INSTITUTO POLITÉCNICO DE LEIRIA – ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO DE LEIRIA

RELATÓRIO ENGENHARIA DO CONHECIMENTO

Ano Letivo 2013/2014

3º Ano – 1º Semestre

David Alecrim, 2110122

Diogo Costa, 2130529

Table of Contents

Compreensão do negócio

O problema a ser abordado neste projeto e relatório consiste numa loja *online* que comercializa produtos, e que tem como objetivo a obtenção do maior volume de vendas com o maior lucro possível. O principal critério de sucesso do negócio é o lucro, mas também atrair o maior número de clientes possível para efetuarem encomendas na loja.

O negócio fornece alguns recursos, tais como, oferecer cupões de desconto a clientes e recomendar determinados produtos àqueles que tenham algum grau de probabilidade de realizar encomendas desses mesmos produtos. Oferecer cupões acarreta custos para a loja, pois é determinado valor em percentagem de um produto que a loja perde, no entanto é favorável à mesma, pois influencia o cliente a realizar encomendas mais facilmente e em maior número.

Os objetivos deste projeto de Data Mining são os de determinar com a maior certeza possível, se um cliente vai efetuar uma encomenda na sua sessão atual, e personalizar as sessões de cada cliente ou com cupões de desconto ou com sugestões de compras para adaptar a loja a cada cliente e para aumentar o número de encomendas efetuadas. O projeto de Data Mining tem sucesso, se aumentar o número de encomendas efetuadas pelos clientes e se previr com sucesso se um cliente vai ou não efetuar uma encomenda na sessão.

A metodologia a utilizar neste projeto é a CRISP-DM e as ferramentas a utilizar neste projeto são o Microsoft Excel, o RapidMiner 5, o Weka 3.7, o Notepad++ e o SublimeText 3.

Compreensão dos dados

Os dados iniciais para a realização deste projeto foram fornecidos em formato de texto simples (.txt) pelo docente da cadeira de Engenharia do Conhecimento, e disponibilizados na página da cadeira no Moodle.

Existem 24 caraterísticas nos ficheiros de dados disponibilizados, sendo que 23 são atributos comuns e o último é o atributo objetivo (class), estando este apenas no ficheiro de treino. Este é o atributo que indica se há alguma compra na sessão ou não. Todos estes atributos à exceção do atributo objetivo (*order*), do *availability* e *onlineStatus* são atributos numéricos, sejam eles naturais, inteiros ou reais. Estas exceções são do tipo nominal.

De seguida apresenta-se informação sobre os atributos dos Data Sets, estatísticas e restrições aos campos dos Data Sets.

***sessionNo**: É o primeiro atributo* e indica o número da sessão do cliente. Este valor é único para cada sessão e o seu domínio é o domínio dos números naturais ([1,50000]). Apesar de nos dados se encontrarem várias transações por cada sessão, a sessão continua a ser única.

Nome do Ficheiro	Valores em Falta	Mínimo	Máximo	Média	Desvio Padrão	Número de Valores Distintos	Número de Valores Únicos (Sessão apenas com uma transação)
DataSet_Treino	0	1	50000	25274.631	14441.366	50000	5492
DataSet_Validacao	0	1	5111	2385.701	1426.207	5111	533

Estatísticas de sessionNo 1

startHour: Apresenta a hora em que a sessão se iniciou, toma valores entre 0 e 23, sendo o seu domínio os números inteiros não negativos ([0,23]).

Nome do Ficheiro	Valores em Falta	Mínimo	Máximo	Média	Desvio Padrão	Número de Valores Distintos	Número de Valores Únicos
DataSet_Treino	0	0	23	14.617	4.486	24	0
DataSet_Validacao	0	0	23	18.59	5.544	12	0

Estatísticas de startHour 2

startWeekday: Define o dia da semana em que se realizou a sessão e toma valores de 1 a 7 inclusive, em que 1 corresponde a Segunda-Feira e 7 a Domingo. O domínio destes valores é os números naturais ([1,7]).

Nome do Ficheiro	Valores em Falta	Média	Desvio Padrão	Número de Valores Únicos
DataSet_Treino	0	5.925	0.791	0
DataSet_Validacao	0	6.458	1.719	0

Estatísticas de startWeekday 3

***duration**:* Indica o tempo que passou, em segundos, desde o início da sessão actual do cliente. Os valores apresentados inserem-se nos números reais não negativos ([0,21320.113]).

