**Machine Learning interview task:**

**What am I predicting?**

You are predicting the binary variable labelled ”target” (note - no information is provided about this variable). We want you to build a predictive model for this binary outcome without overfitting to the minimal data provided.

There is a small sample size to build your model on (“Dataset\_1.csv”) and a large number of cases to predict the “target” on (“Dataset\_2.csv”). Please note all cases are in rows and variables in columns in each of the datasets. Only dataset\_1 has class labels for the “target” variable (*supervised*), dataset\_2 has no class labels (*unsupervised*), we have hidden the “target” labels for the purposes of this exercise in dataset\_2.

We will be assessing how well your model predicts the “target” variable within dataset\_2 (where you have no class labels for the “target” variable). The aim of this task is to try and get the best accuracy on dataset\_2, from building your model and assessing it on dataset\_1.

**There are two data sets:**

**Files**

* Dataset\_1.csv - used to build your model. 250 rows.
* Dataset\_2.csv – to predict “target” variable on. 19,750 rows.

**Columns**

* id- sample id
* target- a binary target.
* 0-299- continuous variables.

**What we expect back:**

* Predictions for “target” assigned to **all ids** in *Dataset\_2*, we have the class membership for “target” in Dataset\_2, which will be used to assess the quality of your model.
* Your code, delivered in either R or python with commentary on how you approached this task (we will be looking at this to see how you thought about the task, types of models used etc).

Time expected to complete: You will have up to 4 days to complete the assignment

Any questions please get in touch with s.pesarra@bonamyfinch.com.