UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ADMINISTRAÇÃO

GUILHERME BRITZ FONTOURA

DATA SCIENCE NO APOIO À TOMADA DE DECISÃO COM DADOS EM UM MACROMERCADO

PORTO ALEGRE
2021

GUILHERME BRITZ FONTOURA

DATA SCIENCE NO APOIO À TOMADA DE DECISÃO COM DADOS EM UM MACROMERCADO

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciências Administrativas da Universidade Federal do Rio Grande do Sul, como requisito parcial para a obtenção do grau de Bacharel em Administração.

Orientador: Daniela Brauner

PORTO ALEGRE

2021

SUMÁRIO

1 INTRODUÇÃO

O consumo faz cada vez mais parte da vida do brasileiro, impulsionado tanto pelo modo econômico vigente quanto pelo consumo conspícuo, termo que primeiro aparece nos escritos de Thorteins Veblen(1899) e indica um consumo por ostentação, em termos práticos, a diferença entre querer comprar um carro e escolher o carro mais bonito e luxuoso ao invés de um carro econômico.

Um setor que se beneficia do aumento de consumo é o varejo, mais especificamente o varejo restrito, que em 2017 teve uma participação de 20,25% do PIB brasileiro, segundo informações da Sociedade Brasileira de Varejo e Consumo (2018). O IBGE(2020) define o varejo restrito como todas as atividades de consumo excetuando-se materiais de construção e veículos e se refere a ele apenas como varejo. Dentro deste pedaço do setor, existem os comércios que vendem bens de consumo, como os supermercados, atacados, bazares, lojas de vestuários e outras variações de consumo, com o propósito de compra ou produção e futura venda destes produtos de consumo, lucrando na margem.

O varejo em geral, assim como seu subsetor mais especializado de macroatacado, é um setor que, especialmente no Brasil, necessita que haja uma maior exploração sobre seus dados captados. O setor capta muitos dados de comportamento de clientes, que por consequência podem acabar virando insumo para modelos de predição e/ou para tomadas de decisões.

Como diz Rogers(2020), o processo de digitalização é importantíssimo para todas as empresas que querem concorrer em vantagem sobre as outras. No seu livro, ele traz um exemplo do Walmart que se utiliza de dados meteorológicos para poder prever a demanda, pois em dias que chove há um menor fluxo de visitas às lojas físicas e um maior fluxo às lojas online.

A tomada de decisão para compor esta estratégia de digitalização é um processo que necessita de informações, quanto mais melhor. “A informação é um recurso efetivo e inexorável para as empresas, especialmente quando planejada e disseminada de forma personalizada, com qualidade inquestionável e

preferencialmente antecipada para facilitar as decisões.” (REZENDE, 2005 p.247). Estas informações podem vir de diversas fontes, como pesquisas do consumidor, análise de mercado, análise da concorrência ou até de dados internos de consumo (PROVOST e FAWCET, 2016).

Com o crescente volume de dados, para análise destas informações, tornou-se essencial o uso de algoritmos que automatizam a extração de conhecimento de dados, auxiliando os gestores a tomarem decisões fundamentadas em dados e qualificados de acordo com o perfil dos seus clientes (PROVOST e FAWCET, 2016). Esta nova cultura de decisão é chamada de *data driven decision making* (DDDM), sendo adotada com sucesso por empresas como o Grupo Pão de Açúcar (Folha UOL, 2018) e Zara (Tera, 2019). Ela é muito benéfica para empresas com foco em bens de consumo, pois precisam de uma resposta rápida e efetiva contra as estratégias dos concorrentes, ou até para responder às necessidades de seus clientes, muitas vezes antecipando seus próprios desejos (Stern Speakers, 2021).

Um ponto de apoio para o gestor é conhecer a jornada do consumidor dentro de seu estabelecimento. Segundo Kotler(2010), no marketing 1.0 o cliente era visto como massa, não sendo tratado de forma customizada, pois eventualmente iria acontecer o consumo, ou seja, sem customização. No marketing 2.0 o autor exemplifica demonstrando que agora o consumidor é considerado “portador de desejos”, sendo considerado um indivíduo e portanto, segmentado em grupos pelos profissionais de marketing, neste ponto surgindo as propagandas orientadas a grupos específicos. Já no 3.0 o consumidor possui algum tipo de comunicação com a empresa, permitindo feedbacks construtivos.

Em um artigo mais recente, Kotler(2017) acrescenta o conceito do marketing 4.0 em que o consumidor participa ativamente do processo de venda. Este processo de conhecimento do seu público, identificação dos desejos e como o consumo ocorre por parte dos clientes individuais é chamado de jornada do consumidor. Esse conceito é crucial para entender como o consumidor se comporta no processo da compra, o que ele prioriza, quais produtos ele está interessado e qual o tempo dedicado a cada processo da compra.

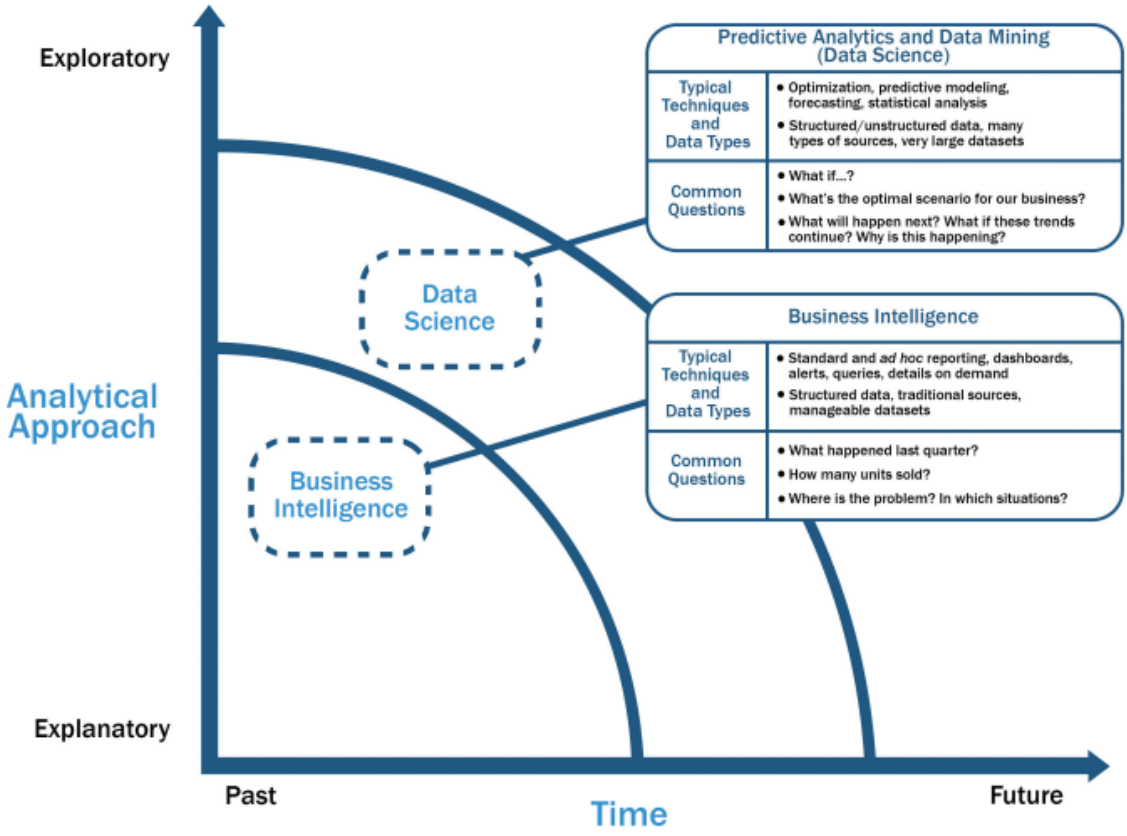
Para resolver o problema de encontrar respostas rápidas, efetivas e preditivas, que se utilizam de informações, criou-se um ecossistema inteiro de ferramentas de análise de dados, entre elas algoritmos de *machine learning* especializados em fornecer respostas e predições de mercado (PROVOST e FAWCET, 2016). A integrante do time de comunidade cloud da Azure, Zorovich(2021), que faz parte de uma das maiores infraestruturas *cloud* do mundo, recomenda, entre algumas opções, utilizar os algoritmos de predição para reestocar prateleiras e sempre garantir que os melhores produtos estejam nos lugares mais visíveis.

O termo *Machine Learning* (aprendizado de máquina) foi cunhado em um artigo de 1959, de autoria de A.L.Samuel(1959), em que ele descreve como um algoritmo que ele criou pode aprender rapidamente a jogar um jogo de damas. Em uma definição mais recente, no seu livro *Machine Learning: A guide to current research*, Mitchell(1977) define como algoritmos que são capazes de se aprimorar por experiência.

Esses algoritmos, ainda segundo Mitchell(1977), aprendem quanto mais dados eles consomem e ficam cada vez melhores. Para empresas de bens de consumo, que podem acumular muitos dados de compra, venda, devolução e entrega, o uso de aprendizado de máquina pode ser interessante para extrair *insights* (conhecimento aplicado) para melhorias de processos operacionais ou de marketing. .

Na administração o seu uso é mais comumente ligado a análises de processos de negócios. Antes uma área chamada de *Business Inteligence* (Inteligência do Negócio) que era dominada por administradores realizando análises e retirando *insights* para melhorar o alvo destas análises, agora algumas empresas já migraram para um novo conceito de uma área chamada apenas de *Dados*.

Figura 1 - Business intelligence and Data Science



Fonte: Dietrich et. al. (2015)

Como podemos ver na figura de Dietrich et. al. (2015), os antigos administradores que eram chamados de analistas de negócio transformam-se em *Data Scientist*. Esta nova leitura dos cargos possibilita uma ênfase maior no capital humano que se apropria do dado e o transforma em algo útil, utilizando as ferramentas para automatizar o trabalho da captura dos dados.

A apropriação e valorização desses dados, seja ele comportamental, geográfico ou de outra categoria, é o que torna empresas que adotam a cultura DDDM mais eficientes em sua tomada de decisão. Segundo a empresa Frost & Sullivan(2018), o mercado de *Big Data* e *Analytics* na América Latina movimentou $2.9 bilhões em 2017, sendo que o Brasil compôs 47.6% dessa receita e a empresa ainda projeta que a receita no ano de 2023 será de $8.5 bilhões, se manter a mesma proporção o Brasil gerará cerca de $4.04 bilhões apenas provenientes de tecnologias de análise de dados.

Segundo Rogers(2020), tanto dados quanto clientes são eixos estratégicos essenciais para a transformação digital. O uso de dados para entender o comportamento dos clientes e apoiar a tomada de decisão, se torna estratégia essencial para garantir a competitividade no mundo digital.

Neste sentido, o presente trabalho explora a seguinte questão: o uso de algoritmos de *machine learning* consegue melhorar a capacidade de previsão da demanda para auxiliar a tomada de decisão de um macroatacado?

2 OBJETIVOS

2.1 Objetivo Geral

O presente estudo tem como objetivo validar o apoio à tomada de decisão com dados para a estratégias de negócio de macromercados através da aplicação de data science e algoritmos de inteligência artificial para supermercados. O estudo será realizado junto a um supermercado, mais especificamente do seu setor de delivery, localizado no interior do Rio Grande do Sul.

