Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Alan Geovani Mazeto

Aplicação de técnicas de aprendizado de máquina na previsão de tempo de estadia de contêineres

Curitiba
2023

Alan Geovani Mazeto

Aplicação de técnicas de aprendizado de máquina na previsão de tempo de estadia de contêineres

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. DR ANDERSON ARA

Curitiba
2023

Aplicação de técnicas de aprendizado de máquina na previsão de tempo de estadia de contêineres

Applying machine learning techniques on container dwell time forecasting

Alan G. Mazeto

A previsão do tempo de estadia de contêineres desempenha um papel crucial na gestão eficiente das operações portuárias e na otimização da cadeia logística. Este artigo aborda diferentes modelos de regressão com a intenção de identificar a possibilidade da utilização de técnicas de aprendizagem de máquina para prever o tempo de permanência dos contêineres nos portos. Foi realizada a coleta dos dados históricos de contêineres importados com estadia no Terminal de Contêineres de Paranaguá (TCP), foi realizada a análise exploratória e aplicadas técnicas para a seleção dos principais atributos. A base final para criação e teste dos modelos possui 926605 registros e 11 covariáveis que se mostraram relevantes na previsão de tempo de estadia. Com a base final foram aplicadas as seguintes técnicas: regressão linear múltipla, árvore de decisão, floresta aleatória, redes neurais, regressão linear generalizada e gradient boosting. Para comparar e avaliar as diferentes técnicas, analisamos os resultados utilizando como métrica a raiz do erro quadrático médio (RMSE), para determinar a eficácia de cada método na previsão do tempo de estadia de contêineres. Os resultados obtidos demonstram que o uso de técnicas de aprendizagem de máquina pode trazer melhorias para o processo logístico, dando fundamento para o planejador portuário posicionar os contêineres com base nas previsões de tempo de estadia calculadas. Foi identificado que as técnicas de arvore de decisão se saíram melhor, dada a base fornecida e a métrica utilizada e o modelo de gradient boosting foi o que teve melhor desempenho nos dados de treino e teste. Este artigo destaca o potencial das técnicas de aprendizado de máquina na previsão do tempo de estadia de contêineres nos terminais portuários. Ao aplicar tais técnicas, os portos podem obter melhorias na eficiência operacional, reduzir custos e otimizar a gestão da cadeia logística.
Palavras-chave: Terminal de Contêineres, Tempo de Permanência, Aprendizado de Máquina, Regressão

The prediction of container dwell time plays a crucial role in efficient port operations and logistics chain optimization. This article addresses different regression models with the intention of identifying the possibility of using machine learning techniques to predict container dwell time at ports.

Historical data of imported containers with dwell time at the Terminal de Contêineres de Paranaguá (TCP) was collected, and an exploratory analysis was conducted, along with techniques applied to select the key attributes. The final dataset for model creation and testing consists of 926,605 records and 11 covariates that proved to be relevant in predicting dwell time.

The following techniques were applied to the final dataset: multiple linear regression, decision tree, random forest, neural networks, generalized linear regression, and gradient boosting.

To compare and evaluate the different techniques, the results were analyzed using the root mean square error (RMSE) as the metric to determine the effectiveness of each method in predicting container dwell time.

The results obtained demonstrate that the use of machine learning techniques can bring improvements to the logistics process, providing a basis for the vessel planner to position the containers based on the calculated dwell time predictions. It was identified that the decision tree based techniques performed better given the provided dataset and metric, and the gradient boosting model had the best performance in the training and testing data.

This article highlights the potential of machine learning techniques in predicting container dwell time at port terminals. By applying these techniques, ports can achieve improvements in operational efficiency, cost reduction, and optimization of logistics chain management.
Keywords: Container Terminal, Dwell Time, Machine Learning, Regression

