HW4 - KNIME Data Mining I

Moises Marin

A20349918

mmarinm@okstate.edu

3/3/2022

—

MSIS 5633

Predictive Analytics Technologies

1. Business Understanding

Data are related to customer churn analysis, this type of analysis is used to explore customers that are likely to change of service provider. Such analysis can help in identifying customers in risk of leaving and take actions to revert the scenario

2. Data Understanding

Data analyzed come from 1000 customers, it has the following variables (the target variable is marked in bold.)

region	wiremon
tenure	longten
age	tollten
marital	equipten
address	cardten
income	wireten
ed	multline
employ	voice
retire	pager
gender	internet
reside	callid
tollfree	callwait
equip	forward
callcard	confer
wireless	ebill
longmon	loglong
tollmon	lninc
equipmon	custcat
cardmon	**churn**

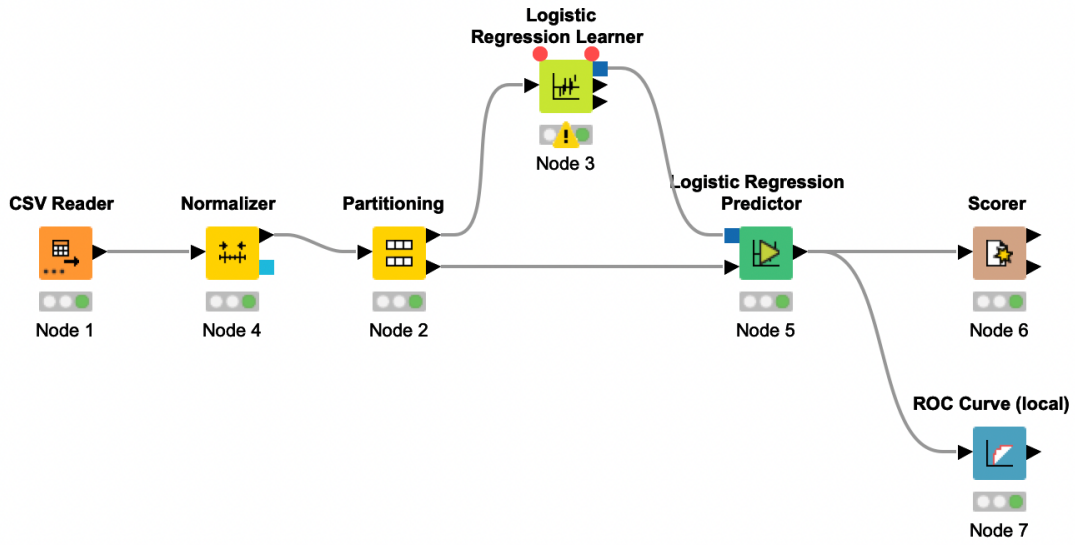
Variable cust_id is excluded from analysis because it is an id of customer.
There are no missing values in data set

3. Data Preparation

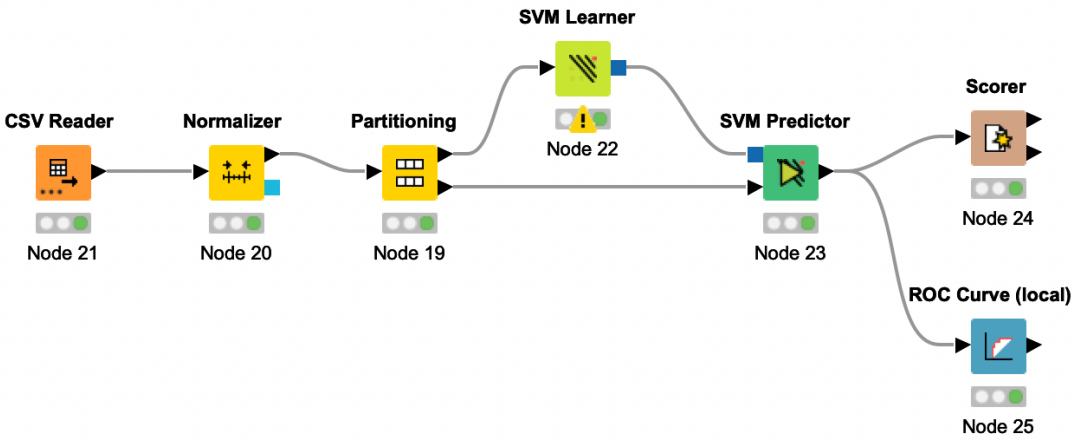
Knime nodes Normalizer and Partitioning were used to prepare the data for model building.

