Desafio ML# Data Preparation

1. Desafio lançado pelo prof.
   - Identificar padrões de clientes
2. PKDD'99 Discovery Challenge- Guide to the Financial Data Set. This database was prepared by Petr Berka and Marta Sochorova

Domínio

Era uma vez um banco que oferecia serviços a particulares. Os serviços incluem gerenciamento de contas, oferta de empréstimos etc. O banco deseja melhorar seus serviços encontrando grupos interessantes de clientes (por exemplo, para diferenciar entre bons e maus clientes). Os gerentes dos bancos têm apenas uma vaga idéia de quem é um bom cliente (a quem oferecer alguns serviços adicionais) e quem é um cliente ruim (a quem observar atentamente para minimizar as perdas do banco). Felizmente, o banco armazena dados sobre seus clientes, as contas (transações dentro de vários meses), os empréstimos já concedidos, os cartões de crédito emitidos. Assim, os gerentes dos bancos esperam encontrar algumas respostas (e perguntas também) analisando esses dados.

Descrição da tarefa

A tarefa do desafio de descoberta é:

Desafio_1:

1- Limpeza e tratamento de dados.

2- Definir um problema que possa ajudar o banco a melhorar seus serviços (por exemplo, definir a noção de um cliente bom ou ruim, sugerir um serviço novo / atual que possa ser oferecido a um grupo de clientes etc.); do ponto de vista do descoberta de conhecimento, o problema pode ser classificação, previsão ou descrição, e justificar a sua escolha.

3- Mostre como o ML pode ser usado para resolver o problema (mesmo que os resultados não sejam muito significativos do ponto de vista do ML, eles podem ser importantes para o banco).

Desafio_2:

1. Utilizam a mesma base de dados 'PKDD Discovery Challenge de 1999', para **prever o saldo médio por conta bancária.** O objetivo do projeto é prever o saldo médio por conta bancária, de forma a ajudar o banco a compreender quais são os clientes mais lucrativos, e para que consiga fazer marketing dos seus serviços direcionado a esses clientes.....

2. Podem ainda :
   -dizer quais os clientes possui cartões de crédito
   -quem pediu empréstimos ao banco
   -clientes menores de idade
   - numero de clientes por sexo

   - os tipos de cartão o banco oferece.

Descrição de dados

Os dados sobre os clientes e suas contas consistem nas seguintes relações:

- relation **account** (4500 objects in the file ACCOUNT.ASC) - each record describes static characteristics of an account,
- relation **client** (5369 objects in the file CLIENT.ASC) - each record describes characteristics of a client,
- relation **disposition** (5369 objects in the file DISP.ASC) - each record relates together a client with an account,
- relation **permanent order** (6471 objects in the file ORDER.ASC) - each record describes characteristics of a payment order,
- relation **transaction** (1056320 objects in the file TRANS.ASC) - each record describes one transaction on an account,
- relation **loan** (682 objects in the file LOAN.ASC) - each record describes a loan granted for a given account,
- relation **credit card** (892 objects in the file CARD.ASC) - each record describes a credit card issued to an account,
- relation **demographic data** (77 objects in the file DISTRICT.ASC) - each record describes demographic characteristics of a district.

Cada conta possui características estáticas (por exemplo, data de criação, endereço da agência) fornecidas em relação à "conta" e características dinâmicas (por exemplo, pagamentos debitados ou creditados, saldos) dadas em relação à "ordem permanente" e "transação". A relação "cliente" descreve as características das pessoas que podem manipular as contas. Um cliente pode ter mais contas, mais clientes podem manipular com uma única conta; clientes e contas são relacionados juntos em relação à "disposição". Relações "empréstimo" e "cartão de crédito" descrevem alguns serviços que o banco oferece a seus clientes; mais cartões de crédito podem ser emitidos para uma conta, no máximo um empréstimo pode ser concedido para uma conta. A relação "dados demográficos" fornece algumas informações publicamente disponíveis sobre os distritos (por exemplo, a taxa de desemprego); informações adicionais sobre os clientes podem ser deduzidas disso.

Relation account

Relation account

item	meaning	remark
account_id	identification of the account	
district_id	location of the branch	
date	date of creating of the account	in the form YYMMDD
frequency	frequency of issuance of statements	"POPLATEK MESICNE" stands for monthly issuance "POPLATEK TYDNE" stands for weekly issuance "POPLATEK PO OBRATU" stands for issuance after transaction

Relation client

item	meaning	remark
client_id	record identifier	
birth number	identification of client	the number is in the form YYMMDD for men, the number is in the form YYMM+50DD for women, where YYMMDD is the date of birth
district_id	address of the client	

Relation disposition

item	meaning	remark
disp_id	record identifier	
client_id	identification of a client	
account_id	identification of an account	
type	type of disposition (owner/user)	only owner can issue permanent orders and ask for a loan

Relation permanent order

item	meaning	remark
order_id	record identifier	
account_id	account, the order is issued for	
bank_to	bank of the recipient	each bank has unique two-letter code
account_to	account of the recipient	
amount	debited amount	
K_symbol	characterization of the payment	"POJISTNE" stands for insurrance payment "SIPO" stands for household "LEASING" stands for leasing "UVER" stands for loan payment

Relation Transaction

item	meaning	remark
trans_id	record identifier	
account_id	account, the transation deals with	
date	date of transaction	in the form YYMMDD
type	+/- transaction	"PRIJEM" stands for credit "VYDAJ" stands for withdrawal
		"VYBER KARTOU" credit card withdrawal "VKLAD" credit in cash
operation	mode of transaction	"PREVOD Z UCTU" collection from another bank "VYBER" withdrawal in cash "PREVOD NA UCET" remittance to another bank
amount	amount of money	
balance	balance after transaction	
		"POJISTNE" stands for insurrance payment "SLUZBY" stands for payment for statement
k_symbol	characterization of the transaction	"SANKC. UROK" sanction interest if negative balance "SIPO" stands for household "DUCHOD" stands for old-age pension "UVER" stands for loan payment
bank	bank of the partner	each bank has unique two-letter code
account	account of the partner	

Relation Loan

item	meaning	remark
loan_id	record identifier	
account_id	identification of the account	
date	date when the loan was granted	in the form YYMMDD
amount	amount of money	
duration	duration of the loan	
payments	monthly payments	
status	status of paying off the loan	'A' stands for contract finished, no problems, 'B' stands for contract finished, loan not payed, 'C' stands for running contract, OK so far, 'D' stands for running contract, client in debt

Relation Credit card

item	meaning	remark
card_id	record identifier	
disp_id	disposition to an account	
type	type of card	possible values are "junior", "classic", "gold"
issued	issue date	in the form YYMMDD

Relation Demographic data

item	meaning	remark
A1 = district_id	district code	
A2	district name	
A3	region	
A4	no. of inhabitants	
A5	no. of municipalities with inhabitants < 499	
A6	no. of municipalities with inhabitants 500-1999	
A7	no. of municipalities with inhabitants 2000-9999	
A8	no. of municipalities with inhabitants >10000	
A9	no. of cities	
A10	ratio of urban inhabitants	