



Universidade do Minho  
Escola de Engenharia

# Previsão e Otimização de Produtos Bancários

Mestrado Integrado em Engenharia e  
Gestão de Sistemas de Informação  
2016/2017

## Sistemas Adaptativos para a Inteligência do Negócio

Docente: Prof. Dr. Paulo Alexandre Ribeiro  
Cortez

## Constituição da Equipa



Pamela Coelho 71179  
a71179@alunos.uminho.pt



Ricardo Araújo 71206  
a71206@alunos.uminho.pt



João Gomes 72357  
a72357@alunos.uminho.pt



David Vila Nova 71782  
a71782@alunos.uminho.pt

## Índice

Introdução.....	5
Metodologia Crisp-DM.....	6
1. Estudo do Negócio .....	6
1.1. Determinar os Objetivos do Negócio .....	6
1.2. Avaliação da situação atual .....	7
1.3. Determinar objetivos de data mining .....	9
1.4. Produzir o plano do projeto .....	9
2. Estudo dos dados .....	10
2.1. Recolha inicial dos dados .....	10
2.2. Descrição dos dados.....	11
2.3. Exploração dos dados.....	12
2.4. Verificação da qualidade dos dados.....	13
3. Preparação dos dados .....	14
3.1. Seleção dos dados .....	14
3.2. Limpeza dos dados .....	15
3.3. Construção dos dados .....	15
4. Modelação.....	16
4.1. Seleção das técnicas de modelação .....	16
4.2. Criação de um cenário teste .....	16
4.3. Construção do modelo .....	17
4.4. Avaliação dos modelos.....	17
5. Avaliação .....	19
5.1. Avaliar resultados.....	19
5.2. Revisão do processo .....	19
5.3. Determinação dos próximos passos.....	20
Otimização.....	21
Representação da solução .....	21
Função de Avaliação.....	21
Modelos de otimização utilizados e cenários escolhidos .....	23
Avaliação de resultados .....	24
Sistema construído em Shiny .....	25
Conclusão .....	26
Bibliografia .....	27
Anexos.....	28



---

Atas.....	28
Avaliação .....	37

## Índice de figuras

Figura 1 - OBS .....	7
Figura 2 - Diagrama de Gantt (print1) .....	9
Figura 3 - Diagrama de Gantt (print2) .....	10
Figura 4 - Diagrama de Gantt (print3) .....	10
Figura 5 - Sistema Shiny .....	25

## Índice de tabelas

Tabela 1 - Tabela de riscos .....	8
Tabela 2 - Ferramentas .....	10
Tabela 3 - Descrição dos dados iniciais .....	11
Tabela 4 - Descrição dos dados tratados .....	14
Tabela 5 - Tabela de comparação entre modelos comparativos .....	18
Tabela 6 - Comparação entre modelos de otimização.....	24

## Introdução

No âmbito da unidade curricular de Sistemas Adaptativos para a Inteligência do Negócio, do 4º ano de Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação, foi-nos proposto um projeto onde se pretende aplicar modelos de previsão e realizar a sua posterior otimização.

Na previsão classificativa serão testados sete modelos preditivos em oito cenários diferentes. No total serão criadas 56 previsões diferentes e, através de métricas definidas, será possível escolher o modelo com melhores resultados. Cada elemento do grupo ficará responsável por, no mínimo, um modelo preditivo. No final, os modelos serão organizados e manipulados de forma conveniente para o seu uso na otimização.

Na otimização serão testados quatro modelos: um modelo de Blind Search, dois modelos de Local Search e por último, um modelo de Population Based Search. Cada um será testado em quatro cenários: um cenário sem restrições que utiliza *death penalty*; outro que utiliza *death penalty* na restrição; um terceiro sem restrições e que utiliza o *repair* para realizar correções nas soluções apresentadas; e, por último, um cenário que utiliza *repair* na restrição.

Por fim, será criada uma aplicação no Shiny, que permite, ao utilizador, escolher qual o modelo de previsão que pretende testar, entre outros casos. Com isto, o projeto tornar-se-á um pouco mais dinâmico e permitirá apresentar um output mais intuitivo.

Para a elaboração do projeto, o grupo recorreu ao livro do docente Paulo Cortez, aos slides disponibilizados pelo mesmo e aos exemplos fornecidos nas aulas.

## Metodologia Crisp-DM

Será utilizada a metodologia Crisp-DM, para a execução do projeto. Esta metodologia apresenta o ciclo de vida de um projeto de data mining e consiste nas seguintes seis fases (baseadas no manual específico da metodologia):

Estudo do negócio: Esta fase consiste na abordagem do âmbito do projeto, segundo a perspetiva de negócio. Permite perceber o porquê da aplicação do data mining, quais os objetivos do negócio, bem como a evolução do projeto ao longo do tempo do projeto para a sua execução.

Estudo dos dados: Para ser perceptível o tipo de previsão e otimização a efetuar, há que compreender os inputs que serão recebidos nos modelos. Para tal, esta fase permite analisar o conjunto de dados fornecidos, de modo a perceber quais as alterações a efetuar, se necessário, quais os dados que influenciarão a target, entre outros. Ou seja, serve esta fase para levantamento dos aspetos positivos e negativos dos dados existentes.

Preparação dos dados: De acordo com o registo efetuado na etapa anterior, os dados serão organizados e preparados para a respetiva utilização na previsão e otimização do projeto. Nesta fase efetuar-se-á o tratamento necessário ao conjunto de dados, obtendo-se o dataset final.

Modelação: As técnicas de modelação são selecionadas e aplicadas e os parâmetros adaptados. As técnicas possuem requisitos específicos para o formato de dados existentes e, por isso, pode ser necessário voltar à etapa anterior.

Avaliação: Com base nos modelos escolhidos, várias previsões serão obtidas. Para perceber qual o modelo com melhor previsão, serão aplicadas medidas de análise, de forma a comparar os modelos entre si.

Implementação: Esta fase consiste na organização e apresentação do conhecimento, de forma a que o cliente possa utilizar o output a seu favor.

### 1. Estudo do Negócio

#### 1.1. Determinar os Objetivos do Negócio

No âmbito da atribuição de créditos, por parte de uma instituição bancária, pretende-se prever, tendo em conta um conjunto de dados, quais os clientes que, no próximo mês, estarão em incumprimento no retorno dos créditos. Com base nesta previsão, será possível compreender

o leque de clientes favoráveis à apresentação dos produtos financeiros existentes, otimizando o negócio da instituição.

## 1.2. Avaliação da situação atual

A equipa que desenvolverá o projeto é constituída por quatro elementos, tal como se pode ver na seguinte OBS (*Organization Breakdown Structure*):

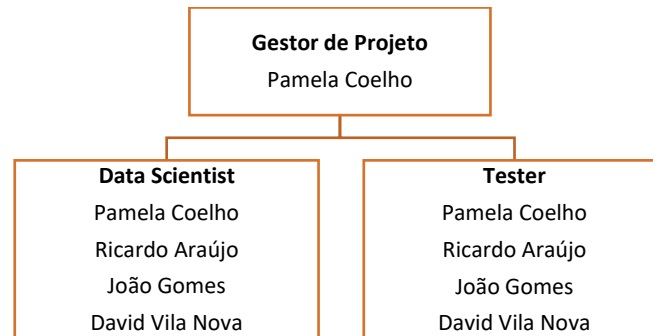


Figura 1 - OBS

Tendo em conta a unidade curricular de Sistemas para a Inteligência do Negócio e Organização, estes possuem conhecimentos, na área do data mining, que serão aplicados no decorrer do trabalho. Cada elemento possui um computador portátil onde executará as tarefas que lhe forem assinaladas. De forma a atingir os objetivos definidos, a equipa debroçar-se-á nos dados fornecidos pelo docente. As ferramentas a utilizar encontram-se descritas no ponto 1.4 do Estudo do Negócio.