1. Introdução

Existe um problema comum entre os portos que operam contêineres, o espaço disponível para armazena-

gem é limitado. De forma geral, essas empresas necessitam operar grandes quantidades de unidades diariamente e a acertividade na forma como os contêineres são armazenados influenciam diretamente na produtividade dos trabalhadores e maquinários. A previsão precisa de tempo na estadia de contêineres nos portos desempenha papél crucial na gestão eficiente da cadeia logística e no planejamento operacional, haja visto que as unidades são armazenadas em pilhas e para que seja realizada a retirada de contêineres que se encontram em baixo da pilha, os contêineres de cima devem ser remanejados. Essas movimentações são consideradas improdutivas pois não há cobrança realizada sobre contâineres que se encontram no caminho. Identificar as posições ótimas para empilhamento, e as movimentações ideias para que sejam feitos o menor numero possíveis com o maior resultado logístico é um desafio. Um ponto inicial para melhorar a eficiencia operacional e otimizar a logística pode ser identificar o numero de dias que uma unidade deve permanecer no pátio, dessa forma pode-se distribuir manualmente os contêineres com uma base factual do tempo provável, Existem incertezas quanto ao tempo de estadia de contêineres dado que não existe obrigatoriedade de tempo de retirada, e cada produto importado passa por processos específicos de liberação de carga no Brasil. Dessa forma, não se sabe quando ela será retirada, e portando a forma como os contêineres são alocados no pátio é feita de forma empírica. A aprendizagem de máquina, pode oferecer métodos e algorítmos que permitam a realização de tal previsão. A capacidade de aprendizagem não supervisionada torna essa abordagem particularmente adequada para a previsão de tempo de estadia dos contêineres, uma vez que os dados disponíveis são em sua maioria qualitativos e apresentam padrões complexos. Este trabalho traz como proposta uma análise de diferentes técnicas de aprendizagem de máquina, como arvores de decisão, floresta aleatória e redes neurais com a intenção de avaliar as melhores estratégias e discutir a aplicabilidade da inteligência artificial para a predição de tempo de estadia de contêineres em portos, utilizando como caso de uso, o Terminal de Contêineres de Paranaguá (TCP). No segmento a seguir serão apresentados os materiais, os métodos e as abordagens utilizadas para a predição de tempo de estadia de contêineres no pátio portuário, seguido de estudos de casos e análise dos resultados obtidos.

2. Coleta e tratamento de dados

O processo de coleta de dados foi realizado primeiramente identificando o volume dos dados exitentes e quais variáveis seriam passíveis de obter relevância para algorítmos de regressão. Identificamos que os registros com formato válido se iniciam em 2015 e se extendem até o último dia da coleta dos dados. Estes registros possuem algumas características básicas, são essas: São registros de importação, possuem conteúdo e tiveram o processo completo finalizado, ou seja, possuem data de entrada e saída do terminal. Das covariáveis selecionadas, foram escolhidos 100 atributos que poderiam dar base para a predição de tempo de estadia, dessas variáveis, grande parte correspondia a processos internos dos contêineres dentro do pátio, porém essas variáveis de mostraram ineficases por falta de amostragem relevante, muitas delas apareciam com pouca frequencia e de forma aleatória, não justificando assim a coleta de amostragem com balanceamento. Após seleção de covariáveis por exclusão, identificamos 11 atributos que justificam em parte o tempo de estadia dos contêineres e que foram utilizados em outros trabalhos, como o de Gaete, Myriam, et al. (2017 p. 128-139). Essas variáveis são: Mês de chegada do contêiner, serviço prestado, proposito do container, cliente, porto de origem, produto, tamanho do contêiner, remanejamento do container no pátio, container danificado, direção da porta não encontrada e data de abertura de ordem de serviço. Essas variáveis foram escolhidas por, em partes serem utilizadas em trabalhos semelhantes e outras por possuir volume significativo em comparação com a quantidade de dados. A base final, após tratamentos, possui 926.605 registros e 11 covariáveis. A variável de resposta por sua vês é o número de dias que o contêiner esteve dentro do pátio, essa variável possui valores de 0 a 300 dias e como pode ser visto nas figuras 1 e 2, possui pouco mais de 10% dos registros como outliers. esses regsitros não foram retirados pois entende-se que fazem parte da distribuição dos containeres no pátio e não são anomalias.

