Faculdade IMPACTA

DS OTIF Detection

Faculdade IMPACTA

SOLID Gestão Empresarial

ZT ZENATUR

AGENDA

- Visão Geral das Empresas
- A DOR da ZT
- Contexto dos Dados
- EDA Descritivo
- EDA Inferencial & Feature Engineering
- Pré-Processamento & Eliminando Vazamento
- Modelos Clássicos e Métricas Fundamentais
- Threshold de Decisão
- Matriz de Confusão, ROC e Precision-Recall

VISÃO GERAL DAS EMPRESAS

- **SOLID Gestão Empresarial**: Empresa especializada em otimização de processos, aceleração de resultados e desenvolvimento de soluções corporativas. Atualmente presta consultoria e fornece sistemas de gestão para a Zenatur.

- **A Zenatur**: Empresa focada em soluções completas de logística promocional, atuando de ponta a ponta no fluxo operacional dos clientes.

- **Tiago Lima**: Consultor de dados da SOLID, atuando dentro da Zenatur como responsável por automações, BI e projetos de Data Science para melhoria contínua dos processos operacionais.

Faculdade IMPACTA

SOLID Gestão Empresarial

ZT ZENATUR

A DOR DA ZT

- **Processos Longos e Complexos**: A Zenatur opera uma cadeia logística estruturada em múltiplas fases, cada uma com impacto direto no cumprimento dos prazos acordados com os clientes.

- **Problemas de OTIF**: Atualmente, não existe visibilidade antecipada sobre quais pedidos estão em risco de atrasar, dificultando ações preventivas, o resultado é a dificuldade de cumprir a OTIF gerando prejuízos financeiros.

- **Solução para OTIF**: Prever com antecedência quais pedidos têm alta probabilidade de violar o OTIF, permitindo atuação proativa e redução de prejuízos operacionais.

CONTEXTO DOS DADOS

- **O DATASET**: O dataset contém o histórico de pedidos da Zenatur, com informações de datas planejadas e reais, duração das fases logísticas, atrasos, ocorrências e outros atributos operacionais. A variável-alvo é binária, um flag_atraso onde 1 = Verdadeiro e 0 = Falso

- **AMOSTRA**: O conjunto amostral são suficientemente consistentes para modelagem, porém apresentam lacunas temporais e comportamentos irregulares em etapas específicas, sobretudo nas fases Q e W.

EDA DESCRITIVO

- **Forte Assimetria a Direita**: A EDA revelou assimetria significativa nos tempos das fases, presença de outliers e concentração elevada de casos dentro dos intervalos inferiores.

- **Alta Dispersão**: Os histogramas mostram dispersões diferentes entre as etapas.

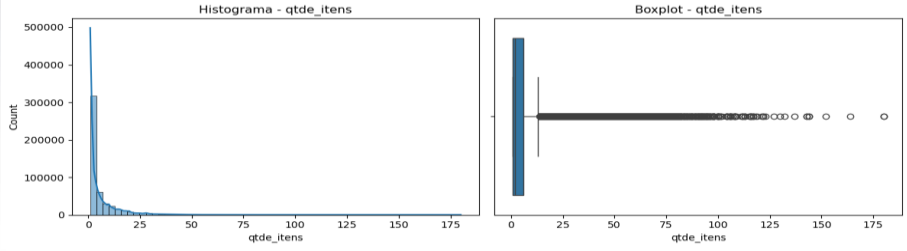
- **Altos Valores Extremos**: Os boxplots evidenciam a existência de valores extremos que impactam a distribuição geral e o comportamento do modelo.

