Previsão de Vendas: Uma análise dos resultados obtidos através de AutoML

Trabalho de Conclusão de Curso

Engenharia da Computação

Marcus Vinicius Baracho Da Silva
Orientador: Alexandre Magno Andrade Maciel

Universidade de Pernambuco
Escola Politécnica de Pernambuco
Graduação em Engenharia de Computação

MARCUS VINICIUS BARACHO DA SILVA

Previsão de Vendas: Uma análise dos resultados obtidos através de AutoML

Monografia apresentada como requisito parcial para obtenção do diploma de Bacharel em Engenharia de Computação pela Escola Politécnica de Pernambuco – Universidade de Pernambuco.

Recife, dezembro de 2021.

Silva, Marcus Vinicius Baracho Da

Previsão de Vendas: Uma análise dos resultados obtidos através de AutoML / Marcus Vinicius Baracho Da Silva. – Recife - PE, 2021.

xii, 35 f. : il. ; 29 cm.

Trabalho de Conclusão de Curso (Graduação em Engenharia de Computação) Universidade de Pernambuco, Escola Politécnica de Pernambuco, Recife, 2021.

Orientador: Prof. Dr. Alexandre Magno Andrade Maciel.

Inclui referências.

1. AutoML. 2.Previsao de Vendas. 3. Regressão de Dados –Recife (PE). I. Título. II. Maciel, Alexandre Magno Andrade. III. Universidade de Pernambuco.

ESCOLA POLITÉCNICA
DE PERNAMBUCO

MONOGRAFIA DE FINAL DE CURSO

Avaliação Final (para o presidente da banca)*

No dia 17/12/2021, às 10h00min, reuniu-se para deliberar sobre a defesa da monografia de conclusão de curso do(a) discente **MARCUS VINICIUS BARACHO DA SILVA**, orientado(a) pelo(a) professor(a) **ALEXANDRE MAGNO ANDRADE MACIEL**, sob título Previsão de Vendas: Uma análise dos resultados obtidos através de AutoML, a banca composta pelos professores:

JOABE BEZERRA DE JESUS JÚNIOR (PRESIDENTE)
ALEXANDRE MAGNO ANDRADE MACIEL (ORIENTADOR)

Após a apresentação da monografia e discussão entre os membros da Banca, a mesma foi considerada:

X Aprovada Aprovada com Restrições* Reprovada

e foi-lhe atribuída nota: 10 (dez)

*(Obrigatório o preenchimento do campo abaixo com comentários para o autor)

O(A) discente terá 7 dias para entrega da versão final da monografia a contar da data deste documento.

Joabe Bezerra de Jesus Júnior

AVALIADOR 1: Prof (a) **JOABE BEZERRA DE JESUS JÚNIOR**

AVALIADOR 2: Prof (a) **ALEXANDRE MAGNO ANDRADE MACIEL**

AVALIADOR 3: Prof (a)

* Este documento deverá ser encadernado juntamente com a monografia em versão final.

Dedico este trabalho aos meus pais e avós que me ajudaram e apoiaram durante toda a minha vida. Agradeço também à minha namorada Anna Julia e aos meus amigos Marcio, Renato e Wictor, que estiveram comigo durante todo o percurso e nunca deixaram de me incentivar e acreditar em mim. Ao meu irmão Bruno, que não está mais aqui em vida, mas que com certeza está vibrando do céu por mais essa conquista.

Agradecimentos

Em primeiro lugar gostaria de agradecer a Deus, que me deu forças para poder ultrapassar todas as dificuldades que existiram durante todos os anos da graduação. A minha família e amigos que me incentivaram nos momentos difíceis e compreenderam a minha ausência enquanto eu me dedicava à realização deste trabalho. Aos professores, em especial a Alexandre Maciel, pelos ensinamentos que me permitiram apresentar um melhor desempenho no meu processo de formação profissional ao longo do curso. e os meus colegas de curso, com quem convivi intensamente durante os últimos anos, pelo companheirismo e pela troca de experiências que me permitiram crescer não só como pessoa, mas também como formando. Gostaria de agradecer também a todo grupo GPCDA e aos meus colegas de turma Ana Catarina e Daniel Almeida que foram de fundamental importância durante todo o período do trabalho.

ESCOLA POLITÉCNICA
DE PERNAMBUCO

Autorização de publicação de PFC

Eu, **Marcus Vinicius Baracho Da Silva** autor(a) do projeto de final de curso intitulado: **Previsão de Vendas: Uma análise dos resultados obtidos através de AutoML;** autorizo a publicação de seu conteúdo na internet nos portais da Escola Politécnica de Pernambuco e Universidade de Pernambuco.

O conteúdo do projeto de final de curso é de responsabilidade do autor.

Marcus V B Da Silva

Marcus Vinicius Baracho Da Silva

Orientador(a): **Alexandre Magno Andrade Maciel**

Coorientador(a):

Prof, de TCC: **Daniel Augusto Ribeiro Chaves**

Data: 17/12/2021

Resumo

A previsão de vendas é o processo de organizar e analisar informações de forma a permitir estimar como serão as vendas. Esse processo é normalmente realizado por pessoas que possuem conhecimento na área de mineração de dados, utilizando modelos de regressão que geram resultados que auxiliam na tomada de decisões a partir de dados históricos, permitindo assim uma melhoria na tomada decisões e auxiliando no planejamento estratégico. O objetivo deste estudo é validar se os resultados obtidos através do aprendizado automatizado de máquina (AutoML) é capaz de gerar ótimos resultados se comparados com modelos que foram gerados seguindo o CRISP-DM. Os resultados foram comparados seguindo o teste de hipóteses e mostrou respostas significativas em busca da democratização do acesso ao estudo de mineração de dados.

