**Otimizando o abastecimento de produtos no varejo com Random Forest**

Guilherme Lima de Sousa¹\*; Vanilde de Castro2

1 Rua Vicente Carducho, 249 – Vila Remo. 05864220. São Paulo – SP – Brasil

2 Prof.ª Dr.ª Orientadora- Convidada - "MBA Data Science e Analytics" ESALQ /USP - Piracicaba – SP- Brasil.

\*autor correspondente: guilhermelima.edsp@hotmail.com

**Otimizando o abastecimento de produtos no varejo com Random Forest**

**Resumo**

Ao passo que atender os seus clientes com agilidade se torna um objetivo de uma empresa, estocar os seus materiais em pontos estratégicos para distribuição torna-se parte fundamentalmente importante neste processo. Neste trabalho foi abordado recursos e métodos para criação de um modelo de machine learning randomforest, para classificação de produtos que possuem risco de ruptura (falta de produto) ou risco de alta cobertura (produto parado), frente a um cenário simulado de vendas. Foram feitas três simulações de cenários, ou seja, três bases de dados afim de aplicar o modelo, bem como ajustes nos parâmetros dos modelos e também as suas validações. Sendo um problema multiclasse, o modelo apresentou bons resultados, quando aplicado o ajuste no balanceamento das classes, a acurácia cai -2,64%, mas este comportamento reflete sobre o quanto os dados desbalanceados podem induzir e o como o ajuste se faz necessário. Por fim, foi exibido os resultados desta classificação dos produtos, realizado a leitura destes resultados e foi sugerida a tomada de decisão.

**Palavras-chave:** Distribuição; Parâmetros; Machine Learning.

**Introdução**

A atividade econômica conhecida por Varejo é citada na obra de Varotto (2006) como estando presente no Brasil desde a época colonial, chegando a representar um terço do Produto Interno Bruto [PIB] do ano de 2006. É uma das principais fontes geradoras de emprego, aliado à diversas tecnologias aplicadas nas rotinas das empresas e das pessoas, o potencial a ser descoberto neste ramo de atividade é ainda desconhecido, pois todo ano temos inovações e evoluções em toda a sua estrutura, tendo algumas vertentes mais inovações do que outras.

Com a acirrada competitividade entre os grandes players do mercado varejista, um dos principais objetivos das empresas é fornecer o produto certo no momento certo e para realizar isso a melhoria de suas estruturas de abastecimento e distribuição tornam-se uma pauta recorrente nas áreas de logística. O excesso de materiais ou a falta deles, nas lojas, impacta diretamente os resultados da empresa, sejam eles fiscais ou comerciais e considerando uma empresa varejista de produtos eletrônicos, esses impactos são ainda maiores.

Varotto (2006) cita sobre a estratégia comercial utilizada pelas empresas que oferecem produtos acessíveis para os mercados de baixa renda, com objetivo de impulsionar a demanda e consequentemente o crescimento econômico. Essas empresas entendem a potencialidade que as grandes massas possuem no âmbito aquisitivo e exploram suas possibilidades oferecendo produtos tecnológicos (entre outros) que competem um curto ciclo de vida, fazendo-se necessário um upgrade a cada período de tempo.

Aliado a temática do produto certo no momento certo, em sua pesquisa de doutorado Gianesi e Biazzi (2009), comentam o fato de serem recentes as tecnologias para o tratamento individualizado dos materiais de um estoque e o quanto isso impacta algumas operações em empresas.

O difícil acesso a profissionais ou tecnologias de logística diante essas empresas se refletiu de tal forma em que houvesse uma normalização dos produtos dentro do estoque, gerando cálculos agrupados, sem nenhuma individualidade (Gianesi e Biazzi,2009).Individualidade essa que é observada diante da sazonalidade com as datas comemorativas, cujo alguns produtos podem ser mais procurados do que outros.

Ao passo que atender os seus clientes com agilidade se torna um objetivo de uma empresa, estocar os seus materiais em pontos estratégicos para distribuição torna-se parte fundamentalmente importante neste processo. No tocante ao cenário de uma empresa no ramo de produtos eletrônicos, os canais de vendas físicas não são os únicos meios para comunicar com o cliente, existindo outros canais como o: Business to Business [B2B], Business to Consumer [B2C], Vendas eletrônicas [Online] que permitem o atendimento e satisfação de seus clientes por diferentes meios (Cunha, 2023).

