Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Отчет по случайному лесу и градиентному бустингу

Выполнил:

Студент 3 курса
Н. А. Панин

Москва
2022

1 Введение

В данном отчете рассматриваются одни из самых мощных алгоритмов классического машинного обучения, а именно *Random Forest* (RF) и *Gradient Boosting Random Forest* (GBRF).

Для исследования влияния гиперпараметров на эти алгоритмы в качестве данных брались данные с соревнования Kaggle о продаже недвижимости **House Sales in King County, USA**. В ходе экспериментов для случайного леса изучалась зависимость RMSE и времени обучения от количества деревьев в ансамбле, размерности подвыборки признаков для узла дерева, максимальной глубины дерева. Для изучения зависимости RMSE и времени обучения для градиентного бустинга помимо печисленных выше гиперпараметров рассматривался еще темп обучения (*learning rate*).

2 Предобработка данных

В любой задаче машинного обучения всегда приходится делать предобработку данных в противном случае можно получить результаты, не удовлетворяющие нашим ожиданиям. В данном случае колонка с датами в силу своего формата (строки вида <yyyymmddT000000>, где y - год, m - месяц, d - день) были удалены и вместо них добавлены три новых признака: год, месяц, день недели. Далее из датасета была удалена целевая переменная и id, чтобы модель на этих данных не переобучилась. Также была проведена проверка на пропущенные значения (NaN), в результате чего оказалось, что таких значений нет. Выборка была разделена на обучающую и валидационную в отношении 7 : 3 соотвественно.

3 Список экспериментов

В ходе данных экспериментов исследовалось влияние на RMSE и время следующих гиперпараметров:

- n_estimators - количество деревьев в ансамбле
- max_depth - максимальная глубина дерева при его построении
- feature_subsample_size - размерность подвыборки признаков для одного дерева
- learning_rate - темп обучения для GBRF

3.1 Эксперимент 1. Исследование Random Forest.

3.1.1 Дизайн эксперимента

В данном эксперименте в качестве значений по умолчанию для изучаемых гиперпараметров брались следующие значения:

- n_estimators - 100
- max_depth - None (максимальная глубина не ограничена)
- feature_subsample_size - $\lceil \frac{1}{3} \rceil$ от всего признакового пространства

В этом эксперименте при исследовании влияния на модель значений гиперпараметров в качестве уже обработанных параметров фиксировались оптимальные. Выбор наилучшего параметра определялся из значений *RMSE* на валидационной выборке. Таким образом, в ходе эксперимента получалась модель с лучшими параметрами, подобранными жадно.

3.1.2 Результаты эксперимента

1. **max_depth**

   Все параметры по умолчанию фиксировались и перебирался max_depth (см. рис. 1). Из графика видно, что с ростом глубины дерева RMSE

