Trabalho Final – Turma 12

Caso de Uso: Olist

05/06/2020 (data da entrega)

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

GRUPO 10:

- **João Paulo Ribeiro dos Santos**

LABDATA | FIA FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

Agenda

1. Objetivo do Trabalho
2. Contextualização do Problema
3. Base de Dados
    i. Bases originais
    ii. Filtros
    iii. Principais variáveis
    iv. Processo de redução de variáveis
4. Análise Exploratória de Dados
5. Modelagem com Estatística Tradicional e Inteligência Artificial
6. Clusterização dos Vendedores
7. Aprimoramento do modelo de regressão
8. Desafios encontrados
9. Conclusões

1. Objetivo do Trabalho

O objetivo do trabalho é **projetar o faturamento dos clientes lojistas** da Olist, nesse caso os **clientes que vendem seus produtos por meio da Olist**. Também é de interesse entender **quais as principais características deles**, mais precisamente seus **perfis.**

A predição será realizada utilizando **dados históricos transacionais** e alguns **modelos estatisticos e de inteligência artificial**, onde conseguirão selecionar quais são as **características utilizadas para predizer o faturamento.**

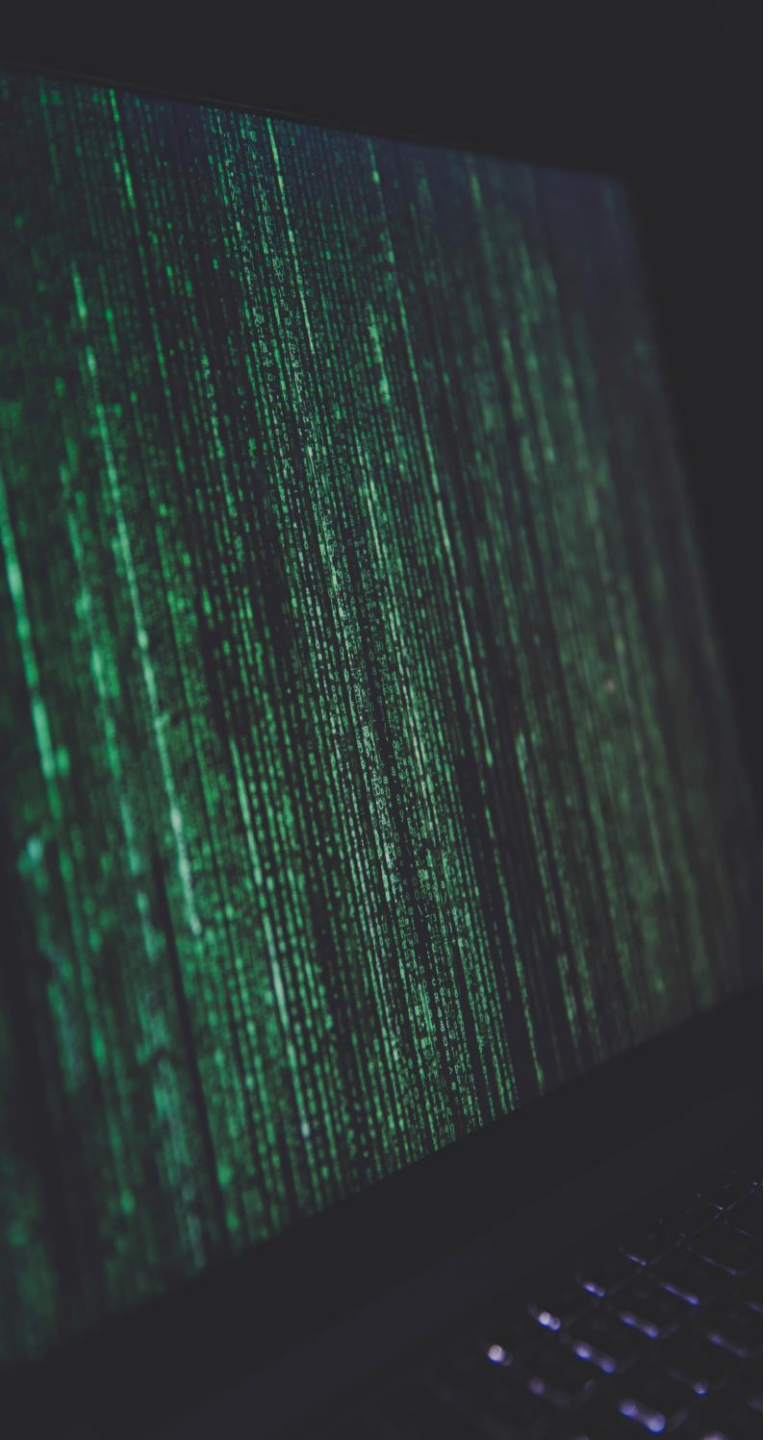
Com todos os processos realizados nesse trabalho, a empresa poderá traçar **estratégias de relacionamento** com seus clientes, **alcançar novos prospect's**, iniciar **ações para aumentar o faturamento de seus clientes** e consequentemente agregar mais valor a si mesmo.

