



UNIVERSIDADE
AUTÓNOMA
DE LISBOA



TRABALHO FINAL DE SAD – PROJETO DE DATA MINING

Utilização da Data Mining para Estudo da Avaliação de Carros:

Utilizando o Método CRISP-DM

Bruno Saraiva 20160782

Diogo Palos 30001058

Miguel Nunes 30000814

Ricardo Melo 30000486

Universidade Autónoma de Lisboa, Engenharia Informática, 1169-023 Lisboa, Portugal

Abstract

O nosso data set em estudo é derivado do modelo de decisão hierárquico simples. O conjunto de dados extraídos, permite a uma empresa de venda de automóveis, analisar as preferências dos seus clientes e dos seus futuros clientes, no que toca ao preço de venda (buying), o custo da manutenção (maint), o número de portas (doors), o número de pessoas que o carro pode transportar (persons), o tamanho da bagagem (lug_boot) e a segurança total estimada do carro (safety). Estes são os atributos de entrada.

Keywords: empresa; dados; extração; preferência; atributos; data set; modelo de decisão hierárquico simples;

TRABALHO FINAL DE SAD – PROJETO DE DATA MINING

Utilização da Data Mining para Estudo da Avaliação de Carros:

Utilizando o Método CRISP-DM

Este trabalho prático consiste no desenvolvimento de um projeto de *Data Mining* com dados reais utilizando o método CRISP-DM (Cross Industry Standard Process for Data Mining). Recorreremos ao programa *RStudio* para explorar e analisar o dataset da Avaliação de Carros.

O *RStudio* é um *software* de desenvolvimento integrado no R, sendo que esta é uma linguagem de programação que inclui cálculos estatísticos, elaboração de gráficos e análise de dados. É considerado um dos melhores ambientes computacionais para a análise de dados.

Para realizarmos este trabalho fomos ao site recomendado pela docente: “UCI Machine Learning” e escolhemos um tema do nosso agrado, relacionado com a venda de automóveis e a preferência dos clientes.

CRISP-DM

“CRISP-DM é a abreviação de Cross Industry Standard Process for Data Mining, que pode ser traduzido como Processo Padrão Inter-Indústrias para Mineração de Dados. É um modelo de processo de mineração de dados que descreve abordagens usualmente usadas por especialistas em mineração de dados para atacar problemas.” In: Wikipédia: a enciclopédia livre.

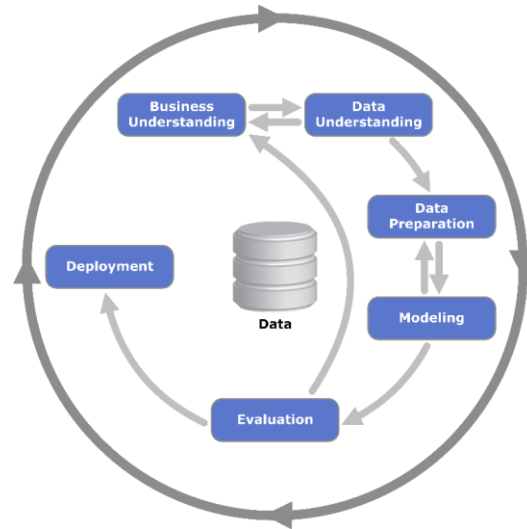


Fig. 1 – Esquema do CRISP-DM

Compreensão do Negócio

Prende-se com o entendimento dos objetivos e requerimentos do projeto numa perspectiva de negócio. Os dados que constituem o dataset utilizado foram extraídos de um modelo de decisões hierárquicas. Os estudos e análises efetuadas neste projeto no âmbito da avaliação de carros, têm como objetivo orientar as vendas dos automóveis para o que os clientes estão mais predispostos a comprar, ou seja, a partir dos dados dos clientes, verificar quais os carros ou as características dos mesmos que são mais populares e geram mais lucro e, prever o que estão a pensar em comprar futuramente, podendo assim corresponder melhor à necessidade e expectativa do cliente. Desta forma é ainda possível reforçar os sectores responsáveis pela construção,

compra e venda dos carros preferidos dos clientes uma vez que, podemos a partir da sobreposição verificar que ao conhecer as preferências do consumidor, conseguimos orientar a venda do modelo ideal para o mesmo. É ainda possível reforçar os setores anteriormente referidos através do foco na produção e venda dos modelos automóveis com maior procura, podendo assim obter um número de vendas superior e uma margem de lucro também mais elevada e, obter novos dados de preferência e popularidade para a criação de novos modelos automóveis ou adaptação dos modelos existentes. Não só este último fator permite um aumento de lucro, como também evita a venda, revenda e produção de modelos não populares, melhorando assim o custo/benefício do negócio.

