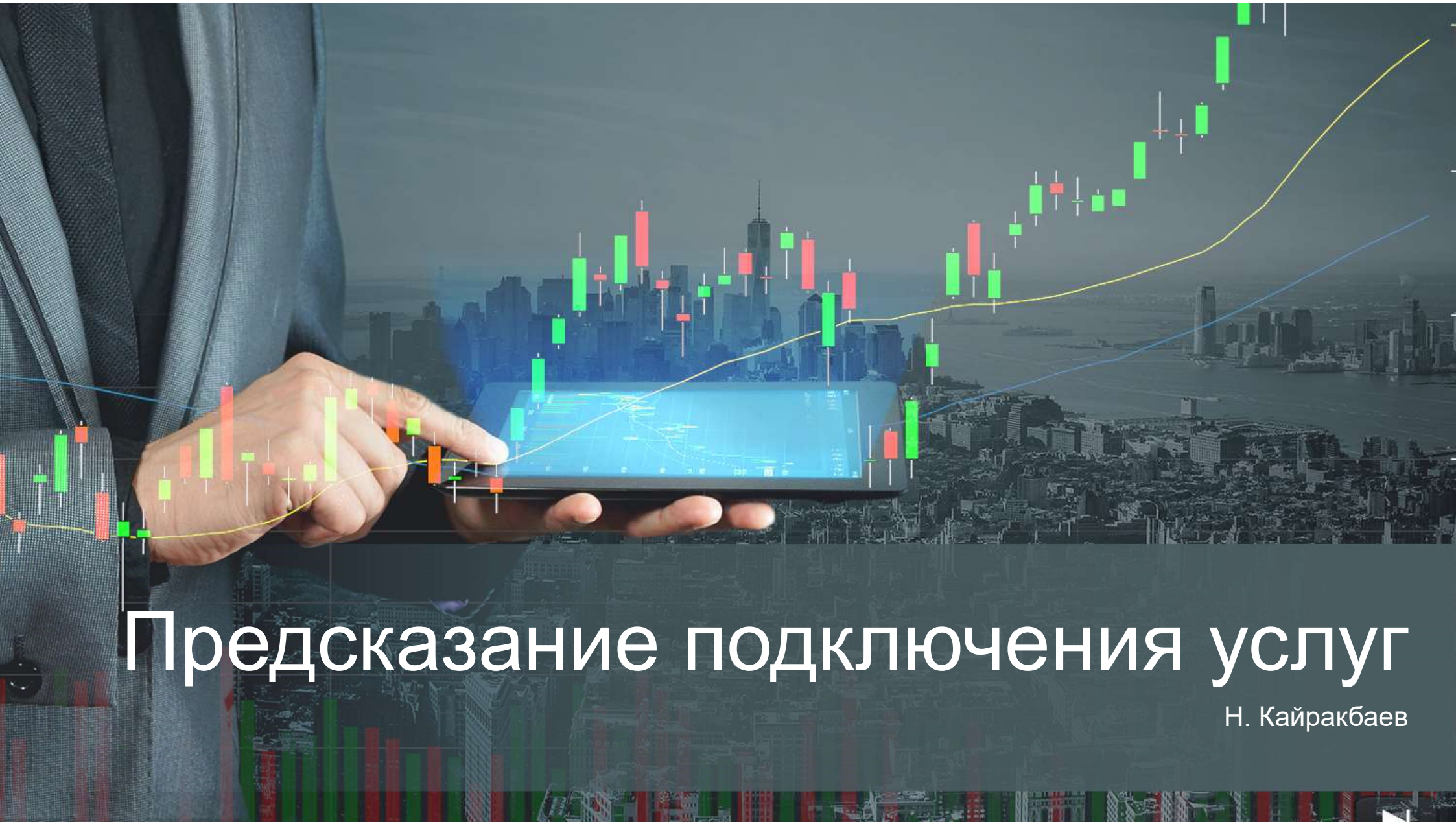


Предсказание подключения услуг

Н. Кайракбаев

Задача Проекта

Построить и обучить модель-классификатор для предсказания подключения услуги

Ключевые вопросы

01 Использование расширенных данных по пользователям

02 Эффективное преобразование входных данных

03 Имеются ли кластеры?

