Business Intelligence and Predictive Analytics

KNIME Data Mining

Due Date

April 23, 2023

By

Chad 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.

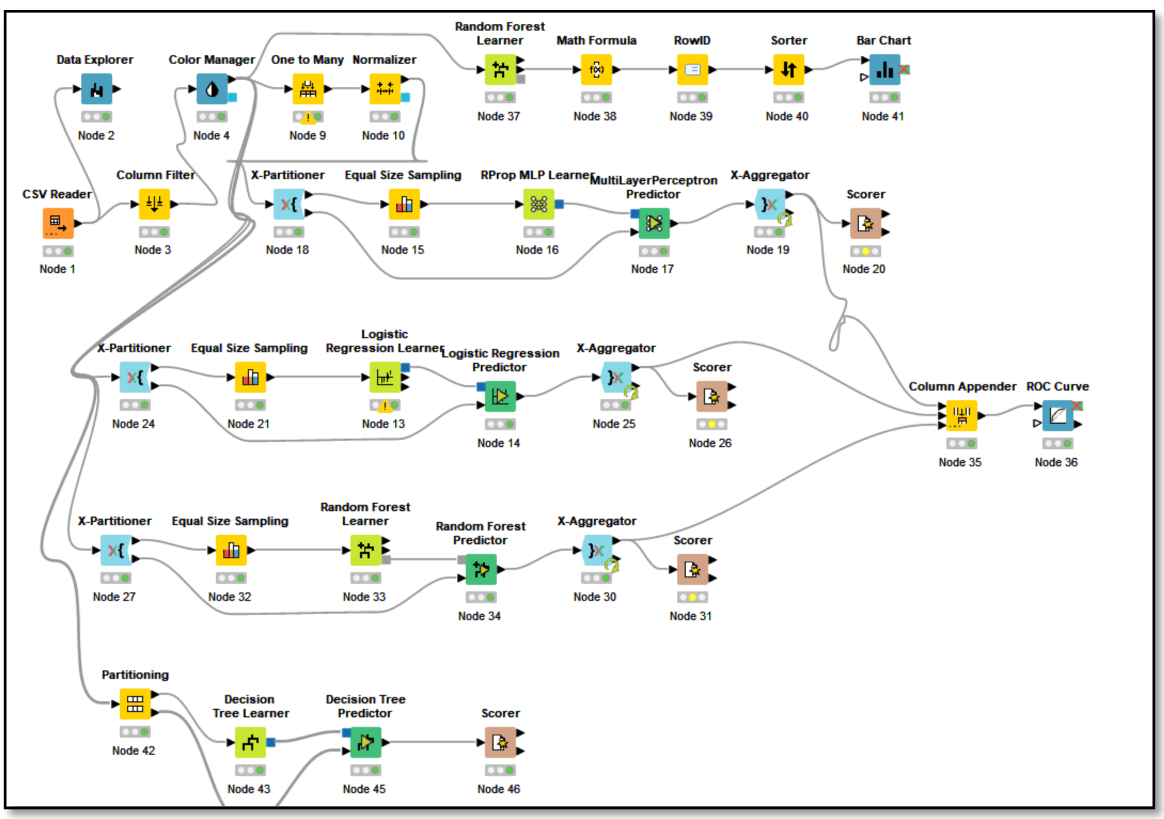

Figure 1: KNIME Churning Analysis Overview

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

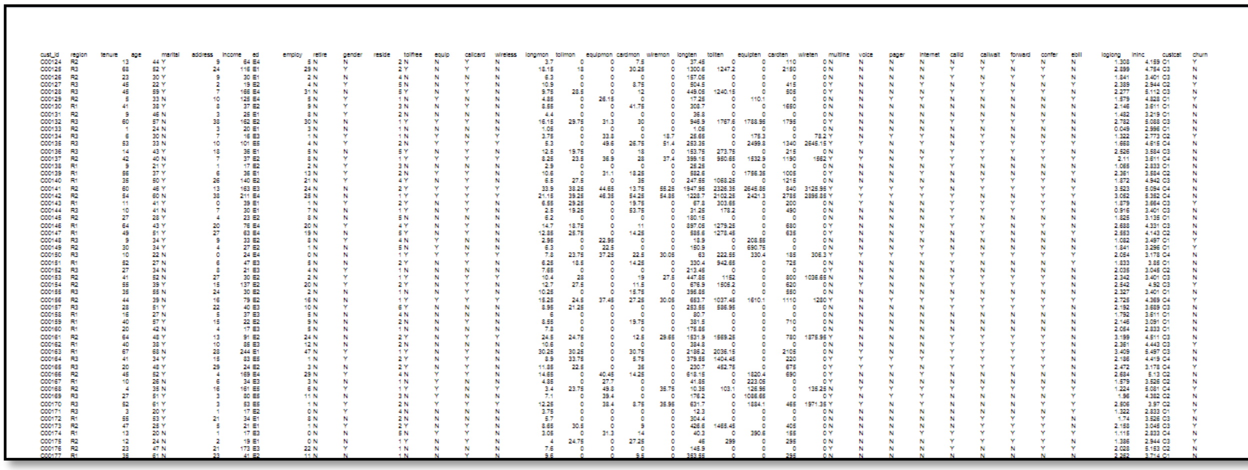

Figure 3: Sample Data from Customer Churn Dataset

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.

