Trabalho individual de Verificação suplementar

MBA Big Data e Business Analytics – Análise Preditiva (BSB T10)

**Nome\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Ao entregar este trabalho final no e-class Você estará declarando:**

**"Estou ciente das razões pelas quais estamos fazendo este trabalho de forma remota. Assim, declaro que utilizei somente formas éticas de resolver as questões e não consultei ninguém, de nenhuma forma, para respondê-las ”**

Escrevi para que Você leia, por favor!!!

**Instruções**:

* Este trabalho tem **4 páginas**, incluindo esta.
* Consulta a notas de aula, apontamentos, internet, scripts etc. é permitida.
* **A consulta a colegas ou outras pessoas, qualquer que seja o meio, é proibida**.
* Responda as perguntas digitando as respostas **neste** arquivo.
* Se possível, para facilitar a correção, coloque as respostas em azul ou verde escuro.
* Ao terminar, **não esqueça de colocar seu nome**. Verifique cuidadosamente se não esqueceu nenhuma resposta **nos espaços indicados**.
* **Salve como .pdf** e faça upload no ECLASS na caixa de **entrega de atividades**, na pasta correspondente, dentro do prazo abaixo. Não serão aceitos trabalhos em outros formatos.
* Por favor, não cole os scripts do R nem os outputs do R a menos que explicitamente solicitados.
* Se tiver dificuldades ou dúvidas com o ECLASS, por favor, envie-me um mail [alsicsu@gmail.com](mailto:alsicsu@gmail.com).
* Dúvidas quanto à interpretação das questões não serão respondidas. A interpretação faz parte da prova. Se Você achar que uma questão é impossível, diga o porquê (não recomendo pois já fiz o gabarito)
* Veja que a data de entrega é  **02/09/2021 até as 23h47min** . **Repetindo**: faça upload do arquivo **em .pdf** no ECLASS na caixa de **entrega de atividades**, na pasta correspondente. Por favor, não envie por e-mail ou WhatsApp (não receberei).
* Bom trabalho

**Questão 1:** Considere a planilha ***questoes1e2***  *da*  base de dados **BSB T10 AP P1.xlsx**

A variável alvo é STATUS, as demais são variáveis previsoras (exceto a variável FUNCIONARIO)

* Note que Vc não deve:
  + Transformar (log, sqrt, ...) ou discretizar nenhuma variável. Utilize-as como estão.
  + Dividir a amostra em duas partes ( treinamento e teste). Utilize a mesma amostra para desenvolvimento e teste.
  + Eliminar observações mesmo que lhe pareçam discrepantes

**Parte I:** Sem dividir a amostra em duas partes, obtenha a árvore de classificação com o software R**,** função rpart**, sem podar! Utilize set.seed(123)**

A1) Qual a P(bom) estimada pelo modelo para Indivíduos com o perfil especificado no quadro seguinte?

Resp: \_\_\_\_\_\_\_\_\_\_\_\_ *(3 casas decimais)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| IDADE | ECIV | DIST\_EMP | TIPORESID | PRIM\_EMP | TESTE |
| 25a45 | solt | média | própria | não | 82 |

A2) Considere o ponto de corte o valor 0,40, ou seja de P(bom) > 0,40 classifique como **bom**. Qual a proporção de verdadeiros **bom** que seriam classificados como **mau** utilizando a árvore obtida no **item a?**Resp:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

A3) Calcule a área AUROC para a árvore obtida no **item a**.

Resp: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (3 casas decimais)

A4) Obtenha a árvore (sem podar) fixando 200 como número mínimo de indivíduos em um nó terminal. Qual a probabilidade estimada de que o funcionário A1 seja **bom**? **Utilize set.seed(123)**

Resp: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (3 casas decimais)

**Questão 2:** Considere a planilha ***questoes1e2***  *da*  base de dados **BSB T10 AP P1.xlsx**

* Note que Vc não deve:
  + Transformar (log, sqrt, ...) ou discretizar nenhuma variável. Utilize-as como estão.
  + Dividir a amostra em duas partes ( treinamento e teste). Utilize a mesma amostra para desenvolvimento e teste.
  + Eliminar observações mesmo que lhe pareçam discrepantes

Obtenha a regressão logística para prever STATUS utilizando todas as variáveis previsoras (exceto a variável FUNCIONARIO). **Não** **selecione variáveis:**  utilize todas as previsoras.

Com este modelo responda as questões seguintes:

B1) Qual a probabilidade estimada de que o indivíduo A1 seja “**mau**” . Resp: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (3 casas decimais) (cuidado: veja qual o evento resposta que Você adotou)

B2) Calcule a probabilidade estimada de que o indivíduo seguinte seja “**mau**”

Resp: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (3 casas decimais)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| IDADE | ECIV | DIST\_EMP | TIPORESID | PRIM\_EMP | TESTE |
| 25a45 | solt | média | própria | sim | 76 |

B3) Se a probabilidade estimada de ser **mau** for superior a 0.50, um funcionário será classificado como “**mau**”. Caso contrário será classificado como “**bom**”. Qual a taxa de erro de classificação estimada com essa amostra?

Resp: \_\_\_\_\_\_\_\_\_\_\_\_\_\_ (3 casas decimais)

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

Rode o algoritmo randomForest para estimar a probabilidade de ser **mau** de um indivíduo, considerando B=400 (número de árvores) e set.seed(789) e 4 variáveis para cada partição. Se Vc achar que necessita outros hiper parâmetros adote-os e especifique-os a seguir:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

B4) Qual a probabilidade estimada de que o indivíduo A1 seja “**mau**” ? Resp:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (3 casas decimais)

B5) Qual o valo da área AUROC para esse modelo?   
Resp:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (3 casas decimais)

**Questão 3** Considere a planilha ***questao3***  *da*  base de dados ***BSB T10 AP P1.xlsx***

* Note que Vc não deve:
  + Transformar (log, sqrt, ...) ou discretizar nenhuma variável. Utilize-as como estão.
  + Dividir a amostra em duas partes ( treinamento e teste). Utilize a mesma amostra para desenvolvimento e teste.
  + Eliminar observações mesmo que lhe pareçam discrepantes

C1) determine um modelo de regressão linear múltipla, **sem eliminar variáveis**, para prever a variável EXPN a partir das demais variáveis. Qual o coeficiente de HIGH?

Resp: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (3 casas decimais)

C2) calcule o erro percentual para o **segundo indivíduo** da planilha fornecida para a prova. Note que o valor observado de EXPN desse indivíduo é igual a 772,77  
Resp: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ . (em porcentagem, 3 casas decimais).

C3) Algum ponto da amostra é um outlier em Y? (sim /não)\_\_\_\_\_\_\_\_\_. Em caso positivo que ponto(s) é (são) esse(s) ? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

C4) Selecione as variáveis utilizando o comando step. Alguma variável foi removida? Qual ou quais? Resp: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_