

Predictive Modeling Fundamentals I

Lab 2: Load, understand and prepare the data for modeling

Contents

<i>Load, understand and prepare the data for modeling.....</i>	<i>3</i>
1.1 Download the dataset from Kaggle website.....	3
1.2 Load the data in IBM SPSS Modeler.....	4
1.3 Exploring a dataset.....	6
1.4 Preparing the data for modeling.....	8
Summary	12

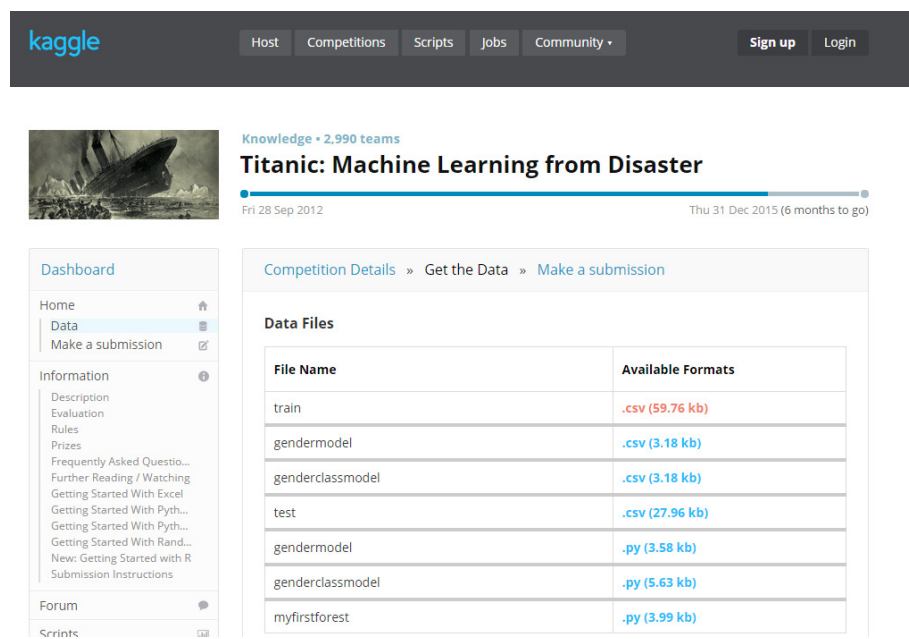
Load, understand and prepare the data for modeling

1.1 Download the dataset from Kaggle website

_1. Go to the Kaggle website and open the “Titanic: Machine Learning from Disaster” in the following URL:

<https://www.kaggle.com/c/titanic/data>

_2. Download the “train.csv” and the “train.csv” datasets. You will need to create an account in the platform beforehand.



