**Lab for Chapter 3:**

**Regression Techniques Part II: Logistic Regression**

Contents

[Exercise 0: Overview **3**](#_bookmark0)

[Exercise 1: Create the Model **4**](#_bookmark1)

[Step 1: Get Data 4](#_bookmark2)

[Step 2: Preprocess Data 5](#_bookmark3)

[Exercise 2: Train the Model **8**](#_bookmark4)

[Exercise 3: Test the Model **9**](#_bookmark5)

[Step 1: Predict Over New Data 9](#_bookmark6)

# **Exercise 0: Overview**

Identify what kind of machine learning problem this is. **“Predict if a customer will buy a bike or not?”**

This exercise will create a Classification model to predict if a customer will buy a bike or not. The simplest process to create any predictive model is

**Create a Model**

1. Get data
2. Preprocess data
3. Define features

**Train the Model**

1. Choose and apply a Machine Learning algorithm   
     
   **Test the Model**
2. Predict over new data

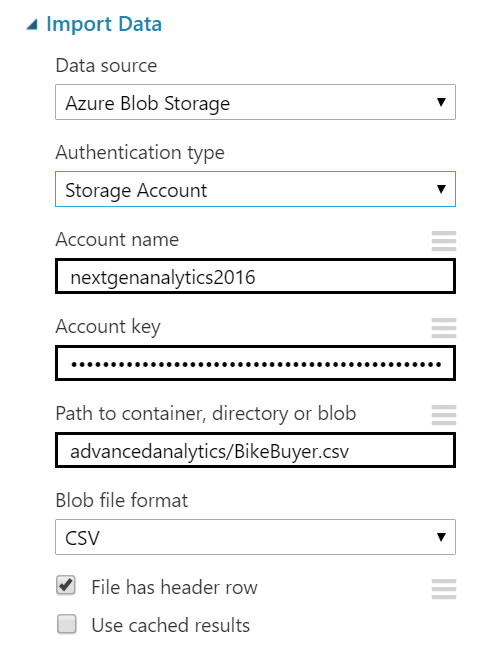
# **Exercise 1: Create Model**

**Step 1: Get Data**

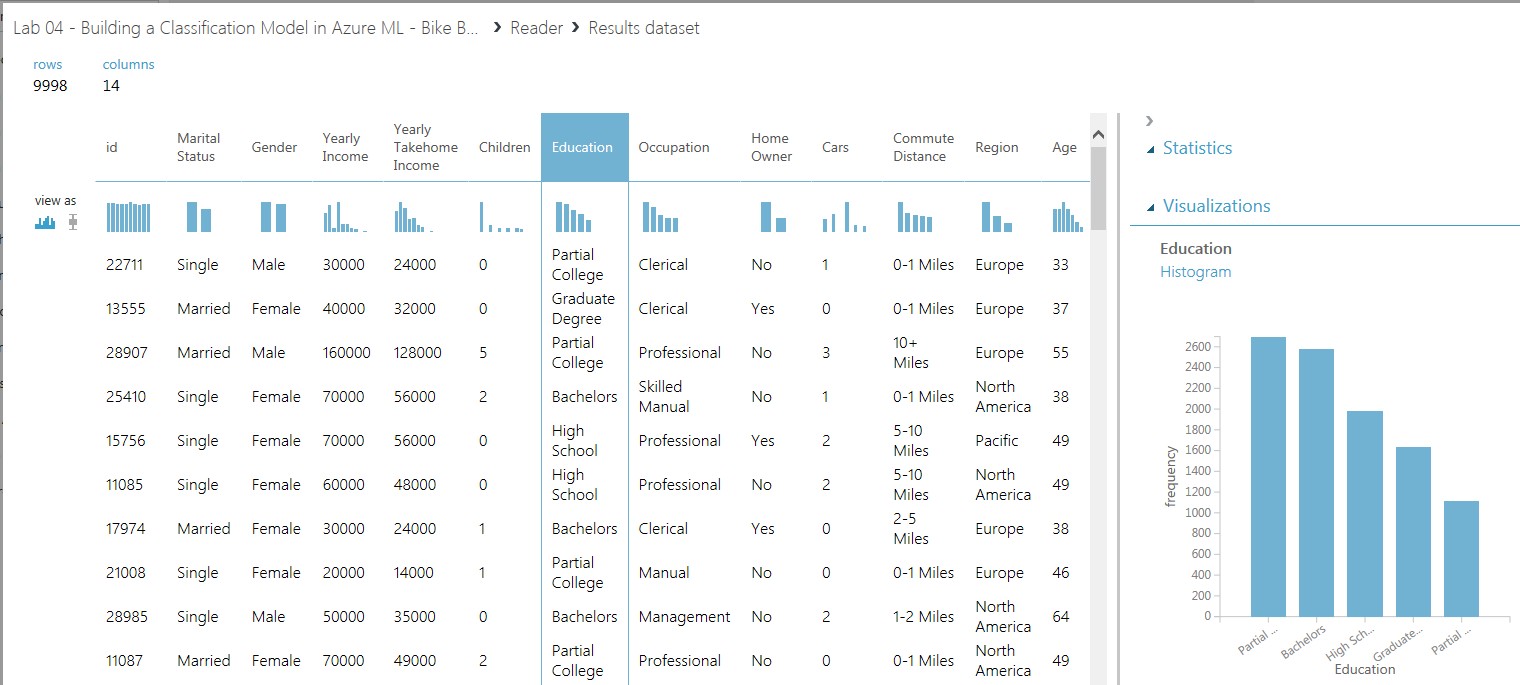
