Universidade do Minho

Mestrado em Engenharia Informática

Ciência de Dados

Aprendizagem Automática I

João Pimentel A80874
José Carvalho A80424
Ricardo Martins A78914

11 de Janeiro de 2020

Resumo

O presente projeto consistiu na análise de um conjunto de dados referente a campanhas de *telemarketing* por parte de um banco, de modo a angariar clientes para aderirem às mesmas. Para tal, foi necessário, numa fase inicial, realizar um tratamento de dados para permitir poder relacionar mais facilmente os preditores associados com a variável resultado. Posteriormente, criaram-se três modelos supervisionados de regressão logística, aos quais se foi retirando, gradualmente, variáveis menos significativas. Estes modelos foram testados com *validation set approach*. Além disso, todos os modelos foram comparados com *k-fold cross validation*. Numa fase de análise de resultados, observou-se que o modelo capaz de responder às questões de interesse com maior exatidão consistia naquele que continha um menor número de variáveis associado. Através deste trabalho, foi possível constatar que os preditores que mais impacto tiveram no desempenho do modelo consistiam nos seguintes: *contact, month, day_of_week, duration, campaign, poutcome, emp.var.rate* e *cons.price.idx*.

Conteúdo

Lista de Figuras

1 Introdução

O presente trabalho tinha como objetivo a análise de um conjunto de dados à escolha do grupo, utilizando as metodologias lecionadas na Unidade Curricular de Aprendizagem Automática I. Assim, desenvolveram-se modelos de aprendizagem que, com base nos dados de entrada, conseguiriam aprender as caraterísticas mais relevantes do conjunto de dados para, posteriormente, ser capaz de prever o resultado dado um outro conjunto, dentro da mesma temática. Para tal, recorreu-se à linguagem de programação R.

1.1 Contextualização

A base de dados escolhida é constituída por dados referentes a campanhas de *marketing* de uma instituição bancária portuguesa, sendo que estas campanhas foram efetuadas através de chamadas telefónicas. Por vezes, foi necessário mais do que um contacto com o cliente, de modo a determinar o interesse ou falta deste no produto [1].

Assim, o objetivo da classificação é prever se o cliente pretende subscrever um depósito a prazo, componente que se refere à variável *y*.

1.2 Variáveis

O conjunto de dados em estudo possui 41188 registos e 21 variáveis, sendo uma delas a variável resultado.

Seguidamente, serão enumeradas as variáveis relativas ao cliente, ao último contacto com campanha decorrente e ainda a atributos extra relacionados com os dois aspetos anteriores. Ademais, serão mencionados os preditores relativos aos contextos social e económico. Por fim, será introduzida a variável de decisão.

1.2.1 Cliente

1. *Age*: idade do respetivo cliente (variável numérica);
2. *Job*: profissão (variável categórica: *admin.*, *blue-collar*, *entrepreneur*, *housemaid*, *management*, *retired*, *self-employed*, *services*, *student*, *technician*, *unemployed*, *unknown*);
3. *Marital*: estado conjugal do cliente (variável categórica: *divorced*, *married*, *single*, *unknown*; *divorced* pode corresponder a divorciado ou viúvo);
4. *Education*: nível de escolaridade (variável categórica: *basic.4y*, *basic.6y*, *basic.9y*, *high.school*, *illiterate*, *professional.course*, *university.degree*, *unknown*);
5. *Default*: caso tenha um crédito padrão associado (variável categórica: *no*, *yes*, *unknown*);
6. *Housing*: caso tenha pedido um empréstimo de habitação (variável categórica: *no*, *yes*, *unknown*);
7. *Loan*: caso tenha pedido um empréstimo pessoal (variável categórica: *no*, *yes*, *unknown*).

1.2.2 Último Contacto com a Decorrente Campanha

1. *Contact*: via de comunicação (variável categórica: *cellular*, *telephone*);
2. *Month*: último mês onde se realizou contacto (variável categórica: *jan*, *feb*, *mar*, ..., *nov*, *dec*);
3. *Day_of_week*: último dia da semana em que se realizou contacto (variável categórica: *mon*, *tue*, *wed*, *thu*, *fri*);
4. *Duration*: duração do último contacto em segundo (variável numérica).

1.2.3 Outros Atributos

1. *Campaign*: número de contactos realizados com um determinado cliente durante a decorrente campanha (variável numérica e inclui o último contacto);
2. *Pdays*: número de dias que passaram depois do cliente ter entrado em contacto com a campanha anterior (variável numérica; *999* significa que o cliente não foi contactado anteriormente);
3. *Previous*: número de vezes que se efetuou contacto com um mesmo cliente antes desta campanha (variável numérica);
4. *Poutcome*: resultado da campanha de marketing anterior (variável categórica: *failure*, *nonexistent*, *success*).

1.2.4 Contextos Social e Económico

1. *Emp.var.rate*: taxa de variação emprego, indicador trimestral (variável numérica);
2. *Cons.price.idx*: índice de preço no consumidor, indicador mestral (variável numérica);
3. *Cons.conf.idx*: índice de confiança no consumidor, indicador mestral (variável numérica);
4. *Euribor3m*: taxa da euribor a três meses, indicador diário (variável numérica);
5. *Nr.employed*: número de empregados, indicador trimestral (variável numérica).

