PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS (PUC GOIÁS)

X CONGRESSO DE CT&I DA PUC GO – 2024/2

MESTRADO EM ENGENHARIA DE PRODUÇÃO E SISTEMAS (PPGEPS)

NÚCLEO DE MATEMÁTICA DO NEPE

ROTEIRO

VII DESAFIO EM CIÊNCIAS DE DADOS
"Modelagem de Dados para Predizer Tempestades Geomagnéticas".
BASE DE DADOS: Coletadas de Satélites da NASA e NOAA – 11 anos

Apesar dos desafios, este problema oferece uma excelente oportunidade de aprendizado prático e aplicação de conceitos teóricos em um problema real e relevante. A experiência de trabalhar em equipe em um problema desafiador com dados reais será valiosa para a formação dos participantes das equipes, independentemente do resultado final em termos de desempenho do modelo.

Base de dados pública: (pasta VII desafio em Ciências de dados – sala do desafio em ciências de dados - na sala teams)

FORMAÇÃO DAS EQUIPES/ ENTREGA DA BASE DE DADOS/ CONTEXTUALIZAÇÃO/ FASE DE PREPARAÇAO – 01/10 A 08/10

DESENVOLVIMENTO DA SOLUÇÃO: 03 A 16/10

Ver pasta VI DESAFIO – *Recordings* – as oficinas estão gravadas.

Arquivos:

1. solar_wind.csv:

Este é o seu principal conjunto de dados de entrada (features).

Contém as medições do vento solar coletadas pelos satélites ACE e DSCOVR.

As colunas representam variáveis como velocidade, densidade, temperatura e componentes do campo magnético interplanetário (IMF) em diferentes sistemas de coordenadas (GSE e GSM).

A coluna "source" indica qual satélite forneceu os dados em cada momento. Isso é importante porque a posição e os instrumentos dos satélites podem influenciar as medições.

Variáveis

Essas variáveis são essenciais para o estudo do clima espacial, especialmente para entender como o campo magnético interplanetário e o vento solar interagem com o campo magnético da Terra, levando a eventos como tempestades geomagnéticas. Essas medições são usadas para prever os efeitos no espaço próximo à Terra, incluindo as potenciais perturbações em

satélites, comunicações, e redes de energia elétrica.

GSE (Geocentric Solar Ecliptic): Este sistema de coordenadas é centrado na Terra e alinhado com o plano eclíptico (o plano da órbita da Terra em torno do Sol). As componentes representam as direções X, Y, e Z do campo magnético interplanetário nesse sistema.
bx_gse - Componente X do campo magnético interplanetário (IMF) nas coordenadas eclípticas solares geocêntricas (GSE), medido em nanoteslas (nT).
by_gse - Componente Y do campo magnético interplanetário (IMF) nas coordenadas GSE, medido em nanoteslas (nT).
bz_gse - Componente Z do campo magnético interplanetário (IMF) nas coordenadas GSE, medido em nanoteslas (nT).
theta_gse - Latitude do campo magnético interplanetário nas coordenadas GSE (definida como o ângulo entre o vetor magnético B e o plano eclíptico, sendo positivo quando B aponta para o Norte), medido em graus.
phi_gse - Longitude do campo magnético interplanetário nas coordenadas GSE (o ângulo entre a projeção do vetor IMF no plano eclíptico e a direção Terra-Sol), medido em graus.

GSM (Geocentric Solar Magnetospheric): Este sistema de coordenadas também é centrado na Terra, mas é alinhado com o campo magnético terrestre em vez do plano eclíptico. As componentes representam as direções X, Y, e Z do campo magnético interplanetário nesse sistema.
bx_gsm - Componente X do campo magnético interplanetário (IMF) nas coordenadas magnéticas solares geocêntricas (GSM), medido em nanoteslas (nT).
by_gsm - Componente Y do campo magnético interplanetário (IMF) nas coordenadas GSM, medido em nanoteslas (nT).
bz_gsm - Componente Z do campo magnético interplanetário (IMF) nas coordenadas GSM, medido em nanoteslas (nT).
theta_gsm - Latitude do campo magnético interplanetário nas coordenadas GSM, medida em graus.
phi_gsm - Longitude do campo magnético interplanetário nas coordenadas GSM, medida em graus.

