Relatório Final: Precificação de Fretes

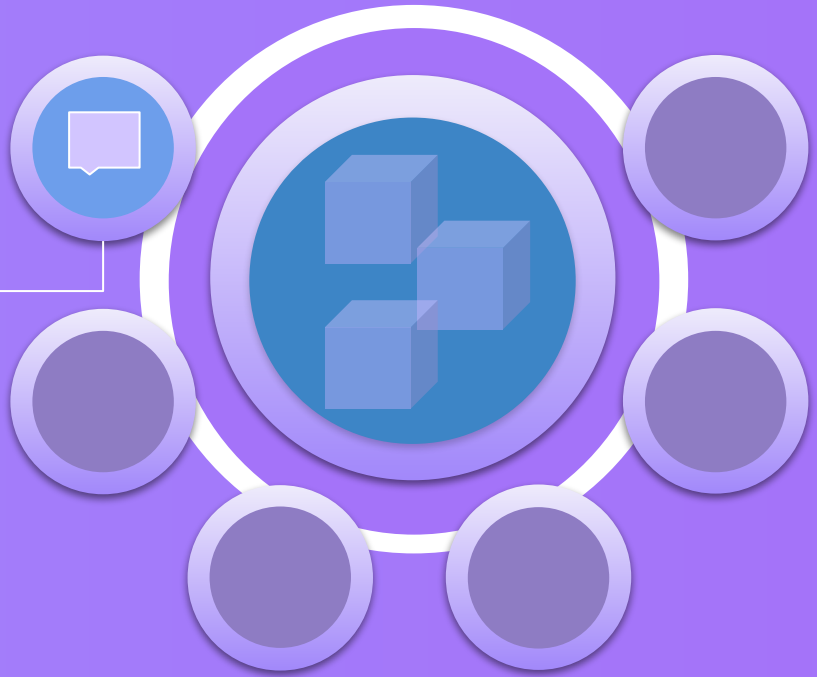
Andréia Castanharo
Arthur Gebhard
Luan de Brito

Jornada de Desenvolvimento

Introdução

Briefing

Imersão Temática



Briefing

“*A partir dos dados cruzados, avaliar estatisticamente quais são os fatores que mais podem impactar no frete de um produto, e se possível, construir um modelo que consiga prever o frete a partir dessas características*”

Imersão Temática

Quais as características relevantes para a precificação de um frete?

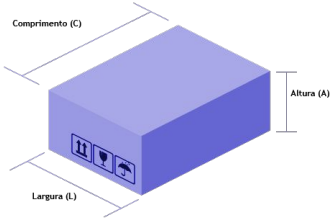
Peso

Dimensões

Seguro da Nota Fiscal

Tempo e Distância

Tributos



O peso e as dimensões são relevantes. As maiores operadoras de frete expresso possuem limites em ambos os aspectos.

O valor na nota fiscal, e características como frágil, perecível, alto valor agregado, impactam no valor do frete dado o fornecimento de seguro.

Além disso o prazo de entrega e a distância fazem parte da precificação pois agregam aos custos logísticos. Existe uma diferença entre pagar por Sedex e por PAC.

Imersão Temática

Quais as formas de frete?

CIF

FOB

Expresso

CIF: Responsabilidade e custo do envio da encomenda é do próprio fornecedor do produto. Preço do frete embutido no preço final do produto. Comum entre empresas ou realizado por empresas que trabalham com entregas de produtos próprios de alto valor e peso.

FOB: Quem compra o produto assume toda a responsabilidade pelo transporte da mercadoria. Normal entre empresas, já que o processo para buscar e acompanhar o produto pode necessitar do suporte de uma equipe de logística.

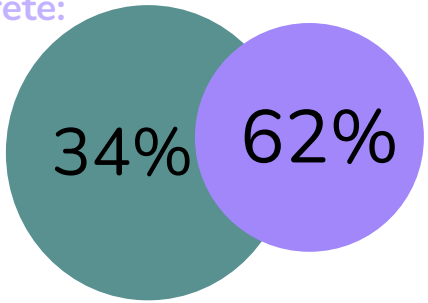
Expresso: O pedido chega até o cliente no menor prazo possível. Entretanto, para que isso ocorra será preciso pagar mais. É o modal mais comum entre comércios e clientes finais. Um exemplo prático são o PAC e o SEDEX, onde os limites de volume e peso são os mesmos. Já

Valor do frete

Uma boa previsão e gestão de frete pode salvar seu E-Commerce da falência aliada a um diagnóstico de ticket médio adequado.

Uma estratégia é trabalhar além do ticket, promoções que reduzem o valor do frete para o cliente.Nos últimos 2 anos, a google identificou um aumento de 109% na busca por fretes gratuitos.

Desistências de compra por causa do valor de frete:



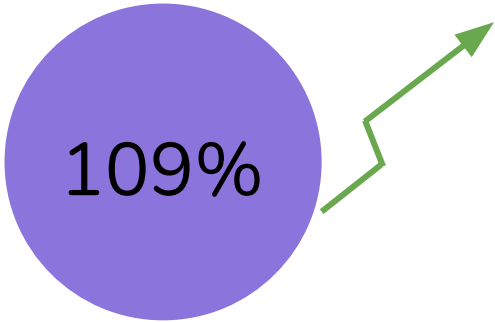
Dessas,

são devido às taxas

Desistências de compra por causa do prazo de frete:

Dimensione seu Ticket e evite perdas

Busca por: “Frete Grátis”

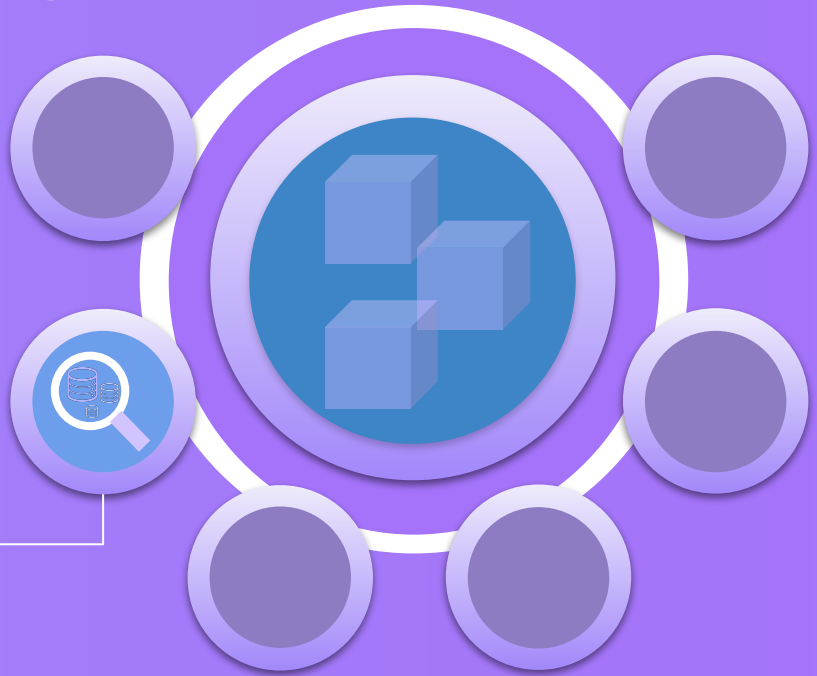


Análise Exploratória

Dados recebidos

Criação de variáveis

Construção do data frame principal



Dados recebidos

O que foi recebido e analisado?

Os seguintes data frames foram recebidos:

- *order_id_df*
- *orders_df*
- *products_df*
- *seller_df*
- *customer_df*
- *geo_df*

Data frame:

- *order_id_df*
- *orders_df*
- *products_df*
- *seller_df*
- *customer_df*
- *geo_df*

- *order_id*
- *order_item_id*
- *product_id*
- *seller_id*
- *shipping_limit_date*
- *freight_price*
- *value*

contém os dados de pedidos

É o principal conjunto de dados para o projeto, contendo os pedidos, valor dos produtos e valor do frete pago.

Dados de: 3095 Lojas

Data frame:

- *order_id_df*
- *orders_df*
- *products_df*
- *seller_df*
- *customer_df*
- *geo_df*

- *order_id*
- *customer_id*
- *order_status*
- *order_purchase_timestamp*
- *order_approved_at*
- *order_delivered_carrier_date*
- *order_delivered_customer_date*
- *order_estimated_delivery_date*

contém dados do status de entrega dos pedidos

Esse data frame contém o status de entrega dos pedidos, desde o pedido, aprovação, acompanhamento do frete e entrega estimada.

O maior interesse por esses dados foi a data de aprovação do pedido e estimativa de entrega, com eles foi possível determinar o prazo de entrega estimado para o produto, em dias.

Data frame:

- *order_id_df*
- *orders_df*
- *products_df*
- *seller_df*
- *customer_df*
- *geo_df*

- *product_id*
- *product_category_name*
- *product_name_lenght*
- *product_description_lenght*
- *product_photos_qty*
- *product_weight_g*
- *product_length_cm*
- *product_height_cm*
- *product_width_cm*

contém dados com as propriedades dos produtos

Esse data frame contém a listagem dos produtos disponíveis com suas características dimensionais e peso.

73 categorias de produto

32000+ produtos

O peso e o número de categorias foram utilizados mais adiante para formulação de hipóteses.

Data frame

- *order_id_df*
- *orders_df*
- *products_df*
- *seller_df*
- *customer_df*
- *geo_df*

- *seller_id*
- *seller_zip_code_prefix*
- *seller_city*
- *seller_state*

dados de vendedores

O data frame apresenta dados referentes a localização dos vendedores, como código postal, cidade e estado de origem.

Data frame

- *order_id_df*
- *orders_df*
- *products_df*
- *seller_df*
- *customer_df* → *customer_id*, *customer_zip_code_prefix*, *customer_city*, *customer_state*
- *geo_df*

contém dados de localização de clientes

O data frame apresenta dados referentes a localização dos clientes, similar aos dados dos vendedores.

Data frame:

- *order_id_df*
- *orders_df*
- *products_df*
- *seller_df*
- *customer_df*
- *geo_df*

- *geolocation_zip_code_prefix*
- *geolocation_lat*
- *geolocation_lng*
- *geolocation_city*
- *geolocation_state*

contém dados de geolocalização

O data frame apresenta dados referentes a coordenadas e localização com base nos prefixos do código postal.

Criação de Variáveis

Novo data frame:
time_df

Idéia: determinação do prazo (dias)

nome da variável criada: prazo

$$Prazo = D.Estimada\ de\ Entrega - Data\ de\ Aprovação$$

Para a construção dessa variável, os dados de datas de aprovação de pedido e a data limite de entrega foram utilizados e vieram dos dataframes *order_id_df* e *order_df* respectivamente, gerando os dados salvos em *time_df*. Esses novos dados foram utilizados para gerar o data frame final posteriormente.

Criação de Variáveis

Volume (cm^3)

nome da variável criada:
product_volume_cm3

$$Volume = Largura \cdot Comprimento \cdot Altura$$

Variáveis utilizadas:

- Largura = *product_width_cm*
- Comprimento = *product_length_cm*
- Altura = *product_heigth_cm*

todas as variáveis dentro do data frame *products_df*

O volume é uma das variáveis de maior influência no valor dos fretes.

