MSIS – 5633 PREDICTIVE ANALYTICS TECHNOLOGIES

Homework Assignment #4

KNIME Data Mining I

Customer Churn Prediction and Explanation

Due Date

October 15, 2023

By

Rishitha Ganagoni

A20398497

Table of Contents

Executive Summary…………………………………………………………….…….2

CRISP-DM

1. Business Understanding……………………………………………………….4

2. Data Understanding………………………………………………........................4

3. Data pre-processing……………………………………………........................5

4. Modelling………………………………………………………………………...6

5. Evaluation……………………………………………………...................12

6. Deployment…………………………………………………………………...13

Variable Importance…………………………………………………………..13

Comparing the results of a single model……………………………………...15

Comparing results of Balanced Data with Unbalanced Data…………………………16

Conclusion……………………………………………………………………….17

Executive Summary

The main objective of this report is to assess the customer churn rate within the Tele Communication Office. To achieve this, we have utilized a dataset comprising 1,000 rows and 39 columns. We have developed machine learning models to predict customer churn, and in this endeavor, we have adopted the widely endorsed CRISP-DM methodology for Data Mining.

Within this report, we have created six distinct models, namely ANN, DT, RF, LR, BT, and SVM, taking into account a variety of factors to determine the most effective model for pattern recognition. Our evaluation process involves considering sensitivity, specificity, and ROC curve values to identify the model with the highest accuracy.

Furthermore, a significant portion of our effort has been dedicated to data pre-processing, which entails the removal of specific values and the replacement of missing data points with the median.

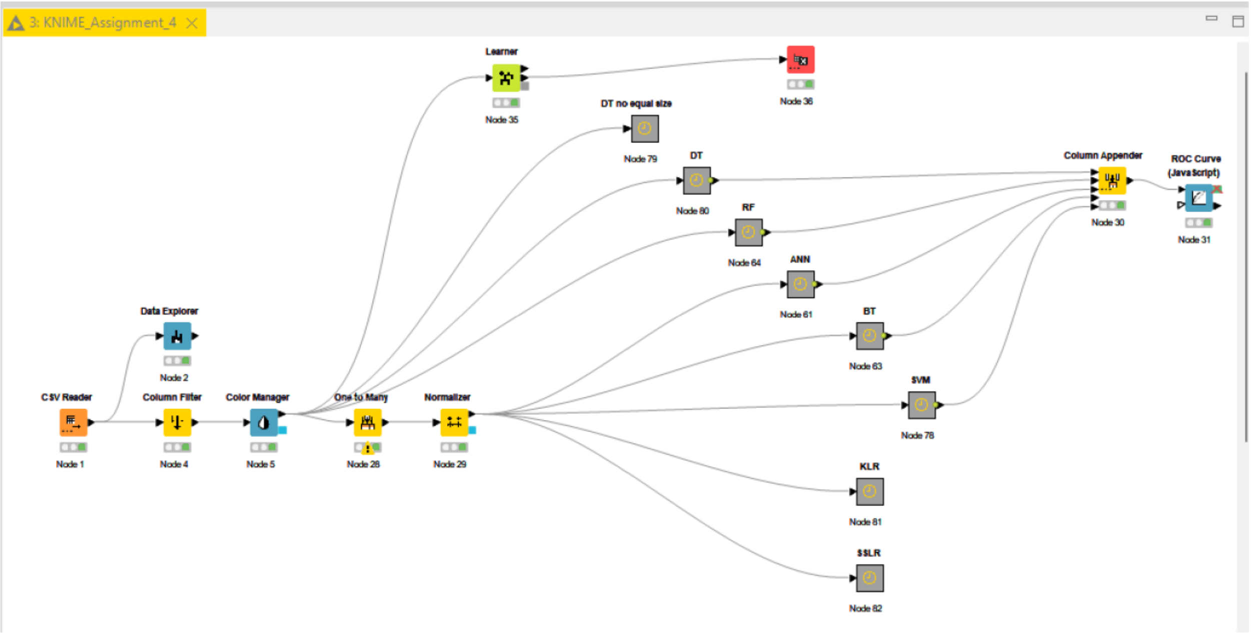


Figure 1: Overview of Workflow of all six models

CRISP-DM Methodology:

Debuted in 1999, it is known by its acronym, CRISP-DM, which stands for CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING, is widely respected and frequently employed as one of the most prevalent data mining methodologies. CRISP-DM serves as a process model that offers a comprehensive view of the data mining life cycle. CRISP-DM aims to address the following key stages in data mining:

