Отчет о практическом задании «Ансамбли алгоритмов для решения задачи регрессии»

Практикум 317 группы, ММП ВМК МГУ

Овсиенко Олеся Павловна

Декабрь 2024

Содержание

1 Введение **1**

2 Эксперименты **2**

2.1 Предобработка данных 2

2.2 Исследование алгоритма RandomForest 6

2.3 Исследование алгоритма GradientBoosting 8

3 Заключение **10**

1 Введение

Данное практическое задание направлено на глубокое исследование свойств ансамблей и композиций алгоритмов в сфере машинного обучения, используя в качестве примеров случайный лес и градиентный бустинг в контексте задачи предсказания стоимости недвижимости.

Основными целями исследования являются:

- Детальное изучение используемых данных
- Разработка собственных реализаций рассматриваемых алгоритмов
- Изучение зависимости ошибки RMSE от различных параметров этих алгоритмов

2 Эксперименты

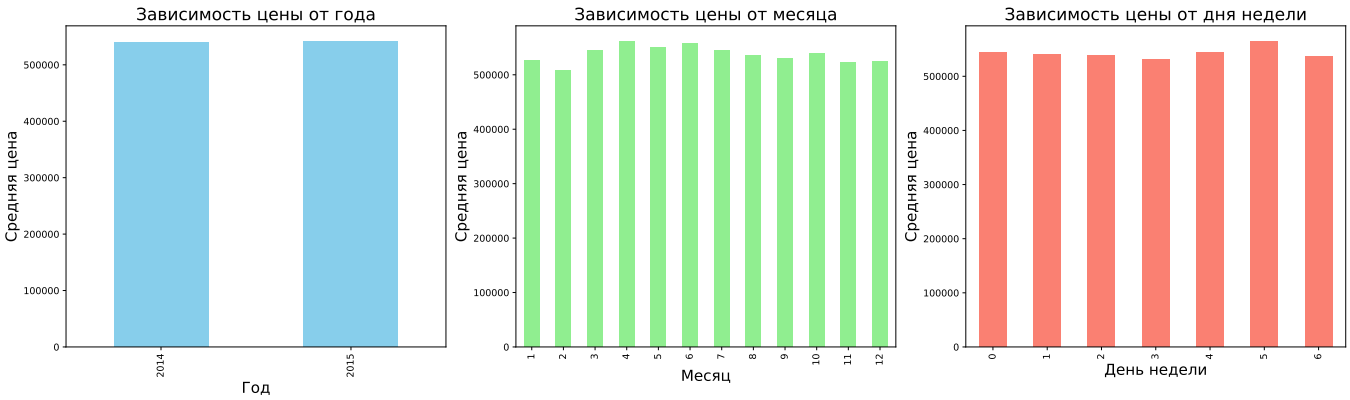
2.1 Предобработка данных

Датасет "House Sales in King County, USA"— это популярный набор данных, который используется для различных задач машинного обучения, включая регрессионный анализ. Он содержит информацию о продажах домов в округе Кинг (штат Вашингтон, США) в период с мая 2014 года по май 2015 года.

Датасет содержит 21 признак:

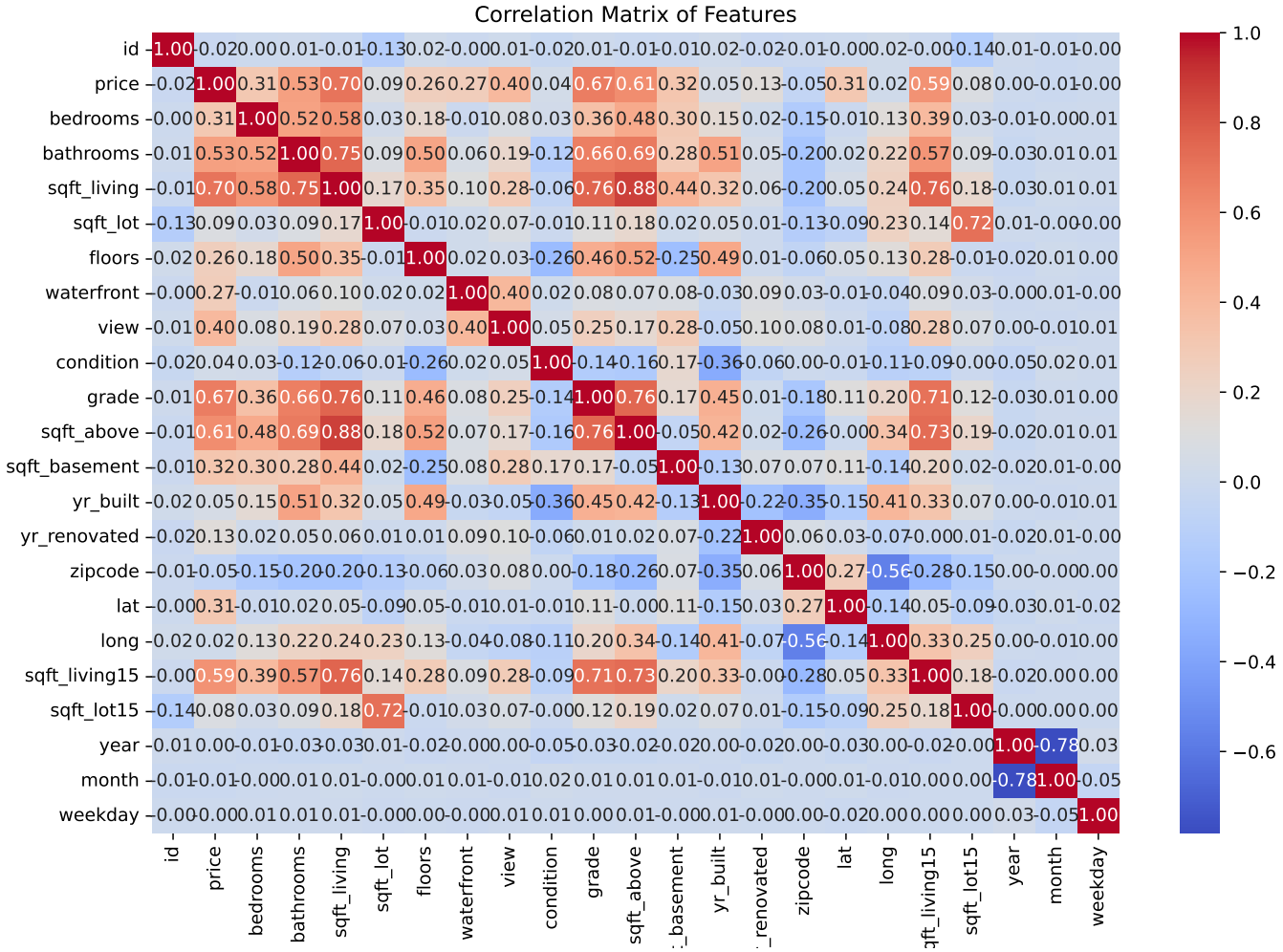
- id — уникальный идентификатор для каждого дома
- date - дата продажи дома
- price — стоимость продажи дома
- bedrooms — количество спален в доме
- bathrooms — количество ванных комнат(где .5 - это туалет без душевой)
- sqft_living — площадь дома в квадратных футах
- sqft_lot — площадь участка в квадратных футах
- floors — количество этажей в доме
- waterfront — наличие вида на набережную (1 — дом с видом на воду, 0 — без)
- view — рейтинг вида с дома (от 0 до 4)
- condition — состояние дома (от 1 до 5, где 1 — плохое состояние, 5 — отличное)
- grade — оценка дома (12 возможных значений)
- sqft_above — площадь дома, расположенная над землей (без подвала)
- sqft_basement — площадь подвала
- yr_built — год постройки дома
- yr_renovated — год последнего ремонта дома
- zipcode — почтовый индекс
- lat — географическая широта расположения дома
- long — географическая долгота расположения дома
- sqft_living15 — площадь дома в квадратных футах, измеренная в 2015 году
- sqft_lot15 — площадь участка в квадратных футах, измеренная в 2015 году

Пропуски в данных отсутствовали. В отличие от остальных признаков признак **date** был типа **object**, что не позволяло работать с признаком, как с остальными. Данный признак был заменён на 3 новых численных признака: **year, month, weekday**. Зависимость цены от новых признаков:



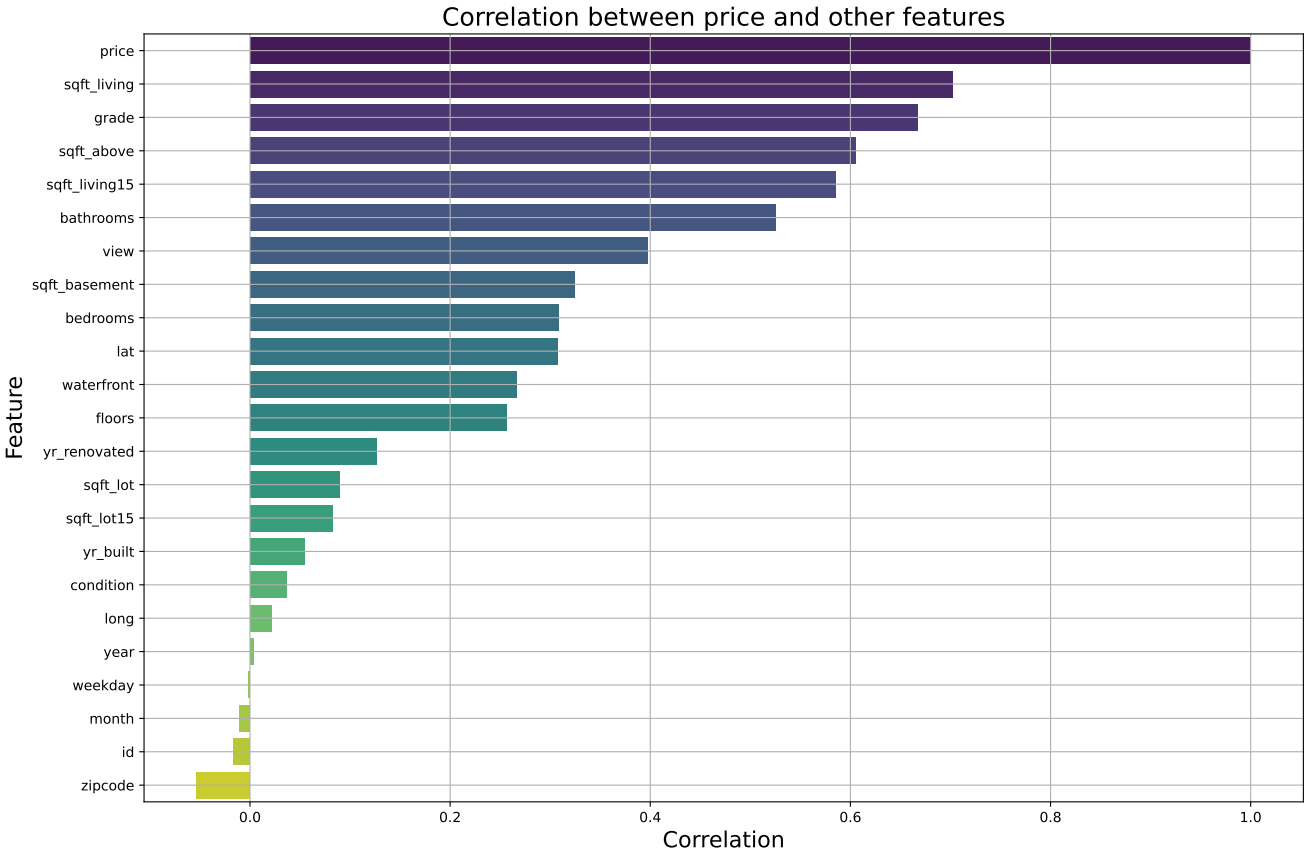
Как видно из графиков, небольшая зависимость цены присутствует только от месяца.