04 Выбор оптимальной модели и гиперпараметров

Исходные данные

Данные представлены 2 файлами (data_train and data_test), содержащими основные признаки (buy_time, id, id_vas), а также файлом features, содержащим детальные данные по каждому пользователю с анонимизированными признаками

Основные данные

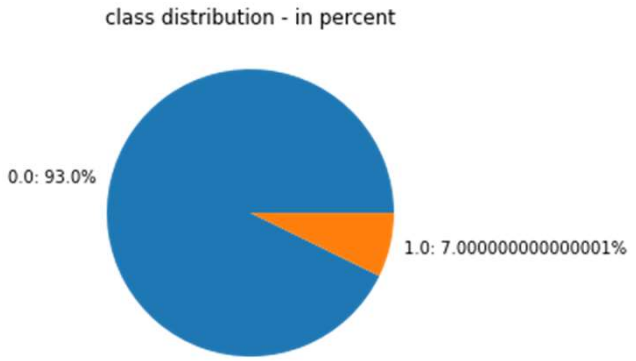
- Количество записей
  - Train: ~ 831,7 тыс. (в т.ч. ~806,6 тыс.users)
  - Test: ~ 71,2 тыс. (в т.ч. ~70,2 тыс.users)
- Основные признаки:
  - Пользователи (id)
  - Время (4 месяца + 1 месяц)
  - Предложения (i_vas) – 10 видов
- Особенности:
  - Пользователи на train и test – практически не пересекаются (число пересечений – 4,2 тыс.)
  - Отсутствуют пропуски

