

MSIS-4673-26637 – Predictive Analytics Technologies

Section – In - class

Homework Assignment #4

Customer Churn Prediction and Explanation with KNIME Analytics Platform

Due Date

March 10, 2023

By

Rhoda Alawiye

## CONTENTS

EXECUTIVE SUMMARY .....	3
BUSINESS UNDERSTANDING .....	3
DATA UNDERSTANDING .....	4
DATA PREPARATION .....	5
MODELING .....	6
EVALUATION.....	9
DEPLOYMENT .....	10

## EXECUTIVE SUMMARY

The advancement in the world of machine learning techniques has brought powerful tools that have revolutionized customer churn predictions for anticipation of customer attrition. Machine learning algorithms can use the learning patterns from past/historical data and make predictions on future data. By leveraging historical data, including demographic information, transaction records, and behavioral insights, machine learning algorithms can help forecast the likelihood of customer churn with remarkable accuracy. Customer churn significance transcends industries, impacting businesses in banking and finance, healthcare, and e-commerce. Aside from the major disadvantage of customer churn, which is revenue loss, churn also undermines the company's reputation.

This research paper aims to predict customer churn rates using real-world customer churn dataset and compare the performance of various machine learning classifiers for optimized prediction accuracies. KNIME open-source tool was used, and the classifiers used in this study are decision tree, random forest, neural network, and logistic regression.

The paper is structured to provide an organized comprehensive understanding of the customer churn prediction processes. The CRISP-DM approach was used for the analysis which commenced with the business understanding, data understanding, data preparation, model building, selection, and evaluation methodologies, all culminated in a detailed analysis of results and actionable insights for mitigating customer churn.

By harnessing the predictive capabilities of machine learning, businesses can proactively address customer attrition, safeguard revenue streams, and ultimately enhance customer satisfaction. This paper contributes valuable insights to strategic decision-making processes, empowering businesses to thrive in an increasingly competitive landscape.







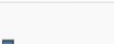
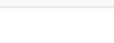
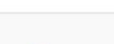
## BUSINESS UNDERSTANDING

The issue of customer churn poses a significant concern for organizations in a variety of industries, including healthcare, utilities, financial institutions, retailers, telecommunication companies and others. The consequences are severe, as losing customers not only negatively affects revenue but it also affects a company's reputation spanning both current and future customers. Hence, organizations need to proactively identify customers that are likely to churn and implement effective retention strategies. Although machine learning techniques are useful prediction tools, it is worth noting that they are not designed to provide final solutions to all the problems; rather, they help provide a framework for informed decision-making supported by empirical data. Machine learning algorithms analyze previous customer datasets to identify trends and contributing reasons to customer attrition. With this information, firms may implement focused activities to reduce churn risk, such as customizing tailored incentives, improving customer support channels, or refining product and service offerings. Whether the focus is on strengthening current customer







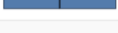
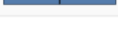

connections through service innovations or exploring tactics to recruit new consumers, these results provide vital direction for strategic maneuvering within the competitive environment.

## DATA UNDERSTANDING

There are two datasets whereby the first dataset comprises of customer attributes information such as demographics, transaction history, and usage patterns and the second dataset comprise of the customer churn target information. The aim is to evaluate the performance of the selected machine learning classifiers and identify the model that is most suitable for predicting customer churn. The first dataset consists of 38 unique predictor variables, of which 20 are nominal variables and the other 18 are interval variables. The data was split into social-demographic attributes (age, income, education, gender, and others) and behavioral attributes (usage hours and account selected services). These were the predictive modeling input variables. **Figure 1, Figure 1b** and **Figure 2** entail information about the variables and histograms for all the numeric and nominal variables respectively.

