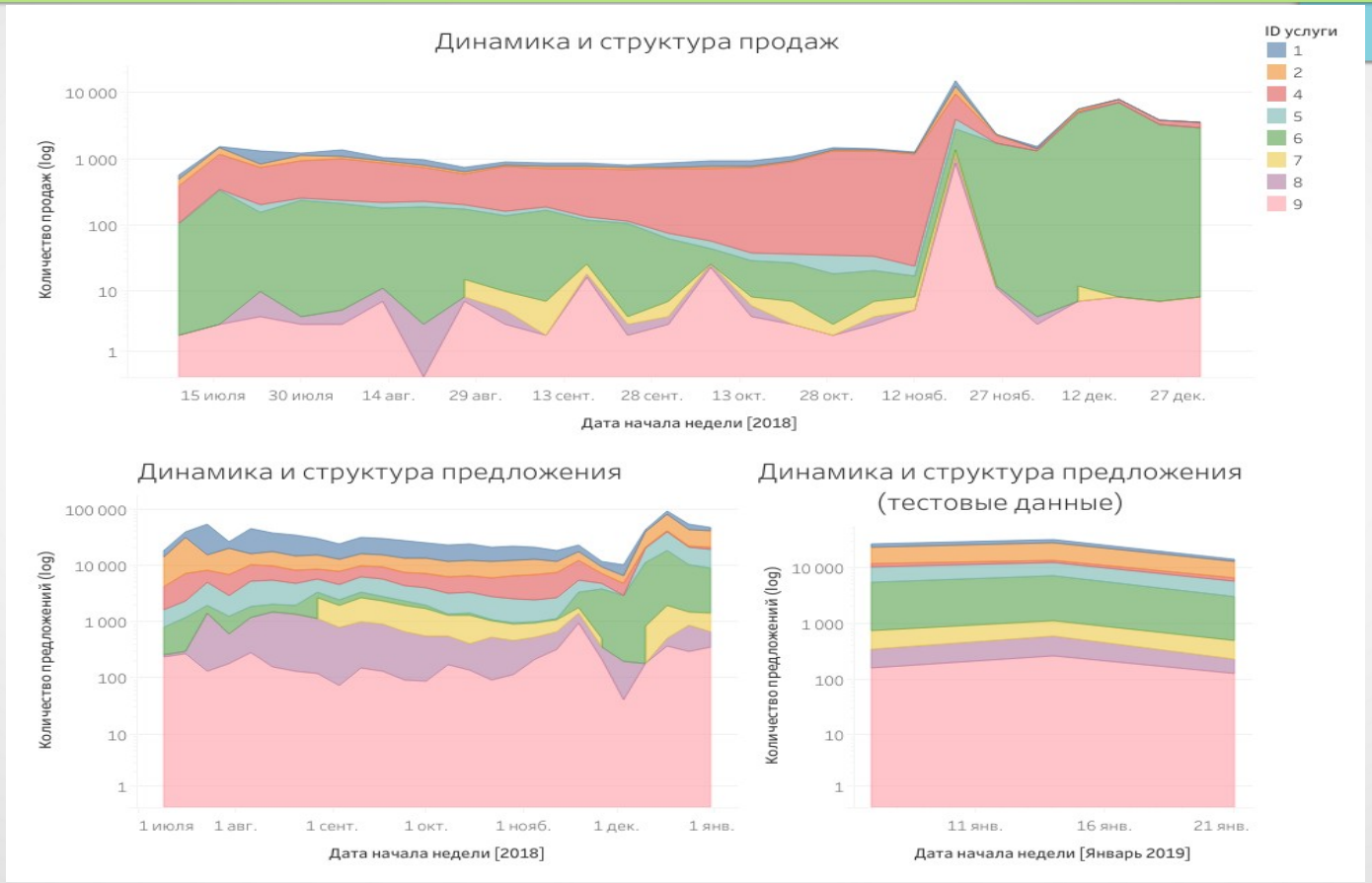
Прогнозирование вероятности подключения услуги абонентом сети «Мегафон»