Business Comprehension

Data Exploration

Data Cleansing and Preparatory Procedures

Application of Modeling Techniques

Model Assessment

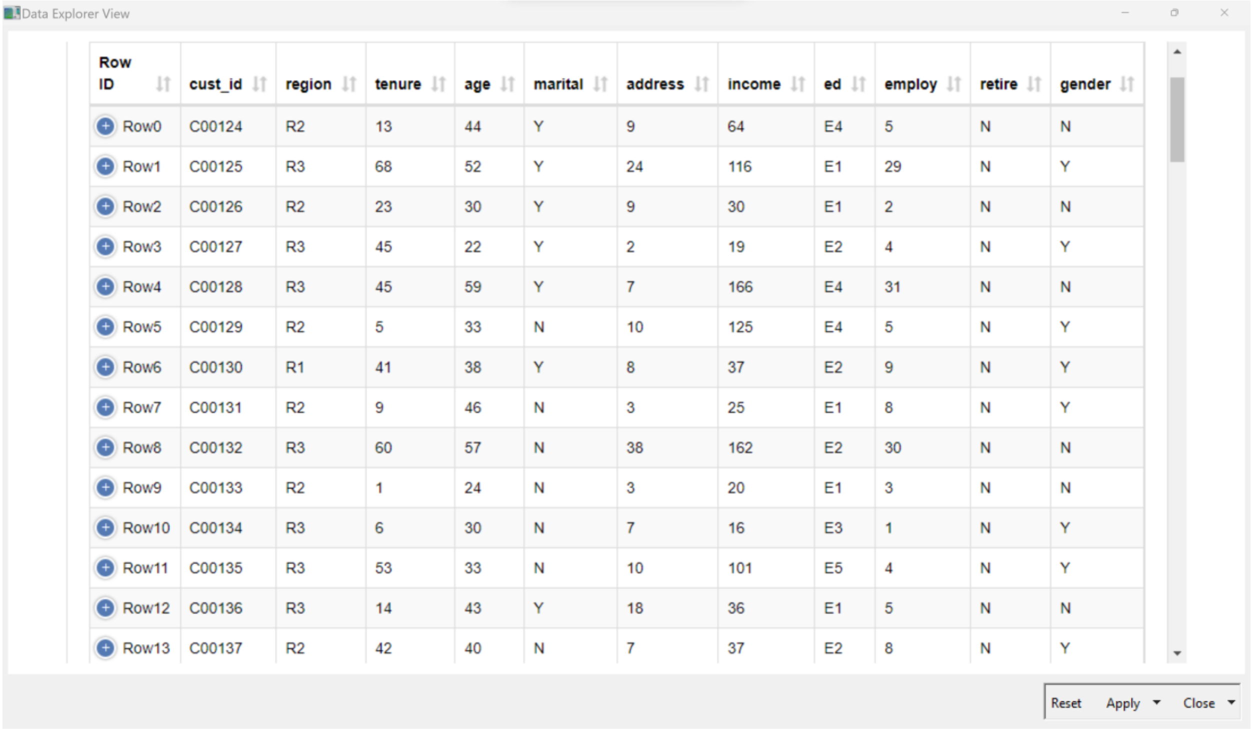
Implementation

Business Comprehension:

There is no project without any business need. If there is no business requirement means no application of Data Mining techniques. For any project, understanding the customer requirements and drill downing according to it and fulfilling the customer requirement is the main motto. Firstly, We need to gather and analyze the data and according to that we have to take further steps. Here, in this project, we are finding whether the customers are interested or not in the Tele Communication Service, we are predicting the future of the Telecommunication whether the customers are interested or not in further by using predicting models.

Data Exploration:

To do anything first we have to understand the given data which is the key step in any project. Removing the unwanted data and analyzing it, improves the richness of the Data.



Row ID	cust_id	region	tenure	age	marital	address	income	ed	employ	retire	gender
Row0	C00124	R2	13	44	Y	9	64	E4	5	N	N
Row1	C00125	R3	68	52	Y	24	116	E1	29	N	Y
Row2	C00126	R2	23	30	Y	9	30	E1	2	N	N
Row3	C00127	R3	45	22	Y	2	19	E2	4	N	Y
Row4	C00128	R3	45	59	Y	7	166	E4	31	N	N
Row5	C00129	R2	5	33	N	10	125	E4	5	N	Y
Row6	C00130	R1	41	38	Y	8	37	E2	9	N	Y
Row7	C00131	R2	9	46	N	3	25	E1	8	N	Y
Row8	C00132	R3	60	57	N	38	162	E2	30	N	N
Row9	C00133	R2	1	24	N	3	20	E1	3	N	N
Row10	C00134	R3	6	30	N	7	16	E3	1	N	Y
Row11	C00135	R3	53	33	N	10	101	E5	4	N	Y
Row12	C00136	R3	14	43	Y	18	36	E1	5	N	N
Row13	C00137	R2	42	40	N	7	37	E2	8	N	Y

Figure 2: Cleaned data after Data Exploration

Data Cleansing and Preparatory Procedures:

Data pre-processing involves the data reduction, cleaning the data according to the requirements and transforming it. This step plays a critical role in analyzing the data because removing the unwanted data and analyzing it as the further steps are depended on this step itself. Here the data is given in the excel sheet. Importing the excel sheet into cvs reader and exploring the data in the Data Exploartion. Here the missing values are replaced with the median values. Used the Column Filter where the Cust_ID column has been removed because this is unique where it is not used to combine with anything. Used the Color Manager to differentiate the percent of Yes and No (Customer churning) for easy identification.

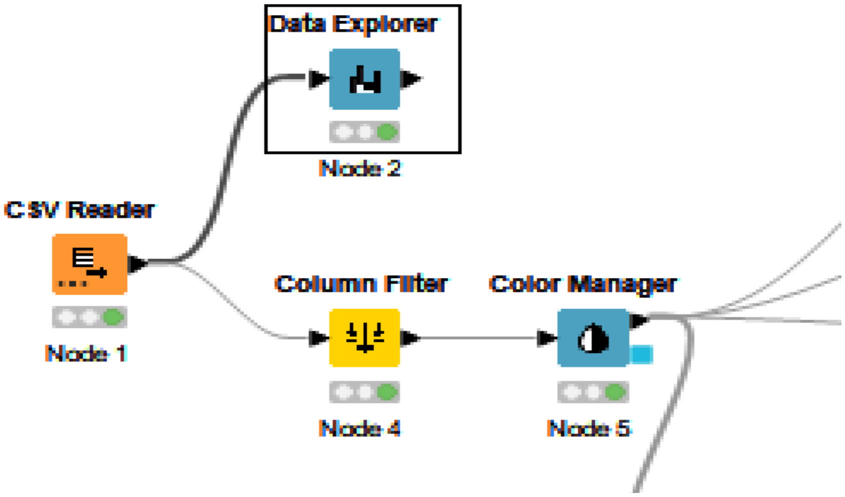
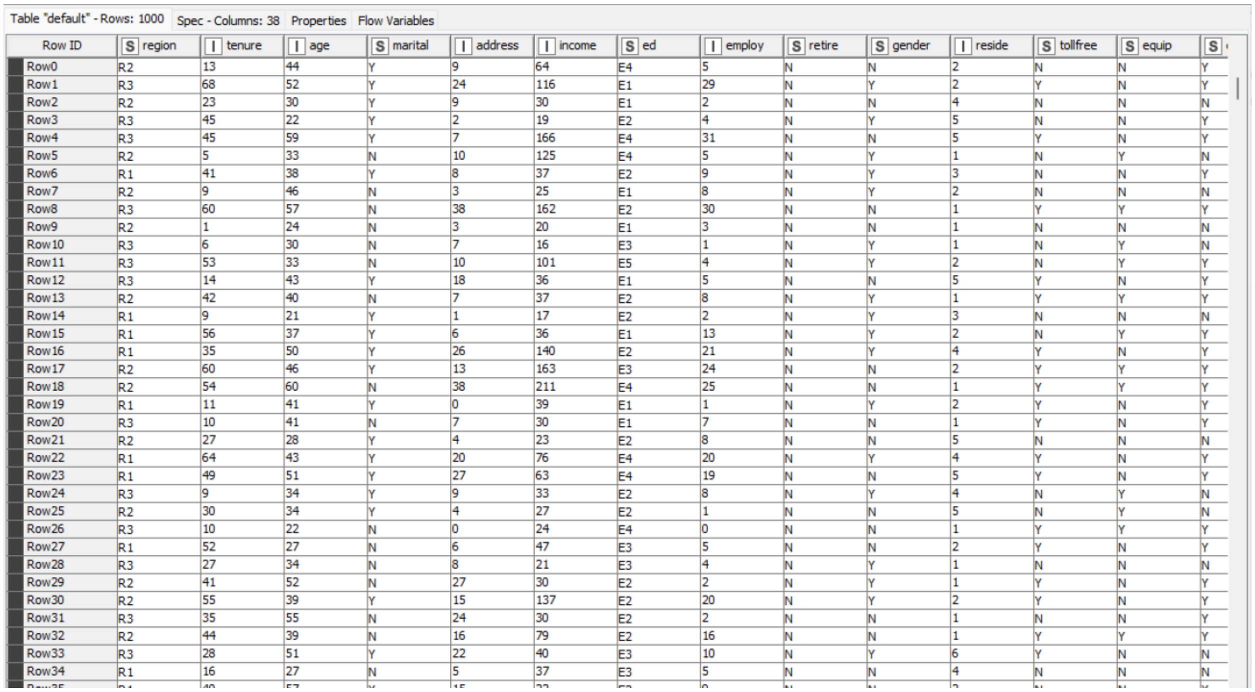


