2013

**Grupo 12**

**Relatório Loja Online**

**Modelo de vendas de uma loja online**

**André Rosado – 2110125;Paulo Vieira - 2110117**

Índice

[Compreensão do Negócio 3](#_Toc376970925)

[Determinar Objectivos de Negócio: 3](#_Toc376970926)

[Avaliar o cenário 3](#_Toc376970927)

[Objectivos de Data-Mining 5](#_Toc376970928)

[Plano de projecto 5](#_Toc376970929)

[Compreensão dos Dados 6](#_Toc376970930)

[Recolha dos dados iniciais 6](#_Toc376970931)

[Descrição dos dados 6](#_Toc376970932)

[Tratamento de dados 12](#_Toc376970933)

[Modelização 13](#_Toc376970934)

[Avaliação 14](#_Toc376970935)

Compreensão do Negócio

Determinar Objectivos de Negócio:

**Descrição do Negócio:**

**Através deste tópico vamos tentar compreender e analisar os objectivos do data-mining.**

**Este projecto consta numa loja de vendas onde se pretende analisar se um cliente irá comprar algum produto ou não, onde iremos analisar essa acção do cliente através do modelo de data-mining, podendo através disso, posteriormente, avaliar quais os melhores produtos a apresentar/recomendar a cada cliente.**

**O sucesso será garantido quando as nossas previsões se conseguirem aproximar com o que irá acontecer na realidade e realmente conseguirmos prever se um utilizador irá comprar algum produto ou não.**

**Objectivos:**

1. Definir qual/quais os atributos que mais se relacionam com a característica objectivo, isto é, se o cliente efectua uma compra ou não;

**Critérios de Sucesso:**

1. **Verificar se um cliente numa determinada sessão irá comprar ou não algum produto;**

Avaliar o cenário

**Inventário de Recursos, requisitos, assunções e restrições:**

* Computador;
* RapidMiner;
* Weka;
* Excel;
* Enunciado do Projecto;

**Riscos e contingências:**

* Desistência de um ou ambos os membros do grupo;
* Não conseguir atingir os objectivos agendados;

**Terminologia:**

**Custo e Benefícios:**

Para a realização deste projecto foi-nos fornecido um dataset com dados de clientes de uma loja virtual online. O Data set apresenta dados de 50.000 sessoes, sendo constituído por dados da sessão.

Para este cenário definimos a seguinte terminologia:

* **Verdadeiro Positivo (VP)**, acontece quando o modelo sugere que o cliente quer efectuar uma compra e, este, realmente efectua a compra;
* **Verdadeiro Negativo (VN)**, surge quando o modelo declara que o cliente não quer comprar e ele não efectua essa compra.
* **Falso Positivo (FP)**, o sistema afirma que o cliente quer comprar mas de facto não compra.
* **Falso Negativo (FN)**, quando o cliente quer comprar mas o sistema não o sugere.

Para uma percepção mais real dos valores, geramos a seguinte grelha com os valores que nos constaram mais apropriados.

|  |  |  |
| --- | --- | --- |
|  | Verdadeiro | Falso |
| Verdadeiro | (VP)-> -110 | (VN)-> -90 |
| Falso | (FN)-> 100 | (FP)-> 0 |

Tab.1 – Matriz de Custos;

Objectivos de Data-Mining

**Determinar objectivos de Data-Mining:**

Melhorar as vendas das lojas em 10%;

**Critérios de sucesso de Data-Mining:**

Taxa de Acerto;

Custo Negativo;

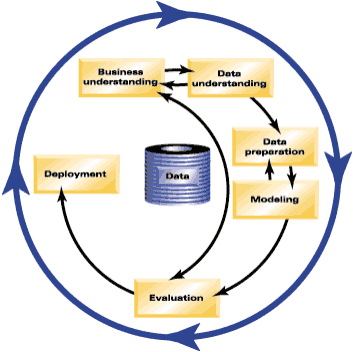
Maior Valor de ROC Area;

Plano de projecto

**Plano e metodologia:**

Para a resolução do problema optou-se pela metodologia **Cross Industry Standard Process for Data Mining (*CRISP-DM*).** Estando o CRISP-DM estruturado em 6 fases distintas, torna-se, desta forma, mais simples fazer a organização do trabalho e também existe uma fácil movimentação entre as diversas fases.

O seguinte diagrama ilustra esta metodologia:



**Determinação inicial de ferramentas e técnicos:**

Excel, RapidMiner, Weka, Notepad++.