Compreensão dos Dados

Inicia-se com a obtenção dos dados, e posteriormente a realização de atividades a fim de se familiarizar com os dados, identificar os problemas de qualidade (integridade dos dados, redundância), detetar subconjuntos de dados e formar hipóteses sobre possíveis informações escondidas.

Os dados foram retirados da UCI Machine Learning Repository que derivam de um modelo de decisão hierárquico simples, estes dados mostram várias avaliações de carros.

Foram consideradas 6 variáveis: (buying, maint, doors, persons, lug_boot, safety) e 4 valores da classe (unacc, acc, good, vgood). Não existem valores de atributos nulos neste dataset.

- buying é o preço de compra do carro, tem o valor de vhigh, high, med e low, (preço muito alto, preço alto, preço médio e preço baixo, respectivamente);

- maint é o preço da manutenção do carro, tem o valor de vhigh, high, med e low, (preço muito alto, preço alto, preço médio e preço baixo, respetivamente);
- doors é o número de portas, tem o valor de 1, 2, 3, 4 e 5more, (1 porta, 2 portas, 3 portas, 4 portas, 5 ou mais portas, respetivamente);
- persons é a capacidade total de pessoas, tem o valor de 2, 4 e more, (2 pessoas, 4 pessoas, mais do que 4 ou variadas, respetivamente);
- lug_boot capacidade da mala de bagagens, tem o valor de small, med, big, (capacidade pequena, capacidade média, capacidade grande, respetivamente);
- safety é a segurança total efetiva do carro, tem o valor de low, med, high, (segurança baixa, segurança média e segurança alta, respetivamente).

Tabela 1

Valores dos Atributos

buying	maint	doors	persons	lug_boot	safety
vhigh	vhigh	2	2	small	low
high	high	3	4	med	med
med	med	4	more	big	high
low	low	5more	////////////////	////////////////	////////////////

Os atributos de entrada são impressos em minúsculas. Além do conceito de alvo (CAR),

o modelo inclui três conceitos intermediários: PREÇO, TECNOLOGIA, CONFORTO. O

modelo original relacionado aos descendentes de nível inferior, segue em anexo:

CAR car acceptability

. PRICE overall price

. . buying buying price

. . maint price of the maintenance

. TECH technical characteristics

.. COMFORT comfort
 ... doors number of doors
 ... persons capacity in terms of persons to carry
 ... lug_boot the size of luggage boot
 .. safety estimated safety of the car

Temos também uma coluna do valor das classes que se traduz desde unacc to vgood, a tabela com os valores da classe segue em anexo:

Tabela 2

Valores das Classes

distr
unacc
acc
good
vgood

Tabela 3

Distribuição das Classes

class	N	N[%]
unacc	1210	70,03%
acc	384	22,222%
good	69	3,993%
vgood	65	3,762%

Olhando para a tabela de distribuição das classes podemos concluir que a maioria dos registos não foram aceitáveis. Esta informação estava contida no data folder da Avaliação de carros.

Preparação dos Dados

É chave para todo o processo, podendo consumir mais de metade do tempo gasto. É nesta fase que são definidos os dados sobre os quais serão aplicados métodos de data mining com justificação da inclusão/exclusão dos dados, realização de testes de significância e correlação, amostragem da base de dados, limpeza dos dados seleccionados, produção de novos dados,

criação de novos registos para os dados construídos, agregação da informação e reorganização dos atributos.

Em relação ao Data Set, o nosso ficheiro é do tipo “data”, podemos importar o ficheiro para o RStudio através do seguinte código:

```
> car <- read.csv("car.data")
```

Fig. 2 – Import do DataSet

Estamos a atribuir à variável “car”, a leitura do Data Set “car.data”.