Рассмотрим корреляцию между признаками:



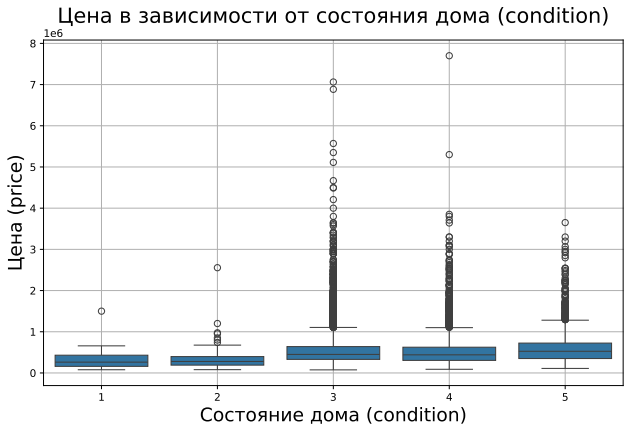
Как видно из корреляционной матрицы, признаки **sqft_living** и **sqft_above** сильно коррелируют между собой. Следовательно, мы можем оставить один из них.

Отдельно рассмотрим корреляцию признака **price** со всеми остальными признаками:



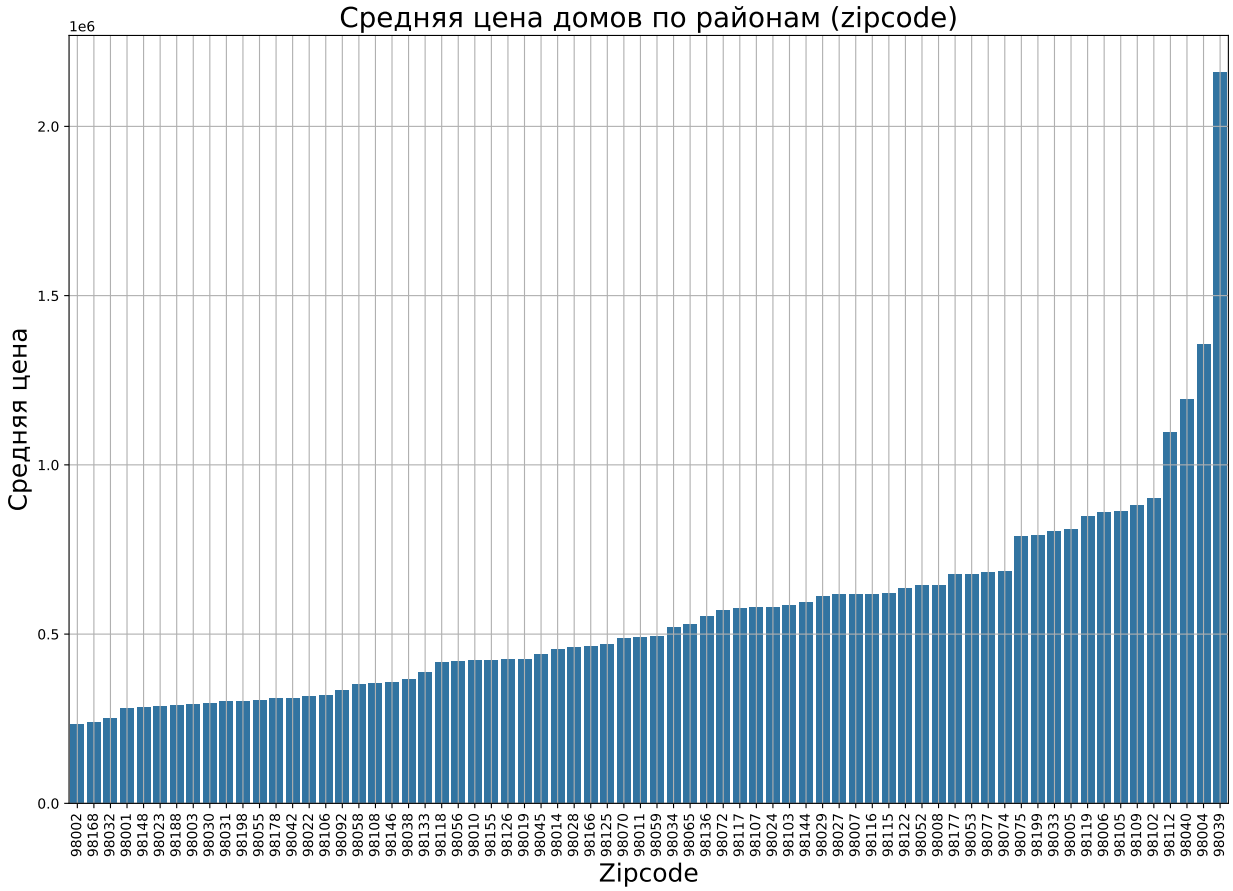
Рассмотрим признаки, которые слабо коррелируют с признаком **price**, а именно **zipcode, id, month, weekday, year, long, condition**:

- Признаки: **id, month, weekday, year** - настолько слабо коррелируют с признаком **price**, что было принято решение их удалить из датасета
- Признак **condition**: Для лучшего понимания данного признака рассмотрим boxplot график:



Как видно из графика, если распределения цен при **condition** = { 1; 2}, совпадают, то при **condition** = { 3, 4, 5} распределения уже отличаются и присутствует больше выбрасов. Следовательно, данный признак может быть полезным.

- Признак **zipcode**:



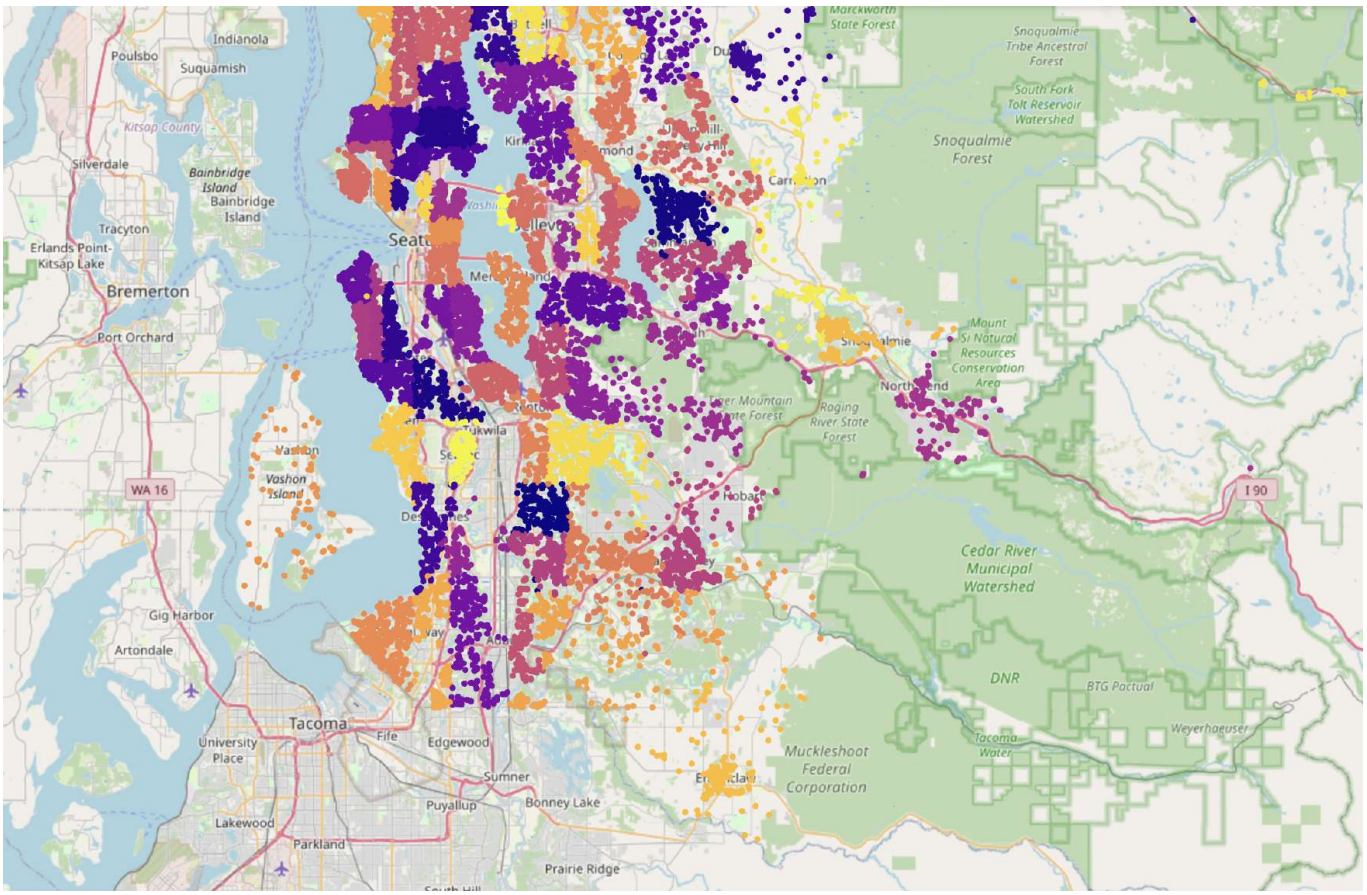
Как видно из графика, от района зависит средняя цена недвижимости, что логично, так как бывают "богатые"районы, а бывают "бедные".