Os requisitos para o projeto serão a utilização da metodologia Crisp-DM; a revisão semanal, com o docente, das tarefas realizadas; uso de mais do que um modelo de data mining; utilização de métricas e métodos de avaliação para mensuração da qualidade dos modelos.

A equipa pressupõe que os dados são reais e que descrevem a situação do negócio na sua totalidade.

O facto de existirem outras unidades curriculares, o limite de tempo para conclusão do projeto, o número de elementos e as suas competências, bem como o tempo de execução de certos modelos, são algumas das restrições ao projeto.

A lista de riscos do projeto pode ser visualizada na seguinte tabela:

Tabela 1 - Tabela de riscos

Descrição	Prob [1-5]	Imp [1-5]	Seriedade [PxI]	Ação Atenuante [AA]	Prob [AA]	Imp [AA]	Seriedade [AA]
Complexidade do Projeto	3	5	15	Analisar melhor o enunciado; Esclarecer potenciais dúvidas com o docente; Dedicar mais tempo à compreensão do projeto.	2	5	10
Dificuldades na análise dos dados	3	4	12	Reunir a equipa e resolver problemas.	2	4	8
Dificuldades na utilização das ferramentas	3	4	12	Pesquisar e efetuar tutoriais sobre as mesmas.	2	3	6
Atraso na execução do plano do projeto	2	5	10	Gerir melhor a execução das tarefas, para cumprir prazos.	1	4	4
Má gestão da equipa de trabalho	2	4	8	Rever plano e reunir com a equipa de forma a solucionar o problema.	1	4	4
Má interpretação dos objetivos do trabalho	2	4	8	Reanalisar enunciado em equipa e esclarecer dúvidas com o docente.	1	3	3
Sobrecarga de outras unidades curriculares	2	4	8	Reajustar as tarefas atribuídas, consoante a carga de trabalho dos elementos.	1	3	3
Falhas de comunicação na equipa	2	3	6	Agendar reuniões que possibilitem a partilha de opiniões; Estimular o trabalho de equipa.	1	3	3
Perda de informação	1	5	5	Efetuar backups do projeto; Garantir a utilização de	1	3	3



ferramentas de  
armazenamento.

Sendo este um projeto académico, os elementos não são remunerados pelo seu trabalho. Para além das três horas presenciais, cada elemento do grupo terá de dispensar cinco horas semanais para o desenvolvimento do projeto. No final, o projeto será avaliado pelo docente e cada elemento terá uma nota consoante o trabalho que efetuou. O projeto desenvolvido será fundamental para que a instituição bancária consiga otimizar o seu negócio.

### 1.3. Determinar objetivos de data mining

O principal objetivo de data mining consiste na utilização de modelos que atuam sobre um conjunto de dados e irão trazer informações relevantes para o negócio, como, por exemplo, previsões e otimizações.

A previsão terá sucesso se a métrica AUC for superior a 75% e possa ser escolhido um modelo preditivo, que apresente estes resultados, para uso na otimização.

### 1.4. Produzir o plano do projeto

Para elaborar o plano do projeto foram registadas atas, aquando de cada reunião do grupo. Estas podem ser visualizadas em anexo. Desta forma, segue abaixo o diagrama de Gantt do projeto de Sistemas Adaptativos para a Inteligência de Negócio.

1	Projeto SAIN	64,5 hrs	Sex 24/03/17	Sex 02/06/17	
2	Análise do enunciado	4 hrs	Sex 24/03/17	Qua 29/03/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
3	Previsão	34,5 hrs	Sex 31/03/17	Qua 03/05/17	
4	Análise dos modelos de previsão classificativa	6 hrs	Sex 31/03/17	Sex 07/04/17	2 David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
5	Estudo do negócio	0,5 hrs	Sex 31/03/17	Sex 31/03/17	2
6	Determinação dos objetivos do negócio	0,5 hrs	Sex 31/03/17	Sex 31/03/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
7	Descrição da situação atual	0,5 hrs	Sex 31/03/17	Sex 31/03/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
8	Determinação dos objetivos de Data Mining	0,5 hrs	Sex 31/03/17	Sex 31/03/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
9	Plano do projeto	0,5 hrs	Sex 31/03/17	Sex 31/03/17	Pamela Coelho
10	Estudo dos dados	10 hrs	Sex 31/03/17	Qua 19/04/17	5
11	Recolha de dados inicial	1 hr	Sex 31/03/17	Sex 31/03/17	
12	Descrição dos dados	10 hrs	Sex 31/03/17	Qua 19/04/17	Pamela Coelho
13	Exploração dos dados	10 hrs	Sex 31/03/17	Qua 19/04/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
14	Verificação da qualidade dos dados	10 hrs	Sex 31/03/17	Qua 19/04/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
15	Preparação dos dados	8 hrs	Qua 19/04/17	Ter 25/04/17	10
16	Seleção dos dados	1,5 hrs	Qua 19/04/17	Qua 19/04/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
17	Limpeza dos dados	8 hrs	Qua 19/04/17	Ter 25/04/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
18	Construção dos dados	1,5 hrs	Qua 19/04/17	Qua 19/04/17	
19	Integração dos dados	1,5 hrs	Qua 19/04/17	Qua 19/04/17	

Figura 2 - Diagrama de Gantt (print1)



19	Integração dos dados	1,5 hrs	Qua 19/04/17	Qua 19/04/17	
20	Formatação dos dados	8 hrs	Qua 19/04/17	Ter 25/04/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
21	Modelação	10 hrs	Ter 25/04/17	Qua 26/04/17	15
22	Seleção das técnicas de modelação	5 hrs	Ter 25/04/17	Ter 25/04/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
23	Criação de uma conceção de teste	10 hrs	Ter 25/04/17	Qua 26/04/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
24	Construção do modelo	10 hrs	Ter 25/04/17	Qua 26/04/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
25	Revisão do modelo	10 hrs	Ter 25/04/17	Qua 26/04/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
26	Avaliação	6 hrs	Qua 26/04/17	Qua 03/05/17	21
27	Avaliação dos resultados	6 hrs	Qua 26/04/17	Qua 03/05/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
28	Revisão do processo	6 hrs	Qua 26/04/17	Qua 03/05/17	
29	Determinação dos passos seguintes	6 hrs	Qua 26/04/17	Qua 03/05/17	
30	Otimização	26 hrs	Qua 03/05/17	Sex 02/06/17	3
31	Análise de modelos de otimização	6 hrs	Qua 03/05/17	Sex 05/05/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
32	Escolha da representação da solução	6 hrs	Qua 03/05/17	Sex 05/05/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
33	Criação da função de avaliação	8 hrs	Qua 03/05/17	Qua 10/05/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
34	Aplicação dos modelos	20 hrs	Qua 03/05/17	Qua 31/05/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
35	Construção do sistema em R	26 hrs	Qua 03/05/17	Sex 02/06/17	David Carvalho;Ricardo Araújo
36	Avaliação de resultados	4 hrs	Qua 03/05/17	Sex 05/05/17	David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo

Figura 3 - Diagrama de Gantt (print2)

37	Sistema em Shiny	16 hrs	Sex 02/06/17	Sex 16/06/17	30
38	Exploração do Shiny	16 hrs	Sex 02/06/17	Sex 16/06/17	João Gomes
39	Criação do sistema	6 hrs	Sex 02/06/17	Sex 09/06/17	João Gomes;Ricardo Araújo
40	Produção do relatório	28 hrs	Sex 31/03/17	Qua 26/04/17	2 Pamela Coelho
41	Revisão do relatório	4 hrs	Qua 26/04/17	Sex 28/04/17	40 David Carvalho;João Gomes;Pamela Coelho;Ricardo Araújo
42	Entrega do projeto	0 hrs	Sex 02/06/17	Sex 02/06/17	

Figura 4 - Diagrama de Gantt (print3)

Para a evolução do projeto foram utilizadas as seguintes ferramentas:

Tabela 2 - Ferramentas

Ferramenta	Justificação de uso
R Studio	Tratamento de dados; Desenvolvimento da previsão classificativa; Elaboração do sistema de otimização.
OneDrive for Business	Compartilhamento de ficheiros entre os elementos do grupo.
MS Project 2013	Elaboração do diagrama de Gantt.
Word 2016	Elaboração do relatório do projeto.
Facebook	Contacto entre os elementos do grupo.

## 2. Estudo dos dados

### 2.1. Recolha inicial dos dados

O dataset inicial foi fornecido pelo docente Paulo Cortez, mas pode ser adquirido no site UCI – Machine Learning Repository (<http://archive.ics.uci.edu/ml/>). Os dados foram disponibilizados, no site, pelo Prof. I-Cheng Yeh do Departamento de Engenharia Civil da Universidade de Tamkang, Taiwan. Este dataset diz respeito ao registo dos reembolsos, a uma entidade bancária, de créditos



atribuídos a 5000 clientes. Os reembolsos apresentados dizem respeito aos meses de Abril a Setembro de 2005. Caso o cliente não cumpra com o reembolso mensal, este estará em incumprimento para com a instituição bancária.

## 2.2. Descrição dos dados

Antes da preparação dos dados, estes têm de ser compreendidos pela equipa. O dataset recebido contém 5000 linhas, respetivas a 5000 clientes da instituição bancária. Para melhor compreensão do tipo de dados presentes no dataset, segue a seguinte tabela:

*Tabela 3 - Descrição dos dados iniciais*

Atributos	Descrição	Tipo dos dados
id	Número que identifica a linha (cliente) do dataset (1-5000)	Numérico
X1_LIMIT_BAL	Valor de crédito atribuído ao cliente (individual e crédito suplementar) – NT DOLLAR	Numérico
X2_SEX	Género do indivíduo (2 - feminino ou 1 - masculino)	Numérico
X3_EDUCATION	Grau de educação do(a) cliente (1 - Pós-graduação, 2 - Universidade, 3 - Secundário e 4 - Outros)	Numérico
X4_MARRIAGE	Estado civil do(a) cliente (1 - Casado(a), 2 - Solteiro(a), 3 - Divorciado(a), 0 - Outro)	Numérico
X5_AGE	Idade, em anos, do(a) cliente	Numérico
X6_PAY_0	Histórico do estado dos pagamentos dos créditos, à instituição, por parte dos clientes em Setembro	Numérico
X7_PAY_2	(...) em Agosto	Numérico
X8_PAY_3	(...) em Julho	Numérico
X9_PAY_4	(...) em Junho	Numérico



X10_PAY_5	(...) em Maio	Numérico
X11_PAY_6	(...) em Abril	Numérico
X12_BILL_AMT1	Valor/Montante da fatura (NT DOLLAR) em Setembro	Numérico
X13_BILL_AMT2	(...) em Agosto	Numérico
X14_BILL_AMT3	(...) em Julho	Numérico
X15_BILL_AMT4	(...) em Junho	Numérico
X16_BILL_AMT5	(...) em Maio	Numérico
X17_BILL_AMT6	(...) em Abril	Numérico
X18_PAY_AMT1	Valor/Montante pago em Setembro	Numérico
X19_PAY_AMT2	(...) em Agosto	Numérico
X20_PAY_AMT3	(...) em Julho	Numérico
X21_PAY_AMT4	(...) em Junho	Numérico
X22_PAY_AMT5	(...) em Maio	Numérico
X23_PAY_AMT6	(...) em Abril	Numérico
Y	Variável de estudo: Cumprimento (0) ou Incumprimento(1), por parte do(a) cliente, para com o pagamento do crédito à instituição bancária	Numérico

### 2.3. Exploração dos dados

A análise ao dataset foi efetuada na ferramenta R. Como tal, para serem perceptíveis os intervalos de valores assumidos pelas 23 variáveis existentes, foi lido o dataset e executada a função *summary()*.

`summary(d)`

```
##      id      X1_LIMIT_BAL      X2_SEX      X3_EDUCATION
## Min.   : 1    Min.   : 10000    Min.   :1.000    Min.   :1.000
## 1st Qu.:1251  1st Qu.: 50000    1st Qu.:1.000    1st Qu.:1.000
## Median :2500  Median : 140000    Median :2.000    Median :2.000
## Mean   :2500  Mean   : 165640    Mean   :1.573    Mean   :1.811
## 3rd Qu.:3750  3rd Qu.: 230000    3rd Qu.:2.000    3rd Qu.:2.000
## Max.   :5000  Max.   :1000000    Max.   :2.000    Max.   :4.000
## X4_MARRIAGE  X5_AGE      X6_PAY_0      X7_PAY_2
## Min.   :1.000  Min.   :21.00    Min.   :-2.000    Min.   :-2.0000
## 1st Qu.:1.000  1st Qu.:28.00    1st Qu.: -1.000    1st Qu.: -1.0000
```



```
## Median :2.000 Median :34.00 Median : 0.000 Median : 0.0000
## Mean :1.572 Mean :35.37 Mean : 0.005 Mean : -0.1484
## 3rd Qu.:2.000 3rd Qu.:41.00 3rd Qu.: 0.000 3rd Qu.: 0.0000
## Max. :3.000 Max. :75.00 Max. : 8.000 Max. : 7.0000
## X8_PAY_3 X9_PAY_4 X10_PAY_5 X11_PAY_6
## Min. : -2.0000 Min. : -2.0000 Min. : -2.000 Min. : -2.0000
## 1st Qu.: -1.0000 1st Qu.: -1.0000 1st Qu.: -1.000 1st Qu.: -1.0000
## Median : 0.0000 Median : 0.0000 Median : 0.000 Median : 0.0000
## Mean : -0.1594 Mean : -0.2452 Mean : -0.261 Mean : -0.2892
## 3rd Qu.: 0.0000 3rd Qu.: 0.0000 3rd Qu.: 0.000 3rd Qu.: 0.0000
## Max. : 7.0000 Max. : 7.0000 Max. : 7.000 Max. : 8.0000
## X12_BILL_AMT1 X13_BILL_AMT2 X14_BILL_AMT3 X15_BILL_AMT4
## Min. : -14386 Min. : -30000 Min. : -15000 Min. : -170000
## 1st Qu.: 3208 1st Qu.: 2941 1st Qu.: 2436 1st Qu.: 1799
## Median : 21321 Median : 20579 Median : 19532 Median : 17988
## Mean : 50216 Mean : 48281 Mean : 45364 Mean : 40788
## 3rd Qu.: 62634 3rd Qu.: 60543 3rd Qu.: 56638 3rd Qu.: 49304
## Max. :964511 Max. :983931 Max. :578971 Max. : 891586
## X16_BILL_AMT5 X17_BILL_AMT6 X18_PAY_AMT1 X19_PAY_AMT2
## Min. : -28335 Min. : -339603.0 Min. : 0 Min. : 0.0
## 1st Qu.: 1491 1st Qu.: 967.5 1st Qu.: 1000 1st Qu.: 639.5
## Median : 17352 Median : 15859.0 Median : 2113 Median : 2000.0
## Mean : 39544 Mean : 37963.6 Mean : 5577 Mean : 5431.3
## 3rd Qu.: 49160 3rd Qu.: 48026.5 3rd Qu.: 5001 3rd Qu.: 5000.0
## Max. :927171 Max. : 961664.0 Max. :368199 Max. :344261.0
## X20_PAY_AMT3 X21_PAY_AMT4 X22_PAY_AMT5 X23_PAY_AMT6
## Min. : 0.0 Min. : 0.0 Min. : 0.0 Min. : 0
## 1st Qu.: 222.8 1st Qu.: 229.8 1st Qu.: 208.5 1st Qu.: 0
## Median : 1403.0 Median : 1500.0 Median : 1500.0 Median : 1316
## Mean : 4604.0 Mean : 4740.2 Mean : 4758.8 Mean : 5272
## 3rd Qu.: 4000.0 3rd Qu.: 4000.0 3rd Qu.: 4000.0 3rd Qu.: 4000
## Max. :896040.0 Max. :497000.0 Max. :332000.0 Max. :528666
## y
## Min. :0.0000
## 1st Qu.:0.0000
## Median :0.0000
## Mean :0.2214
## 3rd Qu.:0.0000
## Max. :1.0000
```

