UNIVERSIDADE FEDERAL DO PARANÁ

ALEXANDRE AUGUSTO LACO DZAZIO

PREVISÃO DAS VENDAS DE UMA EMPRESA MULTINACIONAL ATRAVÉS DE MODELOS DE SÉRIES TEMPORAIS

CURITIBA, PR
2023

ALEXANDRE AUGUSTO LACO DZAZIO

PREVISÃO DAS VENDAS DE UMA EMPRESA MULTINACIONAL ATRAVÉS DE MODELOS DE SÉRIES TEMPORAIS

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Especialista em Inteligência Artificial Aplicada no Programa de Pós-Graduação em Inteligência Artificial Aplicada, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná.

Orientador: Prof. Dr. Razer Anthom Nizer Rojas Montaño.

CURITIBA, PR
2023

MINISTÉRIO DA EDUCAÇÃO
SETOR DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
CURSO DE PÓS-GRADUAÇÃO INTELIGÊNCIA ARTIFICIAL APLICADA - 40001016348E1

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação INTELIGÊNCIA ARTIFICIAL APLICADA da Universidade Federal do Paraná foram convocados para realizar a arguição da Monografia de Especialização de **ALEXANDRE AUGUSTO LACO DZAZIO** intitulada: **Previsão das Vendas de uma Empresa Multinacional Através de Modelos de Séries Temporais**, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua _____ no rito de defesa.

A outorga do título de especialista está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 09 de Novembro de 2023.

RAZER ANTHOM NIZER ROJAS MONTAÑO
Presidente da Banca Examinadora

JAIME WOJCIECHOWSKI
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Previsão das Vendas de uma Empresa Multinacional Através de Modelos de Séries Temporais

Alexandre Augusto Laco Dzazio
Setor de Educação Profissional e Tecnológica
Universidade Federal do Paraná
Curitiba, Brasil
alexandreldzazio@gmail.com

Prof. Dr. Razer Anthom Nizer Rojas Montaño
Setor de Educação Profissional e Tecnológica
Universidade Federal do Paraná
Curitiba, Brasil
razer@ufpr.br

***Resumo*—As previsões de venda de uma empresa desempenham um importante papel para o planejamento estratégico e gestão de estoques, influenciando diretamente nos seus resultados. Este trabalho visa obter o melhor modelo para previsão das vendas de uma de uma multinacional da região de Ponta Grossa. Realizou-se a comparação entre algoritmos que usaram apenas dados históricos de faturamento, com modelos onde foram acrescentados dados derivados dos arquivos de *Electronic Data Interchange* (EDI), recebidos dos clientes. Foram testados modelos séries temporais univariadas e regressões lineares, através de algoritmos ARIMA, Redes Neurais Recorrentes, XGBoost, SVR e Florestas Aleatórias. Observa-se que, para todos os clientes, foram obtidos modelos de aprendizado de máquina que apresentaram valores menores de RMSE que as demandas previsionais enviadas pelos clientes.**

***Palavras-chave*—projeção de vendas, EDI, aprendizado de máquina**

***Resumo*—Sales forecasting plays an important role in every company budget and inventory planning, impacting directly its results. This study aims to obtain the best sales forecasting model for a multinational company located in Ponta Grossa. Algorithms trained using only historical sales data were compared within models where data from Electronic Data Interchange (EDI) received from customers were also used as an input for the training models. Unidimensional time series models and linear regressions were tested, including ARIMA, Recurrent Neural Networks, XGBoost, SVR, and Random Forests. The Artificial Intelligence models achieved lower RMSE values than the demand forecast sent by the customers.**

***Index Terms*—sales forecasting, EDI, machine learning**

I. DESENVOLVIMENTO

O estoque de uma empresa é um assunto crucial para a saúde financeira do negócio. Itens em inventário não possuem liquidez, sendo prejudicial para o capital de giro. Além disso, tem-se ainda os custos de armazenamento, somados à gastos com transporte e obsolescência [1]. Todavia, baixos níveis de inventário podem acarretar distúrbios nas linhas de produção, atrasos na entrega e, até mesmo, problemas nos atendimentos aos clientes [2].

A pandemia de COVID-19 acentuou um cenário que já era incerto, tornando a cadeia de suprimentos ainda mais imprevisível. A escassez de matérias primas, o aumento dos tempos de trânsito para importação e o maior custo de *commodities* são apenas alguns dos exemplos. As empresas foram assim obrigadas a aumentar os estoques de componentes [3]. Em paralelo, modelos de previsão de demandas tem obtidos excelentes resultados como ferramentas para redução dos estoques e obtenção de níveis ótimos [2].

Muito comum na indústria automotiva, protocolos de comunicação via *Eletronic Data Interchange* (EDI) são utilizados para otimizar o tráfego de informações logísticas entre clientes e fornecedores (B2B). Nos arquivos de EDI os clientes enviam para os fornecedores uma previsão dos itens que serão comprados nos próximos meses. Entretanto, as quantidades informadas nestes arquivos tratam-se de estimativas. Desta forma, estes dados não são precisos, havendo alterações nas quantidades informadas constantemente, o que acarreta em problemas de planejamento e faltas de componentes ou excesso de estoques [4].