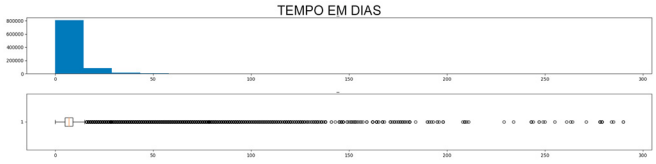

Figura 1: Distribuição e boxplot da variável dependente

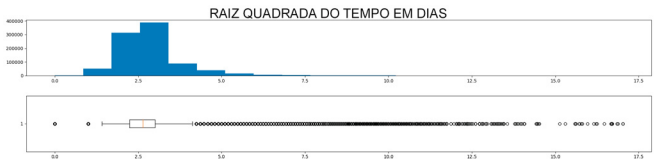

Figura 2: Distribuição e boxplot da raiz quadrada da variável dependente

3. Criação dos modelos

Com os dados coletados e após análise exploratória, iniciou-se o processo de criação dos modelos de aprendizagem de máquina para posteriormente ser realizada a comparação entre as técnicas e análisar os resultados dadas as características dos os dados fornecidos.

3.1. Regressão Lineal Multipla

Pelo tipo dos dados encontrados na base, entendemos que os modelos de regressão linear não se sairiam bem em comparados com outros modelos, principalmente os que possuem embasamento em árvores de decisão. A escolha por realizar a regressão por este método foi para ter como base inicial a eficácia em um modelo mais simples e ter base para nivelamento e comparação. A aplicação da tecnca foi realizada utilizando a biblioteca sklearn do python e as bibliotecas numpy e pandas para manejamento da base de dados. Os resultados encontrados foram coletados e comparados com os valores reais. a distribuição abaixo apresenta a relação entre pevisto e observado.

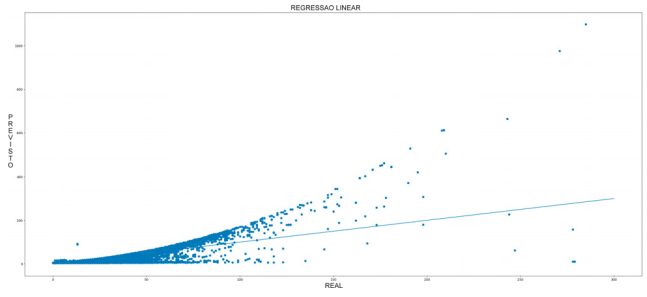

Figura 3: Dispersão da resposta real em relação à resposta predita pela regressão linear múltipla

A análise aponta que a previsão é prejudicada quando o tempo de estadia no pátio se prolonga. junto a isso, observamos a métrica de raiz do erro quadratico medio (RMSE) e score do modelo, dado pela ferramenta utilizada, esses valores são: 6,14 e 0,66, o que indica que o modelo erra em média 6 dias e as variáveis explicam 66% do resultado.

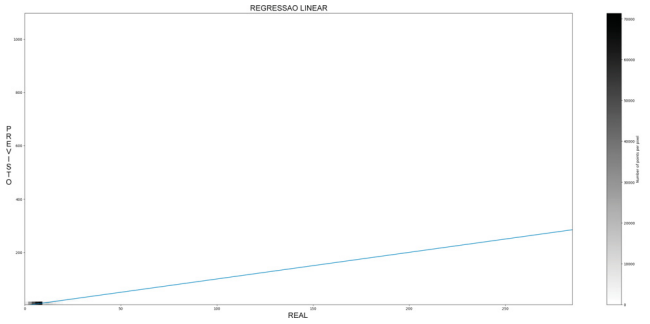

Figura 4: Densidade da resposta real em relação à resposta predita pela regressão linear múltipla

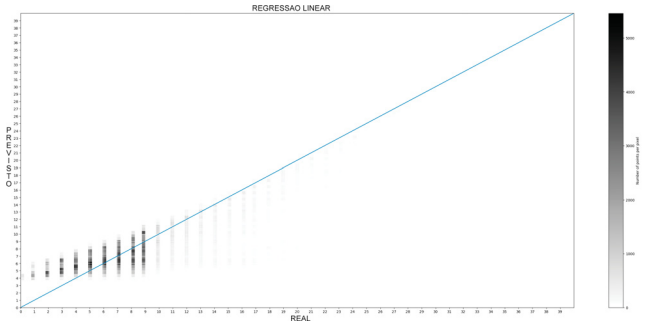

Figura 5: Densidade da resposta real em relação à resposta predita pela regressão linear múltipla ampliada nos primeiros 45 dias

