

ПРОГНОЗИРОВАНИЕ ВЫБОРА КЛИЕНТОМ ОТЕЛЯ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Голубева Л. Л., Мурашко А. С.

*Белорусский государственный университет, Минск, Беларусь,
e-mail: goloubeva@bsu.by, aliaksandra.murashka@gmail.com*

В работе рассматриваются вопросы выявления внутренних закономерностей больших объемов данных методами машинного обучения в применении к задаче мультиклассификации на выборке данных, которая представляет собой информацию о поисковых запросах пользователя относительно конкретного отеля при бронировании на сайте Expedia. Выборка данных содержит 40 198 536 образцов, из них 2 528 243 тестовых образца, для каждого образца указаны 24 признака. Согласно условию имеется 100 кластеров отелей, к одному из них и требуется отнести отель.

Процесс моделирования включал в себя настройку виртуальной машины для выполнения вычислений на облаке, подготовку данных (создание и отбор признаков с помощью статистических процедур [1], обработку пропущенных значений, приведение данных к подходящему формату). Были построены модели на основании различных алгоритмов машинного обучения (Decision Tree, Random Forest, Gradient Boosted Trees, Stacking). Проведено исследование поведения моделей на различных наборах признаков, выбран наиболее информативный набор. Точность прогноза различными алгоритмами оценивалась с помощью метрики MAP@5. Специфика данной метрики заключается в том, что для нее необходимо, чтобы предсказание состояло не из одного значения предполагаемого кластера, а из пяти наиболее вероятных. Для моделей, показавших более высокую точность на начальном этапе, подобраны параметры (число классификаторов, входящих в ансамбль, максимальная глубина и др.) [2], которые позволили улучшить первоначальный результат.

В приведенной сводной таблице (см. Табл. 1) указаны характеристики точности и времени обучения методов, показавших лучшие результаты в своей категории.

Табл. 1. Сравнение точности и времени обучения моделей

Модель	Точность на тестовом наборе	Map@5	Время обучения
Решающее дерево	0.185124	0.235172	39.4 с
Случайный лес	0.369408	0.338504	9 ч 34 мин
Градиентный бустинг	0.363490	0.385587	11 ч 37 мин 2 с
Стекинг (градиентный бустинг + случайный лес + решающее дерево)	0.186890	0.241722	1ч 58 мин 3 с

Библиографические ссылки

1. Hastie T. The elements of statistical learning. Data mining, Inference and Prediction second edition/ T. Hastie, R. Tibshirani, J. Friedman. -Springer, 2008, 764 p.
2. Jain A. Complete Guide to Parameter Tuning in XGBoost (with codes in Python) [Electronic source]. Mode of access: <https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/>. Date of access: 10.03.2017