Alpine Miner Product Training

REVISION HISTORY			
NUMBER	DATE	DESCRIPTION	NAME

Contents

Chapter 1

Overview

In the set of activities described in this document we are going to explore a subset of Alpine Data Miner's rich feature set. We will use Alpine Data Miner to create two different prediction models for a moderately large data set. We will then evaluate these prediction models and compare them using a graph called an ROC.

The data set we will use consists of details on airline arrivals and departures at Chicago's O'Hare International Airport in the year 2008. We will use this data setto create two predictive model one using Naive Bayes and the other using Logistic Regression to predict the probability of a flight being delayed.

Before we set about this task we need to be able to reliably execute certain elementary atomic tasks that will be needed repeatedly.

Note
I need formal names for Alpine, Miner, proper set of logos etc.

1.1 Elementary Tasks

- Log in to Alpine Miner
- Create a connection
- Create a Flow
- Set properties on a module
- Create a Data source by importing from a CSV file
- Join two data sources
- Run a flow

Note
Note to reviewers. Should the above list of tasks be here, or in an appendix in this document or in a separate document called "Intro. . . ." or something else?

Note
Somewhat off topic but v. important. Currently the documentation that comes with the Alpine download does not mention the username and password needed to create a connection. This is a huge showstopper and after downloading one simply can't do anything useful besides running canned flows. The download doc needs to be fixed.

Chapter 2

Data Ingest and Data Exploration

In this stage we will import our data into an Alpine Data Miner DataSource. Then we will do an initial exploration and simple cleanup of the data in preparation for modeling. This stage will introduce us to some of the visualization capabilities

2.1 Import data from file ORD_2008

- Expand the zip file
- Create an Alpine Miner DataSource by importing this file (see Elementary Tasks)

Note
Figure with properties dialog for Import File

When done you should see the following icon in your workspace.

Note
Figure of DataSource icon

2.2 Set Up Initial flow

Once we have created a DataSource we will use it to do an intial exploration of the data. The very first thing we do, something that is almost always the first step, is running Summary Statistics on teh data. The Summary Statistics module in Alpine Data Miner does an inventory and sanity check of the data provided by the Data Source. It then produces a tabular report with the results of it's efforts.

2.2.1 Summary Statistics

We use the Summary Statistics module operating on data from the Data Source.

Create a flow with the following schematic (see Appendix for schematic syntax)

```
ORD 2008 |------> Summary Statistics
```

Note
Figure for flow

Then run the flow.

Once the messages show that the flow has completed running, click on the Summary Statistics module. You will see the results as

Note
Figure with Summary Statistic run output

Placeholder

Note
Question to reviewers - how much to discuss about the content of the output. What is important to discuss?

2.2.2 Null Value Replacement

Summary statistics output will tell us how many null values there were and where. We need to replace null values with some concrete value such as 0 for numeric columns and ” for character columns. The Null Value Replacement module does this en masse over the whole Data Source.

We set the values in the dialog as follows

Note
Figure with dialog for setting replacement values

Schematic for flow.

Now we run the flow again and when the flow has completed the values will have been replaced.

We are now ready to do some exploratory visualization.

2.2.3 Box plot

Schematic for flow.

