Повышение качества предсказания оттока абонентов оператора сотовой связи

Выполнил: Долголев Филипп, 371гр.
Научный руководитель: Невоструев Константин

Мотивация

- Отток абонентов порядка 10 процентов в квартал
- Привлечение новых абонентов дороже удержания

Обзор

- Предсказанием оттока занимаются многие
- Универсального подхода нет
- Чаще всего лучшие результаты достигаются с использованием:
  - XGBoost
  - Случайный лес
  - Нейронные сети
- Лучший результат дипломной работы прошлого года:

	ROC AUC	Precision	Recall
XGBoost с ручной группировкой данных	0.90	0.75	0.66

Цель

Повысить качество классификации уходящих абонентов

- Подготовить данные для классификации
- Провести кластеризацию абонентов
- Построить классификатор на основе кластеризации
- Оценить оптимальное значение метрик для построенного классификатора

Метрики

	Ушедшие	Оставшиеся
Предсказанные ушедшими	TP	FP
Предсказанные оставшимися	FN	TN

- Precision - TP / (TP + FP)
- Recall - TP / (TP + FN)
- ROC-кривая - соотношение между TP / (TP + FN) и 1 - TN / (TN + FP)
- ROC AUC - площадь под ROC-кривой
- PR-кривая - соотношение между Precision и Recall
- PR AUC - площадь под PR-кривой

Оценка каждой метрики бралась как средняя от всех итераций перекрёстной проверки при разбиении данных на 5 частей

Формат данных

- Персональная информация (пол, возраст, дата подключения, регион)
- Помесячная активность абонента, разбитая по сервисам:
  - Входящие звонки
  - SMS
  - GPRS
  - Исходящие звонки, разбитые на различные группы

