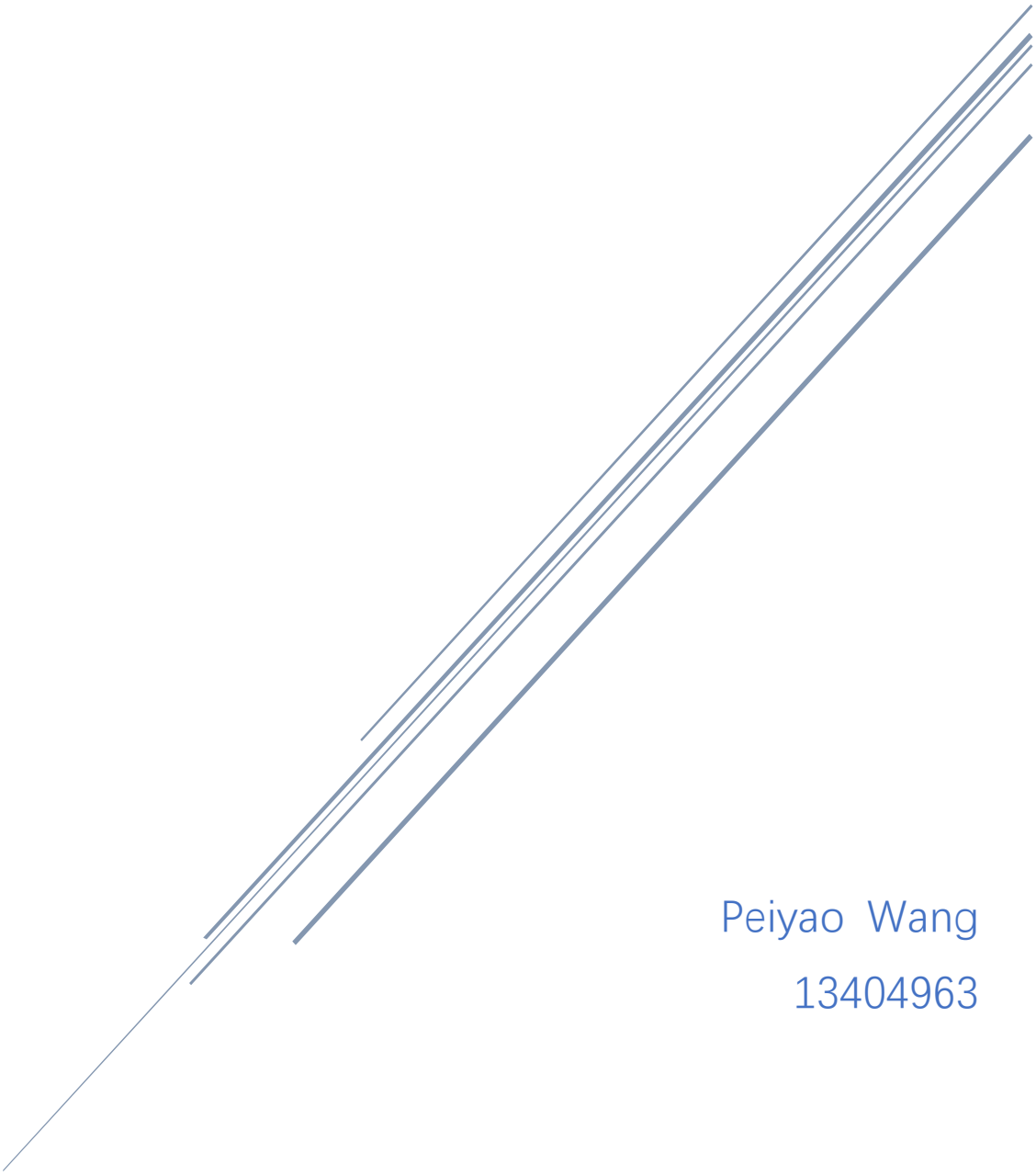
DATA MINING IN ACTION

32130 Fundamental of Data Analytics – Assignment 3



Peiyao Wang

13404963

Contents

Introduction

In this report we will model the data based on the data in the training data set provided in the material and use the model to analyze the data in the test data set and predict the "Final_Y" data. The main method used in this data analysis is K-nearest neighbor, and the decision tree is used as a comparison.

