PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS (PUC GOIÁS)

9 CCTI PUC GO 2023/2

MESTRADO EM ENGENHARIA DE PRODUÇÃO E SISTEMAS (PPGEPS)

NÚCLEO DE MATEMÁTICA DO NEPE

V DESAFIO EM CIÊNCIAS DE DADOS
PREDIÇÃO DE PREÇOS NO MERCADO DE HOSPEDAGEM VIRTUAL
BASE DE DADOS: AIRBNB 2018 A 2020 RIO DE JANEIRO

Base de dados pública: Airbnb Rio de Janeiro 2018 a 2020 (pasta V desafio na sala teams)

CONTEXTUALIZAÇÃO

De acordo com García et al, (2020), atualmente, a maior parte dos processos de informações turísticas são realizados eletronicamente, incluindo o aluguel de hospedagens. Os clientes deixam suas impressões digitais em grande parte das atividades realizadas tanto no planejamento da viagem como durante e após, mas também através de comentários sobre diferentes plataformas. Consequentemente, uma grande quantidade de dados sobre as necessidades e comportamentos dos clientes, bem como a sua percepção dos serviços, são armazenados em várias fontes. Com isso, surgem as aplicações como o Airbnb, que fornecem as mais variadas opções de hospedagens para os viajantes, trazendo facilidade e substituindo parte dos intermediários do plano, tais como as agências de turismo ou corretores imobiliários.

O Airbnb é, desde sua origem, uma plataforma online para mediação de hospedagens de particulares em suas próprias propriedades. Sua história, do nascimento e crescimento até se tornar a gigante do mercado de hospedagem que é hoje, foi registrada por Gallagher (2019). A ideia surgiu em 2007 quando os jovens Brian Chesky e Joe Gebbia, morando em São Francisco (Califórnia, EUA) precisavam de uma renda extra para pagar o aluguel. Com a realização de um congresso internacional na cidade perceberam a oportunidade de oferecer três colchões de ar em casa para congressistas visto que a demanda no mercado hoteleiro estava muito alta. Nascia então a AirBed&Breakfast (Brossat, 2019; Gallagher, 2019). A ideia de oferecer um lugar para viajantes ficarem em casa não era nova. O Couchsurfing já existia com esse mesmo conceito desde 1999, porém sem a necessidade de pagamento.

Referências

BROSSAT, I. (2019). Airbnb, la ciudad uberizada (1a ed., Vol. 1). Pamplona: Katakrak Liburuak.
GALLAGHER, L. (2019). La historia de Airbnb: como tres chicos comunes transtornaron

una industria, ganaron miles de milones y crearon gran controversia. (1a ed,. Vol 1). Bogotá: Conecta.
GARCíA, J. Álvarez et al. Big data and tourism research: Measuring research impact. Quality and Quantity, 09 2020.

Neste projeto iremos analisar os dados referentes à cidade do Rio de Janeiro e ver quais insights podem ser extraídos a partir desses dados brutos.

DICIONÁRIO DE DADOS

O conjunto de dados "listings.csv" contém uma série de campos que fornecem informações sobre as listagens de propriedades no Airbnb no Rio de Janeiro. Aqui está uma descrição dos campos presentes no conjunto de dados:

id: Um número inteiro que representa o identificador único da listagem no Airbnb.

listing_url: Um campo de texto que contém o URL da listagem no Airbnb.

scrape_id: Um número grande (bigint) que faz parte do processo de "raspagem" (scrape) do Airbnb. Este campo está relacionado ao processo de coleta de dados.

last_scraped: Uma data e hora (datetime) no formato UTC que indica a data e hora em que a listagem foi "raspada" (scraped) ou coletada.

source: Um campo de texto que indica a fonte da listagem, podendo ser "neighbourhood search" (pesquisa por bairro) ou "previous scrape" (raspagem anterior).

name: O nome da listagem, geralmente um título descritivo.

description: Uma descrição detalhada da listagem.

neighborhood_overview: Uma descrição do bairro onde a listagem está localizada, fornecida pelo anfitrião.

picture_url: Um URL que aponta para a imagem regular da listagem hospedada no Airbnb.

host_id: Um número inteiro que representa o identificador único do anfitrião ou usuário no Airbnb.

host_url: Um campo de texto que contém o URL da página do anfitrião no Airbnb.

host_name: O nome do anfitrião, geralmente apenas o primeiro nome.