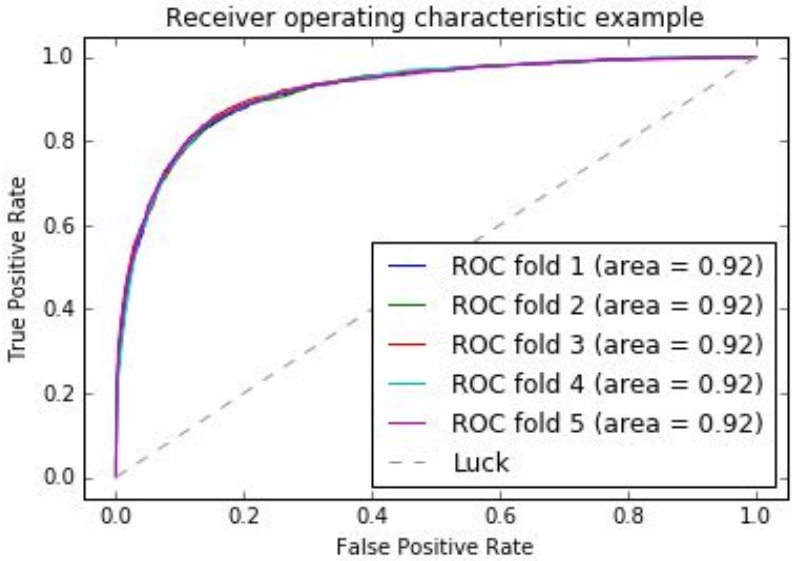
Подготовка данных

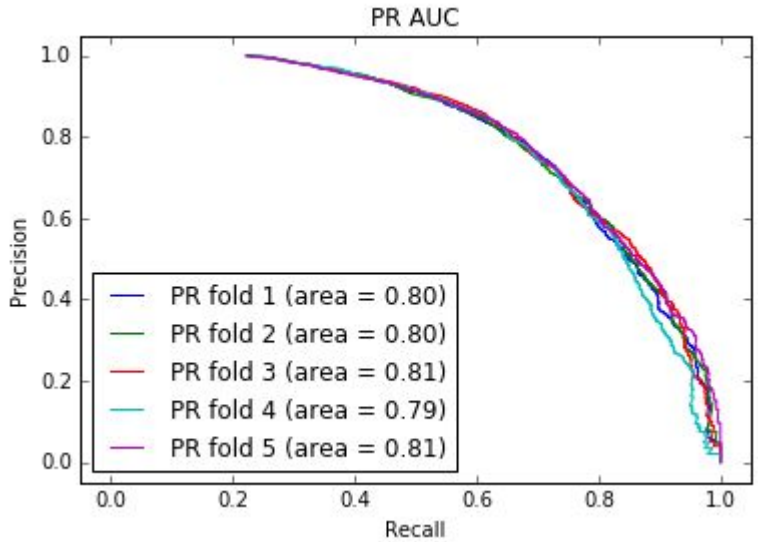
- Сформированы временные серии активности абонентов по 3 месяца
- Данные очищены
- Рассмотрен был только Санкт-Петербург, т.к. из всех данных на него приходится около половины, а остальное на 7 разных городов
- Получены новые признаки на основе активности:
  - Для звонков по разным сервисам получены их доли в минутах
  - Для всех сервисов и их долей были получены арифметические и геометрические отношения относительно временного ряда
  - Дисперсия, Математическое ожидание
  - Коэффициенты асимметрии и эксцесса
- Исключены признаки с количеством минут

Итог: 50 000 объектов со 148 признаками, ушедших порядка 10 000

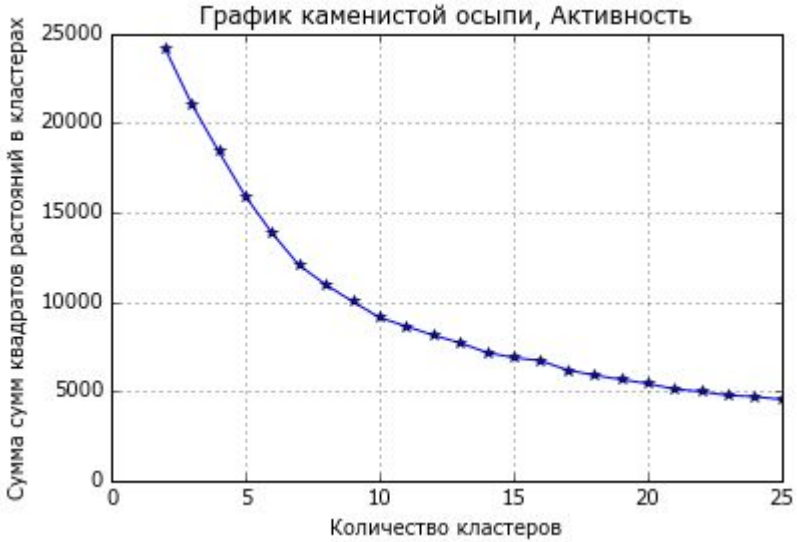
Результаты построение модели XGBoost

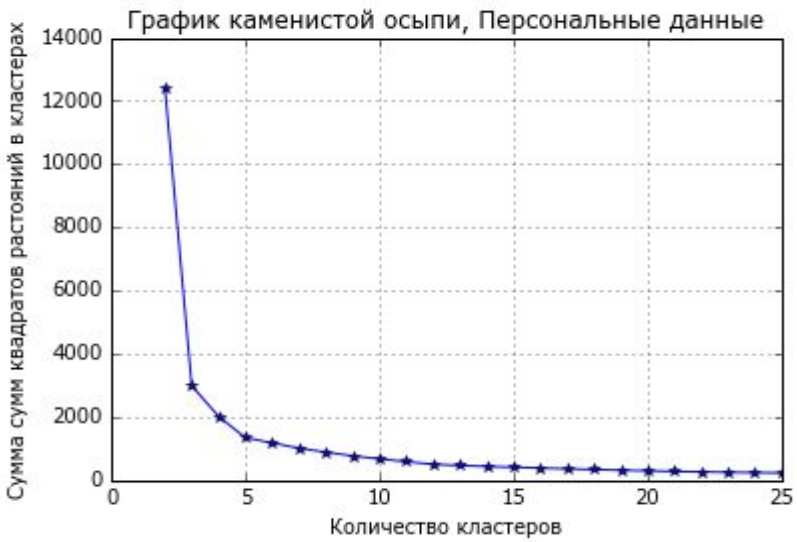
	ROC AUC	PR AUC	Precision	Recall
XGBoost	0.92	0.80	0.74	0.69





