

# Business Intelligence and Predictive Analytics

## Homework 4 – KNIME Data Mining

Due Date

April 23, 2023

By

Meagan Windler

## Executive Summary

The enclosed report explores a recent data mining project I conducted within the KNIME platform. The project was designed to use four different prediction models (two number-based and two set-based) to discover the determinants of the customer churning, which is the percentage of customers who stop using the company during a particular timeframe. The analysis used the following logistical regression and artificial neural network for our number-based models. The set-based models used were random forest and a decision tree. This analysis determined that the random forest model was most accurate for discovering the determinants of customer churning. This paper will walk through the CRISP-DM process followed for this analysis.

Throughout the analysis I will refer to the four models conducted using the KNIME platform. Figure 1 below shows an overview of the four models used in this analysis in KNIME.

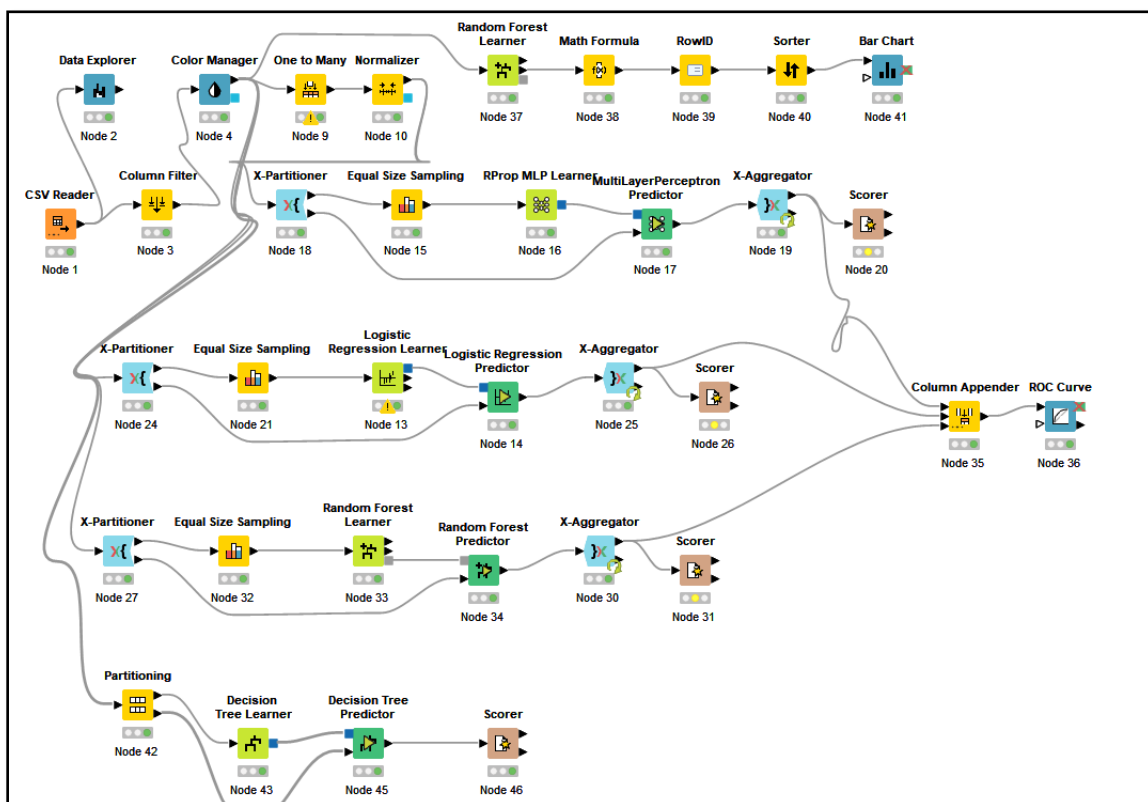


Figure 1: KNIME Churning Analysis Overview

## Cross-Industry Standard Process for Data Mining (CRISP-DM)

### Step 1: Business Understanding

The business goal of this project was to answer: “Which customers are most likely to churn (or leave) the company?” In order to do that, I utilized the Customer Churn Data, which is churn/ attrition behavior for 1,000 of the company’s customers. Figures two and three show the data dictionary and a few sample rows from the data. In my analysis, I had to consider the data as a whole and then dive deeper to determine the most optimal number of variables to consider.

Therefore, the project plan was to investigate the factors that are determinants of customer churning utilizing two types of models: number-based and set-based. The analysis will use the following logistical regression and artificial neural network for our number-based models. The set-based models will be a random forest and a decision tree. I will then compare their results together to determine the statistically significant factors as well as how to deploy this information for my business use.

### Step 2: Data Understanding

To fully dive into the data, please reference Figure 1 on the next page, which is the data dictionary for the customer churning dataset. It includes whether or not the customer is a churner along with socio-demographic attributes like age, marital status, geographic region, and education and behavioral attributes like services used and hours of usage. Figure 3 gives a sample of the customer churn data. As you can see from the data, the decision variable of churn is a binary data type, with 0 indicating a “No” and 1 being “Yes” a churner.

Class of variable		Variable	Description	Type
Socio-demographic attributes		Region	The region where the customer lives	Nominal
		Age	The age of customer	Numeric
		Marital	Marital status: 1: Yes, 0: No	Binominal
		Address	The number of years of residence in current location	Numeric
		Income	The customers' income	Numeric
		Education	The customers' education: 1-Diploma, 2: AS 3: BS 4:MS, 5: PhD	Nominal
		Employment	Years of employment	Numeric
		Retire	Retired or not?: 1: Yes, 0: No	Binominal
		Gender	Gender of customer: 1: Male, 0: Female	Binominal
Behavioral attributes	Hours of usage	Longmon	Hours of using service 1 per month	Numeric
		Tollmon	Hours of using service 2 per month	Numeric
		Equipmon	Hours of using service 3 per month	Numeric
		Cardmon	Hours of using service 4 per month	Numeric
		Wiremon	Hours of using service 5 per month	Numeric
	Selected services	Multiline	Is customer has a multiline phone: 1: Yes, 0: No	Binominal
		Voice	Has voice service or not?: 1: Yes, 0: No	Binominal
		Pager	Has pager or not?: 1: Yes, 0: No	Binominal
		Internet	Has internet or not?: 1: Yes, 0: No	Binominal
		Callid	Has caller ID or not?: 1: Yes, 0: No	Binominal
		Callwait	Has call waiting service or not?: 1: Yes, 0: No	Binominal
		Forward	Has call forwarding service or not?: 1: Yes, 0: No	Binominal
		Confer	Has conference service or not?: 1: Yes, 0: No	Binominal
	Callcard	Has contact card or not?: 1: Yes, 0: No	Binominal	
	Wireless	Has wireless system or not?: 1: Yes, 0: No	Binominal	
Label		Churn	Churner or Non-churner?: 1: Yes, 0: No	Binominal

Figure 2: Data Dictionary for the Customer Churn Dataset

### Figure 3: Sample Data from Customer Churn Dataset

To further explore the data, I utilized a bar chart node, which analyzed the variable importance measures of all the variables, which is shown in Figure 4 below.

