UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE QUIXADÁ

CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

MAURÍCIO OLIVEIRA DOS SANTOS

ESTUDO COMPARATIVO DE TÉCNICAS DE APRENDIZAGEM DE MÁQUINA PARA PREDIÇÃO DE DEMANDA DE ESTOQUE

QUIXADÁ

2022

MAURÍCIO OLIVEIRA DOS SANTOS

ESTUDO COMPARATIVO DE TÉCNICAS DE APRENDIZAGEM DE MÁQUINA PARA PREDIÇÃO DE DEMANDA DE ESTOQUE

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Sistemas de Informação do Campus de Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Fábio Carlos Sousa Dias

QUIXADÁ

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S236e Santos, Maurício Oliveira dos.
Estudo comparativo de técnicas de aprendizagem de máquina para predição de demanda de estoque / Maurício Oliveira dos Santos. – 2022.
35 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Sistemas de Informação, Quixadá, 2022.
Orientação: Prof. Dr. Fábio Carlos Sousa Dias.

1. Previsão de demanda. 2. Aprendizagem profunda. 3. Algoritmos computacionais. I. Título.
CDD 005

MAURÍCIO OLIVEIRA DOS SANTOS

ESTUDO COMPARATIVO DE TÉCNICAS DE APRENDIZAGEM DE MÁQUINA PARA PREDIÇÃO DE DEMANDA DE ESTOQUE

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Sistemas de Informação do Campus de Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Sistemas de Informação.

Aprovada em: 10 de dezembro de 2022

BANCA EXAMINADORA

Prof. Dr. Fábio Carlos Sousa Dias (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Me. Francisco Erivelton Fernandes de Aragão
Universidade Federal do Ceará - UFC

Prof. Dr. Wladimir Araujo Tavares
Universidade Federal do Ceará - UFC

Prof. Dr. Alberto Sampaio Lima
Universidade Federal do Ceará - UFC

À Deus, à minha mãe e a minha família.

AGRADECIMENTOS

À minha mãe Raimunda Oliveira, por sempre incentivar meus estudos, por sempre me apoiar em todos os momentos que passamos durante minha tragetória acadêmica, por sempre me ajudar a ir atrás de meus objetivos.

Ao meu pai Guilherme e meu irmão Marcelo, “In Memorian”, que sempre tiveram cuidado e zelo por mim.

À minha irmã Mônica que sempre me apoiou, ajudou, incentivou, cobrou durante minha tragetória acadêmica.

Aos demais membros da minha família, pelo apoio direto e indireto durante o período da graduação.

Agradeço a todos meus amigos que fizeram parte da minha formação, em especial Adson, João Paulo, Luis Siqueira, que apoiaram nos momentos que mais precisei.

À minha namorada Nágila Kaline, por sempre me apoiar e ajudar, estar do meu lado nos momentos fáceis e difíceis, por sempre me cobrar a terminar a graduação.

Ao Prof. Dr. Fábio Carlos Sousa Dias meu orientador, pela oportunidade e apoio na elaboração deste trabalho.

Aos Professores Me. Francisco E. F. de Aragão, Dr. Wladimir Araujo, Dr. Alberto Sampaio, pelo seu tempo dedicado a estarem presentes na banca examinadora e também pelos conhecimentos passados através de suas aulas, durante minha graduação.

Agradeço aos demais professores e também aos servidores da UFC - Quixadá, por todo o conhecimento transmitido e por tantas experiências boas proporcionadas. Todos os aprendizados nesse período foram de extrema importância para minha formação, tanto acadêmica como profissional e também pessoal.

"O sonho é que leva a gente para frente. Se a gente for seguir a razão, fica aquietado, acomodado."

(Ariano Suassuna)

RESUMO

O uso de aprendizado de máquina vem crescendo à cada dia nas mais diversas áreas do cotidiando. Dentre elas, o controle de estoque de produtos de alta perecibilidade é uma área sensível para empresas. Com isso essas empresas recorrem a técnicas para conseguir realizar a predição das demandas dos produtos, afim de reduzir perdas e fazer um melhor controle de seu estoque. E atualmente o aprendizado de máquina é uma maneira que as empresas utilizam para realizar essa predição. Este trabalho busca realizar um estudo comparativo técnicas de predição na área de aprendizado máquina, voltado a construção de modelos preditivos utilizando os algoritmos Random Forest, Xgboost e Catboost. Para pode predizer a demanda dos produtos e realizar uma comparação entre os resultados obtidos com a execução dos modelos..

Palavras-chave: Previsão de Demanda; Aprendizado de máquina; Técnicas de predição;

ABSTRACT

The use of machine learning is growing every day in the most diverse areas of everyday life. Among them are companies with stock of highly perishable products. With this, these companies resort to techniques to be able to forecast product demands, in order to reduce losses and make better control of their stock. And today machine learning is one way that companies use to make that prediction. This work seeks to use prediction techniques to predict the demand for products, performing a comparative study between the approached techniques.

Keywords: Demand Forecast; Machine learning; Prediction techniques;

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Em busca da redução de custos e prevenção de perdas, empresas do setor do varejo que comercializam produtos com alta perecibilidade, ou seja, produtos com curta data de vencimento, procuram ter um melhor controle do seu estoque. Com essa finalidade as empresas buscam métodos e formas de melhorar o controle de seu estoque e, umas das ferramentas que são usadas para esse fim é através da previsão da demanda dos produtos. A previsão de demanda permite que os administradores destas empresas antecipem o futuro de forma mais conveniente a suas ações (TUBINO, 2000).

A previsão de demanda de estoque de produtos com alta perecibilidade é complexa, pois ao mesmo tempo que visa evitar prejuízos com a perda dos produtos, deseja-se também que a disponibilidade dos produtos aos clientes seja afetada da menor forma possível. Enquanto de um lado a empresa diminui as perdas dos produtos, por outro lado ela pode ter perdas devido a falta de disponibilidade dos produtos aos clientes. Essa equação deve ser equilibrada de tal forma que resulte em um maior lucro para a empresa. Dias (2010) define demanda prevista como uma estimativa antecipada do volume de vendas num período determinado, com uma margem de erro determinada.