Para visualizar os dados, aplicamos o código “View(car)” que nos levará à construção de uma tabela com os dados presentes no Data Set, como podemos visualizar :

	buying	maint	doors	persons	lug_boot	safety	distr
1	vhigh	vhigh	2	2	small	low	unacc
2	vhigh	vhigh	2	2	small	med	unacc
3	vhigh	vhigh	2	2	small	high	unacc
4	vhigh	vhigh	2	2	med	low	unacc
5	vhigh	vhigh	2	2	med	med	unacc
6	vhigh	vhigh	2	2	med	high	unacc
7	vhigh	vhigh	2	2	big	low	unacc
8	vhigh	vhigh	2	2	big	med	unacc
9	vhigh	vhigh	2	2	big	high	unacc
10	vhigh	vhigh	2	4	small	low	unacc
11	vhigh	vhigh	2	4	small	med	unacc
12	vhigh	vhigh	2	4	small	high	unacc
13	vhigh	vhigh	2	4	med	low	unacc
14	vhigh	vhigh	2	4	med	med	unacc
15	vhigh	vhigh	2	4	med	high	unacc

Fig. 3 – Screenshot Parcial do DataSet

O nosso Data Set é composto por 1728 observações e 7 variáveis.

O objeto “buying” refere-se ao custo do veículo, este varia entre low, medium, high e very high

O objeto “maint” refere-se ao custo de manutenção do veículo, este também varia entre low, medium, high e very high

Em relação ao objeto doors refere-se ao número de portas que o automóvel possui, varia entre 1, 2, 3, 4 e 5more (5 ou mais).

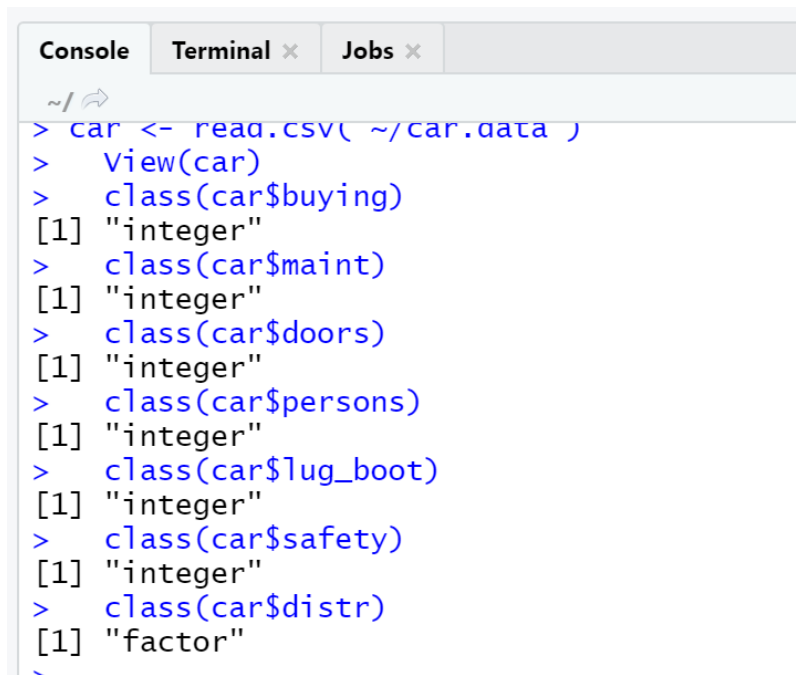
O objeto persons refere-se ao número de pessoas que o veículo suporta, também varia entre 1, 2, 3, 4 e 5more (5 ou mais).

O objeto lug_boot refere-se ao tamanho da bagagem pode conter os valores small, medium e big.

O objeto safety tem como função detalhar o nível de segurança dos veículos.

De seguida vamos visualizar o tipo de dados presente na nossa Data Set

Com a utilização do comando `class(car$objetos)`, este retorna o tipo, podemos verificar que são todas do tipo “factor”.



```
Console Terminal x Jobs x
~/
> car <- read.csv( ~/car.data )
> View(car)
> class(car$buying)
[1] "integer"
> class(car$maint)
[1] "integer"
> class(car$doors)
[1] "integer"
> class(car$persons)
[1] "integer"
> class(car$lug_boot)
[1] "integer"
> class(car$safety)
[1] "integer"
> class(car$distr)
[1] "factor"
```

Fig. 4 – Screenshot dos Tipos de Variável

Para conseguirmos efetuar a análise detalhada dos valores mínimos, máximos e dos quartis é necessário transformar os valores dos objetos em número, para isso podemos definir uma escala lógica de associação.

Em todos os objetos exceto o objeto “lug_boot” temos uma classificação de low a very high, logo podemos associar a seguinte escala:

- Low = 1;
- Med = 2;
- High = 3;
- Very High = 4.

