Bank Marketing

GRUPO 8

Manuel Monteiro

Vitor Peixoto

Tiago Alves

Universidade do Minho

Perfil de Ciência de Dados
Aprendizagem Automática I

Conteúdo

1. Análise exploratória dos dados
2. Tratar problemas do *dataset*
3. Construção do modelo
4. Discussão dos resultados
5. Trabalho futuro

Análise exploratória dos dados

Bank Marketing Dataset

Apresentação do *dataset*

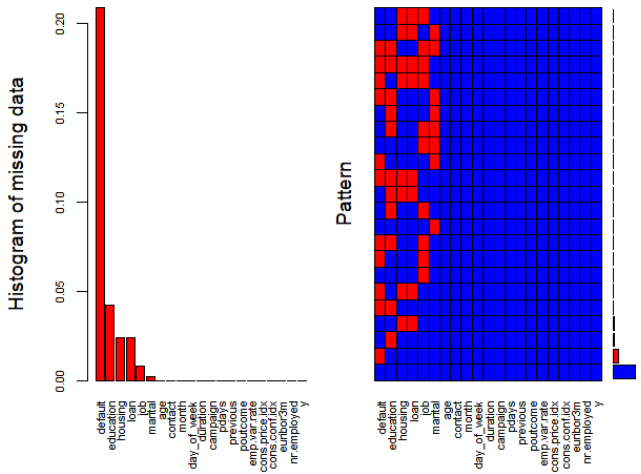
- *Dataset* de uma instituição bancária portuguesa.
- *Dataset* com 41188 entradas e 21 variáveis (10 numéricas e 11 categóricas).
- Objetivo: prever se o cliente vai subscrever depósito a prazo.
- Problema de classificação binária.

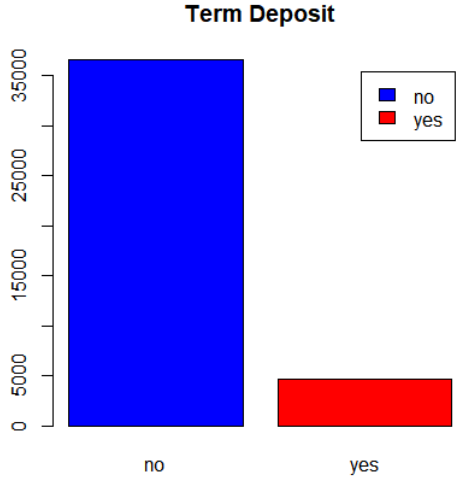
Questões Pertinentes

- Quais as variáveis que mais influenciam a decisão final do cliente?
- Que fator possui maior relação com a decisão final?

Tratar problemas do *dataset*

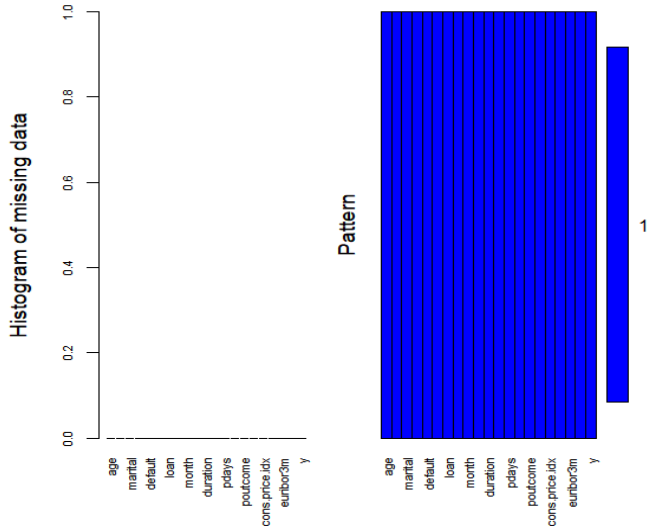
Tratar problemas do *dataset*





- Presença de dados ‘unknown’ para variáveis categóricas tratadas como *missing values*.
- Váriavel de interesse: *imbalanced aprox.* 88% 'no’ e 12% 'yes'.

