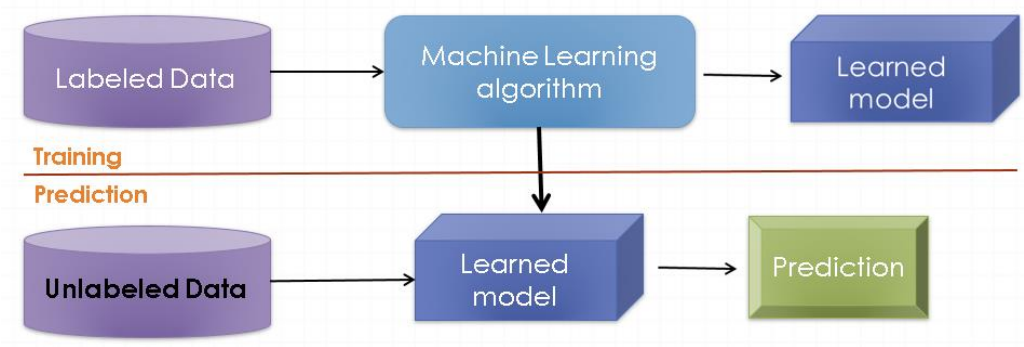
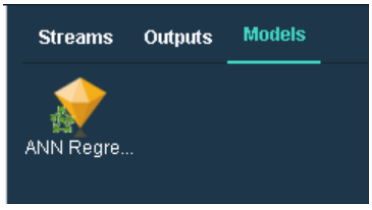
Building a Machine Learning Model and using the model on unlabeled data



The reason we build ML algorithms is to use them on unlabeled data to predict the target variable. The above diagram shows the Training and Prediction (using the model) phases. Once the model is trained, and the model evaluation satisfactory, you are ready to use the model for prediction. In the SPSS modeler, the trained model can be found in the **right-north** window. In order to make predictions, you can drag the model icon onto the canvas, and then connect it to the input **unlabeled** data. Next attach an output table and run. The predictions on the target variable will be in the output table.



Warning: When you use SPSS Modeler to make predictions, make sure that the file type of labelled data and unlabeled data is the same. If you created the training data as a **csv** file, then the unlabeled data for prediction must also be a **csv** file.

1. **Linear Regression:** In HW 2 Q1, the regression model is as follows

Sales = 0.471846 +0.06116*children TV + 0.062 * newspaper

For **Children TV = 31, Newspapers = 12**; substitute the values into the equation to get **Sales = $3.1118 million**

2. **Applying a Classification (Decision Tree) model to unseen data:**

The following stream shows how to first build a classification model and then apply (or use the model) for business purposes. The file **score customers for churn.csv** contains information on current customers of the company. If you open the file, you will see that it does not have the Churn column, since we do not know if these customers are likely to leave or will stay as current customers. The

Business wants to know which of these customers are likely to leave. To find that out, we make a copy of the model built in the train/test phase of model building. So, this model has all the rules generated by the decision tree. When you connect score customers file to the model, the model uses the rules and assigns a class to each row in the input file. Next, we are only interested in the class 'Vol', the customers who are likely to leave. Next, we sort the results by the Confidence of the classification ($CC-Churned), in descending order. Connect to an output table and you can see results, where the customers with a high probability to leave are on top. Similarly you can also select "Invol" or "Current" to understand these classes better.