Dentre as várias estratégias de realizar a previsão da demanda de estoque, a aprendizagem de máquina tem ganho destaque, que a partir de um conjunto de dados devidamente tratado e de sofisticados algoritmos, permite em certos casos obter previsões precisas e com alto índice de acurácia. Mohri *et al.* (2012) definem aprendizagem de máquina como métodos computacionais, que utilizam experiências para melhorar o desempenho ou para fazer previsões precisas, onde estas experiências refere-se às informações fornecidas para o algoritmo aprendiz, que normalmente são dados coletados e disponíveis para análise.

O problema abordado neste trabalho consiste em evitar prejuízo para empresas que tenham seus produtos estragando nas prateleiras, por que já estão vencidos ou por que estão com a data de vencimento próxima de expirar, bem como com a frustração de clientes com a falta de produtos nas prateleiras das lojas. Atualmente, essa previsão de demanda é muitas vezes realizada de forma imprecisa pelos funcionários de entrega, a partir de uma fórmula que determina a demanda para uma próxima semana. Uma fórmula bastante utilizada é: D = V – R onde uma demanda D é determinada pela subtração das vendas V desta semana com o retorno de produtos R, onde esses produtos retornados são os produtos que se estragaram. Com isso o Grupo Bimbo disponibilizou um conjunto de dados relativos ao seu fluxo de vendas referente a

nove semanas de vendas na plataforma web Kaggle. Este conjunto de dados contém informações sobre as lojas e suas respectivas localizações, e ainda quais são os produtos e suas respectivas quantidades nas vendas.

Alguns trabalhos têm focado a previsão de demanda de estoque como conceito principal e alguns destes utilizam-se de conceitos de aprendizagem de máquina para fazer as previsões. Entre estes trabalhos estão o de Higucho (2006) que relata as técnicas de predição de demanda e um estudo de caso em três ambientes distintos. Já Delgado Filho (2020) e Roza (2016) utilizam conceitos de aprendizagem de máquina para propor um modelo de predição. Delgado Filho (2020) utiliza os conceitos de aprendizagem de máquina para construir modelos de predição para prever a demanda da produção semanal de uma distribuidora de bebidas. Roza (2016) utiliza os conceitos de aprendizagem de máquina para auxiliar o apoio à tomada de decisão em vendas do varejo.

Considerando o contexto atual, o presente trabalho tem como objetivo realizar um estudo comparativo de modelos de predição utilizando conceitos e técnicas de aprendizado de máquina para predizer a demanda de estoque dos produtos, e com isso determinar qual o melhor modelo construído utilizando os algoritmos Random Forest, XGBoost e Catboost, utilizando o conjunto de dados disponibilizado na competição de nome *Grupo Bimbo Inventory Demand* na plataforma web Kaggle.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta Seção, serão abordados os principais conceitos relacionados ao presente trabalho e qual a contribuição de cada conceito para o seu desenvolvimento .

2.1 Previsão de Demanda

Dias (2010) define previsão de demanda como uma estimativa antecipada do volume de vendas num período determinado, com uma margem de erro a ser considerada. Para o autor existem três tipos de demanda: demanda regular, que acontece quando a necessidade de materiais é constante ao longo do tempo ou tem oscilações, de tal forma que podemos identificar um comportamento regular ao longo do tempo; demanda crescente (decrescente), que ocorre quando identifica-se um crescimento (decrescimento) do consumo ao longo do tempo; demanda irregular, que ocorre quando há influência da sazonalidade.

Segundo Mancuzo (2003), a previsão de demanda pode ser utilizada com vários objetivos nas empresas. É uma estratégia crucial para o controle e planejamento do estoque. Quanto maior o erro na previsão da demanda, maior é a dificuldade da empresa planejar estratégias nas diversas áreas que atua, dificuldade esta, impõe perdas financeiras às empresas, ocasionando a redução de sua competividade com seus concorrentes (GERBER *et al.*, 2013).

2.2 Aprendizagem de Máquina

Aprendizagem de máquina pode ser definida como métodos computacionais, que utilizam experiência para melhorar o desempenho ou para fazer previsões precisas, onde esta experiência refere-se às informações fornecidas para o algoritmo aprendiz, que normalmente são dados coletados e disponibilizados para análise. Geralmente, as técnicas de aprendizagem de máquinas são métodos que combinam conceitos fundamentais da ciência da computação com ideias da estatística, da probabilidade e da otimização (MOHRI *et al.*, 2012)

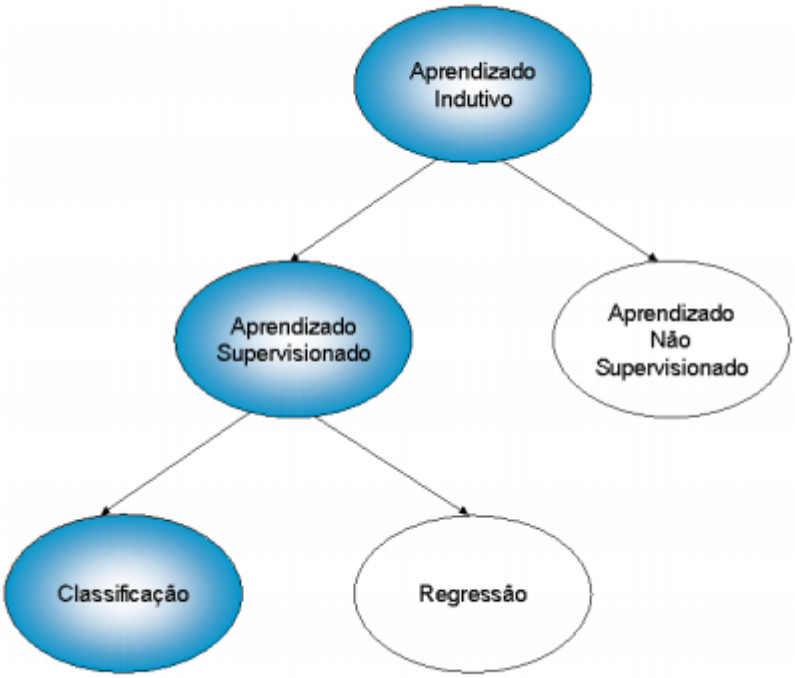
Para Shalev-Shwartz e Ben-David (2014), o termo aprendizagem de máquina refere-se à detecção automatizada de padrões nos dados. Ou seja, é uma área da computação que através dos padrões dos dados consegue obter resultados preditivos utilizando algoritmos que aprendem a partir dos dados. A aprendizagem de máquina é classificada em dois tipos: supervisionado e não supervisionada.