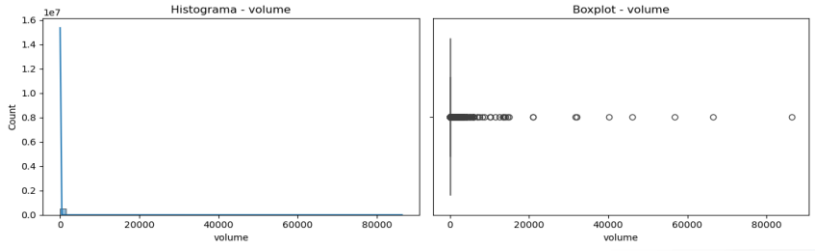
Faculdade IMPACTA

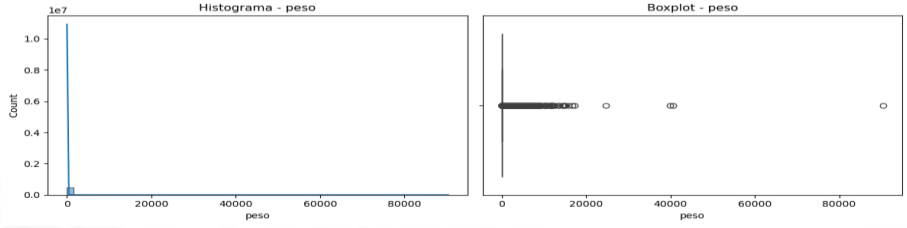
SOLID Gestão Empresarial

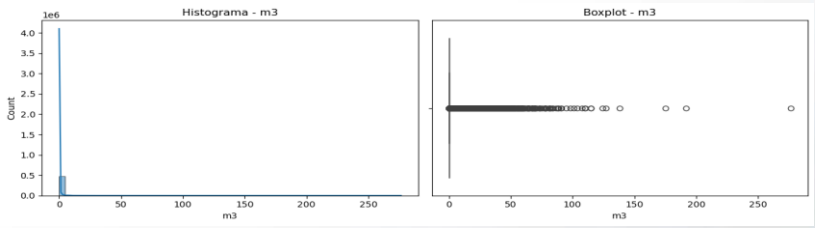
ZT ZENATUR

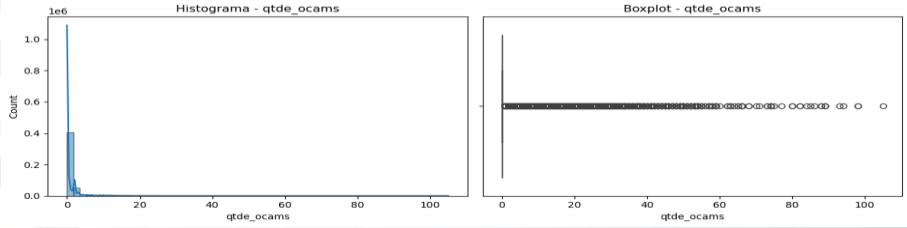
EDA DESCRITIVO - Gráficos

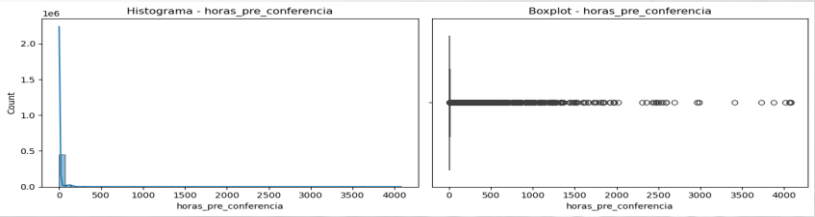












EDA INFERENCIAL & FEATURE ENGINEERING

- A análise mostrou que atrasos iniciais e prolongamentos nas fases são causados por grandes volumes de itens em um pedido, o que acarrega em complexas divisões de OCAMs, estes eventos são os ofensores que aumentam exponencialmente a probabilidade de violação do OTIF.

- Para capturar esses padrões, foram criadas as features derivadas como:
  - **Lead Time Total em Horas:** soma de todas as horas operacionais
  - **Complexidade Operacional:** Definido pela quantidade de itens e OCAMs
  - **Flag de Pedido Grande:** Pedidos acima do terceiro quartil ou seja 75%
  - **Flag de Processo Longo:** Lead Time das Horas muito superiores as medias

Faculdade IMPACTA

SOLID Gestão Empresarial

ZT ZENATUR

EDA INFERENCIAL & FEATURE ENGINEERING

2. Correlação de Pearson (linear)

As maiores correlações lineares com atraso foram:

Variável	Correlação	Interpretação
horas_divisao_ocam	0.108	pedidos com muitas divisões tendem a atrasar
horas_planejamento	0.101	planejamento acima do normal impacta o downstream
horas_conferencia	0.085	conferências mais longas elevam risco de atraso
horas_coleta	0.082	dificuldade na coleta afeta fluxo
qtde_itens	0.059	pedidos maiores tendem a ser mais lentos