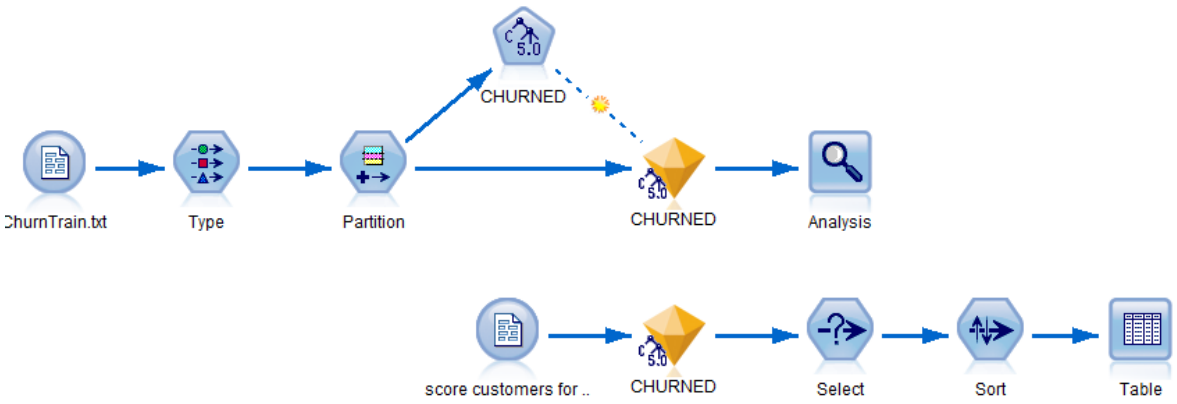


Figure 1: Train and test a model and apply the model on new data, that was not seen by the model, to make class predictions.

Once you execute the second stream, open and look at the output table created, it has a classification **($C-Churned**) and a classification confidence **($CC-Churned**) added to each customer record. We are only interested in those customers that may leave voluntarily, and want to know the probability that they may leave. Sorting on Descending order w.r.t. the probability ($CC-Churned), we have

	ID	LONGDIST	International	LOCAL	DROPPED	PAY_MTHD	LocalBillType	LongDistanceBillType	AGE	SEX	STATUS	CHILDREN	Est_Income	Car_Owner	$C-CHURNED	$CC-CHURNED
1	1067	17.339	0.000	9.316	1	CH	Budget	Intnl_discount	44	F	S	2	17689.200	N	Vol	0.913
2	661	17.795	0.000	12.565	1	CC	FreeLocal	Standard	58	F	M	2	29597.100	N	Vol	0.913
3	3953	6.957	0.000	11.030	3	CC	Budget	Standard	87	F	S	0	70492.400	N	Vol	0.913
4	3259	12.760	0.000	16.420	1	CC	Budget	Intnl_discount	73	F	M	1	19621.600	N	Vol	0.913
5	4644	26.120	9.527	37.886	1	CH	Budget	Standard	54	F	S	1	11659.500	Y	Vol	0.913
6	1193	19.780	0.000	57.618	3	CC	FreeLocal	Intnl_discount	73	F	S	0	23832.000	N	Vol	0.913
7	2387	12.059	0.000	27.388	1	Auto	FreeLocal	Intnl_discount	41	F	M	1	3920.430	Y	Vol	0.913
8	1596	14.806	0.000	121.406	1	Auto	FreeLocal	Intnl_discount	50	F	M	2	96821.100	N	Vol	0.913
9	3583	19.325	0.000	10.673	1	CC	Budget	Standard	88	F	M	2	66906.600	N	Vol	0.913
10	849	25.653	0.813	81.738	1	CH	FreeLocal	Intnl_discount	69	F	M	0	49188.900	Y	Vol	0.913
11	4962	19.652	0.000	146.405	0	CC	Budget	Standard	80	F	M	2	83890.700	N	Vol	0.889
12	972	14.046	4.429	114.323	0	CC	Budget	Standard	92	F	S	2	47396.200	N	Vol	0.889

Figure 2: sample list of Customers who are predicted to be of class 'Vol', listed in descending order of @@CC-Churned.

Thus, we have identified the customers that most likely to leave voluntarily. Next, management has come up with a scheme to reach out to these customers, and make them stay.

The input file is **score customers for churn.csv,** but the same data stored as an EXCEL file will run into difficulties.

The same process of building a model and applying the model to a different dataset can be used, except the analysis will vary after applying the model to a new dataset.

In this example, a new dataset is being presented to a clustering algorithm. The model in this case adds a cluster ID to each record in the input data. Next it is up to you to use all the available tools to do further analysis and derive insight. If you want to select each cluster for further analysis, you may do that. You may analyze all the clusters together as we did in class.