The method used in this study is CRISP-DM, which is called CRoss-Industry Standard Process for Data Mining. It contains six stages, namely business understanding, data understanding, data preparation, modeling, evaluation, deployment, except for the deployment stage, other stages will appear in this report.

Business understanding

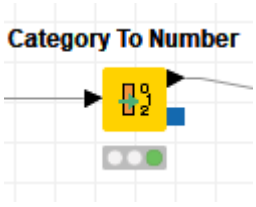
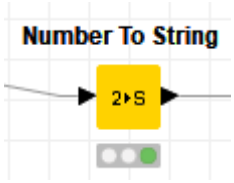
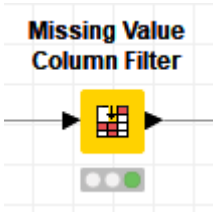
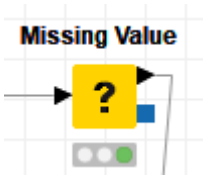
In this project, as a data scientist in a consulting company, I need to help customers. The clients of this project hope to provide them with commercial assistance through our analysis. They need to know who will order their time deposits and what attributes these people have.

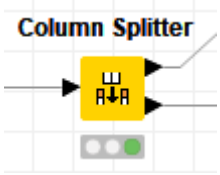
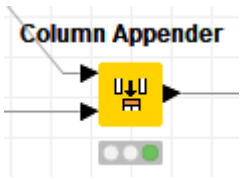
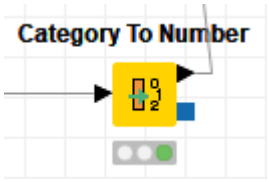
Data understanding

This project involves two data sets, one is the training data set and the other is the test data set. The data in the dataset has unknown data, and the reliability of the data is yet to be verified. There is no "Final_Y" data in the test dataset, which is our target data. In summary, the data needs to be further pre-processed before it is officially started and modeled.

Data preparation

In order to conduct more effective analysis and mining of data, the first thing to do before the formal analysis and mining of data is to preprocess the original data. General data preprocessing methods include data cleaning, data integration, and data changes.
In the source data of this project, it is necessary to quantize the data of the String type other than "Final_Y" into integer type data and ensure that the data type of "Final_Y" is String. In addition, the Missing value in the source data is also removed to ensure the smooth analysis of the data. The following is the data processing process.

Training Data		
Steps	**Nodes**	**Description**
1		Import training data into KNIME.
2		Convert the string data to number data.
3		Convert the type of “Final_Y” to string.
4		Remove the column with 90% value are missing.
5		Assign values to missing data by calculation. Replace missing values with mean values for integers and doubles. Replace the missing value with the most frequency value for the string.

Test Data		
Steps	**Nodes**	**Description**
1		Import training data into KNIME.
2		The table is divided into two parts, one for "Final_Y" and the other for the rest.
3		For parts other than "Final_Y", assign values to missing data by calculation. Replace missing values with mean values for integers and doubles. Replace the missing value with the most frequency value for the string.
4		Combine the two tables into one.
5		Converts data which data type is string to number in addition to "Final_Y".

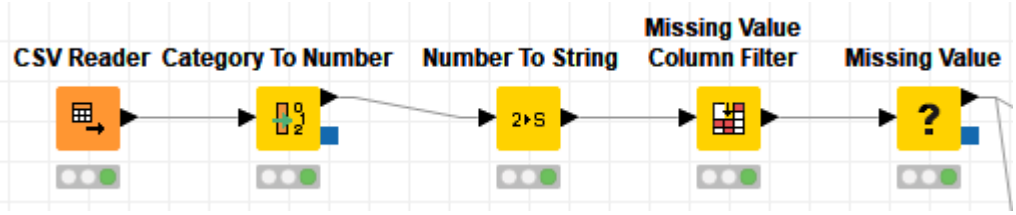
After preprocessing, the data types are as follows:

Attribute	Data type (before)	Data type (after)
age	Integer	Double
job	String	Double
marital	String	Double
education	String	Double
default	String	Double
housing	String	Double
loan	String	Double
contact	String	Double
month	String	Double
day_of_week	String	Double
duration	Integer	Double

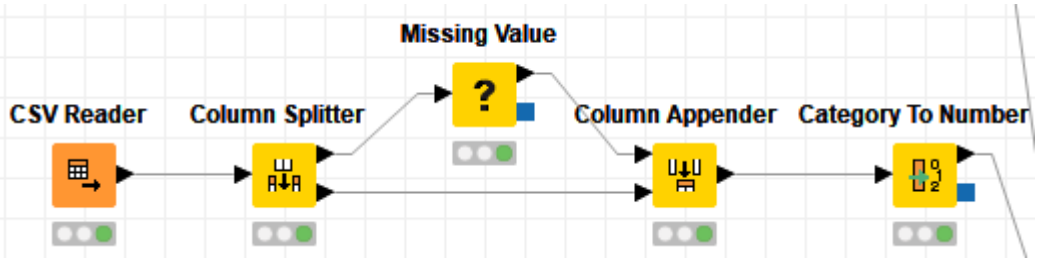
campaign	Integer	Double
pdays	Integer	Double
previous	Integer	Double
poutcome	String	Double
emp.var.rate	Double	Double
cons.price.idx	Double	Double
cons.conf.idx	Double	Double
euribor3m	Double	Double
nr.employed	Double	Double
Final_Y	Integer	String

The workflows are:

Training Data:



Testing Data:



Modeling

K Nearest Neighbor

Description

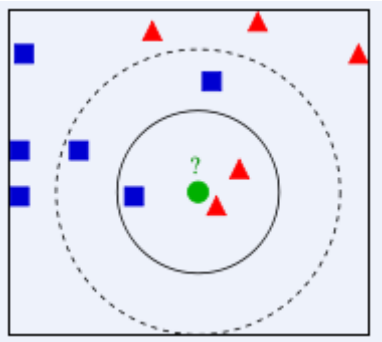


Figure 1

KNN, also known as K-Nearest Neighbor, is a very simple and basic classification algorithm used to predict an unknown value. When the KNN performs prediction, as shown in Figure 1, the known value around the predicted target is selected, and the known value with the highest repetition rate is assigned to the unknown value. The K in the KNN is the number of known values around the predicted target. When the K value is too small, the known value for analysis is too small, resulting in inaccurate results. When the K value is too large, the range of known values used for analysis is too large, and too many interference factors may result in inaccurate results. Therefore, the value of K is very critical in the KNN algorithm.

Modeling

The modeling tool used this time is KNIME, and the workflow is shown in Figure 2.

