Modelo de Recomendações Olist

Ellaynne Christine R. de Moraes Sousa

13/11/2019

Entendimento do problema

O conjunto de dados a seguir apresenta informações referentes à vendas realizadas na Olist, uma loja de departamentos dos marketplaces. A proposta deste trabalho é criar um Modelo de Recomendações (SR) para sugerir produtos que os clientes possam gostar. A abordagem trabalhada para esse SR foi a Filtragem Baseada em Conteúdo, já a maioria dos dados disponibilizados possuem mais características sobre os produtos à venda. A Filtragem Baseada em Conteúdo analisa os itens que foram avaliados pelo usuário para lhe sugerir itens parecidos. A target, isto é, a variável alvo do modelo é a nota de avaliação que os clientes deram para produtos já adquiridos. Assim, será realizada uma análise exploratória sobre os dados, olhando-se questões importantes como a correlação entre variáveis, que mostrem, por exemplo, como a categoria e dimensões do produto podem influenciar na avaliação dos clientes, e, posteriormente a criação de uma modelo que possa dizer qual possível nota um cliente daria para determinado produto, analisando as características desses produtos.

Entendimento dos Dados

Os dados aqui utilizados são resultado de uma extração realizada a partir de um conjunto de datasets disponibilizado pela Olist. As variáveis foram escolhidas pela própria autora desse trabalho, que se baseou nos conceitos da Filtragem Baseada em Conteúdo, levando em consideração as informações disponíveis sobre os produtos. As avaliações consideradas na extração foram aquelas de pedidos que possuíam apenas um produto.

Primeira visualização dos dados

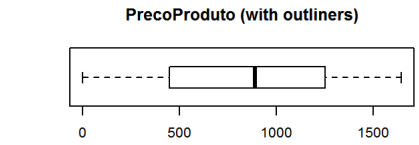
O dataset é composto por 13 variáveis, sendo a variável dependente (target) a “NotaAvaliacao”. As demais variáveis são referentes ao Id da Avaliação, Status do Pedido, Preço do Produto, Categoria, Quantidade de Fotos, Peso, Comprimento, Altura, Largura, Cidade do Vendedor, Dias para Envio e Dias Previstos para Entrega.

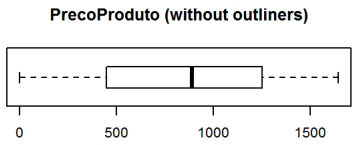
```
summary(train)
```

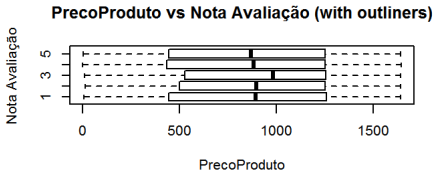
```
##                           IdAvaliacao      StatusPedido  NotaAvaliacao
##  3242cc306a9218d0377831e175d62fbf:   2   delivered:7709   Min.   :1.000
##  0005bc906dfe3b73261020bf8e9b76cf:   1                   1st Qu.:4.000
##  000688fa7009b3c9fafe4950b37cbc81:   1                   Median :5.000
##  0010388b006db42c9457d7148035db0e:   1                   Mean   :4.191
##  0016d2d10b5c26607549a29446578edf:   1                   3rd Qu.:5.000
##  001ce468e08e9ee56753421128bec6f4:   1                   Max.   :5.000
##  (Other)                         :7702
##   PrecoProduto                    Categoria    QuantidadeFotos
##  Min.   :   1.0   cama_mesa_banho       : 701   Min.   : 1.000
##  1st Qu.: 444.0   beleza_saude          : 663   1st Qu.: 1.000
##  Median : 888.0   esporte_lazer         : 594   Median : 2.000
##  Mean   : 864.3   informatica_acessorios: 538   Mean   : 2.272
##  3rd Qu.:1251.0   moveis_decoracao      : 496   3rd Qu.: 3.000
##  Max.   :1646.0   utilidades_domesticas : 484   Max.   :15.000
##                   (Other)               :4233   NA's   :97
##       Peso          Comprimento         Altura          Largura
##  Min.   :    0.0   Min.   :  7.00   Min.   :  2.00   Min.   :  9.00
##  1st Qu.:  286.8   1st Qu.: 18.00   1st Qu.:  8.00   1st Qu.: 15.00
##  Median :  700.0   Median : 24.00   Median : 13.00   Median : 20.00
##  Mean   : 2077.4   Mean   : 29.77   Mean   : 16.33   Mean   : 23.02
##  3rd Qu.: 1813.0   3rd Qu.: 37.00   3rd Qu.: 20.00   3rd Qu.: 30.00
##  Max.   :30000.0   Max.   :105.00   Max.   :105.00   Max.   :100.00
##  NA's   :1         NA's   :1        NA's   :1        NA's   :1
##         CidadeVendedor   DiasEnvio      DiasPrevistoEntrega
##  sao paulo     :1959   Min.   : 2.000   Min.   : 3.00
##  ibitinga      : 468   1st Qu.: 5.000   1st Qu.:19.00
##  curitiba      : 225   Median : 6.000   Median :24.00
##  santo andre   : 203   Mean   : 6.604   Mean   :24.23
##  belo horizonte: 196   3rd Qu.: 8.000   3rd Qu.:29.00
##  ribeirao preto: 166   Max.   :48.000   Max.   :91.00
##  (Other)       :4492
```