bt - Magnitude do componente do campo magnético interplanetário (IMF), medida em nanoteslas (nT). Esta variável representa a magnitude total do vetor campo magnético interplanetário (IMF), independente da direção. É uma medida escalar que indica a força do campo magnético.

density - Densidade de prótons do vento solar, medida em partículas por centímetro cúbico (N/cm^3). Refere-se à concentração de prótons no vento solar, indicando quantas partículas estão presentes por unidade de volume.

speed - Velocidade do vento solar, medida em quilômetros por segundo (km/s). Velocidade (speed): Mede a rapidez com que o vento solar está se movendo, o que pode influenciar como ele interage com a magnetosfera da Terra.

temperature - Temperatura dos íons do vento solar, medida em Kelvin (K). Mede a rapidez com que o vento solar está se movendo, o que pode influenciar como ele interage com a magnetosfera da Terra. Temperatura (temperature): Refere-se à temperatura dos íons no vento solar, que pode afetar a pressão do plasma e a dinâmica do vento solar.

source - Indica a fonte dos dados do vento solar, a partir de 2016. Pode ser de dois satélites: "ac" indica que os dados foram obtidos do satélite ACE (Advanced Composition Explorer), e "ds" indica que foram obtidos do satélite DSCOVR (Deep Space Climate Observatory), dependendo da qualidade dos dados.

2. satellite_pos.csv:

Fornece informações contextuais importantes sobre a posição dos satélites.

Contém as coordenadas diárias (em coordenadas GSE) dos satélites ACE e DSCOVR.

A hipótese aqui é que a posição do satélite em relação à Terra e ao Sol pode influenciar as medições do vento solar e, consequentemente, a previsão do Dst.

3. sunspots.csv:

Oferece dados adicionais sobre a atividade solar.

Contém informações sobre o número de manchas solares observadas.

Manchas solares são um indicador da atividade magnética do Sol, que impulsiona as tempestades geomagnéticas.

Incorporar essa informação em seu modelo pode ajudar a capturar padrões de longo prazo e melhorar as previsões, especialmente para eventos extremos.

4. labels.csv:

Aqui estão os seus dados de saída (rótulos) para treinar seus modelos.

Contém os valores do índice Dst em intervalos de uma hora, correspondendo aos mesmos períodos dos outros arquivos.

O objetivo do seu modelo é aprender a relação entre as features (vento solar, posição dos satélites, manchas solares) e o Dst, para que você possa prever o Dst apenas com base nas features.

ETAPAS DO DESAFIO:

1. **Entendimento do desafio**
2. **Avaliação do projeto/área**
3. **Extração e obtenção de dados**
4. **Preparação dos dados (limpeza)**
5. **Análise exploratória**
6. **Modelagem e algoritmos**
7. **Interpretação do resultado**
8. **Conclusão**

ENTENDIMENTO DO DESAFIO

Leia o artigo: Nair, M., Redmon, R., Young, L.-Y., hulliat, A., Trotta, B., Chung, C. et al. (2023). MagNet—A data-science competition to predict disturbance storm time index (*Dst*) from solar wind data. *Space Weather*, *21*, e2023SW003514. https://doi.org/10.1029/2023SW003514

Foram disponibilizados outros artigos na pasta do VII desafio.

Compreensão do domínio: A física por trás de tempestades geomagnéticas, vento solar e o índice Dst não é trivial. Será necessário dedicar tempo para pesquisa e entendimento do contexto.

Limpeza e pré-processamento de dados: Os dados reais geralmente apresentam ruídos, outliers e dados faltantes. Os alunos precisarão aplicar técnicas de limpeza e escolher métodos adequados de preenchimento/tratamento para garantir a qualidade dos dados.

Engenharia de *features*: A criação de novas features a partir dos dados brutos, como médias móveis, variáveis derivadas e informações sobre a orientação do campo magnético, pode ser crucial para melhorar a performance do modelo.

Seleção e treinamento de modelos: A escolha do modelo de aprendizado de máquina mais adequado (redes neurais, RNN, modelos baseados em árvores, etc.), ajuste de hiperparâmetros e validação requerem *expertise* e experimentação.

Avaliação e interpretação: A avaliação da performance do modelo vai além da simples acurácia. Métricas específicas para séries temporais, análise dos erros e interpretação dos resultados são importantes para garantir a robustez da solução.

Trabalho em equipe: A colaboração eficaz entre alunos de diferentes níveis de conhecimento e com habilidades complementares é crucial para o sucesso.

