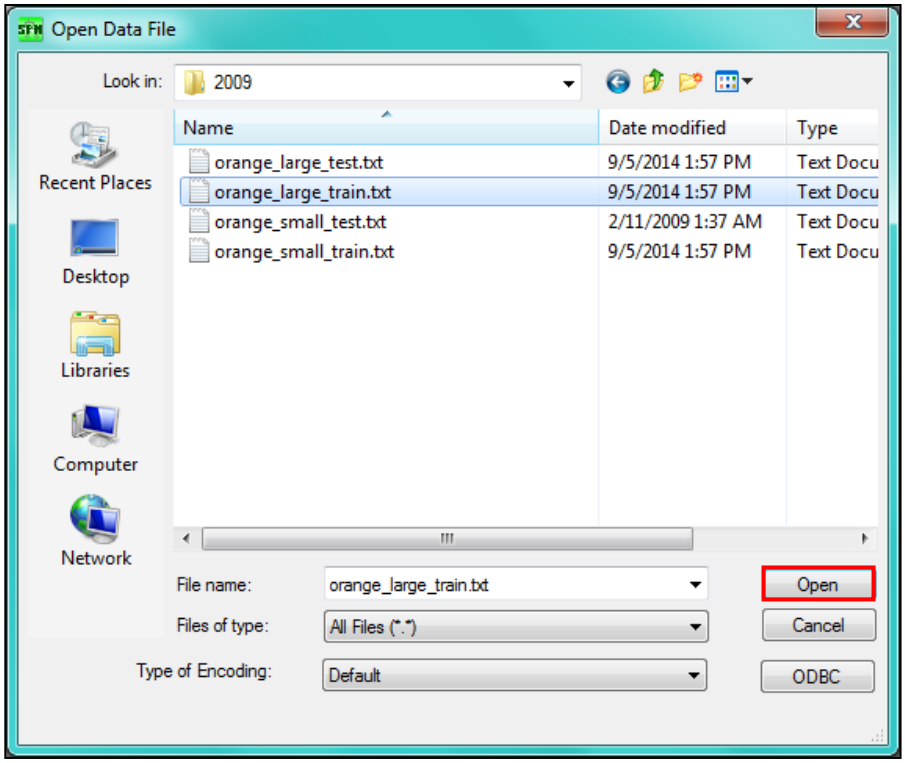
Tutorial: KDD Cup 2009*

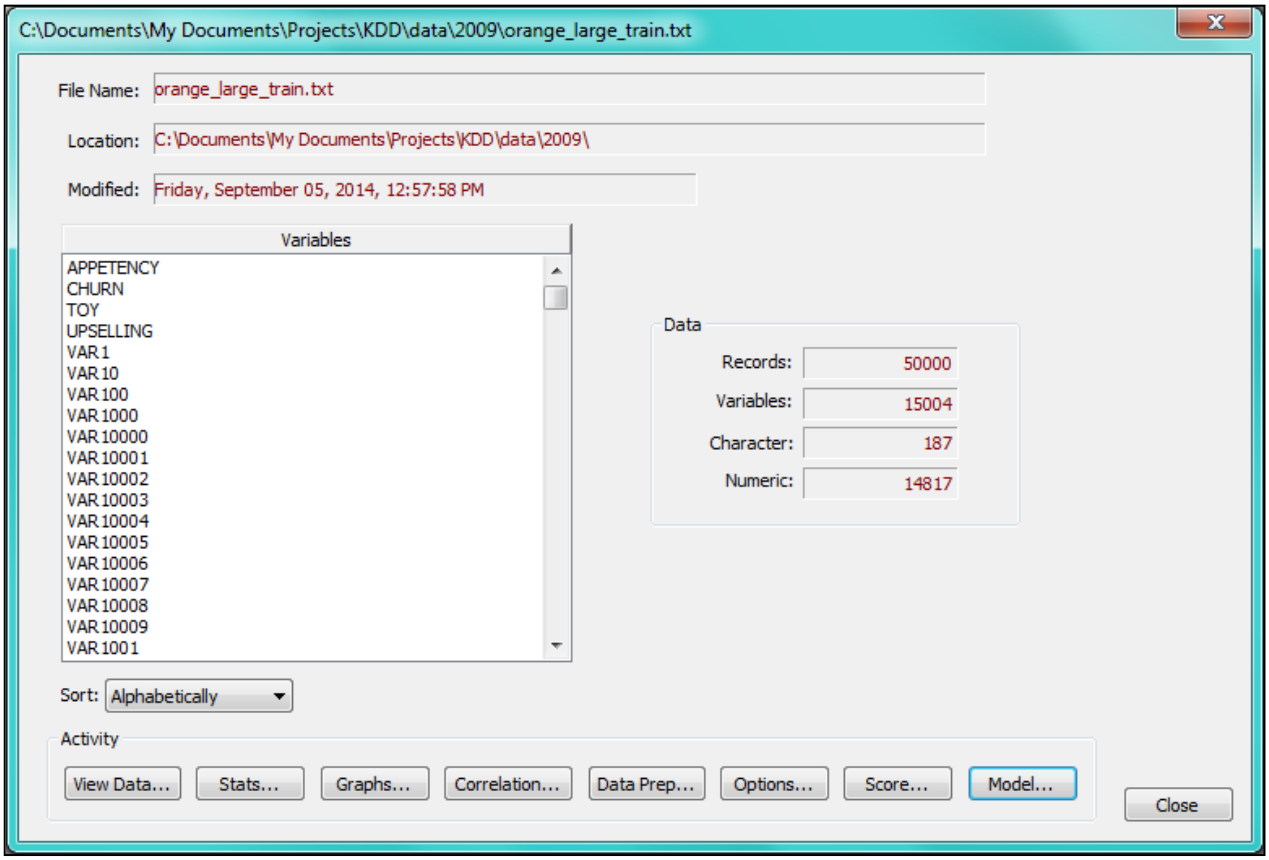
Not a data scientist? Don't sweat it! This is a tutorial simple enough for users of all levels to achieve results comparable to the experts.

Open the data file, orange_large_train.txt**:

File > Open > Data File…

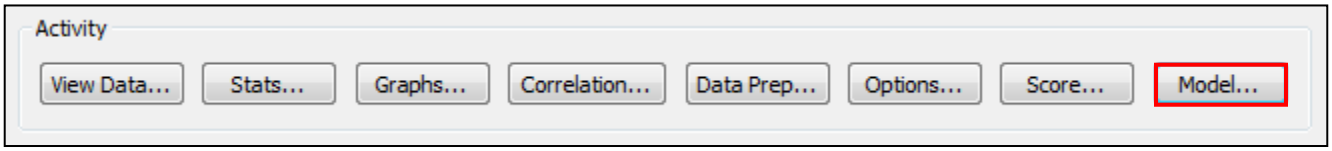


The Activity Window, pictured below, will appear:

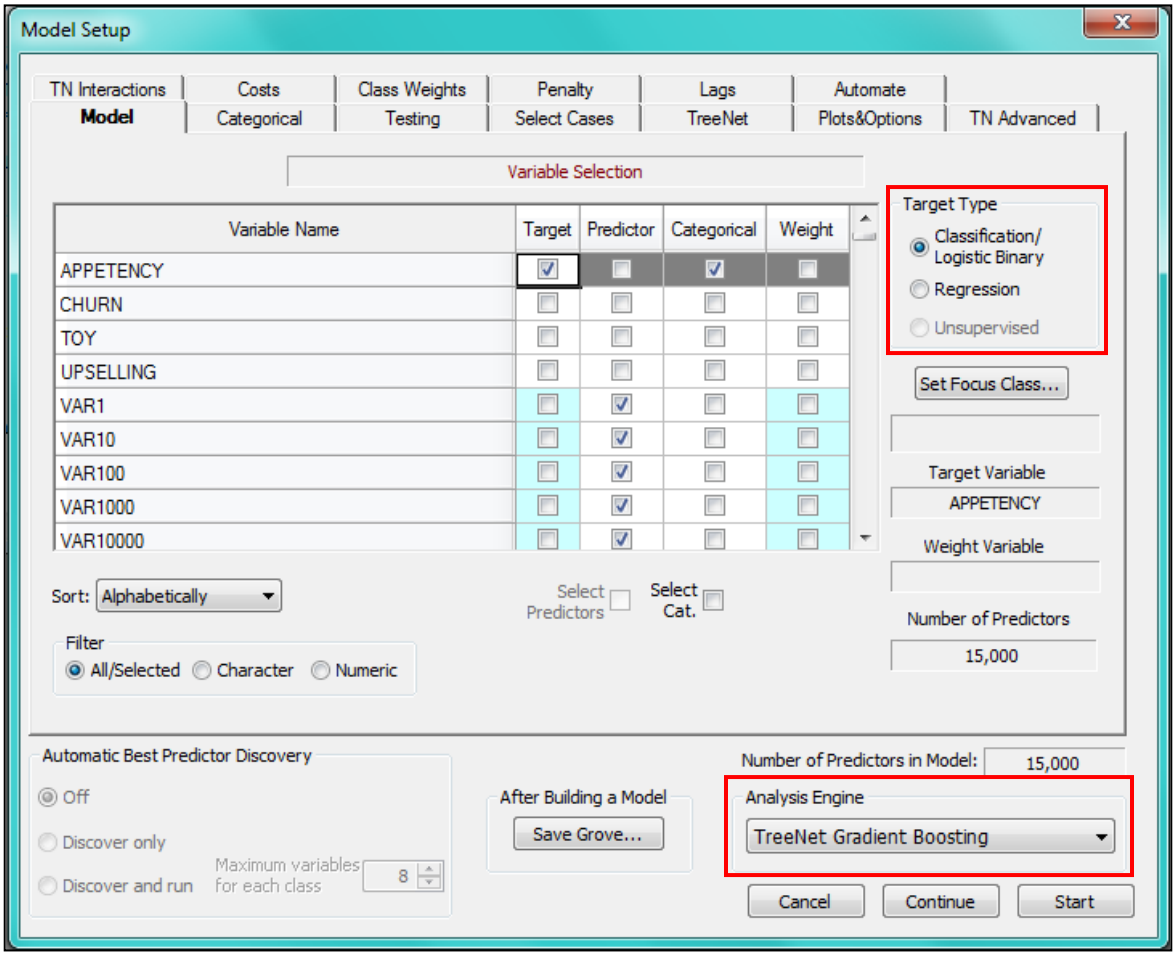


The list on the left contains the 3 target variables (Appetency, Churn, Upselling) and the 15,000 predictor variables. To the right, you can see there are 50,000 total observations.

