Altair Analytics Workbench: EXERCISES

Contents

Introduction

Introduction

All lessons are interactive and must be viewed in sequence such that subsequent lessons become available once previous lessons are complete. Note that some lessons have associated exercises. Data used for exercises is available from the files section, *data.zip* contains all data, output and projects referenced throughout lessons and it is advised to unzip this to a known and accessible location.

Chapter 2: Introduction to Altair Analytics Workbench

Introduction

Contents for this lesson include an introduction, a look at Altair Analytics Workbench, the SAS language and workflow perspectives and common elements. We'll then speak about preferences and help options prior to getting onto to a demonstration and then returning to slides to highlight important points before wrapping up with a summary.

- Introduction
- Two perspectives
- SAS Language
- Workflow
- Common perspective elements
- Preferences
- Help
- Demonstration
- Things to remember

Exercises

1. Open Altair Analytics Workbench and close the Welcome screen.
   - Create a new project called Project_1.
   - From the File Explorer view locate and open the program file: *Ex_1.sas.*
   - Run the import part of the code.

Does the code run correctly?

2. Can the code be modified to run correctly?

3. How many variables and observations are there in the file data_.xlsx?

14 and 18539 respectively.
The data file is empty.
13 and 18538 respectively.
14 and 18538 respectively.

4. What is the median *capital_gain*? Hint: Highlight and run all remaining code. This will generate charts and statistics accessible from the HTML node.

726.2
0
5303281
162104

5. Save the SAS program file to the project folder: *Project_1*. How can the project path be known so that the file can be saved to it? Select all correct options.

By double-clicking the project folder from the Project Explorer view.
Write the path down when a project is created.
Using File > Save As will automatically navigate to the project folder.
Select the project folder from the Project Explorer view. Its path is visible from the Properties view.
Right-click Project_1 and select either Properties or Show in System Explorer.

6. How many variables and observations are in the file?

- Confirm the program has been saved in the Project_1 folder and it has a *.sas* extension.
- Create a Workflow called Workflow_1 in the project folder.
- Ensure the Workflow perspective is active.
- Import the file data_.xlsx into Workflow_1 by either:
    - Using an Excel Import block from the Import group.
    - Locating the file from the File Explorer and dragging it onto the Workflow canvas.
- Do not import the variables: *Customer_ID* and *weight*.
- Rename the resulting dataset on the Workflow canvas to: Filtered.
- Export the dataset, including headers, as a *.wpd* dataset called Filtered to the Project_1 folder.
    - HINT: Right-click the dataset on the Workflow canvas and select the option: Save Dataset.
    - NOTE: This option is only available when using a desktop version of the software. If using a server version of ALTAIR, the ALTAIR Dataset Export block, found in the Export group can be used to accomplish the same results.
- Drag the dataset: Filtered, from Project_1 and drop it onto the Workflow canvas.

14 and 18538 respectively.
12 and 18538 respectively.
14 and 18539 respectively.
12 and 18539 respectively.

7. How can it be known that a block has a comment?

8. What length of time is the cache set to and can it be changed?

30 days, no.
30 days, yes.
20 days, no.
20 days, yes.

9. Can the language of SAS code for the Excel Import block only be accessed, and how?

10. How can Contextual Help be displayed for the Excel Import block?

For the brave!

11. Is it possible to export the Workflow to a SAS language program and run this in the SAS Language perspective error free?
Hint:

- Ensure that only the originally imported dataset is on the canvas.
- Modify its configuration to import all variables.
- Add a Select block from the Data Preparation group and connect to the Workflow dataset.
- Remove the variables *Customer_ID* and *weight.*
- Rename the resulting dataset to *New.*

Chapter 4: Data exploration and profiling

Introduction

Contents for this lesson include focus on the functionality available in Altair Analytics Workbench to explore and profile data.

The Dataset File Viewer and Data Profiler are introduced as well as the Chart Builder block. The use of the SAS Language perspective for data exploration and profiling is also covered.

Exercises

1. Using the Workflow created in the previous exercise, import the file *data_.xlsx*. Open the data with the Dataset File Viewer. Add a filter: AGE GE 40. (NOTE: GE means Greater than or Equal to)
How many observations are greater than or equal to 40?

1181
931
4660
8141

2. Is the number of observations equal to that in the file: *data_.xlsx* with the filter on? What does this mean? Hint: From the Dataset File Viewer and with the filter still in place, export the dataset as a delimited file named *dlm_export*, including headers, to the project folder. HINT: right-click any variable name and choose the option: Export. Drag the resulting file onto the Workflow canvas.

3. What is the average age, minimum education_num? Do any variables have missing values?
Hint: Remove the file dlm_export from the Workflow. Open *data_* with the Data Profiler. The number of variables and observations in the dataset can be ascertained from the Summary View tab.