1.2.5 Decisão

y: caso o cliente tenha subscrito um depósito a termo (variável binária: *no*, *yes*).

1.3 Questões de Interesse

Tendo em conta as diversas variáveis incluídas nesta base de dados, considerou-se apropriado estudar três questões principais:

1. De entre as variáveis, quais são as que mais influenciam a decisão do cliente?
2. Quais as estratégias a tomar para aumentar as chances de sucesso?
3. Que clientes deverão ser contactados?

2 Metodologia

2.1 Análise Exploratória dos Dados

O primeiro passo para a compreensão do problema trata-se de uma análise exploratória dos dados. Assim, é possível detetar problemas associados aos registos e resolver os mesmos antes da geração do modelo preditivo.

Tendo em conta a grande quantidade de variáveis associadas a cada registo, serão apresentados apenas os gráficos mais relevantes, referentes aos preditores que mais interesse possuem. O primeiro método analítico utilizado foi a apresentação gráfica de relações entre os dados e deles próprios.

Assim, o primeiro possível problema encontrado nos dados foi a discrepância entre a quantidade de registos por cada categoria da variável de resultado, como se pode ver pela Figura 1.

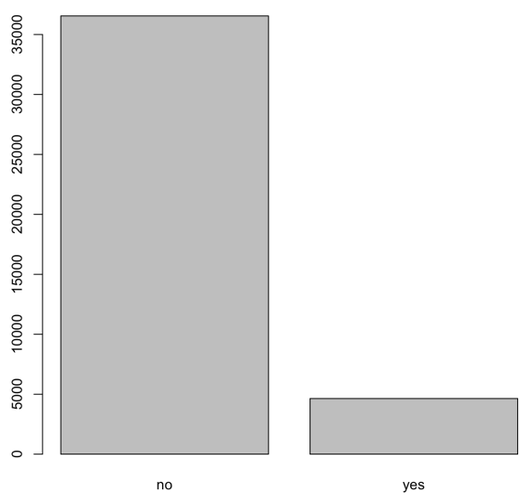

Figura 1 - Variável de resultado.

Em seguida, a ideia passou por ver como eram compostos alguns dos atributos associados aos clientes. Como se pode observar pela Figura 2, apesar de não existirem valores nulos, existem campos desconhecidos, sendo que estes podem afetar o resultado final da predição negativamente.

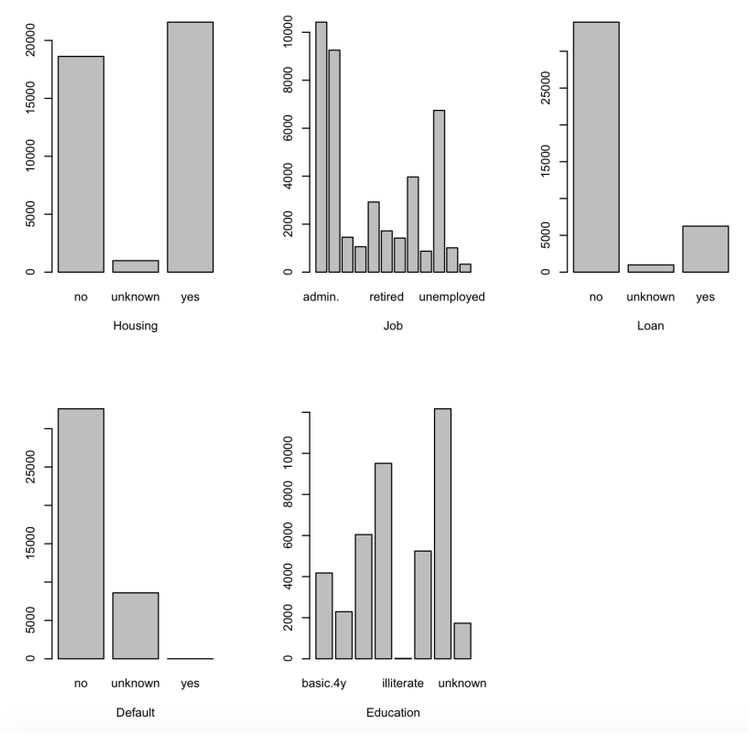

Figura 2 - Existência de valores indefinidos.

Outro aspeto merecedor de destaque é a existência de *outliers* nas variáveis. Pela análise da Figura 3, pode-se ver que existe uma grande quantidade de *outliers* nas variáveis em questão, apesar de que, após alguma análise, maioria destes mostrou-se válida. No entanto, no que toca ao preditor referente aos dias desde a última chamada (*pdays*), o valor que mais vezes é visto nos dados é o valor associado, por defeito, a nunca ter sido efetuada uma chamada, ou seja, o valor 999. Isto faz com que valores reais deste atributo sejam vistos como *outliers*.

