Tinkoff Data Science Challenge

Задача 1. Выбор кредита Tinkoff.ru

Постановка задачи и данные

- Tinkoff.ru работает с сетью магазинов электроники, в которой присутствуют и другие банки. Заявка на кредит от покупателя поступает сразу в несколько банков. Необходимо предсказать, выберет ли покупатель кредит от Tinkoff.ru.

- Задача бинарной классификации, метрика качества – **AUC**

- Выборка:
  - Train 170746 наблюдений / Test 91940 наблюдений
  - Таргет бинарный. Средняя вероятность отклика по train p = 0.176
  - 13 переменных: 6 числовых, 7 категориальных

Предобработка данных: living_region

Исходная переменная – 317 уникальное значение, 308 пропущенных значения

- Обработка:
  - Выкидываем мусорные слова ['ОБЛ','ОБЛАСТЬ', 'РЕСП'…]
  - Оставшиеся кейсы обрабатываем вручную (около 20). Итого 84 уникальных значения – регионы России + группа для NaN

- Сильный признак – частота грязного **living_region** / частота чистого **living_region**
- Добавляем поле **big_region** (федеральный округ)
- Признак – частота чистого **living_region** / частота **big_region**

Предобработка данных: новые признаки

- Все, что связано с кредитным риском: ovd_rate, pti…
- Группа признаков: var / mean(var) by [cat_col1, cat_col2..], примеры:
  - **score_shk** / mean(**score_shk**) by **tariff_id**
  - **age** / mean(**age**) by [**job_position, education**]

- Замена NaNs:
  - Для деревьев -1
  - Для остального – по логике (0, либо медиана) + флаг isnull, если пропуски частые
  - Объединение редких категорий для **job_position, tariff_id**

Признаки, улучшающие модель: средние

Среднее значение таргета в группах категориальных переменных и различных их комбинациях:

- Считается **out-of-fold** (использовался **5 StratifiedKFold**)
- Добавляется регуляризация:

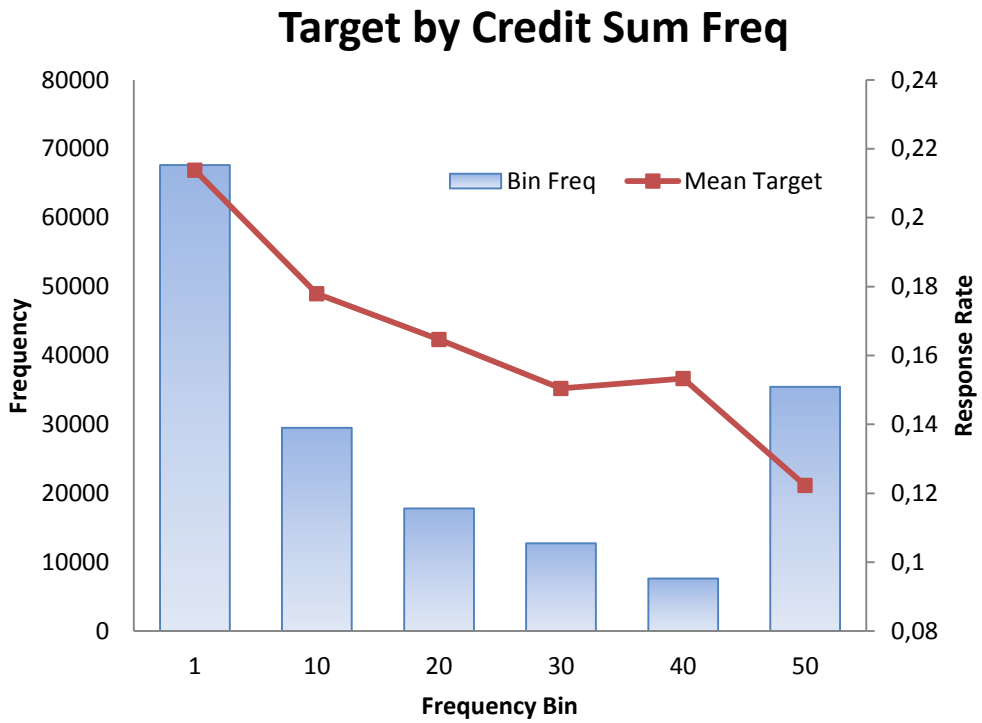
$$\frac{\boldsymbol{Resp} + \boldsymbol{\alpha} * \boldsymbol{Prior}}{\boldsymbol{Freq} + \boldsymbol{\alpha}}$$