Segundo Monard e Baranauskas (2003), no aprendizado supervisionado é fornecido

ao algoritmo de aprendizado, um conjunto de dados onde o rótulo da classe associada é conhecida para o treinamento. Sendo que o objetivo do algoritmo é construir um classificador que possa determinar corretamente a classe de novos exemplos não rotulados. Onde esses rótulos podem ser de classificação ou regressão.

Shalev-Shwartz e Ben-David (2014) define que no aprendizado não supervisionado não existe uma distinção entre os dados de treinamento e teste. O algoritmo processa os dados de entrada com o objetivo de apresentar algum resumo ou versão compactada dos dados agrupando o conjunto de dados em subconjuntos de objetos semelhantes. Na Figura 1, mostra a hierarquia do aprendizado indutivo.

Figura 1 – Hierarquia do Aprendizado



Fonte: Shalev-Shwartz e Ben-David (2014)

O conceito de aprendizado de máquina no presente trabalho foi utilizado para construção dos modelos preditivos com foco no aprendizado máquina supervisionado voltados para regressão, pois a característica do problema proposto é uma predição de valores com base em dados temporais. Algoritmos de regressão são eficazes quando o objetivo é prever demanda (DELGADO FILHO, 2020).

2.3 Algoritmos de aprendizado de máquina

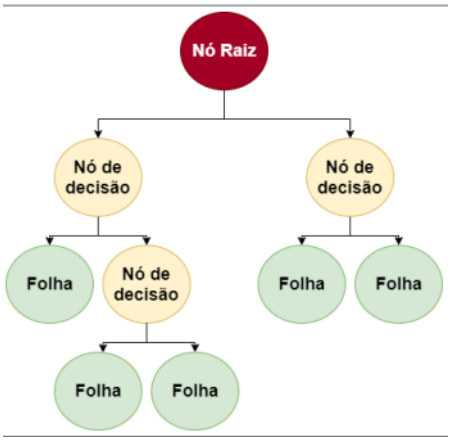
Na aprendizagem de máquina existe uma grande quantidade de algoritmos onde cada um tem seus pontos positivos e negativos. Nesta Seção serão abordados os algoritmos utilizados neste trabalho. A escolha dos algoritmos ocorreu pela popularidade da utilização deles para resolução do tipo do problema proposto neste trabalho. Os algoritmos utilizados foram: Random Forest, CatBoost Regressor, XGBoost Regressor.

2.3.1 Árvores de Decisão

Segundo Aurélien (2019) Árvores de Decisão são algoritmos versáteis de Aprendizado de Máquina que podem executar tarefas de classificação, regressão. São algoritmos muito poderosos capazes de moldar conjuntos complexos de dados.

> O algoritmo Árvore de Decisão é uma ferramenta simples e eficaz utilizando um esquema de decisão em vários estágios, podendo ser eles hierárquicos ou estruturas de árvore. A estrutura é composta por um conjunto de nós que simbolizam a tomada de decisão. Os nós recebem um term de acordo com seu posicionamento, podendo ser chamados de nó raiz, que contém todos os dados, nós internos ou nós de decisão, que são as divisões e os nós folhas, que são os nós terminais (DELGADO FILHO, 2020).

Figura 2 – Estrutura de uma Árvore de Decisão



Fonte: Delgado Filho (2020)

Apesar de ser uma ferramenta simples e eficaz, as Árvores de Decisão tem um problema principal que elas são sensíveis a variações dos dados de treinamento e também o modelo pode sofer um ajuste excessivo, prejudicando sua capacidade de generalização, conhecido

com *overfitting*. Para evitar o overfitting pode se utilizar um conjunto de validação ou definir restrições aos parâmetros do modelo (AURÉLIEN, 2019) (DELGADO FILHO, 2020).

2.3.2 Random Forest

Segundo Breiman (2001), Random forest ou Florestas aleatórias são uma combinação de árvores de decisão de tal forma que cada árvore depende dos valores de um vetor aleatório amostrado independentemente e com a mesma distribuição para todas as árvores da floresta. Esse processo de usar vários modelos, treinados sobre os mesmos dados, calculando a média dos resultados de cada modelo é conhecido como "Ensemble Learning". Da técnica Ensemble Learning as abordagens mais populares são: *bagging* e *boosting*. Para Aurélien (2019)

> O *bagging* treina modelos individuais de maneira paralela a fim de obter um conjunto diversificado de classificadores. Já o *boosting* combina o treinamento de uma sequência de modelos, onde cada modelo de forma individual aprende com os erros cometidos pelo modelo anterior.

Já Breiman (2001), propôs que o Random Forest uma modificação na técnica de Bagging para que fosse adicionada uma camada de aleatoriadade e também uma alteração na forma que as árvore de decisão são construídas. Segundo Delgado Filho (2020) nas Florestas Aleatórias os nós se separam escolhendo o melhor subconjunto dos preditores escolhidos aleatoriamente .

2.3.3 XGBoost

XGBoost é uma implementação eficiente e escalável do algoritmo de *gradient boosting* por (FRIEDMAN, 2001). Gradient Boosting refere-se a uma classe de algortimos de aprendizado de máquina que podem ser usados para problemas de classificação ou regressão. Os conjuntos são construídos a partir de árvores de decisão. As árvores são adicionadas uma de cada vez ao conjunto e ajustadas para corrigir os erros de previsão feitos pelos modelos anteriores. Este é um tipo de modelo de aprendizado de máquina de conjunto conhecido como boosting. Os modelos são ajustados usando qualquer diferença arbitrária. Os modelos são ajustados usando qualquer função de perda diferenciável arbitrária e algoritmo de otimização de gradiente descendente. Isso dá à técnica seu nome, “aumento de gradiente”, pois o gradiente de perda é minimizado à medida que o modelo é ajustado.

