Kaggle. Home Credit Default Risk

Евгений Патеха (3 место)

Конкурс на Kaggle Home Credit Default Risk

- Задача – предсказание вероятности невозврата кредитов
- Данные:

  Train / Test – **307 511 / 48 744**

  Доп. данные: **previous_application**

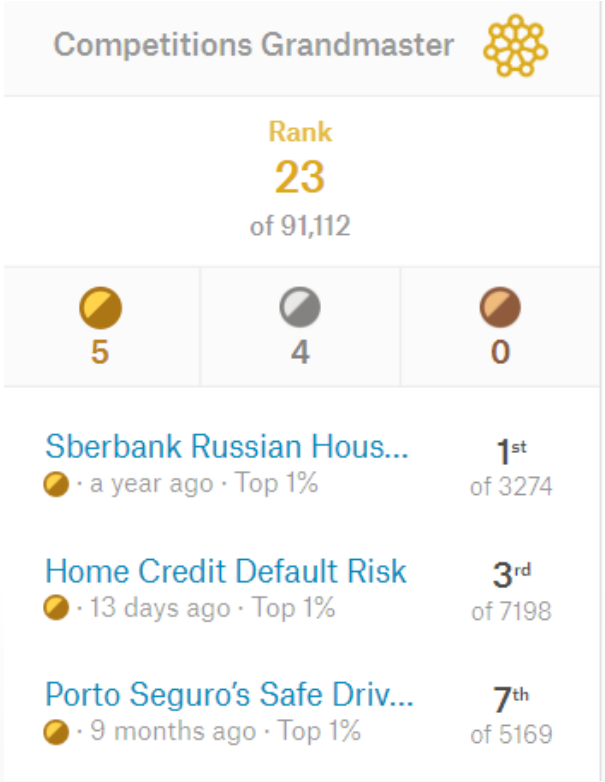
  POS_CASH_balance, credit_card_balance,

  installments_payments, bureau, bureau_balance
- Метрика – **ROC AUC**

  public lb – **20**%, private lb – **80**%

1	Home Aloan	0.80570
2	ikiri_DS	0.80561
3	**alijs & Evgeny**	**0.80511**
4	Quad Machine	0.80474
5	Kraków, Lublin i Zhabinka	0.80449
6	silver	0.80419
7	A.Assklou _ Aguiar	0.80396
8	七上八下	0.80376
9	International Fit Club	0.80374
10	Best Friend Forever: CV	0.80354