ANOVA

feature	f_statistic	p_value	significativo_5pct
qtde_itens	1714.602583	0.000000e+00	True
fl_base	1819.566140	0.000000e+00	True
horas_divisao_ocam	5783.234460	0.000000e+00	True
horas_planejamento	5037.615296	0.000000e+00	True
horas_conferencia	3564.702534	0.000000e+00	True
horas_coleta	3330.829576	0.000000e+00	True
horas_exped_minuta	1222.285906	1.879743e-267	True
horas_analise_producao	1162.383648	1.823463e-254	True
m3	1113.866668	6.038817e-244	True
peso	582.790162	1.103748e-128	True
qtde_ocams	535.505503	2.075628e-118	True

3. Correlação de Spearman (monotônica)

Aqui surgiram padrões muito mais expressivos:

Variável	Spearman	Interpretação
m3	0.202	cargas volumosas elevam o risco de atraso
peso_cubado_rodoviario	0.196	cubagem alta = gargalo operacional
horas_minuta	0.194	minutas longas → impacto no prazo
peso	0.167	pedidos muito pesados = operação mais lenta
volume	0.158	confirma padrão relacionado ao tamanho do pedido

Variável	χ^2	p-value	Conclusão
sigla_cliente	38457	< 1e-300	Fortemente dependente
tipo_veiculo	49719	< 1e-300	Fortemente dependente
uf	6800	< 1e-300	Dependente
modalidade	40397	< 1e-300	Fortemente dependente
flag_entrega_agendada	1231	6e-270	Dependente

Faculdade IMPACTA

SOLID Gestão Empresarial

ZT ZENATUR

EDA INFERENCIAL & FEATURE ENGINEERING

- Com base nos resultados inferenciais criou-se as variáveis: lead_time_total_horas, complexidade_operacional, pedido_grande_flag e processo_longo_flag

	lead_time_total_horas	complexidade_operacional	pedido_grande_flag	processo_longo_flag
0	229.0	1	0	0
1	234.0	1	0	0
2	249.0	1	0	0
3	266.0	7	1	0
4	542.0	4	1	1

PRÉ-PROCESSAMENTO – ELIMINANDO DATA LEAKAGE

- Imputação de valores ausentes: SimpleImputer(strategy="median") para aplicar a mediana nas variáveis numéricas, pois é robusto contra outliers. Para categóricas foi aplicado SimpleImputer(strategy="most_frequent")(moda), pois mantem a coerência de padrões minimizando viés.

- Codificação Categórica (OneHot Encoding): É o método mais seguro e universal para modelos clássicos, como Regressão Logística, Árvores, Random Forest e Gradient Boosting.

- Scaling (Padronização das Variáveis Numéricas): A padronização é uma etapa fundamental na preparação dos dados, especialmente para modelos sensíveis à escala das variáveis. Neste projeto, utilizamos o StandardScaler, que transforma todas as variáveis numéricas para uma escala comum, com média 0 e desvio padrão 1.

PRÉ-PROCESSAMENTO – ELIMINANDO DATA LEAKAGE

- **ColumnTransformer — Unindo o Processo:** Consolidação de todos os elementos dentro de um único bloco de pré-processamento usando **ColumnTransformer**, garantindo que cada transformação seja aplicada apenas ao tipo correto de variável, preservando a integridade do pipeline.

- **Split Train/Test**: Esta etapa divide o dataset em **70% treino e 30% teste**. Usou-se o **parâmetro stratify=y** com o objetivo de manter as proporções originais de atrasos/não atrasos garantindo que o conjunto de treino e testes representem adequadamente o problema. A separação é importante como um dos fatores para evitar o data leakage.

- **Pipeline Final de Pré-processamento:** Nos passos anteriores criou-se ferramentas, nada foi processado, todo o processamento será feito nesta etapa usando o **Pipeline(setps=[()])**

MODELOS CLÁSSICOS e METRICAS FUNDAMENTAIS

- **Modelos avaliados:** Foram avaliados modelos clássicos de classificação como:
  - **DummyClassifier:** por ser um modelo mais simplista, o DummyClassifier foi usado para definir a baseline
  - **LogisticRegression:** capaz de prever probabilidade com base em variáveis independentes
  - **DecisionTreeClassifier:** modelo supervisionado que usa uma estrutura de árvore de previsões usando um fluxograma de regras de decisões.
  - **RandomForestClassifier:** modelo robusto, pois combina múltiplas arvores de decisão para melhorar a precisão e a estabilidade das previsões
  - **GradientBoostingClassifier:** o queridinho dos Cientistas de Dados por ser muito forte e preciso combinando vários modelos mais fracos de forma sequencial