48.72, 1.Yes, capital_gain.
38.72, 1. Yes, capital_gain.
17, 1, No.
48.72, 1. No.

4. From the Summary View take note of the Type and Classification columns. Type is determined by whether the variable contains numbers or strings. Variables with numbers are Numeric and variables with strings are Character. Classification for both is determined by the number of Distinct Values reaching a threshold. For categorical variables, classification is Discrete if the number of distinct values if >=50 and Categorical otherwise. For Numeric variables, Classification is Continuous f the number of distinct values if >=50 and Categorical otherwise.

Can this threshold be changed and what effect will it have?

Yes, little to no effect.
No, no effect.
Yes, it may change the Classification of some Numeric variables only.
Yes, it may change the Classification of some variables.

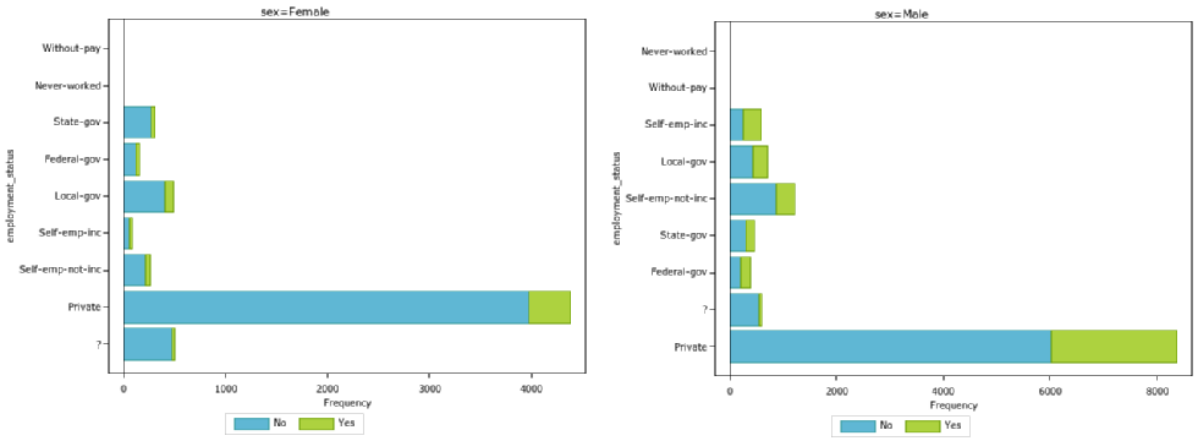
5. View the distribution of the variable D*V.* As can be seen there are approximately 76% *No* and 24% *Yes*. Excluding the variables *weight* and *customer_id*, what are the top two useful variables that could be used to predict *DV*?

Customer_ID and *relationship*
Relationship and *marital_status*
marital status and *capital_gain*
Customer_ID and *capital_gain*

6. Are there any variables that are highly correlated. i.e. where: r >= 0.6?

Yes
No

7. Use the Chart Builder block to create charts for each category of the variable: *sex*. Modify the chart such that it displays the distribution of the variable: *employment_status*, across the categories of the variable: *DV*. Multiple charts should result and resemble the following illustration:



Approximately how many females in the *Private employment_status* category are *Yes*?

500
4000
6000
2200

For the Brave!

7. Use the SAS Language perspective to write a SAS program to access the file *data_.xlsx*. What is the proportion of females in comparison to males in the *Yes* category of *DV*?
HINT: Create a folder called temp in the c:\ directory. Write code to export the file using the SAS code block. From the SAS Language perspective, import the data and analyse it. The following code snippets may come in useful!

```sas
libname out sas7bdat 'c:\temp';

data out.outfile;
set &Input_1;
run;
```

```sas
libname out sas7bdat 'c:\temp';

data in;
set out.outfile;
run;

proc freq data = in;
tables sex*DV /nocol nopercent;
run;
```

Chapter 5: Data preparation and manipulation

Introduction

This lesson includes a focus on data preparation capabilities available from the Altair Analytics Workbench Workflow perspective.

Exercises

1. Create a new Workflow or use an already existing one but delete all blocks. Using the files: *mar_Altair.csv* & *demographics_csv.* Profile the datasets and use them to generate the dataset depicted. Note: *ID* should be unique. Retain all observations from *mar_Altair.csv* and matching observations from *demographics_csv*. How many observations and variables are in the resulting dataset?

	ID	num_products_AVG	num_products_SUM	tot_price_AVG	tot_price_SUM	ID_COUNT	Month_MIN	Marital_Status	Sex	DV	age	Years_Education	hours_worked	Occupation
1	106	4.8	72	190.795333333...	2861.93	15	Mar	Married	Male	0	51	15	25	Other
2	106	4.8	72	190.795333333...	2861.93	15	Mar	Married	Male	0	51	15	25	Other
3	484	4.46666666666667	67	285.564	4283.46	15	Mar	Married	Fe...	0	36	17	33	Other
4	484	4.46666666666667	67	285.564	4283.46	15	Mar	Married	Fe...	0	36	17	33	Other

131 and 14.
140 and 14.