host_since: Uma data que indica quando o anfitrião ou usuário foi criado no Airbnb.

host_location: A localização auto-relatada do anfitrião.

host_about: Uma descrição sobre o anfitrião.

host_response_time: O tempo médio de resposta do anfitrião.

host_response_rate: A taxa de resposta do anfitrião.

host_acceptance_rate: A taxa de aceitação de pedidos de reserva pelo anfitrião.

host_is_superhost: Um campo booleano (t=true; f=false) que indica se o anfitrião é um "superhost".

host_thumbnail_url: Um URL que aponta para a imagem em miniatura do anfitrião.

host_picture_url: Um URL que aponta para a imagem do anfitrião.

host_neighbourhood: O bairro onde o anfitrião está localizado.

host_listings_count: O número de listagens que o anfitrião possui, conforme calculado pelo Airbnb.

host_total_listings_count: O número total de listagens que o anfitrião possui, também conforme calculado pelo Airbnb.

host_verifications: As verificações feitas pelo anfitrião.

host_has_profile_pic: Um campo booleano que indica se o anfitrião possui uma foto de perfil.

host_identity_verified: Um campo booleano que indica se a identidade do anfitrião foi verificada.

neighbourhood: O bairro onde a listagem está localizada.

neighbourhood_cleansed: O bairro, geocodificado usando latitude e longitude.

neighbourhood_group_cleansed: O grupo de bairros, geocodificado usando latitude e longitude.

latitude: A latitude da localização da listagem.

longitude: A longitude da localização da listagem.

property_type: O tipo de propriedade, selecionado pelo anfitrião.

room_type: O tipo de quarto, que pode ser "Entire home/apt" (Casa/apartamento inteiro), "Private room" (Quarto privado), "Shared room" (Quarto compartilhado) ou "Hotel".

accommodates: A capacidade máxima da listagem.

bathrooms: O número de banheiros na listagem.

bathrooms_text: O número de banheiros na listagem, em formato de texto.

bedrooms: O número de quartos na listagem.

beds: O número de camas na listagem.

amenities: As comodidades oferecidas na listagem, em formato JSON.

price: O preço diário em moeda local.

minimum_nights: O número mínimo de noites necessárias para reservar a listagem.

maximum_nights: O número máximo de noites permitido para reservar a listagem.

minimum_minimum_nights: O menor valor de mínimo de noites do calendário (olhando 365 noites no futuro).

maximum_minimum_nights: O maior valor de mínimo de noites do calendário (olhando 365 noites no futuro).

minimum_maximum_nights: O menor valor de máximo de noites do calendário (olhando 365 noites no futuro).

maximum_maximum_nights: O maior valor de máximo de noites do calendário (olhando 365 noites no futuro).

minimum_nights_avg_ntm: O valor médio de mínimo de noites do calendário (olhando 365 noites no futuro).

maximum_nights_avg_ntm: O valor médio de máximo de noites do calendário (olhando 365 noites no futuro).

calendar_updated: A data da última atualização do calendário.

has_availability: Um campo booleano que indica se a listagem possui disponibilidade.

availability_30, availability_60, availability_90, availability_365: A disponibilidade da listagem para 30, 60, 90 e 365 dias no futuro, determinada pelo calendário.

calendar_last_scraped: A data da última raspagem do calendário.

number_of_reviews: O número total de avaliações que a listagem possui.

number_of_reviews_ltm: O número de avaliações que a listagem recebeu nos últimos 12 meses.

number_of_reviews_l30d: O número de avaliações que a listagem recebeu nos últimos 30 dias.

first_review: A data da primeira avaliação.

last_review: A data da última avaliação.

review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value: As pontuações de avaliação da listagem em várias categorias.

license: O número de licença, permissão ou registro associado à listagem.
instant_bookable: Um campo booleano que indica se os hóspedes podem reservar a listagem automaticamente, sem a necessidade de aprovação do anfitrião. Isso pode ser um indicador de listagens comerciais.

calculated_host_listings_count: O número de listagens que o anfitrião possui no momento da raspagem atual, na geografia da cidade/região.

calculated_host_listings_count_entire_homes: O número de listagens de casas/apartamentos

inteiros que o anfitrião possui na raspagem atual, na geografia da cidade/região.

calculated_host_listings_count_private_rooms: O número de listagens de quartos privados que o anfitrião possui na raspagem atual, na geografia da cidade/região.

calculated_host_listings_count_shared_rooms: O número de listagens de quartos compartilhados que o anfitrião possui na raspagem atual, na geografia da cidade/região.

reviews_per_month: O número médio de avaliações que a listagem recebe ao longo de sua vida.