3.2. GLM

O modelo de regressão linear generalizado foi desenvolvido para tentar melhorar o resultado obtido na regressão linear multipla. Ao analizarmos a distribuição da variável dependente, obervamos que não se trata de uma distribuição normal. A distribuição em questão possui uma calda alongada chegando até os 300 dias mas com pouco volume de informação, a grande maioria dos dados encontra-se entre os primeiros 45 dias. O modelo de regressão generalizada foi desenvolvido com base na distribuição gama. O desenvolvimento do modelo foi realizado utilizando o modulo GammaRegressor da biblioteca sklearn e numpy e pandas para o manejamento dos dados.

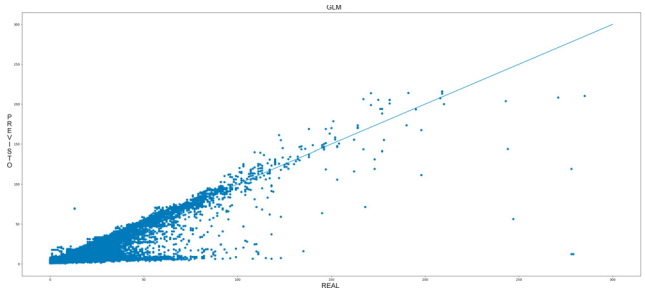

Figura 6: Dispersão da resposta real em relação à resposta predita pela regressão linear generalizada

A regressão foi realizada e foram coletados os resultados do modelo, assim como na regressão linear múltipla, comparamos os resultados do score fornecido pela ferramenta com o valor do RMSE, no caso da regressão linear generalizada, obtivemos o score de 0,70 e o RMSE de 4,2.

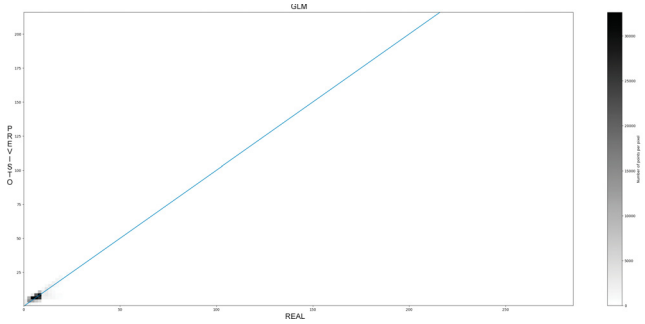

Figura 7: Densidade da resposta real em relação à resposta predita pela regressão linear generalizada

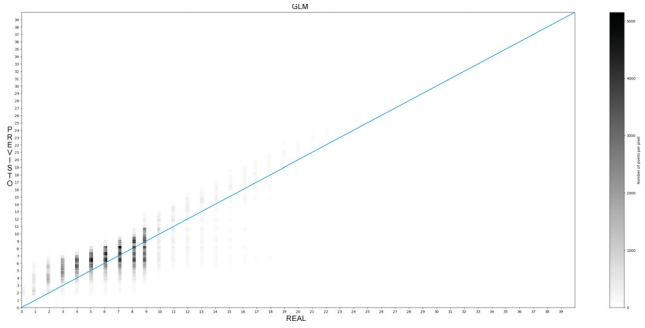

Figura 8: Densidade da resposta real em relação à resposta predita pela regressão linear generalizada ampliada nos primeiros 45 dias

3.3. Arvore de decisão

O desenvolvimento do modelo de árvore de decisão foi realizado utilizando o módulo DecisionTree da biblioteca Sklearn, assim como os outros modelos treinados, coletamos as informações de score e RMSE para comparação. O modelo de árvore de decisão se saiu melhor que o modelo de regressão linear múltipla, já que o tipo das variáveis utilizadas eram em sua maioria categóricas. A densidade dos dados observados se concentram em sua maioria dentro dos primeiros 45 dias de estadia, mas o modelo teve bons resultados para os casos que se afastam desse agrupamento.