3. Applying a Classification Model (Wisc. Breast Cancer Data):

In this example we have a sample of 24 patient tumors, which we do not know if they are Benign or Malignant. We can build a classification model using a classifier of our choice and then apply to the data.

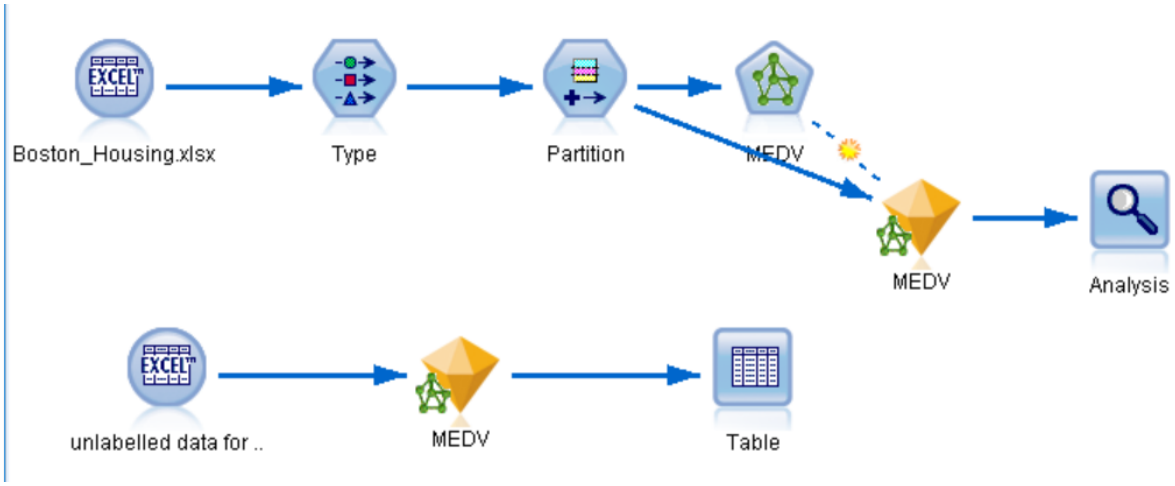
	group	id	clump thickness	size uniformity	shape uniformity	Marginal adhesion	epithelialcell size	bare nuclei	blandchromatin	normalNucleoli	Mitoses	$N-Diagnosis	$NC-Diagnosis
1	10.0...	2000061....	2.000	1.000	1.000	1.000	2.000	1.000	2.000	1.000	1.000	Beneign	0.993
2	10.0...	2000060....	1.000	1.000	1.000	1.000	1.000	1.000	3.000	1.000	1.000	Beneign	0.993
3	10.0...	2000072....	3.000	1.000	1.000	1.000	2.000	1.000	2.000	1.000	1.000	Beneign	0.992
4	10.0...	2000057....	2.000	1.000	2.000	1.000	2.000	1.000	3.000	1.000	1.000	Beneign	0.990
5	10.0...	2000066....	4.000	1.000	1.000	1.000	2.000	1.000	2.000	1.000	1.000	Beneign	0.990
6	10.0...	2000055....	8.000	9.000	9.000	8.000	7.000	10.000	9.000	6.000	1.000	Malignant	0.990
7	10.0...	2000059....	4.000	2.000	1.000	1.000	2.000	1.000	2.000	1.000	1.000	Beneign	0.990
8	10.0...	2000064....	8.000	7.000	5.000	10.000	7.000	9.000	5.000	5.000	4.000	Malignant	0.989
9	10.0...	2000063....	1.000	1.000	1.000	1.000	2.000	3.000	3.000	1.000	1.000	Beneign	0.988
10	10.0...	2000068....	9.000	7.000	7.000	6.000	4.000	9.000	4.000	1.000	2.000	Malignant	0.987
11	10.0...	2000071....	10.000	5.000	5.000	3.000	6.000	7.000	7.000	9.000	2.000	Malignant	0.987
12	10.0...	2000052....	3.000	1.000	1.000	1.000	2.000	2.000	3.000	1.000	1.000	Beneign	0.986
13	10.0...	2000067....	4.000	1.000	1.000	1.000	2.000	1.000	3.000	1.000	1.000	Beneign	0.984
14	10.0...	2000054....	4.000	1.000	1.000	3.000	2.000	1.000	3.000	1.000	1.000	Beneign	0.973
15	10.0...	2000069....	6.000	1.000	1.000	1.000	2.000	1.000	3.000	1.000	1.000	Beneign	0.972
16	10.0...	2000070....	7.000	3.000	2.000	9.000	5.000	10.000	5.000	3.000	4.000	Malignant	0.971
17	10.0...	2000058....	2.000	1.000	3.000	1.000	2.000	1.000	1.000	1.000	5.000	Beneign	0.966
18	10.0...	2000050....	4.000	2.000	2.000	2.000	1.000	3.000	3.000	1.000	1.000	Beneign	0.965
19	10.0...	2000051....	5.000	8.000	4.000	5.000	7.000	10.000	3.000	2.000	1.000	Malignant	0.927
20	10.0...	2000053....	6.000	8.000	8.000	1.000	3.000	4.000	3.000	7.000	1.000	Malignant	0.880
21	10.0...	2000065....	7.000	4.000	6.000	4.000	6.000	1.000	4.000	3.000	1.000	Malignant	0.820
22	10.0...	2000056....	1.000	2.000	2.000	1.000	2.000	9.000	3.000	1.000	1.000	Beneign	0.779
23	10.0...	2000073....	8.000	4.000	5.000	1.000	2.000	0.000	7.000	3.000	1.000	Malignant	0.748
24	10.0...	2000062....	5.000	3.000	3.000	3.000	2.000	3.000	4.000	4.000	1.000	Beneign	0.516