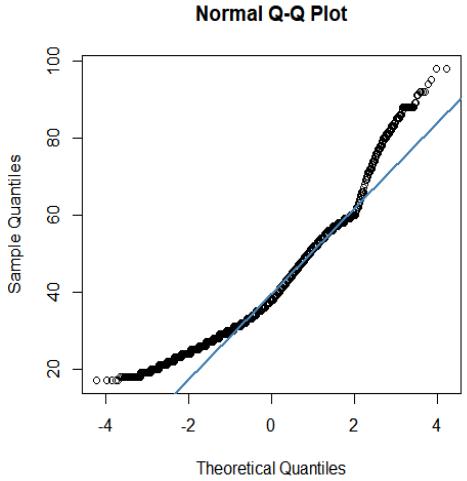
Tratar *missing values*

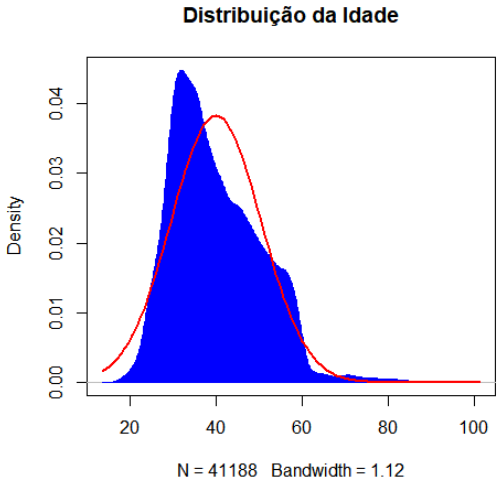


- Técnicas de tratamento de *missing values*:
  - Remoção das linhas
  - Imputação: Substituir por zero, mediana ou média
- *Package* '*mice*' do R; usa um algoritmo que utiliza informações de outras variáveis no *dataset imputar missing values*.

Tratar problemas do *dataset*

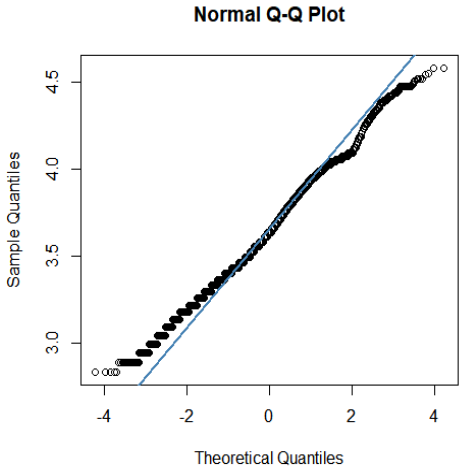
- Variável preditora com distribuição Gaussiana assimétrica.
- Conformidade de um conjunto de dados com a distribuição normal.

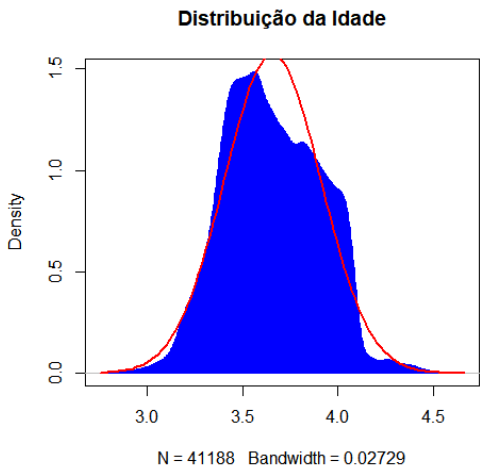




Tratar problemas do *dataset*

- Variável preditora depois de aplicada função estatística logarítmica (resultados traduzem variação relativa).
- Reduz efeito do viês.





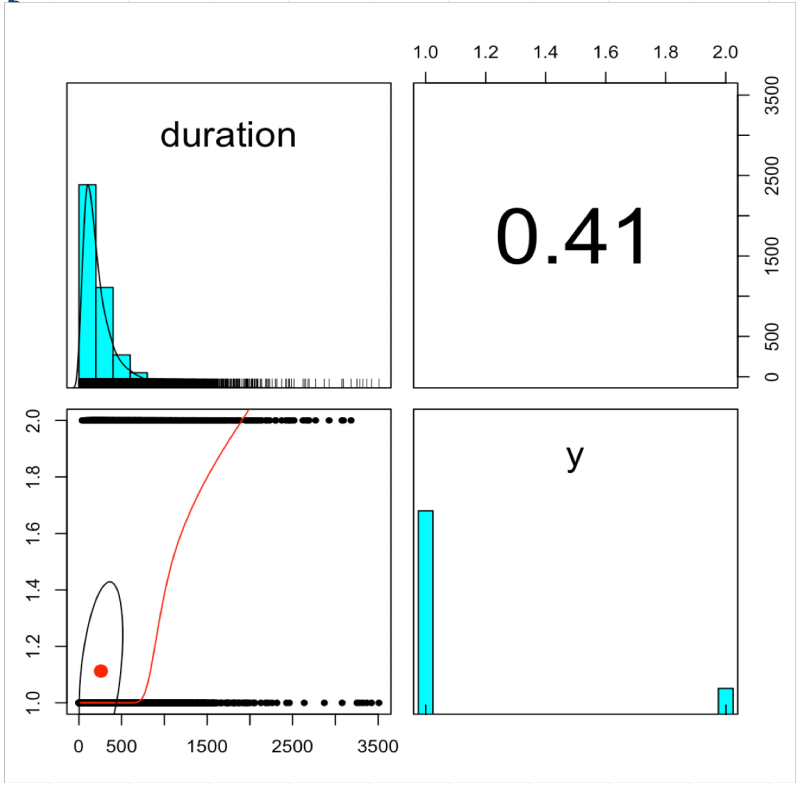
Construção do modelo

Construção do modelo

- Para resolver este problema foi aplicada Regressão Logística.
- Cálculo de correlação entre preditores e variáveis de decisão.
- Aplicou-se validação cruzada (Split: Training data 80% / Test Data 20%).
- Matriz de confusão – *Accuracy*.
- *ROC Curve* e *AUC* (*Area under curve*).

