Кузьмин Никита, ММП, 317.

Отчет по заданию №3

"Композиции алгоритмов для решения задачи регрессии".

Случайный лес и градиентный бустинг.

Содержание

1 Введение

В данном документе представлен отчет о проделанных экспериментах по практическому заданию №3, анализ результатов.

Краткое описание задания:

Необходимо реализовать алгоритмы **RandomForestMSE** и **GradientBoostingMSE**, провести указанные ниже эксперименты на датасете с данными о продажах недвижимости, проанализировать результаты.

2 Эксперименты

В этом блоке приведены все обязательные эксперименты, которые изложены в формулировке задания.

Стандартный дизайн эксперимента:

- Из данных был удален признак «**id**», так как он не несет полезной информации для модели. Так же признак «**date**» был закодирован с помощью **LabelEncoder** таким образом: ранним датам соответствуют меньшие значения. Датасет был разделен на тренировочную (70%) и валидационную выборки (30%).
- Стандартные параметры экспериментов в случае **RandomForestMSE** (если не оговорено обратное):
  - n_estimators = 20
  - max_depth = None (не ограничена)
  - feature_subsample_size = 6
- Стандартные параметры экспериментов в случае **GradientBoostingMSE** (если не оговорено обратное):
  - n_estimators = 20
  - max_depth = 5
  - feature_subsample_size = 6
  - learning_rate = 0.1

2.1 Исследование алгоритма RandomForestMSE

2.1.1 Зависимость RMSE и времени работы алгоритма в зависимости от количества деревьев

Соответствующие графики приведены на рис. 1, 2

Рис. 1:

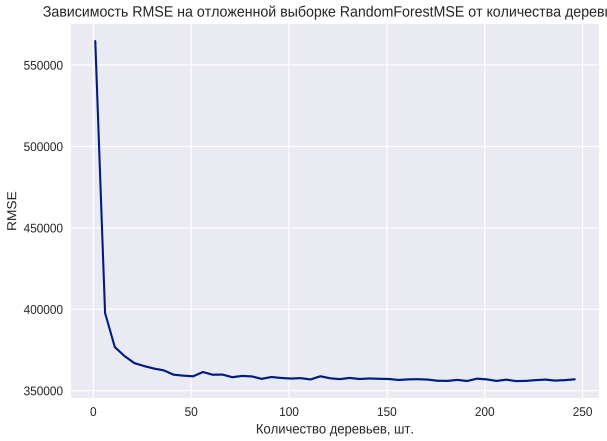
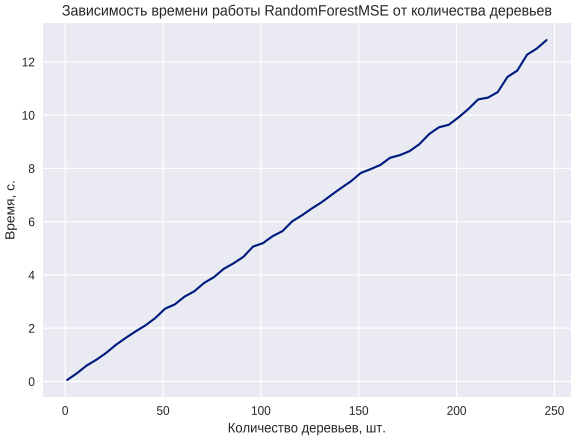


Рис. 2:



Из данных выше графиков видно, что с возрастанием количества деревьев ошибка падает, а время работы линейно растет. Это подтверждает теоретические выкладки, приведенные на лекциях и семинарах.

2.1.2 Зависимость RMSE и времени работы от размерности подвыборки признаков для одного дерева

Соответствующие графики приведены на рис. 3, 4

Рис. 3:

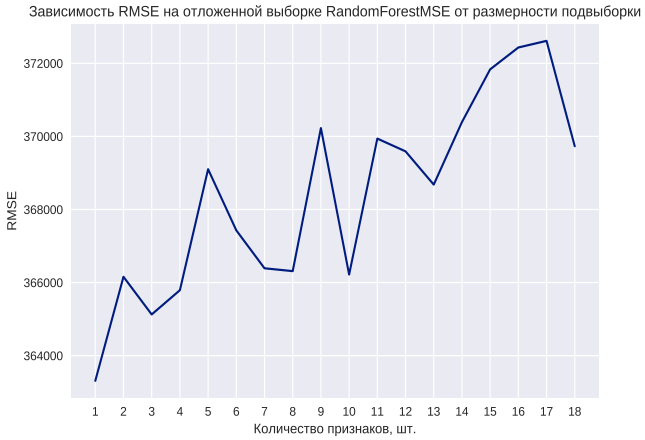
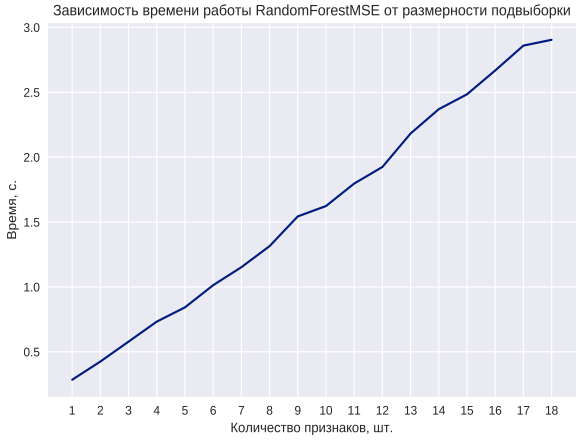


Рис. 4:



Интересно, что оптимальный размер подвыборки признаков равен 1. Это может быть из-за того, что в данных какой-то из признаков значительно полезнее других. Как и ожидалось, время возрастает с ростом размера подвыборки признаков.

2.1.3 Зависимость RMSE и времени работы от размерности максимальной глубины деревьев

Соответствующие графики приведены на рис. 5, 6

Рис. 5:

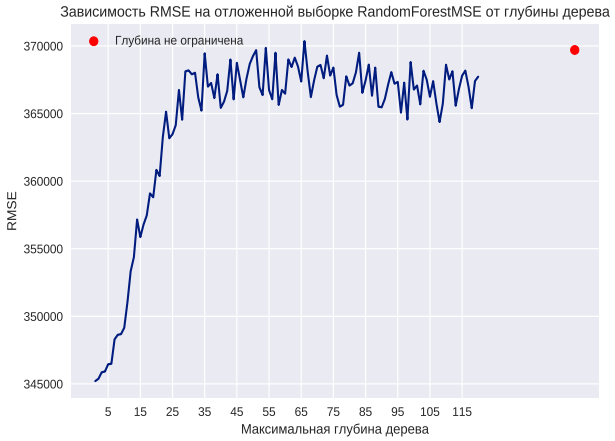
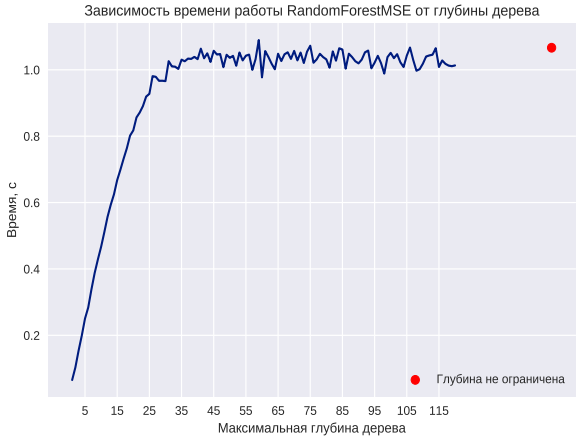


Рис. 6:



Несмотря на то, что рекомендуют брать деревья в случайном лесе переобученными (большой глубины), в условиях данной задачи это правило не выполняется. Возможно, это получается из-за особенности данных. Время работы алгоритма возрастает с ростом максимальной глубины деревьев (до 40 деревьев), затем стабилизируется.

2.2 Исследование алгоритма GradientBoostingMSE

2.2.1 Зависимость RMSE и времени работы алгоритма в зависимости от количества деревьев

Соответствующие графики приведены на рис. 7, 8

Рис. 7:

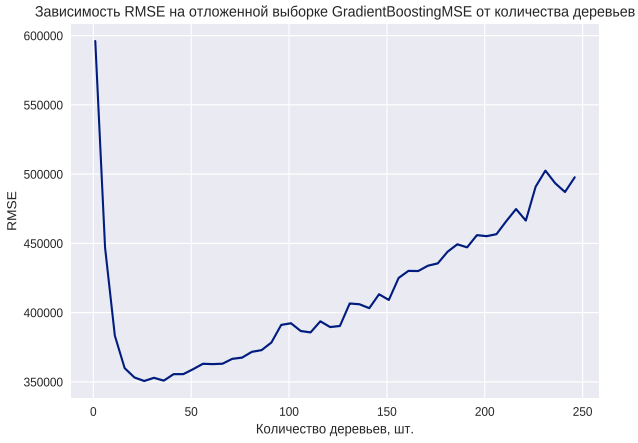
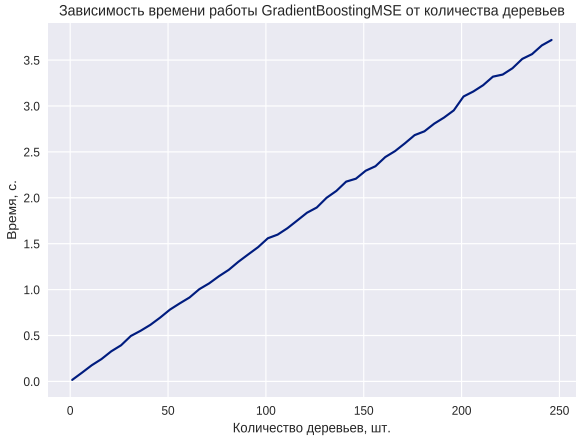


Рис. 8:



Из графика зависимости RMSE от количества деревьев видно, что оптимальное количество деревьев – 20, а далее с увеличением числа базовых моделей растет и ошибка. Это говорит о переобучении, что свойственно для градиентного бустинга. Из графика зависимости времени видно, что время растет линейно.

2.2.2 Зависимость RMSE и времени работы от размерности подвыборки признаков для одного дерева

Соответствующие графики приведены на рис. 9, 10

Рис. 9:

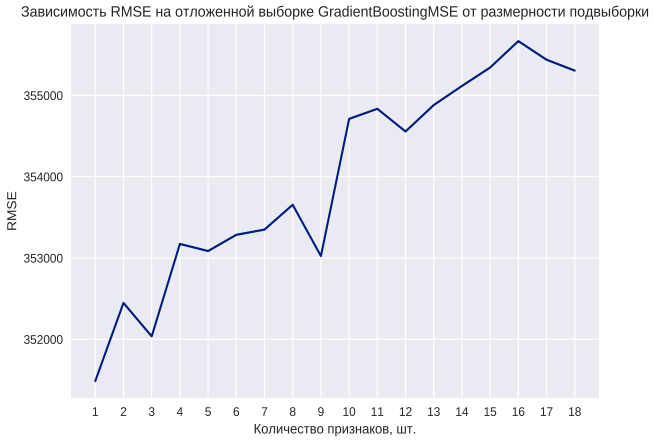
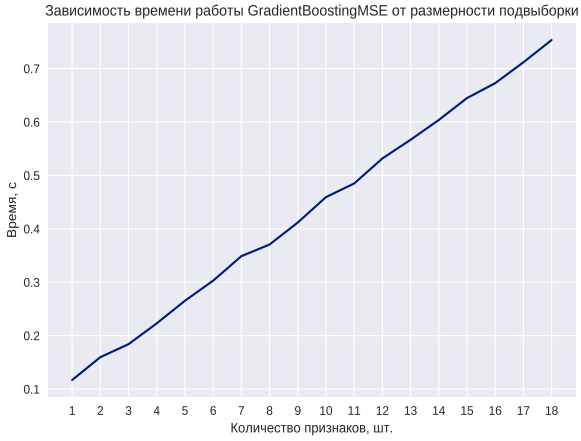


Рис. 10:



Как и для алгоритма **RandomForestMSE**, оптимальное значение размера подвыборки признаков получилось равным 1, а зависимость времени работы алгоритма от размера подвыборки признаков - линейная.

2.2.3 Зависимость RMSE и времени работы от максимальной глубины деревьев

Соответствующие графики приведены на рис. 11, 12

Рис. 11:

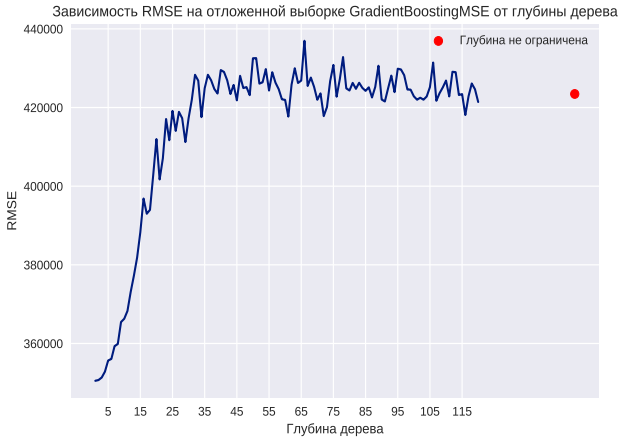
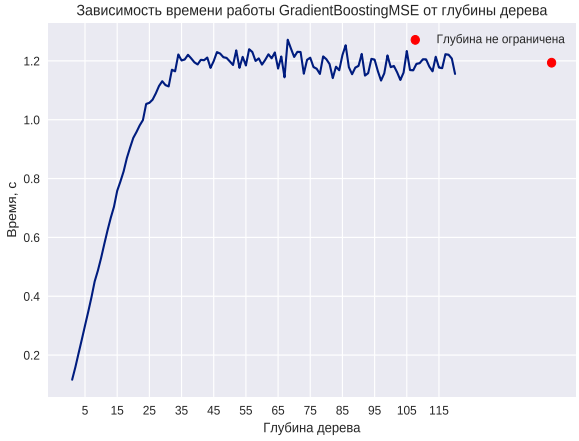


Рис. 12:



Из графиков видно, что с увеличением глубины деревьев увеличивается и ошибка на отложенной выборке. Это возникает из-за того, что алгоритм градиентного бустинга имеет свойство переобучаться.

Время работы алгоритма возрастает до определенного значения глубины (40), затем стабилизируется. Это говорит о том, что при этой глубине достигается один из других критериев останова построения решающего дерева:

- во всех листах (на данном этапе построения) количество объектов $<$ min_samples_split
- для любого листа (на данном этапе построения) количество объектов, которые будут получены в каждом из листов при разбиении $<$ min_samples_leaf
- количество листов $>=$ max_leaf_nodes
- и др. . .

2.2.4 Зависимость RMSE и времени работы от значения learning rate

Соответствующие графики приведены на рис. 13, 14

Рис. 13:

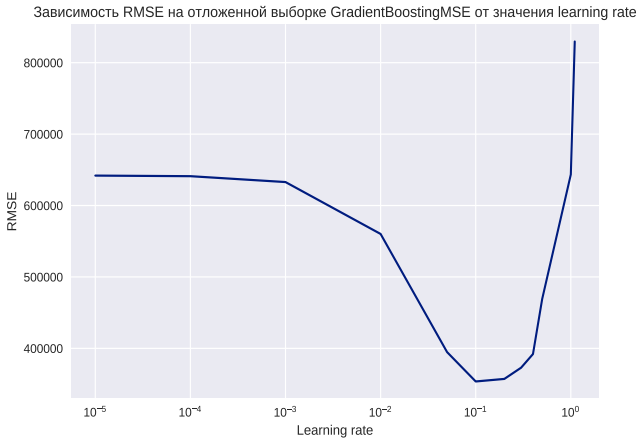
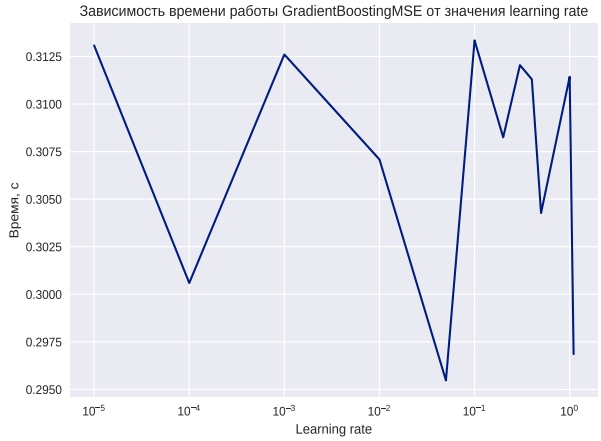


Рис. 14:



Из графика зависимости RMSE от значения learning rate видно, что при:

- learning rate < 0.1 возникает недообучение
- learning rate > 0.1 возникает переобучение
- learning rate $= 0.1$ является оптимальной точкой

Зависимость времени работы от значения learning rate не прослеживается.

3 Выводы

В данном отчете были рассмотрены реализации двух алгоритмов регрессии:

- **RandomForestMSE**
- **GradientBoostingMSE**

Были подтверждены многие теоретические факты, связанные с данными алгоритмами.