Análise Exploratória de Dados

Distribuição dos dados: Qual a distribuição dos valores de Dst? É normal? Assimétrica? Quais os valores mínimo, máximo, média e desvio padrão?

Correlação entre variáveis: Quais as correlações entre as variáveis do vento solar (ex. velocidade, densidade, componentes do campo magnético) e o índice Dst? Existem variáveis mais fortemente correlacionadas com o Dst?

Influência do ciclo solar: Como o índice Dst e as variáveis do vento solar variam ao longo do ciclo solar de 11 anos? As tempestades geomagnéticas mais fortes ocorrem em momentos específicos do ciclo?

Dados Faltantes e Outliers:

Dados faltantes: Existem dados faltantes nas medições do vento solar ou no índice Dst? Qual a porcentagem de dados faltantes para cada variável? Como esses dados faltantes estão distribuídos no tempo?

Outliers: Existem outliers nas variáveis do vento solar ou no índice Dst? Qual a definição de outlier utilizada? Como esses outliers se comportam em relação à ocorrência de tempestades geomagnéticas?

Análise de Variáveis para Elaboração de Modelos:
Engenharia de features: É possível criar novas features a partir das variáveis existentes que melhorem a performance dos modelos? Por exemplo, calcular variáveis derivadas como a pressão do vento solar ou a componente sul do campo magnético (Bz)?

Importância das features: Quais as variáveis mais importantes para a previsão do índice Dst? É possível utilizar técnicas de seleção de features para reduzir a dimensionalidade do problema e melhorar a performance dos modelos?

Modelos de Machine Learning Aplicáveis:

Comparação de modelos: Quais modelos de aprendizado de máquina são mais adequados para a previsão do índice Dst: redes neurais, árvores de decisão, modelos lineares, etc.? Quais as vantagens e desvantagens de cada modelo nesse contexto?

Validação de modelos: Como avaliar a performance dos modelos de previsão do Dst? Quais métricas de erro são mais relevantes: erro absoluto médio, raiz do erro quadrático médio, etc.? Como garantir que o modelo generalize bem para dados não vistos?

Explicabilidade dos Modelos:

Importância das features: Quais as variáveis mais importantes para a previsão do índice Dst segundo os modelos escolhidos? Essa análise pode auxiliar na compreensão da física por trás das tempestades geomagnéticas.

Questões adicionais sobre a base de dados de previsão do Índice Dst:

Análise Exploratória de Dados:

Comparação entre ACE e DSCOVR: Existem diferenças significativas nas medições do vento solar entre os satélites ACE e DSCOVR? Essas diferenças afetam a previsão do Dst?

Análise temporal: As séries temporais do Dst e das variáveis do vento solar apresentam sazonalidade, tendências ou padrões específicos? A autocorrelação e a análise espectral podem revelar informações relevantes?

Influência da posição dos satélites: A posição dos satélites ACE e DSCOVR, registrada em satellite_pos.csv, tem influência na previsão do Dst? Seria útil incorporar essas informações nos modelos?

Modelagem e Previsão:

Previsão multi-step: Qual a melhor abordagem para previsão do Dst em múltiplos horizontes de tempo (t+1, t+2, ...)? Modelos ARIMA, redes neurais recorrentes (RNNs) ou outras arquiteturas seriam mais adequadas?

Tratamento de dados desbalanceados: Considerando que eventos extremos de Dst são raros, como lidar com o desbalanceamento dos dados durante o treinamento dos modelos? Técnicas de *oversampling, undersampling* ou funções de custo ponderadas podem ser úteis?

Incorporação de incerteza: Como incorporar a incerteza nas previsões do Dst, fornecendo intervalos de confiança ou distribuições de probabilidade em vez de valores pontuais?

Considerações Práticas:

Tempo de resposta: Qual o tempo de resposta necessário para um sistema de previsão do Dst em tempo real? As previsões precisam ser geradas com qual antecedência para serem úteis na prática?

Robustez a ruídos: Como garantir que os modelos de previsão sejam robustos a ruídos e erros nas medições do vento solar, que podem ocorrer em situações reais?

Atualização do modelo: Com que frequência os modelos de previsão do Dst precisam ser re-treinados com novos dados para manter sua acurácia ao longo do tempo?

Outras Perspectivas:

Análise de erros: Analisar os erros de previsão do modelo em relação a diferentes magnitudes de tempestades. O modelo apresenta maior dificuldade na previsão de eventos extremos?

