Teste Data Science Cognitivo.Ai

Gledson L. Picharski

21/09/2019

Sumário

[1 Introdução 2](#_Toc19977682)

[2 Verificação inicial de variáveis 2](#_Toc19977683)

[2.1 Análise Exploratória Univariada 3](#_Toc19977684)

[2.2 Análise Bivariada 4](#_Toc19977685)

[3 Modelagem 7](#_Toc19977686)

[4 Respostas 9](#_Toc19977687)

.

# Introdução

O objetivo deste trabalho é demonstrar habilidades técnicas para resolver problemas de Data Science, neste sentido atenderei aos itens propostos (explorar os dados e desenvolver o modelo). Neste trabalho incluo meus comentários para esclarecer sobre a linha de raciocínio que segui.

# Verificação inicial de variáveis

Num primeiro momento verifico as características gerais dos dados, o intuito aqui é entender se os dados foram carregados da forma adequada, se valores numéricos estão representados como numéricos e strings como strings, se não aparece algo estranho logo de início.

## 'data.frame': 6497 obs. of 13 variables:  
## $ type : chr "White" "White" "White" "White" ...  
## $ fixed.acidity : num 7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...  
## $ volatile.acidity : num 0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...  
## $ citric.acid : num 0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...  
## $ residual.sugar : num 20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...  
## $ chlorides : num 0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...  
## $ free.sulfur.dioxide : num 45 14 30 47 47 30 30 45 14 28 ...  
## $ total.sulfur.dioxide: num 170 132 97 186 186 97 136 170 132 129 ...  
## $ density : num 1.001 0.994 0.995 0.996 0.996 ...  
## $ pH : num 3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...  
## $ sulphates : num 0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...  
## $ alcohol : chr "8.8" "9.5" "10.1" "9.9" ...  
## $ quality : int 6 6 6 6 6 6 6 6 6 6 ...

Vemos que a variável “alcohol” está como character quando deveria ser numérica, isso ocorre geralmente quando existem caracteres indevidos no meio dos números, ou mesmo quando alguma formatação numérica é empregada em uma planilha antes de salvar o arquivo como csv, neste caso no R, uma forma de verificar quais dados estão diferentes de números é forçar a transformação para número e verificar o que fica como “NA”.

Usando o objeto linhasNA podemos obter as linhas com valores com caracteres que não permitiram a conversão em valor numérico, ou mesmo os que permitiram esta conversão da variável “alcohol”. Pegando os valores convertidos para numéricos podemos fazer um resumo destes dados.

|  |  |
| --- | --- |
| Min. | 8 |
| 1st Qu. | 9.5 |
| Median | 10.3 |
| Mean | 10.49 |
| 3rd Qu. | 11.3 |
| Max. | 14.9 |

*Tabela 1 -* Resumo da variável alcohol apenas para os valores passíveis de conversão em numérico

Com o resumo observamos que o valor mínimo é de 8 e o máximo é de 14.9. Ao observar a seguir os valores que não foram convertidos em numéricos uma suspeita é de que ocorreu alguma falha de formatação numérica em alguma planilha.

|  |
| --- |
| 128.933.333.333.333 |
| 128.933.333.333.333 |
| 114.666.666.666.667 |
| 100.333.333.333.333 |
| 114.333.333.333.333 |
| 105.333.333.333.333 |

*Tabela 2 -* Primeiros valores da variável alcohol que não puderam inicialmente ser convertidos para numéricos.

Neste caso algumas atitudes poderiam ser tomadas:

* Assumir que houve um erro de formação e forçar que a casa decimal seja colocada nestes números respeitando os valores mínimo e máximo.
* Eliminar estes registros.
* Questionar o cliente sobre estes valores.

Qualquer que seja a escolha, é importante apresentar a escolha e confirmar com o cliente se é a escolha mais adequada. Neste caso, vou seguir com a primeira opção.