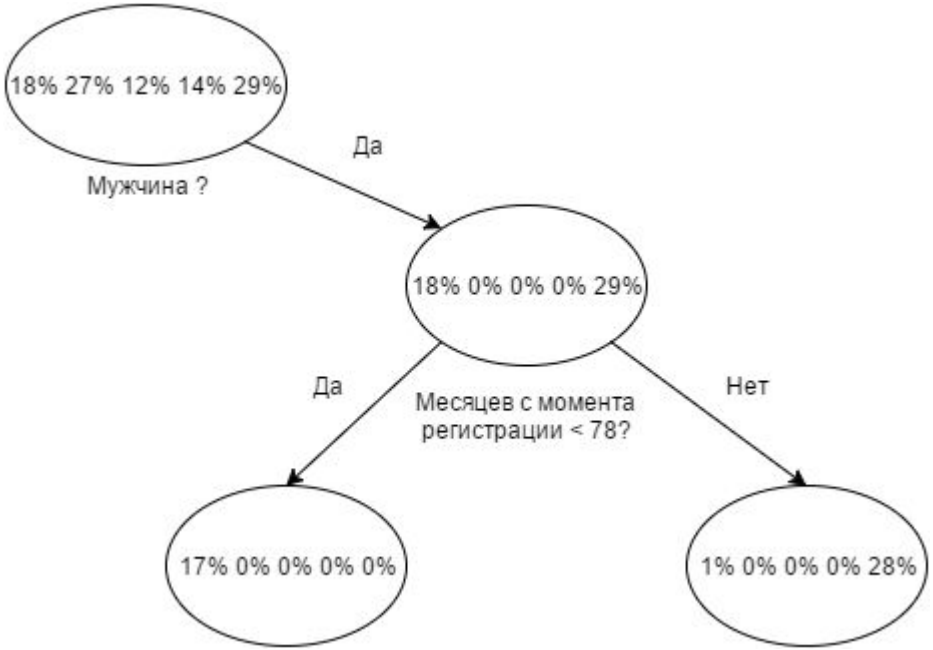
Кластеризация K-Means

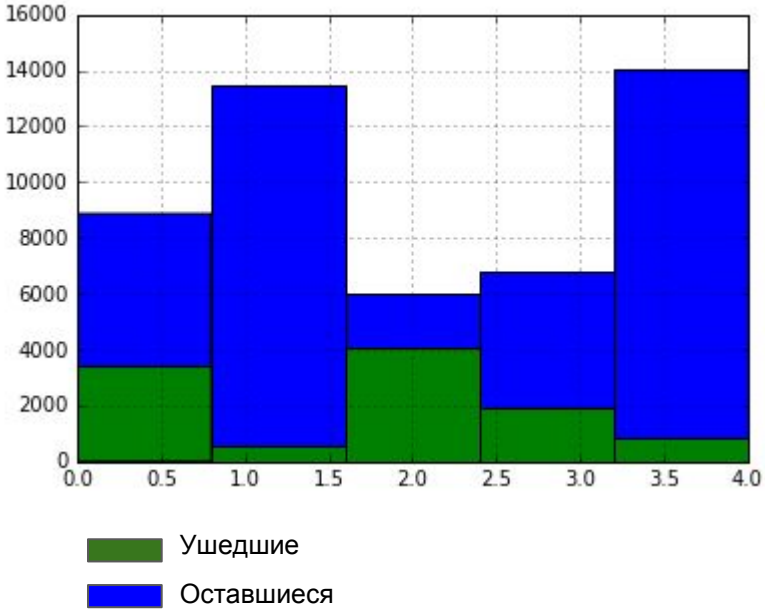




- Вручную выбраны признаки для кластеризации
- Проведено нормирование данных
- Было решено разбить на 7 кластеров по активности, и на 5 по персональным данным.