Validação cruzada com dados externos: É possível validar os modelos utilizando dados de outras fontes, como dados de magnetômetros terrestres ou de outros satélites?

Estas perguntas adicionais exploram aspectos mais complexos da análise de dados e modelagem, com foco em desafios específicos da previsão do índice Dst e na aplicação prática dos modelos.

O índice Dst, ou Disturbance Storm Time Index, é uma medida da intensidade da componente horizontal do campo magnético terrestre em nanoTeslas (nT) na região equatorial. Ele é uma medida fundamental da atividade geomagnética e indica a severidade das tempestades geomagnéticas.

Como funciona?

O índice Dst é calculado a partir de medições do campo magnético terrestre obtidas por uma rede global de observatórios magnéticos localizados próximos ao equador. Essas medições são então processadas para remover as variações regulares do campo magnético, como as causadas pela rotação da Terra. O valor resultante representa a variação do campo magnético causada por correntes elétricas na magnetosfera da Terra, que são fortemente influenciadas pela atividade solar, principalmente pelo vento solar.

Valores e Interpretação:

Valores Positivos: Um índice Dst positivo indica condições geomagnéticas calmas.

Valores Negativos: Um índice Dst negativo indica a ocorrência de uma tempestade geomagnética. Quanto mais negativo o valor, mais intensa a tempestade.

Tempestade Fraca: -30 nT a -50 nT

Tempestade Moderada: -50 nT a -100 nT

Tempestade Forte: -100 nT a -400 nT

Tempestade Severa: Menor que -400 nT

Importância do Índice Dst:

O índice Dst é uma ferramenta crucial para:

Monitoramento da atividade geomagnética: Permite acompanhar a ocorrência e intensidade de tempestades geomagnéticas em tempo real.

Previsão de impactos: As tempestades geomagnéticas podem afetar satélites, redes elétricas, sistemas de navegação e comunicação. O índice Dst auxilia na previsão e mitigação desses impactos.

Pesquisa científica: O estudo do índice Dst e sua relação com a atividade solar são importantes para compreender a física da interação Sol-Terra.

Em resumo, o índice Dst é um indicador fundamental da "saúde" do ambiente espacial próximo à Terra, fornecendo informações valiosas sobre a ocorrência e intensidade de tempestades geomagnéticas.

Modelos de forma geral

Regressão Linear:
Pode ser útil para avaliar a relação linear entre uma variável de entrada (ou um conjunto delas) e a variável de saída.

Regressão de Árvore de Decisão:
É eficaz para lidar com relacionamentos não lineares e interações entre variáveis.

Random Forest:
Uma extensão das árvores de decisão, o Random Forest pode ser útil para capturar padrões mais complexos e aumentar a precisão da previsão.

Gradient Boosting:
O Gradient Boosting tem sido amplamente utilizado em competições de ciência de dados e em aplicações do mundo real devido à sua capacidade de produzir modelos de alta precisão e sua flexibilidade para lidar com uma variedade de tipos de dados e problemas de aprendizado supervisionado. Exemplos populares de bibliotecas que implementam Gradient Boosting incluem XGBoost, LightGBM e CatBoost

Algoritmos como XGBoost ou LightGBM:

São poderosos para lidar com dados complexos e são eficazes na previsão de eventos raros

Redes neurais:
Podem ser usadas para aprender padrões complexos e não lineares nos dados.

> **Redes Neurais Densas (FNN - Feedforward Neural Networks):**
> As redes neurais densas, ou *feedforward neurais networks*, são um dos tipos mais básicos de redes neurais. Elas consistem em uma série de camadas de neurônios, onde cada neurônio em uma camada está conectado a todos os neurônios na camada seguinte. Essas redes são eficazes em aprender representações complexas dos dados e são amplamente utilizadas em uma variedade de problemas de aprendizado supervisionado e não supervisionado.
>
> **Redes Neurais Recorrentes (RNN):**
> As redes neurais recorrentes são projetadas para lidar com dados sequenciais, onde a

ordem e a dependência temporal dos dados são importantes. Elas possuem conexões retroalimentadas que permitem que informações sejam mantidas em memória ao longo do tempo. Isso as torna ideais para tarefas como previsão de séries temporais, processamento de linguagem natural e análise de sequências de dados.

