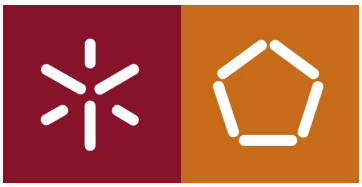


Universidade do Minho
Escola de Engenharia

Aprendizagem Automática I
Ciência de Dados

Bank Marketing

Autores:	**Número:**
Manuel Monteiro	PG37158
Tiago Alves	A78218
Vitor Peixoto	A79175

29 de Dezembro de 2018

Conteúdo

1 Introdução

O *telemarketing* é uma parte crucial da estratégia de *marketing* das organizações corporativas, tendo por finalidade, aumentar as vendas e manter um registo dos clientes.

O estudo dos seus clientes é um instrumento essencial para as empresas e organizações com vista a melhorarem as suas estratégias de *marketing*, permitindo entender melhor os seus clientes e quais os mercados a investir.

O objetivo deste trabalho visa estudar os dados obtidos por uma instituição bancária portuguesa, via conversa telefónica, junto dos seus clientes e determinar se os referidos clientes irão ou não subscrever um determinado produto financeiro.

Neste estudo, serão usadas diversas técnicas estatísticas de modo a melhorar a precisão de um modelo de regressão logística a ser desenvolvido, modelo esse sobre o qual esse estudo foi baseado, de modo a prever as variáveis que mais influenciam a escolha do cliente e de que modo influenciam.

2 Conjunto de dados

2.1 Contextualização

Os dados deste *dataset* foram recolhidos por um banco português, numa campanha de *telemarketing*, que foi realizada de modo a angariar clientes para subscreverem um novo depósito bancário a prazo.

Foram recolhidos os dados de contactos telefónicos a potenciais clientes durante um período de 5 anos, de 2008 a 2013.

Pretende-se prever se as pessoas subscrevem ou não os depósitos a prazo publicitados pela campanha.

2.2 Descrição do Conjunto de Dados

Este *dataset* é composto de 41188 registos e de 20 variáveis preditoras e 1 variável resultado.

Apresentamos de seguida, as variáveis preditoras e uma pequena descrição de cada uma delas:

Atributos relativos ao cliente

1. **age**: idade do cliente (quantitativa).
2. **job**: tipo de trabalho (qualitativa: *admin*, *blue-collar*, *entrepreneur*, *housemaid*, *management*, *retired*, *self-employed*, *services*, *student*, *technician*, *unemployed*, *unknown*).
3. **marital**: estado conjugal (qualitativa: *divorced*, *married*, *single*, *unknown*; nota: *divorced* significa divorciado ou viúvo).
4. **education**: escolaridade (qualitativa: *basic.4y*, *basic.6y*, *basic.9y*, *high.school*, *illiterate*, *professional.course*, *university.degree*, *unknown*).
5. **default**: tem crédito padrão (qualitativa: *no*, *yes*, *unknown*).
6. **housing**: tem crédito à habitação (qualitativa: *no*, *yes*, *unknown*).
7. **loan**: tem empréstimo pessoal (qualitativa: *no*, *yes*, *unknown*).

Atributos relativos ao contacto

8. **contact**: tipo de comunicação de contacto (qualitativa: *cellular*, *telephone*).
9. **month**: mês onde ocorreu último contacto (qualitativa: *jan*, *feb*, ..., *dec*).
10. **day_of_week**: dia da semana onde ocorreu último contacto (qualitativa: *mon*, *tue*, *wed*, *thu*, *fri*).

11. **duration**: duração do último contacto, em segundos (quantitativa). Nota: este atributo afeta altamente o alvo output (e.g., se duration=0 então y='no'). Ainda assim, a duração não é conhecida antes que a chamada seja realizada. Também, depois do fim da chamada y é obviamente conhecido. Portanto, este input apenas deveria ser incluído para propósitos de *benchmark* e deveria ser descartado caso a intenção seja ter um modelo preditivo realista.

Atributos relativos a anteriores contactos

12. **campaign**: número de contactos realizados durante esta campanha e para este cliente (quantitativa, inclui último contacto).
13. **pdays**: número de dias que passaram desde que o cliente foi contactado por uma campanha anterior (quantitativa; *999* significa que o cliente não foi previamente contactado).
14. **previous**: número de contactos realizados antes desta campanha e para o cliente (quantitativa).
15. **poutcome**: resultado do campanha de *marketing* anterior (qualitativa: *failure*, *nonexistent*, *success*).

Atributos de contexto social e económico

16. **emp.var.rate**: taxa de variação de emprego - indicação trimestral (quantitativa).
17. **cons.price.idx**: índice de preços ao consumidor - indicação mensal (quantitativa).
18. **cons.conf.idx**: índice de confidência do consumidor - indicador mensal (quantitativa).
19. **euribor3m**: taxa de 3 meses da *euribor* - indicador diário (quantitativa).
20. **nr.employed**: número de empregados - indicador trimestral (quantitativa)

Variável Resultado

21. **y**: o cliente subscreveu um depósito a prazo? (binária: *yes*, *no*)

2.3 Supervisionado ou Não supervisionado

Estamos perante um problema de aprendizagem supervisionado uma vez que neste tipo de aprendizagem são apresentados dois conjuntos de dados, o conjunto de *input* e o conjunto de *output* esperado.

2.4 Regressão ou Classificação

O objetivo do problema é prever se uma chamada telefónica influencia o cliente a subscrever um depósito a prazo.

Estamos perante um problema de classificação uma vez que o que queremos "prever"é uma variável categórica (*yes* ou *no*).

