Trabalho Final – Turma 12

Caso de Uso: Olist

05/Junho/2020

Coordenadores:
Profª Drª Alessandra de Álvila Montini
Profª Dr. Adolpho Walter Pimazoni Canton

GRUPO 12:

- **Ana Lúcia da Cunha Cox**
- **Thiago Yoshiaki Miyabara Nascimento**

Agenda

1. Objetivo do Trabalho
2. Contextualização do Problema
3. Base de Dados
    i. Bases originais
    ii. Processo de redução de variáveis
    iii. Principais variáveis
4. Análise Exploratória de Dados
5. Modelagem com Estatística Tradicional
6. Modelagem com Inteligência Artificial
7. Desafios encontrados
8. Conclusões

1. Objetivo do Trabalho

O objetivo do trabalho é **predizer o valor do frete** das compras realizadas na plataforma do e-commerce Olist.

A predição será realizada por meio da análise do banco de dados histórico e uso de **modelos estatísticos** e **algoritmos de Machine Learning**, que selecionarão as **características mais relevantes** que explicam o valor da entrega.

Desta forma, a empresa poderá traçar **estratégias de logística**, desenvolver **programas de frete grátis** e **ações preventivas** para minimizar a desistência da compra devido ao valor do frete.

2. Contextualização do Problema

Em 2020 é esperado que **38%** de todas as vendas sejam feitas através de marketplaces como a Olist.

Importância do eCommerce

Estudo feito em todas as capitais pela Confederação Nacional de Dirigentes Lojistas (CNDL) e pelo Serviço de Proteção ao Crédito (SPC Brasil). Os dados mostram que **86% dos consumidores** conectados **realizaram ao menos uma aquisição em lojas online nos últimos 12 meses**.

Comportamento do Cliente

De acordo com uma pesquisa realizada pela Manhattan Associates, mais de **70% dos consumidores brasileiros preferem fazer compras online** ao invés de ir à uma loja, e cerca de **60% pedem para retirar seus itens no local**.

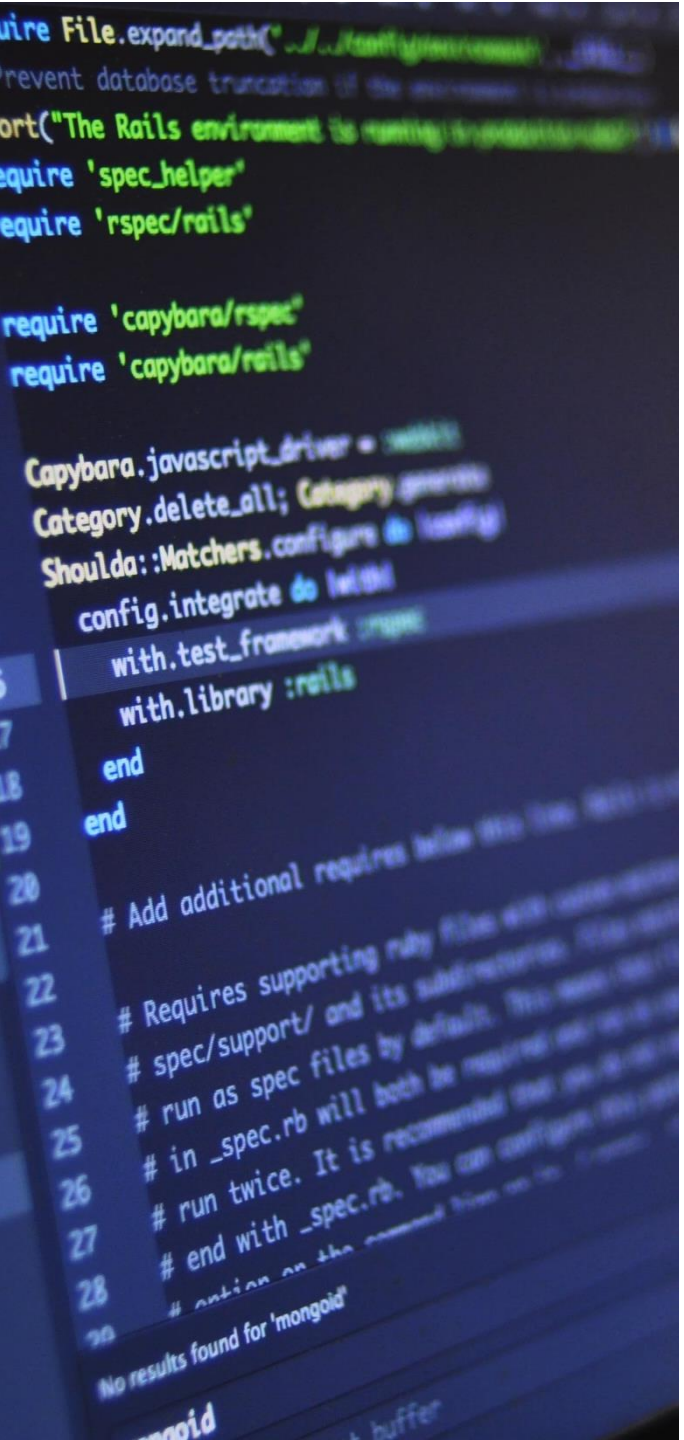
2. Contextualização do Problema

Segundo a pesquisa E-commerce Trends, da empresa de marketing digital Rock Content, o frete caro é responsável por **82,3%** do abandono do carrinho de compras.