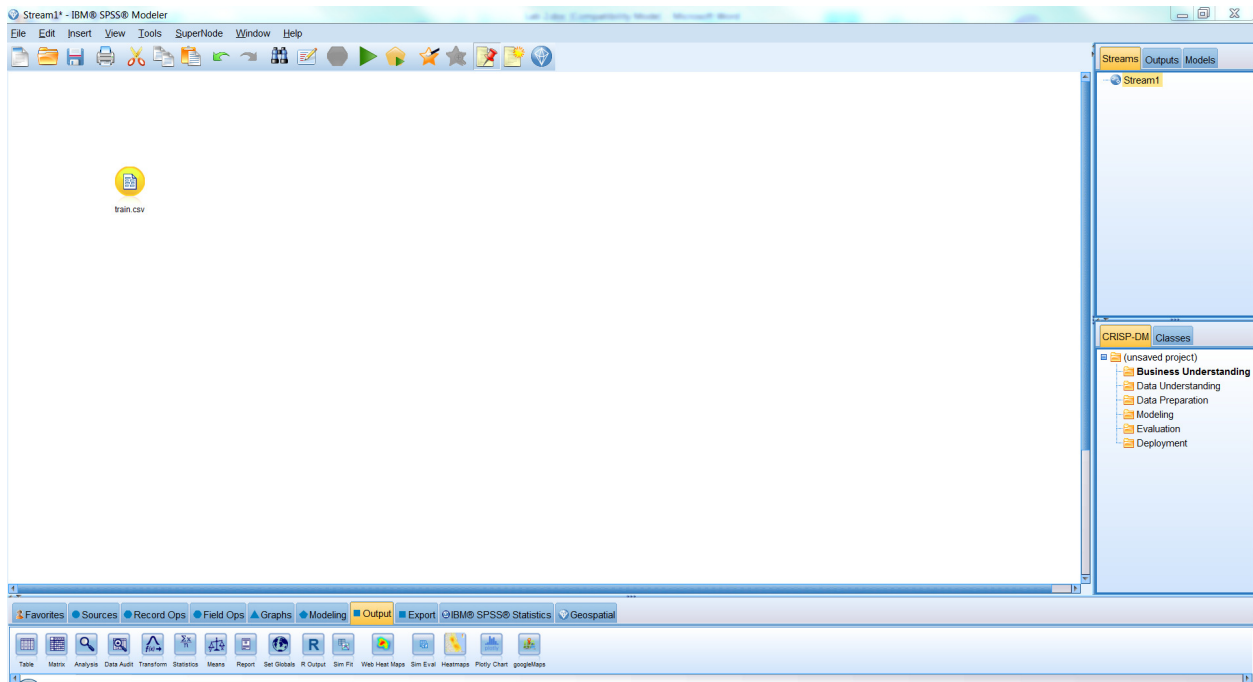
File Name	Available Formats
train	.csv (59.76 kb)
gendermodel	.csv (3.18 kb)
genderclassmodel	.csv (3.18 kb)
test	.csv (27.96 kb)
gendermodel	.py (3.58 kb)
genderclassmodel	.py (5.63 kb)
myfirstforest	.py (3.99 kb)

If you do not wish to create an account in the Kaggle website, you can also get the datasets in the following Github repository:

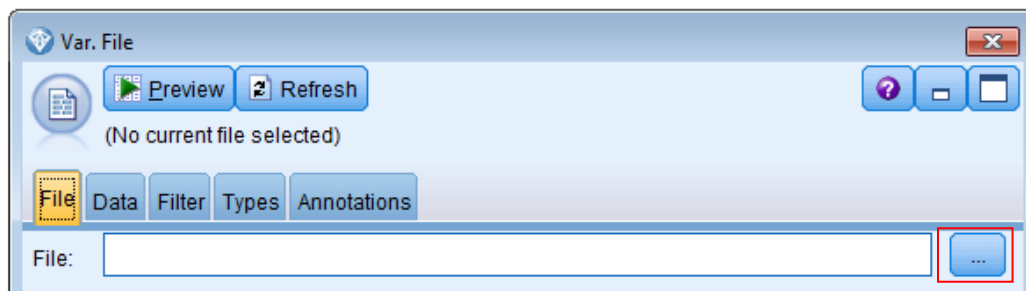
- Train.csv: <https://github.com/aruizga7/TitanicShinyApplication/raw/master/data/train.csv>
- Test.csv: <https://github.com/aruizga7/TitanicShinyApplication/raw/master/data/test.csv>

1.2 Load the data in IBM SPSS Modeler

_1. Start a new stream or remove all the nodes from the current stream. From the **Sources** palette, add as **Var File** node.



_2. Double-click the **Var File** node to open a dialog box. Open the file **train.csv**. Click the button with the ellipsis to select your data file.

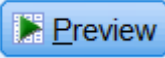


_3. Check that the quotation options are selected as below (to ensure that the fields – such as passenger names - are properly separated).

Quotes

Single quotes: Discard

Double quotes: Pair and discard

Click in the  **Preview** button to have a quick look of your data. This will display the first 10 records of the dataset. If your data matches the screen shot below, click OK to close the node.

