

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

```
/* Nome: Matheus Freitas Martins  
   Matrícula: ES111281  
   Disciplina: CCF 726 */
```

```
// Universidade Federal de Viçosa  
// CCF 726 - Engenharia de Aprendizado de Máquina
```

}

# Sumário {

```
// 1. Introdução
// 2. Metodologia
// 3. Resultados e Discussões
// 4. Conclusão
// 5. Trabalhos Futuros
// 6. Implantação do modelo
// 7. Referências
```

}

# Introdução - Problema?

Dropshipping é um modelo de negócios amplamente adotado no e-commerce, que permite aos vendedores operar sem manter estoque, pois os fornecedores enviam diretamente os produtos aos compradores.

Embora eficiente em muitos aspectos, essa abordagem apresenta desafios significativos. Um dos principais é a dificuldade em prever a demanda e a quantidade de estoque recomendada para um determinado produto em um mês.

Pode resultar em:

- Atrasos na entrega; 🐢
- Insatisfação do cliente; 😡
- Perder espaço para a concorrência. 🙌

# Introdução - Solução?

Nosso trabalho tem como objetivo aliviar esse problema recorrente no dropshipping. Propomos uma metodologia inovadora que ajuda a prever a demanda do produto e a determinar a quantidade ideal de estoque que um fornecedor deve ter para um determinado produto.

Dessa forma, os vendedores podem:

- Otimizar sua operação; 💰📦
- Estocar ou combinar previamente produtos com o fornecedor; 📦
- Minimizar a demora no envio; 🚀
- Melhorar a satisfação do cliente. 😍😁

# Metodologia - Conjunto de Dados

API do Mercado Livre? ✗

Dataset Russo no Kaggle? ✓ (thx Germano)

Companhia de software russa - 1C Company.

Os dados foram combinados em um único `DataFrame` → `combined_sales.csv`

- `sales_train.csv` - o conjunto de treinamento. Dados históricos diários de janeiro de 2013 a outubro de 2015.
- `items.csv` - informações complementares sobre os itens/produtos.
- `item_categories.csv` - informações complementares sobre as categorias dos itens.
- `shops.csv` - informações complementares sobre as lojas.

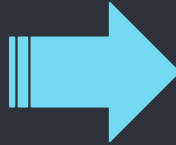
# Metodologia - Conjunto de Dados

- `ID` - um ID que representa uma tupla (Loja, Item) dentro do conjunto de teste
- `shop_id` - identificador único de uma loja
- `item_id` - identificador único de um produto
- `item_category_id` - identificador único da categoria do item
- `item_cnt_day` - número de produtos vendidos.
- `item_price` - preço atual de um item
- `date` - data no formato dd/mm/yyyy
- `date_block_num` - um número de mês consecutivo, usado para conveniência. Janeiro de 2013 é 0, fevereiro de 2013 é 1, ..., outubro de 2015 é 33
- `item_name` - nome do item
- `shop_name` - nome da loja
- `item_category_name` - nome da categoria do item



# Metodologia - Conjunto de Dados

- date\_block\_num
- shop\_id
- item\_id
- item\_price
- item\_cnt\_day
- item\_name
- item\_category\_id
- item\_category\_name
- shop\_name
- year
- month



- year
  - month
  - item\_id
  - item\_category\_id
- } Features  
(variáveis independentes)

- item\_price
  - item\_cnt\_day
- } Target  
(variáveis dependentes)

# Metodologia - Conjunto de Dados

- **Remoção de Duplicatas:** < 5% das linhas eram duplicadas; 🗑️
- **Conversão de Datas:** Conversão o formato de data correto, permitindo a extração das colunas `year` e `month`.
- **Filtragem de Itens:** Somente aqueles `itens` que têm registros em todos os três anos foram mantidos no dataset. Isso ajuda a garantir que haja dados suficientes para cada item para fazer previsões precisas.
- **Agregação de Dados:** Os dados foram agregados em `médias mensais e totais`, para o `preço do item` e a `quantidade de item`, respectivamente. Isso permite que as previsões sejam feitas em uma base mensal.
- **Codificação de Variáveis Categóricas:** A variável `item_category_id` foi transformada por `LabelEncoder`.