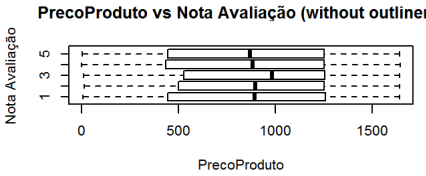
No detalhamento das variáveis mostrado acima, identificá-se que há uma variável, “StatusPedido”, que possui valor constante, e, por isso, pode ser descartada por não agregar valor na análise, já que todas os valores são o mesmo: “delivered”. E há uma variável identificadora, IdAvaliação, que também pode ser descartada por não influênciar na busca de correlação como a variável target. Assim, essas variáveis citadas não serão analisadas. Outra coisa importante a se notar é que as variáveis “QuantidadeFotos”, “Peso”, “Comprimento”, “Altura” e “Largura” possuem missings, isto é, dados nulos.

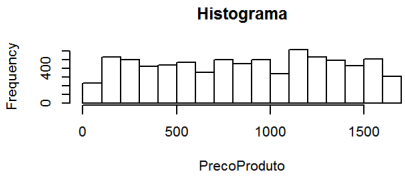
Análise descritiva das variáveis numéricas:

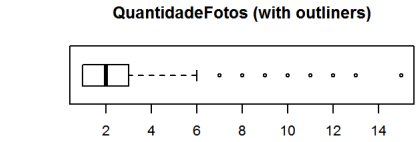

PrecoProduto (with outliners)
0
500
1000
1500

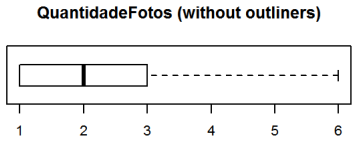

PrecoProduto (without outliners)
0
500
1000
1500

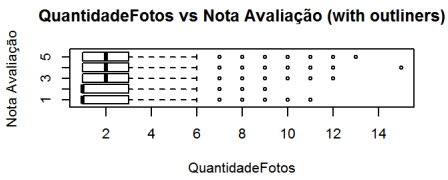

PrecoProduto vs Nota Avaliação (with outliners)
Nota Avaliação
1
3
5
0
500
1000
1500
PrecoProduto

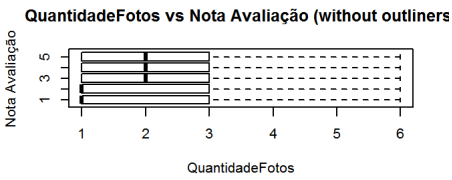

PrecoProduto vs Nota Avaliação (without outliners)
Nota Avaliação
1
3
5
0
500
1000
1500
PrecoProduto

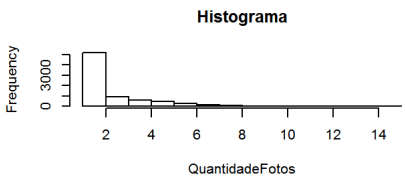

Histograma
Frequency
0
400
0
500
1000
1500
PrecoProduto

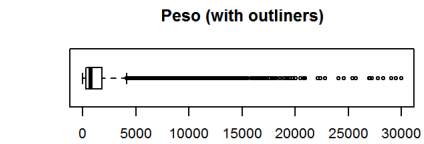

QuantidadeFotos (with outliners)
2
4
6
8
10
12
14

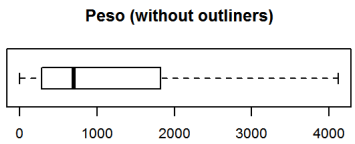

QuantidadeFotos (without outliners)
1
2
3
4
5
6

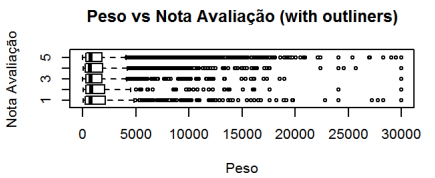

QuantidadeFotos vs Nota Avaliação (with outliners)
Nota Avaliação
1
3
5
2
4
6
8
10
12
14
QuantidadeFotos

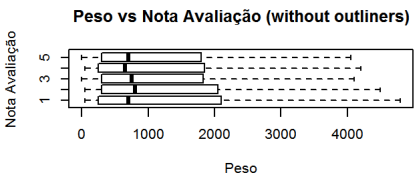

QuantidadeFotos vs Nota Avaliação (without outliners
Nota Avaliação
1
3
5
1
2
3
4
5
6
QuantidadeFotos