Cada modelo de previsão de demandas apresenta vantagens e desvantagens. No contexto de séries temporais, modelos estatísticos apresentam uma implementação mais simples e direta, quando comparados a algoritmos de aprendizado de máquina, onde um maior esforço é necessário. Porém, estes tendem a apresentar um desempenho superior em cenários de maior complexidade e relações não lineares. Portanto, é necessário sempre testar diferentes cenários [5].

Desta forma, este trabalho testa diversas técnicas de previsão, para obter o melhor modelo de predição da quantidade de peças vendida por uma empresa fornecedora de componentes automotivos, da região de Ponta Grossa. Modelos de aprendizado de máquina e modelos estatísticos são comparados. Também avaliou-se os resultados obtidos através de modelos treinados com diferentes fontes de dados. Modelos de séries temporais ARIMA e *Extreme Gradient Boosting* (XGBoost ou XGB), treinados apenas com a quantidade de peças faturadas ao decorrer do tempo, são comparados com modelos de regressão lineares baseados nos arquivos de EDI, recebidos dos clientes. Nestes, foram utilizados os métodos de Redes Neurais Profundas (RNAs), Regressão por Máquina de Vetores de Suporte (SVR), Floresta Aleatória (RF) e XGBoost.

A. Descrição dos dados

Para a realização deste trabalho foram escolhidos, dentre toda carteira de clientes desta multinacional, seis de seus principais clientes. Estes pertencem a dois grupos de montadoras automo-

tivas, para as quais são fornecidos componentes para veículos leves. Todas as seis plantas são independentes entre si, apresentando comportamentos distintos. Suas fábricas encontram-se espalhadas por todo território Nacional, América Latina e Europa. Ressalta-se que, devido a políticas de confidencialidade da empresa, alguns dos dados foram anonimizados.

1) Base de dados de faturamento: A base de dados de faturamento consiste no histórico de todas as peças vendidas pela empresa ao decorrer de janeiro de 2019 até agosto de 2023. Este arquivo foi disponibilizado em uma planilha eletrônica do Microsoft Excel©. Para cada uma das vendas realizadas têm-se as informações: código do item vendido, quantidade de peças faturada, código do cliente que realizou a compra e a área de negócio deste componente.

Os gráficos da figura 1 foram traçados a fim de se exemplificar o comportamento de cada um dos clientes ao decorrer do tempo. Ao se analisar as quantidades vendidas pode-se observar que o comportamento de cada um dos parceiros é completamente distinto dos demais. Também são observados em 2021 e 2022 meses com comportamentos anômalos, em detrimento da pandemia de COVID-19.

De forma complementar, a tabela I apresenta a quantidade total vendida, para cada um dos seis clientes, nos últimos cinco meses. Observa-se que as grandezas de volume também são diferentes.

A variação das quantidades faturadas entre os diferentes anos e diferentes meses também são informações relevantes para a análise de dados temporais. As figuras 2 e 3, respectivamente, exibem os *box plots* das distribuições das quantidades vendidas nos anos e meses. Nota-se que a quantidade peças vendidas em 2019 foi inferior aos demais anos. Também pode-se observar uma queda no faturamento de 2021, havendo ainda uma grande variância nos dados. Neste ano tiveram-se vários efeitos adversos da pandemia de COVID-19, o que resultou na elevada variância. A partir do gráfico mensal, pode-se observar que também existem grandes variâncias, tanto nas quantidades vendidas dentro do próprio mês, bem como entre os diferentes meses do ano. O mês de outubro é o único com comportamento estável.

2) Base de dados de EDI: Os arquivos das demandas previsionais, recebidos por EDI dos clientes, também foram exportados como planilhas eletrônicas do Microsoft Excel©. Nestes, a quantidade prevista para a compra nas próximas 30 semanas subsequentes são informadas. Ou seja, são disponibilizadas estimativas do que o cliente planeja comprar nos próximos meses, em quantidade de peças.

Na empresa em que o presente estudo foi realizado cada um dos clientes envia um novo arquivo de EDI semanalmente. Uma vez que toda a previsão é atualizada a cada novo arquivo de EDI, há uma sobreposição de predições. Diferentes arquivos preveem o consumo de uma mesma data, porém informando quantidades diferentes.

Dentre o período de janeiro de 2021 e agosto de 2023 cerca de 140 arquivos foram gerados. Os atributos desta base de dados são: quantidade prevista para a compra; data prevista para a compra desta quantidade (dispostas como colunas das tabelas); código do cliente; item vendido e a área de negócio.

B. Métodos

Uma vez que os dados brutos disponibilizados pela empresa não se encontravam nos formatos ideais para o processamento dos modelos de aprendizado de máquina, foi necessário submetê-los a uma etapa de pré-processamento.

1) Pré-processamento da base de dados de faturamento: Foram filtrados os clientes relevantes para este estudo, de forma a manter apenas os seis parceiros relevantes para este estudo, onde os demais foram removidos. Para o modelo de negócio da empresa em questão é indiferente em qual dia da semana o faturamento acontece. Desta forma, todas as datas de faturamento foram ajustadas para a segunda-feira anterior. A área de negócio em que a venda foi realizada foi convertida em uma informação mais relevante do ponto de vista de negócio, ajustada para a família do produto que foi vendido. E, para reduzir o custo computacional necessário para o processamento de dados, o código do item vendido foi desconsiderado e um agrupamento por cliente, família e semana de faturamento foi realizado. Por fim, uma vez que a família é um atributo de classe, foi convertida, através do método *OneHotEncoder*, da biblioteca *scikit-learn*, para colunas. As semanas em que nenhuma venda é realizada são relevantes para o negócio, uma vez que o modelo deve ser capaz de predizer épocas em que nenhuma venda será efetivada. Visto que as mesmas não existiam no relatório e, para que o modelo possa aprender estes comportamentos, os mesmos devem estar representados nos dados. Estas informações foram inseridas através de iterações por todas as semanas dos anos: caso não existisse nenhum dado na semana em que estava sendo iterada, uma nova linha com quantidade igual a zero era inserida.