Redes Neurais Profundas (DNN):

As redes neurais profundas são modelos com múltiplas camadas ocultas entre a entrada e a saída. Elas são capazes de aprender representações hierárquicas dos dados, o que as torna poderosas em lidar com problemas complexos e de alta dimensionalidade. As DNNs são usadas em uma variedade de domínios, incluindo processamento de linguagem natural, reconhecimento de fala, visão computacional, análise de séries tenporais e muito mais.

Redes Generativas

são aplicáveis em ciência de dados e têm uma ampla gama de aplicações. As redes generativas são uma classe de modelos de redes neurais projetados para aprender a distribuição de probabilidade dos dados de treinamento e, em seguida, gerar novos dados que são semelhantes aos dados de treinamento. Aqui estão algumas aplicações comuns de redes generativas em ciência de dados:

Reconstrução de Dados Ausentes:

Em muitos conjuntos de dados, podem existir lacunas ou valores ausentes. As redes generativas podem ser utilizadas para reconstruir dados ausentes, preenchendo lacunas de forma plausível com base nas informações disponíveis. Isso pode ser útil em várias aplicações, como preenchimento de valores ausentes em conjuntos de dados de séries temporais ou reconstrução de partes danificadas de imagens médicas.

Amostragem de Dados:

As redes generativas podem ser usadas para amostrar novos dados que sigam uma distribuição de probabilidade semelhante à dos dados de treinamento. Isso pode ser útil em várias aplicações, como geração de cenários hipotéticos para análise de risco, criação de conjuntos de dados sintéticos para treinamento de modelos e simulação de dados para testes de algoritmos.

Aprimoramento de Dados:

As redes generativas também podem ser usadas para aprimorar dados existentes, adicionando detalhes ou corrigindo artefatos. Por exemplo, as GANs podem ser usadas para remover ruídos de imagens, aumentar a resolução de imagens de baixa qualidade ou reconstruir informações perdidas em sinais de áudio.

Transformers

É outra arquitetura de rede neural que é aplicável em ciência de dados, especialmente em tarefas de processamento de linguagem natural (PLN) e sequências de dados. Os modelos baseados em Transformers têm sido revolucionários nessas áreas devido à sua capacidade de capturar relacionamentos de longo alcance em sequências de dados e lidar com tarefas complexas com alto desempenho. Aqui estão algumas das principais aplicações dos Transformers em ciência de dados:

Modelagem de Sequências Temporais:

Além de PLN, os Transformers também podem ser aplicados em tarefas de modelagem de séries temporais, como previsão de séries temporais, detecção de anomalias e preenchimento de lacunas em dados temporais.

Regressão LASSO ou Ridge (modelos de regressão com reguladores).

Úteis para lidar com problemas de multicolinearidade ou reduzir a dimensionalidade das variáveis. Por exemplo, você pode aplicar regressão LASSO para selecionar as variáveis mais importantes relacionadas à segurança da aviação e construir um modelo mais simples e interpretável.

Regressão Logística: A Regressão Logística é um modelo de aprendizado supervisionado usado para problemas de classificação binária, onde a variável dependente é categórica. A Regressão Logística produz uma saída probabilística que pode ser interpretada como a probabilidade de um evento ocorrer.

SVM (Support Vector Machine): SVM é um modelo de aprendizado supervisionado que pode ser usado para classificação ou regressão. No contexto da segurança da aviação, o SVM pode ser aplicado para problemas de classificação, como prever se um determinado evento é um incidente leve, um acidente grave ou outra categoria de evento. O SVM é eficaz na identificação de fronteiras de decisão não lineares, o que pode ser útil para lidar com dados complexos e não lineares.

Validação de Modelos:

A validação de modelos é uma etapa fundamental em qualquer projeto de ciência de dados. Ela nos permite avaliar o desempenho e a eficácia do modelo construído, garantindo que ele seja capaz de fazer previsões precisas em situações do mundo real. A utilização de métricas apropriadas desempenha um papel crucial nesse processo, fornecendo uma base objetiva para a análise dos resultados.