## 2.4. Verificação da qualidade dos dados

Durante a análise do dataset foram encontrados valores impossíveis de justificar, como, por exemplo, valores a NULL, ou valores a negativo. Assim sendo, o docente aconselhou o grupo a continuar o projeto sem limpar dados e a supor que a qualidade dos mesmos era boa.

### 3. Preparação dos dados

#### 3.1. Seleção dos dados

De acordo com a análise dos dados, apenas a variável “id” foi considerada inútil para a elaboração das previsões. Não foram excluídas mais variáveis, mas estas tiveram de ser tratadas. Assim sendo, segue a seguinte tabela com os dados selecionados.

*Tabela 4 - Descrição dos dados tratados*

Atributos	Descrição	Tipo dos dados
X1_LIMIT_BAL	Valor de crédito atribuído ao cliente (individual e crédito suplementar) – NT DOLLAR	Numérico
X2_SEX	Género do indivíduo (2 - feminino ou 1 - masculino)	Factor
X3_EDUCATION	Grau de educação do(a) cliente (1 - Pós-graduação, 2 - Universidade, 3 - Secundário e 4 - Outros)	Factor
X4_MARRIAGE	Estado civil do(a) cliente (1 - Casado(a), 2 - Solteiro(a), 3 - Divorciado(a), 0 - Outro)	Factor
X5_AGE	Idade, em anos, do(a) cliente (21-30, 31-40, 41-75)	Factor
X6_PAY_0	Histórico do estado dos pagamentos dos créditos, à instituição, por parte dos clientes (Normal, Em falta, Muito grave) em Setembro	Factor
X7_PAY_2	(...) em Agosto	Factor
X8_PAY_3	(...) em Julho	Factor
X9_PAY_4	(...) em Junho	Factor
X10_PAY_5	(...) em Maio	Factor
X11_PAY_6	(...) em Abril	Factor



X12_BILL_AMT1	Valor/Montante da fatura (NT DOLLAR) em Setembro	Numérico
X13_BILL_AMT2	(...) em Agosto	Numérico
X14_BILL_AMT3	(...) em Julho	Numérico
X15_BILL_AMT4	(...) em Junho	Numérico
X16_BILL_AMT5	(...) em Maio	Numérico
X17_BILL_AMT6	(...) em Abril	Numérico
X18_PAY_AMT1	Valor/Montante pago em Setembro	Numérico
X19_PAY_AMT2	(...) em Agosto	Numérico
X20_PAY_AMT3	(...) em Julho	Numérico
X21_PAY_AMT4	(...) em Junho	Numérico
X22_PAY_AMT5	(...) em Maio	Numérico
X23_PAY_AMT6	(...) em Abril	Numérico
Y	Variável de estudo: Cumprimento (0) ou Incumprimento(1), por parte do(a) cliente, para com o pagamento do crédito à instituição bancária	Factor

### 3.2. Limpeza dos dados

Na análise dos dados foram verificados valores estranhos como, por exemplo, valores negativos em algumas variáveis. De forma a acelerar a evolução do projeto, o docente aconselhou o grupo a ignorar as alterações e a continuar o trabalho com os dados originais, supondo que não haviam valores inexplicáveis nos mesmos. Posteriormente, se o grupo tivesse tempo de manobra para testes, poderia voltar a analisar os dados e, só aí, corrigi-los.

### 3.3. Construção dos dados

De acordo com o dataset fornecido e tendo em conta que os dados existentes são suficientes para a elaboração da previsão, não houve necessidade de serem criados novos dados de suporte ao projeto.

## 4. Modelação

### 4.1. Seleção das técnicas de modelação

A previsão em estudo é uma previsão classificativa. Deste modo, o grupo teve de explorar os modelos de previsão classificativa existentes, de forma a obter o modelo com melhores resultados para aplicar à otimização. Assim sendo, os modelos escolhidos foram o Random Forest, o KSVM, o MLpe,, MLP, Regressão Logística, Árvores de Decisão e Naive Bayes.

### 4.2. Criação de um cenário teste

Aquando da execução dos modelos foram instalados os packages “rminer”, “randomForest”, “DMwR”, “nnet”, “SMOTE”, “rpart.plot” e “e1071”.

Para a previsão classificativa, os dados, após tratamento, foram divididos em dados de treino(70%) e dados de teste(30%) recorrendo à função *holdout()*. Neste caso houve criação de dois cenários: um *order holdout*, onde os dados de treino permaneciam sempre os mesmos a cada iteração; outro designado *rolling holdout*, com janela de tamanho 3000 (dados de treino), e 200 dados de teste, e a cada iteração o *increment* era de 60. Após criação destes dois primeiros cenários, os dados de treino tiveram de ser balanceados, para que houvesse o mesmo número de clientes em cumprimento e incumprimento. Sendo assim, em cada um dos dois cenários anteriores, foram criados novos quatro cenários: dados de treino não balanceados, dados de treino balanceados através da função *SMOTE()*, dados balanceados por *Oversampling()* e dados balanceados por *Undersampling()*.

Seguidamente, e tendo em conta os oito cenários criados anteriormente, os sete modelos escolhidos foram aplicados aos dados de treino. No total elaboraram-se 56 cenários para comparação de resultados. Após esta fase de aprendizagem, os modelos treinados foram aplicados aos dados de teste, de forma a ser obtida uma nova previsão.