Kaggle и карьера в data science



Coursera + Kaggle

05.2016 Santander Customer Satisfaction
первое соревнование - public #45, private #549

12.2016 Santander Product Recommendation #7

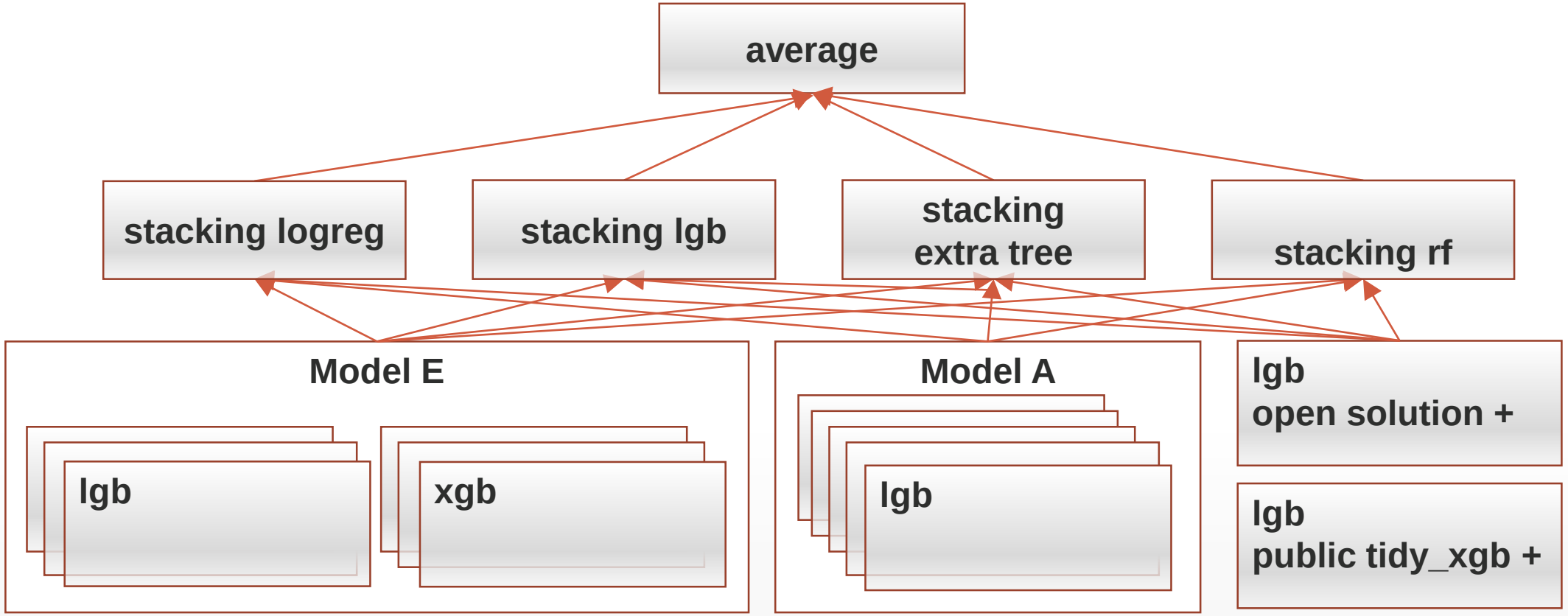
06.2017 Sberbank Russian Housing Market #1

10.2017 Банк Уралсиб. Аналитик

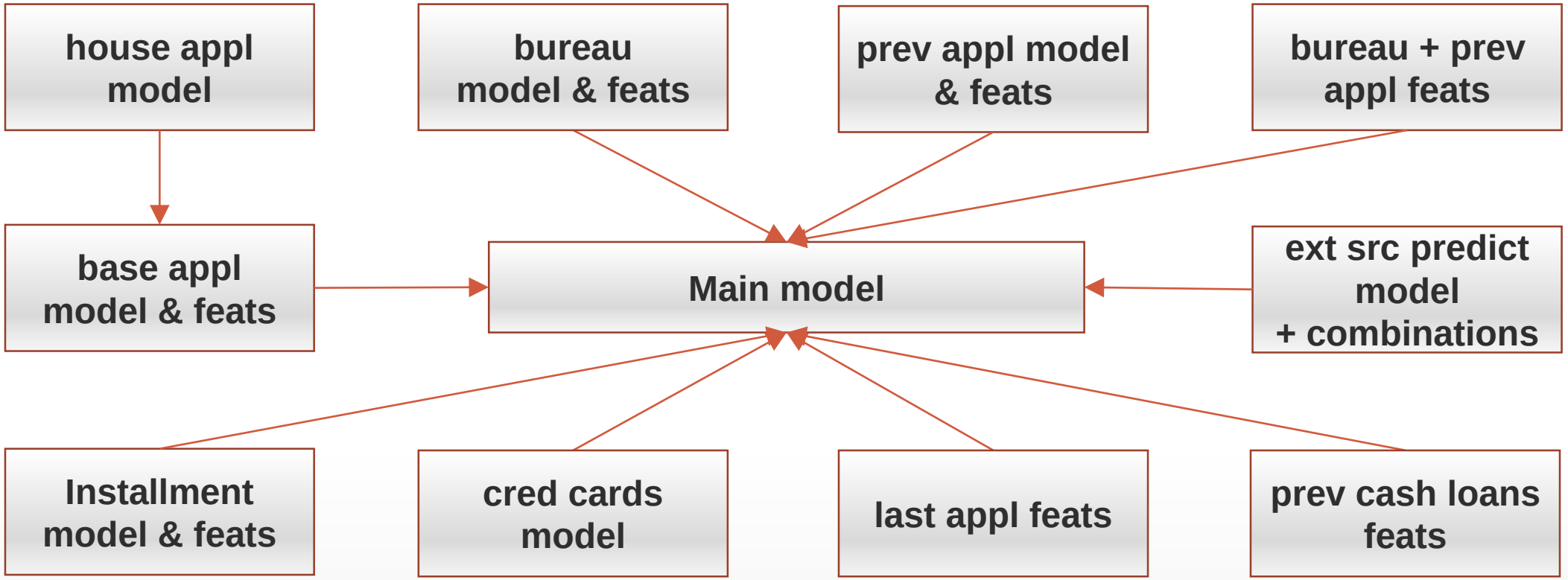
07.2018 QIWI. Аналитик в группе больших данных

