Alpine 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's rich feature set. We will use Alpine to explore a moderately large data set using descriptive statistics. Then we will create two different prediction models for this data set, using a question with a yes/no outcome. We will then evaluate these prediction models and compare them using a graph called an ROC.

The data set consists of details on airline arrivals and departures at Chicago's O'Hare International Airport in the year 2008. We will use this data set to create a predictive model one using Naive Bayes and another using Logistic Regression to predict the probability of a flight being delayed or alternately to answer the question "Will this flight be delayed?"

Before we set about this task we need to be able to reliably execute certain elementary atomic tasks that will be needed repeatedly. These tasks are identified in Appendix A. Please make sure you are familiar with these operations especially Alpine account name, password, and datasource properties for the datasource you will use.

Chapter 2

Data Exploration

In this stage we will import our data into an Alpine 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 of Alpine.

2.1   Set Up Initial flow

We will set up a flow with multiple visualizations incrementally, one module at a time. When we are done we should have a flow in our workspace that looks like this.

Data Ingest and Data Exploration
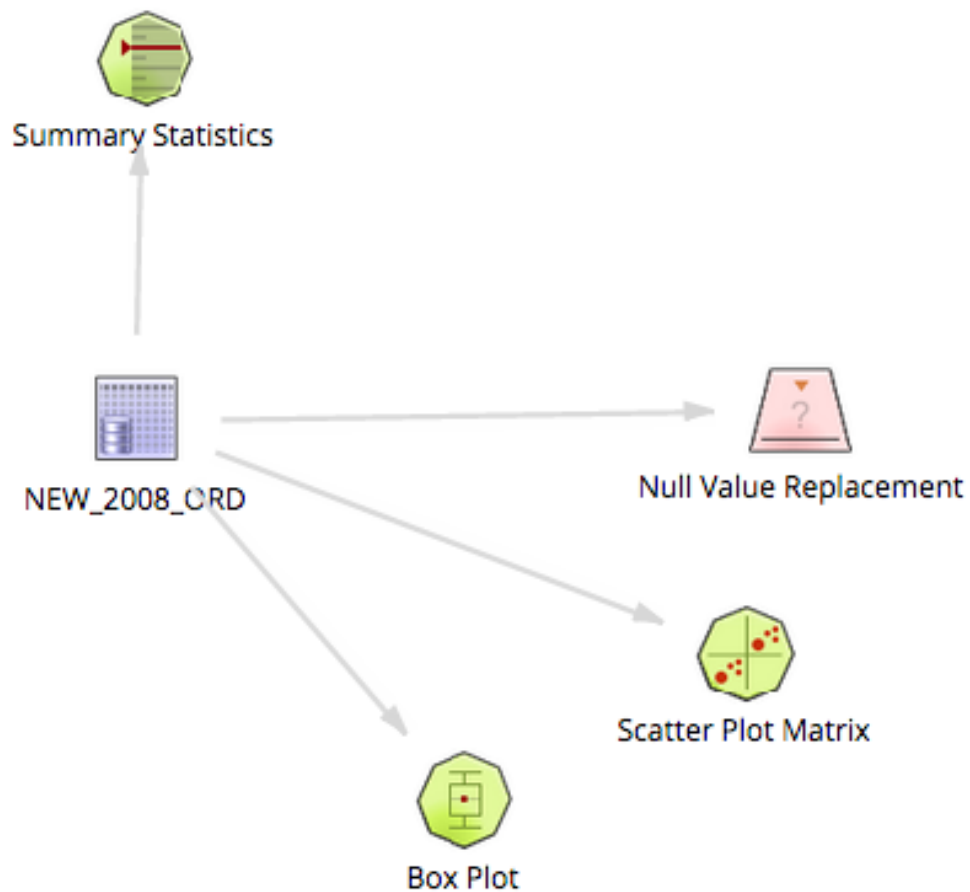
We start by creating a Data Source.

2.2 Import data from csv file

- You should have with you a csv file called 2008_ORD.csv or a zipped version of this file.

- Expand the zip file if necessary.

- Create an Alpine Data Source by importing this file (see Appendix A Elementary Tasks)

- Once you are done you should be able to see a data source called ORD_2008 under ConnectionDemo.

2.2.1 Summary Statistics

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

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

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

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

Note that this requires you to drag and drop the modules for the two end points onto the workspace first and then make a connection between them as per the schematic.