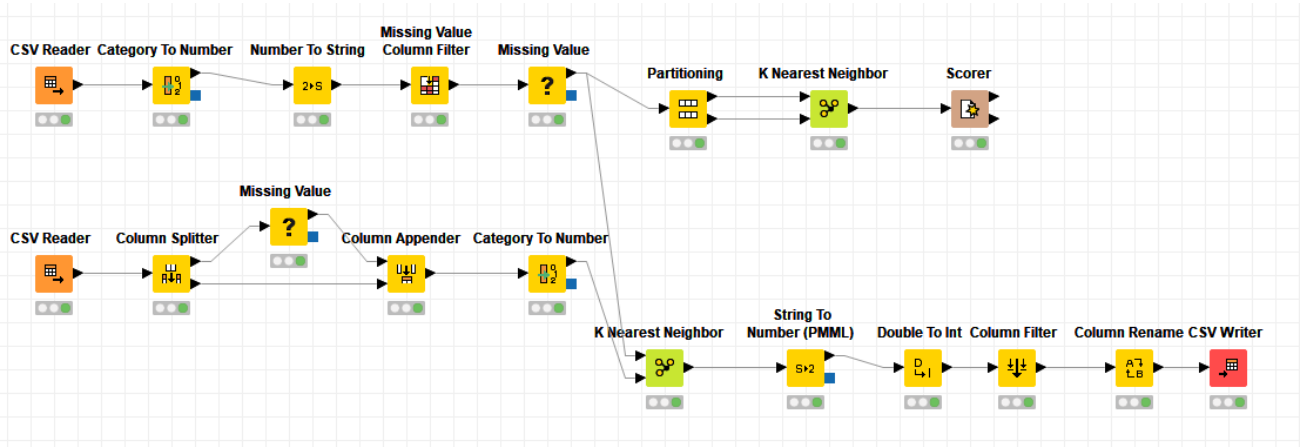
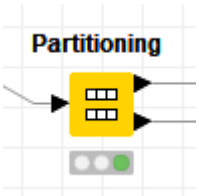
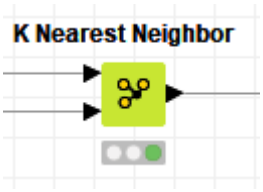
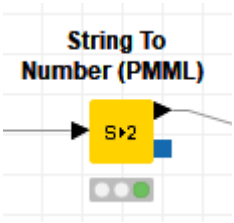
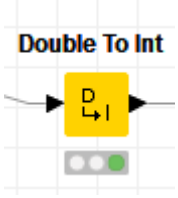
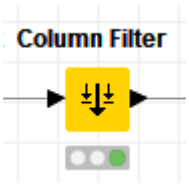
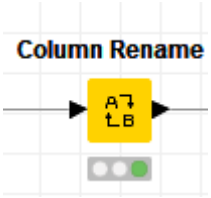


Figure 2

Nodes	Description
	The Training data is divided into two parts, one for generating the model and the other for detecting the correct rate.
	The KNN algorithm is used to model and predict the value of Final_Y.
	The correct rate is scored for the model generated by KNN.
	The type of Final_Y predicted by the KNN model is changed from a string to a number.
	Change the type of Final_Y from double to integer.

	Filter the columns, keep Final_Y and remove the others.
	Name the Final_Y column that has been predicted by KNN as "Final_Y".
	Write the final prediction to a new CSV file.

Decision Tree

Description

A decision tree is a predictive model in machine learning in which each node of the genus represents an object, and each leaf node represents a value, and the branching path represents the possibility. The decision tree is very easy to understand because he can often directly reflect the characteristics of the data.

Modeling

The modeling tool used this time is KNIME, and the workflow is shown in Figure 3.

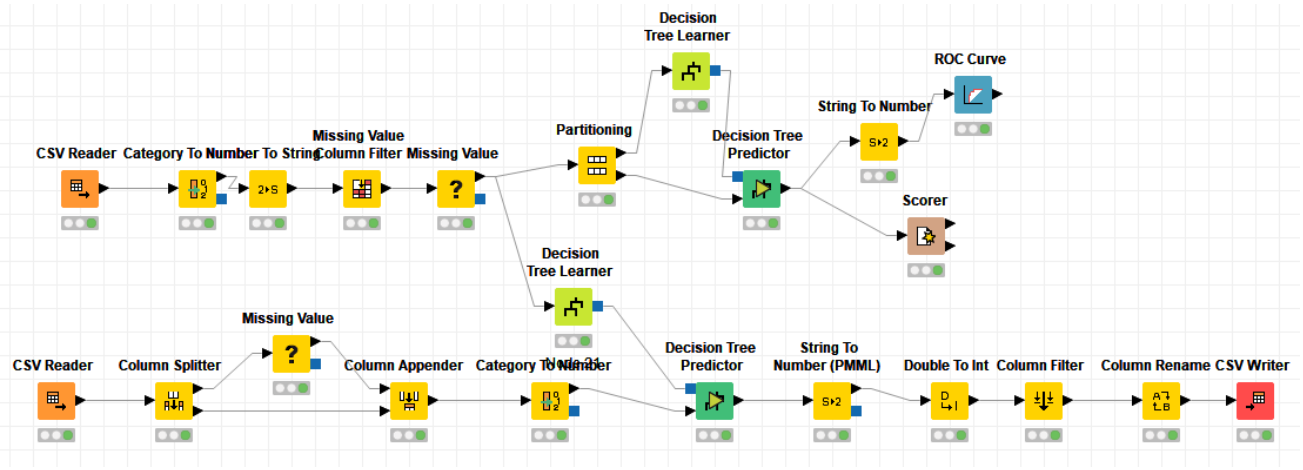
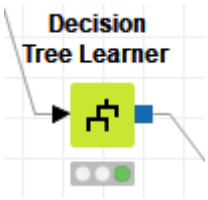
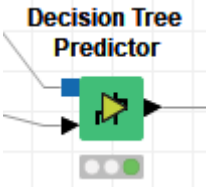
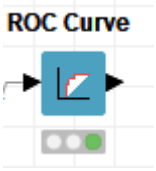


