Website to Download the Zip File and Only use cdph-age & cdph-race-ethnicity files:

URL: <https://github.com/datadesk/california-coronavirus-data>

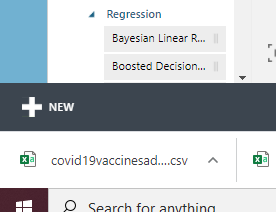
Log in to Azure Machine Learning: [Microsoft Azure Machine Learning Studio (classic) (azureml.net)](https://studio.azureml.net/)

Objective: Check the Accuracy of the data using Linear regression and Bayesian Regression. Check for permutation of Importance and cross Validate to check the data you are training to check for accuracy and error of the model.

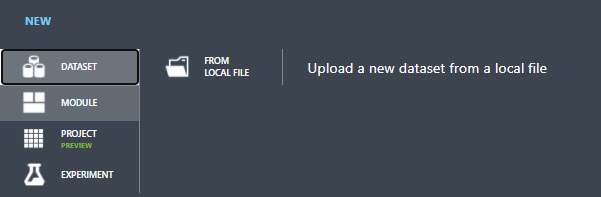
Import the data sets:

Create a Blank experiment and name It COVID-19 Confirmed Cases

On the bottom left of you page click the new button:



Select Dataset on the left and select from local file. Then import the two datasets that you have downloaded. Name the race ethnicity file COVID-19 Confirmed Cases and the Age file name it COVID-19 Age.



Bayesian Linear Regression:

Drag the Confirmed Cases Dataset module to the experiment pane. In the search item section type in “Bayesian Linear Regression” and drag the module to the experiment. Look up the “split data” module as well and drag it into the experiment.

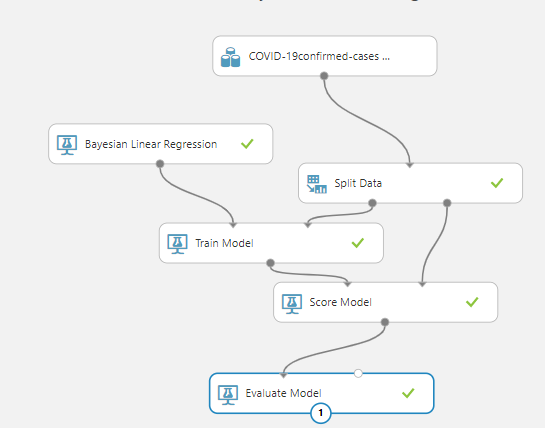
Use the default properties for both modules.

Search for the “train model” module and drag it into the experiment. Launch the column selector and select “Confirmed\_cases\_total” (with rules, included, by column names).

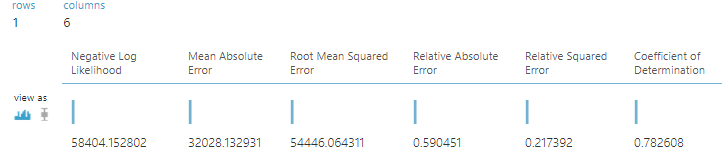
Search for the “score model” module and drag it into the experiment

Search for the “Evaluate Model” module and drag it into the experiment.

Connect the modules like this:



Save and run. When finished, visualize the “evaluate model” module



It tells you the level of errors and accuracy for the specific data column you trained in the train model module.

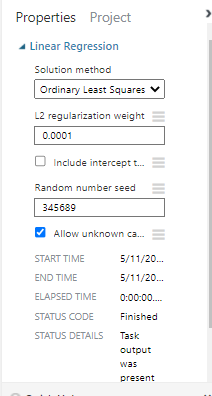
You can launch the column selector for the “train model” module and choose “death\_total” this time to check the accuracy for that column data.

Repeat these steps for the ages data just drag in the COVID-19 Ages dataset instead of the demographic dataset. Check the data level of errors and accuracy.