Compreensão dos Dados

Recolha dos dados iniciais

Os dados para a realização deste estudo de data mining já foram anteriormente recolhidos e fornecidos pelo docente da UC. Estes dados estão em dois ficheiros um de testes e outro de treino.

Descrição dos dados

No dataset atribuído para treino temos, sem qualquer tipo de tratamento, temos 429013 amostras com 24 atributos cada e no dataset de teste temos 45068 amostras e 23 caracteristicas (não contém característica objectivo). Vamos apenas focar no dataset de treino pois será o que iremos usar para construir o nosso modelo.

**SessionNo**

Característica que indica qual o número da sessão da amostra. Existem 50000 sessões diferentes. Característica que será bastante importante na nossa análise pois é o que nos vai fazer diferenciar cada acesso. Este caracteristica é um inteiro.

**StartHour**

Característica que indica a hora de inicio da sessão. Este caracteristica está em horas [inteiros de 0 a 23] e permite-nos verificar que as horas de maior acesso são entre as 9 e as 21 e o pico encontra-se às 17h. Podemos também observar que o valor é igual dentro da sessão. Não parece ser relevante para o caso de uso.

**StartWeekDay**

Esta característica serve para indicar o dia em que se iniciou a sessão. Como só temos valores inteiros entre 5 e 7 podemos partir do principio que o dataset foi retirado num fim de semana, o valor de domingo tem menos amostras pois terminou às 18h e não foi um dia completo. Não varia dentro da sessão. Este caracteristica não parece ser relevante para o caso de estudo.

**Duration**

Característica que serve para indicar a duração de uma sessão. Valor numérico que varia entre 0 e 21580,092 com valor médio de 1573,902 e que varia dentro da sessão. Esta característica parece ser importante nos extremos, quem demora pouco tempo não compra nada, tal como quem demora demasiado tempo (podendo também ser dados incorrectos).

**cCount**

Característica numérico que indica no numero de produtos que o cliente clicou. Este Característica varia entre 0 e 200 no entanto a média é apenas de 24 e o desvio padrão de 30. Este Característica varia dentro das sessões explicando esta média baixa (ao inicio da sessão haverá menos cliques). Será importante guardar o valor máximo dentro de cada sessão pois este será o seu real valor. Característica que parece ser importante.

**cMinPrice**

Característica numérico que indica qual o artigo de menor valor que o cliente visualizou. Este Característica varia dentro da sessão como tal média e desvio padrão não serão certos sem tratamento. Será importante guardar o valor mínimo dentro de cada sessão.

**cMaxPrice**

Característica numérico que indica qual o artigo de maior valor que o vliente visualizou. Tal como o cMinPrice sem tratamento não se consegue tirar grandes conclusões. No entanto deverá-se guardar o maior valor.

**cSumPrice**

Característica numérico que indica a soma dos valores de todos os artigos visitados. Enquadra-se na mesma situação dos dois Característica s anteriores e será importante guardar o maior valor.

nota: Observou-se que os três Característica s anteriores tem o mesmo numero de valores em falta (2760)

**bCount**

Característica numérico que conta o número de artigos colocados no cesto de compras. Varia dentro da sessão logo será incerta uma análise estatística. Deverá guardar-se o maior valor.

**bMinPrice**

Característica numérico que indica o menor dos valores de todos os artigos colocados no cesto de compras. Tal como bCount varia dentro da sessão logo será incerta uma análise estatística. Deverá guardar-se o menor valor.

**bMaxPrice**

Característica numérico que indica o maior dos valores de todos os artigos colocados no cesto de compras. Incerto como o anterior. Importante guardar o maior valor dentro da sessão.

**bSumPrice**

Característica numérico que indica a soma dos valores de todos os artigos colocados no cesto de compras. Varia dentro da sessão logo será incerta uma análise estatística. Deverá guardar-se o maior valor.

nota: Observou-se que os três Característica s anteriores tem o mesmo numero de valores em falta (2760)

**bStep**

Este Característica indica o passo de compra em que se encontra a sessão. Varia entre inteiros de 0 a 5 podendo aumentar ou diminuir ao longo de uma sessão. Curioso que apenas 15706 chegam ao passo 5. Tem 45% valores em falta, no entanto poderá ser reduzido no tratamento de dados (valores a null e atribuídos na mesma sessão). Possivelmente será importante guardar a sessão mais alta que se chega ou então um Característica que indica se chegou a um certo passo.

