**# Introdução**

Você está participando de um processo de seleção para Cientista de Dados em uma grande empresa multinacional. A empresa definiu um problema e quer saber como você pode solucioná-lo explorando o conjunto de dados, realizando análise estatística e desenvolvendo um modelo para previsões.

O conjunto de dados anexado a esta competição contém dados sobre a eficiência energética de edifícios. O dataset foi extraído do web site Open Data for All New Yorkers (https://opendata.cityofnewyork.us/). Alternativamente o dataset pode ser encontrado aqui: https://catalog.data.gov/dataset

Nossa hipótese é que esse conjunto de dados contém variáveis ​​independentes que podem ser usadas para inferir fatos interessantes sobre outros edifícios em Nova York. O principal interesse é a pontuação (score) de Energy Star, pois acreditamos que essa pontuação é usada como uma medida agregada do desempenho energético de um prédio. Nossa pergunta é simples:

Este conjunto de dados contém um conjunto de variáveis ​​independentes que se correlacionam com a classificação "Energy Star" do edifício? Para edifícios sem a pontuação Energy Star, você pode inferir qual a pontuação deles?

Obs: Não sabe o que é Energy Star? Então pesquise, pois conhecimento de negócio faz parte dos skills de um Cientista de Dados. Aqui uma fonte de informação útil:

https://www.energystar.gov/ia/partners/spp\_res/neprs/ENERGY\_STAR\_and\_Automated\_Benchmarking\_Quick\_Facts.pdf

Para responder a esta pergunta, pedimos que você prepare uma investigação fundamentada dessa hipótese. Acreditamos que o melhor método é realizar descrições estatísticas visuais dos dados usando tabelas de resumo ou uma biblioteca de gráficos de sua escolha (análise exploratória). Então treine um modelo e avalie-o. Você pode usar regressão para prever a pontuação numérica (embora outros métodos de aprendizagem de máquina possam ser usados neste dataset, esta competição considera que você usará regressão).

Analisando os dados brutos, podemos ver vários problemas que deverão ser resolvidos antes da modelagem preditiva. Existem 60 colunas e não sabemos o que muitas delas significam! Tudo o que sabemos da declaração do problema é que queremos prever o número na coluna de pontuação. Algumas das outras definições de coluna podem ser razoavelmente adivinhadas, mas outras são difíceis de entender. No aprendizado de máquina, isso não é realmente um problema, porque deixamos o modelo decidir quais recursos são importantes. Às vezes, podemos nem receber nomes de coluna ou saber o que estamos prevendo. No entanto, é sempre importante entender o problema na medida do possível e, como também queremos interpretar os resultados do modelo, seria uma boa ideia ter algum conhecimento das colunas. Para uma descrição completa do dataset, acesse: https://www1.nyc.gov/html/gbee/html/plan/ll84.shtml

Crie um ou mais modelos (você usar diferentes algoritmos de regressão) e nos dê uma interpretação de seu desempenho, que você pode usar para responder à questão de saber se esse conjunto de dados contém informações suficientes para entender a relação de uma pontuação Energy Star com edifícios de diferentes tipos. Para o desempenho do modelo, gostaríamos de ver métricas de avaliação típicas / apropriadas como pontuações de F1, R Squared, resíduos, etc. Esta competição considera como métrica de avaliação o MAE (Mean Absolute Error).

Esse trabalho é uma grande desafio e superá-lo é seu objetivo! Bom trabalho!

Equipe DSA

**# Introduction**

You are participating in a selection process for Data Scientists in a large multinational. The company defined a problem and wants to know how you can solve it by exploring the dataset, performing statistical analyses and developing a predictive model.

The dataset attached to this competition contains data regarding the energy efficiency of buildings in the city of New York. The dataset was extracted from the Open Data For All New Yorkers website (<https://opendata.cityofnewyork.us/>). Alternatively, the dataset can be found here: <https://catalog.data.gov/dataset>.

Our hypothesis is that this dataset contains independent variables which can be used to infer interesting facts about other buildings in New York. The main interest is in the Energy Star score, as this score is used as an aggregate measure of a building’s energy consumption performance. Our question in simple:

Does this dataset contains independent variables which are correlated to the building’s “Energy Star” score? For buildings without an Energy Star score, can you infer their scores?

