Predictive Modeling Fundamentals I

Lab 4: Scoring test data

Contents

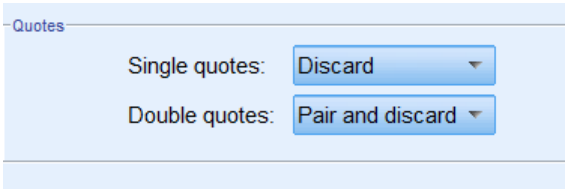
Scoring test data

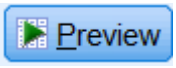
1.1 Using your model

In this lab you will use the model that you created to score the test.csv dataset in order to see which passengers would be predicted to survive.

_1. From the **Sources** palette, add a **Var File** node.

_2. Double-click the **Var File** node to open a dialog box. Open the file **test.csv**. Check that the quotation options are selected as below (to ensure that the fields – such as passenger names - are properly separated).



Click in the  button to have a quick look of your data. This will display the first 10 records of the dataset.

Preview from test.csv Node (11 fields, 10 records)

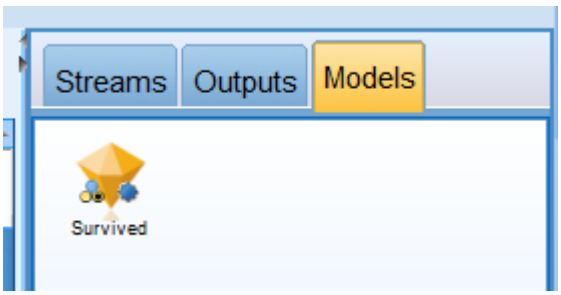
File Edit Generate

Table Annotations

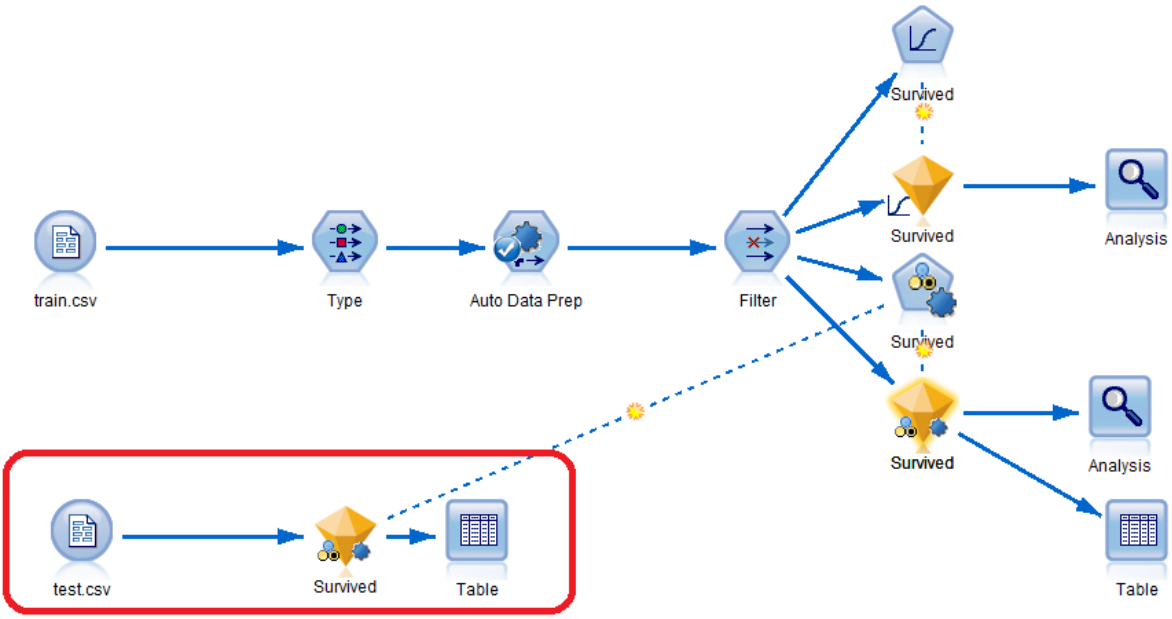
	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	892	3	Kelly, Mr. James	male	34....	0	0	330911	7.829		Q
2	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47....	1	0	363272	7.000		S
3	894	2	Myles, Mr. Thomas Francis	male	62....	0	0	240276	9.688		Q
4	895	3	Wirz, Mr. Albert	male	27....	0	0	315154	8.662		S
5	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22....	1	1	3101298	12.288		S
6	897	3	Svensson, Mr. Johan Cervin	male	14....	0	0	7538	9.225		S
7	898	3	Connolly, Miss. Kate	female	30....	0	0	330972	7.629		Q
8	899	2	Caldwell, Mr. Albert Francis	male	26....	1	1	248738	29.000		S
9	900	3	Abrahim, Mrs. Joseph (Sophie Halaut Easu)	female	18....	0	0	2657	7.229		C
10	901	3	Davies, Mr. John Samuel	male	21....	2	0	A/4 48871	24.150		S