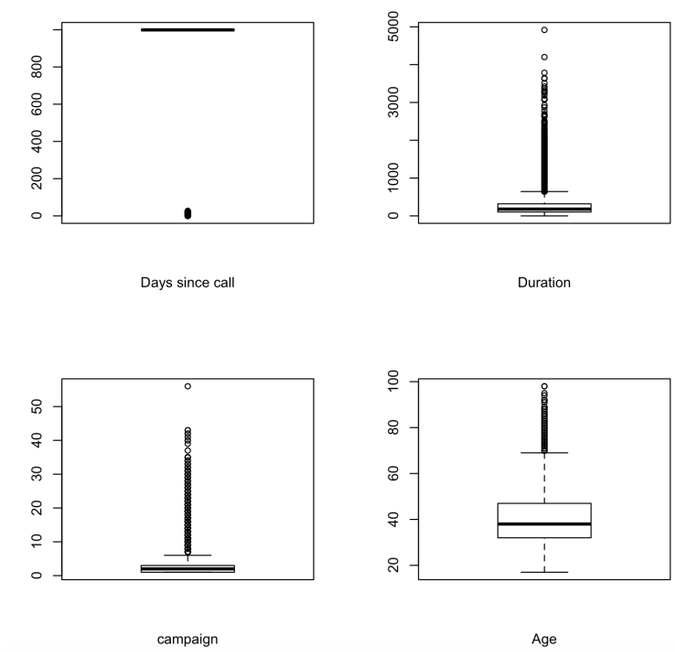

Figura 3 - *Outliers*.

2.2 Tratamento dos Dados

Nesta secção serão apresentadas as formas de resolução dos problemas anteriormente mencionadas, bem como o estado do conjunto de dados após cada tratamento. É de destacar que não foi efetuado tratamento ao conjunto de dados no que toca ao desbalanceamento da variável de decisão, uma vez que esta ação levaria a uma adulteração dos dados e consequente adulteração do modelo.

2.2.1 Valores Indefinidos

O primeiro caso a tratar é a existência de valores indefinidos, sendo que estes podem ser vistos como valores nulos, já que são comuns a vários preditores, como se observa na Figura 4.

age	job	marital	education	default	housing	loan	contact
0	330	80	1731	8597	990	990	0
month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate
0	0	0	0	0	0	0	0
cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y			
0	0	0	0	0			

Figura 4 - Contagem de valores indefinidos por variável.

Tendo em conta o volume do *dataset* em estudo, é possível remover os registos com valores em falta, sem ter grande influência no modelo. Note-se que, apesar desta remoção de registos, o conjunto continua com uma dimensão considerável. Outra solução seria a substituição destes valores nulos por valores concretos, derivados dos registos no *dataset*, mas esta solução não foi necessária.

2.2.2 Categorização dos *pdays*

Outro problema que teve de ser resolvido antes da definição do modelo, é referente à variável *pdays*. Como foi mencionado anteriormente, existe um valor por defeito caso nunca tenha sido efetuada uma chamada ao cliente. No entanto, tendo em conta a definição do mesmo, o seu verdadeiro significado é que a última chamada efetuada ao cliente ocorreu há 999 dias, pelo que a caraterização desta variável como numérica não é a melhor solução, devendo ser categorizada.

Dito isto, optou-se por agrupar os valores em conjuntos, sendo estes referentes a semanas, ou seja, todos os valores entre 0 e 7 dias pertencem à categoria *1W* (*1 week*), os valores entre 8 e 14 pertencem à categoria *2W* (*2 weeks*), os valores por defeito (999) pertencem a *N* e os restantes a *+2W* (*>2 weeks*). Como se pode observar pela Figura 5, o número de clientes que nunca foram contactados pelo banco é o mais comum em todo o *dataset*. Note-se que os contactos ocorrem frequentemente no espaço de uma semana após a primeira interação.

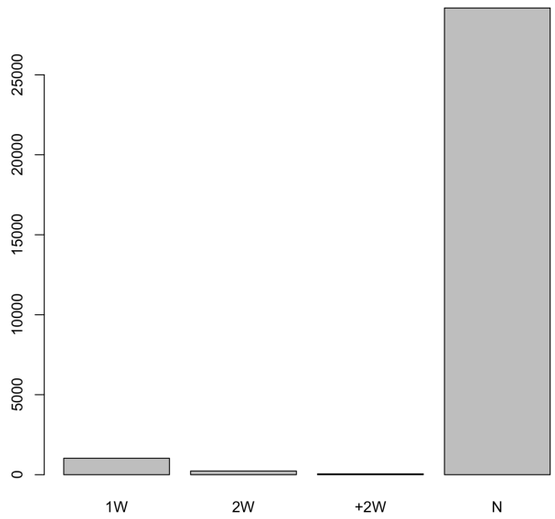

Figura 5 - Contagem de registos por categoria da variável *pdays*.

2.3 Regressão Logística

Tendo em conta que este se trata de um problema de classificação, ou seja, a variável de resposta é qualitativa, o objetivo é prever a probabilidade de resposta desta mesma variável, ou seja, a probabilidade da categoria ocorrer. Assim, tendo em conta as técnicas lecionadas nas aulas da unidade curricular, o grupo optou por implementar um modelo de regressão logística múltipla.

