ASA 2023

ArcelorMittal

Instruções para o Estudo de Caso

Sumário

Introdução

Esse é um estudo de caso fictício que representa o tipo de trabalho que encontramos na ArcelorMittal. Ele servirá para testar suas habilidades para lidar com dados, executar análises estatísticas e de Machine Learning, comunicar suas descobertas e obter *insights* a partir do seu trabalho.

Nós sabemos que o processo produtivo do aço é altamente complexo e ele foi significativamente simplificado para esse teste. Outrossim, lembre-se de demonstrar e explicar tudo que está sendo feito, trazendo sua opinião sobre as análises. **Não é o objetivo desse teste apenas demonstrar sua capacidade de escrever código, treinar modelos e plotar gráficos. Estamos interessados em ver sua capacidade analítica e de intepretação sobre os resultados obtidos.**

Cenário

A ArcelorMittal está procurando formas de melhorar sua identificação de defeitos nas placas de aço. As placas são produzidas no lingotamento contínuo, após um processo de várias etapas que começa nos altos-fornos. Por causa da complexidade do processo, diversos defeitos podem ocorrer na placa produzida. Nosso especialista gerou um conjunto de dados contendo dois defeitos que ele gostaria de distinguir com maior exatidão. Todos os dados foram obtidos a partir de sensores automatizados ou imagens de câmeras, que identificam dimensões e características da placa e do defeito.

Você não precisa entender profundamente o processo produtivo para realizar essa atividade, mas esse vídeo pode te ajudar a entender melhor sobre a parte do lingotamento contínuo (ver a partir do instante 2:57).

ASA 2023

ArcelorMittal

Sua Tarefa

Você deve encontrar *insights* a partir dos dados e auxiliar o time de qualidade da ArcelorMittal a **identificar se o defeito encontrado na placa é do tipo 0 ou do tipo 1**. Você também deve apresentar suas descobertas para o responsável técnico da área. Existe um arquivo (.csv) que será enviado junto com esse documento que contém toda a base de dados coletada. Como auxílio, têm-se no apêndice algumas imagens e uma descrição das colunas do conjunto de dados. **O modelo deve ser desenvolvido em Python, conforme ensinado durante o treinamento**.

Para ajudar na sua tarefa, temos uma sugestão de passos:

- Realizar a carga dos dados e analisar os tipos e características das colunas;
- Realizar a análise exploratória para comunicar as informações relevantes encontradas sobre os dados (utilize textos e gráficos);
- Aplicar métodos de detecção de outliers;
- **Caso seja necessário para o tipo de modelo que você escolheu utilizar**, realizar o pré-processamento, contendo tratamento de nulos, *encoding* e normalização, conforme visto no treinamento;
- Realizar o treino do modelo utilizando validação cruzada (*cross validation*), justificando a escolha do modelo utilizado;
- Calcular todas métricas de desempenho estudadas. **Recomendamos utilizar o F1 como métrica principal.**
- Responder as perguntas propostas na página seguinte;
- Anotar todas as suas descobertas e análises que podem contribuir para melhorar o desempenho do modelo e do processo de tomada de decisão.

Perguntas

Questão 1: O modelo construído será utilizado para reduzir o erro na hora da classificação do defeito. Sabe-se que:

- Cada acerto do modelo significa um custo de R$ 500,00 para a recuperação da placa;
- Identificar o defeito 0 incorretamente como defeito 1 gera um custo de R$ 500,00 para recuperação da placa mais um custo de R$ 3500,00 por conta do custo logístico de ter enviado a placa para o tratamento incorreto;
- Identificar o defeito 1 incorretamente como defeito 0 gera um custo de R$ 500,00 para a recuperação da placa. Além disso, há também um custo de R$ 6213,00 por conta do erro logístico de ter enviado a placa para o tratamento incorreto, sabendo que o tratamento do defeito 1 ocorre numa etapa do processo anterior ao do defeito 0.

Em posse dessas informações, qual seria o custo total do processo com o uso em produção do modelo desenvolvido? **Deixe bem claro todo o passo a passo utilizado para a obtenção do resultado.**

Questão 2: Sem uma ferramenta mais tecnológica ao seu dispor, atualmente a distinção dos dois defeitos é feita de forma manual por um especialista. Para esse conjunto de dados específico, ele obteve os seguintes resultados:

- 350 placas com defeito tipo 0 foram identificadas corretamente;
- 256 placas com defeito tipo 0 foram identificadas como defeito tipo 1;
- 161 placas com defeito tipo 1 foram identificadas corretamente;
- 200 placas com defeito tipo 1 foram identificadas como defeito tipo 0.

Conhecendo o custo de cada tipo de erro (conforme questão 1), qual seria a economia que a utilização do seu modelo traria para o processo? **Deixe bem claro todo o passo a passo utilizado para a obtenção do resultado.**

ASA 2023

ArcelorMittal

Critérios de Avaliação

1) Desempenho (Peso = 2)
   Resultado do modelo em termos de F1 score.

