**Applied Statistics – Turma 9DTS**

Atividade prática aula 09

**Habilidades desenvolvidas: Modelos preditivos usando modelos de regressão logística. Cluster Analysis.**

**Parte 1**

Utilize o texto 1 para completar os exercícios a seguir.

1. Use o resultado da Regressão Logística Múltipla apresentado na Tabela 3 da página 112 do artigo para calcular o escore das pessoas a seguir:

|  |  |  |
| --- | --- | --- |
| ID | Características | Escore de crédito |
| 1 | Mulher, casada, primeira faixa de idade, primeira faixa de tempo de emprego, primeira faixa de valor da parcela, tipo de crédito carnê, primeira categoria de cep, primeira categoria de profissão, primeira faixa de empréstimo e primeira aquisição de empréstimo. | Y = 0+0-0.4848-0.2035-1.34.7429-0.3549+0.3033+0.2491-0.6513+0.5868  Y = EXP(-2,6492)/1+EXP(-2,6492))  Y = 0.0622 |
| 2 | Homem, casado, primeira faixa de tempo de emprego, primeira faixa de valor da parcela, tipo de crédito carnê, primeira faixa de idade, primeira categoria de cep, primeira categoria de profissão, primeira faixa de empréstimo e primeira aquisição de empréstimo. | Y = 0.5868 -0.314 -0.4848 -0.2035 -1.34 -0.7429 -0.3549 + 0.3033 + 0.2481 -0.6513 = -2.9632  Y = EXP(-2.9632) / (1 + EXP(-2.9632))  Y = 0.0491 |
| 3 | Homem, casado, primeira faixa de idade, primeira faixa de tempo de emprego, primeira faixa de valor da parcela, tipo de crédito carnê, primeira categoria de cep, primeira categoria de profissão, primeira faixa de empréstimo e primeira aquisição de empréstimo. | Y = 0.5868 -0.314 -0.1707 -0.4848 -0.2035 -1.34 -0.7429 -0.3549 -0.6513 =  -3.6753  Y = EXP(-3.6753) / (1 + EXP(-3.6753))  Y = 0.0247 |

1. Utilize os dados da página 116 para calcular o acerto do modelo da Regressão Logística na amostra de treinamento. Preencha o quadro a seguir e responda

|  |  |  |  |
| --- | --- | --- | --- |
| Observado | Predito | | Total |
| Mau | Bom |
| Mau | 2913 | 1087 | 4000 |
| Bom | 1184 | 2816 | 4000 |
| Total | 4097 | 3903 | 8000 |

1. Qual o percentual de acerto da resposta Mau pagador?
   * (2913 / 4000) \* 100 = 72,83%
2. Qual o percentual de acerto da resposta Bom pagador?
   * (2816 / 4000) \* 100 = 70,40%
3. Qual o percentual de acerto médio do modelo?
   * ((2913 + 2816) / 8000) \* 100 = 71,64%
4. Utilize os dados da página 116 para calcular o acerto do modelo de Redes Neurais na amostra de treinamento. Preencha o quadro a seguir e responda

|  |  |  |  |
| --- | --- | --- | --- |
| Observado | Predito | | Total |
| Mau | Bom |
| Mau | 3000 | 1000 | 4000 |
| Bom | 1280 | 2720 | 4000 |
| Total | 4280 | 3720 | 8000 |

1. Qual o percentual de acerto da resposta Mau pagador?
   * (3000 / 4000) \* 100 = 75%
2. Qual o percentual de acerto da resposta Bom pagador?
   * (2720 / 4000) \* 100 = 68%
3. Qual o percentual de acerto médio do modelo?
   * ((3000 + 2720) / 8000) \* 100 = 71,5%
4. Qual modelo você escolheria para implementar na empresa?
   1. O modelo de Regressão Logística apresentou um desempenho superior ao de Redes Neurais de forma geral. Ele se mostrou mais eficaz na identificação de bons pagadores, enquanto o modelo de Redes Neurais foi mais eficiente para identificar maus pagadores. Diante disso, recomendaria a adoção do modelo de Regressão Logística, pois, além de ter tido um desempenho um pouco melhor, oferece maior facilidade na interpretação das decisões e tende a ser mais simples de manter ao longo do tempo.

**Parte 2**

1. Utilizando a base descrita e disponibilizada de vinhos desta região de Portugal com as variáveis de características (composição) dos vinhos.

Contexto dos dados:

**Source:**

Paulo Cortez, University of Minho, Guimarães, Portugal, <http://www3.dsi.uminho.pt/pcortez>   
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal @2009

**Data Set Information:**

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult: [[Web Link]](http://www.vinhoverde.pt/en/) or the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).   
  
