* [Classifying Telecommunications Customers (Discriminant Analysis)](http://127.0.0.1:54857/help/topic/com.ibm.spss.modeler.tutorial/clementine/example_telco_custcat_discriminant.htm)

# Classifying Telecommunications Customers (Discriminant Analysis)

Discriminant analysis is a statistical technique for classifying records based on values of input fields. It is analogous to linear regression but takes a categorical target field instead of a numeric one.

For example, suppose a telecommunications provider has segmented its customer base by service usage patterns, categorizing the customers into four groups. If demographic data can be used to predict group membership, you can customize offers for individual prospective customers.

This example uses the stream named telco\_custcat\_discriminant.str, which references the data file named telco.sav. These files are available from the Demos directory of any IBM® SPSS® Modeler installation. This can be accessed from the IBM SPSS Modeler program group on the Windows Start menu. The telco\_custcat\_discriminant.str file is in the streams directory.

The example focuses on using demographic data to predict usage patterns. The target field custcat has four possible values which correspond to the four customer groups, as follows:

| **Value** | **Label** |
| --- | --- |
| 1 | Basic Service |
| 2 | E-Service |
| 3 | Plus Service |
| 4 | Total Service |

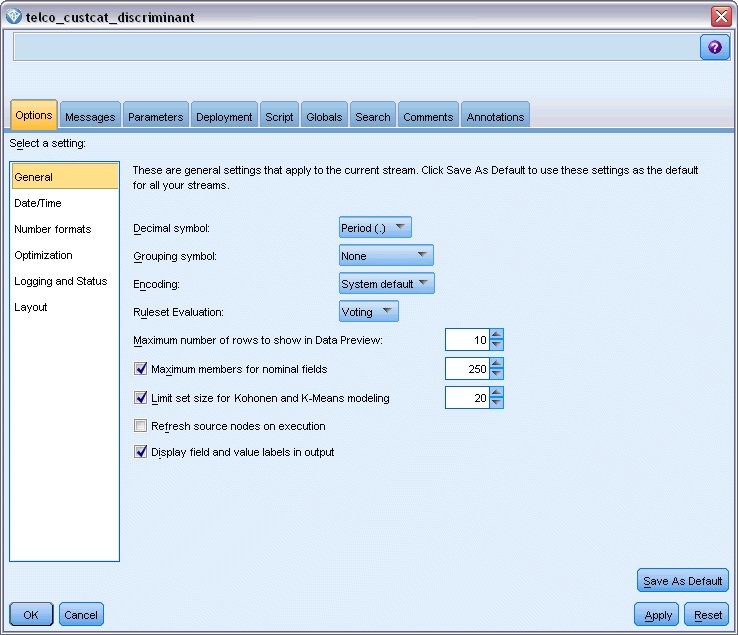
**Creating the Stream**

1. First, set the stream properties to show variable and value labels in the output. From the menus, choose:

**File** > **Stream Properties...** > **Options** > **General**

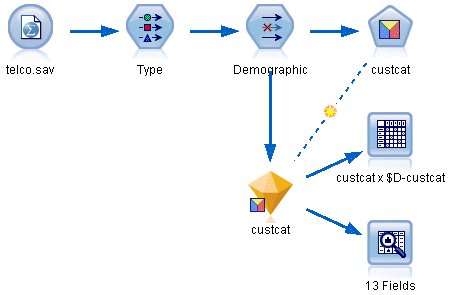
1. Make sure that **Display field and value labels in output** is selected and click **OK**.

*Figure 1. Stream properties*



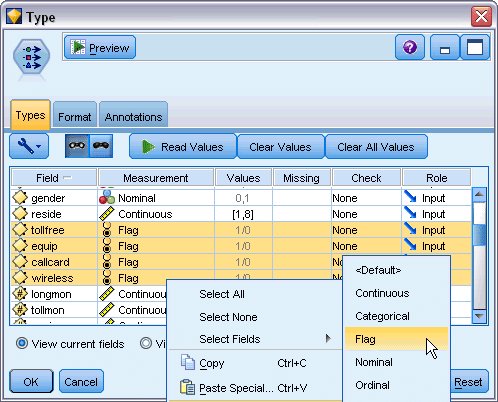
1. Add a Statistics File source node pointing to *telco.sav* in the *Demos* folder.