Nome do Ficheiro	Valores em Falta	Mínimo	Máximo	Média	Desvio Padrão	Número de Valores Distintos	Número de Valores Únicos
DataSet_Treino	0	0	2580.092	1573.902	2427.123	369334	338539
DataSet_Validacao	0	0	21320.113	1645.291	2279.004	41870	41346

Estatísticas de duration 4

cCount: Apresenta o número de produtos no qual o cliente clicou na loja, o que corresponde a um intervalo de números inteiros não negativos ([0,200]). Neste caso vamos admitir que 0 faz parte do intervalo, já que o utilizador pode ter iniciado sessão e estar a fazer outras tarefas, como por exemplo actualizar os seus dados, não incrementando o número de cliques.

Nome do Ficheiro	Valores em Falta	Mínimo	Máximo	Média	Desvio Padrão	Número de Valores Distintos	Número de Valores Únicos
DataSet_Treino	0	0	200	24.14	30.398	201	0
DataSet_Validacao	0	0	200	27.317	32.549	201	0

Estatísticas de cCount 5

cMinPrice: Mostra o preço mínimo do produto na lista de produtos em que o cliente clicou, estando estes valores no domínio dos números reais não negativos ([0,5999.99]).

Nome do Ficheiro	Valores em Falta	Mínimo	Máximo	Média	Desvio Padrão	Número de Valores Distintos	Número de Valores Únicos
DataSet_Treino	2765 (1%)	0	5999.99	55.289	148.88	725	59
DataSet_Validacao	326 (1%)	0	1999.99	53.299	146.775	354	40

Estatísticas de cMinPrice 6

cMaxPrice*:* Contrariamente ao atributo anterior, este apresenta o preço mais elevado do produto que foi clicado pelo cliente. O domínio é o dos números reais não negativos ([0,6999.99]).

Nome do Ficheiro	Valores em Falta	Mínimo	Máximo	Média	Desvio Padrão	Número de Valores Distintos	Número de Valores Únicos

DataSet_Treino	2765 (1%)	0	6999.99	146.663	283.218	873	51
DataSet_Validacao	326 (1%)	0	4799	149.135	272.25	434	38

Estatísticas de cMaxPrice 7

***cSumPrice**:* Indica a soma dos preços dos produtos em que o cliente clicou. Os valores apresentados pertencem ao domínio dos reais não negativos ([0,117310.7]).

Nome do Ficheiro	Valores em Falta	Mínimo	Máximo	Média	Desvio Padrão	Número de Valores Distintos	Número de Valores Únicos
DataSet_Treino	2765 (1%)	0	115742	1189.248	3371.174	72989	34367
DataSet_Validacao	326 (1%)	0	117310.7	1240.986	3523.665	13255	7683

Estatísticas de cSumPrice 8

bCount: Mostra o número de produtos que o cliente tem no carrinho de compras. O domínio destes valores pertence aos números inteiros não negativos ([0,108]).

Nome do Ficheiro	Valores em Falta	Mínimo	Máximo	Média	Desvio Padrão	Número de Valores Distintos	Número de Valores Únicos
DataSet_Treino	0	0	108	4.135	4.452	109	29
DataSet_Validacao	0	0	43	4.367	4.334	44	1

Estatísticas de bCount 9

bMinPrice: É igual ao *cMinPrice*, mas relativamente aos produtos que estão no carrinho de compras. O seu intervalo está compreendido entre [0,6999.99].

Nome do Ficheiro	Valores em Falta	Mínimo	Máximo	Média	Desvio Padrão	Número de Valores Distintos	Número de Valores Únicos
DataSet_Treino	5130 (1%)	0	6999.99	67.625	174.986	747	63
DataSet_Validacao	589	0	1999.99	65.936	172.341	370	36

	(1%)						

Estatísticas de bMinPrice 10

***bMaxPrice**:* É igual ao *cMaxPrice*, mas também relativamente aos produtos que estão no carrinho de compras. O seu intervalo está compreendido entre [0,6999.99].

