Customer Churn Prediction Exercise Using KNIME Data Analytics

by Devang Maniar

What is KNIME?

- KNIME is an open-source platform for data analytics, reporting, and integration. It allows users to manipulate, analyze, and visualize data using a visual programming approach, making it particularly useful for data blending, data preparation, and advanced analytics.
- KNIME is a lightweight version in the Market
- Other Open Source Platforms are: Altericks, DataRobot, DataIQ, Knime

What is KNIME Data Lab?

- "KNIME Data Lab" may refer to a specific module, extension, or functionality within the KNIME platform that is designed for conducting data analytics related to customer behavior and customer-centric insights.

What is TRAINING Dataset?

- Training data is the data you use to TRAIN a machine learning algorithm or model to ACCURATELY PREDICT a particular outcome, or answer, that you want your model to predict.

What is TEST Dataset?

- Test data is used to measure the performance, such as accuracy or efficiency, of the algorithm you are using to train the machine.
- Test data will help you see how well your model can predict new answers, based on its training.
- Both Training and Test data are important for improving and validating machine learning models.

What are we going to do in this Exercise?

- Present a real case study with some real data
- Use an analytics platform (KNIME)
- Build a Segmentation Model to understand customer segments
- Generate a Predictive Model based on customer Data
- Analyze the Data
- Gather Customer Insights from the Prediction – Understand Customer Segments, Identify why Customers are likely to churn or stay, and how do we improve the lifetime value of these customers

Overview

In this exercise, we will use the KNIME platform to understand steps in creating a machine learning model for churn prediction.

Why KNIME?

- Desktop version is available free of charge (Open Source)
- Modular platform for building and executing workflows using predefined components called nodes
- Out-of-the-box functionality for tasks such as standard data mining, data analysis and data manipulation
- Extra features and functionality available through KNIME extensions

Installing and working with KNIME:

The following resources will help you get started with KNIME – from installing the platform and few extensions, to building familiarity with the interface and workflow tools. You can follow these links or use the more guided approach (KNIME Primer) that is included in the Resource Library.

Install KNIME	https://www.knime.com/knime-introductory-course/chapter1/section1/installation-guide
Install KNIME Extensions	https://www.knime.com/knime-introductory-course/chapter1/section1/install-knime-extensions
Get to Know the KNIME Workbench	https://www.knime.com/knime-introductory-course/chapter1/section1/knime-workbench
Create a KNIME Workflow	https://www.knime.com/knime-introductory-course/chapter1/section2/workflow-coach

Other Resources

Additionally, since KNIME is a very popular open source platform, its community of users have contributed a rich set of resources including tools, ML libraries and training material. Many of these are available through the links below.

KNIME Learning Hub	https://www.knime.com/learning-hub
KNIME TV Channel	https://www.youtube.com/user/KNIMETV
KNIME Cheat Sheets	https://www.knime.com/learning/cheatsheets

Nodes are basic processing units of KNIME workflows. The chart below shows what various parts of a node represent.