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness	Kurtosis	Overall Sum	No. zeros	No. missings	No. NaN	No. +∞	No. -∞	Histogram
Address	<input type="checkbox"/>	0	55	11.571	10.128	102.572	1.150	1.092	11548	56	0	0	0	0	
Age	<input type="checkbox"/>	18	77	41.536	12.687	160.963	0.392	-0.566	41453	0	0	0	0	0	
Cardmon	<input type="checkbox"/>	0	109.250	13.865	14.382	206.843	1.753	5.606	13837.750	314	0	0	0	0	
Cardten	<input type="checkbox"/>	0	7515	614.449	874.981	765592.123	2.927	13.006	613220	314	0	0	0	0	
Employ	<input type="checkbox"/>	0	47	10.783	10.150	103.015	1.118	0.665	10761	106	0	0	0	0	
Equipmon	<input type="checkbox"/>	0	77.700	13.599	18.860	355.697	0.939	-0.474	13572.200	628	0	0	0	0	
Equipten	<input type="checkbox"/>	0	4758.050	441.176	848.871	720582.434	2.216	4.614	440293.200	628	0	0	0	0	
Income	<input type="checkbox"/>	9	732	71.844	80.349	6455.986	3.875	21.248	71700	0	0	0	0	0	
InInc	<input type="checkbox"/>	2.197	6.596	3.925	0.779	0.606	0.600	0.170	3917.414	0	0	0	0	0	

*Figure 1: Showing Histogram and Information about the numeric variables*

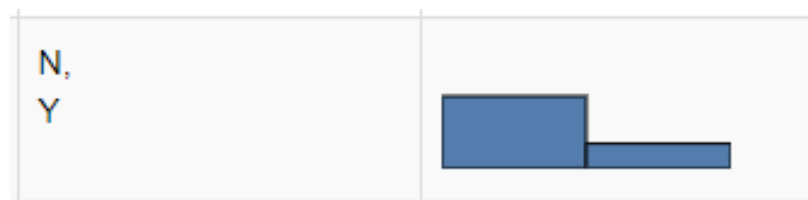
Marital	<input type="checkbox"/>	0	2	Y, N	
Multiline	<input type="checkbox"/>	0	2	N, Y	
Pager	<input type="checkbox"/>	0	2	N, Y	
Region	<input type="checkbox"/>	0	3	R3, R2, R1	
Retire	<input type="checkbox"/>	0	2	N, Y	
Tollfree	<input type="checkbox"/>	0	2	N, Y	
Voice	<input type="checkbox"/>	0	2	N, Y	
Wireless	<input type="checkbox"/>	0	2	N, Y	
Churn (Y/N)	<input type="checkbox"/>	0	2	N, Y	

*Figure 2: Showing Histogram and Information about the nominal variables*

## DATA PREPARATION

Data preparation is essential for the accuracy and ease of the data analysis process. This step entails steps taken to collect, clean, and organize the data for analysis. The dataset was a real-world customer churn dataset encompassing information of 1000 customers for the model development.

The dataset did not require specific data cleaning or manipulation. There was no missing data in any of the numeric or nominal variables. The two “cust\_id” variable that resulted from the joining of the two datasets were dropped as they were only there as identifiers and did not have a relevance to the analysis. Color target was added to the target variable to ensure that the classes “Y” and “N” were distinguishable. Note that the exploratory analysis showed that there was class imbalance in the target variable as shown in **Figure 3**.



*Figure 3: showing the imbalance in the distribution of the target variable “churn”*

## MODELING

The data was first partitioned into the learning (training data) and the predictor (test data). Due to the high imbalance in the target variable which raised concerns about whether the rest of the dataset was imbalanced, the dataset was balanced using the “exact sampling” method in KNIME to ensure that the target variable was balanced. The four machine learning algorithms that were used for the analysis are illustrated below:

### Decision Tree

The decision tree classification algorithm, which predicts target labels based on the values of other characteristics in the dataset, was utilized. It divides the dataset depending on the relevance of each variable, which is then utilized to produce predictions about the target variables. **Figure 4** depicts the results of the decision tree split.

