**Relevância de Indicadores no Varejo e Predição de Vendas**

João Pedro Tonello de Souza¹;Juliano Domingues da Silva2

1 Zurich Brasil Seguros S.A.. Cientista de Dados II. Av. Jornalista Roberto Marinho, 85 – 20º andar – Brookling Novo; 04576-010 São Paulo, SP, Brasil

2 Universidade Estadual de Maringá. Professor Adjunto do Departamento de Administração. Av. Colombo, 5790 - Zona 7, 87020-900, Maringá - PR, Brasil.

\*autor correspondente: tonello.jpedro@gmail.com

**Relevância de Indicadores no Varejo e Predição de Vendas**

**Resumo**

Este estudo teve como objetivo avaliar estatisticamente a significância de variáveis do varejo e como elas impactam as vendas nas lojas da Walmart. Foram avaliados os indicadores de preço do combustível, índice de preços no consumidor (CPI), presença de feriados, desemprego e temperatura. Por fim, o estudo também propôs um modelo preditivo capaz de prever as vendas semanais com base nos indicadores relevantes, comparando o desempenho deste a um modelo *naive* baseado apenas em valores médios de cada loja. Foi demonstrada a significância dos indicadores apresentados e o modelo preditivo que leva em consideração todas as variáveis foi obteve um desempenho superior ao modelo *naive*, que previa apenas com base na média de valores anteriores para dada loja. Os resultados foram comparados com conclusões de pesquisas anteriores acerca do impacto individual de cada indicador nas vendas varejistas.

**Palavras-chave:** Machine Learning; Regressão; Estatística; Análise.

**Introdução**

Embora o impacto de indicadores do varejo já tenha sido estudado anteriormente, a maioria das pesquisas foca em variáveis individuais e em modelos clássicos de regressão linear e não linear (Badorf et al., 2020; Oh et al., 2009).

É conhecido através de pesquisas anteriores que há correlações não lineares entre as condições climáticas e as vendas em lojas de materiais de construção, sabendo-se que é possível aumentar a capacidade preditiva de um modelo adicionando detalhes climáticos, como temperatura, precipitação e incidência solar. Ao adicionar indicadores de condições climáticas no modelo, aumenta-se a variância explicada entre 4,1% e 9,2%, dependendo do nível de detalhe das informações adicionadas (Badorf et al., 2020).

Assim como a temperatura, é conhecido que feriados impactam na eficácia de promoções de preços e aumenta o volume de vendas offline, o preço da gasolina tem impacto significativo nas vendas de produtos alimentícios e o desemprego e o índice de preços ao consumidor afetam diretamente as vendas, onde a queda do índice implica no crescimento das vendas (Oh et al., 2009; Gicheva et al., 2010; Anagboso et al., 2009).

Todos os indicadores supracitados possuem relações de interdependência com o montante de vendas na indústria varejista quando analisados isoladamente e possuem capacidade preditiva sobre o valor faturado, indicando a possibilidade da predição através de modelos de *machine learning*.

Na década de 1950, surgia o campo de *machine learning* como uma ramificação da inteligência artificial. Ao longo das últimas décadas, diversos desenvolvimentos foram realizados com o objetivo de prever a demanda. Para tal, muitas vezes é necessária a utilização de um grande volume de dados históricos gerados por empresas e consumidores. Com um crescente aumento no volume total de dados gerados, projeta-se que o *machine learning* tem um potencial de causar um impacto profundo no varejo. (Wang et al., 2021).

O foco da inteligência artificial é um “agente inteligente” ou máquinas capazes de perceber o ambiente e realizar tarefas que a inteligência humana pode exercer (Poole et al., 1998). A IA se ramifica em campos da visão computacional intimamente relacionados, como o processamento de linguagem natural, reconhecimento de fala e *machine learning*.

*Machine learning* consiste em métodos ou algoritmos projetados para aprender padrões subjacentes nos dados e realizar predições baseadas nesses padrões. Especificamente, o aprendizado de máquina ocorre em uma experiência de treinamento a respeito de uma tarefa. Além disso, o desempenho do modelo pode ser estabelecido através de métricas.

