Использование методов машинного обучения для предсказания оттока клиентов телекоммуникационной компании

Кутькин Никита Андреевич,

444 группа

Научный руководитель:

д.ф.-м.н, проф. А.Н. Терехов

Рецензент:

К. Н. Невоструев, ведущий разработчик

ООО “НМТ — Новые Мобильные Технологии"

Предсказание оттока клиентов

- Цель – удержание клиентов
- Это задача классификации
- Различные методы работают по-разному в зависимости от специфики задачи и имеющихся данных

Постановка задачи

- Провести предварительную обработку и анализ данных
- Реализовать систему для дальнейшей обработки данных и тестирования классификаторов (или использовать уже существующую)
- Сравнить результаты различных алгоритмов построения классификаторов

Используемые в работе средства, алгоритмы

- Язык программирования – Python
- Библиотека машинного обучения – scikit-learn
- Алгоритмы: Метод опорных векторов, Случайный лес, Gradient boosting, Ada boost
- Метрики: AUC, precision, recall, $\mathrm{F}_{0.5}$

Данные

- 411802 абонентов в выборке
- Данные на одного абонента:
  - Возраст
  - Пол
  - Регион подключения
  - Дата регистрации
  - Помесячные данные по активности (всего 15 месяцев)

Обработка данных

- Форматирование признаков:

  «муж.» → <1,0,0,0> «2001-07-14» → 4401

- Анализ активности:

  <0,0,1,1,1,1,1,1,1,1,1,0,0,0,0>

- Добавление новых признаков:

  Минимум, максимум, мат. ожидание, дисперсия …

Вариативность обработки данных

- Количество месяцев для предсказания
- Использование новых признаков (каких?)
- Нормализация/стандартизация
- Размер выборки
- Использование персональных данных

Система тестирования классификаторов

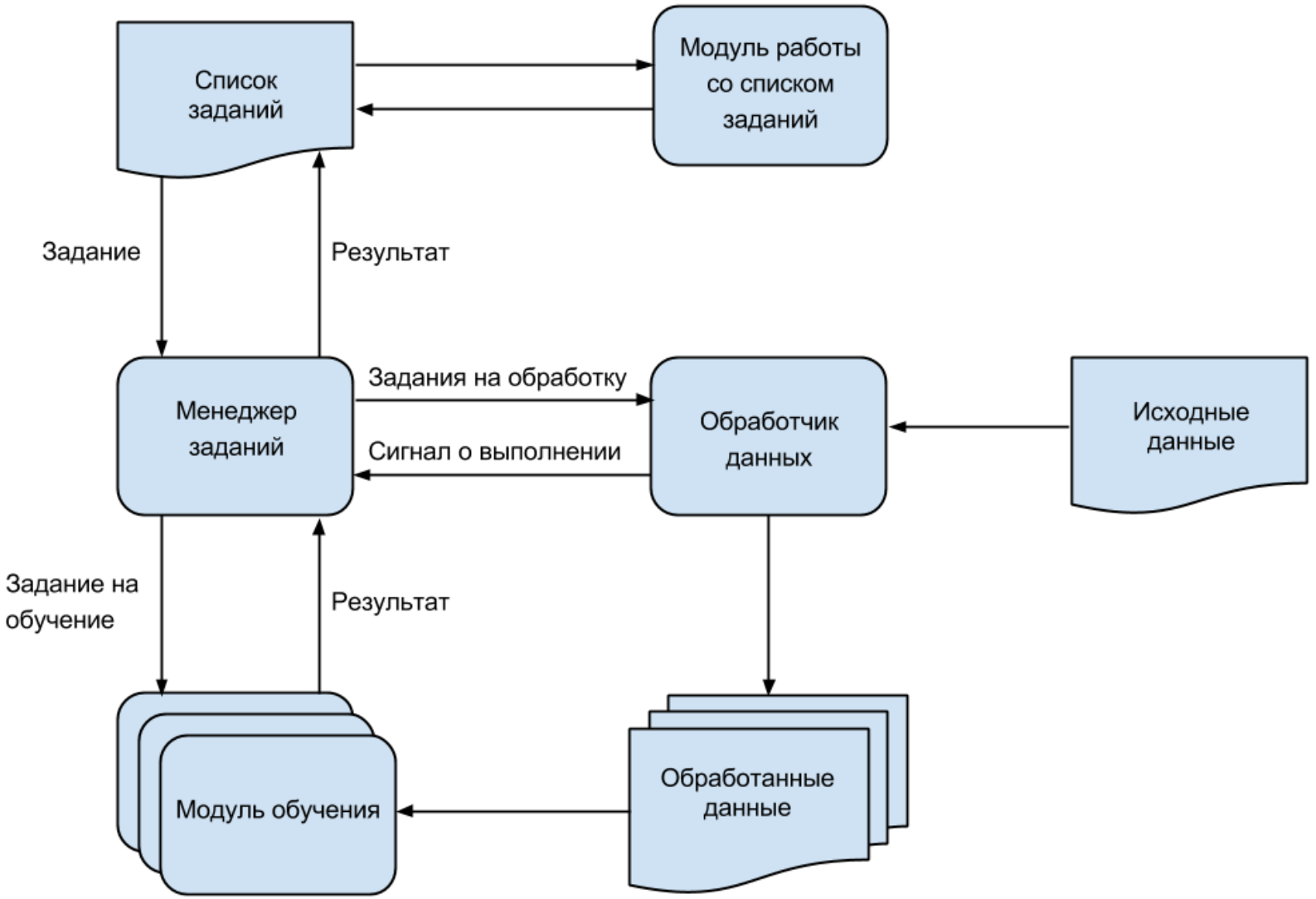
Требования:

- Возможность перебора параметров обработки данных
- Возможность перебора алгоритмов и их параметров
- Возможность остановки и возобновления работы, просмотр промежуточных результатов
- Распараллеливание

Существующие решения:

- Модуль grid search (библиотека scikit-learn)
- Auto-WEKA (библиотека WEKA)

Система тестирования классификаторов



Сравнение классификаторов

- Использование более 3 месяцев для предсказания не улучшает результат
- Использование персональных данных абонента (пол, возраст и т.д.) улучшает результат
- Использование новых признаков, полученных из временных рядов улучшает результат

Сравнение результатов различных алгоритмов

	AUC	**Precision**	**Recall**	$\mathbf{F_{0.5}}$
Gradient boosting	**0.88**	0.72	0.65	0.70
ADA boost	**0.86**	0.68	0.60	0.66
Случайный лес	**0.85**	0.62	0.72	0.64
Метод опорных векторов	**0.80**	0.64	0.57	0.62

$$\mathbf{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\mathbf{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathbf{F_\beta} = (1 + \beta^2)\frac{\mathrm{precision} \cdot \mathrm{recall}}{\beta^2 \cdot \mathrm{precision} + \mathrm{recall}}$$

Результаты работы

- Проведены предварительная обработка и анализ данных
- Реализована система для дальнейшей обработки данных и тестирования классификаторов
- Проведено сравнение результатов различных алгоритмов построения классификаторов