4. Model Building

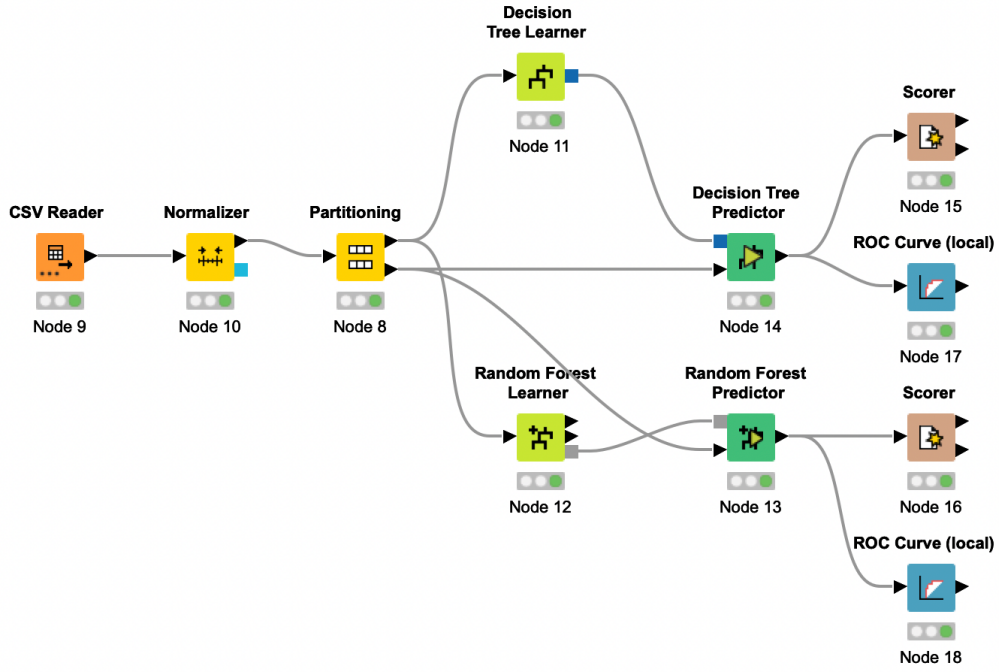
Model 1 Logistic Regression



Model 2 State Vector Machine



Models 3 and 4 Decision Tree and Random Forest

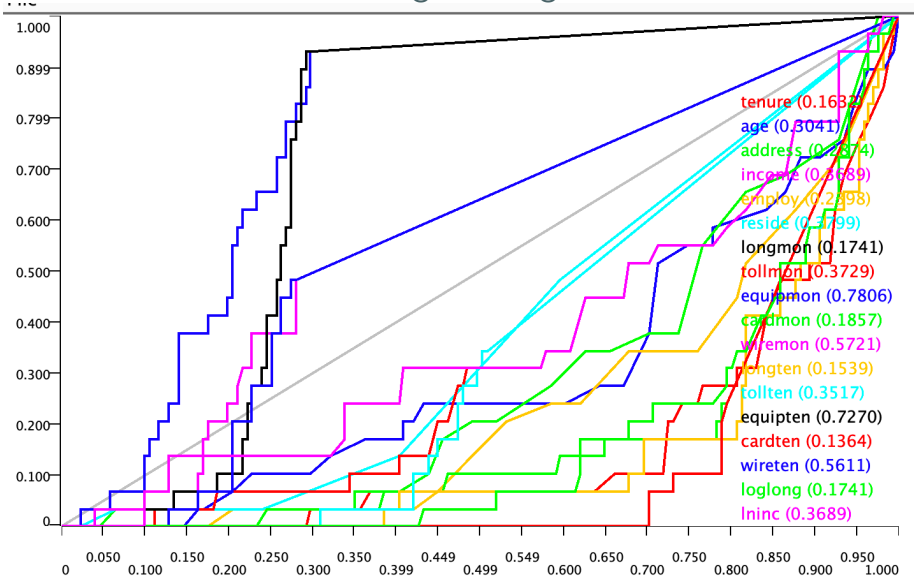


5. Testing and Evaluation

Confusion Matrix of Model 1 Logistic Regression

Row ID	Y	N
Y	19	28
N	10	143

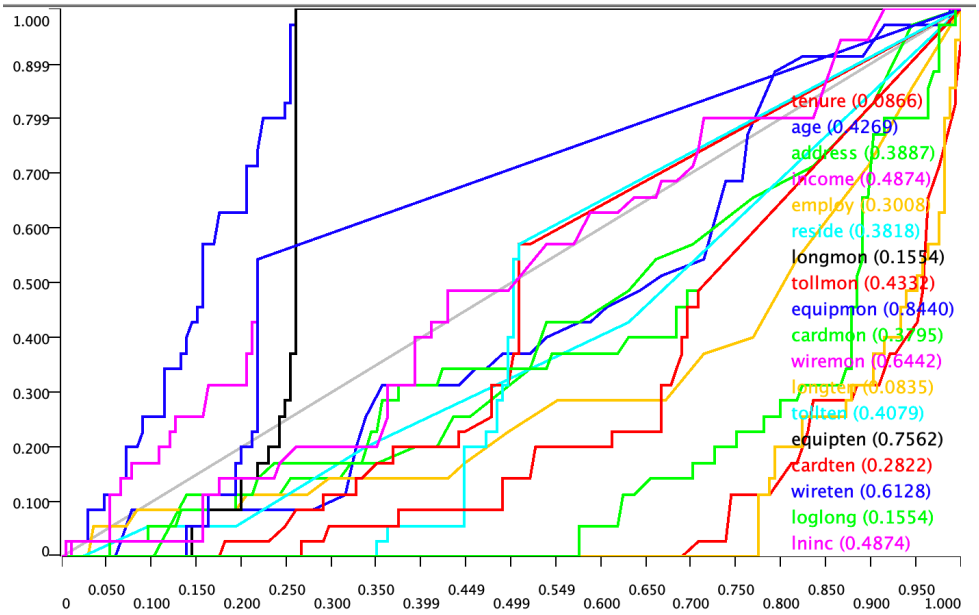
ROC Curves of Model 1 Logistic Regression