The classifier added two columns, namely $N-Diagnosis (classification), and confidence of classification ($NC-Diagnosis). The stream above sorted these records w.r.t. the classification confidence in descending order.

4. Applying an ANN Regression model.

The process is very similar to using a Decision Tree. The only difference is you will get predicted values for the target variable. For example, if you apply the Regression model for Boston Housing data, you will get the predicted Median house Price (MEDV).

In the following example, an ANN Regression model is built using the data in the input file Boston Housing Data.xlsx. After model is built, the model is used to predict the MEDV for about 20 test records.



The table on the next page shows the results after the model added the predicted value **($N-MEDV**) for each test record. Note that in the input data set, there is no MEDV column, since we do not know what that value is and want to predict using the model

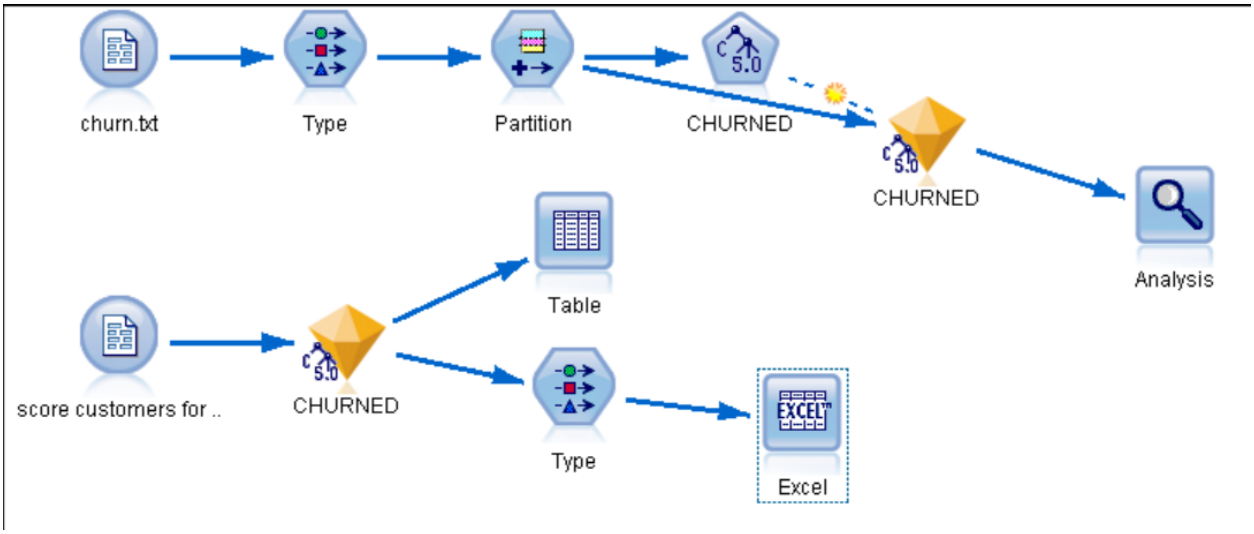
	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	$N-MEDV
1	0.006	18.000	2.310	0.000	0.546	6.575	65.200	4.090	1.000	296.000	15.300	396.900	4.980	27.613
2	0.027	0.000	7.070	0.000	0.469	6.421	78.900	4.967	2.000	242.000	17.800	396.900	9.140	25.329
3	0.726	0.000	8.140	0.000	0.538	5.727	69.500	3.796	4.000	307.000	21.000	390.950	11.280	17.402
4	0.064	0.000	5.960	0.000	0.499	5.933	68.200	3.360	5.000	279.000	19.200	396.900	9.680	20.366
5	0.028	75.000	2.950	0.000	0.428	6.595	21.800	5.401	3.000	252.000	18.300	395.630	4.320	30.371
6	0.229	0.000	6.910	0.000	0.448	6.030	85.500	5.689	3.000	233.000	17.900	392.740	18.800	22.355
7	0.254	0.000	6.910	0.000	0.448	5.399	95.300	5.870	3.000	233.000	17.900	396.900	30.810	20.946
8	0.220	0.000	6.910	0.000	0.501	5.602	62.000	6.088	3.000	233.000	17.900	396.900	16.200	21.423