2.2 Objetivos específicos

- Compreender como o data science pode apoiar à tomada de decisão dos supermercados;
- Identificar dados e algoritmos que possam ser aplicados para auxiliar a previsão de demanda no delivery
- Analisar os gráficos e análises retornadas pelos algoritmos comparado com as vendas reais para comprovar sua eficácia.

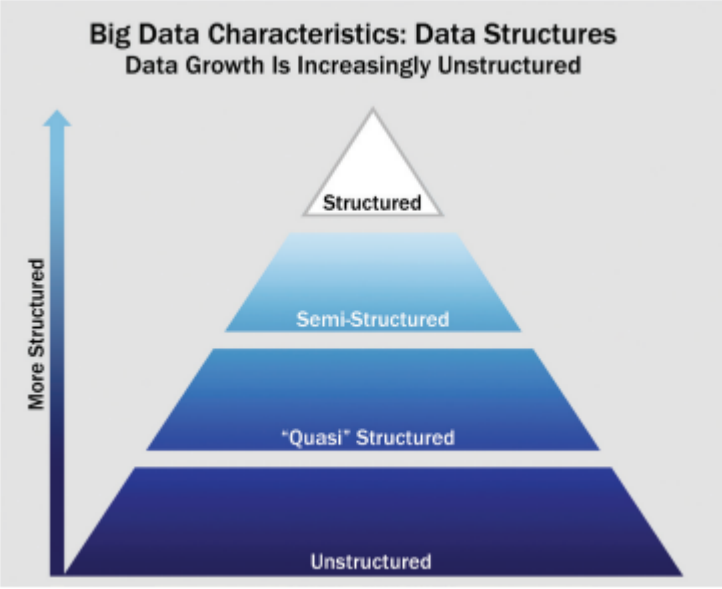
3 REFERENCIAL TEÓRICO

3.1 Estrutura de dados

Todo processo de análise depende, como seu principal fator, de dados. Estes dados podem ser infinitas coisas, como a quantidade de vezes que um motorista vira à direita em sua cidade, quanto a porcentagem de flores polinizadas durante o verão.

Na sociedade atual, praticamente todo eletrônico gera algum tipo de dado. Segundo o Dietrich et. al.(2015), é estimado que 80 a 90% do crescimento futuro da disponibilização do dado será de forma não estruturada, sendo este o dado no formato mais puro.

Figura 2 - Estrutura de dados



Fonte: EMC Education Services

Ainda segundo a organização, os dados possuem, em sua maioria, 4 estados, descritos na figura 2:

- Estruturado: o dado em sua forma final, geralmente são necessários pouca ou nenhuma alteração nele para que os dados contidos na estrutura sejam consumidos e virem informações. Possuem uma estrutura bem definida, como definições de tipos de colunas e são geralmente disponibilizados em um arquivo oriundo de banco de dados, arquivos CSV (*Comma-separated values*) ou até simples planilhas de dados.
- Semi-estruturado: são arquivos ou textos que possuem uma estrutura definida a ponto de ser possível mapear o formato que o dado está sendo passado, desta forma automatizando a estruturação do mesmo. Exemplos são arquivos XML, como o código fonte de sites.
- Quasi estruturado: este é um formato diferente, pois possui um formato errático e requer um tempo e esforço considerável para torná-lo em informação útil, geralmente possuindo inconsistências e falhas em sua coleta. Um exemplo seria a jornada do consumidor perante a navegação do site, todas as URLs que ele clicar gerará um log de navegação que irá depois ser explorado para mapear como o usuário se comporta.
- Não estruturado: qualquer dado que não possui nenhum formato definido. Exemplos seriam vídeos, textos e pdfs sem formatação e imagens.

Então, quando se fala de dados pode-se falar de uma infinidade de conjuntos, tanto de formato quanto de adequação ao padrão textual. O segredo está em investigar a fundo o dado, transformando e estruturando ele enquanto está trabalhando e retirando alguma informação útil dele.

3.2 Data Driven Decision Making

A tomada de decisão dentro de uma empresa é sempre algo que tem que ser respaldado por diversas fontes, quanto mais conhecimento o tomador de decisão tem, mais informado e qualificado será sua decisão. Diante disso, vê-se necessário cada vez mais obter informações do ambiente ao seu redor, saber os padrões de compras e suas sazonalidades e até conhecer o perfil completo do seu cliente.

Uma forma automatizada desse processo de decisão passa por algoritmos extremamente complexos e que dependem de uma quantidade massiva de dados, inteligência artificial e informações úteis para que se possa prever comportamentos olhando apenas para os dados estatísticos.

Com novas tecnologias surgindo a cada ano, o acesso a esse tipo de informação se democratizou cada vez mais, impulsionado por profissionais que o seu único intuito é prover, manter e fazer análises sobre dados comportamentais como esses. Os *insights* gerados por estes profissionais podem então ser revertidos para dentro da empresa, ajudando-a a vender mais, colocar produtos em lugares estratégicos dependendo de onde o cliente mais olha, maior eficiência de reestocagem entre outros benefícios (PROVOST e FAWCET, 2016).

Ainda segundo os autores, empresas que focam nesses tipos de decisões apoiadas em dados possuem uma cultura de *Data Driven Decision Making*, ou seja, suas decisões provêm de dados captados de diversas fontes e que enriquecem o poder decisório, seja ele manual ou automático. Quanto mais dados e informações, melhor um algoritmo será treinado para reconhecer padrões e acertar em suas decisões. Quanto maior a qualidade do dado, melhor o gestor pode tomar decisões de mercado.

Uma das categorias que mais pode se beneficiar desta digestão de informação é o varejo. Ele é definido pelo BNDES(1999) como tendo as características gerais de: “procura e seleção de produtos, aquisição, distribuição, comercialização e entrega”, ao mesmo tempo que as empresas do setor disponibilizam crédito ao comprador, sendo esta uma característica marcante, podendo ser supermercados, vendas, pet shops, feiras, restaurantes e afins.

Segundo o indicador da Pesquisa Mensal do Comércio do IBGE(2020), o consumo no setor de varejo aumentou em 1,2% no ano de 2020, ocorrendo uma queda apenas nos meses iniciais da epidemia do novo coronavírus, que acabou ocasionando uma quarentena em que partes do comércio foram fechadas para retardar a proliferação do vírus. Este indicador demonstra o poder de venda do setor como um todo, pois são produtos necessários para a sobrevivência do ser humano, especialmente se olharmos para os indicadores de mercados em geral. É importante ressaltar que neste presente trabalho, quando se está falando do setor de varejo, fala-se especificamente de supermercados de pequeno porte e referenciando a eles como minimercados.

3.3 Clientes e dados

Rogers(2021) define em seu livro sobre transformação digital de negócios como a parte dos dados que são importantes para qualquer negócio prosperar na era digital. O autor comenta que todas as empresas têm de possuir uma estratégia de uso, governança e estruturação dos dados, de forma a se manter competitiva perante a concorrência e não infringir as leis de proteção de dados.

É necessário mapear todo o ecossistema de dados que a empresa possui para que se possa fazer uma transformação digital de sucesso. Rogers(2021) comenta que é necessário identificar como os clientes chegam até o negócio, qual o meio de publicidade mais adequado, como a plataforma gera o dinheiro e outros conhecimentos chaves para poder montar uma estratégia robusta.

Qualquer empresa hoje que atua no mundo digital tem que tratar seus dados como um ativo intangível, pois ele pode gerar um valor muito grande se usado. Porém, para que isso ocorra, é necessário trazer uma estratégia de dados coerente para a empresa, que se assemelhe com o seu negócio.

Ainda segundo o autor, esta revolução na tratativa de dados fez com que a The Weather Company (TWC) fosse de apenas uma empresa de meteorologia para uma das maiores provedoras de dados meteorológicos dos Estados Unidos, ao alavancar o

valor do seu dado e distribuí-lo para grandes empresas utilizarem em suas análises. Tudo isso ocorreu graças aos *Data Scientist* da empresa na época.

3.4 Digitalização de macromercados

Neste trabalho será usado a definição de Badin(1997) para atacarejo: “São lojas de auto-serviço pelos clientes (pegue e carregue), com alguns setores oferecendo serviços, e com linha completa de itens alimentares e não alimentares.” Os serviços podem incluir açougues acoplados dentro dos minimercados, ou podem oferecer outros tipos de serviços como padarias e afins.

O resultado obtido por Barros(2020), ao realizar uma análise de previsão de demandas em alguns mercados e atacados, demonstra que um controle maior do estoque é essencial para o crescimento sustentável de mercados. Percebe-se a diferença nítida entre administradores do estabelecimento que tem o perfil de viver sobre os resultados dos mercados, focando então na maximização do resultado financeiro e entre os administradores que tem um foco muito maior no resultado operacional, permitindo à empresa crescer, reinvestindo no estabelecimento com aquisição de sistemas gerenciais e outras soluções tecnológicas.

A automação desses processos decisórios como controle de estoque, previsão de venda e precificação dos produtos é cada vez mais frequente no mercado de varejo e portanto necessária para os macroatacados se manterem competitivos frente a competidores cada vez maiores e mais agressivos na região. Considerando que há frequentes avanços tecnológicos nessas áreas, os softwares que possibilitam o controle de estoque acabam ficando cada vez mais baratos, permitindo que o acesso seja democratizado a todos que possuírem o interesse, inclusive empresas que não possuem muito foco em tecnologia.

3.5 Apoiando a tomada de decisão em casos reais

Há diversas maneiras de explorar essas informações não captadas e transformá-la em dados úteis: primeiro é necessário que seja criada a infraestrutura necessária para captar essas informações, seguida por um tratamento delas por meio

de algoritmos criados especialmente para limpeza de dados e detecção de anomalias para enfim deixar o algoritmo de *machine learning* consumir esses dados e transformá-los em informações e *insights* úteis para plano de negócios da empresa.

Um exemplo desse processo é a melhora da consistência: a empresa Karna AI (KARNA, 2021) fornece um produto que promove a consistência de informações dos *SKUs* (*Stock Keeping Unit* - Unidade de Manutenção de Estoque) de todo o seu estoque através de todas as filiais da sua empresa. O algoritmo é treinado através de reconhecimento da imagem do produto e pode fornecer insights detalhados em tempo real de como o produto específico está performando nas lojas do controlador, além de detectar se o lançamento de novos produtos segue as regras descritas pelo setor de marketing, promovendo consistência e um *awareness* (como a marca é lembrada) da marca.

A Zara, uma rede de lojas que vende roupas e acessórios com foco no público feminino, possui uma intensa infraestrutura de captura, leitura e análise de dados. Segundo um artigo de Uberoi(2017), a empresa instala um microchip RFID em todas as peças de roupas que saem dos seus armazéns de produção, acompanhando em tempo real onde está cada peça de roupa individualmente. Este chip permite a Zara descobrir quanto tempo uma peça individual leva do armazém de produção até ser vendida e promove um controle de estoque e vendas padronizados a um nível organizacional muito eficiente.