O XGBoost é bastante reconhecido e utilizados por suas características. De acordo com Chen *et al.* (2015) suas características são:

- Capacidade de realizar computação paralela em diferentes sistemas operacionais.

- Capacidade de poder receber vários tipos de dados de entradas: matriz densa, matriz esparsa, arquivo de dados.
- aceita entrada esparsa para booster de árvore e booster linear, é otimizado para entrada esparsa.
- Suporta funções personalizadas.
- Melhor desempenho em vários conjuntos de dados diferentes.
- Capacidade de transformar um aprendiz fraco em forte, por meio da etapa de otimização para cada árvore implementada no sistema.
- Validação cruzada como recurso interno.

2.3.4 CatBoost

Catboost é uma implementação de *gradient boosting*, que utiliza árvores de decisão binárias como preditores de base (PROKHORENKOVA *et al.*, 2018). O Catboost é um algoritmo que traz melhorias em sua implementação para evitar que ocorra o *overfitting* e também para lidar de uma melhor forma com o paralelismo. Catboost tem como característica utilizar árvores esquecidas como preditores básicos, nessas árvores, o mesmo critério de divisão é usado em todo nível da árvore. Essas árvores são equilibradas e menos propensas a overfitting. De acordo com Dorogush *et al.* (2018) as vantagens do catboost são:

- **Category features:** capacidade de utilizar todos conjunto dados para o treinamento.
- **Feature combinations:** capacidade de conter todos os recursos de combinação e classificação na árvore atual do conjunto de dados com todos os recursos categóricos.
- **Unbiased boosting:** Várias permutações dos dados de treinamento são empregadas para aumentar a robustez. E diferentes permutações serão utilizadas para treinar modelos distintos para lidar com overfitting
- **Fast scorer:** utiliza recursos binários armazenados em vetor contínua para calcular as previsões do modelo, como esses vetores podem ser construídos de maneira paralela proporciona um aumento de velocidade em até 3x.

2.4 Métricas para avaliação dos modelos

Na aprendizagem de máquina, para avaliar a qualidade do modelo, pode-se utilizar várias métricas: Acurácia, F1 Score, Precisão, Recall, Log Loss, MSE, RMSE, MAE. Este trabalho irá utilizar a métrica RMSE e MAE.

2.4.1 Root Mean Square Error (RMSE)

RMSE é a medida da magnitude média dos erros estimados. Tem valor sempre positivo e quanto mais próximo de zero, maior a qualidade dos valores medidos ou estimados, calculado pela Equação 2.1.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_i - O_i)^2} \tag{2.1}$$

Onde, E_i e O_i são os valores estimados e observados (medidos), respectivamente, e n é o número de observações (WILLMOTT; MATSUURA, 2005).

2.4.2 Mean Absolute Error (MAE)

Mean Absolute Error (MAE): erro médio absoluto, definido pela Equação 2.2 é a medida média do erro absoluto. Essa métrica mede a magnitude média dos erros em um conjunto de previsões, sem considerar sua direção.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(E_i - O_i)^2 \tag{2.2}$$

3 TRABALHOS RELACIONADOS

Este trabalho baseia-se em técnicas, informações, métodos, conceitos coletados em livros, artigos científicos na área de previsão de demanda de estoque e na área de aprendizagem de máquina. Os trabalhos de Higucho (2006), Delgado Filho (2020) e Roza (2016) são os principais contribuidores para o desenvolvimento deste trabalho.

Assim como nos trabalhos citados acima, o trabalho aqui proposto também utilizará, os conceitos e técnicas de previsão de demanda de estoque e de aprendizagem de máquina. Higucho (2006) faz um estudo de caso em três locais distintos utilizando técnicas de previsão de demanda de estoque. Delgado Filho (2020) utiliza conceitos e técnicas de aprendizagem de máquina para a realização de uma predição de demanda de uma produção de bebidas de uma distribuidora. Roza (2016) utiliza a aprendizagem de máquina como auxílio a tomada de decisões em vendas no varejo a partir de registros de vendas. Este trabalho por meio de técnicas de aprendizagem de máquina visa construir modelos de predição de demanda de estoque de produtos perecíveis.

3.1 A previsão de demanda de produtos perecíveis: três estudos de caso

Higucho (2006) defende que a previsão de demanda e administração de estoques possuem a tarefa de diminuir os custos e ao mesmo tempo manter o nível do serviço em patamares que não comprometam a imagem da organização. Com isso, ele propôs uma revisão literária sobre previsão de demanda e administração de estoques, tomando como base para análise dados de três organizações de ramos distintos, que compartilham algo em comum, todas trabalham com produtos alimentícios de alta perecibilidade. As motivações para este trabalho foi analisar os modelos de previsão de demanda e controle de estoques utilizadas pelos administradores das três organizações distintas que compartilham o mesmo aspecto em comum.

A coleta de dados foi realizada em julho de 2005, e os dados foram analisados e colocados em uma tabela comparativa, levando em consideração somente o modelo adotado e validade de suas previsões.

O resultado obtido através da análise dos dados foi que nas organizações A e B existe uma complementaridade entre os modelo quantitativos e qualitativos enquanto na organização B o escopo dos itens estudados obscureceu a interação entre os dois modelos.

3.2 Aprendizagem de máquina para apoio à tomada de decisão em vendas do varejo utilizando registros de vendas

Roza (2016) propôs em seu trabalho empregar técnicas de aprendizado de máquina para criar modelos que representam relações entre clientes e produtos nas vendas no varejo, pois constatou que a quantidade e qualidade dos dados disponíveis das lojas físicas do varejo para análise são inferiores em relação as lojas virtuais. Foi utilizado um conjunto de dados referentes a cadastro de clientes e registros de vendas de cerca de 100 lojas de uma rede de lojas do varejo de todo Brasil, do período dos anos de 2015 e meados de 2016. Foram utilizados para a criação dos modelos de classificação algoritmos de árvore de decisão e k-Neighbors, e para o agrupamento dos dados dos produtos foi apresentado um algoritmo genético como solução. Com o intuito de validar os resultados foram utilizadas técnicas de avaliação de modelos de classificação.