2. Contextualização do Problema

A instituição vem acompanhando uma tendência mundial de informatizar alguns processos, principalmente na **área de vendas**, contudo existem muitos clientes, **principalmente de pequeno porte**, que encontram **muitas dificuldades em migrar para a venda online**. Sendo assim, a empresa analisada possui uma **loja onde muitos clientes podem divulgar seus produtos**, que ficam **visiveis nos maiores marketplaces do país**.

Com o passar dos anos muitos **clientes vem aderido a plataforma**, clientes dos **mais variados perfis**, que vendem os **mais diversos produtos**. Com isso, seria de interesse saber quais os perfis desses clientes afim de **entender suas principais caracteristicas** e iniciar campanhas para melhorar a sua adesão a loja.

Uma vez sabendo o perfil de seus clientes e as principais caracteristicas de seu negócio, será possivel **predizer o faturamento que o cliente terá na plataforma**, que é deveras util para ele **saber onde chegará no ritmo atual**, e ao mesmo tempo é util para **a aquisição de prospects**, cosiderando o **faturamento que podem ter na plataforma**, e o **diferencial em relação aos concorrentes.**



3. Base de Dados

A base de dados utilizada está hospedada em um bucket na Amazon, e todo o processo de analise, detecção e elaboração dos modelos estatísticos foi feito na plataforma databricks

Amazon S3

databricks

3.i. Bases originais

- **8 Bases de Dados**, sendo essas correspondentes aos registros de **pedidos, produtos, clientes finais, vendedores, dados geográficos, pagamentos** e **reviews**;
- Existem ao todo **48 variáveis/ colunas/ parâmetros nativos**;
- O periodo analisado é **Setembro/2016** até **Outubro/2018**;

Clientes
+ 99.000

Vendedores
+ 3.000

Pedidos
+ 99.000

Produtos
+ 32.000

Review
100.000

3.ii. Filtros

Base Inicial

Base original com as informações do negócio, nesse caso com o registro de **3.095 vendedores**

3.iii. Principais variáveis

Variáveis cadastrais

- Geolocalização do vendedor;
- Quantidade de meses na plataforma;

Variáveis dos Pedidos

- Data do primeiro pedido atendido;
- Data do último pedido atendido;
- Quantidade de pedidos;
- Quantidade de produtos que vende;
- Preço médio de seus produtos;
- Preço médio do frete;
- Distância média entre o vendedor e seus clientes;

Variáveis relacionadas a review

- Média das avaliações do vendedor, que vão de 0 a 5

Variável Resposta

Total Faturado: Corresponde ao faturamento total que o vendedor teve na loja

3.iv. Processo de redução de variáveis

Quantidade original

De acordo com as características do negócio, foram recebidas 48 variáveis

Novas Variáveis

Foram criadas algumas variáveis, para complementar a analise e ao mesmo tempo para a construção do algoritmo

Explorando os Dados

Conforme foi feito a analise exploratória foram criadas algumas variáveis relacionadas aos pedidos, principalmente as relacionadas as datas.

Variaveis do modelo

Para o modelo de regressão foram utilizadas 4 variáveis, e para o modelo de clusterização 7

Aprimorando as informações

Para o modelo de regressão utilizado, no final 5 variáveis conseguiram explicar com maior assertividade o evento de estudo

4. Análise Exploratória de Dados

O processo de analise exploratoria utilizou desde métricas estatisticas tradicionais, até detecção de outliers, valores nulos e distribuição dos dados

4. Análise Exploratória de Dados

Variáveis cadastrais

- Geolocalização do vendedor;
- Quantidade de meses na plataforma;

Persona

- A maior parte deles está concentrada nas regiões **Sudeste (74%)** e **Sul (22%);**
- Os **estados** que possuem **mais vendedores** são **SP (60%)** seguido de **PR (11%)** e **SC (6%);**
- Considerando o período analisado de **Set/ 2016** a **Out/ 2018**, podemos dizer que **a base de vendedores é jovem,** sendo que **metade dos vendedores** utiliza a plataforma **a 4 meses ou menos**

Detalhes das análises

4. Análise Exploratória de Dados

Variáveis dos Pedidos

- Data do primeiro pedido atendido;
- Data do último pedido atendido;
- Quantidade de pedidos;
- Quantidade de produtos que vende;
- Preço médio de seus produtos;
- Preço médio do frete;
- Distância média entre o vendedor e seus clientes;

Persona

- **Metade** dos vendedores vendeu **até 8 pedidos;**
- A maior parte dos vendedores vendeu até **10 produtos (75%);**
- A **média do preço** dos produtos vendidos pela **metade dos vendedores** é de **R$ 95,00**

