**Project Part 3**

Durgaprasad Koppala, Sri Sai Sathya Keshav Arigela, Sree Harshita Reddy Gangasani, NikhilVishnu Teja Madasu , Jyothi Naga Lakshmi Addala and Dileep Durgam

Northern Illinois University

OMIS 681: Predictive Business Analytics

Prof. Karah Osterberg

October 16, 2023

1. This builds off of project-part 1 and project-part 2. If I mentioned anything you need to fix about part 2 of the project, please do so before proceeding here.

No changes made to the project-part 1 and 2.

1. Go through all the **tools for data modification** nodes we went over (Drop, Replacement, Impute, Interactive Binning, and Principal Components). Would it make sense to use any of these nodes for your data? If so, please add it onto the diagram and explain why the node was added and what was done.

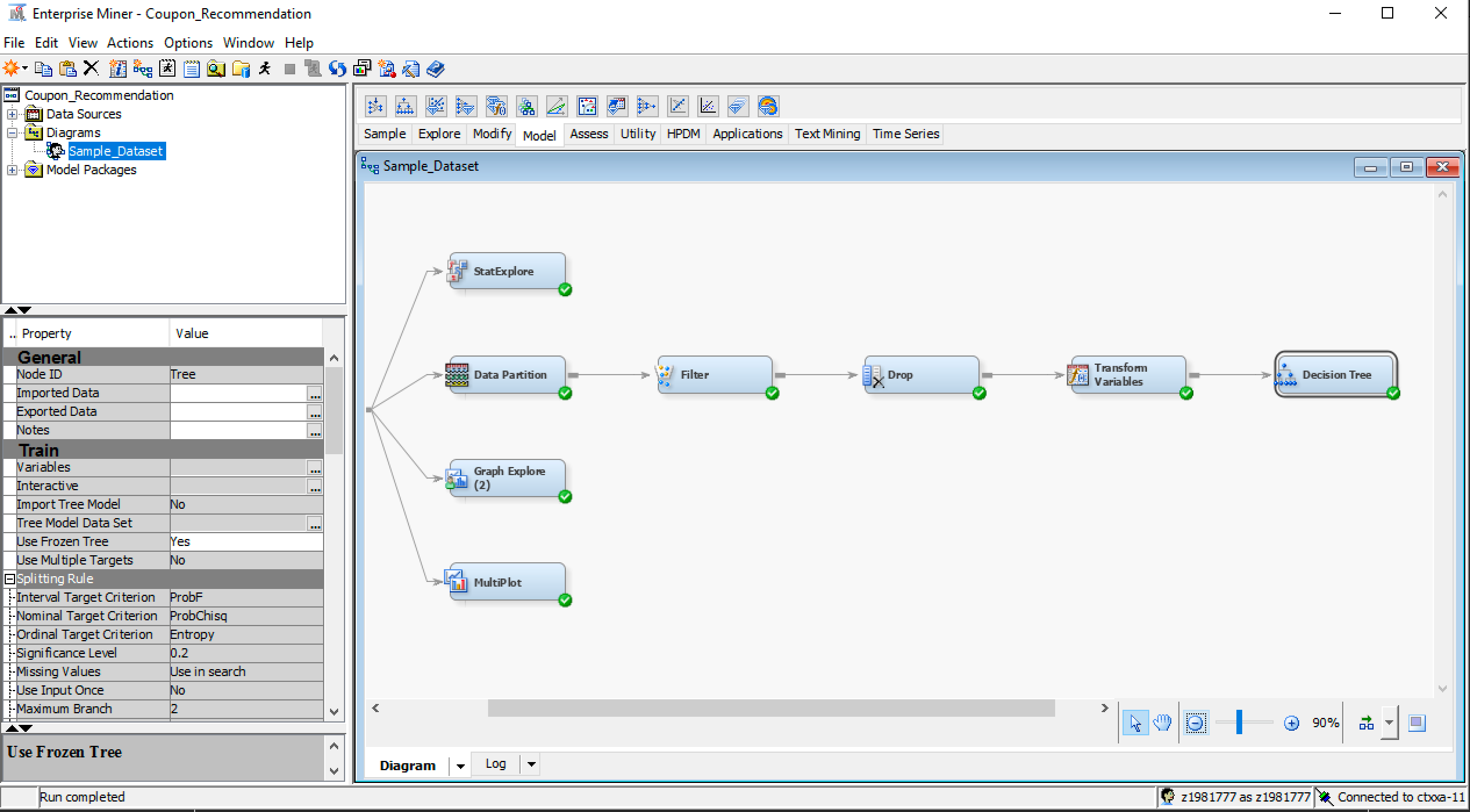
No changes made to the project-part 1 and 2. We did not used any of the data modifications nodes (Drop, Replacement, Impute, Interactive Binning, and Principal Components) in the project-part 3

1. Are you choosing to do variable selection or transformation of variables? What is your reasoning behind your decision? If you do choose to do either of these, add the nodes to the diagram and set the appropriate settings.

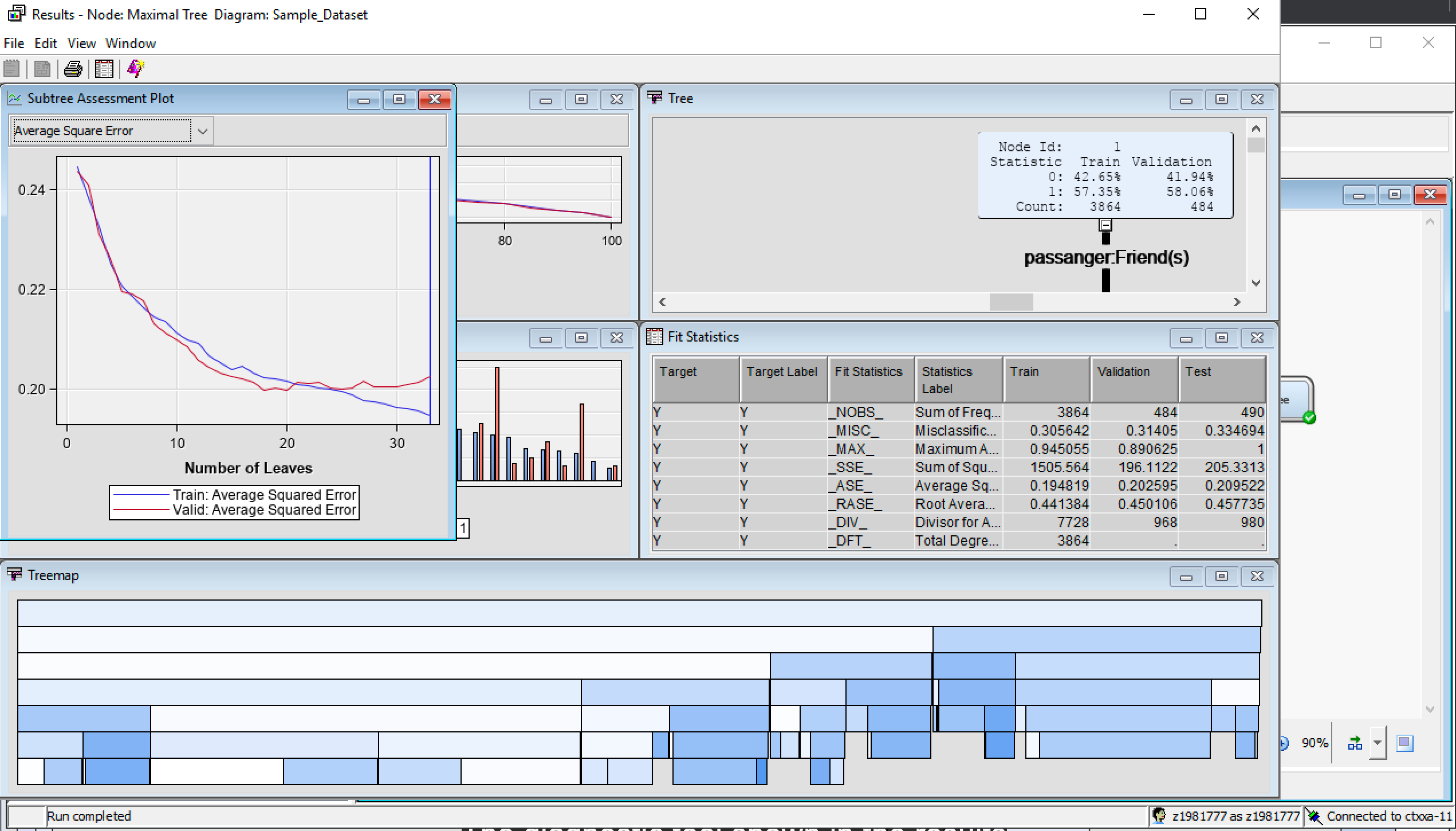
We decided to use transformation of variables instead of variable selection. Since variable selection was omitting most of the important variables when we ran the model.

