14/10/2022

Relatório

Bootcamp - Precificação dinâmica

Ana Chaves / Paulo Angellotti/ Victor Mitsuo

SPRINT 3

BOOTCAMP - PRECIFICAÇÃO DINÂMICA

MERCARI PRICE SUGGESTION CHALLENGE - KAGGLE

Com base no dataset da Mercari, será criado um produto de sugestão de preços aos vendedores, que será oferecido a sites de e-commerce.

“O preço do produto fica ainda mais difícil em escala, considerando quantos produtos são vendidos online. As roupas têm fortes tendências de preços sazonais e são fortemente influenciadas por marcas, enquanto os eletrônicos têm preços flutuantes com base nas especificações do produto.”

ENTENDENDO O NEGÓCIO: MERCARI

Entre as muitas plataformas de comércio eletrônico online, Mercari é um famoso site de compras e venda online, uma escolha popular para encontrar itens a preços mais baratos do que em outras plataformas. Tudo começou em 2013 e, devido aos seus métodos de compra e venda muito fáceis, rapidamente ganhou fama e agora tem cerca de 16 milhões de usuários ativos por mês. Mercari expandiu seus serviços e agora possui Mercari USA e Mercari UK.

A Mercari tem várias lojas de conveniência parceiras onde o produto pode ser enviado. Neste caso, a taxa de envio está incluída na venda do item e o vendedor não paga nenhuma taxa para mandar o mesmo.

Os vendedores da Mercari Japão estão todos localizados no Japão. Uma pequena parte deles já enviou encomendas para o estrangeiro. De fato, a maioria dos vendedores na Mercari são indivíduos. Além disso, uma envio internacional não tem os mesmos custos que um envio doméstico. Portanto, mais difícil para o vendedor calcular o seu custos. É por isso que a maioria deles não se dá ao trabalho de negociar com compradores estrangeiros. Podemos observar que é um aplicativo mais restrito a venda local.

Os produtos ofertados podem ser novos ou usando, respeitando um sistema de classificação, no dataset chamado de , cujo vai de 1 a 5, sendo 1 ruim e 5 ótimo.

BRAINSTORM

Durante a análise exploratória, foram observados alguns pontos e levantado algumas dúvidas de como funciona o negócio e como pode ser feito um modelo para identificar o preço de uma maneira mais precisa e rápida.

- Onde estão o grande volume de vendas?
- Como o mercari ganha dinheiro?
    - *O valor pelo qual você vendeu o item será creditado na sua conta Mercari após dedução de 10% da taxa Mercari.*

- Comissão? Porcentagem? Anúncio?
    - *10% da taxa*

- Mercado restrito? Ou generalista?
- Quão importante é a métrica?
- Quão importante a margem de erro?
- O frete interfere no preço?
    - A maioria dos itens do Mercari é vendida com custos de envio incluídos no preço do item. Você pode optar por cobrar o frete separadamente, mas suas chances de venda são menores. O preço mais baixo que você pode cobrar por um item é 300 Yen

- As marcas influenciam no preço?
- Rede neural poderia ser utilizada, entretanto um cliente não iria esperar 90 minutos pelo resultado.
- Verificar para o vendedor não perder dinheiro e deixar de utilizar o aplicativo.
- Como será normalizado os dados?
- Dataset pesado, qual melhor método para limpar e processar?

DICIONÁRIO DE DADOS

Name	Texto	O título da listagem. Obs: Observe que limpamos os dados para remover textos que parecem preços (por exemplo, US$ 20) para evitar vazamentos. Esses preços removidos são representados como [rm]
Item_condition_id	1 - 2 – 3 – 4 – 5	A condição dos itens fornecidos pelo vendedor. Categoria da condição do item, entre 1 que significa ruim até 5 ótimo.
Category_name	Texto	Categoria da listagem
Brand_name	Texto	Marca
Price	Números	O preço pelo qual o item foi vendido. Esta é a variável de destino que você irá prever.
Shipping	0 ou 1	Frete: 1 se a taxa de envio for paga pelo vendedor e 0 pelo comprador.
item_description	Texto	A descrição completa do item.
Date	Date	Criado para este desafio.
Stock	Números	Criado para este desafio, é o que tem em estoque disponível.
gen_cat		Categoria principal do item.
sub1_cat		Subcategoria nível 1
sub2_cat		Subcategoria nível 2

ANÁLISE EXPLORATÓRIA

Ao abrir o *dataset*, podemos observar que os produtos contém nome, condição do item, a qual categoria pertence, sendo a primeira categoria principal e as demais subcategorias.

Alguns produtos possuem marca, mas em sua grande maioria não possui. O atributo de envio, possui duas categorias 1 e 0, cujo 0 é o envio por conta do comprador, e 1 envio pago pelo vendedor, e a descrição do item. Foi solicitado para colocarmos data e número de estoque aleatório.

É um *dataset* com mais de um milhão e meio de linhas e sete atributos, sendo solicitados a inclusão de datas aleatórias e estoque.

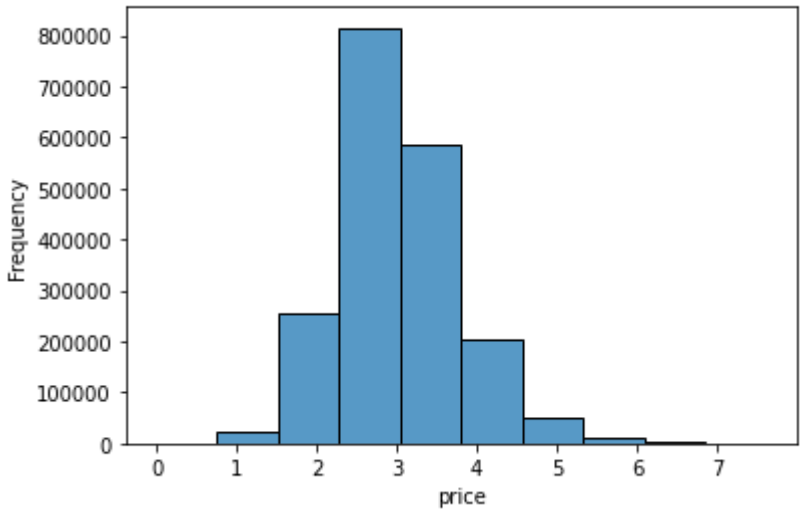
Notamos que o valor mínimo das vendas é de US$ 0, esse valor sugere que produtos são doados através do serviço de *ecommerce* e não são uteis para nosso modelo preditivo.

Também notamos uma mediana de US$ 17 e uma média de US$ 26,7 indicando uma dispersão assimétrica dos valores.