Criação de Variáveis

A distância possui potencial de influenciar no valor dos fretes já que aumenta os custos logísticos.

nome da variável criada:
distance

- **Distância (km):**
  através da função: *geopy.distance.geodesic*

- É calculada a menor distância entre dois pontos, levando em consideração a curvatura da Terra. Os pontos são definidos pela Latitude e Longitude dos clientes e vendedores

Construção de Dataframe

objetivo: filtragem de dados e categorias relevantes para a precificação de fretes com base na imersão teórica

O que foi utilizado para cruzar os dados?

- order_id_df
- orders_df
- products_df
- seller_df
- customer_df
- geo_df

dados em comum

Os dados foram cruzados através da variável em comum que identifica o pedido.
Dados em comum entre data frames como o *order_id* (identificação do pedido), o *client_id* (identificação do cliente), .

Ao todo 100.000 linhas de dados, sujeitas a mais filtragens.

Impacto de características

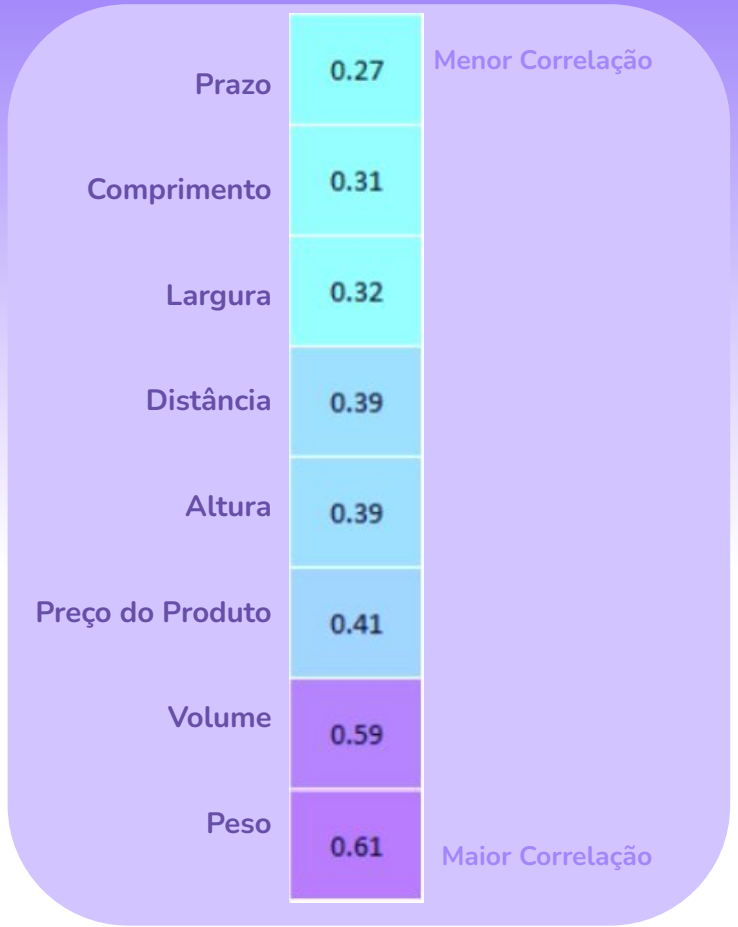
Correlação de dados relevantes

Análises de distribuição, dispersão e características

Correlação em Heatmap

Correlação com o valor do Frete

Esse é o grupo de características correlatas ao frete que a equipe determinou como as mais relevantes.. Optou-se por ordenar as categorias de correlação com o frete em ordem crescente para facilitar a visualização.



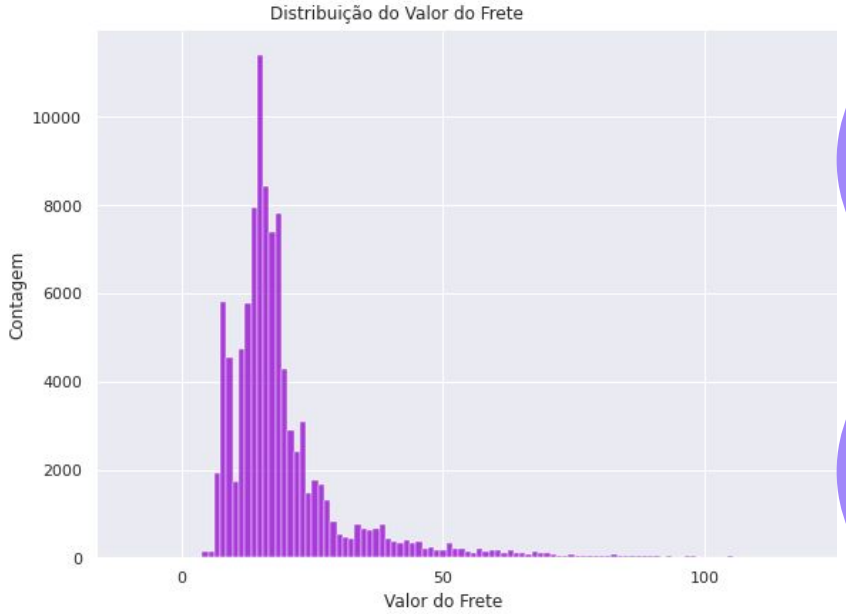
Correlação Peso x Volume = 0.8

Análises de Distribuição

Valor do frete (R$)

Contagem do número de pedidos por valor de frete. Método de Bins: Scott

Distribuição: Valor do frete (R$)



Remoção de fretes abaixo de 5 reais.

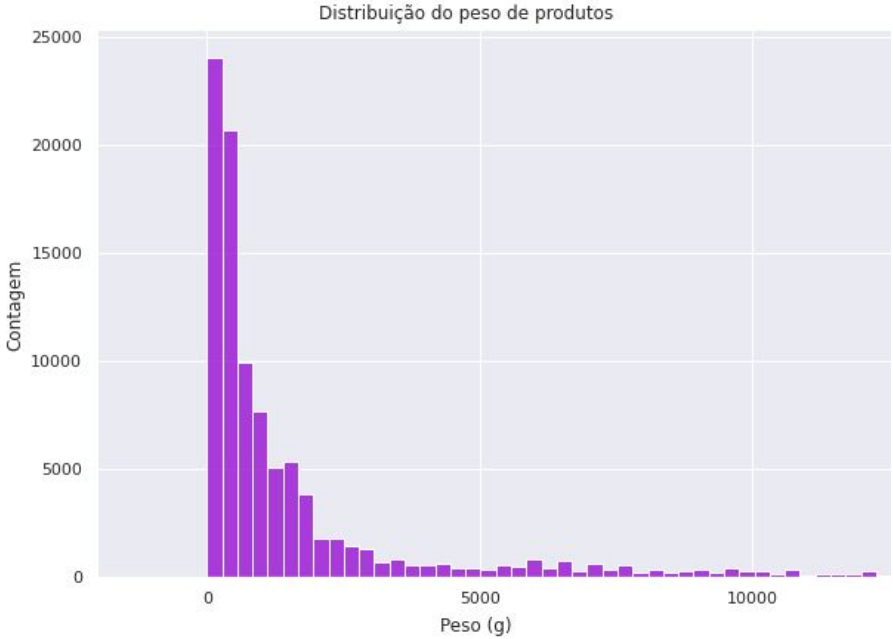
Mais de 50% dos Valores: 9 a 24 reais

Para remoção de valores de frete mínimo a equipe pegou valores presentes de frete e percebeu que dentro de uma mesma cidade e com baixa distância, um produto leve e barato ainda assim possui um valor de frete mínimo por volta de R$10,00. Levando esse valor para a época dos dados (2017-2018) seria o equivalente a aproximadamente R$5,90. Dessa forma optou-se por remover os fretes com valores abaixo de R$5,00, por considerar que eles não podiam representar fretes integrais, e sim fretes passíveis de descontos.

Gráfico de distribuição de produtos por valor de frete - Fonte : Autores - Cálculo de Bins: Método de Scott

Análise do Peso

Peso do produto (g)



Mais de 50% dos fretes abaixo de 1kg

Outliers com faixa de peso de até 40kg

Peso x Preço do Frete

Análises do Peso

Há nesse gráfico de dispersão uma faixa de peso limite visível. Essa faixa é condizente com o peso limite de operação dos Correios e algumas outras empresas de frete, para entregas para outros estados. Existe também a faixa de peso limite para entregas dentro do mesmo estado, para os correios essa faixa é até 50kg, razão que determinou a não remoção de outliers.

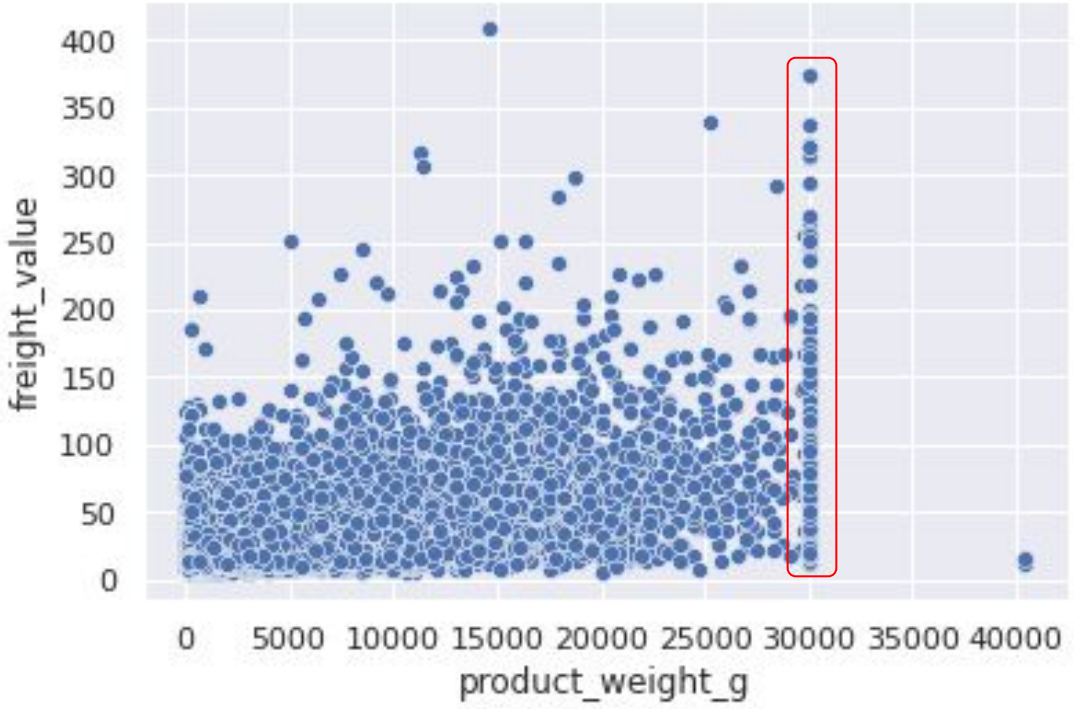
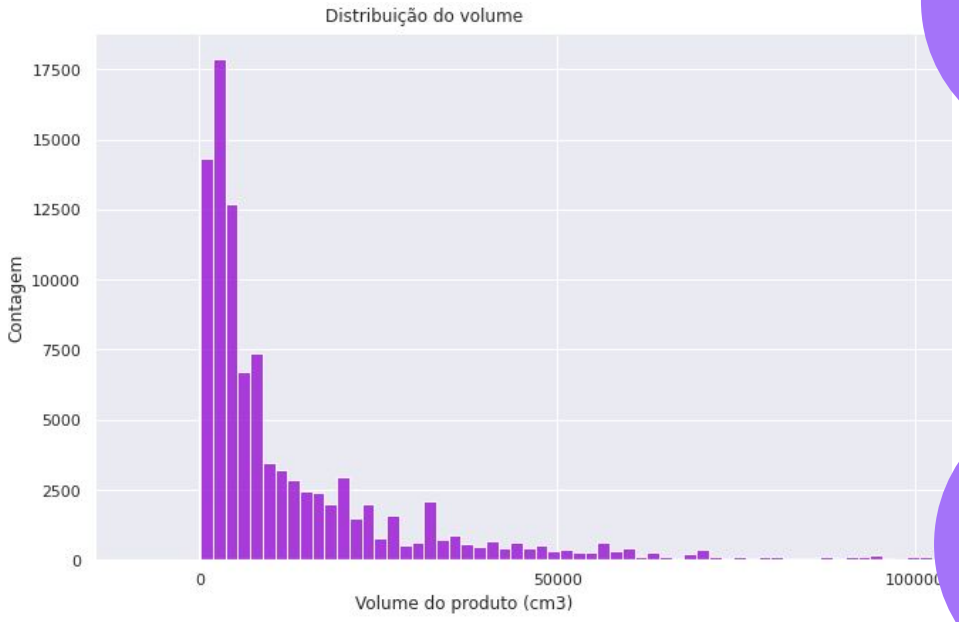