* 1. Create a new experiment and rename appropriately (Lab 04 - Building a Classification Model in Azure ML – Bike Buyer)
  2. **In Previous lab** exercise we used pre-loaded saved dataset. For this exercise, we will get data from Azure Blob Storage.
  3. Drag an “**Import Data**” module onto the experiment canvas and configure it with sample blob storage account we have set up for you.
     1. **Data Source**: Azure Blob Storage
     2. **Authentication type**: Storage Account
     3. **Account name**: analyticsbootcamposlo
     4. **Account key:**

WxHhL/+EhKva80Y3x25Id4gYndW0H6hzo1ChikRzaD21rtf3Dy2JIy4mNcSWw7ohqqDA1UPB29OTfoG7ogP2+w==

* + 1. **Path to container, directory or blob**: advancedanalytics/BikeBuyer.csv
    2. **Blob file format**: CSV
    3. **File has header row**: Checked

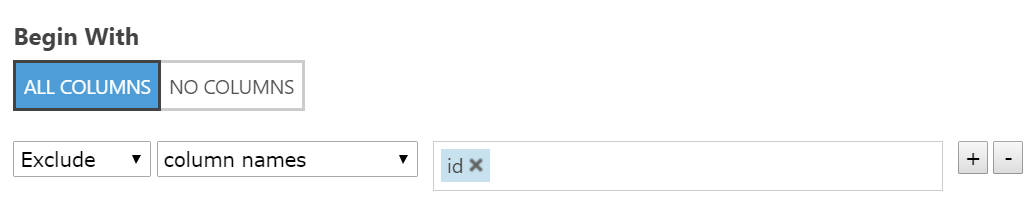


* 1. Run the experiment.
  2. By clicking the output port of the Import Data module, select ‘Visualize’ to see what we just downloaded from Azure and see if there are any issues in the dataset.