Após a construção de um modelo para previsão a etapa seguinte é avaliar o quão bem ele se comporta em relação aos dados de teste ou dados não vistos. A validação é um processo crítico para determinar a capacidade do modelo de generalizar informações a partir dos dados de treinamento para novos dados. Aqui estão alguns aspectos-chave a serem considerados:

1. **Conjunto de Teste Adequado**: Reserve uma porção dos seus dados para criar um conjunto de teste. Isso deve ser feito de forma a representar adequadamente os cenários do mundo real que o modelo enfrentará. O conjunto de teste deve ser independente dos dados de treinamento. **Avalie a Cross-Validation (Validação Cruzada):** técnicas de validação cruzada, como k-fold cross-validation, para avaliar a robustez do seu modelo pode trazer resultados melhores. Isso envolve dividir seus dados em k subconjuntos, treinar e testar o modelo k vezes e calcular métricas médias de desempenho.

Métricas de Avaliação: Escolha métricas apropriadas para avaliar o desempenho do modelo. Cada métrica fornece uma perspectiva diferente sobre o desempenho do modelo. Cada algoritmo preditivo pode ser avaliado por meio de métricas específicas que fornecem *insights* valiosos sobre diferentes aspectos do seu desempenho. Aqui estão algumas das métricas mais comuns e suas aplicações em diferentes algoritmos:

- **Acurácia (*Accuracy*):**
- Est métrica mede a proporção de previsões corretas em relação ao total de previsões feitas pelo modelo. É uma métrica geralmente utilizada para avaliar modelos de classificação binária e multiclasse.
- **Precisão (**Precision**):**
- A precisão mede a proporção de exemplos classificados como positivos que realmente são positivos. É particularmente útil quando o foco está na minimização de falsos positivos. É calculada como a proporção de verdadeiros positivos (TP) em relação ao total de previsões positivas (TP + Falsos positivos (FP)).
- **Revocação (*Recall ou Sensibilidade*):**
- A revocação mede a proporção de exemplos positivos que foram corretamente identificados pelo modelo. É especialmente importante quando a identificação de todos os exemplos positivos é uma prioridade. É calculada como a proporção de verdadeiros positivos (TP) em relação ao total de exemplos positivos na base de dados (TP + Falsos

negativos (FN)).

- **F1-Score:**
- O F1-score é a média harmônica entre precisão e revocação. Ele fornece uma métrica única que equilibra essas duas medidas e é útil quando não há uma clara preferência entre elas. É calculado como 2 * (Precisão * Revocação) / (Precisão + Revocação).
- **Área sob a Curva ROC (ROC AUC):**
- Esta métrica é comumente usada para avaliar a capacidade de um modelo de classificação em distinguir entre classes positivas e negativas. A curva ROC é uma representação gráfica da taxa de verdadeiros positivos (Revocação) em relação à taxa de falsos positivos (1 - Especificidade), e a área sob a curva (ROC AUC) é calculada para medir a qualidade geral do modelo.
- **Erro Médio Absoluto (MAE) e Erro Quadrático Médio (MSE):**
- Essas métricas são comumente usadas para avaliar modelos de regressão. O MAE mede a média das diferenças absolutas entre as previsões do modelo e os valores reais. O MSE mede a média dos quadrados dessas diferenças. Ambos fornecem uma medida da qualidade das previsões do modelo.
- **Erro Médio Percentual Absoluto (MAPE):**
- O MAPE mede a média das proporções de erro absoluto em relação aos valores reais. É uma métrica útil para entender a precisão relativa do modelo em relação às observações reais, especialmente em contextos em que a magnitude das previsões é importante.
- **R^2 (Coeficiente de Determinação):**
- O R^2 é uma medida que indica a proporção da variabilidade dos valores de saída que é explicada pelo modelo. Ele fornece uma medida da adequação do modelo aos dados observados e é uma métrica importante em problemas de regressão.
- **Comparação com *Benchmark***: Compare o desempenho do seu modelo com um *benchmark* ou baseline. Isso pode ser um modelo simples, como uma média dos preços, que fornece um ponto de referência para avaliar se o seu modelo está realmente trazendo melhorias.

1. ***Overfitting e Underfitting***: Esteja atento a sinais de overfitting (modelo se ajustando demais aos dados de treinamento) ou underfitting (modelo muito simplificado). A curva de aprendizado, que mostra o desempenho do modelo em relação ao tamanho do conjunto de treinamento, pode ser útil para identificar esses problemas.