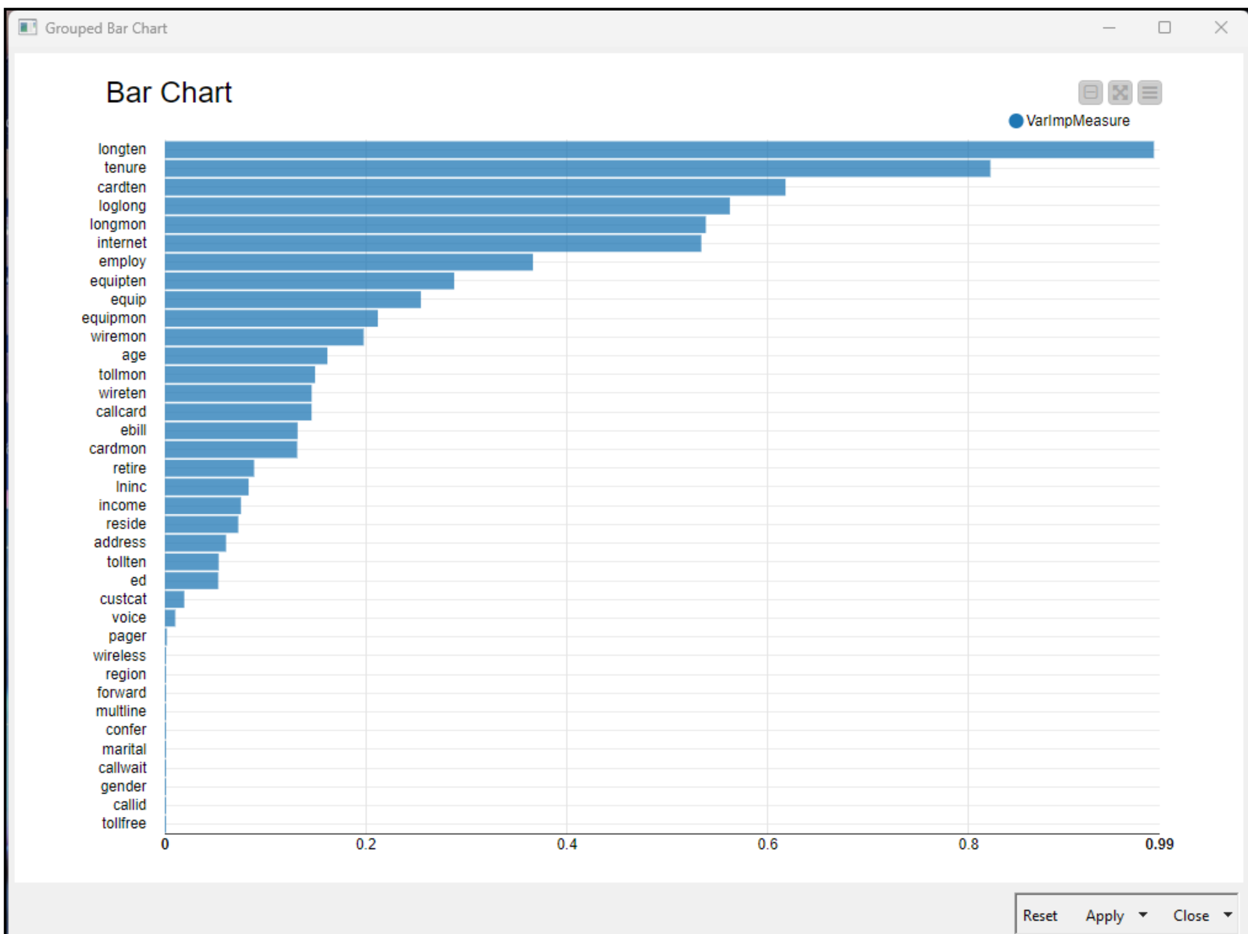

Figure 4: Bar Chart Node Output

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.

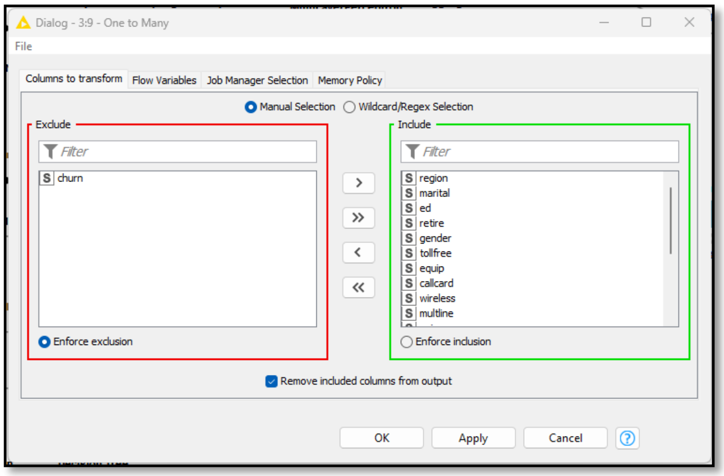

Figure 5: One-to-many Node

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.