**OnlineStatus**

Característica nominal que indica se o cliente está online ou não. Dentro de uma sessão este Característica varia. No entanto apenas 3009 amostras atingem o não. Tem 37% de valores em falta. Característica que não parece ser relevante.

**Availability**

Característica de valor nominal que pode ser completely orderable, completely not orderable, mixed, mainly not orderable, completely nor determinable ou mainly not determinable. Valor que indica se o produto que o cliente está a visualizar está disponível para entrega ou não. Este Característica varia dentro da sessão. Será importante guardar quais as “availability” que foram visitadas dentro da sessão. Este Característica tem 39% de valores em falta, no entanto dentro de uma sessão pode ter valores atribuídos e valores a null fazendo este valor reduzir quando do tratamento de dados. Característica importante.

**CostumerNo**

Característica numérico que indica o número de cliente. Característica totalmente inútil para o caso de estudo.

**MaxVal**

Característica numérica que indica o valor máximo admissível para o preço de uma compra. Valor que varia entre 0 e 50000 com média 2486 e o mesmo valor em todas os Características de uma sessão

**CostumerScore**

Caracteristica que indica a avaliação do cliente do ponto de vista da loja. Esta característica é numérica e varia entre 0 e 683 com média de 483 e desvio padrão de 131, mostrando que a maioria das amostras se encontram nos valores finais. Caracteristica importante.

**accountLifetime**

Esta característica representa o tempo de vida das contas dos utilizadores das sessões em valores numéricos inteiros (supomos que sejam dias). Podemos observar que a maioria dos dados se encontram com valores menores, no entanto entre os 219 e os 223 é onde encontramos o maior número de sessões. Este atributo varia dentro de uma sessão e tem 36% dos valores em falta.

**Payments**

Número de pagamentos que um cliente realizou até ao momento. Este é um valor numérico inteiro e apesar do máximo ter um valor alto (868) a sua média é relativamente baixa (34.89) e o desvio padrão pequeno (15.22),logo podemos verificar que a maioria realizou um número baixo de pagamentos.

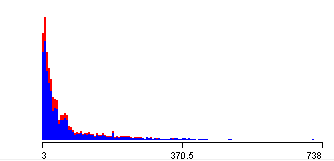
**Age**

Atributo onde se verifica a idade dos clientes da sessão. A maioria dos utilizadores estão entre os 30 e os 50 anos. Atributo não parece ser determinante para categorizar a característica objectivo.

**Address**

Esta característica indica o sexo do cliente ou se é uma empresa. O valor varia entre 1 2 ou 3, sendo que 1 corresponde a Mr. (masculino), 2 a Mrs (feminino) e 3 a Company (empresa). Com esta característica podemos observar que o publico alvo é o publico feminino, no entanto, não parece ser uma característica determinante.

**lastOrder**

Indica o tempo decorrido em dias desde a última compra. Este atributo varia entre 3 e 738 (nenhum cliente com uma compra à menos de 3 dias) em valores inteiros e tem uma média de 78.88, no entanto através de observação do gráfico podemos reparar que a moda é bastante mais baixa e que se encontra entre os 8 e os 13 dias.

**Order**

É a característica objectivo para o caso de uso. Pode ter os valores de y ou n. Podemos verificar que existem bastantes mais valores y do que n (290030 contra 138983, mais do dobro) no entanto como existem várias sessões repetidas e as sessões com valor y parecem repetir-se mais vezes estes dados com certeza serão uma falácia e na realidade teremos mais n do que y.

Observações:

Através da observação dos dados chegámos à conclusão que os dados de: costumerNo, maxVal, customerScore, accountLifeTime, Payments, Age, Address e lastOrder (as informações pessoais) apenas existem quando existe lastOrder.

Existem lastOrder sem payments (poderia ser possível no entanto o lastOrder já tem bastante tempo) e lastOrder superior a accountlifetime.