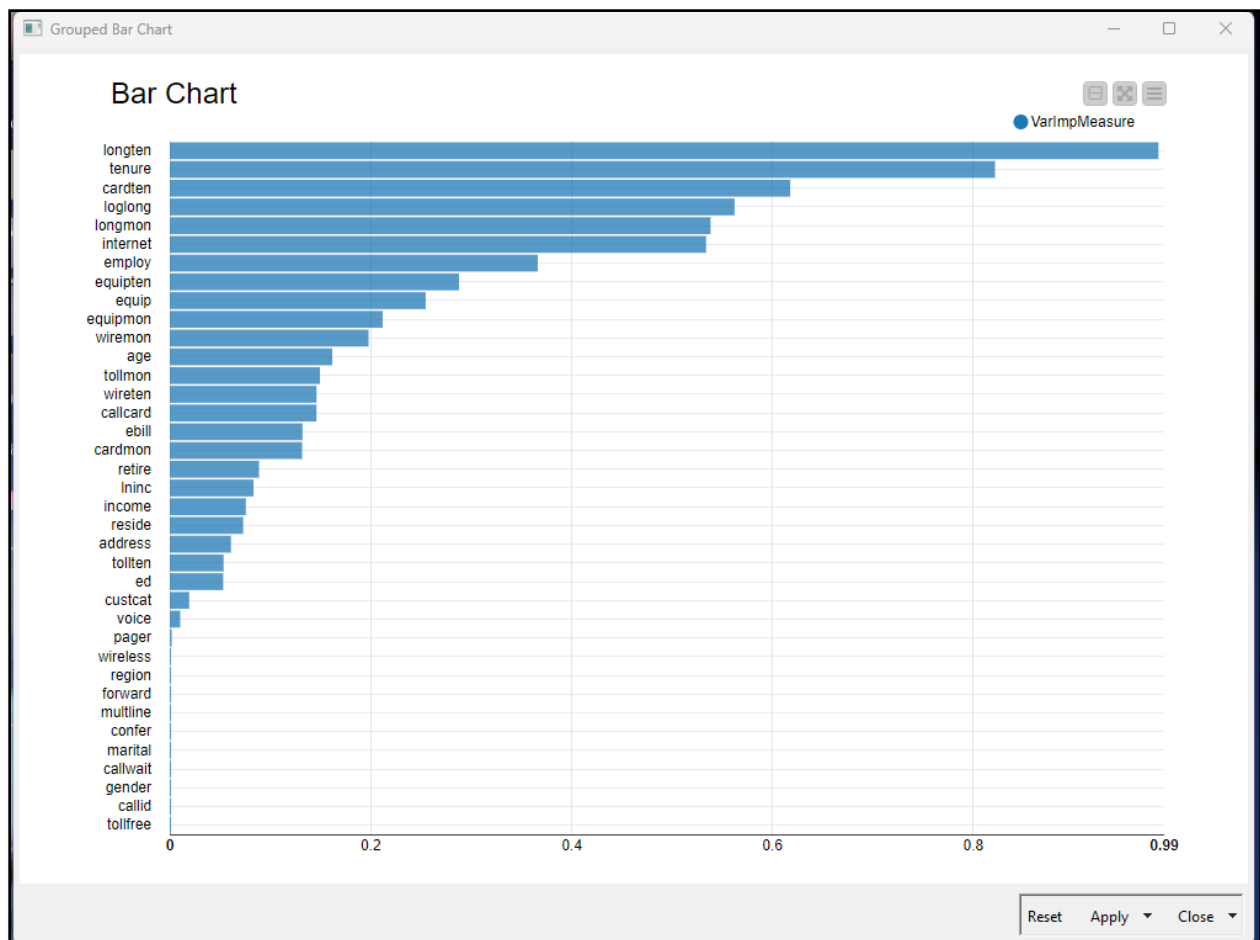


Figure 4: Bar Chart Node Output

### Step 3: Data Preparation

In my data pre-processing, I had to exclude the churn variable from the analysis using the one-to-many node, pictured in Figure 5 below.

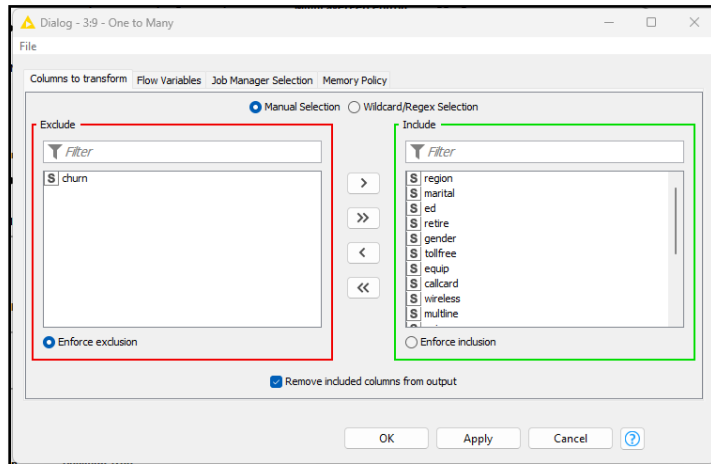


Figure 5: One-to-many Node

I noticed from the bar chart that the data included several irrelevant fields. To rectify this, I added a Column filter to exclude the variables that were not statistically important from the bar chart. This will make the models more efficient, as it is best practice to use the least number of relevant variables. Figure 6 shows the variables that are excluded from the analysis using the column filter. Figure 7 shows the bar chart node with the irrelevant variables excluded.