08.2018 Home Credit Default Risk #3 **Grandmaster**

Архитектура решения. Стэкинг



Архитектура решения. Основная модель



Топ-20 переменных. Основная модель

1. ext_source_1_3_bki_prev_prod	11. bkiext_sum_lim_consum_cred
2. baseappl_pred	12. amt_annuity
3. ext_source_2_bki_prod	13. days_birth
4. inst_pred	14. prev_cash_open_bal_sum
5. prev_pred	15. ext_src2_err
6. ext_source_3	16. bkiext_avg_cur_debt_to_cred
7. product_combination	17. ext_source_1
8. bki_cur_debt_to_cred_max	18. days_employed
9. ext_src1_err	19. ext_src1_pred
10. bki_pred	20. bkiext_max_debt_to_cred

Топ-20 переменных. Base application

1. annuity_to_cred	11. name_education_type
2. days_employed	12. amt_annuity
3. goods_price_to_cred	13. region_population_relative
4. organization_type	14. region_rating_client_w_city
5. occupation_type	15. own_car_age
6. days_last_phone_change	16. code_gender
7. days_birth	17. amt_credit
8. days_id_publish	18. flag_doc1
9. house_pred	19. amt_income_total
10. amt_goods_price	20. amt_req_credit_bureau_year

Топ-20 переменных. Bureau

1. cur_debt_to_cred	11. days_credit_update_after_end_fact
2. days_credit	12. amt_credit_sum_limit
3. amt_credit_sum	13. amt_credit_sum_overdue
4. cred_len	14. total_dpd
5. max_overdue_to_cred	15. total_payments36
6. days_credit_enddate	16. total_dpd48
7. days_credit_update	17. total_dpd_we
8. days_enddate_dif	18. total_payments24
9. credit_type	19. total_dpd36
10. amt_credit_sum_debt	20. total_payments12

Топ-20 переменных. Previous applications

1. product_combination	11. days_last_due_1st_version
2. days_decision	12. inst_pmt_max_pmt_dif2_12
3. goods_price_to_cred	13. cnt_payment
4. inst_pmt_prc_adv_pmt	14. channel_type
5. code_reject_reason	15. annuity_to_cred
6. inst_pmt_max_pmt_dif	16. inst_pmt_sd_pmt_plan
7. cc_util	17. name_goods_category
8. hour_appr_process_start	18. name_yield_group
9. name_client_type	19. inst_pmt_oldest_plan_pmt
10. amt_annuity	20. inst_pmt_min_pmt_plan

Топ-20 переменных. Installments

1. inst_pmt_oldest_plan_pmt	11. inst_pmt_tot_pmt
2. inst_pmt_sum_pmt_fact	12. inst_pmt_max_pmt_plan_12
3. inst_pmt_max_pmt_dif2_12	13. inst_pmt_days_sd_dif_we_12
4. inst_pmt_status_m500_max_36	14. inst_pmt_status_m500_cnt5p_48
5. inst_pmt_min_pmt_plan	15. inst_pmt_sum_pmt_fact_6_12_dif1
6. inst_pmt_sd_pmt_plan	16. inst_pmt_avg_pmt_dif3_12
7. inst_pmt_max_pmt_dif	17. inst_pmt_prc_exact_pmt
8. inst_pmt_prc_adv_pmt	18. inst_pmt_amt_pmt_sumdif_6
9. inst_pmt_cnt_notlate_pmt_6	19. inst_pmt_max_nun_inst_vers_12
10. inst_pmt_avg_pmt_fact	