A tecnologia RFID permite um controle muito forte sobre seu produto. Uberoi(2017) continua em seu artigo comentando sobre acompanhar as vendas de seus produtos em tempo real é crítico para a Zara. Seu time de engenheiros recebe *feedback* em tempo real de clientes comentando sobre como o zíper é muito longo de tal peça, ou pedindo um tamanho que não existe na loja e em alguns dias um novo modelo de roupa está criado.

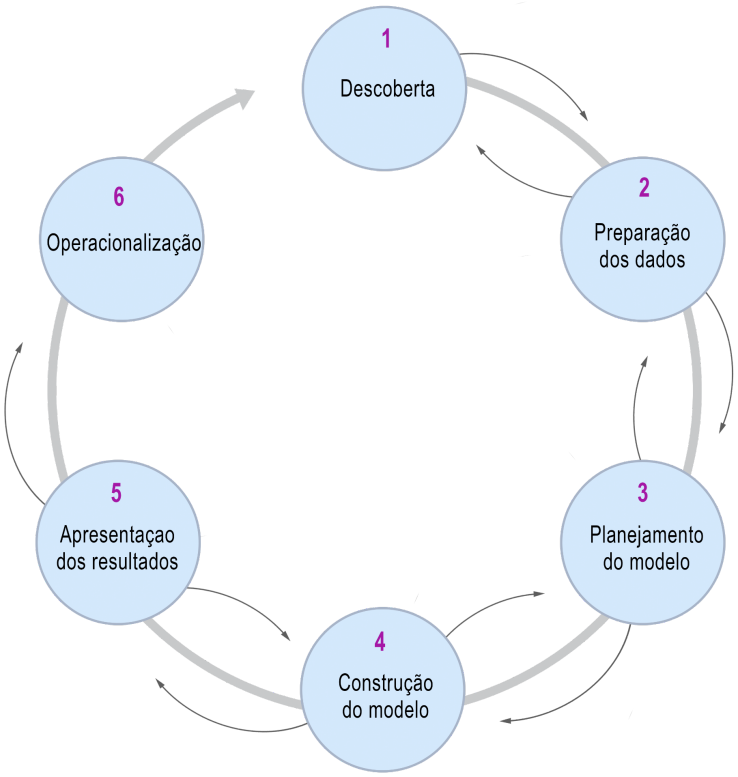
Este tipo de controle sobre os dados comportamentais é extremamente importante para a indústria do varejo, em especial mercados, pois há muitos dados não sendo explorados de maneira correta, resultando numa experiência não muito satisfatória por parte dos clientes, como falta de produtos na prateleira por insuficiência de estoque, ou preços muito altos se comparados aos seus concorrentes.

3.6 Ciclo de vida de um projeto de dados

Toda matéria no mundo tem um ciclo de vida finito. No caso de dados, Dietrich et. al.(2015) define que há certas etapas do ciclo de vida de um dado, em especial aqui falando de um projeto inteiro, que se retroalimentam até formar um novo ciclo de vida.

A retroalimentação do sistema é uma etapa essencial da vida no projeto que irá se utilizar do *insight*, pois ela que dita o valor do dado em si, pois somente via múltiplas etapas de validação podemos refinar o dado ao ponto dele virar uma informação, ou seja, ao ponto de virar uma métrica mensurável que agrega valor ao negócio.

Figura 3 - Ciclo de vida do projeto



Fonte: Taylor (2016) - adaptado pelo autor

No ciclo descrito na figura 3, a primeira etapa é a da “descoberta”. Esta etapa é importante no sentido de que dá toda a fundamentação do trabalho futuro do projeto. É nele que são definidas as questões do objetivo de negócio, etapa fundamental para a entrega de valor no futuro. Também são aprendidas as questões de negócio que se planejam responder com o projeto, assim como se desenvolve a primeira hipótese do que o dado pode responder.

A partir disso, a próxima etapa é a de “preparação dos dados”. Esta é uma das etapas que mais consome o tempo do projeto, pois nela ocorre qualquer transformação nos dados pertinentes à análise. Requer o uso de um *sandbox* analítico, ou seja, algum *software* que consiga carregar esse dado e possua processamento suficiente.

Durante a preparação dos dados ocorrem diversos processos de ETL (*Extract*, *Transform and Load)* ou em portugues - Extrair, transformar e carregar. Estes processos são importantíssimos para que o projeto possa funcionar de uma forma ordenada, pois usualmente ele é automatizado e consome uma boa parte do tempo para que se faça de uma forma correta e que atenda as necessidades do modelo. Importante nesta etapa mencionar que podemos voltar no ciclo do projeto caso necessitemos refinar a etapa de “descoberta”, a fim de termos uma base melhor em como preparar o dado.

Começando os modelos algoritmos, temos a fase 3 que é o “planejamento do modelo”. Nesta etapa de planejamento, define-se as ferramentas, *frameworks* e métodos que irão ser utilizados para a construção do modelo. Aqui também é descoberto a relação entre as variáveis dos dados, usando funções estatísticas para conhecer a correlação destes dados. Caso não tenha dados suficientes para o planejamento eficaz do modelo, pode-se voltar para a etapa 2 e preparar mais o dado.

Durante a próxima etapa, “construção do modelo”, ocorre a execução das ideias propostas durante o planejamento. Nela, ocorre todo o tipo de teste para que se assegure que as máquinas que irão rodar o modelo aguentem a carga de processamento, assim como os testes para verificar se o modelo planejado realmente condiz com o dado em sua forma final. Caso negativo, volta-se ao planejamento para refinar as ferramentas utilizadas.

Por fim, como última etapa real de execução, é apresentado o resultado para as partes interessadas. Essa é a parte mais importante do projeto como um todo, pois é

nela que é decidido se o projeto vai em frente ou não. É necessário quantificar as métricas que o modelo gera, qualificar o retorno destes dados ao negócio e construir uma narrativa em que os tomadores de decisão possam entender o que se passou no projeto sem necessariamente ter o conhecimento técnico.

Caso tenha dado tudo certo até aqui, o modelo é posto em prática em um ambiente real, muitas vezes realizando um projeto piloto para que o modelo comece em uma pequena escala de dados. Nela, também se realizam todas as documentações técnicas necessárias para o entendimento do projeto, assim como as reuniões finais e ajustes no código para o sustentamento do projeto.

3.7 Algoritmos de *machine learning*

Previsão de demanda é uma métrica muito importante para saber como está a saúde do seu negócio e como será daqui para frente. Barros(2020) afirma que é possível encontrar resultados satisfatórios com uma amostra pequena de dados, como será o caso das vendas de macromercados.

Portanto, para realizar essa previsão irão ser utilizados dois algoritmos de inteligência artificial e comparar entre eles: o modelo LSTM (*Long Short-Term Memory*) e modelos de regressão. Os dois se utilizam do conceito de rede neural, que é todo o algoritmo que possui uma etapa de entrada de dados (*input*), uma etapa de saída de dados (*output*) e uma camada oculta (aos olhos do usuário) entre eles, que são conectados através de nós (*nodes*) formando uma rede, um comportamento similar aos nossos neurônios (SAS, 2021).

LSTM é o nome dado para um modelo de rede neural que é muito utilizado em análises de dados sequenciais, isto é, todo dado que tem uma sequência lógica a ser seguida. Uma base de dados que oferece os dados de venda de determinado produto ou mercado é um dado sequencial pois possui uma sequência a ser seguida, neste caso a data da venda. Este tipo de modelo pode ser simples ou complexo, dependendo das variáveis escolhidas e inseridas dentro do algoritmo, como sazonalidade dos produtos.

Uma parte importante ao trabalhar com algoritmos de inteligência artificial é que eles precisam necessariamente ser treinados. Para isso, é necessário que haja dados históricos de venda para que o modelo possa consumir estes dados e depois conseguir gerar uma previsão de como serão as vendas a partir da data escolhida. Quanto mais dados o algoritmo consumir, mais complexo e enriquecido ele vai ficar, mas também mais acurado ele será. O LSTM, nesse caso, será treinado com dados históricos da venda do macromercado escolhido, para então projetar as vendas do mesmo através do último ano de dado, podendo assim comparar se o modelo acertou nas vendas ou não.

O outro modelo escolhido para comparação são os de regressão, redes neurais auto-regressivas que se utilizam de médias móveis para fazer previsões sobre as vendas. É um modelo muito usado na economia para fazer análises de regressões sobre dados, obtendo dados confiáveis. Ele será treinado com os mesmos dados históricos mas fará previsões dinâmicas e agrupará os resultados em médias.

Os modelos serão comparados no seu grau de acuracidade contra os dados reais de vendas, sendo testados os seus índices de raiz do erro quadrático médio (*root mean squared error,* RMSE) e erro médio absoluto (mean absolute error, MAE).

4 PROCEDIMENTOS METODOLÓGICOS

A seguir irão ser dispostos os procedimentos metodológicos utilizados nesta pesquisa para conseguir atingir os resultados propostos.

4.1 Tipo de pesquisa

O presente trabalho é uma pesquisa bibliográfica para alinhamento conceitual sobre o tema, seguido de uma análise exploratória usando dados reais de vendas fornecidas pelo macromercado estudado. Esta análise permitirá explorar a fundo os padrões de comportamento dos consumidores evidenciados pelos dados.

4.2 Captura dos dados

O macroatacado alvo do estudo é situado numa cidade do interior do Rio Grande do Sul com 343 mil habitantes (IBGE, 2021). O negócio implantou um *delivery* em 2017 e com as restrições impostas pela pandemia da covid 19 foi percebido um aumento da demanda.

Para o estudo foi solicitado à empresa um *dataset* contendo seu histórico de vendas, para que se pudesse obter o maior número de dados históricos, podendo assim ter uma maior acuidade nos modelos propostos.

O dataset enviado contém dados de novembro de 2017 até outubro de 2020.

4.3 Dicionário de dados

No arquivo principal enviado pela empresa, foi enviado um *dataset* contendo as seguintes colunas:

- DATA: consta a data da venda de determinado produto, com o formato de dd/mm/yy. Ex: 14/06/19.
- NOME: classificação de nome dado ao produto da venda. Ex: MIST BOLO FLEISCHMANN 450G MILHO.
- EAN: número identificador único atrelado ao produto. Ex: 7898409950889
- EMBALAGEM: identificador que diz ao leitor se o produto é UN 1, ou seja, vendido por uma unidade, ou KG 1, quando é vendido fracionado.
- LITROS: identificador que deveria reportar quantos litros o produto (caso aplicável) tenha.
- QUANTIDADE: quantidade total do produto vendido em determinado dia. Ex: 5.
- QUANTIDADEUNIT: quantidade total do produto vendido em determinado dia, unitariamente. Ex 1 para unidades inteiras e 0.366 para fracionários.
- VALOR_VENDA: valor total da venda do produto no determinado dia. Ex: 86.

4.3 Limpeza dos dados

Foram realizados diversos tratamentos no *dataset* recebido, a começar por todas as entradas nulas serem excluídas.

Em seguida, verificou-se o estado de cada uma das colunas, tentando encontrar comportamentos anômalos nelas. A coluna ‘LITRO’ chegou totalmente zerada, então removeu-se ela do *dataset*. Do mesmo jeito, a coluna ‘EMBALAGEM’ não trazia informações pertinentes à análise em questão, então também foi removida.