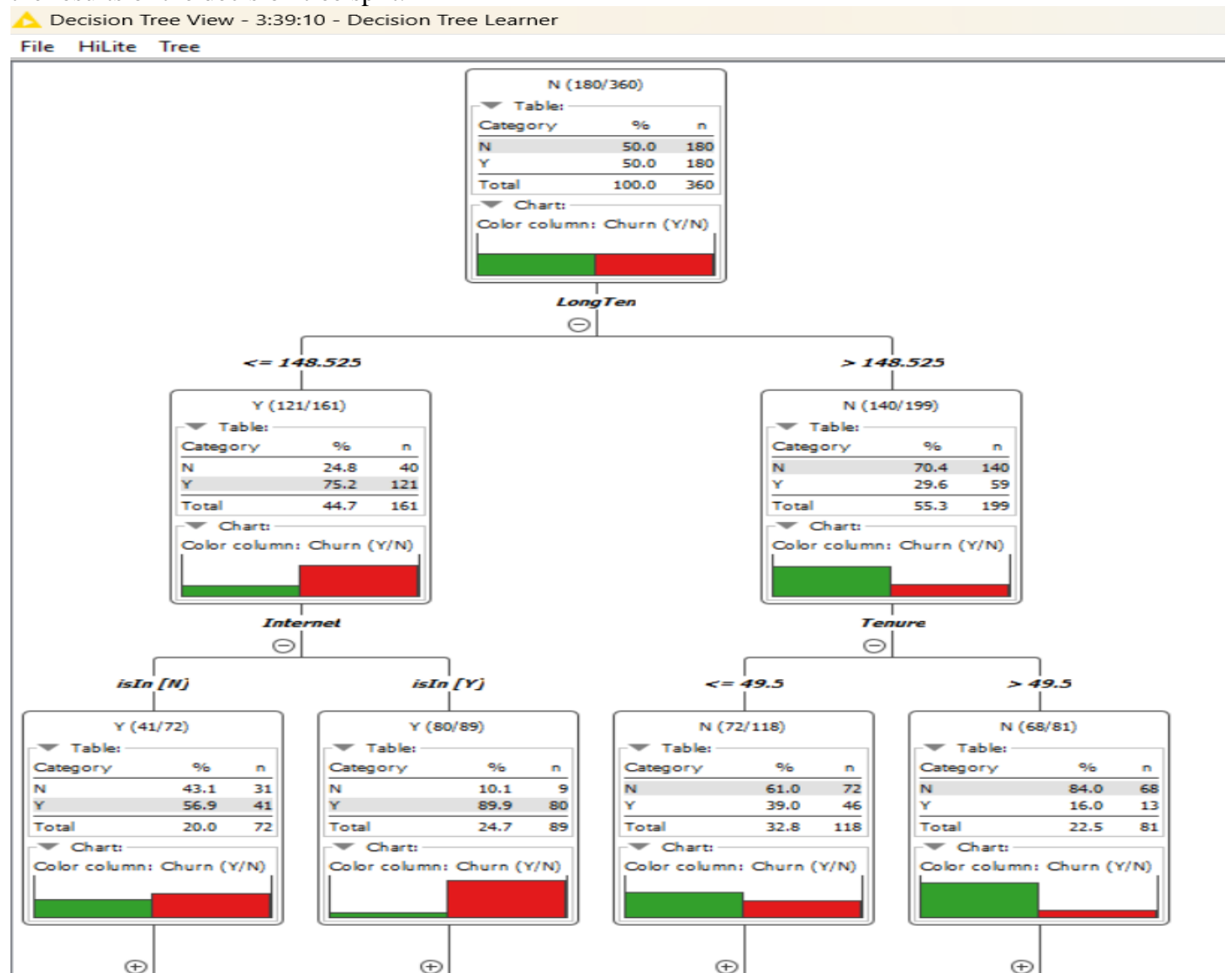