Note-se que, como se pode deduzir pelas Equações 1 e 2, o incremento de, por exemplo, $X1$ em uma unidade leva a um aumento das *log-odds* em $\beta 1$, sendo que esta alteração não corresponde à alteração em $Pr(Y = cat1|X)$. Esta informação será relevante posteriormente para a compreensão do modo como cada variável afeta a predição dos modelos.

$$Pr(Y = cat1|X) = \frac{e^{\beta 0+\beta 1X1+...+\beta pXp}}{1 + e^{\beta 0+\beta 1X1+...+\beta pXp}} \tag{1}$$

$$log(\frac{Pr(Y = cat1|X)}{1 - Pr(Y = cat1|X)}) = \beta 0 + \beta 1X1 + ... + \beta pXp \tag{2}$$

2.3.1 Amostragem dos Dados

De modo a avaliar a exatidão do modelo, foi necessário separar o conjunto de dados em dados de treino e dados de teste. Esta separação foi efetuada de forma simples, amostrando de forma aleatória registos, tendo por base uma percentagem que dita quantos registos serão de treino e quantos serão de teste. Assim, foi decidido que se utilizariam 80% dos dados para treino e os restantes 20% para teste do modelo.

2.3.2 Modelo Geral

De modo a gerar um modelo de Regressão Logística na linguagem *R*, foi utilizada função *glm.fit*, sendo que esta recebia como argumento o conjunto de dados de treino.

Tendo em conta que ainda não existe uma significância de cada variável no que toca ao seu peso na decisão, o primeiro passo passou por utilizar todos os preditores no modelo. Assim, efetuado um sumário do modelo gerado, visível na Figura 6, é possível constatar que as variáveis cujos *p-values* são menores que 0.05 possuem maior significância no resultado final do que as restantes.