2. Add data from the files: *Jan_Altair.csv* and *Feb_Altair.csv* to the process. The resulting dataset should have 355 observations and 14 variables. Comment on the distribution of *age_3* across the categories of the variable: DV.

NOTE: If this did not succeed the file: *dp_.xlsx* can be used from this point.
Using an appropriate block from the Data Preparation group: create a new variable: *age_3* using the variable: *age*. Create 3 equal height bins. Make sure the variable age is Numeric and Continuous!

Older people are generally in the 1 category of DV.
Older people are generally in the 0 category of DV.
Younger people are generally in the 1 category of DV.
Missing observations are generally in the 1 category of DV.

3. Can the bins created previously be applied to new data? Select all correct answers.

No.
Yes, by coping the code and using it in a SAS code block.
Yes, using the binning code in the SAS Language perspective.
Yes, by generating a Binning Model and using the Score block.

Chapter 7: Introduction to decision trees

Introduction

This lesson will include decision tree algorithms and their common ground.

Exercises

1. Create a new Workflow or use an already existing one but delete all blocks. Import and profile the dataset *data_.xlsx*. Create two partitions with a 70/30 split using the random seed 1000.
Name the partitions *dev* and *test* respectively. Question: How many observations and variables are in there in each partition?

12920 and 5618
12977 and 5561

2. Create a decision tree model using the Dev partition. Use the following settings:

- Select *DV* as the Dependent variable with *Yes* as the Target category.
- Use all other variables as independent variables.
- Set the Treatment for *education_num* to Interval, and *education* to Ordinal.
- Use the Grow Tree option with BRT selected as the Growing algorithm.
- Set Minimum node size(%) to 1%.
- Select the option to Merge missing bin.

How many leaf nodes in the tree and what is the minimum leaf node size?

67 and 131 respectively.
65 and 131 respectively.
60 and 130 respectively.
97 and 37 respectively.

3. Is the tree acceptable and why?

Yes, it captures the *Yes* categories very well.
Yes, it predicts the *Yes* category ok.
No, it does not predict the dependent variable very well.
No, many reasons including questionable leaf nodes and variables.

4. Is there any way to use automated methods to grow an acceptable tree?

Yes, but settings must be set very carefully and some manual pruning may be required.
No, the target category will not be predicted well.
Yes, just select the correct algorithm for the data.
No, leaf node sizes will always be too small.

5. Ensure the following setting are applied:

- Set the Default bin count to 5 from Binning Preferences.
- Set the Minimum node size(%): 1% from Growth Preferences.
- Grow the tree 1 level using Grow tree 1 Level (BRT).

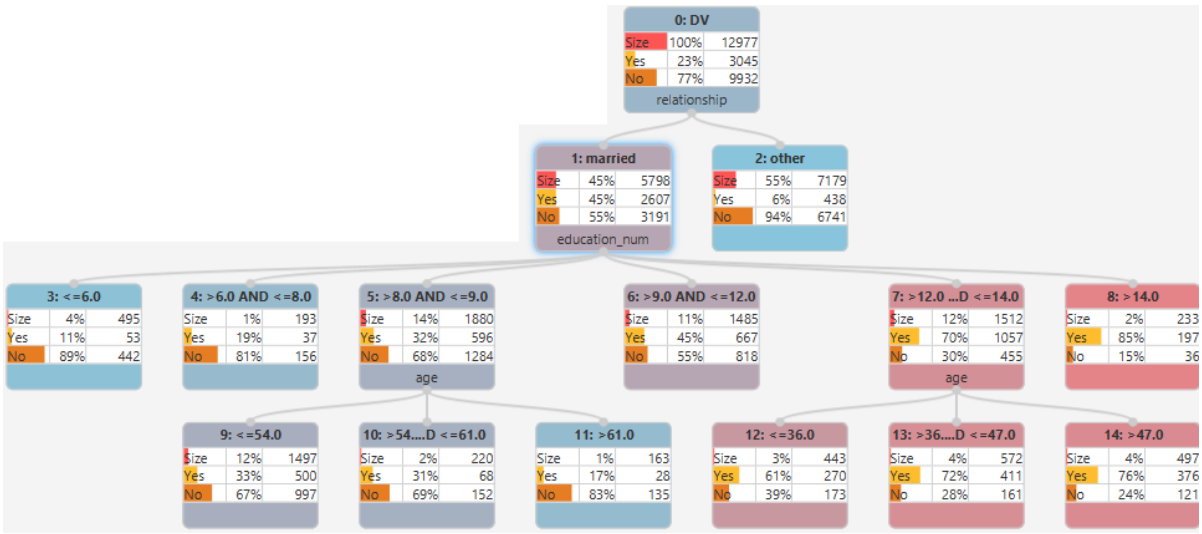
Is the split acceptable, why?