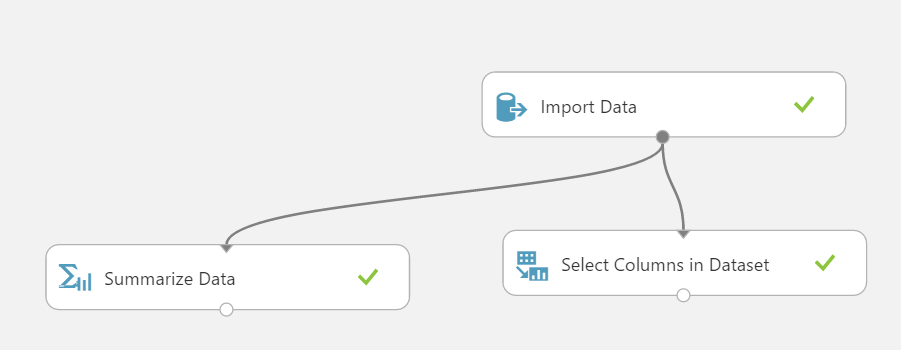


**Step 2: Preprocess Data**

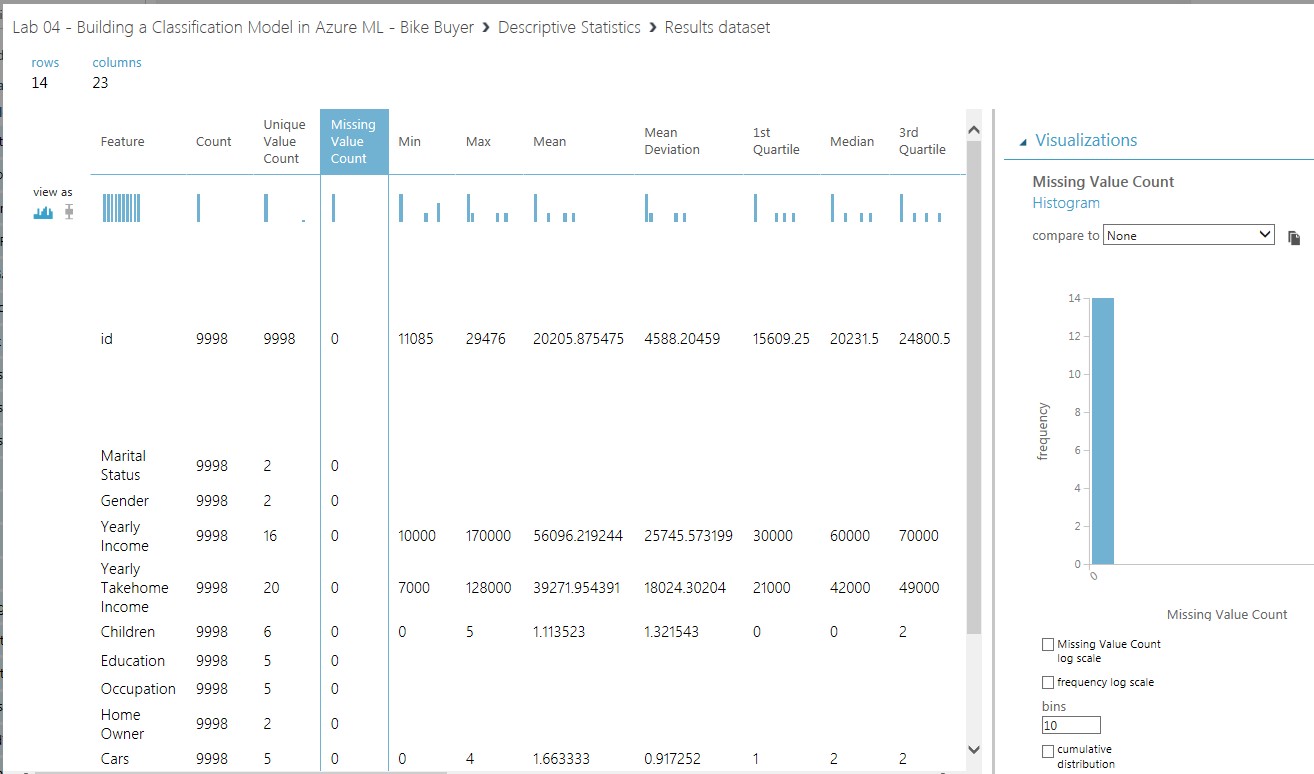
1. In many cases raw data needs to be processed before being sent as input to train a model.
2. From the previous analysis we see that **id** field will not be of any help in analysis and can be ignored.
3. Drag “**Select Columns**” module and link it to the Import Data module above.



