”Machine Learning in Dropshipping: The Revolutionary Approach of Drop Master”

Matheus Freitas Martins
Instituto de Ciências Exatas e Tecnológicas
Universidade Federal de Viçosa
Minas Gerais, Brasil
Email: matheus.f.martins@ufv.br

***Resumo*—O comércio moderno enfrenta desafios significativos na gestão de estoques e na precificação sazonal. O armazenamento excessivo de estoque pode levar a custos desnecessários, enquanto a falta dele pode resultar em perda de vendas. Da mesma forma, a flutuação sazonal dos preços requer análise cuidadosa para maximizar os lucros. A modalidade de *dropshipping* surgiu como uma solução, eliminando a necessidade de manter estoques. Contudo, para agilizar a entrega, especialmente considerando que a maioria dos produtos é proveniente da China, ter um estoque mínimo é uma estratégia viável. O desafio então é prever qual produto é o mais rentável para manter em estoque. Neste contexto, este trabalho propõe uma abordagem baseada em Machine Learning para auxiliar vendedores na tomada de decisões. Utilizando um modelo de regressão com Random Forest, conseguimos prever com precisão de 92.5% quais produtos terão a melhor performance. Esta abordagem revoluciona o *dropshipping*, permitindo estratégias mais eficientes e lucrativas.**

***Palavras-chave*—** *dropshipping*, Gestão de estoque, Previsão de demanda, Machine Learning, Regressão com Random Forest.

1. Introdução

O comércio eletrônico tem crescido de forma significativa na última década, apresentando uma série de oportunidades e desafios para varejistas e consumidores [1]. Dentre as muitas estratégias de comércio eletrônico que se desenvolveram, o *dropshipping* surge como uma inovação na gestão de estoques e cadeia de suprimentos [2]. Nesse modelo de negócios, os vendedores não mantêm estoques de produtos, mas dependem de terceiros – geralmente atacadistas ou fabricantes – para enviar produtos diretamente aos consumidores [3].

Apesar de suas vantagens óbvias, como a redução de custos associados ao armazenamento e manutenção de estoques, o *dropshipping* não está isento de desafios. A dependência de fornecedores externos pode levar a longos tempos de entrega, especialmente quando a maioria dos produtos é proveniente de países distantes, como a China [4]. Além disso, a falta de controle sobre o estoque pode resultar em situações onde a demanda dos consumidores não pode ser atendida de maneira oportuna [5].

Assim, surge a necessidade de ferramentas que possam auxiliar os vendedores na previsão dos produtos que provavelmente terão alta demanda, permitindo-lhes manter um estoque mínimo desses produtos e melhorar a eficiência das operações de *dropshipping*. Em resposta a esse desafio, este trabalho apresenta uma abordagem baseada em Machine Learning, especificamente um modelo de regressão com Random Forest, para prever quais produtos terão a melhor performance em termos de vendas [6].

A regressão com Random Forest é uma técnica robusta que combina a força de múltiplas árvores de decisão para fazer previsões mais precisas e confiáveis [7]. Ao aplicar esta técnica ao problema do *dropshipping*, somos capazes de fornecer aos vendedores informações valiosas que podem ajudá-los a otimizar suas estratégias de negócios e operações.

2. Metodologia

Nossa metodologia consistiu em várias etapas, desde a coleta e preparação dos dados até a modelagem, avaliação e implementação do modelo final. Iniciamos buscando um conjunto de dados adequado para a tarefa em mãos. Embora inicialmente tivéssemos planejado usar a API do Mercado Livre, acabamos por utilizar um conjunto de dados disponível no Kaggle que continha as vendas diárias de produtos em várias lojas [8].

Os dados foram inicialmente carregados e combinados em um único DataFrame, utilizando a biblioteca pandas, e consistiam em informações diversas sobre vendas de produtos, incluindo a data de venda, preço do produto, identificador único do produto, entre outras informações.

Nossa base de dados é composta por um total de 1.622.434 registros. A distribuição dos dados ao longo dos anos apresenta uma certa distinção, com 48,5% dos registros correspondendo ao ano de 2013, 37,7% ao ano de 2014 e 13,8% ao ano de 2015. Embora a distribuição dos dados não seja igualmente balanceada ao longo desses anos, optamos por não realizar uma equalização dos volumes anuais.