Nome do Ficheiro	Valores em Falta	Mínimo	Máximo	Média	Desvio Padrão	Número de Valores Distintos	Número de Valores Únicos
DataSet_Treino	5130 (1%)	0	6999.99	107.505	212.916	761	43
DataSet_Validacao	589 (1%)	0	2299.99	105.75	204.395	372	27

Estatísticas de bMaxPrice 11

bSumPrice: Mostra a soma dos preços dos produtos no carrinho de compras. Os valores obtidos pertencem aos reais não negativos ([0,23116.88]).

Nome do Ficheiro	Valores em Falta	Mínimo	Máximo	Média	Desvio Padrão	Número de Valores Distintos	Número de Valores Únicos
DataSet_Treino	5130 (1%)	0	23116.88	213.261	459.39	20247	7418
DataSet_Validacao	589 (1%)	0	8948.96	209.595	414.708	4834	1938

Estatísticas de bSumPrice 12

bStep: Define o passo do processo de compra em que o cliente se encontra no momento e toma valores entre 1 e 5 inclusive, sendo o seu domínio os números naturais ([1,5]).

Nome do Ficheiro	Valores em Falta	Mínimo	Máximo	Média	Desvio Padrão	Número de Valores Distintos	Número de Valores Únicos
DataSet_Treino	191333 (45%)	1	5	2.292	1.307	5	0
DataSet_Validacao	20766 (46%)	1	5	2.241	1.285	5	0

Estatísticas de bStep 13

onlineStatus: Indica se o cliente está online (Y) ou não (N). Como se pode observar, estes valores são nominais.

Nome do Ficheiro	Valores em Falta	Contagem de y (online)	Contagem de n (offline)	Número de Valores Distintos	Número de Valores Únicos
DataSet_Treino	160379 (37%)	265625	3009	3	0
DataSet_Validacao	17355 (39%)	27494	219	2	0

Estatísticas de onlineStatus 14

***Availability*:** Mostra o estado do produto para entrega, sendo estes valores nominais.

Nome do Ficheiro	Valores em Falta	Contagem de Completamente Encomendável	Contagem de Completamente Não Encomendável	Contagem de Maioritariamente Encomendável	Vários	Contagem de Maioritariamente Não Encomendável	Contagem de Completamente Não Determinável	Contagem de Maioritariamente Não Determinável	Número de Valores Distintos	Número de Valores Únicos
DataSet_Treino	165255 (39%)	253692	1491	5756	1284	320	1017	198	8	0
DataSet_Validacao	17757 (39%)	26283	144	654	102	36	90	2	7	0

Estatísticas de availability 15

customerNo: Indica o número (no intervalo dos números naturais) do cliente na sessão. Este valor é único para cada cliente e o seu domínio está compreendido entre [1,27318].

Nome do Ficheiro	Valores em Falta	Mínimo	Máximo	Média	Desvio Padrão	Número de Valores Distintos	Número de Valores Únicos

DataSet_Treino	151098 (35%)	1	25038	12184.131	7297.774	25037	1645
DataSet_Validacao	17264 (38%)	47	27318	25236.189	4087.158	2440	170

Estatísticas de customerNo 16

maxVal: Apresenta o valor máximo que o cliente pode gastar na loja. O domínio destes valores pertence aos inteiros não negativos ([0,50000]).

Nome do Ficheiro	Valores em Falta	Mínimo	Máximo	Média	Desvio Padrão	Número de Valores Distintos	Número de Valores Únicos
DataSet_Treino	153740 (36%)	0	50000	2486.358	3038.426	178	1
DataSet_Validacao	17452 (39%)	0	25000	2039.006	2157.554	80	3

Estatísticas de maxVal 17

customerScore: São valores que avaliam o cliente do ponto de vista da loja, sendo o domínio o dos números inteiros não negativos ([0,638]).

Nome do Ficheiro	Valores em Falta	Mínimo	Máximo	Média	Desvio Padrão	Número de Valores Distintos	Número de Valores Únicos
DataSet_Treino	153740 (36%)	0	638	485.298	131.028	307	2
DataSet_Validacao	17452 (39%)	0	614	481.639	141.709	228	5

Estatísticas de customerScore 18

accountLifeTime: Indica o tempo de vida da conta do cliente, e é representado em meses cujo domínio dos valores é o dos números inteiros não negativos ([0,600]).

