Exemplo de Aplicação do DataMinig

Felipe E. Barletta Mendes

19 de fevereiro de 2008

INTRODUÇÃO AO DATA MINING

A mineração de dados (Data Mining) está inserida em um processo maior denominado Descoberta de Conhecimento em Bancos de Dados, o KDD. Rigorosamente o Data Mining se restringe a obtenção de modelos, ficando as etapas anteriores como coleta e seleção dos dados, pré-processamentos dos dados, transformação dos dados e o próprio DM como instâncias do KDD.

Sem a estatística não seria possível termos o Data Mining, visto que a mesma é a base a partir da qual o Data Mining é construído. Existe muito conhecimento escondido na imensa quantidade de dados disponíveis nos bancos de dados das empresas. Com o Data Mining , pode-se transformar esses dados brutos em informação valiosa para auxiliar o processo decisório. Data Mining não substitui técnicas estatísticas tradicionais. Ao invés disto, Data Mining é uma extensão dos métodos estatísticos. O poder cada vez maior dos computadores com custos mais baixos, aliado à necessidade de análise de enormes conjuntos de dados com milhões de linhas, permitiu o desenvolvimento de técnicas baseadas na exploração de soluções possíveis. As técnicas de DM podem ser aplicadas em diversas áreas:

- Vendas e Marketing
- Finanças
- Seguros e Planos de Saúde
- Transporte
- Medicina
- Telecomunicações
- Mercado Financeiro

Vantagens do Data Mining

O uso de Data Mining pode trazer as seguintes vantagens:

- Modelos são de fácil compreensão: pessoas sem conhecimento estatístico (por exemplo, analistas financeiros ou pessoas que trabalham com data base marketing) podem interpretar o modelo e compará-lo com suas próprias idéias. O usuário ganha mais conhecimento sobre o comportamento do cliente e pode usar esta informação para otimizar os processos dos negócios.
- Grandes bases de dados podem ser analisadas: grandes conjunto de dados, de até vários gigabytes de informação podem ser analisados com Data Mining.
- Data Mining descobre informações não esperadas: como muitos modelos diferentes são validados, alguns resultados inesperados podem surgir. Em diversos estudos, descobriu-se que combinações de fatores particulares apresentaram resultados inesperados.

- Variáveis não necessitam de recodificação: Data Mining lida tanto com variáveis numéricas (quantitativas) quanto categóricas (qualitativas). Estas variáveis aparecem no modelo exatamente da mesma forma em que aparecem na base de dados.

- Modelos são precisos: os modelos obtidos por Data Mining são validados por técnicas de estatística. Desta forma, as predições feitas por modelos são precisas.

CLASSIFICAÇÃO DE CRÉDITO EM UM BANCO

Mil clientes de um banco solicitaram o uso de crédito ao banco, e de acordo com a fidelidade de pagamento dos clientes, eles receberam um rótulo de inadimplentes ou não, ou seja, criou-se uma variável dicotômica, com as categorias crédito bom e crédito ruim.

Iniciamos o trabalho com o objetivo de ajustar um modelo que fosse capaz de predizer a categoria dos clientes, para o banco poder decidir se ia conceder ou não crédito ao ou não. Para tal, foram consideradas cerca de vinte variáveis acerca dos clientes do banco tais como:

- Saldo da conta do cliente

- Histórico do Crédito

- Propósito do crédito

- Quantidade de crédito disponível

- Dinheiro em poupança

- Tempo no presente emprego
- Sexo, idade e estado civil
- Tempo na atual residência
- Número de créditos no banco
- Profissão
- Idade da conta
- Outros planos de parcelamento
- Taxa de parcelamento em relação à renda líquida

Vale ressaltar que da amostra total, 70% dos clientes tinham crédito bom e 30% tinham crédito ruim.

Portanto após seleção dos dados aplicamos o método da Regressão Logística, pois a variável de interesse é categórica, neste caso dicotômica. Em mineração de dados é comum ajustar diversos tipos de modelos, aqui poderia ajustar também modelos baseados em métodos de Árvores de Classificação, mas aqui vamos analisar apenas o primeiro.

Após definir alguns parâmetros para se ajustar o modelo como método de seleção de variáveis, ponte de corte para classificação, etc e validação do modelo obtemos uma matriz de confusão que nos dá a análise de quão bom o modelo ficou para se classificar corretamente os clientes, ou seja, porcentagem de erro de classificação, assim constatamos se o poder de generalização ou predição do modelo foi satisfatório. Outra forma de verificar a qualidade do modelo é o ROI (Return Of Investiment).

A matriz de confusão apresentou um erro de 27% quando o ponto de corte foi o menor possível, pois pontos de cortes grandes resultavam em erros maiores.

O ROI, após definição de alguns parâmetros como, taxa de juros ao mês, custo fixo, etc, apresentou um maior retorno de investimento quando o ponto de corte foi o maior aceitável, cerca de $48,42\%$ de retorno do dinheiro investido.

A partir desses dois resultados vemos que para errar menos o banco deve emprestar mais dinheiro (classificar clientes como bom pagador). No caso do ROI, é melhor emprestar menos para não desperdiçar dinheiro, e assim, maximizar o lucro com relação às despesas.

Referências

[1] BRAGA, L. P. V. **Introdução à Mineração de Dados**. Rio de Janeiro, E-papers Serviços Editoriais, 2005