*Figure 2. Sample stream to classify customers using discriminant analysis*



* 1. Add a Type node and click **Read Values**, making sure that all measurement levels are set correctly. For example, most fields with values 0 and 1 can be regarded as flags.

*Figure 3. Setting the measurement level for multiple fields*

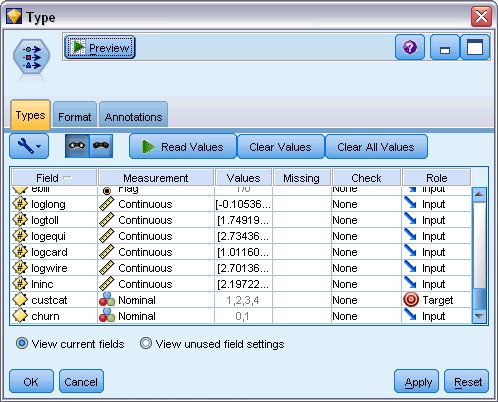


*Tip*: To change properties for multiple fields with similar values (such as 0/1), click the *Values* column header to sort fields by value, and then hold down the shift key while using the mouse or arrow keys to select all the fields you want to change. You can then right-click on the selection to change the measurement level or other attributes of the selected fields.

Notice that *gender* is more correctly considered as a field with a set of two values, instead of a flag, so leave its Measurement value as **Nominal**.

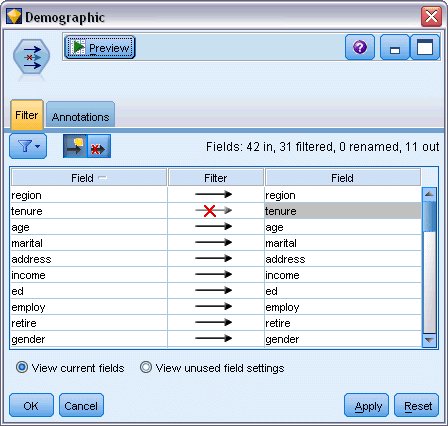
* 1. Set the role for the *custcat* field to **Target**. All other fields should have their role set to **Input**.

*Figure 4. Setting field role*



Since this example focuses on demographics, use a Filter node to include only the relevant fields (*region*, *age*, *marital*, *address*, *income*, *ed*, *employ*, *retire*, *gender*, *reside*, and *custcat*). Other fields can be excluded for the purpose of this analysis.

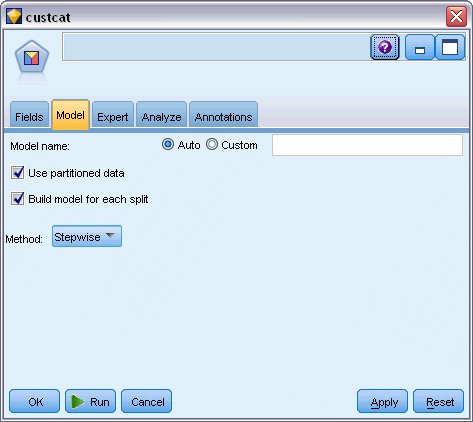
*Figure 5. Filtering on demographic fields*



(Alternatively, you could change the role to **None** for these fields rather than exclude them, or select the fields you want to use in the modeling node.)

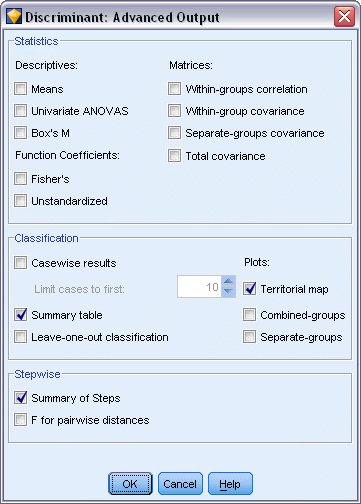
1. In the Discriminant node, click the Model tab and select the **Stepwise** method.

*Figure 6. Choosing model options*



1. On the Expert tab, set the mode to **Expert** and click **Output**.
2. Select **Summary table**, **Territorial map**, and **Summary of Steps** in the Advanced Output dialog box, then click **OK**.