## Análise Exploratória Univariada

Inicialmente verificamos as características gerais das variáveis do estudo individualmente, começamos então pelas variáveis numéricas. A seguir são apresentadas as medidas de resumo de cada variável numérica, a quantidade (N) é igual em todos por não termos valores faltantes em nenhuma das variáveis, o das variáveis são positivos, o que faz sentido no caso destes testes fisico-químicos. O desvio-padrão (DP) indica a variabilidade dos dados, a variação vai sempre ocorrer de acordo com cada tipo de variável.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | N | Min | 1º Quartil | Mediana | 3º Quartil | Max | Média | DP |
| fixed.acidity | 6497 | 3.8 | 6.4 | 7 | 7.7 | 15.9 | 7.22 | 1.3 |
| volatile.acidity | 6497 | 0.08 | 0.23 | 0.29 | 0.4 | 1.58 | 0.34 | 0.16 |
| citric.acid | 6497 | 0 | 0.25 | 0.31 | 0.39 | 1.66 | 0.32 | 0.15 |
| residual.sugar | 6497 | 0.6 | 1.8 | 3 | 8.1 | 65.8 | 5.44 | 4.76 |
| chlorides | 6497 | 0.009 | 0.038 | 0.047 | 0.065 | 0.611 | 0.06 | 0.04 |
| free.sulfur.dioxide | 6497 | 1 | 17 | 29 | 41 | 289 | 30.53 | 17.75 |
| total.sulfur.dioxide | 6497 | 6 | 77 | 118 | 156 | 440 | 115.7 | 56.52 |
| density | 6497 | 0.9871 | 0.9923 | 0.9949 | 0.997 | 103.9 | 1.71 | 7.64 |
| pH | 6497 | 2.72 | 3.11 | 3.21 | 3.32 | 4.01 | 3.22 | 0.16 |
| sulphates | 6497 | 0.22 | 0.43 | 0.51 | 0.6 | 2 | 0.53 | 0.15 |
| alcohol | 6497 | 8 | 9.5 | 10.3 | 11.3 | 14.9 | 10.49 | 1.19 |
| quality | 6497 | 3 | 5 | 6 | 6 | 9 | 5.82 | 0.87 |

*Tabela 3:* Resumo das variáveis numéricas.

A seguir temos as quantidades e percentuais da variável Type.

|  |  |
| --- | --- |
|  | **Quantidade (%)** |
| Type | \ |
| Red | 1599 (24.6) |
| White | 4898 (75.4) |

*Tabela 4:* Quantidades e percentuais da variável categórica Type.

## Análise Bivariada

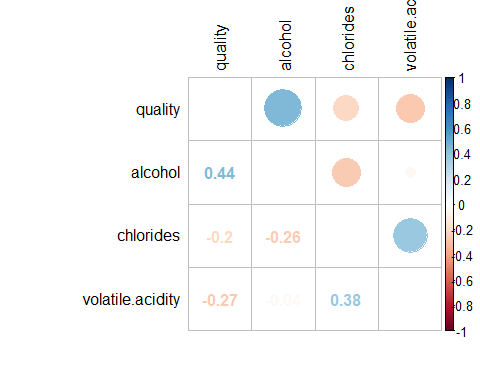
Sabendo que o intuito é estimar a qualidade do vinho, temos a variável “quality” como variável resposta do modelo, assim devemos entender a relação das variáveis com a quality. Cada tipo de relação pode ser representada por um conjunto diferente de formas.

A seguir temos a tabela de correlação entre quality e as variáveis numéricas.