Confusion Matrix of Model 2 State Vector Machine

Row ID	Y	N
Y	16	31
N	19	134

ROC Curves of Model 2 State Vector Machine

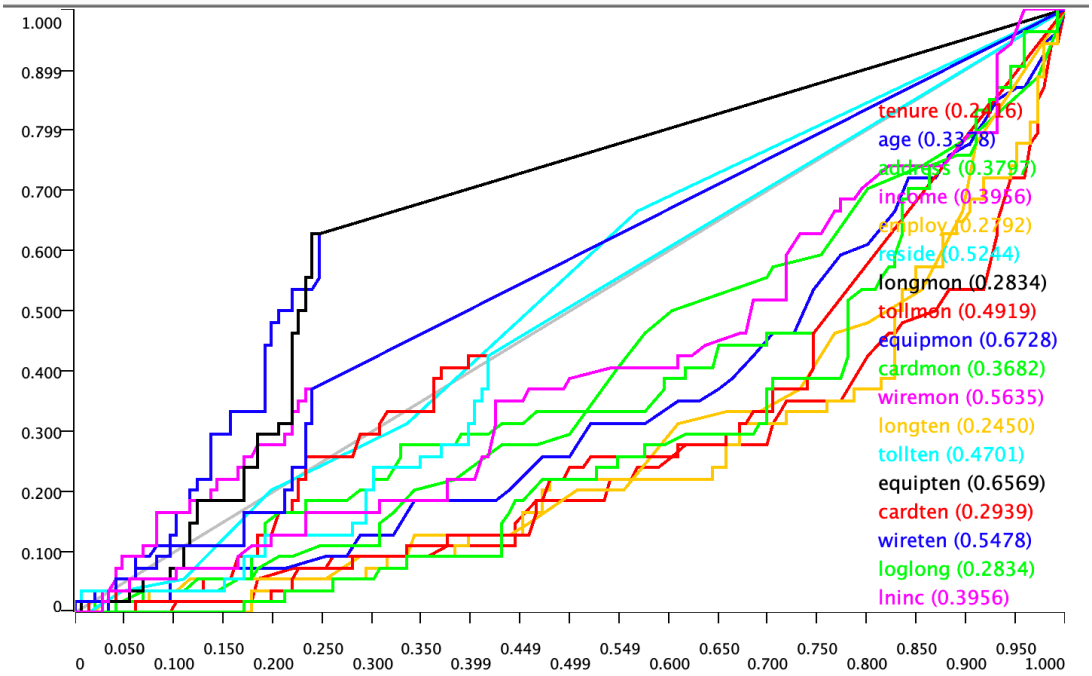


Confusion Matrix of Model 3 Decision Tree

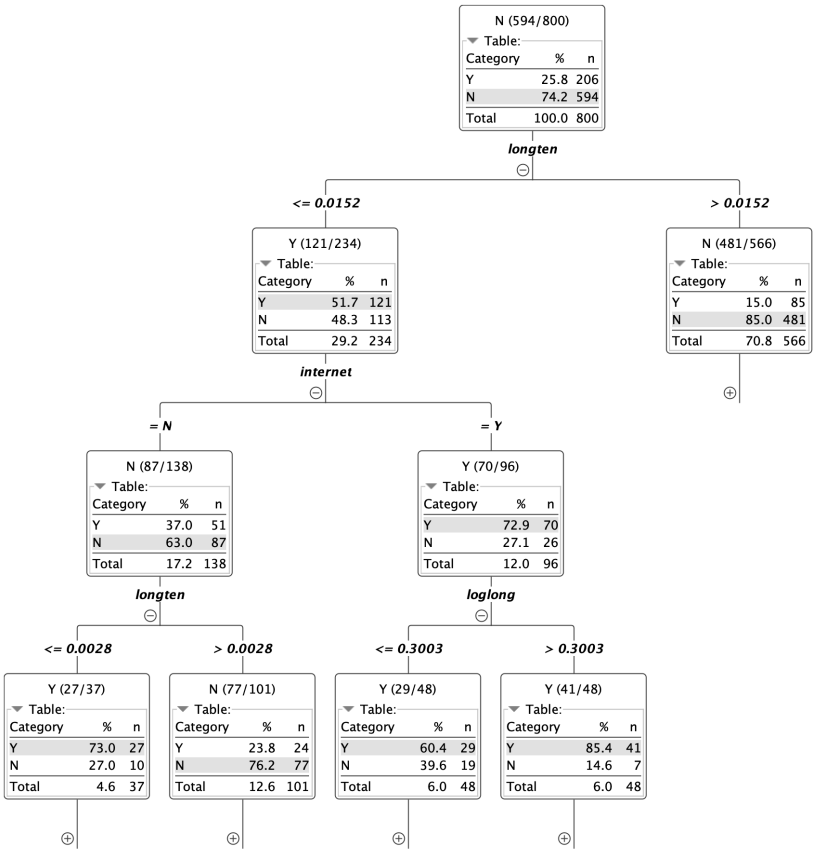
churn \ Prediction (churn)	Y	N
Y	52	0
N	2	146

Correct classified: 198	Wrong classified: 2
Accuracy: 99%	Error: 1%
Cohen's kappa (κ): 0.974%	

ROC Curves of Model 3 Decision Tree



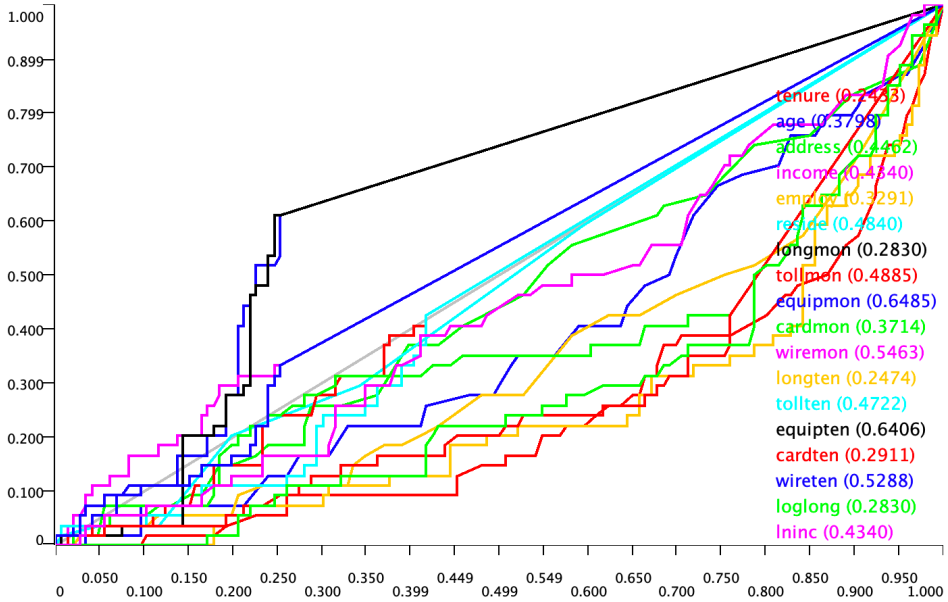
First 3 levels of Decision Tree



Confusion Matrix of Model 4 Random Forest

Row ID	Y	N
Y	52	0
N	2	146

ROC Curves of Model 4 Random Forest



Accuracy, Sensitivity and Specificity Table

Model	Accuracy	Sensitivity	Specificity
Logistic Regression	0.81	404	0.935
State Vector Machines	0.75	0.34	0.876
Decision Tree	0.9	0.827	0.926
Random Forest	0.99	1	0.986

Based on previous table the best model is Randeom Forest with 0.99 accuracy. The perfect score of sensitivity makes me think that this model may be too specific for this data set and may not be good for new data.

6.Deployment

Models were created in Knime and can be executed in this tool.