In the row of buttons at the bottom of the Activity Window, click **Model**:

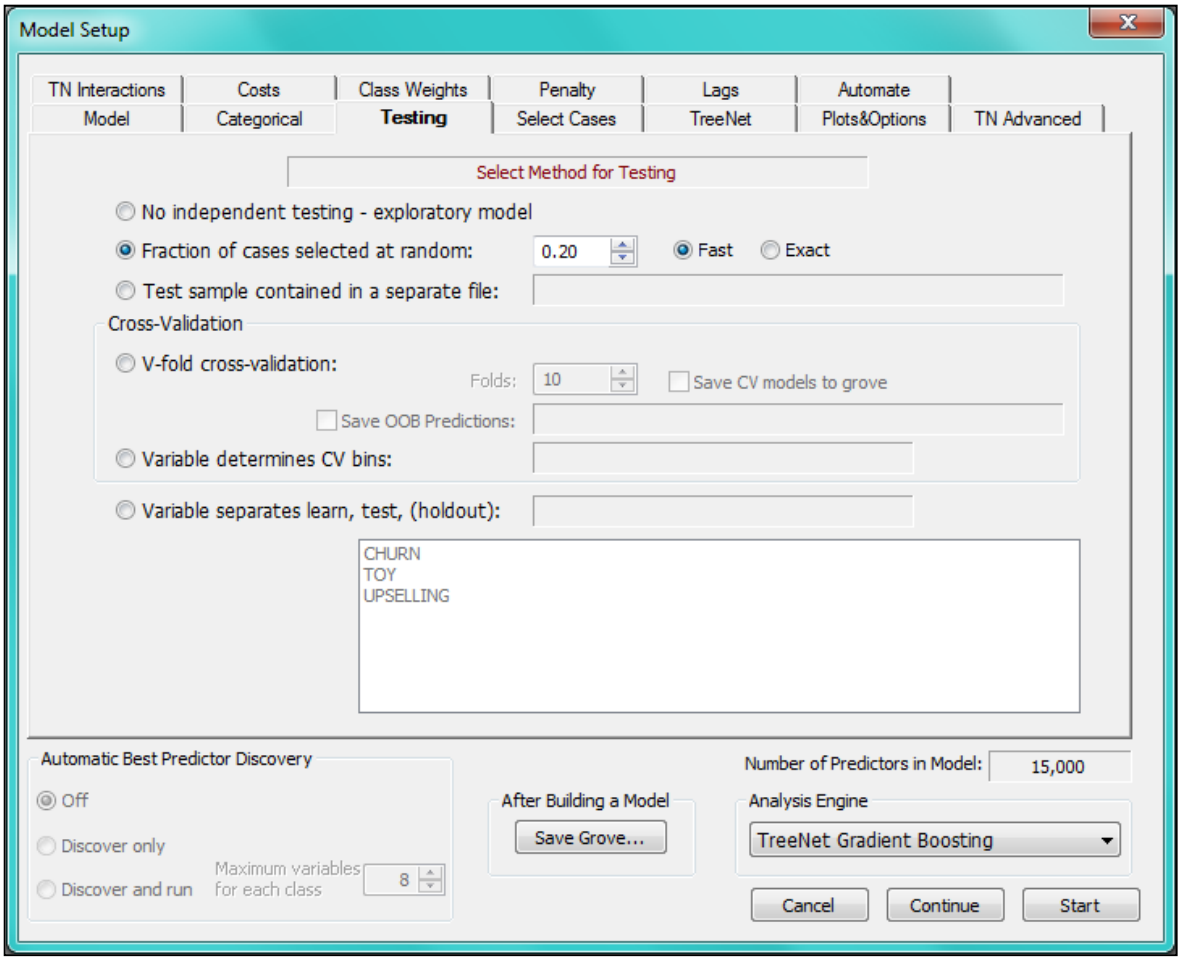


The Model Setup window, pictured below, will appear:



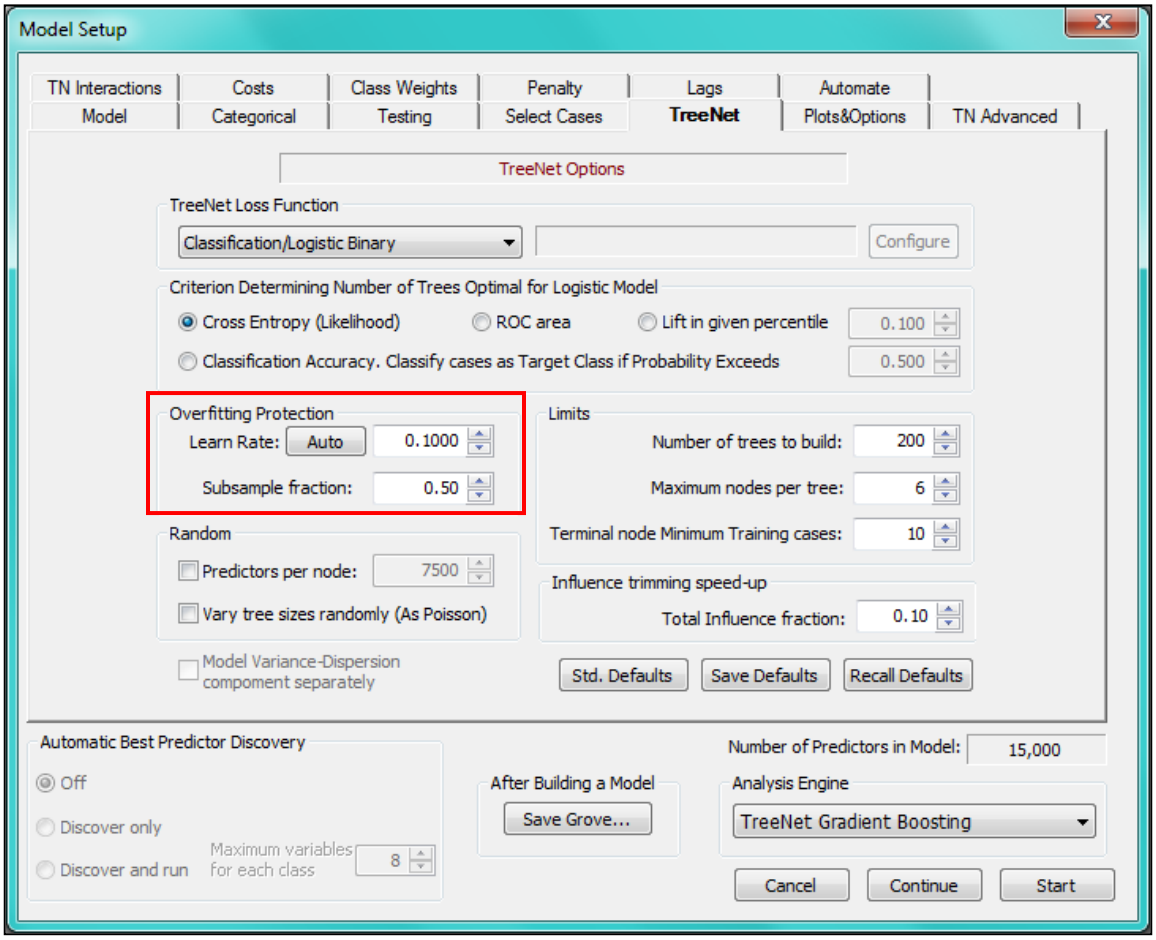
In the Variable Selection pane, choose Appetency in the “Target” column and VAR1-VAR15000 in the “Predictor” column. Choose TreeNet Gradient Boosting as the Analysis Engine and Classification/Logistic Binary as the Target Type. The Model tab should now match the picture above.