Nome do Ficheiro	Valores	Mínimo	Máximo	Média	Desvio	Número	Número

	em Falta				Padrão	de Valores Distintos	de Valores Únicos
DataSet_Treino	153740 (36%)	0	600	135.557	109.577	462	0
DataSet_Validacao	17452 (39%)	0	524	129.781	104.308	367	4

Estatísticas de accountLifeTime 19

payments: Mostra o número de pagamentos previamente feitos pelo cliente, sendo os valores representados como números naturais não negativos ([0,868]).

Nome do Ficheiro	Valores em Falta	Mínimo	Máximo	Média	Desvio Padrão	Número de Valores Distintos	Número de Valores Únicos
DataSet_Treino	151098 (35%)	0	868	15.278	34.893	227	1
DataSet_Validacao	17264 (38%)	0	278	9.951	13.389	71	2

Estatísticas de payments 20

age: Apresenta a idade do cliente. O domínio é os números naturais ([17,99]).

Nome do Ficheiro	Valores em Falta	Mínimo	Máximo	Média	Desvio Padrão	Número de Valores Distintos	Número de Valores Únicos
DataSet_Treino	151396 (35%)	17	99	44.92	11.936	78	0
DataSet_Validacao	17282 (38%)	17	90	43.282	11.527	68	0

Estatísticas de age 21

address: Indica a forma de endereçamento ao cliente, e toma valores pertencentes ao domínio dos números naturais ([1,3]). Estes estão compreendidos entre 1 que corresponde a *Mr*, 2 que indica *Mrs* e 3 que é para *Company*.

Nome do Ficheiro	Valores	Mínimo	Máximo	Média	Desvio	Número	Número

	em Falta				Padrão	de Valores Distintos	de Valores Únicos
DataSet_Treino	151098 (35%)	1	3	1.735	0.444	3	0
DataSet_Validacao	17264 (38%)	1	3	1.744	0.439	3	0

Estatísticas de address 22

lastOrder, Mostra o tempo decorrido em dias (domínio dos números naturais, [3,738]) desde a última encomenda efectuada.

Nome do Ficheiro	Valores em Falta	Mínimo	Máximo	Média	Desvio Padrão	Número de Valores Distintos	Número de Valores Únicos
DataSet_Treino	151098 (35%)	3	738	79.884	113.202	677	4
DataSet_Validacao	17264 (38%)	4	733	94.915	125.051	401	10

Estatísticas de lastOrder 23

order: Apresenta o resultado da sessão do cliente, e toma valores nominais de *y* se existir uma compra ou *n*, se não existir compra nessa sessão.

Nome do Ficheiro	Valores em Falta	Contagem de *y* (encomenda)	Contagem de *n* (não encomenda)	Número de Valores Distintos	Número de Valores Únicos
DataSet_Treino	0	290030	138983	2	0

Estatísticas de order 24

Os valores dos intervalos aqui identificados abrangem os valores máximos e mínimos dos ficheiros de teste e de validação.

Os dados iniciais possuem vários valores indefinidos, que se encontram com "?" no local desses valores e vai-se partir do princípio que os dados estão corretos. A duração das sessões tem de ser sempre crescente de linha do Data Set para linha, pois o tempo em cada transação gravada está sempre a contar, e não pode voltar atrás no tempo de duração, a não ser que crie uma nova sessão depois de terminar a presente sessão.

O atributo *cCount* é um atributo que de transação para transação apenas pode ficar com o mesmo valor ou aumentar, nunca pode diminuir. Os valores mínimos de produtos clicados ou em carrinho de compras nunca podem superar os valores máximos, e os valores de Soma de valores (Sum) têm de ser sempre iguais ou superiores aos valores máximos. O tempo de vida da conta do cliente tem de ser sempre igual ou superior ao tempo de vida da conta da sessão anterior. O atributo *lastOrder* terá que ter sempre valores iguais ou superiores aos valores que possuía na sessão anterior.

Preparação dos dados

É necessário filtrar os 429013 registos de forma a perder-se o mínimo de informação. Para tal, existem três possíveis estratégias:

Eliminar-se todas as linhas correspondentes ao mesmo número de sessão deixando a última. Isto faz com que haja uma grande perda de informação sem que sejam examinados os dados;

Agregar-se todas as linhas cujo número de sessão é igual numa única. Aqui já há uma verificação de dados porque existem valores cuja soma fará adquirir-se informação errada. Contudo, existem dados que serão perdidos e que poderão ser úteis para o futuro;

Juntar toda a informação numa única linha que corresponda ao mesmo número de sessão e adicionar-se novos atributos para caracterizarem a informação que não se deseja perder na eliminação das linhas. Esta será a estratégia que vamos desenvolver ao longo deste projeto por ser aquela que no nosso entender menos perde informação.