Figure 3: Data Pre-Processing done through Data Explorer, Column Filter and Color Manager nodes.



Row ID	region	tenure	age	marital	address	income	ed	employ	retire	gender	reside	tollfree	equip	
Row0	R2	13	44	Y	9	64	E4	5	N	N	2	N	N	Y
Row1	R3	68	52	Y	24	116	E1	29	N	Y	2	Y	N	Y
Row2	R2	23	30	Y	9	30	E1	2	N	N	4	N	N	N
Row3	R3	45	22	Y	2	19	E2	4	N	Y	5	N	N	Y
Row4	R3	45	59	Y	7	166	E4	31	N	N	5	Y	N	Y
Row5	R2	5	33	N	10	125	E4	5	N	Y	1	N	Y	N
Row6	R1	41	38	Y	8	37	E2	9	N	Y	3	N	N	Y
Row7	R2	9	46	N	3	25	E1	8	N	Y	2	N	N	N
Row8	R3	60	57	N	38	162	E2	30	N	N	1	Y	Y	Y
Row9	R2	1	24	N	3	20	E1	3	N	N	1	N	N	N
Row10	R3	6	30	N	7	16	E3	1	N	Y	1	N	Y	N
Row11	R3	53	33	N	10	101	E5	4	N	Y	2	N	Y	Y
Row12	R3	14	43	Y	18	36	E1	5	N	N	5	Y	N	Y
Row13	R2	42	40	N	7	37	E2	8	N	Y	1	Y	Y	Y
Row14	R1	9	21	Y	1	17	E2	2	N	Y	3	N	N	N
Row15	R1	56	37	Y	6	36	E1	13	N	Y	2	N	Y	Y
Row16	R1	35	50	Y	26	140	E2	21	N	Y	4	Y	N	Y
Row17	R2	60	46	Y	13	163	E3	24	N	N	2	Y	Y	Y
Row18	R2	54	60	N	38	211	E4	25	N	N	1	Y	Y	Y
Row19	R1	11	41	Y	0	39	E1	1	N	Y	2	Y	N	Y
Row20	R3	10	41	N	7	30	E1	7	N	N	1	Y	N	Y
Row21	R2	27	28	Y	4	23	E2	8	N	N	5	N	N	N
Row22	R1	64	43	Y	20	76	E4	20	N	Y	4	Y	N	Y
Row23	R1	49	51	Y	27	63	E4	19	N	N	5	Y	N	Y
Row24	R3	9	34	Y	9	33	E2	8	N	Y	4	N	Y	N
Row25	R2	30	34	Y	4	27	E2	1	N	N	5	N	Y	N
Row26	R3	10	22	N	0	24	E4	0	N	N	1	Y	Y	Y
Row27	R1	52	27	N	6	47	E3	5	N	N	2	Y	N	Y
Row28	R3	27	34	N	8	21	E3	4	N	Y	1	N	N	N
Row29	R2	41	52	N	27	30	E2	2	N	Y	1	Y	N	Y
Row30	R2	55	39	Y	15	137	E2	20	N	Y	2	Y	N	Y
Row31	R3	35	55	N	24	30	E2	2	N	N	1	N	N	Y
Row32	R2	44	39	N	16	79	E2	16	N	N	1	Y	Y	Y
Row33	R3	28	51	Y	22	40	E3	10	N	Y	6	Y	N	N
Row34	R1	16	27	N	5	37	E3	5	N	N	4	N	N	N

Figure 4: Filtered Data after go through the Column Filter.

Application of Modeling Techniques:

This step also plays a vital role in Data Mining. Applying the Data model according to the data present is the key step in designing a model. Here, according to the data I am using six distinct data models namely ANN, DT, RF, LR, BT, and SVM.

Artificial Neural Networks(ANN):

ANN is a modeling technique inspired from Human Nervous system. It represents the data by a physical phenomenon or from a decision process. There is a unique feature for this modeling is to establish a relationship between the dependent variables and independent variables that extracts complex knowledge from the data sets.