2.5 Questões

Conhecidas as variáveis a explorar, tornou-se pertinente colocar as seguintes questões a este conjunto de dados:

- Quais os fatores (preditores) que mais influenciam a decisão final do cliente, após contactado pela campanha de marketing?
- Entre as varíaveis obtidas como resposta à questão anterior, qual aquela que mais influência exercita sobre o cliente?
- Geralmente, no futuro, qual o tipo de clientes que deverão ser contactados?

3 Análise Exploratória dos Dados

Uma análise inicial do conjunto de dados que temos em mãos é essencial para analisar as variáveis que o compõem e a variância apresentada pelas mesmas. A análise das variáveis existentes permite detetar problemas inerentes ao conjunto de dados que terão de ser resolvidos antes de gerar um modelo preditivo.

Este conjunto de dados é constituído por um volume considerável de variáveis preditoras, pelo que iremos apresentar gráficos apenas para as variáveis mais interessantes em termos de apresentarem problemas para a geração de um modelo preditivo. Os gráficos das restantes variáveis estão disponíveis em anexo.

Como método de análise, foram computacionados gráficos (*boxplot* e *barplot*) para as diversas variáveis categóricas e quantitativas que compõem este *dataset*.

Foram encontrados vários dados que considerámos poderem apresentar problemas para a geração de um modelo preditivo. Apesar de este conjunto de dados não apresentar dados nulos (*missing data*), verificamos a existência de um atributo *unknown* comum a várias variáveis. De certa forma, pode-se considerar este atributo como *missing data*.

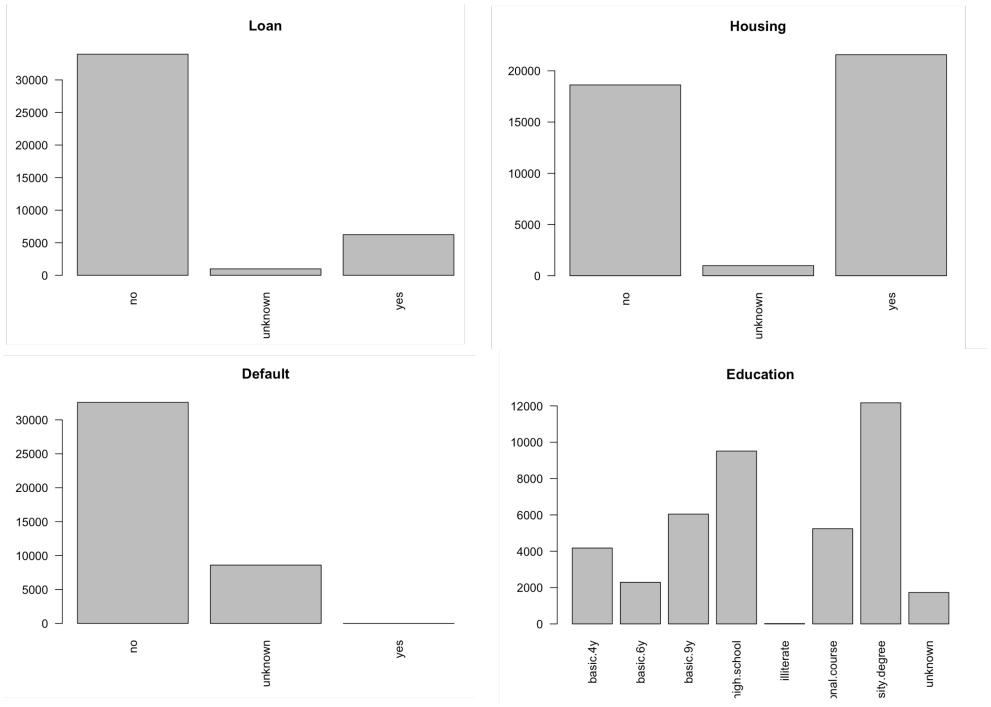


Figura 1: Existência de atributos *unknown* em várias variáveis.

Outro fator importante na análise exploratória dos dados (EDA) é a deteção de *outliers*.

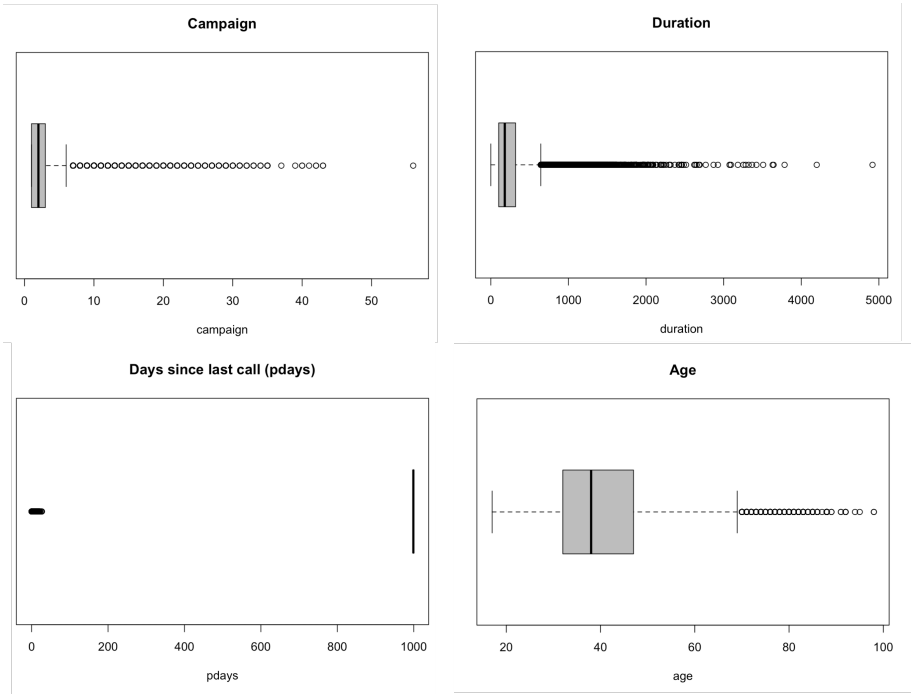


Figura 2: Existência de *outliers* em várias variáveis.