Esta decisão foi tomada levando em conta a natureza temporal dos dados. Dados de séries temporais possuem características e padrões únicos que podem se perder ou serem distorcidos caso seja aplicada uma técnica de balanceamento. Além disso, o balanceamento poderia introduzir um viés artificial que poderia comprometer a integridade dos dados e a precisão de nossas análises.

Em seguida, realizamos uma série de etapas de pré-processamento para limpar e preparar os dados para a modelagem. Isso incluiu a remoção de linhas duplicadas, a conversão da data para o formato correto, a extração do ano e do mês das datas e a filtragem dos dados para incluir apenas os itens que tiveram registros em todos os três anos (2013, 2014 e 2015) cobertos pelo conjunto de dados. Além disso, realizamos a agregação dos dados em uma média mensal e soma, e também efetuamos a codificação de variáveis categóricas.

Com os dados devidamente preparados, dividimos o conjunto de dados em dados de treino e de teste utilizando a proporção 80-20. Essa abordagem demonstrou ser superior em relação a divisão temporal, uma vez que nosso conjunto de dados cobre um período

limitado de tempo e queríamos treinar nosso modelo em uma variedade maior de condições.

Para modelagem, escolhemos um modelo de floresta aleatória (*Random Forest*) por sua capacidade de lidar com uma ampla variedade de tipos de dados e pela robustez em face a *outliers* e variáveis irrelevantes. Combinamos este modelo com um escalonador de recursos (*StandardScaler*) em um *pipeline* e realizamos a validação cruzada com 10 folds para evitar o ajuste excessivo e garantir que o modelo seria capaz de generalizar bem para novos dados e utilizamos a métrica R^2 Score como medida de avaliação. Quanto aos hiperparâmetros utilizados no modelo, definimos o número de estimadores como 100, o que significa que 100 árvores de decisão diferentes serão criadas no modelo de floresta aleatória. Além disso, definimos o parâmetro de semente aleatória como 42.

Avaliamos o desempenho do modelo utilizando três métricas diferentes: o erro médio absoluto (MAE), a raiz do erro quadrático médio (RMSE) e o coeficiente de determinação (R^2 Score). Enquanto o MAE e o RMSE nos fornecem uma medida da magnitude do erro do modelo, o R^2 Score nos dá uma ideia de quão bem as variáveis independentes do modelo são capazes de explicar a variação nos dados de saída.

Finalmente, tentamos implementar nosso modelo em um ambiente de produção. Inicialmente tentamos utilizar o Azure, mas enfrentamos problemas para fazer o upload do arquivo de modelo, que pesava cerca de 1.5 GB. Após algumas tentativas frustradas, optamos por usar o ngrok, uma ferraria que permite expor servidores locais à internet. Embora esta não seja a solução ideal, nos permitiu colocar nosso modelo em produção e atingir nossos objetivos.

2.1. *Features* e Variáveis-Alvo

Para o treinamento do nosso modelo, várias *features* extraídas do conjunto de dados original foram utilizadas. Estas *features*, juntamente com as suas descrições, são ilustradas na Tabela 1.

Feature	**Descrição**
year	O ano em que os dados foram coletados.
month	O mês em que os dados foram coletados.
item_id	Identificador único para cada item.
item_category_id	Identificador único para a categoria do item.

TABLE 1. *Features* USADAS NO TREINAMENTO DO MODELO

Além dessas características, as variáveis ***item_cnt_day*** e ***item_price*** foram usadas como as variáveis dependentes, ou seja, dois alvos (*multi-target*) [9] que nosso modelo pretende prever.

2.2. Plataforma Drop Master

Uma parte fundamental deste trabalho foi a criação de uma plataforma de previsão personalizada, denominada ***Drop Master*** [10]. Essa plataforma foi desenvolvida para permitir que vendedores de dropshipping obtenham previsões de quantidade e preço recomendados para seus produtos com base em dados históricos.

A ideia principal por trás do ***Drop Master*** é fornecer aos vendedores uma ferramenta intuitiva e de fácil utilização para tomar decisões informadas sobre quais produtos promover e como precificá-los, a fim de maximizar seus lucros. A plataforma utiliza um modelo de regressão baseado em *Random Forest*, treinado em dados históricos de vendas, para realizar as previsões.

Ao acessar a plataforma, os usuários podem inserir os dados do produto, como ano, mês, ID do produto e ID da categoria do produto, por meio de um formulário. Esses dados são então utilizados para gerar uma previsão de quantidade recomendada e preço recomendado para o produto selecionado.