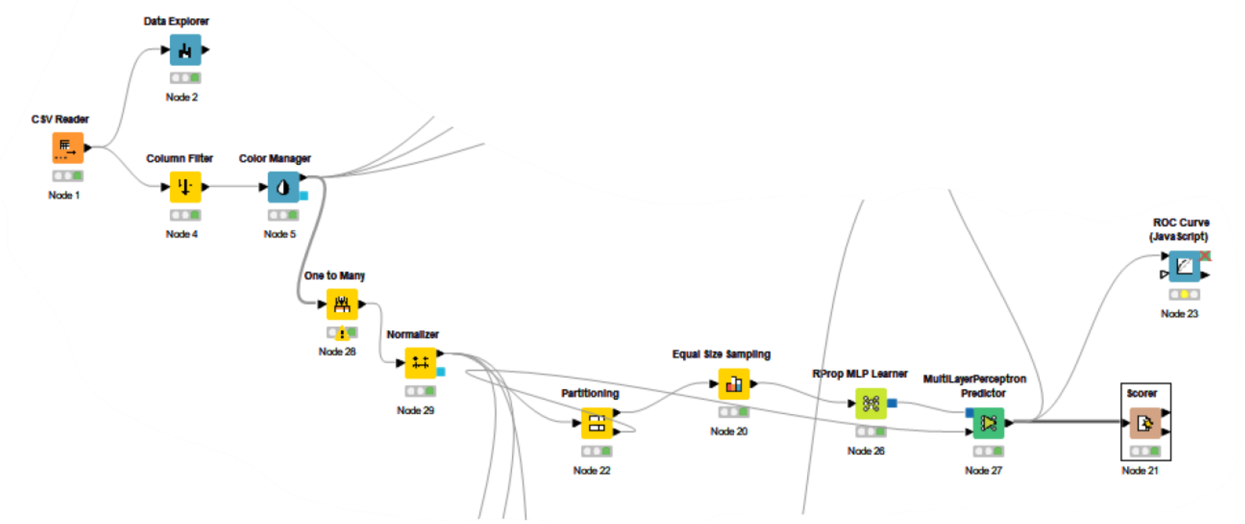


Figure 5: Designing Model for ANN

Support Vector Machine(SVM):

This extensively utilized machine learning algorithm partitions data points into distinct categories by creating a hyperplane or a series of hyperplanes within a multi-dimensional space. It is applicable to tasks involving both classification and regression.

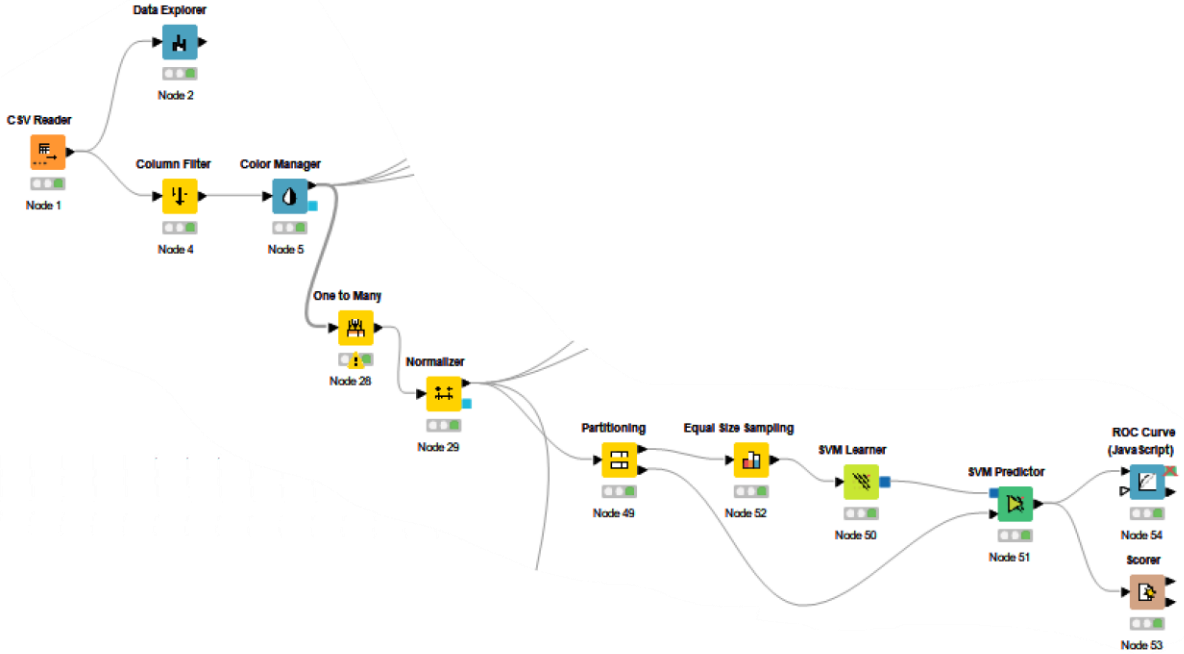


Figure 6: Designing Model for SVM

Decision Tree:

This algorithm is most widely used for classification and regression models. It splits the data into subsets based on the features of Data set and creates a tree-like structure of decision to predict the target variable. This process is repeated recursively for each subset until the criteria is met.

