Модели машинного обучения для предсказания ухода пользователей

Курбанов Рауф Эльшад оглы

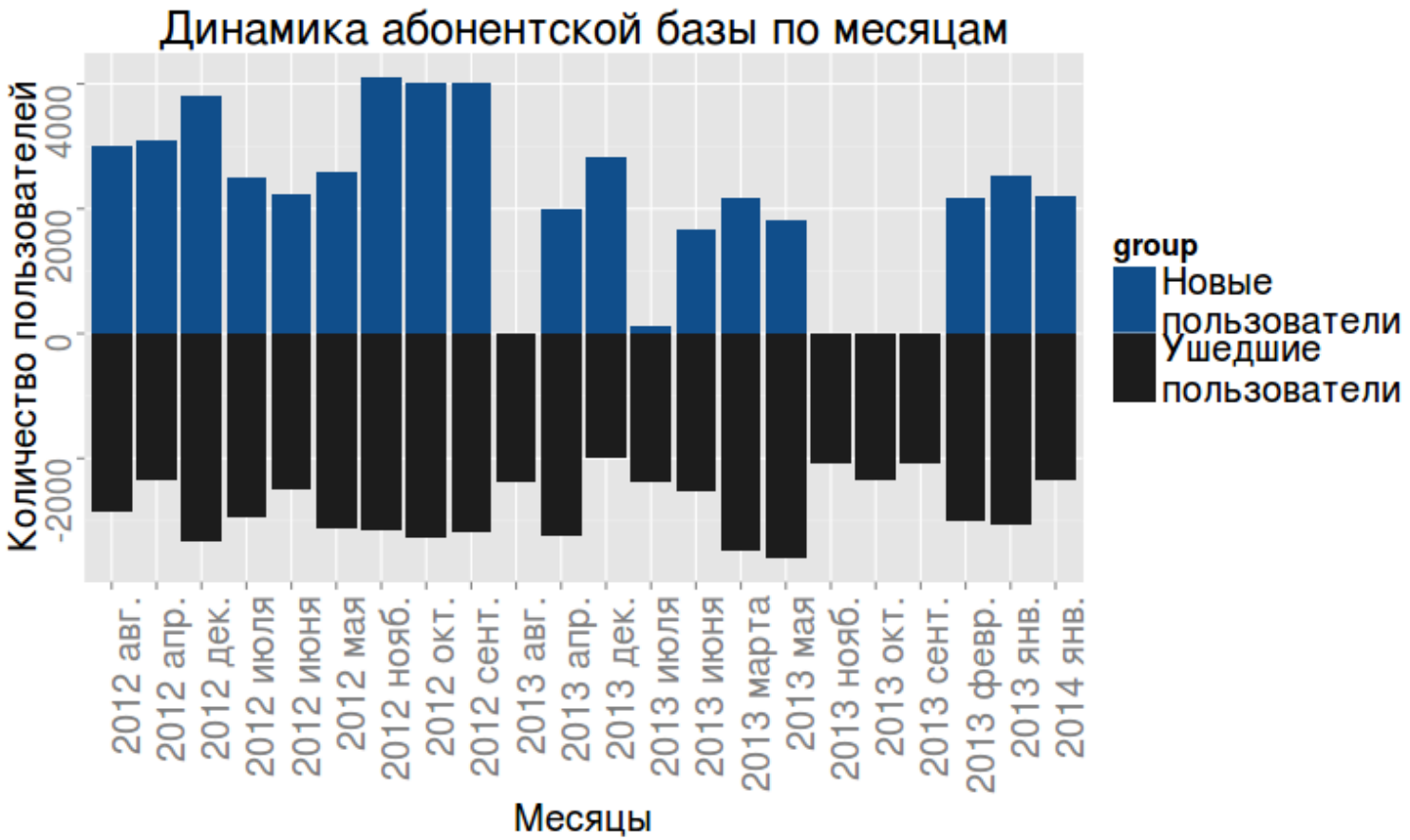
Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра Системного программирования

Научный руководитель: д. ф.-м. н., профессор А. Н. Терехов
Рецензент: технический директор ООО "Лаборатория Анализа Данных" А. Г. Натекин

Санкт-Петербург

2015г.