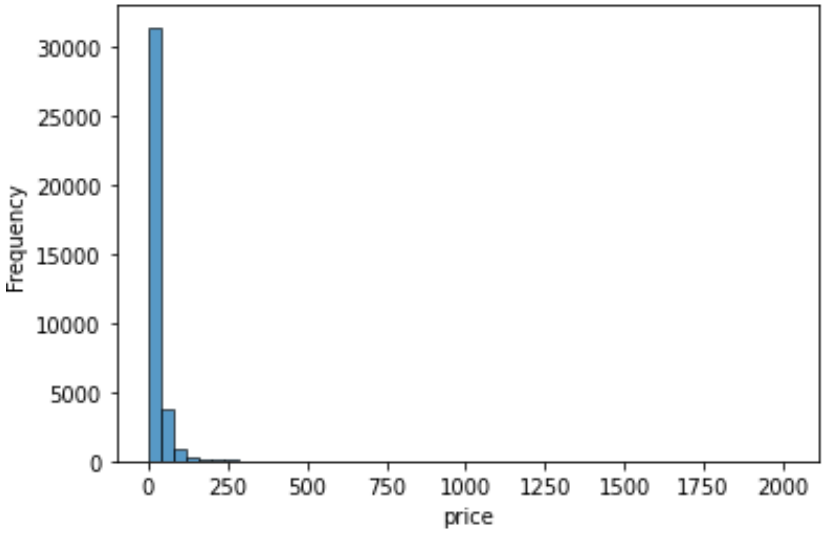
Há uma grande amplitude entre os valores de média e mediana e o valor máximo do *dataset*, esse tipo de comportamento pode gerar problemas na visualização dos valores e pode ser corrigido com a aplicação de uma transformação logarítmica na variável.

A função log1p do *numpy* realiza essa transformação (adicionando 1 a variável envolvida para evitar valores negativos e o 0).

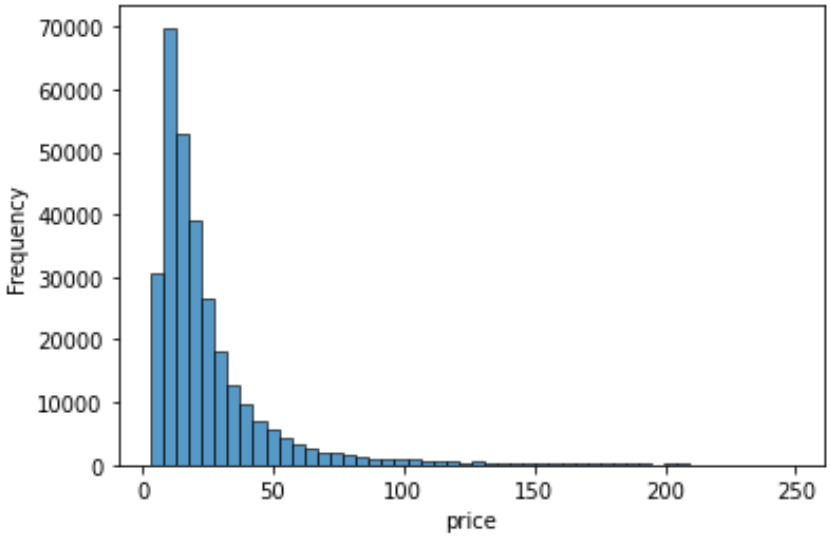
ANÁLISE DA VARIÁVEL 'PRICE'



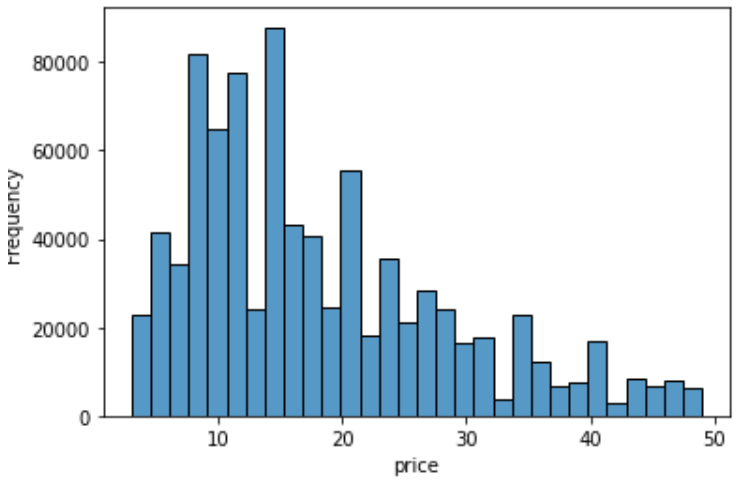
Vemos que a distribuição do preço transformado tende a seguir uma distribuição normal, sendo muito mais adequado para a aplicação em um modelo de *machine learning*.



Este histograma ilustra o problema, há uma grande amplitude entre os valores mínimos e máximos além dos valores serem distribuídos de forma assimétrica. Como a grande maioria dos valores está concentrado na faixa entre $0 e $250, a análise a seguir se concentrará nesta faixa.



Mesmo faixa de valores entre $0 e $250 ainda notamos uma forte distribuição assimétrica, porém, a visualização dos dados começa a ficar mais clara, deixando mais explicito o perfil de vendas das pessoas que utilizam este tipo de serviço.

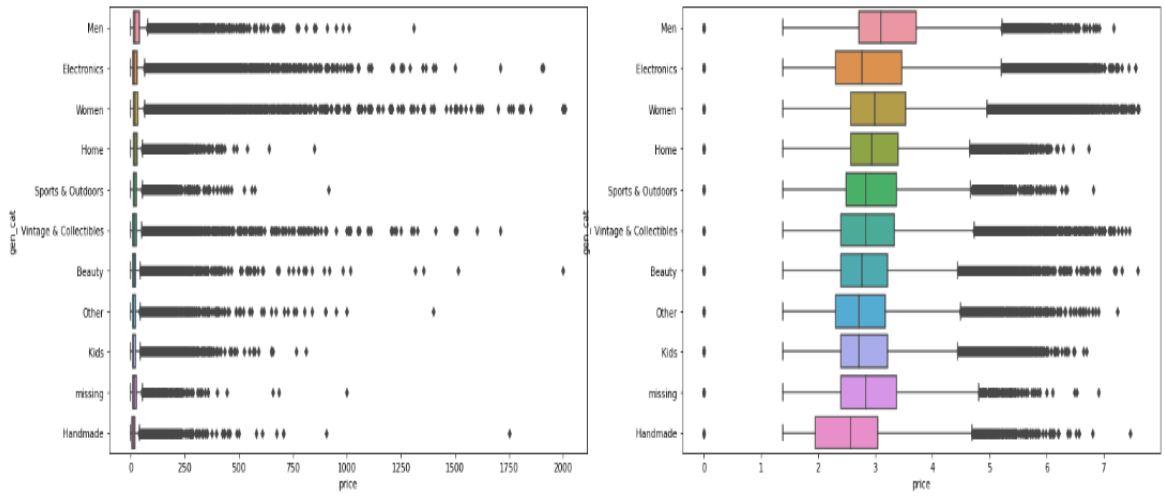


Total de observações: 1.482.535
Total de observações com preço até $250: 1.475.215
Total de observações com preço até $50: 1.322.308

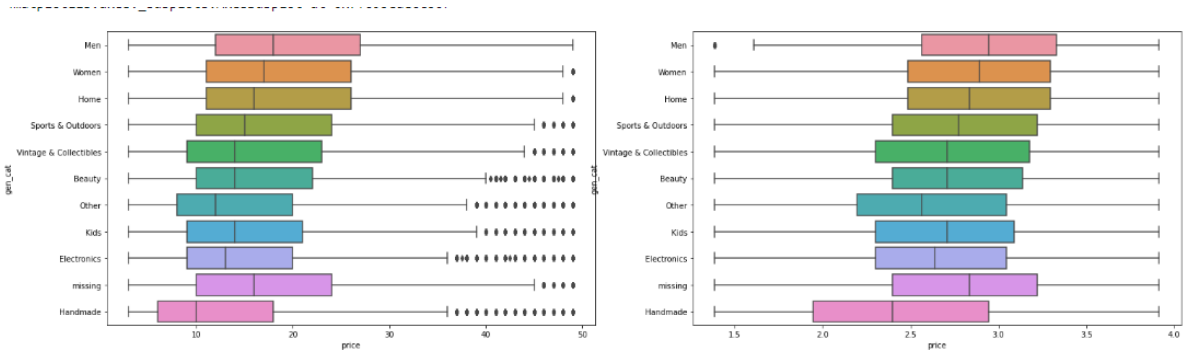
% de observações com preço até $250: 99.51 %
% de observações com preço até $50: 89.19 %

Visto a diferença mínima no nº de observações nas faixas de preço estudadas, outra possibilidade é a utilização de um modelo descartando todos os valores acima de $250 ou $50, este tipo de abordagem nos permite focar na faixa onde temos maior volume de usuários, minimizando o erro especificamente neste público.