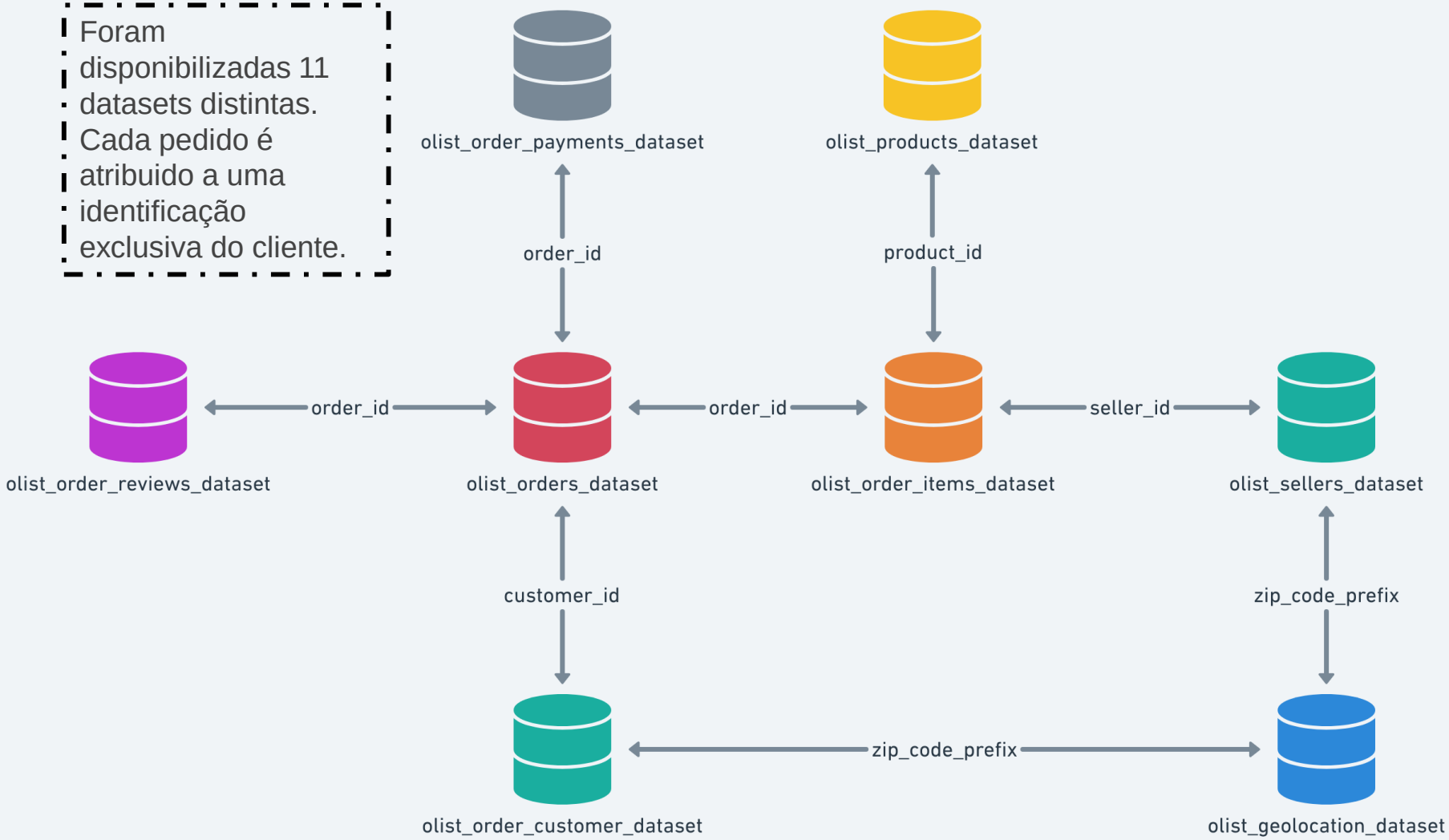
Principais fatores que influenciam no comportamento de compra:

- Relação entre altos custos do frete e abandono do carrinho de compras;
- Demora ou indisponibilidade no valor do frete;
- Diferença entre preços da loja física e e-commerce;
- Defasagem no preço entre marketplaces;
- Limitação de preços e prazos de entrega.