De facto, em algumas das variáveis quantitativas encontra-se um grande número de *outliers*. Porém analisando alguns dos *outliers*, eles revelam-se válidos. No entanto, na variável *pdays* verificamos uma disparidade muito grande entre a média e os *outliers*. Isto deve-se ao facto de representar o número de dias desde a última chamada, sendo que se for a primeira chamada, o valor é 999.

Outro facto curioso é o desequilíbrio entre os resultados recolhidos da ação de *marketing* uma vez que o número de resultados negativos é bastante inferior ao número de resultados positivos.

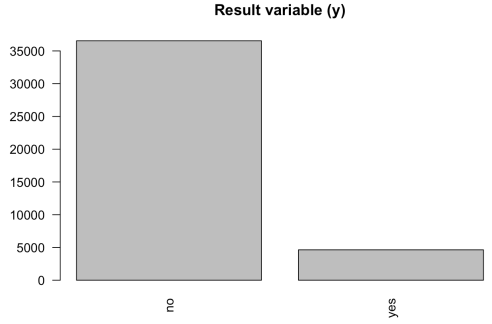


Figura 3: Desequilíbrio na variável de resposta.

Foram assim detetados alguns problemas no conjunto de dados, que deverão ser corrigidos de seguida.

4 Tratar problemas do *dataset*

Antes de passarmos para a criação dos modelos preditivos, torna-se necessário tratar os problemas que este *dataset* apresenta. Esses problemas apresentam-se em vários locais, desde a variável *y imbalanced* até a uma variável preditora não-normalizada. Todos esses fatores afetam negativamente o resultado do modelo e não reproduzem a realidade do contexto de onde o *dataset* foi recolhido.

4.1 Variável de interesse *imbalanced*

Um problema observado no nosso conjunto de dados era o desequilíbrio acentuado na variável de resultado. Este desequilíbrio provoca uma viciação no modelo obtido uma vez que a existência de um grande número de classes *no* permite construir um modelo com grande precisão, uma vez que basta prever sempre para a classe *no*. Este problema chama-se *overfitting* e traduz-se um modelo preciso mas incorreto.

Torna-se então necessário equilibrar os dados relativos à variável de interesse *y*.

Um dos métodos existentes para balancear os dados é o aumento da amostragem da classe minoritária, denomidado por *oversampling*. Deste modo obtemos um *dataset* balanceado o que leva a modelos equilibrados e mais corretos, apesar de poder traduzir-se numa redução da precisão.

A técnica de *oversampling* pode apresentar desvantagens como por exemplo o *overfitting* devido à duplicação de instâncias. No entanto, o *overfitting* causado pelo *oversampling* é menor do que o *overfitting* causado por dados desequilibrados. Aplicando assim *oversampling* ao conjunto de dados:

```
bank_marketing_data_full <- ovun.sample(y ~ ., data = bank_marketing_data_full, method = "over",N = 53000)$data
```

Como resultado temos um equilíbrio nos atributos da variável resultado, reduzindo assim a quantidade de *overfitting* para *no*.

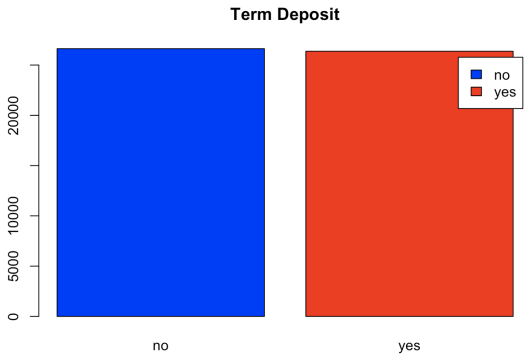


Figura 4: Variável resultado equilibrada.

4.2 *Missing Values*

Outro problema encontrado na EDA é a existência não de *missing values*, mas de atributos *unknown*. Estes poderão ser tratados como *missing values* uma vez que é um atributo comum a várias variáveis. De facto, se assumirmos os atributos *unknown* como *missing data* (NA), descobrimos que são bastante frequentes em algumas das variáveis.

```
bank_marketing_data_full[bank_marketing_data_full=="unknown"] <- NA
```

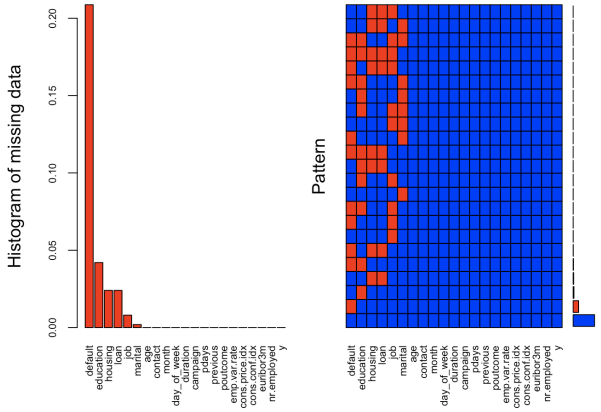


Figura 5: Frequência de *missing data* nas variáveis preditivas.

Sendo o nosso conjunto de dados volumoso, isso permite-nos remover as linhas com *missing data* visto a influência no modelo preditivo ser reduzida, pois ainda ficamos com um volume de dados significativo.

No entanto foi efetuada pesquisa sobre como substituir *missing data* caso a sua remoção influenciasse o conjunto de dados. O *R* tem um pacote denominado *MICE* que permite a substituição de *missing values* por valores sintéticos plausíveis usando o método de *sampling Gibbs*.

4.3 *Outliers*

De facto, como observado na EDA, existem diversos *outliers* em diversas variáveis.