ANÁLISE DA VARIÁVEL 'PRICE'



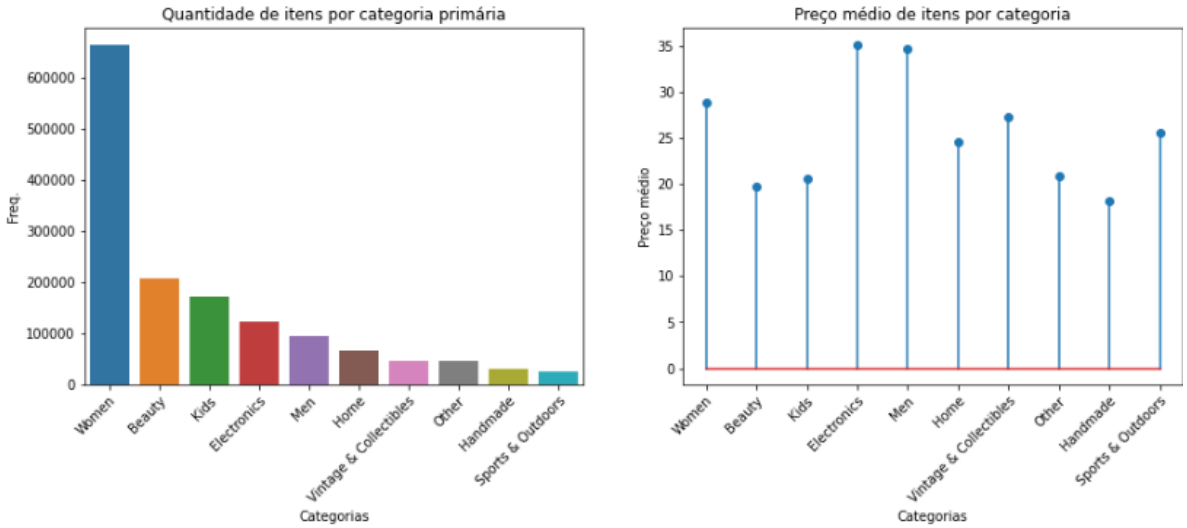
Em ambas as visualizações, notamos uma grande presença de outliers. Porém, por tratar-se de um serviço de vendas de diversos tipos de produto, espera-se que tais discrepâncias sejam naturais e por hora serão mantidas na baseline inicial.



Ao analisar conjuntos reduzidos de preços, notamos uma redução expressiva no número de outliers.

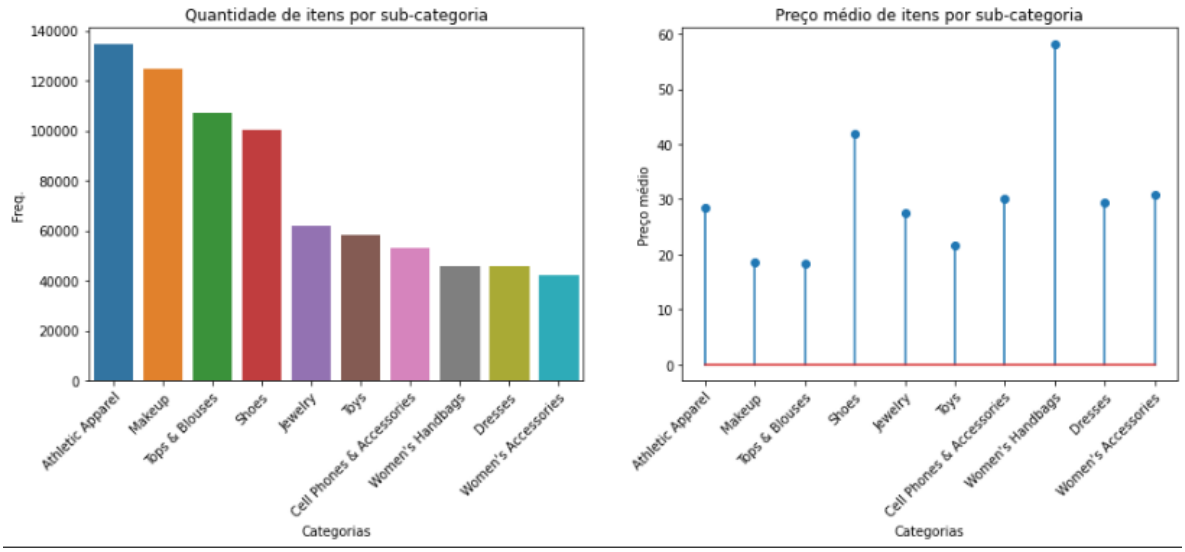
ANÁLISE DAS CATEGORIAS

Aqui apresentou a categoria que mais tem volume no dataset, que são os produtos femininos, e o outro o preço médio dos produtos dessas mesmas categorias.

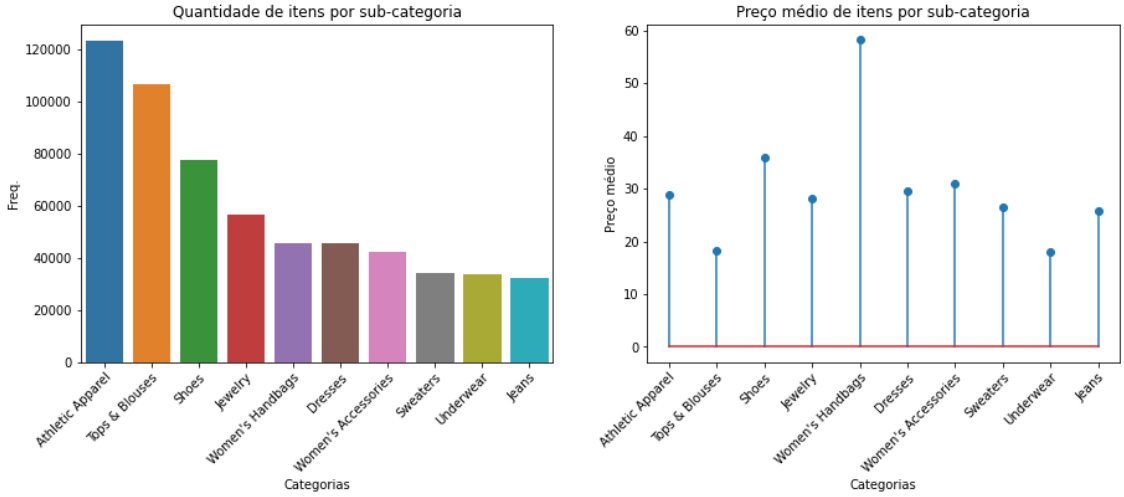


A frequência permite a retirada de alguns insights sobre o comportamento dos dados. Podemos ver que a variável ‘Women’ aparece numa frequência muito acima das demais e pode vir a gerar ruídos nos resultados do modelo de predição a ser desenvolvido. Além disso vemos que a distribuição de preços entre as categorias não apresenta uma grande variação e a maioria está presente numa mesma faixa.