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -2.409e+02  4.841e+01  -4.976 6.48e-07 ***
age                           -3.841e-03  3.126e-03  -1.229  0.21918
jobblue-collar                -2.320e-01  1.030e-01  -2.252  0.02430 *
jobentrepreneur               -1.484e-01  1.573e-01  -0.943  0.34565
jobhousemaid                  -6.416e-03  1.902e-01  -0.034  0.97309
jobmanagement                 -1.208e-01  1.072e-01  -1.126  0.25996
jobretired                     2.971e-01  1.375e-01   2.160  0.03079 *
jobself-employed              -1.117e-01  1.413e-01  -0.791  0.42903
jobservices                   -1.942e-01  1.086e-01  -1.788  0.07381 .
jobstudent                     1.955e-01  1.435e-01   1.362  0.17318
jobtechnician                  8.776e-03  8.726e-02   0.101  0.91989
jobunemployed                 -1.647e-01  1.636e-01  -1.007  0.31399
maritalmarried                -6.305e-03  8.696e-02  -0.073  0.94220
maritalsingle                  3.659e-02  9.793e-02   0.374  0.70867
educationbasic.6y             -8.504e-02  1.662e-01  -0.512  0.60893
educationbasic.9y             -4.176e-03  1.253e-01  -0.033  0.97341
educationhigh.school          -3.242e-02  1.204e-01  -0.269  0.78765
educationilliterate            2.267e+00  9.123e-01   2.485  0.01297 *
educationprofessional.course   1.487e-02  1.315e-01   0.113  0.90999
educationuniversity.degree     1.240e-01  1.206e-01   1.028  0.30414
defaultyes                    -7.243e+00  1.133e+02  -0.064  0.94903
housingyes                    -2.731e-02  5.150e-02  -0.530  0.59588
loanyes                       -7.357e-02  7.152e-02  -1.029  0.30365
contacttelephone              -6.440e-01  9.579e-02  -6.723 1.78e-11 ***
monthaug                       8.938e-01  1.499e-01   5.964 2.46e-09 ***
monthdec                       1.450e-01  2.637e-01   0.550  0.58231
monthjul                       1.258e-01  1.220e-01   1.031  0.30238
monthjun                      -6.098e-01  1.547e-01  -3.940 8.13e-05 ***
monthmar                       1.962e+00  1.785e-01  10.992  < 2e-16 ***
monthmay                      -4.114e-01  1.025e-01  -4.012 6.03e-05 ***
monthnov                      -4.324e-01  1.531e-01  -2.825  0.00473 **
monthoct                       2.589e-01  1.944e-01   1.332  0.18286
monthsep                       4.123e-01  2.267e-01   1.818  0.06900 .
day_of_weekmon                -7.492e-02  8.407e-02  -0.891  0.37281
day_of_weekthu                 1.443e-01  8.148e-02   1.771  0.07649 .
day_of_weektue                 1.398e-01  8.392e-02   1.666  0.09568 .
day_of_weekwed                 2.547e-01  8.306e-02   3.067  0.00216 **
duration                       4.601e-03  9.713e-05  47.375  < 2e-16 ***
campaign                      -4.634e-02  1.491e-02  -3.108  0.00188 **
pdays2W                       -3.422e-01  2.116e-01  -1.617  0.10591
pdays+2W                      -5.218e-01  3.835e-01  -1.361  0.17359
pdaysN                        -1.188e+00  2.898e-01  -4.100 4.13e-05 ***
previous                      -5.941e-02  7.319e-02  -0.812  0.41693
poutcomenonexistent            4.719e-01  1.166e-01   4.048 5.16e-05 ***
poutcomesuccess                7.779e-01  2.715e-01   2.865  0.00417 **
emp.var.rate                  -1.816e+00  1.737e-01 -10.458  < 2e-16 ***
cons.price.idx                 2.232e+00  3.162e-01   7.060 1.66e-12 ***
cons.conf.idx                  1.568e-02  9.730e-03   1.611  0.10713
euribor3m                      3.626e-01  1.691e-01   2.144  0.03203 *
nr.employed                    5.518e-03  3.992e-03   1.382  0.16686
```

Figura 6 - Modelo geral.

2.3.2.1 Variáveis Significativas

Como mencionado, a análise da significância das variáveis dá-se pelos seus *p-values*, ou seja, a probabilidade da hipótese nula. Esta hipótese representa a não influência da variável no resultado. Tipicamente, assume-se que as variáveis cujo *p-value* é superior a 0.05 não são significantes para o problema.

De modo a auxiliar à compreensão destes valores, foi efetuada uma representação gráfica da probabilidade da rejeição da hipótese nula, ou seja, de $H1$. Quer isto dizer que se efetuou o complemento dos *p-values*, de modo que as variáveis mais significativas sejao aquelas cujo valor seja superior ou igual a 0.95. Dito isto, pela análise da Figura 7 é possível observar que os preditores com mais peso no resultado final são: *contact*, *month*, *day_ of_ week*, *duration*, *campaign*, *pdays*, *poutcome*, *emp.var.rate* e *cons.price.idx*.

Note-se que, no caso de certas variáveis categóricas, apenas alguns valores são influentes no resultado final, como é o caso do preditor *month*. Assim, tendo em conta que continuam a ser variáveis significativas e a sua remoção iria diminuir a precisão do modelo, foi decidido que seriam mantidas para futura análise.

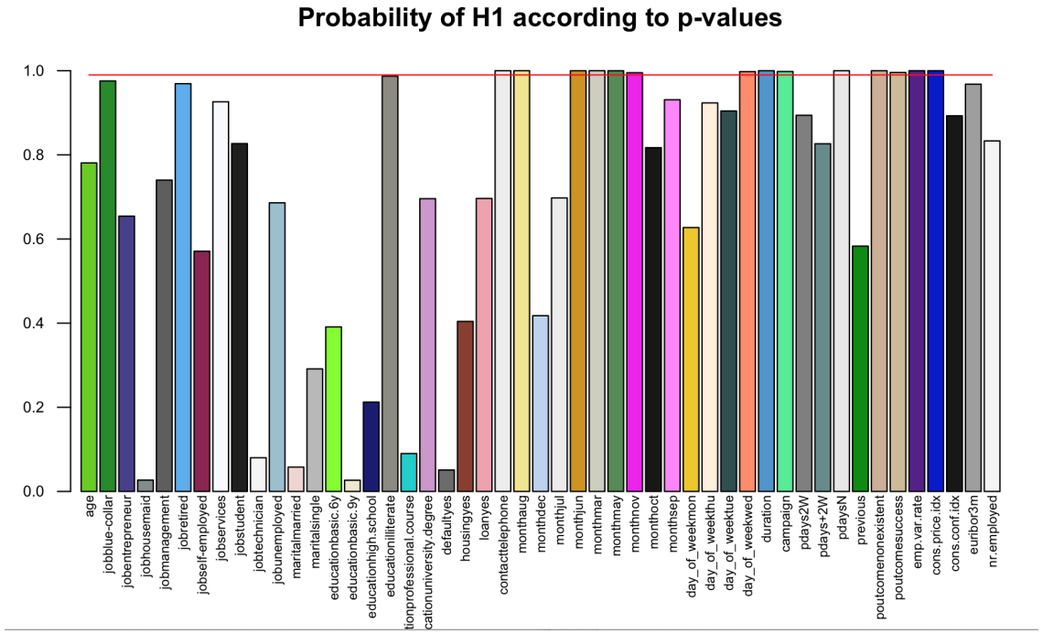

Figura 7 - Análise da significância das variáveis.

Além de se determinarem as variáveis mais significativas do modelo, é, ainda, possível determinar a influência que estes têm na probabilidade de cada categoria do resultado.

No que toca a variáveis quantitativas, se o valor da coluna *Estimate* for positivo, um aumento nesse preditor produz um efeito positivo na probabilidade da variável resultado ser positiva, sendo que o contrário acontece se for um valor negativo. É de notar que, quanto maior for o módulo do valor na coluna, maior será o efeito na variável resultado.

Já no caso das variáveis qualitativas, um valor positivo implica que, na hipótese de a variável ser dessa categoria, a probabilidade do cliente subscrever o depósito é maior. O inverso acontece se o valor for negativo, ou seja, a probabilidade diminui.

Tenha-se, a título de exemplo, *day_ of_ weekmon* tem uma influência de -0.07492, querendo isto dizer que se o contacto for efetuado numa segunda-feira, a probabilidade do cliente aderir ao depósito é menor que se for efetuada numa terça-feira. Outro caso interessante é o do preditor *duration*, em que se verifica que, quanto maior a duração da chamada, maior a probabilidade da subscrição ao depósito.

2.3.3 Modelo Significativo

Tendo por base as variáveis significativas obtidas anteriormente, a ideia passou por gerar um modelo que apenas tivesse em consideração estas. Assim, através da análise da Figura 8, é possível constatar que todas as variáveis são em certo modo significativas para o resultado final, apesar de algumas categorias das mesmas não o serem.

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.178e+02  6.296e+00 -18.709  < 2e-16 ***
contacttelephone    -3.735e-01  7.839e-02  -4.764 1.90e-06 ***
monthaug             9.167e-01  1.056e-01   8.679  < 2e-16 ***
monthdec             3.573e-01  2.334e-01   1.531 0.125851
monthjul             4.164e-01  1.085e-01   3.837 0.000124 ***
monthjun            -2.164e-02  1.086e-01  -0.199 0.842071
monthmar             1.770e+00  1.381e-01  12.812  < 2e-16 ***
monthmay            -5.397e-01  9.151e-02  -5.898 3.69e-09 ***
monthnov            -9.311e-02  1.091e-01  -0.854 0.393228
monthoct             5.232e-01  1.314e-01   3.981 6.86e-05 ***
monthsep             4.148e-01  1.427e-01   2.907 0.003654 **
day_of_weekmon      -8.998e-02  8.337e-02  -1.079 0.280474
day_of_weekthu       1.235e-01  8.078e-02   1.530 0.126129
day_of_weektue       1.382e-01  8.316e-02   1.662 0.096437 .
day_of_weekwed       2.386e-01  8.255e-02   2.890 0.003851 **
duration             4.587e-03  9.688e-05  47.344  < 2e-16 ***
campaign            -4.896e-02  1.491e-02  -3.285 0.001021 **
pdays2W             -3.916e-01  2.077e-01  -1.885 0.059426 .
pdays+2W            -5.467e-01  3.764e-01  -1.452 0.146414
pdaysN              -1.222e+00  2.662e-01  -4.589 4.45e-06 ***
poutcomenonexistent  5.732e-01  7.825e-02   7.325 2.39e-13 ***
poutcomesuccess      7.595e-01  2.564e-01   2.963 0.003051 **
emp.var.rate        -1.002e+00  2.936e-02 -34.118  < 2e-16 ***
cons.price.idx       1.223e+00  6.685e-02  18.292  < 2e-16 ***
```