Abstract

Sales forecasting is the process of organizing and analyzing information in a way that allows you to estimate how sales will be. This process is normally carried out by people who have knowledge in the data mining area, using regression models that generate results that help in decision making based on historical data, thus allowing for an improvement in decision making and assisting in strategic planning. The aim of this study is to validate whether the results obtained through automated machine learning (AutoML), are capable of generating robust results when compared to models that were generated following the CRISP-DM. The results were compared following the hypothesis test and showed significant answers in search of democratization of access to the data mining study.

Lista de Figuras

Lista de Tabelas

Lista de Siglas

AutoML – Aprendizado de máquina Automatizado

AdaBoost – Adaptive Boosting Regressor

LGBM – Light Gradient Boosting Regressor

RMSE - Raiz do erro médio quadrático

MSE - Erro Médio Quadrado

CRISP-DM - Processo padrão entre setores para mineração de dados

FMDEV - Framework de Mineração de Dados Educacionais

Sumário

Capítulo 1

Introdução

Segundo [1], o número de empresas abertas no ano de 2020 aumentou 6% e o de empresas fechadas apresentou uma queda de 11,3%. De acordo com [2], a competição entre empresas está cada vez mais acirrada e para enfrentar os desafios da concorrência os gestores estão utilizando com maior frequência sistemas de informações que ajudam na tomada de decisões. Essas, se erradas, sejam estratégicas, táticas ou operacionais, podem custar o futuro da empresa, assim como uma correta, definir sua sobrevivência ou sua expansão [3]. Apesar disso, muitas empresas não utilizam mecanismos adequados para prever as vendas e fazem uso normalmente de aproximações que passam por calcular a média de vendas mensais ou a média daquele mês em um historial de alguns anos [4].

Essas imprecisões no cálculo de previsões de vendas podem ocasionar diversos problemas no planejamento estratégico das empresas, como por exemplo: reserva extensa de produtos, ruptura de estoque, clientes insatisfeitos, entre outros. Nesse contexto, os dados surgem como uma importante fonte de informações para obter vantagem competitiva [5]. Para isso, as empresas contam com sistemas que lhes permitem saber como os recursos devem ser usados. Esses sistemas possibilitam extrair informações por meio de grandes volumes de dados para fornecer percepções antes não vistas com facilidade. Assim, a mineração de dados auxilia na extração de informações do banco de dados, sendo uma grande aliada no planejamento de estratégias corporativas, tendo a capacidade de permitir aos gestores um entendimento mais adequado da situação atual da empresa, permitindo assim, a geração de estratégias para aumentar as vendas, bem como, determinar a direção a ser seguida [6]. Portanto, é extremamente importante desenvolver uma previsão de vendas eficaz, de modo a gerar previsões mais assertivas e resultados robustos. [7].

No entanto, o processo para realizar essas previsões nas empresas sofre diversos desafios devido à ausência de profissionais com conhecimentos

especializados em mineração de dados. Implementar o pré-processamento de dados, definir os algoritmos a serem utilizados e seus hiperparâmetros, avaliar os resultados obtidos de maneira adequada não são tarefas simples. Como alternativa para solucionar esse problema surgiu o aprendizado de máquina automatizada, mais conhecido como AutoML. De acordo com [8], o campo do aprendizado de máquina automatizado visa tomar essas decisões de forma orientada por dados, objetiva e automatizada com o objetivo de simplificar os processos repetitivos de Mineração de Dados, os quais não exigem conhecimento técnico do usuário. O mesmo simplesmente fornece dados, e o sistema AutoML determina automaticamente a abordagem com melhor desempenho para o problema. Isso pode ser visto como uma democratização do aprendizado de máquina.

1.1 Objetivos

Este trabalho tem como principal objetivo realizar uma análise dos resultados obtidos através da previsão de vendas realizada através de um AutoML que foi implementado e disponibilizado dentro da plataforma FMDEV, onde não é necessário ter um conhecimento técnico, a fim de determinar se os resultados são equiparáveis aos resultados obtidos por modelos que seguiram a metodologia e fases (descritas na figura 1) do CRISP-DM (processo de mineração de dados que descreve abordagens comumente usadas por especialistas em mineração de dados para atacar problemas)[9] e contaram com pessoas com conhecimento técnico em sua implementação.

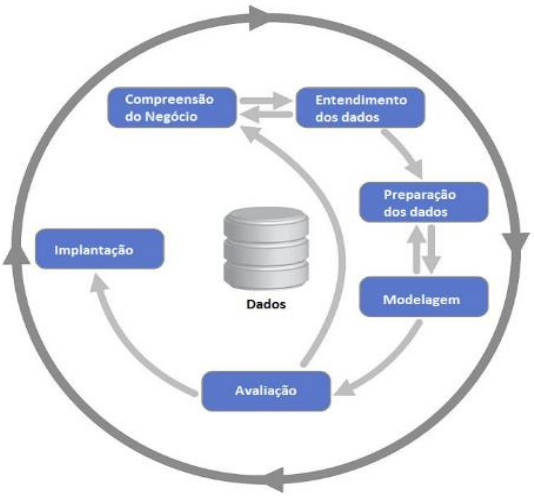

Figura 1. Fases do CRISP-DM

1.2 Justificativa

Espera-se através desse trabalho que os resultados das regressões de dados realizadas através de um AutoML sejam precisos e robustos de modo a validar a utilização de ferramentas que ajudem na democratização do estudo de mineração de dados, auxiliando desde empresas a qualquer usuário que necessite realizar previsões de acordo com sua necessidade e sua base de dados.

Capítulo 2

Fundamentação teórica

Este capítulo aborda a definição dos conceitos utilizados na construção do trabalho. Além disso, também são apresentados alguns algoritmos de regressão que serão utilizados neste trabalho.

2.1 Previsão de vendas

A previsão de vendas consiste no cálculo ou sequência de etapas que o tomador de decisão executa, com base no histórico de vendas e no entendimento do consumidor, para prever o resultado de suas vendas dentro de um período. Segundo [10] uma boa previsão das vendas ajuda os gestores a tomar decisões sobre novos investimentos e são de extrema importância para a operação eficiente da empresa.

