**Statistic with R – Turma 7DTS**

Projeto Integrado – final

**A QuantumFinance está acompanhando um crescimento de inadimplência entre seus clientes e solicita a consultoria para desenvolver uma análise de inadimplência.**

**Os resultados apresentados pela consultoria mostram que o modelo de *credit scoring* implementado utilizou 70 variáveis e apresentou uma acurácia de 26% (R2 ajustado = 0.2593609). Consequentemente, a QuantumFinance está oferecendo crédito para mau pagador e deixando de oferecer para o bom pagador.**

**Para que a QuantumFinance tome decisões mais precisas sobre concessões de crédito, ela precisa aprimorar seu modelo de concessão de crédito.**

**Desafio: Melhorar a acurácia do modelo preditivo mediante uso do valor target disponível na base de dados train.csv.**

**3ª etapa: Conhecer os dados**

A base de dados contém uma coluna chamada id que identifica exclusivamente cada linha, várias colunas identificadas por strings hexadecimais e um destino de coluna que gostaríamos que você previsse. As colunas que contêm hashes SHA-256 para seus valores representam variáveis ​​categóricas, enquanto o restante das variáveis ​​é numérica. O arquivo test.csv tem os mesmos nomes de coluna e tipos de dados que train.csv, mas está faltando a coluna da variável resposta. Não há valores ausentes ou problemas de corrupção de dados em nenhum desses arquivos. Não se preocupe com os significados das variáveis ​​ou dos metadados - este é um conjunto de dados artificial.

Dicionário de variáveis

|  |  |
| --- | --- |
| **Variável** | **Tipo da variável (natureza)** |
| 016399044a | Quantitativa discreta |
| 023c68873b | Qualitativa nominal |
| 0342faceb5 | Quantitativa discreta |
| 04e7268385 | Quantitativa discreta |
| 06888ceac9 | Qualitativa nominal |
| 072b7e8f27 | Quantitativa contínua |
| 087235d61e | Quantitativa discreta |
| 0b846350ef | Quantitativa contínua |
| 0e2ab0831c | Quantitativa contínua |
| 12eda2d982 | Quantitativa contínua |
| 136c1727c3 | Quantitativa contínua |
| 173b6590ae | Quantitativa contínua |
| 174825d438 | Quantitativa discreta |
| 1f222e3669 | Quantitativa contínua |
| 1f3058af83 | Quantitativa discreta |
| 1fa099bb01 | Quantitativa discreta |
| 20f1afc5c7 | Quantitativa contínua |
| 253eb5ef11 | Quantitativa contínua |
| 25bbf0e7e7 | Quantitativa discreta |
| 2719b72c0d | Quantitativa contínua |
| 298ed82b22 | Quantitativa contínua |
| 29bbd86997 | Quantitativa contínua |
| 2a457d15d9 | Quantitativa discreta |
| 2bc6ab42f7 | Quantitativa contínua |
| 2d7fe4693a | Quantitativa contínua |
| 2e874bc151 | Quantitativa contínua |
| 361f93f4d1 | Qualitativa nominal |
| 384bec5dd1 | Qualitativa nominal |
| 3df2300fa2 | Quantitativa contínua |
| 3e200bf766 | Quantitativa discreta |
| 3eb53ae932 | Quantitativa contínua |
| 435dec85e2 | Quantitativa contínua |
| 4468394575 | Quantitativa contínua |
| 49756d8e0f | Quantitativa contínua |
| 4fc17427c8 | Quantitativa contínua |
| 55907cc1de | Quantitativa contínua |
| 55cf3f7627 | Quantitativa contínua |
| 56371466d7 | Quantitativa discreta |
| 5b862c0a8f | Quantitativa discreta |
| 5f360995ef | Quantitativa discreta |
| 60ec1426ce | Quantitativa contínua |
| 63bcf89b1d | Quantitativa contínua |
| 6516422788 | Quantitativa contínua |
| 65aed7dc1f | Quantitativa discreta |
| 6db53d265a | Quantitativa discreta |
| 7734c0c22f | Quantitativa contínua |
| 7743f273c2 | Quantitativa contínua |
| 779d13189e | Quantitativa contínua |
| 77b3b41efa | Quantitativa contínua |
| 7841b6a5b1 | Quantitativa contínua |
| 789b5244a9 | Quantitativa contínua |
| 7925993f42 | Quantitativa contínua |
| 7cb7913148 | Qualitativa ordinal |
| 7fe6cb4c98 | Quantitativa contínua |
| 8311343404 | Quantitativa contínua |
| 87b982928b | Quantitativa contínua |
| 8a21502326 | Quantitativa contínua |
| 8c2e088a3d | Quantitativa discreta |
| 8d0606b150 | Qualitativa nominal |
| 8de0382f02 | Quantitativa discreta |
| 8f5f7c556a | Quantitativa discreta |
| 91145d159d | Qualitativa nominal |
| 96c30c7eef | Quantitativa contínua |
| 96e6f0be58 | Quantitativa contínua |
| 98475257f7 | Quantitativa contínua |
| 99d44111c9 | Quantitativa discreta |
| 9a575e82a4 | Qualitativa nominal |
| 9b6e0b36c2 | Quantitativa contínua |
| a14fd026ce | Quantitativa discreta |
| a24802caa5 | Quantitativa contínua |
| aa69c802b6 | Quantitativa contínua |
| abca7a848f | Quantitativa discreta |
| ac826f0013 | Quantitativa contínua |
| ae08d2297e | Quantitativa discreta |
| aee1e4fc85 | Quantitativa contínua |
| b4112a94a6 | Quantitativa contínua |
| b709f75447 | Quantitativa contínua |
| b835dfe10f | Qualitativa nominal |
| b9a487ac3c | Quantitativa contínua |
| ba54a2a637 | Quantitativa contínua |
| bdf934caa7 | Quantitativa contínua |
| beb6e17af1 | Quantitativa discreta |
| c0c3df65b1 | Quantitativa discreta |
| c1b8ce2354 | Quantitativa discreta |
| c58f611921 | Quantitativa contínua |
| d035af6ffa | Quantitativa discreta |
| d2c775fa99 | Quantitativa discreta |
| d4d6566f9c | Quantitativa discreta |
| dcfcbc2ea1 | Quantitativa discreta |
| e0a0772df0 | Quantitativa contínua |
| e16e640635 | Qualitativa nominal |
| e5efa4d39a | Quantitativa contínua |
| e7ee22effb | Quantitativa discreta |
| e86a2190c1 | Quantitativa discreta |
| ea0f4a32e3 | Quantitativa contínua |
| ed7e658a27 | Quantitativa discreta |
| ee2ac696ff | Quantitativa contínua |
| f013b60e50 | Quantitativa contínua |
| f0a0febd35 | Quantitativa contínua |
| f1f0984934 | Qualitativa nominal |
| f66b98dd69 | Quantitativa contínua |
| fbf66c8021 | Quantitativa contínua |
| fdf8628ca7 | Quantitativa discreta |
| fe0318e273 | Quantitativa contínua |
| fe8cdd80ba | Quantitativa contínua |
| ffd1cdcfc1 | Quantitativa contínua |
| id | Id |
| target | Quantitativa contínua |