Sua execução foi iniciada com a coleta dos dados. Para implementação das técnicas de aprendizado de máquina foi utilizada a linguagem de programação Python, as bibliotecas Pandas no auxílio do tratamento dos dados e a Scikit-learn que auxilia na implementação dos algoritmos de aprendizagem de máquina. Foi possível perceber que a complexidade dos problemas a serem abordados e a qualidade dos resultados depende totalmente da base de dados.

3.3 Análise de Métodos de Regressão para Previsão de Demanda de Curto Prazo

Delgado Filho (2020) propôs em seu trabalho investigar a aplicabilidade de algoritmos de Apendizagem de Máquina para realizar uma previsão de demanda de curto prazo. Para este trabalho foram utilizados conjuntos de dados de uma grande empresa de distribuição de bebidas no mercado brasileiro. O objetivo do trabalho consiste em realizar a predição da produção semanal dessa empresa, através da análise de Séries Temporais. Utilizando algoritmos de regressão para atingir esse objetivo. Sendo eles: Linear Regression (LR), Support Vector Regression (SVR), Stochastic Gradient Descent (SGD), Decision Tree, Multilayer Perceptron (MLP), Random Forest (RF) e XGBoost.

Os autores também tiveram como objetivo realizar uma comparação dos resultados apresentados pelos modelos de regressão desenvolvidos, com modelos desenvolvidos por outras empresas. Também foi utilizado variáveis externas temporais, como dados econômicos e dados de condições climáticas, nos dados de entrada dos modelos com o objetivo de analisar qual influência esses dados externos podem impactar nas previsões.

Com isso foi possível concluir que, os métodos de regressão são uma opção eficiente para realização de previsão de séries temporais, alcançando resultados superiores na maioria dos casos, quando comparados aos já existentes.

Tabela 1 – Trabalhos Relacionados

Trabalhos	Aprendizado de Máquina	Previsão de Demanda
Higuchi (2006)	Não utilizou	Produtos Perecíveis
Filho (2020)	Regressão Linear, Árvore de Decisão, Random Forest, XGBoost	Distribuição de Bebidas
Roza (2016)	Árvore de decisão, k-Neighbors	Vendas Varejo
Este Trabalho	Árvore de Decisão, Random Forest, XGBoost, CatBoost	Produtos Perecíveis

Fonte: Elaborado pelo autor

4 PROCEDIMENTOS METODOLÓGICOS

Nesta seção serão apresentados todos os procedimentos necessários para execução do trabalho.

4.1 Realização da coleta e preparação dos dados

Os dados foram coletados do site Kaggle, especificamente da competição *Grupo Bimbo Invetory Demand*. Esses dados foram divididos em dados de treino e teste. Foi feita uma verificação dos dados para saber se eles estão em um formato utilizável. Esses dados possuem informações sobre as vendas, produtos, clientes.

4.2 Análise e pré-processamento dos dados

Após a coleta e análise dos dados , foi necessário realizar um trabalho de pré-processamento para retirada de algumas inconsistências dos dados, como dados duplicados e dados com valores faltantes. Também ocorreu nesta etapa a criação de novos atributos para utilização nos modelos.

4.3 Seleção das técnicas de aprendizagem de máquina

Como foi definido na proposta do trabalho, foram aplicadas técnicas de predição utilizando aprendizagem de máquina. Foram escolhidas as técnicas para comparação: *Random Forest, XGBoost, CatBoost*.

4.4 Construção dos modelos preditivos

Esta etapa consiste na criação dos modelos de predição. Os modelos foram construídos utilizando os atributos da base de dados de treino disponíveis.

4.5 Análise comparativa dos resultados

Os resultados obtidos na aplicação dos algoritmos sobre os dados disponibilizados pela competicação *Grupo Bimbo Inventory Demand* da plataforma online *Kaggle*, foram comparados e classificados os resultados utilizando as métricas de avaliação de desempenho *Root*

Mean Squared Error (RMSE) e *Mean Absolute Error* (MAE).

5 RESULTADOS

Nesta seção serão descritos todos os procedimentos executados neste trabalho. Para execução foi utilizada a biblioteca *Scikit-Learn* [1] , onde os mesmo foram executados em máquinas virtuais da plataforma *Google Colab* [2].

5.1 Realização da coleta e preparação dos dados

Os conjuntos de dados foram coletados do site da competição do Kaggle *Grupo Bimbo Inventory Demand* [3]. Esses conjuntos de dados estavam divididos em arquivos *Comma-Separated Values (CSV)*, onde nesses arquivos tem informações referentes as vendas semanais dos produtos, chegando a um total de nove semanas. As tabelas abaixo mostram a descrição do dos atributos encontrados nos dados coletados.

A Tabela 2 mostra a descrição dos atributos do arquivo cliente_tabla.csv. Referentes as informações dos clientes do Grupo Bimbo.

Tabela 2 – Informações sobres os clientes

Atributo	Tipo	Descrição
Cliente_ID	int64	id do cliente
NombreCliente	string	nome do cliente

Fonte: Elaborado pelo autor

A Tabela 3 mostra a descrição dos atributos do arquivo producto_tabla.csv. Com informações sobre os produtos vendidos pelo grupo Bimbo.

Tabela 3 – Informações sobres os produtos

Atributo	Tipo	Descrição
Producto_ID	int64	id do produto
NombreProducto	string	nome do produto

Fonte: Elaborado pelo autor

A Tabela 4 mostra a descrição dos atributos do arquivo town_state.csv. Com informações referentes a localização das lojas abastecidas pelo grupo Bimbo.

A Tabela 5 abaixo contém as informações referentes as vendas das nove semanas coletadas do arquivo train.csv.