Distribuição por subcategorias.



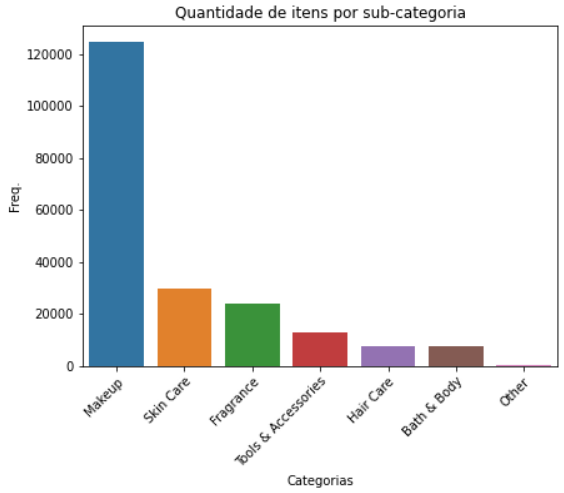
A análise geral das sub-categorias mostra um equilíbrio muito mais saudável em relação ao equilíbrio das categorias principais, porém, seus preços mostram variações mais extremas. Isso mostra que as sub-categorias possuem uma capacidade maior de discriminação de preço do que as categorias gerais.

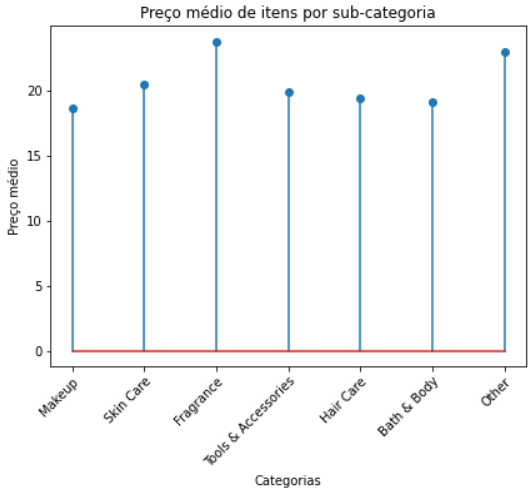


Aprofundando-se dentro da categoria 'Women', podemos notar o perfil dentro desta categoria geral. Dentro do top 10 de itens mais vendidos dentro da categoria temos: Vestuário esportivo, tops e blusas, sapatos, joias, bolsas de mão, vestidos, acessórios, suéter, roupa íntima e jeans. Também notamos a grande amplitude, já esperada, dos preços médios dessas sub-categorias.

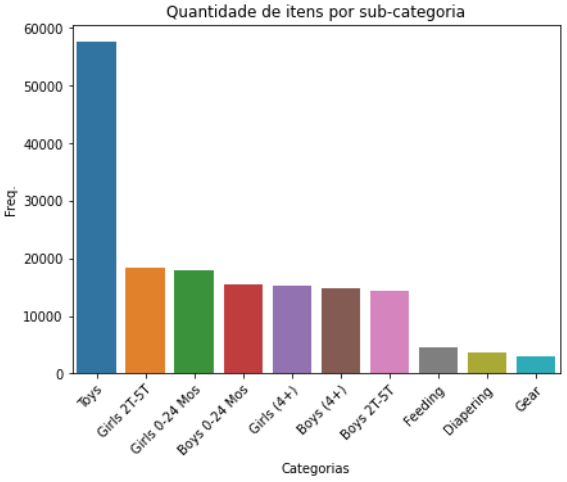
Aprofundando-se dentro da categoria 'Beauty', a grande maioria das observações pertence a sub-categoria 'Makeup'. Dentro do top 10 de itens mais vendidos dentro da categoria temos: Maquiagem, cuidados com a pele, fragrancias, acessórios, cuidados com o cabelo, banho e corpo e outros. Dentro dessa categoria os preços são mais homogeneos, sem as grandes

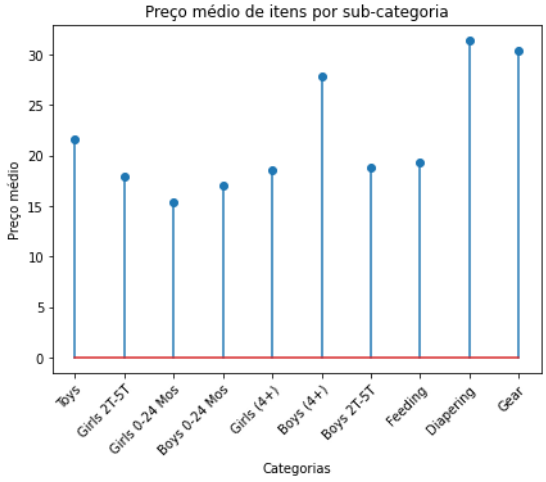
amplitudes vistas anteriormente.





Aprofundando-se dentro da categoria 'Beauty', a grande maioria das observações pertence a sub-categoria 'Makeup'. Dentro do top 10 de itens mais vendidos dentro da categoria temos: Maquiagem, cuidados com a pele, fragrâncias, acessórios, cuidados com o cabelo, banho e corpo e outros. Dentro dessa categoria os preços são mais homogêneos, sem as grandes amplitudes vistas anteriormente.

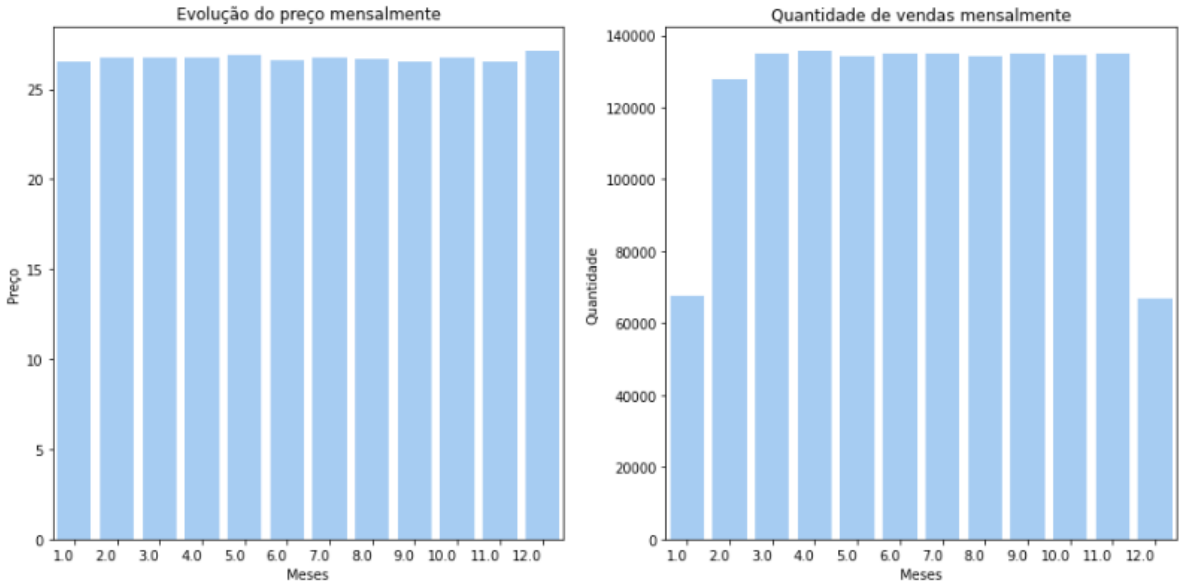




Dentro da categoria 'Kids', também há grande predominancia de uma sub-categoria. Os preços dentro dessa categoria são bem distribuídos e apesar de não apresentarem grandes variações como visto anteriormente, também não são altamente uniformes.