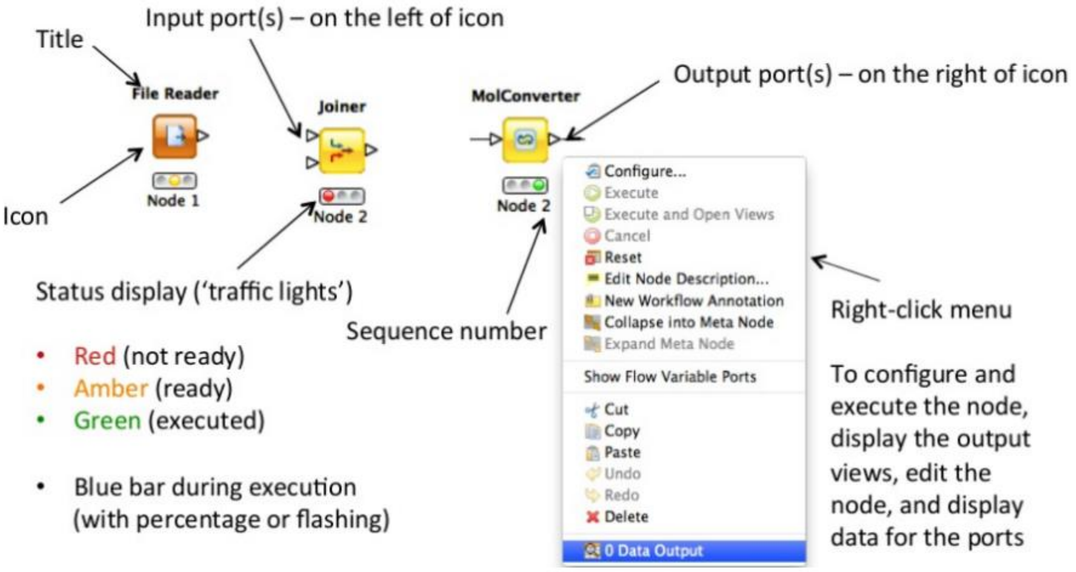


Figure 2.1 Overview of Nodes

Business Use Case:

Customer churn is a critical business metric and risk for many companies, ranging from SaaS providers to telecom operators. In this assignment, we will import customer **contract information** and **call information** to predict churn. Understanding the drivers of churn gives teams valuable insights, so they can take measures such as targeted offers and proactive issue resolution to prevent customers from churning.

Model Overview:

At a high-level the model shown below executes the following steps –

- Read contract and calls data (the dataset is already attached within the two "reader" nodes)
- Perform basic data manipulations / pre-processing
- Partition the dataset into 80% training data and 20% test data
- Apply a decision tree model to generate churn predictions
- Evaluate model performance and calculate model accuracy

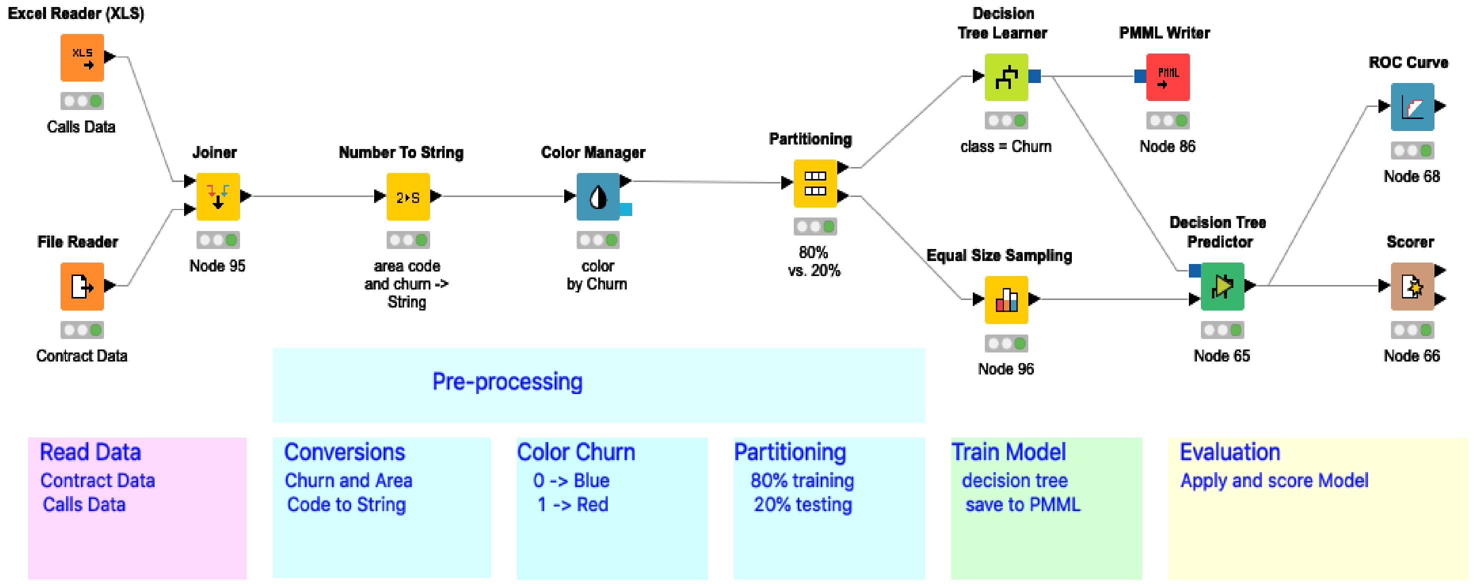


Figure 2.2 Churn Prediction Model in KNIME

Glossary of nodes and their functions:

Excel Reader (XLS)	Reads a spreadsheet and provides it at its output port.
File Reader	Reads data from an ASCII file or URL location. It can be configured to read various formats.
Joiner	Joins two tables in a database-like way. The join is based on the joining columns of both tables.
Number to String	Converts numbers in a column (or a set of columns) to strings.
Color Manager	Assigns colors for either nominal values or numeric columns – churned customers with value 1 colored red, others with value 0 colored blue.
Partitioning	Splits the input table into two partitions (i.e. row-wise), e.g. train and test data. The two partitions are available at the two output ports.
Decision Tree Learner	This node implements a classification decision tree. The target attribute must be nominal (e.g., YES or NO). The other attributes can be either nominal or numerical. Numeric splits are always binary (two outcomes), dividing the domain in two partitions at a given split point. Nominal splits can be either binary (two outcomes) or they can have as many outcomes as nominal values. In the case of a binary split the nominal values are divided into two subsets.
Equal Size Sampling	Removes rows from the input data set such that the values in a categorical column are equally distributed. This can be useful, for instance if a learning algorithm is prone to unequal class distributions and you want to downsize the data set so that the class attributes occur equally often in the data set. The node will remove random rows belonging to the majority classes. The rows returned by this node will contain all records from the minority class(es) and a random sample from each of the majority classes, whereby each sample contains as many objects as the minority class contains.
PMML Writer	This nodes writes a PMML (Predictive Model Markup Language) model from a PMML model port into a PMML v4.0 compliant file or to a remote location denoted by an URL.
Decision Tree Predictor	This node uses an existing decision tree (passed in through the model port) to predict the class value for new patterns.
ROC (Receiver Operating Characteristics) Curve	This node uses an existing decision tree (passed in through the model port) to predict the class value for new patterns. As a rule of thumb, the greater the area under the curve, the better the model. The light gray diagonal line in the diagram is the random line which is the worst possible performance a model can achieve.
Scorer	Compares two columns by their attribute value pairs and shows the confusion matrix, i.e. how many rows of which attribute and their classification match. Additionally, the second out-port reports a number of accuracy statistics such as True-Positives, False-Positives, True-Negatives, False-Negatives, Recall, Precision, Sensitivity, Specificity, etc.

Scenario:

- You are a telecom company, and the typical data set that you have about your customers is,
  - you have account information,
  - you know what contract they're on,
  - what plan they have, and what,
  - what their bills have looked like,
  - and how long they've been a customer.
- You have some demographic information, their gender, their age, their location, and so on.
- You might also know, what value-added services they have bought from you, how many lines do they have, how many minutes and data that they've used, and what did their internet usage look like, and also whether or not they have churn, alright? Which is that have, are they still a customer, or have they lapsed, okay?

<u>**Exercise A:**</u>

1. Typical information available about telecom subscribers includes their demographics, behavioral data and spending history. At the time of renewing contracts, some customers renew while others do not: they churn. It would be extremely useful to know in advance which customers are at risk of churning so that the company can proactively reach out to at-risk customers, particularly for high-value customers.

This model relied on two data types for churn prediction – contract data (account length, state, plan type) and call-based behavioral data (minutes used, monthly charges, international roaming charges).

What other data types would you add to this model to increase its predictive power?
Hint: Think about quality metrics. Think about metrics related to data usage (not just voice calls).