Uma parte importante nessa verificação de anomalias foi que encontrou-se que diversos produtos tinham 2 ou 3 *eans* (identificador único de produto) diferentes atrelados a ele, o que vai contra ao consenso de que o ean deve ser um identificador único de produto. Percebeu-se, porém, que a coluna ‘NOME’ mantinha-se constante durante o histórico, então ao invés de usar o *ean* como identificador principal do produto é utilizaado o nome do mesmo.

Após, transformou-se a coluna ‘NOME’ para que se pudesse usar ela de uma forma padronizada, visto que algumas entradas tinham espaço antes ou após o nome do produto.

Como uma tentativa de clusterizar os dados através da categoria dos mesmos, foi solicitado, como um adendo ao pedido original, um repositório extra de dados que conteria o nome do produto e a categoria do mesmo. Obteve-se apenas o nome do produto e a subcategoria. Através desse repositório extra, foi possível cruzar com a tabela original e obter a subcategoria do produto.

Durante a tratativa acima, percebeu-se que faltavam cerca de 130 produtos a serem categorizados pela tabela enviada. Recorreu-se então a classificar eles manualmente, pesquisando no site da empresa o produto e catalogando a subcategoria do mesmo, para ficar no padrão já enviado.

Por fim, teve-se que realizar uma extensa modificação nos dados, pois a tabela original foi enviada com o delimitador vírgula (,) e os decimais das tabelas numéricas também foram enviados como vírgula, resultando assim em uma não catalogação dos

decimais de valores. A solução foi criar colunas extras que abrigassem temporariamente estes valores e que se juntassem à sua coluna correta no final do tratamento do arquivo.

4.4 Análise dos dados

Será utilizado a plataforma Kaggle, para a aplicação dos algoritmos de *machine learning* e a plataforma Databricks para a limpeza de dados. Ambos são plataformas para a execução de códigos utilizando processamento de máquinas virtuais e foram utilizados em sua versão não paga.

5 RESULTADOS

Com os procedimentos de limpeza e categorização da base realizados, é possível começar a obter valor através dos dados disponibilizados. Iremos utilizar o livro “Introduction to Time Series and Forecasting”, de Brockwell & Davis(2002) como o norteador para as etapas necessárias a serem realizadas para que o algoritmo seja feito de forma eficiente.

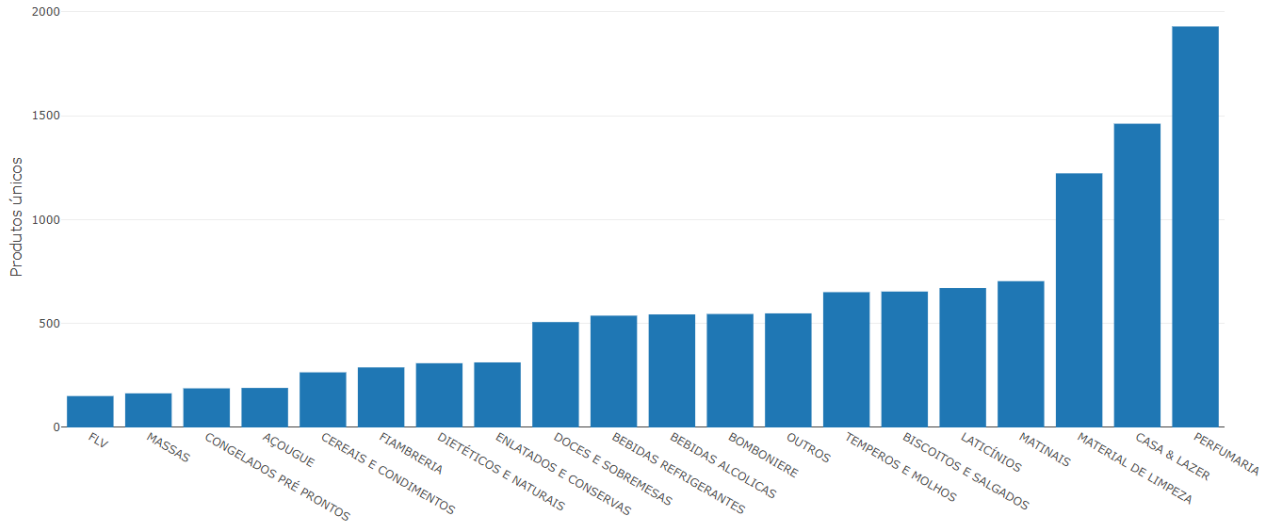
A análise de dados, quando não explicitado contrariamente, se dará sobre o *dataset* total enviado, salvo as transformações relatadas na metodologia. É necessário para manter a análise de forma mais fidedigna aos dados originais.

5.1 Resultados gerais

Primeiramente foram feitas análises mais gerais sobre o estado do *dataset* em si, tentando obter informações sem o uso de um modelo específico para os dados.

Uma informação importante, antes de se realizar a análise sobre estes dados, é que foram agregadas as categorias: “confeitaria”, “preservativos”, “artigos esportivos”, “rotisseria”, “bebe”, “produtos inativos”, “cesta basica”, “bebê & criança”, “luva mascara toucas”, “resfriados pré prontos”, “pet” e “padaria”, pois estas categorias correspondem a menos de 4% das vendas totais dos produtos, tornando a análise delas melhor realizada com a agregação destes produtos dentro da categoria “OUTROS”.

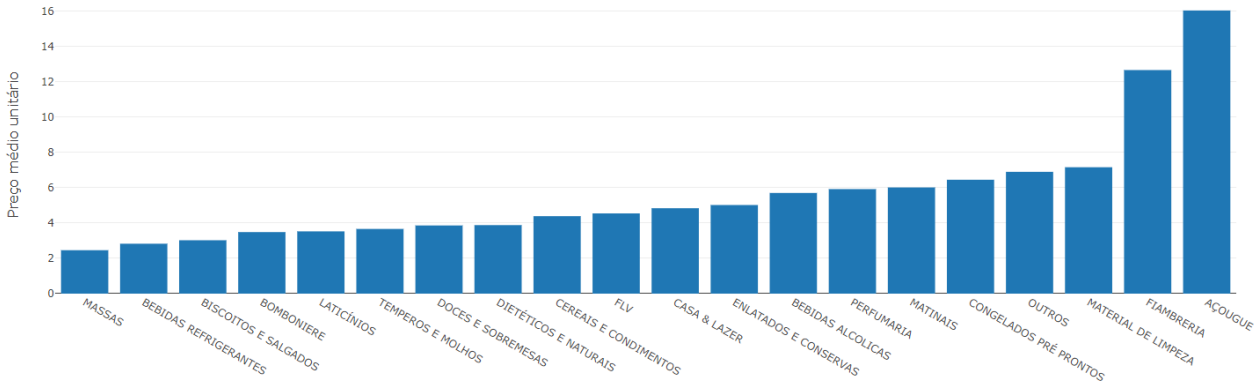
Gráfico 1- Produtos únicos por categoria



Em uma análise da quantidade de produtos únicos em cada categoria, temos a categoria “perfumaria” tendo a maior parte dos produtos catalogados, levando a crer que, como há muitos produtos desta categoria, seria uma das mais vendidas. Irá ser verificado essa hipótese logo adiante.

Em seguida, tem as categorias “casa & lazer” e “material de limpeza” como categorias que possuem uma grande quantidade de produtos. Historicamente, são produtos que possuem concorrência alta, ou seja, há várias opções de produtos diferentes para um mesmo objetivo.

Gráfico 2- Preço unitário médio por categoria

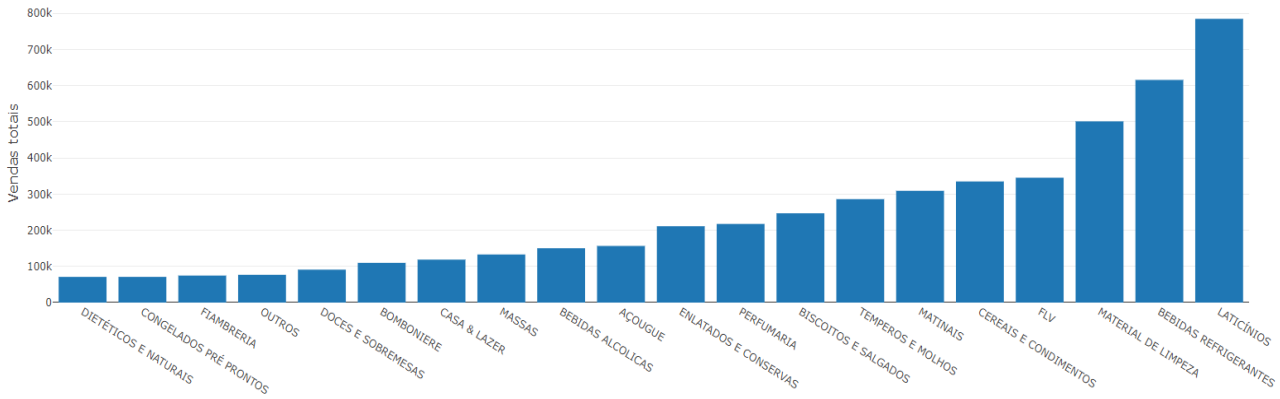


Seguindo a análise, é percebido pelo preço unitário médio de cada categoria, que “açougue” e “fiambreria” possuem o maior valor por quantidade de todas as categorias. Isso deve-se à serem produtos que tem um valor alto por kilo, incluindo aqui carnes, presuntos, queijos e outros derivados.

A informação de preço unitário médio por si só não traz muita informação, pois um grupo de produtos pode possuir um preço unitário médio alto mas mesmo assim não ser vendido em uma quantidade suficiente para que seja uma empreitada lucrativa.

Aqui, caso tivesse o identificador de usuário para cada compra, poderíamos realizar uma análise dos produtos com preço unitário médio alto que mais atraem a compra de outros produtos, realizando uma análise de carrinho. Com isso, é possível dizer quais produtos especificamente o macromercado teria que focar para obter uma melhora significativa em sua receita.

Gráfico 3 - Quantidade total de vendas por categoria

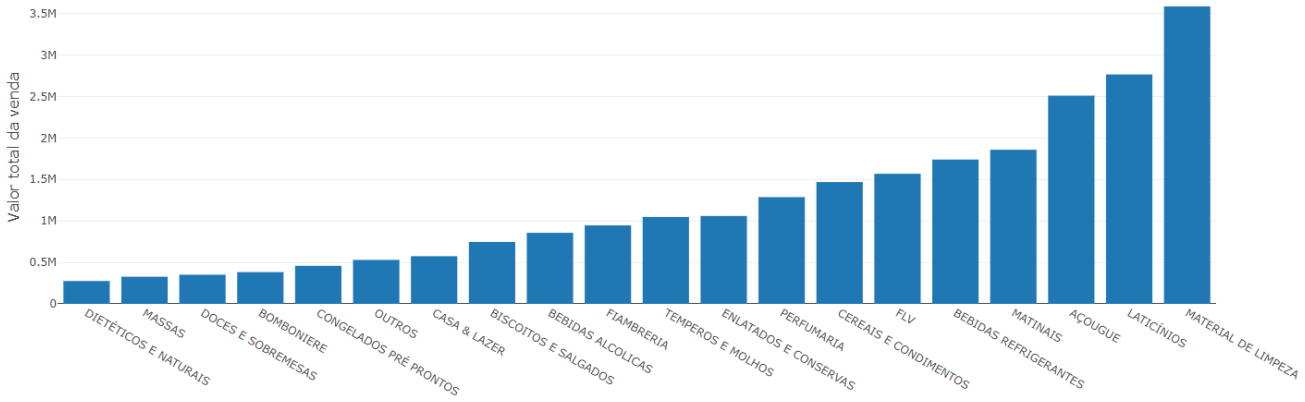


Este gráfico traz informações mais relevantes para uma análise de predição de vendas. É possível perceber que as 3 categorias que mais vendem são “laticínios”, “bebidas refrigerantes” e “material de limpeza”, seguidos mais atrás por “flv” (frutas, legumes e vegetais) e “cereais e condimentos”.