Figure 3

Nodes	Description
	Use the training data and the Decision tree algorithm to build the model.
	Use the established Decision tree model to predict the value of Final_Y in the Test data.
	Use ROC Curve to evaluate the accuracy of the model.

Random forest

Description

The random forest is the advanced of the decision tree. It is not just a collection of multiple decision trees. It also evaluates the decision tree internally.

Modeling

The modeling tool used this time is KNIME, and the workflow is shown in Figure 4.

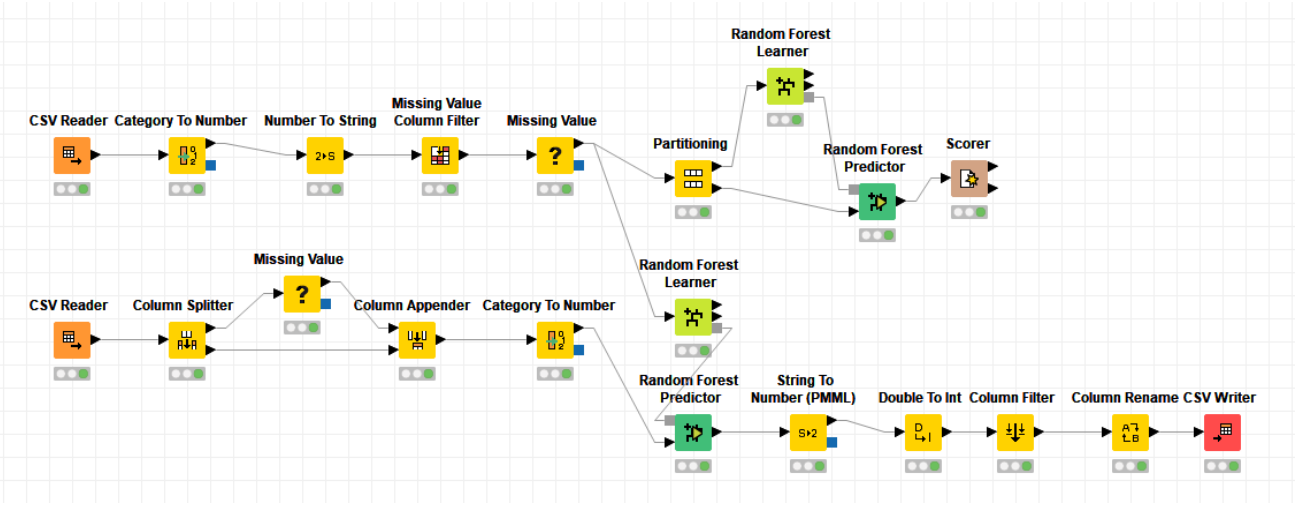
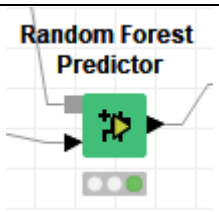


Figure 4

Nodes	Description
	Use the training data and the Random Forest algorithm to build the model.
	Use the established Random Forest model to predict the value of Final_Y in the Test data.

Evaluation

K Nearest Neighbor

The key parameter of the KNN algorithm is the value of K. Then, when the value of K is set to 3, 10, 20, 30, 40, and 50, the evaluation result by the Scorer node is shown in Fig. 5, 6, 7, 8, and 9.