Response:

Other Types of Data Usage:

1) *Data Usage via Text Messages & Multimedia Messages*

2) *Data Usage via Internet Browsing Data (Download/Upload Data)*

3) *Data usage on Video Calls, WhatsApp, Facetime, Skype, etc.*

4) *Data usage associated with Social Media Activity like Facebook, Twitter, Instagram, etc.*

5) *Data usage by mobile apps running in the background*

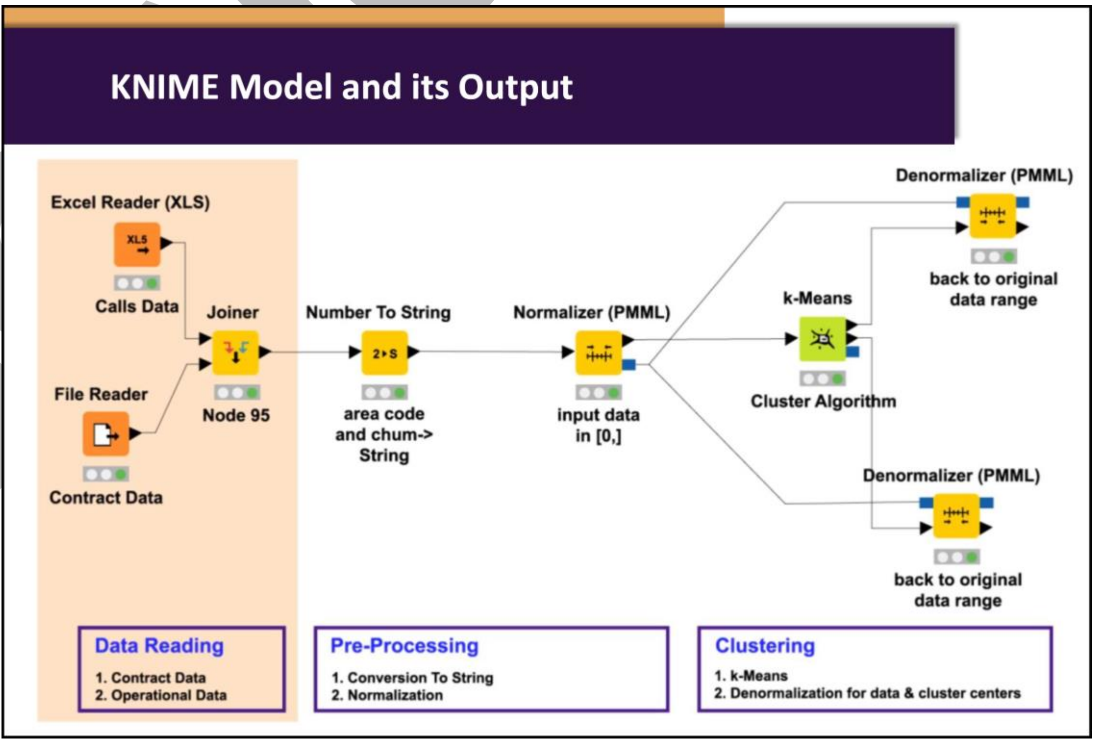
6) *Data usage when you use your mobile phone as a Hotspot*

7) *Data usage related to Streaming Videos or Music from apps like YoutTube, Netflix, Spotify, etc.*

Exercise B:

2. Match the following nodes to their function as it relates to the described KNIME model – reference the node glossary table and Figure 2.2 Churn Prediction Model to answer these questions. For those who have downloaded and setup KNIME, you can also see the help section to learn more about each of these nodes.

