Análise e Predição de Métodos de Pagamento nas Plataformas de e-commerce da Olist

Dataset Publicado em:

https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?select=olist_geolocation_dataset.csv

Aprendizado de Máquina 2023.2 - Prof. João Carlos

Apresentado por: **Guilherme Soares e João Victor Gadelha**

Introdução

E-commerce

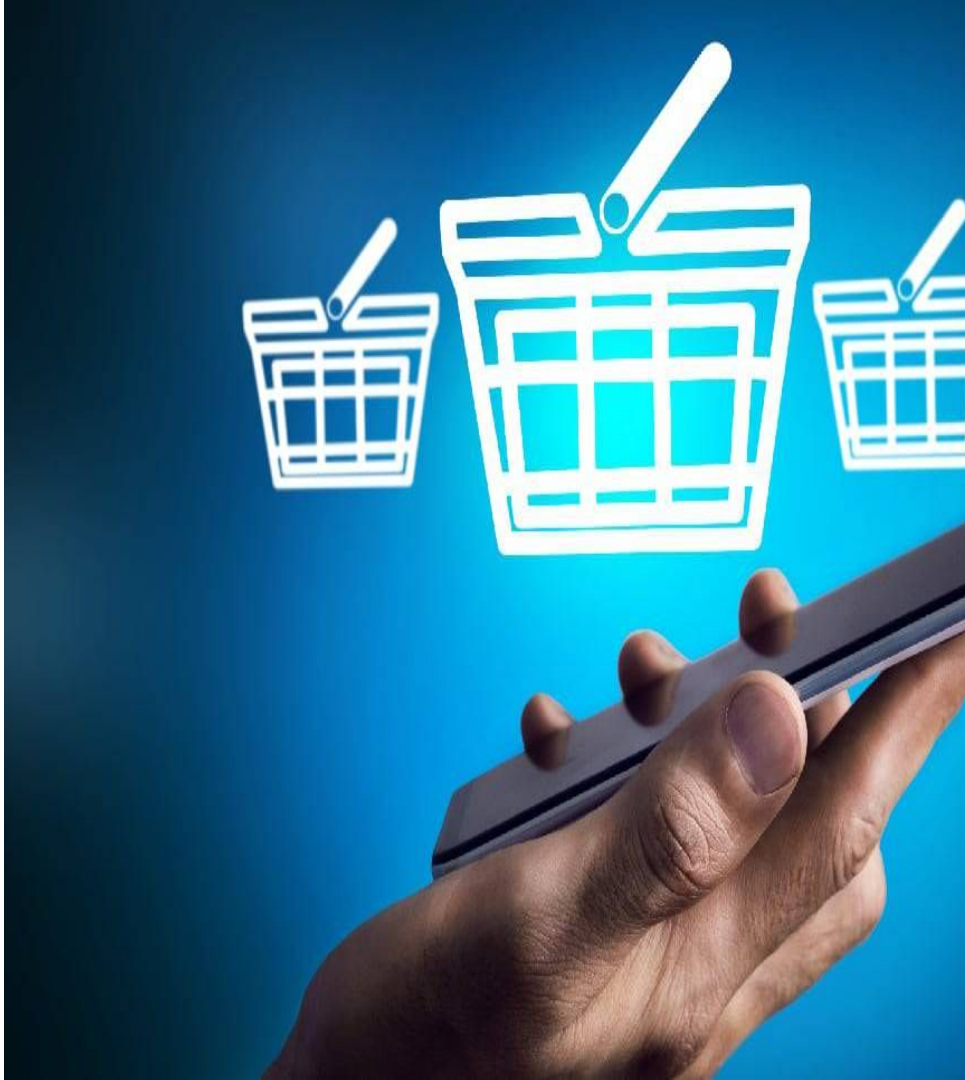
- Acessibilidade da internet impulsionou a realização de compras online
- Presente em peso em todos os cantos da internet através de promoções, anúncios e preços competitivos

E-commerce

- Neste cenário achamos interessante investigar como os brasileiros realizam suas compras
- Em especial qual forma de pagamento é mais comumente usada dependendo do valor da compra e da região

E-commerce

Quem aqui nunca fez uma compra online?



