Data Science Test

Case title: Energy efficiency

These are the instructions for the take home test for candidates for data science positions at Drax. Please do not share these instructions, the accompanying data, or your analysis with anyone other than your contacts at Drax.

You need to return this test by one day after receiving the instruction. However, we estimate that most candidates should be able to complete the tasks in approximately 3 hours and will be able to do so using standard tools. We recommend Python for ease of development and rich data science ecosystem, and because it's the language we use most commonly internally.

Please read the context and the data sections to get familiar with the data provided for you and then follow the instructions to complete the tasks.

Context

The buildings differ with respect to the glazing area, the glazing area distribution, and the orientation, amongst other parameters. The dataset comprises 768 samples and 8 features, aiming to predict two real valued responses

The dataset was created by Angeliki Xifara (Civil Engineer) and was processed by Athanasios Tsanas (Oxford Centre for Industrial and Applied Mathematics, University of Oxford, UK).

Data Set Information:

The dataset contains eight attributes (or features, denoted by $X_1...X_8$) and two responses (or outcomes, denoted by y_1 and y_2). The aim is to use the eight features to predict each of the two responses.

Specifically:

X_1 Relative Compactness

X_2 Surface Area

X_3 Wall Area

X_4 Roof Area

X_5 Overall Height

X_6 Orientation

X_7 Glazing Area

X_8 Glazing Area Distribution

y_1 Heating Load

y_2 Cooling Load

Tasks

1. Try to understand the data by exploratory data analysis and formulate some hypothesis, then test your hypothesis and summarise your understanding by useful visualisations and insights.
2. Build a proof of concept model (using Python) to predict heating and cooling loads (with justification for the choice of a particular model) based on tested hypotheses and identified relationships.
3. Test your model using appropriate metrics and state how you would expect it to perform on unseen data.
4. Conduct a feature importance analysis based on your model and show which features are contributing the most to the heating and cooling loads.

Submission

Please send us all the code you wrote (Jupyter notebook), plus a power point presentation by email in reply to the email originally sent to you. If sending multiple files please zip or archive them first, .zip or .tar is fine.

We are looking for brevity, specificity, accuracy, insight, clarity of thought and style of submitted presentation.

What does 'good' looks like?

- Illustrations with charts and legends, trends, correlations, and relevant confident statements.
- An 'Executive Summary' leading to a storyline in the following support slides
- Clarity of next steps and follow up work
- Submission on time and to deadline
- Use of the Drax PowerPoint template

We hope you will find the exercise challenging and rewarding, and we look forward to your participation in the practical task.