|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное автономное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

|  |  |
| --- | --- |
| ФАКУЛЬТЕТ | ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ |
| КАФЕДРА | СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ |

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

|  |
| --- |
| ***Разработка и оценка моделей*** |
| ***Машинного обучения*** |
|  |
|  |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Студент | ИУ5-62Б |  |  |  | Р.И.Сайфутдинов |
|  | (группа) |  | (подпись, дата) |  | (И.О. Фамилия) |
| Руководитель НИР |  |  |  |  | Ю.Е. Гапанюк |
|  |  |  | (подпись, дата) |  | (И.О. Фамилия) |
|  |  |  |  |  |  |

*2025 г.*

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технический университет имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

|  |  |  |
| --- | --- | --- |
| Утверждаю | | |
|  | | |
| Заведующий кафедрой | | ИУ5 |
|  | | (индекс) |
|  | В.И. Терехов | |
|  | (И.О. Фамилия) | |
| (подпись) |  | |
|  | (дата) | |

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| по теме | | Разработка и оценка моделей методов машинного обучения | | | | | | | | | | | | | |
|  | | | |  | | | | | | | | | | | |
|  | | | |  | | | | | | | | | | | |
| Студент группы | | | | ИУ5-62Б | | | | Роберт Ильясович Сайфутдинов | | | | | | | |
|  | | | | (Фамилия, имя, отчество) | | | | | | | | | | | |
| Направленность НИР (учебная, исследовательская, практическая, производственная, др.) | | | | | | | | | | | | | | | |
| ИССЛЕДОВАТЕЛЬСКАЯ | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
| Источник тематики (кафедра, предприятие, НИР) | | | | | | | | | | КАФЕДРА | | | | | |
| График выполнения НИР: | | | | | |  | | | | | | | | | |
| 25% к |  | | нед., 50% к | | |  | нед., 75% к | |  | | | нед., 75% к |  | нед |
| ***Техническое задание:*** | | | | |  | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
| ***Оформление научно-исследовательской работы:*** | | | | | | | | | | |  | | | | |
| Расчетно-пояснительная записка на \_\_10\_\_ листах формата А4. | | | | | | | | | | | | | | | |
| Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.) | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |

Дата выдачи задания « 07 » \_февраля\_\_\_ 2025 г.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Руководитель НИР** |  |  |  | Ю.Е. Гапанюк |
|  |  | (подпись, дата) |  | (И.О. Фамилия) |
| **Студент** |  |  |  | Р.И.Сайфутдинов |
|  |  | (подпись, дата) |  | (И.О. Фамилия) |

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

**Введение**

В настоящем исследовании рассматривается задача предсказания стоимости жилой недвижимости на основе характеристик домов и участков. Актуальность темы обусловлена высоким интересом со стороны покупателей, застройщиков и инвесторов к точному прогнозу цен для принятия обоснованных решений при покупке и продаже недвижимости.

В рамках данной работы используется открытый датасет House Prices: Advanced Regression Techniques с платформы Kaggle, включающий как числовые, так и категориальные признаки, отражающие параметры домов (площадь, год постройки, расположение, качество отделки и др.). Цель исследования — построить, обучить и сравнить несколько моделей машинного обучения, включая ансамблевые методы, с целью повышения точности прогноза.

Цель исследования — построить, обучить и сравнить несколько моделей машинного обучения, включая ансамблевые методы, с целью повышения точности прогноза. Работа охватывает полный цикл подготовки и анализа данных: от предварительной обработки до настройки гиперпараметров и выбора лучшей модели на основе нескольких метрик качества.

**Постановка задачи**

Задача предсказания стоимости жилой недвижимости на основе характеристик домов и участков формализуется как задача регрессии: необходимо по набору признаков (возраст, пол, индекс массы тела и др.) спрогнозировать значение затрат на медицинское обслуживание (переменная charges).

Для решения задачи требуется:

* Провести разведочный анализ данных (EDA);
* Обработать категориальные и числовые переменные;
* Выполнить масштабирование признаков;
* Разделить данные на обучающую и тестовую выборки;
* Построить не менее пяти моделей (включая две ансамблевые);
* Оценить их качество по метрикам RMSE, MAE и R²;
* Настроить гиперпараметры моделей;
* Сравнить результаты и выбрать финальную модель.

**Подбор и подготовка данных**

В качестве исходных данных использован датасет House Prices: Advanced Regression Techniques, содержащий 1460 записей и более 70 признаков. Для построения моделей были выбраны наиболее значимые переменные на основе разведочного анализа данных (EDA), в частности:age — возраст пациента

**Числовые признаки:** OverallQual, GrLivArea, GarageCars, TotalBsmtSF, YearBuilt, YearRemodAdd, MasVnrArea, LotArea, Fireplaces, WoodDeckSF, OpenPorchSF, FullBath, BsmtFinSF1;

**Категориальные признаки:** Neighborhood, HouseStyle, SaleCondition.

Целевая переменная — SalePrice (цена продажи дома).