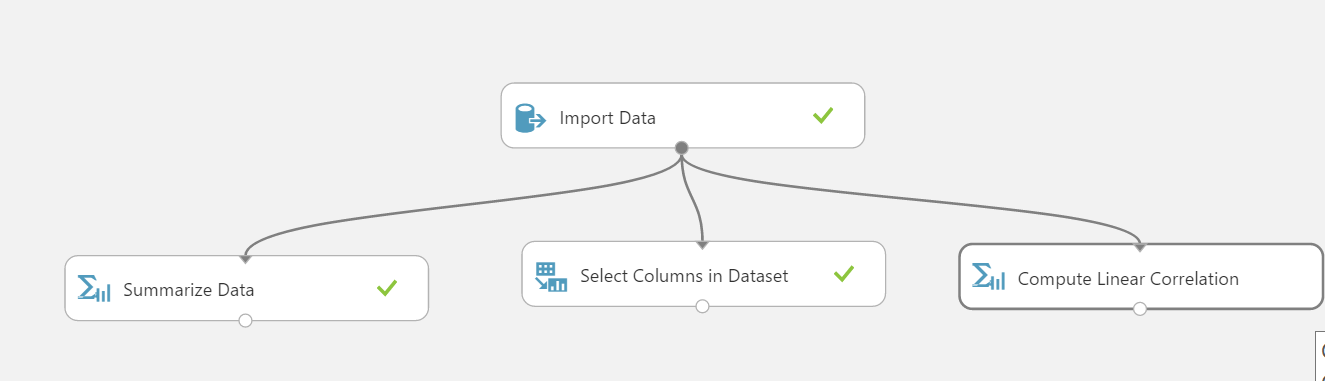
1. Select the **id** column as the one to be excluded.
2. Let’s use “Summarize Data” module to see and analyze key metrics about source data.
3. Drag “**Summarize Data**” module onto the canvas and link the output of Import Data module as input to it.



1. Run the experiment and right click on output of Summarize Data module and select visualize to see the key metrics as shown below.

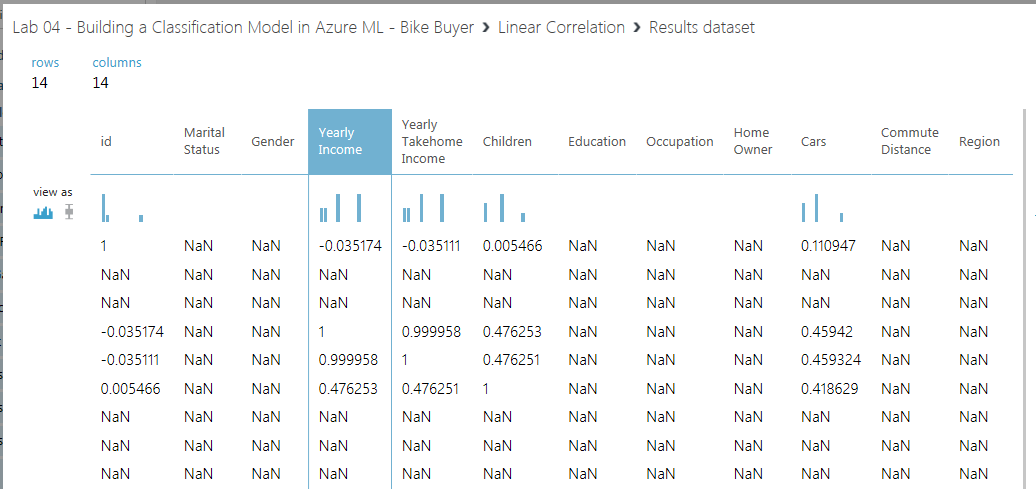


1. You can see that there are no missing values in any of the columns. Also, other metrics can be seen.
2. Drag “**Compute Linear Correlation**” module and link the output of Import Data module as input to it.
3. The “Compute Linear Correlation” module is also useful for computing the correlation of variables in your dataset.
4. Run the experiment and click the output of the Compute Linear Correlation module, and then select Visualize. The tool displays a correlation matrix and in this matrix, you can see the pairwise correlation of all variables in the dataset.