Sendo assim tratar os gargalos de abastecimento, aqueles onde existe o excesso de produtos concentrados em lojas, pode ocasionar uma vantagem competitiva no mercado.

Então, quando concentrado estes produtos nos pontos adequados geograficamente cria-se oportunidade aos demais canais de vendas.

Branquinho Filho (2020) salienta a importância de uma cultura baseada em dados que promove a utilização de modelos de machine learning, contribuindo para as análises da influência que as operações de logística tem sob a satisfação e atendimento dos clientes. Utilizando esses modelos para otimizar os processos da cadeia de abastecimento e garantindo maior acurácia na disponibilidade dos produtos.

Neste trabalho foi apresentado uma análise exploratória dos dados, utilizando algoritmo de previsão RandomForest (Oshiro 2013), com o propósito de verificar a demanda das lojas, prevendo à possibilidade do excesso ou ruptura de produtos e posteriormente propondo um modelo matemático que permita um abastecimento mais assertivo (a coluna de necessidade da base de treino).

Alinhando a disponibilidade de produtos nas lojas para que elas tenham autonomia de vendas, apoiando o processo de distribuição dos produtos e permitindo a concentração destes produtos nos locais estratégicos da empresa, os centros de distribuição [CD].

**Material e Métodos**

A obtenção dos dados para análise exploratória foi por meio de base de dados, criada similar à base empresarial, cujo nome das colunas são os mesmos, mas com valores randômicos entre 0-270 para a coluna quantidade de vendas, entre 0-127 para a coluna de estoque das lojas e por fim, entre 0-10 para a coluna estoque mínimo. Definindo a semente de reprodução, utilizando a biblioteca Numpy na função np.random.seed(), foi garantindo que os resultados dos modelos sejam replicados.

A base possui informações de 20 lojas e suas respectivas demandas ao longo de Janeiro de 2020 até Novembro de 2023, nela está disponível informações como: a quantidade de vendas realizadas, os números de identificação das lojas, números de identificação do produto, ano e mês, quantidade de estoque atual das lojas, quantidade de estoque mínimo necessário de produto, as siglas do estado onde a loja reside, os nomes dos produtos, categoria dos produtos, um status para agrupar os cenários onde a demanda é maior que o estoque, o giro *(resultado do cálculo vendas/estoque\_total)*, a cobertura *(resultado do cálculo estoque\_total/vendas\*30)* e também o nome das lojas. Todo o processo de ETL foi realizado na construção da base, então não existem valores nulos ou infinitos.

O modelo escolhido para realização do objetivo proposto é RandomForest, pois o método Ensemble que é derivado da teoria das multidões, agrupa os preditores (Géron) tornando o trabalho deste grupo superior ao de uma variável preditora individual. O que produz um efeito de maior assertividade do modelo, acurácia.

A tabela a seguir apresenta as variáveis:

|  |  |
| --- | --- |
| Tabela 1. Variáveis explicativas da base de treino do modelo | |
| Variável | Descrição |
| Mês | Mês referente aos dados |
| Ano | Ano referente aos dados |
| SKU | Código identificação do produto |
| Centro | Código identificação da loja |
| Qtd\_estoque\_llpp | Quantidade de estoque atual da loja |
| Qtd\_vendas | Quantidade de vendas realizadas |
| Giro\_llpp | O quanto a loja transformou o estoque em vendas |
| Cobertura\_llpp | O quanto a loja segura a demanda com o estoque atual |
| Sazonalidade | Binário que diz se um produto tem sazonalidade 1 sim ou 0 não |
| Estoque\_minimo\_llpp | Estoque mínimo necessário de cada produto |
| Estoque\_ideal | Estoque necessário para obtermos um giro de 60% |
| Necessidade | O quanto eu preciso abastecer ou remover da loja |
| Status | Flag para dizer se a necessidade é: 0(<0), 1(<16), 2(>=90), 3(>=16e<90) |
| Fonte: Dados originais da pesquisa |  |

Foi utilizado o modelo de RandomForestClassifier da biblioteca sklearn, por meio da linguagem Python com a IDE Jupyter notebook para prever as informações da base de dados, separados 70% dos dados para treino e 30% para teste, a validação do modelo foi por meio da acurácia e da validação cruzada com 5 fold, além das verificações de CURVA ROC e F1-Score. Para o treinamento do modelo foi utilizado apenas uma parte do conjunto de dados: Mês, Ano, SKU, Centro, Qtd\_estoque\_llpp, Qtd\_vendas, Giro\_llpp, Cobertura\_llpp, Sazonalidade, Estoque\_minimo\_llpp, Estoque\_ideal, Status e Necessidade.