PS: Don’t know what Energy Star is? Then research, because business knowledge is part of a Data Scientist’s skillset. Here’s a useful source of information: <https://www.energystar.gov/ia/partners/spp_res/neprs/ENERGY_STAR_and_Automated_Benchmarking_Quick_Facts.pdf>

To answer this question, we ask that you prepare an investigation based on this hypothesis. We believe that the best approach is to perform visual statistic descriptions from the data by using summary tables or a library or your choice (exploratory analysis). Then train and evaluate a model. You can use regression to predict the numeric score (although other machine learning techniques may be used in this dataset, this competition considers that you will use regression).

Analyzing the raw data we identify multiple problems which must be solved before the predictive analysis. There are 60 columns and the meaning of many of them is unknown to us! All we know is that the goal is to predict the score column. The meaning of some of the other columns may be reasonably guessed, but there are yet others for which the meaning is difficult to understand. Sometimes you may not even receive the name of a column or know what you are predicting. Nevertheless, it is always important tot understand the problem to the best of your ability, and, given that the goal is to also interpret the model’s results, it is a good idea to have some knowledge about the columns. For a complete description of the dataset access: <https://www1.nyc.gov/html/gbee/html/plan/ll84.shtml>.

Create one or more models (using different regression algorithms) and give us an interpretation from their performance, which you may use to answer the question of knowing whether the dataset contains sufficient information for the modeling of the relationship between Energy Star Score and different building types. For the model performance, we would like to see typical (and appropriate) evaluation metrics such as F1 scores, R Squared, residuals, etc. This competition considers as evaluation metric the MAE (Mean Absolute Error).

This project is a great challenge and overcoming it is your objective!

Data Science Academy team

**# Métrica de Avaliação**

Há várias métricas usadas em tarefas de aprendizado de máquina e pode ser difícil saber qual delas escolher. Na maioria das vezes, isso dependerá do problema específico e se você tiver uma meta específica para otimização. Um consenso entre os especialistas em Data Science é: em vez de calcular várias métricas e tentar determinar a importância de cada uma delas, devemos usar um único número. Neste caso, porque estamos fazendo regressão, o Mean Absolute Error (MAE) é uma métrica apropriada. Essa métrica também é interpretável porque representa o valor médio de nossa estimativa.

**# Evaluation Metric**

There are many metrics used for supervised learning tasks and it might be difficult to know which to chose. Most times the choice will depend on the specific problem and if there is a specific goal for optimization. A consensus amongst Data Science specialists is: instead of calculating various metrics and trying to determine the importance of each, a single number should be used. In this case, since the problem involved regression, the Mean Absolute Error (MAE) is an appropriate metric. This metric is also interpretable since it represents the mean value of an estimative.

**# Ferramenta**

Você pode usar a ferramenta com a qual se sentir mais confortável, mas recomendamos o uso de R, Python, Scala ou Java. Além do arquivo com as previsões, você deve submeter a solução completa que suporte as previsões. Isso pode ser feito enviando um kernel nesta competição. Você pode trabalhar com seu kernel em modo privado e próximo da data de fechamento da competição deve tornar seu kernel público

**# Tool**

You may use whichever tool you are most comfortable with, but we recommend R, Python, Scale or Java. Besides the submission file, you must also submit your complete solution which supports your predictions. This can be done by sending a kernel in this competition. You can work with your kernel on private mode but near the competition end date you must make it public.

**# Formato do Envio da Solução**

Para cada edifício no conjunto de dados de teste, o arquivo com a solução deve conter duas colunas: PropertyId e score com a previsão feita pelo modelo (a variável score deve ser numérica em um range de 0 a 100).

O arquivo deve conter um cabeçalho e ter o seguinte formato:

PropertyId,Score

5729364,13

**# Solution Submission Format**

For each building in the test dataset the solution file must contain two columns: PropertyId and score with the model’s predictions. The score variable must be numeric and in a range from 0 to 100.

The file should contain a header and the following format:

PropertyId,Score

5729364,13