Resumo das Iterações Conducentes à Iteração final

Ao longo deste trabalho fomos desenvolvendo diversas iterações que nos permitem tirar conclusões relativamente à questão em análise. De seguida passamos a descrever alguns passos e respectivas consequências:

Criação de oito novos atributos para o Availability - Indica a disponibilidade do produto em stock, nomeadamente: completely orderable, completely not orderable, mainly orderable, mixed, mainly not orderable, completely not determinable, mainly not determinable e Availability?. Efectuámos o somatório destes atributos por sessão e quanto mais alto for o valor do produto disponível, maior será a satisfação do Cliente (aumenta a probabilidade de haver compra). Perde-se a ordem em que aparece a disponibilidade dos produtos.

Transformação do atributo OnlineStatus em três novos atributos - É importante saber-se se o Cliente está online ou não em cada transacção. Neste caso para cada sessão conta-se quantas vezes é que aparece o OnlineStatusY, onlineStatusN e o onlineStatus? (indica que não se sabe o estado do cliente). Também é perdida a ordem dos dados.

Análise da Matriz de Correlação – Após a análise atenta da matriz de correlação optámos por não remover quaisquer atributos com relações elevadas porque os atributos em causa poderiam ter grande importância ao longo das iterações.

Iteração Final

Para a iteração que consideramos ter os melhores resultados realizaram-se os seguintes passos na preparação de dados:

Eliminação do *CustomerNo* - Este atributo corresponde a um ID e identifica os clientes. Quando se efetua a análise dos dados deve-se manter sempre a privacidade dos mesmos, para além de que este atributo faz precisamente o mesmo que o atributo *sessionNo*, pelo que resolvemos retirá-lo do dataset (consulte os ficheiros "Treino\01 - CostumerNo.csv" e "Validação\01 - CostumerNo.csv" que contêm esta alteração);

Eliminação dos atributos *cMinPrice*, *cMaxPrice* e *cSumPrice* - Os produtos em que o cliente clica demonstram meramente que há um interesse da parte do comprador num produto. Contudo, este pode não ficar satisfeito devido a algum aspeto (preço elevado, material de má qualidade, entre outros). Para o projeto interessa mais guardar os preços dos produtos que se encontram no carrinho de compras do que aqueles em qual o utilizador clicou. Por este motivo eliminámos estes atributos do dataset (consulte os ficheiros "Treino\02 - cPrices.csv" e "Validação\02 - cPrices.csv" que contemplam estas alterações);

Transformação do *bStep* em seis novos atributos candidatos para substituição do *bStep* original - Cada candidato refere-se a cada registo (*bStep1*, *bStep2*, *bStep3*, *bStep4*, *bStep5* e *bStep?*), sendo cada amostra o somatório de vezes em que o utilizador esteve num determinado passo naquela sessão. Esta transformação é interessante pois vai permitir fazer análises nas quais se conclui que quanto maior o somatório do *bStep4* ou *bStep5*, maior é a probabilidade de o utilizador efetuar uma compra, porque este mesmo se encontra nos últimos passos possíveis de uma encomenda na loja. A desvantagem deste passo é precisamente a perda da ordem dos passos do Cliente. Desta forma, adicionámos seis novos atributos ao dataset (verifique os ficheiros "Treino\03 - bStep.csv" e "Validação\03 - bStep.csv" que contêm as alterações);

Transformação do *Availability* – Face a este atributo original, optámos por colocar o valor mais frequente para cada sessão como resultado no dataset final. A razão pela qual fizemos esta transformação prende-se ao facto de que interessa mais saber o estado que ocorre a maior parte das vezes nas transações de um cliente do que todas as que acontecem, pois se os produtos que o cliente estiver a ver estiverem maioritariamente disponíveis, então nesse caso a probabilidade de compra também aumenta. Nos casos em que existirem valores em que aparece o mesmo número de vezes, é selecionado o primeiro (por predefinição do Microsoft Excel). Isto faz com que a informação apresentada seja um pouco imprecisa e para além disto perde-se a ordem em que aparecem os registos por cada sessão (consulte os ficheiros "Treino\04 - Availability.csv" e "Validação\04 - Availability.csv" para verificar os resultados);