Data Set Treino

Tabela informação valores numéricos

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Caracteristica** | **Média** | **Máximo** | **Minimo** | **Desvio padrão** | **Valores em falta (%)** |
| **sessionNo** | 25274.63 | 50000 | 1 | 14441.37 | 0(0%) |
| **startHour** | 14.62 | 23 | 0 | 4.49 | 0(0%) |
| **startWeekday** | 5.93 | 7 | 5 | 0.791 | 0(0%) |
| **Duration** | 1573.9 | 21580.09 | 0 | 2427.12 | 0(0%) |
| **cCount** | 24.14 | 200 | 0 | 30.4 | 0(0%) |
| **cMinPrice** | 55.23 | 5999.99 | 0 | 148.88 | 2765(1%) |
| **cMaxPrice** | 146.66 | 6999.99 | 0 | 283.22 | 2765(1%) |
| **cSumPrice** | |  |  | | --- | --- | |  | 1189.248 | | |  |  | | --- | --- | |  | 115742 | | 0 | 3371.17 | 2765(1%) |
| **bCount** | 4.14 | 108 | 0 | 4.45 | 0(0%) |
| **bMinPrice** | 67.63 | 6999.99 | 0 | 174.97 | 5130(1%) |
| **bMaxPrice** | 107.505 | 6999.99 | 0 | 212.92 | 5130(1%) |
| **bSumPrice** | 213.26 | 23116.88 | 0 | 459.39 | 5130(1%) |
| **bStep** | 2.29 | 5 | 1 | 1.31 | 191333(45%) |
| **customerNo** | 12184.13 | 25038 | 1 | 7297.77 | 151098(35%) |
| **maxVal** | 2486.36 | 50000 | 0 | 3038.43 | 153740(36%) |
| **customerScore** | 485.3 | 638 | 0 | 131.03 | 153740(36%) |
| **accountLifetime** | 135.56 | 600 | 0 | 109.58 | 153740(36%) |
| **Payments** | 15.22 | 868 | 0 | 34.9 | 151098(35%) |
| **Age** | 44.92 | 99 | 17 | 11.94 | 151396(35%) |
| **Address** | 1.735 | 3 | 1 | 0.44 | 151098(35%) |
| **lastOrder** | 79.88 | 738 | 3 | 113.20 | 151098(35%) |

Tabela informação valores nominais

|  |  |  |  |
| --- | --- | --- | --- |
| **Característica** | **Y** | **N** | **Valores em falta** |
| **onlineStatus** | 265625 | 3009 | 160379(37%) |
| **Order** | 290030 | 138983 | 0(0%) |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Caracteristica** | | **Completely orderable** | **Completely not orderable** | | **Mainly orderable** | | **Mainly not orderable** | | **Completely not determinable** | | **Mainly not determinable** | | **Mixed** | | **Valores em falta** |
| **Availability** | | 253692 | 1491 | | 5756 | | 320 | | 1017 | | 198 | | 1284 | | 165255 (39%) |
| **Caracteristica** | | | **Média** | | **Máximo** | | **Minimo** | | **Desvio padrão** | | **Valores em falta (%)** | |
| **sessionNo** | | | 2385,701 | | 5111 | | 1 | | 1426,191 | | 0(0%) | |
| **startHour** | | | 18,58951 | | 23 | | 0 | | 5,543469 | | 0(0%) | |
| **startWeekday** | | | 6,458418 | | 7 | | 1 | | 1.719 | | 0(0%) | |
| **Duration** | | |  | |  | |  | |  | |  | |
| **cCount** | | | 27.317 | | 200 | | 0 | | 32.549 | | 0(0%) | |
| **cMinPrice** | | | 53.299 | | 1999.99 | | 0 | | 146.775 | | 326(1%) | |
| **cMaxPrice** | | | 149.135 | | 4799 | | 0 | | 272.25 | | 326(1%) | |
| **cSumPrice** | | | |  |  | | --- | --- | |  | 1240.986 | | | |  | | --- | | 117310.7 | | | 0 | | 3523.665 | | 326(1%) | |
| **bCount** | | | 4.367 | | 43 | | 0 | | 4.334 | | 0(0%) | |
| **bMinPrice** | | | 65.936 | | 1999.99 | | 0 | | 174.341 | | 598(1%) | |
| **bMaxPrice** | | | 105.75 | | 2299.99 | | 0 | | 204.395 | | 589(1%) | |
| **bSumPrice** | | | 414.708 | | 8948.96 | | 0 | | 414.708 | | 589(1%) | |
| **bStep** | | | 2.241 | | 5 | | 1 | | 1.285 | | 20766(46%) | |
| **customerNo** | | | 25236.189 | | 27318 | | 47 | | 4087.158 | | 17264(38%) | |
| **maxVal** | | | 2039.006 | | 25000 | | 0 | | 2157.554 | | 17452(39%) | |
| **customerScore** | | | 481.639 | | 614 | | 0 | | 111.709 | | 17452(39%) | |
| **accountLifetime** | | | 129.781 | | 524 | | 0 | | 104.308 | | 17452(39%) | |
| **Payments** | | | 9.951 | | 278 | | 0 | | 13.389 | | 17264(38%) | |
| **Age** | | | 43.282 | | 90 | | 17 | | 11.527 | | 17282(38%) | |
| **Address** | | | 1.744 | | 3 | | 1 | | 0.439 | | 17282(38%) | |
| **lastOrder** | | | 94.915 | | 733 | | 4 | | 125.051 | | 17264(38%) | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Característica** | **Y** | **N** | **Valores em falta** |
| **onlineStatus** | 27494 | 216 | 17355(39%) |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Caracteristica** | **Completely orderable** | **Completely not orderable** | **Mainly orderable** | **Mainly not orderable** | **Completely not determinable** | **Mainly not determinable** | **Mixed** | **Valores em falta** |
| **Availability** | 26283 | 144 | 654 | 144 | 90 | 2 | 102 | 165255 (39%) |