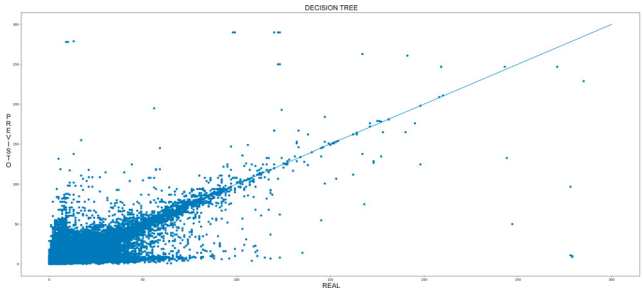

Figura 9: Dispersão da resposta real em relação à resposta predita pela árvore de decisão

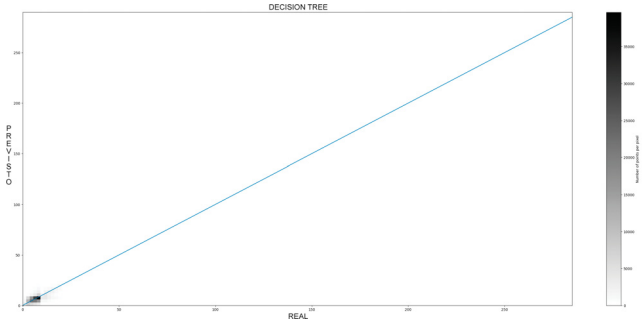

Figura 10: Densidade da resposta real em relação à resposta predita pela árvore de decisão

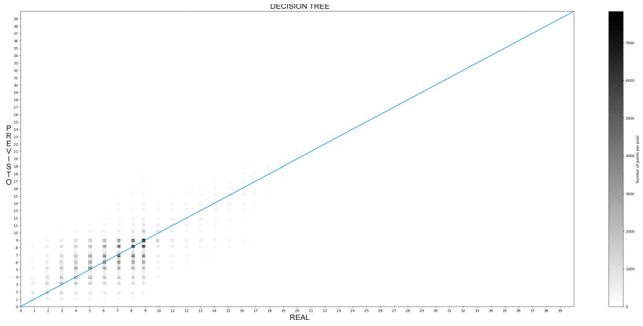

Figura 11: Densidade da resposta real em relação à resposta predita pela árvore de decisão ampliada nos primeiros 45 dias

3.4. Floresta Aleatória

Para o modelo de regressão por floresta aleatória foi necessário realizar o desenvolvimento com uso de técnicas de processamento em unidade de processamento gráfico, dada a quantidade de registros e a quantidade

de modelos de árvores aleatórias estavam sendo criados. Para o desenvolvimento do algorítmos foram utilizados os módulos cuml e cudf da biblioteca Rapids. O resultado obtido foi um dos mais promissores mas deve-se levar em consideração o custo de processamento deste modelo, pois foi também um dos mais demorados e com necessidade de maior infraestrutura computacional.

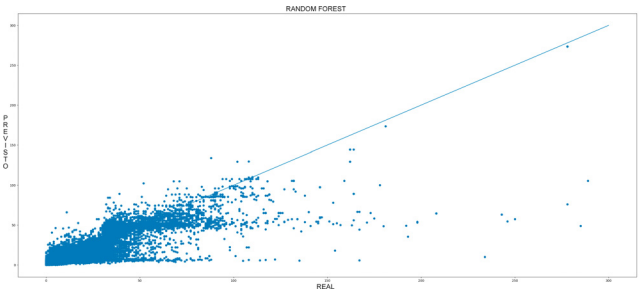

Figura 12: Dispersão da resposta real em relação à resposta predita pela floresta aleatória

- **Tunagem** A tunagem do modelo foi realizada alterando os números de árvores criadas em cada processo, número mínimo de nós e registros para alteração de nó e diferentes tipos de métrica para definição da acurácea do melhor modelo gerado. Ao final foram comparados os valores gerados pelo score e pelo RMSE para entender os resultados obtidos em comparação com os outros modelos criados. Para este modelo em questão, os melhores resultados obtidos após tunagem foram: score de 0,77 e RMSE de 4,1.