Yes. It captures the Yes category very well.
Yes. The split variable selected has 6 nodes so can be grown further.
No. A different variable is needed.
No. For the selected split variable, created nodes have decimal cut-points and there are nodes where all observations are in the same category.

6. Regrow the tree from the root node using Optimal Split by Entropy Variance. Is the split acceptable?

Yes
No

7. Merge appropriate categories of the variable: *relationship*, such that there are two nodes: *married* & *other*. Grow the tree further using available methods and reproduce the tree illustrated.
NOTE: This is also available from the Workflow: *tree_*, which is located in the eLearning data folder.



What is the approximate classification rate for the *Yes* category?

95%
82%
41%

Chapter 9: Evaluating and validating a decision tree model

Introduction

Contents for this lesson focus on aspects associated with evaluating and validating a decision tree model including set up using Workflow blocks.

Exercises

1. Use the Workflow and decision tree created in the previous exercise or, open the workflow: tree_. This is included with course data. Note that the import block used will need to be configured correctly. Score the *Dev* partition with the decision tree model and add an Analyze Models block to evaluate the *Yes* category of *DV*. Once complete, open the Model Analysis Report to access results. What do results look like?

Good. Statistics and charts are acceptable.
Not acceptable. Overall accuracy is misleading, F1 gives a more rounded picture at approx. 52%.
Good. The Gains Chart shows that the target category is captured well.
The model is not acceptable, as the Lift Chart is not monotonically decreasing.

2. Duplicate the decision tree block and connect the Test partition. Compare the structure of the tree across partitions. Does the tree validate structurally?

No, there is not much replication of leaf nodes with comparable DV distributions.
Yes, for the most part, however there is a little instability in one or two leaf nodes.

3. Score the Test partition. Connect it to the Analyse Models block, configure it appropriately and once complete, access results from the Model Analysis Report. Does the model validate well?

No. The model charts and statistics do not compare across partitions.
No. The Yes category Accuracy and F1 values are too low.
Yes. The model charts track well and models statistics are comparable across partitions.
Yes, as Accuracy and the C-Statistic are high on the Test partition.

Chapter 10: Challenging the model

Introduction

This lesson focuses on using Altair Analytics Workbench Workflow modelling capabilities to challenge a model.

Exercises

1. Use the Workflow and decision tree created in the previous exercise or open the Workflow: tree_2. This is included with course data. How many variables are included in the final model? Note that the import block may need to be configured correctly.

Using the *dev* partition; model the *Yes* category of *DV* using logistic regression. Include all variables except *Customer_ID*, *weight*, *capital_gain*, *marital_status*, *education* and *occupation*. Make sure variables are selected as Class appropriately. Use the Stepwise method to ensure only significant predictors are included in the final model.

5
6
7

2. Score the *dev* and *test* partitions with the logistic regression model. Add a separate Analyse Models block and connect both scored partitions. Ensure the block is configured appropriately. Once complete, open the Model Analysis Report. Does the logistic regression model validate?

No, there is not much stability across statistics or charts.
Yes, statistics and charts are comparable across scored partitions.

3. Ensure scored partitions are renamed for ease of identification. Connect all partitions scored with the decision tree and logistic regression models to the same Analyse Models block. Ensure the block is configured appropriately. Once complete, open the Model Analysis Report. Is the logistic regression model a better choice that the decision tree?

From statistics and charts, yes.
From statistics and charts, no.

4. Why is the logistic regression a better model for this data than the decision tree?

Chapter 11: Model deployment

Introduction

This lesson includes an introduction to model deployment and highlights Altair SLC Analytics Hub as a means to easily deploy program code for on-demand, scheduled and real time applications. The scored data can easily be exported to a desired destination format using an appropriate export block available from the Export group.

Exercises

1. Use the Workflow and decision tree created in the previous exercise or open the Workflow: *tree_2*. Note that the import block may need to be configured correctly. Score the dataset *score.csv* with the decision tree. What is the range of *P_Yes* values?

2. The variable *P_Yes* reflects the propensity to default. Using the scored development partition, create three equally sized groups based on the variable *P_Yes.*
Add a strategy such that:

The group with lowest *P_Yes* values is assigned a value: Accept.
The group with intermediate *P_Yes* values is assigned a value: Manually assess.
The group with highest *P_Yes* values is assigned a value: Reject.

Apply the strategy to the imported dataset score.csv
Name the variable *Strategy.*

How many observations have been assigned the value: Manually assess?

3. Score the dataset: *score_.sas7bdat* using a SAS Code block and the decision tree. What is the range of values of *P_Yes*?