1. Add a few decision tree models to the diagram and vary the settings in each one to produce different trees. The number of trees you can create can depend on the type of target variable you chose. (Changing settings may or may NOT change the decision tree output).
   1. What variables are important according to your best decision tree? (Note: You should look at model comparison and then answer this so you know which one it deemed the best model.)

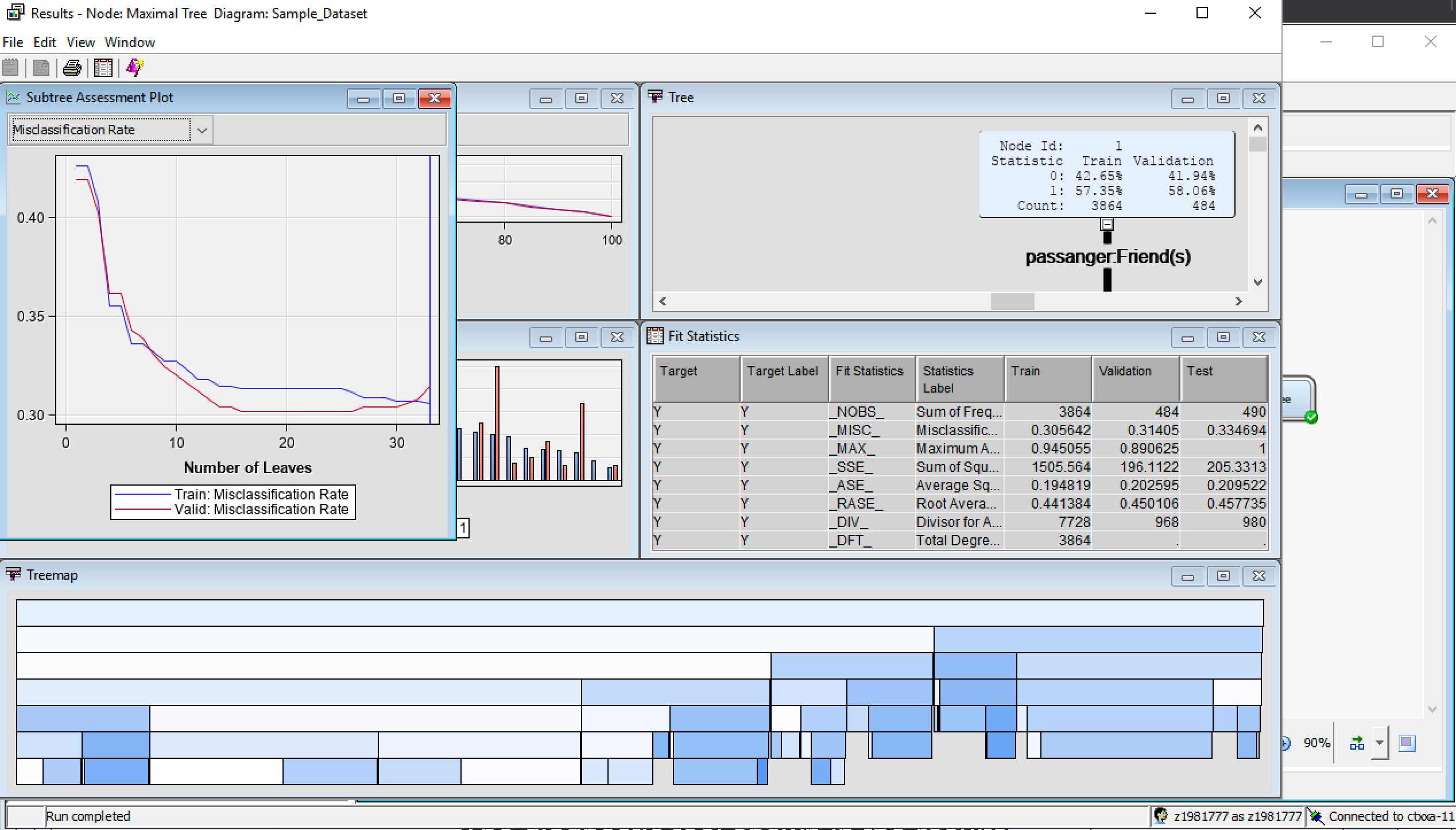
Decision Tree:



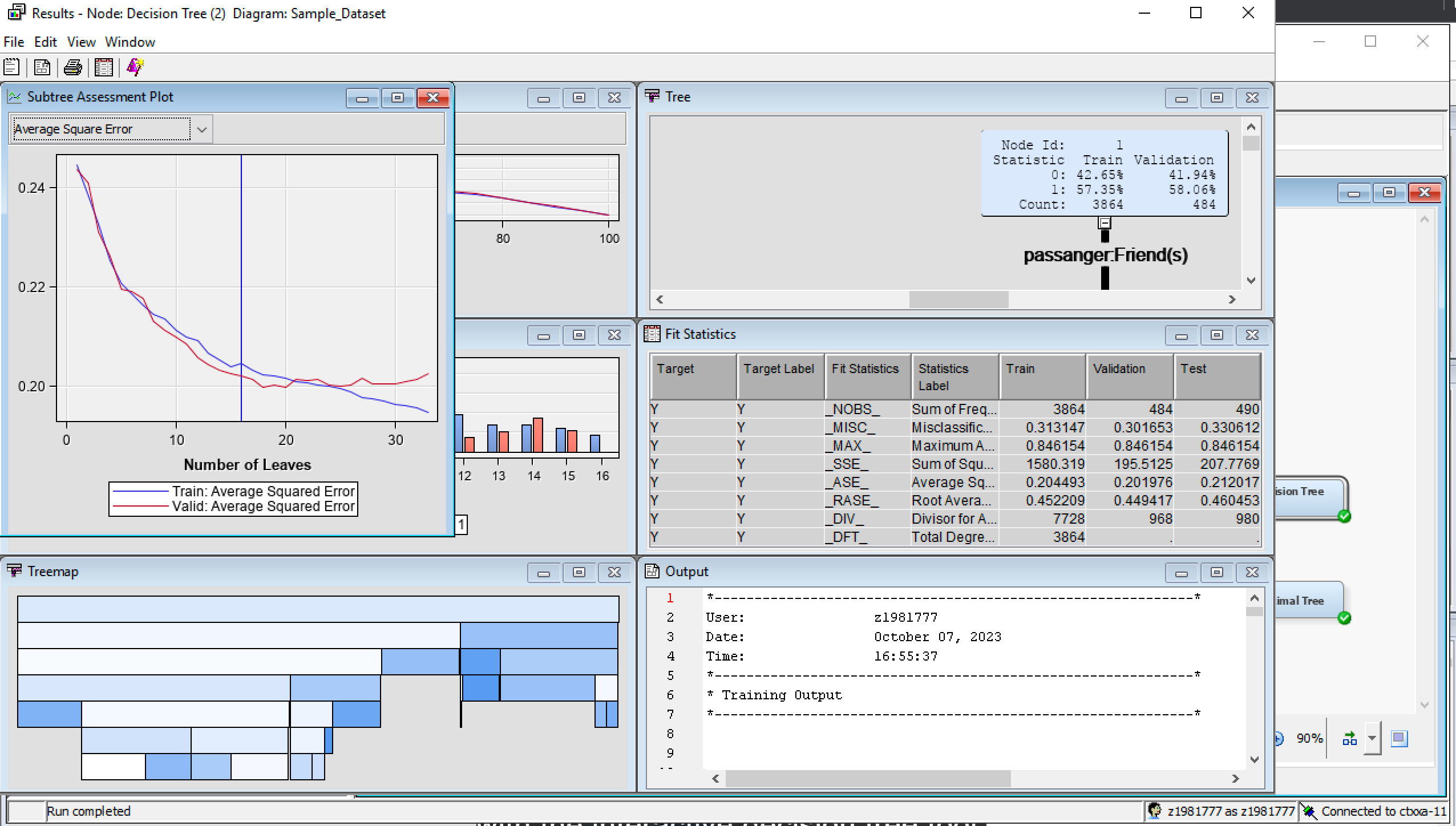
Maximal Tree: Average square error of subtree assessment plot as follows.