Os algoritmos de *machine learning* estão se tornando importantes para o varejo. Uma das empresas que obtiveram sucesso na adoção da inteligência artificial foi a Walmart, que usa *machine learning* para agrupar produtos similares de diferentes comerciantes com base em características do produto, imagens e descrições. Analisando mais de 35 milhões de produtos, o modelo preditivo do Walmart demonstrou uma taxa de erro abaixo de 0,01% (Wang et al., 2021). Na IKEA, entre 10% e 15% dos itens devolvidos foram descartados como resíduos, mas a utilização de *machine learning* reduziu esse desperdício ao prever onde os itens devolvidos deveriam ser reabastecidos (Council, 2020).

Um dos desafios a serem enfrentados ao projetar um modelo de *machine learning* para prever a demanda do varejo é a variação nas vendas devido a fatores externos. O volume de vendas pode ser afetado por fatores sazonais, fatores climáticos, preço do combustível, feriados, fatores macroeconômicos, entre outros fatores (Anagboso e CmLaren, 2009).

**Material e Métodos**

Para o estudo, será utilizada uma base de dados semanais de 45 lojas da Walmart de 2010 a 2012, com as informações necessárias para cada dia, como a quantidade de vendas e as variáveis de estudo. A base de dados é disponibilizada para acesso livre na plataforma Kaggle (<https://www.kaggle.com/datasets/-rutuspatel/retail-analysis-with-walmart-sales-data/code>, 2021). Na Tabela 1 observa-se as variáveis da base de dados e a forma como os dados são apresentados.

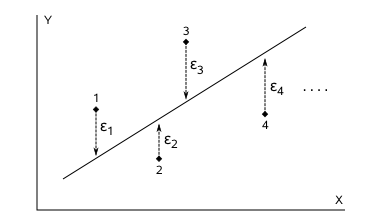
**Tabela 1. Variáveis do modelo**

|  |  |
| --- | --- |
| **Variável** | **Dados Apresentados** |
| Loja | Código numérico variando entre 1 e 45, sendo cada número representante de uma loja da Walmart nos Estados Unidos da América. |
| Data | Data representando o início de cada semana, desde fevereiro de 2010 até dezembro de 2012. |
| Vendas semanais | Quantidade de vendas para a respectiva semana em cada loja. Nos dados analisados, as vendas semanais variam entre 210 mil e 3,8 milhões de dólares americanos. |
| Feriado | Variável binária, indicando 1 para semanas com feriado e 0 para as demais semanas. |
| Temperatura | Temperatura no dia da venda, variando entre -2,06 e 100ºF. |
| Preço do combustível | Preço do combustível em na região, variando entre 2,47 e 4,47 dólares por litro. |
| CPI | Índice de Preço ao Consumidor. É uma métrica utilizada para observar tendências de inflação. Indica o preço médio necessário para comprar um conjunto de bens e serviços num país. No estudo, varia entre 126 a 227 dólares. |
| Desemprego | Taxa de desemprego na região, variando entre 3,88 e 14,3%. |

A partir dos dados fornecidos, serão utilizadas ferramentas de *machine learning* para detectar padrões no comportamento dos dados. Será utilizado um modelo de regressão linear para prever a quantidade mensal de vendas e avaliar a significância estatística de cada variável na composição do modelo.

Após a realização do modelo de regressão linear e de suas inferências estatísticas, outros modelos também serão testados para avaliar qual é a melhor método de prever as vendas semanais de uma loja.

Para uma análise supervisionada da variação das vendas no varejo, algumas técnicas podem ser empregadas de forma eficaz. Algoritmos como regressão linear, árvore de decisão e floresta aleatória são capazes de capturar uma parte da variância das vendas no varejo. Tais modelos, assim como qualquer modelo de *machine learning*, não explicam inteiramente a variação conjunta dos dados, porém, em uma pesquisa no ramo de inteligência artificial, testa-se diferentes modelos para obter a menor taxa de erro. A figura 1 ilustra a dimensão de erros em um modelo de regressão linear.