[1] https://scikit-learn.org/stable/index.html
[2] https://colab.research.google.com/
[3] https://www.kaggle.com/competitions/grupo-bimbo-inventory-demand

Tabela 4 – Informações sobres as lojas

Atributo	Tipo	Descrição
Agencia_ID	int64	id do depósito de vendas
Town	string	endereco da loja
State	string	estado da loja

Fonte: Elaborado pelo autor

Tabela 5 – Informações sobres as vendas

Atributo	Tipo	Descrição
Semana	uint8	número da semana
Agencia_ID	uint16	id do depósito de vendas
Canal_ID	uint8	id do canal de vendas
Ruta_SAK	uint16	id da rota das vendas
Cliente_ID	uint32	id do produto
Producto_ID	uint16	id do produto
Venta_uni_hoy	uint32	venda em unidades
Dev_uni_proxima	uint32	quantidade em unidades proxima semana
Dev_proxima	float16	valores em pesos proxima semana
Demanda_uni_equil	uint32	valor a ser encontrado

Fonte: Elaborado pelo autor

5.2 Análise e pré-processamento dos dados

Após a coleta e entendimento dos dados referente as vendas, foi necessário realizar um pré-processamento para retirada de dados inconsistentes. Onde foi verificado a existência de dados faltantes NaN e dados duplicados nos conjuntos de dados. Esses dados fora do padrão são denomidados de *outliers*.

Posteriormente a eliminação dos *outliers*, foi criado um novo dataset com os dados referentes a semana 9 de vendas encontrados no arquivo train.csv e também com os dados disponibilizados no arquivo test.csv. A escolha pela nona semana de vendas se deu por restrição de hardware da máquina. Como o dataset original (train.csv) contém 7.41×10^7 linhas ficaria inviável a execução no ambiente de trabalho. A criação desse novo dataset foi dividido em 2 etapas.

A primeira etapa foi excluir as colunas que existem no conjunto de treino (train.csv) e que não estão no conjunto de teste (test.csv), que são as colunas *Venta_uni_hoy, Venta_hoy, Dev_uni_proxima, Dev_proxima*.

A segunda etapa foi a criação de novas features nos datasets train e test. Foi criado no dataset train uma nova coluna *id* e no dataset test foi criada a nova coluna *Demanda_uni_equil*. Foram criadas essas novas colunas porque o dataset train contém a coluna *Demanda_uni_equil* que não está presente no dataset test, e também porque a coluna *id* existe no dataset test e não

está presente no dataset train. Também foi criada uma nova coluna em ambos datasets afim de identificar quais são os dados de teste, a nova coluna tem o nome de *teste* e o valor 0 representerá o dataset train e o valor 1 representerá o dataset test.

Após a unificação dos dados no novo dataset chamado de *df_uni*, foram criadas novas features com dados de vendas do intervalo das semanas 4 a 8. Features essas que são a média da demanda ajustada de clientes e a quantidade de registros de clientes por produto. A média da demanda ajustada por produto e a quantidade de registros por produto e a média da demanda ajustada por cliente e a quantidade de registros por produto.

Os dados do novo dataset *df_uni*, foram divididos em conjuntos de treino e predição. Onde o conjunto de treino tem dados referentes ao dataset *train* e o conjunto predição tem dados referentes ao dataset *test*. O dataset de treino ficou com 1.04×10^7 linhas, esse dataset foi dividido em conjunto de treino e teste, onde o conjunto de treino ficou com aproximadamente 80 porcento do tamanho do dataset e com o conjunto de teste ficou com 20 porcento. O que resultou em 8.32×10^6 linhas no conjunto de treino e 2.08×10^6 linhas no conjunto de teste. Essa divisão foi realizada com ajuda da função train_test_split da biblioteca *Scikit-learn*.

5.3 Seleção dos parâmetros e aplicação dos modelos

Esta etapa consistiu na procura dos melhores parâmetros e posteriormente a criação dos modelos de predição já mencionados. Antes da criação de cada modelo buscamos realizar uma seleção dos melhores parâmetros para cada modelo. Para essa busca escolhemos o método *Grid Search*, que é uma das estratégias mais utilizadas para busca de hiper-parâmetros. *Grid Search* é um processo que pesquisa exaustivamente por meio de um subconjunto especificado manualmente do espaço de hiperparâmetros do algoritmo de destino (CHAN; TRELEAVEN, 2015). Para execução do *grid search* iremos utilizar a função *GridSearchCV* da biblioteca *Scikit-Learn*. Na Tabela 6 podem ser visualizados os parâmetros utilizados no *grid search*.

Tabela 6 – Parâmetros para execução do *grid search*

Parâmetro	Descrição
estimator	modelo para qual será estimado os parâmetros
param_grid	conjunto de parâmetros
cv	determina a estratégia de divisão de validação cruzada
n_jobs	número de jobs a serem executados em paralelo
verbose	controle do fluxo de mensagens de saída

Fonte: o autor.

Considerando os modelos escolhidos, buscamos com o *grid search* os melhores valores para parâmetros de execução de ambos. Os parâmetros escolhidos foram:

- n_estimators: número máximo de árvores na floresta. Para escolha do melhor valor do parâmentro para execução dos modelos foram utilizados os seguinte valores: 2, 10, 100, 200, 300, 800.
- max_depth: representa a profundidade de cada árvore na floresta. Foram utilizados em ambos modelos os seguintes valores: 2, 6, 12, 15, 20, 30, 55, 70. Para obtenção do melhor valor do parâmetro para os modelos.

A Tabela 7 apresenta os valores obtidos pelo *grid search* para cada parâmetro para ambos os modelos.

Tabela 7 – Valores obtidos para cada parâmetro após o *grid search*

Parâmetro	Algoritmo		
	RF	XGBoost	CatBoost
n_estimators	100	100	100
max_depth	12	12	6