Разработал и подготовил
Петр Мирзоян

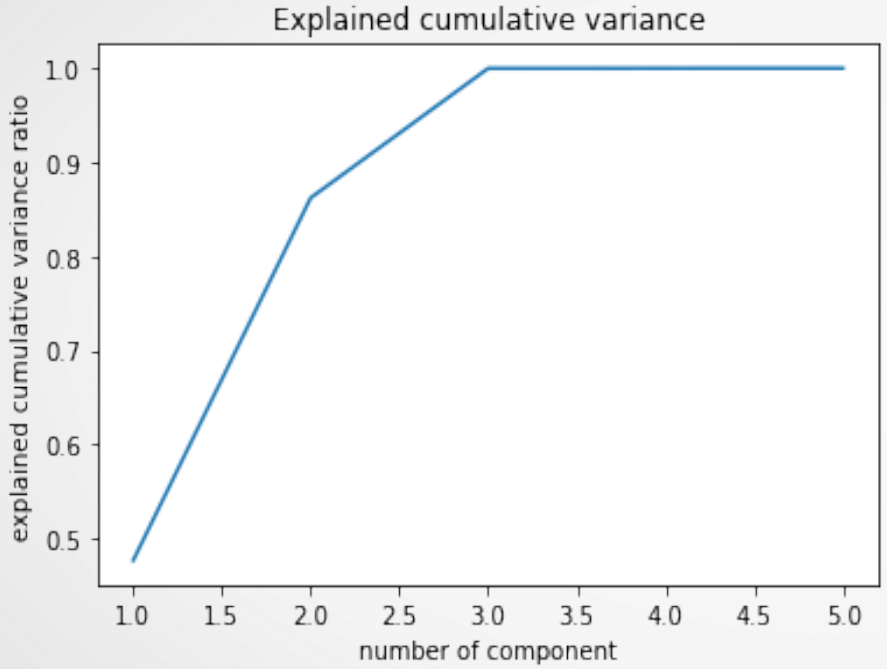
Описание задачи и входные данные

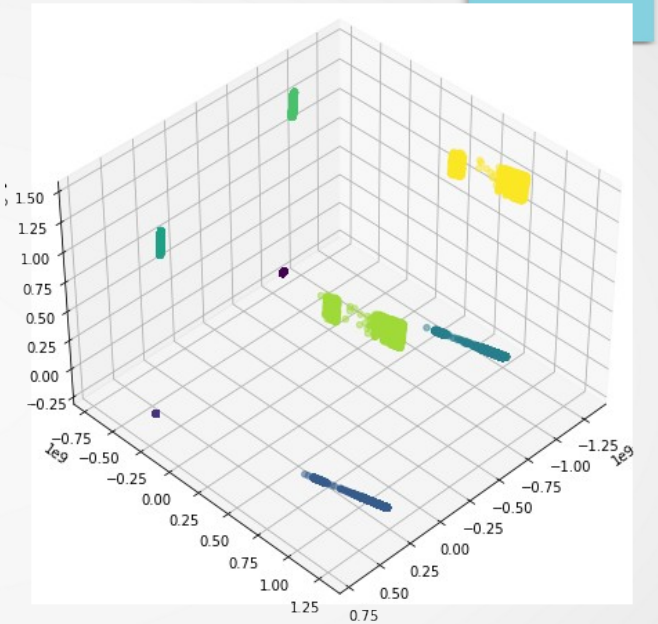
- Задача
  - определить вероятность подключения услуги для пары *пользователь-услуга*
- Данные
  - информация об отклике абонента на предложение
  - анонимизированные признаки профиля абонента, включающие дату агрегации