Detalhes das análises

4. Análise Exploratória de Dados

Variáveis relacionadas a review

- Média das avaliações do vendedor, que vão de 0 a 5

Persona

- A maior parte dos reviews dado aos pedidos e vendedores é muito **positiva (58%)**, possuindo **score máximo;**
- **Metade** dos vendedores possui uma **avaliação média** de **4 estrelas**, em uma escala de **0 a 5 estrelas**

Detalhes das análises

4. Análise Exploratória de Dados

Variável Resposta

Total Faturado: Corresponde ao faturamento total que o vendedor teve na plataforma

Persona

- Dado que **metade dos vendedores** utilizam a loja Olist a **pouco tempo**, o seu faturamento chegou até R$ 820,00;
- Há **vendedores** que utilizam a **plataforma a mais tempo** e alcançaram um **faturamento total**, com uma **média de R$ 147.000,00**

5. Modelagem com Estatística Tradicional e Inteligência Artificial

- Ao logo de todo o projeto foram utilizados os seguintes modelos para o processo de regressão:
  - **Regressão Linear/ Linear Regression**
  - **Árvore de Regressão/ Regression Tree**
  - **Floresta Aleatória/ Random Forest**
  - **Gradient Boosted**

5. Modelagem com Estatística Tradicional e Inteligência Artificial

BASES DE TREINO E TESTE

Tratamento das base de dados para modelagem

1. 70% aleatório para treino e 30% para teste
   - Treino: 2172 vendedores
   - Teste: 923 vendedores

5. Modelagem com Estatística Tradicional e Inteligência Artificial

Regressão Linear

É uma técnica que gera uma equação que descreve a relação estatística entre variáveis de entrada e de saída, basicamente ela nos mostra se existe uma relação entre variáveis preditora e uma variável alvo.

$$Yi = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_n X_{in} + \varepsilon_i$$

Variável	Coeficiente (β)
Intercepto	-1552,21
Quantidade de Pedidos	82.24
Quantidade de Produtos	48.55
Quantidade de Meses na Loja	140.99
Preço médio dos Produtos	9,71

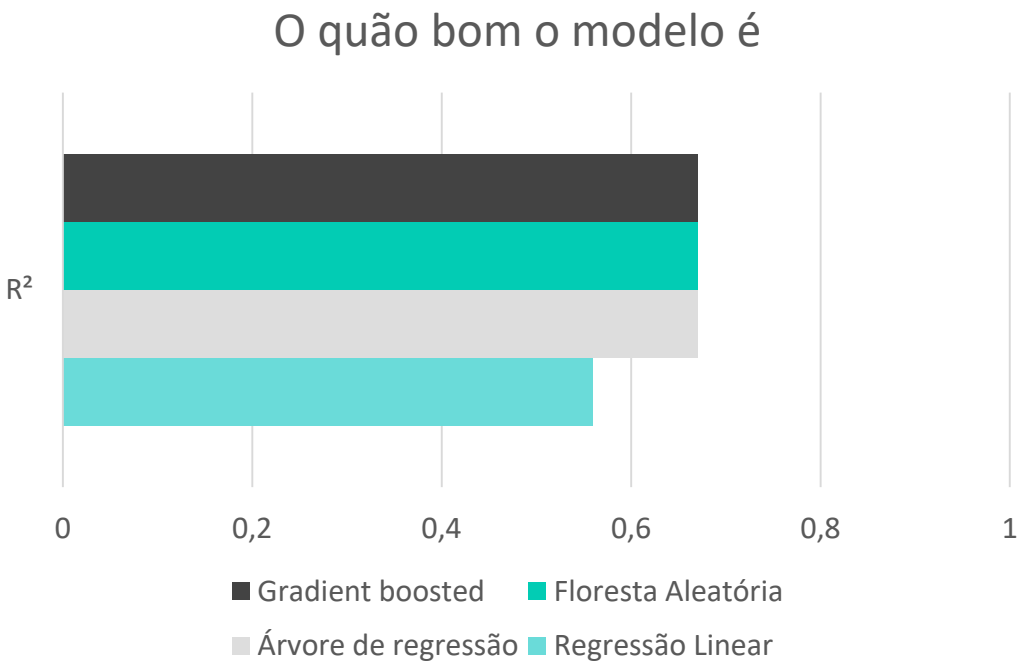
Primeiras Variáveis do Modelo

5. Modelagem com Estatística Tradicional e Inteligência Artificial

Comparação com demais algoritmos

Ao logo de todo o projeto foram utilizados os seguintes modelos para o processo de regressão:

- **Regressão Linear/ Linear regression**
- **Árvore de Regressão/ Tree regression**
- **Floresta Aleatória/ Random Fores**
- **Gradient Boosted**



Com base em todos os modelos apresentados, o que teve maior assertividade e menor processamento foi o de **Arvore de Regressão**, pois **67% da variabilidade** da variável alvo é explicada pelo modelo, e ele **erra em média R$ 1.748** ao predizer o faturamento

6. Clusterização dos Vendedores

A Olist possui **dois tipos de clientes**, os **clientes finais** que compram nos mais diversos marketplaces do pais, e os **clientes lojistas** que vendem na loja da Olist, para os clientes finais.