# Metodologia - Treinamento

1.622.434 registros distribuídos em:

Treinamento

[ 80% ]

Teste

[ 20% ]

2013 (48,5%)

2014 (37,7%)

2015 (13,8%)

Balanceamento? ✖

Divisão temporal? 🤔

# Metodologia - Random Forest

- Capacidade em lidar com outliers e variáveis irrelevantes.

Construção do pipeline: normaliza os dados usando **StandardScaler** → aplica o modelo **Random Forest Regressor**.

```
# Construindo o pipeline
pipeline = Pipeline([
    ('scaler', StandardScaler()),
    ('regressor', RandomForestRegressor(n_estimators=100, random_state=42))
])
```



árvores de  
decisão

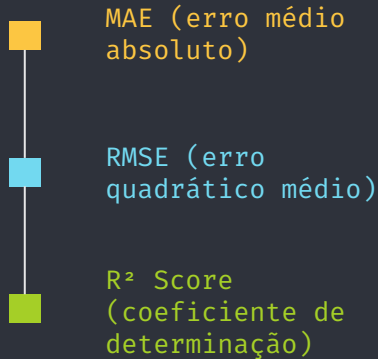


semente  
aleatória

Validação cruzada: Foi realizada a validação cruzada com **10 folds** para avaliar a eficácia do modelo. Avaliando com  $R^2$  Score:

- Scores: [0.90 0.89 0.82 0.74 0.92 0.93 0.82 0.87 0.91 0.94]
- Média: 0.87
- Desvio padrão: 0.060

# Metodologia - Métricas



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.

# Resultados - Random Forest



MAE : 33.105

As previsões do modelo, em média, desviam por cerca de 33.11 unidades do valor real.



RMSE : 104.496

As previsões do modelo desviam em média e em termos quadráticos por cerca de 104.50 unidades do valor real.



R<sup>2</sup> Score : 0.925

O R<sup>2</sup> Score de 92,50% nos diz que a variação nos dados de saída pode ser explicada pelas variáveis independentes no modelo. Indica um bom desempenho do modelo, independentemente da magnitude das saídas.

# Resultados - Baselines



Métrica	Valor
RMSE	104.496
MAE	33.105
R <sup>2</sup> Score	0.925

#1  
Random  
Forest

Métrica	Valor
RMSE	154.369
MAE	76.5619
R <sup>2</sup> Score	0.902

#2  
XGBoost

Métrica	Valor
RMSE	122.947
MAE	29.902
R <sup>2</sup> Score	0.895

#3 Decision Tree

# Resultados - Divisão Temporal

Métrica	Valor
RMSE	211.711
MAE	83.482
R <sup>2</sup> Score	0.638

2013 + 2014 (treino)  
2015 (teste)

Métrica	Valor
RMSE	239.827
MAE	92.539
R <sup>2</sup> Score	0.459

2013 (treino)  
2014 (teste)

Métrica	Valor
RMSE	215.356
MAE	84.383
R <sup>2</sup> Score	0.611

2014 (treino)  
2015 (teste)

# Resultados - Drop Master

	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

# Resultados - Drop Master

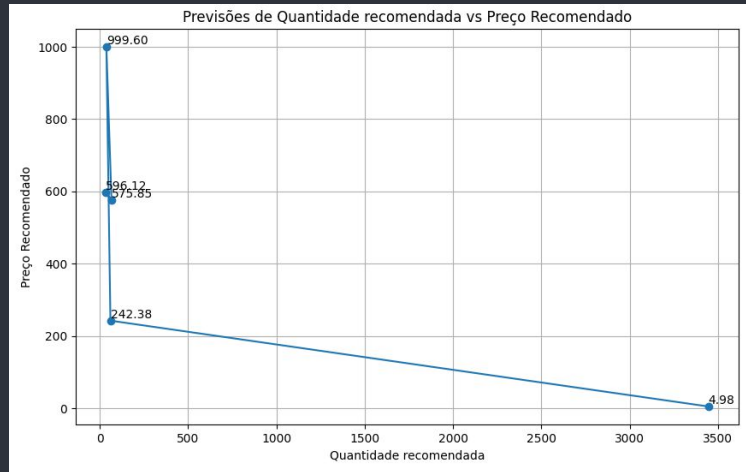


Figura 3. Previsões de quantidade recomendada por preço recomendado de diferentes produtos no mês de Julho.