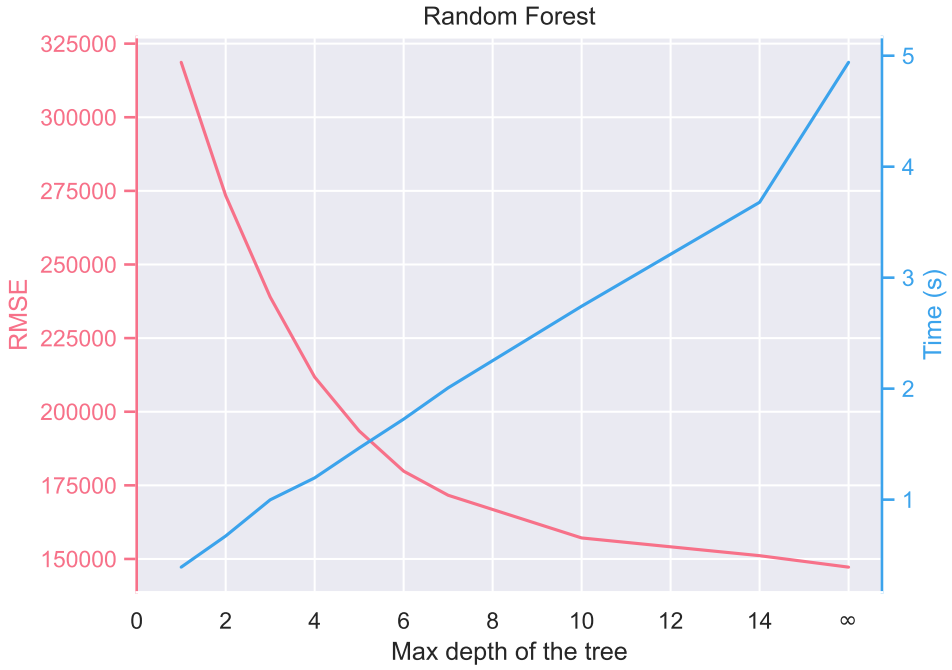


Рис. 1

   уменьшается гиперболически, причем для ∞ (рост дерева в глубину не ограничен) наблюдаем лучшее RMSE $\approx$147215.3. Это не удивительно, поскольку случайный лес хорошо работает для глубоких деревьев. Время при этом увеличивается, причем линейно, если не учитывать max_depth $= \infty$. Увеличение времени связано с тем, что теперь алгоритм при построении очередного дерева для ансамбля может остановиться раньше из-за ограничения по глубине, при том что, возможно, дерево могло расти глубже.

2. **n_estimators**

Фиксировав те же параметры, что и в предыдущем пункте[1], рассматривались различные значения количества деревьев (см. рис. 2). На графике видно, что с ростом количества деревьев в ансамбле

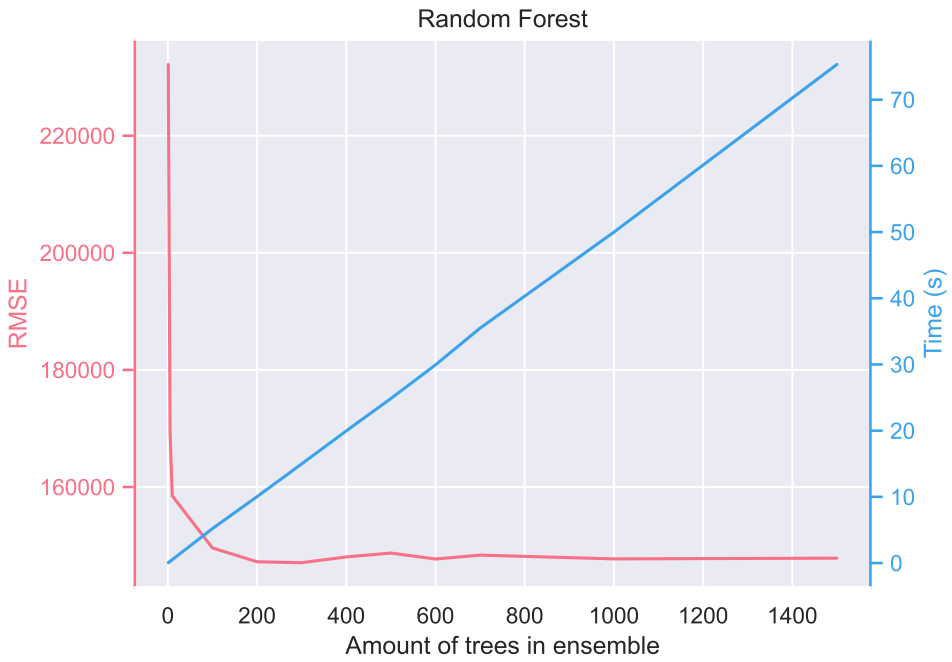


Рис. 2

RMSE уменьшается, достигает минимума при n_estimators = 300 и после этого начинает немного увеличиваться. Такую зависимость можно объяснить разложением RMSE на bias-variance. Так как в этом разложении большую роль играет корреляция базовых моделей, то неудивительно, что с учетом не бесконечной выборки и не бесконечного признакового пространства, что в какой-то момент деревьев будет уже столько, что обязательно множество признаков в каком-нибудь узле пересекутся с множеством признаков в другом узле, а также пересекутся множества обучающих выборок деревьев, что приведет к зависимости этих деревьев, отсюда возрастание RMSE после 300. Уменьшение RMSE в начале можно объяснить тем, что разброс уменьшается с возрастанием количества деревьев при том, что корреляция между деревьями еще мала. Плато же получается в результате компенсирования друг друга факторов, названных выше. Время также увеличивается с ростом числа деревьев.