Assim sendo, às 56 previsões foram aplicadas seis métricas de avaliação: ACC, AUC, PRECISION, CE, TPR e TNR. Para além disso, foram construídos 56 gráficos com a curva de ROC, para comparação de modelos.



#### 4.3. Construção do modelo

Um modelo de previsão permite, através de um conjunto de dados de aprendizagem (designados, ocasionalmente, por dados de treino), aplicar essa aprendizagem a um novo conjunto de dados (dados de teste), de forma a serem previstos valores de uma determinada target. Assim sendo, os sete modelos foram aplicados aos oito cenários e obtiveram-se 56 previsões diferentes.

Tal como referido anteriormente, um conjunto de dados de treino (70% dos dados originais) é criado, para que seja possível treinar um determinado modelo. Seguidamente, o modelo treinado é aplicado nos dados de teste (restantes 30% dos dados originais), obtendo-se o resultado da função *predict()*. Para ser perceptível a modelação, segue um exemplo da aplicação de um modelo preditivo:

```
glm_rolling_smote <- function(dataset) {  
  reais <- character()  
  previsoes <- numeric()  
  iteracoes <- 30  
  for (i in 1:iteracoes){  
    holdout <- holdout(  
      dataset$Y,  
      ratio = 200,  
      mode = "rolling",  
      iter = i,  
      window = 3000,  
      increment = 60)  
    dados_treino <- dataset[holdout$tr, ]  
    dados_teste <- dataset[holdout$ts, ]  
    s <- SMOTE(Y ~ ., dados_treino, perc.over = 250, k = 10, perc.under = 150)  
    glm <- glm(Y ~ ., family = binomial(link = "logit"), data = s)  
    previsoes <- c(previsoes, predict.glm(glm, dados_teste, type = "response"))  
    reais <- c(reais, as.character(dados_teste$Y))  
  }  
  reais <- factor(reais)  
  metrics_glm(reais, previsoes, "GLM Rolling SMOTE")  
}
```

#### 4.4. Avaliação dos modelos

Para avaliar os modelos de previsão escolhidos foram utilizadas seis métricas: ACC, PRECISION, CE, TNR, TPR e AUC. De forma a escolher o modelo de previsão adequado à



otimização do projeto, a métrica AUC foi a escolhida para o efeito. Assim sendo, na seguinte tabela, os modelos encontram-se ordenados por ordem decrescente, de valores, da métrica AUC.

Tabela 5 - Tabela de comparação entre modelos comparativos

	ACC	PRECISION	CE	TNR1	TNR2	TPR1	AUC
<b>glm_rolling_smote</b>	78,05	84,94	21,95	44,71	87,39	87,39	<b>0,750307</b>
glm_smote	78,87	85,56	21,13	46,63	87,82	87,82	0,750034
random_forest_rolling_undersampling	70,70	88,41	29,30	66,34	71,92	71,92	0,748719
random_forest_rolling_smote	76,98	86,36	23,02	52,78	83,76	83,76	0,747457
random_forest_rolling_oversampling	78,85	84,39	21,15	40,90	89,48	89,48	0,745650
ksvm_oversampling	75,93	86,72	24,07	54,91	81,77	81,77	0,742676
glm_rolling_desbalanceados	79,15	81,72	20,85	24,60	94,43	94,43	0,741758
glm_rolling_oversampling	76,95	86,58	23,05	53,85	83,42	83,42	0,741621
random_forest_rolling_desbalanceados	79,92	82,54	20,08	28,87	94,22	94,22	0,740745
random_forest_undersampling	70,07	87,96	29,93	64,72	71,55	71,55	0,738799
glm_rolling_undersampling	75,72	86,99	24,28	56,74	81,03	81,03	0,738009
ksvm_smote	77,33	86,26	22,67	51,53	84,50	84,50	0,737409
glm_desbalanceados	80,00	82,27	20,00	26,38	94,89	94,89	0,737349
naive_bayes_oversampling	34,73	90,79	65,27	93,25	18,48	18,48	0,736881
naive_bayes_desbalanceados	48,53	88,65	51,47	81,90	39,27	39,27	0,736829
glm_oversampling	76,40	86,80	23,60	54,91	82,37	82,37	0,736604
random_forest_smote	77,47	86,28	22,53	51,53	84,67	84,67	0,735683
ksvm_rolling_oversampling	75,92	86,44	24,08	54,07	82,04	82,04	0,731698
random_forest_desbalanceados	80,07	82,72	19,93	29,14	94,21	94,21	0,731417
random_forest_oversampling	78,87	84,58	21,13	41,41	89,27	89,27	0,731195
mlpe_rolling_smote	75,60	85,98	24,40	52,17	82,16	82,16	0,730383
ksvm_rolling_undersampling	74,02	86,83	25,98	57,43	78,66	78,66	0,729013
mlpe_smote	76,47	86,62	23,53	53,99	82,71	82,71	0,726983
ksvm_undersampling	74,73	86,91	25,27	56,75	79,73	79,73	0,725954
ksvm_rolling_smote	76,45	85,80	23,55	50,57	83,70	83,70	0,724365
mlpe_undersampling	66,87	87,65	33,13	65,95	67,12	67,12	0,722338
mlpe_desbalanceados	78,67	82,75	21,33	30,98	91,91	91,91	0,722113
decision_tree_rolling_undersampling	71,55	87,87	28,45	63,67	73,76	73,76	0,721895
decision_tree_smote	78,13	86,85	21,87	53,68	84,92	84,92	0,720596
naive_bayes_rolling_desbalanceados	46,88	89,27	53,12	84,39	36,38	36,38	0,720505
naive_bayes_rolling_oversampling	37,18	90,91	62,82	92,23	21,76	21,76	0,720448
glm_undersampling	75,87	87,32	24,13	57,67	80,92	80,92	0,720391
naive_bayes_undersampling	39,80	89,50	60,20	88,96	26,15	26,15	0,720161
decision_tree_undersampling	73,47	87,38	26,53	59,82	77,26	77,26	0,717241
mlpe_oversampling	74,00	85,83	26,00	52,45	79,98	79,98	0,716913
mlpe_rolling_oversampling	72,80	85,37	27,20	51,87	78,66	78,66	0,714716
decision_tree_rolling_smote	78,33	86,09	21,67	50,27	86,20	86,20	0,713139