MODELOS CLÁSSICOS e METRICAS FUNDAMENTAIS

Métricas fundamentais utilizadas para avaliar os modelos

- **Accuracy:** proporção geral de acertos
- **Precision:** qualidade dos positivos previstos(evita falsos positivos)
- **Recall:** Capacidade de identificar pedidos que realmente irão atrasar
- **F1-Score:** equilíbrio entre Precision e Recall
- **ROC AUC:** separabilidade global entre as classes

Faculdade IMPACTA

SOLID Gestão Empresarial

ZT ZENATUR

MODELOS CLÁSSICOS e METRICAS FUNDAMENTAIS

Os modelos que melhor equilibraram **Recall** e **F1-Score** foram priorizados.

	modelo	accuracy	precision	recall	f1	roc_auc
3	Random Forest	0.8839	0.7276	0.9093	0.8084	0.9576
2	Decision Tree	0.8932	0.7799	0.8407	0.8092	0.8956
4	Gradient Boosting	0.8191	0.7681	0.4699	0.5831	0.8838
1	Logistic Regression	0.7083	0.4729	0.7278	0.5733	0.7879
0	Baseline (Dummy)	0.7308	0.0000	0.0000	0.0000	0.5000

AJUSTE DE DESBALANCEAMENTO

Os ajustes foram realizados no modelo vencedor **RandomForestClassifier**

- **Class Weights:** Atribui peso maior aos exemplos da classe minoritária e peso menor para as classes minoritárias. É geralmente um bom ponto de partida, pois é simples e rápido de aplicar.
- **SMOTE(Synthetic Minority Over-sampling Technique):** Cria novas amostras sintéticas da classe minoritária. É eficaz quando a classe minoritária tem padrões complexos que se beneficiam da geração de dados sintéticos.

Faculdade IMPACTA

SOLID Gestão Empresarial

ZT ZENATUR

AJUSTE DE DESBALANCEAMENTO

Random Forest + SMOTE foi levemente superior, mas o tempo de processamento é alto, então é mais viável optar pelo **Random Forest + Class Weights**, pois o tempo de processamento e menor.

Métrica	Random Forest + SMOTE	Random Forest + Class Weights
Accuracy	0.9211	0.89
Precision	0.8463	0.7295
Recall	0.864	0.94
F1 Score	0.855	0.8215
ROC AUC	0.9704	0.9677