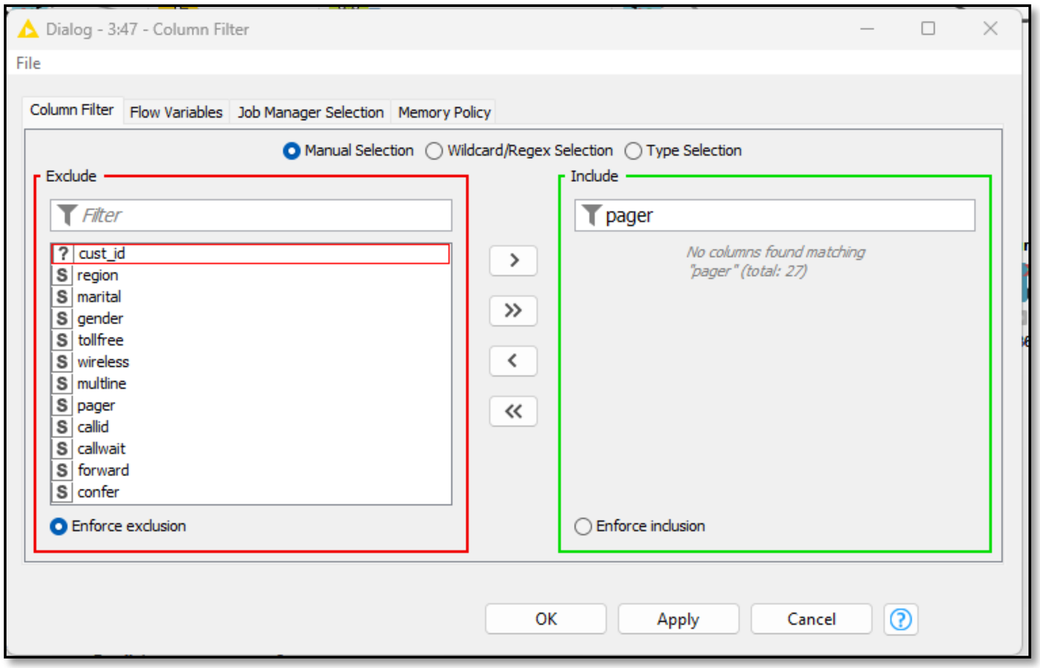

Figure 6: Column Filter

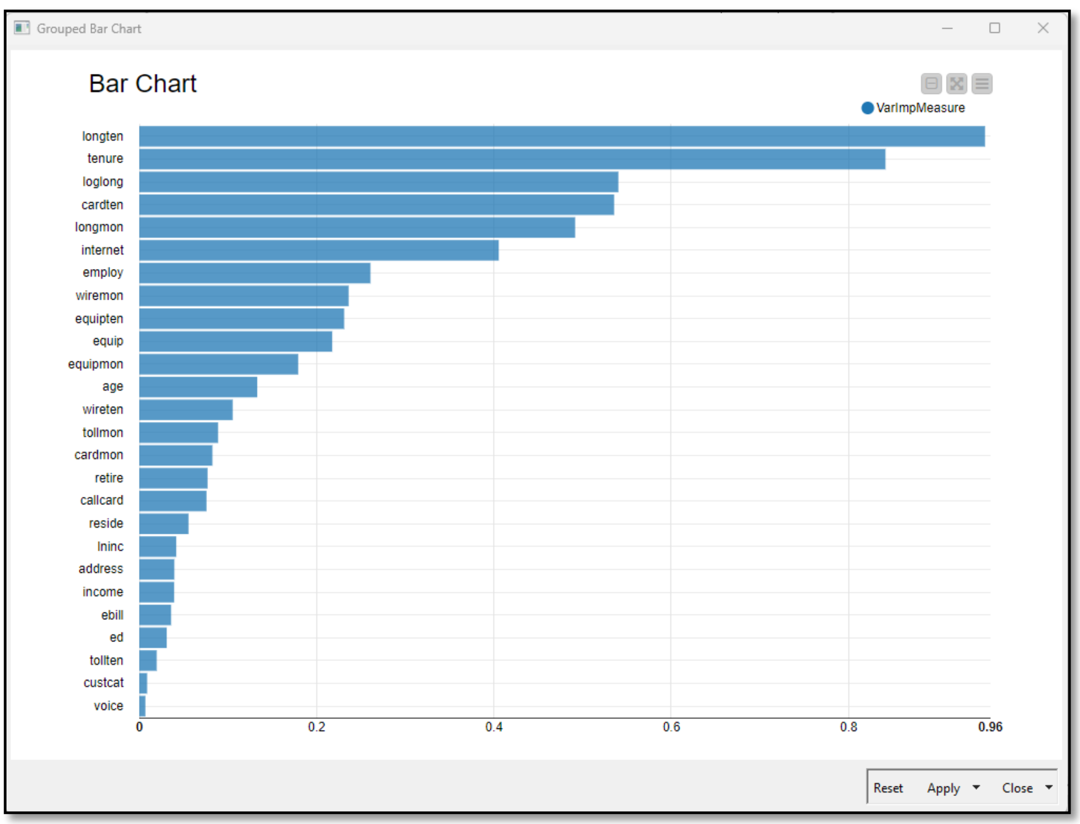

Figure 7: Bar Chart Node after Column Filter

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.

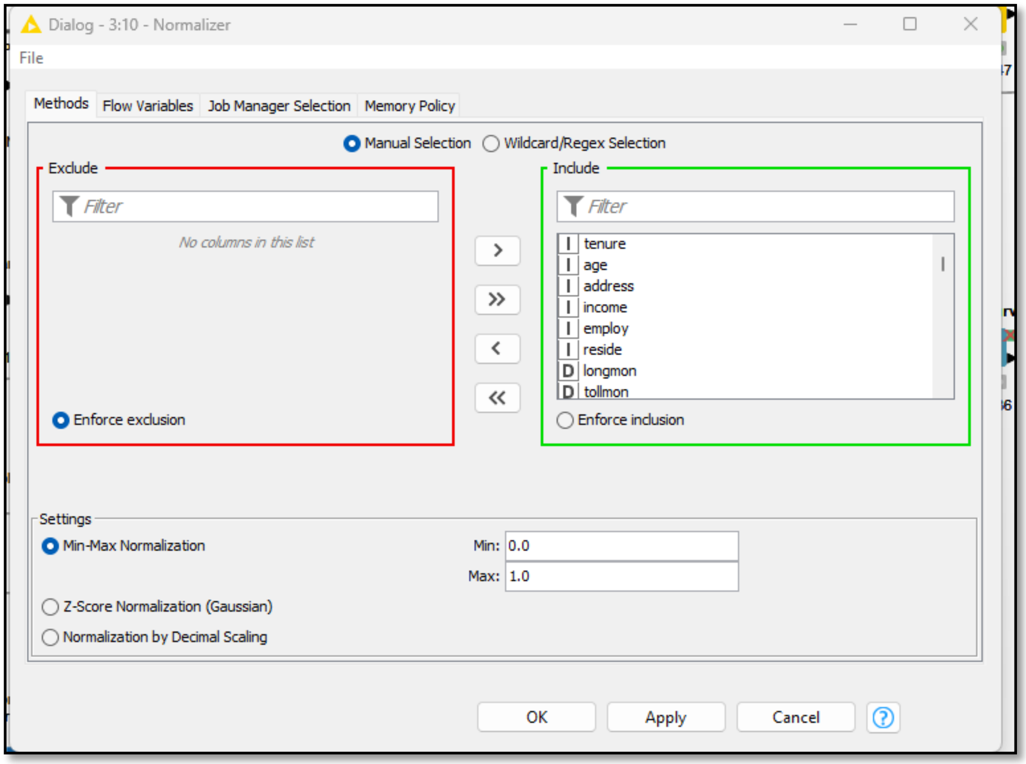

Figure 8: Normalizer Node

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.

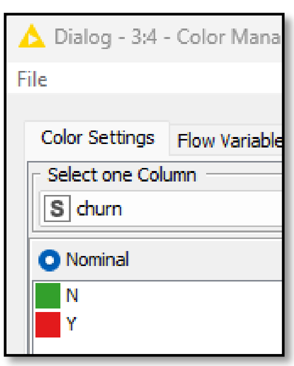

Figure 9: Color Manager Node

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.

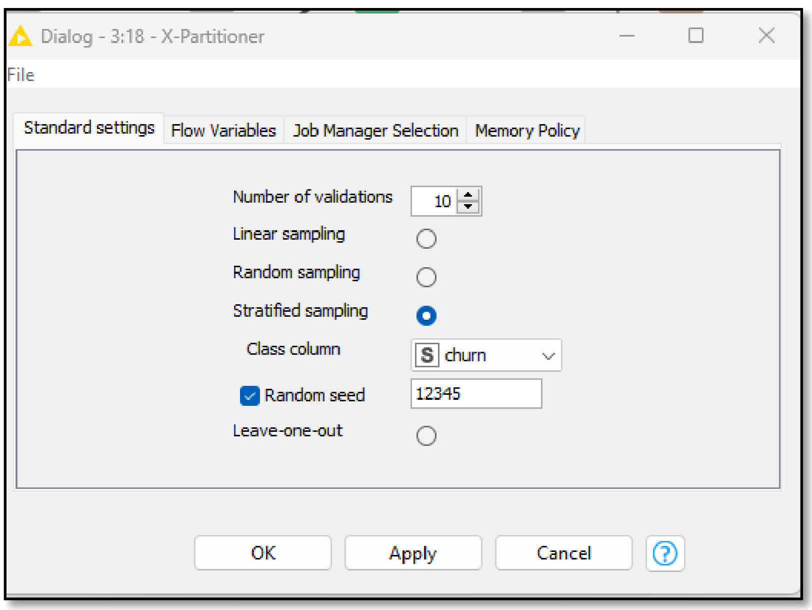

Figure 10: X-Partitioner Node

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.

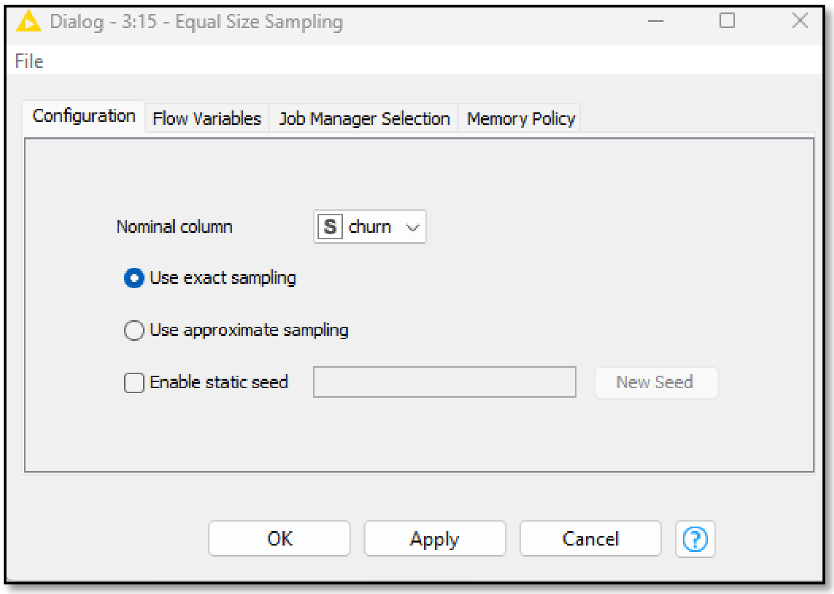

Figure 11: Equal Size Sampling Node

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.