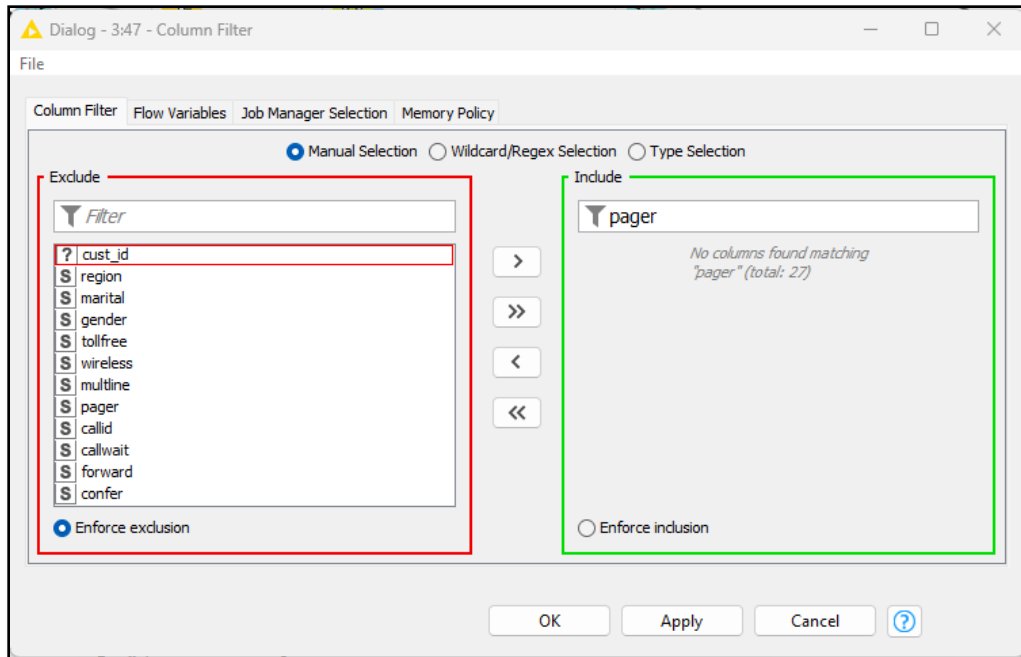


Figure 6: Column Filter

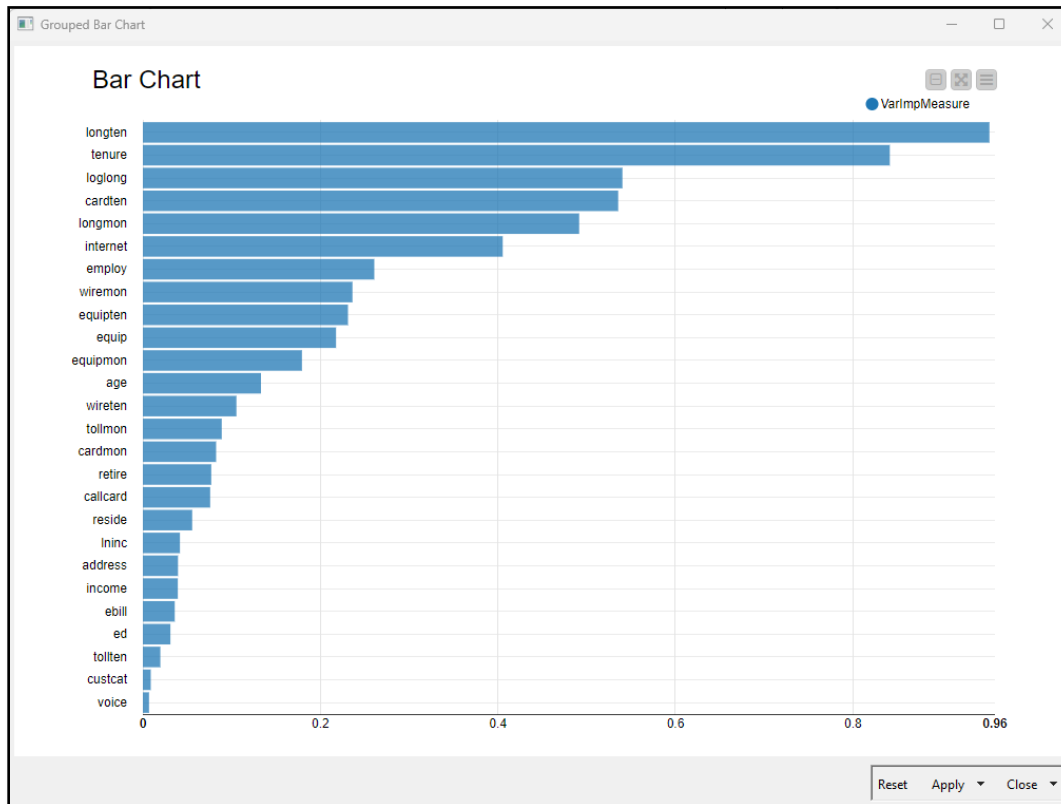


Figure 7: Bar Chart Node after Column Filter

The number-based models (logistical regression and artificial neural network) require nominal variables to be in the numeric form, so I used the Normalizer node (pictured in Figure 8) to transform the data into a numeric binary from nominal data.

