Отчет по заданию №3: Композиции алгоритмов для решения задачи регрессии

Васильев Руслан ВМК МГУ, 317 группа

24 декабря 2020 г.

Содержание

1 Введение

В заключительном практическом задании предлагается реализовать композиции алгоритмов машинного обучения и провести эксперименты, а также спроектировать веб-сервис для взаимодействия с моделью. Весь проект доступен в репозитории[1]. Данный отчет иллюстрирует результаты экспериментов с моделями на датасете данных о продажи недвижимости.

2 Постановка задачи

Итак, рассматривается задача регрессии с метрикой качества RMSE:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)}{N}},$$

где N — размер выборки, y_i — истинное значение целевой переменной на i-м объекте, $\hat{y}_i$ — предсказанное.

Для решения реализованы две модели, представляющие собой ансамбли решающих деревьев: случайный лес и градиентный бустинг. Исследование алгоритмов включает в себя измерение функции ошибки и времени работы при варьировании гиперпараметров (порядок экспериментов соответствует стандартной настройке данных моделей).

3 Эксперименты

3.1 Предобработка данных

Исходные данные о недвижимости были разделены на обучение (80%) и контроль (20%, она же валидационная выборка). И здесь сразу учитывается особенность задачи. Хотя в задании отсутствует описание признаков и целевой переменной, можно с уверенностью предположить, что столбец `date` связан со временем поступления данных (даты имеют небольшой диапазон 2014–2015, монотонно возрастают, дублируются, следуют сразу за `ID`, а столбцы `build_year` и `renovation_year` с ними не связаны). По этой причине было бы некорректно перемешать выборку перед разделением на обучение и контроль — из-за утечки такая стратегия может дать ложную оценку качества моделей и привести к неправильным выводам. В качестве валидационной выборки берутся последние 20% данных, соответствующие хронологическому порядку по столбцу `date`.

[1] https://github.com/artnitolog/mmf_prac_2020_task_3