Детальные данные

- Всего около 4,4 млн.записей
- Количество анонимизированных признаков - 256

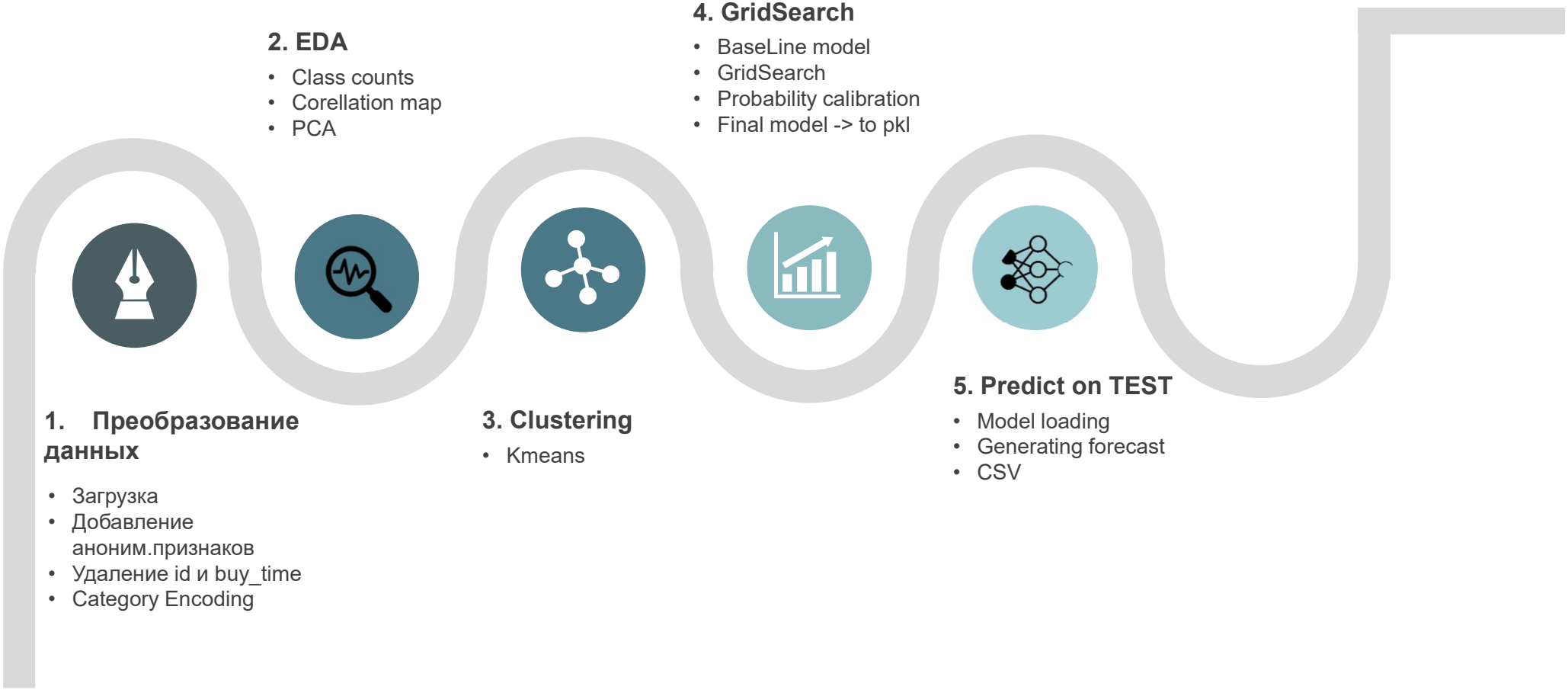
Целевой класс

- Наблюдается сильный дисбаланс классов
  - 0 ~ 93%
  - 1 ~ 7%



Для целей обучения и использования моделей: на основе **id** данные из **features** импортируются в **train** и **test** с усреднением по **id** (в случае нескольких данных), после **id** и **buy_time** – удаляются из датасетов. Категориальные данные определяются, если число уникальных данных не более – 10.

Этапы обучения



1. Преобразование данных

- Загрузка
- Добавление аноним.признаков
- Удаление id и buy_time
- Category Encoding

2. EDA

- Class counts
- Corellation map
- PCA

3. Clustering

- Kmeans

4. GridSearch

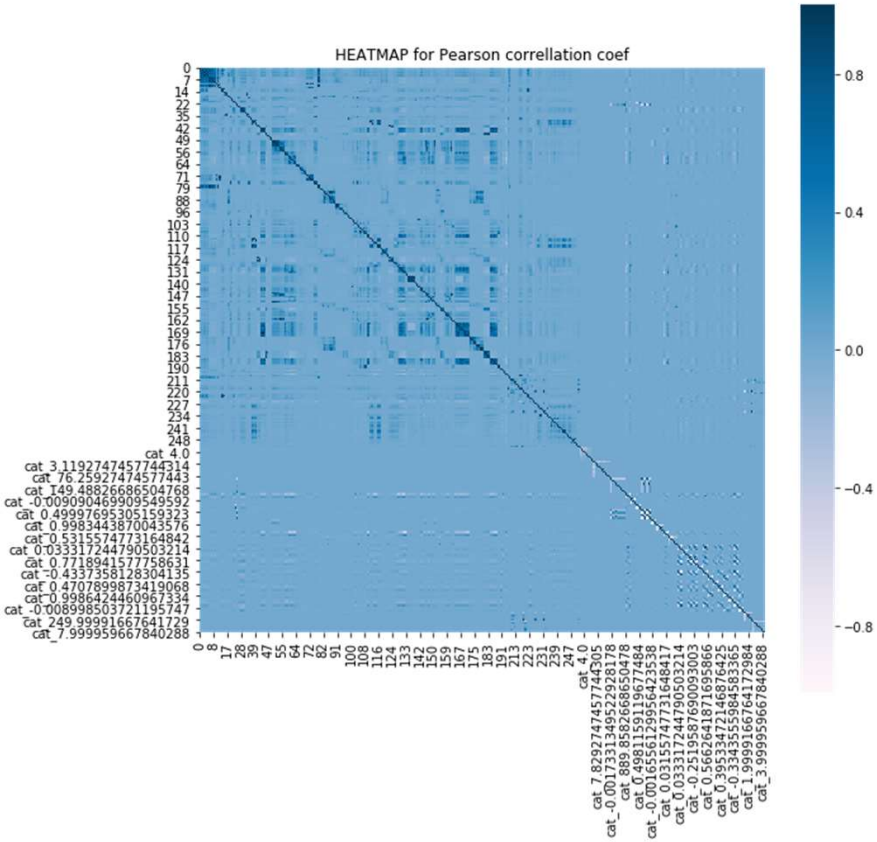
- BaseLine model
- GridSearch
- Probability calibration
- Final model -> to pkl

