Curso de Engenharia de Sistemas e Informática - 5º Ano

Ficha T. Prática n.º 8

Objectivo:

Tomar contacto a descrição relativa a um caso prático de utilização de ferramentas de Data Mining, como preliminar para a efectiva utilização prática de uma ferramenta.

Avaliação de riscos de empréstimo: Um caso de estudo de Data Mining.

Aspectos Genéricos

Data Mining é definido como o "***processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis nos dados***" (Advances in Knowledge Discovery and Data Mining, U.M. Fayad et al., MIT Press, 1996).

Data Mining descobre frequentemente padrões que predizem comportamentos futuros. Desta forma é utilizado em actividades como banca, telecomunicações, retalho e distribuição, marketing e seguros.

O caso em estudo, está relacionado indirectamente com uma das actividades referidas: banca.

Recomenda-se uma leitura prévia do artigo disponível na página da disciplina. Não só é descrito o caso mas são descritas algumas técnicas que foram utilizadas.

Resumo do caso em análise:

O Departamento de Agricultura dos Estados Unidos (USDA) administra um programa de empréstimos sob hipoteca, a pessoas em áreas rurais, em número de 600,000. O departamento mantém informação extensiva acerca de cada caso, num Data Warehouse. Tal como em outros programas de empréstimos, alguns têm melhor desempenho do que outros.

O USDA escolheu o data mining para ajudar a melhor compreender esses empréstimos, melhorando assim a gestão do seu programa de empréstimos e reduzir a incidência de problemas relativos a incumprimento de compromissos relativos a empréstimos concedidos: é necessário o data mining para encontrar padrões que distingam devedores que cumprem os compromissos daqueles que ficam em falta. A esperança é que esses padrões possam predizer quando um devedor vai entrar em problemas.

Neste caso, o objectivo é um tanto diferente da banca comercial: esta última utiliza o data mining para avaliar da concessão ou não do empréstimo, aquando do processo de avaliação (antes de ser concedido o empréstimo); no caso da USDA, o interesse principal, recai na previsão de problemas em empréstimos já concedidos (em vigor) e, assim, devotar mais atenção e assistência a esses possíveis futuros devedores, reduzindo, dessa forma, a possibilidade desses empréstimos se tornarem problemas.

O USDA contratou uma empresa de consultoria para efectuar um estudo preliminar,

utilizando dados extraídos do DW - uma amostra constituída por 12,000 registos relativos a empréstimos para casas unifamiliares (cerca de 2% do número total de registos).

A amostra de dados contém informação acerca:

- do empréstimo, tais como: montante, valor da mensalidade, data de empréstimo e propósito;
- da propriedade, tais como: tipo de moradia e tipo de propriedade;
- da pessoa a que é concedido o empréstimo, tais como: idade, raça, estado civil e categoria de rendimento;
- da região onde o empréstimo é realizado, incluindo o estado e a presença de minorias nesse estado.

Utilização dos Algoritmos

Neste caso, o objectivo era criar um modelo que predissesse a classificação do empréstimo, baseado em informação sobre o empréstimo, pessoa a que foi concedido e propriedade.

Para maximizar a processamento e eficiência da obtenção de resultados, utilizaram-se diversos algoritmos em conjunto. Data a velocidade de execução e interpretabilidade do algoritmo de Naïve-Bayes, utilizou-se para exploração inicial; seguiu-se a aplicação de algoritmos de árvores de decisão e redes neuronais.

Para a criação de um modelo predictivo, as ferramentas de data mining precisam de exemplos: dados que contenham resultados conhecidos. Através do processo chamado de aprendizagem, indução ou treino, faz-se a auto-aprendizagem de como predizer o resultado de um dado processo de transacção.

A coluna de dados que contém o valor resultado - também o valor que eventualmente desejamos prever - tem nomes como: variável dependente, alvo ou de saída. Todas as outras variáveis são denominadas de atributos ou variáveis independentes ou de entrada.

No caso em estudo, a variável dependente do modelo de classificação de empréstimos tinha cinco valores: sem problemas, substandard, de perda, não classificado e não disponível. Cerca de 80% dos casos caem na 1ª categoria.

