Домашнее задание №5

Целью Домашнего задания №5 является освоение алгоритмов и методов прогнозирования для решения учебной задачи анализа данных в условиях, близких к реальным условиям, возникающим при решении прикладных задач анализа данных.

Формулировка задания:

Дан набор данных, в котором описана история предложений клиентам банка застраховать свои вклады. Целевая бинарная переменная INS, содержит признак, согласился ли клиент приобрести такую услугу или нет. Каждый клиент имеет свой уникальный ID. Остальные переменные – входные.

Определения метаданных источника указаны в таблице ниже:

Variable Name	Role	Measurement Level	Label
ACCTAGE	INPUT	INTERVAL	Age of Oldest Account
AGE	INPUT	INTERVAL	Age
ATM	INPUT	BINARY	ATM
ATMAMT	INPUT	INTERVAL	ATM Withdrawal Amount
BRANCH	INPUT	NOMINAL	Branch of Bank
CASHBK	INPUT	INTERVAL	Number Cash Back
CC	INPUT	BINARY	Credit Card
CCBAL	INPUT	INTERVAL	Credit Card Balance
CCPURC	INPUT	INTERVAL	Credit Card Purchases
CD	INPUT	BINARY	Certificate of Deposit
CDBAL	INPUT	INTERVAL	CD Balance
CHECKS	INPUT	INTERVAL	Number of Checks
CRSCORE	INPUT	INTERVAL	Credit Score
DDA	INPUT	BINARY	Checking Account
DDABAL	INPUT	INTERVAL	Checking Balance
DEP	INPUT	INTERVAL	Checking Deposits
DEPAMT	INPUT	INTERVAL	Amount Deposited
DIRDEP	INPUT	BINARY	Direct Deposit
HMOWN	INPUT	BINARY	Owns Home
HMVAL	INPUT	INTERVAL	Home Value
id	ID	NOMINAL	
ILS	INPUT	BINARY	Installment Loan
ILSBAL	INPUT	INTERVAL	Loan Balance
INAREA	INPUT	BINARY	Local Address
INCOME	INPUT	INTERVAL	Income
INS	TARGET	BINARY	Insurance Product
INV	INPUT	BINARY	Investment
INVBAL	INPUT	INTERVAL	Investment Balance
IRA	INPUT	BINARY	Retirement Account
IRABAL	INPUT	INTERVAL	IRA Balance
LOC	INPUT	BINARY	Line of Credit
LOCBAL	INPUT	INTERVAL	Line of Credit Balance
LORES	INPUT	INTERVAL	Length of Residence
MM	INPUT	BINARY	Money Market
MMBAL	INPUT	INTERVAL	Money Market Balance
MMCRED	INPUT	INTERVAL	Money Market Credits
MOVED	INPUT	BINARY	Recent Address Change
MTG	INPUT	BINARY	Mortgage
MTGBAL	INPUT	INTERVAL	Mortgage Balance

NSF	INPUT	BINARY	Number Insufficient Fund
NSFAMT	INPUT	INTERVAL	Amount NSF
PHONE	INPUT	NOMINAL	Number Telephone Banking
POS	INPUT	INTERVAL	Number Point of Sale
POSAMT	INPUT	INTERVAL	Amount Point of Sale
RES	INPUT	NOMINAL	Area Classification
SAV	INPUT	BINARY	Saving Account
SAVBAL	INPUT	INTERVAL	Saving Balance
SDB	INPUT	BINARY	Safety Deposit Box
TELLER	INPUT	INTERVAL	Teller Visits
dataobs	REJECTED	INTERVAL	Observation Number

Необходимо построить модель прогнозирования для бинарного отклика, которая будет наилучшим образом его предсказывать.

Оцениваться качество модели будет:

1) По критерию ROC Index (площадь под ROC кривой).

2) На тестовом наборе, где реальный отклик не будет известен студенту (вам), но будет известен проверяющему (мне).

3) Для зачета по заданию необходимо получить ROC на тестовом наборе больший или равный 0.777 (заметьте, что оценки на тестовом наборах могут сильно отличаться от качества на проверочной выборке, полученной из исходного набора).