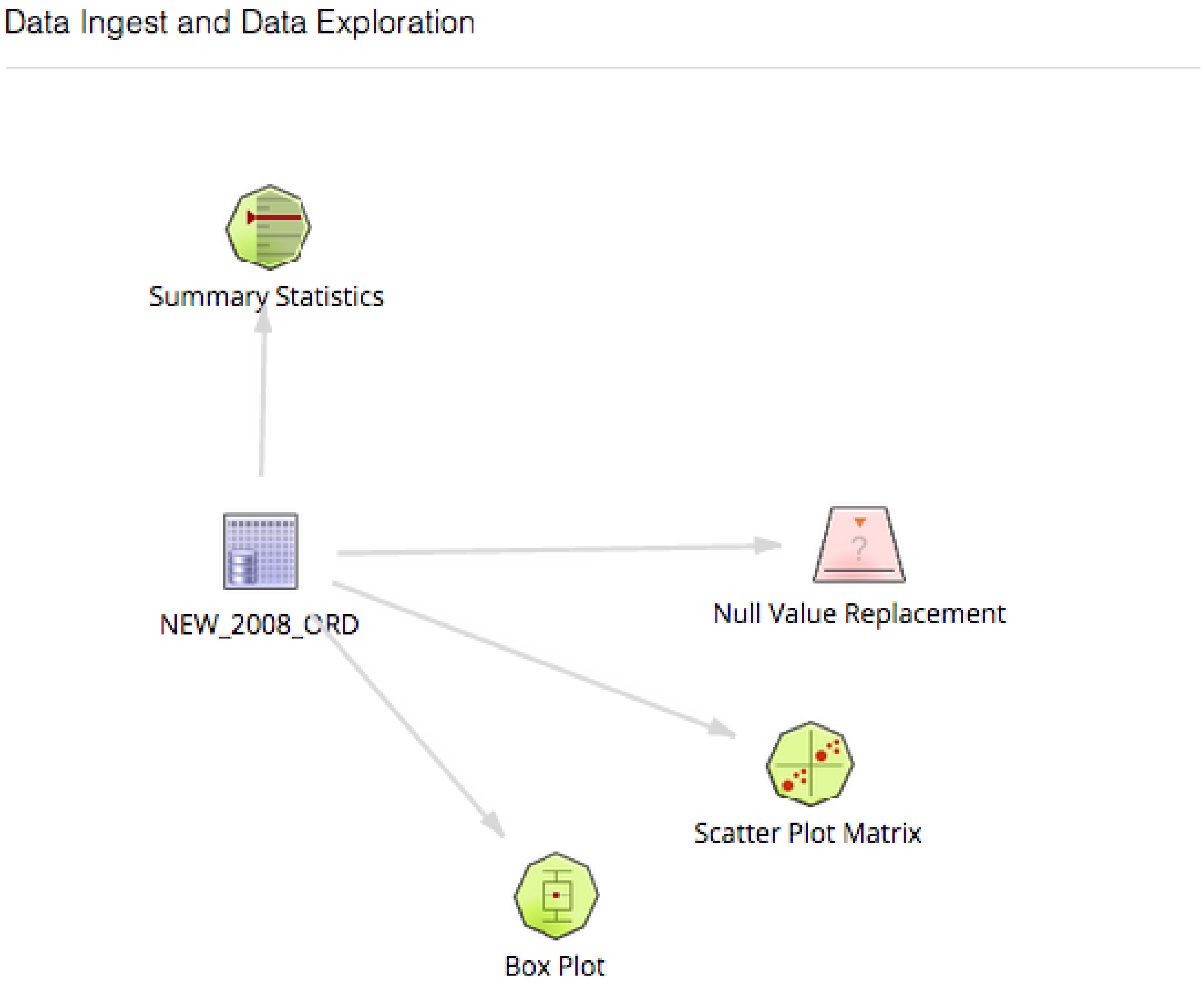
Note
Need property settings for plot

2.2.4 Scatterplot matrix

Schematic for flow.

Note
Need property settings for plot

When we are done we should have a flow in our workspace that looks like this



2.3 Import data from Carriers file

Similar to the ORD_2008 import but for carriers.txt

Note
Where do we get the carriers.txt file? For that matter where do we get the zipped ORD_2008 file?

Chapter 3

Model Creation

3.1 Variable selection

Schematic for flow.

Note
Figure with Variable Selection flow What variables do we pick here?

3.2 Random Sampling

Create two data sets by random sampling from original data.

Schematic for flow.

3.3 Training Set

Note
Do we explain what Training Set is?

Schematic for flow.

3.4 Validation Set

Note
Do we explain what Validation Set is?

Schematic for flow.

3.4.1 Logistic Regression

Note
Do we explain what Logistic Regression is?

Schematic for flow.

3.4.2 Naive Bayes

Note
Do we explain what Naive Bayes is?

Schematic for flow.

Chapter 4

Model Evaluation

4.1 ROC

Schematic for flow.

NB, Logreg, Validation Set as inputs

4.2 LogReg Predictor

Schematic for flow.

Variable Selection, Logreg as inputs

Chapter 5

Appendix A (Schematic syntax)

- A Flow

```
SourceModuleName |---> TargetModuleName
e.g. ORD2008 |---> Summary Statistics
```