Figura 8 - Modelo com variáveis significativas apenas.

2.3.4 Modelo Significativo - Sem *pdays*

Após uma análise aos valores associados à variável *pdays*, é notório que a probabilidade de um depósito ser efetuado, tendo em conta esta variável, não aumenta, como se podia observar na Figura 8. Ora, no caso de o último contacto ter sido feito há uma semana, ou menos, tenha-se em conta o valor de $\beta 0$, o qual é negativo, diminuindo, assim, a probabilidade de aceitação. No caso de duas ou mais semanas, os valores de β seguem o mesmo padrão, além de que a significância das categorias é reduzida. Já no caso da categoria *Never*, como maioria dos registos estão associados a esta, pode existir uma falsa correlação no resultado, pelo que é necessária a definição, treino e teste de um modelo sem o preditor para confirmar a não necessidade do mesmo.

Assim, como se pode ver pela Figura 9, as variáveis do modelo significativo continuam significativas, como desejado, sendo agora possível efetuar uma análise sobre as mesmas.

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.207e+02  6.220e+00 -19.399  < 2e-16 ***
contacttelephone    -3.712e-01  7.823e-02  -4.745 2.08e-06 ***
monthaug             9.523e-01  1.049e-01   9.082  < 2e-16 ***
monthdec             3.620e-01  2.324e-01   1.558 0.119322
monthjul             4.341e-01  1.084e-01   4.005 6.19e-05 ***
monthjun            -1.373e-02  1.084e-01  -0.127 0.899248
monthmar             1.783e+00  1.378e-01  12.938  < 2e-16 ***
monthmay            -5.332e-01  9.137e-02  -5.835 5.37e-09 ***
monthnov            -5.592e-02  1.084e-01  -0.516 0.605980
monthoct             5.399e-01  1.312e-01   4.116 3.85e-05 ***
monthsep             4.226e-01  1.424e-01   2.967 0.003006 **
day_of_weekmon      -8.552e-02  8.330e-02  -1.027 0.304592
day_of_weekthu       1.270e-01  8.073e-02   1.573 0.115617
day_of_weektue       1.437e-01  8.306e-02   1.730 0.083698 .
day_of_weekwed       2.437e-01  8.245e-02   2.956 0.003116 **
duration             4.589e-03  9.683e-05  47.393  < 2e-16 ***
campaign            -4.923e-02  1.491e-02  -3.300 0.000965 ***
poutcomenonexistent  5.044e-01  7.539e-02   6.691 2.22e-11 ***
poutcomesuccess      1.832e+00  1.033e-01  17.740  < 2e-16 ***
emp.var.rate        -1.011e+00  2.915e-02 -34.679  < 2e-16 ***
cons.price.idx       1.241e+00  6.623e-02  18.736  < 2e-16 ***
```

Figura 9 - Modelo sem preditor *pdays*.

Note-se que, posteriormente, poderiam ser retirados mais preditores do modelo, um a um, de modo a perceber qual a variável ou variáveis com mais peso no resultado final. No entanto, esta ação poderia prejudicar a qualidade da predição.

3 Resultados e Discussão

3.1 Modelo Geral

De modo a avaliar a qualidade das predições geradas a partir do modelo, foi utilizada a função *predict* da linguagem, sendo que esta função gera previsões a partir de um modelo de regressão logística. Tendo os valores reais no conjunto de dados de teste, compararam-se os resultados obtidos com estes, de modo a averiguar a exatidão do modelo.

Deste modo, pela visualização da Figura 10, pode-se ver que o modelo acertou 90.01% das vezes na predição, tendo uma sensibilidade de 65.73%, ou seja, foram detetados 65.73% valores verdadeiros corretamente. É de destacar o valor de especificidade elevado, sendo este representativo do número de falsos, ou seja, da categoria negativa, detetados corretamente como falsos.