2.2 Aprendizado de máquina (AutoML)

O aprendizado automatizado de máquina (AutoML) é o processo de automatizar as tarefas de aplicação de aprendizado de máquina a problemas do mundo real, a fim de permitir que pessoas que não são especialistas façam uso dos modelos e técnicas de aprendizado de máquina disponíveis. O AutoML abrange o pipeline completo desde o conjunto de dados bruto até o modelo de aprendizado de máquina implantável e foi proposto como uma solução baseada em inteligência artificial para o desafio crescente de aplicar aprendizado de máquina [11][12].

2.3 Métodos Ensemble

É uma técnica que utiliza vários algoritmos de aprendizagem e os combina para obter melhores resultados no desempenho preditivo. Ao contrário de outras técnicas que irão encontrar um único modelo para fazer previsões para um problema específico em um espaço com múltiplas hipóteses, Ensemble irá combinar um conjunto finito de modelos alternativos para isso. Essa combinação de vários modelos permite a eliminação da variância e, por isso, aumenta a precisão das previsões. Os métodos

de ensemble mais populares são: 1) Impulsionamento, 2) Ensacamento e 3) Empilhamento. Impulsionamento é uma técnica que aprende com os erros cometidos pelo modelo anterior para fazer melhores previsões; Ensacamento melhora os resultados do modelo por meio de árvores de decisão, reduzindo a variância; e Empilhamento é uma técnica que permite a um algoritmo de treinamento agrupar outras previsões de algoritmos de aprendizagem semelhantes.

2.4 Algoritmos

Neste trabalho, alguns algoritmos de mineração de dados foram utilizados e estão descritos a seguir:

2.4.1 Adaptive Boosting Regressor

AdaBoost, abreviação de Adaptive boosting é o algoritmo de Impulsionamento mais popular e foi proposto por [13]. Possui como diferencial o fato de que as predições mais difíceis (previsões mal realizadas) recebem um peso maior no modelo preditor seguinte, buscando uma melhor otimização do algoritmo final. Cada modelo é iniciado com um peso padrão, que irá definir seu poder de decisão no modelo final. A partir daí, conforme o treinamento dos modelos mais simples acontecem, cada "aprendiz fraco" ganha um peso maior para as predições corretas, e um peso menor para as predições em que possui um alto índice de erro. Dessa forma, os aprendizes fracos com maior precisão terão maior poder de decisão no modelo final.

2.4.2 Linear Regression

[17] introduziu a ideia de "regressão" à comunidade de pesquisa em um estudo que examina a relação das alturas dos pais e dos filhos. As regressões são chamadas de lineares quando a relação entre as variáveis preditoras e a resposta segue um comportamento linear. Neste caso, é possível criar um modelo no qual o valor de y é uma função linear de x. Exemplo: y = b + wx, ou seja, considera que a relação da resposta y com a variável x é uma função linear de alguns parâmetros: b e w. Pode-se utilizar o mesmo princípio para modelos com mais de uma variável preditora.

2.4.3 Ridge Regression

A regressão Ridge é um método de estimativa de parâmetro popular, usado para resolver o problema de colinearidade que surge frequentemente na regressão linear múltipla. Sua teoria foi introduzida por [14] e foi desenvolvida como uma possível solução para a imprecisão dos estimadores de mínimos quadrados quando os modelos de regressão linear têm variáveis altamente correlacionadas [15]. Este algoritmo faz com que recursos correlacionados tenham coeficientes semelhantes.

2.4.4 Light Gradient Boosting Machine (LGBM)

É uma estrutura de aumento de gradiente distribuído para aprendizado de máquina originalmente desenvolvido pela Microsoft. [16] Se tornou popular por conseguir lidar com grandes quantidades de dados em um curto espaço de tempo para realizar o processamento. É um método baseado em Impulsionamento que utiliza algoritmos de aprendizado baseados em árvores. Ao invés de escolher o melhor critério de divisão de nós e crescer a árvore em níveis como acontece na maioria dos métodos tradicionais, o LGBM cresce sua árvore em função de suas folhas. O algoritmo escolhe a folha mais heterogênea da árvore para crescer, ou seja, a folha que contém a maior variação de perda. Assim como todos os algoritmos baseados em Impulsionamento, o LGBM foca em melhorar o pior conjunto de divisões sucessivas que caracterizou a previsão para o conjunto de dados.

2.4.5 Bayesian Ridge Regression

A regressão bayesiana permite que um mecanismo natural sobreviva a dados insuficientes ou mal distribuídos. A resposta é extraída de uma distribuição de probabilidades ao invés de ser estimada como um único valor. Matematicamente, para obter um modelo totalmente probabilístico a resposta é assumida como gaussiana distribuída em torno da entrada. Um dos tipos mais úteis de regressão bayesiana é a regressão de Bayesian Ridge, que estima um modelo probabilístico do problema de regressão, utilizando um gaussiano esférico em torno do dado de entrada. O objetivo é encontrar os coeficientes que minimizam a soma dos quadrados dos erros aplicando uma penalidade a esses coeficientes [18].

2.5 Métricas Utilizadas

Foram utilizadas algumas métricas para avaliação dos resultados a quais estão listadas abaixo.

2.5.1 MSE

Mean squared error é a métrica mais utilizada nos problemas de aprendizado de máquina e consiste na média do erro das previsões ao quadrado. É calculada obtendo-se a diferença entre o valor predito pelo modelo e o valor real e em seguida eleva-se o resultado ao quadrado, esse processo é repetido em todos os pontos, soma-os e divide-se pelo número de elementos que foram preditos, quanto maior o valor obtido, pior será o modelo.

2.5.2 RMSE

Root mean squared error surge para melhorar a unidade obtida no MSE, e calcula a raiz quadrática média dos erros para cada ponto predito em relação aos valores reais.