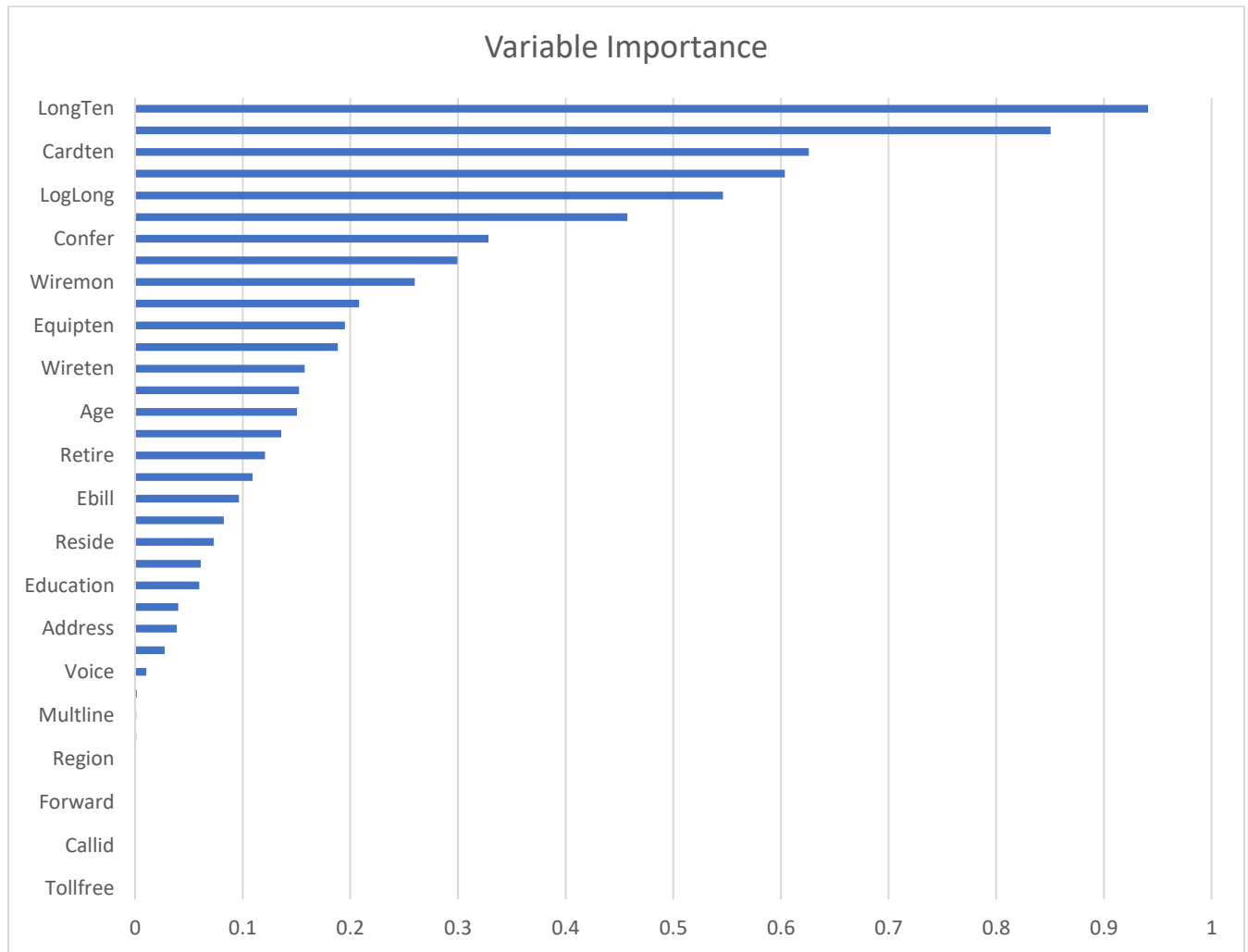