TRUE	FALSE	NODE APPLICATIONS IN CHURN PREDICTION MODEL
	False	The Decision Tree Predictor node is used to build a decision tree model, while the Decision Tree Learner node is used to test the model.
True		The joiner node merges two data tables based on common values in one or more columns
	False	The Color Manager node splits datasets into two colors based on whether the data belongs to training or test data
True		The ROC Curve node generates an accuracy curve for the model along with a confusion matrix that tells the models accuracy level
True		The PMML node makes it easy to move this model from one application to another and deploy it in real-time



KNIME Model and its Output

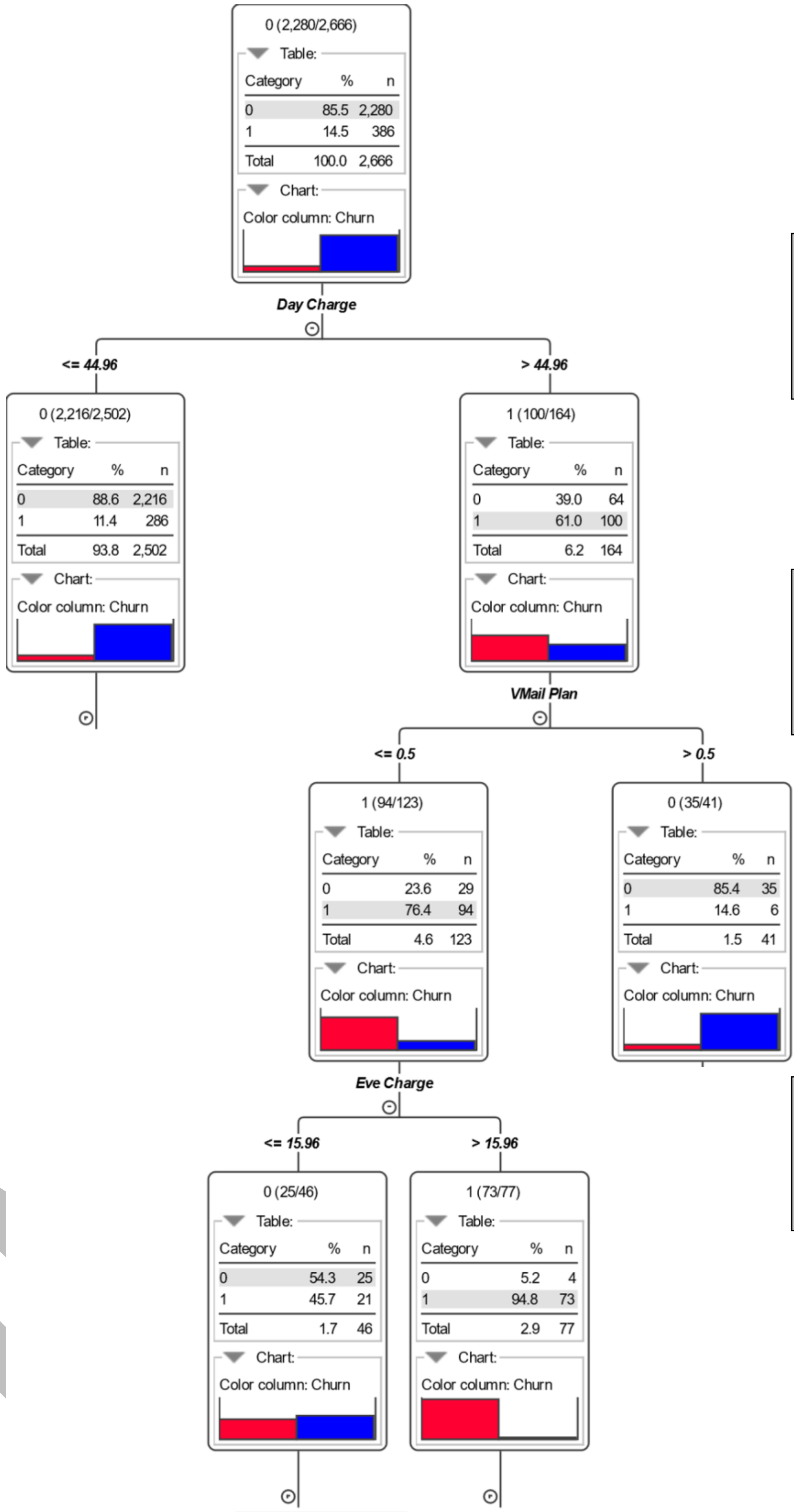
Data Reading
1. Contract Data
2. Operational Data

Pre-Processing
1. Conversion To String
2. Normalization

Clustering
1. k-Means
2. Denormalization for data & cluster centers

What is Churn Prediction?