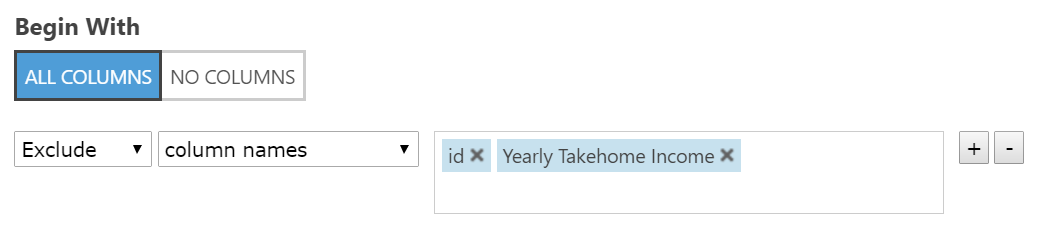


1. Compute Linear Correlation works only with continuous variables and will be showing ‘NaN’ for categorical variables. Any columns which are highly co-related can be ignored.

If you notice, **Yearly Income** and **Yearly Takehome** Income are very highly related and so one of it can be ignored.



1. Edit module “Select Columns” and add “**Yearly Takehome Income**” column to already ignored **Id** column as shown below,



# **Exercise 2: Train the Model**

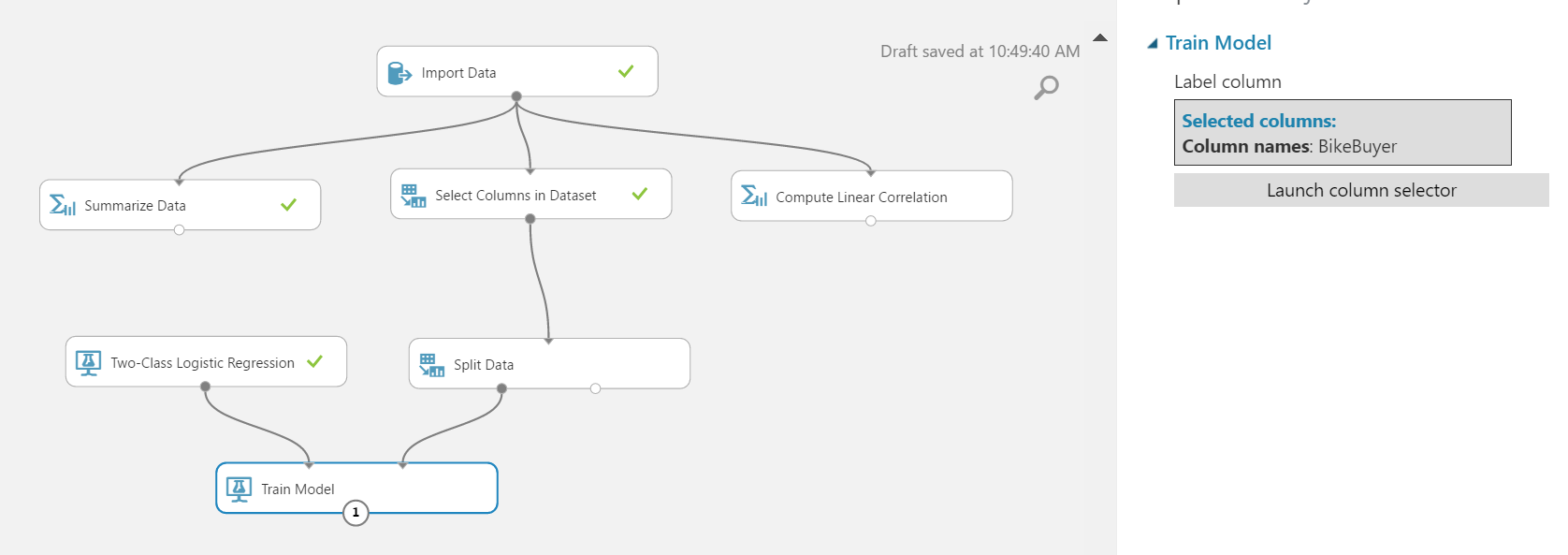
**Step 1: Choose and Apply Machine Learning Algorithms**

When constructing a predictive model, you first need to train the model, and then validate that the model is effective. In this experiment, you will build a Classification model.