O Data Mining consiste num ciclo de geração, teste e avaliação de muitos modelos.

Criação de Modelos e Base de dados de Teste

Criaram-se os modelos utilizando 2/3 dos dados - 8000 registos - deixando-se os restantes como um conjunto independente para teste dos modelos. Os testes revelam quão bem um modelo prediz a variável de saída - neste caso a classificação do empréstimo.

Desta forma, aplicando os casos de teste ao modelo gerado, para realizar a classificação de cada empréstimo e, comparando com o valor real, pode aferir-se da precisão da previsão.

O primeiro modelo criado, tinha um fraco desempenho, com uma precisão de previsão de cerca de 50%. Este resultado desanimador forçou a uma observação mais meticulosa de algumas variáveis não categóricas, como o empréstimo e montante de mensalidade. Descobriu-se que a distribuição distorcida desses valores afectavam negativamente o modelo. O montante do pagamento era um bom exemplo desse efeito: embora muito poucos empréstimos tivessem um valor elevado de mensalidade (até $60,000), a maioria requeria pagamentos inferiores a $400.

Ora o algoritmo de Naïve-Bayes utilizado, requerendo “binning” dos valor numéricos, providenciava a sua efectivação automática em cinco intervalos. Como os valores iam até aos $60,000, resultava em intervalos de $12,000, motivando que quase 99% dos empréstimos caíssem no 1.º intervalo (0-12000), dada a distribuição não normal e não uniforme de valores

das mensalidade. Desta forma, o predictor revelava-se pobre, pois que, embora o data mining se utilizasse para olhar e revelar padrões, neste caso, os intervalos eliminavam realmente um.

Redesenharam-se os intervalos, por forma a que cada um contivesse cerca de 1/5 do total da população. A precisão da previsão melhorou grandemente: 67% geral, elevando-se até 76% na previsão nas categorias sem problemas e de perdas. Esta melhoria mostrava claramente que os intervalos que a ferramenta definia por defeito eliminava padrões importantes.

Poda de valores Irrelevantes

Os valores de precisão de previsão obtidos, eram demasiado bons para durarem: encontrou-se nos dados da amostra um campo "valor total do empréstimo em dívida" que, quando um empréstimo entra em não cumprimento, vai crescendo sucessivamente, à media que mais pagamentos ficam em falta. Claro que o modelo iria utilizar esse campo como um predictor excelente para empréstimos substandard e de perdas. Ou seja, o modelo gerado não era muito útil pois que era baseado em informação pós-incumprimento.

Removendo esse campo, a precisão geral caiu para 46% e a precisão de previsão da categoria perdas, caiu para 37%: passou-se de resultados demasiado bons para serem verdade, para resultados sofríveis.

Nova observação dos dados. Focou-se a atenção nas próprias classes de classificação: duas delas - não identificado e não disponível, ocorriam em menos de 1% dos casos. Não havendo interesse nessa classe de predições, decidiu-se pela sua eliminação, com a correspondente remoção das linhas que os contivessem. Ficaram assim três classes possíveis: sem problemas, substandard e perdas. Mas como o objectivo era apenas prever que empréstimos poderiam requerer atenção especial, combinaram-se as duas últimas, numa única classe: Not OK, ficando a outra OK, para consistência de terminologia.

À primeira vista, o modelo agora gerado - com uma precisão geral de 82% - parecia muito bom. Contudo, um exame mais apertado mostrou que só predizia 20% de todos os empréstimos com problemas, ou seja a classe Not OK, apresentava um precisão desapontadoramente baixa. Sendo a classe mais importante relativamente à tomada de acções (em empréstimos problemáticos), este desempenho mostrava que o nosso modelo requeria novos refinamentos.

Refinamento com Árvore de Decisão

Depois da exploração inicial dos dados com o algoritmo de classificação de Naïve-Bayes, também foi treinado um modelo de árvores de decisão.