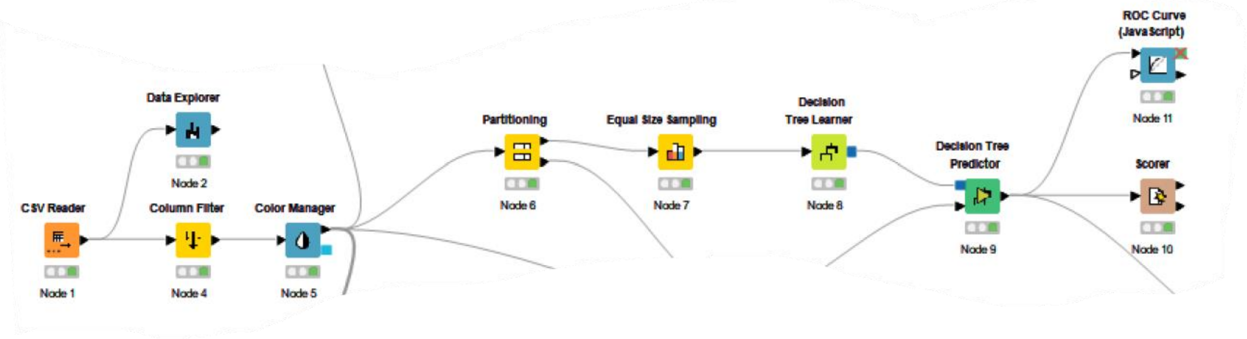
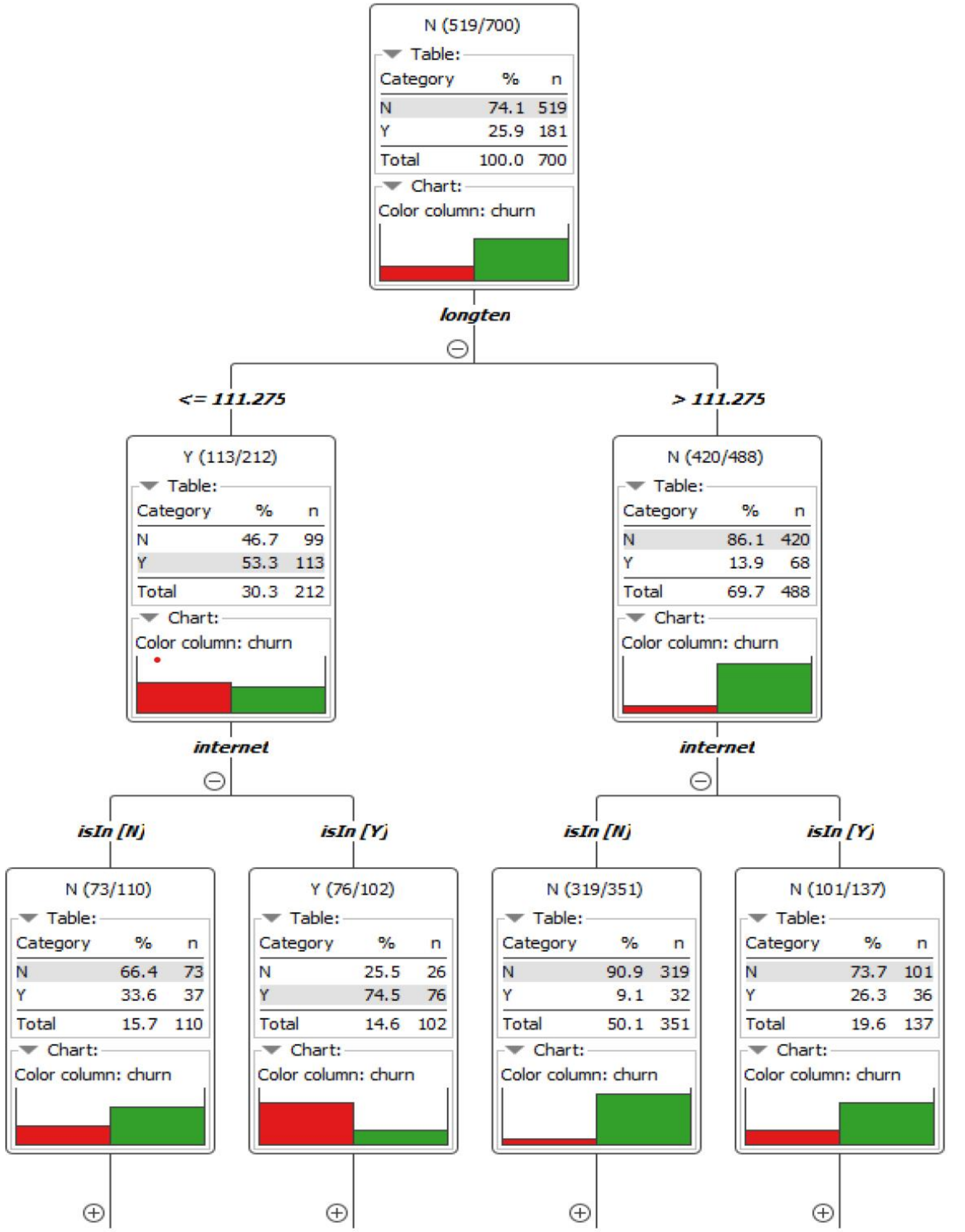


Figure 7: Designing Model for Decision Tree



N (519/700)

Category	%	n
N	74.1	519
Y	25.9	181
Total	100.0	700

Y (113/212)

Category	%	n
N	46.7	99
Y	53.3	113
Total	30.3	212

N (420/488)

Category	%	n
N	86.1	420
Y	13.9	68
Total	69.7	488

N (73/110)

Category	%	n
N	66.4	73
Y	33.6	37
Total	15.7	110

Y (76/102)

Category	%	n
N	25.5	26
Y	74.5	76
Total	14.6	102

N (319/351)

Category	%	n
N	90.9	319
Y	9.1	32
Total	50.1	351

N (101/137)

Category	%	n
N	73.7	101
Y	26.3	36
Total	19.6	137

Figure 8: Shows the 3 level split of Decision Tree.

The 3 level split of decision tree depicts that it is trying to predict the customer will churn (i.e., cancel their subscription) or not upon two factors:

1. Customers in the city with low usage of Internet are more prone to churn.
2. Customers not in the city but having high usage of Internet are more prone to churn.

Random Forests:

This approach involves combining multiple decision trees to enhance the accuracy and robustness of predictions, constituting an ensemble method. It is constructed on different subsets of Datasets. During the training process, each decision tree is trained independently with its own subset of the data set. Once all the trees are trained, they are combined to do the final prediction. This is made by taking the majority of all individual tree predictions.

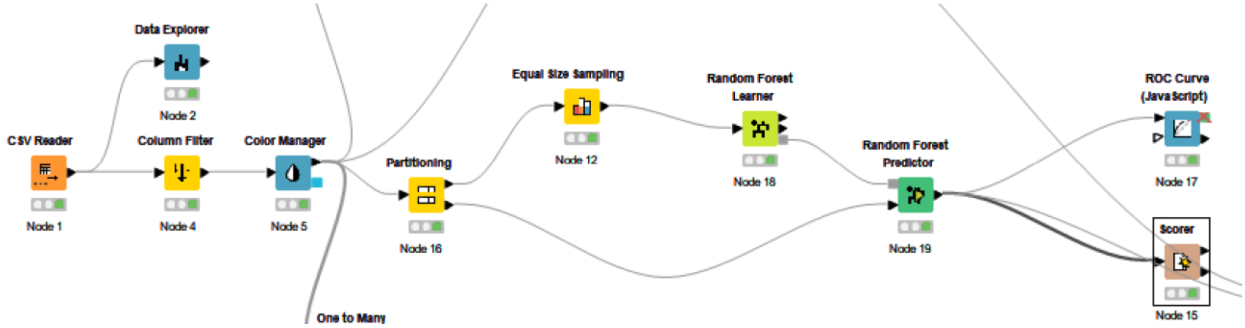


Figure 9: Designing Model of Random Forests

Logistic Regression:

Logistic regression is a statistical method employed to predict binary outcomes, such as a positive or negative outcome, using past observations from a dataset. This model assesses the connection between one or more independent variables within the dataset to anticipate dependent variable.

