K-Nearest Neighbors Classification

Customer churn occurs when a company's customers or subscribers quit their business with the company or unsubscribe from the company's service. Also known as customer attrition, customer churn is an important metric for a company, as it is much less expensive to retain existing customers than it is to acquire new customers. Predicting whether a customer is likely to leave a company or not is crucial, as identifying potential churning customers beforehand can lead to proactive actions to make them stay. In this example we build a machine learning model to predict whether a customer is going to leave a company. More specifically, a k-nearest neighbor's algorithm is used to predict the target variable:

- **Churn:** Whether a customer left the company (yes) or not (no)

By using many predictor variables related to the customer's subscription, like:

- **Monthly Charges:** The fee that the customer paid to the company per month
- **Gender:** The sex of the customer.
- **Internet Service:** Whether the customer's subscription also included internet service.
- **Payment Method:** The method the customer used to pay his or her subscription
- And many more variables.

K-Nearest Neighbors Classification

Nearest neighbors	Weights	Distance	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy
29	rectangular	Euclidean	4500	1126	1406	0.815	0.791

Note. The model is optimized with respect to the *validation set accuracy*.

The *K-Nearest Neighbors model table* shows some of the used parameters for the algorithm, along with the sample sizes for the different subsets of the data. Since the algorithm is optimized, both the validation and test set accuracy are shown.

Data Split

Confusion Matrix

		Predicted	
		No	Yes
Observed	No	891	128
	Yes	166	221

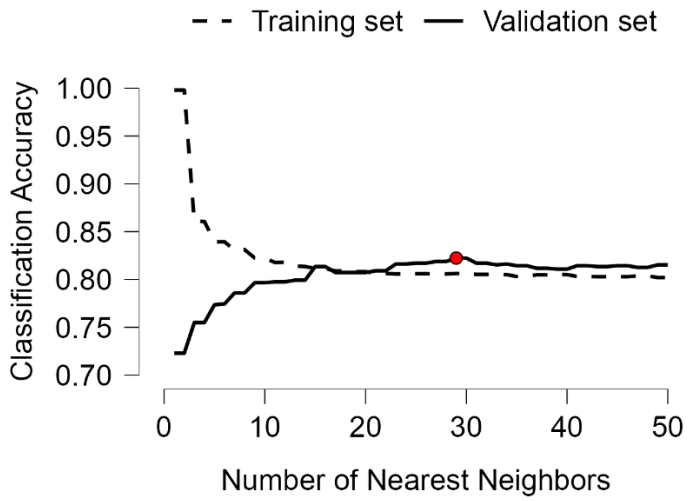
The confusion matrix gives insight into the prediction error on the test set, and shows how the included observations were predicted by the model. It shows that for 641 + 316 = 957 of these observations, the observed class matched the predicted class, resulting in a test accuracy of 957 / 1406 = 0.681.

Class Proportions

	Data Set	Training Set	Validation Set	Test Set
No	0.734	0.736	0.737	0.725
Yes	0.266	0.264	0.263	0.275

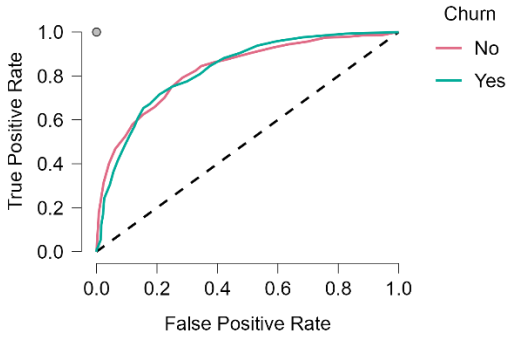
Before the data was imported, a test set indicator variable was made. This variable contained 20 percent of the data and consisted of an equal amount of churning customers and non-churning customers. Our test set indicator variable made sure that we have equal proportions in of both classes in our test set (0.5).

Classification Accuracy Plot



We can see how the performance of the model increases by making a *classification accuracy plot*. The red dot represents the number of nearest neighbors with the highest validation classification accuracy, in our case 21.

ROC Curves Plot



Under plots we can, among other things, plot the corresponding ROC curves for both classes.