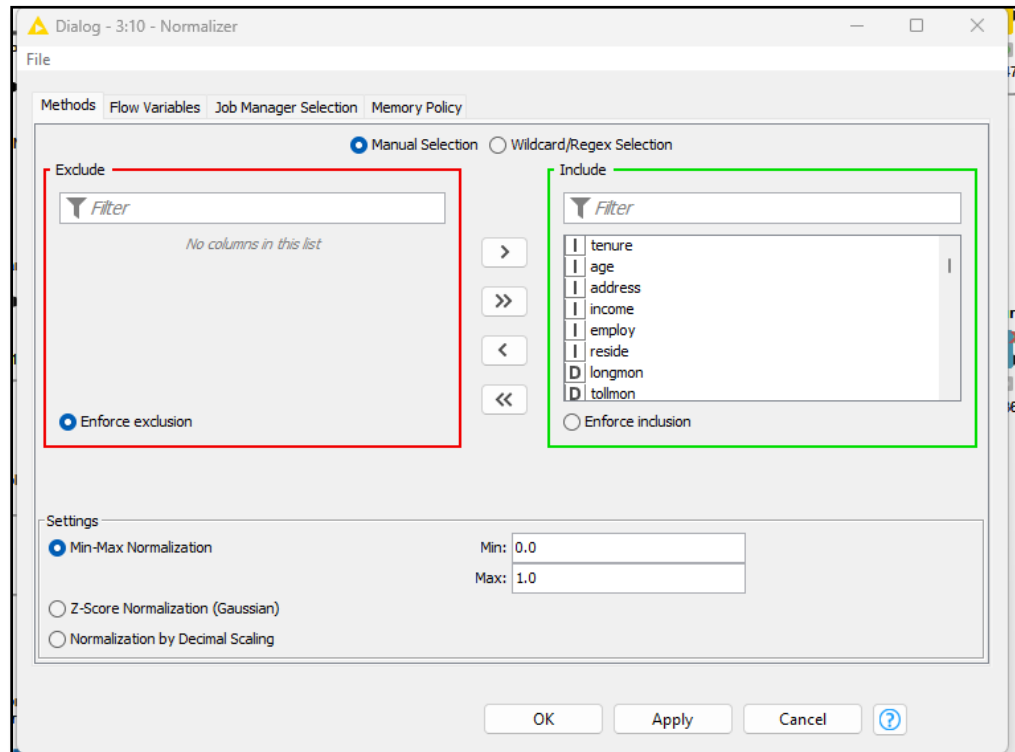


Figure 8: Normalizer Node

The one exception to this is the decision variable of churn, which is excluded and stays a nominal variable. Then we used a Color Manager node to create a color filter to turn rows that are Churners to red and non-churners to green, pictured in Figure 9 below.

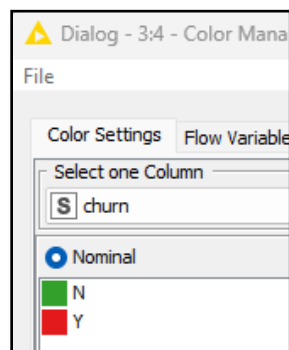


Figure 9: Color Manager Node



## Step 4: Model Building

In my analysis I built models in the following order: artificial neural network, logistical regression, random forest, and a decision tree. This section will explain the general process I followed in my churning analysis.

To build the ANN, logistical regression, and random forest models, I used the X-Partitioner node (shown in Figure 10) to divide the data into two subsets: training and validation testing. For the decision tree model, I used the Partitioning Node to divide the data.

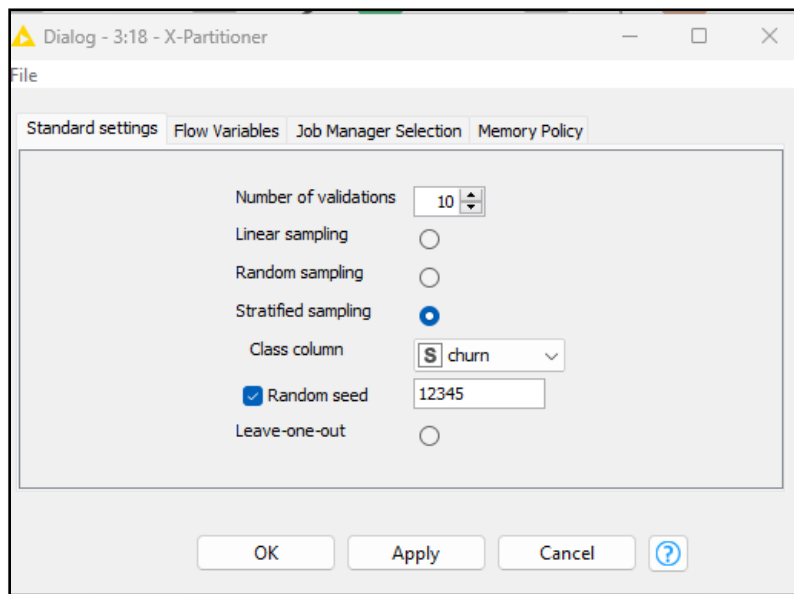


Figure 10: X-Partitioner Node

Then I used an Equal Size Sampling node to balance the sample sizes of data used in the training models, which can be seen in Figure 11.

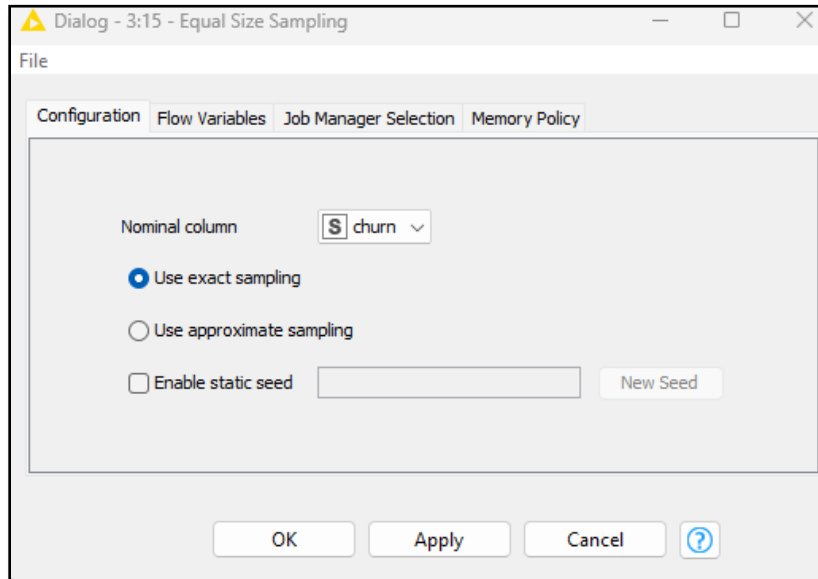


Figure 11: Equal Size Sampling Node

Now was time to build the models. For the Artificial Neural Network, I used the RProp MLP Learner node to analyze the learner data, shown in Figure 12. In this learner node, I made sure to use the random seed of “12345”. I then used the Multilayer Perception Predictor to create the probability column, which we named “\_AN” for Artificial Neural Network, shown in Figure 13.

For the logistical regression model, I employed the Logistic Regression Learner (Figure 14) and named the probability column using the Logistic Regression Predictor. I used the Random Forest Learner (Figure 15) and Random Forest Predictor for the Random Forest model. Figure 16 shows the Decision Tree Learner node, after which I used the Decision Tree Predictor.