2.5.3 R-Quadrado

Métrica que visa expressar a variância dos dados que o modelo construído obteve, assim dizendo o quão próximo do real está o modelo. Esta medida varia de 0 a 1

Capítulo 3

Materiais e Métodos

Neste capítulo serão abordadas a metodologia utilizada para a realização desta pesquisa e as suas ferramentas. Serão abordados pontos como a base de dados utilizada e seu tratamento, ferramentas e a metodologia utilizada.

3.1 Base de Dados

Para efeitos de comparação de maneira fidedigna de modo a analisar se os resultados obtidos através de um AutoML onde não é necessário ter nenhum conhecimento técnico de aprendizado de máquina, foi utilizada a mesma base de dados que [19] utilizou em seu trabalho.

3.2 Ferramentas

3.2.1 FMDEV

Em sua dissertação de mestrado intitulada: Desenvolvimento de uma Solução de Aprendizado de Máquina Automatizado Integrável a Múltiplos Ambientes Virtuais de Aprendizagem, [20] desenvolveu o Framework de Mineração de Dados Educacionais (FMDEV), cujo o objetivo é permitir que usuários independentemente de conhecimentos técnicos, possam construir, validar e disponibilizar baselines de aprendizado de máquina com maior produtividade e menor nível de conhecimento em ciência de dados, trazendo assim a democratização do aprendizado de máquina. Porém, a versão inicial da ferramenta disponibiliza apenas a classificação de dados, que segundo [21] visa identificar a qual classe um determinado registro pertence, podendo ser utilizado por exemplo para problemas como: Determinar quando uma transação de cartão de crédito pode ser uma fraude e identificar em uma escola, qual a turma mais indicada para um determinado aluno. Como a previsão de vendas é um problema solucionado através da regressão de dados, foi implementada e disponibilizada a análise de regressão de dados dentro do ambiente do FMDEV de

modo que a análise do presente trabalho foi realizada utilizando os resultados obtidos. Para a construção da funcionalidade de regressão de dados foi utilizada a linguagem de programação Python e a biblioteca Pycaret no backend e o framework React no frontend.