Descrição da base

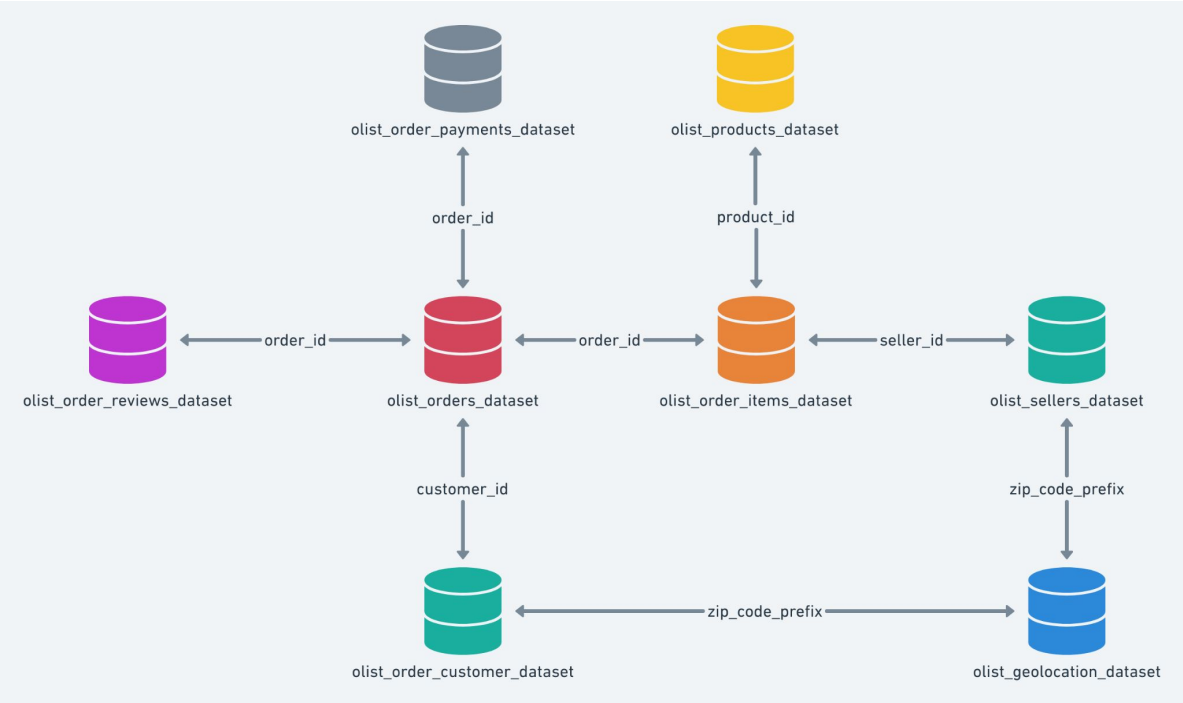
Dados utilizados

- A base pública de dados disponibilizada pela Olist, uma das maiores empresas de e-commerce do Brasil
- Esperamos encontrar alguma relação para prever o tipo de pagamento utilizado

Dados utilizados

- A base utilizada pode ser encontrada no kaggle
- 8 CSVs no total, que podem ser mesclados facilmente através de IDs
    - Pagamentos, Pedidos, Produtos, Vendedores, Localização, Clientes, Avaliações e Itens num Pedido.