Топ-20 переменных. Credit cards

1. cc_util_max	10. cc_sum_payment_dif_minpayment
2. cc_avg_turnover	11. cc_min_lim
3. cc_months_balance_first	12. cc_avg_pay2limit_12
4. cc_max_payment_dif_minpayment	13. cc_sum_amt_drawings_atm
5. cc_min_payment_to_minpayment_12	14. cc_avg_pay2limit
6. cc_avg_oper_drawings_pos	15. cc_max_min_lim2
7. cc_avg_cnt_month_drawings_atm	16. cc_sum_dpd_def
8. cc_avg_oper_drawings_tot	17. cc_cur_to_min_lim
9. cc_max_util	18. cc_sum_payment_dif_minpayment

Топ-20 переменных. Model alijs

1. bki_pred	11. X3_CREDIT_MINUS_ANNUITY_GOODS
2. EXT_SOURCE_2	12. X1_INCOME_VS_EMPLOY
3. prev_pred	13. X3_SCORE_MUL_LOG
4. EXT_SOURCE_3	14. AMT_ANNUITY
5. NEW_EXT_SOURCES_MEAN	15. DAYS_EMPLOYED
6. inst_pred	16. bkiext_avg_cur_debt_to_lim
7. NEW_EMPLOY_TO_BIRTH_RATIO	17. bki_cur_debt_to_cred_max
8. EXT_SOURCE_1	18. bkiext_sum_lim_consum_cred
9. bkiext_avg_cur_debt_to_cred	19. bkiext_avg_days_enddate_dif
10. DAYS_BIRTH	20. DAYS_ID_PUBLISH

Комбинирование переменных. Аутлаеры

Комбинирование переменных с пропусками

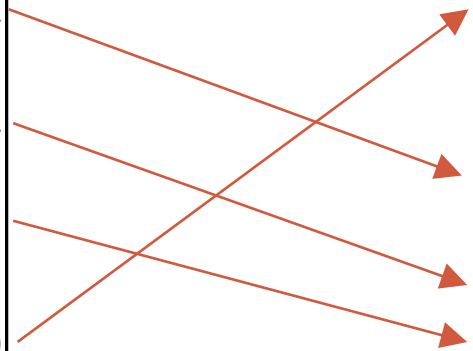
- Идея – создать новые переменные как сочетание действующих (ext_sources, метафичи)
- Проблема – пропуски. Простое сочетание увеличит число пропусков
- Решение – откалибровать переменные на уровень таргета (прогнать через логрег с одной переменной) и заполнить пропуски средним таргетом (для пропусков в этих переменных)

Аутлаеры

- Полезный трюк – выбросить из обучения аутлаеры – те строки, которые модель предсказывает с большой ошибкой (предикт <.01 при таргете 1 или >.9 при таргете 0)
- Результат – модель лучше отрабатывает ‘нормальные’ паттерны, растет качество (на сопоставимых полных данных включая аутлаеры)