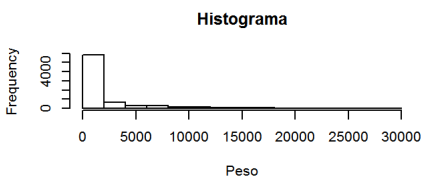

Histograma
Frequency
0
3000
2
4
6
8
10
12
14
QuantidadeFotos

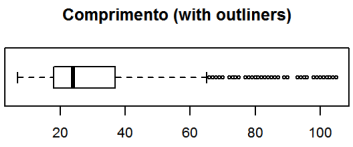

Peso (with outliners)
0
5000
10000
15000
20000
25000
30000

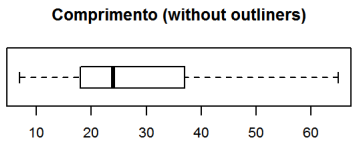

Peso (without outliners)
0
1000
2000
3000
4000

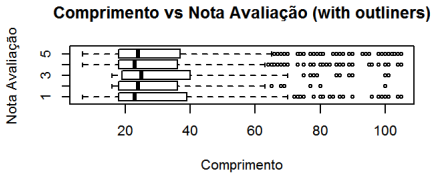

Peso vs Nota Avaliação (with outliners)
Nota Avaliação
1
3
5
0
5000
10000
15000
20000
25000
30000
Peso

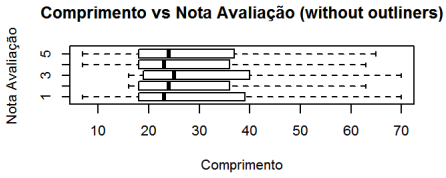

Peso vs Nota Avaliação (without outliners)
Nota Avaliação
1
3
5
0
1000
2000
3000
4000
Peso

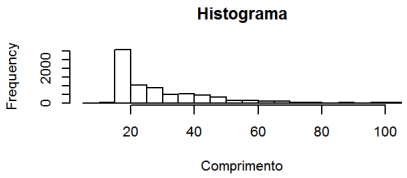

Histograma
Frequency
0
4000
0
5000
10000
15000
20000
25000
30000
Peso

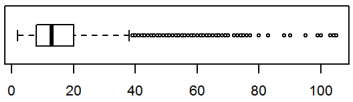

Comprimento (with outliners)
20
40
60
80
100

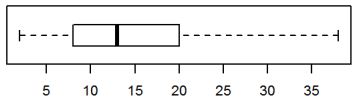

Comprimento (without outliners)
10
20
30
40
50
60

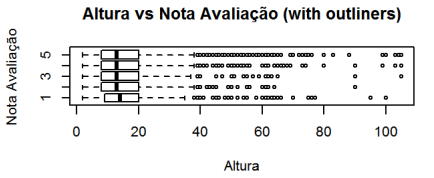

Comprimento vs Nota Avaliação (with outliners)
Nota Avaliação
1
3
5
20
40
60
80
100
Comprimento

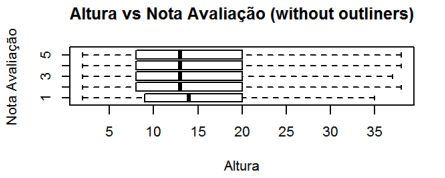

Comprimento vs Nota Avaliação (without outliners)
Nota Avaliação
1
3
5
10
20
30
40
50
60
70
Comprimento

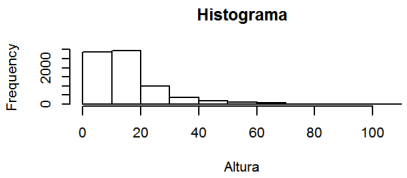

Histograma
Frequency
0
2000
20
40
60
80
100
Comprimento

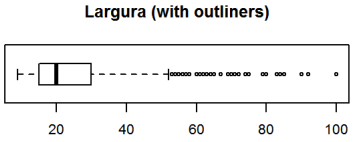

Altura (with outliners)
0
20
40
60
80
100

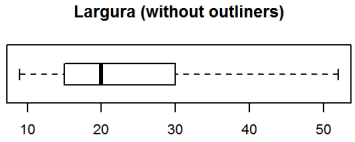

Altura (without outliners)
5
10
15
20
25
30
35

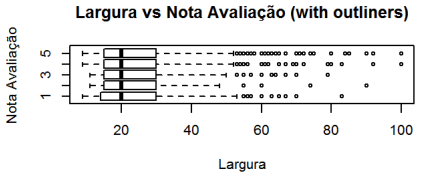

Altura vs Nota Avaliação (with outliners)
Nota Avaliação
1
3
5
0
20
40
60
80
100
Altura