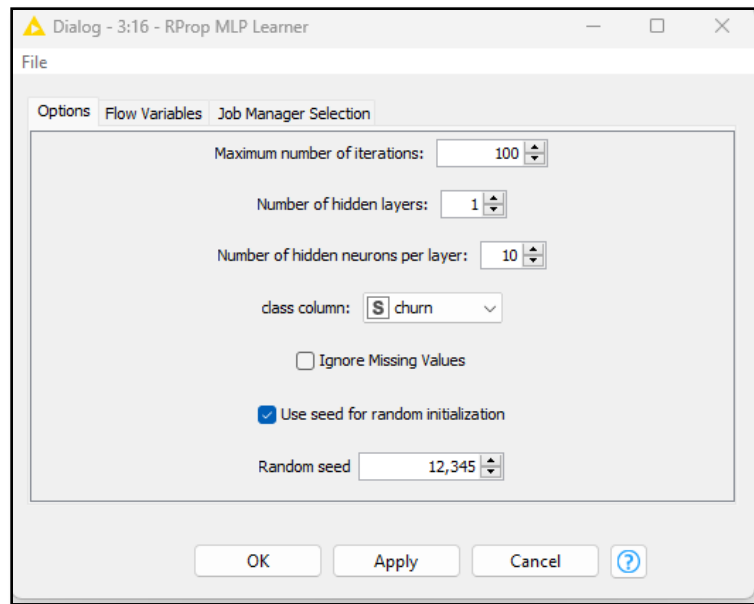


Figure 12: Artificial Neural Network RProp MLP Learner Node

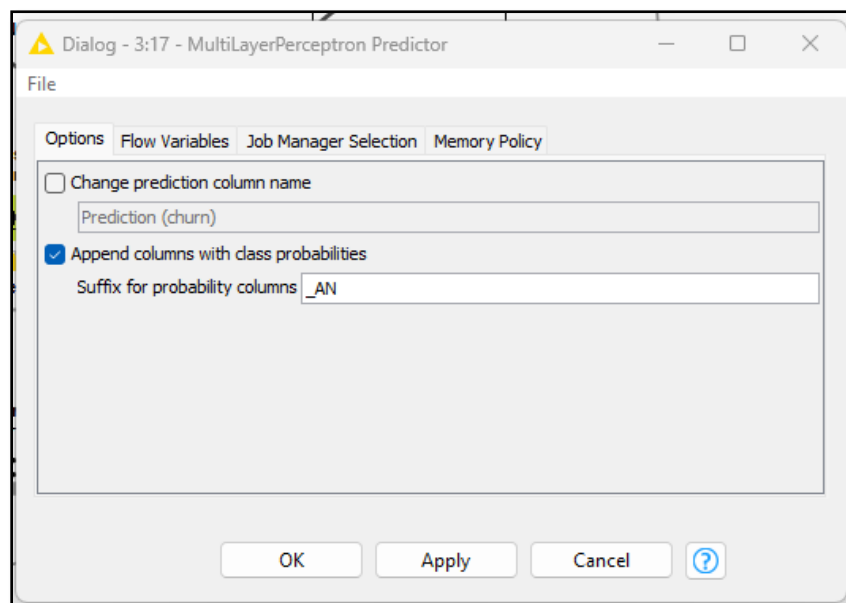


Figure 13: Multilayer Perception Predictor

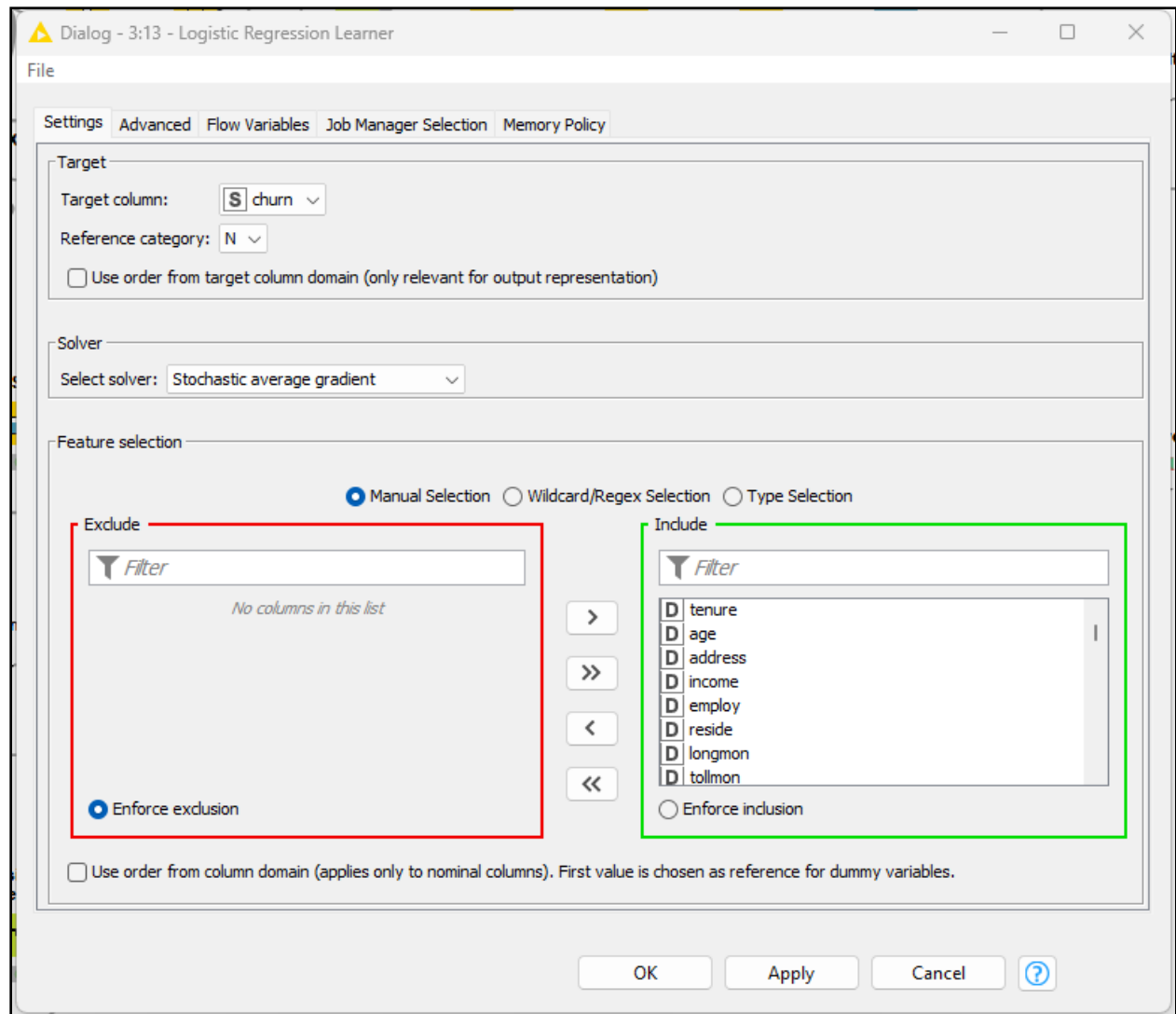


Figure 14: Logistic Regression Learner Node

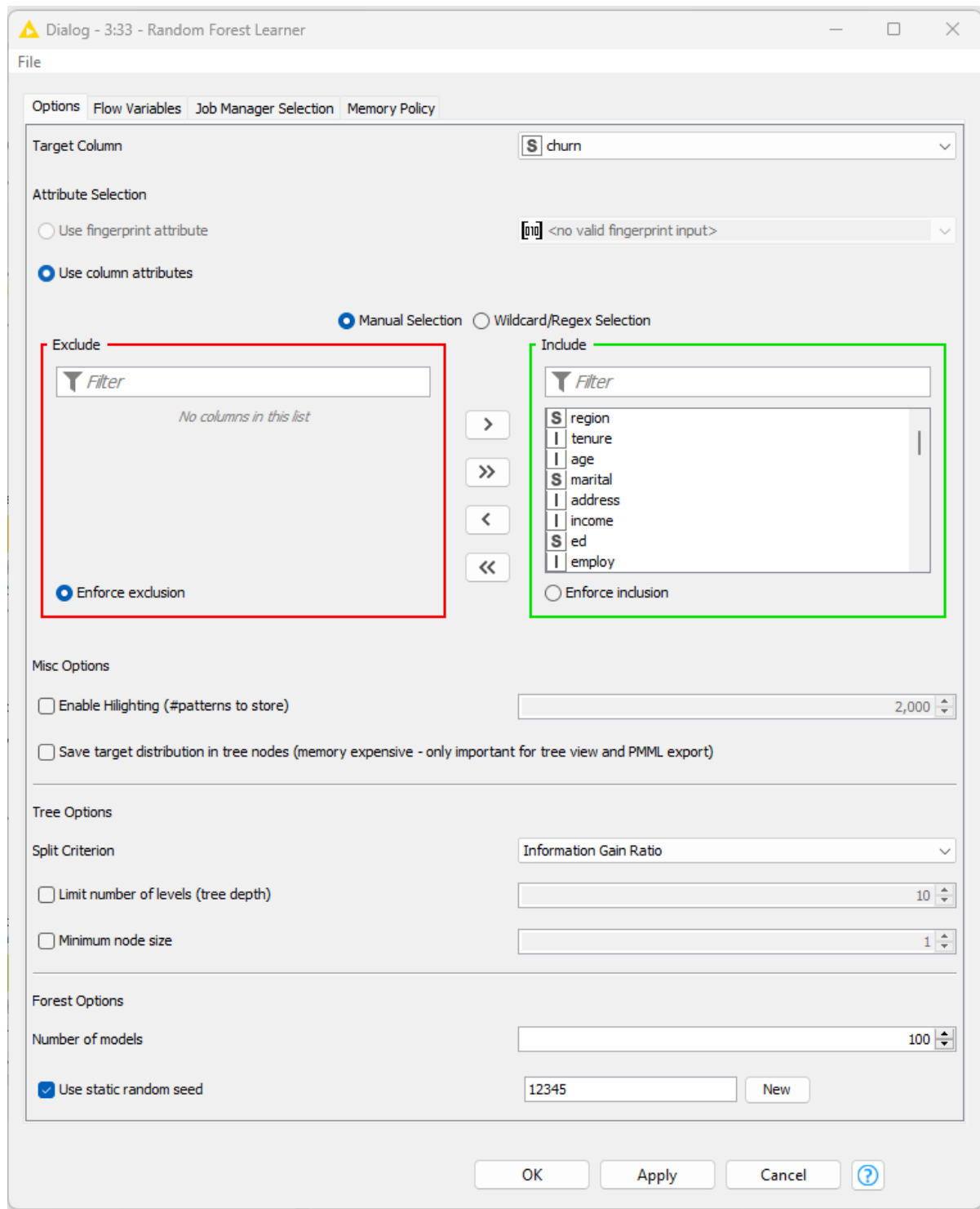


Figure 15: Random Forest Learner Node

Dialog - 3:43 - Decision Tree Learner

File

Options PMMLSettings Flow Variables Job Manager Selection

General

Class column **S** churn ▾

Quality measure Gini index ▾

Pruning method No pruning ▾

☒ Reduced Error Pruning

Min number records per node 2 ▴ ▾

Number records to store for view 10,000 ▴ ▾

☒ Average split point

Number threads 20 ▴ ▾

☒ Skip nominal columns without domain information

Root split

☐ Force root split column

Root split column **S** custcat ▾

Binary nominal splits

☒ Binary nominal splits

Max #nominal 10 ▴ ▾

☐ Filter invalid attribute values in child nodes

OK Apply Cancel ?

Figure 16: Decision Tree Learner Node

For all the models except the decision tree, I added an X-aggregator node to assign the target column as churn and the prediction column as Prediction (churn), which is shown in Figure 17 below.