5. Predict on TEST

- Model loading
- Generating forecast
- CSV

Разведочный анализ данных

Корелляционная матрица



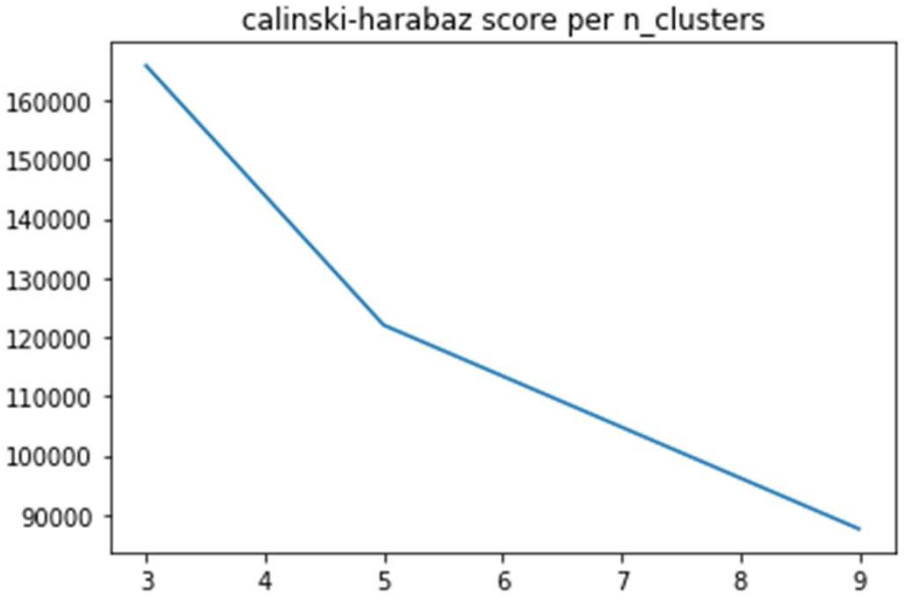
Анализ PCA

n_PCA	VAR portion
13	80%
21	90%
28	95%
36	97,5%
45	99,0%

- Найдено 138 пар признаков с корелляцией более 0,8

Кластерный анализ

Оценка метрики Калински-Харабас



- Для выявления возможных кластеров применен метод Kmeans
- Инициализатор: Kmeans++
- Использованные для анализа данные: X_train с категориальными dummy-признаками
- Найденное оптимальное число кластеров – 3
- Вместе с тем, использование кластеров не было в дальнейшем реализовано