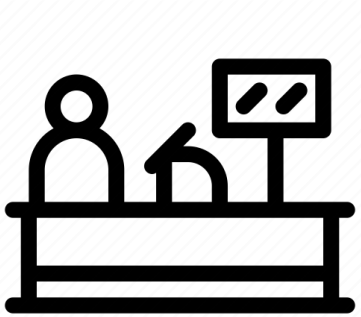
O processo de clusterização vai c**olocar em um mesmo grupo vendedores com características parecidas**, e ao mesmo tempo garantir que tais **grupos sejam diferentes**.

Para tal processo será utilizado o algoritmo K-Means

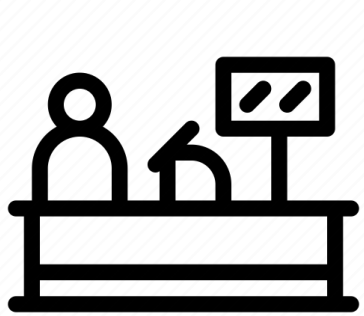
6. Clusterização dos Vendedores

Analise dos Perfis

Ao todo foram identificados **três perfis**, cujo vendedores são muito parecidos em seus respectivos grupos, MAS são diferentes em relação aos outros grupos.



- **19** Vendedores;
- Média de **1.120 pedidos** por vendedor;
- Média de **145** produtos ;
- Média de **faturamento em R$ 146.600,00**
- Utilizam a loja a uma média de 17 meses



- **110** Vendedores;
- Média de **241** pedidos por vendedor;
- Média de **63** produtos ;
- Média de **faturamento em R$ 36.300,00**
- Utilizam a loja a uma média de 14 meses

- **2996** Vendedores;
- Média de **21 pedidos** por vendedor;
- Média de **8 produtos** ;
- Média de faturamento em **R$ 2.296,00**
- Utilizam a loja a uma média de 5 meses