9	0.089	21.000	5.640	0.000	0.438	5.963	45.700	6.815	4.000	243.000	16.800	395.560	13.450	23.027
10	0.014	75.000	4.000	0.000	0.410	5.888	47.600	7.320	3.000	469.000	21.100	396.900	14.800	17.269
11	0.013	90.000	1.220	0.000	0.403	7.249	21.900	8.697	5.000	226.000	17.900	395.930	4.810	31.616
12	0.021	85.000	0.740	0.000	0.410	6.383	35.700	9.188	2.000	313.000	17.300	396.900	5.770	24.329
13	0.014	100.000	1.320	0.000	0.411	6.816	40.500	8.325	5.000	256.000	15.100	392.900	3.950	31.525
14	0.154	25.000	5.130	0.000	0.453	6.145	29.200	7.815	8.000	284.000	19.700	390.680	6.860	21.918
15	0.020	17.500	1.380	0.000	0.420	7.104	59.500	9.223	3.000	216.000	18.600	393.240	8.050	26.839
16	0.036	80.000	3.370	0.000	0.398	6.290	17.800	6.612	4.000	337.000	16.100	396.900	4.670	25.635
17	0.092	0.000	10.810	0.000	0.413	6.065	7.800	5.287	4.000	305.000	19.200	390.910	5.520	23.062
18	0.195	0.000	10.810	0.000	0.413	6.245	6.200	5.287	4.000	305.000	19.200	377.170	7.540	24.159
19	0.079	0.000	12.830	0.000	0.437	6.273	6.000	4.252	5.000	398.000	18.700	394.920	6.780	24.362

The unlabeled data sets for running the models as shown above are on the Black Board.

5. Exporting results from the Model to Excel or other file types

After creating a model, and using it create classification, you can export the results to an excel file If you want to. The following stream shows how to export.

In particular, note that when exporting to a file that needs to know data type, you need to use a type node to specify the data type of all columns. Once you open the Excel file, make sure to specify the number of decimal places you want in each column that is numeric.



6. On building and using an Apriori association rules model and using it.

Please look at the workshop on association rules.