**Figura 1:** Dimensão dos erros em um modelo preditivo

Neste estudo, serão avaliados três modelos de regressão para a previsão de vendas:

1. **Regressão Linear Múltipla:** É a forma mais simples de elaborar um modelo de regressão. O modelo resulta em uma equação que define a combinação linear de variáveis preditoras. Seu resultado é o valor estimado da variável resposta. Além de ser utilizado como modelo matemático preditor, é possível realizar inferências estatísticas para estimar a significância de cada variável e se faz sentido utilizá-las em uma combinação linear para a realização da predição a dado nível de confiança (Su e yan, 2012).
2. **Árvore de Decisão:** Trata-se de uma ferramenta utilizada para a criação de sequências condicionais de decisão para o estabelecimento de valores de saída com base nas variáveis explicativas. Pode ser utilizada para predizer variáveis categóricas ou métricas. Possui uma fácil interpretação das decisões realizadas pelo algoritmo através de uma ilustração gráfica de um diagrama em forma de árvore (Myles et al.,2004).
3. **Floresta Aleatória:** Assim como a árvore de decisão, a floresta aleatória pode ser utilizada tanto para problemas de regressão quanto classificação, podendo ter como saída variáveis categóricas ou métricas. O método utiliza diversas árvores de decisão e combina suas saídas para reduzir a variância. Pode ser capaz de aumentar o desempenho de um modelo em relação às árvores de decisão, mas passam a ser menos interpretáveis do que o modelo anterior (Cutler et al., 2012).
4. **Gradient Boosting:** Os modelos de *gradient boosting* constroem-se através de outros modelos supervisionados preditivos, geralmente baseados em árvores. O objetivo dos modelos *gradient boosting* é obter um modelo F para predizer valores na forma minimizando o erro quadrático médio .

Para um modelo em um *gradient boosting* com M estágios, é adicionado um estimador , que representa a diferença da predição atual com a anterior. Cada etapa m+1 onde m<M, aproxima-se de corrigir os erros do modelo antecessor , utilizando como parâmetro o gradiente negativo da função de perda do erro quadrático médio.

O *gradient boosting* é um modelo de descida do gradiente (Li, 2016).

Para a análise preditiva de vendas da Walmart, o modelo de *gradient boosting* escolhido é o LightGBM (*Light Gradient Boosting Machine*), um algoritmo livre de código aberto desenvolvido originalmente pela Microsoft e baseado em árvores de decisão, tendo seu desenvolvimento focado em desempenho e escalabilidade (KE, Guolin et al, 2017).

A partir dos modelos testados, foi realizada uma comparação dos modelos com base nas seguintes métricas:

* **Erro Quadrático Médio:** Métrica definida pelo valor médio da soma dos erros ao quadrado.
* **Coeficiente de Determinação:** Medida de ajuste de um modelo generalizado. Indica a porcentagem de variância que o modelo é capaz de capturar em suas variáveis explicativas para predizer o valor da variável dependente.

As métricas de desempenho neste estudo foram obtidas através do método *k-fold*, para a obtenção de uma validação cruzada de desempenho, sendo os dados treinados com 10 subdivisões da base de dados.

Para uma avaliação do modelo preditivo, além dos indicadores de desempenho, foram avaliados graficamente os erros de predição, heterocedasticidade, diferença de desempenho entre dados de treino e de teste (sobreajuste), curvas de aprendizado e validação e importância das variáveis analisadas.

**Resultados e Discussão**

Foi carregada a base de dados de vendas da Walmart, com 6435 observações de 45 lojas com as variáveis citadas anteriormente.

A tabela abaixo mostra alguns dados estatísticos das variáveis.