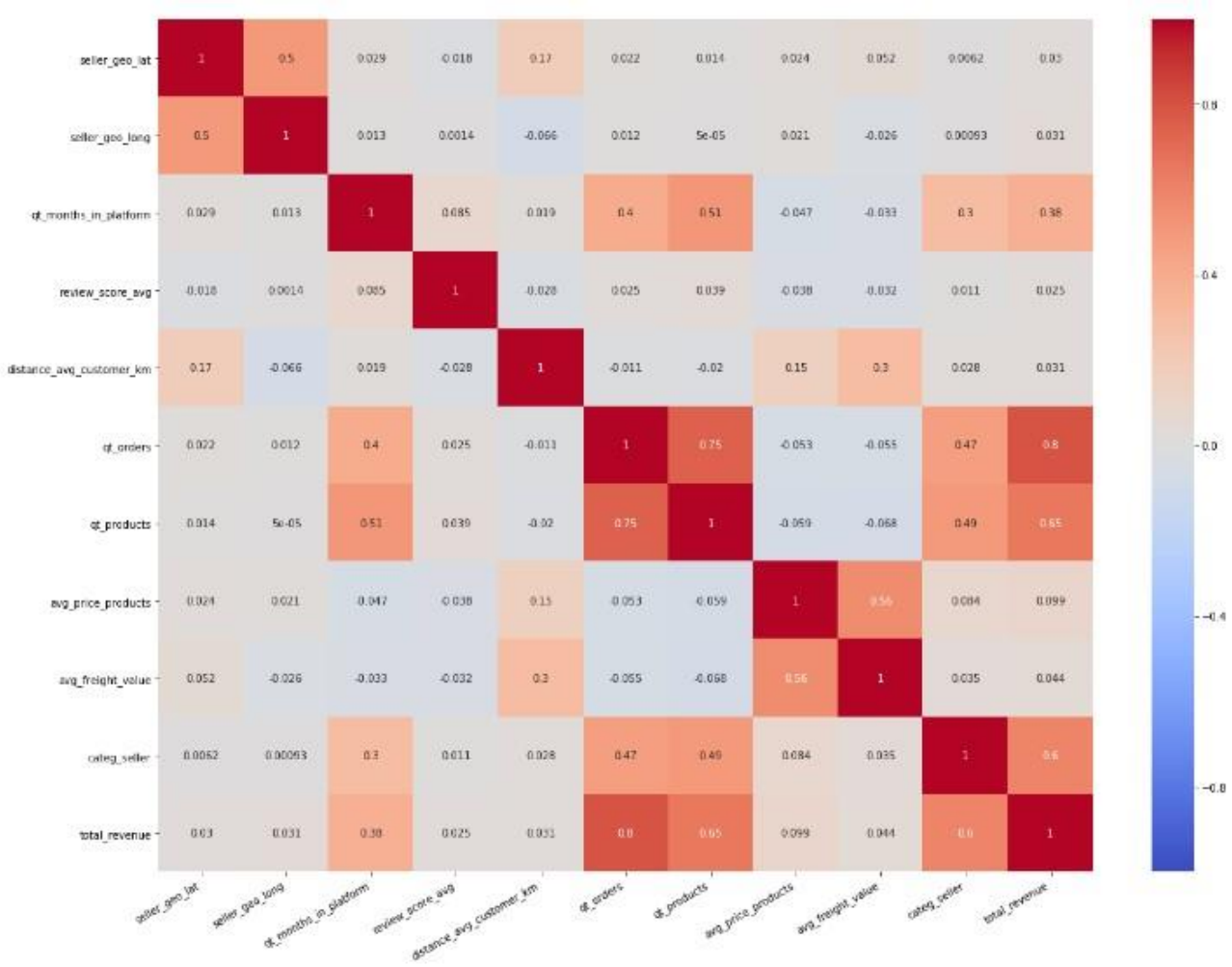
7. Aprimoramento do modelo de regressão

A categoria do vendedor poderia ajudar na predição de seu faturamento ?

7. Aprimoramento do modelo de regressão

Correlação dos Dados

A categoria do vendedor ajuda a explicar com maior assertividade o faturamento do vendedor ?