Gráfico de distribuição de produtos por valor de frete - Fonte : Autores - Cálculo de Bins: Método de Stuart

Análise de Volume

Distribuição: Volume do Produto (cm3)



Cerca de 50% dos fretes tem até 6L

Outliers com faixa de volume de 150L

Gráfico de distribuição de produtos por valor de frete - Fonte : Autores - Cálculo de Bins: Método de Stuart

Análise de Volume

Assim como no peso, há um limite fixado visível para algumas faixas de volume. O que é condizente com as principais operadoras. E também aqui entram limites diferentes dentre de um mesmo estado ou operadoras distintas.

Dispersão: Volume x Preço do Frete

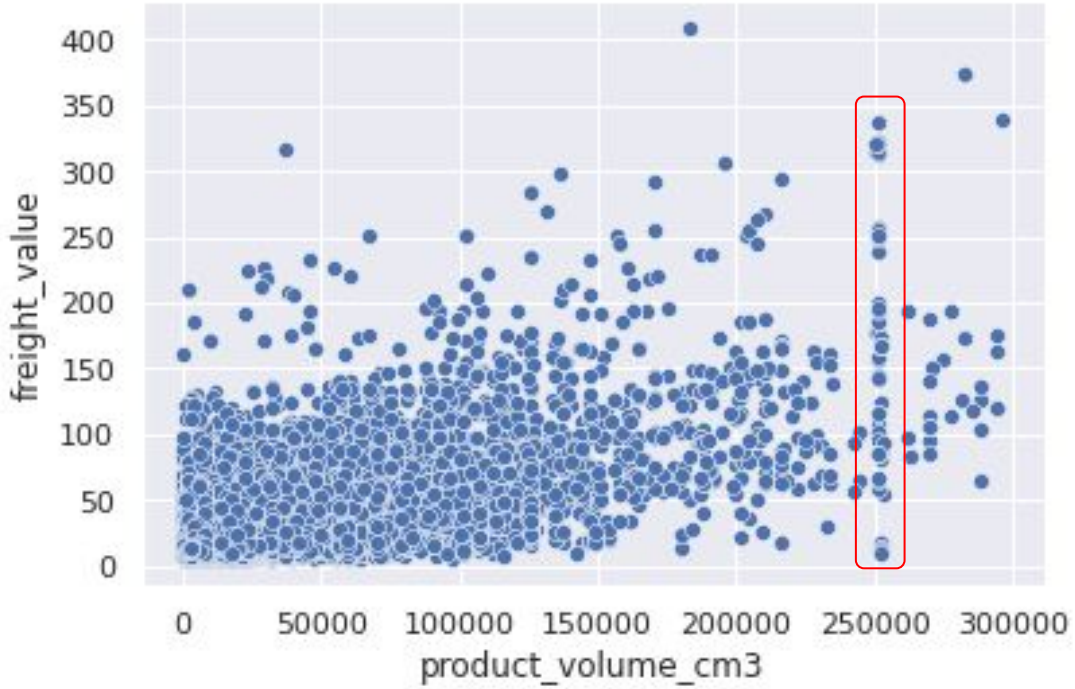
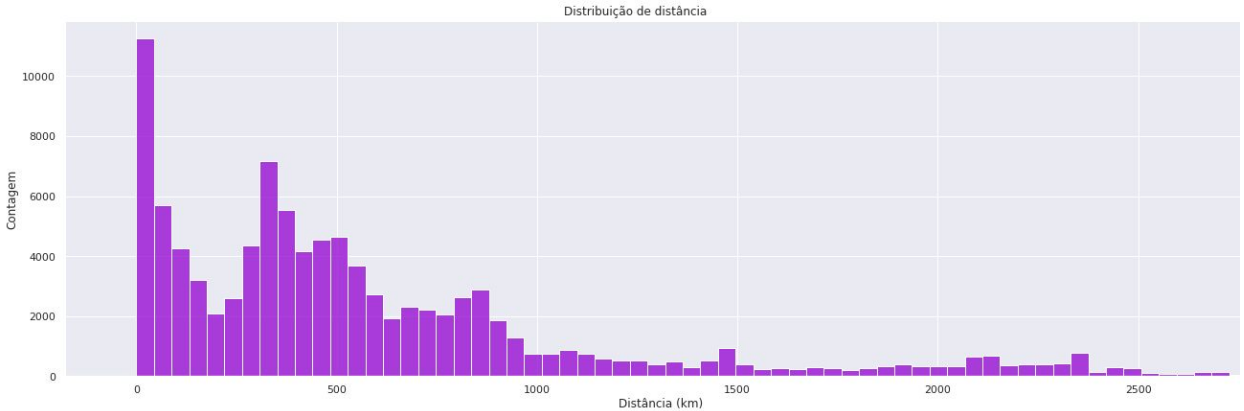


Gráfico de distribuição de produtos por valor de frete - Fonte : Autores - Cálculo de Bins: Método de Stuart

Análise de distância

Distância (km)



Outliers encontrados:

Distâncias acima de 6100 km

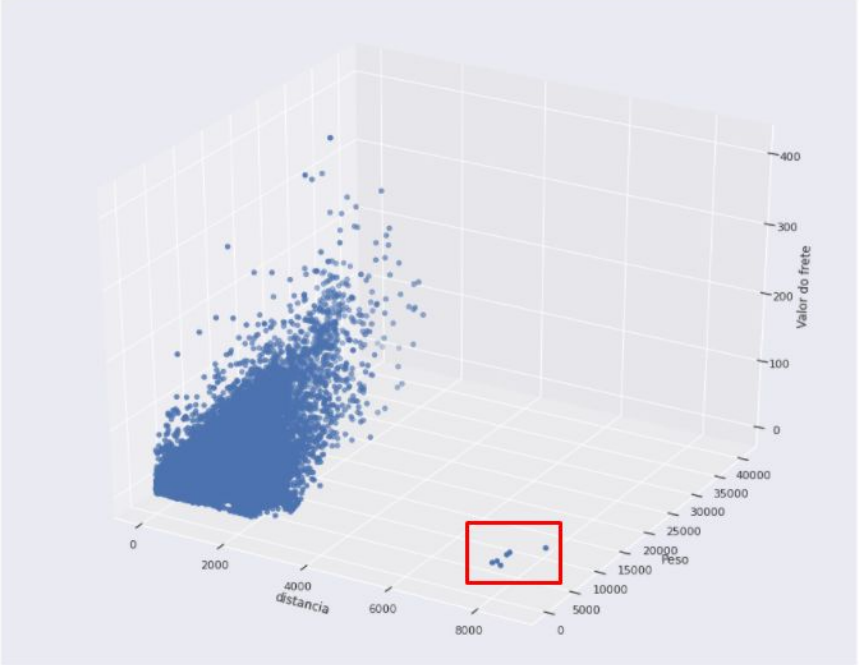
A distribuição integral, sem a filtragem, apresentava outliers com mais de 7000 km de distância. Se o Brasil fosse um quadrado, a maior distância (diagonal) entre dois pontos do país seria de aproximadamente 6100 km. Portanto entende-se que esse grupo de dados era defeituoso e representava algum frete internacional, o que não contribui ao treinamento.

Análise de dispersão 3d

Outliers encontrados:

Distâncias acima de 6100 km

Distância (km)



Análise de Prazo

3 semanas de média.

Distribuição: Prazo de entrega (dias)

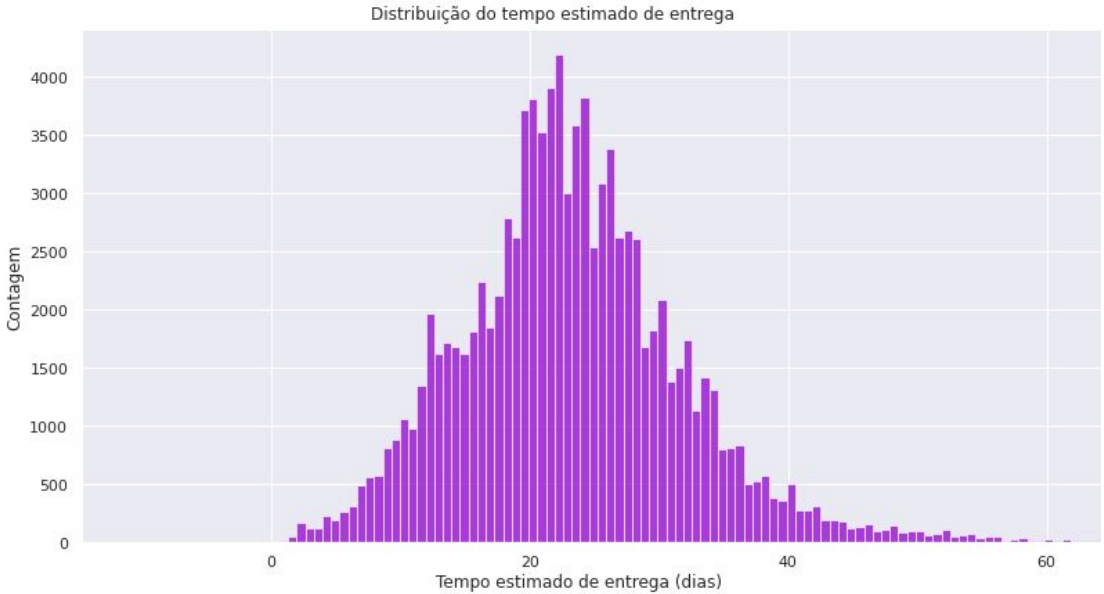
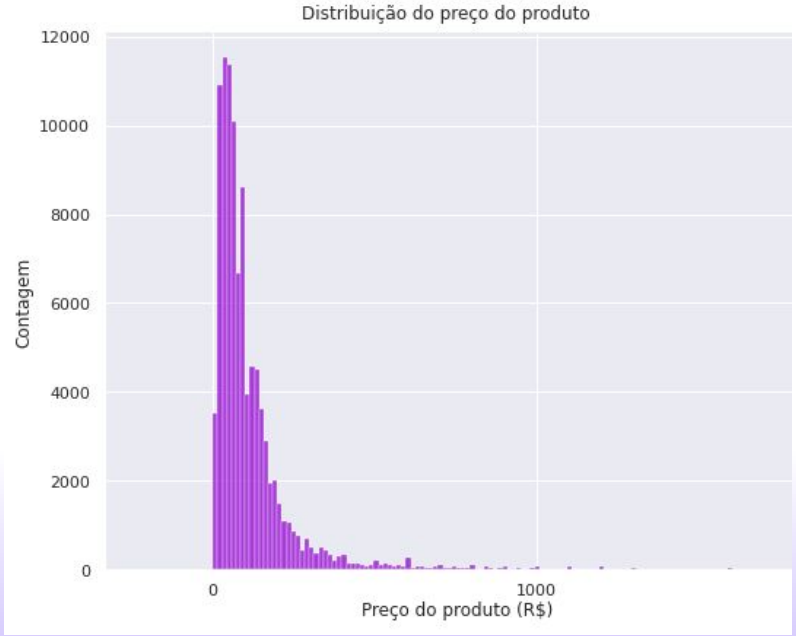