```
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  6137  207
       yes  518  397

               Accuracy : 0.9001
                 95% CI : (0.893, 0.9069)
    No Information Rate : 0.9168
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4695

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.65728
            Specificity : 0.92216
         Pos Pred Value : 0.43388
         Neg Pred Value : 0.96737
             Prevalence : 0.08321
         Detection Rate : 0.05469
   Detection Prevalence : 0.12605
      Balanced Accuracy : 0.78972

       'Positive' Class : yes
```

Figura 10 - Matriz de confusão do modelo geral.

3.2 Modelo Significativo

Já no caso do modelo em que se retiraram as variáveis menos significativas, pela observação da Figura 11, é possível reparar que a exatidão do modelo aumentou ligeiramente para 90.04%. Além disso, a sensibilidade também incrementou para 66.38%, querendo isto dizer que este modelo foi capaz de prever mais valores positivos corretamente. No entanto, a métrica de especificidade diminui ligeiramente, mostrando que o modelo não prevê os registos que não aceitaram o depósito tão bem como o modelo base.

```
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  6147  197
       yes  526  389

               Accuracy : 0.9004
                 95% CI : (0.8933, 0.9072)
    No Information Rate : 0.9193
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4657

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.66382
            Specificity : 0.92117
         Pos Pred Value : 0.42514
         Neg Pred Value : 0.96895
             Prevalence : 0.08073
         Detection Rate : 0.05359
   Detection Prevalence : 0.12605
      Balanced Accuracy : 0.79250

       'Positive' Class : yes
```

Figura 11 - Matriz de confusão do modelo significativo.

3.3 Modelo Significativo - Sem *pdays*

Por fim, falta analisar as predições do modelo significativo sem a variável referente aos dias desde o último contacto. Desta forma, através da leitura da Figura 12, é possível constatar que a exatidão mantém-se comparativamente ao último modelo. A sensibilidade aumenta e a especifidade diminui, ambas com alterações bastante reduzidas.

```
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  6149  195
       yes  528  387

               Accuracy : 0.9004
                 95% CI : (0.8933, 0.9072)
    No Information Rate : 0.9198
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4646

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.66495
            Specificity : 0.92092
         Pos Pred Value : 0.42295
         Neg Pred Value : 0.96926
             Prevalence : 0.08018
         Detection Rate : 0.05331
   Detection Prevalence : 0.12605
      Balanced Accuracy : 0.79294

       'Positive' Class : yes
```

Figura 12 - Matriz de confusão do modelo significativo sem *pdays*.

3.4 *K-Fold Cross Validation*

De modo a confirmar e validar os modelos desenvolvidos, foi definido que seria, também, testado o erro médio em dados de teste, tal como feito anteriormente, mas desta vez com um método mais complexo.

O processo de *K-Fold Cross Validation* implica a divisão aleatória dos dados em K grupos de igual tamanho. Por cada iteração, é escolhido um grupo que será utilizado para testar o modelo, sendo que os restantes serão utilizados para o ajuste do mesmo. Finalizados estes testes, é calculado o erro médio dos grupos, sendo este o valor que será estudado.

No que toca ao primeiro modelo, o resultado depende de todos os preditores, como se pode ver pelo Extrato 1. Assim, aplicando a função *cv.glm* para $K = 10$ foi possível obter um erro médio de teste de 0.06979298.

```
model.kf <- glm(y~., data = BankMarketingComplete, family = binomial)
cv.err <- cv.glm(data = BankMarketingComplete,model.kf, K = 10)
cv.err$delta[1]
```

Extrato 1 - *10-Fold cross validation* do modelo geral.

No segundo modelo, ou seja, o modelo que apenas contém variáveis significativas a 95%, a chamada da função de *cross validation* (Extrato 2) retornou um erro médio de teste de 0.06984794, pouco maior do que o obtido no modelo completo.

```
model.kf2 <- glm(y ~ contact + month + day_of_week + duration + campaign + pdays +
    poutcome + emp.var.rate + cons.price.idx, data = BankMarketingComplete,family =
    binomial)
cv.err2 <- cv.glm(data = BankMarketingComplete,model.kf2, K = 10)
cv.err2$delta[1]
```

Extrato 2 - *10-Fold cross validation* do modelo significativo.

Por fim, o modelo significativo em que se retirou a variável *pdays* apresentou um erro de 0.06990671, ou seja, sensivelmente 7%. Como se pode observar pela comparação dos erros obtidos, os modelos são bastante próximos, pelo que as variáveis utilizadas neste modelo aparentam ser as que mais influência possuem no resultado final.