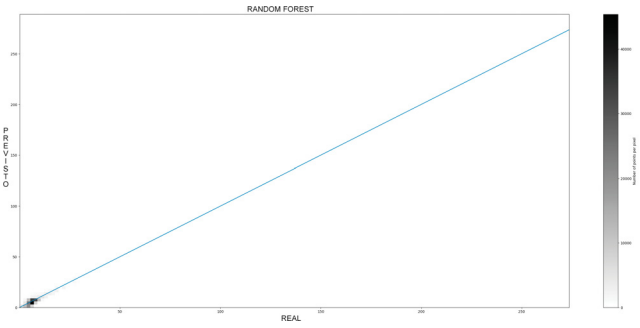

Figura 13: Densidade da resposta real em relação à resposta predita pela floresta aleatória

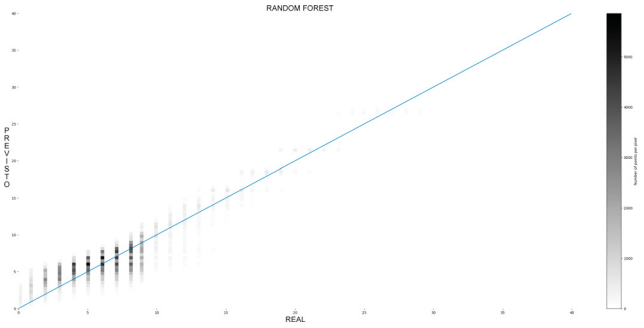

Figura 14: Densidade da resposta real em relação à resposta predita pela floresta aleatória ampliada nos primeiros 45 dias

3.5. Redes Neurais

Foi desenvolvido também o modelo de regressão por redes neurais, para comparação com os outros modelos desenvolvidos e também pois foi um modelo proposto por Ioanna Kourounioti at all. Os resultados obtidos foram uma boa troca entre desempenho e performance, haja visto que foi um modelo relativamente leve mas apresentou score e RMSE melhores que os modelos mais simples. O desenvolvimento do modelo foi realizado utilizando os módulos da biblioteca tensorflow.

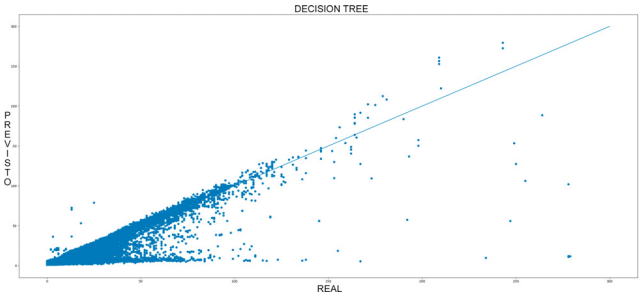

Figura 15: Dispersão da resposta real em relação à resposta predita pela rede neural

- **Tunagem** A melhora no algorítmo desenvolvido se deu a partir de testes com diferentes quantidade de camadas e algoritmos de ativação, além de testes com diferentes funções de perda e otimizadores. O melhor resultado obtido trouxe um score de 0,88 e um RMSE de 4,31, o que indica que o algorítmo erra em média o dia de saida em 4 dias.

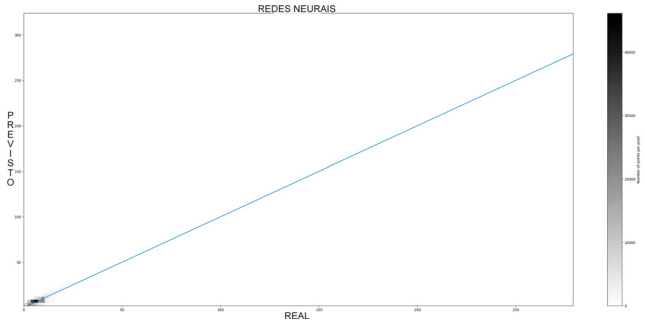

Figura 16: Densidade da resposta real em relação à resposta predita pela rede neural