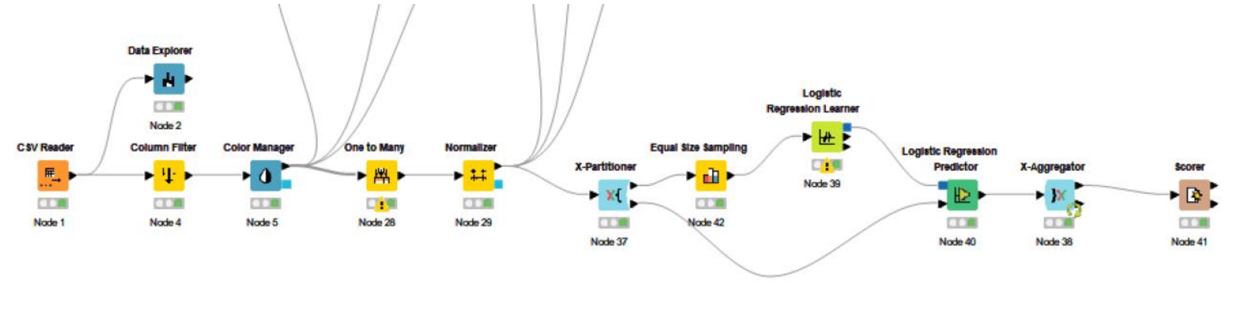


Figure 10: Designing model for Logistic Regression.

Boosted Trees:

Boosted trees are a technique that assembles numerous weaker models to form a robust model. The fundamental concept behind boosting trees is to progressively introduce new models to the ensemble, with each one aimed at rectifying the mistakes of the preceding models. The ultimate model is a weighted combination of all the individual models. This procedure is carried out iteratively, with each subsequent model concentrating on addressing the errors of the prior models.

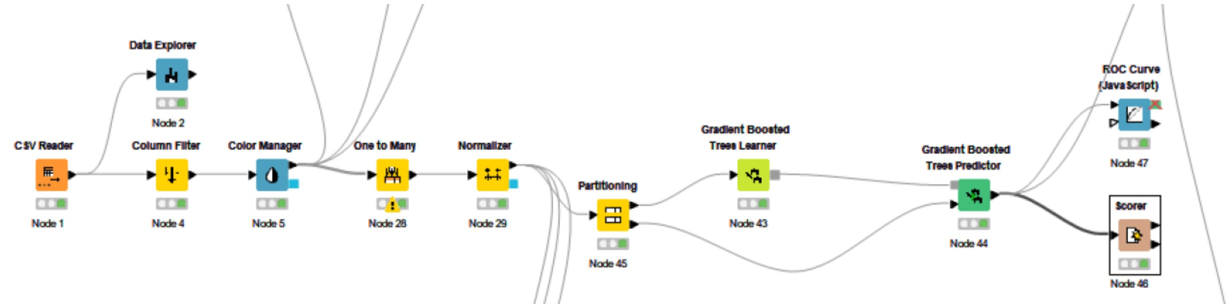


Figure 11: Designing model for Boosted Trees

Evaluating the Models:

For every model we have some particular terms to evaluate the models they are accuracy, Sensitivity, specificity and roc on curve.

Model	Accuracy %	Sensitivity	Specificity	ROC on Curve
ANN	76.66	0.87	0.731	0.838
RF	81.33	0.792	0.821	0.889
DT	72	0.688	0.731	0.759
LR	73.1	0.713	0.737	0.792
SVM	64.66	0.623	0.655	0.755
BT	91.1	0.701	0.982	0.91

Table 1: Showing Accuracy, Sensitivity, Specificity and ROC Value.

Giving all the testing data to ROC Curve to know the best model among them. Combined all the testing data to a Column Appender and then connected to ROC Curve.

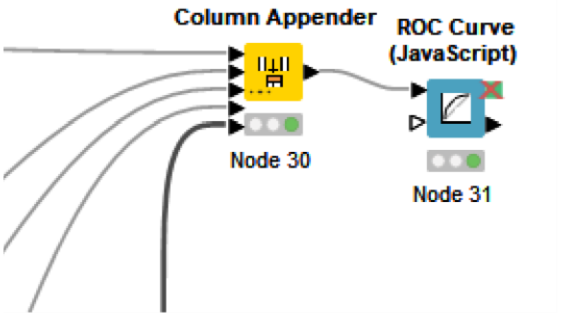


Figure 12: Inputs to Column Appender and ROC Curve.

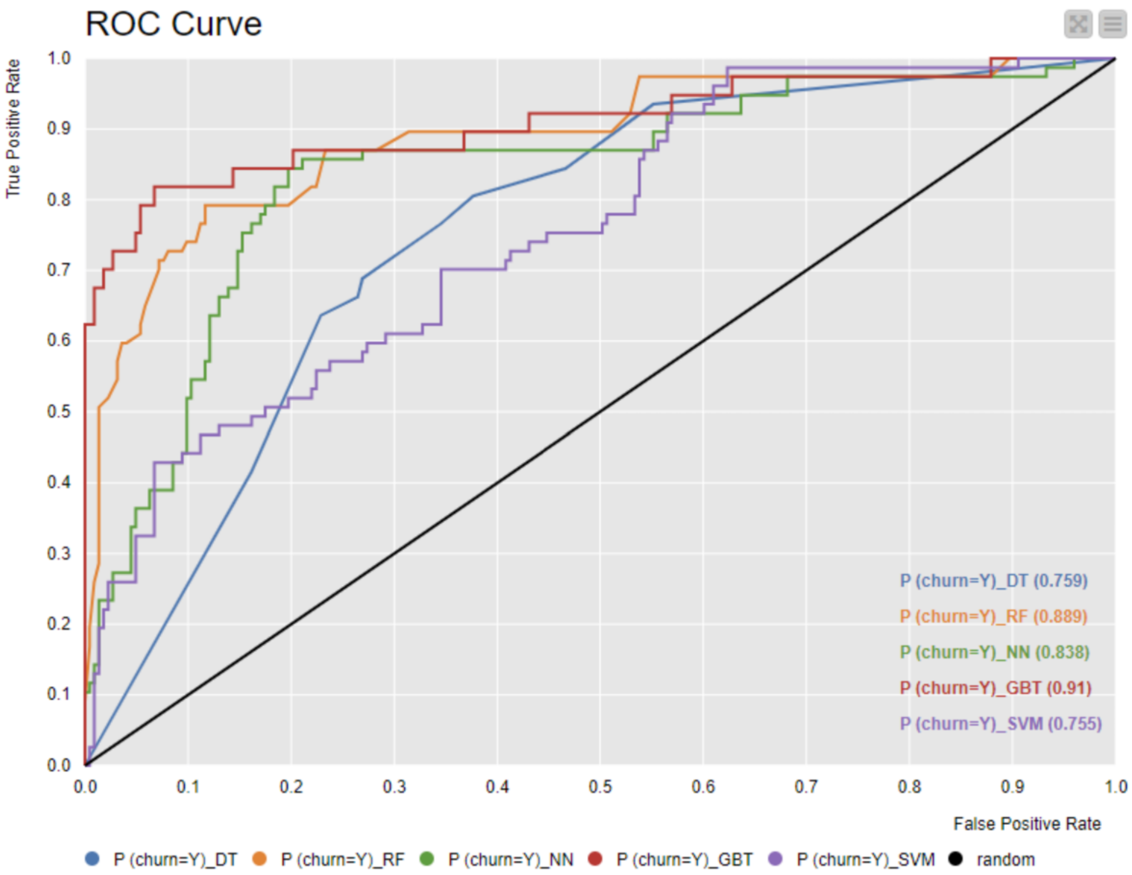
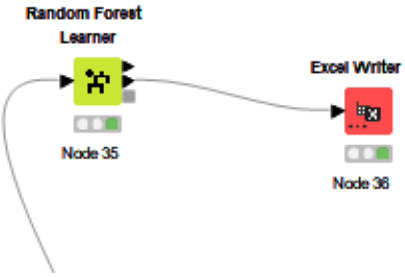