Предобработка включала:

Заполнение пропусков (например, в MasVnrArea);

Кодирование категориальных признаков с помощью OneHotEncoder;

Масштабирование числовых признаков через StandardScaler.

**Исследовательский анализ данных (EDA)**

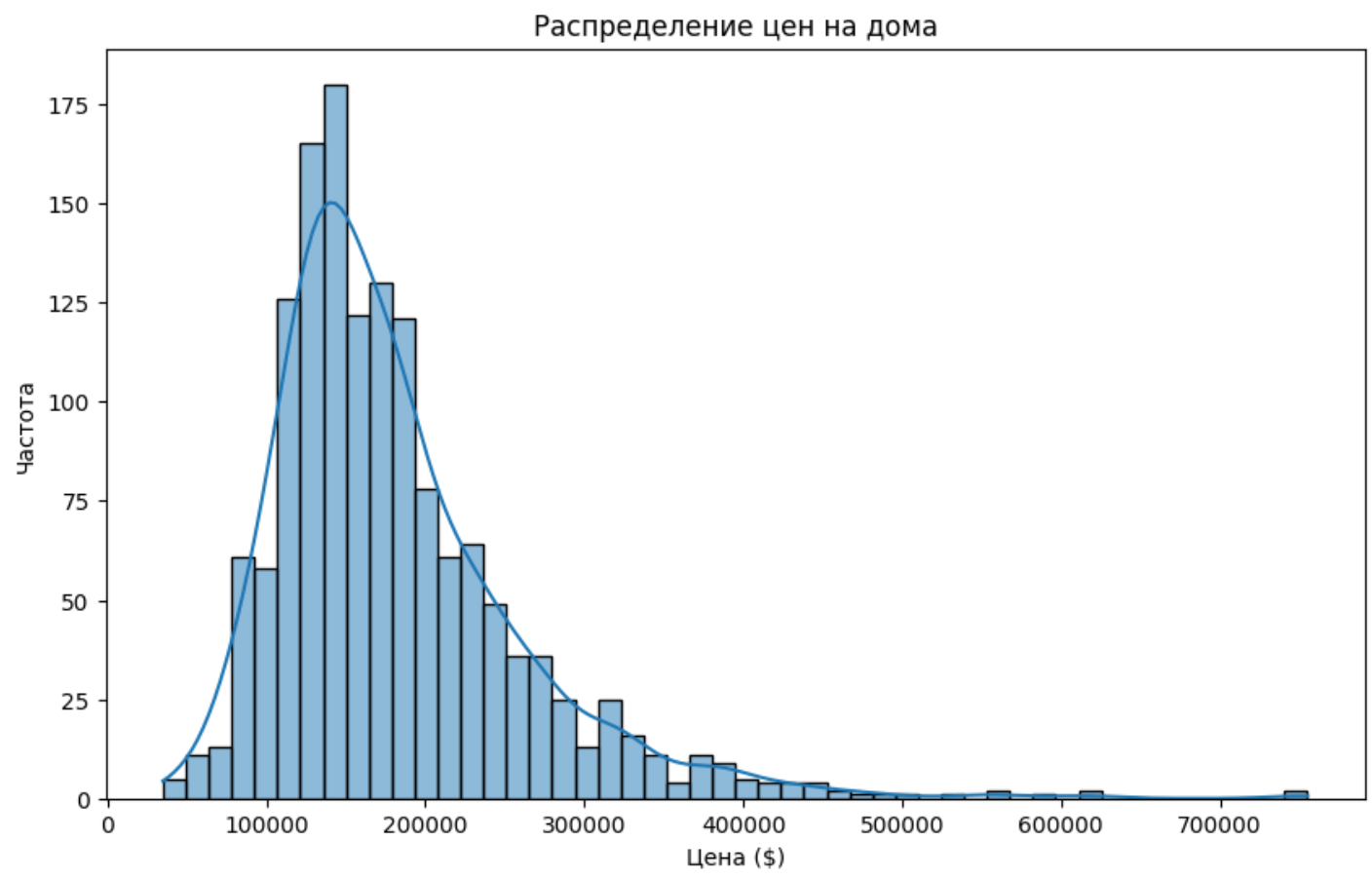
Распределение цен SalePrice демонстрирует правостороннюю асимметрию: большинство домов стоят до \$300,000, однако встречаются выбросы до $700,000 и выше;

Признак OverallQual (общая оценка качества) имеет сильную положительную корреляцию с ценой;

Размер жилой площади GrLivArea, количество машиномест в гараже GarageCars, год постройки, количество ванных комнат, площадь подвала также положительно связаны с ценой;