Следовательно, можно предположить, что цена будет зависеть от этого признака.

- Признак **long**:

На карте, представленной ниже, отображены разными цветами разные районы (в зависимости от значения признака **zipcode**) Как видно на карте, в данном датасете все предложения по недвижимости расположены приблизительно на одной долготе.

Следовательно, для работы с датасетом **"House Sales in King County, USA"** данный признак не нужен.



Таким образом, в результате исследования было выяснено отсутствие необходимости использования таких признаков, как **id, sqft_above, long, year, weekday, month**. Эти признаки были удалены.

Оставшиеся признаки были разделены на численные и категориальные следующим образом:

categorial_features = [**waterfront, floors, view, condition, grade, zipcode**]

numerical_features =

= [**bedrooms, bathrooms, sqft_living, sqft_lot, floors, sqft_basement, yr_built, yr_renovated, lat, sqft_living15, sqft_lot15**]

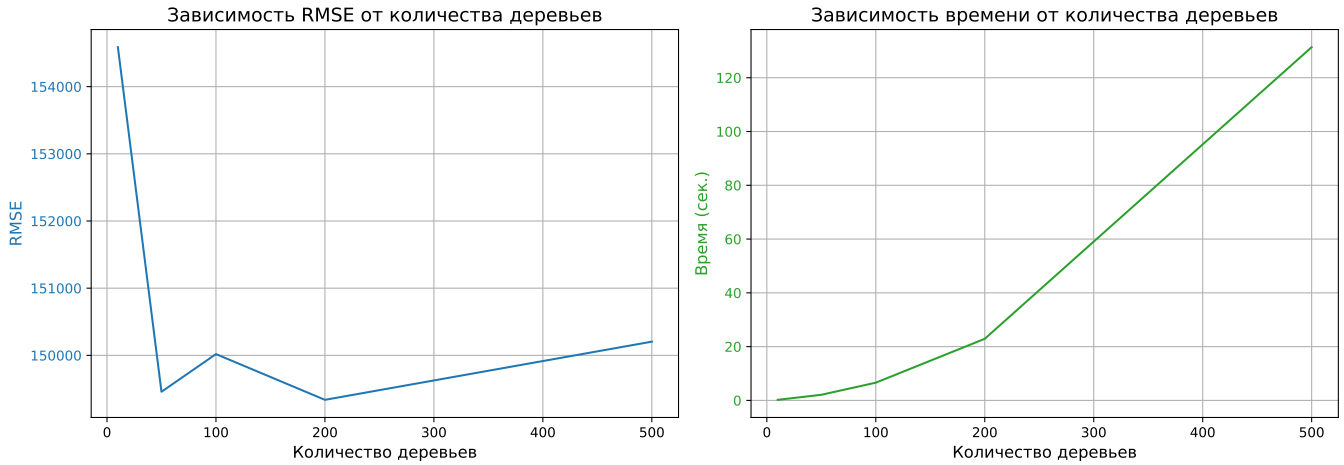
- Для дальнейшей работы с данными:
- к численным признакам был применён **StandardScaler**,
- к категориальным - **TargetEncoder** (не **OneHotEncoder**, так как уникальных значений только признака **zipcode** 70, что очень сильно увеличило бы датасет).

2.2 Исследование алгоритма RandomForest

Было исследовано поведение алгоритма случайный лес, рассматривая **RMSE** на отложенной выборке и **время работы** алгоритма в зависимости от следующих факторов (при изучении зависимости

поведения алгоритма от какого-либо параметра остальные параметры принимали следующие значения: **n_estimators** = 100, **max_features** = 1/3 (как было указано на лекции), **max_depth** = 5):

