Dúvidas aula 02/02/2022

* Variável `Publisher`

Essa variável tem um problema que são as variáveis com caracteres especiais, tipo editoras com nome escrito em Mandarim.

Quando eu tentava transformar a variável para fator essas informações eram perdidas e os nomes ficavam ilegíveis, Marcus disse que poderia tentar transformar esse vetor como caracter ASC.

No entanto, por ter muitas categorias, ele disse que isso poderia dificultar a modelagem, pq no momento que o RF fosse criar as arvores de decisões as ramificações ficariam gigantes, o que computacionalmente iria exigir muito da máquina e provavelmente a gente não tem esse poder computacional, então ele sugeriu removermos essa variável.

* Modelo de regressão ou classificação?

Quando eu fiz primeira limpeza nos dados, foi necessário transformar o valor de `average\_rating` em numérica, no momento de realizar a transformação utilizei a opção `integer`, transformando a variável em inteira, o que fez com que a distribuição da variável apresenta-se apenas dois picos:

Uma imagem contendo Diagrama

Descrição gerada automaticamente

Transformando a variável novamente, mas agora como `double`, ela apresentou uma distribuição mais normal:

Gráfico, Histograma

Descrição gerada automaticamente

Questionei a Marcus se devido a dificuldade em criar as categorias, devido a distribuição da variável e a proximidade dos quartis (abaixo), se deveríamos estar realizando um problema de regressão, ao invés de classificação.

Texto

Descrição gerada automaticamente

Ele disse que devido a proximidade e concentração dos valores, um modelo de regressão pode ser muito sensível e errar muito, as vezes ele vai predizer um valor com um erro de 0,5 , mas isso é suficiente para o modelo errar muito com valores tão próximos, que nesse caso o que ele sugere e é o que ele faria, é trabalhar com um modelo de classificação mesmo, e sugeriu adotarmos categorias subjetivas ao invés de olharmos os quantis.

Dessa forma adotamos o seguinte intervalo:

Ruim: [0,3.5]

Bom: (3.5,4.5]

Ótimo: (4.5,5]

No entanto, verificando melhor como ficaria essa distribuição, os dados serão fatiados assim:

Gráfico, Histograma

Descrição gerada automaticamente

O que geraria uma concentração muito grande de dados na categoria BOM.

Verificando novamente pelos quantis, se trabalhássemos com um quantil de 33%, teríamos uma divisão mais moderada das notas:

Texto

Descrição gerada automaticamente

Texto

Descrição gerada automaticamente

Gráfico, Histograma

Descrição gerada automaticamente

A grande questão é, faz sentido trabalhar com essas notas para ruim e ótimo?

O professor disse que como temos muitos > 11.000 não tem problemas ficarmos com o conjunto de dados desbalanceado, que inicialmente era a preocupação dele com uma categoria mais subjetiva, mas acredito que precisamos definir outros limites, pois acima de 4.5 há muito poucas observações:

Texto

Descrição gerada automaticamente

* Multicolinearidade variável text\_reviews\_count e rating\_count

Professor disse que é bom remover mesmo, mas pensei que podemos fazer uma proporção de resenhas para não perdermos essa informação.

* Apresentação
* Seguir a ficha de avalição:
  + “O que estou apresentando consegue atender bem os tópicos solicitados?”
* Amarrar as pontas: “procurar defeitos” 🡪 cobrir tudo que a gente fizer
* Não se preocupar com uma modelagem perfeita
  + Mas se perguntar e apresentar: “o que poderia ser feito para melhorar”;
* Relatório
  + Introdução
  + Exploratória
  + Modelagem
  + Conclusão

Não precisa de grandes formalidades, mas precisa estar bem escrito.