1 Predictors

With knowledge of the dataset's characteristics, different types of predictors were built to help determine whether a visitor will generate Revenue by making a purchase at Nozama's website. Three predictors, namely Random Forest (RF), Support Vector Machine (SVM), and k-nearest neighbours (k-NN).

1.1 Random Forests

A RF predictor was built first, as seen in fig. 1.1, since its reliability gives a good baseline to evaluate other predictors on.

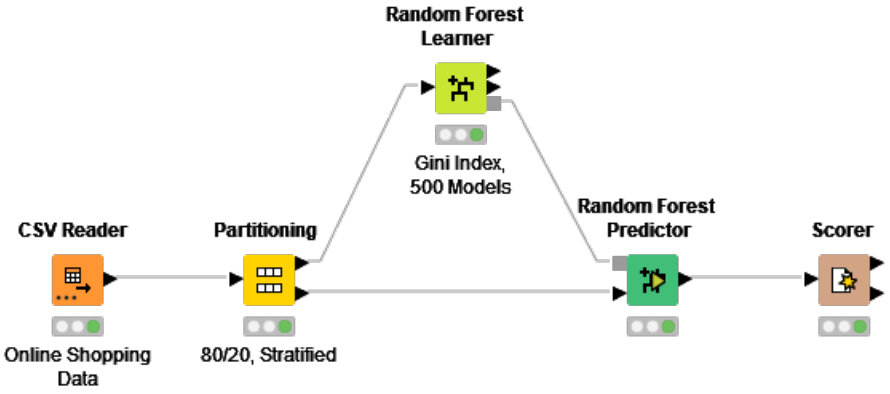

Figure 1.1: KNIME workflow used to build the RF predictor.

Revenue \ Prediction (Revenue)	FALSE	TRUE
FALSE	1996	88
TRUE	160	222

Correct classified: 2,218 Wrong classified: 248

Accuracy: 89.943 % Error: 10.057 %

Cohen's kappa (κ) 0.584

Figure 1.2: Confusion matrix of the RF predictor built.

The prediction accuracy of 89.9 % seen from the confusion matrix shown in fig. 1.2 is greater than the minimum prediction accuracy of 84.5 % determined in chapter 1. Then the RF predictor does provide value as an improvement over naively guessing that no visitor will generate Revenue.

1.2 Support Vector Machines

An SVM predictor was built next, as seen in fig. 1.3, since the prediction attribute has two classes. But using any of the 3 kernel functions (polynomial, hypertangent, radial basis), results in a predictor that's less accurate than the RF predictor and sometimes doesn't meet the 84.5 % threshold of naively guessing. This may be due to the lack of separation between the classes seen in fig. 1.1a, resulting in the difficulty of a kernel function to map the data points into a space where there is sufficient separation between the classes.

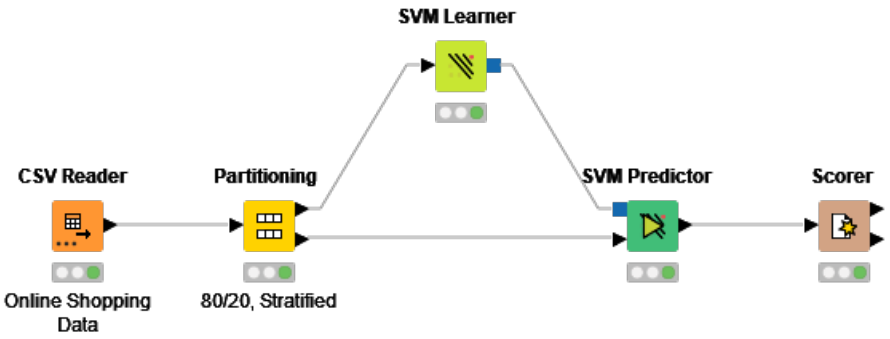

Figure 1.3: KNIME workflow used to build the SVM predictor.

1.3 k-nearest neighbours

k-NN was the last prediction method to consider, as built in fig. 1.4, since cluster results from chapter 1 indicated that it might be feasible to for a k-NN to have a higher prediction accuracy than the RF predictor.

Figure 1.4: KNIME workflow used to build the k-NN predictor.

Revenue \ Class [kNN]	FALSE	TRUE
FALSE	2043	41
TRUE	279	103

Correct classified: 2,146 Wrong classified: 320

Accuracy: 87.024 % Error: 12.976 %

Cohen's kappa (κ) 0.335

Figure 1.5: Confusion matrix of the k-NN predictor built.

The prediction accuracy of 87.0 % observed in the confusion matrix shown in fig. 1.2 is better than naively guessing all FALSE, but doesn't improve in the RF predictor. In retrospect, the non-existent separation between the blue and red data points in fig. 1.1a imply that that the k-NN predictor would have a difficult time dealing with data points close tho the boundary between the two revenue classes.