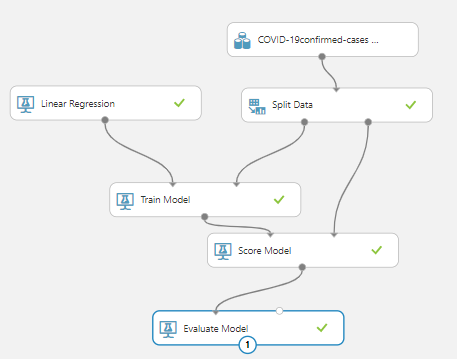
Linear Regression:

Repeat the same steps as above but use the demographic data. Instead of using the “Bayesian linear regression” module, use the “linear regression”

Change the Linear regression module properties like this:

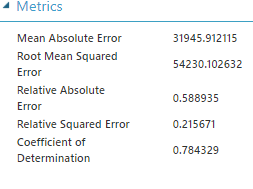


Connect the modules like this:



Save and run.

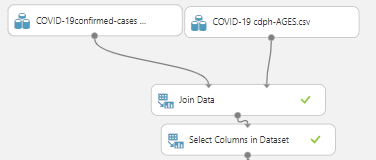
When finished, visualize the “evaluate model” module and check for level of errors and accuracy. It should look like this:



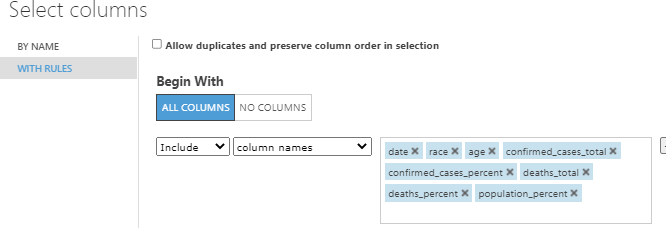
Joining Data:

Drag both datasets, COVID-19 Confirmed Cases and Ages datasets into a new experiment. Name it however you want. Search for the “join data” module drag it into the experiment. Search for “select columns in dataset” module and drag it in the experiment.

Connect the modules like this:

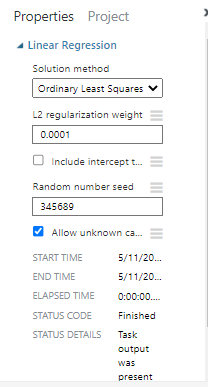


In the select column properties launch the column selector. With rules, all column, include, column name and select all columns. As shown below:

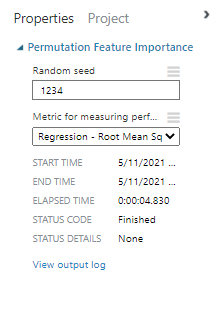


Search for the “split data”, “linear regression”, “train model”, and “permutation of importance” module, then drag it into the experiment.

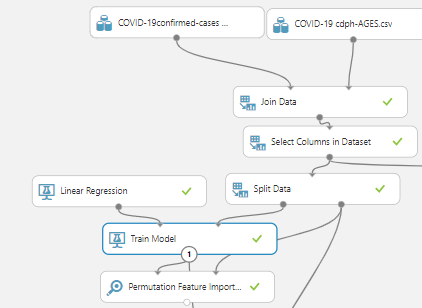
Change the properties for “linear regression” to this:



The properties of “permutation of importance” should look like this:



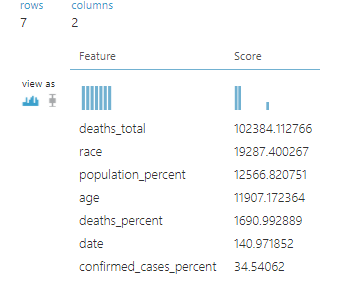
Launch the column selector for “train module”, here you can choose which data column you want to train. In this case select “Confirmed\_cases\_total” and connect the modules like this.



Save and run.

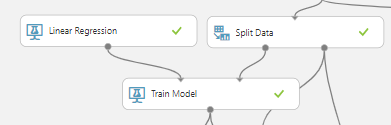
When finished visualize the permutation module and it will list the columns data with most and least importance.