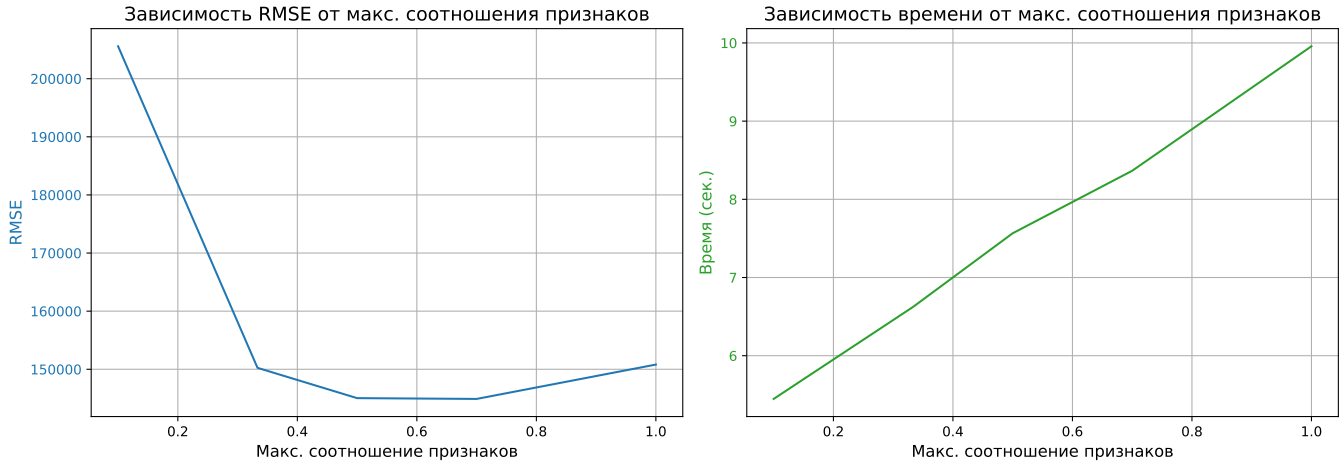
- **Количество деревьев в ансамбле:**



С увеличением числа деревьев в алгоритме увеличивается время обучения.

Минимальное RMSE = 149337.757 при времени обучения = 22.909 было достигнуто при 200 деревьях.

- **Размерность подвыборки признаков для одной вершины дерева:**



С увеличением размерности подвыборки признаков для одной вершины дерева увеличивается время обучения.

Минимальное RMSE = 144897.768 при времени обучения = 8.365 было достигнуто при подвыборке признаков = 0.7 от всех признаков.

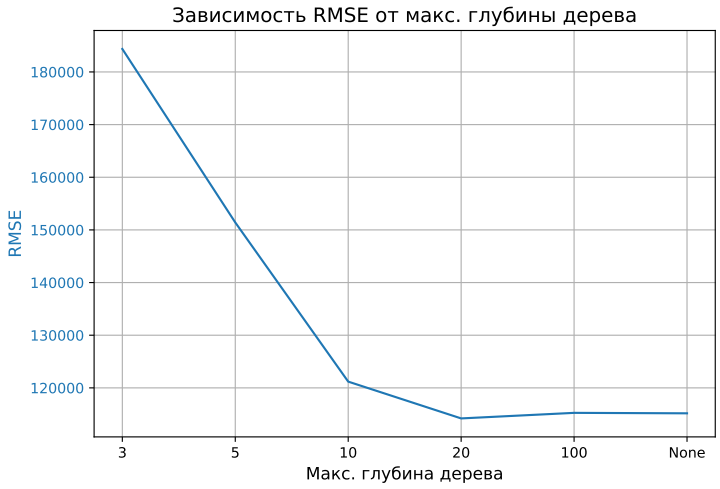
- **Максимальная глубина дерева:**

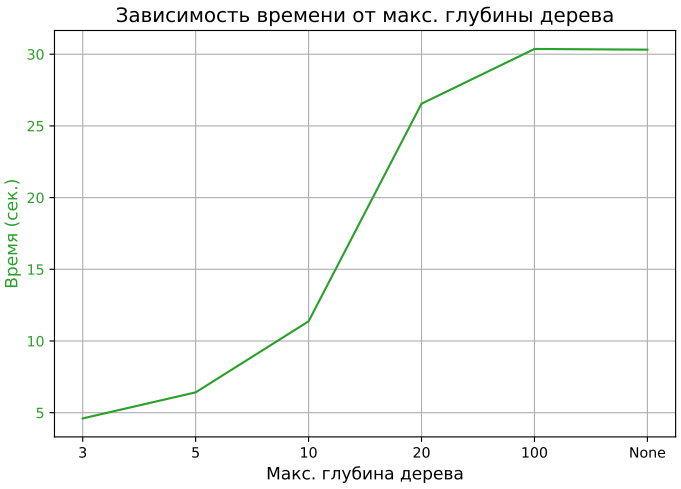
С увеличением максимальной глубины дерева увеличивается время обучения.

Минимальное RMSE = 114198.94 при времени обучения = 26.548 было достигнуто при максимальной глубине дерева = 20.

В случае, когда глубина не ограничена, RMSE = 115167.868 при времени обучения = 115167.868.

Таким образом, при неограниченной глубине значение **RMSE** в градиентном бустинге близко к минимальному значению, однако показатель времени получился сильно больше.