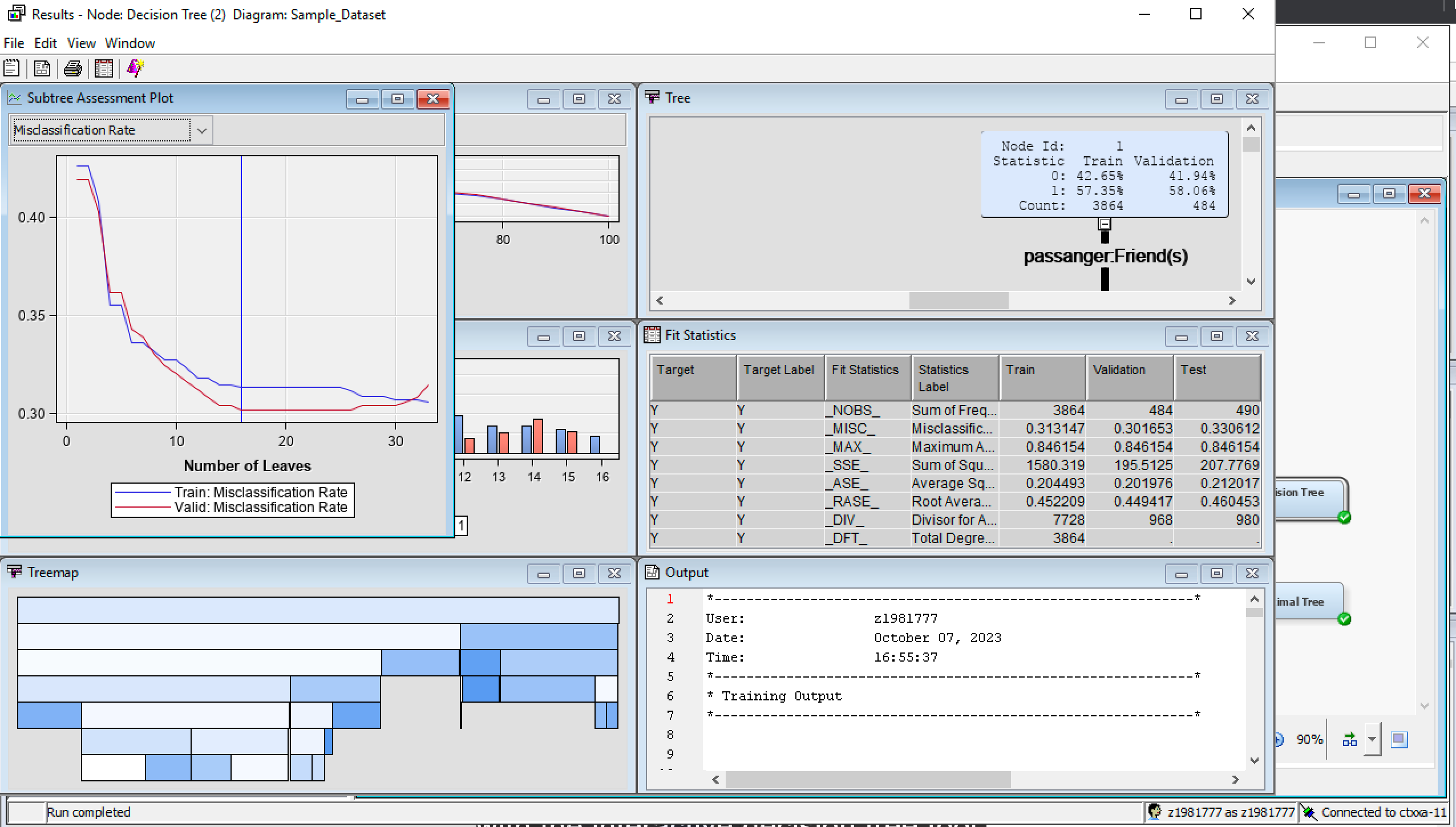
Maximal Tree: Misclassification rate of subtree assessment plot as follows.



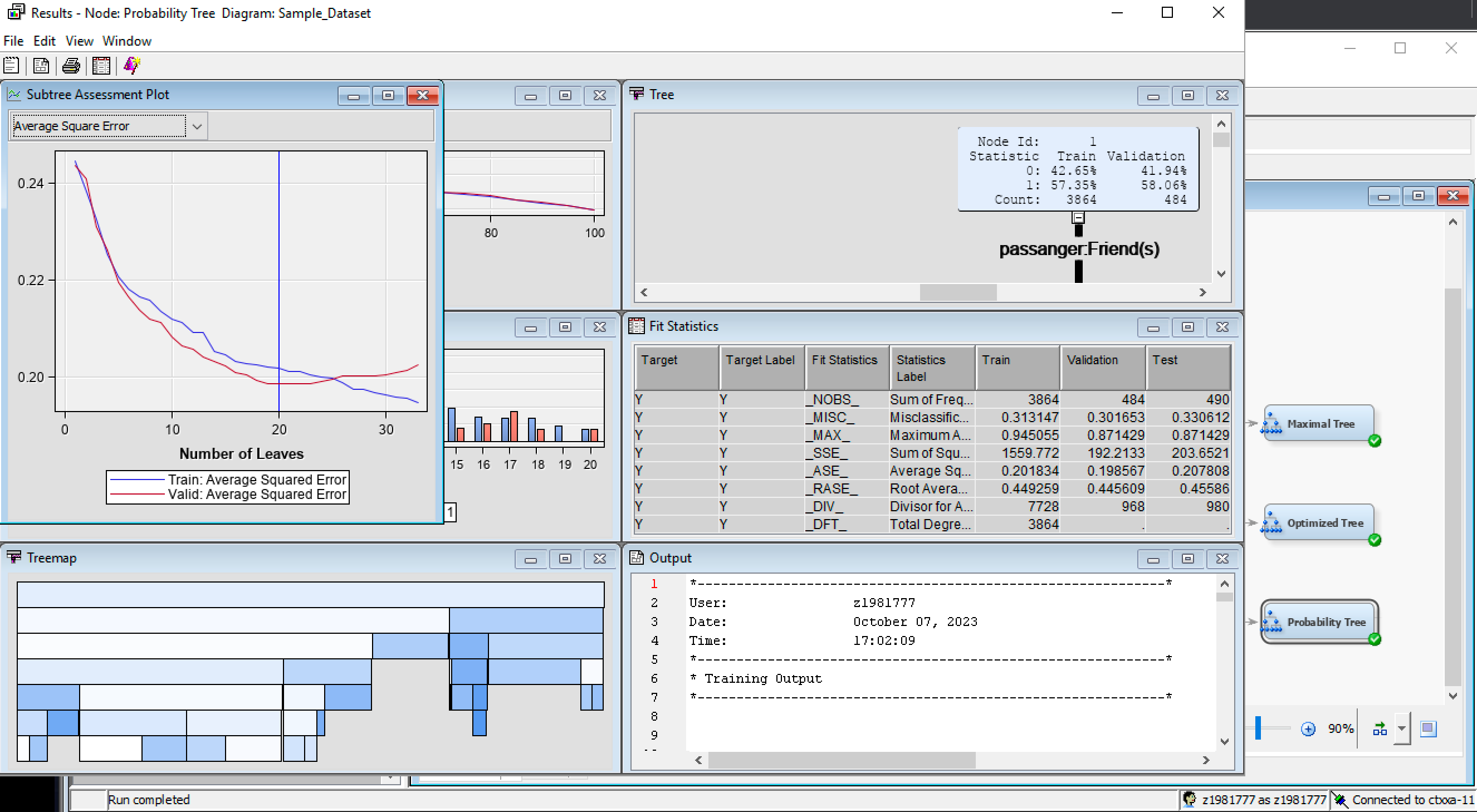
Optimized Tree: Average square error of subtree assessment plot as follows.