Tratamento de dados

Para a resolução deste projecto já tínhamos um dataset preparado para nós iniciarmos a resolução do projecto.

Dos 21 atributos existentes (excluindo o objectivo e o id) apenas foram seleccionadas 8 características e acrescentadas novas 2 caracteristicas. As características seleccionadas foram as seguintes: bSumPrice, bMaxPrice, bCount, bStep, duration, StartHour, age, address, lastOrder. Estes foram os atributos considerados importantes a manter, deste modo conseguimos excluir a correlação (bMaxPrice tinha uma correlação altíssima com vários atributos), manter a coerência do modelo (não fazia muito sentido termos os atributos que o cliente clicou e que o cliente adicionou ao carrinho, ele para adicionar tinha de clicar, no entanto para o caso de estudo pareceu mais pertinente manter os adicionados ao carrinho) e excluir dados com vários missings, incoerência e incorrecções (payments, accountlifetime…).

Para podermos começar a tratar os dados primeiramente tínhamos de juntar os diversos atributos que correspondiam à mesma interacção (sessionId igual). Para isso decidmos manter o maior valor detectado de bMaxPrice, bCount, bStep, bSumPrice e duration, uma vez que nestes dados dentro de uma transacção o que faz sentido é guardar o maior. Foi guardado também a soma dos valores de bStep visto que um utilizador pode variar o bStep e o simples fato de ter chegado a um certo nível de bStep não é suficiente para verificar tudo.

Para melhorar o nosso modelo foram construídos dados na nossa opinião iriam ajudar a atingir a característica objectivo. Estas características foram o num\_transacções e o fez\_encomenda. O num\_transacções foi feito pela contagem de SessionId repetidos e significa o número de transacções(~cliques) que o utilizador fez no site, esta característica surgiu pois verificamos que utilizadores com poucas transacções normalmente não tinham a característica objectivo n. A outra característica adicionada foi criada após observação dos campos de last\_Order, onde outros campos dependiam dela e que apenas existiam se ela existisse(dados como payments, accountlifetime, maxVal, age e address) .

Para optimizar os nossos dados, visto que muitos deles tinham um enorme número de dados diferentes criamos classes de dados, estas classes foram feitas nos dados que achamos relevante esta operação (bMaxPrice, bSumPrice, bCount, Age e startHour). Para definição dos intervalos das classes de bMaxPrice, bSumPrice, bCount calculámos dez percentis e depois verificamos os valores para definir as classes, alguns atributos tinham valores de percentis repetidos então o numero de classes não era alto. Para a idade foi divido em idades conforme o grupo etário (< 25 < 32 < 45 < 60 < 100) e para a hora dividimos em 3 classes que representam aproximadamente a manhã a tarde e a noite.

Modelização

Para o processo de modelização o modelo escolhido foi o j48 do weka com 0.40 de confidence level e com um minNumObj de 10.

O algoritmo de j48 do weka (normalmente chamado C4.5) é uma arvore de decisão que utiliza a razão entre o ganho e a entropia do conjunto de amostras, este modelo. Este algoritmo é benéfico para o nosso projecto pois ele permite lidar com valores em falta e valores nominais e ordinais.