OK

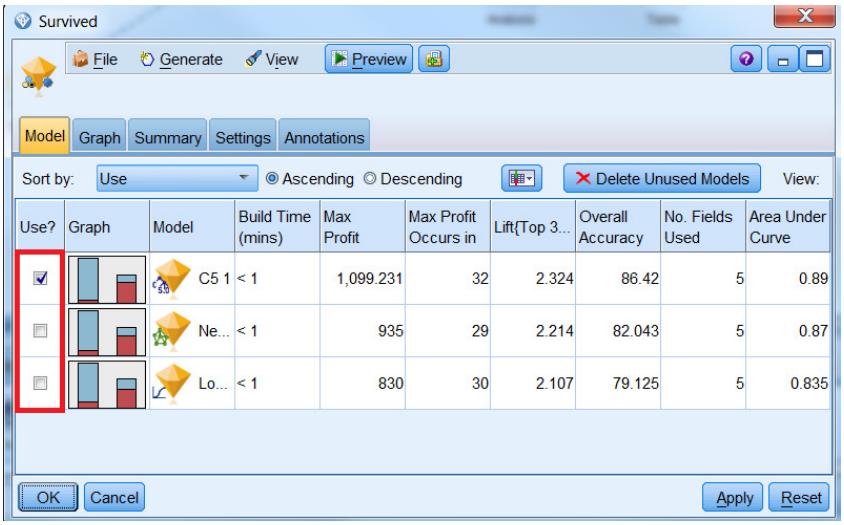
_3. Click on the **Models** tab on the right menu. Here there are all the models that you have created.



_4. Add the **Survived** model to the canvas and connect it to the data source. Connect a **Table** node from the **Output palette**.

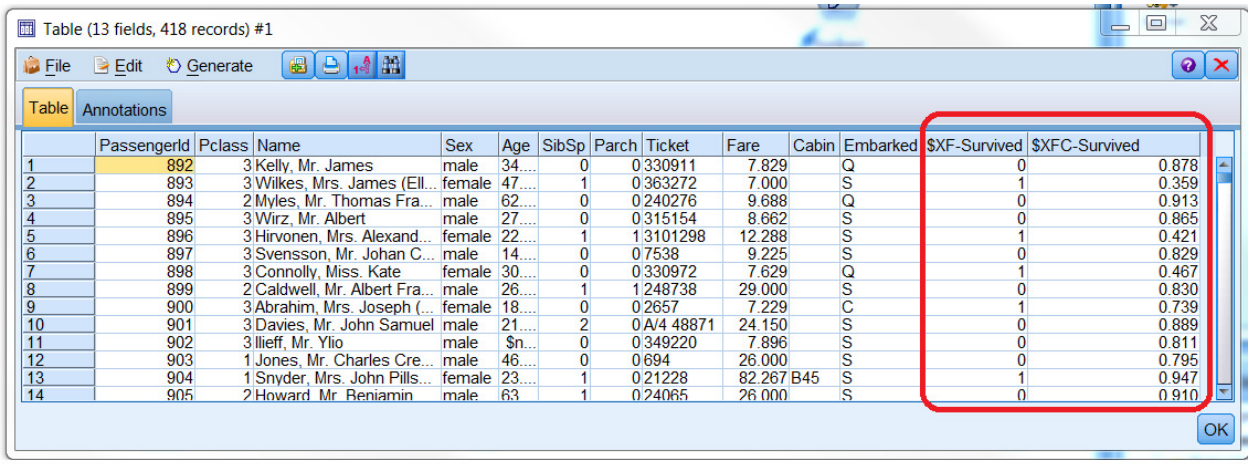


_5. Open the **Survived** model nugget and uncheck the models that you will not be using. In this case we are using **C5.1** as it's the one with highest accuracy.