3. **feature_subsample_size**

[1] max_depth = ∞ оказался оптимальным, поэтому его не меняем

Фиксировав те же параметры, что и в предыдущем пункте, и поменяв n_estimators на 300, рассматривались значения размера подвыборки признаков для одного узла (см. рис. 3). Данный гиперпа-

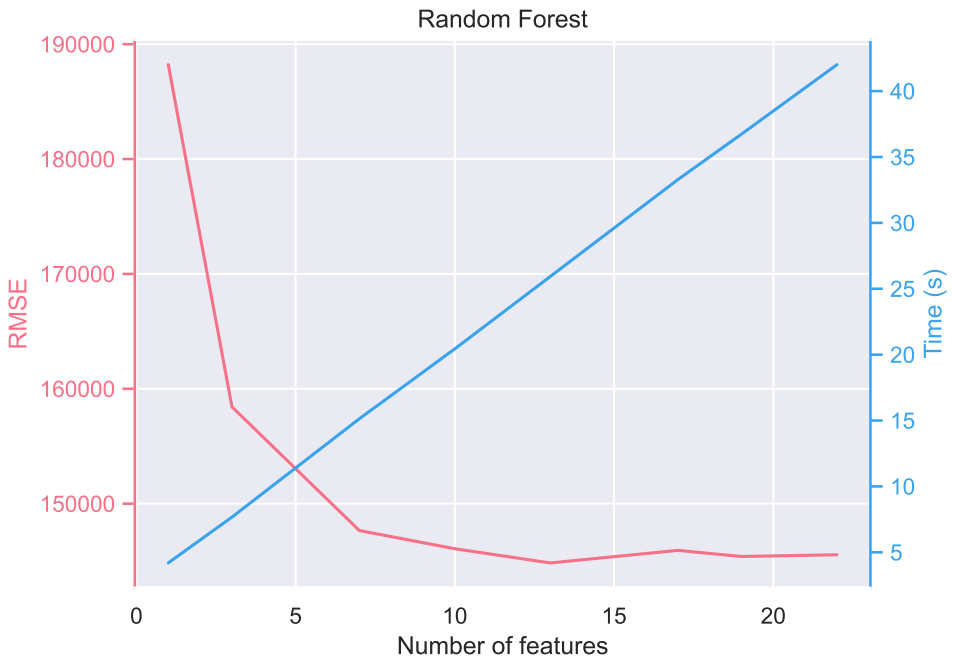


Рис. 3

раметр хоть и отвечает за отличную от n_estimators сущность все же похож с ним, и зависимость, которую видно на графике можно объяснить подобно предыдущему пункту. Действительно, сначала, когда растет количество признаков, у одного дерева больше возможностей подстроиться под лучший признак и так, чтобы коррелированность деревьев во всем ансамбле была малой. После достижения feature_subsample_size = 13 RMSE начинает расти, так как деревья начинают в большей степени зависеть друг от друга. Время увеличивается линейно.

3.1.3 Вывод из эксперимента

Таким образом, в результате эксперимента для исследуемых данных было показано, что для Random Forest:

- глубина деревьев лучше неограниченная
- количество деревьев стоит увеличивать в ансамбле, но не слишком сильно

- количество рассматриваемых признаков в одном узле при построении дерева необходимо уменьшать до определенного значения

3.2 Эксперимент 2. Исследование Gradient Boosting Random Forest.

3.2.1 Дизайн эксперимента

В данном эксперименте в качестве значений по умолчанию для изучаемых гиперпараметров брались следующие значения:

- n_estimators - 100
- max_depth - 3
- feature_subsample_size - $\lceil \frac{1}{3} \rceil$ от всего признакового пространства
- learning_rate - 0.1

В этом эксперименте при исследовании влияния на модель значений гиперпараметров в качестве уже обработанных параметров фиксировались оптимальные. Выбор наилучшего параметра определялся по значениям *RMSE* и по времени на валидационной выборки, то есть, если в множестве моделей есть модель, которая дает небольшой прирост в качестве среди всех, но при этом время увеличивается в разы, то берется, та модель (с определенным гиперпараметром), которая оптимальнее по времени. Таким образом, в ходе эксперимента получалась модель с лучшими параметрами, подобранными жадно.

3.2.2 Результаты эксперимента

1. **n_estimators**

   Фиксировав параметры по умолчанию, рассматривались различные значения количества деревьев (см. рис. 4). На графике видно, что с ростом количества деревьев, RMSE уменьшается по закону похожему на гиперболический, достигает минимума при n_estimators = 600 (RMSE ≈130855.7). Время также увеличивается с ростом числа деревьев, причем линейно.