Por fim, todos os atributos numéricos foram normalizados, conforme a equação (1), onde X_{norm} é o novo dado normalizado; X é o dado real da instância; X_{min} é menor valor do atributo e X_{max} é o maior valor existente entre todos X.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

Conforme observado nas descrições dos dados, um comportamento de compra totalmente distinto é observado para cada um dos clientes. Desta forma, optou-se por treinar modelos individuais para cada um dos clientes. Assim, a base de dados foi dividida em subconjuntos, originados a partir da segmentação, ao filtrar individualmente cada parceiro.

2) Pré-processamento da base de dados de EDI: Para esta etapa foi necessário unificar todos os 140 arquivos em uma única tabela. Foram iterados por todos os arquivos, onde em cada iteração as datas de entrega, dispostas como várias colunas, eram transformados em um único atributo, através da função *melt*, da biblioteca *pandas*. A data em que o arquivo foi recebido também foi adicionada como um novo atributo. Todos os dados foram empilhados todos em uma tabela. Em seguida, processos similares aos relatados no pré-processamento dos dados de

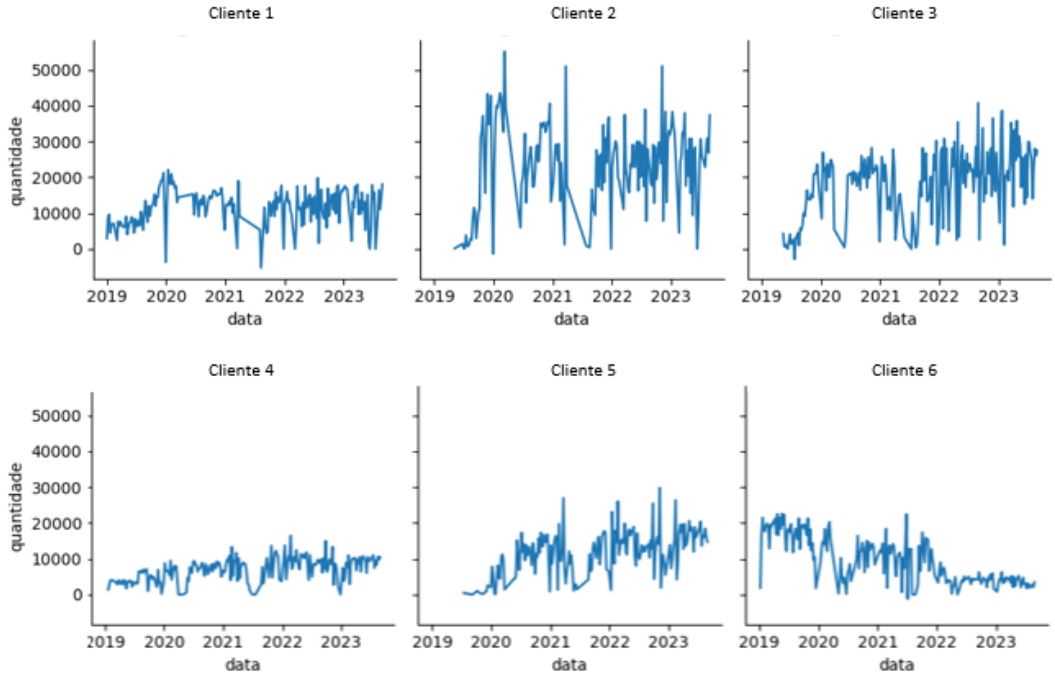

Figura 1: Gráficos das quantidades vendidas, entre 2019 e 2023, para cada um dos clientes.
Fonte: O Autor (2023)

Tabela I: Quantidade total vendida por cada um dos cliente, entre abril de 2023 até agosto de 2023.

Cliente	Quantidade Vendida
1	228.706
2	458.284
3	539.303
4	205.202
5	366.571
6	62.893

Fonte: O autor (2023)

faturamento foram realizados. Os clientes relevantes foram filtrados, removendo os demais; todas as datas ajustadas para a segunda-feira anterior; a área de negócio convertida em família do produto; os dados agrupados por cliente, família e semana de faturamento; *OneHotEncoder* utilizado para converter as classes em atributos; semanas inexistentes, sem nenhuma quantidade prevista, adicionadas com valor igual a zero; atributos numéricos normalizados, fazendo-se uso da equação (1) e subconjuntos para cada um dos cliente gerados.