Here just for clarification we will show the association between the schematic and the actual image in the workspace. The schematic is above and the corresponding flow diagram is below.

Flow for Summary Statistics

Henceforth we'll just use the schematic for brevity and show the state of the full flow only at a few intermediate checkpoints.

Now we will need to run the flow.

For this click on the "RUN" item on the top right of the menu bar.

Now you will see status messages at the bottom in a console.

RESULTS - Status

[10:43:44]	ORD_2008 started running......
[10:43:44]	ORD_2008 finished
[10:43:44]	Summary Statistics started running......

Once the messages show that the flow has completed running,

RESULTS - Status

[10:43:44]	ORD_2008 started running......
[10:43:44]	ORD_2008 finished
[10:43:44]	Summary Statistics started running......
[10:48:31]	Summary Statistics finished
[10:48:31]	Analytic Flow finished
	Click on individual operators to see their output.

click on the Summary Statistics module icon in the flow.

You will see the results that look similar to the following

name	data type	count	unique	null value	empty	zero
taxiout	integer	333331	229	0	N/A	0
cancelled	integer	333331	1	0	N/A	333331
cancellationcode	integer	0	0	333331	N/A	0
diverted	integer	333331	1	0	N/A	333331
carrierdelay	integer	95961	522	237370	N/A	73309
weatherdelay	integer	95961	312	237370	N/A	91535
nasdelay	integer	95961	455	237370	N/A	26818
securitydelay	integer	95961	53	237370	N/A	95842
lateaircraftdelay	integer	95961	396	237370	N/A	51116

2.2.2 Null Value Replacement

The output of the Summary Statistics run 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 add a segment to our existing flow. The new segment has the following schematic

```
ORD 2008 |------> Null Value Replacement
```

Double click on the Null Value Replacement module and see a dialog box for configuration. We set the values in the dialog as follows

Null Value Replacement Configuration

☑	Select Columns	Value
☑	year	0
☑	month	0
☑	dayofmonth	0
☑	dayofweek	0
☑	deptime	0
☑	crsdeptime	0
☑	arrtime	0
☑	crsarrtime	0
☑	uniquecarrier	"
☑	flightnum	0
☑	tailnum	"
☑	actualelapsedtime	0
☑	crselapsedtime	0

CANCEL   OK

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 Step run vs Run

When running complicated flows with large data, it is time effective to run each step of the flow separately once we know that the previous steps have been successful. Otherwise all the previous steps will be run redundantly and with large data sets this will be wasteful of resources.

To run a single module in the flow (assuming all the previous inputs have been computed and are ready) alt-click on the module and pick the "Step run" option.

image:images/nullvalsteprun.png

2.2.4 Box plot

We now explore arrival delay on different days of the week and their distribution using the Box Plot module.

Add the following segment to the flow

```
ORD 2008 |------> Box Plot
```

Set the Box Plot configuration by double-clicking on the Box Plot module.

image:images/boxplotcreate.png

Step run the Box Plot module. You will see a result similar to the following.

image:images/boxplotresults.png

2.2.5 Scatterplot matrix

We want to explore correlations between variables to see which ones we might use in our predictive model.

Add the following segment to the flow.

```
ORD 2008 |------> Scatter Plot Matrix
```

Set the Scatter Plot Matrix configuration as follows

Double click on the Scatter Plot Matrix module icon and select the following columns,

dayofweek, arrdelay, depdelay, distance.

Then save.

Now step run the Scatter Plot Matrix module. This may take a while. Look at the console messages to tell you when you will be done.

When done you should see a diagram similar to this.

image:images/scatterplot.png

To be able to see the diagram fully, click on the arrow at the far right of the console menu. This is an arrow pointing up and to the right and on clicking will expand the window to give you a full view of the Scatter Plot Matrix.

This tells us that depdelay and arrdelay are strongly correlated, i.e. if the plane took off late it is mostl likely to arrive late at the next destination. This is insightful but we don't need data science to tell us this. However this is a very useful training exercise so we continue using this correlation for the predictive model. In reality we would want to predict the delay long before the flight took off late from the previous destination.

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 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

Note
Figure with properties dialog for Import File

- Join two data sources

- Run a flow

Chapter 6

Appendix B (Schematic syntax)

• A Flow

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