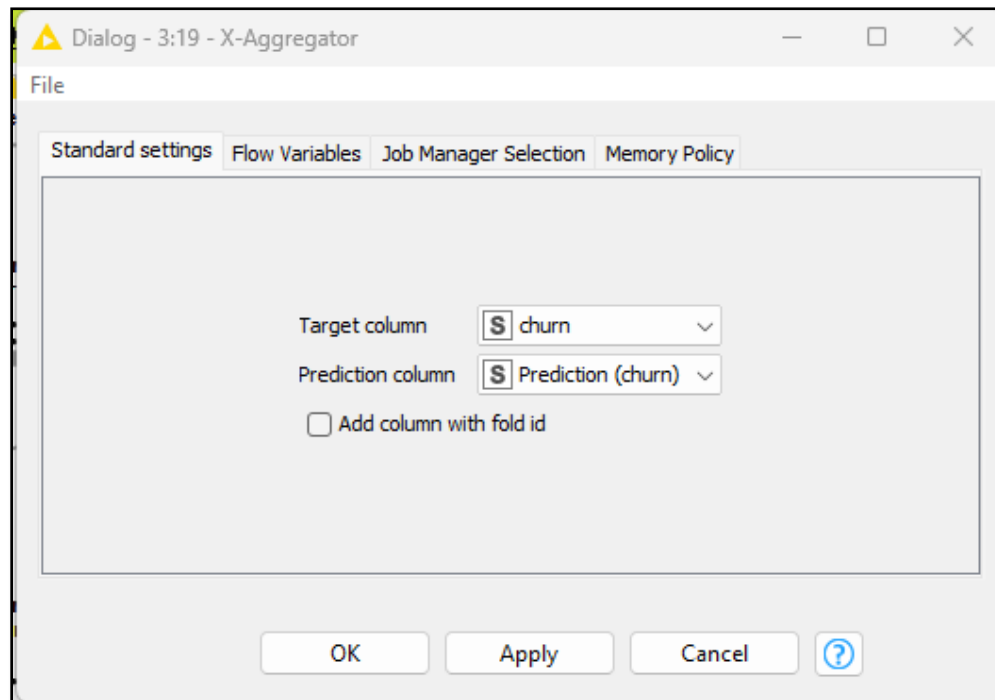


Figure 17: X-Aggregator Node

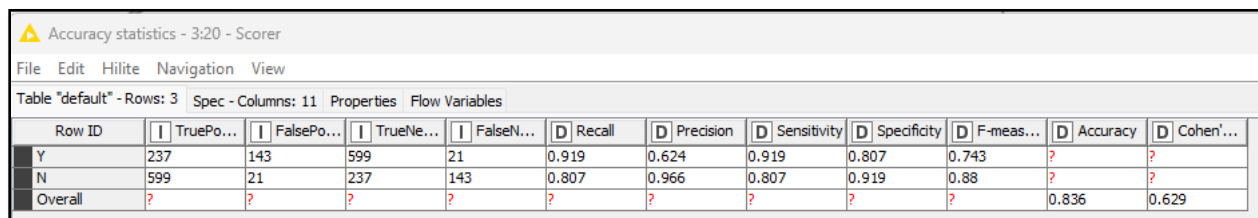
## Step 5: Testing and Evaluation

Based on what we learned in this course, I predicted that the random forest model would be the most accurate model for this use case for the following reasons: random forests use both nominal and numeric data, and they use multiple decision trees to make the predictions using every variable equally. Figure 18 shows a compiled list of the Scorer Node outputs of the various models. The data confirmed my hypothesis that the random forest model would be the most accurate. Its accuracy is at 88% compared to 83.6%, 80.7%, and 72.3% for the ANN, Decision Tree, and Logistical Regression models respectively. Figures 19 through 22 show the specific Scorer node outputs for all four models used.

Model Type	Accuracy	Precision	Sensitivity	Specificity
Random Forest	0.88	0.699	0.938	0.86
Artificial Neural Network	0.836	0.624	0.919	0.807
Decision Tree	0.807	0.656	0.519	0.906
Logistical Regression	0.723	0.476	0.74	0.717

Figure 18: Model Scorer Node Outputs

Figure 18 lists the aggregated data for the four models, with the rows sorted by the accuracy statistic in descending order. As you can see, the random forest model is the most accurate, precise, and sensitive. Therefore, the Random Forest model is my top choice for the project.



Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
Y	237	143	599	21	0.919	0.624	0.919	0.807	0.743	?	?
N	599	21	237	143	0.807	0.966	0.807	0.919	0.88	?	?
Overall	?	?	?	?	?	?	?	?	?	0.836	0.629

Figure 19: Artificial Neural Network Scorer Output



Accuracy statistics - 3:26 - Scorer											
File Edit Hilite Navigation View											
Table "default" - Rows: 3 Spec - Columns: 11 Properties Flow Variables											
Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
Y	191	210	532	67	0.74	0.476	0.74	0.717	0.58	?	?
N	532	67	191	210	0.717	0.888	0.717	0.74	0.793	?	?
Overall	?	?	?	?	?	?	?	?	?	0.723	0.387

Figure 20: Logistical Regression Scorer Output

Accuracy statistics - 3:31 - Scorer											
File Edit Hilite Navigation View											
Table "default" - Rows: 3 Spec - Columns: 11 Properties Flow Variables											
Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
Y	242	104	638	16	0.938	0.699	0.938	0.86	0.801	?	?
N	638	16	242	104	0.86	0.976	0.86	0.938	0.914	?	?
Overall	?	?	?	?	?	?	?	?	?	0.88	0.718

Figure 21: Random Forest Scorer Output

Accuracy statistics - 3:46 - Scorer											
File Edit Hilite Navigation View											
Table "default" - Rows: 3 Spec - Columns: 11 Properties Flow Variables											
Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
Y	40	21	202	37	0.519	0.656	0.519	0.906	0.58	?	?
N	202	37	40	21	0.906	0.845	0.906	0.519	0.874	?	?
Overall	?	?	?	?	?	?	?	?	?	0.807	0.456

Figure 22: Decision Tree Scorer Output

Now let's review the Decision Tree graphical model output, shown in Figure 23, which shows the top variable in the decision tree is longten, with the next variable being Internet.