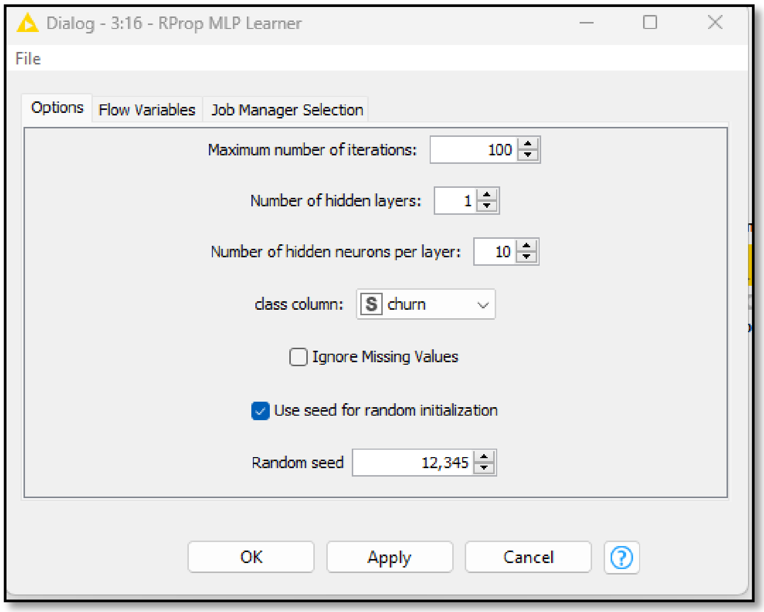

Figure 12: Artificial Neural Network RProp MLP Learner Node

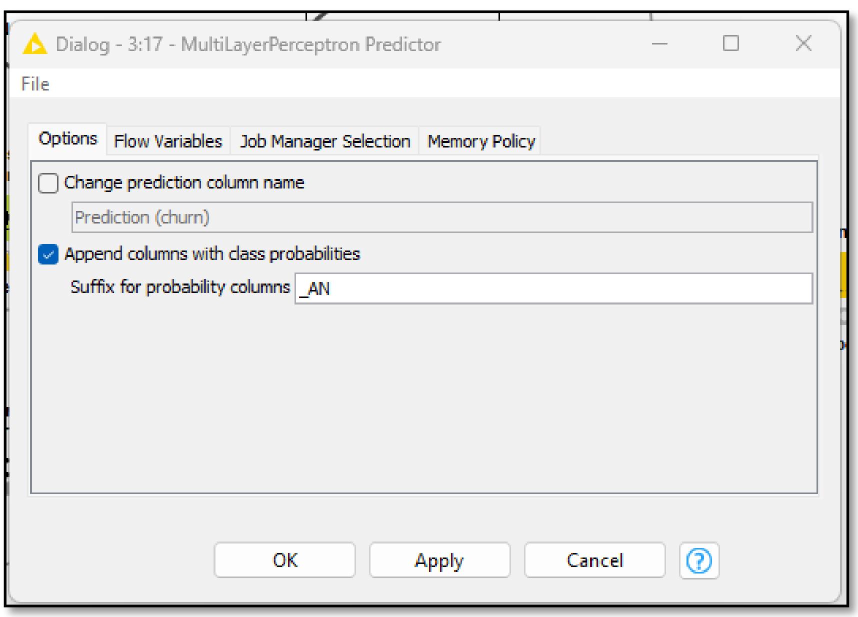

Figure 13: Multilayer Perception Predictor

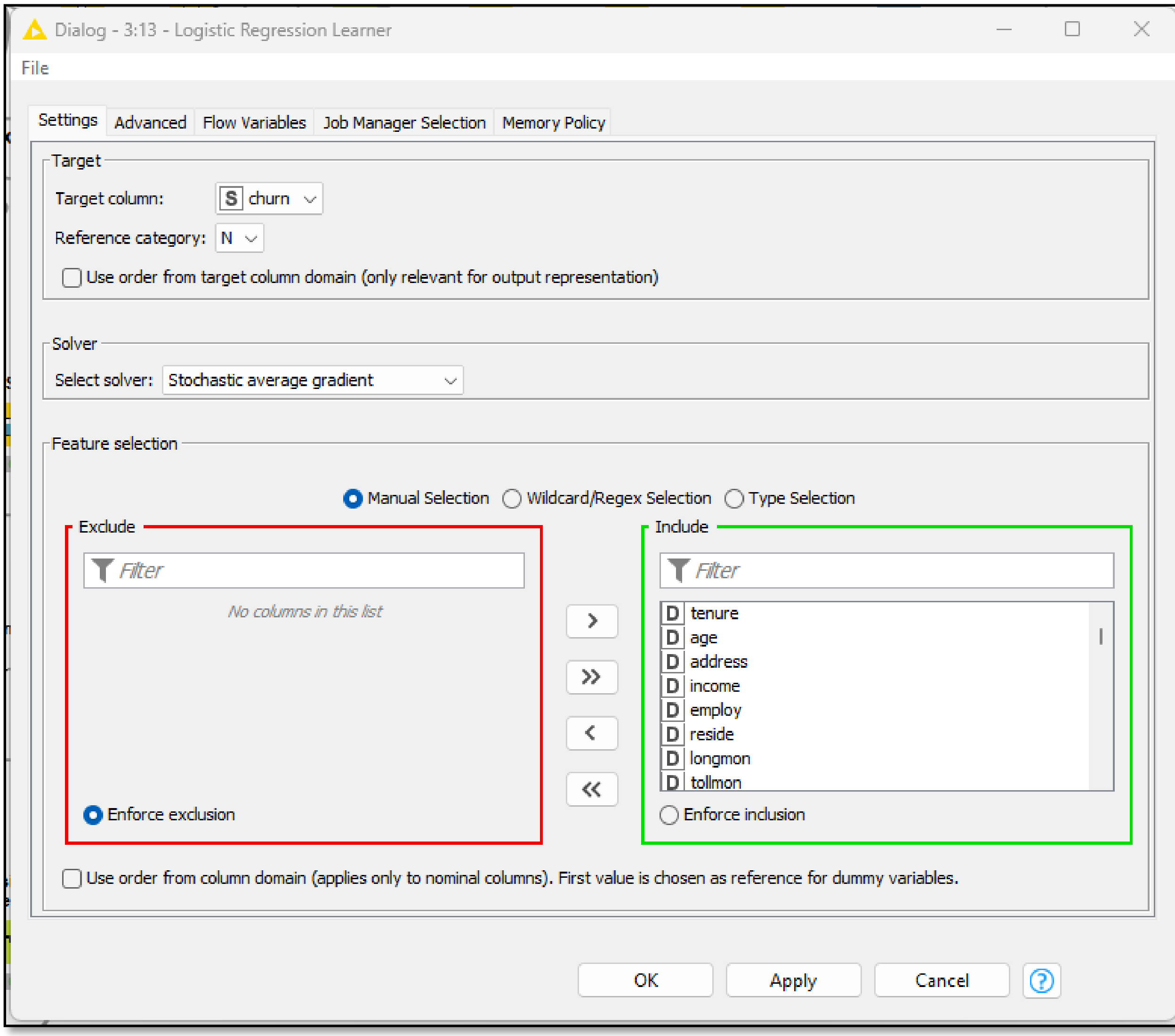