Com essas informações, é possível inferir que o perfil de consumidor do macromercado utiliza o serviço para comprar produtos de categorias com um preço unitário médio e médio-baixo, visto que as categorias mais vendidas começam a aparecer apenas na metade do gráfico de preço unitário médio.

Um outro *insight* que há nas vendas é de que a única exceção ao que foi comentado acima são os materiais de limpeza. Uma das possibilidades é de que é mais cômodo comprar eles diretamente no *delivery*, visto que provavelmente os consumidores não possuem outro local de compra destes produtos, elevando os preços dos mesmos pela precificação da demanda alta sem concorrência.

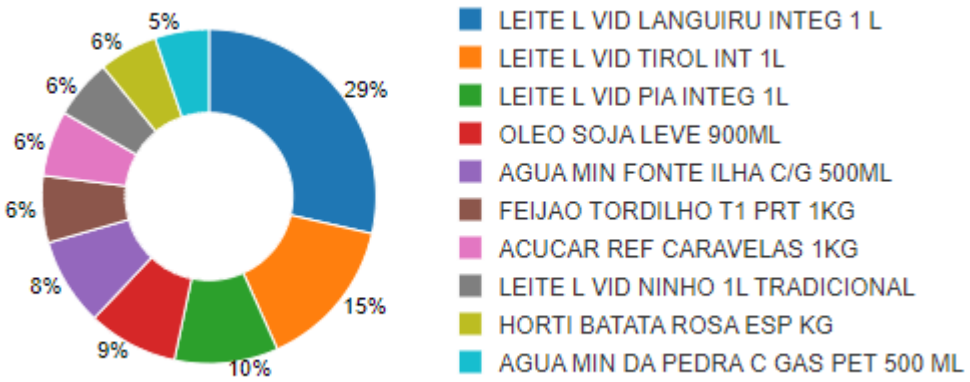
Gráfico 4- Valor total das categorias



É possível ver pelo valor que a categoria “material de limpeza”, embora tenha um preço unitário médio de apenas R$ 7,50 reais, dispara como a que atrai mais receita para a empresa. Como não há os custos da aquisição dos produtos, não podemos calcular o lucro de cada um, mas acreditamos que o lucro deve seguir o mesmo padrão da receita.

É destaque também a categoria “açougue”, que embora seja a 11º categoria mais vendida, aparece na 3º posição do valor total de vendas, consolidando o seu *status* de um produto acima da média em valor.

Gráfico 5 - Produtos mais vendidos (quantidade)



Por fim, olhando individualmente os produtos mais vendidos é visto que condizem com os gráficos mostrados durante esta análise, no ponto que os três

produtos mais vendidos são leites, que entram na categoria dos "laticínios", a mais vendida das categorias. Um ponto importante aqui é que as porcentagens se referem aos 10 produtos em si, não ao total de vendas de todos os produtos.

Essa análise traz algumas informações importantes sobre o comportamento de compra no macromercado. O consumidor históricamente compra mais produtos de carácter diário, ou seja, produtos que irão ser comprados e consumidos durante os próximos dias, como os produtos da categoria "laticínios", “cereais e condimentos” e “FLV”

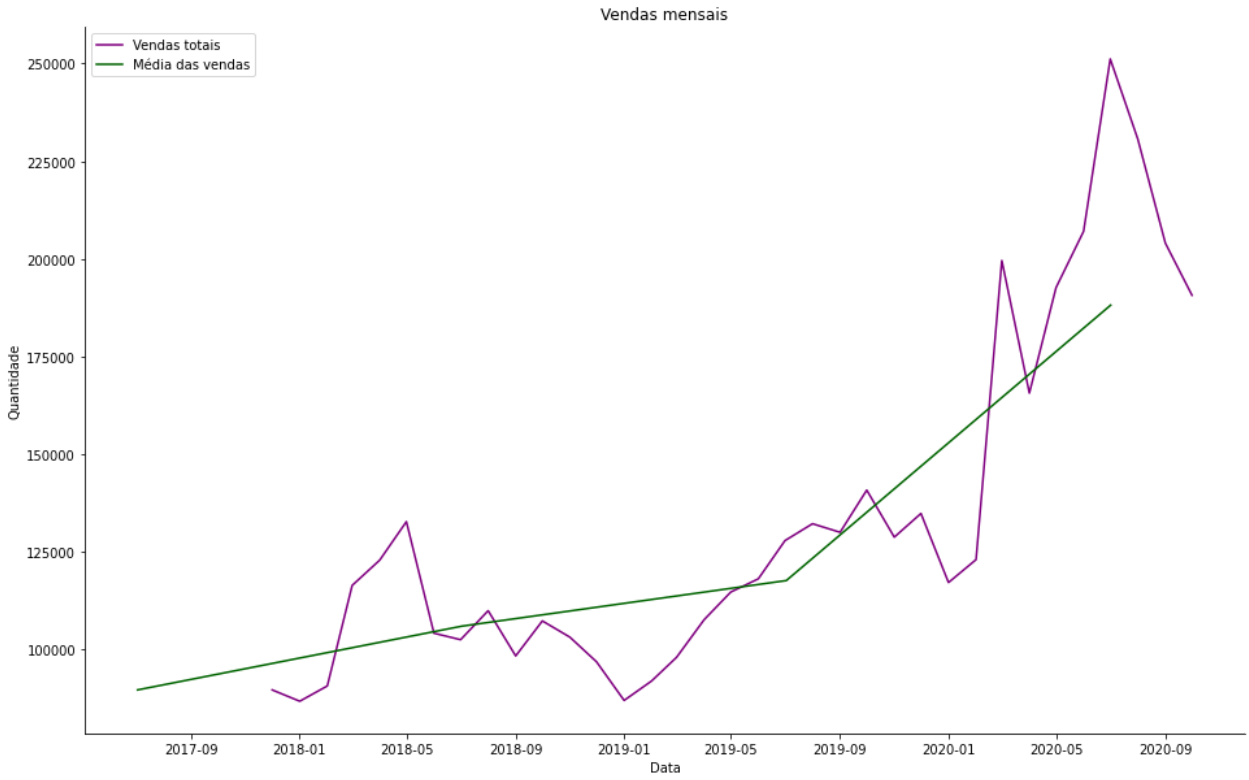
Ele também não se interessa muito por produtos feitos localmente pelo macromercado, visto que as categorias de “confeitaria” e “padaria” não venderam produtos suficientes para entrar no corte final da análise.

Embora os produtos de “açougue” não vendam tanto em questão de quantidade total de produtos vendidos, eles possuem um valor agregado bem elevado, portanto recomenda-se investir em uma maior visibilidade deste produto, tendo em vista que ele é geralmente comprado com temperos e outros produtos para completar a refeição.

5.2 Análise de predição temporal

Para que se realize as próximas etapas desta análise dos dados, é necessário que se realize um teste de estacionariedade e diferenciação do dado. Ele é utilizado para descobrir se a série de dados que se está trabalhando apresenta tendências (ou crescimento) ou não.

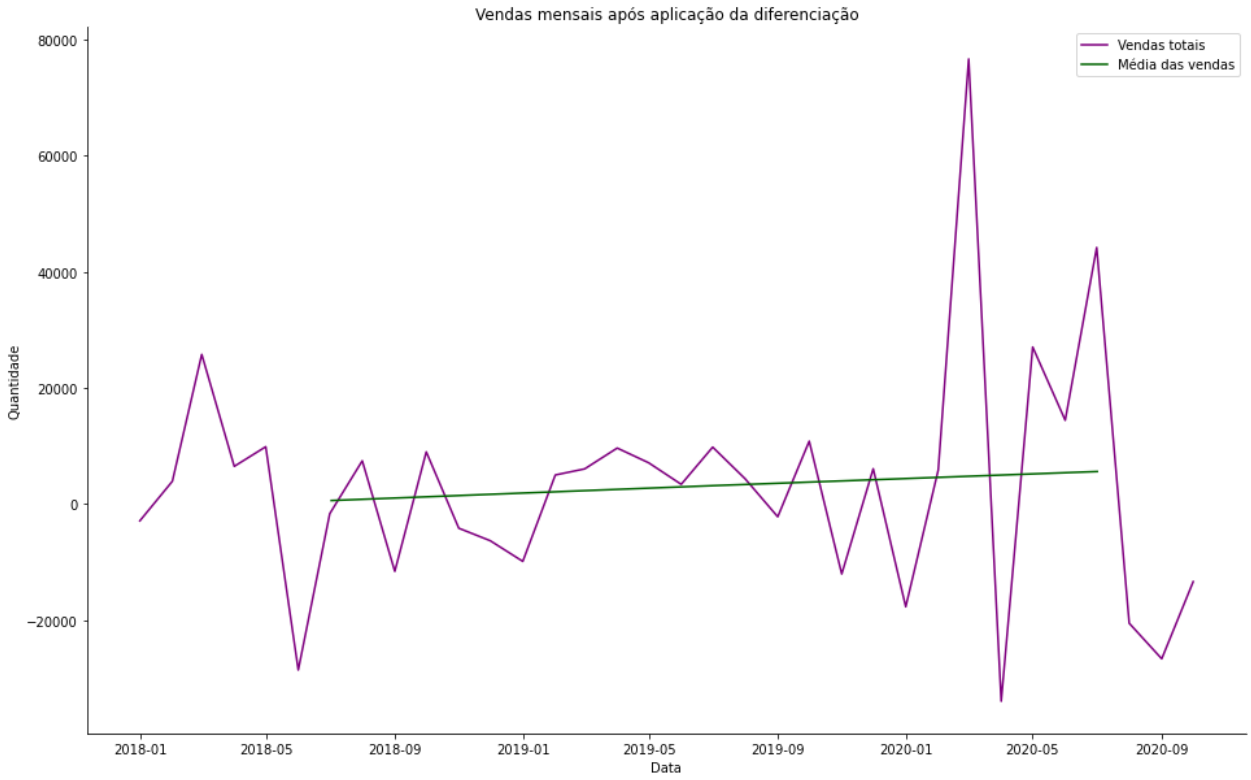
Gráfico 6 - Vendas totais e média móvel de vendas



Com o teste realizado, é evidenciado que a média móvel de vendas ao longo do tempo tem a tendência de crescimento até o segundo semestre de 2019 e após esta data vai aumentando linearmente. Esse comportamento é esperado por dois fatores: se tratar de um novo segmento da empresa, então a base de clientes vai aumentando e o efeito do início da quarentena, chamando especial atenção ao primeiro pico de vendas, logo em março de 2020, período em que foram realizadas as primeiras etapas da quarentena.