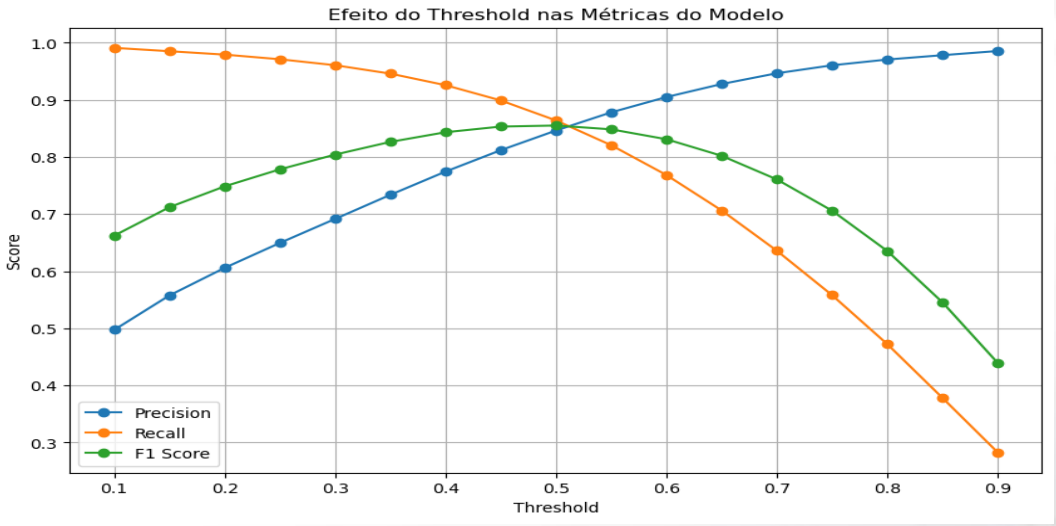
Faculdade IMPACTA

SOLID Gestão Empresarial

ZT ZENATUR

THRESHOLDS DE DECISÃO

	threshold	precision	recall	f1
0	0.10	0.497183	0.991026	0.662167
1	0.15	0.557481	0.985145	0.712032
2	0.20	0.605768	0.979213	0.748496
3	0.25	0.649414	0.971076	0.778321
4	0.30	0.691439	0.960632	0.804104
5	0.35	0.733661	0.945903	0.826372
6	0.40	0.774371	0.925700	0.843300
7	0.45	0.811859	0.898956	0.853190
8	0.50	0.846274	0.863973	0.855032
9	0.55	0.878113	0.820549	0.848355
10	0.60	0.904884	0.768353	0.831048
11	0.65	0.927744	0.706297	0.802015
12	0.70	0.946522	0.635774	0.760634
13	0.75	0.960506	0.557950	0.705867
14	0.80	0.970611	0.472191	0.635311
15	0.85	0.978220	0.377991	0.545282
16	0.90	0.985501	0.282575	0.439213



threshold	precision	recall	f1
0.30	0.691	0.960	0.804
0.40	0.774	0.925	0.843
0.50	**0.846**	**0.863**	**0.855**
0.60	0.904	0.768	0.831
0.70	0.946	0.635	0.760

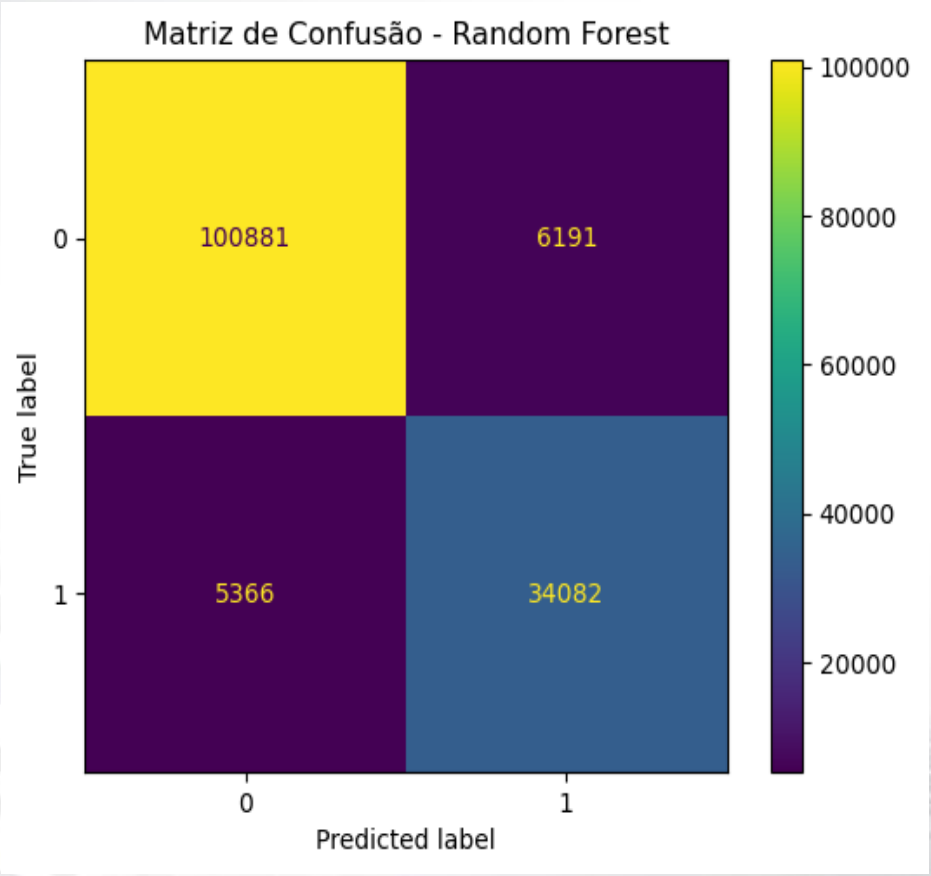
USAR: y_pred = (proba >= 0.55).astype(int)