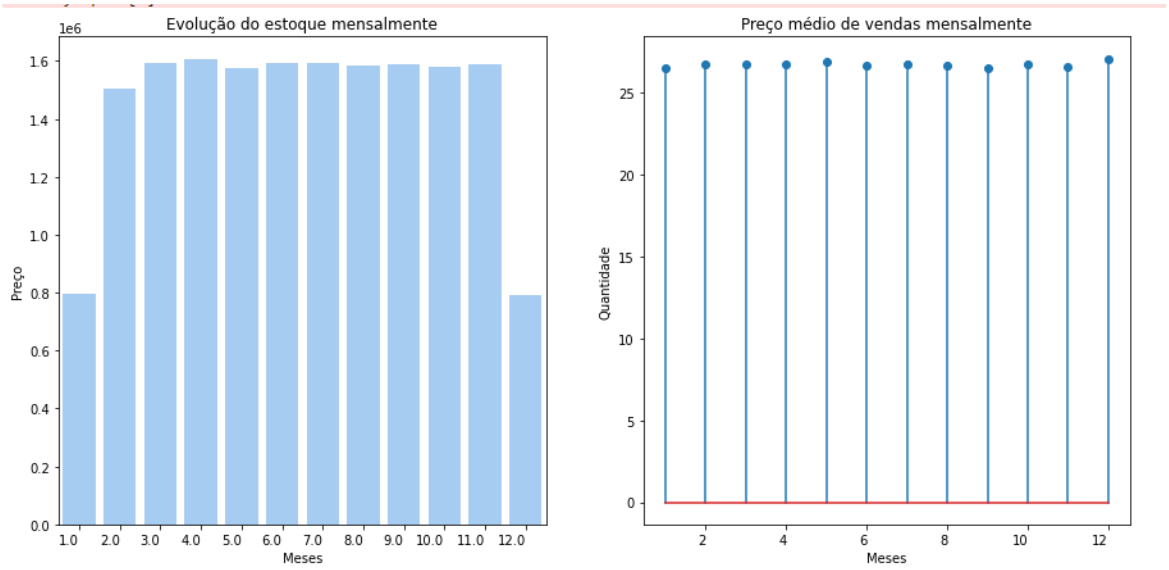
ANÁLISE TEMPORAL

E a evolução dos preços por mês, o que não houve diferença, ou algo que se pode apurar.

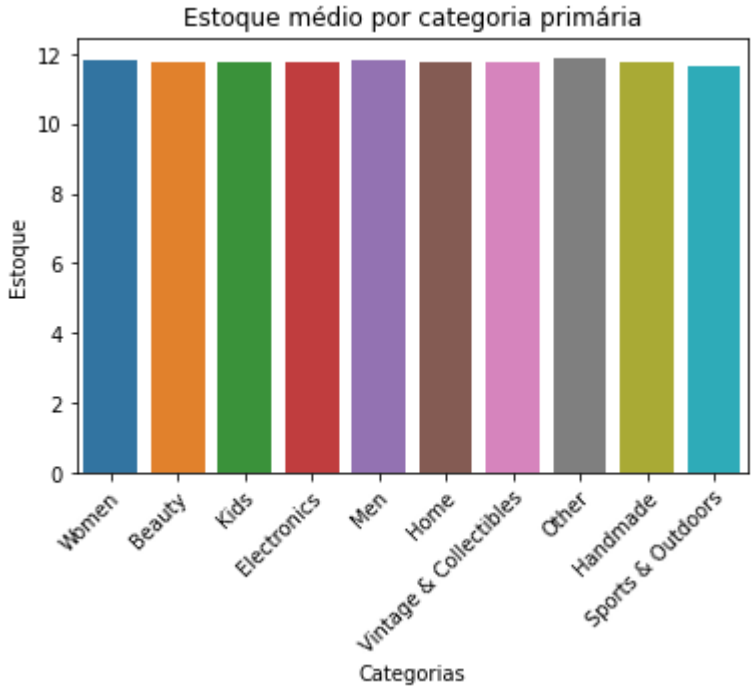


A análise da variação do preço no tempo nos mostra um comportamento atípico dos dados. Não há grande variação na média geral dos preços no decorrer do ano e além disso há grandes quedas no primeiro mês e no último mês.

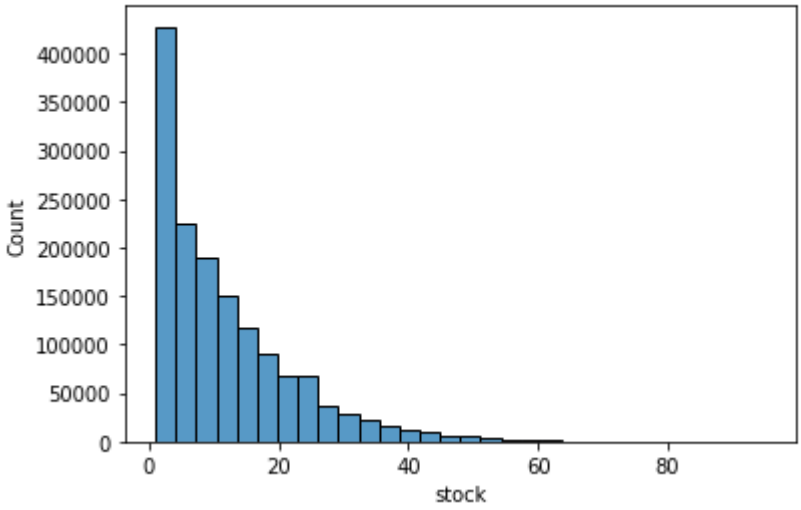
ANÁLISE DO ESTOQUE



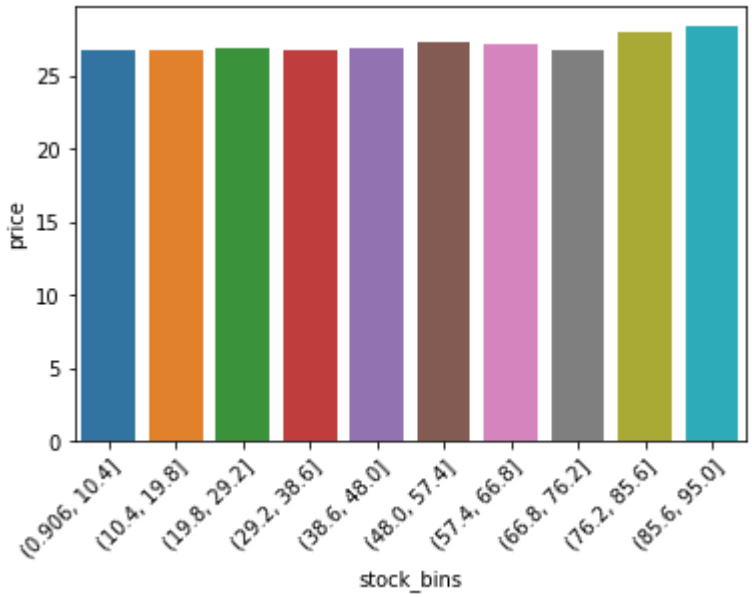
O estoque não apresenta grandes variações durante o ano, exceto por quedas bruscas no começo e no final do ano, que seria justificável por um grande volume de vendas (que não ocorreu, segundo a visualização de quantidade de vendas mensalmente). Apesar dessas duas grandes variações, a média de preços permanece inalterada.