O elemento que se pretende prever com este trabalho é a quantidade real faturada. Porém, este atributo não existe na base de dados de EDI, existindo apenas na base de dados do faturamento. Desta forma, foi necessário mesclar a consulta do EDI com a base de faturamentos. Foi realizado uma iteração para que em cada uma das previsões enviadas pelos clientes também existisse um atributo referente a quantidade real faturada naquela data. Assim, em todas as repetições, foram consultadas informações da tabela de faturamentos e adicionadas à tabela de dados do EDI. Conforme mencionado anteriormente, para uma mesma data do ano existem vários arquivos de EDI informando diferentes quantidades previstas. A diferença na quantidade prevista, em uma mesma data, ao decorrer dos arquivos de EDI enviados em datas diferentes é relevante para o negócio e pode agregar valor ao modelo. Desta forma, para cada uma das instâncias de EDI, também foi consultada a quantidade que estava sendo prevista para esta mesma data nos arquivos de uma, duas, três e quatro semanas antes a data do EDI da instância. A tabela II contém a lista com todos os atributos presentes nesta fonte de dados.

3) Métricas de avaliação: Neste trabalho será adotado como o melhor modelo aquele que apresentar a menor Raiz Do Erro Quadrático Médio (RMSE) na previsão de vendas realizada. O RMSE pode ser calculado a partir da equação (2), onde y_j é o valor real; $\hat{y}_j$ é o valor previsto e n é o número de instâncias.

Tabela II: Atributos disponíveis na base de dados derivada dos arquivos de EDI.

Atributo	Descrição
$DataPrevisao$	Data que o EDI está prevendo a quantidade faturada (não utilizado durante o treinamento do modelo)
$EDI\_DataRef$	Data em que o EDI desta instancia foi recebido (não utilizado durante o treinamento do modelo)
$DataDiff$	Diferença, em dias, entre a data do EDI (*EDI_DateRef*) e a data a ser prevista (*DataPrevisao*)
$Cliente$	Atributo referente ao código do cliente desta instância (não utilizado durante o treinamento do modelo)
$QTD\_Faturada$	Quantidade real de peças vendidas na *DataPrevisao* (*target* do modelo)
EDI_L0	Quantidade prevista na *DataPrevisao* no EDI recebido na data *EDI_DateRef*
EDI_L1	Quantidade prevista na *DataPrevisao* no EDI recebido uma semana antes da data *EDI_DateRef*
EDI_L2	Quantidade prevista na *DataPrevisao* no EDI recebido duas semanas antes da data *EDI_DateRef*
EDI_L3	Quantidade prevista na *DataPrevisao* no EDI recebido três semanas antes da data *EDI_DateRef*
EDI_L4	Quantidade prevista na *DataPrevisao* no EDI recebido quatro semanas antes da data *EDI_DateRef*
$Familia1$	Atributo binário referente a família do item desta instância
$Familia2$	Atributo binário referente a família do item desta instância
$Familia3$	Atributo binário referente a família do item desta instância
$Month$	Atributo referente ao mês da *DataPrevisao*
$DayOfMonth$	Atributo referente ao dia do mês da *DataPrevisao*
$WeekOfYear$	Atributo referente a semana do ano da *DataPrevisao*

Fonte: O autor (2023)

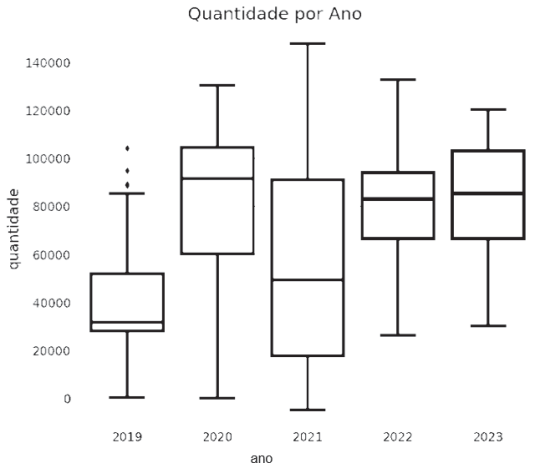

Figura 2: Gráfico de *boxplot* das quantidades de peças vendidas em cada um dos anos de 2019 à 2023.
Fonte: O autor (2023)

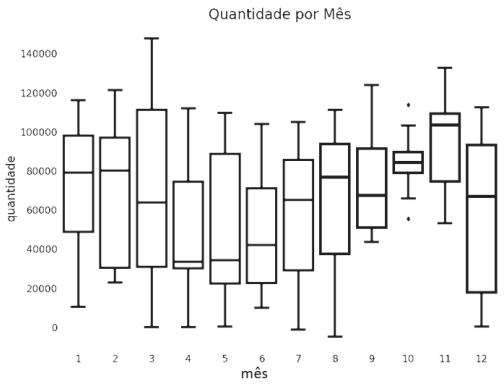

Figura 3: Gráfico de *boxplot* das quantidades de peças vendidas em cada um dos meses do ano.
Fonte: O autor (2023)

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \quad (2)$$

Devido as características do negócio, foi definido que o tempo ideal para predição seria de 5 meses. Portanto, visando uma simulação prática, os últimos 5 meses disponíveis nas bases de dados - 01 de abril de 2023 até 31 de agosto de 2023 - foram segmentados para validação dos dados.

Uma vez que os arquivos de EDI já são uma previsão do que será vendido para cada um dos clientes nos próximos meses, para que os modelos de aprendizado de máquina possuam aplicação prática, o RMSE dos modelos deve ser inferior ao encontrado nos arquivos de EDI. Assim, foi calculado o RMSE das previsões informadas pelos clientes no EDI recebido no final de março, onde estavam contidas as previsões de consumo entre abril e agosto. Estes dados foram denominados valores de referência.