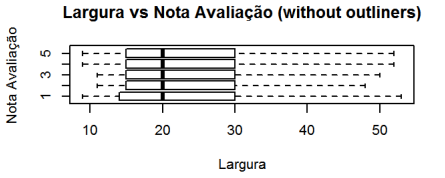

Altura vs Nota Avaliação (without outliners)
Nota Avaliação
1
3
5
5
10
15
20
25
30
35
Altura

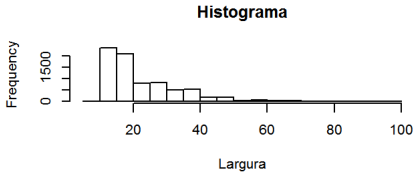

Histograma
Frequency
0
2000
0
20
40
60
80
100
Altura

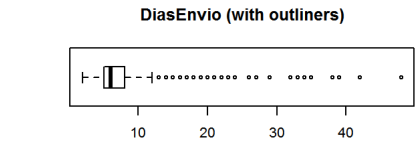

Largura (with outliners)
20
40
60
80
100

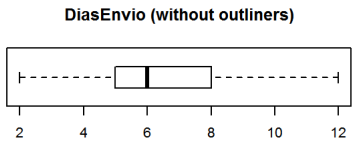

Largura (without outliners)
10
20
30
40
50

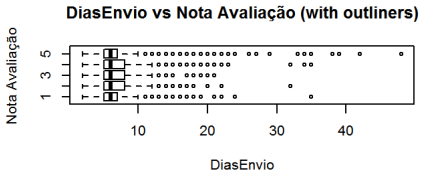

Largura vs Nota Avaliação (with outliners)
Nota Avaliação
1
3
5
20
40
60
80
100
Largura

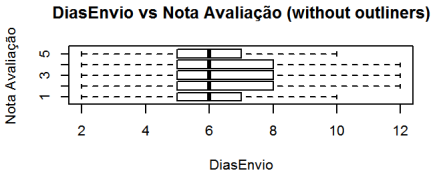

Largura vs Nota Avaliação (without outliners)
Nota Avaliação
1
3
5
10
20
30
40
50
Largura

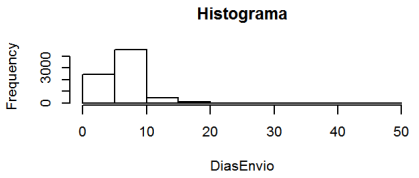

Histograma
Frequency
0
1500
20
40
60
80
100
Largura

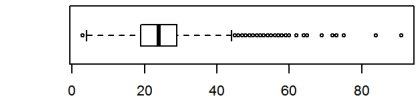


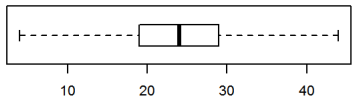


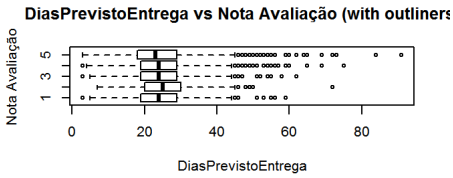


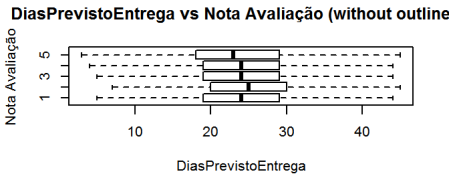


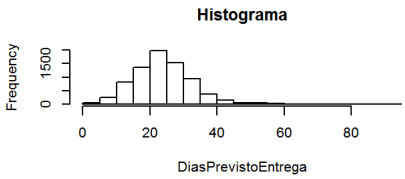












• Preço

A faixa de preços dos produtos vai de 1 a 1646 e a média é de 864,30. Metade dos produtos custam entre 444 e 1.251. Os produtos que receberam nota 3 são os que possuem maior mediana de preço. As demais notas não demonstram uma grande variação em relação ao preço. Isso pode significar que o preço não possui um correlação com a nota atribuída.

• Quantidade de Fotos

A quantidade de fotos do produto no anúncio varia entre 1 e 15 e a maioria dos produtos possui entre 1 e 3 fotos, sendo a média de 2 fotos por produto. Há outliers,isto é, valores distantes dos quatro quartis. Isso quer dizer que alguns produtos possuem mais de 6 fotos no seu anúncio.

Através do boxplot dessa variável e da target(NotaAvaliacao), nota-se que os produtos que tiveram menores notas são os que possuem poucas fotos.

• Peso

O peso médio dos produtos é de 2077,4g, sendo o mínimo de 0g e o máximo de 30000g. Em análise ao boxlot bivariado formado com essa variável e a target, nota-se um pequena variação no valor da nota em relação aos pesos dos produtos. Então, não há uma relação muito significativa entre Peso e Nota.