Além das previsões individuais, a plataforma também armazena todas as previsões anteriores em um DataFrame chamado "predictions". Os usuários têm a capacidade de visualizar todas as previsões anteriores, bem como a receita total estimada com base nessas previsões.

A plataforma oferece recursos adicionais para análise e visualização dos resultados. Os usuários podem baixar todas as previsões anteriores em um arquivo .csv para uma análise mais detalhada. Além disso, a plataforma permite visualizar um gráfico das previsões de quantidade recomendada em relação aos preços recomendados, fornecendo uma representação visual dos padrões identificados pelo modelo.

Por fim, a plataforma também disponibiliza um gráfico de pizza que mostra a distribuição da receita estimada entre diferentes produtos. Isso fornece uma visão rápida e intuitiva dos produtos que contribuem mais significativamente para a receita geral.

No geral, o ***Drop Master*** oferece aos vendedores de *dropshipping* uma plataforma poderosa para obter previsões precisas e tomar decisões estratégicas com base em dados. Com a capacidade de prever a quantidade e o preço recomendados para os produtos, os vendedores podem otimizar suas estratégias de precificação e estoque, reduzir riscos e aumentar sua lucratividade.

A Figura 1 abaixo, não é uma previsão, mas sim uma representação gráfica de dados históricos do nosso conjunto de dados que serve para exemplificar nosso ponto, oferece uma representação visual clara do volume de vendas mensais realizadas ao longo do ano de 2013 para a categoria de produtos PS3. A análise deste gráfico é crucial para compreender os picos e vales nas vendas e correlacioná-los com fatores externos que possam ter influenciado essas flutuações.

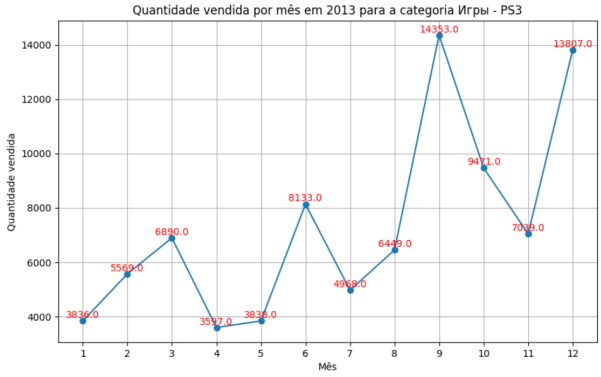

Figura 1. Valores reais de vendas mensais para a categoria de produto PS3 ao longo do ano de 2013.

3. Resultados e discussão

Para analisar o desempenho do modelo, foram utilizadas três métricas: Erro Médio Absoluto (MAE), Erro Quadrático Médio da Raiz (RMSE) e Coeficiente de Determinação (R^2 Score).

Métrica	**Valor**
RMSE	104.496
MAE	33.105
R^2 Score	0.925

TABLE 2. MÉTRICAS DO MODELO RANDOM FOREST

1) O MAE foi de 33.105, o que significa que as previsões do modelo, em média, desviam por cerca de 33.11 unidades do valor real. Esse valor é significativo para a ”Quantidade recomendada”, pois a previsão está sendo desviada por uma margem alta em comparação com o valor real. No entanto, é menos impactante para o ”Preço recomendado”, pois representa uma pequena porcentagem do valor real.
2) O RMSE foi de 104.496, o que significa que as previsões do modelo desviam, em média e em termos quadráticos, por cerca de 104.50 unidades do valor real. Mais uma vez, esse valor é significativo para a ”Quantidade recomendada”, mas menos para o ”Preço recomendado”, devido à diferença nas magnitudes dessas variáveis.
3) Finalmente, o R^2 Score foi de 0.925, o que significa que 92.50% da variação nos dados de saída pode ser explicada pelas variáveis independentes no modelo. Isso é um sinal de bom desempenho do modelo, independentemente da magnitude das variáveis de saída.

Neste caso, o R^2 parece ser a métrica mais adequada para avaliar o desempenho do modelo, pois leva em consideração a variação nos dados de saída, independentemente de sua magnitude. Além disso, o alto valor de R^2 indica que o modelo está fazendo um bom trabalho em capturar a variação nos dados.