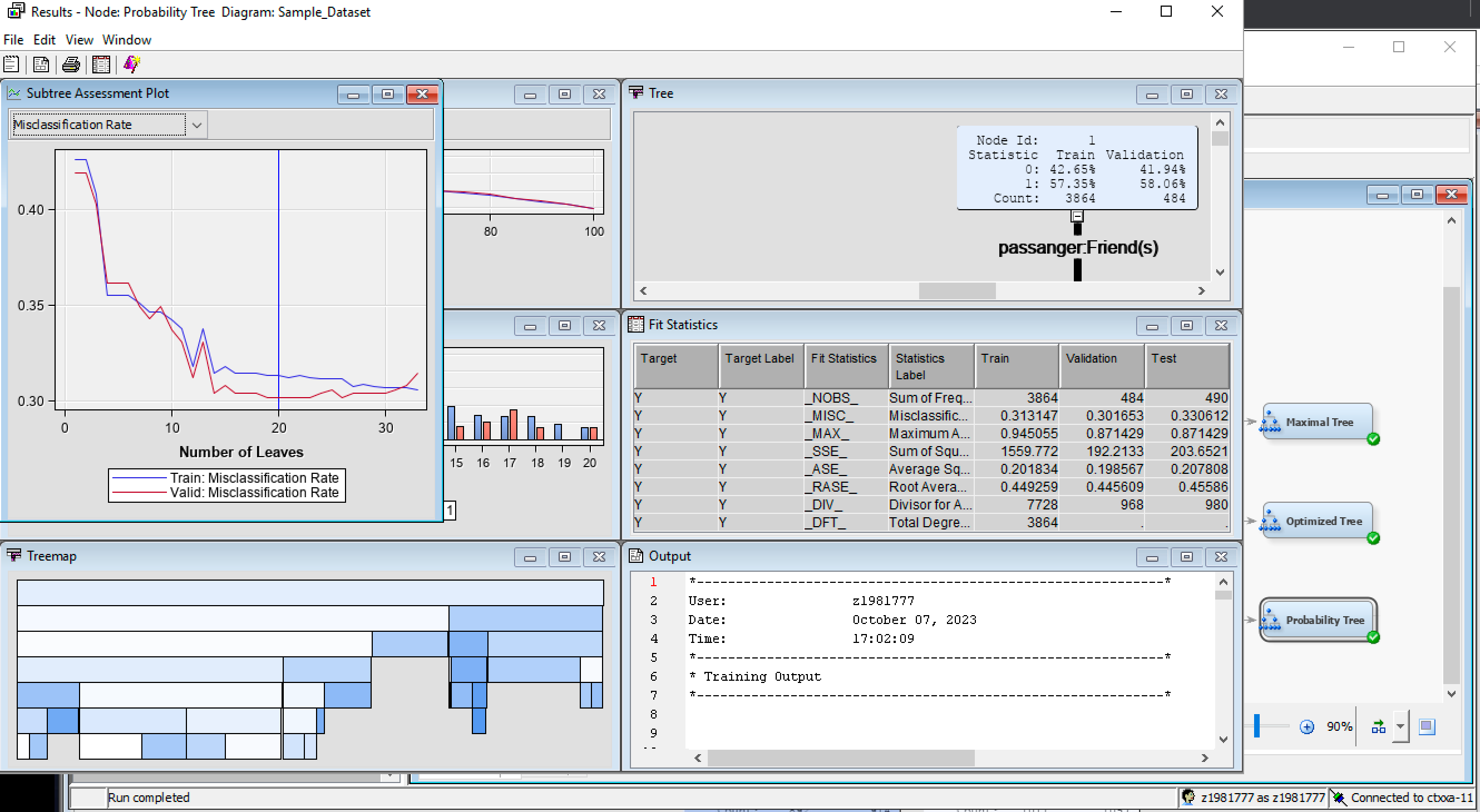
Optimized Tree: Misclassification rate of subtree assessment plot as follows.



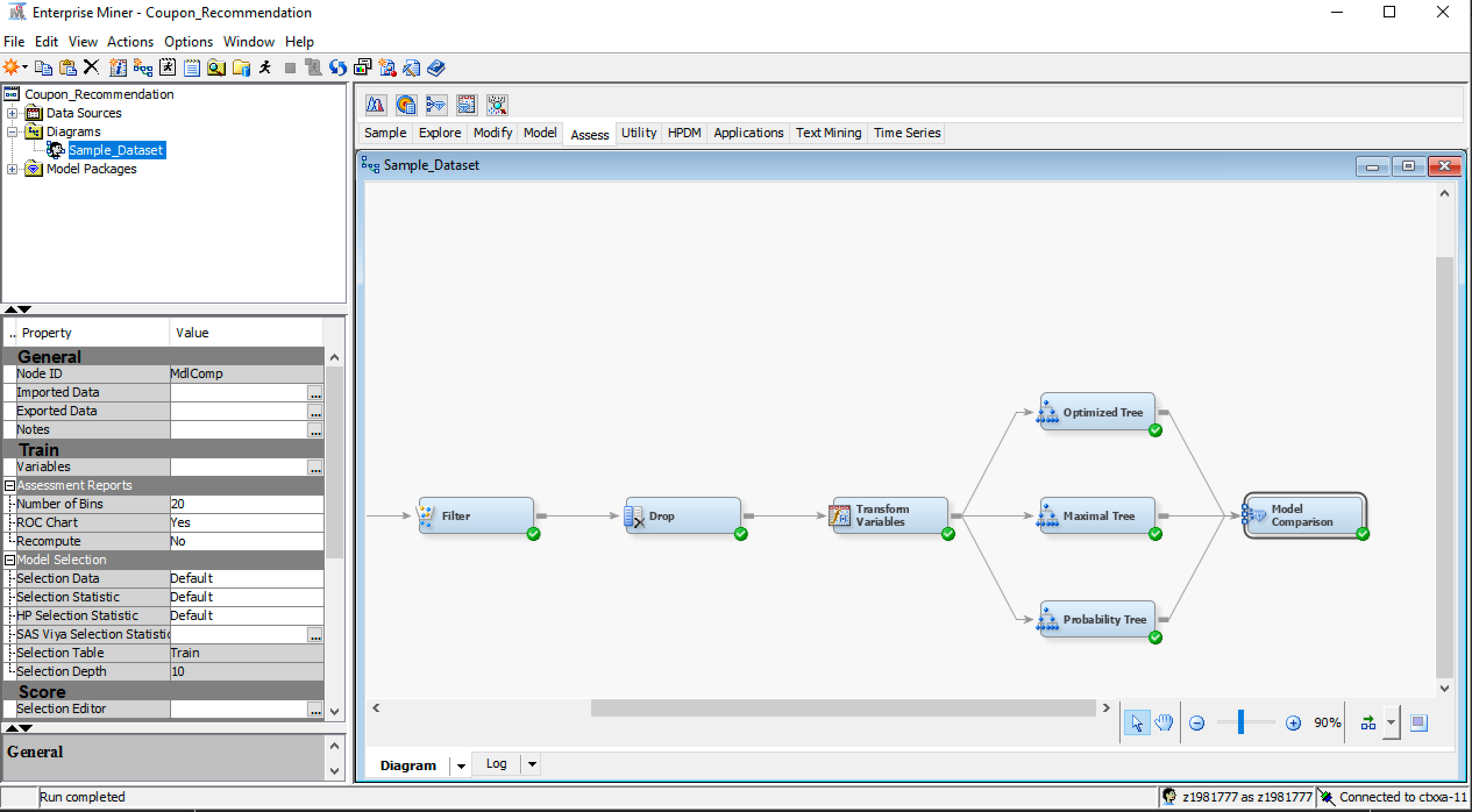
Probability Tree: Average square error of subtree assessment plot as follows.