Final_Y...	0	1
0	2776	154
1	193	172

Correct classified: 2,948 Wrong classified: 347

Accuracy: 89.469 % Error: 10.531 %

Cohen's kappa (K) 0.439

Figure 5

Final_Y...	0	1
0	2821	109
1	194	171

Correct classified: 2,992 Wrong classified: 303

Accuracy: 90.804 % Error: 9.196 %

Cohen's kappa (K) 0.48

Figure 6

Final_Y...	0	1
0	2818	112
1	201	164

Correct classified: Wrong classified: 313

Accuracy: 90.501 % Error: 9.499 %

Cohen's kappa (K) 0.46

Figure 7

Final_Y...	0	1
0	2823	107
1	199	166

Correct classified: Wrong classified: 306

Accuracy: 90.713 % Error: 9.287 %

Cohen's kappa (K)

Figure 8

Final_Y...	0	1
0	2827	103
1	191	174

Correct classified: Wrong classified: 294

Accuracy: 91.077 % Error: 8.923 %

Cohen's kappa (K)

Figure 9

Final_Y...	0	1
0	2832	98
1	190	175

Correct classified: Wrong classified: 288

Accuracy: 91.259 % Error: 8.741 %

Cohen's kappa (K)

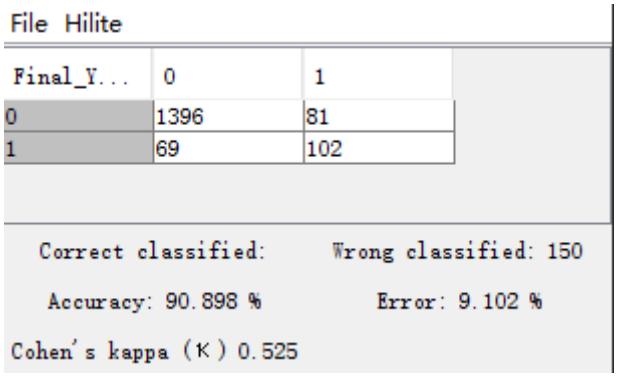
Figure 10

When we rise the K value, the accuracy of the model will be higher.

Decision Tree

When the Decision Tree model is created, the value of the Min number records per node

parameter is very important. When the value is 20, the results of Scorer's evaluation are shown in Figure 11.


File Hilite

Final_Y...	0	1
0	1396	81
1	69	102

Correct classified: Wrong classified: 150

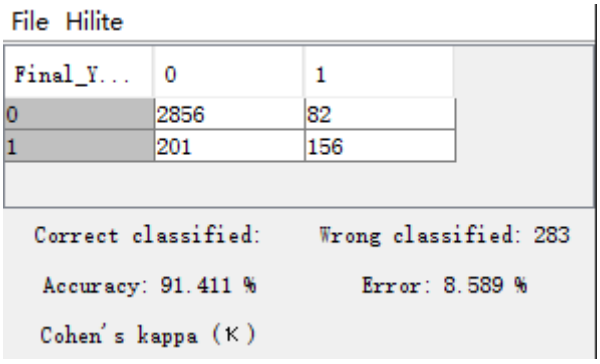
Accuracy: 90.898 % Error: 9.102 %

Cohen's kappa (K) 0.525

Figure 11

Random Forest

When the data was modeled by Random Forest, Scorer's evaluation results are shown in Figure 12.


File Hilite

Final_Y...	0	1
0	2856	82
1	201	156

Correct classified: Wrong classified: 283

Accuracy: 91.411 % Error: 8.589 %

Cohen's kappa (K)

Figure 12

Summary

The results obtained by the three modeling methods which are K Nearest Neighbor, decision tree and random forest are shown in Figures 5 to 12. The model constructed by the random forest method has the highest correct rate, reaching 91.411%. The model established by the decision tree has the lowest accuracy rate of only 90.898%. However, since the value of Final_Y is 1 in this data analysis, it has the greatest impact on the customer's revenue. Therefore, the decision tree with the lowest correct rate is the most priority method because it has the most accurate prediction when the Final_Y value is 1.