Gráfico de Distância - Fonte : Autores

Análises de distribuição

Preço do Produto (R$)



80% dos produtos vendidos: até R$170

Ticket Médio Comercializado: R$ 124,40

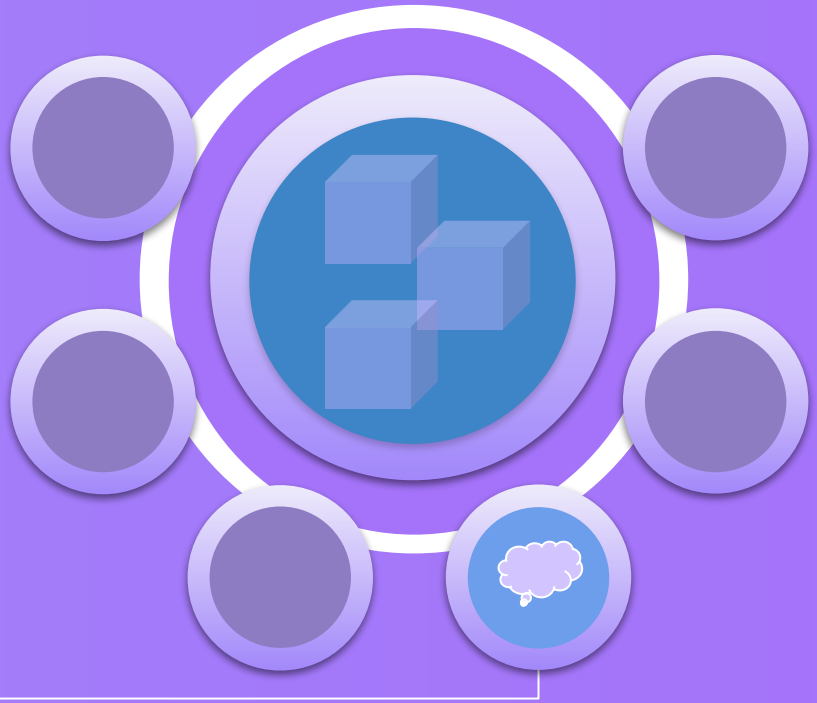
Gráfico de distribuição de preço do produto x número de produtos- Fonte : Autores

Hipóteses Testadas

Hipótese 1

Hipótese 2: Peso

Hipótese 3: Recategorização



Hipótese I

Primeiros testes de modelagem

final_df:
Aproximadamente
99.500 linhas

Realizar a modelagem por regressão com as principais características correlatas

Características:
Tempo
Peso
Preço do Produto
Distância
Volume
Largura
Altura
Comprimento

Quais os melhores modelos testados?

Regressão Polinomial (grau 3)
Random Forest
KNN
Ridge
Decision Tree
Ransac
Redes Neurais

Hipótese I

Técnicas aplicadas junto aos modelos:

Melhor Split: 80% Treino e 20% Teste
K-Fold: **5 folds**
Grid Search

Critérios de análise de resultado

- **MAE** - Erro Médio Absoluto
- **MSE** - Erro Quadrático Médio
- **R²** - Coeficiente de Determinação

Para as técnicas, aplicou-se 80% de treino e 20% de teste nos modelos, e verificou-se que essa faixa apresentava os melhores resultados de MAE.

O número de folds escolhido também apresentou a melhor faixa de treino.

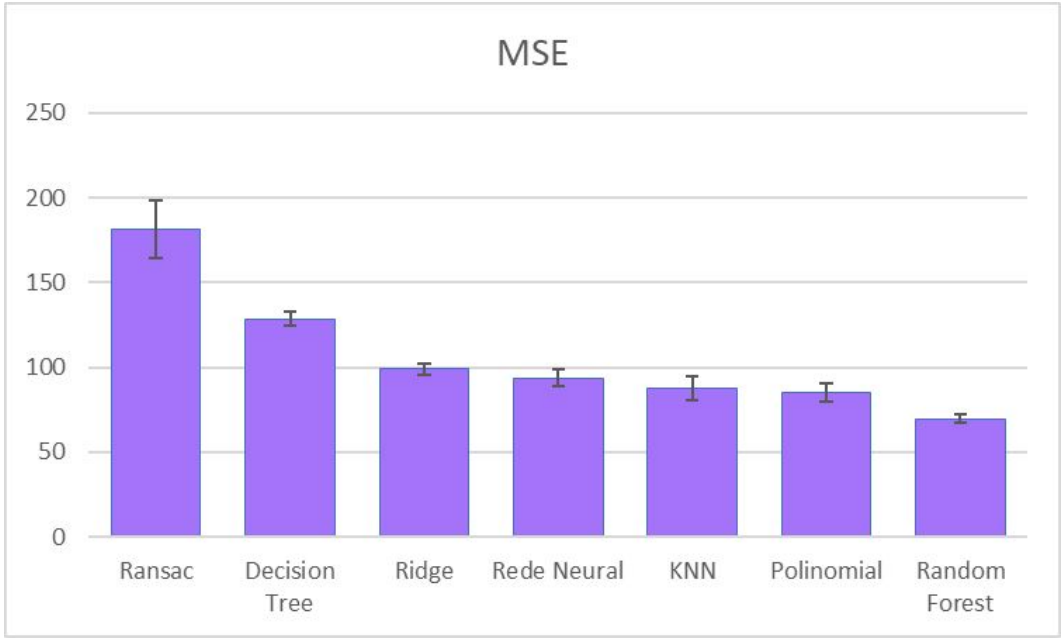
Para os modelos cuja a alteração de parâmetros poderia ser interessante (KNN e Random Forest), foi utilizado o Grid Search.

Tempo para treinar um Grid Search: ~35 minutos

Hipótese I Resultados

Erro médio quadrático

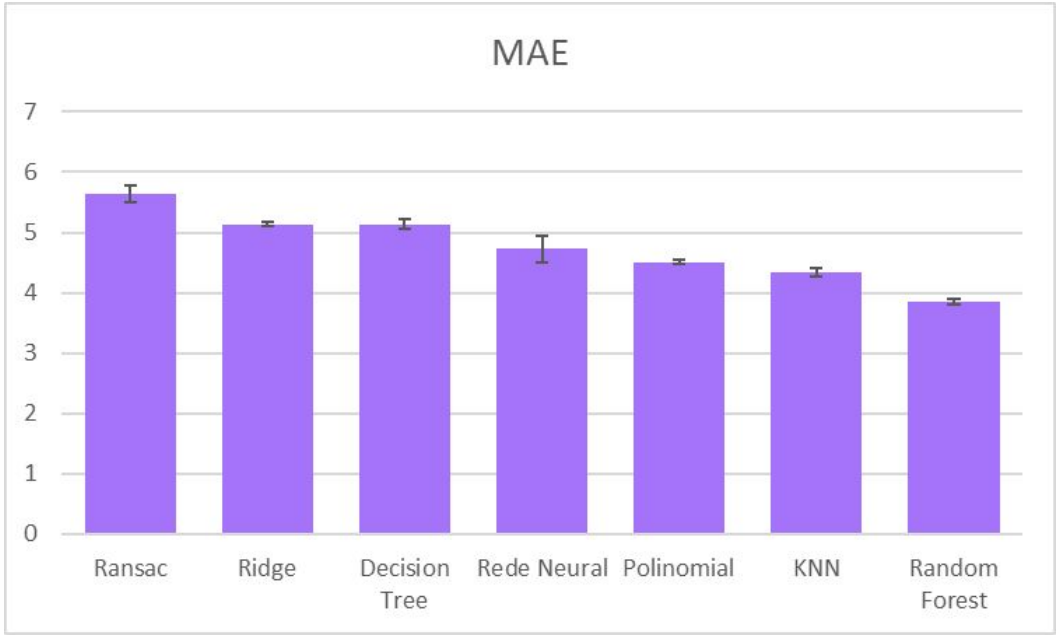
Resultados:
Ransac: 181,49 +- 16,99
Random Forest: 68,41 +- 3,62



Analisando o MSE e seus desvios, os três melhores modelos são Random Forest, Polinomial e KNN