4) Treinamento dos modelos de séries temporais: Para o treinamento dos modelos de séries temporais (ST) foi utilizado a base de dados de faturamento. Foram mantidos apenas os atributos de data de venda e quantidade de peças vendidas, desconsiderando-se as demais colunas. Conforme descrito anteriormente, os últimos 5 meses foram reservados para teste, enquanto o período de janeiro de 2019 até março de 2023 foi empregado como base de treino. Totalizou-se uma proporção de 90% para treino e 10% para teste. A técnica de *holdout* foi

utilizada para validar e testar os dados.

Uma vez que modelos ARIMA estão entre os modelos estatísticos mais amplamente utilizados para previsão de séries temporais [6] optou-se por treinar uma série temporal através deste método. Visando a comparação dos resultados obtidos em modelos estatísticos com técnicas de aprendizado de máquina, escolheu-se treinar também um modelo XGBoost.

O ARIMA consiste em prever os valores futuro a partir da relação dos valores do passado e seus erros. É composto por três partes: Auto Regressiva (AR), Média Móvel (MA) e Integrada (I) [7]. O treinamento foi realizado utilizado a biblioteca *statsmodels*, onde os melhores hiperparâmetros *p*, *d* e *q* foram obtidos através do método *auto_arima* da biblioteca *pmdarima*. Neste caso, uma busca é realizada com base em uma heurística, onde diversos valores de *p*, *d* e *q* são testados e comparados automaticamente, com base no Critério de Informação de Akaike (AIC).

Já para o XGBoost utilizou-se da biblioteca *xgboost*, por meio do *booster gbtree*. A busca em *grid* foi empregada, variando os seguintes hiperparâmetros: número de estimadores (NE), taxa de aprendizado (TA) e profundidade máxima (PM). Os valores para cada um destes encontram-se disponíveis na tabela III.

Em ambos os casos foram utilizadas janelas únicas, com todos os dados de treino, para aprendizado dos padrões temporais, não sendo empregadas técnicas de janelas deslizantes. Uma previsão em vários passos (*multi-step*) foi realizada, onde todos os valores futuros foram gerados em uma única etapa.

5) Treinamento dos modelos de regressão linear: Os modelos de regressão linear (RL) consistiram no uso da base de dados derivada do EDI, descrita na tabela II. O treinamento consistiu em um aprendizado de máquina onde os valores de faturamento futuro são previstos com base nos arquivos que contém a previsão do consumo enviada pelo próprio cliente. Uma vez que esta fonte de dados só continha dados entre 2021 e 2023, o período de janeiro de 2021 até março de 2023 foi empregado como base de treino e abril de 2023 até setembro de 2023 foram utilizados para teste. Assim, tem-se uma proporção de 75% para treino e 25% para teste. De forma similar aos modelos de série temporal, também se utilizou da técnica de *holdout*.

As regressões lineares foram desenvolvidas através de técnicas de RNA, SVR, XGBoost e RF. Em todos os casos uma busca em *grid* para encontrar os melhores hiperparâmetros foi realizada, onde a RMSE foi utilizada como função de avaliação. Os parâmetros utilizados nestas buscas também se encontram disponíveis na tabela III. No algoritmo RNA foram comparados os resultados variando o tamanho de camada oculta (TCO), função de ativação (FA) e o *alpha*. Já no modelo SVR avaliou-se diferentes funções de *kernel* (kr), *C* e epsilon (e). Similar ao modelo de série temporal, o XGBoost foi treinado variando os números de estimadores (NE), a taxa de aprendizado (TA) e a profundidade máxima (PM). Por fim, para o modelo de RF optou-se por avaliar hiperparâmetros semelhantes, analisando o número de estimadores (NE), a profundidade máxima (PM) e a quantidade mínima de amostras por ramificação (QMS).

Conforme mencionando anteriormente, foram treinados modelos distintos para cada um dos seis clientes. Uma vez que os modelos série temporal e regressão linear totalizam seis técnicas, foram treinados no total 36 modelos distintos.

C. Tecnologias

A base de dados de faturamento e EDI foram extraídas do sistema SAP R/3, versão 7.70, sendo exportados para planilhas eletrônicas do Microsoft Excel©. A versão 2208 foi a utilizada para visualização destas planilhas.

Os modelos foram treinados na nuvem Microsoft Azure, através da plataforma Azure Machine Learning, em um workspace com uma máquina virtual do tipo `Standard_E4ds_v4`, com 4 núcleos, 32GB de memória RAM e processador *Intel Xeon Platinum 8272CL*. Foi utilizado a linguagem de programação Python, versão 3.8.5, em um notebook Jupyter, versão 5.2. O cliente Azure encontrava-se na versão 1.26.3. Quanto aos Frameworks de aprendizado de máquina ressalta-se o uso do *xgboost* versão 1.7.5; *statsmodels* versão 0.13.2; *Scikitlearn* versão 1.1.3 e *pmdarima* versão 2.0.3. Ainda se utilizou das bibliotecas *Pandas*, versão 1.1.5, *Numpy*, versão 1.21.6, *matplotlib* versão 3.7.2, *Seaborn* versão 0.12.2. As bibliotecas *datetime* e *math* também foram utilizadas, estas na versão padrão já disponível no ambiente padrão do *python*.