Porém aquela que mais se destaca e que mais problemas poderá representar será a variável *pdays*. Relembrando o que foi referido nesse capítulo, o valor 999 significa que esta é a primeira vez que se efetua uma chamada para o referido cliente e não que a última chamada foi há 999 dias. Logo é incorreto caraterizar esta variável como numérica.

Assim sendo, optamos por transformá-la numa variável categórica. Tal permite que não haja um intervalo tão grande entre os valores "reais"(1,4,20,...) e o 999. Removemos ainda os valores para 20, 25, 26 e 27 dias pois só tinham uma instância e estavam a provocar erros na divisão dos dados para treino e teste (a ser abordado mais à frente).

```
bank_marketing_data_full$pdays <- factor(bank_marketing_data_full$pdays)
bank_marketing_data_full<-bank_marketing_data_full[!(bank_marketing_data_full$pdays==20 |
                                                     bank_marketing_data_full$pdays==25 |
                                                     bank_marketing_data_full$pdays==26 |
                                                     bank_marketing_data_full$pdays==27),]
barplot(table(bank_marketing_data_full$pdays), main="Days since last call (pdays)", col=16, las=2)
```

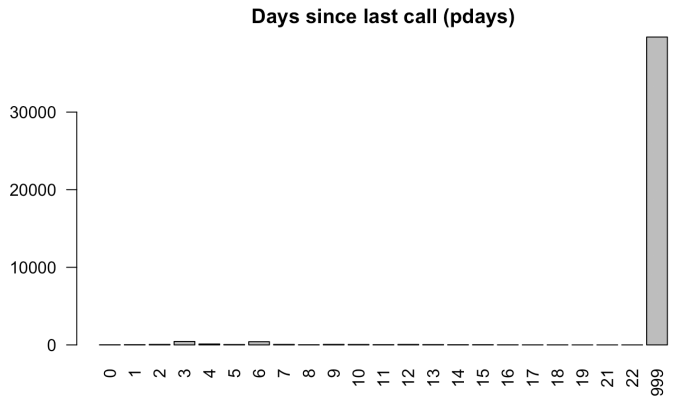


Figura 6: Gráfico da variável *pdays* após a transformação.

5 Regressão Logística

Num caso de classificação, existem vários algoritmos de geração de modelos, tais como a Regressão Logística, K Vizinhos Mais Próximos ou a Análise Discriminante Linear e Quadrática. Porém optámos por abordar apenas a Regressão Logística, uma vez que é o algoritmo considerado mais preciso na geração de modelos na sua globalidade. No entanto, nunca deixamos de parte a futura implementação de mais algoritmos para testar a possível melhoria na precisão do modelo gerado.

5.1 Dados de teste

De modo a poder avaliar a precisão do modelo a ser desenvolvido, é necessário haver a separação do conjunto de dados para treino e teste do modelo.

Aqui tentámos inicialmente a aplicação de *k-fold cross validation*, porém sem sucesso e a gerar erros não relacionados com a própria aplicação de *cross validation*. Resolvemos assim deixar isto para uma etapa futura, aplicando agora uma simples separação dos dados, tendo noção que este método implica um modelo menos preciso do que com *k-fold cross validation*.

```
set.seed(123)
sample = sample.split(bank_marketing_data_full,SplitRatio = 0.80)
train_data = subset(bank_marketing_data_full, sample==TRUE)
test_data =  subset(bank_marketing_data_full, sample==FALSE)
```

Efetuámos então um *split* simples dos dados usando um rácio de 80% dos dados para o conjunto de treino e 20% para o conjunto de teste.

5.2 Gerar modelo

Para gerar um modelo usando Regressão Logística em *R* usamos o método *glm.fit* utilizando como conjunto de dados os dados de treino separados anteriormente do conjunto original. Nesta primeira etapa, vamos usar como variáveis preditoras todas as que compõem o conjunto de dados, exceto y que será usada como variável resultado. O modelo é gerado automaticamente pelo algoritmo de regressão e os resultados podem ser observados aplicando *summary* ao modelo gerado.

```
model<-glm(y~.,data = train_data,family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = y ~ ., family = binomial, data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.7089  -0.3859  -0.1297   0.5053   3.0889
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -2.744e+02  2.773e+01  -9.895  < 2e-16 ***
## age                         -3.132e-03  1.900e-03  -1.648  0.09927 .
## jobblue-collar              -2.879e-01  6.136e-02  -4.693 2.70e-06 ***
## jobentrepreneur             -1.419e-01  9.272e-02  -1.531  0.12588
## jobhousemaid                 5.029e-03  1.168e-01   0.043  0.96564
## jobmanagement               -1.081e-01  6.383e-02  -1.694  0.09032 .
## jobretired                   5.098e-01  8.841e-02   5.767 8.07e-09 ***
```

Figura 7: Geração do modelo usando Regressão Logística.

Observando os coeficientes podemos analisar alguns dados interessantes relativamente à significância das variáveis preditoras.

5.3 Variáveis significativas

A análise das variáveis dá-se pela análise do *p-value* para o qual se rejeita a hipótese nula. Neste caso a hipótese nula é que a variável preditora não apresenta influência na variável resultado. Tipicamente *p-value=0.05*, logo podemos afirmar como insignificantes as variáveis cujo *p-value* se situe acima desse valor.

O *R* ajuda também ao atribuir um número de estrelas, de zero a três, para cada variável.

De modo a auxiliar a compreensão do funcionamento do *p-value*, criámos uma lista com os valores deste para cada variável e invertemos o seu *p-value* subtraindo 1 por este (*1-pvalue*). Deste modo, as variáveis mais significativas serão aquelas que têm um valor de significância superior a 0.95.