Hipótese I Resultados

Erro absoluto médio



Resultados:
Ransac: 5,63 +- 0,14
Random Forest: 3,84 +- 0,03

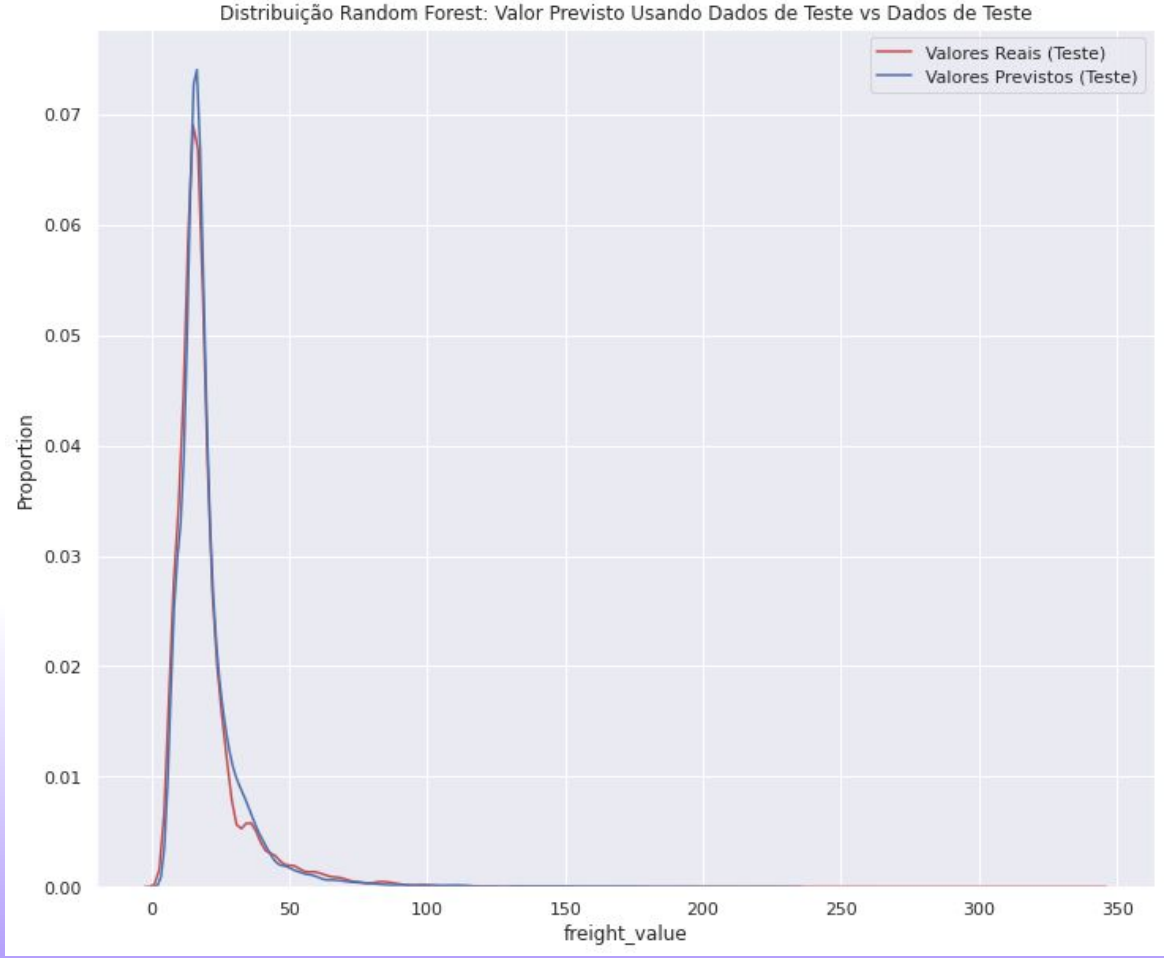
Da mesma forma que o MSE, analisando o MAE e seus desvios, os três melhores modelos são Random Forest, KNN e Polinomial, com o Random Forest se consolidando como melhor modelo.

Hipótese I Resultados

Especificações do Treino:

Random Forest: K-Fold
5 Folds

Resultados:
MSE: 68,41 +- 3,62
MAE: 3,84 +- 0,03
R^2: 0,71 +- 0,02



Hipótese II

Modelagem por faixas de peso.

Quais modelos utilizados nessa hipótese?

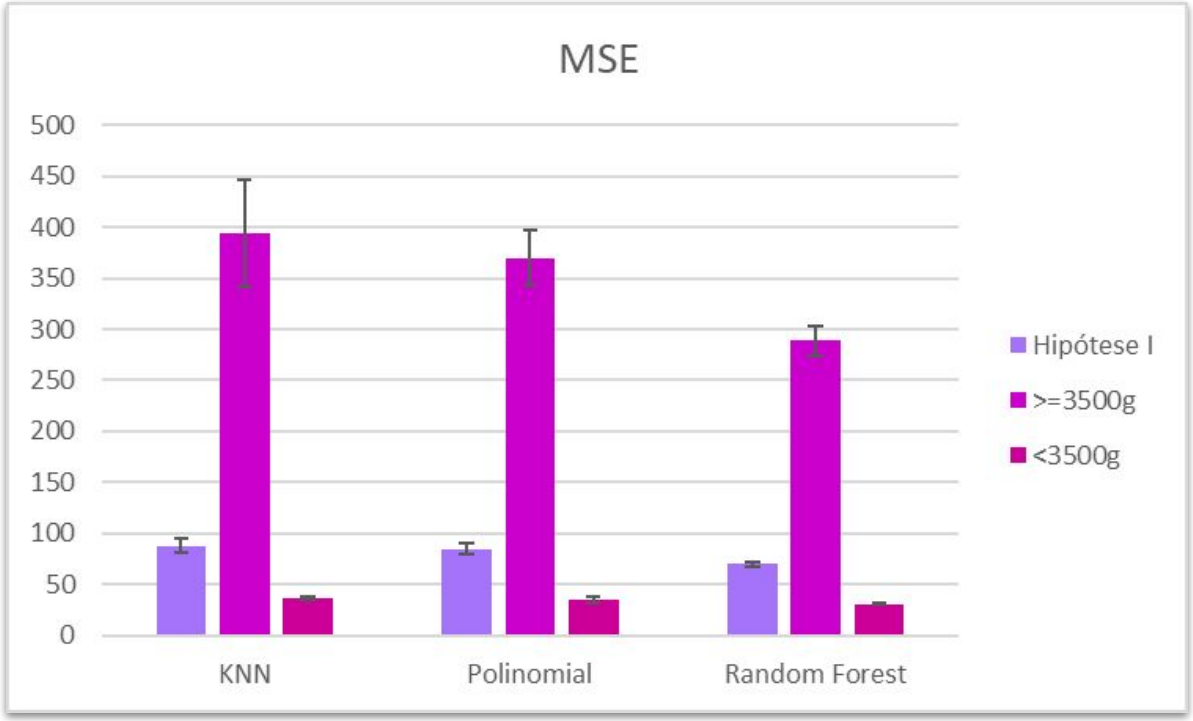
- Random Forest
- KNN
- Polinomial

Validação: K-Fold

~85% dos dados tem menos de 3.5kg

Hipótese II Resultados

Erro médio quadrático

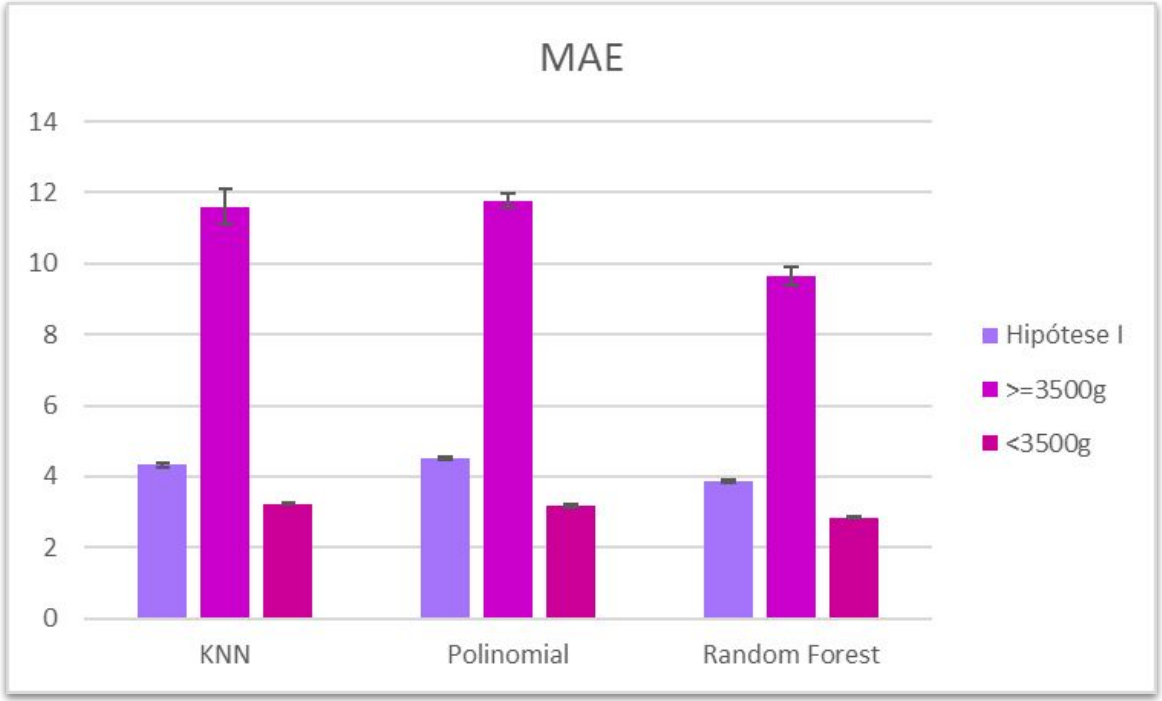


Resultados Random Forest:
Hip 1: 69,65 +- 2,36
>=3500kg:: 289,02 +- 14,5
<3500kg: 30,77 +- 1,19

Para o modelo com melhor desempenho, que é a Random Forest, e entre as Hipóteses testadas. a que apresenta melhor desempenho de acordo com o MSE é a Hipótese 2 < 3,5kg, porém está só pode operar nesta faixa de peso.

Hipótese II Resultados

Erro absoluto médio



Resultados Random Forest:
Hip 1: 3,84 +- 0,03
>=3500kg:: 9,63 +- 0,26
<3500kg: 2,85 +- 0,03

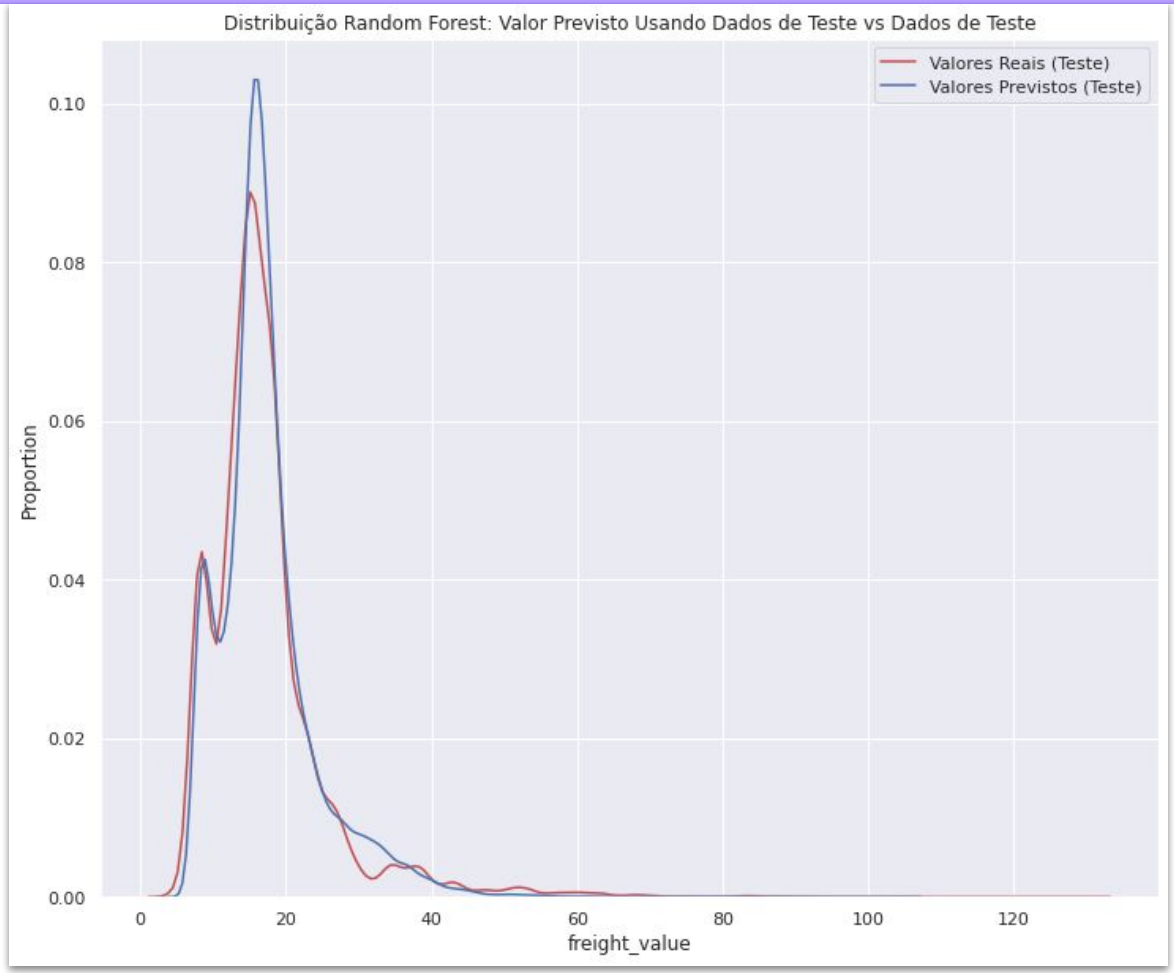
Assim como na análise para o MSE, analisando o MAE, a Random Forest da Hipótese 2 < 3,5kg apresenta o melhor resultado,, porém esta só pode operar nesta faixa de peso.