Figure 4: Showing the Decision Tree Split

The decision tree split in **Figure 4** shows that variable “Longten” is likely the most important variable in the customer churn prediction. To further confirm this suspicion, a variable of importance variable was created using Excel and the results confirms “Longten” as the most important variable followed by variables “Cardten”, “LogLong”, “Confer” and “Wiremon”, thereby making them the top 5 variables of importance as shown in **Figure 5**.



*Figure 5: Showing the Variable of Importance*

## Random Forest

The random forest which is an ensemble learning method that uses multiple decision trees to make decisions, often provides improved predictive performance that are less prone to overfitting unlike the decision tree. **Figure 6** shows the accuracy statistics table for the random forest performance.

RowID	TruePositives Number (integer) ▾	FalsePositives Number (integer) ▾	TrueNegatives Number (integer) ▾	FalseNegatives Number (integer) ▾	Recall Number (double) ▾	Precision Number (double) ▾	Sensitivity Number (double) ▾	Specificity Number (double) ▾	F-measure Number (double) ▾
Y	62	38	184	16	0.795	0.62	0.795	0.829	0.697
N	184	16	62	38	0.829	0.92	0.829	0.795	0.872
Over...	②	②	②	②	②	②	②	②	②

*Figure 6: Showing the Random Forest Accuracy Statistics Table*

## Artificial Neural Network

An artificial neural network (ANN) is made up of several interconnected processing nodes known as neurons. ANN determines the correlations between the input data and the target label and then it uses the relationships to forecast new data. Neural networks are widely used to solve complex classification issues and may achieve extraordinarily high accuracy when it is trained on large datasets. **Figure 7** depicts the accuracy statistics table for artificial neural network performance.

RowID	TruePositives Number (integer) ▾	FalsePositives Number (integer) ▾	TrueNegatives Number (integer) ▾	FalseNegatives Number (integer) ▾	Recall Number (double) ▾	Precision Number (double) ▾	Sensitivity Number (double) ▾	Specificity Number (double) ▾	F-measure Number (double) ▾
Y	64	48	174	14	0.821	0.571	0.821	0.784	0.674
N	174	14	64	48	0.784	0.926	0.784	0.821	0.849
Over...	②	②	②	②	②	②	②	②	②

*Figure 7: Showing the ANN Accuracy Statistics Table*

## Logistic Regression

Logistic regression is used to predict binary outcome, like the customer churn dataset that we have. Linear regression finds the linear decision boundary most suitable for separating the two classes. **Figure 8** shows the accuracy statistics table for the logistic regression performance.

RowID	TruePositives Number (integer) ▾	FalsePositives Number (integer) ▾	TrueNegatives Number (integer) ▾	FalseNegatives Number (integer) ▾	Recall Number (double) ▾	Precision Number (double) ▾	Sensitivity Number (double) ▾	Specificity Number (double) ▾	F-measure Number (double) ▾
Y	49	44	178	29	0.628	0.527	0.628	0.802	0.573
N	178	29	49	44	0.802	0.86	0.802	0.628	0.83
Over...	②	②	②	②	②	②	②	②	②