Criação de *numTransactions* - Este novo atributo é importante para se saber quantas transações é que fez um cliente em cada sessão. Quanto mais alto for o valor, mais ativo está o Cliente no site, o que por um lado pode mostrar uma maior indecisão na compra, mas que maioritariamente mostra que o cliente está ativo no site, e procura aquilo que realmente quer comprar. Foi por este facto que se decidiu adicionar este novo atributo (consulte os ficheiros "Treino\05 - numTransactions.csv" e "Validação\05 - numTransactions.csv" onde consta esta informação);

Eliminação de todas as transações exceto a última por cada sessão - Pretende-se apresentar o último valor de cada atributo constante ou crescente de transação para transação, bem como do *onlineStatus* por sessão de forma a obtermos 50000 amostras no ficheiro de treino e 5111 no de validação. Após isto adicionámos todos os atributos e registos dos atributos constantes nos ficheiros mencionados anteriormente, o que dá origem ao dataset final preparado pelo Microsoft Excel (consulte os ficheiros "Treino\06 - Treino.csv" e "Validação\06 - Validação.csv").

Modelização

Dada por concluída a fase de preparação de dados, é necessário obter o modelo para implementar neste projeto. Para construir e aplicar este mesmo modelo utilizou-se o Weka 3.7.

Após a obtenção dos datasets finais de treino e de validação, procedeu-se à operação de discretização dos atributos numéricos que apresentavam um elevado número de valores distintos. Procedeu-se a esta discretização de valores devido à má ambientação dos algoritmos a um grande número de valores distintos de valores numéricos. Os atributos do dataset final que foram discretizados foram os atributos *duration*, *bMinPrice*, *bMaxPrice*, *bSumPrice*, *maxVal*, *customerScore*, *accountLifetime*, *payments* e *lastOrder*. Depois de aplicarmos a discretização efetuou-se a construção do modelo usando o algoritmo J48, parametrizado com 0.25 fator de confiança, 3 numFolds, *useLaplace* a true e *useMDLcorrection* a true com opções de teste *Cross-*

Validation 10 folds. Este foi o melhor resultado obtido após três iterações de projeto, e como tal foi selecionado para modelo final.

O resultado deste algoritmo deu um *ROC Area* (AUC) de 0.957. As instâncias classificadas corretamente foram 92.314% dos casos (46157).

Avaliação

Como foi dito na secção 2 de Compreensão do Negócio, o critério de sucesso principal deste projeto é o lucro e como tal é necessário efetuar uma análise de custos do modelo obtido no ponto anterior.

Com vista a satisfazer o critério de sucesso deste projeto, calculou-se a matriz de custos real do modelo. No contexto do problema em mãos, e tendo em conta também os valores dos preços dos produtos que o utilizador coloca no carrinho de compras, decidiu-se criar a seguinte matriz de custos real:

	TRUE	FALSE
TRUE	-50 (VP)	-20 (FN)
FALSE	50 (FP)	0 (VN)

Matriz de custos real

Os verdadeiros positivos desta matriz (VP), significam que o modelo faz um previsão de compra, e o utilizador de facto efetua uma compra. O valor de -50 para esta célula da matriz significa que o modelo ao pensar o que foi descrito acima, dá um lucro de 50€ à loja online. Os verdadeiros negativos desta matriz (VN), dizem que o modelo faz uma previsão de não compra, e o utilizador não efetua uma compra. Do ponto de vista da loja online, pensar que não se vende um produto e não o vender, não acarreta nem prejuízo nem lucro e portanto o valor desta célula é 0. Os falsos negativos desta matriz (FN) dizem que o modelo faz uma previsão de não compra e o utilizador faz uma compra. Esta situação é rentável para a loja online, no entanto, nunca vai ser um valor superior nem igual ao valor da célula dos verdadeiros positivos (VP), pois para que a venda tenha sido efetuada, pode ter sido executada qualquer tipo de promoção, desconto, oferta, que baixa o lucro da venda associada, devido aos gastos que se tem em tais ofertas. Portanto, o valor da célula é -20. Os falsos positivos (FP) significam que o modelo prevê que se vai efetuar uma compra e essa mesma compra não é na realidade efetuada. Esta é a situação mais prejudicial para a loja online, pois perde-se dinheiro investido na suposta venda. Como tal esta célula da matriz é a mais penalizada, com um valor de 50.