No objeto “lug_boot” temos uma classificação de small a big, logo, utilizando uma lógica semelhante, vamos associar a escala da seguinte forma:

- Small = 1;
- Med = 2;
- Big = 3;

Através do seguinte código, será executada a conversão dos valores apresentados na tabela, efetuando assim, com base na escala apresentada, a substituição dos valores de texto para formato numérico.

```
gsub("low", "1", car)
```

Fig. 5 – Screenshot da Instrução de Substituição de Valores

Fazendo este tipo de conversão para todos os valores low, med, high, vhigh.

Ficando assim a nossa tabela:

	buying	maint	doors	persons	lug_boot	safety	distr
1	4	4	2	2	1	1	unacc
2	4	4	2	2	1	2	unacc
3	4	4	2	2	1	3	unacc
4	4	4	2	2	2	1	unacc
5	4	4	2	2	2	2	unacc
6	4	4	2	2	2	3	unacc
7	4	4	2	2	3	1	unacc
8	4	4	2	2	3	2	unacc
9	4	4	2	2	3	3	unacc
10	4	4	2	4	1	1	unacc
11	4	4	2	4	1	2	unacc
12	4	4	2	4	1	3	unacc
13	4	4	2	4	2	1	unacc
14	4	4	2	4	2	2	unacc
15	4	4	2	4	2	3	unacc

Fig. 6 – Screenshot do DataSet com os valores das variáveis modificadas

Para obtermos uma análise completa dos dados em relação aos valores mínimos, máximos, quartis, mediana, média, podemos obter essas informações com o comando `summary(car)`, assim vamos obter as seguintes informações:

```
> summary(car)
      buying      maint      doors      persons      lug_boot      safety      distr
Min.   :1.00  Min.   :1.00  Min.   :2.00  Min.   :2.000  Min.   :1    Min.   :1    acc : 384
1st Qu.:1.75  1st Qu.:1.75  1st Qu.:2.75  1st Qu.:2.000  1st Qu.:1    1st Qu.:1    good : 69
Median :2.50  Median :2.50  Median :3.50  Median :4.000  Median :2    Median :2    unacc:1210
Mean   :2.50  Mean   :2.50  Mean   :3.50  Mean   :3.667  Mean   :2    Mean   :2    vgood: 65
3rd Qu.:3.25  3rd Qu.:3.25  3rd Qu.:4.25  3rd Qu.:5.000  3rd Qu.:3    3rd Qu.:3
Max.   :4.00  Max.   :4.00  Max.   :5.00  Max.   :5.000  Max.   :3    Max.   :3
```

Fig. 7 – Screenshot do Sumário dos Valores Máximos e Mínimos

É importante referir que nesta distribuição de dados, todos os casos são completos e não há nenhum dado classificado como NA, ou seja, null.

Podemos observar pela imagem que conseguimos obter os dados mínimos, máximos, média e mediana para cada Objeto e assim fazer uma análise.

Modelação

A modelação ou aplicação de algoritmos de data mining, constitui outro passe chave, sendo esta a etapa onde se revela a nova informação. Esta inclui a seleção de técnicas de modelação dos dados, a definição de procedimentos de treino e de teste, a construção de modelos e a sua avaliação.

Vamos utilizar o modelo de regressão linear simples e o modelo de regressão linear múltipla. Os modelos de regressão linear pretendem representar uma variável target (numérica) a partir da combinação linear de um conjunto de variáveis explicativas.

Vamos utilizar o modelo de regressão linear simples para prever o preço da compra total do veículo com o seu preço de manutenção.

Vamos utilizar o modelo de regressão linear múltipla para prever o preço da compra total do veículo de acordo com a sua segurança, capacidade de pessoas e custo de manutenção.

Obtivemos os seguintes gráficos:

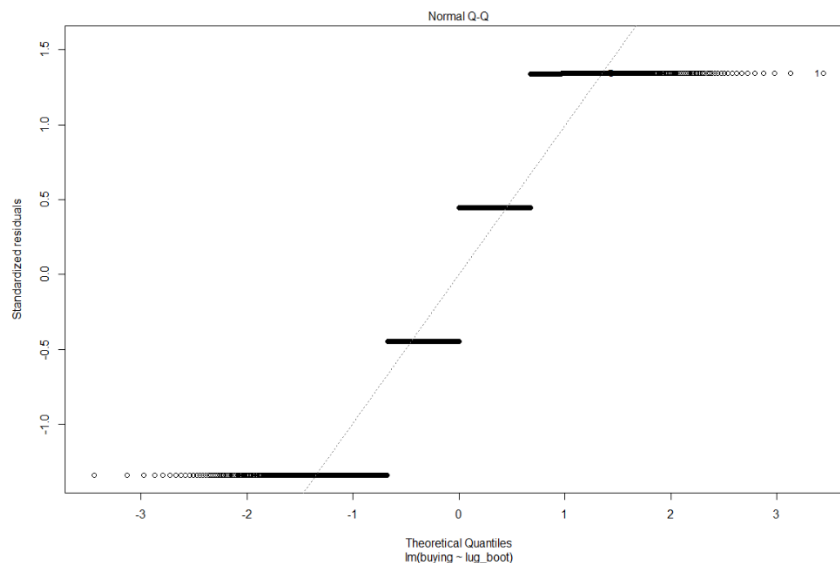


Fig. 8 – Screenshot do plot da relação buying (Teórica VS Residual)

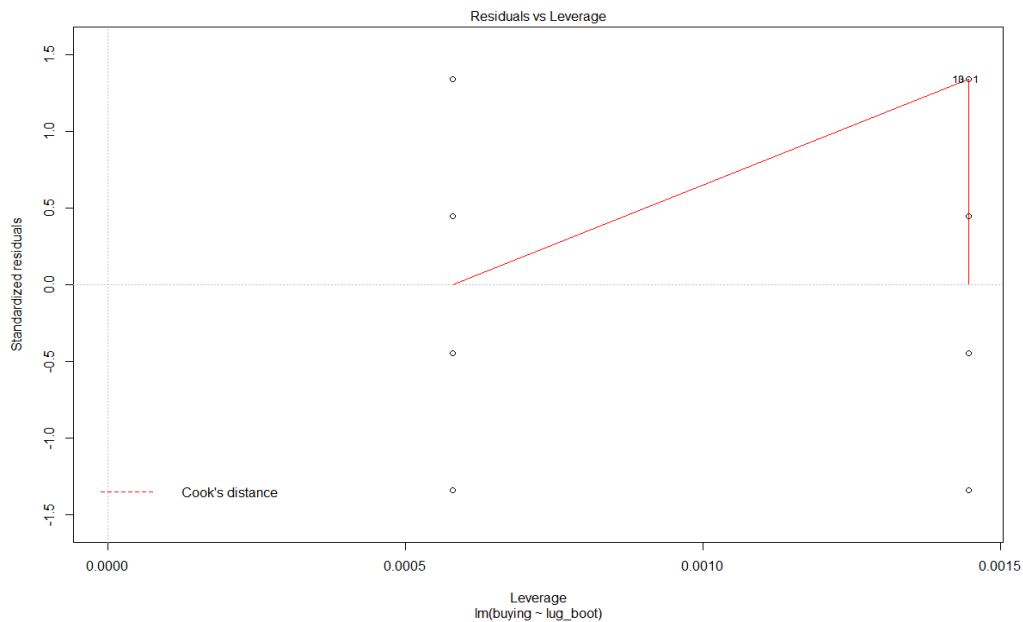


Fig. 9 – Screenshot do plot da relação buying Alavancagem VS Residual

Teste e Avaliação

Em relação à análise dos dados escolhemos os objetos mais importantes e interessantes para análise, ou seja, escolhemos os objetos buying, maint, persons e safety.

```
summary(car[c("buying", "maint", "persons", "safety")])
```

Fig. 10 – Screenshot da Instrução para sumarizar as variáveis mais importantes

Através do código apresentado, conseguimos obter a análise aos objetos e às informações pretendidas:

buying	maint	persons	safety
Min. :1.00	Min. :1.00	Min. :2.000	Min. :1
1st Qu.:1.75	1st Qu.:1.75	1st Qu.:2.000	1st Qu.:1
Median :2.50	Median :2.50	Median :4.000	Median :2
Mean :2.50	Mean :2.50	Mean :3.667	Mean :2
3rd Qu.:3.25	3rd Qu.:3.25	3rd Qu.:5.000	3rd Qu.:3
Max. :4.00	Max. :4.00	Max. :5.000	Max. :3

