1º Trabalho de Data Mining II (Modelos Preditivos)

Mestrado em Gestão de Informação, Mestrado em Estatística e Gestão de Informação

O problema da Companhia de Seguros

A companhia de seguros "Sinto&Such Pensórios" contratou a sua empresa de consultadoria em datamining (a "ED - Eigealunos Duiseji") para desenvolver uma aplicação informática que permita aos vendedores ter uma estimativa do "valor" de um cliente. Um cliente é tão mais valioso quanto menos probabilidade tiver de se ver envolvido num acidente durante o ano seguinte. Depois de ter perdido muito tempo a pensar sobre o que é que aumenta ou diminui a probabilidade de uma dada pessoa ter um acidente de automóvel, descobre que a companhia de seguros já tem uma base de dados relativa aos seus clientes, onde para além de uma série de dados quanto às suas características, tem um campo que indica se tiveram ou não um acidente em que a companhia de seguros teve despesas.
Para cada cliente, a companhia tem na sua base de dados, para além da idade, os seguintes campos binários:

m35	1 se o cliente tem mais de 35 anos (0 em caso contrário)
m65	1 se o cliente tem mais de 65 anos
cas	1 se o cliente é casado
tf	1 se o cliente tem filhos
tc5	1 se o cliente tem carta há mais de 5 anos
ts3	1 se o cliente tem seguro nesta companhia há mais de 3 anos
tsr	1 se o cliente tem seguro contra roubo
tst	1 se o cliente tem seguro contra todos os riscos
sm	1 se o cliente é do sexo masculino
tcs	1 se o cliente tem curso superior
est	1 se o cliente é estudante
tcp	1 se o cliente tem casa própria
tmt	1 se o cliente tem múltiplos telemóveis
fum	1 se o cliente fuma
ta	1 se o cliente teve um acidente

Quere-se que o programa permita ao vendedor introduzir rapidamente as informações que dispõe sobre a pessoa a quem está a tentar vender uma apólice, e que o programa, usando a base de dados da empresa, preveja se essa pessoa vai ou não dar prejuízo e, se possível, qual a probabilidade de isso acontecer. Nos casos em que precisa de saber os custos de decisões erradas, a companhia informa-o que o custo (em lucros perdidos) por não tentar vender uma apólice a uma pessoa que seria um bom cliente é de 500, enquanto o custo de vender uma apólice a uma pessoa que é um mau cliente é de 600.
Para testar a sua capacidade, a companhia de seguros facultou-lhe uma base de dados com 1000 clientes (chamada "seguros"), e outra com 20 (chamada "prova"), onde ocultou o campo "ta".

1) Aparece um cliente que tem claramente menos de 35 anos, mas que usa aliança (é casado), e vem com um rapaz a que chama filho. Deve tentar vender-lhe uma apólice ?

2) Aparece um senhor que preenche a ficha de inscrição, e através dela fica a saber que ele tem 26 anos, não é casado, não tem filhos, tem a carta há menos de 5 anos, não tem nenhum seguro, tem curso superior e já não estuda, não tem casa própria mas tem múltiplos telemóveis, e não fuma. Decida se lhe deve ou não vender uma apólice, usando 6 dos seguintes tipos de classificadores:

    a. Um classificador MAP (sem naive Bayes)
    b. Um classificador MAP (com naive Bayes)
    c. Um classificador de Máxima Verosimilhança (com ou sem estimativas naive)
    d. Um classificador de vizinho mais próximo
    e. Um classificador de k-vizinhos mais próximos, com k=3
    f. Um classificador de k-vizinhos mais próximos, com k=4
    g. Um classificador de k-vizinhos mais próximos, com k=20
    h. Um classificador Bayesiano com custos
    i. Um classificador Bayesiano com "estimativas m"
    j. Um classificador por regressão logística
    k. Um classificador com um perceptrão simples
    l. Um classificador com uma rede neuronal multicamada
    m. Um classificador com uma árvore de decisão
    n. Um classificador linear de Fisher
    o. Outro, ou outros 2 classificadores que queira escolher

3) No caso da alínea anterior, qual seria a sua decisão final: tentava ou não vender a apólice ?

Bom trabalho !