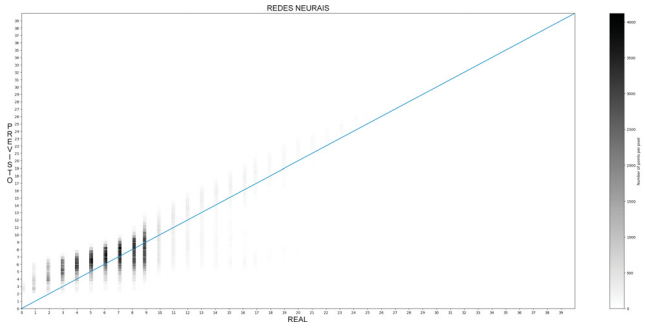

Figura 17: Densidade da resposta real em relação à resposta predita pela rede neural ampliada nos primeiros 45 dias

3.6. Extreme Gradient Boosting

Por último o último modelos desenvolvido foi o de extreme gradient boosting. Esse modelo foi desenvolvido utilizando os mesmos recursos do modelo de floresta aleatório, portanto foi processado em GPU e utilizou os modulos cudf e cuml da biblioteca rapids. Os resultados obtidos foram os mais promissores trazendo a melhor estimativa de tempo de estadia e não sendo o algorítmo mais custoso entre todos os observados.

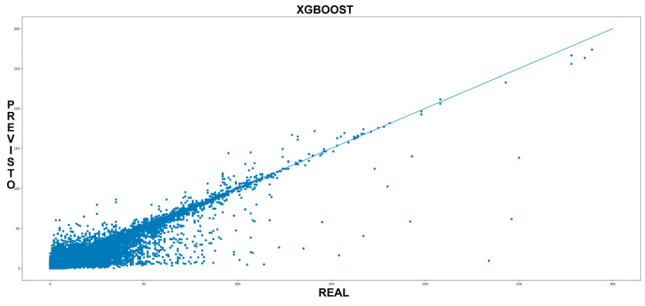

Figura 18: Dispersão da resposta real em relação à resposta predita pelo extreme gradient boosting

- **Tunagem** A tunagem do algorítmo de gradient boosting foi realizada testando diferentes valores para os parametros de métrica de avaliação, objetivo, profundidade máxima e taxa de aprendizagem. Após testes com diferentes parametros, os melhores resultados obtidos foram: score igual a 0,89 e RMS igual a 3,02

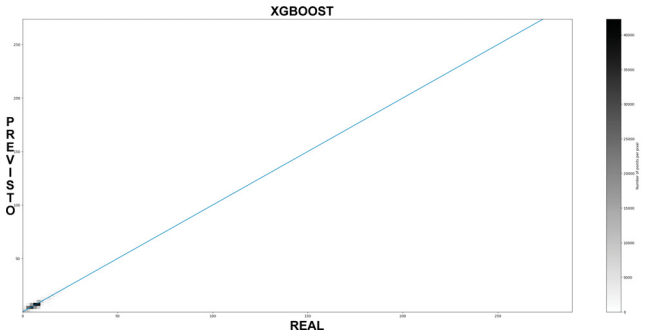

Figura 19: Densidade da resposta real em relação à resposta predita pelo extreme gradient boosting

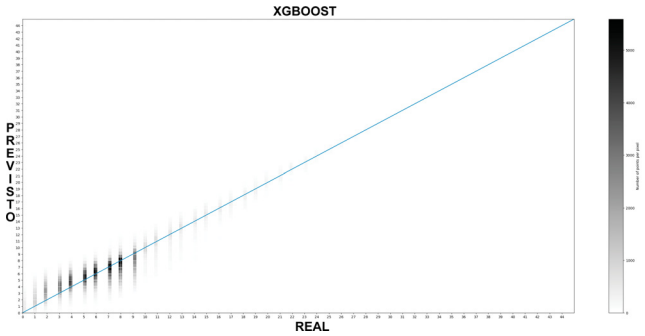

Figura 20: Densidade da resposta real em relação à resposta predita pelo extreme gradient boosting ampliada nos primeiros 45 dias