3. Bases de Dados



Foram disponibilizadas 11 datasets distintas. Cada pedido é atribuido a uma identificação exclusiva do cliente.



3.i. Base Original

Base de dados	Quantidade de Registros	Quantidade de Variáveis	Quantidade de Duplicadas	Quantidade de Nulos	Período Inicial	Período Final
olist_closed_deals_dataset.csv	842	14	0	804	2017-12-05 02:00:00	2018-11-14 18:04:19
olist_customers_dataset.csv	99.441	5	0	0	Não Aplicável	Não Aplicável
olist_geolocation_dataset.csv	1.000.163	5	261.831	0	Não Aplicável	Não Aplicável
olist_marketing_qualified_leads_dataset.csv	8.000	5	0	60	2017-06-14 00:00:00	2018-05-31 00:00:00
olist_order_items_dataset.csv	112.650	7	0	0	2016-09-19 00:15:34	2020-04-09 22:35:08
olist_order_payments_dataset.csv	103.886	5	0	0	Não Aplicável	Não Aplicável
olist_order_reviews_dataset.csv	105.189	7	94	96.151	Não Aplicável	Não Aplicável
olist_orders_dataset.csv	99.441	8	0	2.980	2016-09-04 21:15:19	2018-10-17 17:30:18
olist_products_dataset.csv	32.951	9	0	611	Não Aplicável	Não Aplicável
olist_sellers_dataset.csv	3.095	4	0	0	Não Aplicável	Não Aplicável
product_category_name_translation.csv	71	2	0	0	Não Aplicável	Não Aplicável

3.ii. Processo de redução de variáveis

Base Original
117.604 registros

Análise Exploratória

- Verificação e tratamento de valores faltantes;
- Criação do dicionário de dados;
- Entendimento da base.

Filtro e Tratamento Aplicado à Base Original

Seleção de colunas de acordo com a regra do negócio

- Filtro de valores zerados;
- Filtro dos registros duplicados;
- Vendas canceladas;
- Conexão com API do IBGE para obter os códigos UF.

Base Final com Outliers e sem Outliers
103.268 registros *versus* 86.741 registros

Modelos Preditivos com 6 modelos distintos

- Criação da variável “Volume”;
- Cálculo da distância entre vendedor e comprador com base na geolocalização;
- Exclusão de *outliers*.

3.iv. Principais variáveis

Variáveis do Produto

- Preço;
- Peso;
- Altura;
- Largura;
- Comprimento;
- Volume.

Variáveis Temporais

- *Timestamp* de criação do pedido;
- *Timestamp* de aprovação do pedido;
- *Timestamp* de postagem;
- *Timestamp* de previsão de entrega;
- *Timestamp* da entrega.

Variáveis de Localização

- UF do vendedor;
- UF do cliente;
- Calculo da distância entre vendedor e comprador com base na geolocalização.

Variável Resposta

- **Preço do frete**

4. Análise Exploratória de Dados

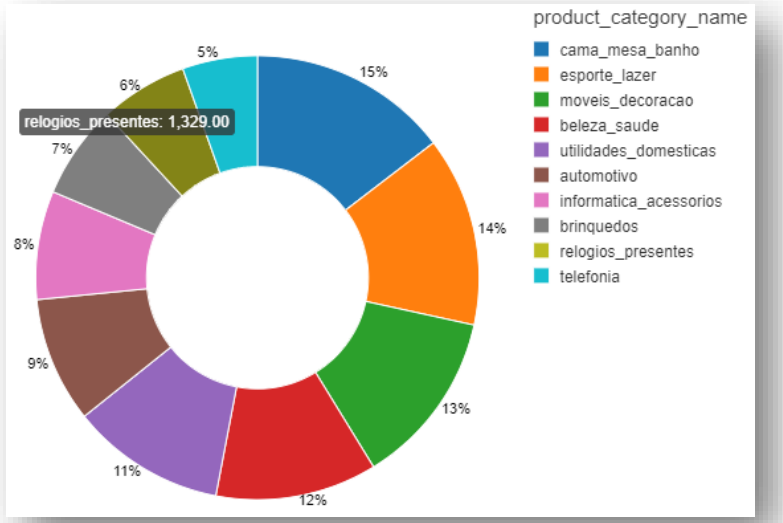
Na **Entrega 1** foi feita uma detalhada análise exploratória das 71 variáveis que as 11 base de dados apresentavam. Posteriormente, refizemos essa análise com **ênfase** no nosso ***business case***.