Основа модели: структура спроса и предложения



PCA сжатие признаков пользователей





- Количество признаков можно сократить до 3 практически без потери информации
- Сжатые признаки образуют явно различимые 8 кластеров

Сравнение моделей

	GradientBoosting	LGBM	RandomForest	SGD	XGB
[NEAREST]	0.578	0.738	0.720	0.450	0.622
[BACKWARD+ MEAN]	0.607	0.753	0.739	0.425	0.639
[BACKWARD+ NEAREST]	0.580	0.738	0.720	0.438	0.623

LGBM

- показывает лучшую метрику на старте
- работает ощутимо быстрее, чем случайный лес

Ключевые особенности

- Автоматизация процессов (airflow)
  - обучения и валидации
  - подбора гиперпараметров
- Не требует распределенных вычислений для прогноза
- Простота настройки

Конфигурация модели

- Параметры модели (`parameters.conf`)
  - список игнорируемых признаков пользователей
  - дата «отсечки» обучающих данных
  - кол-во разбиений для кросс-валидации
- Параметры обучения (`fit_params.json`)
  - стандартные параметры `LGBMClassifier`

Результаты

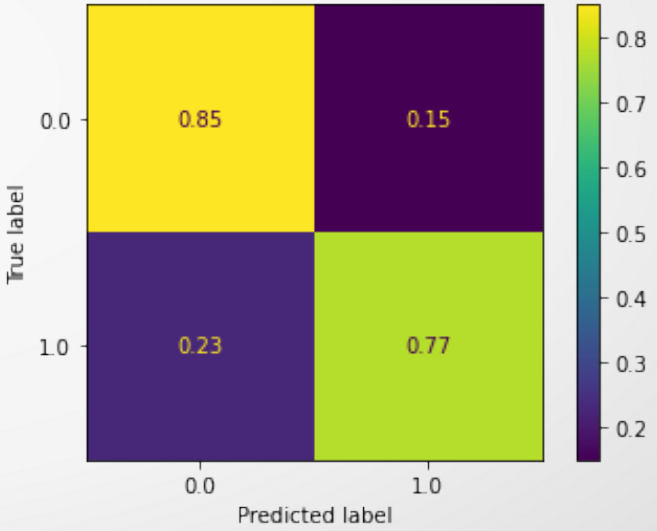
- Средняя метрика (`f1_score`) ≈ 0.738

Правильно классифицированы

- 77% подключивших услугу
- 85% не подключивших услугу

- 84% всех объектов

Матрица ошибок
(доли по классам)



