EST0133 - Projeto II

Marcus Nunes

12 de Janeiro de 2022

Instruções

- A data limite de entrega é 24/01/2022, às 23:59, via SIGAA
- O R é o único software permitido para coleta, limpeza e análise dos dados
- O trabalho deve ser feito em R Markdown, utilizando o arquivo `modelo.Rmd`
- Renomeie-o para `NomeSobrenome.Rmd`, em que `Nome` é o seu primeiro nome e `Sobrenome` é um de seus sobrenomes
- Envie conjuntamente seus arquivos `NomeSobrenome.Rmd` e `NomeSobrenome.pdf` para avaliação em um arquivo chamado `NomeSobrenome.zip`
- Respostas em arquivos que não estejam nos formatos Rmd e pdf não serão consideradas
- Identifique corretamente os eixos dos gráficos produzidos
- Respostas numeradas incorretamente não serão corrigidas
- Respostas com códigos ou outputs supérfluos para a sua resolução, como pacotes desnecessários para as análises realizadas, terão pontuação descontada
- Não é permitido reportar resultados como capturas de tela
- Nos exercícios em que for necessário ajustar uma semente aleatória, rode o código `set.seed(x, kind = "Mersenne-Twister", normal.kind = "Inversion")`, em que `x` é o número da semente pedida

Parte I - Classificação

O arquivo `ataques_cardiacos.csv` traz informações a respeito de 299 pacientes que sofreram ataque cardíaco em algum momento de suas vidas. Eles foram acompanhados durante algum tempo e As colunas presentes são

- `idade`: idade do paciente (anos)
- `anemia`: se o paciente está anêmico ou não
- `cpk`: nível da enzima CPK no sangue (µg/L)
- `diabetes`: se o paciente possui diabetes
- `fracao_ejecao`: percentual de sangue saindo do coração a cada batida
- `pressao_alta`: se o paciente é hipertenso
- `plaquetas`: quantidade de plaquetas no sangue (em milhares/mL)
- `creatinina_sangue`: nível de creatinina no sangue (em mg/dL)
- `sodio`: nível de sódio no sangue (em mEq/L)
- `genero`: gênero do paciente
- `fumante`: se o paciente é fumante
- `morte`: evento de morte do paciente, isto é, se ele faleceu durante o acompanhamento médico

Queremos criar um modelo preditivo para o evento de morte do paciente, baseando-nos nas outras variáveis do conjunto de dados.

Questão 1

(*05 pontos*) O primeiro passo será preparar o conjunto de dados para análise. Para isso, crie um objeto chamado `coracao` com o conteúdo do arquivo `ataques_cardiacos.csv`. Transforme a coluna `morte` de modo que `sim` seja o nível de referência.

Questão 2

(*05 pontos*) Utilize a semente 1201 para criar os conjuntos de treino e teste. O conjunto de treino deve ser criado com 78% das observações.

Questão 3

(*05 pontos*) Crie gráficos de dispersão em duas dimensões entre todas as variáveis quantitativas do conjunto de dados de treino. Informe também o valor da correlação de Spearman entre estas variáveis. Existe alguma suspeita de multicolinearidade entre estas variáveis? Justifique.

Questão 4

(*05 pontos*) Crie boxplots comparando os valores das variáveis preditoras quantitativas entre os níveis de `morte`. Alguma (ou mais de uma) variável quantitativa poderia ser considerada como uma boa preditora para discriminar entre os níveis de `morte`? Qual (ou quais) e por quê?

Questão 5

(*05 pontos*) Pré-processe os dados com apenas 3 transformações:

i) Balanceie o número de observações para cada classe da variável resposta;

ii) Deixe a média das variáveis preditoras igual a zero;

iii) Faça com que a variância das variáveis preditoras seja igual a um.

Não é necessário realizar nenhum outro tipo de pré-processamento para essa análise. Aplique as transformações nos conjuntos de treino e teste.

Questão 6

(*05 pontos*) Defina a validação cruzada com 6 grupos para avaliar o desempenho dos algoritmos que aplicaremos a esses dados. Utilize a semente 2022 para isso.

Questão 7

(*05 pontos*) Crie grids de procura para os hiperparâmetros dos métodos CART e Random Forest. Encontre o melhor valor de `cost_complexity` para o CART entre os valores 10^{-5} e 10^{-1}, `tree_depth` entre 1 e 5 e `min_n` entre 10 e 100. Utilize 5, 5 e 10 valores diferentes, respectivamente, para cada um destes hiperparâmetros (ou seja, ajuste 250 modelos diferentes). Para o random forest, encontre o melhor valor de `mtry` 1 e o máximo permitido, `trees` entre 500 e 1000 e `min_n` entre 10 e 100. Utilize 4, 2 e 10 valores diferentes, respectivamente, para cada um destes hiperparâmetros (ou seja, ajuste 80 modelos diferentes).