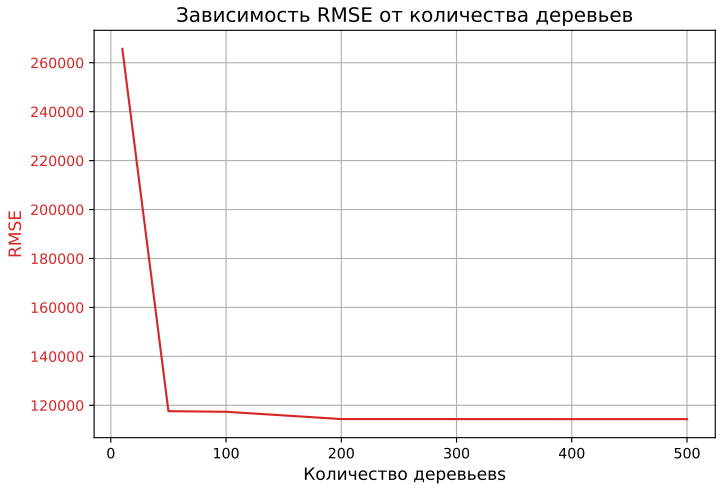


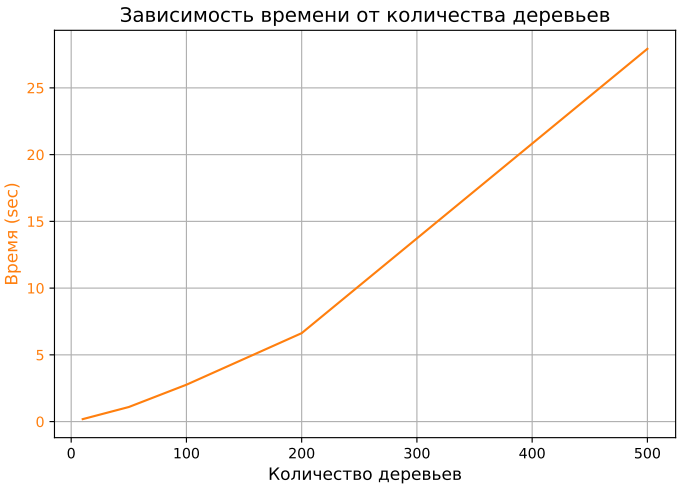


2.3 Исследование алгоритма GradientBoosting

Было исследовано поведение алгоритма градиентный бустинг, рассматривая **RMSE** на отложенной выборке и **время работы** алгоритма в зависимости от следующих факторов (при изучении зависимости поведения алгоритма от какого-либо параметра остальные параметры принимали следующие значения: **n_estimators** = 100, **learning_rate** = 0.1, **max_features** = 1/3 (как было указано на лекции), **max_depth** = 5):

- **Количество деревьев в ансамбле:**





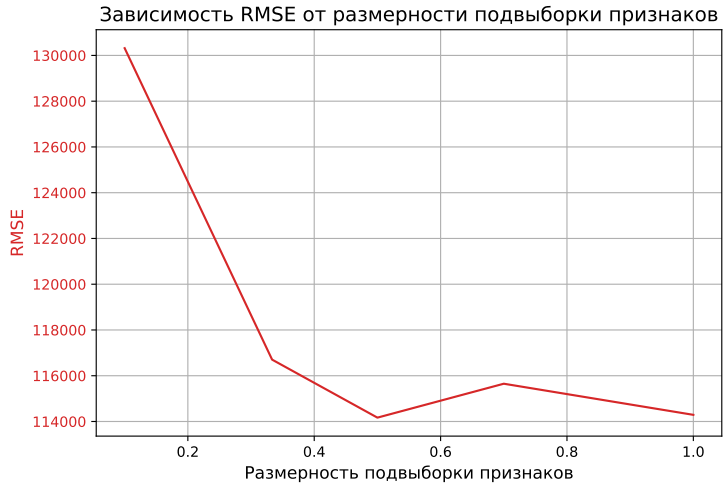
С увеличением числа деревьев в алгоритме увеличивается время обучения.

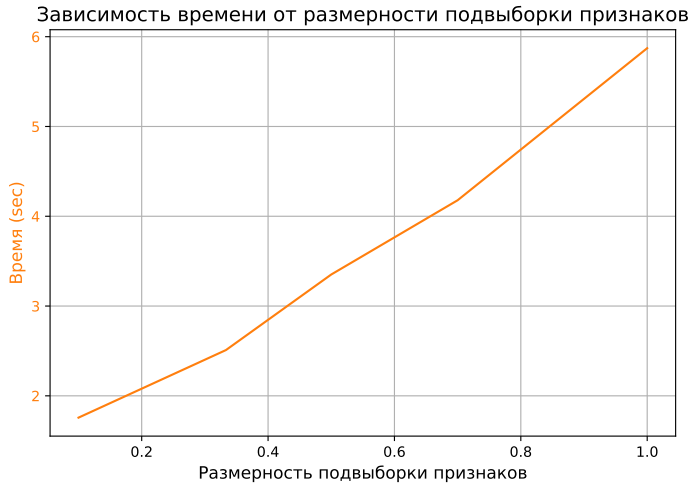
Минимальное RMSE = 114345.708 при времени обучения = 27.925 было достигнуто при 500 деревьях.

- **Размерность подвыборки признаков для одной вершины дерева:**

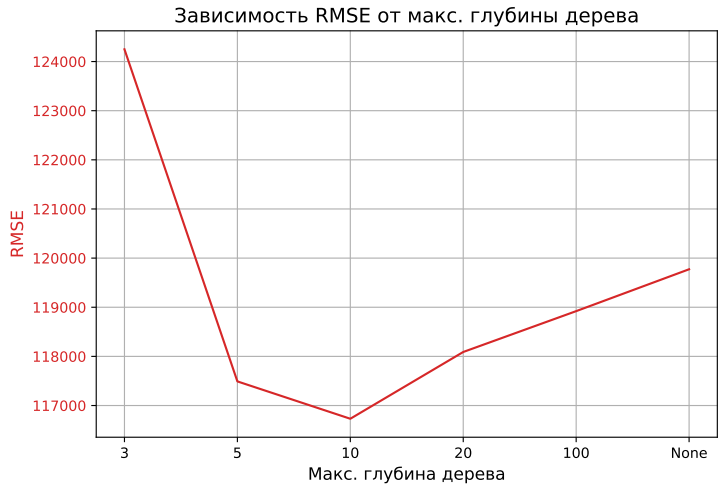
  С увеличением размерности подвыборки признаков для одной вершины дерева увеличивается время обучения.

