Home Credit - Credit Risk Model Stability

Aim: To predict whether clients will default on their loan or not ⇒ Binary Classification.

Given: Train & Test Directories

↙ 32 csv files ↘ 36 csv files.

Depth Values

depth = 0 → Static features
depth = 1 → case_id has historical records indexed by num_group1
depth = 2 → case_id has historical records indexed by num_group1 & num_group2

num_group1 → } They define various different parameters of the
num_group2 → } same data

Internal Data Source Files

File Name	Dimensions	Null Columns	Col with Max NULL	File Description
Static_0_0	10L x 168	121	clientscnt_136L	
Static_0_1	5.2L x 168	119	clientscnt_136L	
applprev_1_0	38L x 41	32	revolvingaccount_394A	
applprev_1_1	26L x 41	17	revolvingaccount_394A	
other_1	51K x 7	0	N.A.	

File Name	Dimensions	NULL Columns	Col with max NULL	File Description
deposit_1	1.4L x 5	1	contractenddate_99D	Deposit Account opening data
person_1	29L x 37	20	housingtype_772L	Personal info of applicants like address, employment, education, etc
debitcard_1	1.5L x 6	4	last180dayturnover_1134A	D card opening date & turnovers
applper_2	1.4Cr x 6	3	credacc_card_status_52L	C. card status of prev. applications like card status, blocking reason
person_2	16L x 11	3	empls_employedfrom_796D	more personal info like employment status

External Data Source Files.

tax_registry_a_1	32L x 5	0	N.A.
tax_registry_b_1	11L x 5	0	N.A
tax_registry_c_1	33L x 5	0	N.A
credit_bureau_a_1_0	41L x 79	66	contractsum_5085717L
a_1_1	60L x 79	66	contractsum_5085717L
a_1_2	37L x 79	66	prolongationcount_599L
a_1_3	20L x 79	66	interestrate_508L

credit_bureau_2_0	52L X 19	10	collater_value of guarantee_876 L
_2_1	78L X 19	10	" " " "
_2_2	178L X 19	10	collater_value of guarantee_1124 L
_2_3	265L X 19	10	" " " "
_2_4	270L X 19	10	" " " "
_2_5	3.3Cr X 19	10	" " " "
_2_6	2.55Cr X 19	10	" " " "
_2_7	80.52 X 19	10	" " " "
_2_8	1.39Cr X 19	10	" " " "
_2_9	1.87Cr X 19	10	" " " "
_2_10	43.8L X 19	10	" " " "

Models That Can Be Used

1) Bagging (Random Forest) - Multiple DTs and then voting to predict the output

2) Boosting (Gradient Boosting) - LGBM or XGBoost - Builds successive DTs that improves performance of prev. models

3) Neural Networks - ANN or LSTM for time-series