• Comprimento

A maioria dos itens vendidos possuem entre 7cm e 70cm, aproximadamente. Há uma variância entre as medianas das notas em relação ao comprimento do produto, que pode indicar uma correlação média entre essas variáveis. Todos os produtos que receberam notas 2 ou 3 possuem mais de 15cm.

• Altura

A altura varia entre 2cm e 105cm, sendo que 50% dos itens têm entre 7cm e 20cm. Os outliers mostram que há itens que vão de 40cm a 100cm. No boxplot bivariado, a única variância que se difere das outras foram as avaliações com valor 1.

• Largura

A largura mínima é 9cm e a máxima 100cm. A maioria dos produtos possuem até 55cm de largura,aproximadamente. A diferença apresentada entre os boxplots de relação entre avaliação e largura não sugerem um correlação significativa.

• Dias para Envio

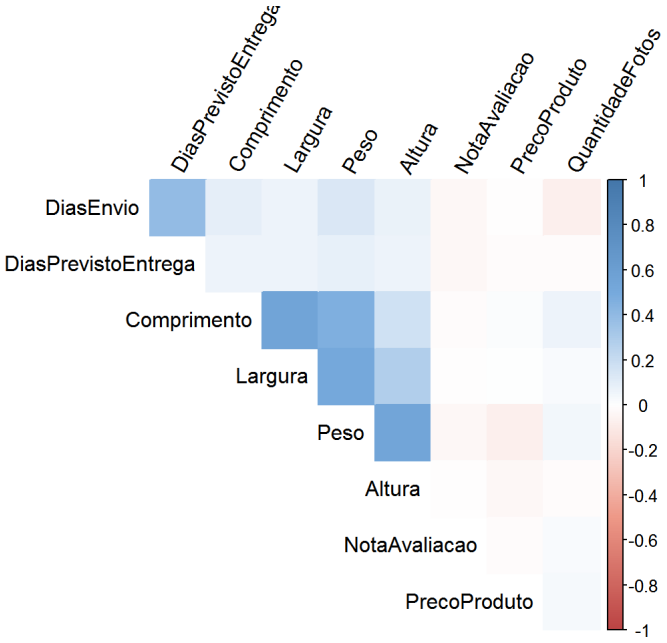
A quantidade de dias para envio do produto pode variar de 2 a 48 dias, mas a maioria deles é despachada com até 8 dias. Os boxplots dessa variável em relação às notas mostram que não há uma variância significativa, com exceção dos produtos que tiveram notas 1 ou 5, o que se pode inferir que essa variável não tem grande influência na nota.

• Dias Previsto Entrega

Metade dos pedidos têm uma previsão de entrega de 19 à 29 dias.

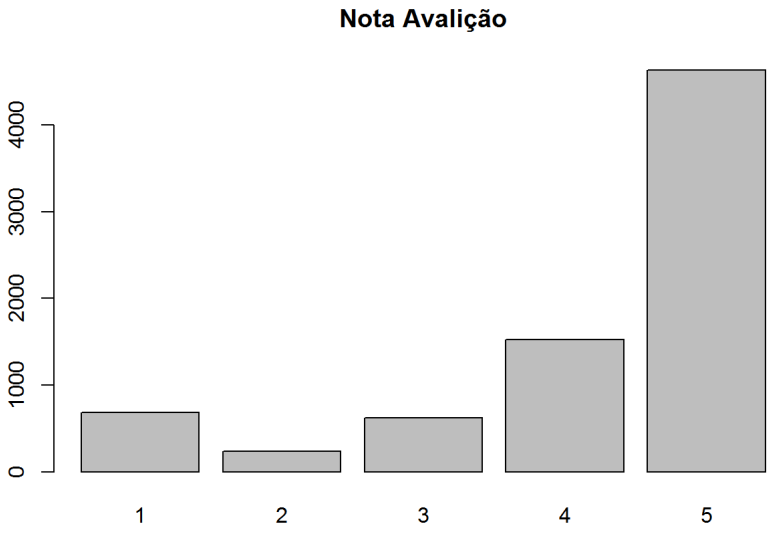
A variância entre os boxplots de notas em relação ao dia previsto de entrega não é grande, porém pode indicar uma certa correlação.