Hipótese II

Análise gráfica - Random Forest
Especificações do Treino:
Peso < 3500 gramas

K-Fold

Resultados:
MSE: 30,77 +- 1,19
MAE: 2,85 +- 0,03
R²: 0,58 +- 0,01

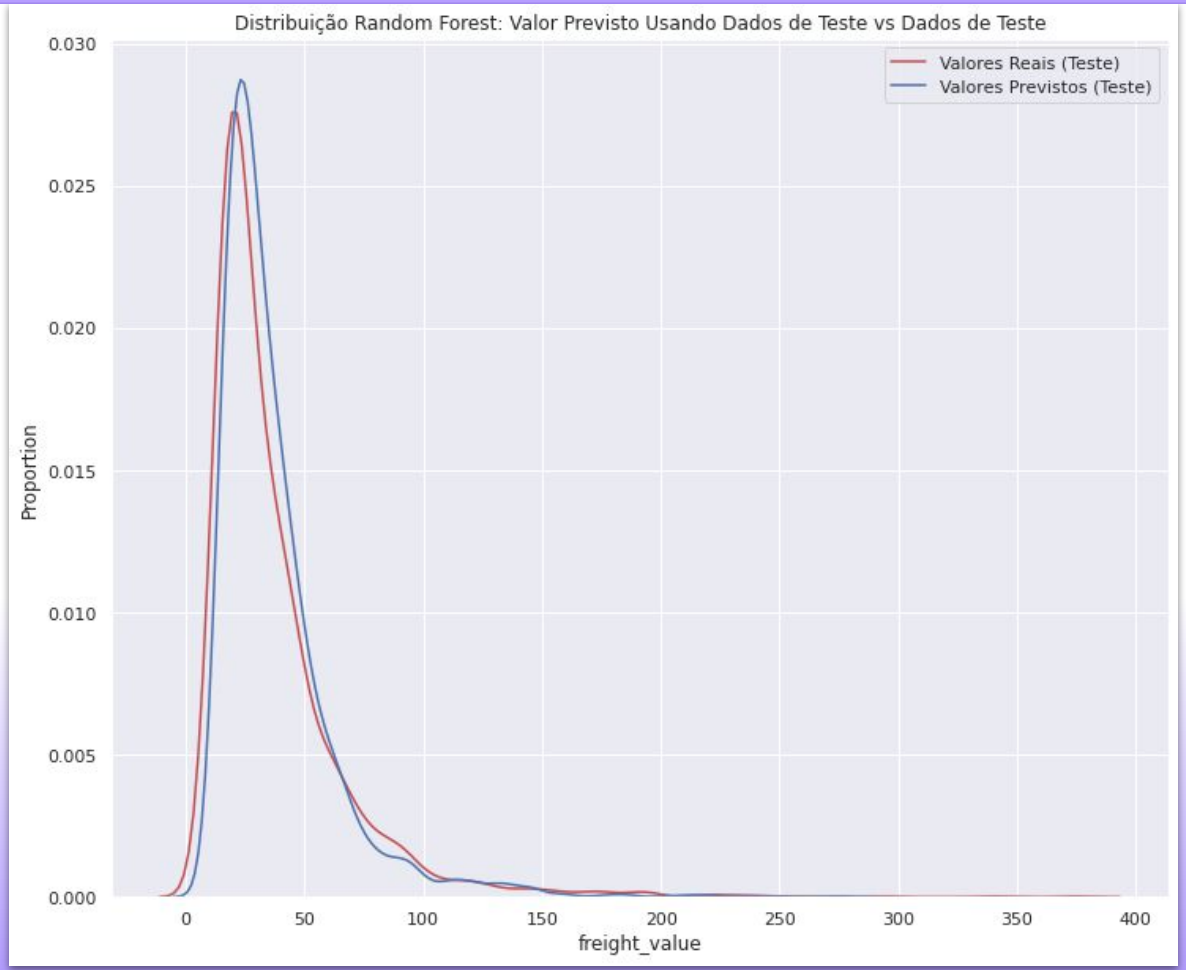


Hipótese II

Análise gráficas
Especificações do Treino:
Peso > 3500 gramas

K-Fold

Resultados:
MSE: 289,02 +- 14,5
MAE: 9,63 +- 0,26
R^2: 0,68 +- 0,01



Hipótese III

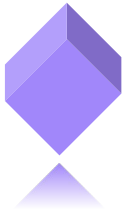
Modelagem por reorganização de características dos produtos.

Quais modelos utilizados nessa hipótese?

- Random Forest
- KNN
- Polinomial

Validação: K-Fold

Criação de dummies através de *pd.get_dummies*



Hipótese III - a

Qualificador por tipo de produto

Recategorização de 73 categorias para 4

- Frágil
- Perecível
- Eletro Eletrônicos
- Comum

Hipótese III - b

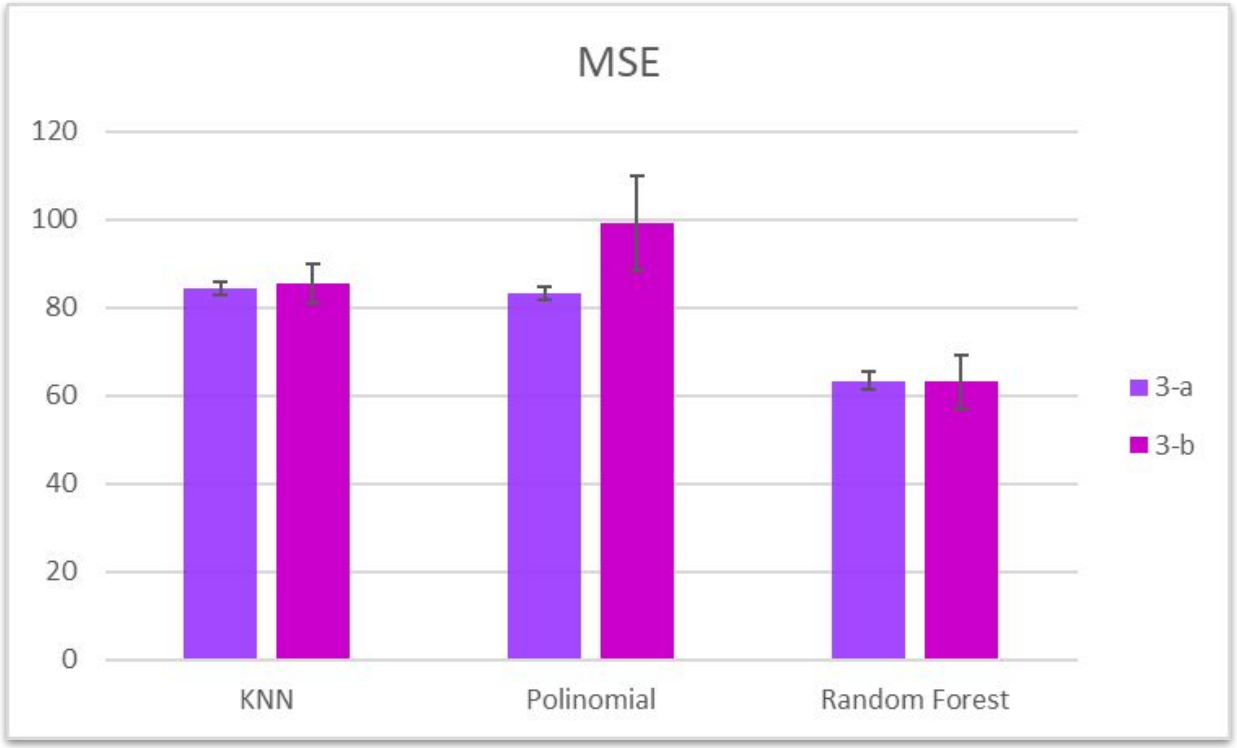
Ampliação do qualificador por tipo de produto

Recategorização de 73 categorias para 14

- Artes
- Livros e papelaria
- Cosméticos
- Casa e Decoração
- Agro e Animais
- Roupas
- Esporte
- Ferramentas e Construção
- Carro e Viagens
- Cozinha
- Infantil
- Serviços e Segurança
- Eletrônicos
- Elétricos

Hipótese III Resultados

Erro médio quadrático

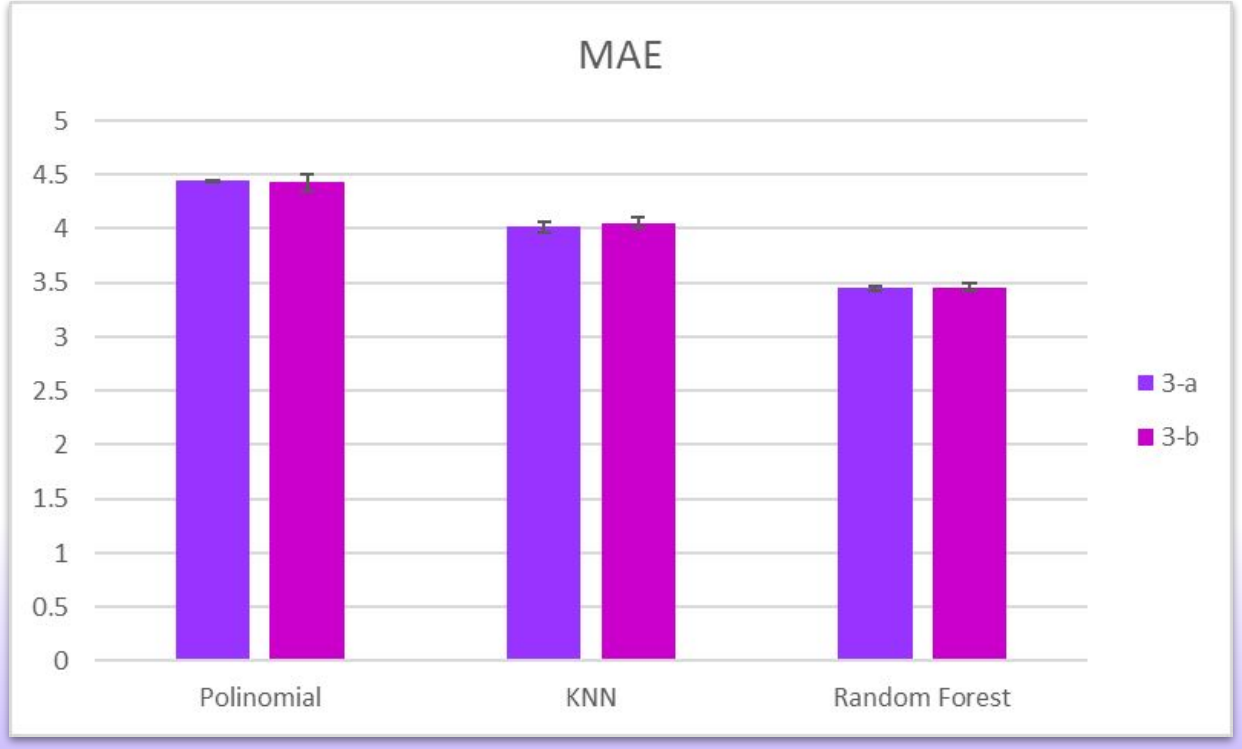


Resultados Random Forest:
Hip 3-a: 63,37 +- 2,06
Hip 3-b:: 63,1 +- 6,03

Para o melhor modelo treinado, as Hipóteses 3-a e 3-b apresentam resultados com variações em casas decimais, porém o desvio padrão para a Hipótese 3-a é menor.