<b>ksvm_desbalanceados</b>	79,33	80,99	20,67	18,71	96,17	96,17	0,711335
<b>mlpe_rolling_desbalanceados</b>	78,77	82,65	21,23	30,92	92,17	92,17	0,709485
<b>naive_bayes_smote</b>	42,27	88,89	57,73	86,50	29,98	29,98	0,709294
<b>mlp_smote</b>	71,87	85,27	28,13	51,84	77,43	77,43	0,708074
<b>naive_bayes_rolling_undersampling</b>	42,92	89,59	57,08	87,36	30,47	30,47	0,704934
<b>mlp_rolling_smote</b>	72,75	85,59	27,25	52,93	78,30	78,30	0,702446
<b>naive_bayes_rolling_smote</b>	41,05	87,43	58,95	85,30	28,65	28,65	0,700811
<b>mlp_undersampling</b>	59,47	87,04	40,53	69,63	56,64	56,64	0,696156
<b>mlpe_rolling_undersampling</b>	65,78	86,32	34,22	62,22	66,78	66,78	0,694865
<b>decision_tree_rolling_oversampling</b>	75,72	86,67	24,28	55,29	81,44	81,44	0,694797
<b>ksvm_rolling_desbalanceados</b>	79,45	81,16	20,55	20,49	95,97	95,97	0,692422
<b>mlp_rolling_oversampling</b>	70,23	84,94	29,77	52,40	75,23	75,23	0,691128
<b>mlp_desbalanceados</b>	76,27	82,20	23,73	30,67	88,93	88,93	0,686925
<b>decision_tree_rolling_desbalanceados</b>	79,93	83,14	20,07	32,52	93,22	93,22	0,685509
<b>decision_tree_desbalanceados</b>	80,07	83,47	19,93	33,74	92,93	92,93	0,682366
<b>mlp_oversampling</b>	74,47	84,54	25,53	45,71	82,45	82,45	0,677159
<b>mlp_rolling_desbalanceados</b>	75,43	82,65	24,57	34,96	86,77	86,77	0,676002
<b>decision_tree_oversampling</b>	76,73	86,03	23,27	50,92	83,90	83,90	0,674107
<b>mlp_rolling_undersampling</b>	62,80	85,65	37,20	62,38	62,92	62,92	0,669128

## 5. Avaliação

### 5.1. Avaliar resultados

Os sete métodos foram aplicados a oito cenários, resultando num total de 56 previsões. Tal como referido anteriormente, as métricas escolhidas foram seis, em que a AUC foi a escolhida para selecionar o modelo. Desta forma, o top 3 dos modelos, aplicados a cenários, é constituído pelo modelo de Regressão Logística, em cenário de *rolling window* e balanceado pela função *SMOTE()* (AUC = 0,750307), em primeiro lugar; em segundo lugar, o modelo de Regressão Logística balanceado pela função *SMOTE()* (AUC = 0,750034); e, por último, o modelo Random Forest em cenário *rolling window* e balanceado pela função *undersampling()* (AUC = 0,748719).

O pior resultado obtido advém do modelo de Redes Neurais em cenário de *rolling window* e balanceado pela função *undersampling()* (AUC = 0,669128).

### 5.2. Revisão do processo

Tendo em conta os objetivos de Data Mining, referidos anteriormente, concluiu-se que o modelo preditivo escolhido cumpre o requisito de 75% na métrica AUC.



Devido à falta de tempo para manobra e limpeza de dados, bem como a possibilidade de acrescentar novos dados, as previsões foram efetuadas com a totalidade dos dados originais.

### 5.3. Determinação dos próximos passos

Com a obtenção das 56 previsões, e definido o limite de AUC ótimo, o grupo decidiu abordar o problema de duas formas distintas. Na primeira opção, o programa escolherá e executará o melhor modelo preditivo, tendo em conta a métrica definida como sendo ideal. No segundo caso, o grupo decidiu permitir que o utilizador escolhesse o modelo de previsão que irá correr, para obtenção de resultados.

## Otimização

### Representação da solução

Para realização da otimização, um modelo preditivo foi escolhido: Regressão Logística no cenário *rolling window* e balanceado pela função *SMOTE()*. Os resultados desta previsão serviram de base à otimização. Com estes valores, a instituição bancária teria acesso a soluções que lhe permitissem otimizar o lucro. Quer isto dizer que, tendo em conta a previsão dos clientes que estivessem em incumprimento/cumprimento, no próximo mês, a instituição saberia a que clientes poderia apresentar os seus cinco produtos financeiros.

Assim sendo, a solução trata-se de um vector de  $N$  clientes, em que, a cada cliente ( $N[i]$ ), é associado um produto (1 a 5). Esta associação de produto ao cliente é feita de forma a ser obtido o máximo de lucro para a instituição bancária.

De acordo com o enunciado, uma restrição terá de ser implementada. Esta restrição terá de permitir que apenas  $\text{round}(N/2)$  dos clientes tenham, associados a si, um dos 5 produtos existentes. Os restantes  $N/2$  não terão nenhum produto associado.

### Função de Avaliação

De acordo com o enunciado e com os dados do dataset, as funções relativas ao lucro de cada produto foram definidas. Seguidamente foi desenvolvida a função de avaliação, designada *profit()*, que permite calcular o lucro, para a instituição bancária, da venda de produtos financeiros a  $N$  clientes. Isto é, será otimizada a associação de cada um dos 5 produtos financeiros, a cada um dos clientes, de forma a maximizar o lucro da instituição bancária.

Os métodos de otimização escolhidos vão comparar soluções e seleccionar o modelo que apresentar a solução de maior valor de lucro. Assim sendo, segue abaixo a função de avaliação.

```
profit <- function(solucao) {  
  profit <- 0  
  
  for (i in 1:length(solucao)){  
    if (solucao[i] == 1)  
      profit <- profit + produto_1(i)  
    else if (solucao[i] == 2)  
      profit <- profit + produto_2(i)  
    else if (solucao[i] == 3)  
      profit <- profit + produto_3(i)
```



```
    else if (solucao[i] == 4)
      profit <- profit + produto_4(i)
    else if (solucao[i] == 5)
      profit <- profit + produto_5(i)
  }

  profit
}

lucro_base <- function(i) {
  cliente <- d[tempo + i, ]
  limit_bal <- cliente$X1_LIMIT_BAL

  if (cliente$Y == 0)
    lucro_base <- limit_bal * 0.01
  else
    lucro_base <- limit_bal * 0.005

  lucro_base
}

produto_1 <- function(i) {
  cliente <- d[tempo + i, ]

  if (cliente$X2_SEX == 1)
    lucro <- lucro_base(i) * 1.02
  else if (cliente$X2_SEX == 2)
    lucro <- lucro_base(i) * 0.98

  lucro
}

produto_2 <- function(i) {
  cliente <- d[tempo + i, ]

  if (cliente$X2_SEX == 2)
    lucro <- lucro_base(i) * 1.03
  else if (cliente$X2_SEX == 1)
    lucro <- lucro_base(i) * 0.97

  lucro
}

produto_3 <- function(i) {
  cliente <- d[tempo + i, ]

  if (cliente$X5_AGE == "21-30")
    lucro <- lucro_base(i) + 10
  else if (cliente$X5_AGE == "31-40")
    lucro <- lucro_base(i) + 30
  else if (cliente$X5_AGE == "41-75")
    lucro <- lucro_base(i) + 50

  lucro
}

produto_4 <- function(i) {
  cliente <- d[tempo + i, ]

  if (cliente$X4_MARRIAGE == 1)
    lucro <- lucro_base(i) * 1.03
  else
    lucro <- lucro_base(i)

  lucro
}
```

```
produto_5 <- function(i) {  
  d[tempo + i, ]$X1_LIMIT_BAL * 0.0075  
}
```

## Modelos de otimização utilizados e cenários escolhidos

Para a realização da otimização foram escolhidos quatro modelos: o Monte Carlo Search, um modelo de Blind Search; o Simulated Annealing e o Tabu Search, dois modelos de Local Search; e, por último, o Genetic and Evolutionary Algorithm, um modelo de Population Based Search. O objetivo foi ter, no mínimo, um modelo de cada método, de procura de soluções, a funcionar. O Blind Search é definido por uma procura exaustiva num determinado espaço de procura fornecido pelo utilizador, até que seja encontrada a melhor solução. Este espaço de procura não pode ser muito grande, pois corre o risco de não ser eficaz. O Local Search tem a capacidade de produzir novas soluções, com base numa solução válida existente. Baseia-se numa procura pela vizinhança definida pelo utilizador, até encontrar a melhor solução. O Population Based Search tem a capacidade de produzir soluções, tendo em conta milhares de outras soluções. Basicamente, permite a mutação de partes da solução, por partes de soluções com um, ou mais, pontos de procura semelhantes. Com isto, uma maior diversidade de soluções pode ser alcançada (baseado no livro de CORTEZ, Paulo; 2014).

