**1-2– Sampling**

In Statistics, **sampling** is a process of generating a subset from a whole population. Its purpose is to provide various types of statistical information of a qualitative or quantitative nature about the whole by examining a few selected units. For example, you can use the sample to train a machine learning model, and use the rest of the data for testing the trained model. In the beginner course, we have shown various usage of the sample tool which ranges from removing excess rows to non-probability sampling. For this module, we are going to discuss other ways of sampling in alteryx.

In this workflow, we are going to generate sample data from this credit card customers dataset. Since our file does not have any IDs available, it is best to add a Record ID tool before any other process. Under the Preparation tool tab, drag Record ID and place it after the Select tool. We will just leave the default configuration as is. To create our first sample, we are going to use the Random % Sample Tool.

**Random % Sample Tool** returns an expected number of records that results in a random sample of the incoming data stream.

From the Preparation tool tab, drag Random % Sample tool and place it after the Record ID tool. Same as the Sample Tool, the Random % Sample has a single input and output anchor. The output for this tool will only contain the sample records and remove the rest of the records, so you can simply join back the results by using the Record ID that we inserted earlier. For this tool’s configuration, you can choose to either set a specific amount of random sample records to be generated using “Random N Records” or set a percentage of records to output as sample using “Random N% of Records”. First, lets set Random N Records to 50 and run the workflow. The resulting sample output chooses 50 random rows from the original dataset. Next, change the configuration to “Random N% of Records” and set it to 10% then run the workflow. Since our input dataset originally has 1000 records, the tool outputs 100 random records which is 10% of the population. Whichever configuration you chose, the random % sample tool will generate a different set of sample records each time the workflow is run. However, if you wish for the output to stay consistent on each run, you need to enable the “Deterministic Output” option. Ticking this function lets you set a “Random Seed” which ensures the same set of random results is returned. Changing the random seed will generate a different set of results. Let’s enable the function and set “1234” the default random seed. Then set the tool to “Random N Records” and output 3 sample records for checking. Once done, run the workflow. This generated 3 random records with IDs 707,735, and 741. Running this workflow again will not change the output. But if we change the seed to “5678”, it will output a different set.

**Create Samples Tool** is used to split the input records into 2 or 3 random samples. In the tool, you specify the percentage of records that are in the estimation and validation samples. If the total is less than 100%, the remaining records fall in the holdout sample.

In our workflow, navigate to the Preparation tool tab and drag Create Samples tool to the canvas. Place it below the Random % Sample and connect it to the output of the Record ID tool. Unlike Random % Sample, Create Sample has 1 input anchor and 3 output anchors which pertains to 3 different streams. We will discuss each as we configure the tool. In its configuration tab, the first part “Record Allocation” lets you set the percentage of records for *Estimation* or the percent of output on the E Anchor, and the *Validation* or output of the V anchor. You can indicate a percentage value between 1% and 99% for each textbox. Ideally, the total for estimation and validation should be equal to 100%. If its less, the residual record will be on the *Holdout* sample or the H anchor. The last configuration is the Random Seed where you can indicate a value between 1 to 1000. Changing this value will alter the sample that an individual row of the data is placed in. Unless necessary, it is recommended to just stick to the default 1.

Let’s set Estimation to 50 and Validation to 45 then run the workflow. Based on the output, 500 records were sampled on Estimation, 450 for validation, and the remaining 5% or 50 records were on the Holdout stream. If ever the total number of records in your dataset is odd and can’t be equally separated between the estimation and validation percentage, the extra 1 record will be placed on estimation. This tool is great for generating different sample streams which is needed when working on models. You can use the Estimation anchor for training the model and Validation for testing.

**Classification Predictive Models assigns a class label to a dataset. For example, medical records of patients can be grouped as either “Diabetic” or not. This type of model that classifies between 2 groups is called binary classification.** It is important to observe that the data is spread evenly on the outcome or target variable of the model. When the data used for the model has higher proportion of positive response rather than negative, or vice versa, the dataset can be considered “**imbalanced**”. **Imbalanced datasets can greatly influence an algorithm as it can result to ignoring the minority group.** For example, if 89% of our patient records are healthy, and only 11% were diabetic, the model can overlook the 11% and might simply tag most cases as healthy and not diabetic. 2of the methods used to fix this is oversampling and undersampling.

Alteryx has the **oversample field tool** which samples incoming data so that there is equal representation of data values to enable effective use in a predictive model. This tool undersamples the majority class or removes random records in the majority instead of oversampling the minority.

The dataset we are going to use is a list of credit card transactions which has the Time or the seconds elapsed, features V1 to V28, transaction amount, and finally Class, which is our target variable. As you can observe from the output of the summarize tool, Class only has 492 records tagged as 1 or fraudulent, as compared to 0 which has 284,315 records. To create an even distribution of data on each class, we will apply the oversample field tool. From the preparation tool tab, drag Oversample field tool and place it on the canvas. Connect it to the output anchor of the select tool. To configure the tool, first we need to select the field you want to base the oversampling on or typically the target variable, click the dropdown and select our target variable “Class”. Next, type the field value you wish to oversample. Since 1 or fraudulent is our minority class in this dataset, type 1 on the textbox. Finally, you need to set the desired ratio or proportion of the output. If you wanted fraudulent records to be half of the sample amount, you can retain the default 50 percent. If you wish for fraudulent records to only cover a third of the whole population you can type 33%. You can input an integer value between 1 to 100. Let’s retain 50% then add a copy of the summarize tool after the oversample field to check the total number of records per class. Once done, run the workflow. As we can see from the output, the oversample field limited the class 0 to 492 records in order to create a 50/50 distribution. It did not create duplicates of the Class 1, but instead minimized the majority class to create an equal proportion.

**Select Records Tool** returns records and ranges of records that are specified, including discontinuous ranges of records. This tool is very useful for troubleshooting and sampling.

Let’s use the credit card customers dataset to explain how this tool works. From the Preparation tool tab, drag Select Records tool and place it below the create samples tool. Connect it to the output anchor of the Record ID tool. The Select Records tool only has 1 multi-line textbox as its configuration. This is where you will input the range of records that you want to output. You can input a single digit, a range of numbers, or any combination of numeric ranges.

For example, lets change the range number to a single digit “5” then run the workflow. This returns a single record which is row 5 on our dataset.

Next, lets change it to a single digit with negative or minus sign. Change the range to “-5” then run the workflow. This outputs the first 5 records of our dataset, from Records 1 to 5. Adding a minus sign on the number will return row 1 up to the entered row.

Next, lets set a specific range of numbers. Input “5-10” as the range then run the workflow. This outputs records 5 up to 10.

Now lets try a number followed by a plus sign. Input “900+” then run the workflow. This returns the record starting from 900 up to the end of our dataset which is record 1000.

You can also create a combination of different records and ranges by adding a new line. For example, if wanted to get the first 100 records, records 300 to 400, and all records after 800, type in “-100” next line “300-400” and the next line “800+”. This returns the selected records of each line and combines them into 1 output anchor with a total of 402 records.