**Tabela 2. Estatísticas das variáveis explicativas**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Mínimo** | **1º quartil** | **Mediana** | **3º Quartil** | **Máximo** | **Média** | **Desvio padrão** |
| **Vendas semanais (USD)** | 209.986 | 553.350 | 960.746 | 1.420.159 | 3.818.686 | 1.046.965 | 564.366 |
| **Temperatura** | -2,06ºF | 47,46ºF | 62,67ºF | 74,94ºF | 100,14ºF | 60,66ºF | 18,4ºF |
| **Preço do combustível (USD)** | 2,47 | 2,93 | 3,45 | 3,74 | 4,47 | 3,36 | 0,46 |
| **CPI** | 126 | 132 | 183 | 213 | 227 | 172 | 39 |
| **Desemprego** | 3,9% | 6,9% | 7,9% | 8,6% | 14,3% | 8,0% | 1,8% |

A figura abaixo ilustra a distribuição das vendas semanais em USD.

**Gráfico, Histograma

Descrição gerada automaticamente**

**Figura 2:** Distribuição das vendas semanais em todas as lojas em USD

A tabela abaixo ilustra os valores médios das variáveis agrupados por loja.

**Tabela 3. Valores médios das variáveis agrupados por loja**

****

A figura 3 representa as vendas semanais por loja, onde cada ponto em amarelo representa vendas de uma semana e a linha em roxo representa a venda média semanal de cada loja. É possível verificar graficamente a diferença do faturamento entre as lojas.

**Gráfico, Histograma

Descrição gerada automaticamente**

**Figura 3:** Distribuição das vendas médias por loja

Com o objetivo de obter resultados preliminares sobre a relevância estatística de indicadores nas vendas do varejo, foi criado um modelo linear multinível com interceptos aleatórios, variando contextualmente para cada loja. As figuras 4 e 5 ilustram os valores interpretáveis de intercepto e inclinações de cada variável e a significância estatística do efeito aleatório de loja, respectivamente.

**Tabela

Descrição gerada automaticamente**

**Figura 4:** Sumarização do modelo linear gerado

**Texto

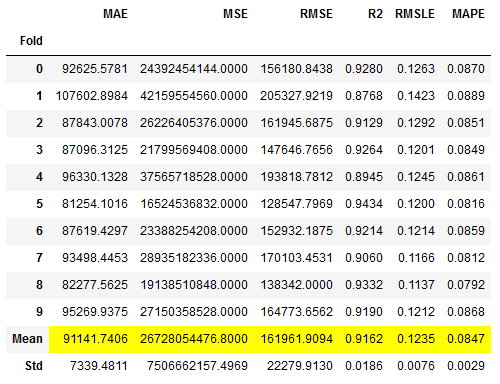
Descrição gerada automaticamente**

**Figura 5:** Significância estatística do efeito aleatório de loja

A figura 5 mostra que há uma significância estatística da variação do intercepto, indicando que existe uma diferença estatisticamente significante nas vendas de loja para loja.

Para todas as variáveis explicativas analisadas (desemprego, CPI, preço do combustível, temperatura e presença de feriado na semana), há uma significância estatística com mais de 95% de confiança, pois todos os valores p do teste estatístico ANOVA são menores que 0,05, indicando que para o nível de confiança estabelecido, afirma-se que a variável loja é estatisticamente relevante para a predição de vendas semanais.

Após a análise exploratória e avaliação estatística das variáveis preditoras, foi realizada uma análise através da ferramenta PyCaret para a obtenção das métricas de desempenho de um modelo de regressão linear que considera somente a variável loja dividida em variáveis *dummy* e observou-se as seguintes medidas:



**Figura 6:** Desempenho do modelo de regressão linear apenas com a variável loja

Nota-se que o erro quadrático médio (MSE) mensurado é de 26,7e+9 e . Após obter este desempenho, foram adicionadas todas as outras variáveis e suas métricas foram comparadas através da ferramenta PyCaret, utilizando os modelos de árvore de decisão, floresta aleatória e LightGBM. O desempenho médio de cada modelo validado pelo método *k-fold* estão ilustrados na figura abaixo.