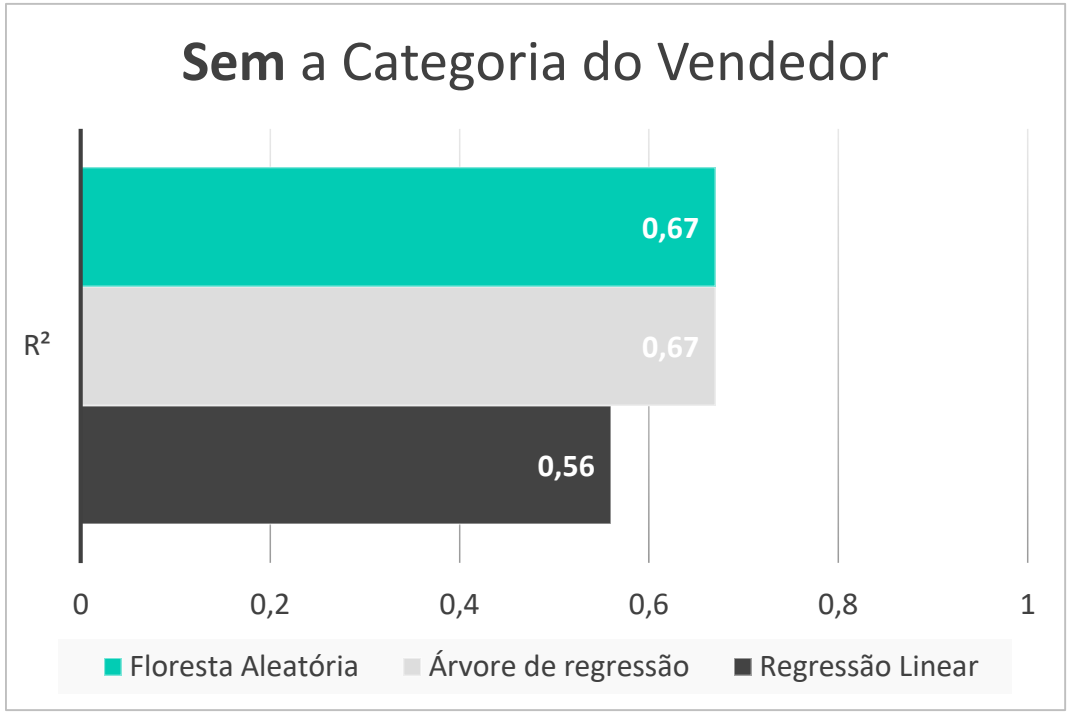


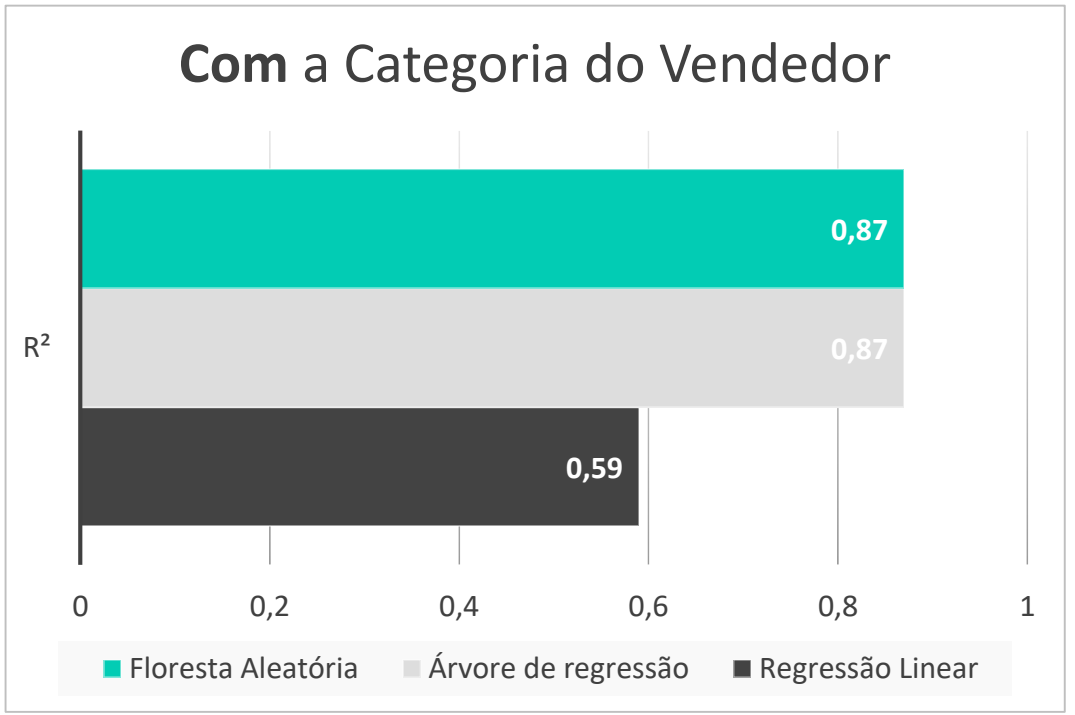
A correlação ajudou a identificar que **a categoria do vendedor seria uma boa variável** para os modelos, dado que sua **correlação foi boa** em relação não somente a **variável alvo**, com **as demais variáveis preditoras**

7. Aprimoramento do modelo de regressão

Comparação dos modelos

A categoria do vendedor ajuda a explicar com maior assertividade o faturamento !!!





8. Desafios encontrados

Os maiores desafios encontrados no projeto foram:

- **Entendimento do Negócio**
- **Interpretação e Analise dos Dados**
- **Ligeiras dificuldades com a plataforma DataBricks, no que diz respeito a performance**

Embora tenha havido essas dificuldades, o trabalho foi muito gratificante

9. Conclusões

- Ao todo existem três perfis de vendedores, sendo eles resumidamente:
  - **Grandes Vendedores**
  - **Médios Vendedores**
  - **Pequenos Vendedores**

- A **Árvore de Regressão** apresentou **melhor desempenho** com comparação com os demais algoritmos, onde **87% da variabilidade foi explicada pelo modelo.**

- **5 variáveis** compõe a **Árvore de Regressão** e predizem o faturamento do vendedor:
  - **Quantidade de pedidos**
  - **Quantidade de produtos**
  - **Quantidade de meses vendendo na loja**
  - **Média de preço dos produtos**
  - **Categoria do vendedor**