4. Resultados e Discussões

Após a aplicação das técnicas de aprendizado de máquina, foi evidente uma performance notavelmente superior nos modelos baseados na lógica de árvores de decisão em comparação com as outras abordagens. Tanto o score do modelo criado quanto o valor do RMSE reforçam essa constatação. Dentre essas técnicas, o Extreme Gradient Boosting (XGBoost) se destacou significativamente, apresentando um score de 0,89 e um RMSE de 3,02. Esses resultados indicam que, em média, o modelo erra a data de saída do contêiner em aproximadamente três dias. Essa performance do modelo XGBoost é atribuída à sua capacidade de capturar relações complexas e não lineares presentes nos dados, além de sua habilidade de lidar com sobreajuste (overfitting) e otimizar a precisão das previsões. O modelo XGBoost demonstrou um desempenho superior em

comparação com as outras técnicas utilizadas tanto na avaliação interna (score) quanto na estimativa do erro médio (RMSE). No entanto, é importante destacar que a métrica do RMSE, embora seja um indicador útil da precisão do modelo, precisa ser interpretada com cautela. Um valor de 3,02 para o RMSE indica que, em média, o modelo erra em cerca de três dias na previsão da data de saída do contêiner. Essa margem de erro pode ser considerada aceitável dependendo do contexto e das necessidades operacionais do porto. Esses resultados reforçam a eficácia das técnicas de aprendizado de máquina, especialmente o modelo XGBoost baseado em árvores de decisão, na previsão do tempo de estadia de contêineres nos portos. Essas abordagens avançadas proporcionam aos planejadores portuários uma valiosa ferramenta para tomar decisões informadas e estratégicas na alocação e gerenciamento dos contêineres, melhorando a eficiência operacional e otimizando o uso do pátio.

Tabela 1: RMSE das técnicas utilizadas

Avaliação das técnicas		
Técnica	Score	RMSE
Extreme Gradient Boosting	0,89	3,02
Floresta Aleatória	0,77	4,10
Regressão Linear Generalizada	0,70	4,20
Redes Neurais	0,88	4,31
Árvore de Decisão	0,92	5,66
Regressão Linear Múltipla	0,65	6,14

5. Conclusão

A aplicação de técnicas de aprendizagem de máquina para auxiliar na operação logística portuária mostrou-se promissora com base nos dados fornecidos. Cada técnica utilizada apresentou vantagens e desvantagens em termos de assertividade e custo. As técnicas com maior assertividade também exigiram ajustes mais complexos e tempo de execução mais longo. Em resumo, a utilização dessas técnicas pode trazer benefícios significativos para a operação logística portuária. Ao utilizar técnicas de aprendizagem de máquina, os portos podem obter previsões mais precisas e confiáveis para o tempo de estadia dos contêineres, o que resulta em melhorias na eficiência operacional. Com previsões mais precisas, os planejadores portuários podem tomar decisões mais informadas e estratégicas sobre o posicionamento e gerenciamento dos contêineres, otimizando o uso dos recursos disponíveis. No entanto, é importante ressaltar que a implementação dessas técnicas requer um cuidadoso processo de ajuste e configuração, além de considerações sobre o tempo de execução e custos associados. Cada técnica tem suas próprias características e requisitos, e é fundamental avaliar a viabilidade e o impacto antes da implementação em larga escala. Em geral, os benefícios potenciais da aplicação de técnicas de aprendizagem de máquina na operação logística portuária são significativos. Com a utilização adequada dessas técnicas, os portos podem melhorar a eficiência, reduzir custos, otimizar a gestão da cadeia logística. É importante continuar pesquisando e explorando abordagens inovadoras para maximizar os benefícios dessas técnicas e enfrentar os desafios específicos da indústria portuária.

Agradecimentos

O autor agradece ao Terminal de Contêineres de Paranaguá pela disponibilização dos dados e ao Professor Dr Anderson Ara pela orientação.

Referências

0.8

GAETE, Myriam et al. A dwell time-based container positioning decision support system at a port terminal, International Conference on Operations Research and Enterprise Systems, p. 128-139, 2017

CHEN, T. GUESTRIN C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining, p. 785–794, 2016

MERCKX, Filip. The issue of dwell time charges to optimize container terminal capacity, Proceedings IAME 2005 Annual Conference, p.23-25, 2005.

BREIMAN, Leo. Random forests. Machine learning, v. 45, n. 1, p. 5-32, 2001.

HAYKIN, Simon. Neural Networks: A Comprehensive Foundation (2 ed.). Prentice Hall, 1998.

MCCULLAGH, Peter; NELDER, John A. Generalized linear models. Routledge, 1986.