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

Seu objetivo é construir um modelo de previsão de preço o qual permita que uma pessoa comum que possui um imóvel saber quanto ela deve cobrar pela diária do seu imóvel. Ou ainda, para o locador comum, dado o imóvel que ele está buscando, ajudá-lo a saber se aquele imóvel está com preço atrativo (abaixo da média para imóveis com as mesmas características) ou não.

A única informação que você possui são algumas basesde dados que possuem os preços dos imóveis obtidos e suas respectivas características em cada mês. As bases vão de abril de 2018 a maio de 2020, com exceção de junho de 2018 que não possui base de dados.

FAÇA LEITURAS DO MATERIAL INDICADO NA PASTA V DESAFIO – SALA TEAMS

PREPARAÇÃO DOS DADOS

Nesta etapa, deve-se realizar a preparação dos dados e consequentemente a limpeza, realizando a integração, formatação e construção de novos dados, para que possamos selecionar os que serão utilizados na construção de nosso modelo. Verificar a qualidade de um *dataset* está diretamente relacionada à quantidade de valores ausentes. É importante entender logo no início se esses valores nulos são significativos comparados ao total de entradas. Pode-se analisar **do tipo de distribuição das variáveis numéricas, fazer avaliação de** ***outliers*** **presentes.**

O curso de introdução a ciência de dados com python 02/10/2023 Prof. Wanderlei, com gravação e github

Link para o minicurso

https://pucdegoias.sharepoint.com/sites/DesafioemCinciasdeDados1/Documentos%20Compartilhados/General/Recordings/Reuni%C3%A3o%20em%20_Geral_-20231002_201309-

Grava%C3%A7%C3%A3o%20de%20Reuni%C3%A3o.mp4?web=1

github

https://wmpjrufg.github.io/MCOMP002/

ANÁLISE EXPLORATÓRIA

Nesta fase, deve-se analisar os dados e ver quais *insights* podem ser extraídos a partir desses dados brutos.

Pode-se identificar os comportamentos médios e discrepantes dos preços das acomodações, realizaremos a comparação desses valores, investigaremos a interdependência entre as variáveis, tipos de acomodações e procuraremos identificar tendências por bairros.

Utilizar técnicas da análise exploratória, como por exemplo:

- Histogramas;
- Box Plot;
- Matriz de correlação;
- Heatmap;
- Entre outras.

A partir dessas análises encontraremos parâmetros para cada uma das variáveis do nosso projeto.

O curso de introdução a ciência de dados com python 02/10/2023 Prof. Wanderlei, com gravação e github

Link para o minicurso

https://pucdegoias.sharepoint.com/sites/DesafioemCinciasdeDados1/Documentos%20Compartilhados/General/Recordings/Reuni%C3%A3o%20em%20_Geral_-20231002_201309-Grava%C3%A7%C3%A3o%20de%20Reuni%C3%A3o.mp4?web=1

github

https://wmpjrufg.github.io/MCOMP002/

SELEÇÃO DE VARIÁVEIS

Segue abaixo algumas variáveis do arquivo:

Variáveis Numéricas (escolha um método para avaliar o impacto das variáveis)

accommodates: Esta variável pode ser importante, pois indica a capacidade máxima da acomodação, o que geralmente afeta o preço. Propriedades maiores tendem a ter preços mais altos.

availability_30, availability_60, availability_90, availability_365: Essas variáveis podem indicar a disponibilidade da acomodação em diferentes períodos e podem influenciar os preços, pois a disponibilidade afeta a demanda.

number_of_reviews: O número de avaliações que uma acomodação recebeu pode ser um indicador da popularidade e da qualidade, o que pode afetar os preços.

review_scores_rating: A pontuação média de avaliação dos hóspedes pode ser uma medida da qualidade da acomodação e influenciar os preços.

bedrooms e beds: O número de quartos e camas na acomodação pode influenciar os preços, pois está relacionado à capacidade de acomodação.

bathrooms: O número de banheiros também pode ser um fator relevante, pois está relacionado ao conforto dos hóspedes.

minimum_nights e maximum_nights: As políticas de estadia mínima e máxima podem afetar os preços, já que podem limitar a flexibilidade dos hóspedes.