- This churn prediction model was built using a decision tree algorithm. The decision tree output is shown below. Three key variables are described alongside the branches of the tree. Refer to this model output to answer the next question.



0 (2,280/2,666)

Table:

Category	%	n
0	85.5	2,280
1	14.5	386
Total	100.0	2,666

Chart:
Color column: Churn

0 (2,216/2,502)

Table:

Category	%	n
0	88.6	2,216
1	11.4	286
Total	93.8	2,502

Chart:
Color column: Churn

1 (100/164)

Table:

Category	%	n
0	39.0	64
1	61.0	100
Total	6.2	164

Chart:
Color column: Churn

1 (94/123)

Table:

Category	%	n
0	23.6	29
1	76.4	94
Total	4.6	123

Chart:
Color column: Churn

0 (35/41)

Table:

Category	%	n
0	85.4	35
1	14.6	6
Total	1.5	41

Chart:
Color column: Churn

0 (25/46)

Table:

Category	%	n
0	54.3	25
1	45.7	21
Total	1.7	46

Chart:
Color column: Churn

1 (73/77)

Table:

Category	%	n
0	5.2	4
1	94.8	73
Total	2.9	77

Chart:
Color column: Churn

The Day Charge variable represents Daytime charges – customers incur additional costs when they exceed their allotted daytime minutes

The VMail plan variable indicates if customers have a voicemail plan (value > 0.5) or do not have a voicemail plan (value < 0.5)

The Eve Charge variables refers to Evening Charges – customers incur additional costs when they exceed their allotted evening minutes

Figure 2.3 Decision Tree Output

Exercise C:

3. This model used the decision tree algorithm to predict churn. Figure 2.3 Decision Tree Output shows the output of the decision tree and describes the key variables at each branch.

Provide an interpretation for this decision tree model articulating:

a) What is the churn rate at each level of the tree?
b) How do the data variables impact churn rates?
c) What can a business executive do with this information?

My Response:

Churn Customer = 1 (Red Color)

a) Below are the Churn Rates for each level:

1. At the Top, the overall Churn Rate is 14.5% for Customers (Total People = 2,666)
2. Next Level, if daytime charge < 45, then Churn Rate is 11.4%. If daytime charge > 45, then Customer Churn Rate is 61%
3. Next Level below that, if daytime charge > 45 and voicemail plan <= 0.5, then Customer Churn Rate is 76.4%. However, if voicemail plan > 0.5 then Customer Churn Rate is 14.6%
4. Next level below that, if daytime charge > 45 and voicemail plan <= 0.5 and if evening charge <=16 then customer Churn Rate is 45.7%. However, if evening charge > 16 then customer Churn Rate is 94.8%

b) If customers are incurring higher daytime charges (>$45) when they exceed their minutes, it seems that they are dropping out from the service because they may not be liking the extra charges on their bill. Furthermore, if customers do carry a voicemail plan, then they are less likely to churn or drop out of service as oppose to customers who do not carry a voicemail plan. So, the rate of customers dropping out of service is higher when they do not carry a voicemail plan. Lastly, if customers are incurring higher evening charges (>$16), then they are even more frustrated and churning because not only are they incurring higher daytime charges but now they are also incurring higher evening charges.

c) Business can re-evaluate their subscriber plans, and offer more flexible plan options upfront to customers so they can pick a plan that truly fits to their needs without being incurred with additional $charges. Also, it seems the highest churn rate is when customers are exceeding their evening minutes. Perhaps, business can include additional free evening minutes, so less customers will get incurred with those evening charges which in turn can hopefully retain those frustrated customers. Another strategy is that business can send warning alerts to customers right before they are violating their threshold minutes, so customers are pre-warned and will likely alleviate frustration.

Reference: *AI Applications for Growth Business Executive Course, Professor Mohanbir Sawhney, Northwestern Kellogg Business School of Management*