LABDATA FIA – Laboratório de Análise de Dados

Unidade Pinheiros

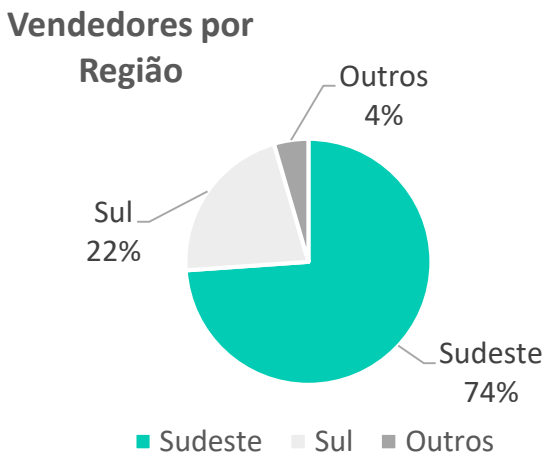
Unidade Paulista

Detalhes da Analise Exploratória

Variáveis cadastrais

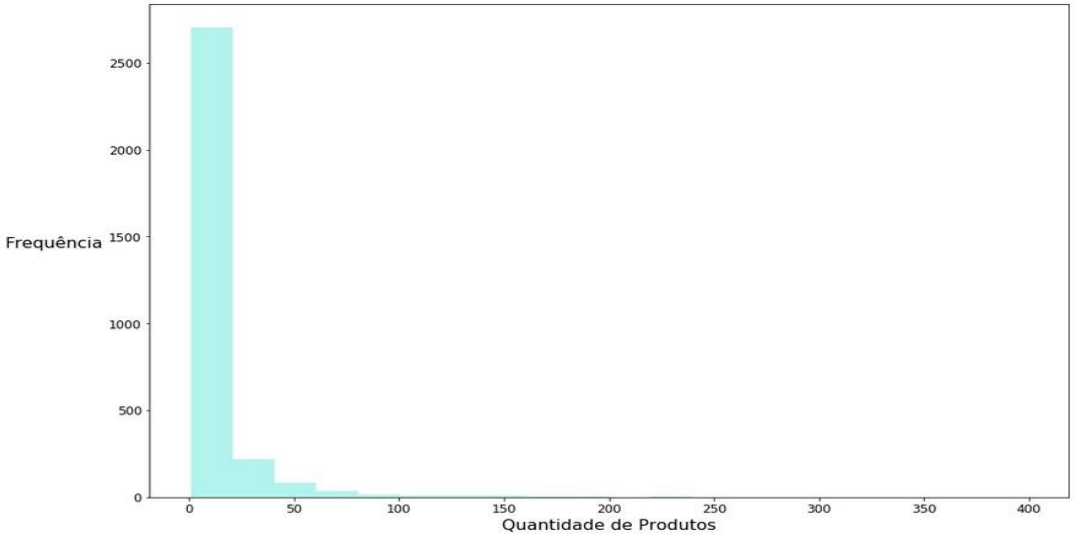
Quantidade de Meses na Loja	
Medida	Valor
Mínimo	0
1º Quartil	1
Mediana	4
Média	5
3º Quartil	9
Máximo	23

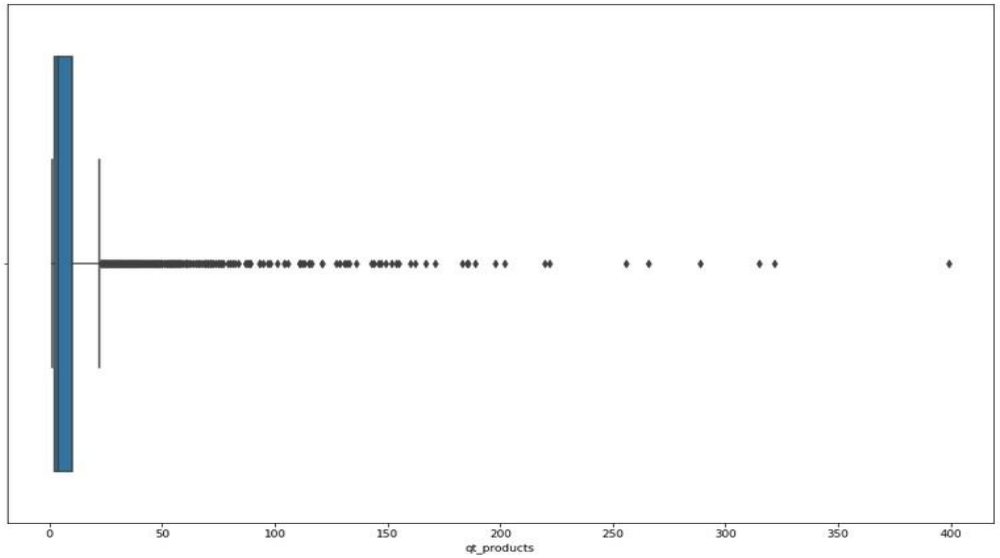
SP e **PR** compreendem sozinhos **71%** dos vendedores



Detalhes das Análises

Variáveis dos Pedidos





Quantidade de Pedidos por Vendedor	
Medida	Valor
Mínimo	1
1º Quartil	2
Mediana	8
Média	36
3º Quartil	24
Máximo	2033

Detalhes das Análises

Variáveis relacionadas a review

Média dos Reviews Por Vendedor	
Medida	Valor
Mínimo	1
1º Quartil	3
Mediana	4
Média	4
3º Quartil	4.5
Máximo	5