Click the Testing tab at the top of the Model Setup window:



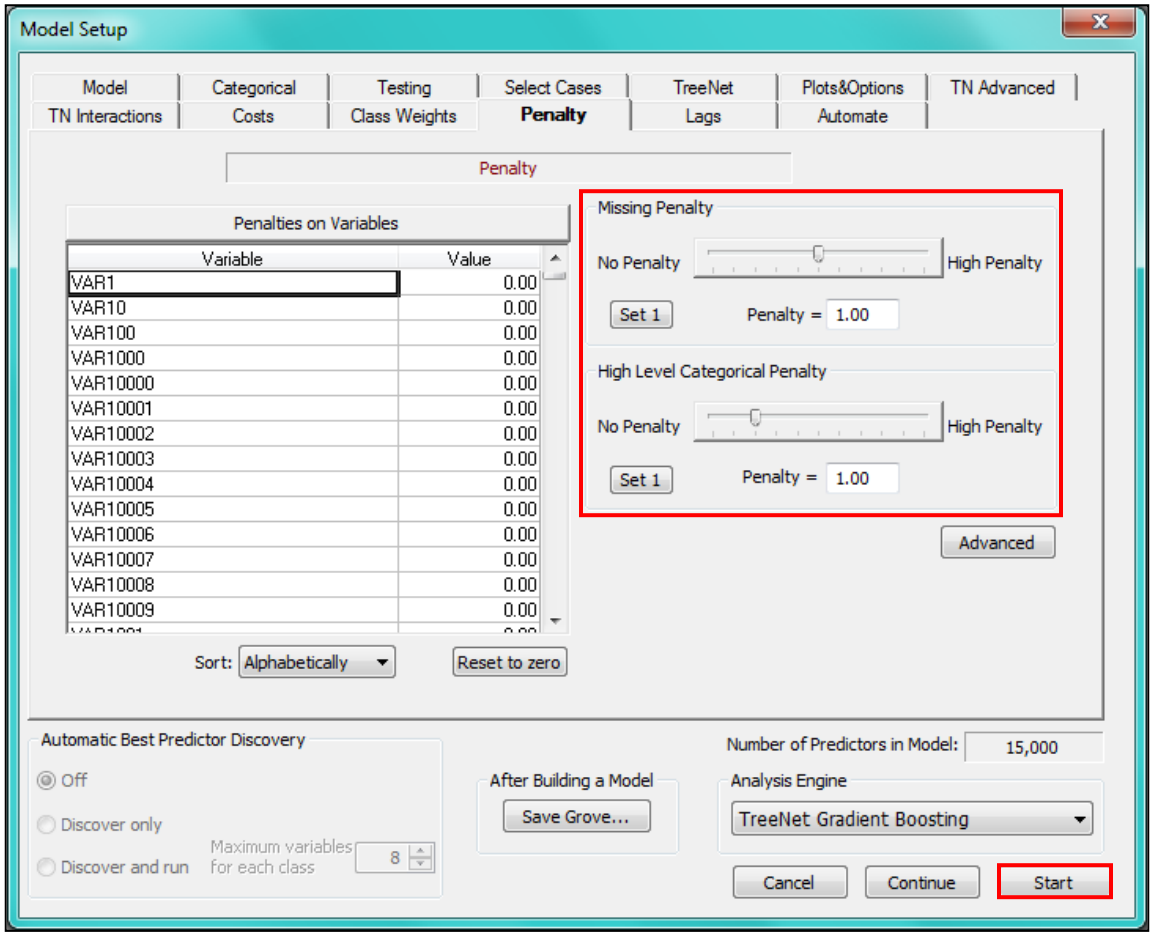
With such a large data set, cross-validation will be computationally expensive. Instead, choose “Fraction of cases selected at random” and enter 0.20 in the box. This will randomly partition the data into 80% for a training sample and 20% for testing.

Click the TreeNet tab:

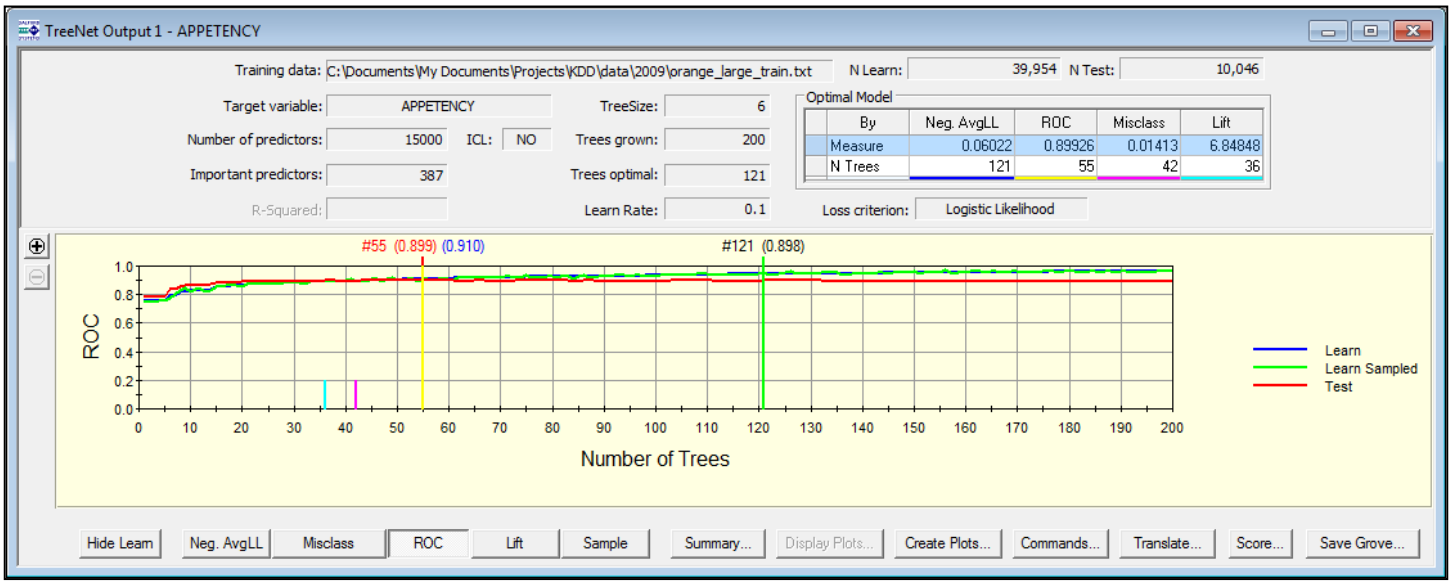


Set the Learn Rate to 0.1 for overfitting protection.

Click the Penalty tab:

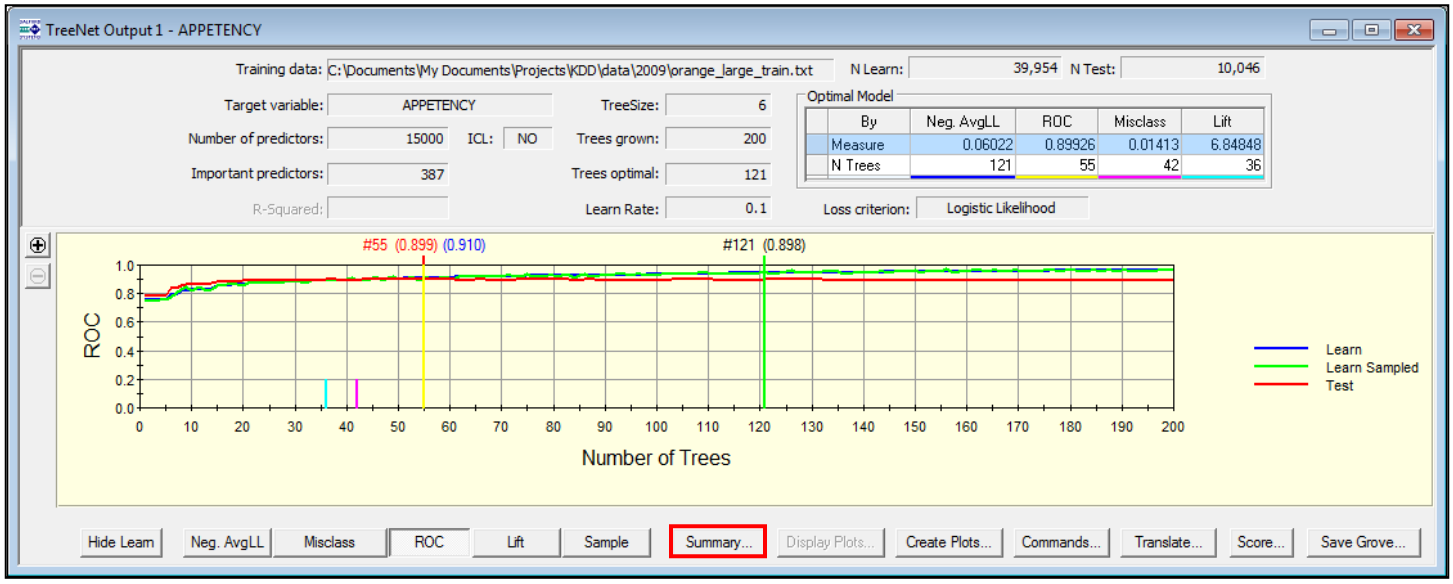


Click the Set 1 button for both missing and high level categorical variables. Click **Start**.