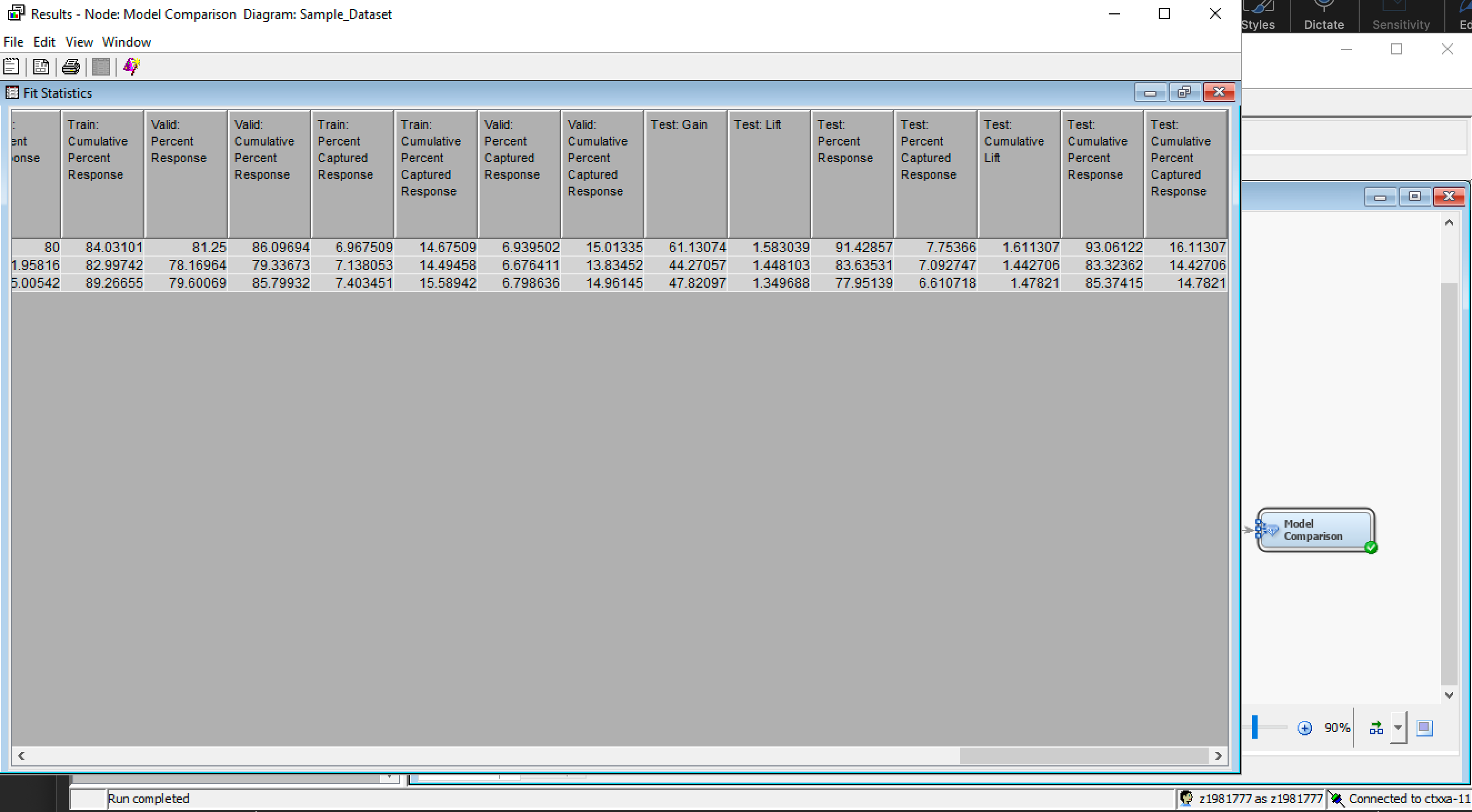
Probability Tree: Misclassification rate of subtree assessment plot as follows.