2) Análise exploratória (Peso = 2)
   Entendimento dos dados, explorar as variáveis, plotar gráficos, tirar insights, criar novas variáveis, etc.

3) Pré-processamento (Peso = 2)
   Executar de forma correta todas as etapas de pré-processamento necessárias para o modelo que você escolheu, exemplo: remoção de outliers, encoding, normalização, tratamento de nulos, etc.

4) Clareza do código (Peso = 2)
   Como cientista de dados, nosso código deve ser capaz de ser lido e entendido por outra pessoa. Use comentários e escreva da forma mais simples e clara possível.

5) Entendimento de negócio (Peso = 2)
   Levaremos em conta a sua resposta dada às questões 1 e 2, bem como suas justificativas e conclusões sobre a possibilidade de aplicação do seu modelo em produção.

6) Apresentação (Peso = 10)
   As 10 participantes com melhor resultado no Testdome serão chamadas para apresentar de forma online seu estudo de caso aos avaliadores. O intuito da apresentação não é fazer um teste oral, apenas queremos que vocês nos expliquem o que fizeram, para garantir que não tiveram uma ajuda injusta ao longo do desenvolvimento da tarefa.

Apêndice 1: Ilustrações

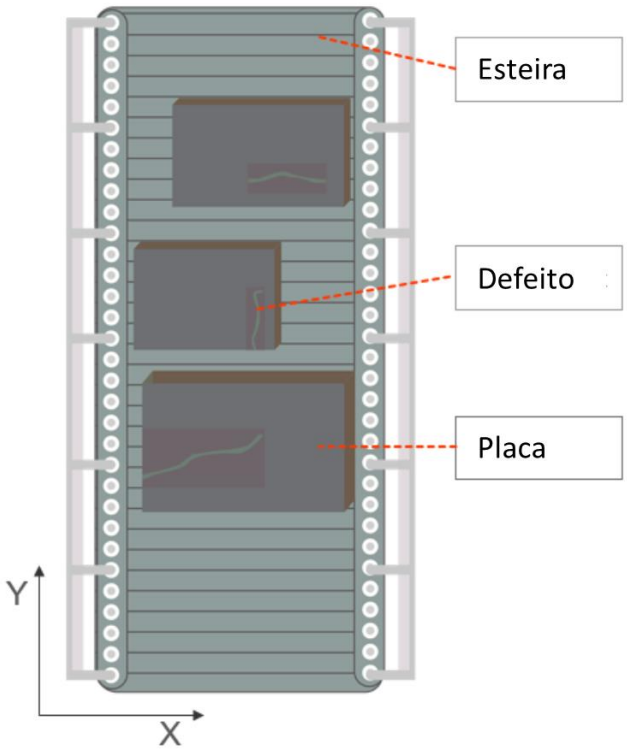
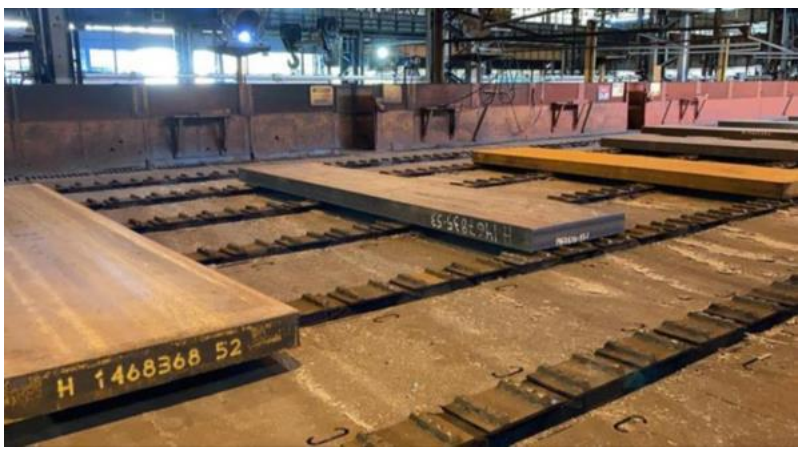


Ilustração das placas e defeitos

Placas de Aço na saída do Lingotamento Contínuo



Placas de Aço

Apêndice 2: Dicionário de dados

- min_x_defect – Coordenada x inicial do defeito
- max_x_ defect – Coordenada x final do defeito
- min_y_ defect – Coordenada y inicial do defeito
- max_y_ defect – Coordenada y final do defeito
- area_pixels – Total de pixels presentes na placa
- slab_width – Largura da placa (eixo X)
- slab_length – Comprimento da placa (eixo Y)
- sum_pixel_luminosity – Soma da luminosidade dos pixels
- min_pixel_luminosity – Mínima luminosidade dos pixels
- max_pixel_luminosity – Máxima luminosidade dos pixels
- conveyer_width – Largura da esteira (correia) transportadora (eixo X)
- type_of_steel – Identifica a classe do aço: pode pertencer à classe A300 ou A400
- defect_type – Tipo de defeito da classe. Pode ser do tipo 0 ou do tipo 1.