Como é comum, este revelou uma melhor precisão, aumentando a precisão geral para 85% e a da classe Not OK para os 23%.

Apesar da precisão ser bastante modesta para a classe Not OK, o resultado global do estudo preliminar não foi totalmente desanimador. A precisão não é, em si própria, o objectivo do estudo. Quando se pretende baixar os custos relativos às perdas, mesmo uma precisão baixa, pode traduzir-se em benefícios significativos. Senão vejamos:

Assuma-se que, em média, cada problema custa $5000 e que são encontrados 50,000 problemas / ano. Se a intervenção atempada puder prevenir 30% desses casos, e cada intervenção custar $500, o USDA pode poupar $11.5 milhões / ano, mesmo no caso do data mining antecipar apenas 23% dos casos Not OK. Este valor é um pouco menor, pois que há que levar em consideração os custos de intervenção nos casos em que não se revelariam problemas e que foram na mesma objecto de intervenção especial. Considerando estes, ficaríamos com uma poupança de $9.1 milhões, ainda um valor considerável, apesar da baixa

precisão da previsão.

Neste estudo preliminar, os modelos iniciais mostraram quais os factores importantes a considerar nos empréstimos. Também demonstrou o potencial da tecnologia como capacidade de previsão e de aprendizagem.

No futuro próximo, o departamento planeia expandir a número limitado de atributos disponíveis para data mining, em particular, incluindo história de pagamentos no DW, esperando-se que se possa assim melhorar a precisão do modelo.

Adaptado de Assessing Loan Risks: A Data Mining Case Study, Rob Gerritsen
http://www.exclusiveore.com/CaseStudies/DM at USDA (ITPro).pdf
Também disponível na página da disciplina.

Questões

1. O que distingue, na sua essência, o objectivo do caso de estudo dos casos gerais, relativos a empréstimos? Justifique essa abordagem diferente.

2. Além da classificação, uma melhor compreensão dos casos é, também, um resultado das técnicas de Data Mining, sendo até um óbice em algumas. Fundamente um e outro, referindo o óbice apontado.

3. Em sua opinião, porque enveredou o USDA pela contratação de serviços a uma empresa de consultoria e qual a razão de ser encetado o estudo preliminar? Fundamente a sua resposta.

4. Foi disponibilizada uma amostra de 12,000 registos, dos 600,000 possíveis. Terá constituído uma estratégia correcta ou amostras porventura maiores teriam obtido maior desempenho? Fundamente a sua resposta, eventualmente tecendo algumas considerações acerca da amostragem, suas técnicas e defensores / detractores do seu uso.

5. É dito no texto que foram utilizados os algoritmos Naïve-Bayes, árvores de decisão e redes neuronais. Porquê os três em conjunto e não apenas um (aquele que, em teoria, se mostrasse mais adaptado ao problema em questão)?

6. Neste caso trata-se de aprendizagem supervisionada. Qual o significado de supervisionada, neste contexto? Será por obrigar a um envolvimento profundo do analista no processo de data mining? Fundamente a sua opinião.

7. A certo ponto do texto, é dito que “o data mining consiste num ciclo de geração, teste e avaliação de muitos modelos”. Comente a afirmação.

8. Os 12,000 registos disponibilizados foram separados em dois sets: 2/3 e 1/3. Qual o intuito e importância da separação?

9. Um dos recuos do processo da criação do modelo predictivo foi motivado pela utilização não lícita de um campo predictor relativo a “valor total do empréstimo em dívida”. Porque era ilícita a sua utilização para a indução do modelo?

10. Também o “binning” se revelou de suma importância. Porquê?

11. Apesar da precisão global ser bastante boa, a precisão relativamente aos empréstimos Not OK, a mais importante neste âmbito, sendo baixa, traduziu-se num desapontamento final? Dê a sua opinião, fundamentando-a.

12. O processo de criação de modelos e teste foi conhecendo vitórias e reveses sucessivos. Que lição poderia tirar desses avanços e recuos?

13. Que vantagens se obtiveram deste estudo preliminar?

14. Especule acerca de previsíveis desenvolvimentos futuros.