It should look like this:



The column on least importance is confirmed\_cases\_percent.

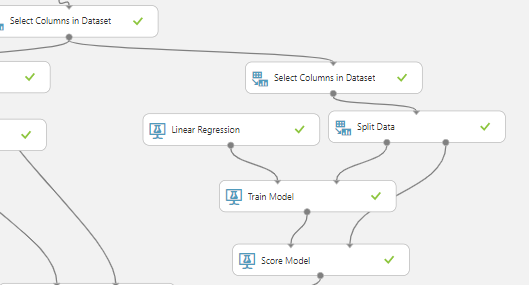
Copy these modules:



and paste them to the right side for now.

Search for a new “project columns” and “score model” module and drag it the right side.

Connect the modules like this:

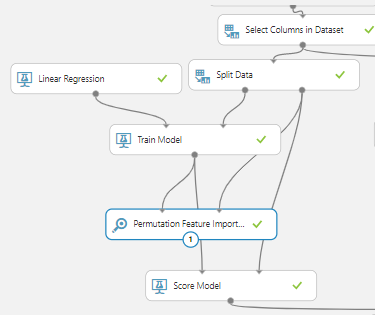


We will train a new model. In the new “project columns” module, launch column selector and exclude the “confirmed\_cases\_percent” column.

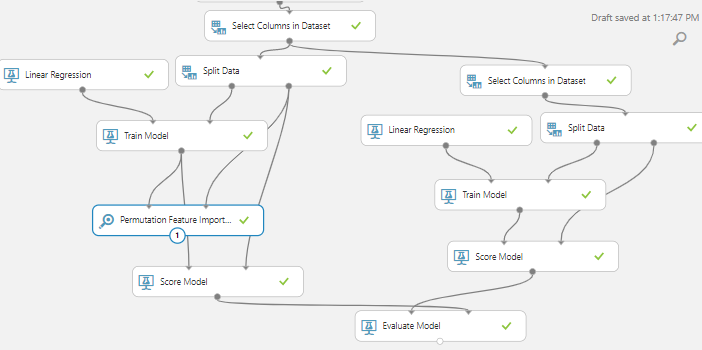
For the new “train model” launch the column selector and choose “Confirmed\_total\_cases”. We want to be able to compare if excluding the “confirmed\_cases\_percent” column made the model better or worse in terms of accuracy.

Search for “evaluate model”, and a new “score model” module, then drag it to the experiment.

Connect the new score module like this (left side):

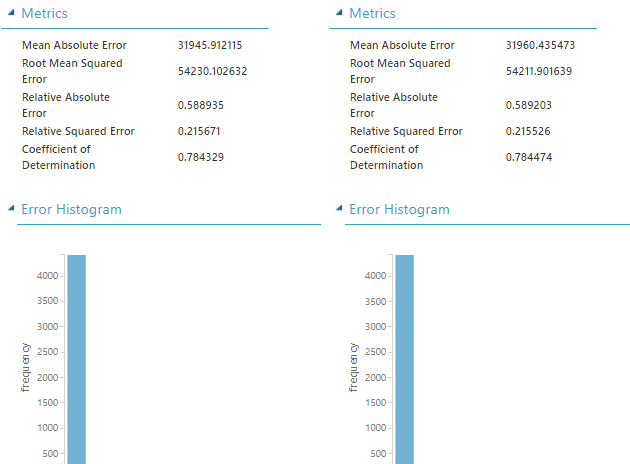


Then connect the “evaluate model” module like this:



Save and Run the model.

When finished visualizing the evaluate module. It should look like this:



The chart on the right side is the model where we excluded the “confirmed\_cases\_percent” column and the chart on the left is where we did not exclude any column. The coefficient of determination for the model on the right made increase a bit. Which means the accuracy improved. The closer it is to one the better the accuracy of the model.

You can train different data columns and compare them to each other to test the level or errors and accuracy.

You can also include/exclude columns to see if it improves or worsens your model.

END