ESCOLA DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE MATEMÁTICA

Primeiro Trabalho de Análise Categórica de Dados (2017/18)
Mestrado em Modelação Estatística e Análise de Dados

Entrega até dia 27 de Abril

1. O ficheiro ***Bankloan.xlsx*** apresenta oito variáveis referentes a 500 pessoas que são clientes de um banco.
   a) Categorize a variável idade em duas categorias, correspondendo a segunda categoria aos clientes com 35 ou mais anos. Existe associação entre a variável idade categorizada e o estar ou não em incumprimento com o banco? Calcule e interprete o valor do risco relativo e do rácio das chances (*Odds ratio*).
   b) Recorrendo a uma eventual recodificação e/ou categorização da variável Nível de educação estratifique a amostra por Nível de educação categorizado. Aplique o teste de Mantel-Haenszel e o teste de Breslow-Day e retire as conclusões que entender convenientes.
   c) Calcule a correlação entre a variável idade categorizada e o Endividamento (% rendimento).

   Através de uma regressão logística, responda às seguintes questões:

   d) Quais as variáveis significativas para se elaborar uma boa previsão de risco de incumprimento? Interprete os coeficientes do modelo referentes às covariáveis (estimativas pontuais acompanhadas de estimativas intervalares a 95%).
   e) Ajuste um modelo, sem as variáveis que apresentam problemas de significância.
   f) Interprete os *outputs* da técnica.
   g) Avalie a bondade do ajustamento do modelo obtido e elabore o diagnóstico dos resíduos que lhe permita investigar observações influentes e/ou outliers determine ainda a distância de Cook.
   h) Elabore uma curva de ROC e interprete-a.
   i) Caso pretenda usar o modelo com objetivos de classificação, como avalia a sua capacidade discriminativa? Encontre um valor de corte que julgue adequado neste contexto.
   j) Calcule a probabilidade de incumprimento de um indivíduo com as seguintes características:
      - Idade = 40 anos
      - Nível de educação = 3
      - Emprego atual = 3 anos
      - Endereço atual = 5 anos
      - Rendimento familiar anual (em milhares) = 60
      - Endividamento = 17%
      - Dívida do cartão de crédito (em milhares) = 70
      - Outras dívidas (em milhares) = 3.

2. Suponha que um banco que financia um *stand* de automóveis está interessado em investir numa campanha de *marketing* direto e, para tal, precisa de identificar na sua base de dados o perfil dos clientes que:
   1) não desejam trocar de carro (y = 0);
   2) desejam trocar de carro, mas pagariam a pronto (y = 1);
   3) desejam trocar de carro, mas financiariam o pagamento (y =2).
   O ficheiro ***Multinomial. xlsx*** contém os dados de uma amostra de 89 observações. As variáveis explicativas disponíveis na base de dados são:
      - *dif_ano*: corresponde à diferença entre o ano base e o ano do veículo;
      - *sexo*: sendo 0 para indicar o sexo feminino e 1 para indicar o sexo masculino;
      - *classesocial*: classes A, B e C.

   a) Obtenha um modelo de regressão multinomial que explique o perfil dos clientes (y) em função das variáveis *dif_ano*, *sexo* e *classesocial*. Retire as conclusões que entender convenientes.
   b) Refaça a análise de regressão multinomial, excluindo a constante. Analise os resultados.
   c) Refaça a análise de regressão logística, selecionando somente os registos cujos valores de Y são 0 ou 2. Compare com os resultados obtidos.