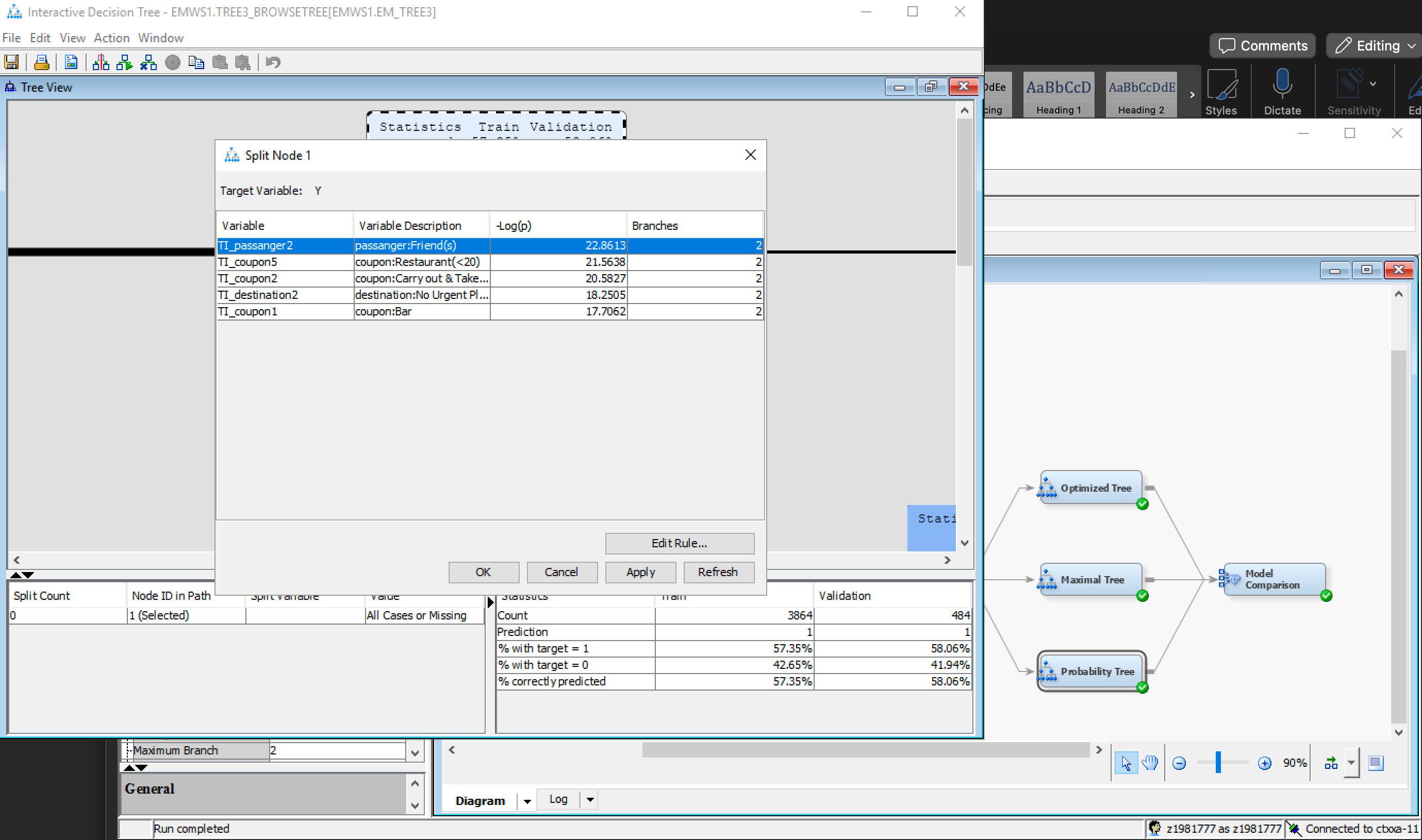


Model Comparison: A model comparison node is used for identifying the best decision tree.



As per the following results, the probability tree has the highest cumulative percent captured response in comparison with the other decision trees.



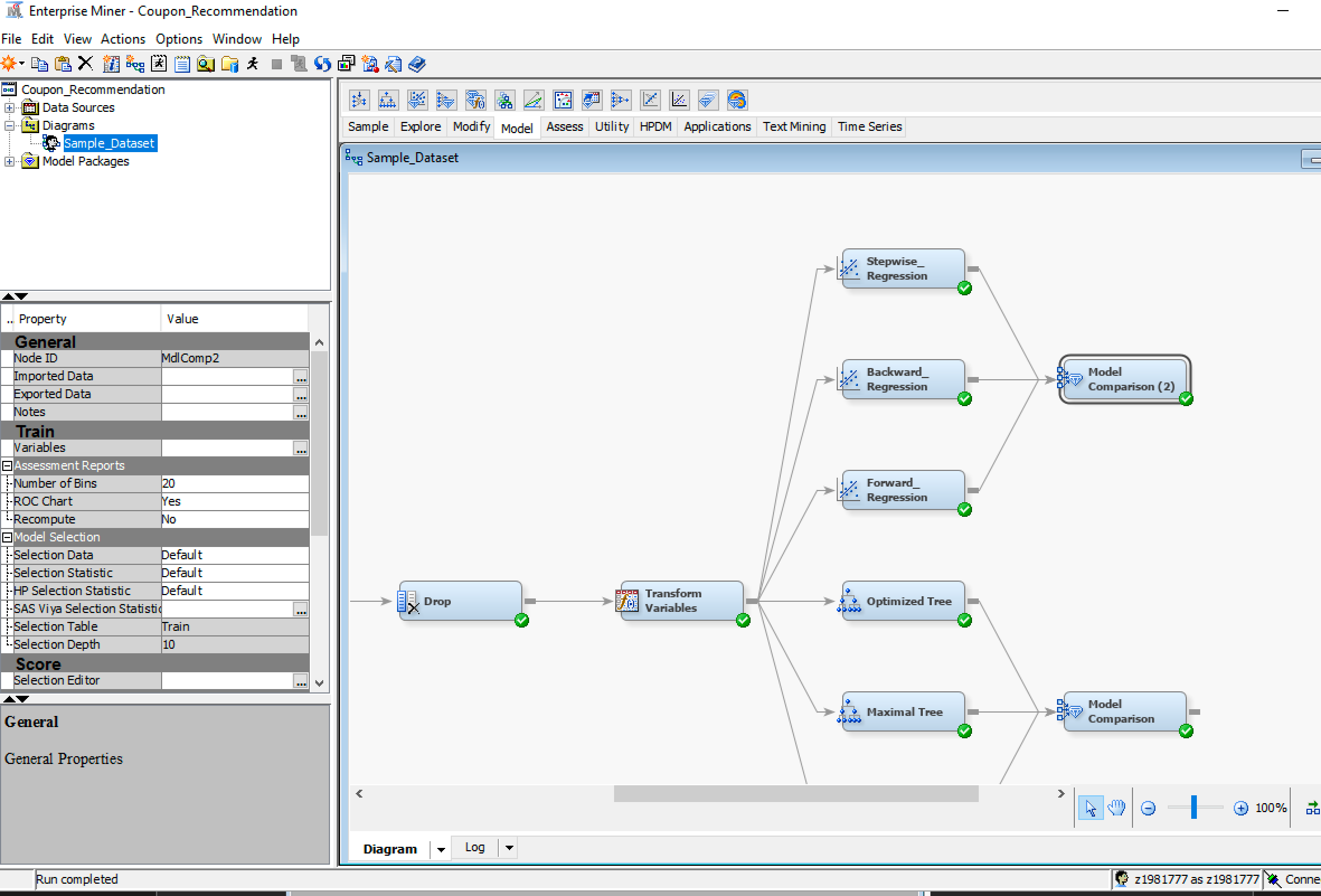


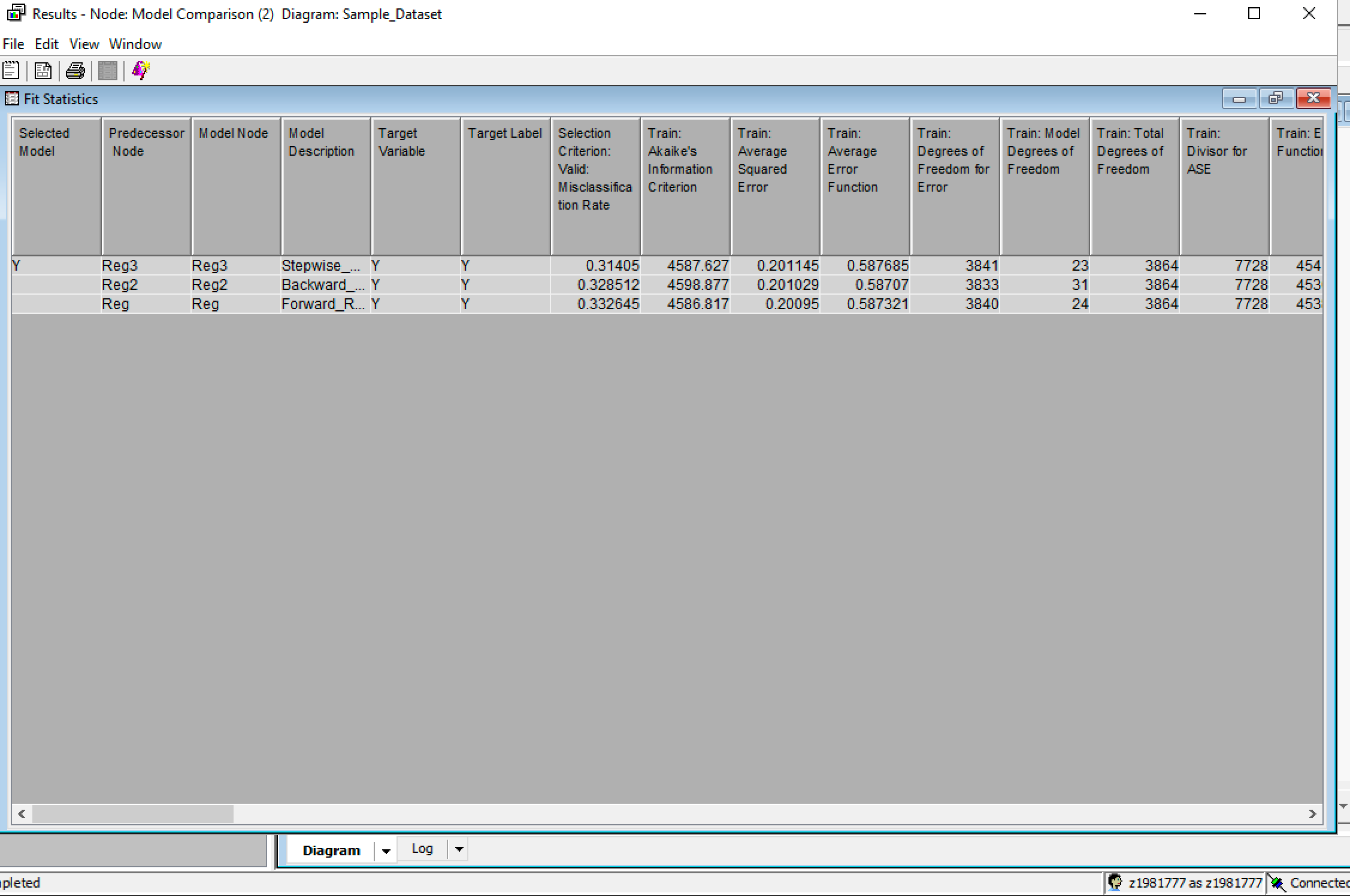
The best decision tree is probability tree, and the following are the important variables as per the probability tree.