A opção de confidence factor é o valor usado no pruning, basicamente o pruning irá remover os nós estatisticamente inúteis, para isto ele avalia de baixo para cima a probabilidade de cada nó verificando as probabilidades dos nós mais superiores através da probabilidade dos nós filhos. Posto isto podemos concluir que o confidence factor quanto mais baixo (reduz a tolerância) menos folhas irão existir. Já a opção de minNumObj define um numero mínimo de instâncias por folha, logo quanto maior mais instâncias serão necessárias para cada folha tornando a árvore mais simples.

Para testar o modelo foi utilizado cross-validation 10 folds.

O resultado foi uma árvore com 473 folhas e um tamanho de 648. Como esta árvore é demasiado complexa iremos apenas apresentar o maior ramo da árvore.

Transactions\_num >4 🡪Maximum bStep >3 🡪Sum bStep >2 🡪 Efetou\_encomenda > 0 🡪Maximum bStep <= 4 🡪 Sum bStep > 11 🡪Transactions\_num >8 🡪Sum bStep > 20 🡪Transactions\_num > 14 🡪 Sum bStep > 29 🡪Transactions\_num > 18 🡪 Maximum bMaxPrice = ]3,4] 🡪Maximum bCount = ]8,9] 🡪 Transactions\_num > 25 🡪 y || < 25 🡪n

Obtendo os seguintes resultados

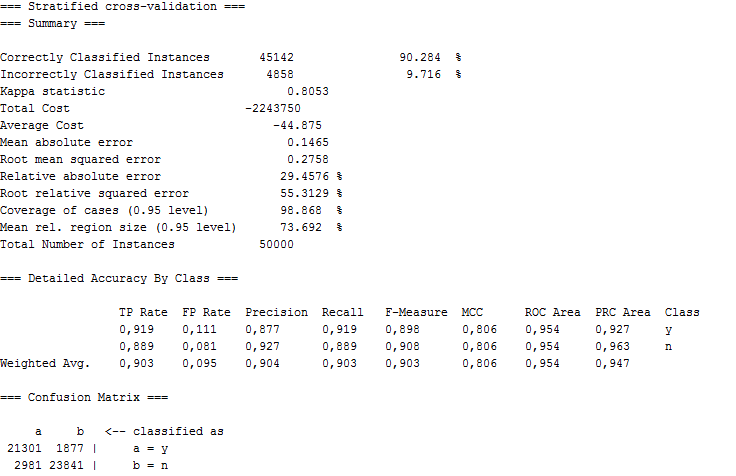
Este modelo apesar de ser possível analisar e compreender é muito complexo, no entanto é possível visualizar um ramo caso seja preciso

Avaliação

De acordo com os critérios de sucesso definidos nos objectivos do processo de negócio, entende-se por sucesso quando se consegue classificar se um cliente efectua uma compra ou não.

Pode, agora, afirmar-se que o objectivo em questão foi cumprido, pois o modelo que foi seleccionado satisfaz o requisito que foi estabelecido.

Resumindo o modelo mostra-se um bom classificador que poderá ser aplicado noutros data sets da empresa que o aplicar.



A matriz de custos feita para este dataset foi:

-110 -90

90 0  
  
De seguida uma explicação da escolha dos valores:

VP -> O utilizador comprou logo tivemos lucro, o valor era para ser 100, no entanto com a recomendação o utilizador pode ter comprado outro produto ou um produto mais caro (o contrário também se pode verificar, logo lucro -110.

VN -> O sistema sugeriu que o utilizador não ia comprar e efectivamente não efectou compra, custo 0.

FN -> Era esperado que não existisse uma compra no entanto o utilizador comprou, como lhe foi dado um desconto o utilizador fez uma compra mas deu menos lucro pois obteve um desconto. A diferença é curta pois normalmente estes descontos nunca são muito elevados. Custo de -90.

FP -> Era esperado que o utilizador efectuasse uma compra e no entanto não efetou, logo perdemos o possível lucro de uma compra com desconto caso lhe oferecêssemos um desconto, por isso temos um custo de 90.

Com isto obtivemos o seguinte resultado na nossa matriz

=== Confusion Matrix ===

a b <-- classified as

21301 1877 | a = y

2981 23841 | b = n

Podemos observar que não houve uma grande discrepância nos valores falsos, uma vez que apesar de ser prejudicada no fp não é um valor muito discrepante. Então o modelo não vai pender muito para esse lado, apesar de se notar uma ligeira piora no ROC Àrea.

O nosso resultado de custo foi o seguinte.

Total Cost -2243750

Average Cost -44.875