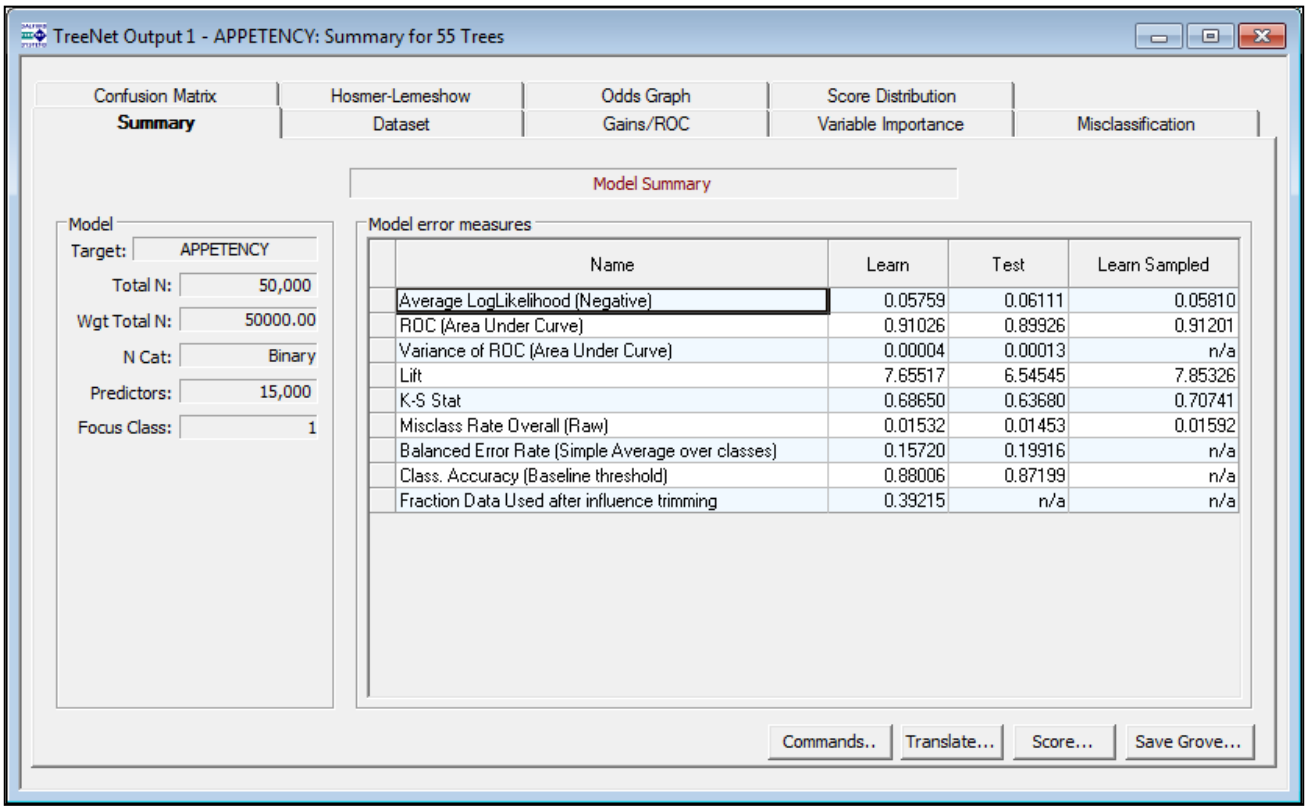


You have now created a preliminary model for Appetency with an ROC of 0.89926. You will now see how a few more steps can drastically improve this result.

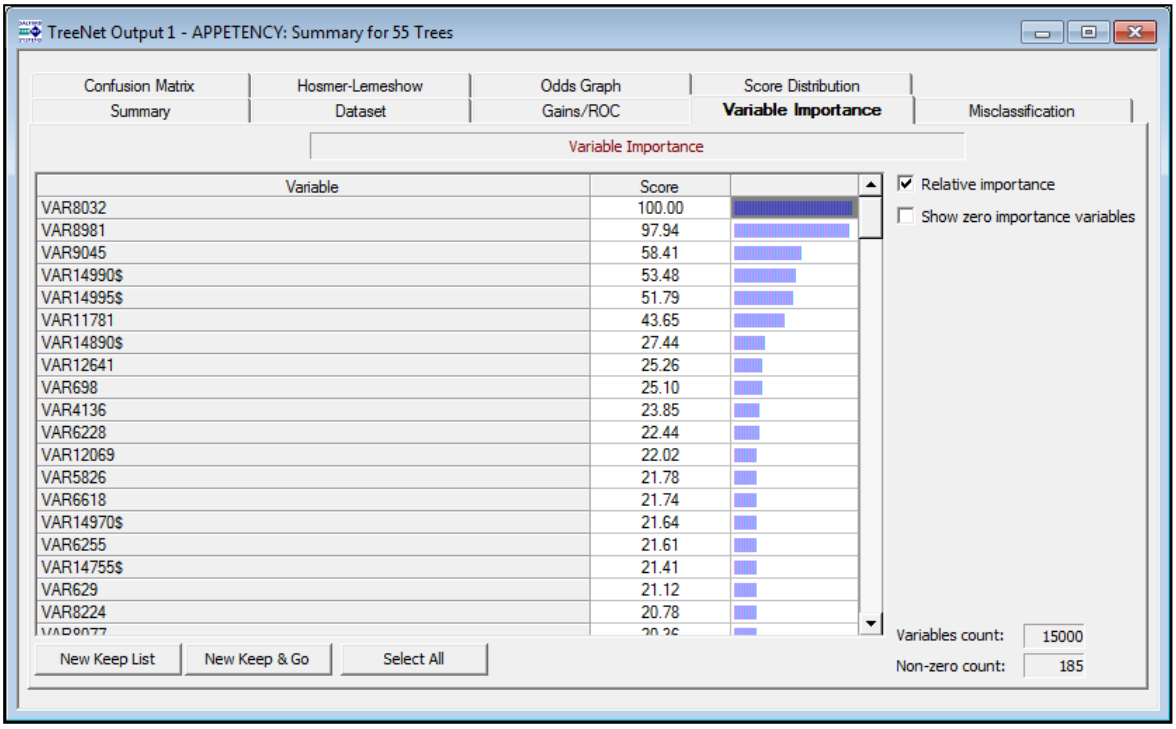
Click the Summary button at the bottom of the Navigator window:



The Summary Results window, pictured below, will appear:

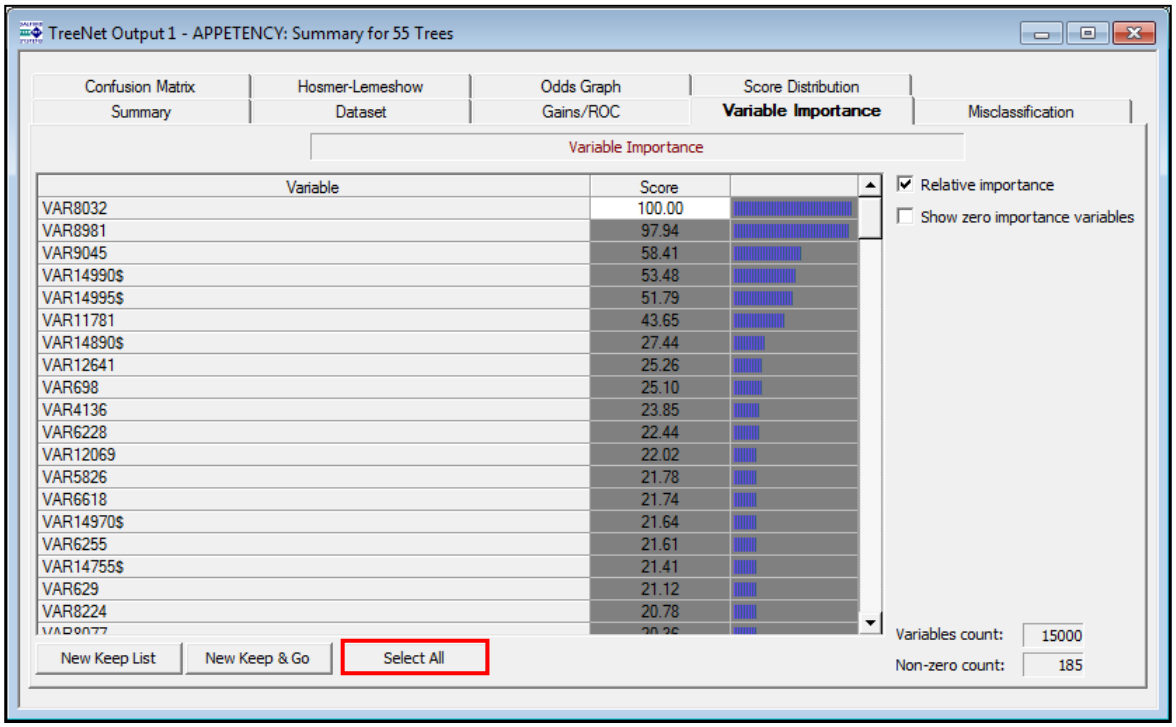


Click the Variable Importance tab:

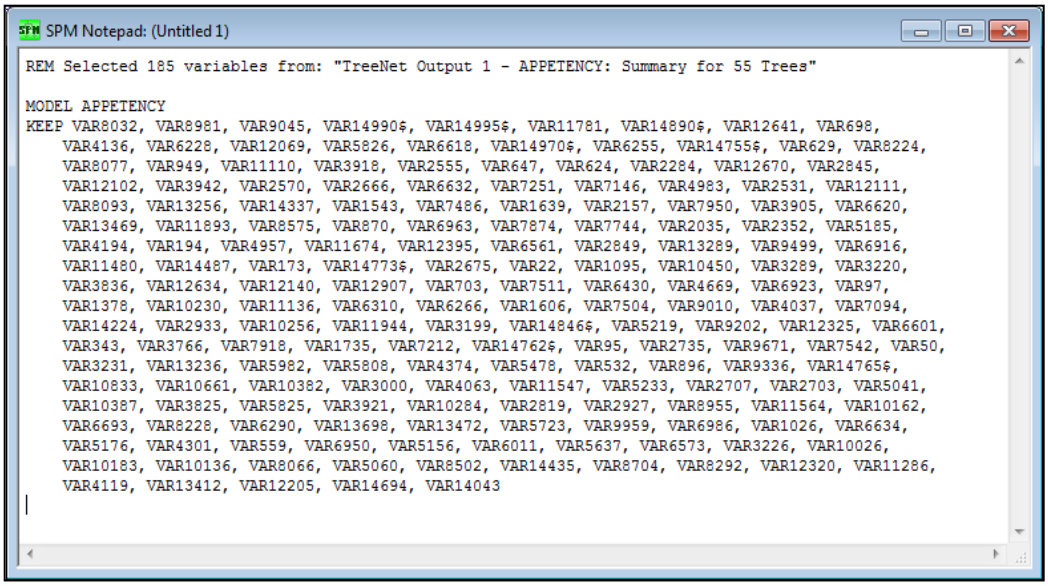


This tab shows which of the 15,000 predictors contributed to your initial model. Not all predictors were important, so you will only keep the necessary ones.

Click Select All to highlight the important predictors:

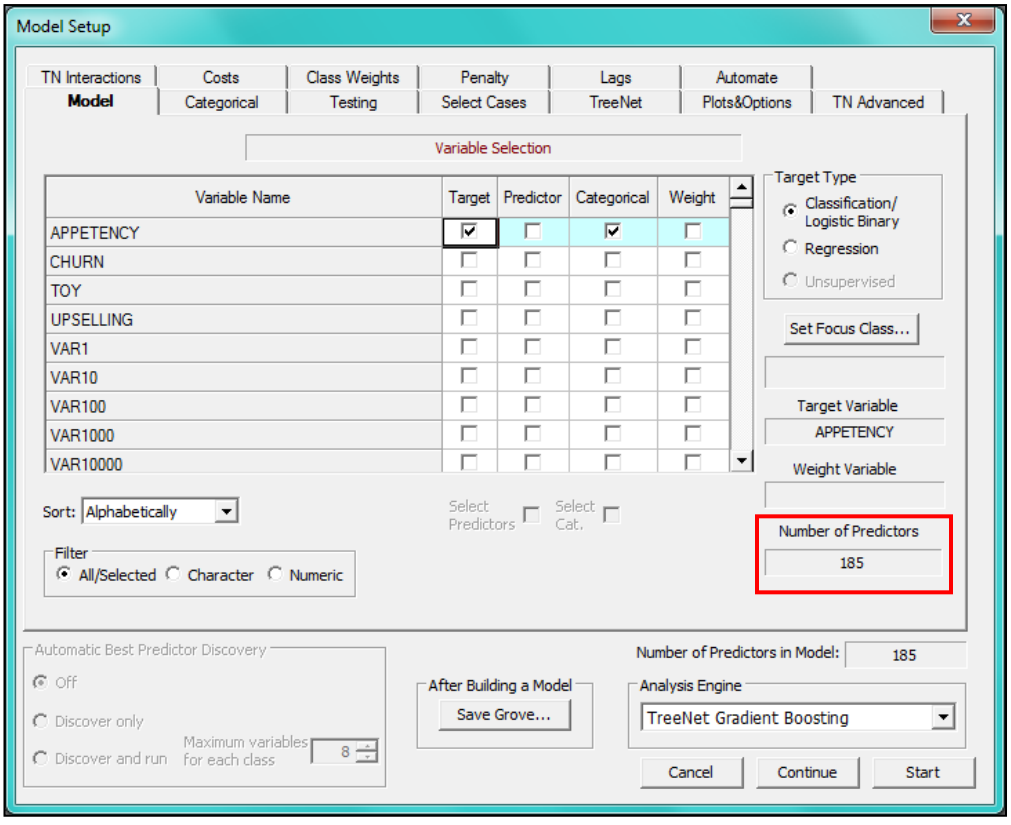


Click New Keep List to bring up a notepad with these variables:

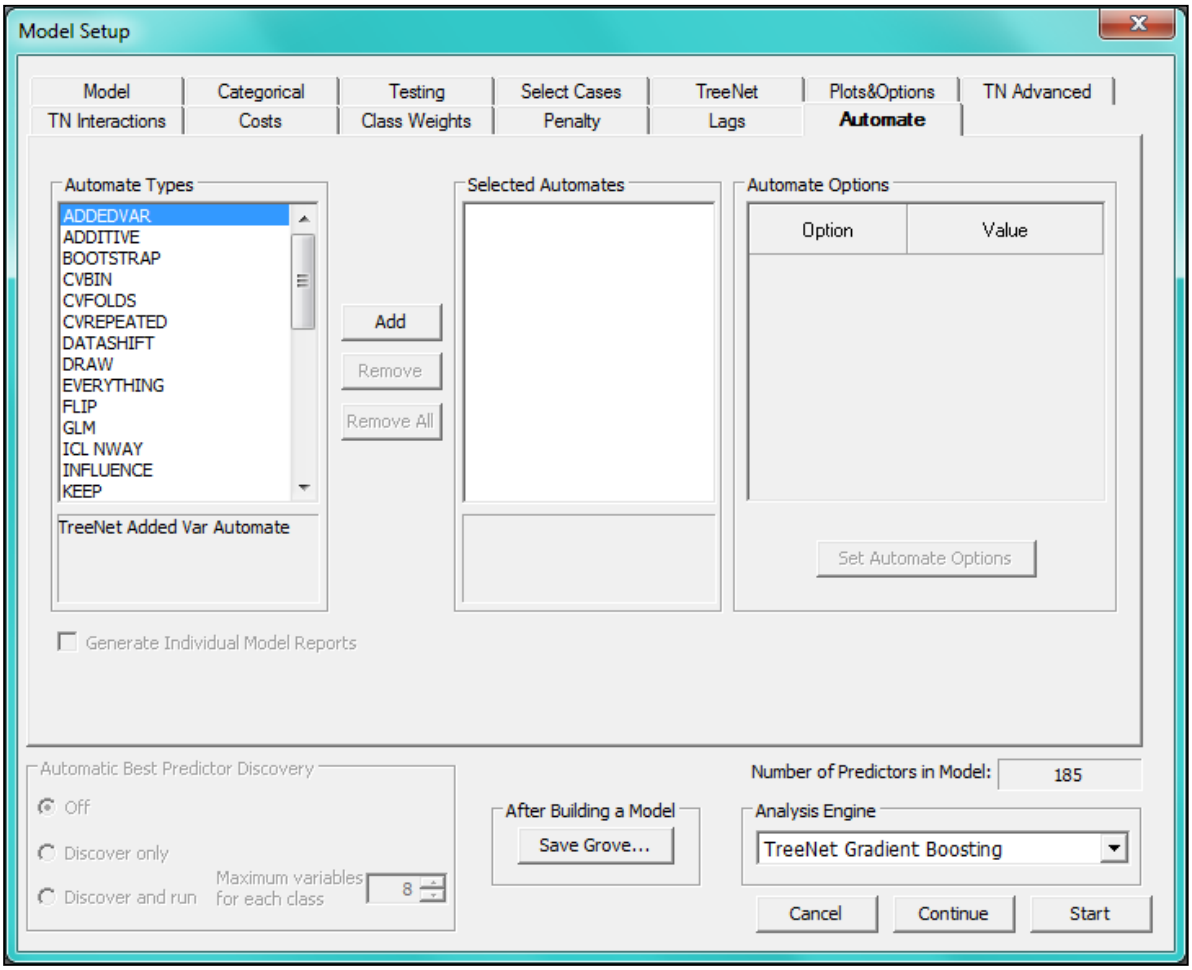


Click **File > Submit Window**. 