*Figure 7. Choosing output options*

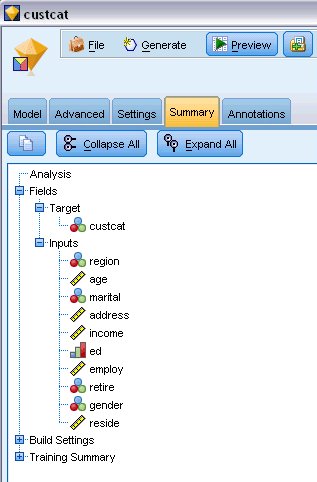


# Examining the Model

1. Click **Run** to create the model, which is added to the stream and to the Models palette in the upper-right corner. To view its details, double-click on the model nugget in the stream.

The Summary tab shows (among other things) the target and the complete list of inputs (predictor fields) submitted for consideration.

*Figure 1. Model summary showing target and input fields*

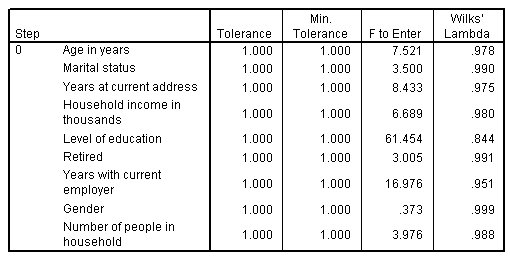


For details of the discriminant analysis results:

1. Click the Advanced tab.
2. Click the "Launch in external browser" button (just below the Model tab) to view the results in your Web browser.

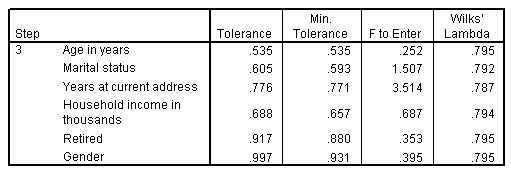
# Stepwise Discriminant Analysis

*Figure 1. Variables not in the analysis, step 0*



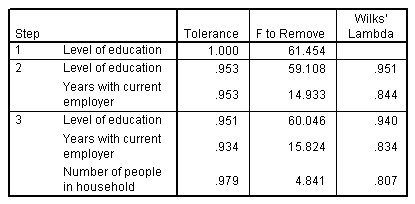
When you have a lot of predictors, the stepwise method can be useful by automatically selecting the "best" variables to use in the model. The stepwise method starts with a model that doesn't include any of the predictors. At each step, the predictor with the largest F to Enter value that exceeds the entry criteria (by default, 3.84) is added to the model.

*Figure 2. Variables not in the analysis, step 3*



The variables left out of the analysis at the last step all have F to Enter values smaller than 3.84, so no more are added.

*Figure 3. Variables in the analysis*



This table displays statistics for the variables that are in the analysis at each step. Tolerance is the proportion of a variable's variance not accounted for by other independent variables in the equation. A variable with very low tolerance contributes little information to a model and can cause computational problems.

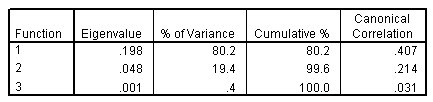
F to Remove values are useful for describing what happens if a variable is removed from the current model (given that the other variables remain). F to Remove for the entering variable is the same as F to Enter at the previous step (shown in the Variables Not in the Analysis table).

# A Note of Caution Concerning Stepwise Methods

Stepwise methods are convenient, but have their limitations. Be aware that because stepwise methods select models based solely upon statistical merit, it may choose predictors that have no practical significance. If you have some experience with the data and have expectations about which predictors are important, you should use that knowledge and eschew stepwise methods. If, however, you have many predictors and no idea where to start, running a stepwise analysis and adjusting the selected model is better than no model at all.

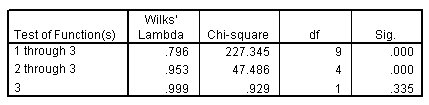
# Checking Model Fit

*Figure 1. Eigenvalues*



Nearly all of the variance explained by the model is due to the first two discriminant functions. Three functions are fit automatically, but due to its minuscule eigenvalue, you can fairly safely ignore the third.