, $\boldsymbol{\alpha}$ подбираем максимизируя метрику

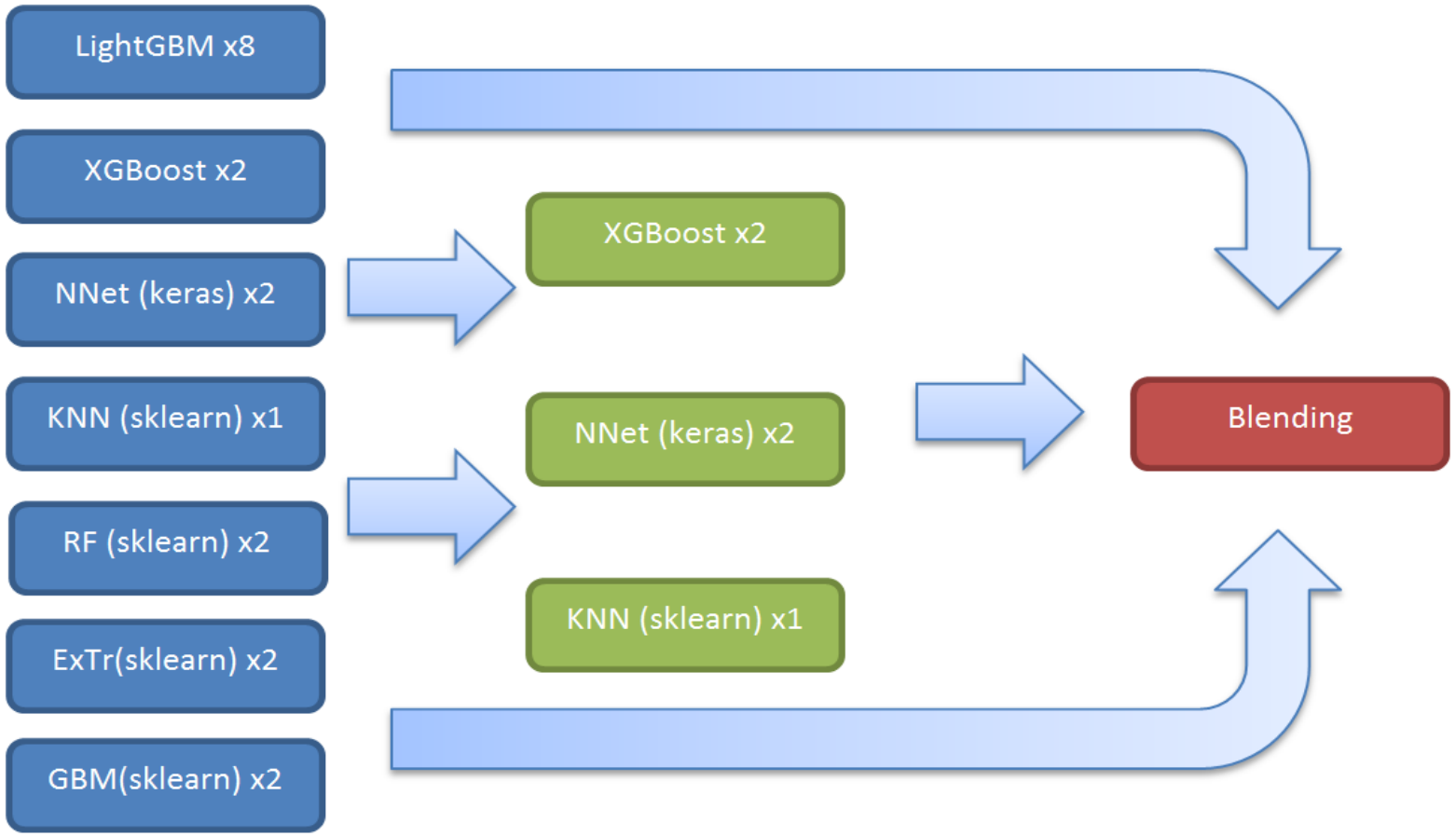
- Делается перебор из всех возможных комбинаций категориальных переменных, макс глубина = 3
- Происходит отбор при помощи линейной модели с **L1** регуляризацией
- К признакам добавляется out-of-fold прогноз линейной модели