Актуальность задачи



Существующие решения

Решения в телекоммуникационной сфере:

- Формат данных спецефичен для области
- Пространство признаков мало
- Сложно переиспользовать

Решения, основанные на модели:

- Посвящены применению модели, а не решению задачи
- Жертвуют точностью предсказаний за счёт универсальности
- Упускают специфику задачи оттока

Постановка задачи

- Цель работы
  - Описать решение проблемы ухода с помощью машинного обучения на примере данных крупного интернет-провайдера
- Постановка задачи
  - Провести подготовку данных о пользователях широкополосного доступа в интернет
  - Сравнить наиболее подходящие модели для предсказания ухода и выбрать лучшую для нашей задачи.
  - Реализовать прототип системы для предсказания ухода абонентов

Подготовка данных

- 257 столбцов
- 6 500 000 записей
- 176 000 пользователей
- 90 000 пользователей активных хотя бы 9 месяцев

Номер договора	Номер месяца	Количество дней с предыдущего платежа	Информация об абоненте	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Построение признаков

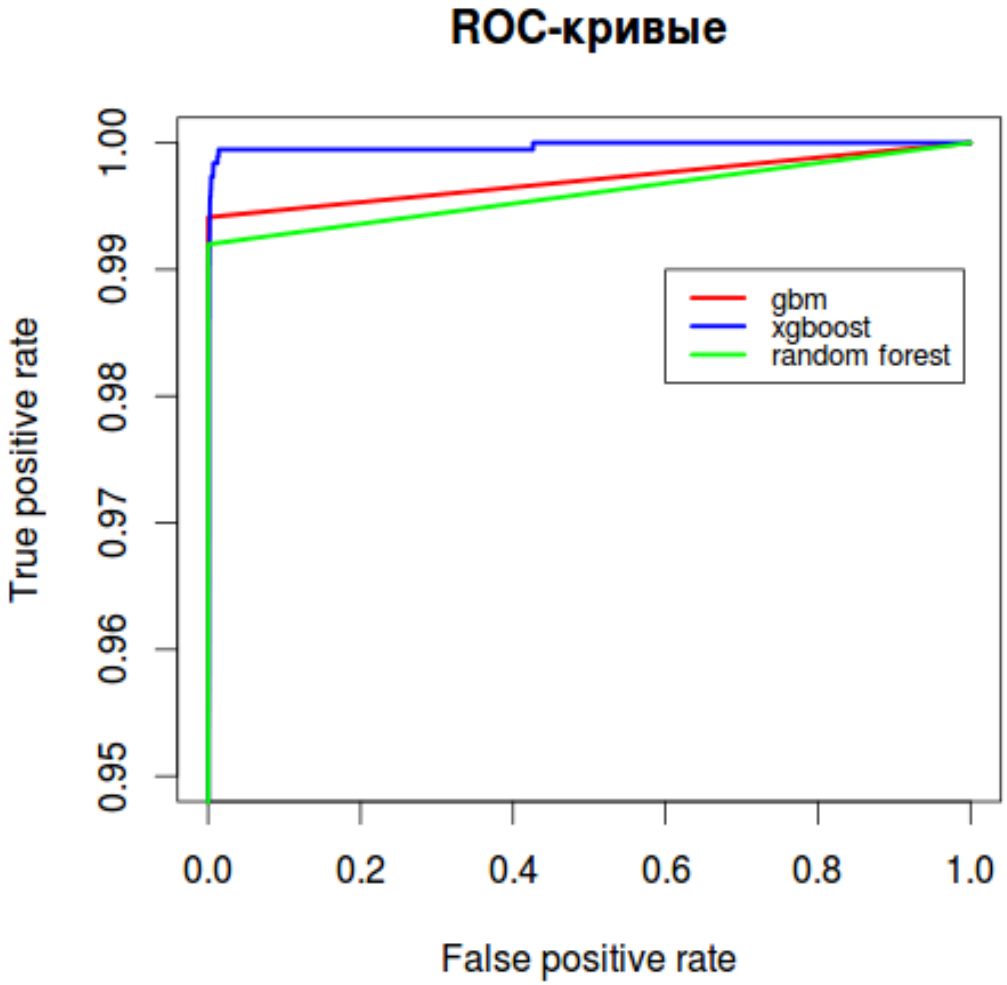
Для каждого параметра X строится следующий какскад признаков:

- X.diff: плавающие разности соседних значений признака X
- X.sum.3: плавающие суммы значений признака X за 3 дня
- X.max.3 плавающий максимум значений признака X за 3 дня
- X.mean.3: плавающее среднее значения признака X за 3 дня
- X.mean3.d: плавающее среднее трёх соседних значений X.diff
- X.max.3.d: плавающий максимум трёх соседних значений X.diff