Principais atividades realizadas:

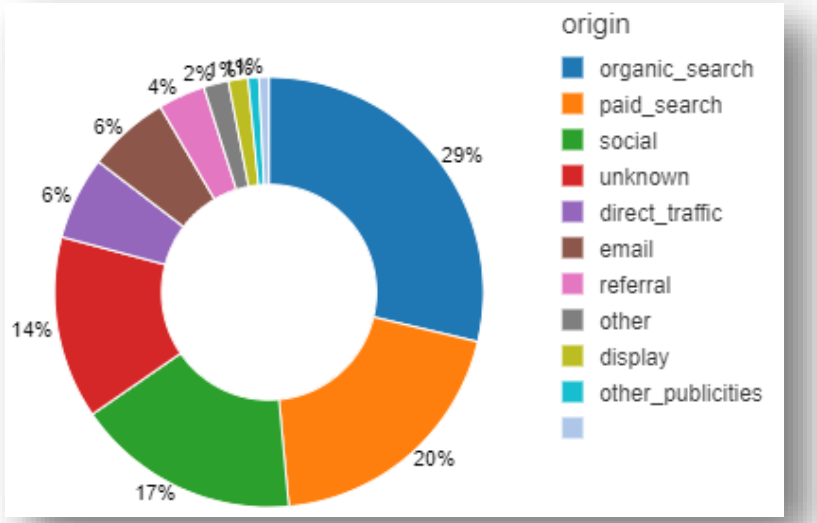
- Contagem e tratamento de valores faltantes e/ou em branco (*missings*);
- Identificação e tratamento de *outliers*;
- Criação de variáveis auxiliares;
- Tratamento de dados inconsistentes;
- Resumo das principais métricas estatísticas da base.

4. Análise Exploratória de Dados

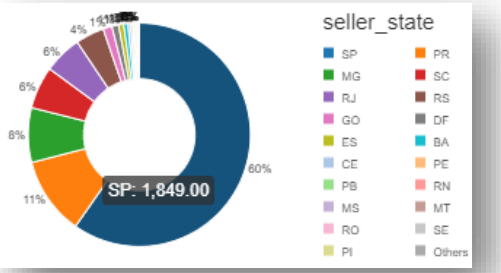
Top 10 categorias de produtos



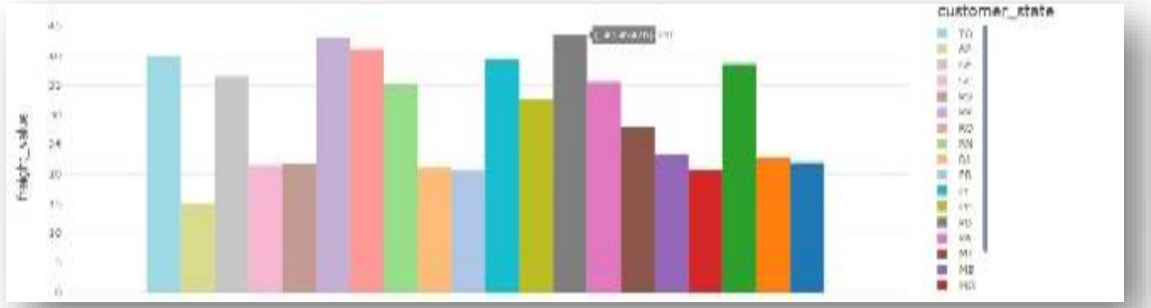
Origem de Mídia



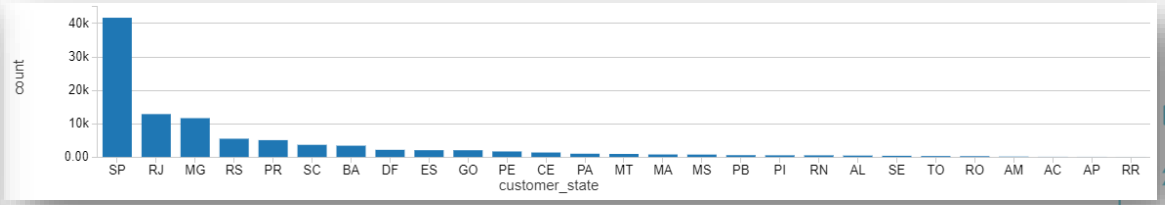
Estado dos Vendedores



Média do Valor do Frete por UF



UF dos Clientes



5. Modelagem com Estatística Tradicional

Depois de definir o ***business case,*** na **Entrega 2,** foi feita uma análise profunda da *feature* **”Valor do Frete”**.

Principais atividades realizadas:

- Resumo das principais **métricas estatísticas**;
- Geração do *heatmap* da **correlação** e **covariância**, relacionando a variável em estudo com as demais variáveis da base;

Analisando os resultados obtidos estatisticamente pudemos:

- Selecionar das variáveis com **maior influência** no valor do frete;
- **Normalizar** da base.