II. RESULTADOS E DISCUSSÕES

A. Valores de Referência

A tabela IV demonstra o RMSE obtido ao comparar as quantidades descritas no último EDI de março com o que de fato foi vendido, para cada um dos parceiros, nos próximos 5 meses.

Importante salientar que, conforme observado na tabela I, a magnitude da quantidade total vendida neste período é diferente para cada um dos clientes. Neste caso, não é possível comparar diretamente o RMSE entre os diferentes parceiros. Assim, é adequado apenas para comparação da performance somente do mesmo cliente entre os diferentes modelos.

B. Modelos de série temporal

O valor de RMSE obtido para cada cliente para os dois modelos de série temporal pode ser visualizado na tabela V. Nesta tabela também se encontra descrito o melhor hiperparâmtro, utilizado para realização desta predição.

Com o treinamento dos modelos ARIMA, nota-se que o RMSE nas amostras de teste variou entre 1.719 e 12.259 ao decorrer dos diferentes clientes. Este é o valor total do período de abril à agosto de 2023. Também observa-se que várias composições diferentes dos componentes *p, q* e *d* foram obtidas. Houve modelos em que o menor RMSE foi encontrado com um valor do parâmetro autorregressivo igual a zero, enquanto em outros a média móvel foi nula. Houve também um modelo de série estacionária, que o valor de integração é igual a zero.

Os modelos XGBoost também tem seus resultados de RMSE e melhores hiperparâmetros exibidos na tabela V. Os valores

Tabela III: Hiperparâmetros utilizados na busca em *grid* dos algoritmos de aprendizado de máquina.

Modelo	Hiperparâmetro
$XGBoostST$	NE: 50, 100, 500 e 1000; TA: 0,001, 0,01 e 0,05; PM: 3, 10, 50 e 100
$XGBoostRL$	NE: 50, 100, 500 e 1000; TA: 0,001, 0,01 e 0,05; PM: 3, 10, 50 e 100
RNA	TCO 50, 100 e 200; FA: *relu* e *tanh*; *alpha*: 0,001, 0,01 e 0,02
SVR	kr: linear e o *radial basis function* (RBF); *C*: 1, 5 e 10; e: 0,1 e 0,5
RF	NE: 50, 100, 500; PM: 5, 10, 20 e 50; QMS: 1, 2 e 4

Fonte: O autor (2023)

Tabela IV: Erro quadrático médio das previsões de vendas informados pelos clientes nos arquivos de EDI para os períodos de abril de 2023 até agosto de 2023.

Cliente	RMSE
1	11.462
2	13.562
3	8.262
4	2.303
5	3.599
6	2.153

Fonte: O autor (2023)

Tabela V: Valores de RMSE resultantes do treinamento dos modelos de série temporal e os melhores hiperparâmetros encontrados.

Cliente	Parâmetros ARIMA	RMSE ARIMA	Parâmetros XGBoost	RMSE XGBoost
1	p: 1, d: 0; q: 2	6.004	TA: 0,01; PM: 3; NE: 500	5.851
2	p: 2, d: 0; q: 0	12.259	TA: 0,01; PM: 3; NE: 100	8.656
3	p: 1, d: 1; q: 1	7.995	TA: 0,01; PM: 3; NE: 500	8.734
4	p: 0, d: 1; q: 1	1.731	TA: 0,01; PM: 3; NE: 100	4.234
5	p: 0, d: 1; q: 1	2.776	TA: 0,01; PM: 3; NE: 100	8.656
6	p: 2, d: 1; q: 1	1.719	TA: 0,01; PM: 3; NE: 500	1.612

Fonte: O autor (2023)

de RMSE variaram entre 1.612 e 8.734. Em todos os casos a melhor taxa de aprendizado encontrada foi de 0,01 e a profundidade máxima foi a menor das disponíveis, sendo igual a 3. Os números de estimadores foram sempre 100 ou 500, dependendo do cliente.

Nota-se que para três dos seis clientes (1, 2 e 6) o desempenho dos algoritmos XGBoost foi superior, enquanto para a outra metade (3, 4 e 5) o ARIMA apresentou um resultado mais satisfatório.

Rožanec *et al.* também compararam modelos estáticos com modelos de aprendizado de máquina, em cenários de uso similares, comparando a previsão da demanda de componentes automotivos do mercado original. Porém, para os casos de Rožanec *et al.* todos os modelos de aprendizado de máquina apresentaram um resultado mais satisfatório que as séries estatísticas [8].

C. Modelos de regressão linear

Os melhores hiperparâmetros encontrados para os modelos de regressão linear, treinados à partir da base de EDI, podem ser visualizados na tabela VI. A partir destes hiperparâmetros os modelos foram treinados e predições para as bases de teste foram geradas, onde o RMSE obtido pode ser visualizado na tabela VII.

Nota-se que para os algoritmos de XGBoost os hiperparâmetros encontrados foram similares ao XGBoost da série temporal. As melhores taxas de aprendizado encontradas também foram iguais a 0,01 e a maioria das profundidades máximas obtidas foi de 3. Já os números de estimadores variaram para alguns dos clientes, quando comparado ao modelo de série temporal. Nos modelos de regressão linear, o melhor número de estimadores encontrado foi sempre de 500.

