Desafio Cientista de Dados

Introdução

Olá candidato, o objetivo deste desafio é testar os seus conhecimentos sobre a resolução de problemas de análise de dados e aplicação de modelos preditivos. Queremos testar seus conhecimentos dos conceitos estatísticos de modelos preditivos, criatividade na resolução de problemas e aplicação de modelos básicos de machine learning. É importante deixar claro que não existe resposta certa e que o que nos interessa é sua capacidade de descrever e justificar os passos utilizados na resolução do problema.

Desafio

Seu objetivo é identificar quais máquinas apresentam potencial de falha tendo como base dados extraídos através de sensores durante o processo de manufatura. Para isso são fornecidos dois *datasets*: um *dataset* chamado *desafio_manutencao_preditiva_treino* composto por 6667 linhas, 9 colunas de informação (*features)* e a variável a ser prevista (“failure_type”).

O segundo dataset chamado de *desafio_manutencao_preditiva_teste* possui 3333 linhas e 8 colunas e não possui a coluna “failure_type”. **Seu objetivo é prever essa coluna a partir dos dados enviados e nos enviar para avaliação dos resultados.**

Você poderá encontrar em anexo um dicionário dos dados.

Entregas

1. Descreva graficamente os dados disponíveis, apresentando as principais estatísticas descritivas. Comente o porquê da escolha dessas estatísticas.
2. Explique como você faria a previsão do **tipo de falha** a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê? Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?
3. Envie o resultado final do modelo em uma planilha com apenas duas colunas (rowNumber, predictedValues).
4. A entrega deve ser feita através de um repositório de código público que contenha:
    a. README explicando como rodar o projeto
    b. Arquivo *requirements* com todos os pacotes utilizados
    c. Relatório de EDA em PDF, Jupyter Notebook ou semelhante conforme passo 1
    d. Códigos de modelagem utilizados no passo 2.
    e. Arquivo final *predicted.csv* conforme passo 3 acima.

Prazo

Você tem até **7 dias corridos** para a entrega, contados a partir do recebimento deste desafio. Envie o seu relatório dentro da sua data limite para o e-mail: **selecao.lighthouse@indicium.tech.**

Bom trabalho!

Dicionário dos dados

The dataset consists of 10 000 data points stored as rows with 8 features in columns:

1. **UID**: unique identifier ranging from 1 to 10000
2. **product ID**: consisting of a letter L, M, or H for low (50% of all products), medium (30%) and high (20%) as product quality variants and a variant-specific serial number
3. **type**: just the product type L, M or H from column 2
4. **air temperature [K]**: generated using a random walk process later normalized to a standard deviation of 2 K around 300 K
5. **process temperature [K]**: generated using a random walk process normalized to a standard deviation of 1 K, added to the air temperature plus 10 K.
6. **rotational speed [rpm]**: calculated from a power of 2860 W, overlaid with a normally distributed noise
7. **torque [Nm]**: torque values are normally distributed around 40 Nm with a SD = 10 Nm and no negative values.
8. **tool wear [min]**: The quality variants H/M/L add 5/3/2 minutes of tool wear to the used tool in the process.

A 'machine failure' label that indicates whether the machine has failed in this particular datapoint for any of the following failure modes is true.

The machine failure consists of five independent failure modes:

1. **tool wear failure (TWF)**: the tool will be replaced of fail at a randomly selected tool wear time between 200 - 240 mins (120 times in our dataset). At this point in time, the tool is replaced 69 times, and fails 51 times (randomly assigned).
2. **heat dissipation failure (HDF)**: heat dissipation causes a process failure, if the difference between air- and process temperature is below 8.6 K and the tools rotational speed is below 1380 rpm. This is the case for 115 data points.
3. **power failure (PWF)**: the product of torque and rotational speed (in rad/s) equals the power required for the process. If this power is below 3500 W or above 9000 W, the process fails, which is the case 95 times in our dataset.

4. **overstrain failure (OSF)**: if the product of tool wear and torque exceeds 11,000 minNm for the L product variant (12,000 M, 13,000 H), the process fails due to overstrain. This is true for 98 datapoints.
5. **random failures (RNF)**: each process has a chance of 0,1 % to fail regardless of its process parameters. This is the case for only 5 datapoints, less than could be expected for 10,000 datapoints in our dataset.