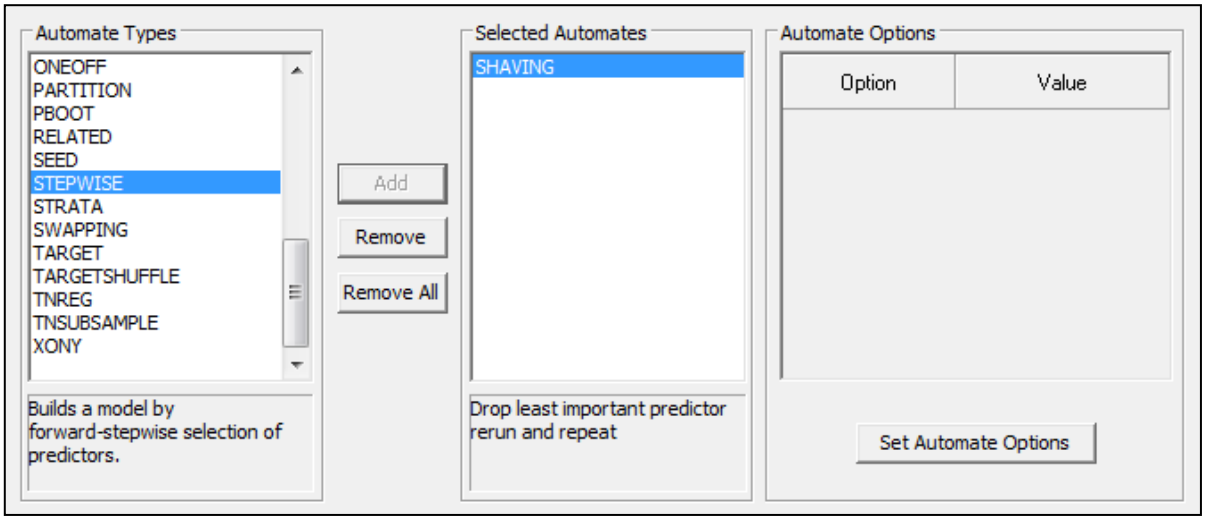
This action tells the engine to include only those submitted predictors in the Model Setup window. Re-open  Model Setup to verify:



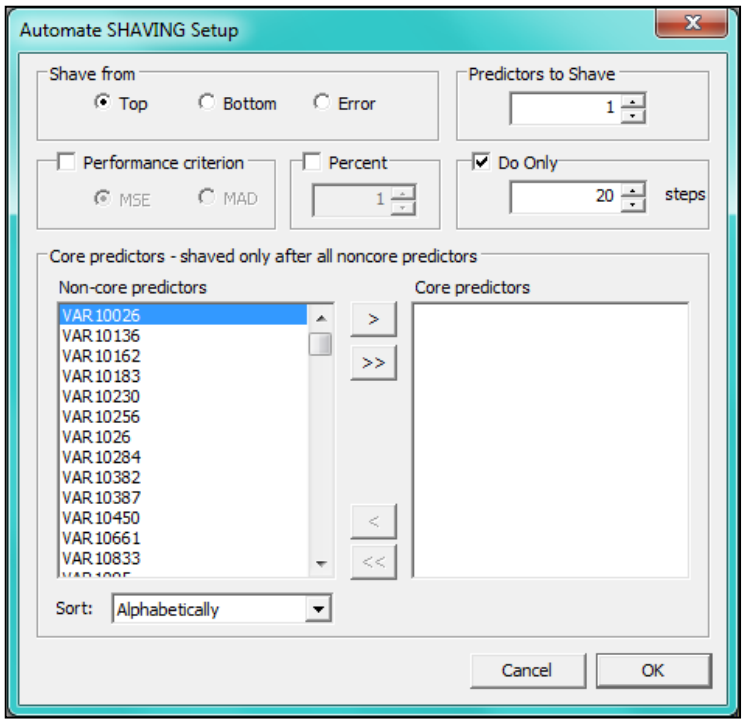
Click the Automate tab:



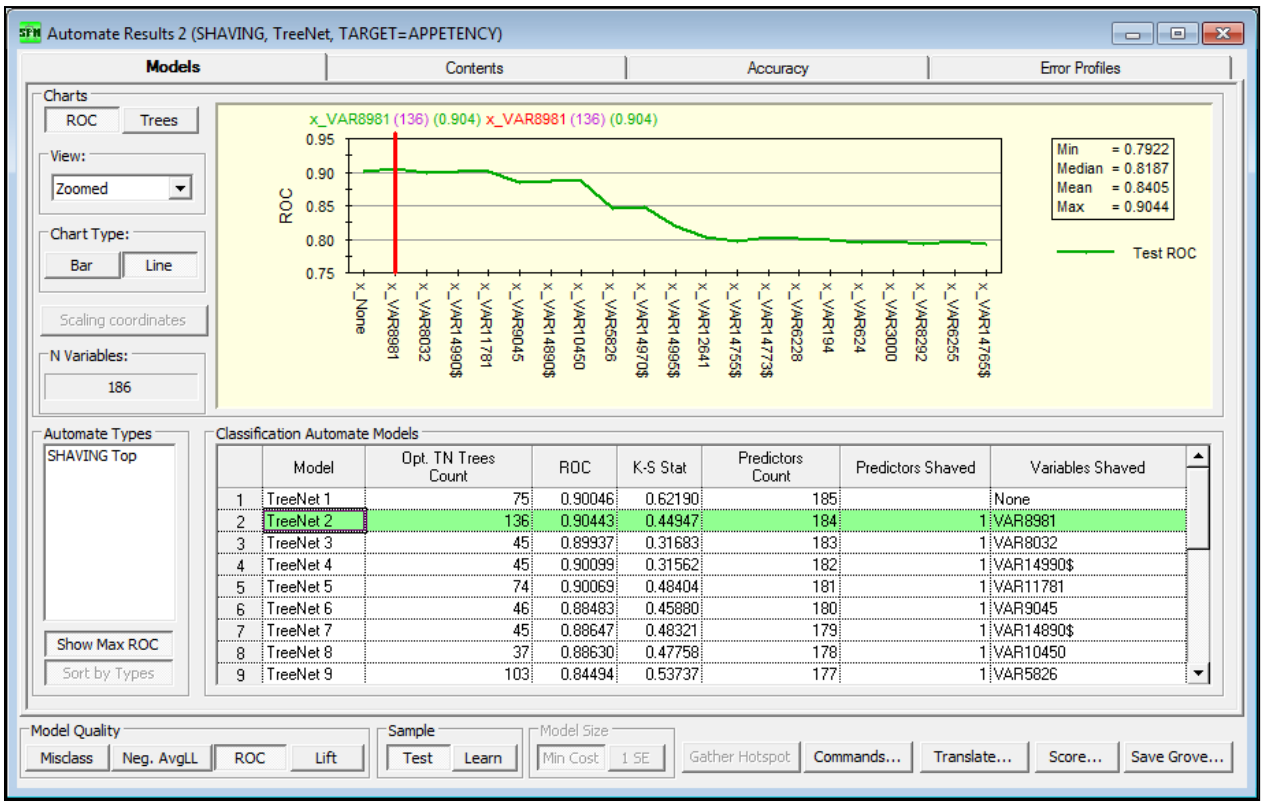
From the list of automates on the left, select SHAVING and click Add:



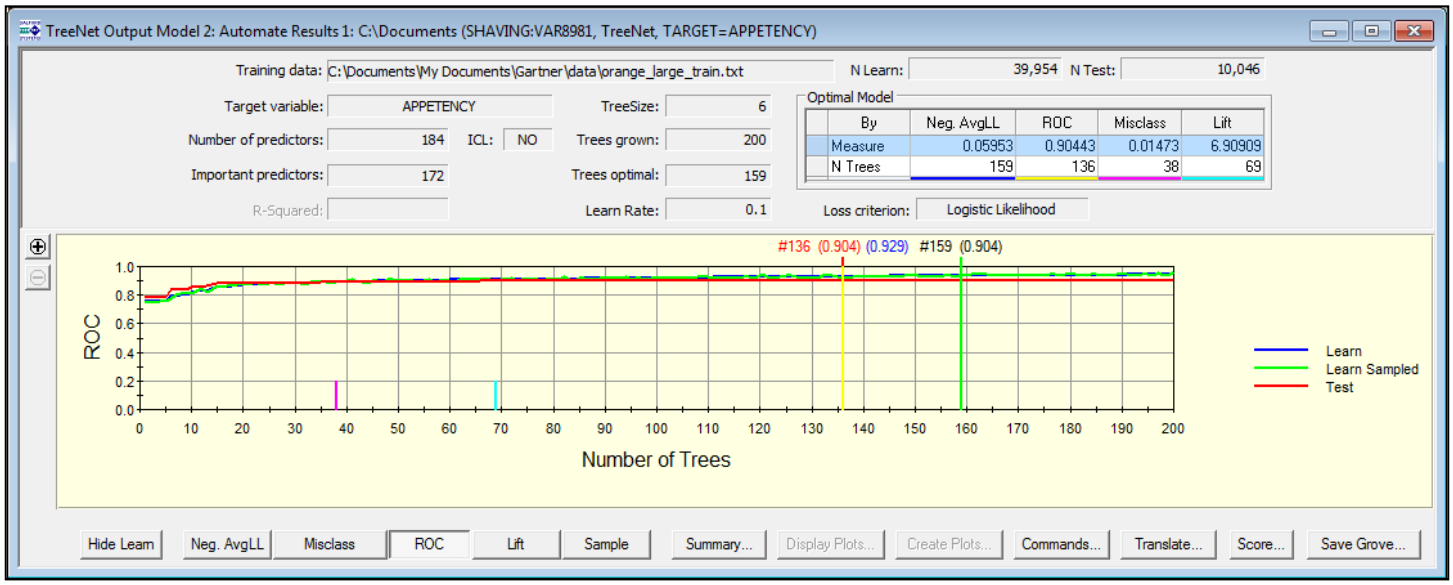
Click Set Automate Options:



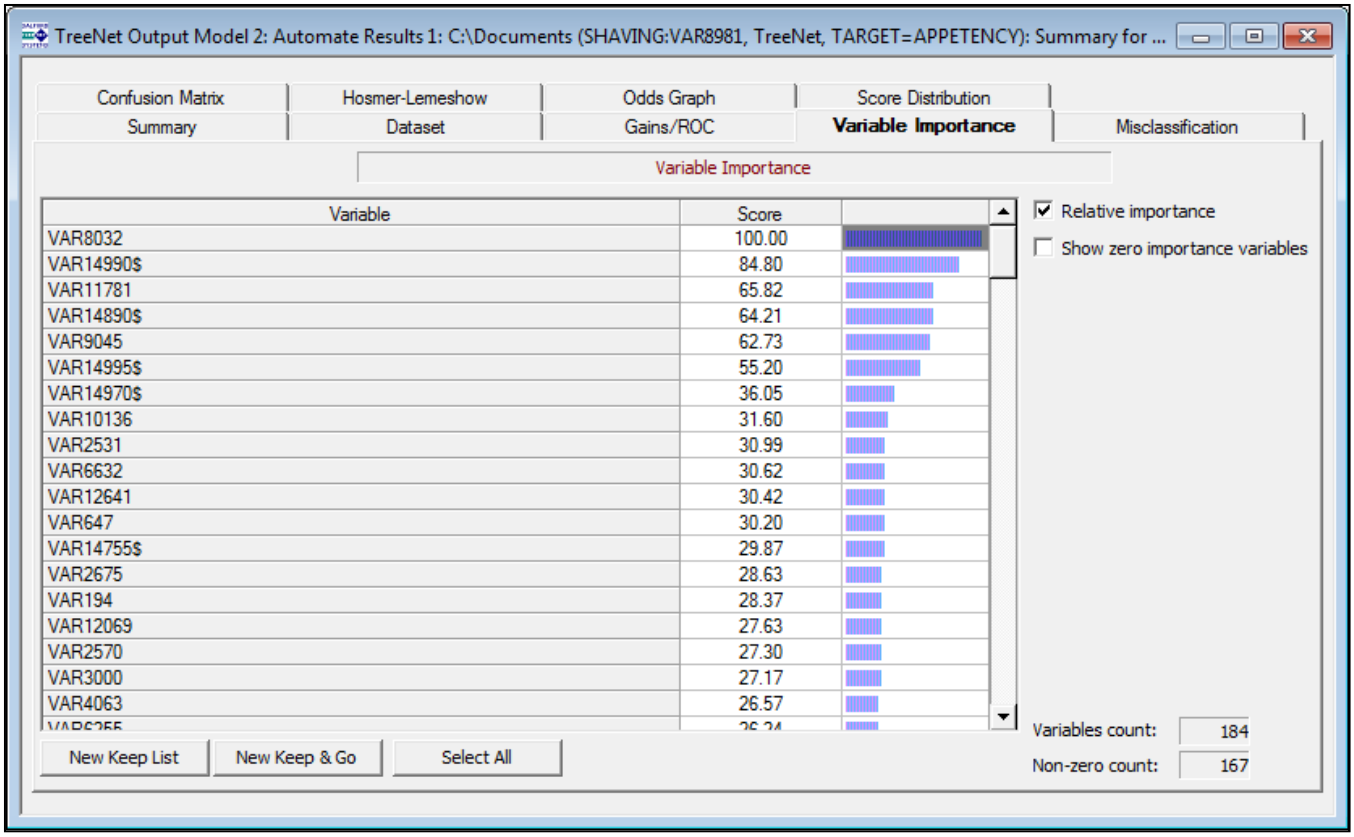
Select shaving from the top for only 20 steps, as above. Click **OK** and **Start.**



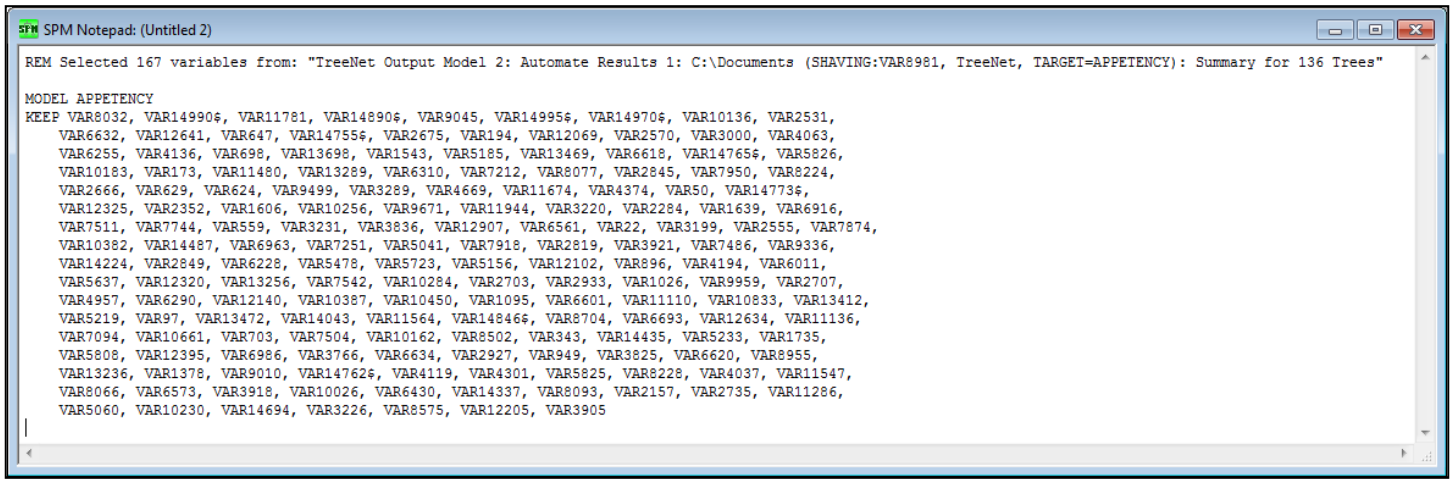
The results window, pictured above, gives a summary of each model created. Find the model with the highest ROC, 0.90443, and click to open:



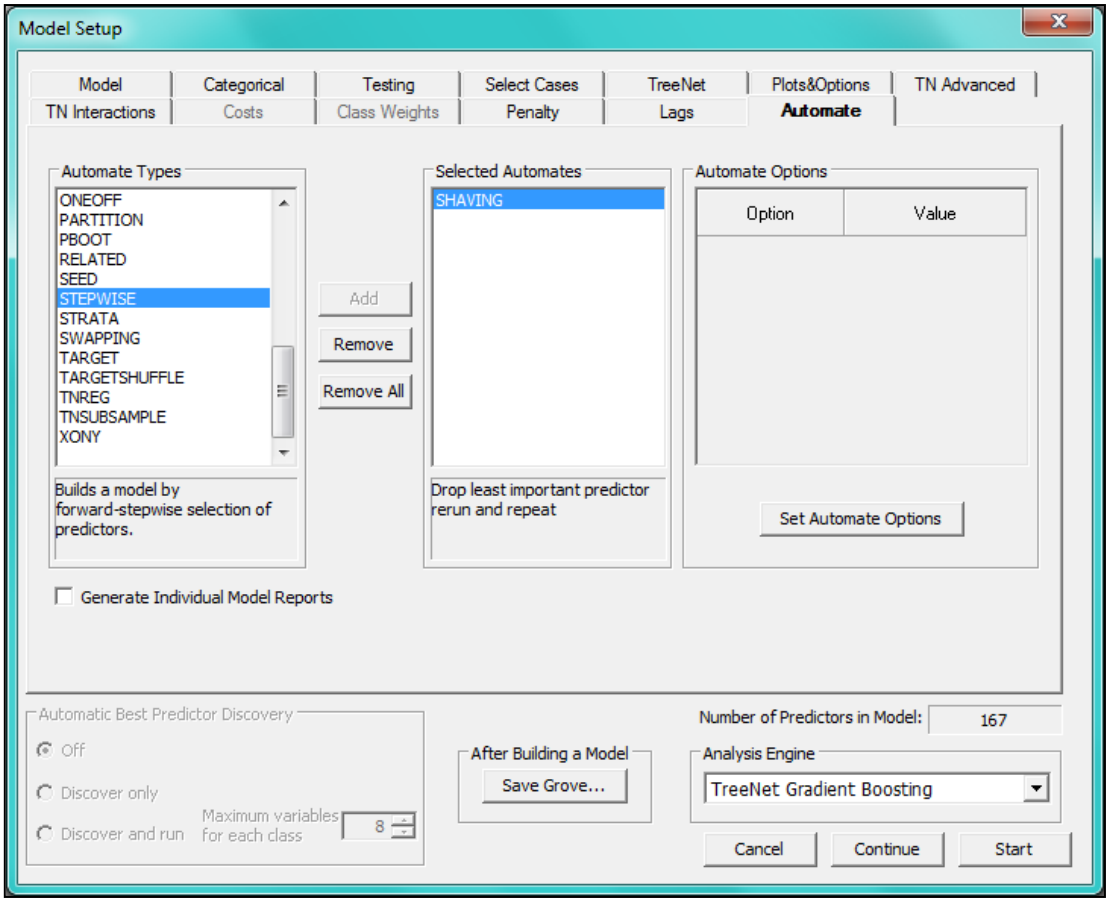
Navigate to the Variable Importance tab via the Summary button:



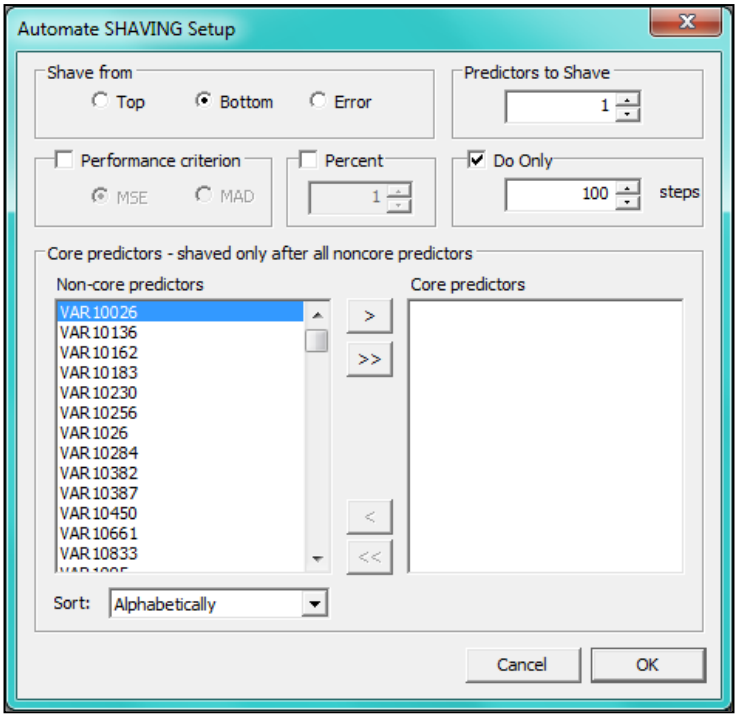
As before, Select All important predictors, create New Keep List, and Submit Window:



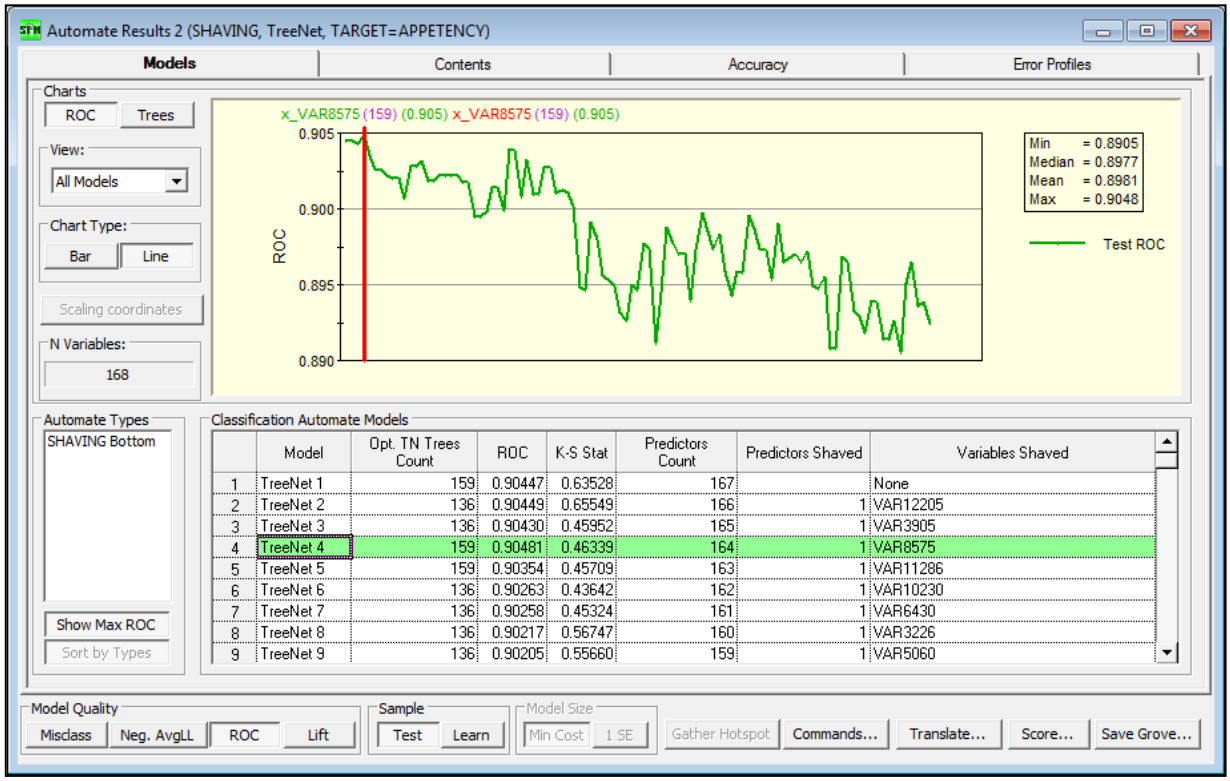
Re-open Model Setup and navigate to the Automate tab:



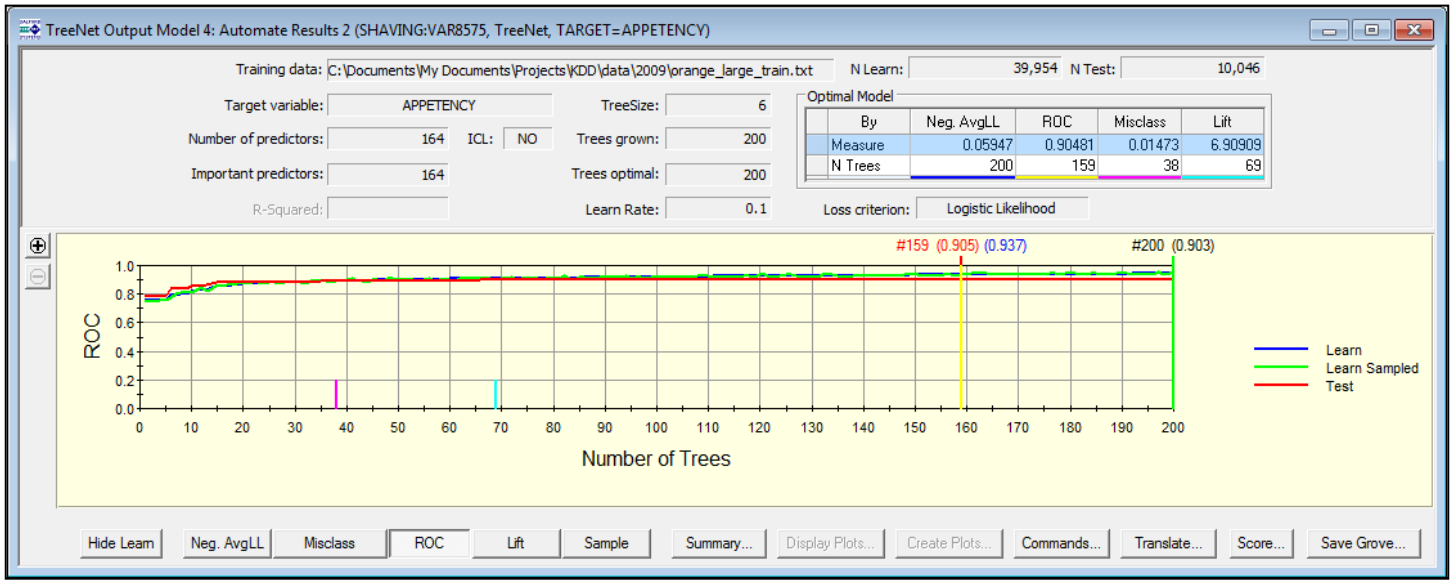
This time, add SHAVING and set the options to shave from the bottom for only 100 steps:



Click **OK** and **Start**.



Open the model with the highest ROC:



Congratulations! In just a few simple steps, you received an ROC of 0.90481 for the Appetency model. Repeat these same steps to create your own Churn and Upselling models. Once finished, average your 3 ROC values and see how you placed below! ***

Fast Track – 5 Days					
Rank	**Team**	**Appetency**	**Churn**	**Upselling**	**Score**
1	IBM Research	0.8830	0.7611	0.9038	0.8493
–	You!	0.9048	0.7320	0.9059	0.8476
2	ID Analytics, Inc.	0.8724	0.7565	0.9056	0.8448
3	Old dogs with new tricks	0.8740	0.7541	0.9050	0.8443
4	Crusaders	0.8688	0.7569	0.9034	0.8430
5	Financial Engineering Group, Inc. Japan	0.8732	0.7498	0.9057	0.8429

Slow Track – 2 Months					
Rank	**Team**	**Appetency**	**Churn**	**Upselling**	**Score**
1	IBM Research	0.8819	0.7651	0.9092	0.8521
2	Uni Melb	0.8836	0.7570	0.9048	0.8484
3	ID Analytics, Inc.	0.8761	0.7614	0.9061	0.8479
4	Financial Engineering Group, Inc. Japan	0.8768	0.7589	0.9074	0.8477
–	You!	0.9048	0.7320	0.9059	0.8476
5	National Taiwan University, Computer Science and Information Engineering	0.8789	0.7558	0.9036	0.8461

* http://www.sigkdd.org/kdd-cup-2009-customer-relationship-prediction

**This data set may be too large for your system to process; try the tutorial with orange_small_train.txt instead.

*** These results are meant to demonstrate quick and easy data mining in SPM for users of all levels. With more time, advanced users were able to reach a score of 0.8610 with TreeNet. See if you can beat them!