Preview from train.csv Node (12 fields, 10 records) #2

File Edit Generate

Table Annotations

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch
1	1	0	3	Braund, Mr. Owen Harris	male	22	1	0
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0
3	3	1	3	Heikkinen, Miss. Laina	female	26	0	0
4	4	1	1	Futelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0
5	5	0	3	Allen, Mr. William Henry	male	35	0	0
6	6	0	3	Moran, Mr. James	male	\$n...	0	0
7	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0
8	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0

OK

_4. To display all the dataset, add a **Table** node from the **Output** palette. Connect the two nodes.



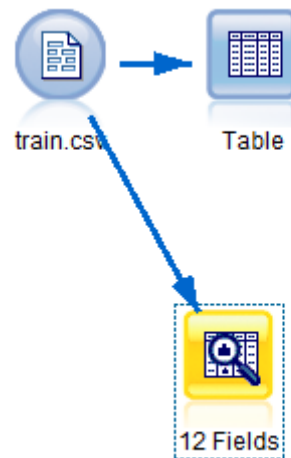
_5. To execute the Stream, select the **Table** node and click the **Run Selection** button in the top menu



. The total number of records in your dataset is 891.

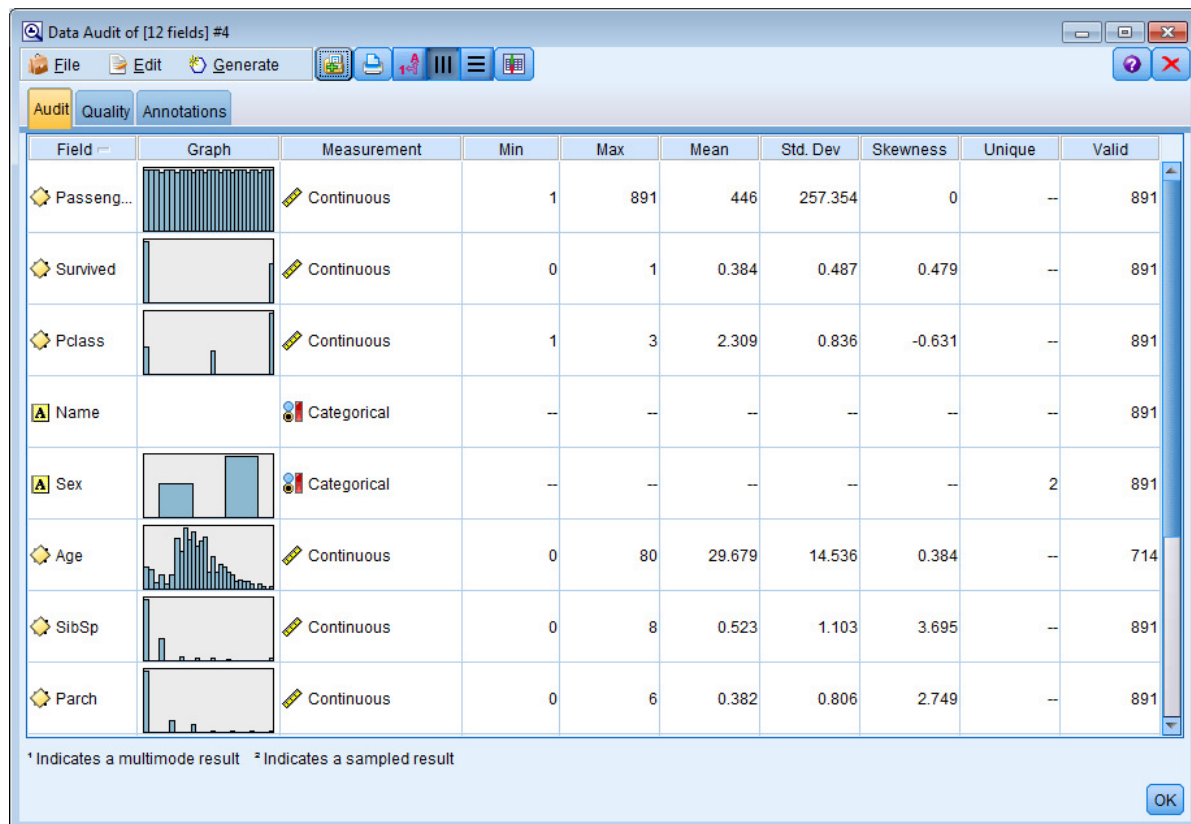
1.3 Exploring a dataset

_1. Add a **Data Audit** node from the **Output** palette and connect it to the **train.csv** node.