Observa-se que os modelos de SVR tiveram o melhor resultado para os clientes 1, 2, 3 e 4, enquanto o modelo de RNA foi o melhor para o cliente 5 e o modelo de XGBoost o melhor para o cliente 6. Porém, mesmo que os modelos de SVR apresentaram um menor resultado de RMSE, pôde-se observar que os mesmos não convergiram, onde os valores preditos foram próximos a uma linha reta, sendo similar a tendência média dos dados. A figura 4 exemplifica este comportamento, ao comparar o modelo de SVR com o modelo de RF para o

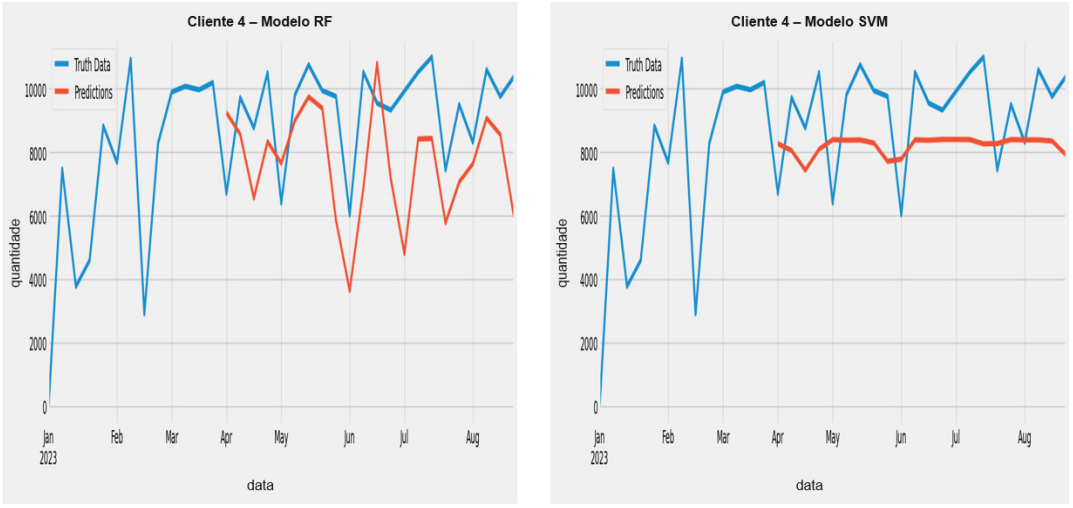

Figura 4: Gráficos da quantidade real faturada e dos valores preditos pelos algoritmos de SVR e RNA, para o Cliente 4. Fonte: O autor (2023)

Cliente 4. Uma vez que a curva de tendência é uma informação relevante para o modelo de negócio e nenhum dos algoritmos de SVR apresentaram generalizações o suficiente para aplicações práticas, os mesmos foram desconsiderados.

A tabela VIII sumariza os melhores resultados encontrados entre os 36 algoritmos. Nota-se que o RMSE de quatro dos seis clientes piorou com a inclusão dos valores de EDI como atributos de treinamento. Este fato pode ser explicado pela dimensionalidade da base de dados. Ainda, nestes outros dois clientes onde o desempenho foi inferior, a diferença entre o RMSE do melhor modelo de regressão linear e do melhor modelo de série temporal não excedeu 5%, sendo relativamente próximos. Este fato pode ser explicado pela inclusão de maiores quantidades de atributos, o que em conjuntos pequenos tende a piorar o desempenho do modelo. Além disto, os modelos de regressão linear foram treinados somente com dados posteriores a 2021, enquanto os algoritmos de série temporal foram treinados com dados desde 2019. Desta forma, observa-se que o maior custo computacional empregado no treinamento destes modelos não é justificado.

Também é possível verificar na tabela VIII que, em todos os cenários, os algoritmos desenvolvidos neste trabalho apresentaram um RMSE inferior ao observado na prática com os arquivos de EDI.

Vineeth, Kusetogullari e Boone (2020) também compararam o RMSE de diversos modelos de aprendizado de máquina para realizar a predição das vendas de peças de componentes automotivos – neste caso, no mercado de veículos pesados. Assim como neste trabalho, foram estudados algoritmos de RF, SVR, XGBoost. Ao se comparar estes três modelos nota-se que o SVR obteve o melhor resultado, enquanto o XGBoost e RF apresentaram tendências de *overfitting* [9]. Resultados diferentes dos obtidos no presente estudo, exibidos na tabela VII. Assim, pode-se observar que a ciência de dados, mesmo que aplicada em cenários similares, apresenta resultados diferentes, sendo sempre necessário testar diferentes alternativas.

Türkbayrağí, Dogu e Albayrak estudaram os efeitos de incorporação de indicadores econômicos em modelos de série temporal, onde os autores obtiveram bons resultados a partir destas informações [5]. Desta forma, recomenda-se o estudo em trabalhos futuros de possíveis otimizações derivadas da incorporação de dados financeiros globais.