Interface gráfica do usuário

Descrição gerada automaticamente com confiança baixa

**Figura 7:** Desempenho dos modelos com todas as variáveis

Observa-se que o MSE diminuiu para 17,5e+9 e subiu para 94,5%, sendo o LightGBM avaliado como melhor modelo para ambas as métricas de desempenho através da ferramenta de autoML. É importante ressaltar que o algoritmo avalia os modelos utilizando os hiperparâmetros padrões da biblioteca Scikit-learn.

Na figura 8, observa-se as configurações do modelo LightGBM.

Texto

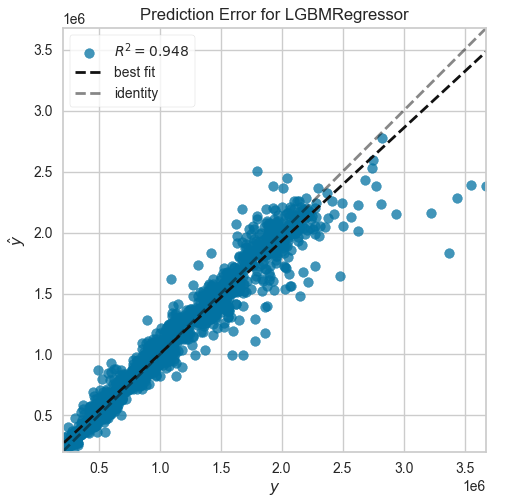
Descrição gerada automaticamente

**Figura 8:** Hiperparâmetros do modelo de regressão *Light Gradient Boosting Machine*

Através dos estudos mencionados anteriormente e do teste estatístico realizado com os dados da Walmart, já se conhecia a relevância dos indicadores. Adicionando-os ao modelo, nota-se que o houve um aumento da capacidade preditiva.

Com a relevância dos indicadores validada e os modelos preditivos criados, escolhendo-se o que possui as melhores métricas de desempenho, foram gerados os gráficos para a realização de análises gráficas do modelo.

As figuras 9 a 13 ilustram graficamente os erros de predição, avaliação de resíduos, curvas de aprendizagem, curvas de validação e importância das variáveis.



**Figura 9:** Erros de predição para o modelo escolhido

No gráfico ilustrado na figura 9, os valores idealmente se aproximam da linha tracejada em cinza claro (identidade). A distância entre o ponto e a linha através do eixo horizontal representa o erro do modelo para a predição da venda semanal. Para este modelo, verifica-se graficamente uma maior densidade de pontos próximos da linha identidade.

A figura abaixo ilustra a magnitude dos erros do modelo nas bases de treino e teste.

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

**Figura 10:** Resíduos do modelo escolhido

No gráfico ilustrado na figura 10, a nuvem de pontos verdes idealmente é semelhante à de pontos azuis para uma menor presença de sobreajuste (*overfitting*). Observa-se visualmente uma maior densidade de resíduos próximos de 0 através do histograma apresentado à direita do gráfico. A semelhança entre os histogramas verde e azul reforça a confirmação de uma baixa presença de *overfitting* no modelo.

Para uma análise mais aprofundada do desempenho, foi verificada também a diferença entre as capacidades preditivas nos dados de treino e nos de teste com a evolução das instâncias de treinamento. A figura 11 ilustra as curvas de aprendizado.

Gráfico

Descrição gerada automaticamente

**Figura 11:** Curvas de aprendizado do modelo

Idealmente, as curvas se aproximam com a evolução das instâncias de treinamento. Para o modelo analisado, verifica-se uma diferença de 0,021 no valor de R². Os valores exatos (0,948 e 0,969) podem ser visualizados na figura 10, mencionada anteriormente.

A figura 12 apresenta uma avaliação análoga através das curvas de validação, representando os valores de R² conforme o aumento da profundidade máxima das árvores de decisão geradas através do algoritmo LightGBM.

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

**Figura 12:** Curvas de validação do modelo

Nas figuras 11 e 12, os indicadores de desempenho nos dados de treino idealmente se aproximam aos indicadores pelo método da validação cruzada.

Ao aumentar a profundidade das árvores do modelo, a curva de validação nos dados de teste acompanha os dados de treino, mantendo a diferença máxima no coeficiente de determinação de 0,021.