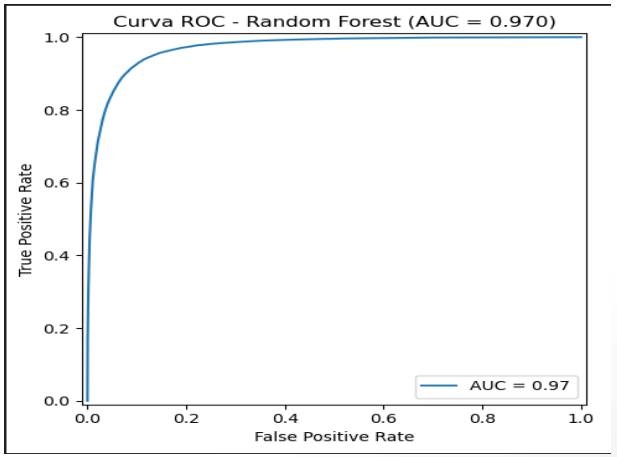
Faculdade IMPACTA

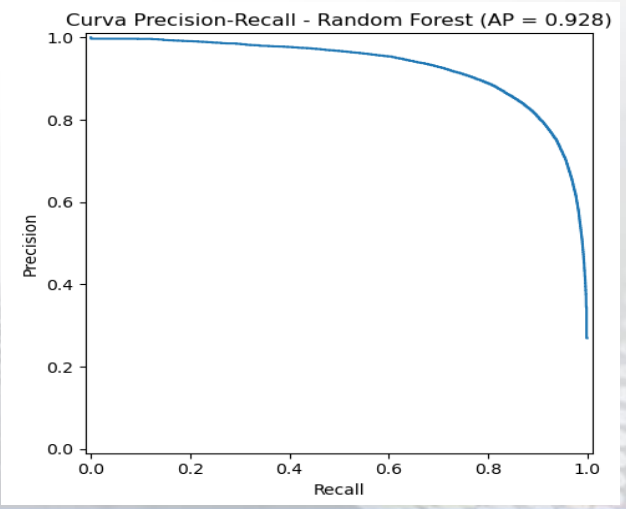
SOLID Gestão Empresarial

ZT ZENATUR

MATRIZ DE CONFUSÃO, ROC e PRECISION-RECALL







CONCLUSÃO

- **O projeto demonstrou que é possível gerar previsibilidade operacional** para a Zenatur utilizando técnicas de Machine Learning aplicadas diretamente ao comportamento histórico da cadeia logística. A análise exploratória revelou assimetrias, fases críticas e padrões específicos que fundamentaram a criação de features altamente relevantes, como *lead_time_total*, complexidade operacional e flags de pedidos grandes e processos longos.

- **O pipeline de pré-processamento foi cuidadosamente estruturado para evitar *data leakage***, garantindo imputação, codificação e padronização feitas exclusivamente sobre o conjunto de treino. Os modelos clássicos foram avaliados de forma comparativa e o **Random Forest com Class Weights** apresentou o melhor equilíbrio entre Recall e F1-Score, fornecendo robustez, estabilidade e excelente capacidade discriminativa (AUC ≈ 0.97).

- **A análise dos *thresholds* permitiu ajustar a decisão final com base nas necessidades reais do negócio**, definindo **0.55** como ponto ótimo para maximizar o F1-Score e reduzir falsos negativos — que representam atrasos reais não identificados. As curvas ROC e Precision-Recall confirmam a forte capacidade do modelo em separar pedidos com risco real de atraso, mesmo em um cenário de classes desbalanceadas.

- **Em síntese, o modelo final entrega previsibilidade, robustez e capacidade prática de suporte à tomada de decisão**, podendo ser integrado diretamente ao fluxo operacional da Zenatur para alertas antecipados de risco OTIF.

PRÓXIMOS PASSOS

- **Criar o artefato em REST API:** Publicar o modelo como endpoint usando FastAPI, incluindo rota de predição, validação de entrada e aplicação do threshold ótimo (0.55).
- **Desenvolver o Frontend com Streamlit:** Interface simples para upload de dados, visualização das probabilidades, ordenação por risco, dashboards e matrizes de decisão.
- **Criar os Containers Necessários (Docker):** Containerizar o backend (API), o frontend (Streamlit) e o modelo pré-processado, garantindo reprodutibilidade e portabilidade.
- **Orquestrar os Containers:** Configurar o Docker Compose com rede interna, variáveis de ambiente e inicialização coordenada dos serviços.
- **Deploy em Ambiente de Demonstração (Heroku):** Publicar a solução integrada em um ambiente acessível via web, permitindo demonstração executiva para diretores, gestores e tomadores de decisão.

Faculdade IMPACTA

Obrigado

- Tiago Pereira lima – 1020325
- LinkedIn: https://www.linkedin.com/in/tiago-lima-935049154/
- Github Projeto: https://github.com/tiago466/ds-otif-detection-ml