No intuito de atingir o padrão de tendência exigido pelo algoritmo de predição temporal, aplica-se uma regra de diferenciação dentro do *dataset*, com o parâmetro de olhar para o mês anterior e considerar a diferença entre os dois meses. Portanto, nesta versão do dado, ao olhar para o mês 2019-09, o dado de quantidade total de vendas que está lá é a diferença entre 2019-08 e 2019-09.

Gráfico 7 - Vendas totais e média móvel de vendas após aplicação da diferenciação



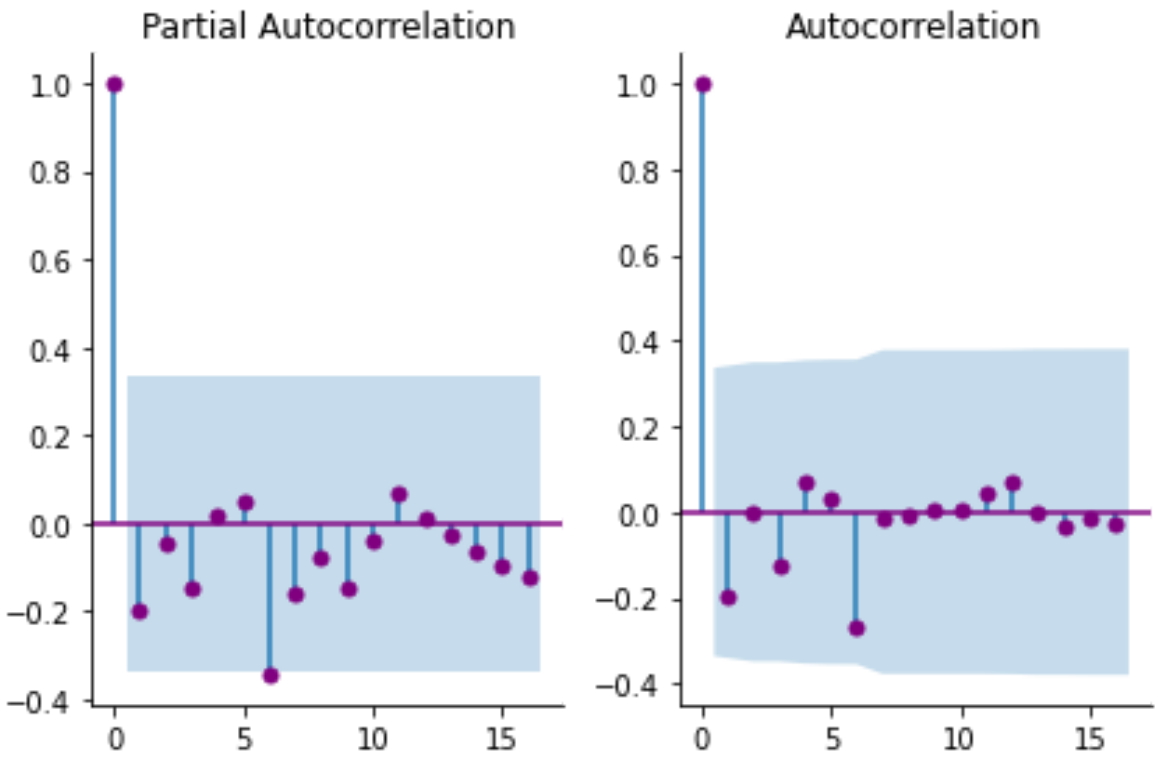
Com a diferenciação aplicada, nota-se que a linha verde está muito mais estacionária, podendo então ser considerada uma base de dados muito mais adequada para se fazer predições temporais. Os valores reais de vendas serão utilizados na parte posterior do algoritmo, esses valores fictícios servirão apenas para o treinamento do algoritmo.

5.3 Seleção do modelo

Existem alguns modelos para predição temporal, nesta presente pesquisa irá se tratar de dois tipos de modelos: os de regressão e *Long Short-Term Memory Networks* (LSTM). Irão ser testados 4 algoritmos dentro destes dois modelos, para que se possa descobrir qual é o mais correto em relação ao dado.

Para auxiliar na escolha do modelo são realizados testes de autocorrelação do dado, segundo os princípios de Brockwell e Davis (2010), que ditam que quando os números de correlação dentro do período escolhido mais se aproximarem do 0, melhor.

Gráfico 8 -Autocorrelação e autocorrelação parcial



Na figura acima é possível visualizar o gráfico de autocorrelação e autocorrelação parcial (o eixo x sendo o coeficiente e o eixo y sendo o período em meses), ambos indicam que se está indo pelo caminho certo, nenhum valor escapa muito dos já previstos picos de venda. Para fins de escolha, escolhe-se usar o *lag* (período) de 12 meses para que possamos fazer uma predição personalizada considerando um período anual.

É importante mencionar que para cada um dos modelos a serem testados serão registrados três características importantes para mensurar a eficácia de cada modelo: erro quadrático médio (RMSE), erro médio absoluto (MAE) e coeficiente de determinação (R2). Estas são as principais características para determinar o melhor

modelo que irá predizer com mais acurácia, todas elas são funções estatísticas calculadas durante a execução de cada modelo.

Para referência, os valores de RMSE e MAE têm que ser baixos e o R2 deverá se aproximar de 1 para ser considerado eficaz.

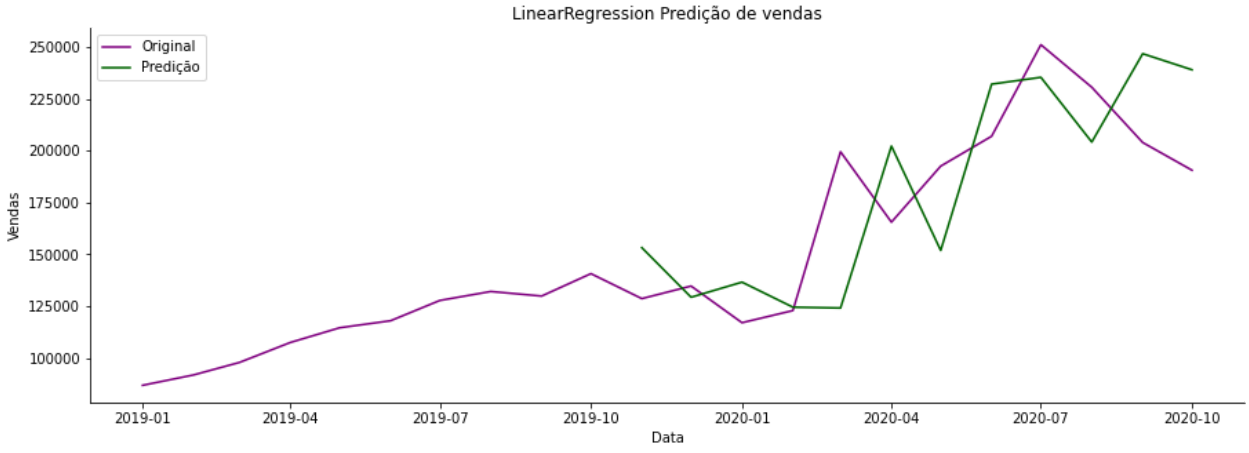
5.3.1 *Algoritmos de regressão*

Estes algoritmos possuem uma característica em comum: todos utilizam alguma técnica de regressão para prever comportamentos futuros. Em princípio, utilizam-se de uma parte dos dados para realizar o treinamento do algoritmo, neste caso será utilizado os últimos 12 meses, conforme constatado pelos coeficientes de autocorrelação.

A ideia aqui é pegar essa parte de dados do treinamento do algoritmo e testar ela contra limites mínimos e máximos das vendas. Testando as duas possibilidades, segundos os parâmetros de cada algoritmo, junta-se os repositórios e cria-se uma predição que tenta se igualar aos últimos 12 meses de dados.

5.3.1.1 Regressão linear

Gráfico 9-Regressão Linear usando quantidade de vendas

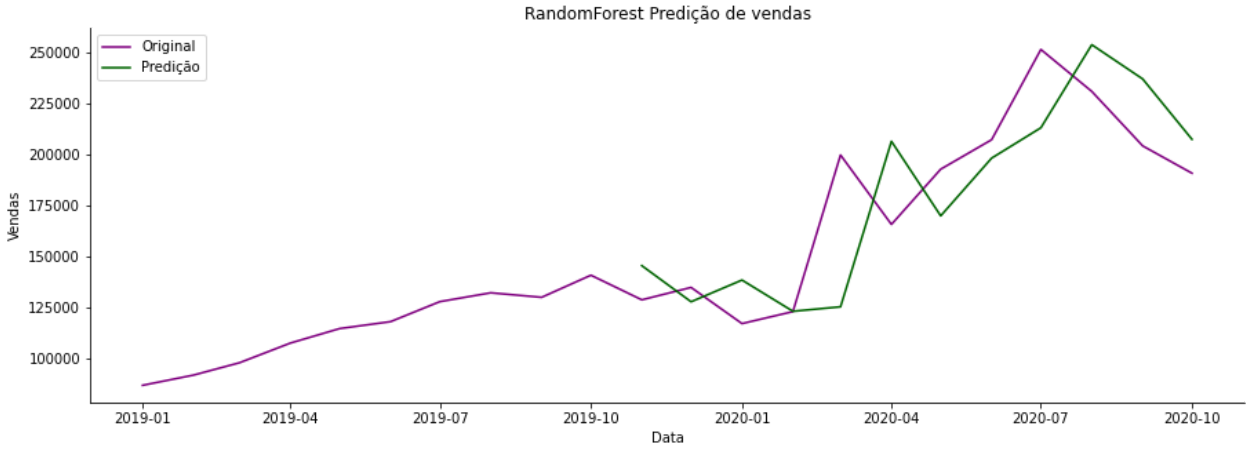


A regressão linear é talvez o mais famoso deles. O algoritmo, assim como seu correspondente matemático, consiste em traçar uma reta entre os possíveis valores do dado original e ver qual predição melhor se encaixa nos valores.

Em sua versão mais básica descrita aqui, o algoritmo alcançou um marco de R2 = 0.28, ou seja, um valor extremamente baixo de acuracidade, como visto no gráfico acima. Suas margens de erros também são grandes, atingindo 35917 para RMSE e 30277 para MAE.