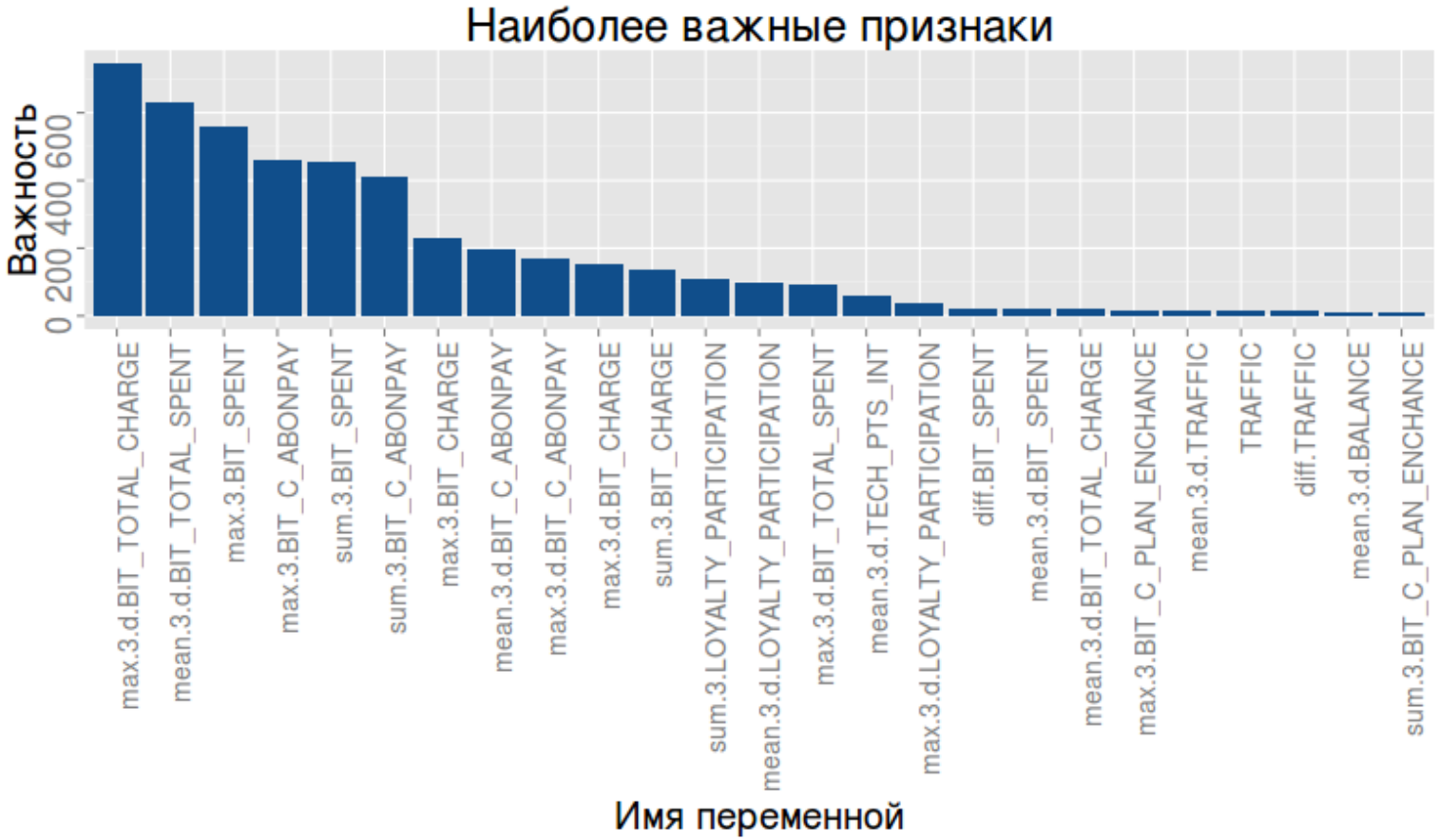
Результаты классификации

Модель	ACC	PREC	RECALL	F1	AUC
xgboost	0.99350	0.99893	0.99986	0.99652	0.99993
gbm	0.99757	0.98829	0.99038	0.98933	0.99757
random forest	0.99890	0.99839	0.99199	0.99518	0.99589

Сравнение моделей



Интерпретация модели



Реализация

В разработанной системе были реализованы следующие функции:

- Параллелизм на этапе обучения
- Оптимизированная работа с разреженными данными
- Автоматизация кросс-валидации гиперпараметров

Результаты

В ходе работы были получены следующие результаты:

- Проведена обработка данных о пользователях широкополосного доступа в интернет
- По результам исследования выбрана и модифицирована эффективная модель для предсказания оттока
- Реализован и протестирован на реальных данных прототип системы для предсказания ухода абонентов для интернет-провайдера