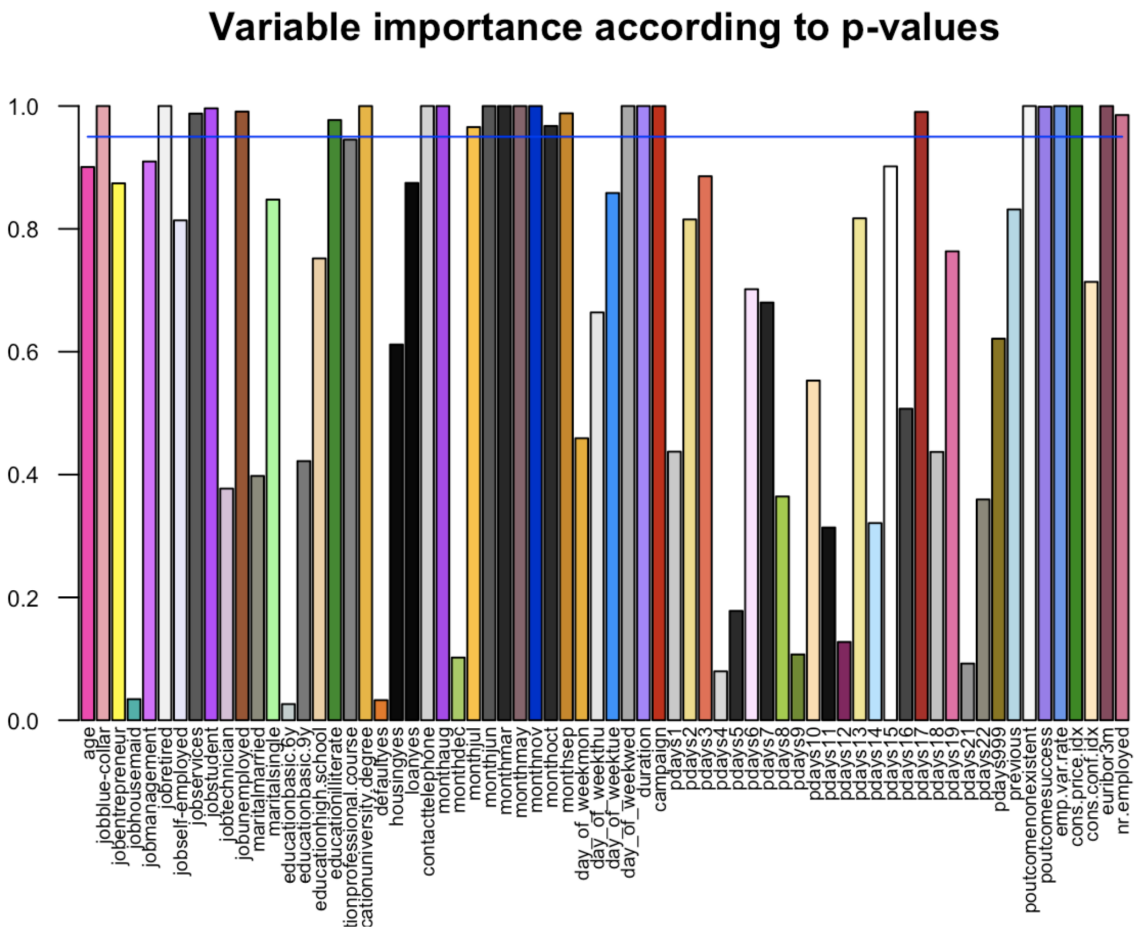


Figura 8: Variáveis significativas.

Analisando então o gráfico acima, conseguimos extrair as variáveis mais significativas, ou seja, aquelas acima da linha azul (0.95): *job*, *education*, *contact*, *month*, *day_of_week*, *duration*, *campaign*, *poutcome*, *emp.var.rate*, *cons.price.idx*, *euribor3m* e *nr.employed*.

No entanto, em algumas variáveis categóricas apenas alguns atributos são mais influentes no *outcome*, como por exemplo *education*. Optámos por mantê-las visto serem significativas na mesma e ao removê-las de um modelo preditivo iríamos perder precisão na classificação.

Para além de descobrir as variáveis preditivas mais significativas do modelo, podemos ainda descobrir a **influência** que estes têm na variável resultado *y*.

A influência é analisada na primeira coluna dos *Coefficients* obtidos no modelo. Podemos então separar as variáveis significativas entre as que influenciam o resultado **positivamente** (*job-retired*, *job-student*, *education-illiterate*, *education-professional.course*, *education-university.degree*, *month-aug*, *month-mar*, *month-oct*, *month-sep*, *day_of_week-tue*, *day_of_week-wed*, *duration*, *poutcome-nonexistant*, *poutcome-success*, *cons.price.idx*, *euribor3m* e *nr.employed*) e as que influenciam o resultado **negativamente** (*job-*