Os valores previstos para quantidade e preço dos itens e suas respectivas categorias são apresentados na Figura 2 fornecida. Esses valores representam as previsões geradas pelo modelo de regressão com base nos dados históricos disponíveis. Essa Figura mostra uma única previsão para cinco diferentes categorias de produtos em um determinado mês. Essa visualização permite verificar como o modelo se comporta ao fazer previsões para diferentes categorias de produtos simultaneamente. Vamos analisar esses dados com mais profundidade:

	Ano	Mês	ID do Produto	ID da Categoria do Produto	Quantidade recomendada	Preço Recomendado	Receita Estimada
0	2016	7	20949	71	3446.58	4.975298	17147.763313
1	2016	7	1905	30	60.59	242.377016	14685.623400
2	2016	7	7071	19	37.75	999.601497	37734.956496
3	2016	7	11921	40	66.60	575.847855	38351.467123
4	2016	7	13881	55	34.74	596.117210	20709.111880

Figura 2. Previsões de produtos das 5 melhores categorias em termos de volume de vendas.

1) **ID do Produto e ID da Categoria do Produto:** Cada linha da tabela representa um item específico identificado pelo seu ID do Produto e ID da Categoria do Produto. Essas informações são importantes para identificar os produtos e suas categorias correspondentes.
2) **Ano e Mês:** O ano e mês indicam o período em que as previsões foram feitas. Nesse caso, as previsões foram realizadas para o mês de julho de 2016.
3) **Quantidade Recomendada:** A coluna ”Quantidade Recomendada” indica a quantidade de itens que é recomendada para serem vendidos nesse período. Essa quantidade é uma estimativa com base no histórico de vendas e outros fatores considerados pelo modelo.
4) **Preço Recomendado:** A coluna ”Preço Recomendado” representa o preço sugerido para os itens. Esse valor é calculado pelo modelo com o objetivo de maximizar a lucratividade. É importante ressaltar que esse preço recomendado pode ser influenciado por fatores como a demanda do mercado, a concorrência e as tendências sazonais.
5) **Receita Estimada:** A coluna ”Receita Estimada” calcula a receita esperada com base na multiplicação da quantidade recomendada pelo preço recomendado. Esse valor representa a estimativa de ganhos que pode ser obtida com a venda dos produtos nesse período.

É importante ressaltar que essas previsões foram consideradas totalmente condizentes com o contexto do problema e com os dados históricos utilizados para treinar o modelo. Uma análise aprofundada foi realizada para comparar essas previsões com os dados reais a fim de reafirmar a precisão e o desempenho do modelo.

Essas previsões podem ser utilizadas pelos vendedores para auxiliar na tomada de decisões estratégicas, como o planejamento de estoque, a definição de preços competitivos e a alocação de recursos de forma mais eficiente. Com informações precisas sobre a quantidade e o preço recomendados, os vendedores podem otimizar suas operações e maximizar seus lucros.

A Figura 3 mostra a previsão da quantidade recomendada em relação ao preço recomendado da Figura 2. Essa visualização permite aos vendedores identificar a relação entre a quantidade recomendada e o preço sugerido, auxiliando na definição de estratégias de precificação.

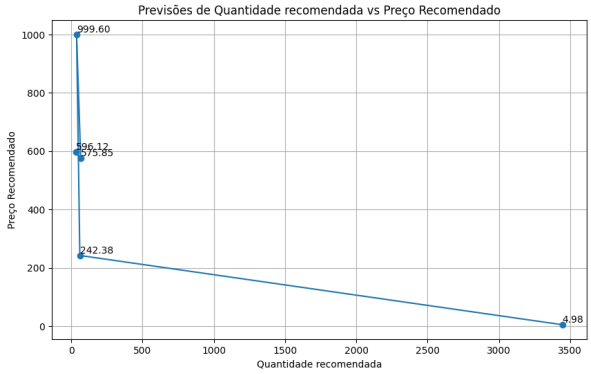

Figura 3. Previsões de quantidade recomendada por preço recomendado de diferentes produtos no mês de Julho.

A Figura 4 apresenta o produto que oferecerá a maior receita percentual da Figura 2. Essa informação é valiosa para os vendedores, pois permite identificar quais produtos têm maior potencial de lucro e direcionar esforços e recursos para maximizar o retorno financeiro.

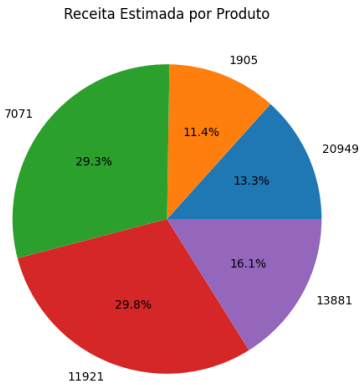

Figura 4. Produtos com as maiores receitas percentuais previstas.