In this experiment, you will train a classification model and use it to predict the likelihood of customer buying a bike.

Specifically, you will train a simple **Two-Class Logistic Regression** model. After the model has been trained, you will use some of the modules available in Machine Learning Studio to validate the model.

1. Split the data into training and testing sets: Select and drag the “**Split**” module to the experiment canvas and connect it to the output of the “Select Columns” module.
2. In the properties pane, select ‘Splitting mode’ to ‘Split Rows’. Next, set ‘Fraction of rows in the first output dataset’ to **0.8**. This way, you will use 80% of the data to train the model and hold back 20% for testing.
3. Run the experiment. This allows the Select Columns and Split modules to pass along column definitions to the modules you will be adding next.
4. To select the learning algorithm, expand the Machine Learning category in the module palette to the left of the canvas and then expand ‘Initialize Model’. This displays several categories of modules that can be used to initialize a learning algorithm.
5. For this example experiment, select the “**Two-Class Logistic Regression**” module under the ‘Classification’ category and drag it to the experiment canvas.
6. Also, find and drag the “**Train Model**” module to the experiment. Click ‘Launch column selector’ and select the ‘**BikeBuyer’** column. This is the feature that your model is going to predict.
7. Connect the output of the “Two-Class Logistic Regression” module to the left input port of the “Train Model” module.
8. Also, connect the training data output (i.e. the left port) of the “Split” module to the right input port of the “Train Model” module.
9. Run the experiment.

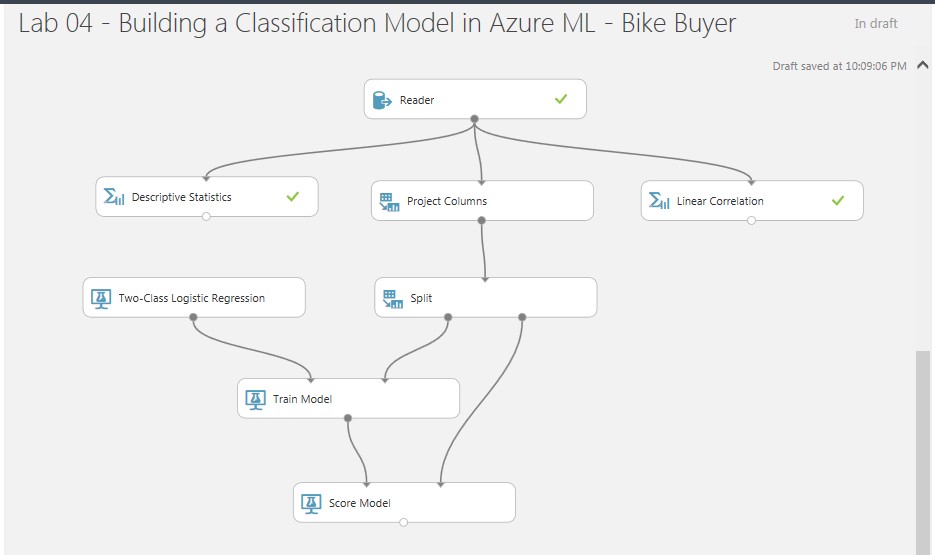


# **Exercise 3: Test the Model**

**Predict Over New Data**

Now that you’ve trained the model, you can use it to score the other 20% of your data and see how well your model predicts on unseen data.

1. Find and drag the “**Score Model**” module to the experiment canvas Connect the output of the “Train Model” module to the left input port and test data output (right port) of the “Split” module to the right input port.



1. Run the experiment and view the output from the “Score Model” module (by clicking the output port and selecting Visualize). The output will show the predicted **Bike Buyers** along with the known values from the test data.
2. Remember, the standard output of any classification is **Scored Labels** and **Scored Probabilities.**