```
model.kf3 <- glm(y ~ contact + month + day_of_week + duration + campaign + poutcome +
    emp.var.rate + cons.price.idx, data = BankMarketingComplete,family = binomial)
cv.err3 <- cv.glm(data = BankMarketingComplete,model.kf3, K = 10)
cv.err3$delta[1]
```

Extrato 3 - *10-Fold cross validation* do modelo significativo sem *pdays*.

3.5 Resposta às Questões de Interesse

Tendo em conta os resultados obtidos, o modelo significativo sem *pdays* parece ser o que mais se adequa ao problema, tendo em conta a enorme redução de variáveis comparativamente ao modelo original e as métricas serem extremamente próximas às obtidas no modelo significativo, bem como os valores de erros médios de teste. Dito isto, as conclusões e respostas às questões de interesse serão efetuadas a partir deste, tendo-se as mesmas em seguida.

- Os contactos devem ser efetuados preferencialmente para os telemóveis dos clientes, em detrimento de contactos para telefone fixo.
- Devem ser publicitados depósitos nos meses de agosto, dezembro, julho, março, outubro e setembro, contrariamente aos meses de abril, maio, junho e novembro.
- Os dias da semana a efetuar contactos com os clientes devem ser, preferencialmente, terça, quarta e quinta. Devem ser evitados contactos às segundas e sextas.
- Quanto maior a duração da chamada, mais provável é a aceitação do cliente.
- Evitar contactar o cliente várias vezes durante a campanha, visto que o aumento neste valor implica uma diminuição na probabilidade de aceitação.
- Clientes sem registo referente às últimas campanhas têm maior probabilidade de aceitação. O mesmo acontece para clientes que aceitaram o depósito em situações anteriores. Tal não se verifica para clientes que recusaram propostas anteriores, sendo que possuem maior probabilidade de a voltar a recusar.
- Momentos em que a taxa de variação de emprego esteja baixa e o índice de preços no consumidor esteja elevado produzem maior chance de sucesso para o cliente aceitar o depósito.

A implementação destas medidas por parte da instituição permitiria uma redução nos custos, uma vez que seriam contactados menos clientes e nas alturas ideais. Desta forma, seria possível aumentar o número de depósitos, bem como a eficácia da campanha, trazendo um retorno financeiro maior ao banco.

4 Conclusões e Trabalho Futuro

Ao longo do presente relatório retrata-se o resultado do projeto prático da unidade curricular de Aprendizagem Automática I. Neste contexto, foram abordados métodos de aprendizagem estatística supervisionada para aproximação de um modelo a um certo conjunto de dados. Assim, tendo em conta os dados e o tipo da variável de resposta, o modelo implementado tratou-se de um modelo de regressão logística.

Um bom modelo está diretamente associado a um bom conjunto de dados, pelo que este tratamento se tratou do primeiro passo. Foi necessário remover registos com valores indefinidos, bem como categorizar através do agrupamento em semanas do tempo desde o último contacto ao cliente.

Tendo finalizado o tratamento dos dados, o objetivo passou por definir modelos, sendo que foram retiradas variáveis que não se apresentassem como significativas a 95%. Após chegar este modelo, foi, ainda, testada a importância da variável *pdays*, sendo que esta não mostrou afetar o resultado de forma elevada, tendo sido retirada para a definição de um novo modelo.

Deste modo, o passo final tratou-se de efetuar uma comparação dos modelos, de forma a ver qual permitia a previsão mais próxima da realidade, bem como uma maior significância das variáveis. Assim, o modelo final, ou seja, o modelo que continha apenas variáveis significativas sem o preditor *pdays*, permitiu responder às questões de interesse previamente definidas.

As variáveis que mais influenciam a decisão do cliente são os preditores do modelo, ou seja, os preditores *contact, month, day_of_week, duration, campaign, poutcome, emp.var.rate* e *cons.price.idx*. As estratégias que devem ser tomadas para aumentar as chances de sucesso devem passar por contactar os clientes através do seu telemóvel, publicitar depósitos nos meses de março, agosto, dezembro, julho, outubro e setembro, efetuar contactos às terças, quartas e quintas, tentar prolongar a chamada o máximo possível, entre outro métodos. Por fim, os clientes que devem ser contactados devem ser clientes sem registo referente às últimas campanhas e clientes que aceitaram o depósito em campanhas anteriores. Note-se que se deve evitar o contacto de clientes várias vezes na mesma campanha, pois este fator prejudica as hipóteses de sucesso da mesma.

Em suma, a realização deste trabalho exigiu a aplicação de todos os conhecimentos lecionados em contexto de aula, permitindo que o grupo cumprisse todos os objetivos propostos no enunciado, conciliando assim a teoria e a prática, bem como compreender a importância de aproximação de modelos a dados reais de modo a compreender os últimos.

Referências

[1] Bank Marketing, Kaggle,
https://www.kaggle.com/henriqueyamahata/bank-marketing