- **Visualização dos Resultados***:* Além das métricas, utilize visualizações, como gráficos de dispersão (scatter plots) das previsões versus valores reais, para entender como o modelo está se comportando em diferentes partes do espaço de recursos.
- **Ajuste de Hiperparâmetros**: Se o desempenho do modelo não atender às expectativas, considere ajustar os hiperparâmetros do modelo e repetir o processo de validação.
- **Interpretação dos Resultados**: Não apenas avalie o desempenho quantitativamente, mas também interprete os resultados qualitativamente. Pergunte-se se as previsões fazem sentido do ponto de vista do negócio e se podem ser úteis para os usuários finais.
- **Documentação Completa** Documente todas as etapas do processo de validação, incluindo as métricas utilizadas, os resultados obtidos e as decisões tomadas com base na validação. Isso é essencial para comunicar os resultados aos stakeholders e para futuras referências.
- **Melhorias Iterativas:** A validação não é uma etapa única; é um processo iterativo. Continue refinando e aprimorando seu modelo com base nos resultados da validação.

Em resumo, a validação de modelos e o uso de métricas são partes essenciais do ciclo de desenvolvimento de modelos de aprendizado de máquina. Essas etapas garantem que seu modelo seja preciso, confiável e útil para tomar decisões informadas no mundo real. Portanto, dedique tempo e atenção a essa fase crítica do projeto de ciência de dados.

Linguagens usadas na ciência de dados

algumas linguagens de programação se destacam devido à sua flexibilidade, eficiência e rica variedade de bibliotecas e ferramentas disponíveis para análise e manipulação de dados. Aqui estão algumas das linguagens mais proeminentes na ciência de dados:

Python:
Python é uma das linguagens mais populares para ciência de dados devido à sua sintaxe simples e legível, grande comunidade de desenvolvedores e vasto ecossistema de bibliotecas especializadas. Bibliotecas como NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, TensorFlow e PyTorch são amplamente utilizadas para análise de dados, visualização, modelagem estatística, aprendizado de máquina e deep learning.
R:
R é uma linguagem de programação especialmente projetada para análise estatística e visualização de dados. Ela oferece uma ampla gama de pacotes e bibliotecas dedicadas à estatística, modelagem, análise de séries temporais e visualização de dados. Pacotes populares incluem ggplot2, dplyr, tidyr, caret, survival e many more.
SQL (Structured Query Language):
Embora não seja uma linguagem de programação completa, o SQL é essencial para trabalhar com bancos de dados relacionais. É amplamente utilizado para consultas, manipulação e agregação de dados em bancos de dados relacionais como MySQL, PostgreSQL, SQLite e muitos outros.
Julia:
Julia é uma linguagem de programação de alto desempenho, projetada para análise numérica e científica. Ela combina a facilidade de uso do Python com o desempenho de linguagens como C e Fortran. Julia é conhecida por sua capacidade de escrever código rápido e eficiente para computação numérica e paralela.
Scala:
Scala é uma linguagem de programação multiparadigma que combina programação funcional e orientada a objetos. Ela é popular entre os cientistas de dados que trabalham com o Apache Spark, uma plataforma de processamento de big data que permite análise distribuída em grandes conjuntos de dados.
Java:
Java é uma linguagem de programação amplamente utilizada em sistemas de grande escala e aplicativos empresariais. Apesar de não ser tão popular quanto Python ou R para ciência de dados, Java é comumente usada em empresas onde há uma infraestrutura existente baseada em Java e para integração de sistemas de dados.

Linguagem Python

Tarefa	Bibliotecas e Comandos Python
Análise Exploratória de Dados (AED)	*Pandas, NumPy, Matplotlib, Seaborn, Plotly*
Pré-processamento de Dados	*scikit-learn.preprocessing (por exemplo, StandardScaler)*
Modelagem Preditiva	*scikit-learn (por exemplo, sklearn.linear_model, sklearn.ensemble, sklearn.neural_network), TensorFlow, PyTorch*
Análise de Sobrevivência	*lifelines (para modelos de risco proporcionais de Cox e Kaplan-Meier), scikit-survival (para modelagem de sobrevivência com scikit-learn), survival (para modelos de risco proporcionais de Cox)*
Validação de Modelo	*scikit-learn.model_selection (por exemplo, train_test_split, cross_val_score)*
Métricas de Avaliação do Modelo	*scikit-learn.metrics (por exemplo, accuracy_score, precision_score, recall_score, f1_score, roc_auc_score para classificação; mean_squared_error, mean_absolute_error, r2_score para regressão)*
Otimização de Hiperparâmetros	*scikit-learn.model_selection.GridSearchCV, scikit-learn.model_selection.RandomizedSearchCV*
Visualização de Dados	*Matplotlib, Seaborn, Plotly, pandas.plotting*
Acesso a Dados	*Pandas, NumPy, scikit-learn.datasets, TensorFlow Datasets*

ENTREGAS E APRESENTAÇÃO DO DESAFIO PARA A BANCA

Os modelos desenvolvidos devem ser apresentados no dia 16/10/2024, das 19h00min às 22h00min, em reunião na sala teams do desafio.