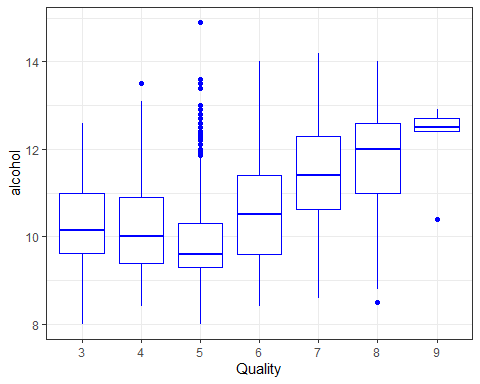
|  |  |
| --- | --- |
| quality | 1 |
| alcohol | 0.4443 |
| citric.acid | 0.08553 |
| free.sulfur.dioxide | 0.05546 |
| sulphates | 0.03849 |
| pH | 0.01951 |
| density | -0.0295 |
| residual.sugar | -0.03698 |
| total.sulfur.dioxide | -0.04139 |
| fixed.acidity | -0.07674 |
| chlorides | -0.2007 |
| volatile.acidity | -0.2657 |

*Tabela 5 -* correlação entre qualidade e variáveis numéricas.

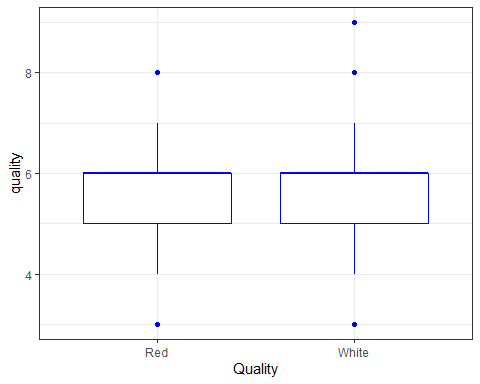
Para representar visualmente estas correlações, temos a seguir as correlações acima de 0,10.



A seguir vemos o boxplot de alcohol para cada nota de qualidade, nota-se que em geral quanto maior os valores de alcohol maior é também a qualidade, de forma que parece existir alguma relação.



A seguir temos o boxplot da qualidade pelo tipo de vinho, aparentemente os dois tipos possuem qualidade similar.



# Modelagem

Existem diversos métodos de machine learning que poderiam ser empregados neste caso, entre eles xgboosting, gradient boosting, regressão, random forest, etc. Antes do desenvolvimento é possível fazer transformações nos dados, seja para mudar a distribuição, pela percepção conceitual da criação de uma nova variável ou mesmo para corrigir algo.

Em cada uma das técnicas é possível investigar uma série de parâmetros em busca do melhor conjunto de parâmetros para cada modelo, com isso compara-se posteriormente os resultados de cada modelo já com os melhores parâmetros possíveis.

É possível também fazer investigações iniciais e variar entre alterar variáveis, parâmetros e técnicas de machine learning, não existe um caminho único, pois as combinações são virutalmente infinitas de tudo que poderia ser feito em busca do melhor modelo.

Optei por fazer uma busca numa grade de parãmetros e apresentar o resultado de cada modelo já com os melhores parâmetros (o código que gera os modelo está configorado para não rodar por conta do tempo de processamento e os modelos resultantes já foram salvos em um objeto do R). As funções de custo visam representar a qualidade dos modelos, em alguns aspectos elas são subjetivas, pois cada contexto trará seu limite prático para estas métricas, o que se pode dizer é que medidas de erro quanto menor melhor (RMSE e MAE) e o R² quanto maior melhor. Como linha geral deseja-se um R² acima de 0,6.

Assim temos na tebela a seguir os resultados de RMSE, R² e MAE para cada um dos modelos no conjunto de treino. Nota-se que o R² está abaixo de 0,5 para maioria das técnicas escolhidas, temos o random forest e o XGboosting com os melhores valores de R².

|  |  |  |  |
| --- | --- | --- | --- |
|  | RMSE | Rsquared | MAE |
| xgb | 0.3627 | 0.8253 | 0.1853 |
| ctree | 0.7108 | 0.3281 | 0.5549 |
| rpart2 | 0.7518 | 0.2481 | 0.5906 |
| rpart | 0.7719 | 0.2072 | 0.6131 |
| rf | 0.5316 | 0.7108 | 0.3918 |