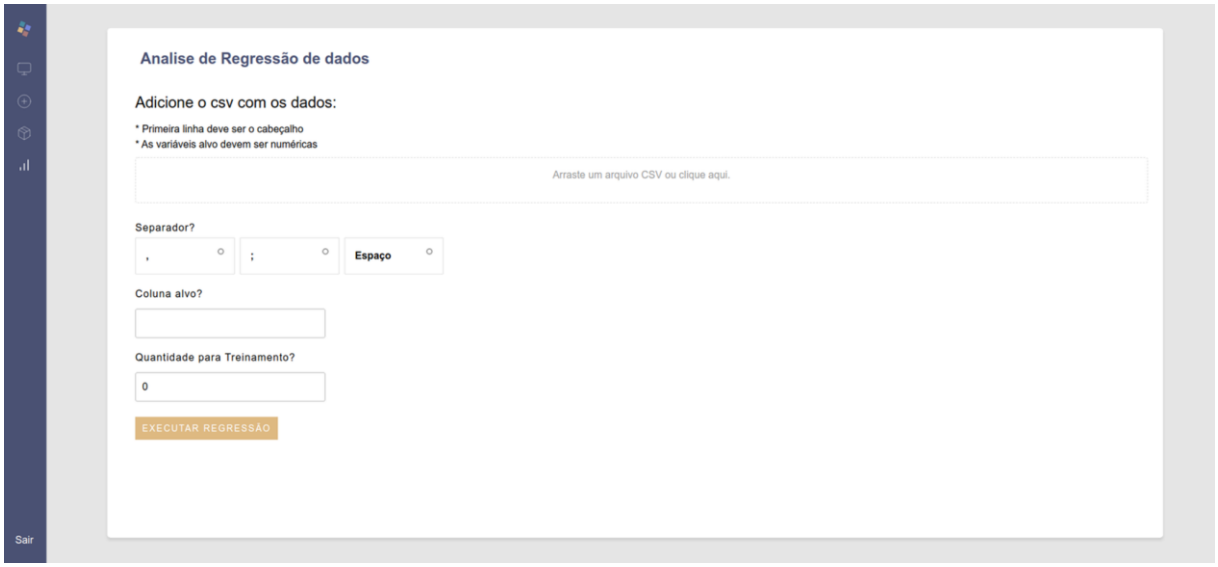

Figura 2. Inserção dos dados e definições dos parâmetros no FMDEV

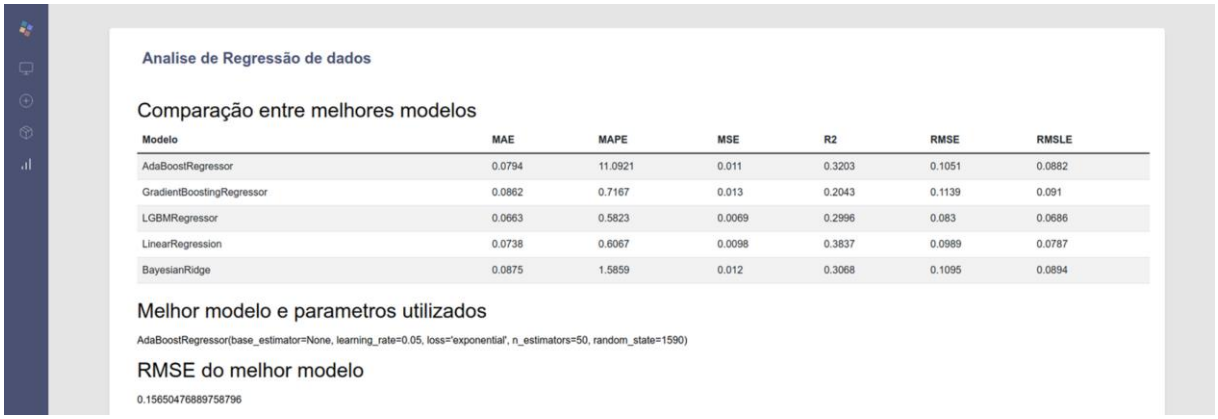

Figura 3. Resultados obtidos pelo FMDEV

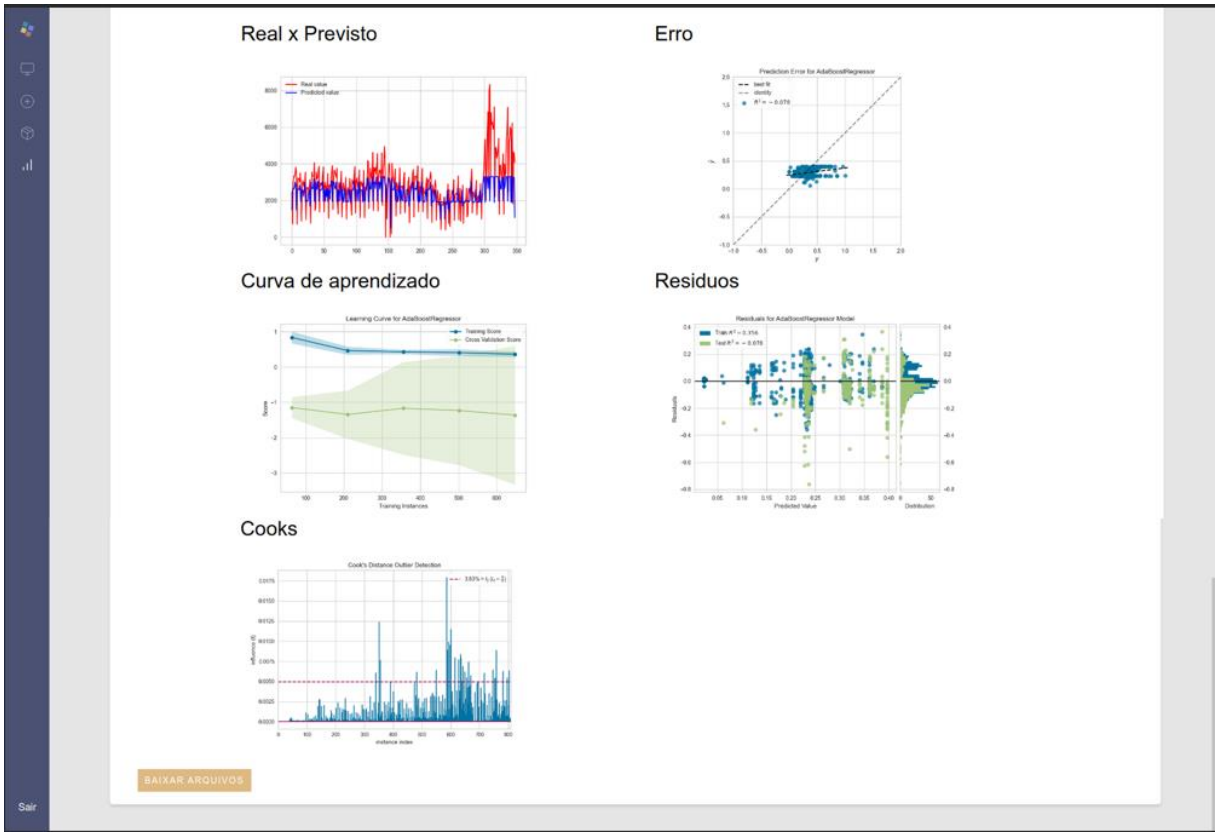

Figura 4. Gráficos obtidos pelo FMDEV

3.2.2 Python

É uma linguagem de programação de alto nível, orientada a objetos que vem ganhando bastante espaço na resolução de problemas de ciência de dados e aprendizado de máquina [22].

3.2.3 Pycaret

É uma biblioteca de aprendizado de máquina de código aberto e de baixo código em Python que automatiza fluxos de trabalho de aprendizado de máquina. É uma ferramenta de aprendizado de máquina e gerenciamento de modelos que acelera o ciclo de experimentos exponencialmente e o torna mais produtivo [23].

3.2.3 React

É uma biblioteca Javascript para criação de interface de usuários [24].

3.3 Metodologia experimental

A metodologia utilizada para a realização deste trabalho está descrita abaixo:

3.3.1 Passo 1

O primeiro passo para execução do estudo foi executar o método de regressão de dados no ambiente da ferramenta FMDEV. Definindo 70% da base de dados para realização do treinamento e 30% para testes.

3.3.2 Passo 2

Utilizar as métricas MSE e RMSE, pois são ótimas métricas para problemas nos quais grandes erros não são tolerados, como é o caso da previsão de vendas. Além disso, foram analisados também a quantidade de variância dos dados para o melhor modelo obtido através da métrica R-Quadrado.

3.3.3 Passo 3

Análise dos gráficos de Erro de Predição, Distância de Cook, Resíduos e Curva de Aprendizado de modo a traduzir os resultados obtidos.

3.3.2 Passo 4

Teste de hipóteses para comparação dos resultados obtidos através do AutoML com o que foi obtido em [21]. De modo a validar se os resultados obtidos são satisfatórios

Capítulo 4

Resultados e Discussões

Neste capítulo é exibido e discutido os resultados obtidos através da execução da previsão de vendas por meio do AutoML

4.1 Resultados

4.1.1 Comparação de performance

Comparação de performance: O primeiro passo executado pela ferramenta FMDEV é a comparação de performance entre todos os algoritmos disponíveis para regressão na biblioteca do Pycaret, utilizando hiperparâmetros padrões para o nosso estudo, os algoritmos da tabela 1 retornaram as melhores métricas para a base de testes.

Modelo	R-Quadrado	MSE	RMSE
LGBMRegressor	0.3817	0.0086	0.0925
LinearRegression	0.1579	0.0151	0.1228
BayesianRidge	0.2002	0.0123	0.1108
AdaBoostRegressor	0.0572	0.0127	0.1129
Ridge	0.3044	0.0108	0.1041

Tabela 1. Comparação de performance

4.1.2 Melhor modelo

Dentre os algoritmos da tabela 2, o AdaBoost Regressor foi o que melhor se encaixou ao nosso problema após o melhoramento dos seus hiperparâmetros.