A Figura 5 apresenta uma previsão para cada mês do ano, focando em um produto específico, no caso, o PlayStation 4. Essa

visualização oferece uma visão mais detalhada das previsões ao longo do tempo para um produto específico, permitindo identificar possíveis padrões sazonais ou tendências de demanda.

	Ano	Mês	ID do Produto	ID da Categoria do Produto	Quantidade recomendada	Preço Recomendado	Receita Estimada
0	2016	1	6674	12	204.42	25725.032619	5.258711e+06
1	2016	2	6674	12	173.22	25557.052354	4.426993e+06
2	2016	3	6674	12	134.52	26554.467180	3.572107e+06
3	2016	4	6674	12	92.44	27590.369256	2.550454e+06
4	2016	5	6674	12	136.93	26748.784602	3.662711e+06
5	2016	6	6674	12	154.24	25931.775230	3.999717e+06
6	2016	7	6674	12	36.79	27134.832322	9.982905e+05
7	2016	8	6674	12	19.91	27069.412745	5.389520e+05
8	2016	9	6674	12	14.61	25290.347381	3.694920e+05
9	2016	10	6674	12	14.61	25146.847381	3.673954e+05
10	2016	11	6674	12	28.88	25128.230130	7.257033e+05
11	2016	12	6674	12	27.82	25163.330913	7.000439e+05

Figura 5. Previsão anual do produto PlayStation 4.

Ao observar a relação entre a quantidade recomendada e o preço recomendado, é possível identificar possíveis padrões ou tendências. Por exemplo, pode-se verificar se há uma relação direta entre a quantidade e o preço, ou se existem variações sazonais ou flutuações significativas em determinados meses.

A partir da Figura 6 nota-se a relação entre essas duas variáveis, é possível encontrar um ponto ideal que equilibre a quantidade de produtos a serem recomendados e o preço a ser cobrado, considerando fatores como a demanda do mercado e a competitividade.

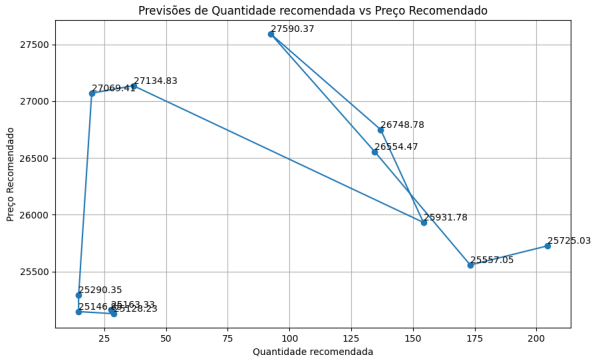

Figura 6. Previsões de quantidade recomendada por preço recomendado para o PlayStation 4 em 12 meses.

No entanto, é importante ressaltar que as previsões são baseadas em modelos estatísticos e podem estar sujeitas a incertezas e variações no mundo real. Portanto, é recomendado que os vendedores utilizem essas previsões como uma orientação e combinem-nas com sua expertise e conhecimento do mercado para tomar decisões informadas e adaptáveis às condições específicas de seus negócios.

Vale ressaltar que, optamos por utilizar o modelo de regressão Random Forest como abordagem principal devido à sua capacidade de lidar com dados complexos e não lineares, além de sua capacidade de capturar relações não lineares entre as variáveis. Testamos também o XGBoost e a Árvore de Decisão, mas ambos demonstraram resultados ligeiramente inferiores em termos do coeficiente de determinação (R^2), com 0,902 e 0,895, respectivamente, em comparação ao desempenho do *Random Forest*. Embora esses modelos tenham apresentado resultados próximos, optamos pelo *Random Forest* para nosso estudo principal devido à sua performance superior e à robustez inerente do modelo.

Também realizamos uma série de testes usando uma divisão temporal dos dados, para avaliar o desempenho do modelo em diferentes condições temporais. Os dados foram divididos de acordo com os anos, com os anos de 2013 e 2014 sendo usados para treinamento e 2015 para teste, resultando em um score R^2 de 0.638. Adicionalmente, dividimos os dados utilizando apenas 2013 para treinamento e 2014 para teste, obtendo um score R^2 de 0.459. Por último, utilizamos 2014 para treinamento e 2015 para teste, resultando em um score R^2 de 0.611.