Figure 14: Logistic Regression Learner Node

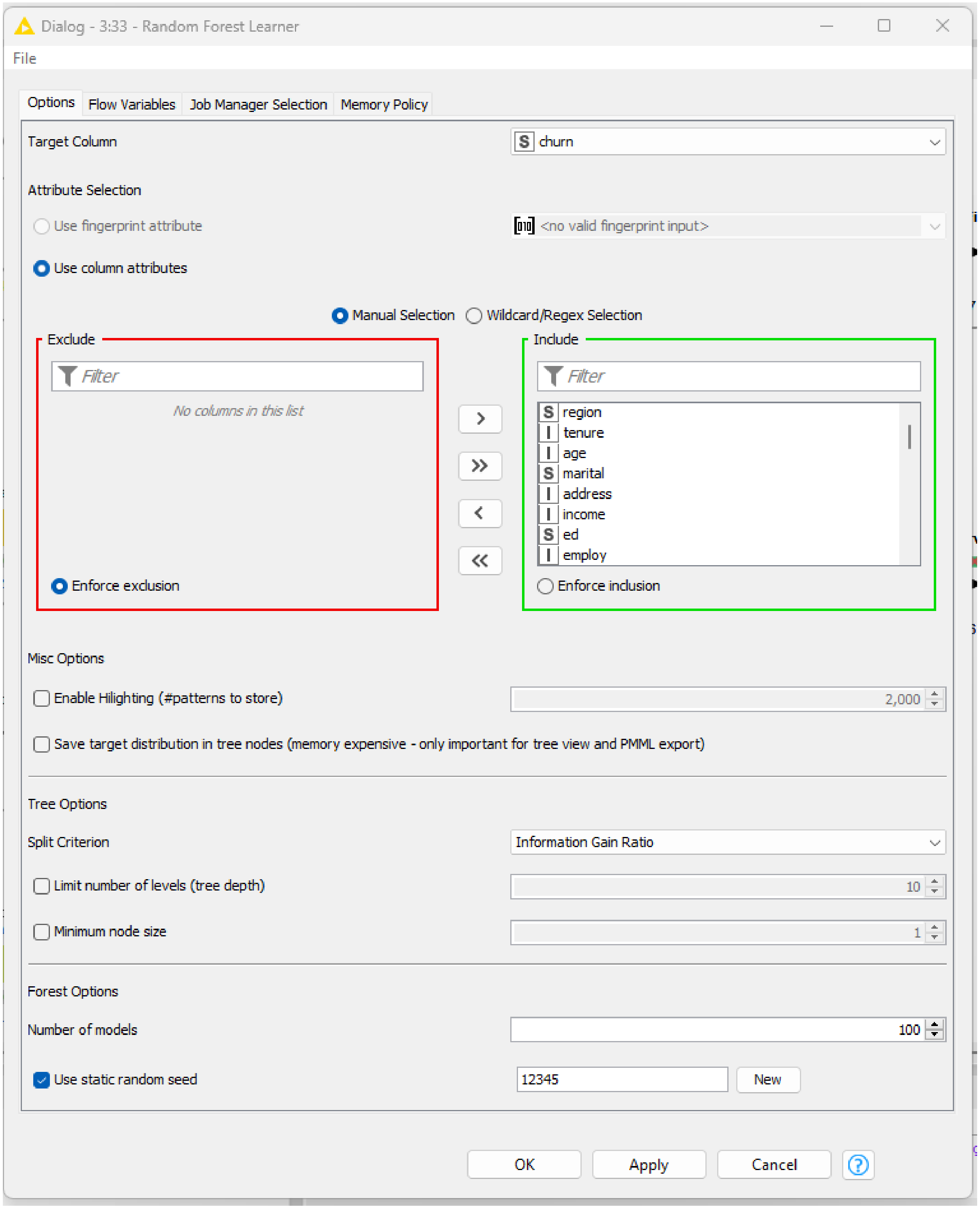

Figure 15: Random Forest Learner Node

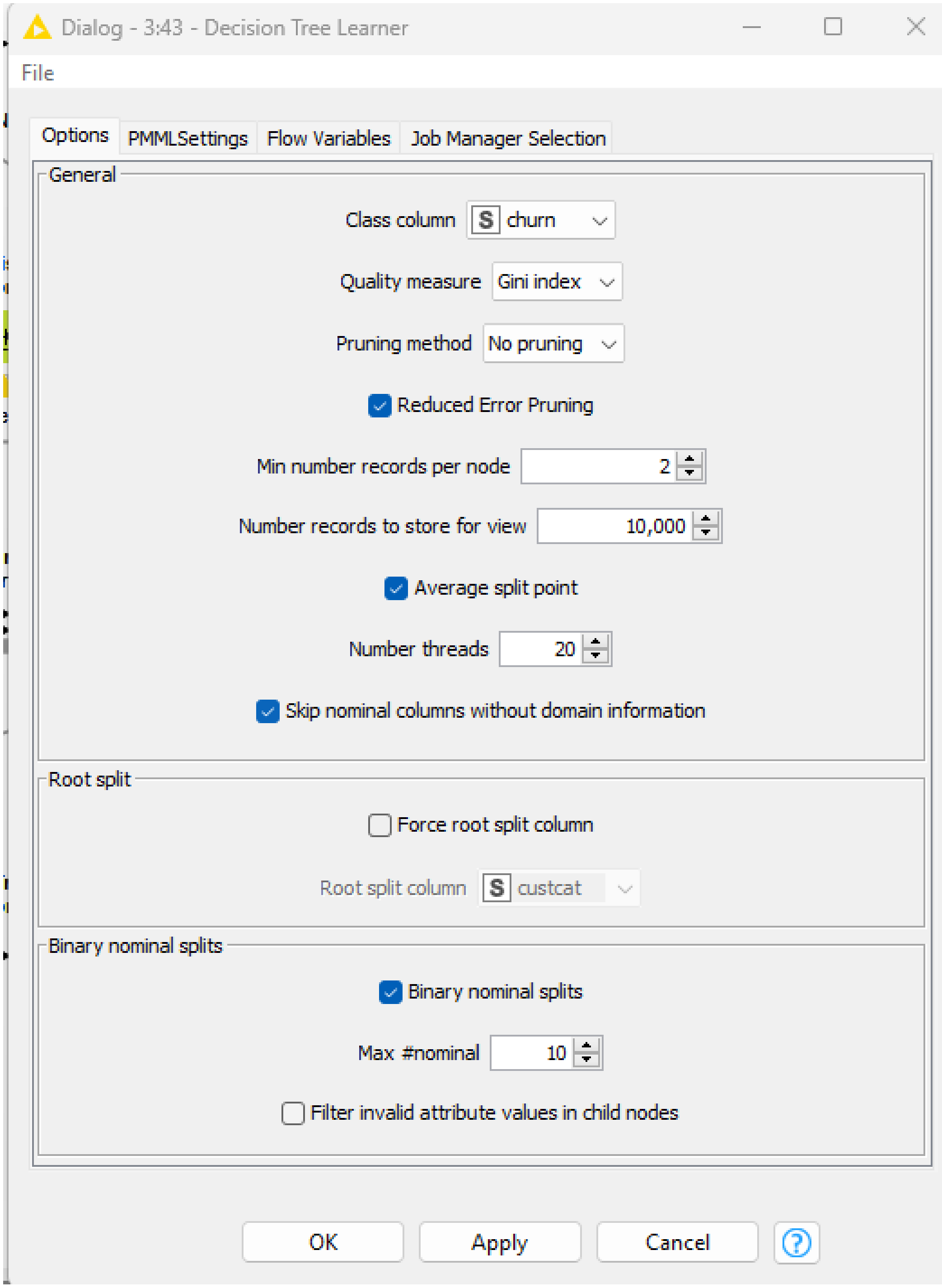

Figure 16: Decision Tree Learner Node

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.