Hipótese III
Resultados

Erro absoluto médio



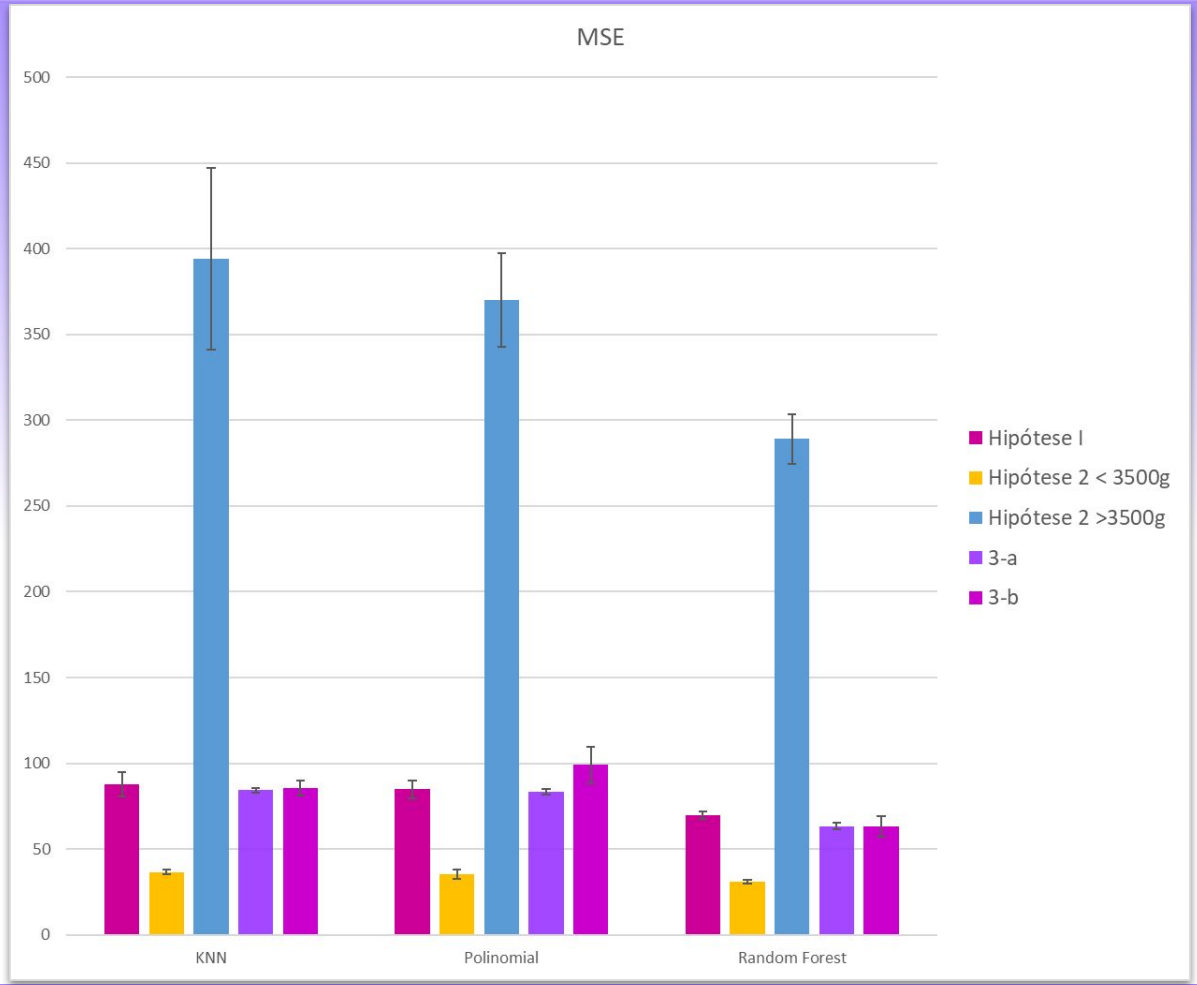
Resultados Random Forest:
Hip 3-a: 3,45 +- 0,02
Hip 3-b:: 3,46 +- 0,04

Analisando o MAE, também ocorrem variações em casas decimais entre as Hipóteses 3-a e 3-b, e da mesma forma que o MSE, o desvio padrão para a hipótese 3-a é menor.

Análise de Resultados

Análise de Resultado

Erro médio quadrático

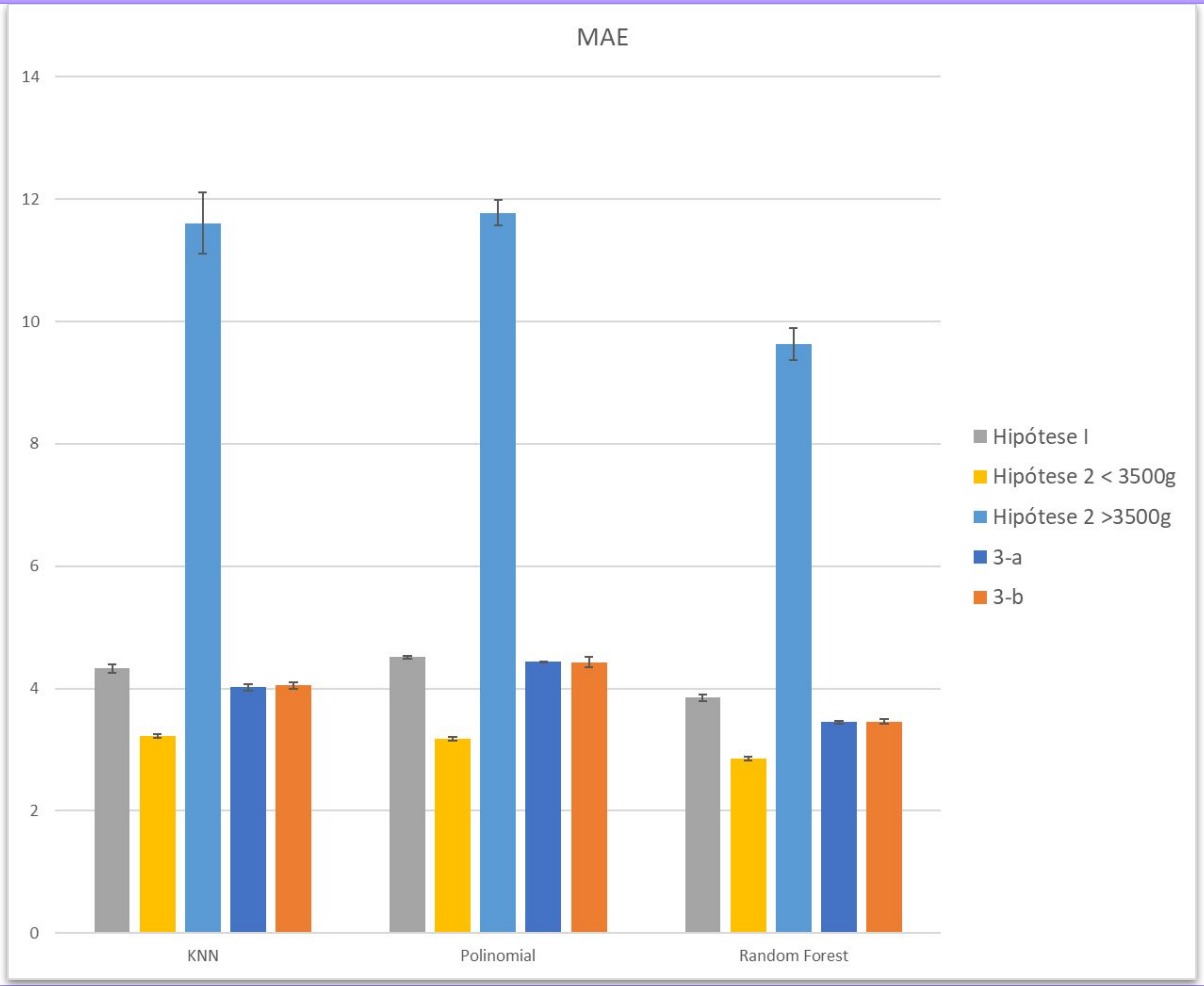


Resultados [MSE] para Random Forest:
Hip 1: 68,41 +- 3,62
Hip 2 < 3,5kg: 30,77 +- 1,19
Hip 2 >= 3,5kg: 289,02 +- 14,5
Hip 3-a: 63,37 +- 2,06
Hip 3-b:: 63,1 +- 6,03

Análise de Resultado

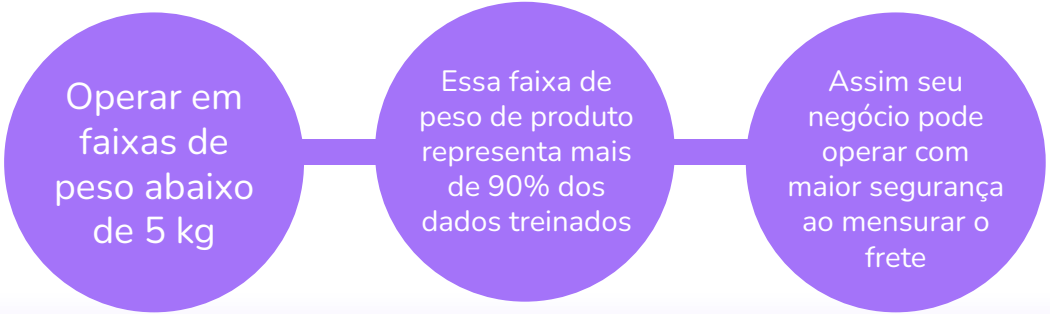
Erro absoluto médio

Resultados Random Forest:
Hip 1: 3,84 +- 0,03
Hip 2 >= 3,5kg:: 9,63 +- 0,26
Hip 2 < 3,5kg: 2,85 +- 0,03
Hip 3-a: 3,45 +- 0,02
Hip 3-b:: 3,46 +- 0,04



Análise de Resultado

Os modelos de precificação obtidos são recomendados ao cliente para:



Além dessa faixa de peso sugerida é importante lembrar que os dados indicam uma média de R$124,40 para o ticket.

Manter-se nessa faixa de preço para os produtos assegura uma melhor relação do consumidor com o preço a ser pago pelo frete.