_6. Run the stream in the **Table** node (double click the node and hit run). The output will be a table with two new fields:

- $XF-Survived: This is the survival prediction that we wanted
- $XFC-Survived: This is the confidence of the prediction

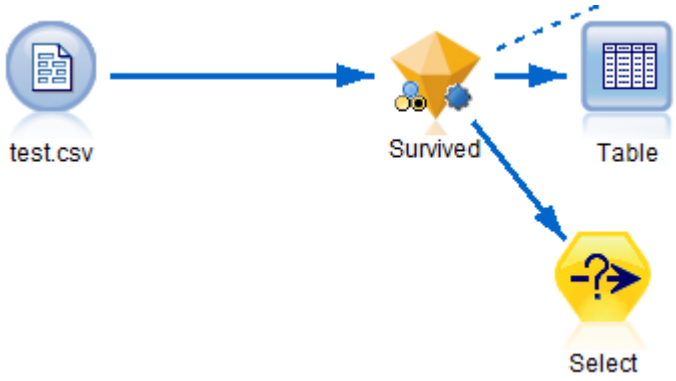


	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	$XF-Survived	$XFC-Survived
1	892	3	Kelly, Mr. James	male	34....	0	0	330911	7.829		Q	0	0.878
2	893	3	Wilkes, Mrs. James (Ell...	female	47....	1	0	363272	7.000		S	1	0.359
3	894	2	Myles, Mr. Thomas Fra...	male	62....	0	0	240276	9.688		Q	0	0.913
4	895	3	Wirz, Mr. Albert	male	27....	0	0	315154	8.662		S	0	0.865
5	896	3	Hirvonen, Mrs. Alexand...	female	22....	1	1	3101298	12.288		S	1	0.421
6	897	3	Svensson, Mr. Johan C...	male	14....	0	0	7538	9.225		S	0	0.829
7	898	3	Connolly, Miss. Kate	female	30....	0	0	330972	7.629		Q	1	0.467
8	899	2	Caldwell, Mr. Albert Fra...	male	26....	1	1	248738	29.000		S	0	0.830
9	900	3	Abrahim, Mrs. Joseph (...	female	18....	0	0	2657	7.229		C	1	0.739
10	901	3	Davies, Mr. John Samuel	male	21....	2	0	A/4 48871	24.150		S	0	0.889
11	902	3	Ilieff, Mr. Ylio	male	$n...	0	0	349220	7.896		S	0	0.811
12	903	1	Jones, Mr. Charles Cre...	male	46....	0	0	694	26.000		S	0	0.795
13	904	1	Snyder, Mrs. John Pills...	female	23....	1	0	21228	82.267	B45	S	1	0.947
14	905	2	Howard, Mr. Benjamin	male	63	1	0	24065	26.000		S	0	0.910

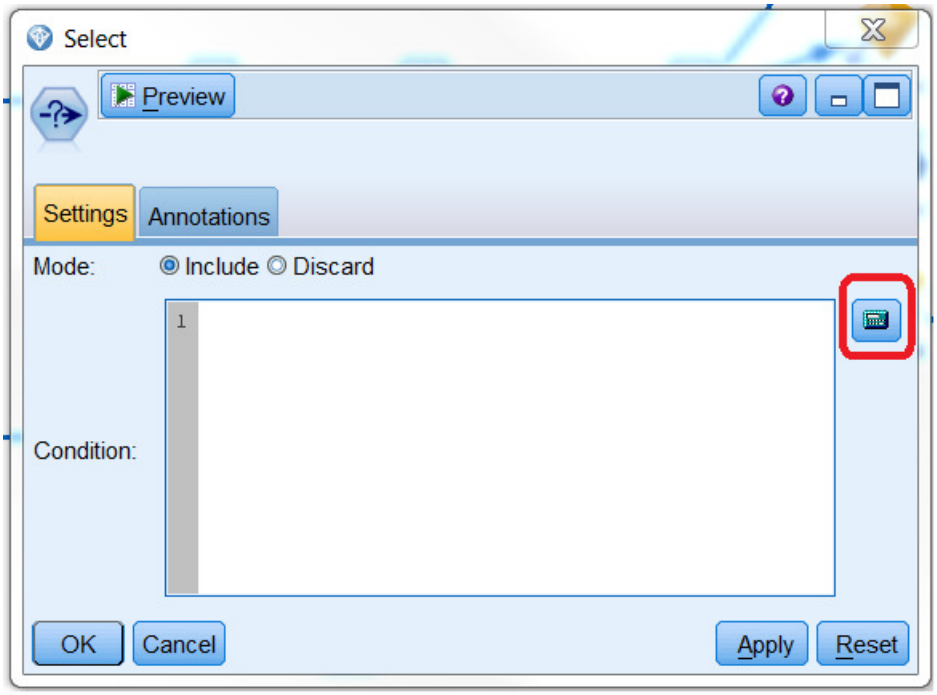
1.2 Select the prediction with 80% confidence