AdaBoost Regressor	
base_estimator	none
learning_rate	0.005
loss	linear
n_estimators	240
random_state	7495

Tabela 2. Melhor modelo e seus parâmetros

4.1.3 Erro de Previsão

Um gráfico de erro de previsão faz uso da métrica R-quadrado e mostra os alvos reais do conjunto de dados em relação aos valores previstos pelo nosso modelo. Isso nos permite ver a quantidade de variância que está presente no nosso modelo. Na figura 5 é possível observar que o valor do r-quadrado obtido foi de 0.417, ou seja, nosso modelo é capaz de explicar 41,7% de variância. Como nosso problema de previsão discorre de uma decisão de compra de cupom em um ponto de venda (compra impulsiva), é esperado uma variação maior nos dados, portanto o valor obtido está dentro do esperado.

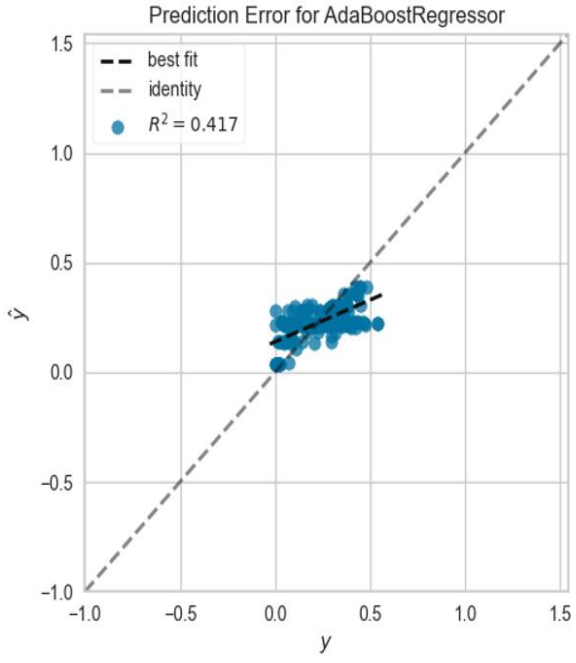

Figura 5. Erro de Previsão

4.1.4 Resíduos

O gráfico de resíduos mostra a diferença entre os resíduos no eixo vertical e a variável independente no eixo horizontal, isso permite detectar regiões que podem ser suscetíveis a mais ou menos erros. Na figura 6 é possível observar que os pontos são dispersos aleatoriamente ao redor do eixo horizontal, assim o AdaBoost Regressor, por ser um modelo que lida bem com problemas lineares pode ser considerado apropriado para nosso caso. Além disso, ainda é possível observar no histograma que nosso erro é normalmente distribuído em torno de zero, o que indica que o modelo é bem indicado.

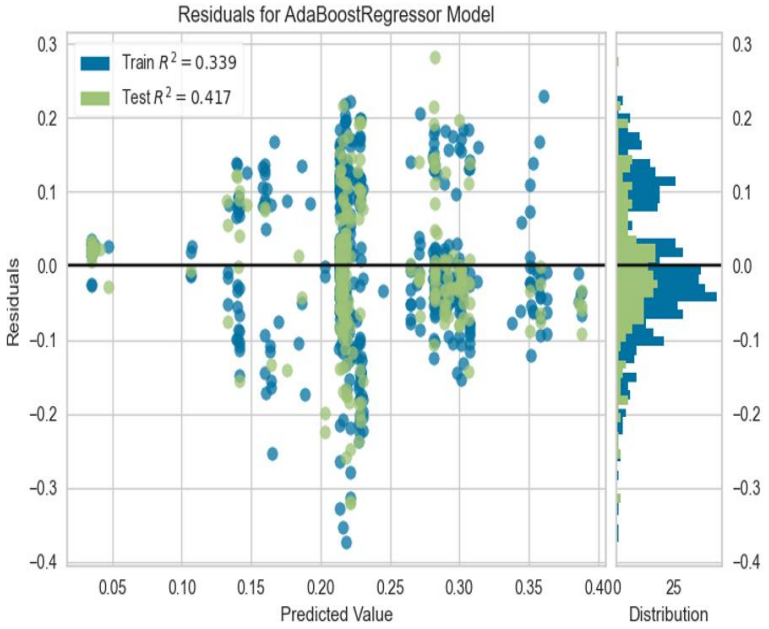

Figura 6. Resíduos

4.1.5 Distância de Cook

É a medida que indica a influência de ponto sobre o modelo executado. Os casos que possuem uma alta influência pode ser considerados outliers. Na figura 7 os pontos em que a distância de cook ultrapassam a linha tracejada em vermelho podem ser considerados outliers. A imputação ou remoção desses pontos podem trazer ganhos para a previsão.

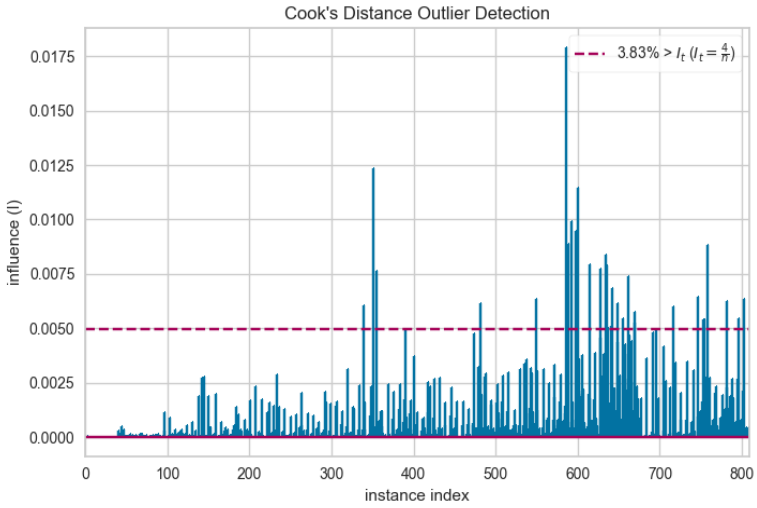

Figura 7. Distância de cook

4.1.6 Curva de aprendizado

O gráfico de curva de aprendizado mostra a relação de treinamento x pontuação do teste de validação cruzada. Através dele é possível saber se o modelo se beneficiaria de mais dados. É possível observar na figura 8, que a pontuação de treinamento é maior que a pontuação de validação e que as duas curvas estão no sentido de convergência, portanto podemos concluir que o modelo provavelmente requer mais exemplos de treinamento para generalizar o problema de maneira competente.