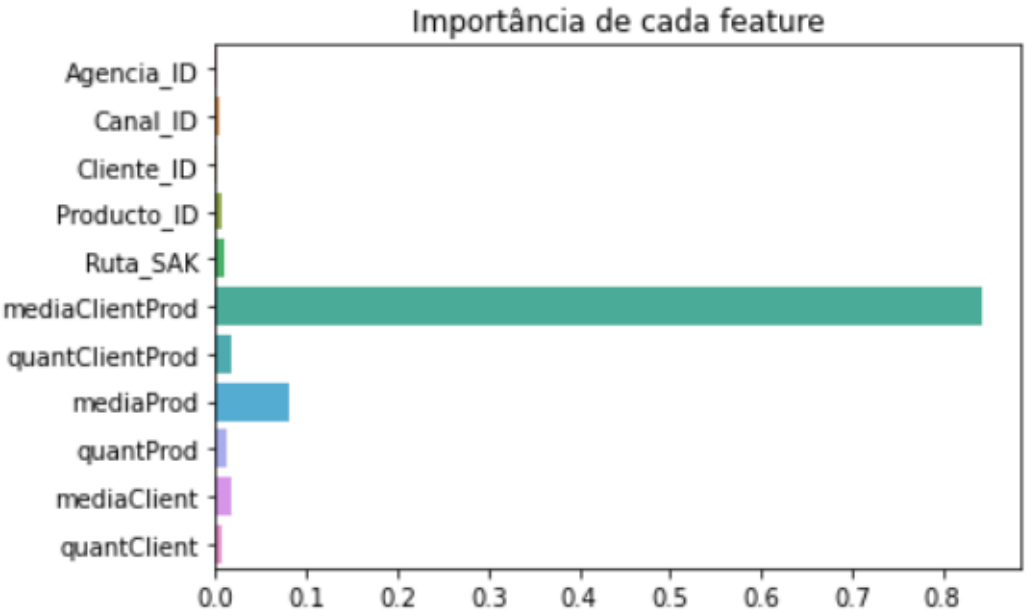
Fonte: o autor.

Observando o resultado obtido para os parâmetros pelo *grid search* para os dados em análise, percebe-se que cada modelo necessita de valores diferentes para sua melhor execução.

Após a criação dos modelos utilizando as mesmas *features* e seus respectivos parâmetros. Foi analisada a importância de cada feature para os modelo. As figuras Figura 3, Figura 4, Figura 5 a seguir mostram a importância das features para cada modelo.

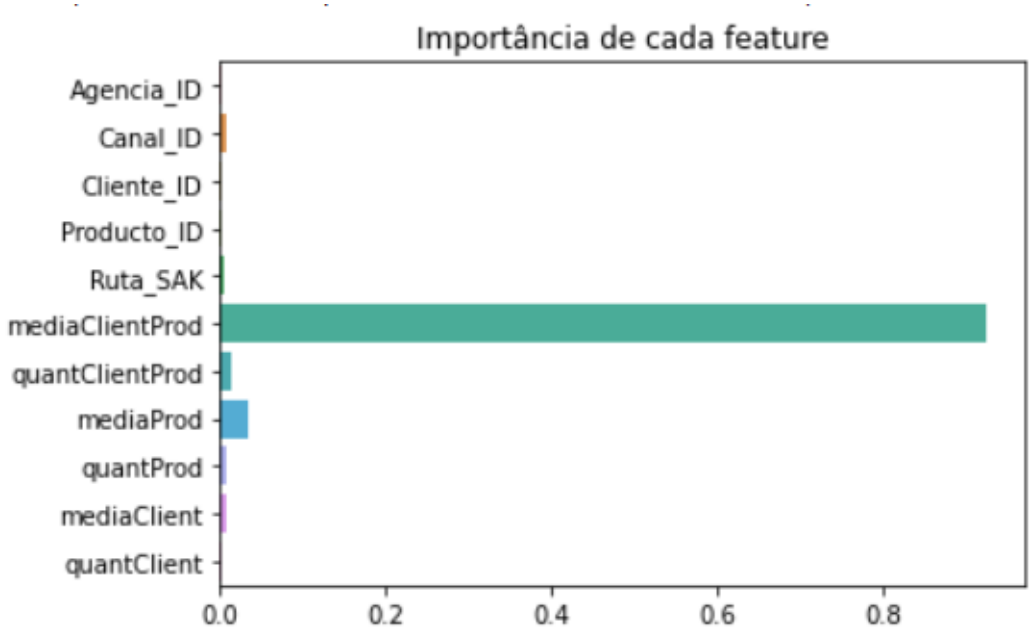
Com esse levantamento da importância das features de cada modelo, foi possível perceber que cada modelo tem um conjunto de melhores features. As Figuras 3,4 e 5 mostram que a feature *mediaClientProd* teve um grau maior de importância em ambos modelos. Isso nos diz que essa feature tem um alto grau de correleção com a variável a ser predita.

Figura 3 – Importância de cada feature para o modelo RF



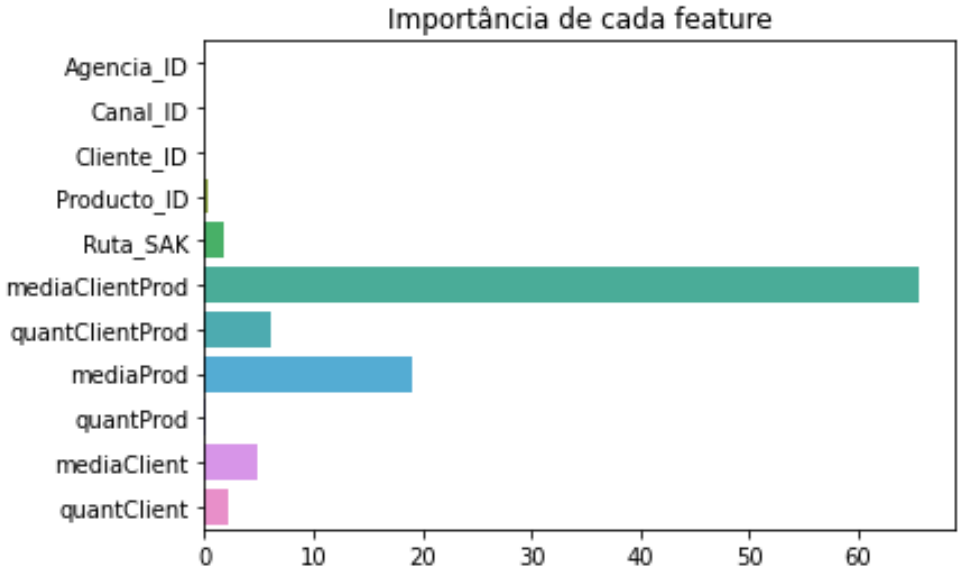
Fonte: o autor

Figura 4 – Importância de cada feature para o modelo XGBoost



Fonte: o autor

Figura 5 – Importância de cada feature para o modelo CatBoost



Fonte: o autor

5.4 Análise Comparativa dos Modelos

Os resultados obtidos no processo de aplicação dos métodos sobre os dados da competição, foram comparados e classificados do melhor para o que obteve pior desempenho, levando em consideração as métricas de avaliação RMSE e MAE. Na tabela Tabela 8 traz os resultados obtidos nos modelos para cada métrica.