Валидация и отбор переменных

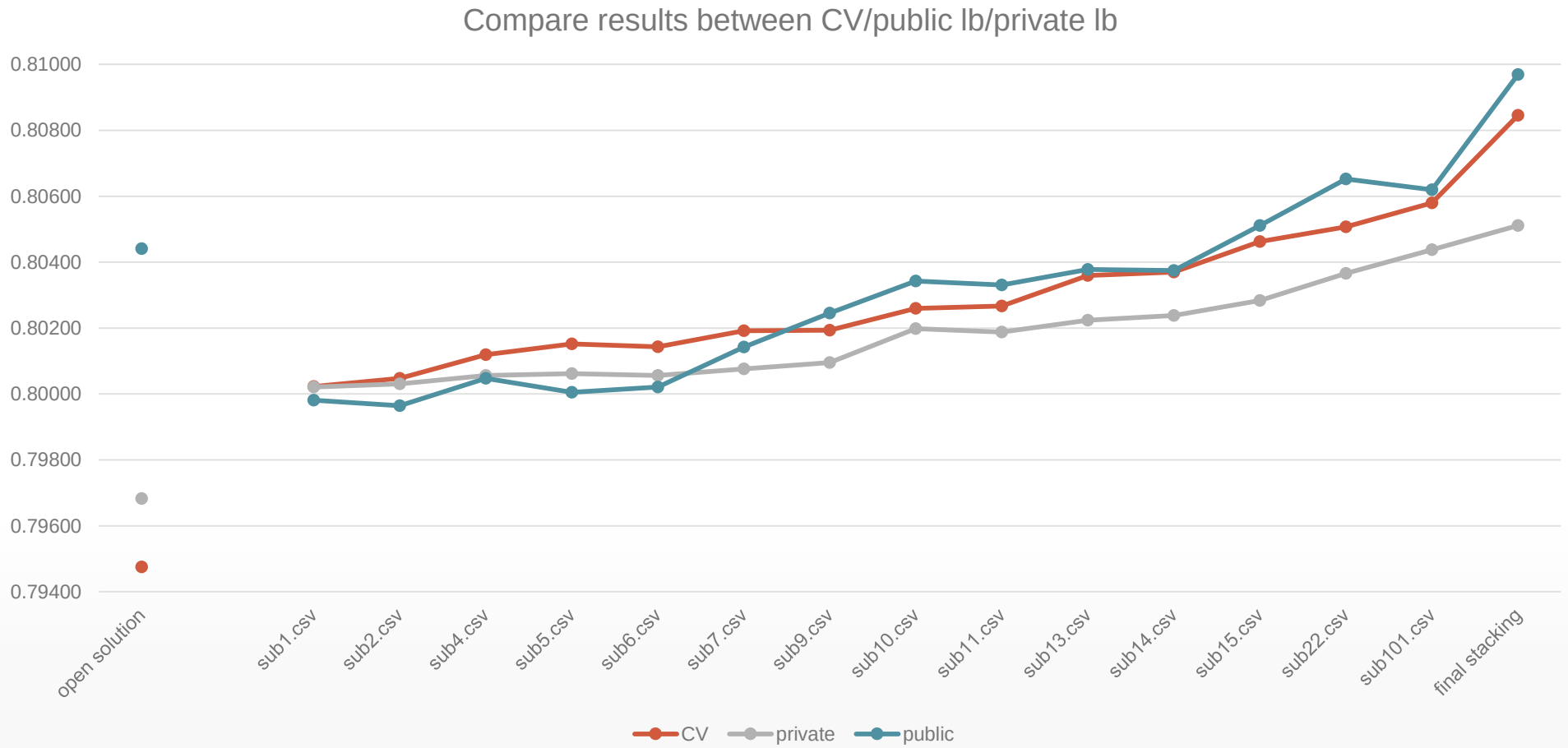
Public LB		
1	Kraków, Lublin i Zhabinka	0.81724
2	ikiri_DS	0.81241
3	circlecircle	0.81124
4	alijs & Evgeny	0.81086
5	Large hypothesis space	0.81041
…		
11	Home Aloan	0.80920



Private LB		
1	Home Aloan	0.80570
2	ikiri_DS	0.80561
3	alijs & Evgeny	0.80511
5	Kraków, Lublin i Zhabinka	0.80449
…		
11	circlecircle	0.80336
35	Large hypothesis space	0.80032

- Валидация – случайное стратифицированное разбиение на фолды
- Тестирование добавления / удаления переменных по одной
- В случае малых приростов общего скора обязательно проверялось, что **прирост имел место для большинства фолдов**

Сравнение CV и лидерборда



Другие интересные идеи

- Перенос данных на заданную временную шкалу (96 мес) + нейросеть
- Использование нескольких агрегатов (помимо среднего) для предиктов из моделей, где для многих строк один таргет

Спасибо за внимание!

e-mail: Evgeny.Patekha@gmail.com

slack: johnpateha