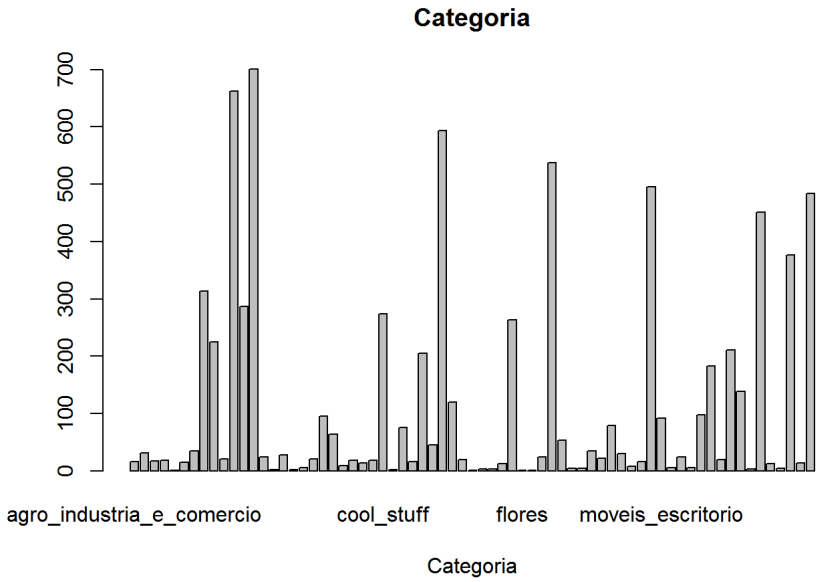
• Correlação entre as variáveis numéricas:

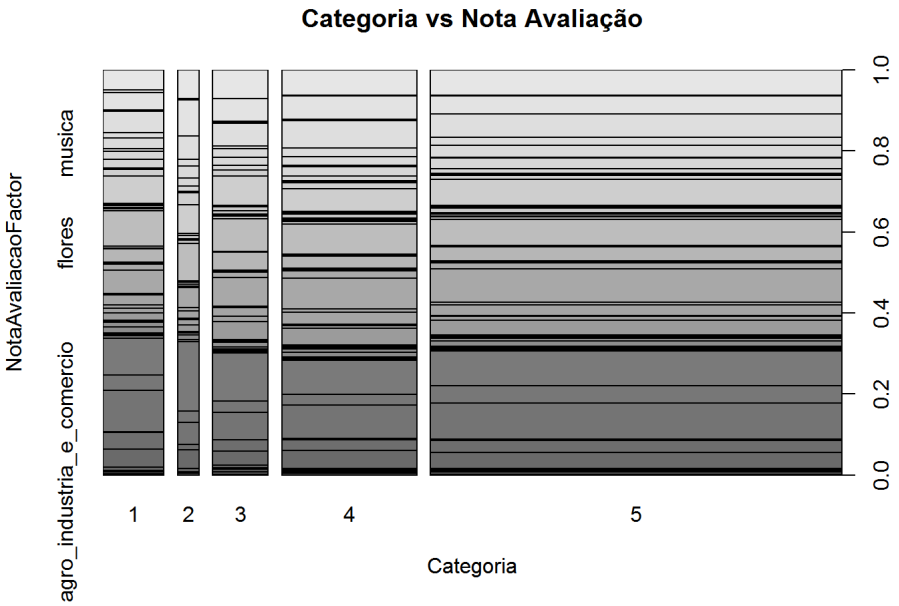


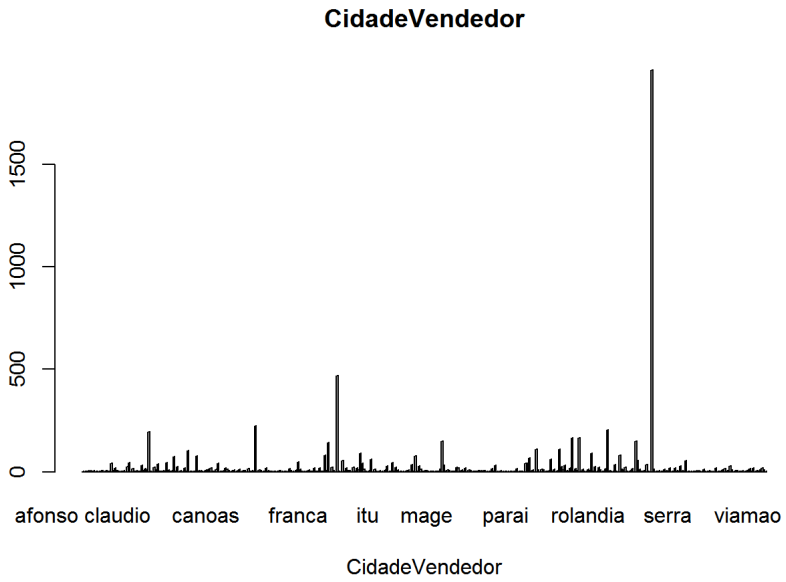
Através da análise do gráfico de correlações é possível identificar que poucas variáveis têm correlação com a target(cores mais claras). Entretanto, há variáveis independentes que têm uma correlação bem alta entre si, como por exemplo Peso e Altura, que possuem uma correlação positiva.