Составление индивидуальных предложений

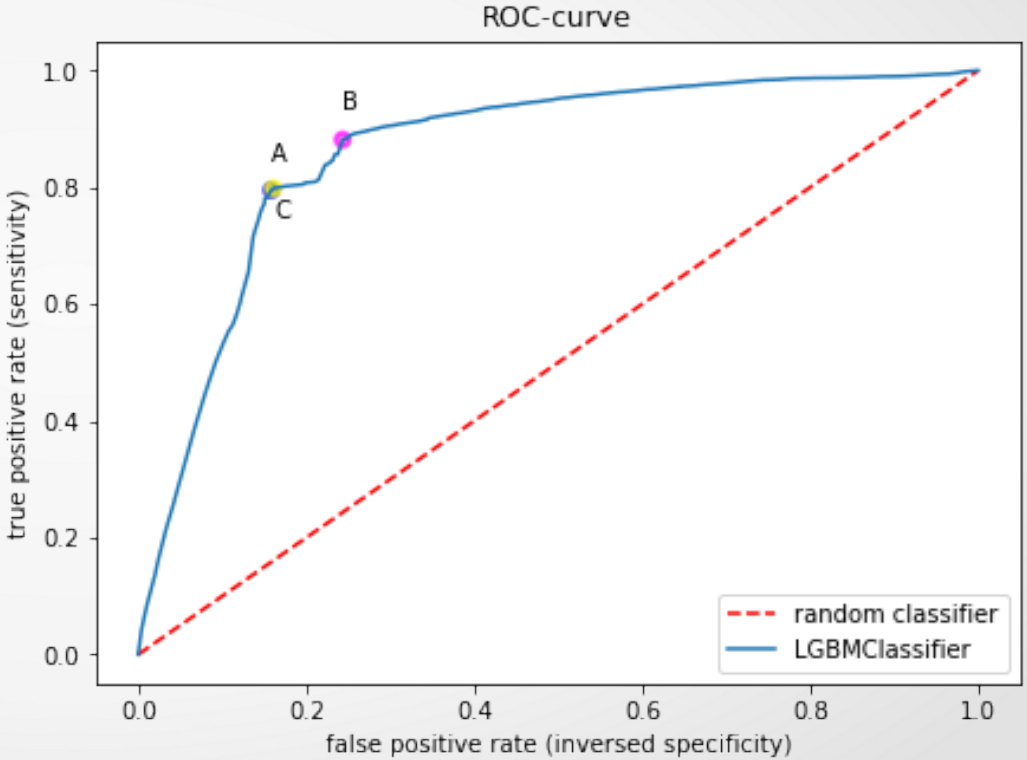
- Получить прогноз модели для каждой пары *пользователь-услуга*
- Отбросить услуги с прогнозной вероятностью ниже условленного порога
- Отсортировать услуги в порядке убывания прогнозной вероятности

Выбор оптимального порога вероятности

- *Max* TPR и *min* FPR (A)
- *Max* сумма sensitivity и specificity (B)
- *Max* G-mean sensitivity и specificity (C)

Оптимальный порог ≈ 0,2121

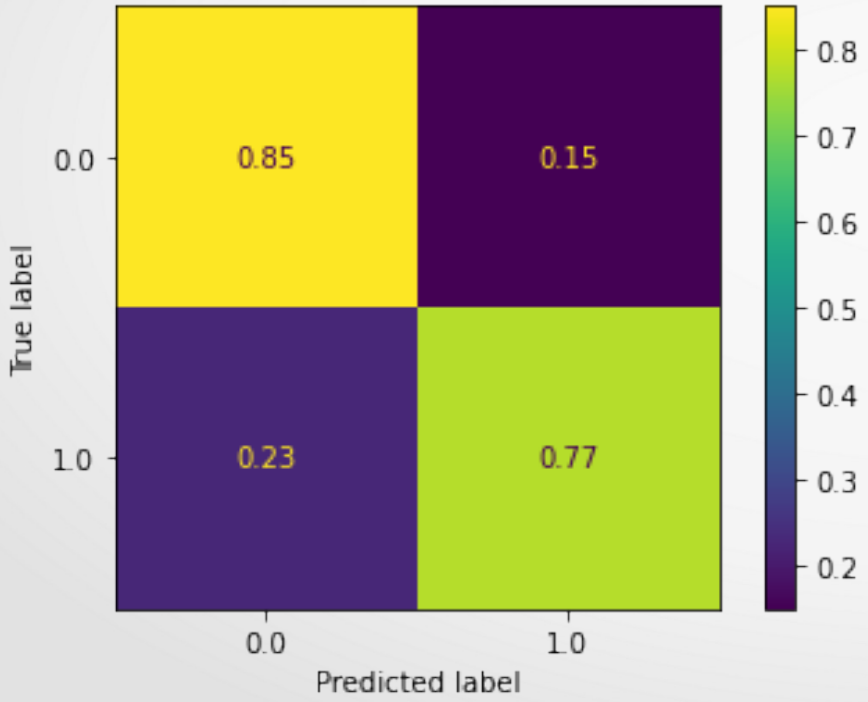
f1_score ≈ 0,737



Спасибо за внимание

Дополнительный материал

Матрица ошибок
(доли по классам)



Матрица ошибок
(доли от общего)