**Relevant Papers:**

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties.   
In Decision Support Systems, Elsevier, 47(4):547-553, 2009.   
  
Available at: [[Web Link]](http://dx.doi.org/10.1016/j.dss.2009.05.016)

|  |  |  |
| --- | --- | --- |
| **Característica** | **Descrição** | |
| * **fixedacidity**  *(Acidez fixa)* | | a maioria dos ácidos presentes no vinho ou fixos ou não voláteis (não evaporaram prontamente) |
| * **volatileacidity** *(Acidez volátil)* | | a quantidade de ácido acético no vinho, que em níveis muito altos pode levar a um gosto desagradável de vinagre. |
| * **Citricacid** (Ácido cítrico) | | encontrado em pequenas quantidades, o ácido cítrico pode adicionar “frescura” e sabor aos vinhos. |
| * **Residualsugar**   *(Açúcar residual)* | | a quantidade de açúcar restante depois que o processo de fermentação (em contato com a levedura, o açúcar da uva vai se transformando em álcool, ou seja, a fermentação vai ocorrendo) |
| * **Chlorides**  *(*Cloretos) | | a quantidade de sal no vinho. |
| * **freesulfurdioxide**   *(Dióxido de enxofre livre)* | | a forma livre de SO2 existe em equilíbrio entre o SO2 molecular (como gás dissolvido) e o íon bissulfito; impede o crescimento microbiano e a oxidação do vinho. Garante condições melhores para os processos de vinificação da bebida, elimina bactérias e leveduras frágeis e indesejáveis, o que permite que apenas as melhores prossigam com o processo fermentativo. Além do mais, melhora o aroma e afina a cor da bebida |
| * **totalsulfurdioxide**   *(Dióxido de enxofre total)* | | quantidade de formas livres e encadernadas de S02; em baixas concentrações, o SO2 é quase indetectável no vinho, mas nas concentrações de SO2 livre acima de 50 ppm, o SO2 se torna evidente no nariz e no sabor do vinho. |
| * **Density**  (*Densidade)* | | a densidade é próxima à da água, dependendo do percentual de álcool e teor de açúcar. |
| * **pH** (*PH)* | | descreve como o vinho é acídico ou básico numa escala de 0 (muito ácido) a 14 (muito básico); a maioria dos vinhos tem entre 3-4 na escala de pH. |
| * **Sulphates** *(Sulfatos)* | | um aditivo de vinho que pode contribuir para os níveis de gás de dióxido de enxofre (S02), que age como um antimicrobiano e antioxidante. |
| * **alcohol** *(Álcool)* | | o teor alcoólico percentual do vinho. |
| * **quality** *(Qualidade)* | | variável de saída (com base em dados sensoriais) que poderiam ser de 0 a 10 sendo zero muito ruim e 10 muito excelente. |

Exemplo da Base (10 linhas)



**Etapa 1**:

Com objetivo de estimar a nota de qualidade dos vinhos (variável “Quality”) em função das características físico-químicas dos vinhos;

1. Qual é a técnica adequada para essa resolver esse objetivo:

|  |  |
| --- | --- |
| X | Técnicas Supervisionadas |
|  | Técnicas Não Supervisionadas |

1. Há variáveis dependentes (D) e independentes (I)? Caso sim, preencha o quadro a seguir.

|  |  |
| --- | --- |
| Variáveis Dependentes: | Quality |
| Variáveis Independentes: | Todas exceto o ID do Vinho |

1. Qual (is) a tarefa(s) de Mining é (são) adequada (s) para solução desse objetivo:

|  |  |
| --- | --- |
| X | Predição |
|  | Classificação |
|  | Regras de Associação |
|  | Agrupamento |
|  | Sumarização |
|  | Detecção de Anomalias |

1. Você pode indicar alguns algoritmos possíveis para essa tarefa?

|  |  |
| --- | --- |
| X | Regressão Linear |
|  | Árvore de Decisão |
|  | Regras de Associação |
|  | Séries Temporais |
|  | Análise de Clusters |
|  | Componentes Principais |
| X | Redes Neurais |
|  | Random Forest, |
| X | Outras: Árvore de Regressão, SVM |

1. Quais cuidados que são indicados a fazer antes da etapa de aplicar os algoritmos/técnicas?
   1. Encontrar valores numéricos correspondentes das categorias de variáveis qualitativas de texto.
   2. Tratativa de valores faltantes
   3. Dividir amostra entre base de treino e validação.