A variável target foi a coluna Status e as demais informações, as features que contribuíram com a previsão. Todos os materiais do estudo podem ser encontrados no meu perfil do GITHUB: https://github.com/iugamil/TCC\_MBA\_USP\_ESALQ\_RANDOMFORESTMULTICLASSE.git

.

**Resultados e Discussão**

São apresentados os resultados preliminares do trabalho, onde foram criados 2 modelos com parâmetros diferentes no quesito de balanceamento de classes. Para verificar os resultados de cada modelo, foi simulado 3 cenários diferentes de vendas, promovendo maior aplicabilidade ao mundo real.

Cada base, destes novos cenários, foi criado a partir de dados do ano de 2022 que já estão dentro da base raiz (usada para treinar o modelo). Sendo necessário o recalculo da maioria das colunas, refazendo a base, tais como: ANO (atribuído 2025), as colunas calculadas como o GIRO\_LLPP, COBERTURA\_LLPP, além de outras.

Abaixo um resumo de cada cenário onde os modelos serão aplicados:

**Cenário um** = As vendas de alguns produtos são 15% menores em alguns meses e anos.

**Cenário dois** = As vendas de alguns produtos são 35% menores em alguns meses e anos.

**Cenário três** = As vendas de alguns produtos são 10% menores em meses sazonais e 55% maiores nos demais meses.

Os meses definidos como sazonais são: Janeiro, Fevereiro, Maio, Junho, Agosto, Outubro, Novembro, Dezembro.

Os produtos escolhidos são: 9511-SAMSUNG GALAXY S22, 9471-IPHONE 13, 9611-MACBOOK PRO, 9681-SAMSUNG GALAXY BUDS PRO.

No tocante ao modelo 1 (sem balanceamento de classes) o atendimento dos pressupostos é positivo, pois o modelo não demostra ocorrência de overfitting (acurácia distoa pouco entre treino e teste), a validação cruzada retorna acurácias próximas entre si, a matriz de confusão retorna uma boa assertividade nas predições e por fim a CURVA ROC tem o número 1 para todas as classes.

O único ponto de atenção é a dificuldade do modelo prever a classe número 1, pode-se observar na medida F1-Score que é a menor classe pontuada. Evidências abaixo:

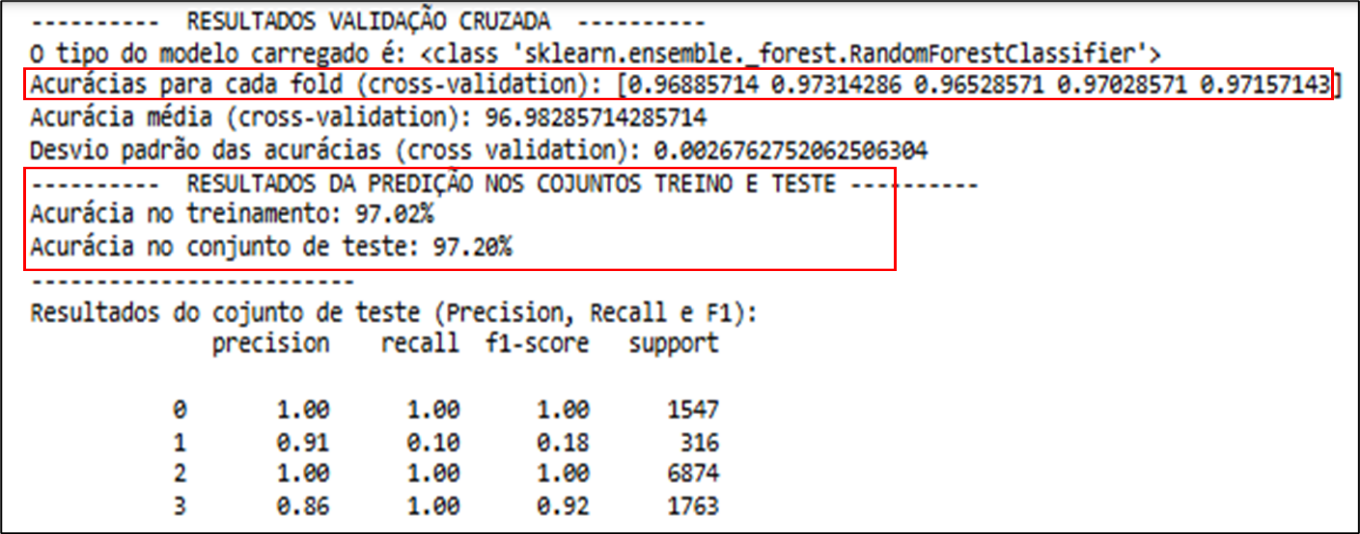


Figura 1. Resultados do treinamento do modelo sem balanceamento de classe.

Fonte: Dados originais da pesquisa.

Ainda sobre o modelo 1 (sem balanceamento de classes) a AUC foi visto que todas as classes tiveram um bom desempenho com resultados iguais a 1, ou seja, o modelo consegue fazer boas predições e distinguir bem cada classe.

Continuando a leitura das evidências do modelo 1 (sem balanceamento):

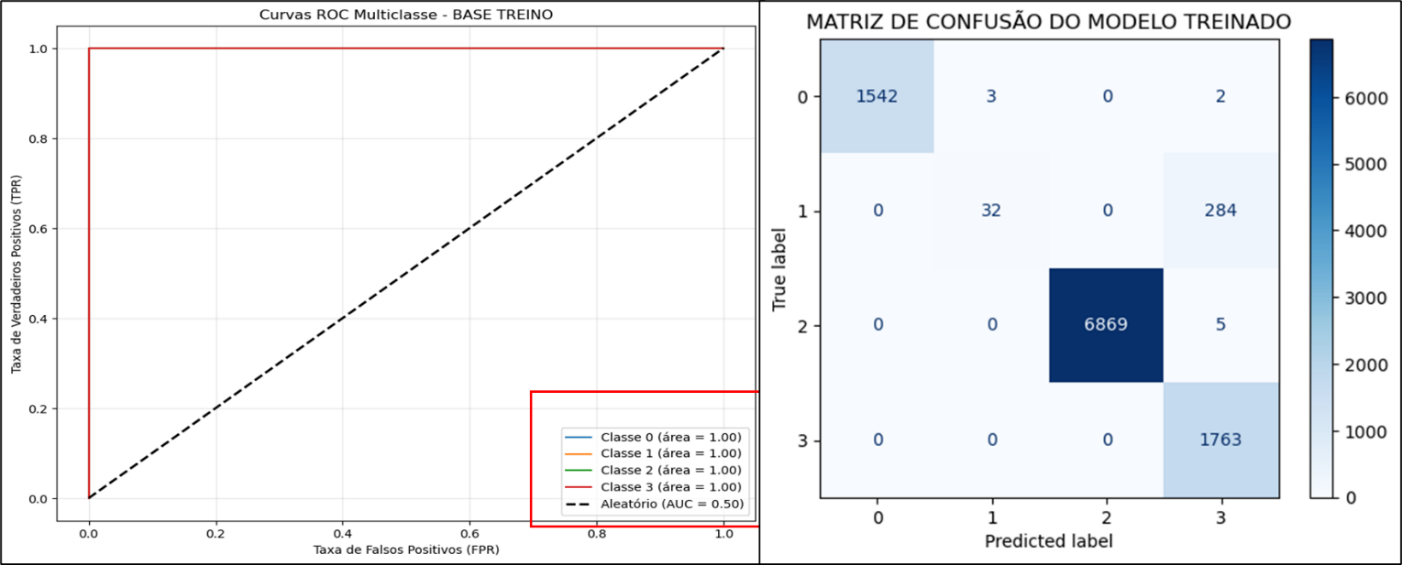


Figura 2. Resultados do treinamento do modelo sem balanceamento de classe.

Fonte: Dados originais da pesquisa.