Cada modelo de otimização apresenta quatro cenários, com a exceção do Tabu Search, que apenas apresenta três, dos quatro cenários aplicados. O primeiro cenário diz respeito à otimização sem restrição e utiliza o método de *death penalty* (se as soluções não forem ideais, ser-lhes-á atribuído um valor extremamente positivo, ou negativo, para que as soluções das iterações não se encontrem fora do intervalo expectável). O segundo cenário diz respeito à otimização com *death penalty* na restrição. O terceiro cenário diz respeito à otimização sem restrição e utiliza o método *repair* (repara os valores das soluções, para que estas se encontrem dentro do intervalo de valores expectável). O quarto cenário diz respeito à otimização com restrição e utiliza o método *repair*. Como referido anteriormente, o segundo cenário não se aplica ao modelo Tabu Search.

## Avaliação de resultados

Após execução dos modelos de otimização, que utilizam *repair*, foi possível compreender que os resultados provenientes da função de avaliação são melhores que os resultados dos modelos que utilizam o método do *death penalty*. Assim sendo, o grupo considerou que seria mais proveitoso apenas implementar os modelos que utilizam o método *repair*. No entanto, é de notar que, os modelos que utilizam *death penalty*, foram implementados e testados.

De forma a perceber qual o modelo de otimização com melhores resultados, segue abaixo a seguinte tabela:

Tabela 6 - Comparação entre modelos de otimização

Modelo	Solução	Lucro
tabu_search_repair	3 2 2 5 4 2 0 2 2 2	18116,5
genalg_dp	3 4 2 5 4 3 1 3 2 5	18214
genalg_repair	5 2 2 5 3 4 4 2 3 5	18110
montecarlo_dp	5 4 2 5 4 4 3 1 3 5	18126
montecarlo_repair	5 2 2 5 4 2 3 3 4 5	18181
sann_dp	5 2 2 5 4 4 3 3 3 4	18041
sann_repair	2 4 2 3 4 2 3 2 3 5	18010,5
tabu_search_repair_restricao	0 1 2 0 0 2 1 0 2 0	13768
genalg_dp_restricao	0 2 2 0 3 0 1 0 2 0	15193
genalg_repair_restricao	0 2 2 0 1 2 0 0 2 0	14702
montecarlo_dp_restricao	3 3 1 0 5 0 3 0 0 0	12193
montecarlo_repair_restricao	3 2 0 0 4 4 0 0 1 0	11071
sann_dp_restricao	0 2 1 0 4 0 1 4 0 0	13396
sann_repair_restricao	0 3 2 0 4 0 1 0 5 0	14534



## Sistema construído em Shiny

O Shiny é um package do R que torna fácil a implementação, e integração, de aplicações web interativas diretamente do R. Todas as aplicações Shiny têm a mesma estrutura: dois ficheiros R guardados na mesma diretoria, um “ui.R” e um “server.R”.

A equipa decidiu recorrer ao Shiny, de forma a criar um interface interativo para mostrar os resultados obtidos do trabalho realizado no projeto.

A nossa aplicação Shiny é bastante simples e intuitiva. Através de *widgets* alocados na *sidebar panel*, o utilizador seleciona o modelo de previsão e otimização que quer utilizar para correr os dados, seleciona a opção “com” ou “sem” restrição de produto, o tempo atual e o número de clientes a analisar. Posto isto, e com base nas opções selecionadas, o utilizador consegue visualizar os resultados no main panel, com base nas opções escolhidas.

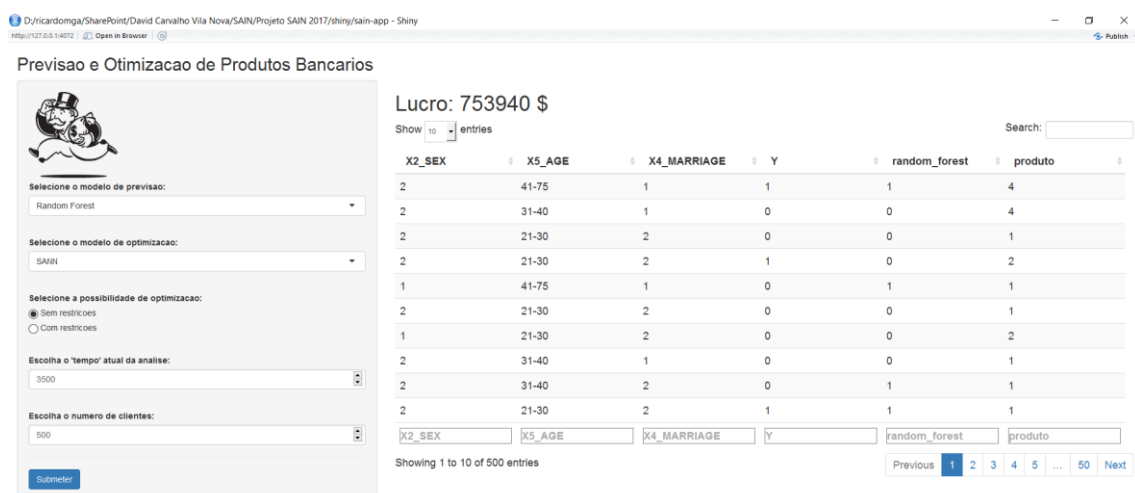


Figura 5 - Sistema Shiny

Como é possível visualizar na imagem acima apresentada, a aplicação Shiny, implementada, permite ao utilizador escolher o modelo de previsão; o modelo de otimização; se pretende que a restrição seja, ou não, implementada; o tempo atual da análise; bem como o número de clientes para venda dos produtos financeiros, da instituição bancária. Após click no botão de “submeter”, o output apresentado ao utilizador contém o lucro (soma do lucro que advém da associação de um determinado produto a N clientes) e os dados relevantes ao cálculo das condições da função de avaliação (X2\_SEX, X5\_AGE, entre outros).

<https://www.youtube.com/watch?v=LqhP3Oip7al&feature=youtu.be>

## Conclusão

De acordo com os objetivos definidos no início do projeto, é possível afirmar que, dentro do prazo estipulado, o grupo conseguiu atingir as expectativas.

Segundo a metodologia Crisp-DM, as previsões foram conseguidas com o uso de sete modelos de previsão classificativa: RandomForest, KSVM, Mlpe, MLp, GLM, Árvores de Decisão e Naive Bayes. Cada modelo foi sujeito a oito tipos de cenários diferentes, para ser possível a comparação entre diferentes ambientes. De forma a ser possível a comparação, seis métricas foram escolhidas: ACC, AUC, CE, TPR, TNR e PRECISION. Destas seis, a métrica que o grupo escolheu para selecionar o modelo foi a AUC. Ou seja, se esta métrica apresentasse valor superior a 75% em resultados de um determinado modelo, este era considerado ótimo. Assim sendo, o melhor modelo de previsão foi a Regressão Logística em ambiente rolling window, balanceado pela função *SMOTE()*.