5. Modelagem com Estatística Tradicional

Resumo das principais Estatísticas com *Outliers*

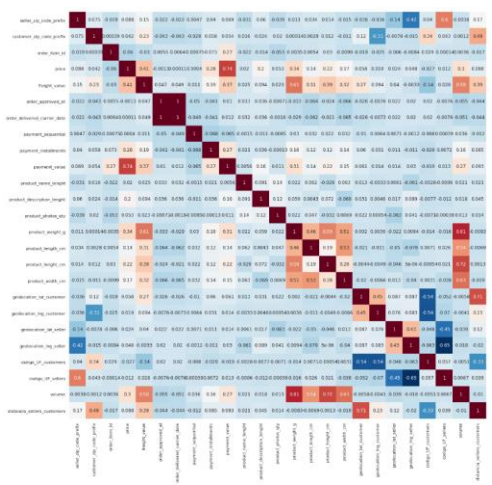
summary	freight_value
count	113930
min	1.0
25%	13.11
mean	20.081679891161265
stddev	15.735210411713263
50%	16.32
75%	21.19
85%	26.63
90%	34.13
95%	45.2
99%	85.59
max	409.68

Resumo das principais Estatísticas sem *Outliers*

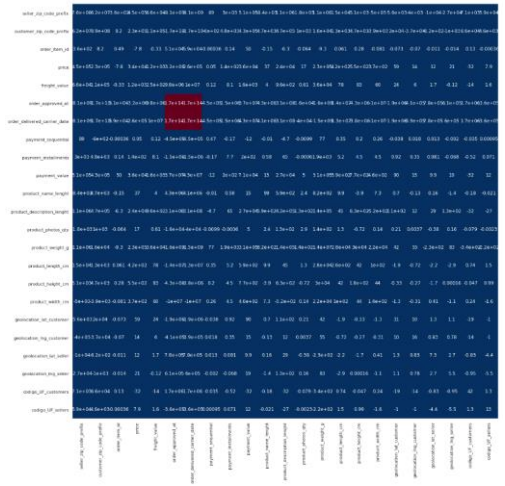
summary	freight_value
count	86741
min	5.0
25%	12.69
mean	15.573961794306335
stddev	4.5732571996071005
50%	15.38
75%	18.3
85%	20.14
90%	22.0
95%	23.63
99%	25.91
max	26.72