No tocante ao modelo 2 (com balanceamento de classes) o atendimento dos pressupostos continua positivo, pois o modelo não demostra ocorrência de overfitting (acurácia distoa pouco entre treino e teste), a validação cruzada retorna acurácias próximas entre si, a matriz de confusão retorna uma boa assertividade nas predições e por fim a CURVA ROC tem o número 1 para todas as classes.

O único ponto de atenção é a dificuldade do modelo prever a classe número 3, pode-se observar na medida F1-Score que é a menor classe pontuada. Evidências abaixo:

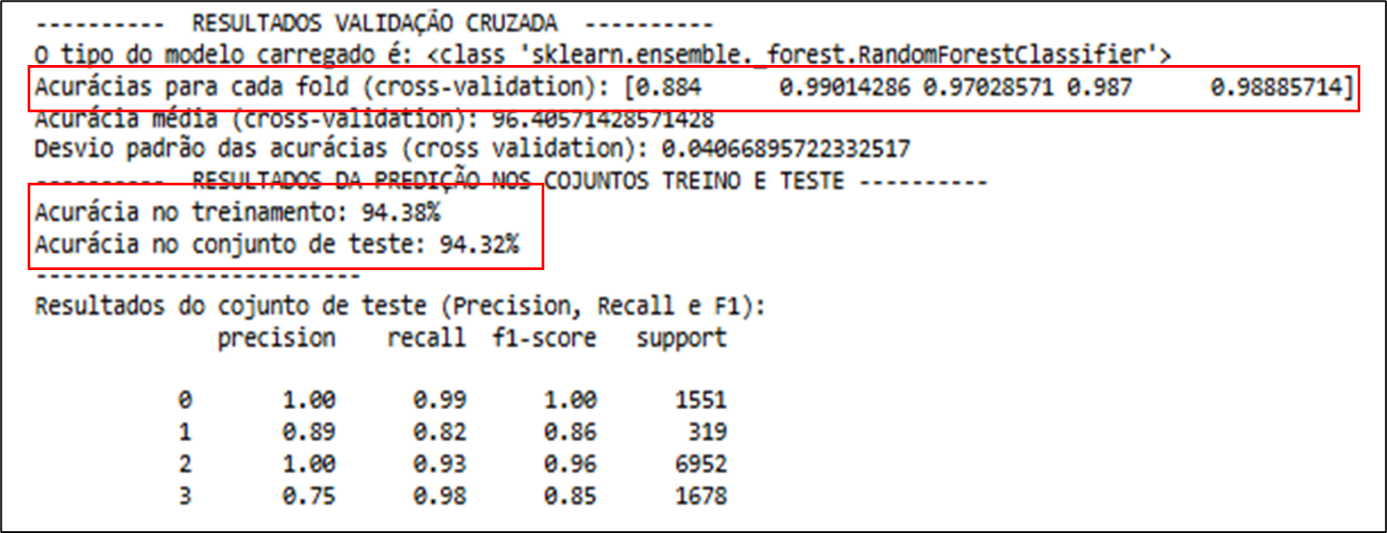


Figura 3. Resultados do treinamento do modelo com balanceamento de classe.

Fonte: Dados originais da pesquisa.

Ainda sobre o modelo 2 (com balanceamento de classes) a AUC foi visto que todas as classes tiveram um bom desempenho com resultados iguais a 1, ou seja, o modelo consegue fazer boas predições e distinguir bem cada classe.

Continuando a leitura das evidências do modelo 2 (com balanceamento):