Figure 13: ROC Curve of all models

Deployment

As of now we are not doing any deployment of the model into the real world.

Variable Importance:



Here, we are giving input to Random Forest from Color Manager and then to Excel Writer. All the Data has been again imported to excel file. There we have calculated the Weighted Sum. Then taken graph of row_id and w_sum.

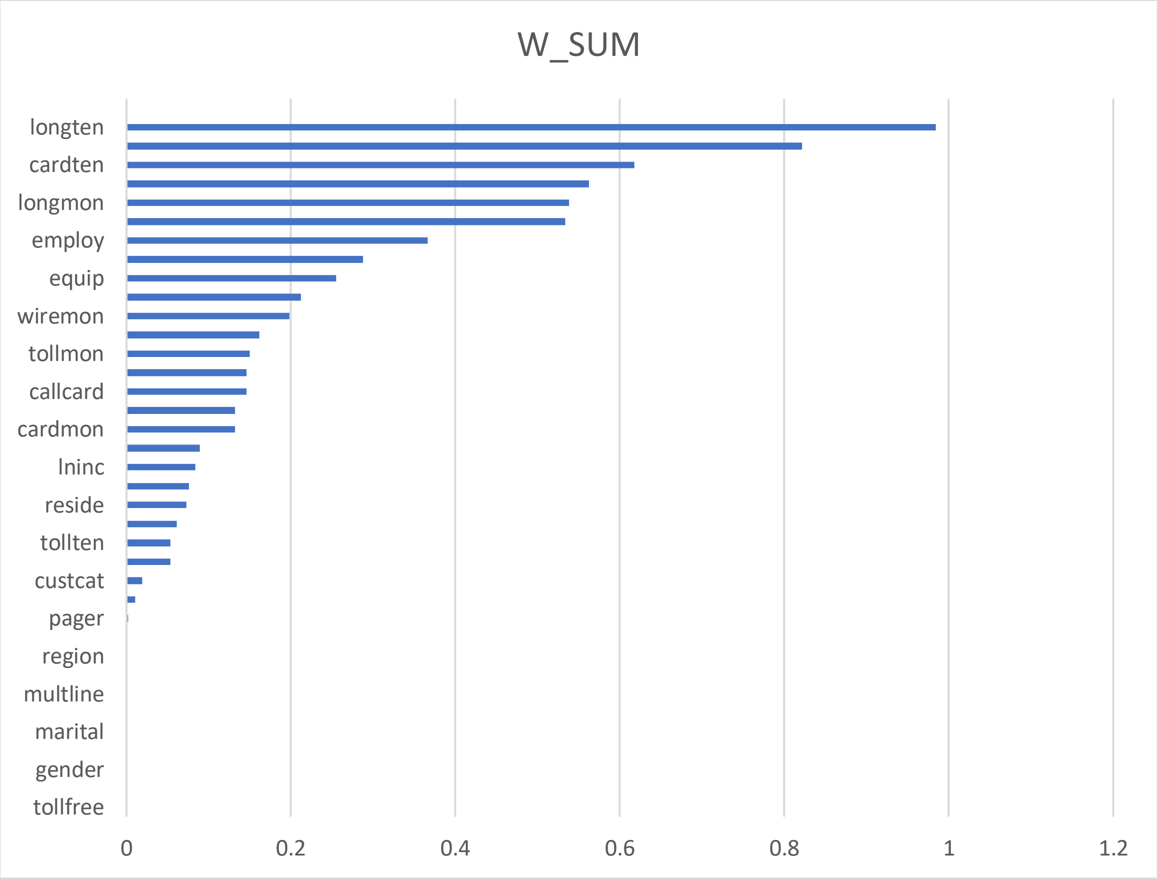


Figure 14: Graph shows the majority of decision tree where the Random Forest taken.

Comparing the results of k-fold of Logistic Regression with single-split of Logistic Regression.

