Data Scientist Test

Data

We provide you with a data set in CSV format.
The data set contains 100,000 instances.
There are 304 input features, labeled x001 to x304.
The target variable is labeled y.

Task

Create a model to predict the target variable y.

Please send us:

1. A report
2. Any custom code you used
3. Instructions for us to run your model on a separate data set

What should be in the report?

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

How to evaluate the model

1. Use the 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".
3. Any other evaluation measure that you believe is appropriate.

Instructions

- 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. Please do not send us Microsoft Word documents for the report. You can use Open Document format or pdf.

- We have a hold-out set that we will use to evaluate your model. Please provide instructions on how we can run your model on our hold-out set, without having to make the hold-out set available to you.

- Send your code in a separate file from the report. It should be in a format that we can execute directly.

- Provide clear and detailed Instructions on how we can run your code on a hold-out data set. Your code must output the RMSE and Accuracy to standard output and the predicted values of y must be written to a text file. Also, state your machine's hardware specifications and, based on this, give an estimate of how much memory (RAM) and time (in seconds) your model requires in order to generate predictions on a hold-out data set of 100,000 data points.

- We are not only evaluating the performance of your model. We are also evaluating your ability to communicate clearly and your ability to write good code.

- 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.

- The hold-out data set will have exactly the same structure and format as the CSV file we have provided you with.

- We will not be able to run your code if it requires interactive or graphical interfaces such as IPython, Jupyter Notebook, or RStudio. 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.

- If your code requires specific versions of software (python, R) or specific libraries then you need to indicate this in your Instructions so that we can install them. Another option is to use Docker. If we find it difficult to run your code then you will automatically fail the test. You therefore need to provide clear and detailed Instructions on how we can run your model.

- We only want to run one of your models. You should choose your best model. We only want to 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.