Apesar das tentativas com a divisão temporal, observamos que os scores R^2 obtidos foram inferiores quando comparados à abordagem de divisão 80-20. Portanto, optamos por prosseguir com a abordagem de divisão 80-20, com a finalidade de treinar nosso modelo.

4. Conclusão e Trabalhos Futuros

Neste estudo, desenvolvemos uma plataforma de previsão de séries temporais para auxiliar vendedores no mercado de *dropshipping*. Utilizando técnicas de machine learning e análise de dados, conseguimos obter resultados promissores na previsão da quantidade de itens vendidos e nos preços desses itens. O modelo de regressão *Random Forest* alcançou um alto desempenho, com um valor de R^2 de 92.5%, demonstrando uma capacidade significativa de explicar a variação nos dados de saída.

As previsões geradas pelo modelo foram consistentes em relação às tendências observadas nos valores anteriores. Isso indica que o modelo foi capaz de capturar padrões e características importantes dos dados de entrada, fornecendo estimativas confiáveis para auxiliar vendedores na tomada de decisões estratégicas.

No entanto, futuros trabalhos podem considerar a comparação com outras baselines e abordagens para uma análise mais abrangente. Ademais, seria extremamente benéfico testar esses modelos em um conjunto de dados recente, maior e mais representativo de um ambiente de produção, preferencialmente com dados históricos de pelo menos uma década. Isso permitiria uma melhor compreensão das tendências e padrões de longo prazo, fornecendo insights mais profundos sobre o desempenho dos modelos e sua aplicabilidade em um cenário do mundo real.

É importante ressaltar que este estudo se destaca por abordar duas variáveis alvo, a quantidade vendida e o preço, simultaneamente. Essa abordagem única permite aos vendedores obter previsões precisas não apenas sobre a quantidade de itens vendidos, mas também sobre os preços recomendados. Essa combinação de informações oferece uma visão abrangente e estratégica para otimizar as operações de negócios e maximizar os lucros.

Em suma, este trabalho demonstrou o potencial e a eficácia do modelo de regressão *Random Forest* na previsão de séries temporais para o mercado de *dropshipping*. As previsões consistentes, o alto valor de R^2 e a abordagem única de duas variáveis alvo fornecem uma base sólida para o uso dessa plataforma como uma ferramenta valiosa para vendedores. Esperamos que este estudo inspire pesquisas futuras e a aplicação prática de técnicas de machine learning para aprimorar as operações de negócios no contexto do *dropshipping*.

Agradecimentos

”Se eu vi mais longe, foi por estar sobre ombros de gigantes.” Frase parafraseada de Isaac Newton.

Referências

[1] V. N. MULLER, “E-commerce: vendas pela internet,” *Fundação Educacional do Municipio de Assis*, 2013.

[2] R. Cui, D. J. Zhang, and A. Bassamboo, "Learning from inventory availability information: Evidence from field experiments on amazon," *Management Science*, vol. 65, no. 3, pp. 1216–1235, 2019.

[3] S. Burt and L. Sparks, "E-commerce and the retail process: a review," *Journal of Retailing and Consumer services*, vol. 10, no. 5, pp. 275–286, 2003.

[4] M. Zhang, Y. K. Tse, B. Doherty, S. Li, and P. Akhtar, "Sustainable supply chain management: Confirmation of a higher-order model," *Resources, Conservation and Recycling*, vol. 128, pp. 206–221, 2018.

[5] S. Cetinkaya and C.-Y. Lee, "Stock replenishment and shipment scheduling for vendor-managed inventory systems," *Management Science*, vol. 46, no. 2, pp. 217–232, 2000.

[6] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[7] R. Genuer, J. Poggi, and C. Tuleau-Malot, "Variable selection using random forests pattern recognition letters, 31, 2225 10.1016," *J. PATREC*, vol. 14, 2010.

[8] i. M. T. u. K. Alexander Guschin, Dmitry Ulyanov, "Predict future sales," 2018. [Online]. Available: https://kaggle.com/competitions/competitive-data-science-predict-future-sales

[9] H. Borchani, G. Varando, C. Bielza, and P. Larranaga, "A survey on multi-output regression," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, no. 5, pp. 216–233, 2015.

[10] M. Freitas, "FreeMarket: A GitHub Repository," https://github.com/mtsfreitas/FreeMarket, 2023.