* TI\_passanger2 – Passenger: Friends
* TI\_coupon5 – Restaurant (<20)
* TI\_passanger2 – Passenger: Friends
* TI\_coupon2 – Carry out & Take Away
* TI\_destinatation2 – destination: No urgent place
* TI\_coupon1 – coupon: Bar

1. Add different types of regression models to the diagram. Vary the selection model property to build the different models. Make sure the regression type property is set correctly for your type of target variable. Explain which variables are important to your target. (Again, changing settings may or may NOT change the regression output).
   1. What variables are important according to your best regression model? (Note: You should look at model comparison and then answer this so you know which one it deemed the best model.)

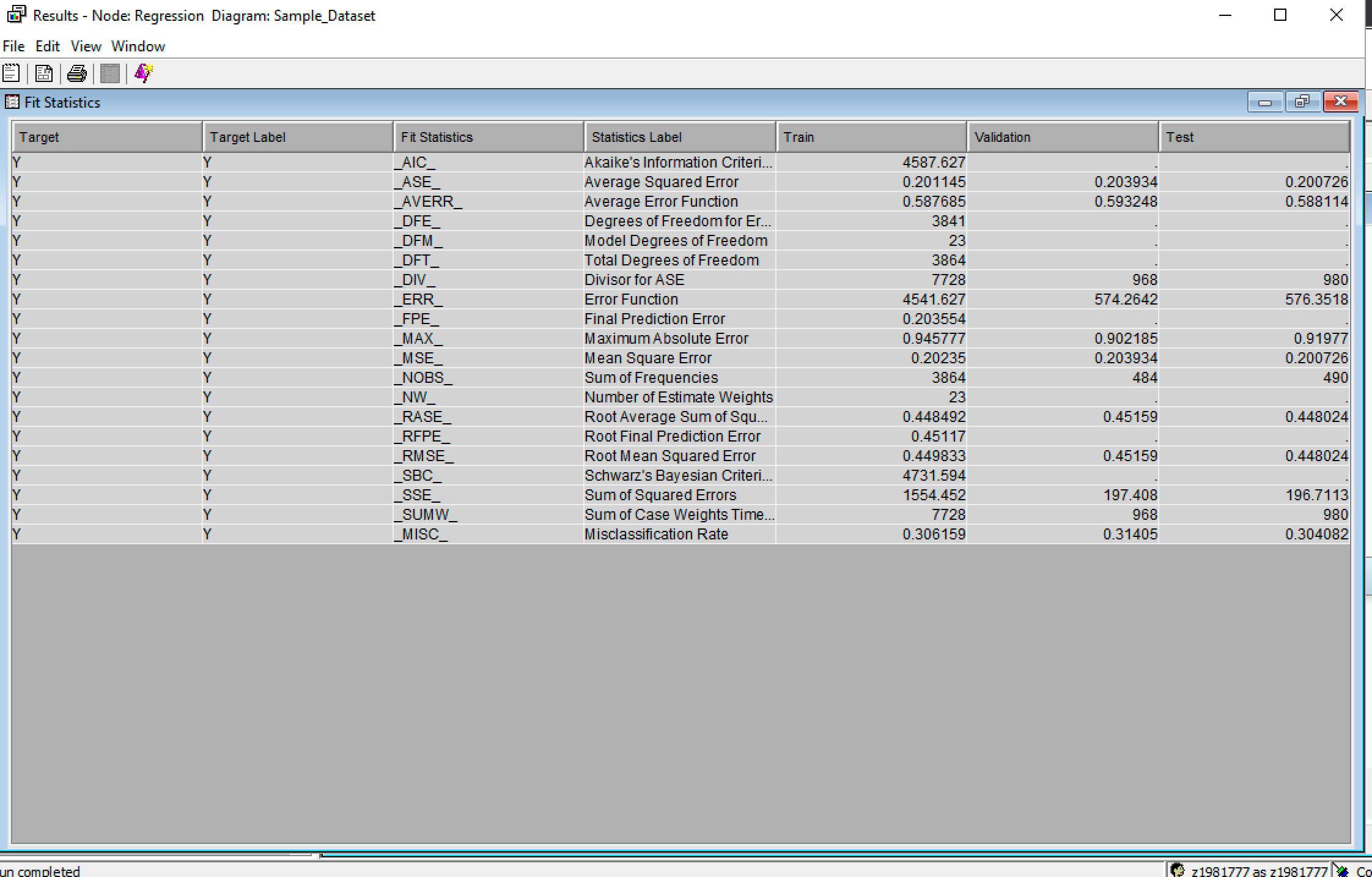
We have selected logistic regression and compared stepwise regression, backward regression, and forward regression using model comparison. We found that stepwise regression is yielding the best results in comparison with other regression models.