Questão 8

(*05 pontos*) Rode o ajuste dos modelos definidos anteriormente. A seguir, utilize os meios necessários para determinar se a acurácia e a área sob a curva dos ajustes com os algoritmos utilizados foram maximizadas em algum momento.

Questão 9

(*05 pontos*) Qual é a sua opção de algoritmo para modelar estes dados? Justifique a sua escolha.

Questão 10

(*05 pontos*) Considerando métricas adequadas aplicadas nos conjuntos de treino e teste, o resultado obtido com a modelagem definitiva é bom o suficiente, na sua opinião? Cite alguma sugestão a ser aplicada nos dados ou na modelagem, que talvez pudesse melhorar o resultado obtido. Não é necessário implementar a sugestão, apenas comentá-la e justificá-la.

Parte II - Regressão

O twitch é um serviço de *streaming* de vídeos ao vivo. É bastante identificado com a comunidade de *esports*, embora possua canais especializados em diversas outras áreas de entretenimento. O arquivo `twitch.csv` possui informações sobre os 1000 canais mais populares em 2020, a saber:

- `channel`: nome do canal
- `watch_time_minutes`: somatório da quantidade total de minutos que o canal foi assistindo, considerando todos os usuários da plataforma
- `stream_time_minutes`: quantidade de minutos que o canal ficou ao vivo durante o ano
- `peak_viewers`: número máximo de espectadores simultâneos do canal
- `average_viewers`: quantidade média de espectadores simultâneos do canal
- `followers`: quantidade de seguidores do canal no final do ano
- `followers_gained`: diferença entre a quantidade de seguidores do canal no final e no começo do ano
- `views_gained`: visualizações ganhas pelo canal durante o ano
- `mature`: variável indicando se o conteúdo do canal é para adultos
- `language`: idioma principal do canal

O objetivo desta tarefa é modelar a variável `followers_gained`, a fim de explicar que fatores são capazes de determinar o número de seguidores que um canal pode arregimentar em um ano.

Questão 11

(*05 pontos*) Importe para o R o conjunto de dados do problema. Retire a coluna com o nome do canal e recodifique a coluna `language`, mantendo apenas o nível `English` original e juntando todas as demais em `Other`.

Questão 12

(*05 pontos*) Utilize a semente 2109 para criar os conjuntos de treino e teste. O conjunto de treino deve ser criado com 70% das observações.

Questão 13

Crie gráficos de dispersão em duas dimensões entre todas as variáveis quantitativas do conjunto de dados de treino. Informe também o valor da correlação linear entre estas variáveis. Alguma correlação entre as variáveis preditoras e a variável resposta se destaca? Existem indícios de multicolinearidade? Justifique.

Questão 14

(*05 pontos*) Pré-processe os dados com apenas 4 transformações:

i) Transforme as variáveis quantitativas (exceto a resposta) utilizando logaritmo;

ii) Crie versões *dummy* das variáveis qualitativas usando a função `step_dummy`

iii) Deixe a média das variáveis preditoras igual a zero;

iv) Faça com que a variância das variáveis preditoras seja igual a um.

Não é necessário realizar nenhum outro tipo de pré-processamento para essa análise. Aplique as transformações nos conjuntos de treino e teste.

Questão 15

(*05 pontos*) Defina a validação cruzada com 5 grupos para avaliar o desempenho dos algoritmos que aplicaremos a esses dados. Utilize a semente 2220 para isso.

Questão 16

Utilize funções do pacote `tidymodels` para ajustar um modelo de regressão linear múltipla aos dados que estamos analisando. Não é preciso realizar o tunning deste modelo.

Questão 17

(*05 pontos*) Utilize o random forest para ajustar um modelo a estes dados. Encontre o melhor valor de `mtry` 1 e o máximo permitido, `trees` entre 500 e 1000 e `min_n` entre 10 e 50. Utilize todos os valores possíveis, 2 e 5 valores diferentes, respectivamente, para cada um destes hiperparâmetros.

Questão 18

(*05 pontos*) Compare os resultados obtidos (no conjunto de treino) entre a regressão linear e o modelo final obtido com random forest utilizando a raiz do erro quadrático médio como critério. Qual é a sua opção de modelagem para estes dados e por quê?

Questão 19

(*05 pontos*) Segundo o random forest, qual é a variável mais importante para o modelo ajustado? Intuitivamente, esse resultado faz sentido? Justifique.

Questão 20

(*05 pontos*) Considerando o conjunto de teste, o resultado obtido com a melhor modelagem é bom o suficiente? Utilize argumentos numéricos e gráficos para justificar a sua resposta.