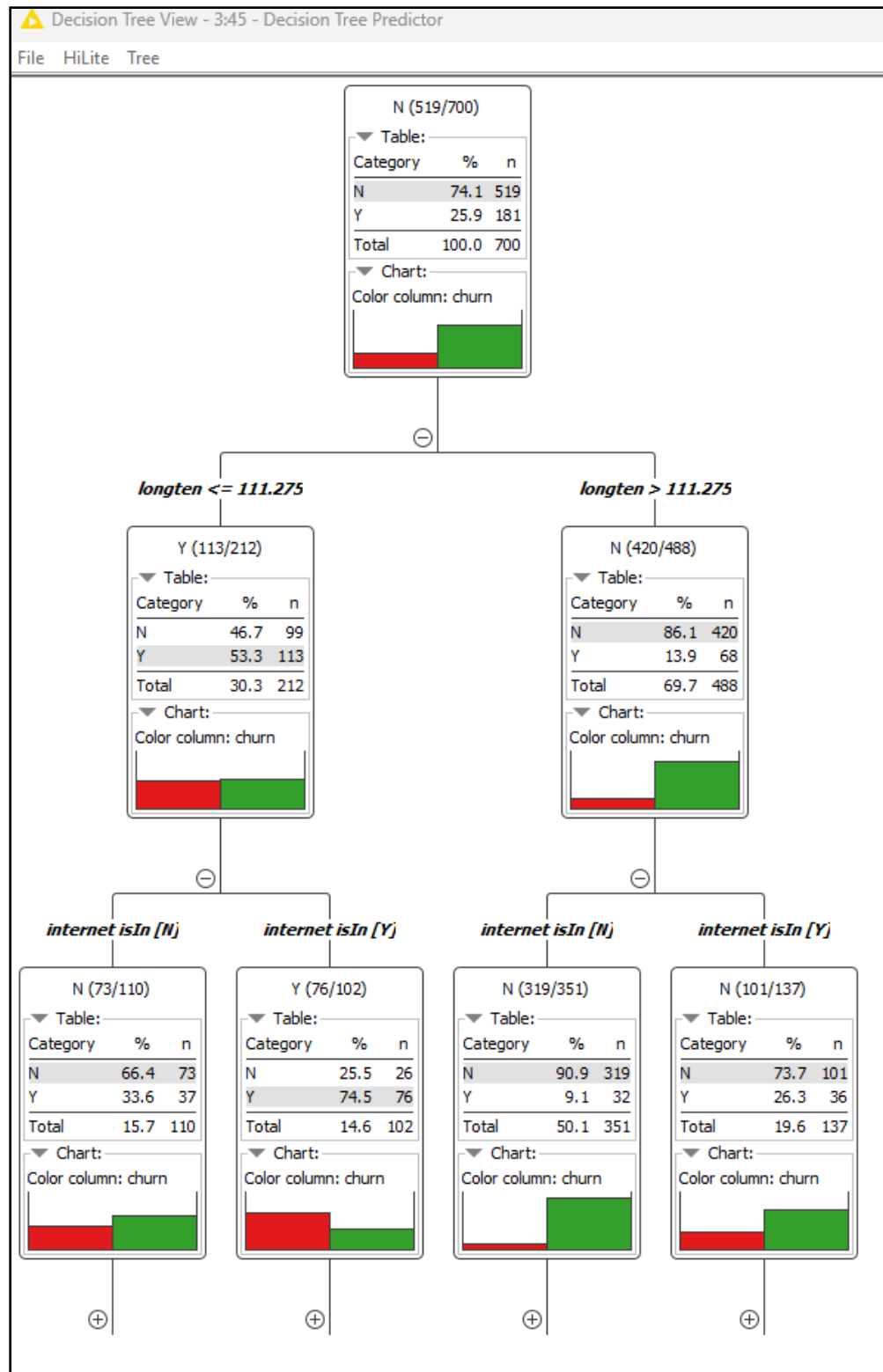


Figure 23: Decision Tree Results

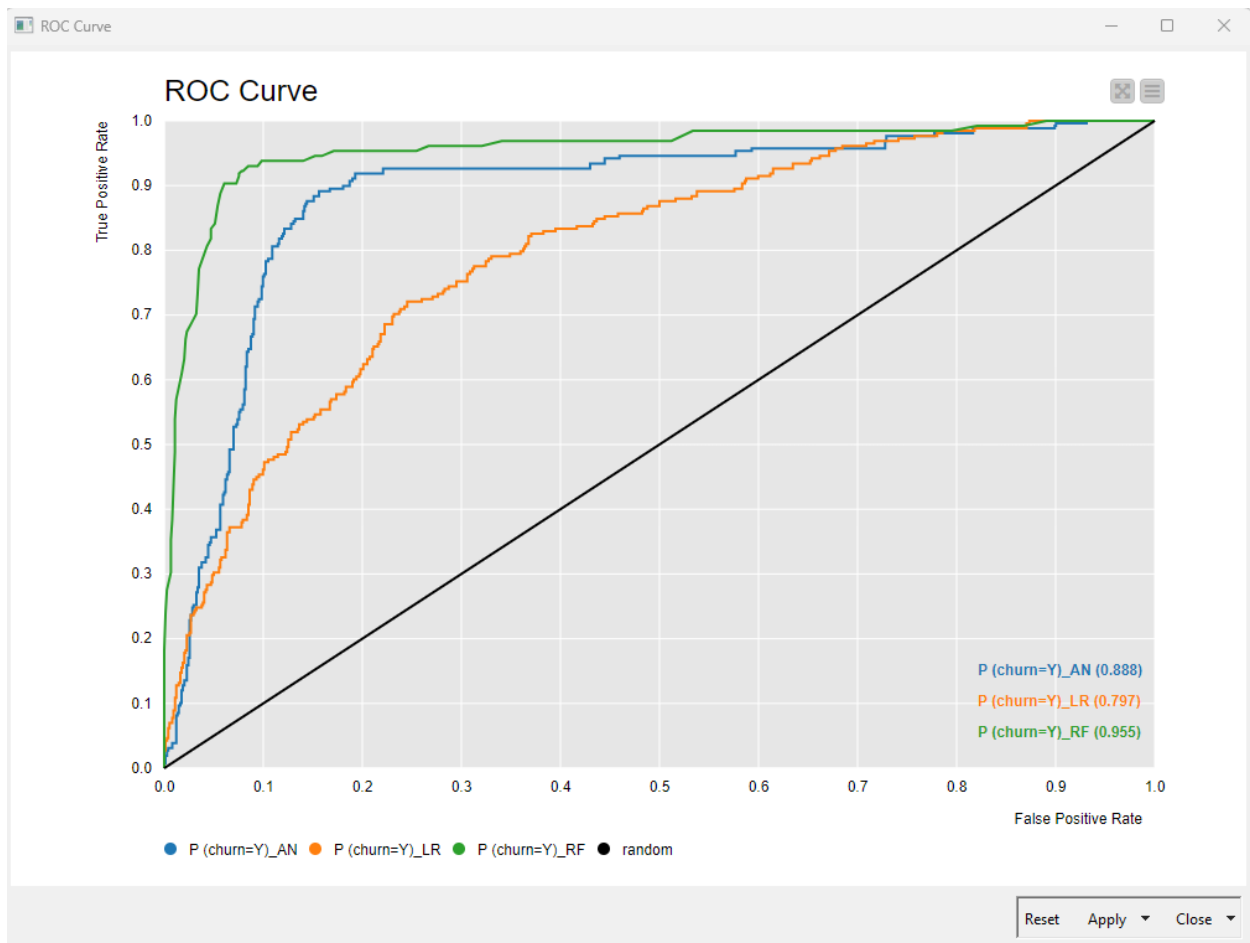


Figure 14: ROC Curve

Figure 14 above shows the ROC curve, which shows that the Random Forest model is the most accurate for this analysis. Figure 15 below shows the Random Forest variable statistics.

[illegible]

Figure 15: Random Forest Predictor

## Step 6: Deployment

The information gleaned from all four models could be used in many business uses. This is helpful for making recommendations to company leaders on how to target customers so that they can prevent them from churning.

## Summary and Conclusion

Utilizing the four different models, Random Forest is the best model to use to accurately predict whether or not a customer will be a churner. This project was a great opportunity to apply what I have learned in this class over the semester. I enjoyed configuring the various nodes and determining which model I preferred. I look forward to using this tool and the CRISP-DM process in my future career as a data analyst and scientist.