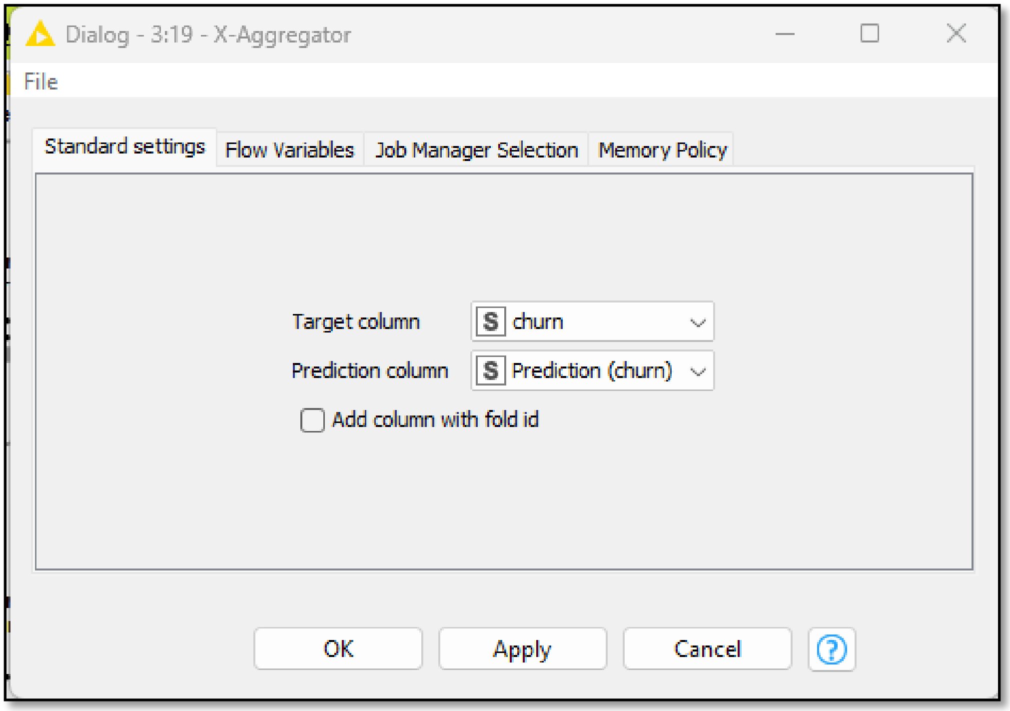

Figure 17: X-Aggregator Node

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.

Accuracy statistics - 3:20 - Scorer

File Edit Hilite Navigation View

Table "default" - Rows: 3 Spec - Columns: 11 Properties Flow Variables

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	TruePo…	FalsePo…	TrueNe…	FalseN…	Recall	Precision	Sensitivity	Specificity	F-meas…	Accuracy	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	TruePo…	FalsePo…	TrueNe…	FalseN…	Recall	Precision	Sensitivity	Specificity	F-meas…	Accuracy	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	TruePo…	FalsePo…	TrueNe…	FalseN…	Recall	Precision	Sensitivity	Specificity	F-meas…	Accuracy	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.