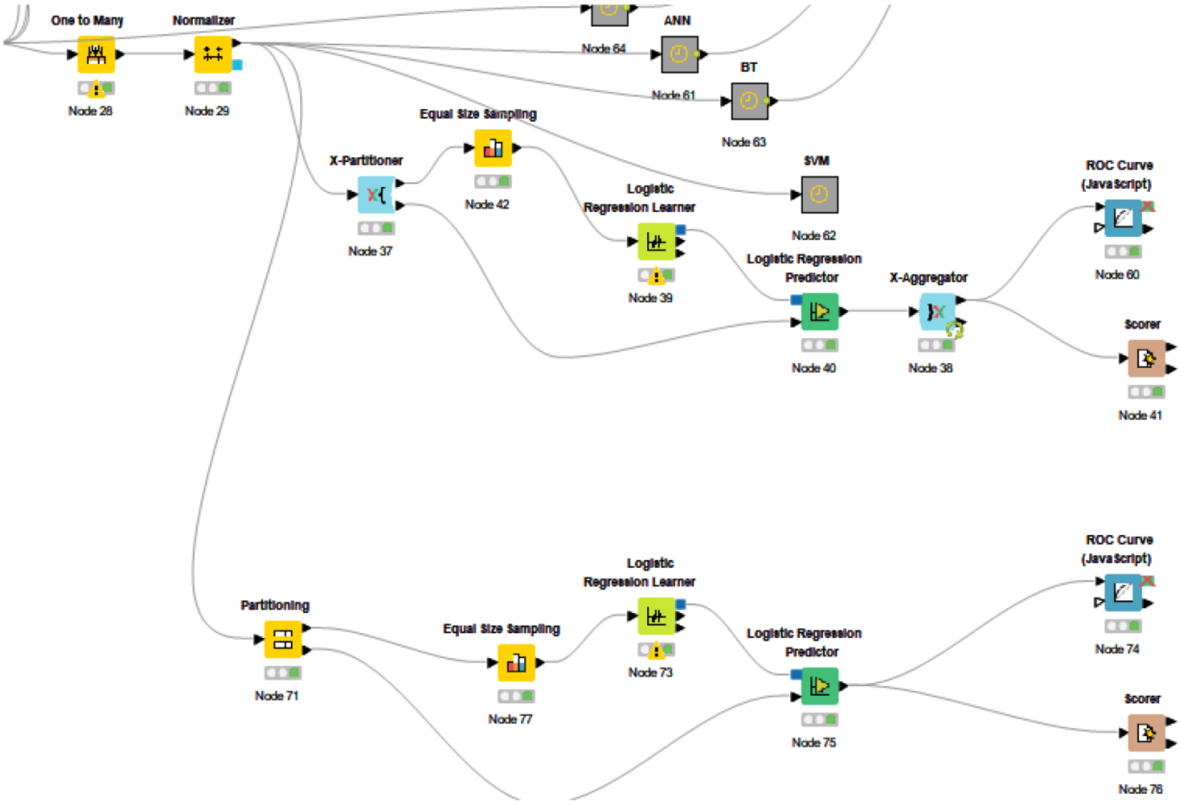


Figure 15: Workflow of Logistic Regression model of k-fold and single-single.

From the above workflow it is clear that the ROC Curve value of k-fold LR is 0.792 and accuracy is 73.1% with sensitivity of 0.71 and specificity of 0.73 And the ROC Curve value of single-split is 0.776 and accuracy is 67% with sensitivity of 0.66 and specificity of 0.67. Hence, it is clear that the model produces high accuracy for k-fold rather than single-split. For better Accuracy k-fold models are preferrable.

Comparing results of Balanced Data with Unbalanced Data

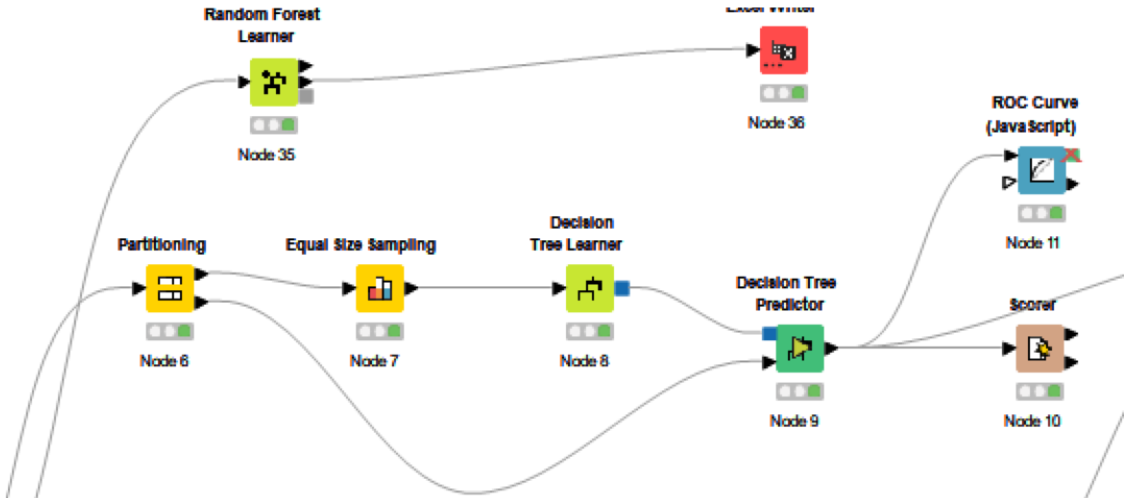


Figure 16: Decision Tree model with equal size sampler

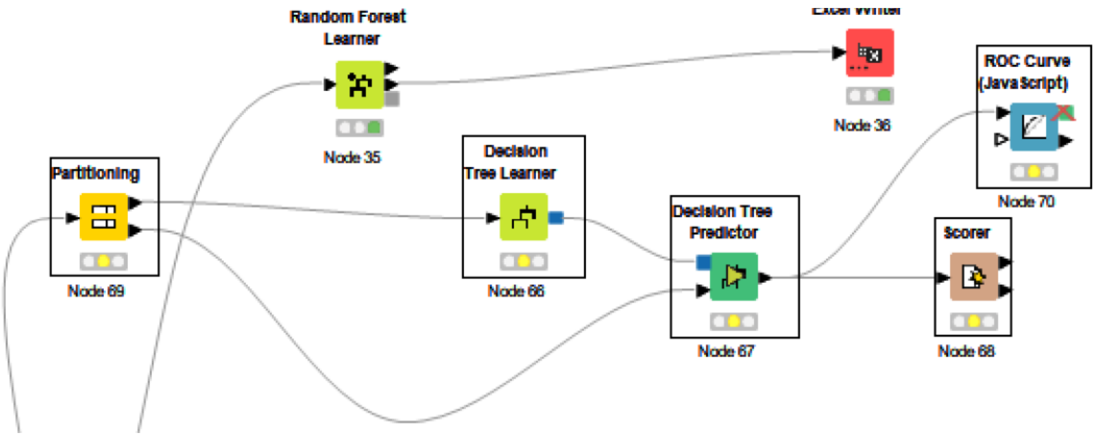


Figure 17: Decision Tree model without equal size sampler

It is clear that with equal size sampler the accuracy is 72% and the values of sensitivity and specificity are 0.68 and 0.73 whereas for without equal size sampler the accuracy is 80.66% and the values of sensitivity and specificity are 0.519 and 0.906. It is clear that without equal size sampler it calculates all the true No's but we are in consideration of true Y's. This is the main difference between them.

Conclusion

To find that people are against or support of Telecommunication. For that we build six predicting models. In consideration of all these accuracy, sensitivity, specificity and ROC on Curve Boosted Trees have highest accuracy followed by Random Forest. The accuracy values are Boosted Tree – 91% and Random Forest – 81.33% respectively. It is good in predicting the actual negative values but in actual positive values.