blue.collar, *job-services*, *contact-telephone*, *month-jun*, *month-may*, *month-nov*, *campaign* e *emp.var.rate*).

```
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -3.030e+02  2.095e+01 -14.466  < 2e-16 ***
## jobblue-collar               -2.301e-01  6.099e-02  -3.773 0.000161 ***
## jobentrepreneur              -4.258e-02  9.031e-02  -0.472 0.637281
## jobhousemaid                  1.665e-01  1.138e-01   1.463 0.143492
## jobmanagement                -6.339e-02  6.193e-02  -1.024 0.306031
## jobretired                    4.345e-01  7.598e-02   5.719 1.07e-08 ***
## jobself-employed             -5.821e-02  8.415e-02  -0.692 0.489114
## jobservices                  -1.355e-01  6.438e-02  -2.104 0.035342 *
## jobstudent                    3.857e-01  8.978e-02   4.296 1.74e-05 ***
## jobtechnician                 9.249e-03  5.326e-02   0.174 0.862139
## jobunemployed                 1.027e-01  9.996e-02   1.028 0.304093
## educationbasic.6y             6.246e-02  9.792e-02   0.638 0.523542
## educationbasic.9y             5.899e-02  7.592e-02   0.777 0.437176
## educationhigh.school          1.163e-01  7.329e-02   1.586 0.112675
## educationilliterate           1.872e+00  6.054e-01   3.092 0.001987 **
## educationprofessional.course  1.967e-01  8.051e-02   2.444 0.014532 *
## educationuniversity.degree    4.059e-01  7.382e-02   5.499 3.83e-08 ***
## contacttelephone             -5.354e-01  5.286e-02 -10.129  < 2e-16 ***
## monthaug                      1.163e+00  9.560e-02  12.168  < 2e-16 ***
## monthdec                     -8.447e-02  1.779e-01  -0.475 0.634893
## monthjul                     -8.886e-02  7.208e-02  -1.233 0.217653
## monthjun                     -1.040e+00  8.421e-02 -12.348  < 2e-16 ***
## monthmar                      2.155e+00  1.073e-01  20.096  < 2e-16 ***
## monthmay                     -7.563e-01  5.972e-02 -12.664  < 2e-16 ***
## monthnov                     -7.494e-01  8.772e-02  -8.543  < 2e-16 ***
## monthoct                      3.651e-01  1.139e-01   3.205 0.001350 **
## monthsep                      5.308e-01  1.318e-01   4.028 5.61e-05 ***
## day_of_weekmon                2.742e-02  4.994e-02   0.549 0.582995
## day_of_weekthu               -1.789e-02  4.972e-02  -0.360 0.718958
## day_of_weektue                1.317e-01  5.031e-02   2.618 0.008840 **
## day_of_weekwed                2.187e-01  4.997e-02   4.376 1.21e-05 ***
## duration                      6.765e-03  7.996e-05  84.606  < 2e-16 ***
## campaign                     -2.797e-02  8.438e-03  -3.315 0.000916 ***
## poutcomenonexistent           4.135e-01  4.566e-02   9.056  < 2e-16 ***
## poutcomesuccess               1.752e+00  7.419e-02  23.619  < 2e-16 ***
## emp.var.rate                 -2.507e+00  9.793e-02 -25.603  < 2e-16 ***
## cons.price.idx                2.794e+00  1.481e-01  18.867  < 2e-16 ***
## euribor3m                     6.503e-01  7.040e-02   9.237  < 2e-16 ***
## nr.employed                   6.972e-03  1.555e-03   4.484 7.32e-06 ***
```

Figura 9: Influência dos preditores no resultado.

Nas variáveis quantitativas, se o valor apresentado na coluna *Estimate* for positivo, indica que um aumento nessa unidade produz um efeito positivo na probabilidade do cliente subscrever (*yes* na variável resultado y) e o inverso caso o valor seja negativo. Quanto maior o módulo desse valor, maior é o efeito (positivo ou negativo) na variável resultado.

Nas variáveis qualitativas, se o valor apresentado na coluna *Estimate* for positivo significa que se a variável for desse valor, a probabilidade do cliente subscrever o depósito é maior e o inverso caso o valor seja negativo, novamente.

Por exemplo, como observamos na imagem, *contact-telephone* tem uma influência de -0.5354. Isto significa que se o contacto for efetuado para o telefone fixo, a probabilidade do cliente aderir ao depósito é menor do que se o fizer para o telemóvel. Noutro caso, para a variável *duration*, verifica-se que quanto maior for a duração da chamada com o cliente, maior a probabilidade deste subscrever o depósito publicitado.

Este tipo de análise é essencial para saber que mercado (p.e. jovens, desempregados, reformados, etc.) a instituição bancária deverá "atacar"e quais as condições que deverá reunir (p.e. chamada de maior duração, contactar o telemóvel, ligar apenas no verão, etc.) para obter resultados positivos.

5.4 Análise do modelo

Para avaliar a capacidade preditiva do modelo gerado, utilizamos a função *predict* do *R* que gera previsões através do modelo de regressão logística e depois geramos a sua matriz de confusão comparando os resultados obtidos com o conjunto de teste.

```
test_result <- predict(model,test_data,type = "response")
test_result <- ifelse(test_result > 0.5,1,0)

test_result<-round(test_result,0)
test_result<-as.factor(test_result)
levels(test_result)<-c("no","yes")
actual1<-test_data[,21]
levels(actual1)<-c("no","yes")

conf1<-confusionMatrix(actual1,test_result,positive = "yes")
conf1
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##        no  5358  982
##        yes  795 5484
##
##                Accuracy : 0.8592
##                  95% CI : (0.853, 0.8652)
##     No Information Rate : 0.5124
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7184
##  Mcnemar's Test P-Value : 1.023e-05
##
##             Sensitivity : 0.8481
##             Specificity : 0.8708
##          Pos Pred Value : 0.8734
##          Neg Pred Value : 0.8451
##              Prevalence : 0.5124
##          Detection Rate : 0.4346
##    Detection Prevalence : 0.4976
##       Balanced Accuracy : 0.8595
##
##        'Positive' Class : yes
##
```

Figura 10: Precisão preditiva do modelo.

Assim sendo o modelo obteve uma precisão de 85.92%, valor semelhante ao obtido

para a *Area Under Curve* (AUC).