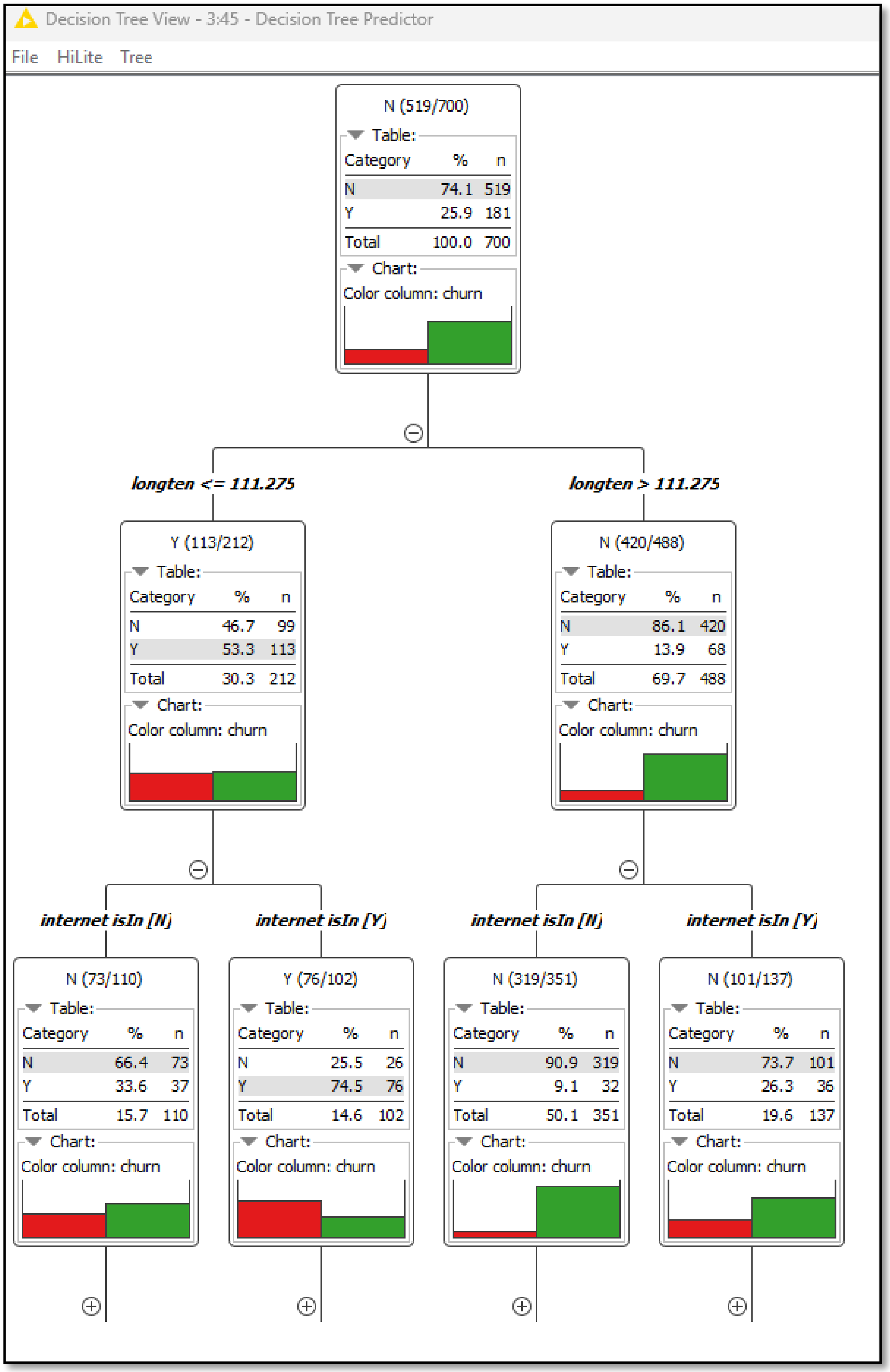

Figure 23: Decision Tree Results

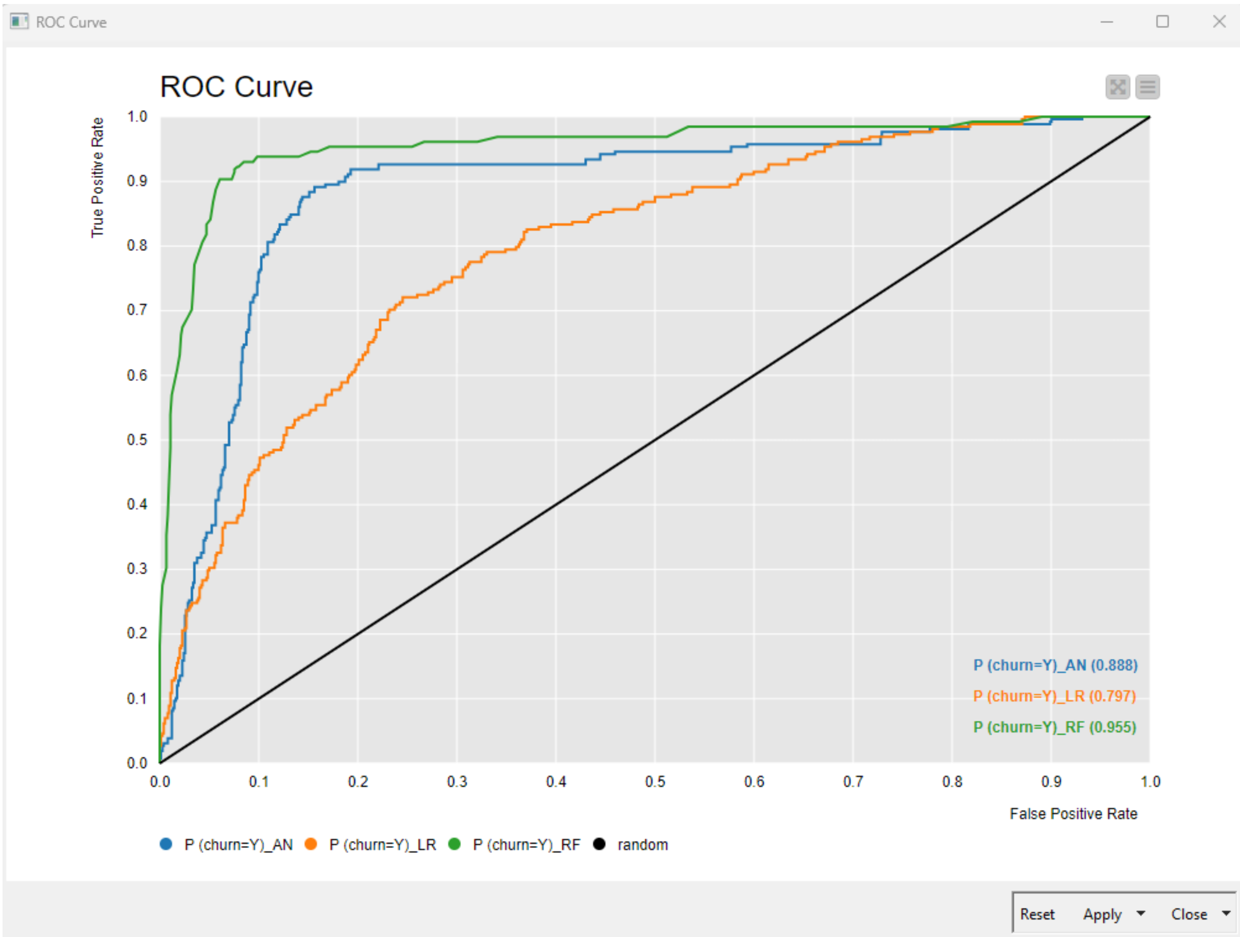

Figure 14: ROC Curve

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.