2. **feature_subsample_size**

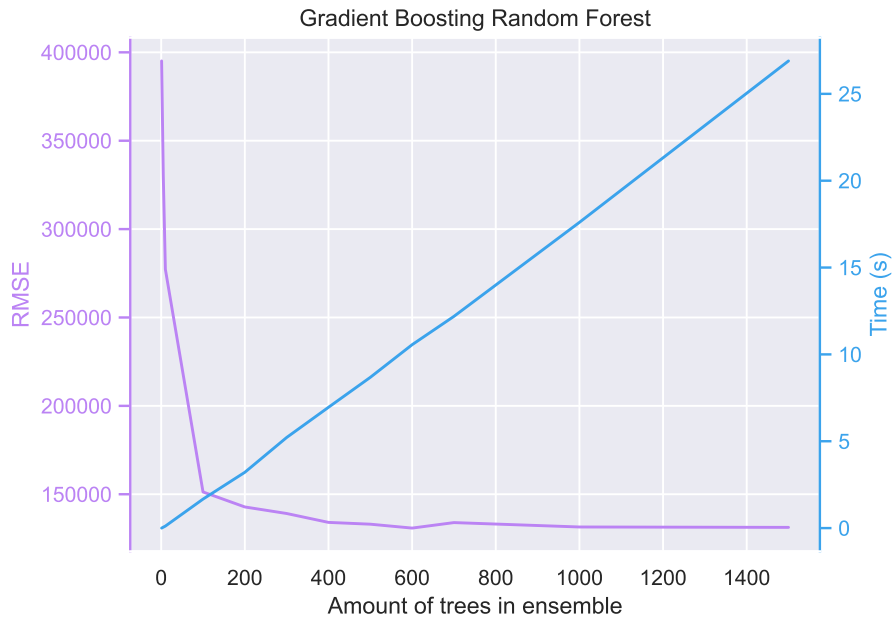


Рис. 4

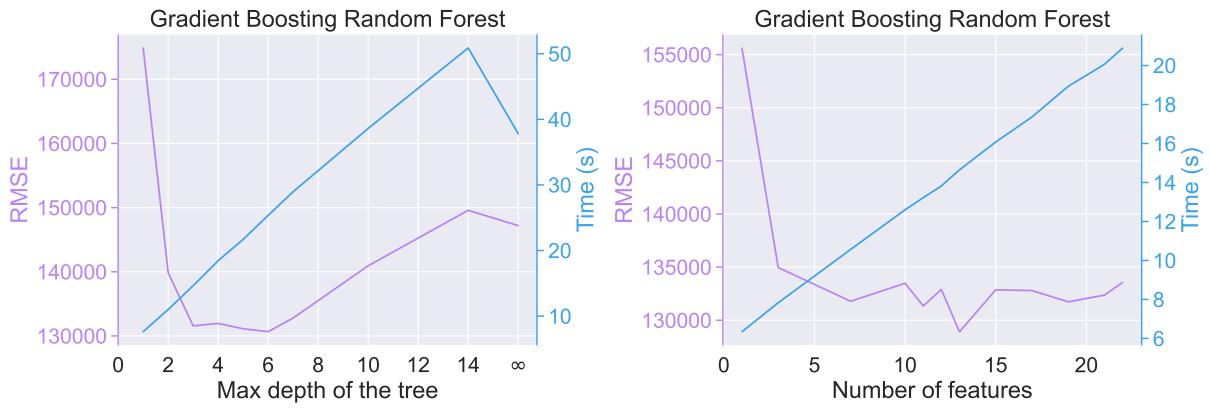
Все параметры из предыдущего пункта фиксировались, а n_estimators брался равным 600. Далее рассматривались разные значения для feature_subsample_size(см. рис. 5b).

Из графика видно, что количество признаков почти ни на что не влияет при их количестве большем 7 (тем не менее лучшее значение это 13 и на этом значении RMSE ≈128913.9). Если количество признаков меньше 7, то RMSE стремительно возрастает.

3. **max_depth**

Фиксировав те же параметры, что и в предыдущем пункте, и поменяв feature_subsample_size на 13, перебирались значения глубины деревьев (см. рис. 5a). Из графика видно, что для бустинга оптимальным выбором max_depth будут не слишком глубокие деревья. Это неудивительно, поскольку бустинг работает несколько иначе чем случайный лес. В нем на каждой следующей итерации алгоритм понижает ошибку композиции. Из-за этого с увеличением количества деревьев bias уменьшается и, если верить, что выполняется bias-variance tradeoff, то при этом увеличивается разброс. Именно поэтому деревья должны быть не глубоким, поскольку у них разброс низкий и при добавлении таких деревьев разброс всей модели будет лишь немного увеличиваться.