A apresentação do desafio para a banca de avaliação é uma etapa crucial para comunicar de forma eficaz os objetivos, a metodologia e os resultados do projeto aos avaliadores. Para garantir que a apresentação seja clara e informativa, aqui estão algumas sugestões sobre o material a ser entregue e a estrutura da apresentação, considerando um limite de 15 minutos por equipe + 5 min banca.

Observação importante: Certifique-se de praticar a apresentação várias vezes para garantir que você se mantenha dentro do limite de tempo. Mantenha o foco nos aspectos mais relevantes e interessantes do projeto para cativar a atenção da banca. Boa sorte com sua apresentação!

Material a ser entregue: dia 16/10, até as 8h00min, em canal privado na sala teams para cada equipe.

Relatório Descritivo: Comece entregando um documento escrito que descreva detalhadamente o desafio, o contexto, os dados disponíveis, as etapas do projeto e as conclusões. Este relatório deve ser claro e conciso, fornecendo uma visão geral abrangente do projeto. Coloque aqui o endereço do gitHub do projeto da equipe. O relatório deve ser colocado na pasta da equipe até 1h antes do início das apresentações.

Código-Fonte: Forneça acesso ao código-fonte do seu projeto, preferencialmente em um repositório online, como **GitHub**. Certifique-se de que o código esteja bem comentado e organizado para facilitar a revisão pela banca.

Os membros da Banca escolherão os 6 melhores trabalhos para a apresentação, até as 12h00min do dia 16/10.

Estrutura da Apresentação (15 minutos) equipe + 5 min banca (sugestão)

Slides de Apresentação: Prepare slides de apresentação que destaquem os principais pontos do projeto. Cada equipe deve ter um conjunto de slides que cubra os aspectos mais importantes do desafio. Certifique-se de que os slides sejam visualmente atraentes e fáceis de seguir. **Os slides devem ser colocados na pasta da equipe até 1h antes do início das apresentações.**

Sugestão da apresentação.

Introdução (1-2 minutos):

Cumprimente a banca e os presentes.

Apresente brevemente a equipe, destacando os membros-chave.

Contextualize o desafio e a importância do problema a ser abordado.

Entendimento do Desafio (1-2 minutos):

Explique em detalhes o problema que está sendo abordado.

Descreva o escopo do projeto e os objetivos específicos.

Apresente os dados disponíveis e qualquer limitação conhecida.

Preparação dos Dados (2-3 minutos):

Fale sobre as etapas de limpeza e preparação dos dados.

Destaque quaisquer desafios enfrentados durante essa fase.

Mostre exemplos de como os dados estão estruturados após a preparação.

Análise Exploratória (2-3 minutos):

Apresente os *insights* obtidos durante a análise exploratória.

Destaque visualizações, gráficos e estatísticas relevantes.

Explique como esses *insights* influenciaram as decisões do projeto.

Modelagem e Algoritmos (2-3 minutos):

Descreva os modelos de aprendizado de máquina ou técnicas utilizadas.

Explique como os modelos foram treinados e avaliados.

Apresente as métricas de avaliação de desempenho.

Validação e Resultados (2-3 minutos):
Compartilhe os resultados obtidos, incluindo métricas de desempenho.
Discuta a validação do modelo e como ele se comporta em situações reais.
Compare o desempenho do modelo com *benchmark*s, se aplicável.

Conclusão (1-2 minutos):
Recapitule os principais resultados e descobertas.
Faça uma breve reflexão sobre as lições aprendidas durante o projeto.
Discuta as implicações práticas e possíveis próximos passos.

Perguntas da Banca (3-5 minutos):
Abra espaço para perguntas e comentários da banca.
Esteja preparado para responder a perguntas técnicas e conceituais.

Agradecimento e Encerramento (1 minuto):
Agradeça à banca e aos presentes pela atenção.

Bom trabalho!
Comissão organizadora
Maria José Pereira Dantas
José Elmo de Menezes