Exemplo – Correlação entre *Duration* e *Y*

- *Duration* apresenta uma correlação de 41% com y, sendo a variável mais influente na decisão final.



Discussão dos resultados

Discussão dos resultados

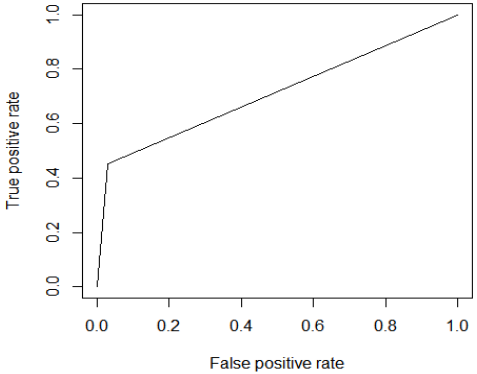
- Aplicação de Regressão Logística sobre todos os preditores (Modelo de *Baseline*)

duration	4.547e-03	9.288e-05	48.960	< 2e-16	***
campaign	-3.620e-02	1.433e-02	-2.527	0.011500	*
pdays	-1.031e-03	2.554e-04	-4.035	5.46e-05	***
previous	-6.361e-02	6.967e-02	-0.913	0.361232	
poutcomenonexistent	4.331e-01	1.118e-01	3.875	0.000107	***
poutcomesuccess	8.423e-01	2.481e-01	3.395	0.000686	***
emp.var.rate	-1.908e+00	1.668e-01	-11.442	< 2e-16	***
cons.price.idx	2.549e+00	3.039e-01	8.389	< 2e-16	***
cons.conf.idx	2.602e-02	9.364e-03	2.778	0.005467	**
euribor3m	2.177e-01	1.619e-01	1.345	0.178722	
nr.employed	9.861e-03	3.820e-03	2.581	0.009838	**
lnage	-1.904e+00	5.989e-01	-3.179	0.001477	**

- Foram escolhidas as variáveis estatisticamente significativas (preditores com *p-values* abaixo de 0.05).

Discussão dos resultados

- *ROC Curve e AUC*




```
> auc
[1] 0.6956938
```

Valor muito baixo!

- Matriz de confusão

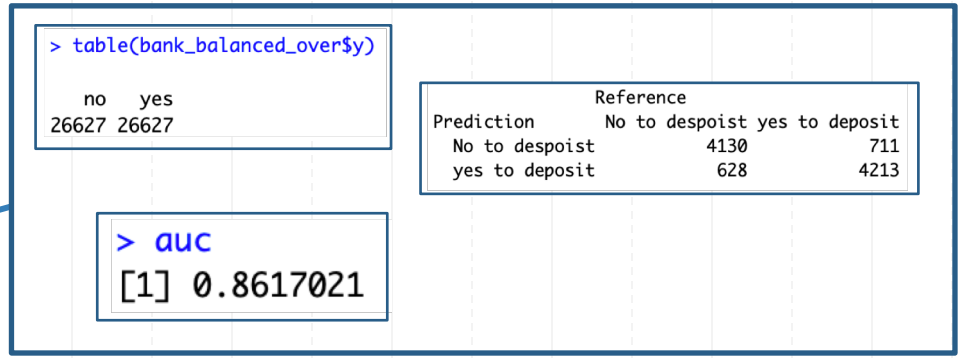
```
Confusion Matrix and Statistics

                Reference
Prediction       No to despoist yes to deposit
  No to despoist           4678             148
  yes to deposit            372             309

               Accuracy : 0.9056
                 95% CI : (0.8975, 0.9132)
    No Information Rate : 0.917
    P-Value [Acc > NIR] : 0.9988
```

Discussão dos resultados

- Não podemos analisar apenas a precisão (*accuracy*) do modelo!
- O valor de *AUC* e a matriz de confusão apresentam resultados que deixam a desejar.
- Porquê? ***Imbalanced data***!
- Como resolver?
  - *Undersampling*
  - *Oversampling*
  - *Cost Sensitive Learning*



Trabalho futuro

- Análise e limpeza mais extensiva dos dados (ex: remoção de *outliers*).
- Utilização de outros classificadores (*Decision trees*, *K-Nearest Neighbours*, *Quadratic Discriminant Analysis*).
- Analisar as consequências das previsões erradas.
- Utilização de métodos de contração (*LASSO* e *Ridge Regression*).

Bank Marketing

GRUPO 8

Manuel Monteiro

Vitor Peixoto

Tiago Alves

Universidade do Minho

Perfil de Ciência de Dados
Aprendizagem Automática I