5.3.1.2 Regressão RandomForest

Gráfico 10 -Regressão RandomForest usando quantidade de vendas



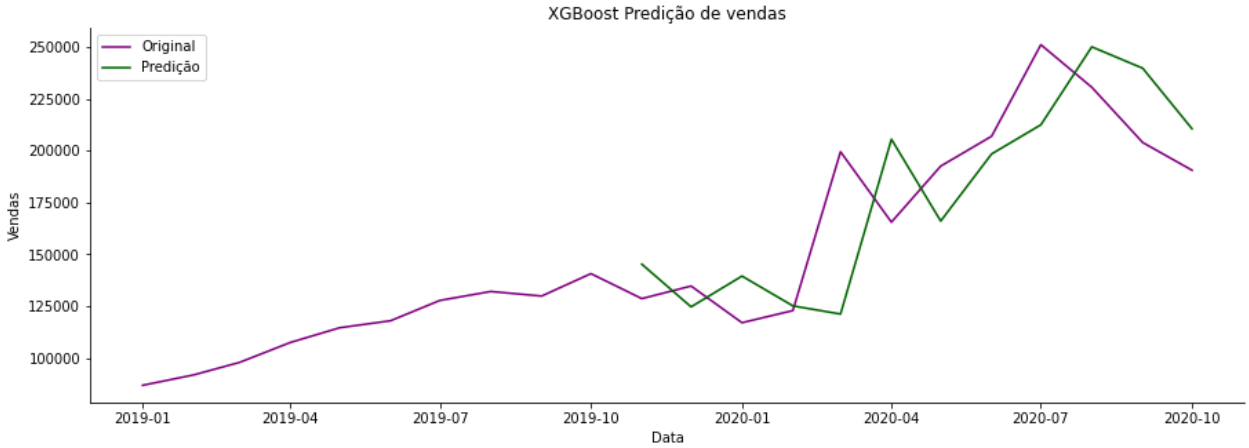
Nesta vertente do algoritmo de regressão, também muito utilizado em algoritmos de clusterização, ao invés de traçar uma reta para achar o melhor valor, são utilizadas "árvores de decisão". Ela também não utiliza a totalidade dos dados de treino, cada árvore será uma parcela dos dados, com cada "galho" ou "nó" da árvore sendo uma decisão entre duas variáveis que são escolhidas aleatoriamente.

Para esse teste inicial, foram escolhidas 100 árvores, cada uma contendo 20 nós, ou seja, os dados de treino foram divididos em 100 pedaços com cada árvore podendo ter 20 decisões em sequência.

Esta versão alcançou um R2 de 0.44, RMSE de 31622 e MAE de 25381. São resultados que estão quase na média de 0.5, um resultado mediano para um algoritmo não otimizado.

5.3.1.3 Regressão XGBoost

Gráfico 11 -Regressão XGBoost usando quantidade de vendas



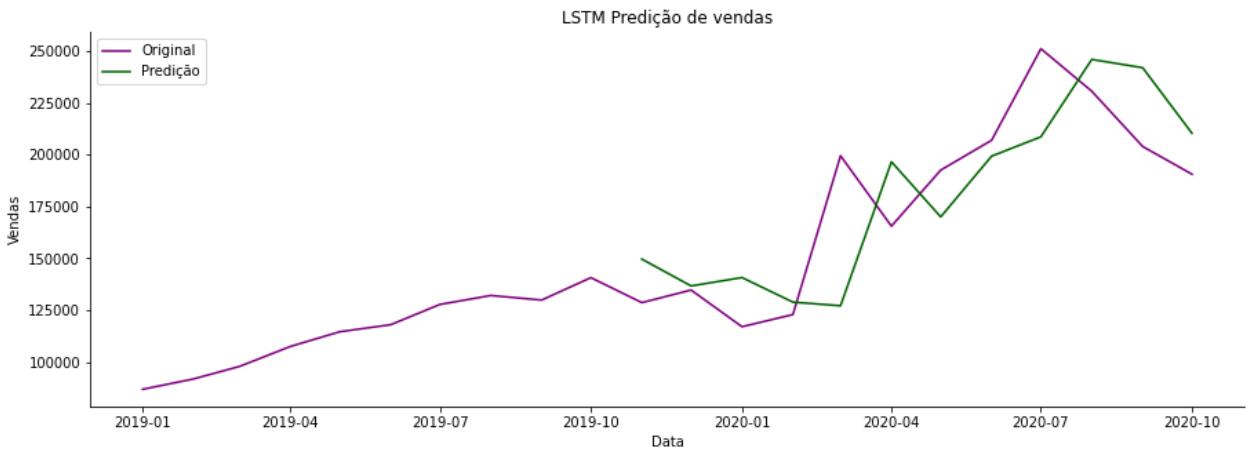
Este usa o mesmo conceito do anterior, porém coloca mais um fator além da aleatoriedade: a tentativa de converter o erro em acerto. Quando ocorre uma decisão na árvore, é calculado a função de erro dela e através deste dado sabemos o quanto aquela decisão influenciou o algoritmo ou não. Fazendo isso repetidas vezes, o algoritmo aprende a fazer alguma função de forma mais eficiente.

O XGBoost entra nesta questão expandindo a função de erro e, ao mesmo tempo, abstraindo partes dela para serem fixas. Deste jeito, sobram apenas duas variáveis que serão consultadas em cada função de erro, otimizando assim o aprendizado.

Mesmo assim, o algoritmo teve um resultado pior que o algoritmo *RandomForest*, com um R2 de 0.4, RMSE de 32868 e MAE de 26573. Parte deste resultado pode ser devido à simplificação do código para fins de teste para escolha do modelo, de forma que ele pode ter sofrido de uma má otimização.

5.3.2 *Algoritmo LSTM*

Gráfico 12 -Rede neural LSTM usando quantidade de vendas



Este é um algoritmo bem diferente dos anteriores. Uma rede neural focada não em regressão, mas recorrência. Ela possui marcos temporais, chamadas de *epoch*, sendo que cada *epoch* roda o algoritmo inteiro em cada uma das suas execuções.

Ao rodar um algoritmo de regressão, o resultado será a soma das execuções. Aqui, cada marco temporal é treinado, executado e aprende recorrentemente, ou seja, cada execução passa a informação que aprendeu para o *epoch* seguinte, portanto o algoritmo não começa a rodar sem informação, ela é sempre um acúmulo das execuções passadas.

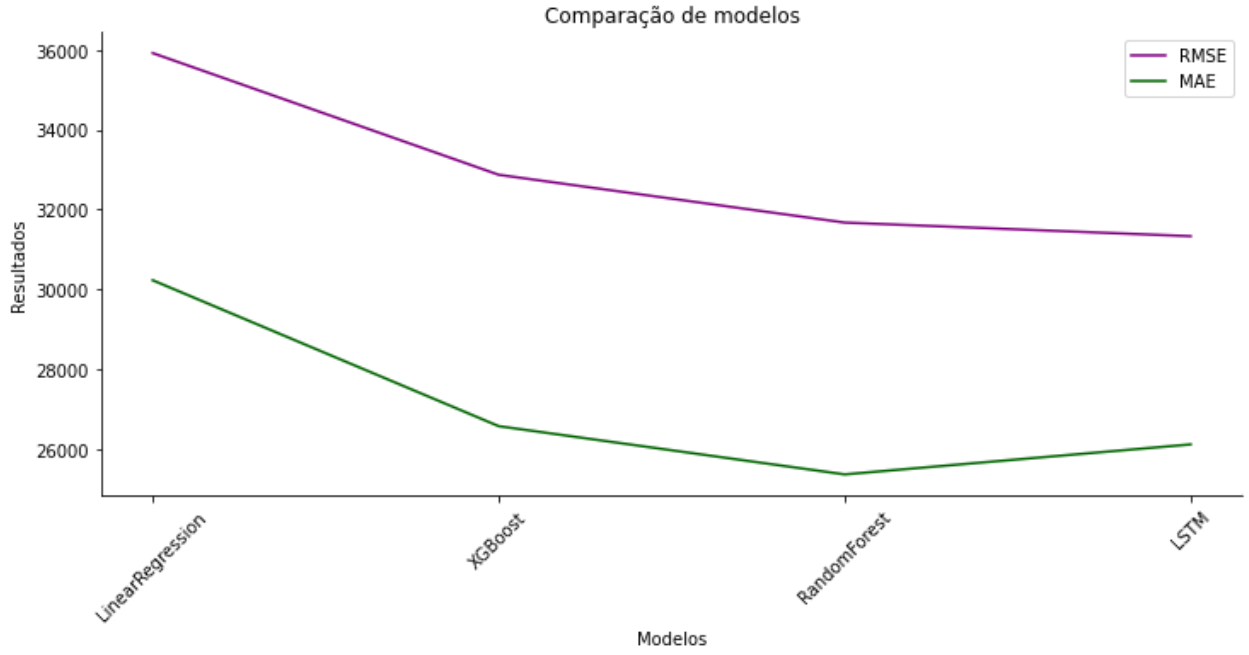
Ele foi pensado para rodar como uma rede neural com memória, daí o nome *Long-Short Term memory*, em português, memória de longo-curto prazo. Por ser executado com *feedbacks* constantes, a sua *loss function* decai com o tempo, melhorando a acuracidade do modelo a cada *loop*.

Sua desvantagem é o taxamento intensivo dos recursos do sistema, para que se possa fazer uma análise adequada com este algoritmo é necessário uma quantidade considerável de CPU e RAM, aumentando exponencialmente conforme a quantidade de linhas no seu repositório de dados.

Para este teste, foram utilizados como parâmetros 200 *epochs*, o que equivale ao mesmo número de execuções no algoritmo, obtendo um R2 de 0.46, um RMSE de 31223 e MAE de 25172.

5.4 Escolha do modelo

Gráfico 13 - Comparação de modelos



Dentre os modelos propostos, os dois que mais performaram foram o *RandomForest*, um algoritmo de regressão e o LSTM, uma rede neural recorrente. Como o próximo passo é tentar melhorar o algoritmo, é escolhido o LSTM pois é o que mais se aproximou na métrica de acerto, o R2.

5.5 Otimização do algoritmo

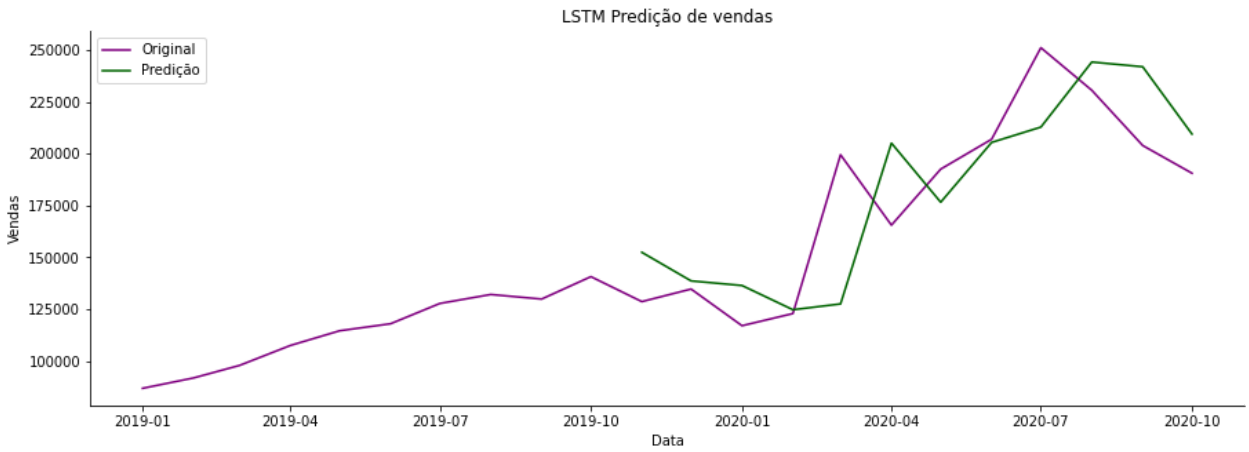
O LSTM possui alguns princípios a mais do que foi escrito acima. Além dos citados na apresentação dele, ainda há a questão de neurônios e tamanhos do teste que são passíveis de mudança para tentar deixar o algoritmo mais fiel ao dado original. O teste original foi feito com 4 neurônios, processadores da decisão, e um tamanho do teste como 1.