A média de estoque de todas as categorias também tende a um mesmo valor



Segundo o histograma, temos muitos valores de 'stock' abaixo de 20 e poucos acima deste valor. Nesse sentido, podemos pensar que um produto com estoque elevado está encontrando dificuldades para ser vendido, enquanto um produto com baixo valor de estoque está em seu preço ideal ou muito barato.

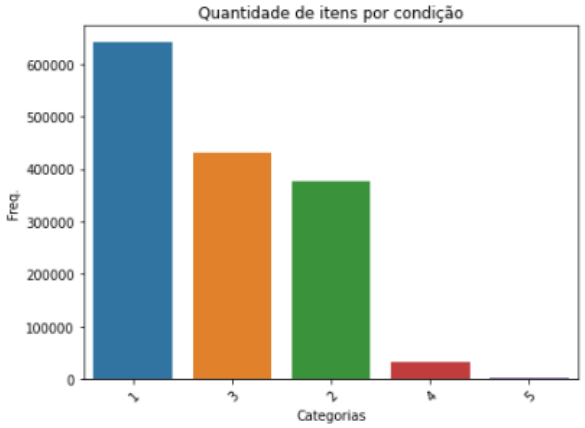


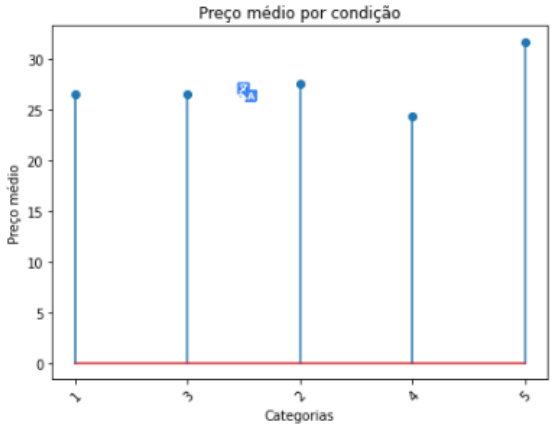
Ao dividir os valores de estoque em diversas faixas e verificar o preço médio, também não notamos variações que justifiquem produtos com alto estoque serem mais raros que produtos com baixo estoque.

Devido ao perfil apresentado, a variável 'stock' não mostra ter um grande impacto na variação do preço. Além do estoque ser praticamente constante durante todo o ano, no momento em que essa variável cai de forma brusca, o preço médio dos produtos permanece inalterado. Sendo assim, podemos dizer que é uma variável que não adicionada nada a um modelo preditivo e sua permanência para os modelos será discutida.

ANÁLISE DA CONDIÇÃO

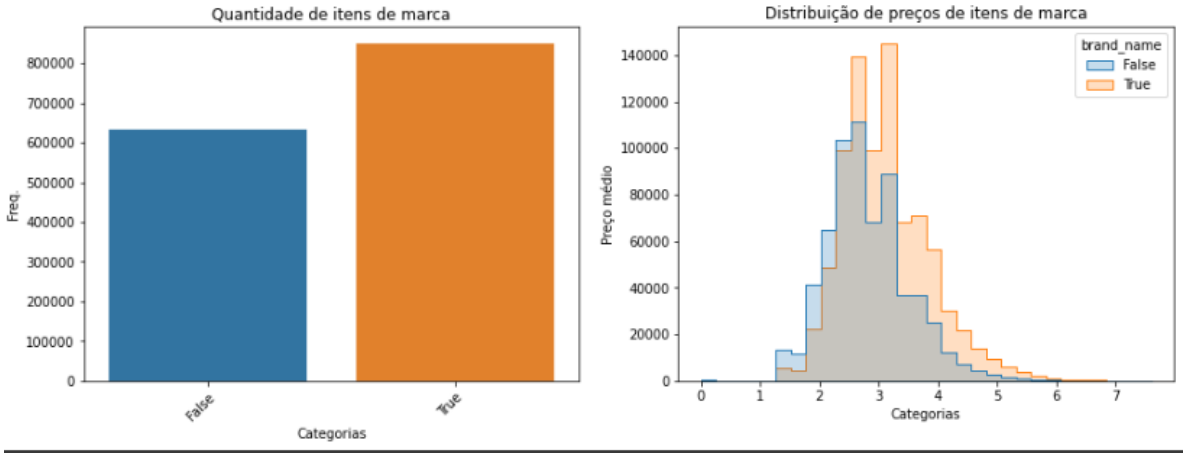
A distribuição de itens por condição nos mostra uma grande valorização de itens com qualidade 5 apesar da grande maioria pertencer a qualidade 1.





ANÁLISE DAS MARCAS

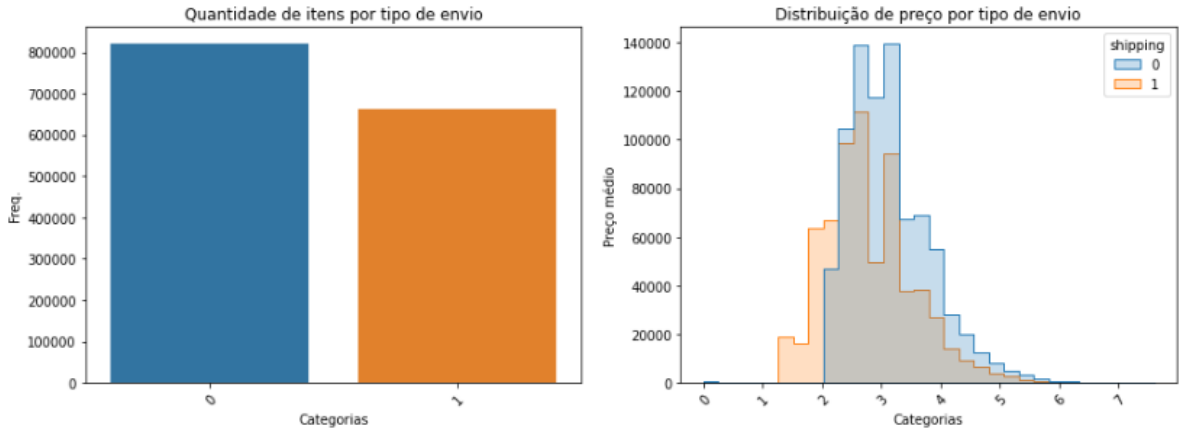
Aqui foram avaliadas as marcas, se itens que possuem marca tem valores maiores.



Apesar de haver marcas diferentes, a análise acima mostra que há uma variação considerável entre itens que possuem marcas. Isso nos abre uma possibilidade de no pré-processamento substituir as marcas por valores que indiquem ausência ou presença.