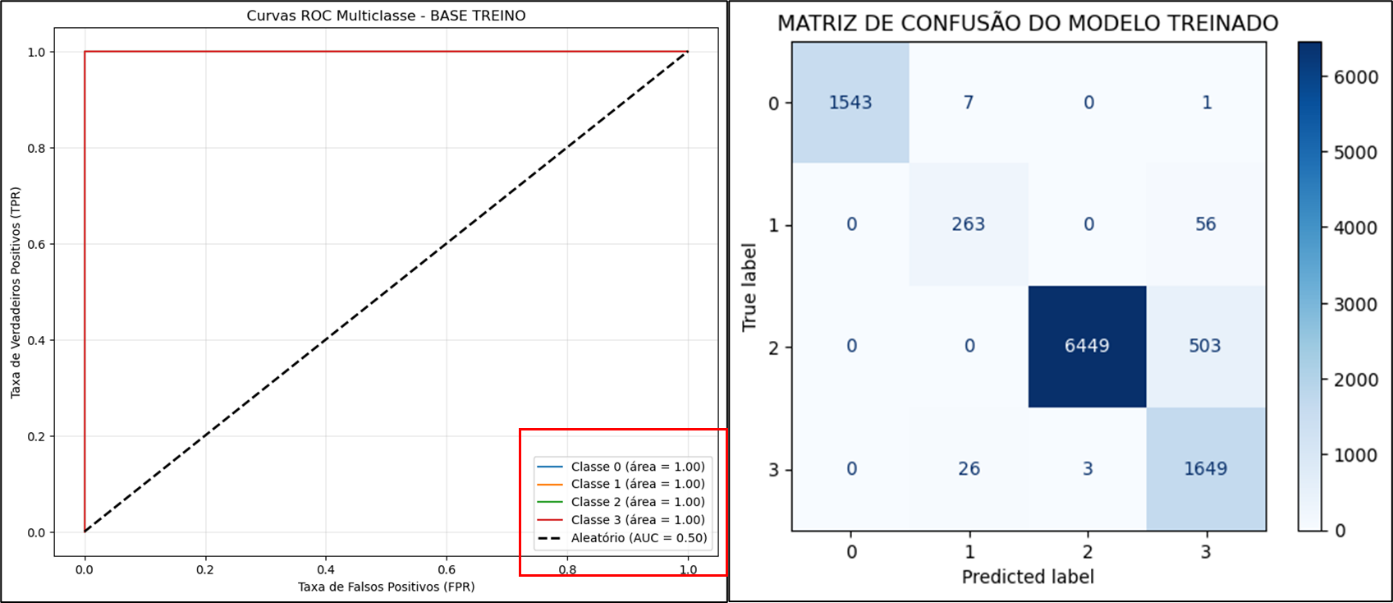


Figura 4. Resultados do treinamento do modelo com balanceamento de classe.

Fonte: Dados originais da pesquisa.

Neste ponto, os modelos já estão treinados e posteriormente, aplicando cada modelo, em três cenários de vendas diferentes.

No modelo número um, na parte de treinamento do modelo, não foram definidos critérios para balanceamento de classes e isso resultou em uma acurácia de 97,02%.

No modelo número dois, na parte de treinamento do modelo, foi definido critério para balanceamento de classes e isso resultou em uma acurácia de 94,38%.

Observe na tabela abaixo os resultados adquiridos para cada modelo, em cada cenário:

|  |  |  |  |
| --- | --- | --- | --- |
| Tabela 3. Comparando os modelos em cada cenário | | | |
| Modelo | cenário um | cenário dois | cenário três |
| Modelo um, acurácia | 96,90% | 96,59% | 96,50% |
| Modelo dois balanceado, acurácia | 93,64% | 93,57% | 93,46% |
| Fonte: Dados originais da pesquisa | |  |  |

Observando a distribuição da coluna Qtd\_vendas da base raiz (usada para treinar o modelo) versus a mesma coluna dos cenários criados. Existem oscilações individuais nas distribuições, mas isso reforça o objetivo de distinção delas:

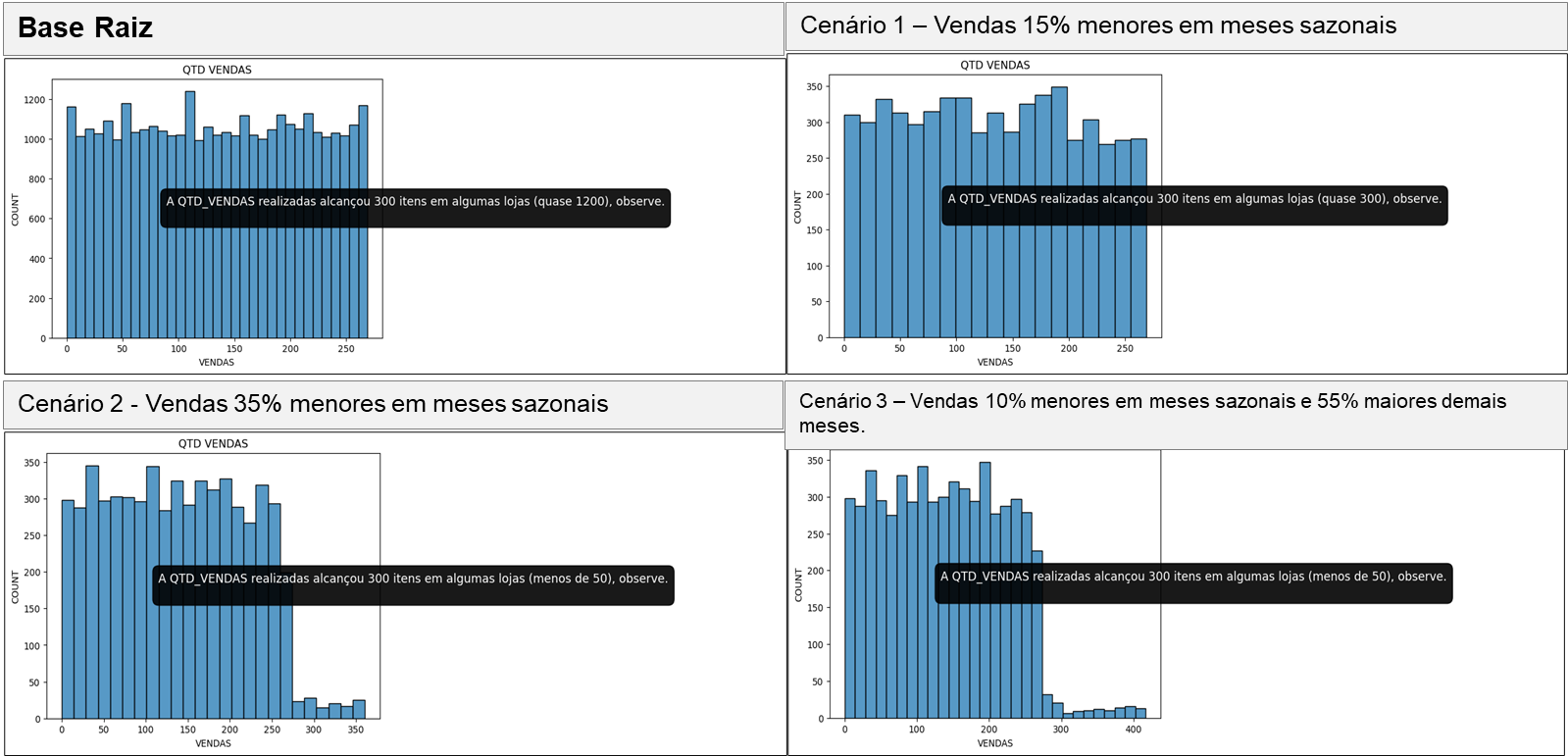


Figura 5. Distribuição da coluna de vendas na base de treino e nas bases de teste (cenários).

Fonte: Dados originais da pesquisa.

Em seu trabalho, Branquinho Filho (2020) alega sobre o uso do balanceamento de classes, demandando atenção sobre o tema. O seu objetivo era prever atrasos nas entregas de produtos de e-commerce, criando um ambiente preditivo para prevenção.

Ele constrói e usa diversos modelos, com apoio do AutoML que é uma função da biblioteca H2O e tem como objetivo a criação de diversos modelos afim de encontrar o melhor, competindo entre eles.

Ele ressalta sobre um bom tratamento de dados, pois assim os modelos poderão operar com a máxima capacidade. Sendo o seu trabalho um problema de classificação binária ele aplicou os métodos para balanceamento em seus modelos escolhidos e estes foram o: xGBoost, Gradiente Boosting, Machine e Dynamic Random Dorest.

Neste trabalho de conclusão de curso é observado que atuar no balanceamento de classes em modelos de Randomforest com variáveis multiclasse, é um processo fundamental para se obter melhor e mais seguros resultados.

**Considerações Finais**

A leitura dos resultados se realiza da seguinte maneira:

Quantidade de produtos que deve estar na loja, para um giro de 60%.

Estoque Ideal = (quantidade de vendas / 0,6)

Quantidade de produtos que deve ser retirado ou enviado à loja.

Necessidade = (Estoque Ideal – Estoque atual)

Lojas agrupadas pela necessidade (Multiclasse).