Prediction output - 3:34 - Random Forest Predictor

File Edit Hilite Navigation View

Table "default" - Rows: 100 Spec - Columns: 42 Properties Flow Variables

Row ID	S region	I tenure	I age	S marital	I address	I income	S ed	I employ	S retire	S gender	I reside	S tollfree	S equip	S callcard	S wireless	D longmon	D tollmon	D equipmon	D cardmon	D wiremon	D longten	D tollten	D equipten	I cardten
Row1	R3	68	52	Y	24	116	E1	29	N	Y	2	Y	N	Y	N	18.15	18	0	30.25	0	1,300.6	1,247.2	0	2150
Row3	R3	45	22	Y	2	19	E2	4	N	Y	5	N	N	Y	N	10.9	0	0	8.75	0	504.5	0	0	415
Row27	R1	52	27	N	6	47	E3	5	N	N	2	Y	N	Y	N	6.25	18.5	0	14.25	0	330.4	942.65	0	725
Row33	R3	28	51	Y	22	40	E3	10	N	Y	6	Y	N	N	N	8.95	21.25	0	0	0	253.55	586.95	0	0
Row34	R1	16	27	N	5	37	E3	5	N	N	4	N	N	N	N	6	0	0	0	0	80.7	0	0	0
Row43	R1	10	26	N	6	34	E3	3	N	N	1	N	Y	N	N	4.85	0	27.7	0	0	41.85	0	223.05	0
Row50	R1	13	20	N	1	17	E3	0	N	N	5	N	Y	Y	N	3.05	0	31.3	14	0	40.3	0	390.6	155
Row53	R1	35	61	N	23	41	E2	11	N	N	1	N	N	Y	N	9.6	0	0	9.5	0	353.55	0	0	295
Row57	R1	59	42	N	1	68	E2	21	N	Y	1	Y	N	Y	N	17.3	21.25	0	19.25	0	997.85	1,183.55	0	1160
Row66	R1	56	47	Y	19	65	E2	22	N	N	3	Y	N	Y	Y	12.4	48.75	0	36.5	58.4	744.45	2,695.95	0	2070
Row88	R1	64	55	N	28	104	E1	26	N	Y	1	N	N	Y	N	15	0	0	39.25	0	960.95	0	0	2360
Row95	R2	57	54	Y	20	21	E2	0	N	Y	3	N	N	Y	N	10.6	0	0	17	0	565.55	0	0	930
Row96	R1	13	21	N	2	19	E3	0	N	Y	1	N	Y	Y	N	7.9	0	25.2	5	0	98.35	0	311.65	40
Row131	R2	61	40	N	15	85	E3	15	N	Y	1	Y	N	Y	N	13.65	29.25	0	22.75	0	831.15	1,823.6	0	1310
Row145	R3	13	43	Y	1	123	E3	21	N	N	5	Y	Y	Y	N	3.85	23	33.05	6.75	0	55.05	317.5	439.3	75
Row149	R1	48	32	N	2	88	E3	9	N	N	1	Y	N	Y	Y	6.15	33.75	0	29.75	44.75	307.65	1,603.8	0	1390
Row182	R1	53	35	N	15	59	E3	5	N	Y	1	Y	N	Y	N	16.85	23.25	0	18.5	0	888.45	1,250.7	0	995
Row183	R1	47	35	Y	13	70	E4	9	N	Y	2	Y	Y	N	Y	10.75	33	47.4	0	45.85	551	1,556.2	2,138.25	0
Row188	R2	35	38	N	7	90	E2	18	N	N	4	Y	N	Y	N	6.8	17.75	0	14	0	245.95	558.35	0	425
Row201	R2	43	27	Y	3	21	E2	1	N	Y	3	N	N	Y	N	12.65	0	0	16.5	0	558.85	0	0	760
Row204	R3	39	34	N	2	54	E3	5	N	N	1	N	N	Y	N	6.5	0	0	13.25	0	265.45	0	0	460
Row207	R2	68	60	N	40	262	E2	39	N	Y	1	N	N	Y	N	31.6	0	0	26	0	2,112.25	0	0	1690
Row218	R1	52	62	N	23	36	E4	17	N	N	1	Y	N	Y	N	13.35	18.5	0	18.5	0	709.25	978.85	0	915
Row224	R2	28	20	Y	1	18	E3	0	N	N	3	N	Y	N	Y	7.7	0	38.9	0	32	190.05	0	1,041.1	0
Row235	R3	37	51	Y	26	24	E4	4	N	N	2	N	N	Y	N	8.2	0	0	13.75	0	288.4	0	0	480
Row257	R2	18	38	N	8	45	E2	2	N	Y	4	Y	Y	Y	Y	2.7	31.5	36.25	15	23.4	45.6	609.2	654.75	270
Row272	R2	48	35	N	5	63	E4	12	N	N	2	N	N	Y	N	13.4	0	0	27.5	0	560.55	0	0	1315
Row279	R2	62	76	N	20	35	E3	18	Y	Y	1	N	N	Y	N	17.25	0	0	17.75	0	1,045.7	0	0	1085
Row280	R1	65	48	N	28	94	E1	25	N	N	1	N	N	Y	N	10.4	0	0	22.5	0	664.65	0	0	1470
Row283	R3	18	28	Y	3	22	E1	2	N	Y	5	N	N	Y	N	3.15	173	0	9.5	0	55.55	0	0	175
Row290	R2	37	40	Y	14	26	E2	0	N	N	5	N	N	Y	N	18.9	0	0	12	0	653.15	0	0	470
Row291	R2	71	70	N	30	153	E3	34	Y	Y	1	Y	N	Y	Y	30.55	46.25	0	26.25	46.7	2,291.1	3,114.95	0	1805
Row297	R2	23	22	Y	1	27	E1	4	N	N	5	Y	N	Y	N	7.9	23	0	15.75	0	210.85	515.8	0	345
Row302	R2	32	34	Y	0	38	E1	10	N	N	4	N	N	Y	N	7.35	0	0	19.25	0	217.45	0	0	600
Row314	R3	69	66	Y	31	49	E2	15	N	N	2	Y	N	Y	N	23.1	29.25	0	25.5	0	1,566.8	2,063.45	0	1750
Row317	R2	9	24	Y	3	26	E4	1	N	Y	3	N	Y	N	N	7.65	0	23.25	0	0	75.25	0	169.75	0
Row324	R3	27	28	Y	0	58	E3	0	N	N	2	Y	N	N	N	5.8	22.5	0	0	0	163.65	566.15	0	0
Row325	R2	4	38	Y	13	54	E2	4	N	N	2	Y	N	Y	Y	4.75	26.25	0	48.5	24.45	21.75	115	0	140
Row331	R1	3	21	Y	1	36	E3	0	N	Y	4	N	N	N	N	3.2	0	0	0	0	13.55	0	0	0
Row346	R3	41	42	Y	12	26	E4	5	N	N	4	N	N	N	N	10.95	0	0	0	0	482.4	0	0	0
Row353	R2	6	50	N	5	37	E2	0	N	N	1	N	Y	N	N	5.35	0	23.55	0	0	38.4	0	131.2	0
Row358	R1	65	70	Y	9	115	E2	39	Y	Y	2	Y	N	Y	Y	26.85	26	0	43	33.75	1,749.6	1,629.65	0	2865
Row361	R1	1	34	Y	6	18	E1	0	N	Y	2	N	N	N	N	1.6	0	0	0	0	1.6	0	0	0
Row366	R1	3	25	N	0	65	E4	0	N	Y	1	N	Y	Y	N	3.3	0	34.25	11.25	0	5.9	0	47.65	5
Row372	R3	32	37	N	9	44	E2	7	N	N	1	N	N	N	N	11.25	0	0	0	0	359.35	0	0	0
Row388	R1	30	57	Y	16	19	E1	1	N	N	4	N	N	Y	N	7.25	0	0	8.5	0	200.1	0	0	285
Row390	R2	17	39	N	12	45	E4	10	N	Y	1	N	Y	N	N	8.65	0	28.95	0	0	133.1	0	450.2	0
Row399	R2	72	45	Y	25	98	E4	20	N	Y	2	N	N	Y	Y	30.25	0	0	13.25	42.85	2,182.65	0	0	900
Row406	R2	39	39	N	19	29	E4	6	N	N	1	N	N	Y	N	5.95	0	0	21.25	0	227.85	0	0	835
Row419	R2	43	49	Y	18	66	E1	22	N	N	4	N	N	Y	N	8.6	0	0	20.75	0	376.7	0	0	875
Row420	R2	6	27	Y	5	26	E4	2	N	N	5	Y	N	N	N	5.6	12	0	0	0	30.35	72.5	0	0
Row449	R2	3	32	N	4	58	E2	11	N	N	4	Y	Y	Y	Y	2.75	15.75	29.5	9.25	28.55	5.7	49.65	50.5	15
Row453	R2	22	33	Y	2	24	E1	3	N	N	4	N	N	N	N	7.1	0	0	0	0	153.4	0	0	0
Row459	R1	62	53	N	14	360	E3	26	N	Y	2	Y	N	Y	N	31.9	53.25	0	16.5	0	1,892.85	3,232.05	0	890
Row465	R2	39	44	Y	14	418	E4	18	N	N	2	Y	N	Y	N	7.05	27.5	0	17.5	0	269	1,046.2	0	640
Row472	R1	71	61	Y	19	155	E3	30	N	N	2	Y	N	Y	N	23.9	27.5	0	15	0	1,684.05	1,935.15	0	1030
Row505	R2	5	33	N	10	125	E4	5	N	Y	1	N	Y	N	N	4.85	0	26.15	0	0	17.25	0	110.1	0

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.