Параметры поиска гиперпараметров

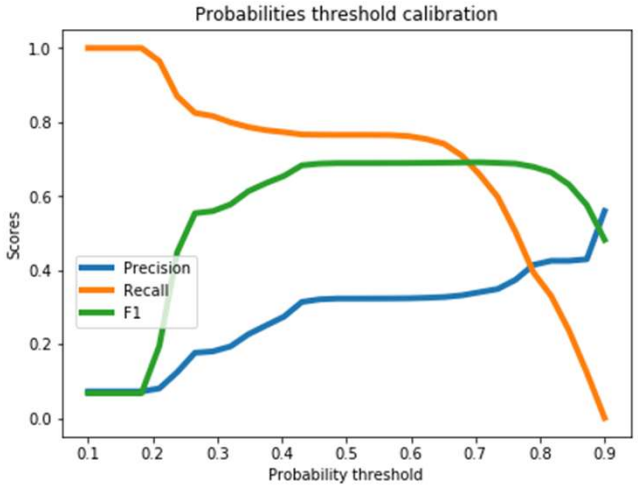
Преобразование данных

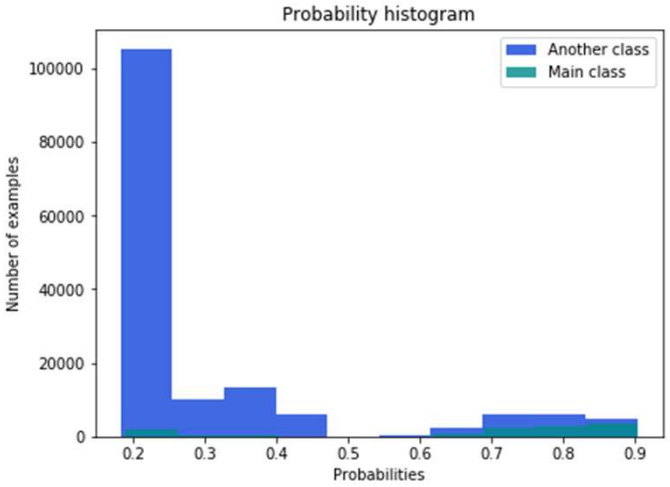
- Масштабирование: MinMaxScaler
- Уменьшение размерности признаков: PCA(n=45)
- Балансировка классов: SMOTE oversampler

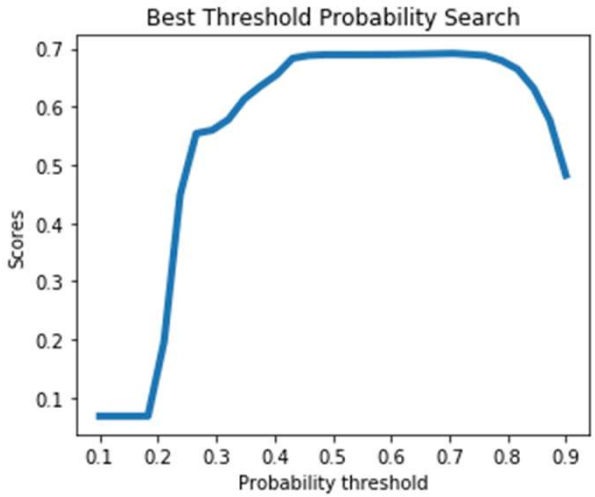
Алгоритм поиска гиперпараметров

- Trainset - > Train_gs (0,8) и Valid (0,2)
- Подбор параметров: GridSearchCV (cv = 3), метрика: F1 (average=macro)
- 2 этап: обучение модели на полной train_gs с опт.параметрами
- Калибровка вероятности и сравнение метрик на y_valid

Калибровка вероятностей. Случайный лес







- Выше представлены прогнозные кривые для алгоритма случайного леса (Выбранный в качестве оптимального алгоритма)
- Вероятность: 0,707
- F1 – на валидационной метрике:
  - До калибровки вероятности: 0,6893
  - После калибровки вероятности: 0,6916

Результаты обучения

Алгоритм	F1_Macro	Threshold Probability
Логистическая Регрессия	**0,6910**	**0,734**
Naïve Bayes	**0,6913**	**0,569**
Random Forest	**0,6916**	**0,707**
XG-boost	**0,6848**	**0,513**
LightGBM	**0,6900**	**0,679**

В качестве оптимального варианта выбран алгоритм случайного леса


Спасибо