*Figure 8: Showing the Logistic Regression Accuracy Statistics Table*

## EVALUATION

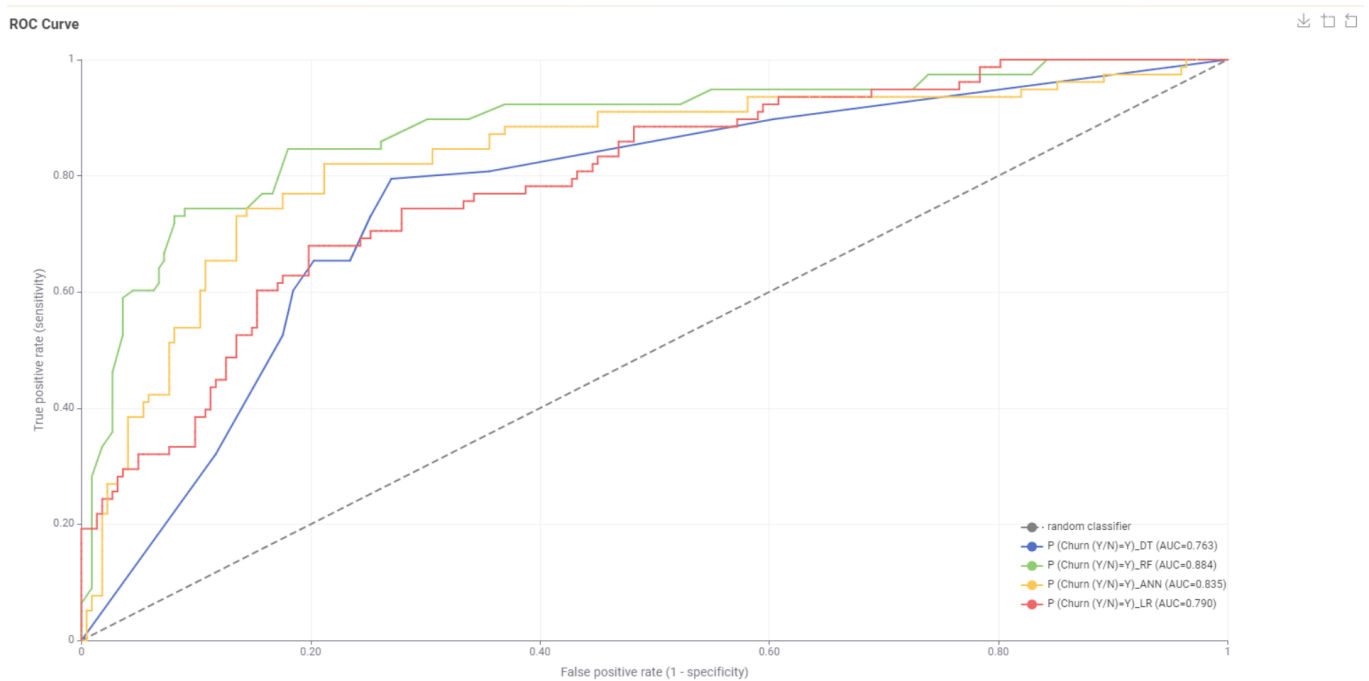
The chosen models' performance was evaluated using the following metrics: sensitivity, specificity, accuracy, and the area under the receiver operating characteristic curve (AUC-ROC).



Accuracy evaluates the fraction of properly identified cases, whereas AUC-ROC assesses the trade-off between sensitivity and specificity at various classification levels. **Table 1** displays the assessment findings for each model, whereas **Figure 9** depicts the ROC curve for all models. The random forest had the highest AUC-ROC value of 0.884, that is 88.4% while the Decision tree had the lowest AUC-ROC value of 0.763 that is 76.3%. Based on the accuracy measure the random forest had the highest accuracy score of 0.82 while the decision tree had the lowest accuracy score of 0.743. The AUC-ROC and Accuracy metric both gave the same performance evaluation results for the models. **Figure 10** below shows the final workflow for the analytical steps taken.

MODEL	SENSITIVITY	SPECIFICITY	ACCURACY	AUC-ROC
Decision Tree	0.731	0.748	0.743	0.763
Logistic Regression	0.628	0.802	0.757	0.790
Neural Network	0.821	0.784	0.793	0.835
Random Forest	0.795	0.829	0.82	0.884