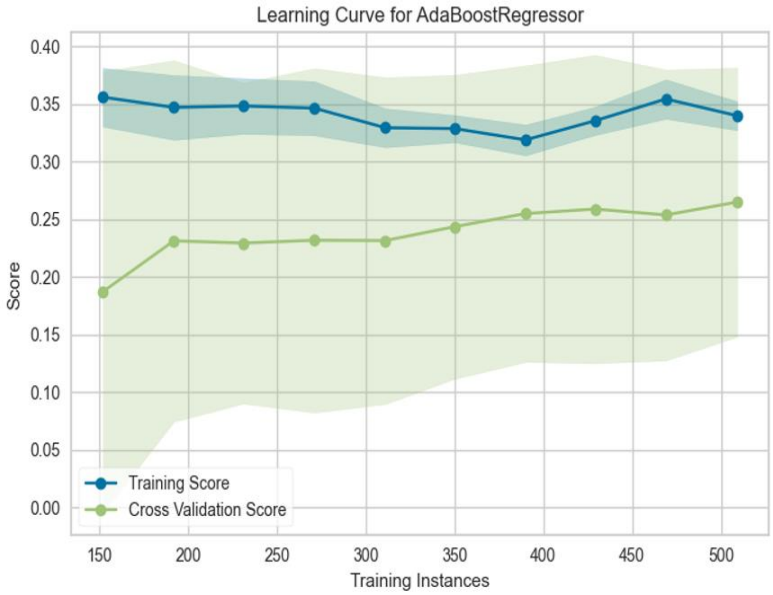

Figura 8. Curva de aprendizado

4.1.7 Real x Predito

O gráfico da figura 9 demonstra a relação da curva real x curva prevista pelo algoritmo AdaBoost para a base de testes. É possível perceber que o algoritmo não soube lidar com altas inesperadas.

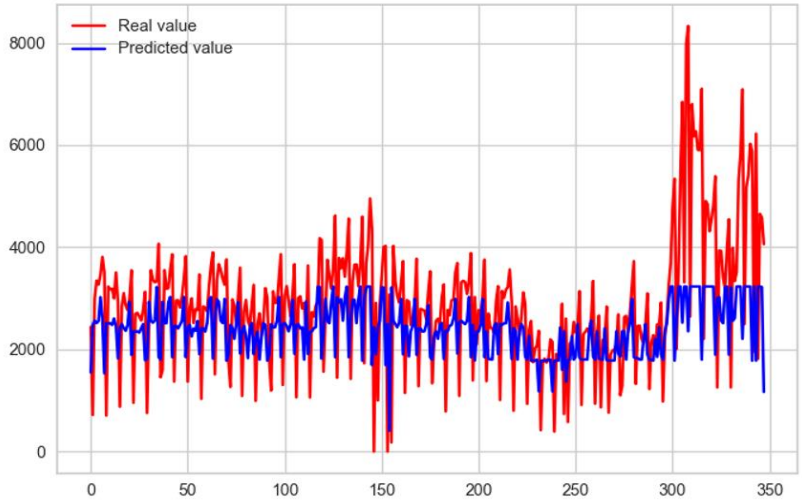

Figura 9. Real x predito

4.1.8 RMSE

Após execução do modelo na base de testes, foi obtido um RMSE de 0.1595 o qual será utilizado em testes de hipóteses com os RMSE da tabela 3 obtidos pelos modelos utilizados por [19] de modo a validar se o modelo obtido através de um AutoML pode ser considerado viável para utilização.

Modelo	RMSE
LSTM	0.99
Arima	0.104
LR	0.131
SVR	0.139
RCV	0.1373
Ensemble	0.1362

Tabela 3. Resultados obtidos em [19]

4.1.9 Teste de Hipóteses

É um procedimento estatístico que permite tomar uma decisão (aceitar ou rejeitar a hipótese nula H0) entre duas ou mais hipóteses, utilizando os dados observados de um determinado experimento. Para o nosso estudo foi utilizada como hipótese nula a ideia de que não existe diferença entre os modelos comparados. Caso a hipótese nula seja rejeitada, é verificado se o modelo AdaBoost regressor apresenta uma melhora com significância de 5% em relação ao modelo que está sendo comprado. Os resultados obtidos presentes na tabela 4 mostram que o modelo obtido através do AutoML demonstrou um melhor desempenho que o LSTM. Em comparação com os demais modelos obteve-se uma evidência estatística de 5% que houve uma piora. Isso não significa que o modelo obtido pelo através do AutoML não possa ser utilizado.

Comparação	**Resultado**
AdaBoost vs LSTM	AdaBoost
AdaBoost vs ARIMA	ARIMA
AdaBoost vs LR	LR
AdaBoost vs SVR	SVR
AdaBoost vs RCV	RCV
AdaBoost vs Ensemble	Ensemble

Tabela 4. Resultados do teste de hipóteses

4.2 Discussões

Apesar do resultado obtido pelo modelo AdaBoost Regressor executado por meio do AutoML ter superado apenas o LSTM de acordo com o teste de hipóteses, podemos considerá-lo robusto para o caso em análise. Além disso, a distância entre a métrica RMSE do AdaBoost Regressor para os demais modelos executados em [19] não é muito grande, evidenciando o bom resultado obtido pelo Aprendizado de Máquina Automatizado. É importante também, levar em consideração que no AutoML, o usuário apenas disponibiliza a base de dados e o sistema gera os

resultados automaticamente sem necessidade de conhecimento técnico na área de mineração de dado

Capítulo 5

Conclusões e Trabalhos Futuros

Neste capítulo é abordado a conclusão dos resultados e do trabalho como um todo, além de explicitar alguns dos possíveis trabalhos futuros.

