30416 – Big Data and Databases

Fraud Prediction Analysis

Consider the datasets *"fraud_train.csv"* and *"fraud_test.csv"* that includes data from a retail bank's clients behavioral data. Some customers included in the datasets are related to fraudulent activity on their banking account (i.e. dummy variable FLG_FRAUD, where 1="fraudster" and 0="otherwise"). The aim of the bank is to predict possible frauds before they happen and take adequate countermeasures to block the most likely fraudsters.

Here's the full list of the fields in both the datasets:

1 ID
2 FLG_FRAUD
3 WEB_AVG_ACCESS_LASTMONTH
4 WEB_WEEK_CHANGE_INDEX
5 WEB_N_COOKIE_LASTDAY
6 WEB_N_COOKIE_LASTFIVEDAYS
7 WEB_N_ACCESS_TOT
8 WEB_N_ACCESS_LASTDAY
9 WEB_AVG_ACCESS_LASTWEEK
10 WEB_N_WEEKDAY_MAX
11 WEB_N_WEEKDAY_MIN
12 TENURE
13 FLG_PRIVACY
14 AGE
15 N_DEVICES
16 COD_PHONETYPE
17 FLG_AREA
18 FLG_PLACE_OF_BIRTH_HIRISK
19 FLG_PLACE_OF_BIRTH_HIRISK_2
20 FLG_EMAIL_DOMAIN_HIRISK
21 FLG_GENDER
22 FLG_FIRSTDEPOSIT_LOW
23 FLG_DEBITCARD
24 CURRENTACCOUNT_STD_DEV_LASTMONTH
25 CURRENTACCOUNT_AVG_TRANSACTIONS_LASTMONTH
26 CURRENTACCOUNT_N_TRANSACTIONS_LASTYEAR
27 CURRENTACCOUNT_AVG_AMOUNT_LASTYEAR

Perform the following tasks using Knime

1) Import the training data ("fraud_train.csv") and explore the FLG_FRAUD distribution in order to identify possible class imbalance problem; proceed to balancing if needed.
2) Train different logistic regression models, with and without regularization, together with one boosted tree model of your choice. Interpret the coefficients/odds ratio of the logit models.
3) Evaluate the models' performances on the test set ("fraud_test.csv") focusing on Sensitivity and AuROC measures.