Variáveis Categóricas:

neighbourhood_cleansed: O bairro onde a acomodação está localizada pode ser um fator importante na determinação do preço, já que alguns bairros podem ser mais valorizados do que outros.

room_type: O tipo de quarto (por exemplo, "Entire home/apt," "Private room," "Shared room," ou "Hotel") pode ter um impacto significativo nos preços, pois a privacidade e o espaço variam entre essas categorias.

property_type: O tipo de propriedade (por exemplo, casa, apartamento, pousada) pode influenciar os preços, uma vez que diferentes tipos de propriedades oferecem diferentes comodidades e tamanhos.

MODELAGEM E ALGORITMOS (MODELOS DE PREVISÃO DE PREÇOS)

Redes neurais, também conhecidas como redes neurais artificiais (RNAs), são modelos computacionais inspirados no funcionamento do cérebro humano. Elas são uma parte fundamental da área de aprendizado de máquina e da inteligência artificial. Aqui está um pequeno resumo sobre redes neurais:

O que são Redes Neurais:

Redes neurais são sistemas de processamento de informações baseados na estrutura e no funcionamento dos neurônios no cérebro humano.

Elas consistem em camadas de unidades computacionais interconectadas, chamadas neurônios artificiais ou unidades, que processam informações de entrada e produzem saídas.

Funcionamento Básico:

As redes neurais são projetadas para aprender a partir de dados. Durante o treinamento, ajustam seus parâmetros internos para fazer previsões ou classificações com base em exemplos de entrada e saída fornecidos.

Cada conexão entre neurônios é associada a um peso que determina a importância da informação transmitida.

As redes neurais podem ser alimentadas com uma grande quantidade de dados e são capazes de identificar padrões complexos e realizar tarefas como classificação, regressão, reconhecimento de padrões e processamento de linguagem natural.

Arquitetura de Redes Neurais:

As redes neurais podem ter diversas arquiteturas, incluindo redes de alimentação direta (feedforward), redes recorrentes (que possuem laços de retroalimentação) e redes convolucionais (especializadas em dados com estrutura espacial, como imagens).

A camada de entrada recebe os dados de entrada, enquanto as camadas intermediárias (ocultas) processam informações e a camada de saída produz os resultados finais.

Redes profundas (com várias camadas intermediárias) são chamadas de redes neurais profundas ou redes neurais profundas (Deep Neural Networks - DNNs).

Aplicações das Redes Neurais:

As redes neurais têm uma ampla gama de aplicações, incluindo visão computacional, processamento de linguagem natural, reconhecimento de fala, previsão do mercado financeiro, previsão do mercado imobiliário, diagnóstico médico, veículos autônomos, jogos de estratégia e muito mais.

Elas têm sido particularmente eficazes em lidar com dados complexos e não lineares.

Desafios e Considerações:

O treinamento de redes neurais pode exigir grandes volumes de dados e poder computacional.

O ajuste dos hiperparâmetros, como taxas de aprendizado e arquitetura da rede, é um desafio crucial.

O overfitting (sobreajuste) é uma preocupação comum, onde a rede se adapta muito bem aos dados de treinamento, mas não generaliza bem para novos dados.

Em resumo, redes neurais são modelos de aprendizado de máquina inspirados na estrutura do cérebro humano que são capazes de aprender e realizar tarefas complexas. Elas têm uma ampla gama de aplicações em diversos campos e são uma ferramenta poderosa para lidar com problemas de análise de dados e reconhecimento de padrões.

No grupo de whatsapp será disponibilizado um link com uma literatura em redes neurais.

Uso de séries temporais

Incorporar o entendimento de séries temporais em seu desafio de previsão de preços de acomodações pode ser uma adição valiosa, especialmente se seus dados contêm informações temporais granulares (o que é verdade, os dados estão disponíveis mensalmente, com um mês faltante). Comece analisando a série temporal dos preços das acomodações ao longo do tempo. Isso pode revelar tendências sazonais, padrões de aumento ou queda de preços em determinadas épocas do ano, feriados, eventos especiais, etc. Use gráficos de séries temporais para visualizar esses padrões.

Integrar a análise de séries temporais em seu desafio de previsão de preços pode enriquecer sua compreensão dos dados e melhorar a precisão das previsões, especialmente se houver padrões temporais significativos nos preços das acomodações. Certifique-se de documentar claramente como as informações temporais estão sendo utilizadas em seu modelo e como elas afetam suas previsões.