**4ª etapa: Preencher o quadro conceitual estatístico.**

|  |  |
| --- | --- |
| COMPONENTES | DESCRIÇÃO |
| 1. Plano Básico de Análise | 1. Importação dos dados. 2. Limpeza e pré-processamento dos dados. 3. Análise exploratória dos dados. 4. Criação de gráficos e visualizações. 5. Teste de hipóteses ou comparação de grupos. 6. Ajuste de modelos estatísticos. 7. Avaliação da qualidade do modelo. 8. Interpretação dos resultados. 9. Comunicação dos resultados de forma clara e concisa. |

**5ª etapa: Faça a análise descritiva das variáveis. Apresente os gráficos e as medidas resumos.**

Entrega do exercício no formato word e o script Rstudio no formato RMarkdown e o relatório com os resultados.

**Análise descritiva no markdown.**

**6ª etapa: Faça a análise bivariada das variáveis qualitativas.**

1. **Tabela de frequência bivariada.**
   1. Tabelas geradas no relatório markdown.
2. **Teste Qui-quadrado.**
   1. Tabelas geradas no relatório markdown.
3. **Gráfico 100% empilhado.**
   1. Imagens geradas no relatório markdown.

**7ª etapa: Faça a análise bivariada das variáveis quantitativas.**

1. **Gráfico de dispersão.**
   1. Gráficos gerados no relatório markdown.
2. **Análise de correlação de Pearson.**
   1. Análise no relatório markdown.
3. **Matriz de correlação de Pearson.**
   1. Tabela gerada no relatório markdown.

**8ª etapa: Preencher o quadro conceitual estatístico.**

|  |  |
| --- | --- |
| COMPONENTES | DESCRIÇÃO |
| 1. Técnica Estatística | Regressão linear Multipla |
| 1. Resultados Estatístico Principal | Equação Matemática (Consta no Simulador do Excel) |

**9ª etapa: Construção do modelo preditivo.**

1. **Selecionar as variáveis preditoras.**
   1. Consta no Markdown
2. **Definir a variável resposta.**
   1. Consta no Markdown
3. **Rodar o modelo de Regressão Linear Múltipla.**
   1. Consta no Markdown
4. **Análise de resíduos.**
   1. Consta no Markdown
5. **Calcular as medidas de erros do modelo na amostra train.csv.**
   1. Consta no Markdown
6. **Calcular a previsão do modelo na amostra validação.csv (30% da tabela de train.csv).**
   1. Consta no Markdown
7. **Calcular as medidas de erros do modelo na amostra test.csv**
   1. Os dados de teste não têm a coluna "target", podemos calcular as métricas de erro, já que elas dependem dos valores reais do target para comparação com as previsões do modelo. Aplicamos o modelo para amostra de teste e geramos a coluna target com os resultados.
8. **Construir o simulador do modelo.**
   1. Excel em anexo

Entrega do script Rstudio no formato RMarkdown e o relatório com os resultados.

Data de entrega: 07/07/2023

Regina Bernal

17/06/2023