Uma vez que o frete médio real esperado é de R$14,40, e nosso erro de preço de aproximadamente R$3,50 é inferior ao frete mínimo (2017-2018) de R$5,90, a perda de vendas por causa de valor do frete seria consideravelmente amortizada.

Na grande maioria dos casos, ao considerar o erro, os fretes estariam na faixa de 11 a 18 reais, o que para um produto perto do valor médio de ticket (124,40) seria algo em torno de 9 a 14% do produto.

Extras

Indicadores

Medidas:

- *Lead Time*
- *Cycle Time*
- *Volumetria*

O que foi medido?

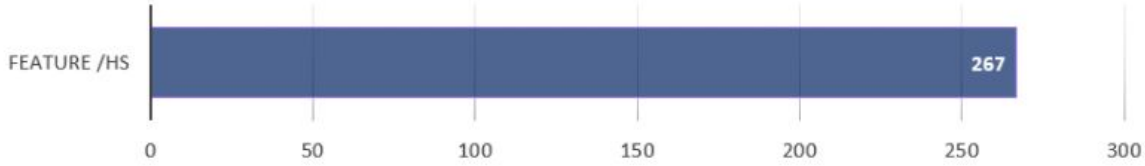
Lead Time:
Tempo da História , desde o 1º dia ate a entrega.

Cycle Time
A qualificação do Lead Time. Como o Lead time foi dividido?

Volumetria:
Quantificação de linhas de código,
Quantidade de gráficos realizados

KPI

Lead Time



Cycle Time

Volumetria

Linhas de código finais	N° de gráficos finais	Tempo testando modelos	N° modelos treinados
800+	50+	20h	200+

Agradecimentos

A todos que participaram e fizeram parte dessa jornada.

Equipe:
Andréia Castanharo, Arthur Gebhard, Luan de Brito

referências imersão

Como calcular fretes:
https://blog.pagseguro.uol.com.br/veja-como-calcular-o-frete-pagar-menos-e-ainda-melhorar-o-envio-de-produtos/

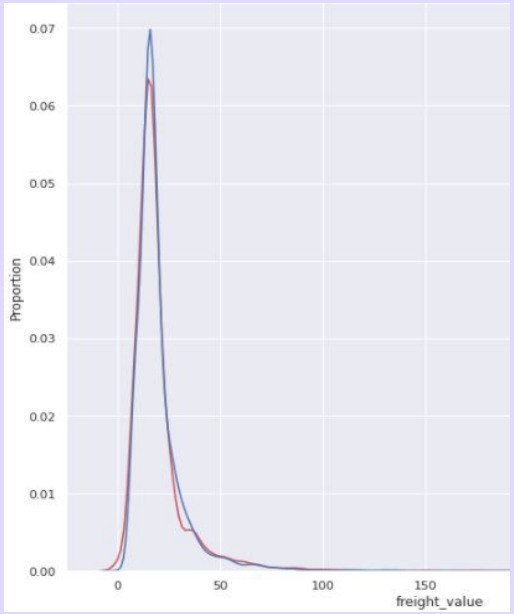
Tipos de Frete:
https://www.nuvemshop.com.br/blog/tipos-de-frete/

Design dos Slides

Elemento 2 - Fonte 12 a 14 Cor intermediária, fonte menor, foco em dar informações secundárias

Element Size 3 - Textual e detalhamento de informações - sem negrito, fonte menor e mais fina

Título do Gráfico



EF II - Compete com Element Size 3

EF II - Compete com Element Size 3

Elemento de Foco I - Compete com Elemento II

EF II - Compete com Element Size 3

EF IIb - maior contraste para elementos maiores

Element Size 3a - Textual e detalhamento de informações - sem negrito, fonte menor e mais fina

Element Size 3b - Fonte 8 a 10 - Textual e detalhamento para maior contraste

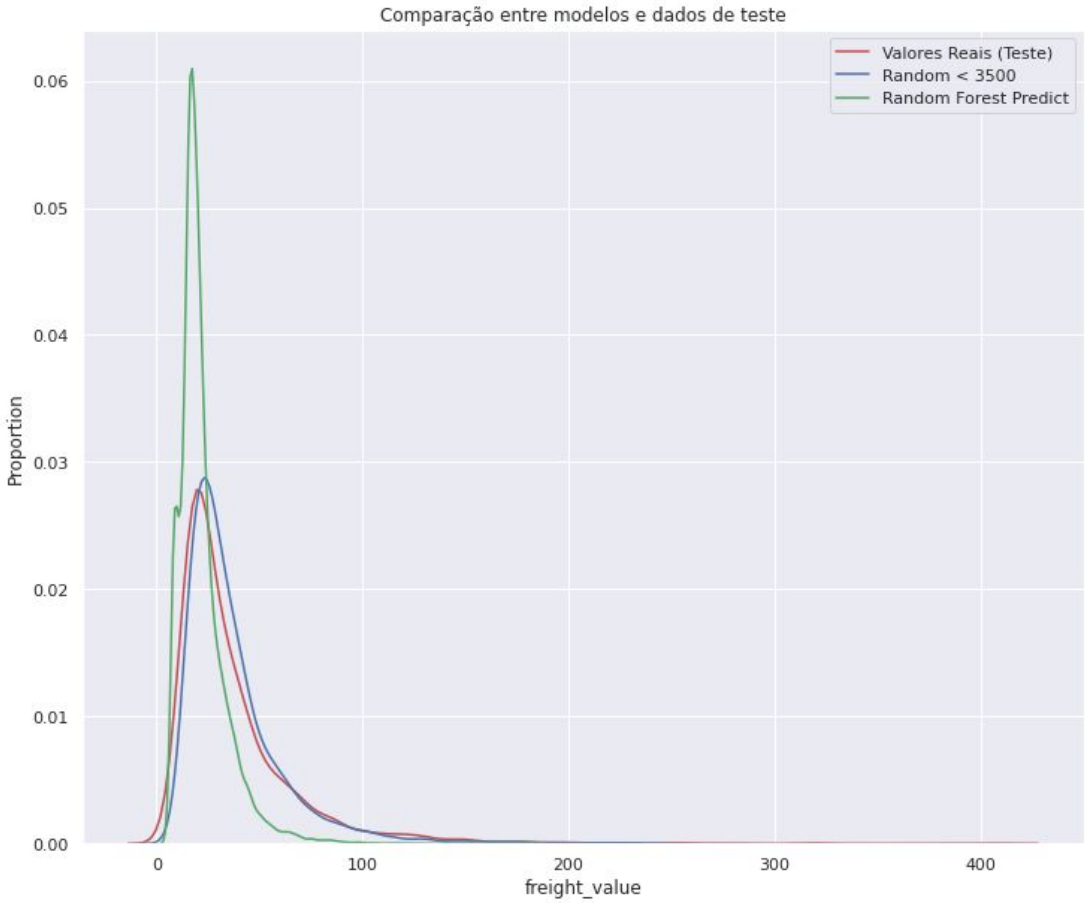
materiais extras produzidos

Hipótese II

Análise gráfica - Random Forest
Especificações do Treino:
Peso < 3500 gramas

K-Fold

Resultados:
MSE: 30,77 +- 1,19
MAE: 2,85 +- 0,03
R²: 0,58 +- 0,01

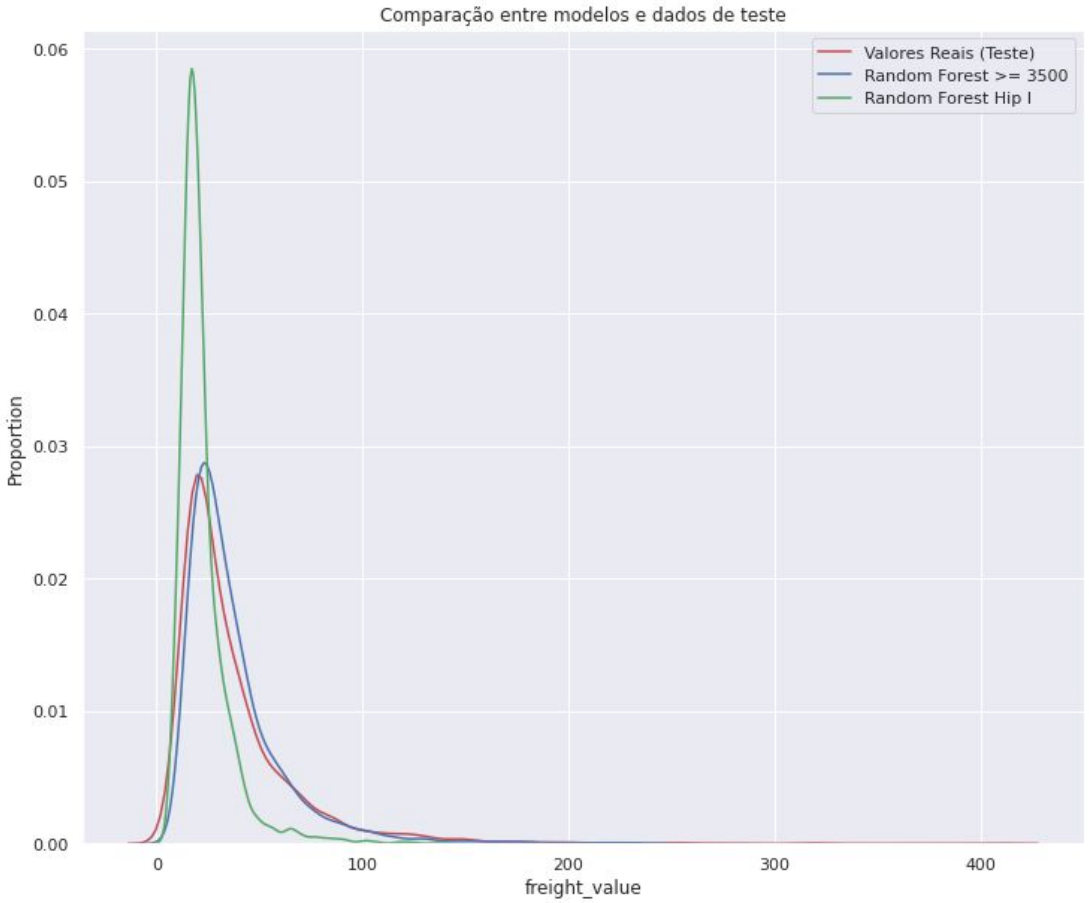


Hipótese II

Análise gráfica - Random Forest
Especificações do Treino:
Peso < 3500 gramas

K-Fold

Resultados:
MSE: 30,77 +- 1,19
MAE: 2,85 +- 0,03
R²: 0,58 +- 0,01



Machine Learning Canvas

Análises Propostas

Influência das características dos produtos no preço

Fontes de Dados

Ordens de venda e características de produtos vendidos

Recursos Técnicos

Python, Modelos de Regressão,Excel, Colab

AI Value

Prever o valor do frete de um produto através de suas características com um baixo erro e variância em relação em relação a maior parte dos dados.

Principais Riscos

Prejuízos para vendedor e comprador
Não conseguir entregar tudo

Principais Stakeholders

Empresa de e-commerce, comprador

Canal de Entrega

Relatório estático, apresentação e notebook do colab

Métricas de Sucesso

Baixo erro de precificação
Determinar melhor faixa de operação