D. Considerações Finais

Os modelos de aprendizado de máquina apresentaram predições melhores que os próprios arquivos de EDI, onde os valores de RMSE foram inferiores aos encontrados no dia a dia destes arquivos. Em quatro dos seis clientes técnicas de séries temporais foram superiores às técnicas de regressão lineares. Ainda, nestes outros dois clientes o valor de RMSE foi similar, diferenciando-se em menos de 5%. Porém, estes desempenhos não foram satisfatórios o suficiente para justificar os riscos atrelados a não usar o EDI como fonte oficial de informação. Visando aplicações práticas seria necessário uma acuracidade melhor do modelo. Portanto, propõe-se para pesquisas futuras a investigação de modelos treinado em nível de item, e não em nível de família, como o desenvolvido neste trabalho. Neste caso haveria o aumento da dimensionalidade da base de dados, o que poderia resultar em melhores predições, principalmente para as regressões lineares, onde tem-se uma quantidade maior de atributos, com um período temporal inferior. Outra possibilidade para melhoria nos resultados das séries temporais seria empregar técnicas de janelas deslizantes nos treinamentos. Também se sugere a análise dos efeitos da pandemia de COVID-19 nos dados. Uma vez que os dados sofreram influências destes efeitos adversos, que possivelmente não serão verificados na prática, e podem prejudicar as predições.

Tabela VI: Melhores hiperparâmetros encontrados durante o treinamento dos modelos de regressão linear.

Cliente	Parâmetros XGB	Parâmetros RNA	Parâmetros RF	Parâmetros SVR
1	TA: 0,01; PM: 3; NE: 500	FA: relu; alfa: 0.01; TCO: (50,)	PM: 50; QMS: 5; NE: 100	C: 10; e: 0.5; kr: rbf
2	TA: 0,01; PM: 3; NE: 500	FA: relu; alfa: 0.01; TCO: (50,)	PM: 20; QMS: 10; NE: 100	C: 10; e: 0.1; kr: rbf
3	TA: 0,01; PM: 50; NE: 500	FA: relu; alfa: 0.001; TCO: (100,)	PM: 50; QMS: 5; NE: 50	C: 10; e: 0.5; kr: rbf
4	TA: 0,01; PM: 3; NE: 500	FA: relu; alfa 0.001; TCO: (50,)	PM: 50; QMS: 5; NE: 50	C: 10; e: 0.5; kr: rbf
5	TA: 0,01; PM: 3; NE: 500	FA: relu; alfa: 0.001; TCO: (50,50)	PM: 20; QMS: 2; NE: 50	C: 10; e: 0.5; kr: rbf
6	TA: 0,01; PM: 3; NE: 100	FA: relu; alfa: 0.001; TCO: (100,)	PM: 20; QMS: 2; NE: 50	C: 10; e: 0.5; kr: rbf

Fonte: O autor (2023)

Tabela VII: Valores de RMSE calculados a partir das previsões realizadas com os modelos de regressão linear.

Cliente	RMSE XGB	RMSE RNA	RMSE RF	RMSE SVR
1	6.354	7.899	8.486	5.807
2	11.303	11.240	13.373	10.897
3	14.134	7.976	12.847	7.779
4	2.658	2.629	2.370	1.786
5	10.162	3.739	12.083	5.281
6	1.549	1.625	2.086	1.835

Fonte: O autor (2023)

Tabela VIII: Melhores valores de RMSE obtidos durante os treinamentos, comparados com o valor de RMSE do EDI dos clientes.

Cliente	RMSE EDI (Referência)	RMSE Melhor Modelo	Melhor Modelo
1	11.462	5.851	XGBoost ST
2	13.562	8.656	XGBoost ST
3	8.262	7.976	RNA RL
4	2.303	1.731	ARIMA ST
5	3.599	2.776	ARIMA ST
6	2.153	1.549	XGBoost RL

Fonte: O autor (2023)

REFERÊNCIAS

[1] A. Viloria and P. V. R. Acuña, "Inventory reduction in the supply chain of finished products for multinational companies," *Indian Journal of Science and Technology*, vol. 9, 2016.

[2] E. E. R. M. S. Rumetna and T. N. Lina, "Designing an information system for inventory forecasting (case study: Samsung partner plaza, sorong city)," *International Journal of Advances in Data and Information Systems*, vol. 1, 2020.

[3] C. Free and A. Hecimovic, "Global supply chains after covid-19: the end of the road for neoliberal globalisation?," *Accounting, Auditing Accountability Journal*, vol. 34, 2021.

[4] M. E. k. B. Jardini and M. Amri, "The complexity of electronic data interchange (edi) compliance for automotive supply chain," *2015 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 2015.

[5] E. D. M. G. Türkbayrağí and Y. E. Albayrak, "Artificial intelligence based prediction models: sales forecasting application in automotive aftermarket," *Journal of Intelligent Fuzzy Systems*, vol. 1, pp. 213–225, 2021.

[6] W. L. T. Weng and J. Xiao, "Supply chain sales forecasting based on lightgbm and lstm combination model," *Industrial Management Data Systems*, vol. 120, 2020.

[7] P. S. S. U. V. S. M. Gurnani, Y. Korke and S. Bhirud, "Forecasting of sales by using fusion of machine learning techniques," in *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, (Pune, India), pp. 93 – 101, 2017.

[8] M. B. F. D. M. J. M.Rožanec, B. Kažič, "Automotive oem demand forecasting: A comparative study of forecasting algorithms and strategies," *Applied Sciences*, vol. 11(15):6787, 2021.

[9] H. K. V. S. Vineeth and A. Boone, "Forecasting sales of truck components: A machine learning approach," in *Proceedings of 2020 IEEE 10th International Conference on Intelligent Systems*, (Varna, Bulgaria), pp. 510 – 516, 2020.