Prova - Ciência de Dados

Consultor de Dados Pleno

Bem-vindo ao desafio ICTS! O Consultor Pleno em Ciência de Dados em nosso time tem um papel muito importante: ele deve desenvolver modelos de machine learning que apresentem bom desempenho, que sejam confiáveis, escaláveis e de fácil manutenção em produção.
Neste desafio, queremos explorar sua capacidade de desenvolver uma solução de machine learning, desde o ETL, EDA e feature selection até a construção de um pipeline completo para classificação de registros em produção.
Apresentamos a seguir um problema para que você possa avaliá-lo e sugerir uma solução. Para tanto você disporá de 2 dias (48 horas) a partir do envio do teste para devolvê-lo. Boa Sorte!

O desafio

O recozimento é um tratamento térmico que tem por finalidade eliminar a dureza de uma peça temperada ou normalizar materiais com tensões internas resultantes do forjamento, da laminação e da trefilação. De acordo com o novo nível de dureza do material tratado, o resultado do processo de recozimento pode ser classificado como: ideal, mediano ou ruim.

O conjunto de dados apresentado traz o resultado da execução de diferentes processos de recozimento em experimentos variados.
Assim sendo, o candidato deve utilizar os dados para treinar um modelo de machine learning que posteriormente deverá ser utilizado para predizer o resultado de novos processos de recozimento.

Importante:

Somente experimentos realizados em agosto de 2020 devem ser considerados.
A tabela acessória `data_experimentos.csv` contém essa informação.

SÃO PAULO	ALPHAVILLE	RIO DE JANEIRO	BELO HORIZONTE
Rua James Joule, 65,	Alameda Araguaia, 2104, 7º andar	Av. Almirante Barroso, 81	Rua Antonio de Albuquerque, 330
5º andar • Cidade Monções	Alphaville Industrial	33º andar • Centro	8º andar • Savassi
São Paulo • SP • Brasil	Barueri • SP • Brasil	Rio de Janeiro • RJ • Brasil	Belo Horizonte • MG • Brasil
CEP: 04576-080	CEP: 06455-000	CEP: 20031-004	CEP: 30112-010

Bases a serem utilizadas:

- train.csv: dados para treino do modelo de ML.
- test.csv: dados a serem classificados no pipeline construído e então submetidos à aprovação.
- data_experimentos.csv: informação sobre data de realização dos experimentos.

Tarefas a serem desenvolvidas pelo candidato:

- Unir o conjunto de dados principal (train.csv) à tabela acessória (data_experimentos.csv) através da chave experimento/exp_id – relacionamento 1x1.
- Realizar análise exploratória e tratamento das variáveis conforme necessário.
- Realizar análises univariadas, bivariadas, multivariadas e testes estatísticos, conforme julgar necessário. O candidato deve produzir insights a partir destas análises.
- Realizar feature selection para treinamento de um modelo de ML
- Treinar um ou mais modelos, e eleger ao fim do processo o modelo mais promissor para aplicar ao conjunto de dados de teste. Observar as melhores práticas para evitar overfitting/viés.
- Documente seu raciocínio e conclusões no notebook, atentando-se ao formato solicitado nos itens 1 e 9 de "Orientações". Aqui simulamos a entrega para o time de Ciência de Dados da nossa área, então seja técnico/detalhista em suas explicações.
- Construir um pipeline de dados/ML e classificar os dados presentes do conjunto de testes através deste pipeline.
- Preparar 4 slides para apresentar o trabalho aos avaliadores (formato livre). Aqui simulamos a entrega ao cliente. Quase regra, é alguém totalmente fora do time de Ciência de Dados, então foque sua apresentação em "insights" encontrados nos dados e explique por que o seu produto resolve o problema. Recursos como storytelling com Dados podem lhe auxiliar.

Orientações:

1. O trabalho deve ser desenvolvido através de **Jupyter Notebooks / scripts python**. As análises devem ser feitas no Notebook, enquanto o pipeline de classificação deve ser desenvolvido em scripts python. Os mesmos devem ser importados em um notebook desenvolvido exclusivamente para classificar os registros presentes no conjunto de dados de teste. A observação de boas práticas e protocolos **PEP8 e PEP257** serão avaliados.
2. A organização do trabalho também será avaliada: cabe ao candidato decidir em quantos notebooks o trabalho será organizado.
3. O candidato deve descrever por que cada análise/teste foi realizado e qual é a conclusão obtida de cada um.
4. O resultado da classificação do conjunto de dados de teste deve ser **exportado em csv** e submetido à avaliação.
5. O candidato terá apenas 48 horas para realizar o trabalho e deverá priorizar quais atividades serão realizadas a fim de poder entregar o trabalho em tempo hábil. O candidato deve explicar como e porque priorizou cada tarefa.