5. Modelagem com Estatística Tradicional

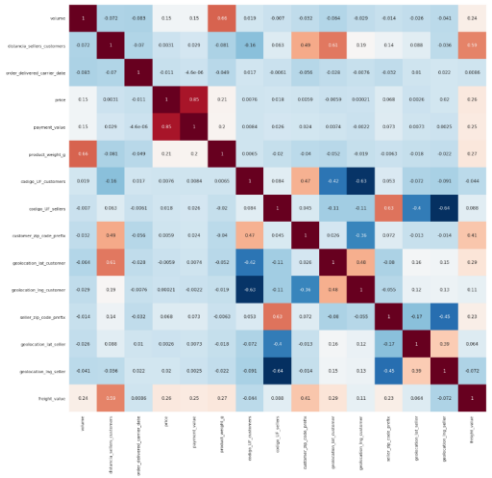
Heatmap da Correlação com *Outliers*



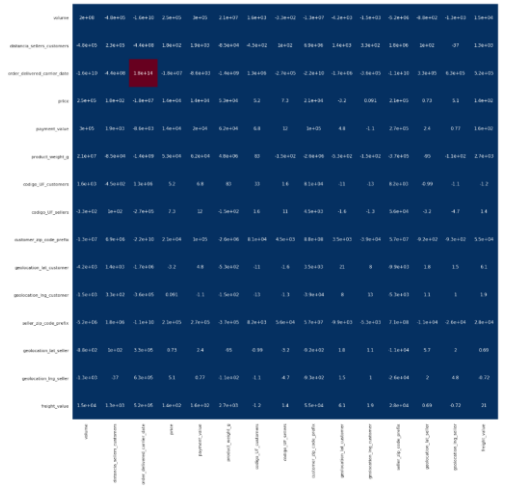
Heatmap da Correlação com *Outliers*



Heatmap da Covariância sem *Outliers*

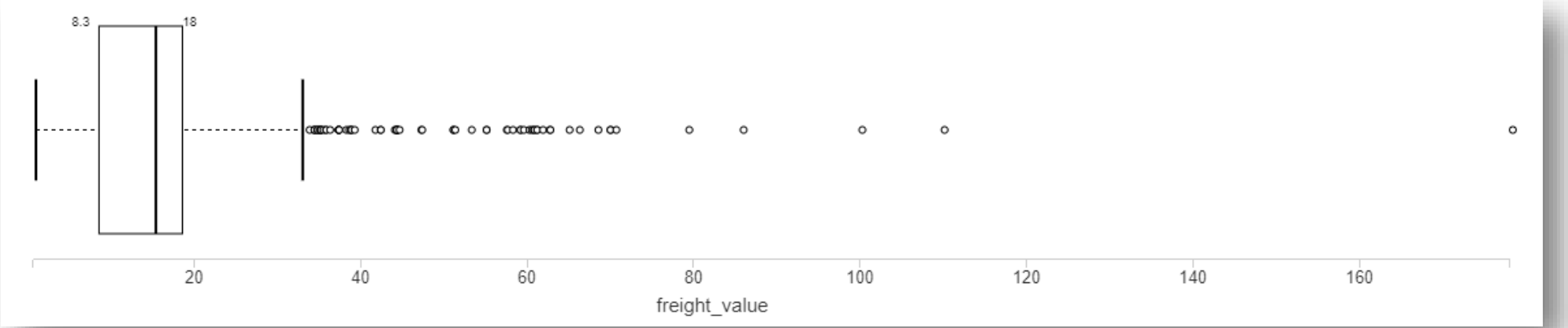


Heatmap da Covariância sem *Outliers*

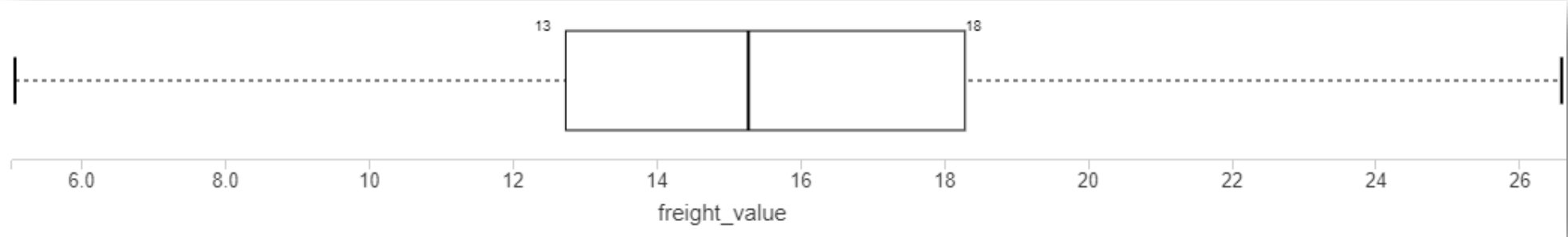


5. Modelagem com Estatística Tradicional

Box-Plot do Valor do Frete com *Outliers*

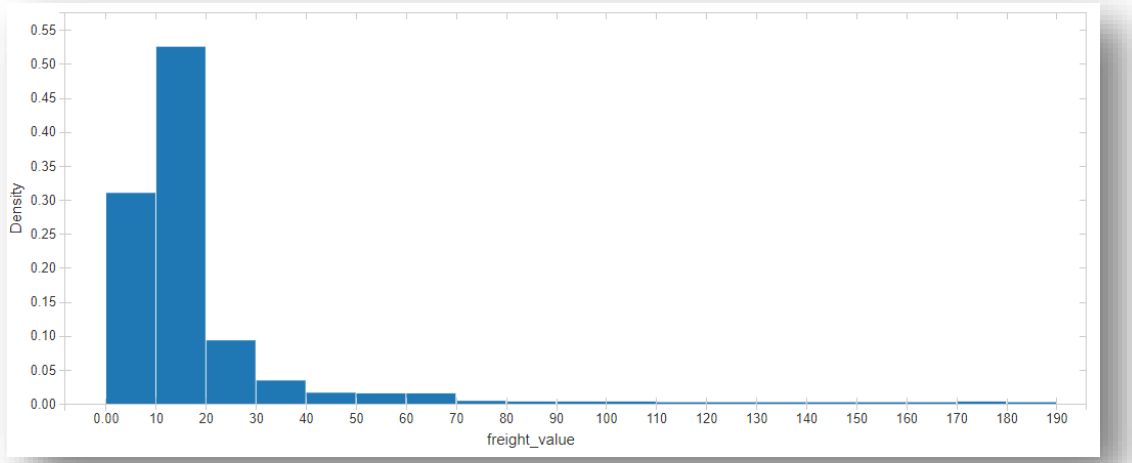


Box-Plot do Valor do Frete sem *Outliers*

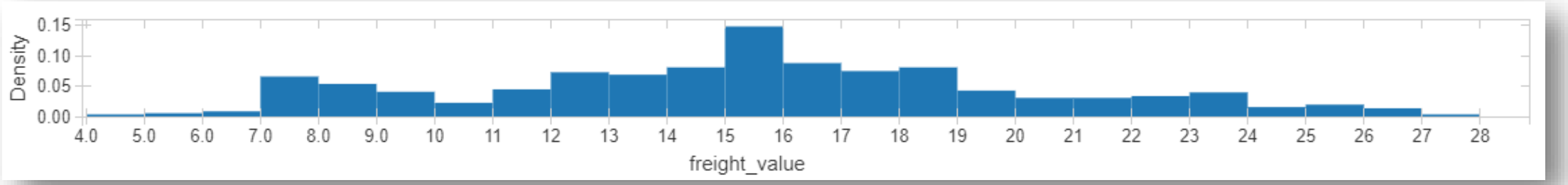


5. Modelagem com Estatística Tradicional

Histograma do Valor do Frete com *Outliers*



Histograma do Valor do Frete sem *Outliers*



6. Modelagem com Inteligência Artificial

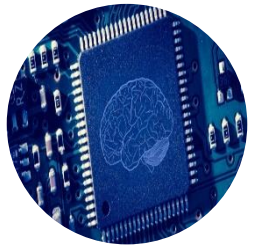
Regressão Linear

Árvore de Regrssão

GBT Regressor

Random Forest Regression

GLR - Gaussian



GBT Regressor – Categoria do Produto

6. Modelagem com Inteligência Artificial

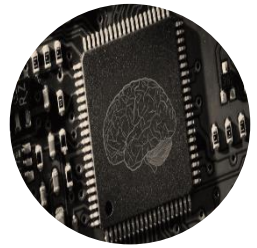
Regressão Linear

Árvore de Regrssão

GBT Regressor

Random Forest Regression

GLR - Gaussian



GBT Regressor – Categoria do Produto

Os resultados do modelo são:

MAE sem *outliers*: 2.301002436385273
MAE com *outliers*: 5.293480047119417
RMSE sem *outliers*: 3.032605165346057
RMSE com *outliers*: 9.862082979772966
R2 sem *outliers*: 55.94360716156693
R2 com *outliers*: 62.003370347318906

6. Modelagem com Inteligência Artificial