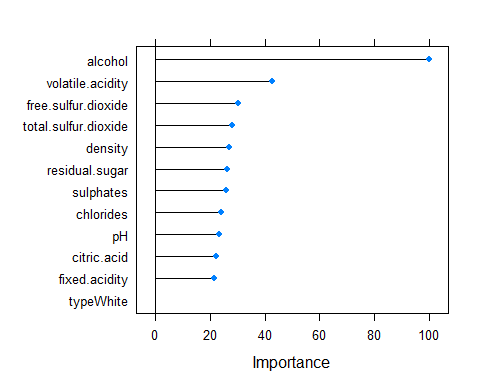
*Tabela 6 -* Resultado dos modelos desenvolvidos para o conjunto de dados de treino.

Ao avaliar as métricas de qualidade no conjunto de teste temos em geral medidas muito parecidas com as do treino para os modelos ctree, rpart e rpart2, já para o xgb e rf temos uma redução expressiva nos valores. Ainda assim o xgb se apresenta como o melhor modeo dentre os avaliados, sendo asism o modelo campeão.

|  |  |  |  |
| --- | --- | --- | --- |
|  | RMSE | Rsquared | MAE |
| xgb | 0.3292 | 0.8638 | 0.1702 |
| ctree | 0.7198 | 0.3428 | 0.5565 |
| rpart2 | 0.7605 | 0.2663 | 0.5936 |
| rpart | 0.784 | 0.2204 | 0.6163 |
| rf | 0.5457 | 0.7098 | 0.4 |

*Tabela 7 -* Resultado dos modelos desenvolvidos para o conjunto de dados de teste.

A seguir podemos var então a importância de cada variável do modelo escolhido.



.

# Respostas

**a. Como foi a definição da sua estratégia de modelagem?**

Inicialmente é importante explorar as variáveis tecnicamente, verificando se os dados foram carregados adequadamente e se estão no formato devido, na sequência vale entender as medidas descritivas de cada avariável para entender como cada uma delas se comporta. Uma análise bivariada ajuda a direcionar quais variáveis possuem mais relevância e por fim na parte de modelagem pode-se fazer a transformação de variáveis, engenharia de variáveis e finalmente a escolha do melhor modelo.

**b. Como foi definida a função de custo utilizada?**

Considerando a qualidade como valor numérico poderíamos definir a função de custo pelo erro da predição em relação ao observado ou pela correlação do predito pelo observado, as duas medidas mais comuns são R² e o RMSE (Root Mean Squart Error - Raiz do Erro quadrático médio). O RMSE é interessante para comparar um modelo em relação ao outro, mas é mais complicado de pensar nele por não ter uma escala específica. O R² por sua vez varia entre 0 e 1, de forma que valores mais próximos de 1 indicam maior qualidade do modelo, alguns podem colocar o ponto de corte de um modelo aceitável em 0,5, 0,6 ou mesmo em 0,7, mas na prática eu optaria pelo melhor modelo que me for possível.

**c. Qual foi o critério utilizado na seleção do modelo final?**

O modelo final é escolhido pela melhor medida de qualidade, neste caso pelo maior R².

**d. Qual foi o critério utilizado para validação do modelo? Por que escolheu utilizar este método?**

Uma forma de validar o modelo, para evitar coisas como overfiting, é verificar medida de qualidade num conjunto que não foi utilizado na modelagem, por isso separa-se em conjunto de treino e de validação.

**e. Quais evidências você possui de que seu modelo é suficientemente bom?**

Utilizando o RMSE, quanto menor o erro melhor o modelo, em alguns aspectos a medida pode ser um pouco subjetiva, então enquanto RMSE estiver elevado (algo acima de 0,20) deveríamos buscar alternativas para melhorar o modelo, seja pela utilização de outras técnicas ou mesmo pela engenharia de variáveis.