5.5 Novo modelo

Para verificar que os preditores não significativos não tinham influência na precisão do modelo, construímos um novo modelo, novamente utilizando o algoritmo de regressão logística, cujas variáveis preditoras são apenas as variáveis mais significativas.

```
model_sig<-glm(y~job+education+contact+month+day_of_week+duration+campaign+poutcome+emp.var.rate+cons.price.idx
+euribor3m+nr.employed, data = train_data,family = binomial)
summary(model_sig)
```

```
test_result_sig <- predict(model_sig,test_data,type = "response")
test_result_sig <- ifelse(test_result_sig > 0.5,1,0)

test_result_sig <- round(test_result_sig,0)
test_result_sig <- as.factor(test_result_sig)
levels(test_result_sig) <- c("no","yes")
actual2 <- test_data[,21]
levels(actual2) <- c("no","yes")

conf2 <- confusionMatrix(actual2,test_result_sig,positive = "yes")
conf2
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##        no  5359  981
##        yes  811 5468
##
##                Accuracy : 0.858
##                  95% CI : (0.8518, 0.864)
##     No Information Rate : 0.5111
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.716
##  Mcnemar's Test P-Value : 6.545e-05
##
##             Sensitivity : 0.8479
##             Specificity : 0.8686
##          Pos Pred Value : 0.8708
##          Neg Pred Value : 0.8453
##              Prevalence : 0.5111
##          Detection Rate : 0.4333
##    Detection Prevalence : 0.4976
##       Balanced Accuracy : 0.8582
##
##        'Positive' Class : yes
##
```

Como se pode comprovar, a remoção das variáveis consideradas insignificantes (cujo *p-value* era inferior a 0.05) teve uma influência mínima na precisão do modelo, cuja percentagem de acerto se verificou nos 85.80%.

6 Conclusões e trabalho futuro

Os resultados que obtivemos sugerem as variáveis que mais contribuirão para o sucesso do estudo, bem como o modo como contribuem, positiva ou negativamente. Podemos concluir que para uma campanha promocional de *marketing* semelhante à desenvolvida no período do conjunto de dados, os agentes da instituição bancária deverão considerar:

- Comunicar preferencialmente reformados e estudantes, em detrimento de trabalhadores de serviços;
- Comunicar preferencialmente analfabetos e pessoas com cursos profissionais ou grau académico;
- Comunicar para o telemóvel do cliente, em detrimento do telefone;
- Publicitar o depósito nos meses de março, agosto, setembro e outubro em detrimento dos meses de maio, junho, novembro e dezembro.
- Contactar os clientes às terças e quartas em detrimento das sextas-feiras.
- Conseguir prolongar a chamada com o cliente o máximo possível.
- Evitar efetuar demasiadas chamadas para o mesmo cliente, visto que a probabilidade de subscrição do depósito desce de acordo com o número de contactos efetuados.
- Clientes que não têm registo ou que subscreveram aos depósitos de campanhas anteriores terão mais probabilidade de aderir do que aqueles que recusaram anteriores ofertas.

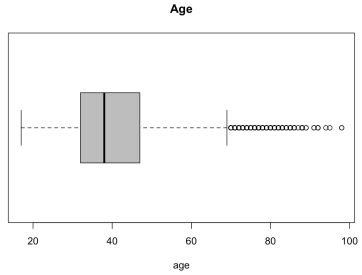
Todas as medidas quando seguidas e implementadas pela instituição financeira no seu processo de *marketing* devem permitir reduzir o custo promocional das campanhas, reduzir o número de clientes contactados por telefone e melhorar a eficácia geral da campanha, aumentando o número de subscrições do depósito e assim o lucro da instituição bancária.

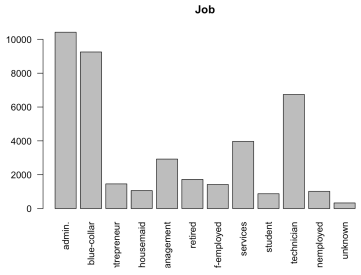
Num estudo deste tipo é sempre possível melhorar o modelo e este certamente. Como falámos anteriormente, a utilização de *k-fold cross validation* que iria melhorar a precisão do modelo, mas infelizmente não conseguimos implementar, seria uma solução bastante interessante para adicionar. A utilização de mais algoritmos de geração de modelos de classificação seria também uma mais valia pois iria permitir comparar a precisão entre os diversos modelos desenvolvidos. Para além destes dois pontos importantes para trabalho futuro, temos outros, tais como: limpeza mais extensiva dos dados; utilização de modelos de contração (*shrinkage*) e análise de componentes principais (PCA).

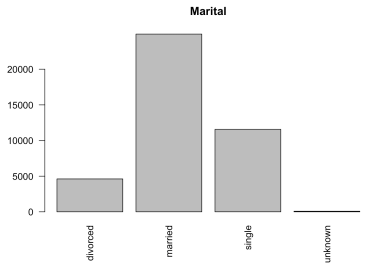
Conclui-se assim que o estudo desenvolvido permitiu responder às questões propostas inicialmente, pelo que podemos considerar que o trabalho desenvolvido foi positivo, mas com espaço ainda para muitas melhorias.