Por fim, observa-se o gráfico de importância das variáveis na figura 13.

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

**Figura 13:** Importância das variáveis para predição de vendas

Observa-se que a variável loja é a preditora mais importante para predizer o montante semanal de vendas das lojas Walmart nos Estados Unidos. Em ordem de importância, as variáveis preditoras entre as analisadas são CPI, preço do combustível, temperatura, desemprego e presença de feriado, sendo a última, a variável com menor capacidade preditiva.

**Conclusões**

Através da análise exploratória e do ajuste dos modelos preditivos, foi possível concluir a um nível de confiança de 95% que todas as variáveis analisadas (índice de preços, preço do combustível, temperatura, desemprego e presença de feriado) afetam significativamente as vendas da Walmart.

No modelo que considera apenas a loja como variável preditora, o valor predito pelo modelo passa a ser o valor médio da loja. Obteve-se um R² de 91,6%, sendo este o percentual de variância que apenas a variável loja é capaz de explicar. No melhor modelo avaliado utilizando todas as variáveis, o valor de R² subiu para 94,5%. O valor do MSE diminuiu de 26,7e+9 para 17,5e+9, representando uma diminuição de 34,5% no valor do erro quadrático médio ao adicionar as outras variáveis.

**Referências**

BADORF, Florian; HOBERG, Kai. The impact of daily weather on retail sales: An empirical study in brick-and-mortar stores. **Journal of Retailing and Consumer Services**, v. 52, p. 101921, 2020.

OH, Hyunjoo; KWON, Kyoung‐Nan. An exploratory study of sales promotions for multichannel holiday shopping. **International Journal of Retail & Distribution Management**, 2009.

GICHEVA, Dora; HASTINGS, Justine; VILLAS-BOAS, Sofia. Investigating income effects in scanner data: do gasoline prices affect grocery purchases?. **American Economic Review**, v. 100, n. 2, p. 480-84, 2010.

ANAGBOSO, Mavis; MCLAREN, Craig. The impact of the recession on retail sales volumes. **Economic & Labour Market Review**, v. 3, n. 8, p. 22-28, 2009.

Council, Jared. IKEA Seeks to Cut Costs, Waste With Smarter Approach to Customer Returns. **Wall Street Journal Pro**, 2020.

Cutler, Adele; Cutler, D. Richard; Stevens, John R. Random forests. In: **Ensemble machine learning**. Springer, Boston, MA, 2012. p. 157-175.

Gicheva, Dora; Hastings, Justine; Villas-boas, Sofia. Investigating income effects in scanner data: do gasoline prices affect grocery purchases?. **American Economic Review**, v. 100, n. 2, p. 480-84, 2010.

Kaggle. **Retail analysis with Walmart sales data**. 2021. Disponível em <https://www.kaggle.com/datasets/rutuspatel/retail-analysis-with-walmart-sales-data?datasetId=1503066>. Acesso em 19 abr. 2022.

Myles, Anthony J. et al. An introduction to decision tree modeling. **Journal of Chemometrics: A Journal of the Chemometrics Society**, v. 18, n. 6, p. 275-285, 2004.

Poole, David, Alan Mackworth, and Randy Goebel. **Computational Intelligence**, 1998.

Schwaighofer, Anton et al. How wrong can we get? A review of machine learning approaches and error bars. **Combinatorial chemistry & high throughput screening**, v. 12, n. 5, p. 453-468, 2009.

Su, Xiaogang; Yan, Xin; Tsai, Chih‐Ling. Linear regression. **Wiley Interdisciplinary Reviews: Computational Statistics**, v. 4, n. 3, p. 275-294, 2012.

Wang, Xin Shane et al. The role of machine learning analytics and metrics in retailing research. **Journal of Retailing**, v. 97, n. 4, p. 658-675, 2021.

KE, Guolin et al. Lightgbm: A highly efficient gradient boosting decision tree. **Advances in neural information processing systems**, v. 30, 2017.