# Data Science Take-home Assignment

## Task

1. Create a model to predict the target variable y from the given CSV file.
2. Productionize the model into inference code so it can be used on to create inference results using other CSV files.

## Data

We have provided you with a dataset in CSV format. The dataset contains:

- 100,000 rows
- 304 input features labeled x001 to x304.
- Target variable labeled y.

## Deliverables:

1. Written Report in PDF format.
2. Python codes, persisted model(s), and Jupyter Notebooks, if any, and any files required to reproduce your results.
3. Detailed instructions on how to run your model on a validation data set.

## What should be in the report

We will evaluate your ability to communicate clearly to business stakeholders and technical peers. Your report should include the following:

1. List of any assumptions that you made
2. Description of your methodology and solution path
3. List of algorithms and techniques you used
4. List of tools and frameworks you used
5. Results and evaluation of your models
6. Your machine's hardware specifications
7. Estimates of inference time to for 100,000 data points
8. Estimate of memory footprint for model inference

## How Will the Code be Evaluated

One key aspect of the job is getting the models ready for production, which means code styles, clarity, documentations, etc are very important, in addition to functional and reproducible code.

1. The code we will run must run from the command-line. Your program must accept the hold-out data set filename as a command-line argument. We should not need to edit your code before running it.
2. Your Jupyter Notebooks, if any, will be evaluated on clarity and reproducibility. We love clear and nicely written self explanatory notebooks!
3. We expect OOP or FP best practices and Pythonic codes, which means
4. Please include any Python virtual environment management files so we can reproduce your environment to run your code. Feel free to use Docker images if you choose to do so.


## How Will the Model be Evaluated

We have a hold-out set that will be used to evaluate your model. Please provide clear and detailed instructions on how we can run your model on our hold-out set, without having to make the hold-out set available to you. The model will be evaluated by the following metrics.

1. Root Mean Square Error (RMSE)
2. We are also interested in the percentage accuracy of the model. If the absolute error of a prediction is greater than 3.0, we regard the prediction as "wrong". Otherwise, it is "correct".

The hold-out data set will have exactly the same structure and format as the CSV file we have provided you with. Your code must output the RMSE and Accuracy to standard output and the predicted values of y must be written to a text file.

## Other Details

- We have purposefully left some things ambiguous and not well-defined. We want to see the assumptions that you make and the solution path that you choose to solve the problem. These should be clearly described in the report.
- Send your code in a separate file from the report. It should be in a format that we can execute directly.
- We will not be able to run your code if it requires interactive or graphical interfaces such as IPython, Jupyter Notebook. You can use such tools when building your model of course. But the code we will run on the hold-out data set should not require them.


- We only want to run one of your models. You should choose your best model. We only want to run the code that makes predictions on a hold-out data set. We do not want to run the training process that builds the model.
- You should send us all your code. We will only run the code that you indicate in your Instructions for making predictions on a hold-out data set. But we want to also see the code that you used for training and building the model.