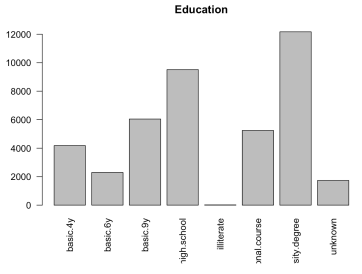
7 Anexos

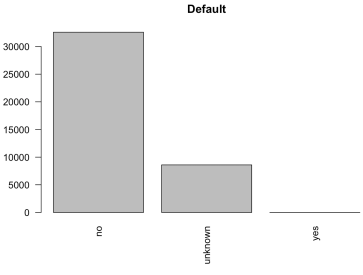
Análise Exploratória dos Dados

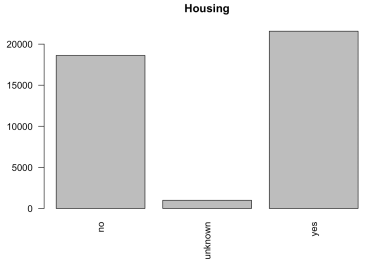


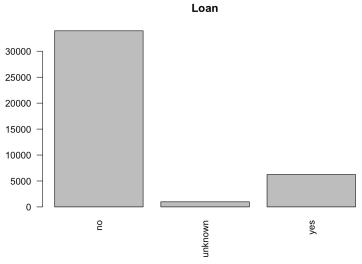


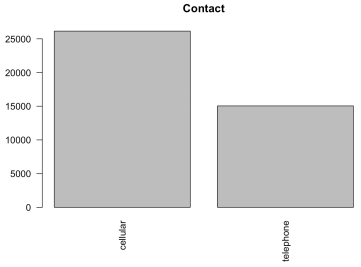


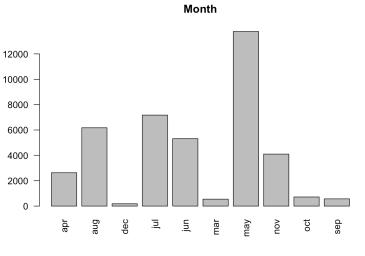


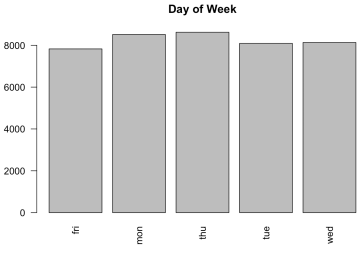


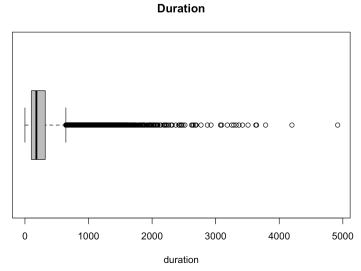


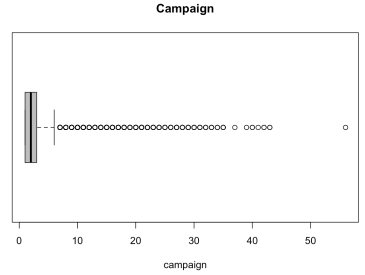


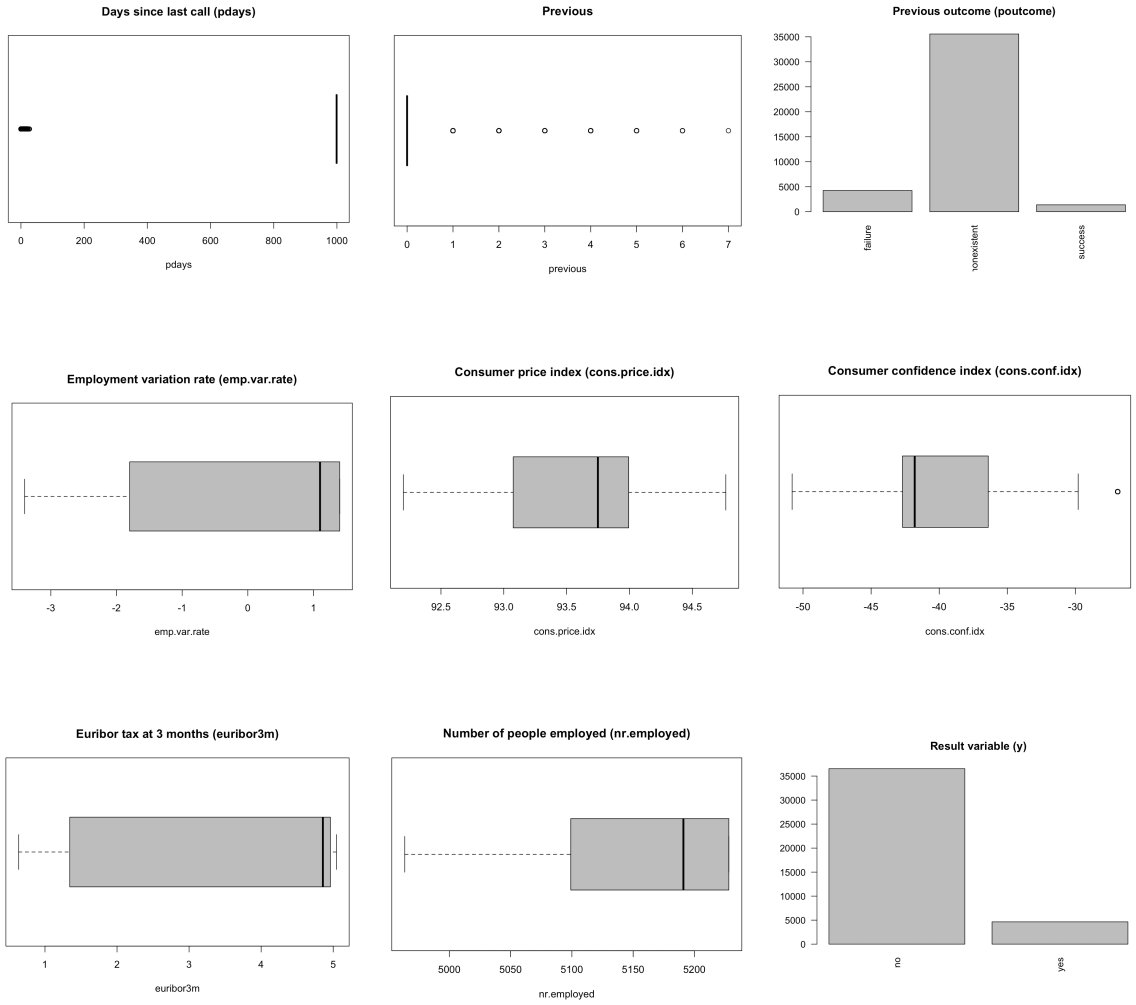












Código *R* e *Slides*

O código R, bem como os *slides* de apresentação deste projeto, encontram-se em anexo na pasta comprimida onde se encontra este relatório também.