Projeto TESTE DE PROFICIÊNCIA SEMANTIX

Resolução das questões

Este documento descreve o procedimento utilizado sobre os dados fornecidos para a resolução das questões.

1. Das ferramentas utilizadas:

Por questões de praticidade, a ferramenta analítica escolhida para a resolução do teste foi a plataforma R.

1. Do código fonte elaborado

Todos os procedimentos descritos neste documento podem ser reproduzidos através do código-fonte disponibilizado junto a este documento. Eventuais processos extremamente elementares que porventura não estejam citados neste documento, constarão do código-fonte, devidamente comentado.

1. Das bibliotecas utilizadas:

As bibliotecas utilizadas para a resolução das questões propostas, além das que são carregadas automaticamente no ambiente, foram:

1. PROCEDIMENTO DE RESOLUÇÃO
   1. Análise exploratória

Para que se possa responder quaisquer questões a respeito dos dados fornecidos, se faz necessária prévia análise exploratória inicial, tanto para a familiarização e possíveis tratamentos requeridos em relação aos dados como para entendimento do contexto o qual será fonte dos questionamentos.

Foram estudadas as distribuições das variáveis numéricas a fim de se entender se possíveis transformações poderiam contribuir no aumento da acuracidade dos modelos a serem utilizados nos próximos passos.

Os gráficos poderão ser gerados no código fonte.

O código fonte da análise exploratória se encontra no arquivo “teste\_Semantix\_analise\_exploratoria.R”

* 1. Aprimorando a qualidade dos dados substituindo os Unknown por dados previstos

Durante a análise exploratória, foi levantado que algumas variáveis possuíam diversos valores “unknow”. Tais ocorrências tendem tanto diminuir a potencialidade dos dados quanto fornecer respostas precisas ao questionamento feito, ou ainda, gerar respostas incorretas.

São essas as variáveis mais impactadas:

|  |  |
| --- | --- |
| VARIÁVEL | % UNKNOW |
| job | 0,8130% |
| marital | 0,1946% |
| education | 4,3871% |
| default | 26,3784% |
| housing | 0,2463% |
| loan | 0,2463% |

Variáveis de extrema importância como “default” ou “job” não poderiam conter números expressivos de valores desconhecidos, dados os efeitos mencionados.

A solução para melhorar a qualidade dos dados foi utilizar algoritmos de decision tree para “predizer” os valores marcados como unknown, do menos relevante para o mais relevante.

Desta forma, tivemos todas as variáveis mencionadas acima sem unknown.

A implementação desse incremento de qualidade dos dados está contido no arquivo “teste\_semantix\_acuracidade\_dataset\_01.R”

* 1. Código de apoio às respostas das questões contidas neste documento:

Por fim, para a confecções das respostas às questões, é fornecido o arquivo na qual as mesmas foram trabalhadas e os códigos e modelos utilizados para a confecção das respostas.

O arquivo com o código de apoio às questões chama-se “teste\_Semantix\_respostas.R”

* 1. Resposta as questões do teste

Questão 1 - Qual profissão tem mais tendência a fazer um empréstimo? De qual tipo?

Para responder esta questão, foi utilizado um simples algoritmo de Nayve Bayes.

O resultado mostrou que a maior tendência de empréstimo é para a categoria de emprego “unemployed”, que mostra a maior tendência entre todas a realizar um empréstimo do tipo House, que é de 56,61%.

Pela simples análise, este é um movimento comum de uma pessoa desempregada, de conseguir recursos hipotecando a casa.

Questão 2 - Fazendo uma relação entre número de contatos e sucesso da campanha quais

são os pontos relevantes a serem observados?

Ainda usando o método de Nayve Bayes, podemos observar com bastante relevância que:

1 – A probabilidade de sucesso da campanha anterior aumenta conforme aumenta a quantidade de contatos feitos

2 – A cada contato feito, é acrescentado uma probabilidade média de 13,42% de sucesso nas campanhas.

3 – Existe forte correlação entre a quantidade de contatos e o sucesso das campanhas anteriores, com taxa de 97,37%.

Questão 3 - Baseando-se nos resultados de adesão desta campanha qual o número médio e o máximo de ligações que você indica para otimizar a adesão?

Analisando graficamente os dados anteriores, podemos concluir que a um número médio de ligações eficientes é o de 4 contatos e o número máximo é de 7 contatos.

Questão 4 - O resultado da campanha anterior tem relevância na campanha atual?

As frequências de cada status de poutcome relacionadas às frequências de cada status da variável Y

sugerem que sim, o resultado da campanha anterior têm relevância no resultado da campanha atual.

Questão 5 - Qual o fator determinante para que o banco exija um seguro de crédito?

O fator determinante para que seja exigido um seguro de crédito é a inadimplência do cliente.

Questão 6 - Quais são as características mais proeminentes de um cliente que possua

empréstimo imobiliário??

São estas:

Idade média: 40 anos

Não possuem empréstimo pessoal (80%)

São casados (67%)

A enorme maioria tem algum tipo de formação escolar, sendo que a maior proporção, em curso básico.

Praticamente a totalidade não é inadimplente (há apenas um caso)