_2. Let's run the **Data Audit** with the default options. Right click this node and click Run to execute the stream

The **Data Audit** node provides a comprehensive first look at the data you bring into IBM SPSS Modeler, presented in an easy-to-read matrix that can be sorted and used to generate full-size graphs and a variety of data preparation nodes.



_3. Take some time to explore the output in the **Audit tab**. There are some interesting facts here, like the average age is 29.679 or that there were more male passengers than female (64.76% over 35.24%). Double-click in the Graphs to open them in full-size.

_4. **The Data Audit** node is also useful to check the **Quality** of your dataset. Click in the **Quality Tab**. You will see that there are only 75% of complete fields and 20.54% of Complete Records. This tab displays information about outliers, extremes, and missing values and offers tools for handling these values. We will not be using these tools in the first course but some of these advanced techniques will be presented in the future trainings.

Data quality has to be taken into account to create accurate models – as the saying goes “garbage in, garbage out.”

Data Audit of [12 fields] #10

File Edit Generate

Audit **Quality** Annotations

Complete fields (%): 75% Complete records (%): 20.54%

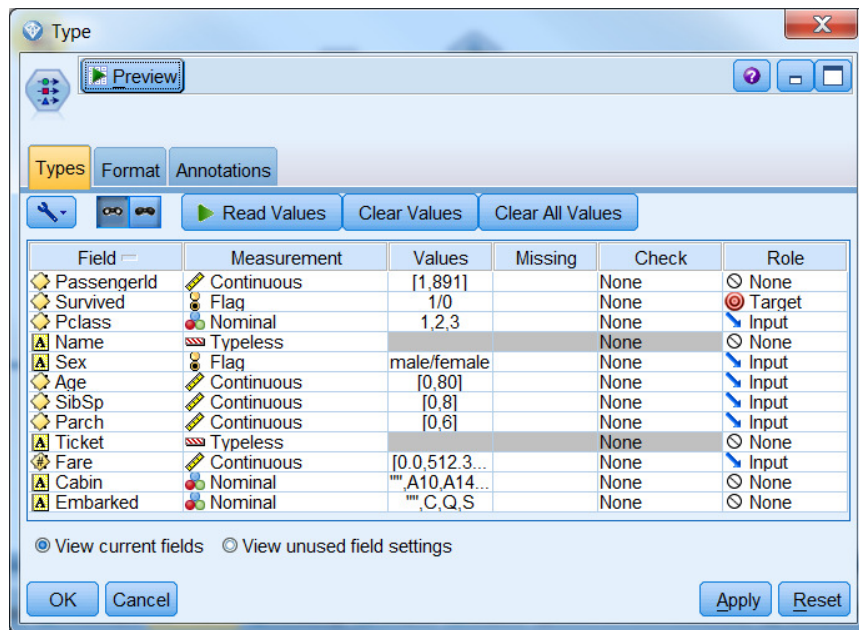
Field	Measurement	Outliers	Extremes
PassengerId	Continuous	0	0 None
Survived	Continuous	0	0 None
Pclass	Continuous	0	0 None
Name	Categorical	--	--
Sex	Categorical	--	--
Age	Continuous	2	0 None
SibSp	Continuous	23	7 None
Parch	Continuous	9	6 None
Ticket	Categorical	--	--
Fare	Continuous	17	3 None
Cabin	Categorical	--	--
Embarked	Categorical	--	--

1.4 Preparing the data for modeling

_1. Add a **Type** node from the **Field Ops** palette and connect it to the **train.csv** node.



_2. Let's open it up and explore by double clicking into it. The **Type** node allows us to understand the attributes (variables) in our data set, review their classes (nominal vs ordinal vs continuous), preview their values and assign roles (Target vs Input vs None).