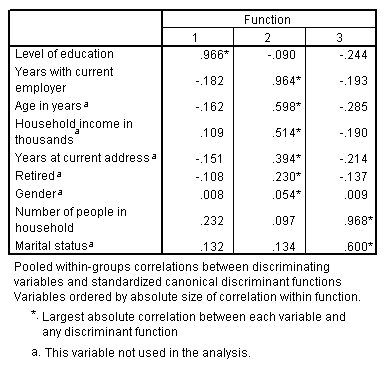
*Figure 2. Wilks' lambda*



Wilks' lambda agrees that only the first two functions are useful. For each set of functions, this tests the hypothesis that the means of the functions listed are equal across groups. The test of function 3 has a significance value greater than 0.10, so this function contributes little to the model.

**Structure Matrix**

*Figure 1. Structure matrix*

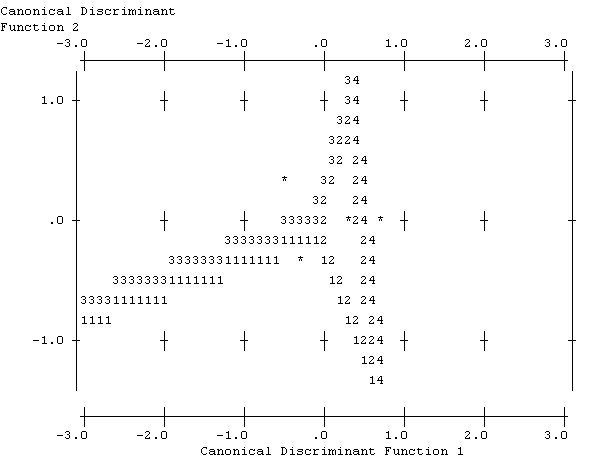


When there is more than one discriminant function, an asterisk(\*) marks each variable's largest absolute correlation with one of the canonical functions. Within each function, these marked variables are then ordered by the size of the correlation.

* *Level of education* is most strongly correlated with the first function, and it is the only variable most strongly correlated with this function.
* *Years with current employer*, *Age in years*, *Household income in thousands*, *Years at current address*, *Retired*, and *Gender* are most strongly correlated with the second function, although *Gender* and *Retired* are more weakly correlated than the others. The other variables mark this function as a "stability" function.
* *Number of people in household* and *Marital status* are most strongly correlated with the third discriminant function, but this is a useless function, so these are nearly useless predictors.

# Territorial Map

*Figure 1. Territorial map*



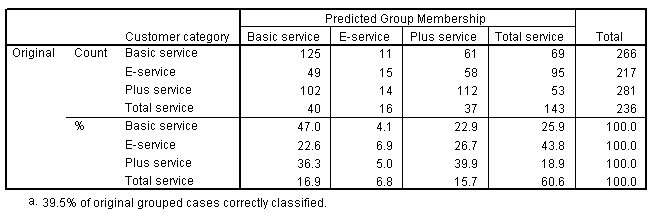
The territorial map helps you to study the relationships between the groups and the discriminant functions. Combined with the structure matrix results, it gives a graphical interpretation of the relationship between predictors and groups. The first function, shown on the horizontal axis, separates group 4 (Total service customers) from the others. Since Level of education is strongly positively correlated with the first function, this suggests that your Total service customers are, in general, the most highly educated. The second function separates groups 1 and 3 (Basic service and Plus service customers). Plus service customers tend to have been working longer and are older than Basic service customers. E-service customers are not separated well from the others, although the map suggests that they tend to be well educated with a moderate amount of work experience.

In general, the closeness of the group centroids, marked with asterisks (\*), to the territorial lines suggests that the separation between all groups is not very strong.

Only the first two discriminant functions are plotted, but since the third function was found to be rather insignificant, the territorial map offers a comprehensive view of the discriminant model.

# Classification Results

*Figure 1. Classification results*



From Wilks' lambda, you know that your model is doing better than guessing, but you need to turn to the classification results to determine how much better. Given the observed data, the "null" model (that is, one without predictors) would classify all customers into the modal group, Plus service. Thus, the null model would be correct 281/1000 = 28.1% of the time. Your model gets 11.4% more or 39.5% of the customers. In particular, your model excels at identifying Total service customers. However, it does an exceptionally poor job of classifying E-service customers. You may need to find another predictor in order to separate these customers.