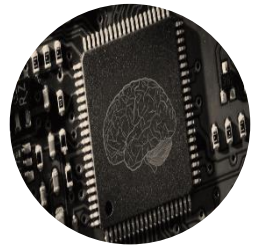
Regressão Linear

Árvore de Regrssão

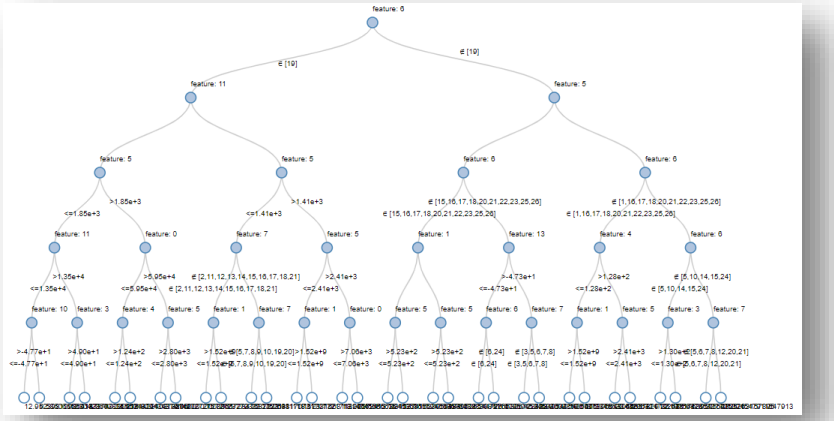
GBT Regressor

Random Forest Regression

GLR - Gaussian



GBT Regressor – Categoria do Produto



Os resultados do modelo são:

MAE sem *outliers*: 1.90685251941861
MAE com *outliers*: 5.050109013400841
RMSE sem *outliers*: 2.652784875178676
RMSE com *outliers*: 10.186470369642779
R2 sem *outliers*: 66.30642631065373
R2 com *outliers*: 59.48098951685474

6. Modelagem com Inteligência Artificial

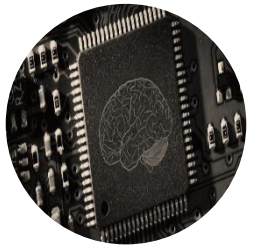
Regressão Linear

Árvore de Regrssão

GBT Regressor

Random Forest Regression

GLR - Gaussian



GBT Regressor – Categoria do Produto

Os resultados do modelo são:

MAE sem *outliers*: 1.6829742177602103
MAE com *outliers*: 4.473540942479851
RMSE sem *outliers*: 2.434614975890428
RMSE com *outliers*: 9.219827125176893
R2 sem *outliers*: 71.62057516443838
R2 com *outliers*: 66.80620295232022

6. Modelagem com Inteligência Artificial

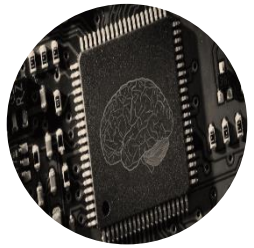
Regressão Linear

Árvore de Regrssão

GBT Regressor

Random Forest Regression

GLR - Gaussian



GBT Regressor – Categoria do Produto

Os resultados do modelo são:

MAE sem *outliers*: 1.8997635232778034
MAE com *outliers*: 4.865857056867012
RMSE sem *outliers*: 2.5946979492735562
RMSE com *outliers*: 9.919719460971242
R2 sem *outliers*: 67.76581988618263
R2 com *outliers*: 61.57532892387621