Tabela 8 – Resultados obtidos através das métricas de avaliação para cada modelo

Métrica	Algoritmo		
	RF	XGBoost	CatBoost
RMSE	0.46	1.15	0.54
MAE	0.59	0.95	0.63

Fonte: o autor.

Os valores obtidos pela métrica MAE mostram que existe um média na taxa de erros das previsões da demanda dos produtos em 0.59 no melhor caso e 0.95 no pior caso, isso indica que para a janela de tempo prevista, houve um erro de 1 produto para mais ou menos. Já na métrica RMSE, vimos que no melhor caso a taxa é de 0.46 no modelo RF e 1.15 no pior caso para o modelo XGBoost.

Avaliando os resultados , pode-se perceber que o modelo RF teve o melhor desempe-

nho tanto na métrica RMSE como na métrica MAE. O modelo que teve melhor desempenho foi CatBoost. Já o modelo XGBoost teve o pior resultado entre os três.

Os códigos fonte estão disponíveis no repositório: https://github.com/mauriciosantos21/tcc.

6 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho, foram realizados estudos para comparação de técnicas de aprendizagem de máquina para a predição de demanda de produtos perecíveis. Para comparação foram utilizados métodos de regressão, a fim de encontrar o melhor modelo para predição da demanda dos produtos da competição *Grupo Bimbo Inventory Demand* na plataforma Kaggle. As features criadas foram utilizadas em ambos modelos de forma igual.

Cada modelo foi construído com as features selecionadas na fase de Análise e pré-processamento dos dados, foi utilizada técnica de seleção de hiperparâmetros, afim de selecionar os melhores parâmetros para execução de cada modelo. Os resultados da execução de cada modelo foram analisados. E o modelo RF foi o que obteve melhor desempenho em ambas às métricas MAE e RMSE. Já o modelo CatBoost teve o segundo melhor desempenho, ficando o resultado bem próximo ao do modelo RF na métrica MAE. Já o modelo XGBoost teve o pior desempenho em relação aos outros dois modelos.

Acreditamos que os resultados que foram obtidos nesse estudo ainda tem condições de serem melhorados, e para estudos futuros fica como sugestão, buscar utilizar análises estatísticas para o auxílio na implementação de novos modelos com novas features que impactem na análise dos dados, buscar de alguma forma de aumentar o tamanho da amostra de dados de treino e teste, porque nesse estudo por falta de recursos computacionais, não foram utilizados em sua totalidade os dados disponilizados pela competição.

REFERÊNCIAS

AURÉLIEN, G. **Hands-on machine learning with scikit-learn and tensorflow**: concepts, tools, and techniques to build intelligent systems. [S.l.]: O'Reilly, 2019.

BREIMAN, L. Random forests. In: **Machine Learning**. [S.l.: s.n.], 2001. p. 5–32.

CHAN, S.; TRELEAVEN, P. Chapter 5 - continuous model selection for large-scale recommender systems. In: GOVINDARAJU, V.; RAGHAVAN, V. V.; RAO, C. (Ed.). **Big Data Analytics**. [S.l.]: Elsevier, 2015, (Handbook of Statistics, v. 33). p. 107–124.

CHEN, T.; HE, T.; BENESTY, M.; KHOTILOVICH, V.; TANG, Y.; CHO, H.; CHEN, K. *et al.* Xgboost: extreme gradient boosting. **R package version 0.4-2**, v. 1, n. 4, p. 1–4, 2015.

DELGADO FILHO, A. J. F. **Análise de métodos de regressão para previsão de demanda de curto prazo**. Dissertação (Mestrado) — UFPE, 2020.

DIAS, M. **Administração de materiais**: uma abordagem logística. [S.l.]: Editora Atlas S.A., 2010. ISBN 9788522459193.

DOROGUSH, A. V.; ERSHOV, V.; GULIN, A. Catboost: gradient boosting with categorical features support. **arXiv preprint arXiv:1810.11363**, 2018.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, JSTOR, p. 1189–1232, 2001.

GERBER, J. Z.; MIRANDA, R. G. de; BORNIA, A. C.; FREIRES, F. G. M. Organização de referenciais teóricos sobre diagnóstico para a previsão de demanda. **GESTÃO. Org**, Universidade Federal de Pernambuco, v. 11, n. 1, p. 160–185, 2013.

HIGUCHO, A. K. A previsão de demanda de produtos alimentícios perecíveis: três estudos de caso. **REA-Revista Eletrônica de Administração**, v. 5, n. 2, 2006.

MANCUZO, F. **Análise e previsão de demanda**: estudo de caso em uma empresa distribuidora de rolamentos. Dissertação (Mestrado) — UFRGS, 2003.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of machine learning**. [S.l.]: MIT press, 2012.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: **Sistemas inteligentes fundamentos e aplicações**. Barueri-SP: Manole, 2003. p. 89–114. ISBN 85-204-168.

PROKHORENKOVA, L.; GUSEV, G.; VOROBEV, A.; DOROGUSH, A. V.; GULIN, A. Catboost: unbiased boosting with categorical features. **Advances in neural information processing systems**, v. 31, 2018.

ROZA, F. S. d. **Aprendizagem de máquina para apoio à tomada de decisão em vendas do varejo utilizando registros de vendas**. Monografia (TCC) — Bacharelado em Engenharia de Controle e Automação, Universidade Federal de Santa Catarina, Florianópolis, 2016.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. **Understanding machine learning**: from theory to algorithms. [S.l.]: Cambridge University Press, 2014.

TUBINO, D. F. **Planejamento e controle da produção**: teoria e prática. [S.l.]: Editora Atlas SA, 2000.

WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. **Climate research**, v. 30, n. 1, p. 79–82, 2005.