Dados utilizados



Dados utilizados

- Dataset criado com a finalidade de utilizar NLPs
- Logo, estamos utilizando o dataset de uma forma que ele não foi projetado

Dados utilizados

- Foram coletados no período entre 2016 a 2018
- Dados Sensíveis foram anonimizados
- Na nossa análise, utilizamos arquivos referentes a:
    - Pagamentos
    - Pedidos
    - Consumidores

Análise Inicial

Análise estatística básica

Através do uso de Python, com suas bibliotecas de dados mais populares, como Pandas e Scikit-learn, conduzimos análise estatística básica inicial nos dados e descobrimos algumas métricas:

Análise Inicial

- A compra média do brasileiro custa em torno de 154 reais
- O número médio de parcelas no cartão é próximo a 3
- Compras acima de mil reais representam por volta de 1% apenas dos dados
- Aproximadamente 80% das compras foram realizadas no crédito
- Máximo valor registrado foi de 14 mil reais

Tratamento de Dados

- Como pretendemos prever o método de pagamento utilizado, precisamos fazer algumas alterações
- Com isso, removemos algumas informações da tabela inicial
- Fizemos o encoding da nossa feature categórica (Estados) utilizando o método *get_dummies*

Métodos Utilizados

Aprendizado de Máquina

- Como o dataset é extremamente desbalanceado optamos por utilizar métodos de classificação com complexidade ascendente
- Separamos os dados em 20% para testes e 80% para treinamento

Métodos Utilizados

- Regressão Logística
- Support Vector Machine
- Random Forest

Fine Tuning

- Feature Engineering
- Grid Search
- Cross Validation

Grid Search

- Permite que o usuário faça escolhas referentes a um parâmetro que pode ser alterado no modelo selecionado
- Iterativamente cria um modelo diferente com cada combinação possível de parâmetros
- Focamos em *n_estimators* e *Bootstrap*
- Também utilizamos validação cruzada

Trabalhos Relacionados

Trabalhos Relacionados

- Como já comentado anteriormente, os principais trabalhos utilizando essa base são em análise de dados e NLP
- Não encontramos outro trabalho que utilizasse o dataset para classificação

Trabalhos Relacionados

- No máximo, encontramos um que, através de clusterização, procurava relações entre as features
- Porém, este notebook utiliza outras tabelas deste dataset

Resultados

Regressão Logística e SVM

- Resultado muito simplista
- Prevê todas as entradas como “Crédito”

payment_type	count
credit_card	15339
boleto	3948
voucher	1168
debit_card	322

0	count
credit_card	20777

SVM

- Resultados foram, classificando a maioria dos dados como Voucher
- Deduzimos que o desbalanceamento de tipos utilizados pode ter atrapalhado a linha de separação

payment_type	count	0	count
credit_card	15339	voucher	16229
boleto	3948	debit_card	2243
voucher	1168	boleto	2028
debit_card	322	credit_card	277

Random Forest

- Acerta os casos de “crédito” em 80%
- Por outro lado, erra as outras classes em mais de 90%
- Apesar disso, foi o modelo mais “balanceado” dentre os utilizados
- A distribuição de escolhas é muita mais próxima do conjunto de teste.

Random Forest

- Matriz de Confusão

Random Forest com nova feature

- Na tentativa de melhorarmos o desempenho do modelo, adicionamos uma nova feature
- Utilizando os dados geográficos disponíveis na tabela, criamos uma coluna booleana para verificar se o usuário fez a compra de uma capital

Random Forest com nova feature

- Matriz de Confusão

Random Forest com Grid Search

- Por ser um método de força bruta, o treinamento foi consideravelmente mais longo
- Com 6 possíveis valores para o número de estimadores e 2 possíveis para o Bootstrap, foram testados 12 modelos diferentes.
- Precisão balanceada foi utilizada como função para avaliar o modelo
- Como vencedor, tivemos o modelo onde o número de estimadores foi igual a 10 e com Bootstrap ativado

Random Forest com Grid Search

- Matriz de Confusão

Conclusão

Conclusão

- Lidar com um dataset desbalanceado se provou um desafio
- Pouquíssimas compras feitas em débito
- Talvez informações adicionais pudessem refinar o modelo

Obrigado!

Análise e Predição de Métodos de Pagamento nas Plataformas de e-commerce da Olist

Dataset Publicado em:

https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?select=olist_geolocation_dataset.csv

Aprendizado de Máquina 2023.2 - Prof. João Carlos

Apresentado por: **Guilherme Soares e João Victor Gadelha**