Em uma série de testes, foi encontrado um parâmetro que deixou o algoritmo um pouco mais fiel ao dado original, sendo este de usar apenas um neurônio e o tamanho de teste como 2.

Como há uma grande discrepância entre o dado pré-quarentena e durante, deixar as escolhas para mais que um neurônio faz com que sejam cometidos mais erros durante o processamento, fazendo assim com que a informação falsa seja passada adiante. Ao limitar o processamento, aumenta-se a chance de acerto e que a informação certa seja aceita.

Aumentou-se também o número de *epochs* para 2000, o que é um número razoável de execuções tendo em vista a diminuição nos neurônios do algoritmo.

Gráfico 14 - Otimização da rede neural LSTM



Com isso, obtém-se o resultado final do algoritmo otimizado, com um aumento no R2 para 0.50, totalizando o coeficiente de determinação em 50%. Não é um resultado satisfatório para qualquer predição real de comportamento das demandas, sendo possível apenas identificar certas tendências de crescimento.

5.6 Resultados finais

Durante o trabalho foram encontrados diversos obstáculos quanto à obtenção dos dados corretos perante a empresa estudada. Este desencontro entre os dados

fornecidos foi o fator determinante em poder aplicar apenas a análise de predição temporal, que depende de uma série temporal em que se possa ao menos mensurar o quanto o efeito sazonal tem perante o dado, o que não foi possível devido à ser uma vertente recente dentro de uma empresa e a quarentena imposta perante a nova pandemia do Coronavírus.

Dado esse viés, não foi possível explorar os dados da forma desejada, pois ocorreram variações de demanda demasiadas grandes, sendo o algoritmo mais otimizado chegando a apenas 0.5 de coeficiente de determinação, sendo impossível o uso dele para predizer qualquer comportamento real.

Porém, foram retirados *insights* importantes quando da realização da análise geral dos dados, percebendo a importância de categorias como “material de limpeza” e “laticínios”, ambas categorias que vendem bastante mas não possuem um valor agregado muito alto.

Uma outra categoria que chamou a atenção foi “açougue”. Os produtos nela contidos possuem um valor elevado, então mesmo com uma quantia baixa de compra unitária, é uma das categorias que mais traz receita ao macromercado. Há também de se levar em consideração que geralmente carnes e derivados são comprados juntos com outros produtos, o que leva a um aumento no preço unitário médio do cliente.

6 CONCLUSÃO

Este trabalho apresentou uma aplicação prática de algoritmos de predição de demanda com o objetivo validar o apoio à tomada de decisão com dados para a estratégias de negócio de macromercados através da aplicação de *Data Science* e algoritmos de inteligência artificial.

O estudo foi realizado junto a um macromercado, mais especificamente do seu setor de delivery, localizado no interior do Rio Grande do Sul. Os resultados demonstraram que os dados disponibilizados pelo mercado não validaram a aplicação de predição satisfatoriamente, visto que não foi possível prever um comportamento de demanda considerando uma série de tempo pequena e os poucos parâmetros recebidos.

Porém, embora a predição temporal esteja fora do alcance da empresa estudada, é possível ser realizada uma análise de carrinho sobre os produtos vendidos, utilizando então algoritmos de clusterização que não sofrem com variações de demandas grandes como foi visualizado neste período.

Data science como um todo pode ajudar em muito a empresa a crescer em seu futuro. São diversos tipos de análises que podem ser feitas se houver um time dedicado a isso, podendo trazer um crescimento muito grande à empresa.

Durante a tratativa com o macromercado, percebeu-se inclusive que seu sistema de TI não é tão robusto. Embora tenham contratado a TOTVS, uma das maiores empresas de tecnologia no brasil, para utilizar seu *software* de vendas, o time em questão parece não ter muito conhecimento sobre seu próprio sistema, o que acarreta em análises defasadas dos dados.

Caso a empresa queira melhorar esta situação, é recomendado a contratação de profissionais especializados em lidar com este tipo de análise e neste tipo de *software*, pois os dados estão sendo captados de acordo com o sistema escolhido, já sendo possível realizar as análises de clusterização que não foi dado acesso devido ao que foi falado acima.

Estudar a forma como o cliente se comporta é fundamental para a empresa poder prosperar, ainda mais num ambiente competitivo e de grande alcance como o

e-commerce. Este deve ser um processo internalizado pela empresa, pois será um grande diferencial competitivo quando a competição acirrar.

Ter o poder de sugerir produtos relacionados ao carrinho do usuário, pode levar a compras assistidas, como no caso dos produtos da categoria de “açougue”. Tendo isto em mente, o comportamento do *delivery* deveria ser de quando o usuário coloca um destes produtos no carrinho, automaticamente sugerir outros produtos complementares, já catalogados pelo algoritmo de clusterização.

Por fim, melhorar a experiência do usuário e fidelizar o cliente através de algoritmos deve ser explorado, argumentado e implementado, especialmente no seu viés de *delivery*. Conforme mencionado, o macromercado possui sua loja física e tem o *delivery* como uma forma de complementar seu alcance na cidade, mas acredita-se que com a contratação de profissionais que tenham conhecimento em *data science* e após a aplicação dos algoritmos mencionados aqui, este formato pode vir a ser o principal desta empresa, devido à facilidade de expansão e escalonamento que os algoritmos provêm.

REFERÊNCIAS

BADIN, Neiva. **Avaliação da produtividade de supermercados e seu benchmarking**. Universidade Federal de Santa Catarina, 1997. Disponível em: https://repositorio.ufsc.br/xmlui/handle/123456789/77089. Acesso em 10 de nov. 2021.

BARROS, Carlos. **Ferramental matemático e computacional para apoio a gestão de pequenos supermercados.** São Paulo: Universidade de São Paulo, 2020. Disponível em: https://tinyurl.com/43h67wzk. Acesso em 25 de abr. 2021.

BNDES - BANCO NACIONAL DE DESENVOLVIMENTO ECONÔMICO E SOCIAL. **Comércio varejista.** Rio de Janeiro: Banco Nacional de Desenvolvimento Econômico e Social, 1999. Disponível em: https://tinyurl.com/5vrpuhwc. Acesso em 16 fev. 2021.

BROCKWELL & DAVIS. **Introduction to Time Series and Forecasting**. Harrisonbourg, Virgínia: RR Donnely and sons, 2002. Versão ebook. Disponível em: http://home.iitj.ac.in/~parmod/document/introduction%20time%20series.pdf. Acesso em 06 de nov. de 2021.

DIETRICH, et. al. **Data Science & Big Data Analytics**. Indianápolis, Indianápolis: EMC Education Services, 2015. Disponível em: http://home.iitj.ac.in/~parmod/document/introduction%20time%20series.pdf. Acesso em 08 de nov. 2021.

ELLER, David. **Previsão de demanda [...]**. Florianópolis: UFSC, 2020. Disponível em https://repositorio.ufsc.br/bitstream/handle/123456789/218772/TCC_DAVID_TELES_ELLER.pdf. Acesso em 18 maio 2021.

FROST & SULLIVAN. **Latin American Big Data and Analytics Market, Forecast to 2023**. [S.l]: Frost & Sullivan, 2018. Disponível em https://tinyurl.com/ecmaj4t9. Acesso em 18 maio 2021.

Folha UOL. **Redes de varejo apostam em big data para atrair consumidores.** Disponível em: https://www1.folha.uol.com.br/mercado/2018/01/1954427-redes-de-varejo-apostam-em-big-data-para-atrair-consumidores.shtml. Acesso em 08 de nov. de 2021

IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Pesquisa mensal do comércio**. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística, c2021. Disponível em: https://tinyurl.com/85z4wrn3. Acesso em 16 fev. 2021.

IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Panorama das cidades.** Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística, c2021. Disponível em: https://cidades.ibge.gov.br/brasil/rs/pelotas/panorama. Acesso em 08 de nov. 2021.

KARNA. **Retail Shelf Monitoring for Perfect In-Store Execution.** [S.l]: Karna, c2019. Disponível em: https://tinyurl.com/w8vu29j4. Acesso em 24 de abr. 2021

KOTLER, Philip. **Marketing 3.0: as forças que estão definindo o novo marketing centrado no ser humano** . Rio de Janeiro: Elsevier, 2010.

_____. **Marketing 4.0: do tradicional ao digital**. Rio de Janeiro: Elsevier, 2017.

LAVRADO, Fernando. **BPM e Transformação digital.** São Paulo: Seminário em administração, nov. 2019. Disponível em: https://tinyurl.com/22wj7wkk. Acesso em 22 de mar. 2021.

MITCHELL, Tom Michael. **Machine Learning: A guide to current research**. Nova Iorque, Estados Unidos: McGraw-Hill, 1977.

PROVOST, Frost e FAWCETT, TOM. **Data Science Para Negócios: O que Você Precisa Saber Sobre Mineração de Dados e Pensamento Analítico de Dados**. Alta Books, 2016.

REZENDE, Denis. **Sistemas de Informações Organizacionais: Guia Prático para Projetos**. Barueri, SP: Editora Atlas, 2005.

ROGERS, David. **Transformação Digital: Repensando o seu negócio para a era digital.** Tradução: Afonso Celso da Cunha Serra. Autêntica Editora, 1. ed; 3ª reimp, São Paulo, 2020.

SAMUEL, Arthur Lee. Some Studies in Machine Learning Using the Game of Checkers. **IBM Journal of Research and Development**, [S.l.] vol. 3, no. 3, p. 210-229, jul. 1959.

SBVC - SOCIEDADE BRASILEIRA DE VAREJO E CONSUMO. **O Papel do varejo na economia Brasileira**. São Paulo: Sociedade brasileira de varejo e consumo, 2018. Disponível em: https://tinyurl.com/b3x5tkxa. Acesso em 03 de mar. 2021.

SCIKIT. **Support Vector Machines**. Scikit, 2021. Disponível em https://scikit-learn.org/0.24/modules/svm.html#regression. Acesso em 01 de nov. 2021.

Stern Speakers. **Anticipating trends planning for the future**. Disponível em: https://sternspeakers.com/portfolio/anticiapting-trends-planning-for-the-future/ Acesso em 04 de nov. de 2021

Tera. **Dados e Zara em relacionamentos**.Disponível em: https://medium.com/somos-tera/dados-e-zara-em-relacionamento-s%C3%A9rio-2bfd943b6986. Acesso em 08 de nov. de 2021

UBEROI, Ravneet. **ZARA: Achieving the “Fast” in Fast Fashion through Analytics.** Cambridge, Estados Unidos: Digital Innovation and Transformation, 2017. Disponível em: https://tinyurl.com/3ft52hff. Acesso em 04 de maio de 2021.

VEBLEN, Thorstein. **The Theory of The Leisure Class**. Oxônia, Reino Unido: Oxford University Press, edição reimpressa, 2009.

ZOROTOVICH, Mariya. **Current use cases for machine learning in retail and consumer goods.** Seattle, Estados Unidos: Microsoft Azure, 2018. Disponível em: https://tinyurl.com/m59h6fez . Acesso em 02 de mar. 2021.