Para a otimização foram escolhidos os modelos Monte Carlo Search (Blind Search); os modelos Tabu Search e Simulated Annealing (Local Search); e Genetic and Evolutionary Algorithm (Population Based Search). Cada modelo foi sujeito a quatro cenários: sem restrições e utilização do *death penalty*; *death penalty* na restrição; sem restrições e utilização do *repair*; *repair* na restrição. O Tabu Search foi o único que não correu no segundo cenário. Concluiu-se que, na otimização, o uso do método *repair* é mais proveitoso, pois permite a obtenção de melhores resultados.



## Bibliografia

- CORTEZ, Paulo; *Modern Optimization with R*; Springer International Publishing Switzerland 2014
- <http://archive.ics.uci.edu/ml/>

## Anexos

### Atas

#### Ata número um

No passado dia vinte e quatro de março, na sala B.2.35, deu-se início ao projeto da unidade curricular de Sistemas Adaptativos para a Inteligência do Negócio.

Tarefa	Agentes	Estado da tarefa
<b>Análise do enunciado do projeto</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta

#### Ata número dois

No passado dia trinta de março, na biblioteca do Campus de Azurém, o grupo reuniu para dar continuidade ao projeto da unidade curricular.

Tarefa	Agentes	Estado da tarefa
<b>Análise do enunciado do projeto</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Completa
<b>Revisão da matéria, leccionada pelo docente, para determinação do tipo de previsão em estudo</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Completa
<b>Análise dos modelos de previsão classificativa para aplicação ao projeto</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta

#### Ata número três

No passado dia trinta e um de março, na biblioteca de Azurém, o grupo reuniu para dar continuidade ao projeto da unidade curricular.

Tarefa	Agentes	Estado da tarefa
<b>Análise dos modelos de previsão classificativa para aplicação ao projeto</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Análise da metodologia Crisp-DM</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Construção do relatório</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta

#### Ata número quatro

No passado dia sete de abril, na biblioteca de Azurém, o grupo reuniu para dar continuidade ao projeto da unidade curricular.

Tarefa	Agentes	Estado da tarefa
<b>Análise dos modelos de previsão classificativa para aplicação ao projeto.</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Completa
<b>Análise da metodologia Crisp-DM</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Completa
<b>Business Understanding</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Completa
<b>Data Understanding</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta

<b>Data Preparation</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Construção do relatório</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta

#### Ata número cinco

No passado dia vinte de abril, na biblioteca de Azurém, o grupo reuniu para dar continuidade ao projeto da unidade curricular.

Tarefa	Agentes	Estado da tarefa
<b>Data Understanding</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Completa
<b>Data Preparation</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Modeling</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Construção do relatório</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta

#### Ata número seis

No passado dia vinte e um de abril, na biblioteca de Azurém, o grupo reuniu para dar continuidade ao projeto da unidade curricular.

Tarefa	Agentes	Estado da tarefa
<b>Data Preparation</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta

<b>Modeling</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Análise dos modelos de otimização para aplicação ao projeto</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Construção do relatório</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta

### Ata número sete

No passado dia vinte e oito de abril, na biblioteca de Azurém, o grupo reuniu para dar continuidade ao projeto da unidade curricular.

<b>Tarefa</b>	<b>Agentes</b>	<b>Estado da tarefa</b>
<b>Data Preparation</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Completa
<b>Modeling</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Análise dos modelos de otimização para aplicação ao projeto</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Evaluation</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Construção do relatório</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta

### Ata número oito

No passado dia quatro de maio, na biblioteca de Azurém, o grupo reuniu para dar continuidade ao projeto da unidade curricular.

Tarefa	Agentes	Estado da tarefa
<b>Modeling</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Análise dos modelos de otimização para aplicação ao projeto</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Completa
<b>Evaluation</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Escolha do formato da solução</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Completa
<b>Construção da função de avaliação</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Construção do relatório</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta

### Ata número nove

No passado dia cinco de maio, na biblioteca de Azurém, o grupo reuniu para dar continuidade ao projeto da unidade curricular.

Tarefa	Agentes	Estado da tarefa
<b>Modeling</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta



<b>Evaluation</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Construção da função de avaliação</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Completa
<b>Aplicação de modelos de otimização</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Exploração do Shiny</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Construção do relatório</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta

#### Ata número dez

No passado dia onze de maio, na biblioteca de Azurém, o grupo reuniu para dar continuidade ao projeto da unidade curricular.

<b>Tarefa</b>	<b>Agentes</b>	<b>Estado da tarefa</b>
<b>Modeling</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Evaluation</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Aplicação de modelos de otimização</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Exploração do Shiny</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta

<b>Construção do relatório</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
--------------------------------	--	------------

#### Ata número onze

No passado dia doze de maio, na biblioteca de Azurém, o grupo reuniu para dar continuidade ao projeto da unidade curricular.

Tarefa	Agentes	Estado da tarefa
<b>Modeling</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Evaluation</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Aplicação de modelos de otimização</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Exploração do Shiny</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Construção do relatório</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta

#### Ata número doze

No passado dia vinte e seis de maio, na biblioteca de Azurém, o grupo reuniu para dar continuidade ao projeto da unidade curricular.

Tarefa	Agentes	Estado da tarefa
<b>Modeling</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta

<b>Evaluation</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Aplicação de modelos de otimização</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Exploração do Shiny</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Construção do relatório</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta

#### Ata número treze

No passado dia trinta de maio, nas salas das residências universitárias de Azurém, o grupo reuniu para dar continuidade ao projeto da unidade curricular.

<b>Tarefa</b>	<b>Agentes</b>	<b>Estado da tarefa</b>
<b>Modeling</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Evaluation</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Aplicação de modelos de otimização</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Exploração do Shiny</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Construção do relatório</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta

**Ata número catorze**

No passado dia trinta e um de maio, nas salas das residências universitárias de Azurém, o grupo reuniu para dar continuidade ao projeto da unidade curricular.

Tarefa	Agentes	Estado da tarefa
<b>Modeling</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Completa
<b>Evaluation</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Aplicação de modelos de otimização</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Exploração do Shiny</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Construção do sistema em Shiny</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Construção do relatório</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta

**Ata número quinze**

No passado dia um de junho, nas salas das residências universitárias de Azurém, o grupo reuniu para dar continuidade ao projeto da unidade curricular.

Tarefa	Agentes	Estado da tarefa
<b>Evaluation</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Completa

<b>Aplicação de modelos de otimização</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Completa
<b>Exploração do Shiny</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Construção do sistema em Shiny</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta
<b>Construção do relatório</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Incompleta

#### Ata número dezasseis

No passado dia dois de junho, nas salas das residências universitárias de Azurém, o grupo reuniu para finalizar o projeto da unidade curricular.

Tarefa	Agentes	Estado da tarefa
<b>Exploração do Shiny</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Completa
<b>Construção do sistema em Shiny</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Completa
<b>Construção do relatório</b>	David Carvalho, João Gomes, Pamela Coelho e Ricardo Araújo	Completa

#### Avaliação

O grupo decidiu autoavaliar-se em 18 valores, sem que houvessem distinções de notas.