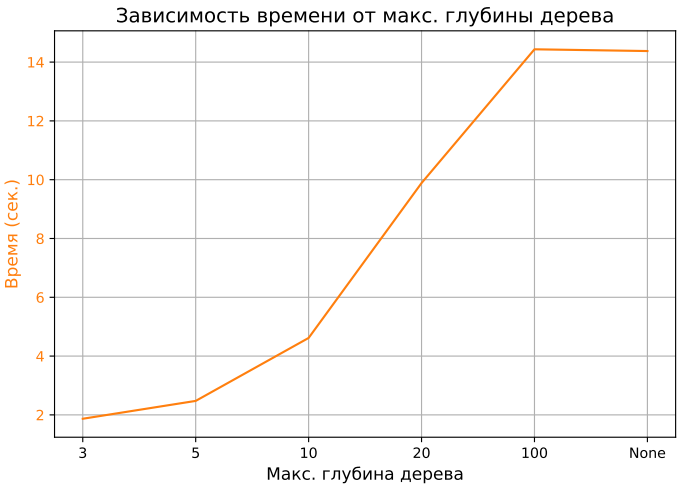
  Минимальное RMSE = 114168.357 при времени обучения = 3.351 было достигнуто при подвыборке признаков = 0.5 от всех признаков.





- **Максимальная глубина дерева:**





  С увеличением максимальной глубины дерева увеличивается время обучения.

  Минимальное RMSE = 116730.819 при времени обучения = 4.617 было достигнуто при максимальной глубине дерева = 10.

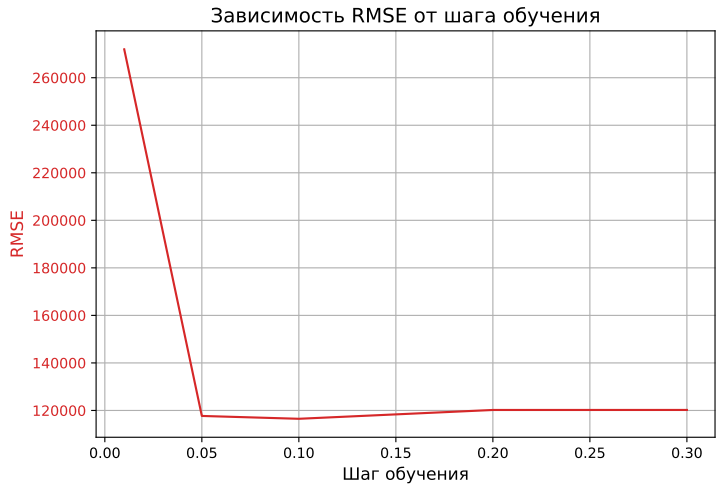
  В случае, когда глубина не ограничена, RMSE = 119772.845 при времени обучения = 119772.845.

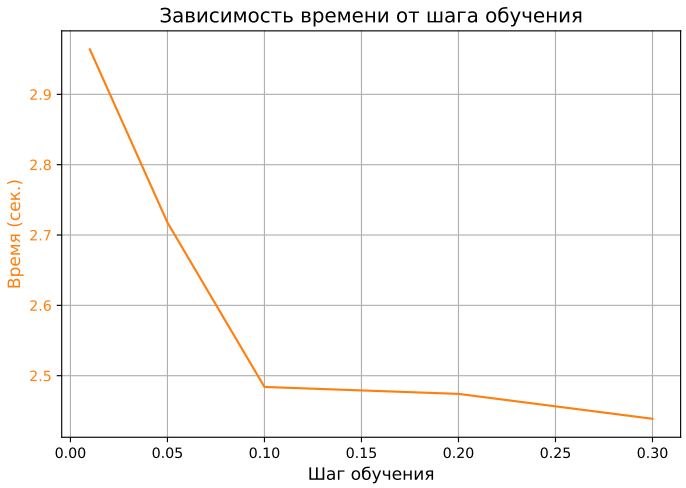
  Таким образом, при неограниченной глубине в градиентном бустинге показатели сильно хуже.

- **Learning_rate:**

  Из графика видно, что с увеличением шага обучения, уменьшается время обучения.

  Минимальное RMSE = 116495.165 при времени обучения = 2.484 было достигнуто при шаге обучения = 0.1.





3 Заключение

Таким образом, в результате экспериментов было выяснено, что:

Для **RandomForest** рекомендуется:

- Около 200 деревьев (необходимо ограничивать число деревьев, так как при большем числе растёт **RMSE**)
- Подвыборку признаков в размере 70% (полезно брать как можно больше признаков)
- Не ограничивать глубину дерева

Для **GradientBoosting** рекомендуется:

- Около 500 деревьев (нужно больше, чем для RandomForest)
- Подвыборку признаков в размере 50%
- Ограничивать глубину дерева до 10 уровней (базовые модели градиентного бустинга должны быть простыми)
- Использовать шаг обучения равный 0.1 (следует брать небольшим)

Подводя итоги исследования, также отметим, что в проведённых экспериментах алгоритм случайный лес показал хуже результаты: как **RMSE**, так и **время обучения** больше, чем у алгоритма градиентный бустинг. Следовательно, можем сделать вывод, что алгоритм градиентный бустинг лучше подходит для работы с датасетом "House Sales in King County, USA".