5.1 Conclusões

Com a disputa entre as empresas cada vez maior, a realização de análises de dados é de extrema importância, visto que impacta a eficiência dos processos internos e externos das empresas. Uma boa previsão, aliada a realização de uma análise eficiente dos dados, pode auxiliar diretamente nas tomadas de decisões e direcionamento das empresas. Porém esse processo necessita de pessoas com expertises nas diretamente ligadas a mineração de dados para obtenção dos modelos que mais se adequam a realidade do seu negócio. Desta forma, o presente trabalho apresentou uma análise dos resultados obtidos por meio do Aprendizado de Máquina Automatizado considerando-os robustos e capazes de colaborar para tomada de decisões mais eficientes, além de gerar direcionamentos para empresas ou pessoas com base no problema que está sendo analisado. Se comparados com resultados obtidos por modelos que passaram por todas as fases do CRISP-DM. Concluindo assim a validação dos resultados obtidos pelo AutoML em busca do objetivo de democratizar o estudo de mineração de dados.

5.2 Trabalhos Futuros

Em busca do aprimoramento da análise de regressão de dados por meio do AutoML, pode-se realizar a implementação de algoritmos de séries temporais de modo a trazer outros algoritmos de regressão de dados como opção, a fim de

aumentar o leque de modelos disponíveis para novas análises em diversos tipos de aplicações.

Referências

[1] Aumenta o número de empresas abertas no país. Publicado em 02/02/2021 - 14:30 Por Luciano Nascimento - Repórter da Agência Brasil - Brasília.

[2] Huang, T. C. -K., Liu, C. -C. and Chang, D. C. An empirical investigation of factors influencing the adoption of data mining tools. In: International Journal of Information Management, pages 257-270, 2012.

[3] Côrte-Real, Nadine & Oliveira, Tiago & Ruivo, Pedro, 2017. "Assessing business value of Big Data Analytics in European firms," Journal of Business Research, Elsevier, vol. 70(C), pages 379-390.

[4] Ribeiro, A., Seruca, I., & Durão, N. (2016). Sales prediction for a pharmaceutical distribution company: A data mining based approach. Em 2016 11th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1–7).

[5] KUBINA, Milan; VARMUS, Michal; KUBINOVA, Irena. Use of big data for competitive advantage of company. Procedia Economics and Finance, v. 26, p. 561-565, 2015.

[6] YU, Xiaodan; QI, Zhiquan; ZHAO, Yuanmeng. Support vector regression for newspaper/magazine sales forecasting. Procedia Computer Science, v. 17, p. 1055-1062, 2013.

[7] Nunnari, G., & Nunnari, V. (2017). Forecasting Monthly Sales Retail Time Series: A Case Study. Em 2017 IEEE 19th Conference on Business Informatics (CBI) (Vol. 01, pp. 1–6).

[8] ELSHAWI, R.; MAHER, M.; SAKR, S. Automated Machine Learning: State-of-The-Art and Open Challenges. 2019.

[9] SHEARER, Colin. The CRISP-DM model: the new blueprint for data mining. Journal of data warehousing, v. 5, n. 4, p. 13-22, 2000.

[10] Armstrong, J. Scott. "Sales forecasting." Available at SSRN 1164602 (2008).

[11] THORNTON, Chris et al. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In: Proceedings of the 19th ACM

SIGKDD international conference on Knowledge discovery and data mining. 2013. p. 847-855.

[12] HUTTER, F. et al. Automatic machine learning (autoML). In: ICML 2015 Workshop on Resource-Efficient Machine Learning, 32nd International Conference on Machine Learning. 2015.

[13] Yoav Freund, Robert E. Schapire. "A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting", 1995

[14] HOERL, Arthur E.; KENNARD, Robert W. Ridge regression: applications to nonorthogonal problems. Technometrics, v. 12, n. 1, p. 69-82, 1970.

[15] JOLLIFFE, Ian T. Principal components in regression analysis. In: Principal component analysis. Springer, New York, NY, 1986. p. 129-155.

[16] KE, Guolin et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, v. 30, p. 3146-3154, 2017.

[17] GALTON, Francis. Hereditary talent and character. Macmillan's magazine, v. 12, n. 157-166, p. 318-327, 1865.

[18] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[19] Dallegrave, T., Silva, M., Neto, D., Junior, P., Filho, J. E., & Santos, W. (2021). Sales Forecast Optimization: Ensemble and Time Series Comparison. Revista De Engenharia E Pesquisa Aplicada, 6(5), 110-119. https://doi.org/10.25286/repa.v6i5.2153

[20] DA SILVA, Raniel Gomes. Desenvolvimento de uma Solução de Aprendizado de Máquina Automatizado Integrável a Múltiplos Ambientes Virtuais de Aprendizagem. Universidade de Pernambuco (UPE), 2020.

[21] Camilo, Cássio Oliveira, and João Carlos da Silva. "Mineração de dados: Conceitos, tarefas, métodos e ferramentas." Universidade Federal de Goiás (UFC) 1.1 (2009): 1-29

[22] PYTHON SOFTWARE FOUNDATION. Python Language Site: Documentation, 2020. Página de documentação. Disponível em: <https://www.python.org/doc/>. Acesso em: 05 de dez. de 2021.

[23] M. Ali, Pycaret: An open source, low-code machine learning library in python, PyCaret version 2.3.1, Apr. 2020. [Online]. Available: https://www.pycaret. org.

[24] GACKENHEIMER, Cory; PAUL, Akshat. Introduction to React. Apress, 2015.