Best regression model – Stepwise regression

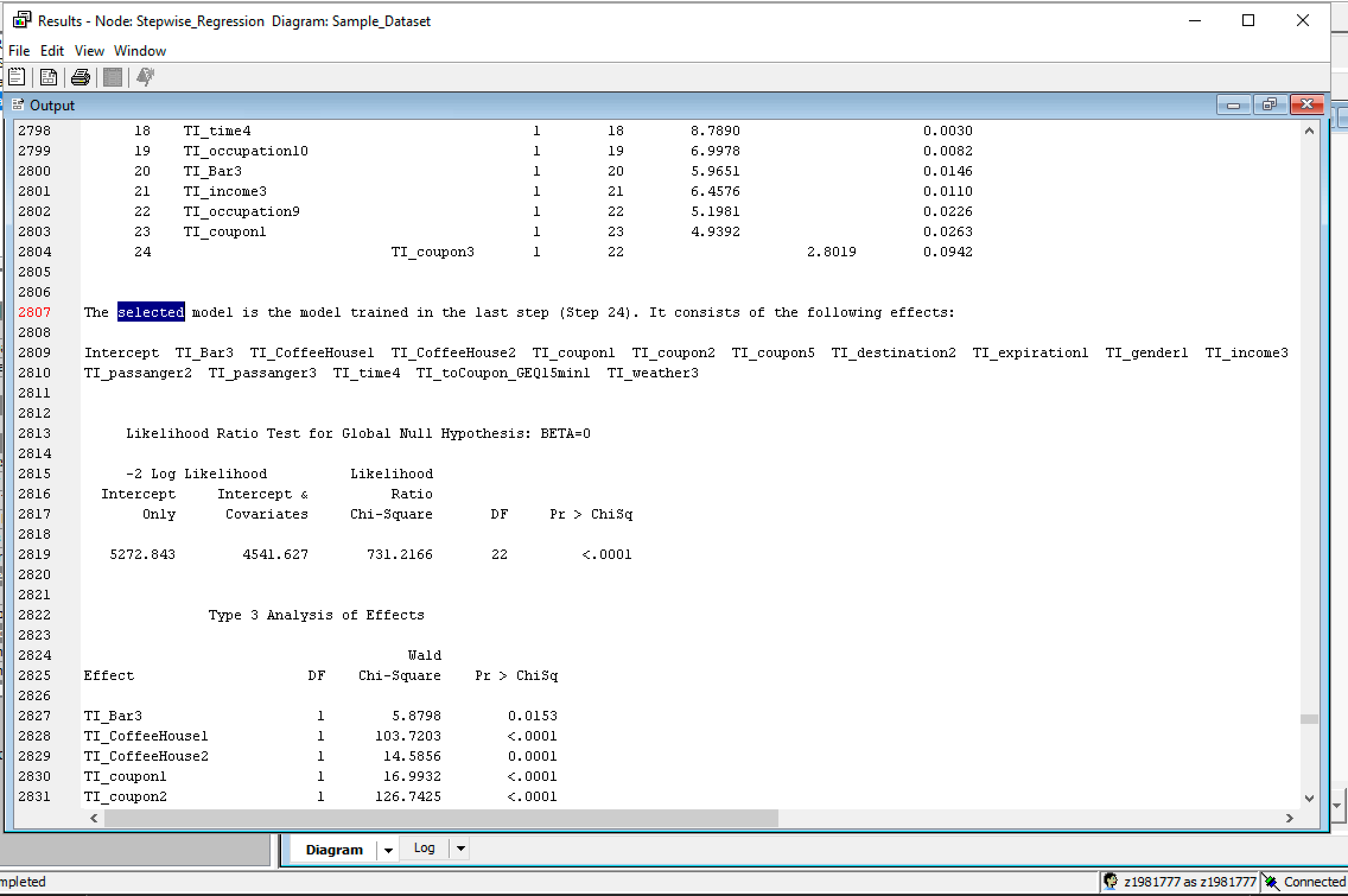
Input Selection: In the properties, the model selection is set to ‘Stepwise’, and the results of fit statistics are explained below.



After selecting the regression model as stepwise, the average squared error for trained, validation is 0.201145 and 0.203934 respectively. Similarly, the misclassification rate for trained and validation is 0.306159 and 0.31405 respectively.

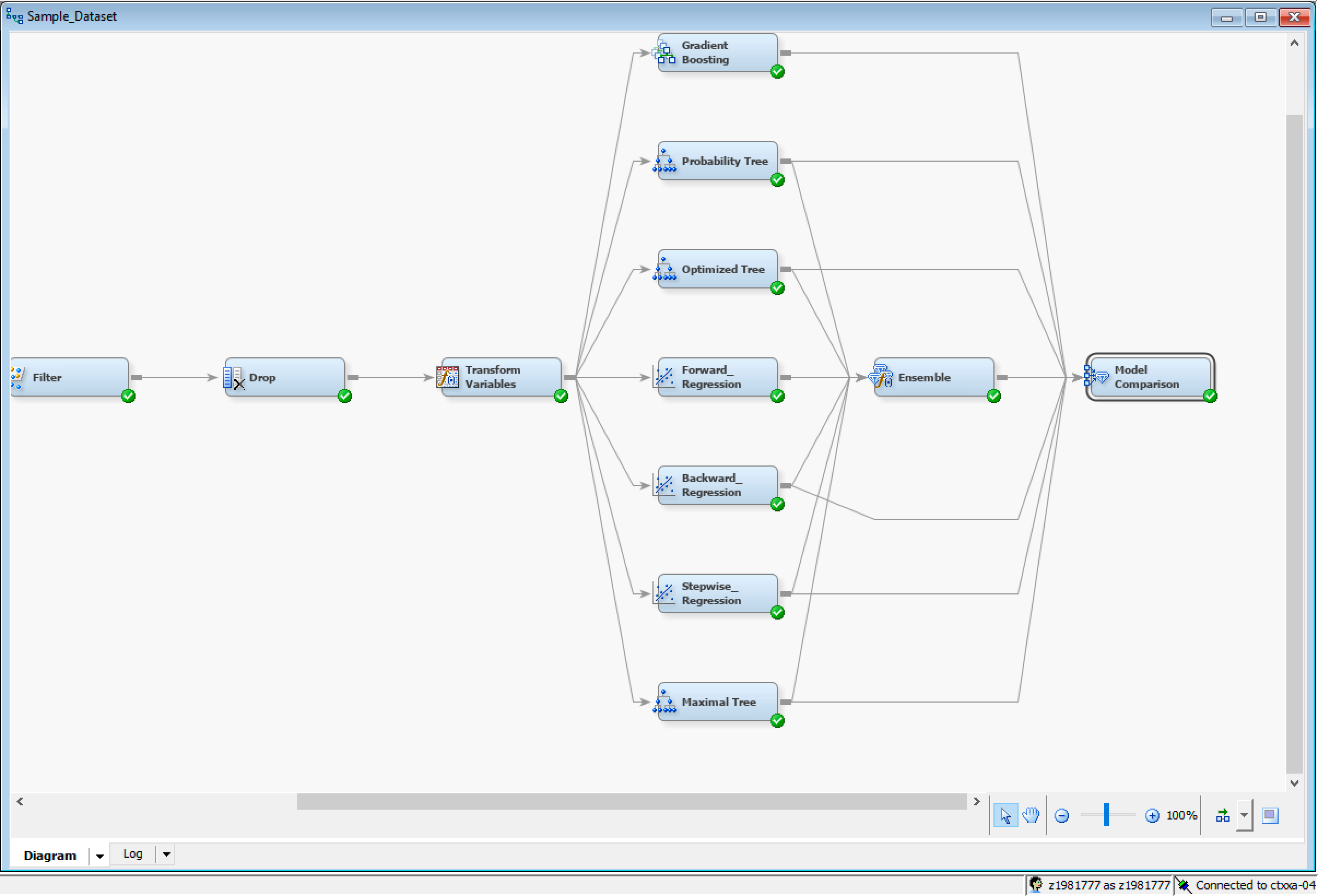
Best variables:

TI\_Bar3 TI\_CoffeeHouse1 TI\_CoffeeHouse2 TI\_coupon1 TI\_coupon2 TI\_coupon5 TI\_destination2 TI\_expiration1 TI\_gender1 TI\_income3 TI\_income5 TI\_income9 TI\_maritalStatus3 TI\_occupation10 TI\_occupation18 TI\_occupation6 TI\_occupation9 TI\_passanger2 TI\_passanger3 TI\_time4 TI\_toCoupon\_GEQ15min1 TI\_weather3



1. Add Gradient Boosting and Ensemble nodes to the diagram.

We have added gradient boosting and ensemble nodes to the diagram



1. Add a model comparison node and feed all of the models you’ve built into the model comparison node. Which model ends up being the best one and why?

As per the model Comparision, we conclude that the ensemble model yeilds the best results in comparison with the other models as per the validation misclassification rate.