SÃO PAULO	ALPHAVILLE	RIO DE JANEIRO	BELO HORIZONTE
Rua James Joule, 65,	Alameda Araguaia, 2104, 7º andar	Av. Almirante Barroso, 81	Rua Antonio de Albuquerque, 330
5º andar • Cidade Monções	Alphaville Industrial	33º andar • Centro	8º andar • Savassi
São Paulo • SP • Brasil	Barueri • SP • Brasil	Rio de Janeiro • RJ • Brasil	Belo Horizonte • MG • Brasil
CEP: 04576-080	CEP: 06455-000	CEP: 20031-004	CEP: 30112-010

6. **Seja simples, sem ser simplista!** Em nossa área trabalhamos com a premissa de que se "calcular a média é suficiente, então calculamos apenas a média". Isso quer dizer, que não dificultamos as soluções mais que o necessário, se ML ou técnicas avançadas de ML não são necessárias, simplesmente não usamos, assim como, não somos simplistas e não subestimamos a real necessidade de utilizarmos soluções mais complexas.
7. Dê o máximo de **detalhe** para justificar suas escolhas.
8. **Justifique** seus argumentos com fundamentos de matemática, estatística ou outros fundamentos técnicos.
9. Tenha em mente que **o que/Qual(is)? Como? E Por quê?** Exigem respostas diferentes.
    a. **O que/Qual(is)?** Esperamos que a sua resposta seja uma citação/menção daquilo que vai ser feito
    b. **Como?** Esperamos que sua resposta seja um processo, etapas que você seguiria para resolver, seu modus operandi, enfim, seu raciocínio.
    c. **Por quê?** Esperamos que você explique com detalhe, embasamento técnico (estatístico, matemático ou de sistemas), fundamente e justifique as suas decisões.
10. Quando enxergar mais de uma solução para uma resposta, escreva quantas soluções quiser, apenas justifique porque usaria a alternativa A em detrimento de B ou C e em que **contexto** você faria essa opção.

SÃO PAULO	ALPHAVILLE	RIO DE JANEIRO	BELO HORIZONTE
Rua James Joule, 65,	Alameda Araguaia, 2104, 7º andar	Av. Almirante Barroso, 81	Rua Antonio de Albuquerque, 330
5º andar • Cidade Monções	Alphaville Industrial	33º andar • Centro	8º andar • Savassi
São Paulo • SP • Brasil	Barueri • SP • Brasil	Rio de Janeiro • RJ • Brasil	Belo Horizonte • MG • Brasil
CEP: 04576-080	CEP: 06455-000	CEP: 20031-004	CEP: 30112-010

Detalhamento dos conjuntos de dados:

```
Train.csv / Test.csv
Number of Attributes: 39
Attribute Information:
    1. family:         --,GB,GK,GS,TN,ZA,ZF,ZH,ZM,ZS
    2. product-type:   C, H, G
    3. steel:          -,R,A,U,K,M,S,W,V
    4. carbon:         continuous
    5. hardness:       continuous
    6. temper_rolling:-,T
    7. condition:      -,S,A,X
    8. formability:    -,1,2,3,4,5
    9. strength:       continuous
   10. non-ageing:     -,N
   11. surface-finish:P,M,-
   12. surface-quality: -,D,E,F,G
   13. enamelability: -,1,2,3,4,5
   14. bc:             Y,-
   15. bf:             Y,-
   16. bt:             Y,-
   17. bw/me:          B,M,-
   18. bl:             Y,-
   19. m:              Y,-
   20. chrom:          C,-
   21. phos:           P,-
   22. cbond:          Y,-
   23. marvi:          Y,-
   24. exptl:          Y,-
   25. ferro:          Y,-
   26. corr:           Y,-
   27. blue/bright/varn/clean:              B,R,V,C,-
   28. lustre:         Y,-
   29. jurofm:         Y,-
   30. s:              Y,-
   31. p:              Y,-
   32. shape:          COIL, SHEET
   33. thick:          continuous
   34. width:          continuous
   35. len:            continuous
   36. oil:            -,Y,N
   37. bore:           0000,0500,0600,0760
   38. packing:        -,1,2,3
   39. recozimento:    ruim, mediano, ideal
   40. experimento:    1 ... n - discrete


   -- The '-' values are actually 'not_applicable' values rather than
      'missing_values' (and so can be treated as legal discrete
      values rather than as showing the absence of a discrete value).

Missing Attribute Values: Signified with "?"

-----------
Data_experimentos.csv

    1. exp_id:         1 ... n - discrete
    2. ano:            1 ... n - discrete
    3. mes:            1 ... 12 - discrete
```


SÃO PAULO	ALPHAVILLE	RIO DE JANEIRO	BELO HORIZONTE
Rua James Joule, 65,	Alameda Araguaia, 2104, 7º andar	Av. Almirante Barroso, 81	Rua Antonio de Albuquerque, 330
5º andar • Cidade Monções	Alphaville Industrial	33º andar • Centro	8º andar • Savassi
São Paulo • SP • Brasil	Barueri • SP • Brasil	Rio de Janeiro • RJ • Brasil	Belo Horizonte • MG • Brasil
CEP: 04576-080	CEP: 06455-000	CEP: 20031-004	CEP: 30112-010