Status =

0 = lojas com necessidade menor ou igual a 0.

1 = lojas com necessidade menor que 16.

2 = lojas com necessidade maior ou igual a 90.

3 = lojas que não se enquadraram nos demais status.

Para cada cenário, onde o modelo classificou o produto como ruptura (status 1), a ação a ser tomada é criar um reforço de estoque no armazém mais próximo das lojas, ganhando velocidade nos transportes e abastecimento delas.

Para os cenários onde o modelo classificou com alta cobertura (status 0 e 2), a ação a ser tomada é a redistribuição destes produtos, removendo-os destas lojas e colocando à disposição onde existem oportunidades/necessidade.

Embora o modelo número um, obtenha uma performance melhor, ele não está preparado para lidar com dados desbalanceados. No modelo número dois, tendo definido o parâmetro para balanceamento de classes, o resultado decai sua acurácia em -2,64% (97,02%para94,38%).

No tocante ao modelo dois, o indicador de acurácia ter baixado pode ser algo preocupante, mas quando percebe-se e evitam-se os impactos inerentes ao desbalanceamento de classes, então agora o modelo se torna mais fluido, sem vícios.

A validação de modelos apenas pela acurácia tende a ser maliciosa, pois aponta sobre características da amostra completa, a acurácia não retorna sobre as características das variáveis. Sendo assim, ao ajustar o balanceamento de classes, tornamos o modelo mais equilibrado. O uso do parâmetro para balanceamento de classes é fundamental para resultados mais seguros, aliado às outras necessárias interpretações dos seus resultados.

Por fim, o modelo proposto demonstrou ser capaz de classificar a necessidade de cada loja, para cada cenário. Sendo considerado coerente com os objetivos esperados.

**Agradecimentos**

Primeiro, eu agradeço por essa energia, que me rodeia e ajuda a ser sempre melhor. Essa energia de transformação que ajuda a melhorar a minha vida profissional e pessoal, alguns chamam de Deus, Orixá, entre outros.

Segundo agradecimento vai para os mestres, vocês professores e a minha orientadora, que me apoiou e me ajudou ao máximo que pode.

Família e amigos, eu escolhi o caminho dos estudos e por este motivo eu sigo acreditando que podemos melhorar, não por quem sabe mais ou menos, mas sim por quem se interessa em movimentar. Gratidão!

**Referências**

Branquinho Filho, MSc Delermando. 2020. Predição de Falhas na Logística de Entregas em E-commerce. Revista de Ubiquidade, v. 3, n. 2, p. 1–14. Disponivel em: <https://revistas.anchieta.br/index.php/RevistaUbiquidade/article/view/1673>.Acesso em: 25 fev. 2024.

Cunha, V.B.C. 2023. Uso de aprendizado de máquina para especificação do tempo de entrega em vendas via e-commerce. Universidade Federal de Outro Preto [UFOP]. 1-42. Disponivel em: <http://www.monografias.ufop.br/handle/35400000/5808>.Acesso em: 25 fev. 2024.

Chen, D.Y. 2018. Analise de dados com python e pandas. Primeira edição. Novatec editora Ltda, São Paulo, SP, Brasil.

Géron, Aurélien. 2021. Mãos à Obra: Aprendizado de Máquina com Scikit-Learn, Keras & TensorFlow. Segunda edição. Alta Books, Rio de Janeiro,RJ, Brasil.

Gianesi, I.G.N.; Biazzi, J.L. 2009. Gestão Estratégica de Estoques. Revista de Administração da Universidade de São Paulo [RAUSP]. 1-15.

Disponivel em: <https://revistas.usp.br/rausp/article/view/44542>.Acesso em: 25 mar. 2024.

Oshiro, T.M. 2013. Uma abordagem para a construção de uma única árvore a partir de uma Random Forest para classificação de bases de expressão gênica. Google Acadêmico. 1-101. Disponivel em: <https://www.teses.usp.br/teses/disponiveis/95/95131/tde-15102013-183234/publico/monografia>.Acesso em: 15 mar. 2024.

Varotto, L.F. 2006. História do varejo.FGV-EAESP.1-5.

Disponivel em: <https://doi.org/10.12660/gvexec.v5n1.2006.34379>.Acesso em: 25 fev. 2024.