6. Modelagem com Inteligência Artificial

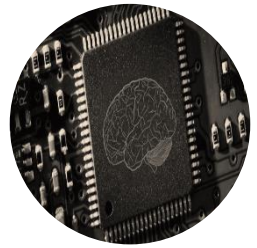
Regressão Linear

Árvore de Regrssão

GBT Regressor

Random Forest Regression

GLR - Gaussian



GBT Regressor – Categoria do Produto

Os resultados do modelo são:

MAE sem *outliers*: 2.3068815598471657
MAE com *outliers*: 5.267211130719013
RMSE sem *outliers*: 3.0397542524028918
RMSE com *outliers*: 9.866085169380018
R2 sem *outliers*: 55.75950674222789
R2 com *outliers*: 61.98971737533314

6. Modelagem com Inteligência Artificial

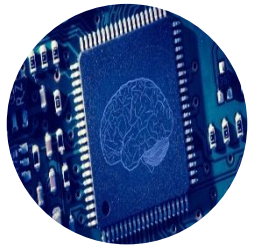
Regressão Linear

Árvore de Regrssão

GBT Regressor

Random Forest Regression

GLR - Gaussian



GBT Regressor – Categoria do Produto

Os resultados do modelo são:

MAE sem *outliers*: 1.737223408500116
MAE com *outliers*: 5.419682592124251
RMSE sem *outliers*: 2.5606264366004683
RMSE com *outliers*: 14.23506543295993
R2 sem *outliers*: 68.65254958422193
R2 com *outliers*: 16.805421774143237

6. Modelagem com Inteligência Artificial

O modelo que apresentou melhores resultados foi o **GBT Regressor** sem a variável do nome da categoria do produto.

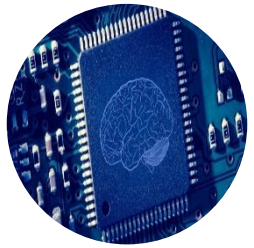
Regressão Linear

Árvore de Regrssão

GBT Regressor

Random Forest Regression

GLR - Gaussian



GBT Regressor – Categoria do Produto

MODELO	MAE	MAE_FULL	RMSE	RMSE_FULL	R2	R2_FULL
GBTRegressor	1.6829742177602103	4.473540942479851	2.434614975890428	9.219827125176893	71.62057516443838	66.80620295232022
RandomForestRegressor	1.8997635232778034	4.865857056867012	2.5946979492735562	9.919719460971242	67.76581988618263	61.57532892387621
DecisionTreeRegressor	1.90685251941861	5.050109013400841	2.652784875178676	10.186470369642779	66.30642631065373	59.48098951685474
LinearRegression	2.301002436385273	5.293480047119417	3.032605165346057	9.862082979772966	55.94360716156693	62.003370347318906
GeneralizedLinearRegression Gaussian	2.3068815598471657	5.267211130719013	3.0397542524028918	9.866085169380018	55.75950674222789	61.98971737533314

MODELO	MAE	MAE_FULL	RMSE	RMSE_FULL	R2	R2_FULL
GBTRegressor	1.6829742177602103	4.473540942479851	2.434614975890428	9.219827125176893	71.62057516443838	66.80620295232022
GBTRegressor Product_Category_Name	1.737223408500116	5.419682592124251	2.5606264366004683	14.23506543295993	68.65254958422193	16.805421774143237

7. Desafios encontrados

Desafios da 1ª entrega – Entendimento inicial da base
Criação de um dicionário de dados

Desafios da 2ª entrega – Dúvidas em relação ao código (Erros no Pyspark)
Escolha do problema a ser resolvido

Desafios da 3ª entrega – Escolha dos modelos corretos
Análise dos resultados

Desafios da 4ª entrega – Elaboração do resumo executivo dos resultados
Busca incansável por melhores resultados

8. Conclusões

Conclusões

1. O projeto foi uma excelente oportunidade de aplicar na **prática** os conhecimentos teóricos adquiridos durante o curso.

2. O trabalho desenvolvido pode ser utilizado para controle dos preços aplicados pelos fornecedores do serviço de entrega, evitando assim **fraudes;**

3. Com os modelos aplicados é possível saber quais localidades de origem-destino o frete ficará mais barato, possibilitando **ações estratégicas de logística** para a Olist;

4. Os resultados do modelo apresentado podem ser utilizados como uma nova *feature*, enriquecendo **futuras análises** como por exemplo, **ações promocionais**.

Próximos Passos

1. Entendimento dos *outilers*, com foco na obtenção de valor destes dados.

Trabalho Final – Turma 12

Caso de Uso: Olist

05/Junho/2020

Coordenadores:
Profª Drª Alessandra de Álvila Montini
Profª Dr. Adolpho Walter Pimazoni Canton

GRUPO 12:

- **Ana Lúcia da Cunha Cox**
- **Thiago Yoshiaki Miyabara Nascimento**