Fig. 11 – Screenshot do Sumário Executado

Conseguimos aferir que o padrão médio de carro dos clientes tem como capacidade de lugares 4 pessoas, o nível de preço do automóvel é de 2.50 ou seja um custo médio/alto, um nível de manutenção também médio/alto e um nível de segurança 2 ou seja médio. Podemos também observar os valores de variância e desvio padrão através das funções `Var()` e `sd()` sucessivamente, vamos analisar então o custo de manutenção através destas funções:

```
> var(car$maint)
[1] 1.250724
> sd(car$maint)
[1] 1.118358
```

Fig. 12 – Screenshot da Instrução para Aferir o Desvio Padrão

Podemos também fazer uma análise para verificar qual o valor que se repete mais vezes. Vamos então analisar qual o poder de compra mais predominante por parte dos clientes. Aplicando o seguinte código, podemos obter o número que se repete mais vezes:

```
> tail(names(sort(table(car$buying))), 1)
[1] "4"
```

Fig. 13 – Screenshot da Instrução para Obter o Valor Repetido

Ao contrário do que se possa pensar, o valor que mais se verificou foi o 4 que equivale a very high, ou seja, há mais clientes a comprar carros de valor muito alto do que os restantes.

Agora vamos apresentar a relação entre o custo do veículo e o custo de manutenção através de um boxplot, para isso executamos o seguinte código:

```
> boxplot(car$buying, car$maint, main="Comparação custo do automóvel e custo manutenção", col="#58D3F7",  
+         ylab = "Custo de 1 a 4", names=c("Custo do automóvel", "Custo da manutenção"))
```

Fig. 14 – Screenshot da Instrução que Permite Criar uma boxplot

Obtemos então o seguinte boxplot:

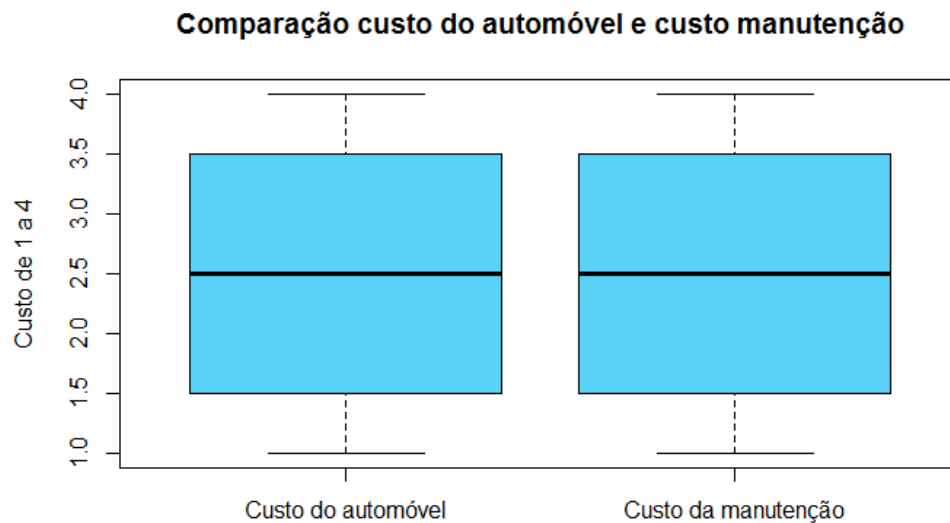


Fig. 15 – Screenshot da boxplot Comparando o Custo do Automóvel e da Manutenção

Podemos assim observar uma relação direta entre o custo do automóvel e o custo da manutenção do mesmo, o que faz sentido visto que, quanto mais caro for um automóvel, maior será o número dos seus componentes e complexidade o que origina uma manutenção mais dispendiosa.

Conseguimos também criar histogramas para visualizar melhor os dados, escolhemos por exemplo visualizar um histograma sobre a segurança dos veículos, aplicamos o seguinte código:

```
> hist(car$safety, breaks=2, main="Histograma sobre a segurança", xlab="Escala de 1 a 3", ylab="Frequência", col = c("blue", "red"))
```

Fig. 16 – Screenshot da Instrução do Histograma

Resultando no seguinte gráfico:

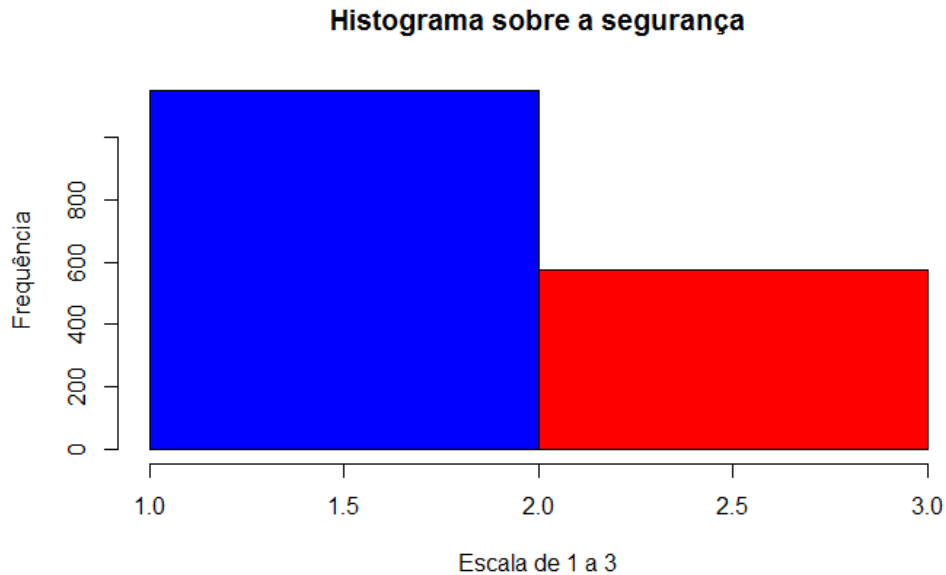


Fig. 17 – Screenshot do Histograma sobre a Segurança

Implementação

O conhecimento adquirido poderá ser utilizado nos processos de tomada de decisão, devendo existir um plano para a sua implementação, sendo que, no final, poderá existir um relatório para sumarizar todos os resultados do processo.

Posto de lado toda a informação analisada, podemos concluir que, na maioria dos casos, os clientes preferem investir em carros com valores de buying, high (4), ou seja, custos mais elevados, como se pode comprovar no summary que é a tendência do valor máximo.

Quanto aos lugares (persons), a maior parte dos clientes preferem carros standard com 4 lugares, como era esperado, no entanto, em relação à segurança verificou-se o oposto, ou seja, a grande maioria dos carros comprados, compreende-se entre níveis de segurança entre médio (2) a baixo (1).

Entre preço de compra e o preço de manutenção obtivemos uma mediana de 2,50, ou seja, a maior parte da avaliação dos carros, os clientes preferem valores médios a altos, tendo em conta que se obteve uma previsão de 2,5 no modelo de regressão linear múltipla, como segue na imagem:

```
> predict(Model2, data.frame(maint = 2, persons = 2, safety = 1))  
1  
2.5
```

Fig. 18 – Screenshot da Instrução de Predição para Comprar um Carro preço de manutenção med, pessoas max=2 e segurança low

Podemos assim concluir que os carros mais caros, que são os que têm mais saída, não possuem necessariamente os maiores níveis de segurança.

Como os clientes preferem investir em valores mais altos em automóveis no preço de compra, sabemos quais modelos devemos implementar no mercado, tendo em conta a sua preferência nos níveis de segurança, capacidade de bagagem e número de pessoas que podem transportar, assim podemos prevenir a perda de valor material, pois quando um carro não é preferido pelo cliente este fica em armazém, perdendo assim o seu valor com o tempo.

Conseguimos então orientar as vendas, deste modo, damos ao cliente o seu modelo preferido e evitamos perdas monetárias.

References

Academia In. (24/10/2018). Data Mining: Você realmente entende o que é a sua importância?.

Acedido a: 25/01/2020, em: <https://blog.academaiain1.com.br/data-mining-voce-realmente-entende-o-que-e-e-sua-importancia/>.

Aprendis. (22/03/2016). Data Mining. Acedido a: 25/01/20 20, em:

http://aprendis.gim.med.up.pt/index.php/Data_Mining.

Wikipedia. (13/01/2020). Cross Industry Standard Process for Data Mining. Acedido a:

23/12/2019, em:

https://pt.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining.

B. Zupan, M. Bohanec, I. Bratko, J. Demsar: Machine learning by function decomposition.

ICML-97, Nashville, TN. 1997 (to appear).

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>].

Irvine, CA: University of California, School of Information and Computer Science.

M. Bohanec and V. Rajkovic: Knowledge acquisition and explanation for multi-attribute decision making. In 8th Intl Workshop on Expert Systems and their Applications, Avignon, France. pages 59-78, 1988.