ANÁLISE PELO TIPO DE ENVIO

Quantidade por tipo de envio:



Além disso, o modo de cobrança do frete também influencia no preço final do produto. E, ao contrário do que se espera, o preço tende a subir quando o frete é pago pelo vendedor. Isso provavelmente acontece porque geralmente o vendedor paga o frete para vendas mais caras como forma de impulsionamento.

OUTRAS ANÁLISES

A Correlação dos demais atributos com o preço.

A partir das correlações acima, observa-se que o frete é inversamente proporcional ao preço.

Filtrando a brand Apple, e analisando sobre a sub-categoria2 ‘Cell Phones & Smartphones’, verificamos vários aparelhos, de modelos distintos, preços variados, pois alguns modelos são obsoletos.

Valores nulos:

* 43% dos registros não tem nomes de marca

* 0,4% dos registros não tem categorias

* 0,0003% não possuí descrição do item

Iremos excluir os valores nulos, exceto a marca iremos avaliar o impacto dos valores nulo no modelo.

Link: EDA

HIPÓTESES

- Qual o melhor modelo para o negócio?

- ~~O Estoque pode ser um fator de alteração do preço de venda, devido a oferta e demanda?~~

- O modelo poderá identificar que o produto é obsoleto/encalhado e irá sugerir um valor mais baixo para ter maior atratividade? (Exemplo Iphone 5 para iphone 10)

- ~~Etapas para comprovação da hipótese 1:~~

    - ~~Qual é a relevância da marca para o modelo~~?

- ~~Usar PCA, irá melhorar o processamento?~~
    - ~~da pra aplicar PCA em matriz esparsa~~

Em desenvolvimento.

HIPOTESE 1

Vamos avaliar o impacto da *"brand"* para o nosso modelo.

A razão de tal análise serve para entendermos melhor qual tratamento deveremos dar para nossos dados 632.682 nulos, um número significativo dentro de nosso *dataset*. Para isso vamos separa-los em dois *datasets*, um contendo apenas os valores nulos*/missing* e outro *dataset* contendo todas as colunas preenchidas.

Após essa etapa vamos gerar modelos probatórios para um teste a/b, onde poderemos analisar o melhor caminho para o tratamento desses valores nulos.

Conforme levantado na hipótese, verificaríamos o quão explicativo a *brand* é no atual projeto, para saber como tratar tantos valores nulos, aproximadamente 42%. O objetivo não foi alcançar a melhor métrica, apenas comparar em condições similares.

Na preparação do treino/teste foi utilizado o mesmo sample de dados, pré-processamento, tratamento e parâmetros de modelo, no caso o lightgbm.

obtendo os resultados:

----------Modelo com brand----------

MAE: 13.470141

RMSE: 1266.732524

----------Modelo sem brand----------

MAE: 10.244517

RMSE: 624.229604

Podemos observar que com as *brands* o modelo perde em ambas as métricas, isso pode ser devido ao grande desbalanceamento entre as marcas.

Contudo, essas informações de *brand* não serão descartadas.

Link com o código: Hipotese 1

HIPOTESE 2

- Usar PCA, irá melhorar o processamento?

  A aplicação do PCA é viável em matrizes esparsas?

Após os tratamentos iniciais no *dataset*, aplicamos um modelo simples de regressão para obter as primeiras predições e um esboço inicial de um modelo preditivo.

Escolhemos uma regressão linear por sua simplicidade e facilidade na aplicação. Também utilizamos uma amostra menor do *dataset* para verificar o funcionamento do modelo. As métricas deste modelo inicial são:

Para esta primeira analise as colunas categóricas foram tratadas de duas formas. A ideia da extração de características em textos é a de estruturar dados não estruturados para que possam ser utilizados em algum modelo de *machine learning*. A técnica TF-IDF (*Term Frequency - Inverse Document Frequency*) mostra o quão importante é uma palavra é em um texto seguindo a seguinte fórmula:

$$TFIDF = (a/b) * \log(\alpha/\beta)$$

onde:
a = Nº de vezes que uma palavra aparece no texto
b = Nº de palavras no documento
α = Total de documentos
β = Nº de documentos com o respectivo termo

Uma das características deste tipo de tratamento é que cada termo se torna uma coluna e, de forma geral, sua saída é uma matriz esparsa. Uma matriz esparsa caracteriza-se pela presença de zeros na maioria dos elementos da matriz, sendo assim para a maioria das observações há muitas colunas com valores iguais a zero.

O tratamento das colunas categóricas foi feito utilizando o *One-Hot Encoding* também presente no *Scitkit-Learn*, neste tipo de tratamento tornamos dados categóricos em dados numéricos transformando cada uma das colunas são transformadas em diversas colunas binárias que representam a presença ou ausência das características presentes nesta coluna. Em colunas com diversas características possíveis, sua saída também pode ser uma matriz esparsa.

Após isso, ajustamos o modelo de regressão linear e então aplicamos a mesma transformação no conjunto de validação separado anteriormente.

As métricas mostram-se aceitáveis e provam que este esqueleto inicial do modelo é funcional, podendo ser escalado para as mais de 1 milhão de linhas presentes no *dataset* de treino completo.

Com este grande número de colunas, técnicas de redução de dimensionalidade começam a ser viáveis, como temos diversas matrizes esparsas, a única técnica de redução de dimensionalidade fornecida pelo *Scikit-Learn* é a *Truncated SVD.* Assim como em técnicas como **o PCA**, as novas componentes são escolhidas baseadas na variação que melhor explica os dados. Por recomendação da própria documentação do *Scikit-Learn,* será utilizado o valor de 100 componentes na redução de dimensionalidade.

Com a confirmação do funcionamento, tentamos utilizar todo o *dataset* para podermos comparar os resultados com modelos futuros, porém, devido ao grande volume de dados, a máquina virtual oferecida na versão gratuita do *google colab* não possui capacidade o suficiente de memória RAM para o processamento.

Ao analisar nosso problema, notamos que as funções de vetorização de texto são as que mais geram colunas na matriz esparsa. Assim, decidimos por adotar dois argumentos para limitar a entrada de um novo termo na matriz do *Tfidf*.

O argumento *min_df* ignora os termos que possuem frequência menor que um certo limite, neste modelo inicial um termo deve aparecer pelo menos 50 vezes para aparecer na matriz esparsa de vetorização.

O argumento *max_df* ignora os termos que são muito frequentes e que não adicionariam nenhuma informação útil ao modelo. O valor 0.5 representa uma proporção entre documentos e o total da contagem de um termo. Com este tratamento, os vetores da coluna de descrição passam a ter 39930 linhas.

Nesta configuração o modelo é capaz de convergir, tendo métricas aceitáveis considerando a distribuição dos valores que estamos tentando prever. Porém, ao comparar a aplicação do PCA com modelos desenvolvidos sem a sua utilização, notamos uma grande deterioração dos resultados, o que significa que não poderemos utilizá-lo nas etapas subsequentes deste trabalho.

Link do código: Hipotese 2

HIPOTESE 3

ESCOLHENDO O MODELO

Colocamos todos os modelos, que até então estavam avulsos em dois notebooks, em apenas um, avaliamos as métrica e juntamente com o tempo que leva para cada modelo executar.