We scored 418 records but some of the predictions have a very low confidence. We shouldn't trust them, that's why we will only select the predictions with a confidence level greater than 80%.

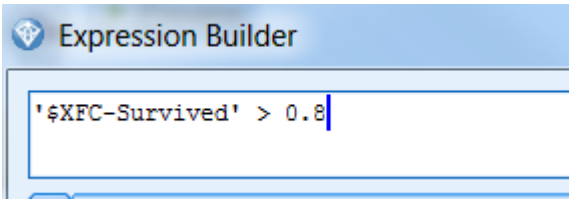
_1. Add a **Select** node from the **Record Ops** palette and connect it to the **Survived** Model.



_2. Double click on the **Select** node. The options of the node will appear. Click on the small Calculator icon.

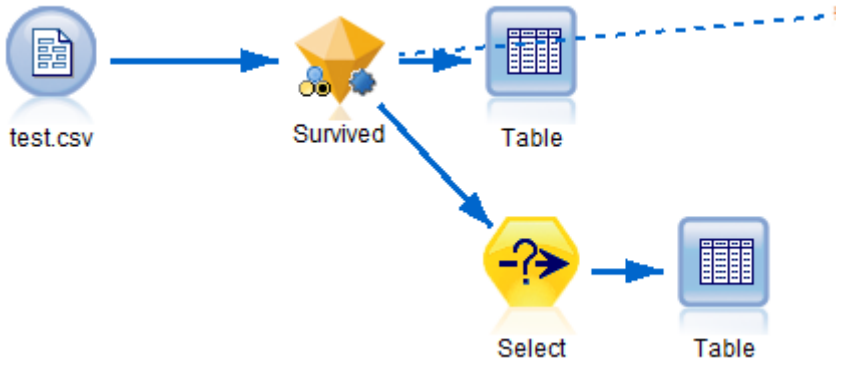


_3. Now you are in the **Expression Builder**. Here you can type CLEM expressions manually or use the Expression Builder, which displays a complete list of CLEM functions and operators as well as data fields from the current stream , allowing you to quickly build expressions without memorizing the exact names of fields or functions. To keep it simple just add this text:

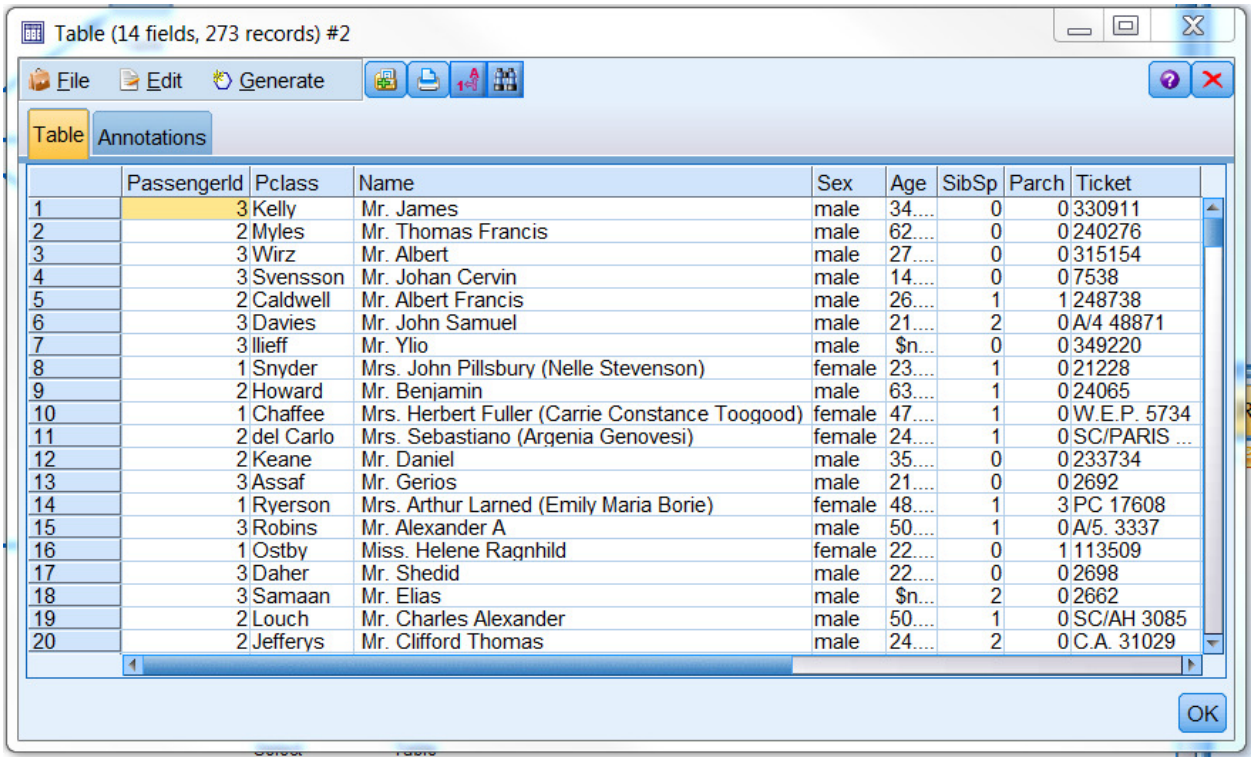


Click **Check** and **Ok**. This expression will select the records with a confidence greater than 0.8.

_4. Add a **Table** node from the **Output palette**.



_5. Run the stream. You will get the same table as before but now with only 273 records. Check that are the records are over 0.8. These are predictions that have higher statistical confidence!



	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
1	3	Kelly	Mr. James	male	34....	0	0	330911
2	2	Myles	Mr. Thomas Francis	male	62....	0	0	240276
3	3	Wirz	Mr. Albert	male	27....	0	0	315154
4	3	Svensson	Mr. Johan Cervin	male	14....	0	0	7538
5	2	Caldwell	Mr. Albert Francis	male	26....	1	1	248738
6	3	Davies	Mr. John Samuel	male	21....	2	0	A/4 48871
7	3	Ilieff	Mr. Ylio	male	$n...	0	0	349220
8	1	Snyder	Mrs. John Pillsbury (Nelle Stevenson)	female	23....	1	0	21228
9	2	Howard	Mr. Benjamin	male	63....	1	0	24065
10	1	Chaffee	Mrs. Herbert Fuller (Carrie Constance Toogood)	female	47....	1	0	W.E.P. 5734
11	2	del Carlo	Mrs. Sebastiano (Argenia Genovesi)	female	24....	1	0	SC/PARIS ...
12	2	Keane	Mr. Daniel	male	35....	0	0	233734
13	3	Assaf	Mr. Gerios	male	21....	0	0	2692
14	1	Ryerson	Mrs. Arthur Larned (Emily Maria Borie)	female	48....	1	3	PC 17608
15	3	Robins	Mr. Alexander A	male	50....	1	0	A/5. 3337
16	1	Ostby	Miss. Helene Ragnhild	female	22....	0	1	113509
17	3	Daher	Mr. Shedid	male	22....	0	0	2698
18	3	Samaan	Mr. Elias	male	$n...	2	0	2662
19	2	Louch	Mr. Charles Alexander	male	50....	1	0	SC/AH 3085
20	2	Jefferys	Mr. Clifford Thomas	male	24....	2	0	C.A. 31029

Summary

Congratulations! You used your model to score new records!

In this lab you learnt how to use your model to score new records. Then we selected the records with a confidence higher than 80% using the Select node.