1	30%	86.30%	84.69%	84.93%	97.31%	170.99%	Avg Ensemble w/ Neural Net & Dec Tree Boost	#3 Ensemble 2
Ensemble vote1	37%	80.13%	92.99%	91.10%	96.48%	173.12%	Vote Ensemble w/ Neural Net & Dec Tree Boost	#3 Ensemble 2
Ensemble max1	47%	81.50%	89.80%	88.60%	96.60%	171.30%	Max Ensemble w/ Neural Net & Dec Tree Boost	#3 Ensemble 2
Ensemble proportion1	40%	82.20%	94.97%	93.11%	96.90%	177.17%	Proportion Ensemble w/ Neural Net & Dec Tree Boost	#3 Ensemble 2

Figure 18: Assignment 3 Ensemble 1 model statistical table. With the Ensemble average model with Logistic Regression and Decision Tree boosting did the best with a goodness score of

179.4%