Признаки, улучшающие модель: частоты

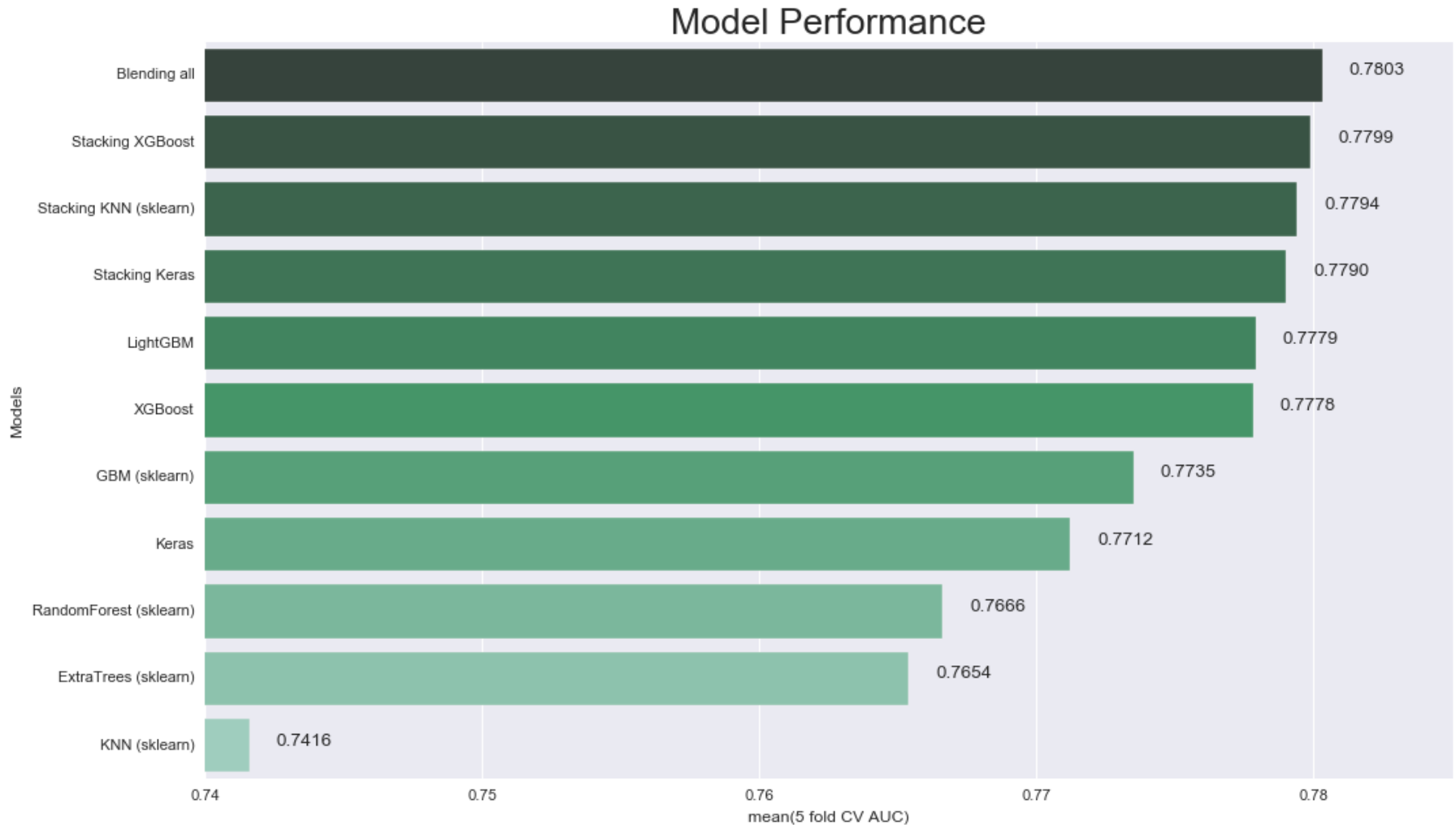
- Идея – наиболее нетипичные условия кредита ведут к повышению вероятности отклика.
- Самые сильные признаки – подсчет частоты появления ровно такой же суммы/срока кредита



Все модели (обучались на logloss и mse)



5 fold CV AUC моделей



Итоговое решение

- Итоговой моделью является линейная комбинация всех построенных моделей
- Веса подбирались таким образом, чтобы максимизировать **AUC**
- Для максимизации **AUC** делался покоординатный спуск
- Дает + .0001 CV AUC по сравнению с лог регрессией с **L2** регуляризацией

Что можно было сделать

- Оптимизация гиперпараметров моделей (**HyperOpt**). На деле настраивал вручную.
- Больше моделей, более разнообразный ансамбль
- Поработать с отбором признаков (в текущей версии добавлял блоками, если стало лучше – оставлял весь блок)
- Делать сабмиты 1 – p

Спасибо за внимание

Ссылка на код:

https://github.com/btbpanda/Tinkoff-boosters-2nd-place-