**Etapa 2:**

Com o objetivo de classificar os vinhos dessa região em vinhos em “Bom” ou “Ruim” em função de suas características físico-químicas;

1. Qual é a técnica adequada para essa resolver esse objetivo:

|  |  |
| --- | --- |
| X | Técnicas Supervisionadas |
|  | Técnicas Não Supervisionadas |

1. Há variáveis dependentes (D) e independentes (I)?

|  |  |
| --- | --- |
| Variáveis Dependentes: | Bom (quality >=8) e Ruim (quality < 8) |
| Variáveis Independentes: | Todas exceto ID |

1. Qual (is) a tarefa(s) de Mining é (são) adequada (s) para solução desse objetivo:

|  |  |
| --- | --- |
|  | Predição |
| X | Classificação |
|  | Regras de Associação |
|  | Agrupamento |
|  | Sumarização |
|  | Detecção de Anomalias |

1. Você pode indicar alguns algoritmos possíveis para essa tarefa?

|  |  |
| --- | --- |
|  | Regressão Linear |
| X | Árvore de Decisão |
|  | Regras de Associação |
|  | Séries Temporais |
|  | Análise de Clusters |
|  | Componentes Principais |
| X | Redes Neurais |
| X | Random Forest, |
|  | Outras: |

1. Quais cuidados que são indicados a fazer antes da etapa de aplicar os algoritmos/técnicas?
   1. De-para das categorias das variáveis qualitativas de texto para número

**Etapa 3:**

Com o objetivo de auxiliar na importação de vinhos dessa região a pergunta que é apresentada é: quantos diferentes tipos de vinhos encontramos essa região.

1. Qual é a técnica adequada para essa resolver esse objetivo:

|  |  |
| --- | --- |
|  | Técnicas Supervisionadas |
| X | Técnicas Não Supervisionadas |

1. Há variáveis dependentes (D) e independentes (I)?

|  |  |
| --- | --- |
| Variáveis Dependentes: | Não há |
| Variáveis Independentes: | Todas exceto ID e Quality |

1. Qual (is) a tarefa(s) de Mining é (são) adequada (s) para solução desse objetivo:

|  |  |
| --- | --- |
|  | Predição |
|  | Classificação |
|  | Regras de Associação |
| X | Agrupamento |
|  | Sumarização |
|  | Detecção de Anomalias |

1. Você pode indicar alguns algoritmos possíveis para essa tarefa?

|  |  |
| --- | --- |
|  | Regressão Linear |
|  | Árvore de Decisão |
|  | Regras de Associação |
|  | Séries Temporais |
| X | Análise de Clusters |
|  | Componentes Principais |
|  | Redes Neurais |
|  | Random Forest |
|  | Outras: |

1. Quais cuidados que são indicados a fazer antes da etapa de aplicar os algoritmos/técnicas?
   1. Normalização das variáveis quantitativas
   2. Escolher qual será a distância utilizada

**Etapa 4:**

O dono dessa importadora que nos pediu para ajudar na importação de vinhos dessa região de Portugal, agora, ele tem um novo objetivo: ele quer fazer uma segmentação comportamental de seus clientes. Como proceder?

1. Qual é a técnica adequada para essa resolver esse objetivo:

|  |  |
| --- | --- |
|  | Técnicas Supervisionadas |
| X | Técnicas Não Supervisionadas |

1. Há variáveis dependentes (D) e independentes (I)?

|  |  |
| --- | --- |
| Variáveis Dependentes: |  |
| Variáveis Independentes: | Variáveis associadas apenas ao cliente |

1. Qual (is) a tarefa(s) de Mining é (são) adequada (s) para solução desse objetivo:

|  |  |
| --- | --- |
|  | Predição |
|  | Classificação |
|  | Regras de Associação |
| X | Agrupamento |
|  | Sumarização |
|  | Detecção de Anomalias |

1. Você pode indicar alguns algoritmos possíveis para essa tarefa?

|  |  |
| --- | --- |
|  | Regressão Linear |
|  | Árvore de Decisão |
|  | Regras de Associação |
|  | Séries Temporais |
| X | Análise de Clusters |
| X | Componentes Principais |
| X | Redes Neurais |
|  | Random Forest |
|  | Outras: |

1. Quais cuidados que são indicados a fazer antes da etapa de aplicar os algoritmos/técnicas?
   1. Tratamento de valores faltantes
   2. Transformar variáveis qualitativas em quantitativas por valores correspondentes
   3. Normalização ou padronização dos dados numéricos
   4. Seleção de variáveis relevantes
   5. Definir métricas de avaliação para clusterização
   6. Definir o número ótimo de clusters

Entrega do exercício no formato word.

Data de entrega: 13/09/2024

Regina Bernal

02/09/2024