Figura 4. Produtos com as maiores receitas percentuais previstas.



# Resultados - Drop Master

	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.

# Resultados - Drop Master

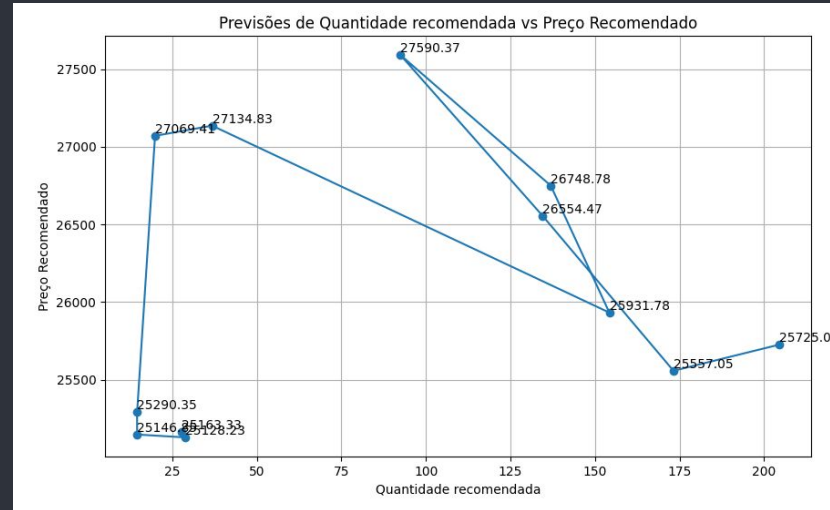


Figura 5. Previsões de quantidade recomendada Vs preço recomendado para o PlayStation 4 em 12 meses.

# Conclusão

Neste estudo, desenvolvemos uma plataforma de previsão de séries temporais para auxiliar vendedores no mercado de dropshipping.

- $R^2$  de 92.5% para Random Forest
- Previsões consistentes e confiáveis
- Se destaca por utilizar duas variáveis alvo (multi-target)

}

# Trabalhos Futuros

- Comparação com outras baselines e abordagens
- Conjunto de dados recente e mais robusto
- Aplicação do modelo em um conjunto de dados de produção real

}

# Implantação do modelo em prod.

- Problema devido ao tamanho do `model.plk` (1.5 GB)

<https://dropmasterportalfinal.azurewebsites.net/> (erro 504.0 GatewayTimeout)

- Ngrok (túnel seguro para o localhost)

1. Inicie seu aplicativo Flask em seu ambiente de desenvolvimento local. Por padrão, o Flask inicia no `localhost:5000`.
2. Comando: `ngrok http 5000`

**Acessar:** <https://fd0c-45-178-248-51.ngrok-free.app>

# Implantação do modelo em prod.

**Drop Master**

Ano:

Mês:

ID do Produto:

ID da Categoria do Produto:

**Prever**

**Download das previsões em .csv**

**Visualizar gráfico das previsões**

**Visualizar gráfico de pizza da receita estimada**



## Resultado da Previsão:

Quantidade recomendada: 3446.58

Preço recomendado: 4.98

## Previsões Anteriores:

	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

## Receita Total Estimada:

Receita total estimada: 17147.76

**Download das previsões em .csv**

**Visualizar gráfico das previsões**

**Visualizar gráfico de pizza da receita estimada**

# Referências

[1] V. N. MULLER, "E-commerce: vendas pela internet," Fundação Educacional do Município 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-sciencepredict-future-sales>

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

Obrigado!!!



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14