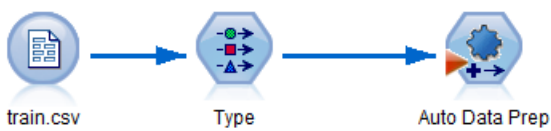
_3. Review the classes (measurements) for our variables and make sure they make sense. Now let's assign the Survived field to a role of **Target** (as this is what we will be predicting) and assign role of **None** to fields PassengerID, Name, Ticket, Cabin and Embarked –as they will not be adding value to us in the analysis (see screenshot above).

_4. Let's preview the data. As we saw earlier, we have some missing values, which we will need to take care of before the modeling stage. Click OK on the data preview, and the main Type node screen to close.

Preview from Type Node (12 fields, 10 records)

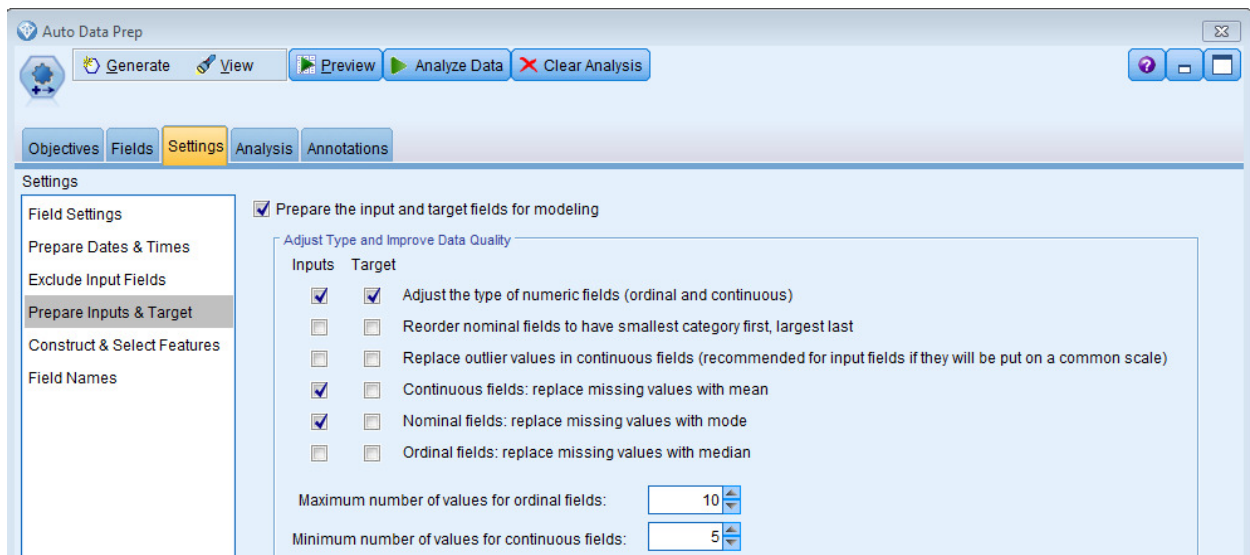
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Par
1	1	0	3	Braund, Mr. Owen Harris	male	22	1	
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	
3	3	1	3	Heikkinen, Miss. Laina	female	26	0	
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	
5	5	0	3	Allen, Mr. William Henry	male	35	0	
6	6	0	3	Moran, Mr. James	male	\$null	0	
7	7	0	1	McCarthy, Mr. Timothy J	male	54	0	
8	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	

_5. Add an **Auto Data Prep** node from the **Field Ops** palette and connect it to the **Type** node.



_6. Double click into the Data Prep node to explore the options. We will not be using all of the advanced settings in this lesson and will return to more complex techniques in further classes. In the **Objectives** tab we keep the 'Balance for speed and accuracy' as our objective.

_7. Click into the **Settings** tab. This is where we can preprocess the fields in our data set. As we saw in the **Data Audit** node, we have some missing values. So let's click into **Prepare Inputs & Target**.



_8. Make sure that we have the boxes checked for replacing missing values with mean for continuous fields - in our data, there are missing values in the Age field (which is continuous). Uncheck the box to reorder nominal field to have the smallest category first.