Para além da matriz de custos real, é necessário obter a matriz de custos normalizada, e para tal basta colocar a 0 as células dos verdadeiros (positivos e negativos), somando ou subtraindo um determinado valor

aos dois valores de cada linha da matriz de custos real. Neste problema em concreto foi necessário adicionar 50 à primeira linha, e na segunda não se mexeu. O resultado da matriz de custos normalizada foi o seguinte:

	TRUE	FALSE
TRUE	0 (VP)	30 (FN)
FALSE	50 (FP)	0 (VN)

Matriz de custos normalizada

O próximo passo passou por utilizar o algoritmo *CostSensitiveClassifier*, para se poder analisar os custos associados a este modelo. A razão de ter sido efetuada esta análise deve-se ao facto que um modelo para ser posto em produção, necessita de dar o maior lucro possível à empresa, e tem também de possuir uma percentagem de acerto de encomendas boa ou muito boa. Como parâmetros do algoritmo acima citado, utilizou-se o J48 parametrizado da mesma forma como foi imediatamente antes. Foi associada também a matriz de custos normalizada do projeto a este algoritmo. Depois de parametrizar o algoritmo foi necessário adicionar a matriz de custos real na opção *more options*, usando o *cost-sensitive evaluation*. Os resultados desta análise de custos deram num custo total associado ao modelo de -982.580, e num *ROC Area* de 0.956, com 91,9% de instâncias corretamente classificadas.

Conclusão

Após as etapas da metodologia CRISP-DM estarem realizadas, e se ter obtido o algoritmo que melhor satisfazia o critério de sucesso (lucro), pode-se concluir que o melhor modelo foi obtido usando o algoritmo J48, em detrimento de outros algoritmos que foram corridos como o Multilayer Perceptron. Conclui-se também que em relação à preparação de dados, o dataset que produziu melhores resultados foi o dataset onde foram eliminados os campos *customerNo*, *cMinPrice*, *cMaxPrice*, *cSumPrice*, e onde foram alteradas as colunas *Availability* (para o valor mais frequente) e *bStep* (separado em 6 colunas) e adicionada a coluna *numTransactions*. Discretizar os dados do dataset é importante para este projeto, porque existem alguns campos numéricos que possuem muitos valores distintos, mas no entanto quando se discretizou demasiados campos do dataset, os resultados pioraram. Ou seja, deve-se discretizar as colunas que possuem muitos valores numéricos distintos, mas apenas as necessárias, o que neste projeto se traduziu em mais do que 200 valores distintos.

Glossário

Ficheiro de treino: As redes neurais são treinadas apenas através do ficheiro de treino. Contudo, existe a necessidade de se saber o momento de parar com a fase de treino, pois existe o perigo de um "sobre-treino" da rede e, neste caso, a rede perde o seu poder preditivo.

Ficheiro de validação: Para contornar o perigo de memorização por parte da rede, confronta-se o modelo gerado com a base de validação.

Dados nominais: Os dados são expressos numa escala nominal quando cada um deles for identificado apenas pela atribuição de um nome que representa uma categoria. Qualquer dado pertence a uma categoria, estas devem ser exaustivas, mutuamente exclusivas (cada dado pertence a uma só categoria) e não ordenáveis (não é estabelecida preferência de uma classe em relação a outra). Exemplos de características definidas em escalas nominais são: a religião, a raça, a localização geográfica, o sexo, a profissão, tipo de residência, preferências, ocupações, etc.

Números inteiros não negativos: São os números positivos incluindo o zero. Na sua representação devemos colocar o + ao lado do Z. Z + = { 0,1 ,2 ,3, 4,...} O Conjunto Z + é igual ao Conjunto dos N Números naturais.

Bibliografia

Dados nominais:
http://liveeducation.wordpress.com/2007/02/21/escalas-de-medidas-de-variaveis/

Números inteiros não negativos:
http://www.mundoeducacao.com/matematica/conjunto-dos-numeros-inteiros.htm

Ficheiros de teste, validação e treino:
http://webcache.googleusercontent.com/search?q=cache:LTI15YxU3CcJ:www.inf.ufsc.br/~ogliari/arquivos/capitulo9masupervisionados.ppt+&cd=3&hl=pt-PT&ct=clnk&gl=pt&client=firefox-a