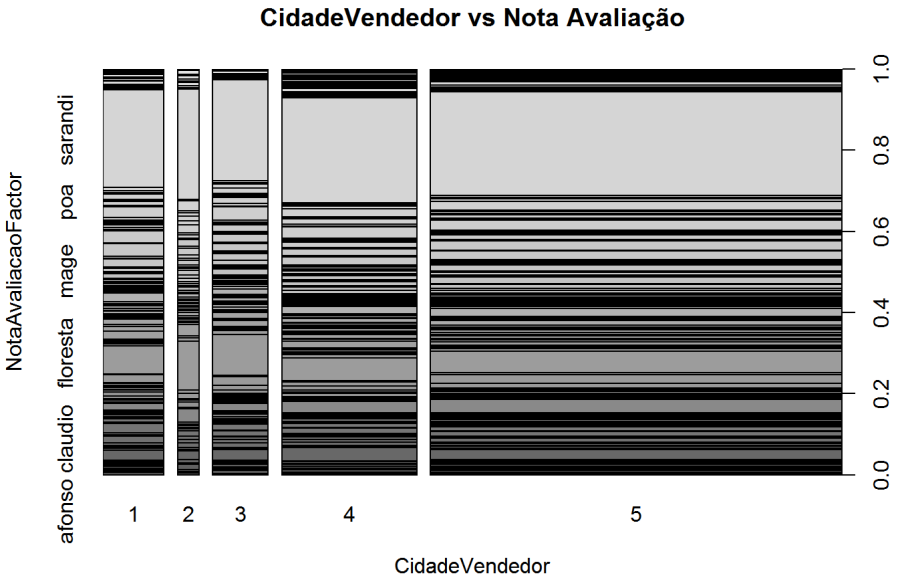
Análise das variáveis categóricas


Nota Avalição
0
1000
2000
3000
4000
1
2
3
4
5


Categoria
0
100
200
300
400
500
600
700
agro_industria_e_comercio
cool_stuff
flores
moveis_escritorio
Categoria


Categoria vs Nota Avaliação
NotaAvaliacaoFactor
agro_industria_e_comercio
flores
musica
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
Categoria





• Nota Avaliação

A maioria das notas foram 5, seguidas de 4. Isso quer dizer que a maioria dos produtos e dos serviços agradaram seu comprador.

• Categoria

De um total de 69 categorias, essas foram as que tiveram maior quantidade de vendas: cama_mesa_banho(701), beleza_saude(663), esporte_lazer(594) e informatica_acessorios(538). As melhores avaliações(nota 5) foram para as categorias beleza_saude, cama_mesa_banho, esporte_lazer e moveis_decoracao. As notas mais baixas (nota 1) foram para as categorias beleza_saude, cama_mesa_banho, informatica_acessorios e moveis_decoracao.

• Cidade Vendedor (LOja)

As três cidade que mais tiveram avaliações nos dados aqui avaliados foram São Paulo (com 1959 avaliações), Ibitinga (468) e curitiba (225) de um total de 385 cidades.

As cidades que mais avaliaram seus pedidos com nota 5 são as mesmas que tiveram maior número de avaliações. E as cidades que mais avaliaram com nota 1 foram São Paulo, Ibitinga e Ribeirão Preto.

Poucas cidades deram nota 2.

Preparação dos Dados

Após descartar variáveis que não agregam valores à análise, conforme já apresentado no tópico anterior, ficaram 10 variáveis independentes e a target (NotaAvaliacao). Das 10 variáveis não dependentes, 8 são numericas e 2 são categóricas.

Apesar de a quantidade dados missings serem bem pequenas, decidiu-se utilizar a técnica de input para esses dados nulos. Para as variáveis QuantidadeFotos, Comprimento, Altura, Largura e Peso os valores missings foram substituídos pelas respectivas médias, pois esse valor não inflenciaria nas demais medidas de posição do conjunto.

Para melhor treinamento do algoritmo, adotou-se a técnica de cross-validation(k-folds) que consiste em dividir o conjunto de dados e realizar o treinamento e validação com essas partes, alternando entre elas o processo de treinamento e validação, aumentando, assim, as possibilidades de aprendizado do algoritmo.

Modelagem

A abordagem de machine learning a ser utilizada nesse trabalho é a de aprendizagem de máquina supervisionado, pois no conjunto de dados utilizado para treinamento já existem as classes de saída já definidas, que são os valores das notas de avaliação : 1, 2, 3, 4 ou 5. E a tarefa a ser executada é uma Regressão, pois o conjunto de labels ou classes esperado são valores continuos, ao contrário de uma tarefa de classificação, onde o resultado esperado possui saída discreta (normalmente Verdadeiro ou Falso).