Etapas do modelagem

Nesta etapa, deve-se utilizar as técnicas de modelagem em Data Science (DS). Construir possíveis modelos para a DS, com a análise exploratória dos dados (tabelas, gráficos), outras técnicas disponíveis, e por fim, o modelo será avaliado quanto à qualidade das suas previsões (métricas de avaliação do modelo).

Como sugestão

A seleção do modelo de previsão depende da natureza dos seus dados e dos objetivos específicos do seu projeto. Aqui estão alguns modelos que você pode considerar para prever os preços de hospedagem:

Regressão Linear: É um modelo simples que assume uma relação linear entre as variáveis de entrada e a variável de saída (preço). Pode ser um bom ponto de partida.

Regressão de Árvore de Decisão: Este modelo pode lidar com relacionamentos não lineares e interações entre variáveis. Pode ser útil para capturar padrões complexos nos dados.

Random Forest: É uma extensão das árvores de decisão que combina várias árvores para melhorar a precisão da previsão. Pode ser robusto e eficaz.

Gradient Boosting: Algoritmos como o Gradient Boosting (por exemplo, XGBoost ou LightGBM) são poderosos e frequentemente vencedores em competições de ciência de dados. Eles podem lidar bem com dados complexos.

Redes Neurais: Se você tiver um grande volume de dados, as redes neurais profundas podem ser uma opção. Elas podem aprender padrões complexos, mas também requerem um grande conjunto de dados.

Regressão LASSO ou Ridge: Se você deseja lidar com problemas de multicolinearidade ou reduzir a dimensionalidade das variáveis, a regressão LASSO (L1) ou Ridge (L2) pode ser útil.

Se você tiver várias variáveis de entrada relevantes, a regressão linear múltipla permite incorporar essas variáveis para criar um modelo de regressão mais complexo.

Redes Neurais Densas (FNN):

Você pode criar uma rede neural densa (também conhecida como feedforward neural network) com várias camadas ocultas. Isso permite que o modelo aprenda representações mais complexas dos dados e capture relacionamentos não lineares entre as variáveis de entrada e os preços.

Redes Neurais Recorrentes (RNN):

Se você estiver trabalhando com dados sequenciais, como avaliações de hóspedes ao longo do tempo, as redes neurais recorrentes (por exemplo, LSTM ou GRU) podem ser úteis para capturar dependências temporais e criar um modelo de previsão.

Redes Neurais Convolucionais (CNN):

Se você estiver trabalhando com imagens das acomodações, as redes neurais convolucionais podem ser usadas para extrair características relevantes das imagens e incorporá-las ao modelo de regressão.

Redes Neurais Profundas (DNN):

Redes neurais profundas com várias camadas podem ser usadas para aprender representações complexas dos dados. Isso pode ser útil se você tiver muitas variáveis de entrada.

Redes Neurais de Atenção:

As redes neurais de atenção podem ser usadas para dar mais peso a determinadas características ou variáveis de entrada que são mais relevantes para a previsão de preços.

Redes Neurais Generativas Adversariais (GANs):

Se você quiser gerar visualizações sintéticas das acomodações com base nas características de entrada, pode considerar a utilização de GANs para criar imagens sintéticas e, em seguida, prever os preços com base nessas imagens.

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 prever os preços das acomodações no Airbnb, 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.

2. **Métricas de Avaliação**: Escolha métricas apropriadas para avaliar o desempenho do modelo. No contexto de previsão de preços de acomodações, métricas comuns incluem Erro Médio Absoluto (MAE), Erro Quadrático Médio (MSE), Erro Médio Percentual Absoluto (MAPE) e R^2. Cada métrica fornece uma perspectiva diferente sobre o desempenho do modelo.

3. **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.

4. **Cross-Validation (Validação Cruzada):** Utilize técnicas de validação cruzada, como k-fold cross-validation, para avaliar a robustez do seu modelo. Isso envolve dividir seus dados em k subconjuntos, treinar e testar o modelo k vezes e calcular métricas médias de desempenho.

5. *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.

6. 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.

7. **Ajuste de Hiperparâmetros**: Se o desempenho do modelo não atender às expectativas, considere ajustar os hiperparâmetros do modelo e repetir o processo de validação.

8. **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.

9. **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.

10. 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.

Entrega e apresentação: dia 18/10/2023, entre 19 e 22h,

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:

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.

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.

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

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 benchmarks, 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