*Table 1: Showing Model Performance Comparison*



*Figure 9: Showing the ROC curve for all the models.*

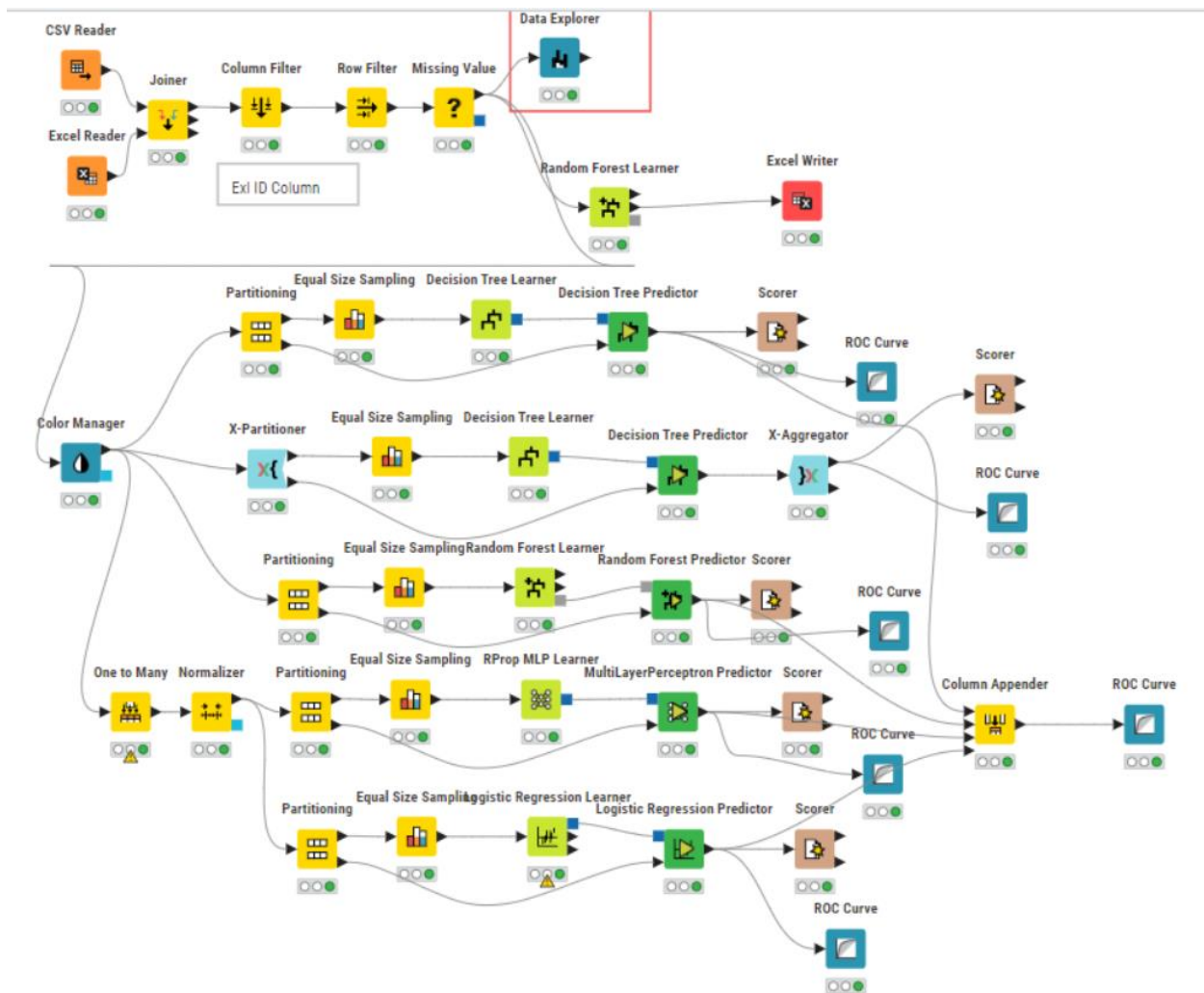


Figure 10: Showing the KNIME Workflow.

## DEPLOYMENT

According to the accuracy and the AUC\_ROC values, the random forest was the model best suited for predicting customer churn. The logistic regression model and the neural network model also had good performance scores with varying performance based on the specificity and sensitivity. The analysis provided insights into the performance of these models as they were applied to real-time dataset. The variable of importance showed that variables “Longten”, “Cardten”, “LogLong”, “Confer” and “Wiremon” were the top five most important variables in predicting customer churn rate. Companies should investigate these factors to improve customer retainment and ultimately increase customer size.