Анализ кластеризации

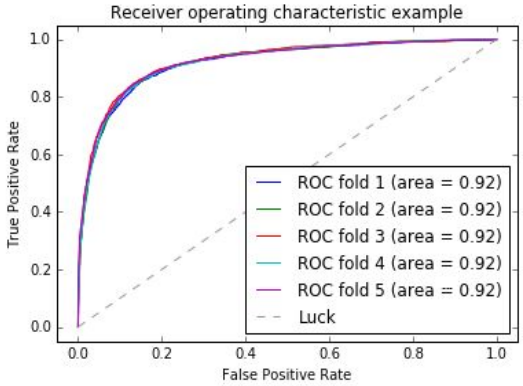


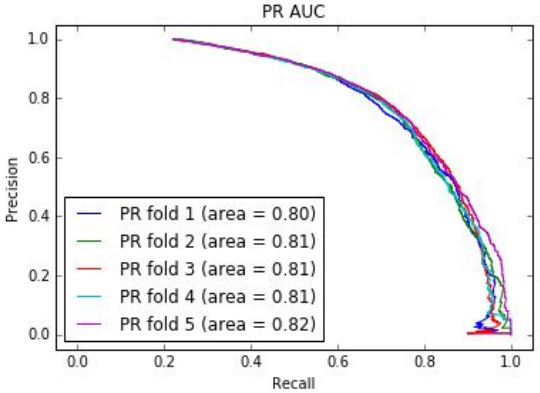


Построение ансамбля на кластерах

- На каждом кластере обучен независимый классификатор XGBoost
- Классификаторы объединены логистической регрессией

	ROC AUC	PR AUC	Precision	Recall
Ансамбль	0.92	0.81	0.75	0.72



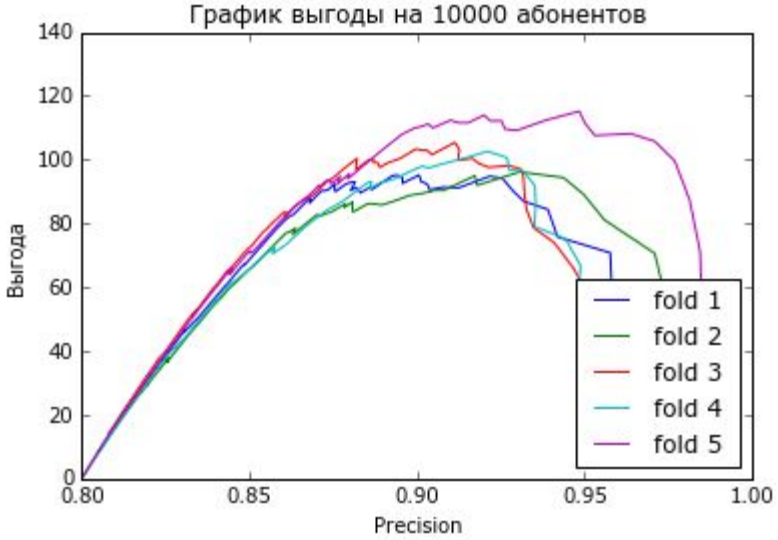


Сводная таблица классификаторов

	ROC AUC	Precision	Recall
Результат прошлого года	0.90	0.75	0.66
XGBoost на подготовленных данных	0.92	0.74	0.69
Ансамбль на кластерах	0.92	0.75	0.72

Оценка оптимального значения Precision

- Предположения:
  - 80% прибыли от удержанного - затраты на его удержание
  - Ушедших порядка 20%
- Выгода = TP - 0.8 * (TP + FP)
- Наибольшая выгода: 94.8
  - Precision = 0.92
  - Recall = 0.33



Результаты

Повышено качество классификации уходящих абонентов

- Подготовлены данные для классификации
- Проведена кластеризация абонентов
- Построен классификатор на основе кластеризации
- Оценены оптимальные значения метрик для построенного классификатора