Detalhes das Análises

Variáveis cadastrais

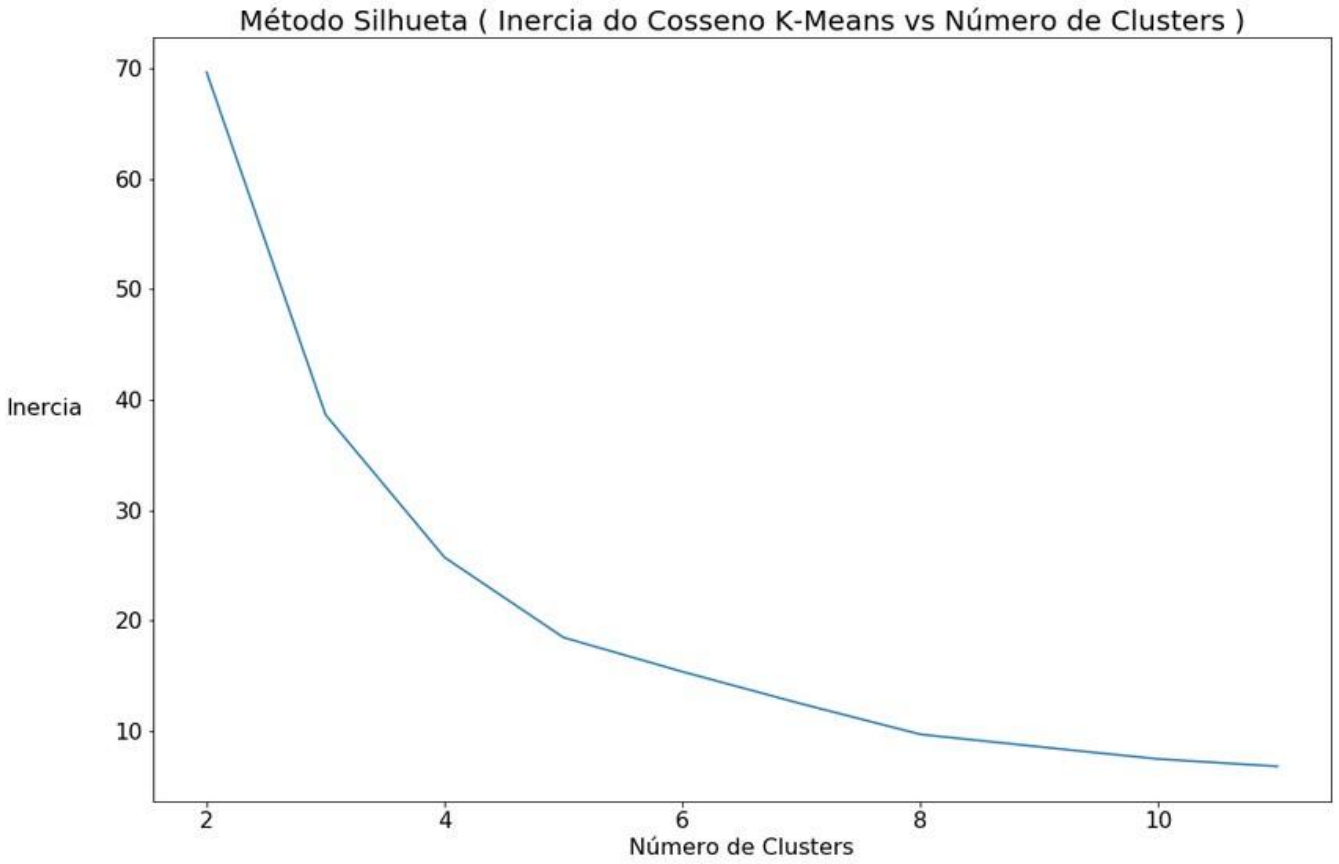
Modelo	R^2	Erro médio absoluto (MAE)	Raiz do erro quadrático (RMSE)
Regressão Linear	0,58	2405,07	7346,15
Árvore de Regressão	0,67	1748,18	6368,17
Floresta Randomica	0,67	1748,18	6368,17
Gradiente Boosted	0,67	1748,18	6368,17

Comparação entre modelos após clusterização

SEM a categoria do vendedor			
Modelo	**R^2**	**Erro médio absoluto (MAE)**	**Raiz do erro quadrático (RMSE)**
Regressão Linear	0,56	2405,07	7346,15
Árvore de Regressão	0,67	1748,18	6368,17
Floresta Randomica	0,67	1748,18	6368,17

COM a categoria do vendedor			
Modelo	**R^2**	**Erro médio absoluto (MAE)**	**Raiz do erro quadrático (RMSE)**
Regressão Linear	0,59	2450,86	10678,38
Árvore de Regressão	0,87	1771,36	5973.88
Floresta Randomica	0,87	1771,36	5973.88

Quantidade adequada de clusters



Com a escolha de **3** clusters foi possível obter uma **silhueta com um valor de aproximadamente 0,95**

5. Regressão Linear

Primeira Submissão de variáveis ao Modelo

$$Yi = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_n X_{in} + \varepsilon_i$$

Variável	P-Valor
Intercepto	
Estado	0.028515
Quantidade de meses na Plataforma	0.000089
Média de Reviews	0.560851
Distancia média Clientes	0.367613
Quantidade de Pedidos	0.000000
Quantidade de Produtos	0.000007
Preço médio dos Produtos	0.000000
Preço médio do Frete	0.747639