4. **learning_rate**



(a) (b)

Рис. 5

После изменения max_depth на 6 перебирались по логарифмической сетке от 10^{-5} до 1 значения learning_rate (см. рис. 6)[2]. Из графика

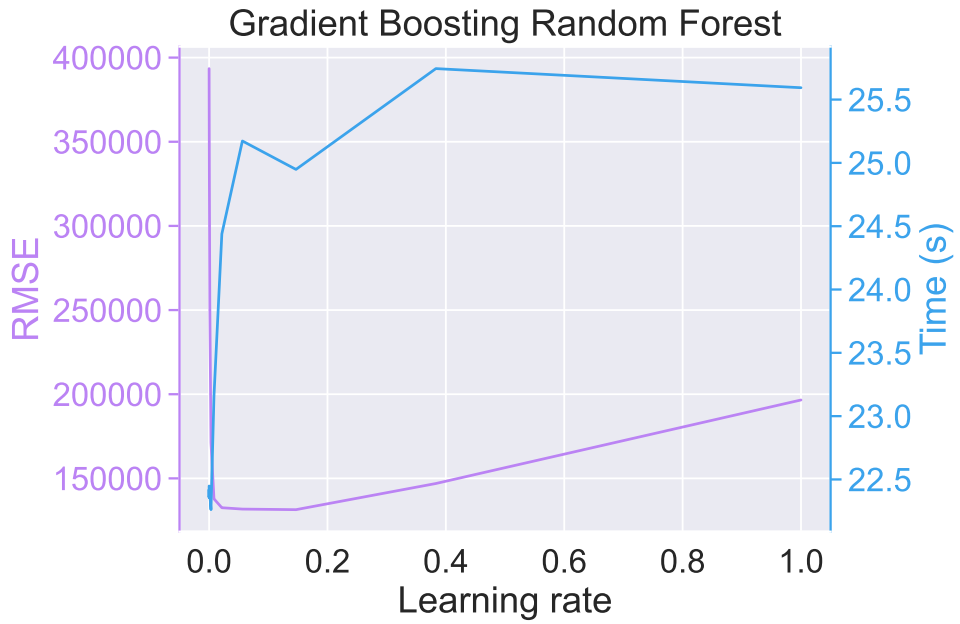


Рис. 6

видно, что слишком маленькие значения (ближе к нулю) и слишком большие (ближе к 1) увеличивают RMSE. Оптимальным значением является learning_rate = 0.147. На нем RMSE = 131412.8. Это согласуется с ожиданиями, так как смысл learning_rate в градиентном бустинге как раз в том, чтобы уменьшить большую способность бу-

[2]большие значения не брались, потому что тогда теряется смысл в регуляризации, RMSE начнет стремиться к бесконечности

стинга к переобучению. Время обучения увеличивается и, достигнув определенного значения почти не изменяется.

3.2.3 Вывод из эксперимента

Таким образом, в результате эксперимента для исследуемых данных было показано, что для Gradient Boosting Random Forest:

- глубина деревьев должна быть не большой (4-6 уровней)
- количество деревьев имеет смысл увеличивать в бустинге для повышения обобщающей способности
- количество рассматриваемых признаков почти не влияет на работу бустинга
- темп обучения стоит подбирать, поскольку, он оказывает положительный эффект на обобщающую способность. learning_rate брать примерно в пределах от 0.09 до 0.15 для исследуемых данных.

4 Общие выводы

В рамках проведенного исследования были достигнуты поставленные цели и решены сформулированные в начале исследования задачи. Особенно хотелось бы выделить следующие результаты:

- В случайном лесе, как и в бустинге, большую роль в обучении играет количество базовых моделей. Хотя этот гиперпараметр надо подбирать отдельно, все же можно сформулировать общее правило: больше деревьев равно лучше.
- Размерность подвыборки признаков для одного узла является существенным параметром в случае со случайным лесом в отличие от бустинга, на который этот параметр почти не влияет.
- Для бустинга следует использовать не глубокие деревья (4-6 уровней), а для случайного леса глубокие (не ограниченный рост)
- Темп обучения оказывает положительное воздействие на бустинг. Общая тенденция следующая: приближаясь к нулю (до $\approx$0.09) RMSE уменьшается.