Para realizada a análise dos dados apresentadas acima, utilizou-se o RStudio, que é um ambiente de desenvolvimento integrado (IDE) para a leitura da linguagem R, uma linguagem de programação para gráficos e cálculos estatísticos. E para realizar a modelagem de machine learning, será utilizado o pacote H2O,que possui vários algoritmos de aprendizagem de máquina já implementados, integrado com o RStudio.

```
myX <- setdiff(colnames(train), c("NotaAvaliacao", "NotaAvaliacaoFactor", "CidadeVendedor", "DiasEnvio"))

myY <- "NotaAvaliacaoFactor"

Model.GBM <- h2o.gbm(
                     x= myX,
                     y = myY,
                     training_frame = train.h,
                     balance_classes = TRUE,
                     nfolds = 5,
                     seed = 1234,
                     model_id = "GBM",
                     ntrees = 35,
                     max_depth = 25,
                     learn_rate = 0.001
                    )
```

```
##
  |
  |                                                                |   0%
  |
  |=                                                               |   1%
  |
  |====                                                            |   6%
  |
  |======                                                          |   9%
  |
  |==========                                                      |  16%
  |
  |===============                                                 |  23%
  |
  |==================                                              |  28%
  |
  |=====================                                           |  32%
  |
  |=========================                                       |  37%
  |
  |================================                                |  50%
  |
  |================================                                |  51%
  |
  |===================================                             |  53%
  |
  |====================================                            |  55%
  |
  |=====================================                           |  57%
  |
  |======================================                          |  59%
  |
  |========================================                        |  61%
  |
  |=========================================                       |  63%
  |
  |==========================================                      |  65%
  |
  |==========================================                      |  67%
  |
  |===========================================                     |  68%
  |
  |============================================                    |  70%
  |
  |============================================                    |  71%
  |
  |=============================================                   |  73%
  |
  |==============================================                  |  75%
  |
  |================================================                |  78%
  |
  |================================================                |  80%
  |
  |=================================================               |  81%
  |
  |================================================                |  83%
  |
  |================================================                |  84%
  |
  |=================================================               |  86%
  |
  |=================================================               |  88%
  |
  |==================================================              |  90%
  |
  |===================================================             |  92%
  |
  |====================================================            |  94%
  |
  |=====================================================           |  96%
  |
  |======================================================          |  98%
  |
  |================================================================| 100%
```

Matriz de confusão da validação

```
## Confusion Matrix: Row labels: Actual class; Column labels: Predicted class
##        1 2 3 4     5  Error              Rate
## 1      0 0 0 0  4641 1.0000 =   4.641 / 4.641
## 2      0 0 0 0  4642 1.0000 =   4.642 / 4.642
## 3      0 0 0 0  4635 1.0000 =   4.635 / 4.635
## 4      0 0 0 0  4631 1.0000 =   4.631 / 4.631
## 5      0 0 0 0  4635 0.0000 =       0 / 4.635
## Totals 0 0 0 0 23184 0.8001 = 18.549 / 23.184
```

Precision: 0.60 Recall: 1.00

Importância das variáveis no treinamento

As variáveis que tiveram mais importância para o modelo na hora de realizar as predições foram Dias Previsto para Entrega, Preço do Produto e Categoria

```
## Variable Importances:
##              variable relative_importance scaled_importance percentage
## 1 DiasPrevistoEntrega        86387.296875          1.000000   0.208342
## 2        PrecoProduto        71470.796875          0.827330   0.172367
## 3           Categoria        66681.078125          0.771885   0.160816
## 4             Largura        43286.554688          0.501075   0.104395
## 5              Altura        43220.730469          0.500313   0.104236
## 6                Peso        41404.683594          0.479291   0.099856
## 7         Comprimento        37790.058594          0.437449   0.091139
## 8     QuantidadeFotos        24401.347656          0.282465   0.058849
```

Conclusão

Muitos testes foram realizados até chegar no algoritmo apresentado: diferentes valores de parâmetros de configuração para a construção dos modelos, além de diferentes combinações de variáveis (independentes) no dataset de treinamento.

O modelo encontrado foi o que obteve o resultado mais interessante dentre outros que não apresentaram bons valores para as métricas de avaliação. Esse modelo é baseado em um algoritmo de árvore de decisão, o GBM (Gradient Boosting Machines). O GBM é um algoritmo de aprendizagem supervisionada que consegue trabalhar tanto com problemas de classificação, como os de regressão. Duas métricas de avaliação apresentaram um bom resultado no modelo para a classe 5: a precisão (60%) e a sensibilidade(100%). A precisão é o número de acertos divido pelo número total de exemplos. E a sensibilidade ou recall é a proporção de acertos de verdadeiros positivos em cima dos exemplos que realmente pertencem. A precisão e o recall com bom valores, temos outra métrica que também apresenta um bom valor: a F-measure(F1), que nada mais é que a média harmonica entre a precisão e o recall.

O modelo acertou todos os exemplos que pertencem à classe 5, apesar de ter errado ao predizer todas as outras classes. Mesmo com esses erros nas classes diferentes de 5, o modelo consegue recomendar ao cliente aqueles produtos que seriam avaliados com nota máxima.