_9. We can also normalize our continuous variables using several techniques (such as z-score or min max transformations). This technique transforms the data to give attributes equal weight, which is particularly useful for certain classification and clustering algorithms, which we will not be getting into in this part of the course. So let's make sure to uncheck the box 'put all continuous input fields on a common scale.'

Maximum number of values for ordinal fields:

Minimum number of values for continuous fields:

Outlier cutoff value: (standard deviations)

Method for replacing outliers: ☒ Replace with cutoff value ☐ Delete value

Transform Continuous Field

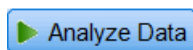
☒ Put all continuous input fields on a common scale (highly recommended if feature construction will be performed)

Rescaling method: Final mean: Final standard deviation:

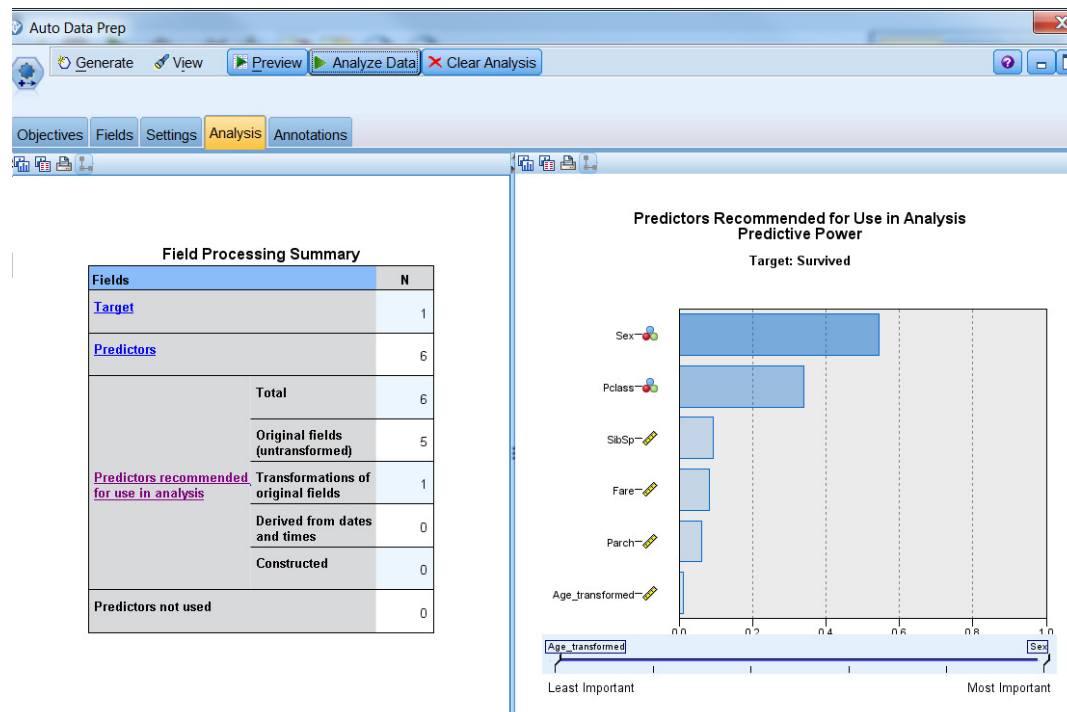
☐ Rescale a continuous target with a Box-Cox transformation to reduce skew

Final mean: Final standard deviation:

_10. Now we are ready to prepare our data and we click **Analyze Data**.



_11. Click in the Analysis tab in the node. The software shows us the summary of the data processing step and recommends fields to be used in the analysis.



_12. After we analyze the data, click OK

_13. The red triangle changes to a blue checkmark and we are now ready for the next step in the process.



Summary

Congratulations! Your dataset is ready to create a predictive model!

In this lab you learned how to load a dataset file in IBM SPSS Modeler and how to perform exploratory analysis using the Data Audit node. There are many other advanced tools available to explore your data.

We also learned how to do simple data preparation. Remember that the process of Data Mining is iterative, and if the results of modeling are not satisfactory you might have to come back and prepare your data differently to get more accurate results.