Pelo modelo MLP, obtivemos os seguintes resultados:

```
MLP
MAE: 10.546949903299131
RMSE: 28.168716160371442
RMSLE: 0.2264574632599648
```

Pelo modelo XGBoost, obtivemos os resultados abaixo:

```
XGBoost
MAE: 11.842656
RMSE: 31.02
RMSLE: 0.279296
```

Pelo modelo LigthGBM, obtivemos os seguintes resultados:

```
LigthGBM:
MAE: $ 10.85
RMSE:  $ 21.20
RMSLE: 0.274586
```

Pelas métricas, temos uma leve vantagem da MLP sobre os outros modelos. Devido ao escopo apresentado no início do projeto, devemos prosseguir com modelos que apresentem métricas e que além disso se beneficiem do aumento da CPU/GPU. Sendo assim, optamos por prosseguir com a MLP.

- Link com o código: Hipotese 3

HIPOTESE 4

O Estoque pode ser um fator de alteração do preço de venda, devido à oferta e demanda?

O estoque no *dataset* apresenta somente os valores que o vendedor possui em inventário, não há informação sobre a quantidade de produto vendido, sendo difícil mensurar seu valor ou conseguir fazer uma gestão de estoque ou sazonalidade.

Iremos averiguar após a limpeza das categorias e nomes, se o estoque que o *dataset* possui, será capaz de ter interferência no preço como se ele fosse um fator de oferta e demanda.

Juntamente com as datas de venda, será possível verificar alguma sazonalidade nos produtos ofertados.

Conforme apresentado no EDA, o estoque não apresenta grandes variações durante o ano, exceto por quedas bruscas no começo e no final do ano, que seria justificável por um grande volume de vendas (que não ocorreu, segundo a visualização de quantidade de vendas mensalmente). Apesar dessas duas grandes variações, a média de preços permanece inalterada.

A média de estoque de todas as categorias também tende a um mesmo valor.

Temos muitos valores de 'stock' abaixo de 20 e poucos acima deste valor. Nesse sentido, podemos pensar que um produto com estoque elevado está encontrando dificuldades para ser vendido, enquanto um produto com baixo valor de estoque está em seu preço ideal ou muito barato.

Ao dividir os valores de estoque em diversas faixas e verificar o preço médio, também não notamos variações que justifiquem produtos com alto estoque serem mais raros que produtos com baixo estoque.

Devido ao perfil apresentado, a variável 'stock' não mostra ter um grande impacto na variação do preço. Além do estoque ser praticamente constante durante todo o ano, no momento em que essa variável cai de forma brusca, o preço médio dos produtos permanece inalterado. Sendo assim, podemos dizer que é uma variável que não adicionada nada a um modelo preditivo e sua permanência para os modelos será discutida.

Para o modelo julgamos ser uma variável irrelevante e será excluída.

RELATÓRIO MODELAGEM EM DESENVOLVIMENTO

Primeira tentativa de *baseline* tivemos dificuldade de memória, e serão realizados pré-processamentos para tentar contornar esse problema.

Erro corrigido, pois estávamos transformando matriz esparsa em densa, após essa correção, a questão de memória foi sanada.

Separamos o *dataset* em treino e teste para começarmos as transformações, protegendo nossos dados de validação de possíveis vazamentos.

Normalizado, com log, o *target (price),* foi realizado de forma separada para não terem dados vazados.

Realizamos o pré-processamento das colunas com a divisão do *dataset* entre treino e teste, foi realizado a vetorização dos textos do *dataset,* e aplicado *dummy*.

No pré-processamento do texto, o *dataset* foi utilizado o *tokenizador* para separar o texto em listas e o a remoção dos *stopwords*.

Não estamos preocupados com a semântica das frases, e sim com as palavras, ate por que ninguém irá falar algo negativo sobre seus produtos à venda, e neste caso, juntamos três colunas o nome, marca e descrição do item, pois facilitará o entendimento da máquina, evitando a duplicação de informação e reduzindo o volume do conjunto de dados.

Montamos três modelos de *baseline* para escolher qual modelo será usado, sendo eles: *XGBoost, LigthGBM* e o CNN.

XGBoost, por ser um algoritmo de aprendizado de máquina, atualmente, mais populares, é considerado um dos modelos mais poderosos para dados tabulares para classificação, bem como regressão.

Escolhemos também para teste o *LigthGBM*, um dos modelos de *Boosting* mais utilizados atualmente.

Ele é um modelo baseado em árvore de decisão, sendo muito mais rápido e utiliza menos memória, possui suporte ao paralelismo, o *LightGBM* faz o crescimento por folha e é amplamente usado em dados grandes.

Um dos maiores perigos dos algoritmos de *boosting* é que eles acabem aprendendo com o ruído nos dados, o que pode levar o modelo a um *overfitting*.

E o MLP *(Multi-Layer Perceptron*), é uma rede neural com uma ou mais camadas ocultas com um número indeterminado de neurônios. Funcionam em quase tudo, são modelos fortes e rede neural resolve qualquer problema.

E analisando os resultados do EDA, decidimos excluir os *outliers* dos valores acima de 250 dólares, pois focaremos na maioria dos usuários do site, do que invés de englobar os que usam esporadicamente com valores altos, pois valores altos podem ter

outro viés, desde marca, coleção, raridade, entre outros, o que possam causar ruído no modelo.

Excluiremos os valores 0 (zero) pois são doações e isso possa dificuldade ou criar ruído. Conforme analisamos o estoque não apresenta grandes diferenças, e também será excluído.

Testando dropar datas, pois acreditamos ser ruído, já que não presentou sazonalidade. Em teste.

Utilizaremos as métricas (*MAE*) Erro Médio Absoluto, (*MSE*) Erro quadrático Médio e (*RMSE*) raiz do Erro Quadrático Médio.

MAE: O erro médio absoluto (*MAE —Mean Absoluto Error*), mede a média da diferença entre o valor real com o predito. Esta métrica não é afetada por valores discrepantes — os denominados *outliers*.

MSE: O erro quadrático médio (*MSE - Mean Squared Error*) é uma métrica que calcula a média de diferença entre o valor predito com o real. Penalizando valores que sejam muito diferentes entre o previsto e o real. Portanto, quanto maior é o valor de MSE, significa que o modelo não performou bem em relação as previsões.

RMSE: A raiz do erro quadrático médio (*RMSE —Root Mean Squared Error*) é basicamente o mesmo cálculo de MSE, contendo ainda a mesma ideia de penalização entre diferenças grandes do valor previsto e o real. Esta métrica pode ser uma boa opção quando é preciso ter uma avaliação mais criteriosa sobre as previsões do modelo.

Não queremos ter a melhor métrica, como numa competição, nosso intuito é criar o modelo que traga o valor mais adequado para o negócio.

LINKS

- Kaban: https://trello.com/b/iUi8g2cW/grupo-3-nan
- GitHub: https://github.com/anachavesv8/Bootcamp_Blue

REFERÊNCIAS

Como comprar e vender suas coisas usando o app Mercari Japão. <https://jn8.jp/pb/life_list/3073/>. Acessado 14/09/2022.

フリマアプリはメルカリ(メルペイ)-フリマアプリ&スマホ決済. IMAGEM<https://play.google.com/store/apps/details?id=com.kouzoh.mercari&hl=pt&gl=US>. Acessado 15/09/2022.

Como comprar da Mercari Japan. <https://www.whiterabbitexpress.com/pt/lojas/mercari-japan/>. Acessado 15/09/2022.

JUNIOR, Clébio. **Prevendo Números: Entendendo as métricas R², MAE, MAPE, MSE e RMSE**. Disponível em: <https://medium.com/data-hackers/prevendo-números-entendendo-métricas-de-regressão-35545e011e70>. Acessado 05/10/2022.