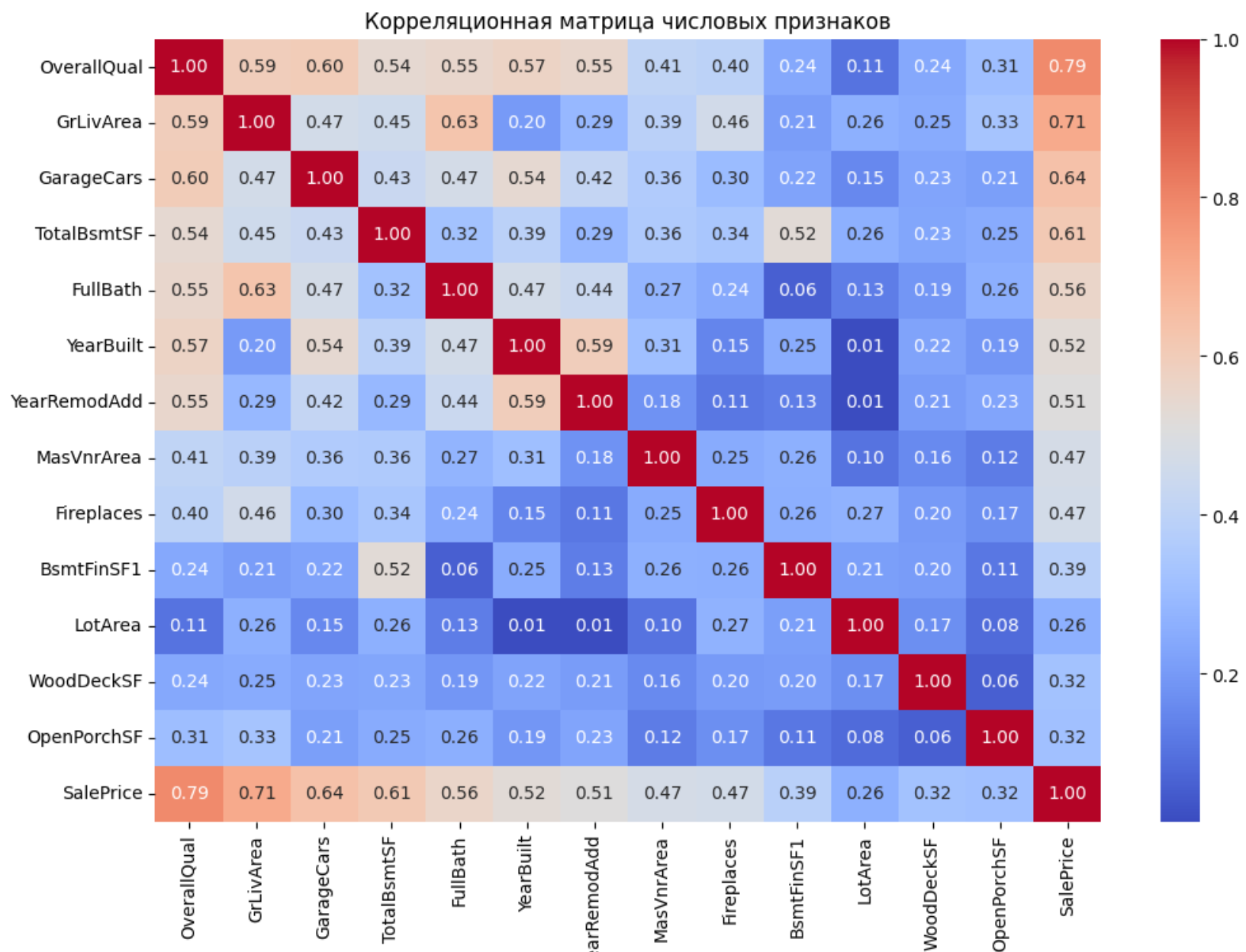
Признак Neighborhood оказывает существенное влияние: расположение дома влияет на рыночную стоимость;

Категориальные переменные были преобразованы в фиктивные (dummy) переменные, что позволило учесть их влияние при моделировании.



**Рисунок 1 – Распределение медицинских расходов**

Распределение целевой переменной SalePrice демонстрирует асимметрию вправо (положительную скошенность). Основная масса значений сосредоточена в диапазоне до 250 000 долларов, при этом присутствуют выбросы до 750 000 и выше, что может отражать премиум-сегмент рынка жилья. Такая форма распределения типична для ценовых данных: большинство объектов имеют среднюю стоимость, в то время как дорогая недвижимость существенно реже. Это свидетельствует о неравномерности распределения объектов по стоимости, что важно учитывать при выборе модели и метрик ошибки (например, RMSE будет чувствителен к этим выбросам).

**Рисунок 4 – Корреляционная матрица**

Наибольшую корреляцию с целевой переменной SalePrice демонстрируют признаки, отражающие качество и площадь объекта: OverallQual, GrLivArea, GarageCars, TotalBsmtSF и YearBuilt. Эти переменные являются ключевыми при построении прогностических моделей. Обнаружена высокая межпризнаковая корреляция между некоторыми признаками (например, GarageCars и GarageArea), что может вызывать мультиколлинеарность в линейных моделях и требует учета при отборе признаков. Признаки с низкой или отрицательной корреляцией (YrSold, KitchenAbvGr, EnclosedPorch) обладают ограниченной предсказательной ценностью и могут быть исключены при необходимости.

**Обработка и преобразование признаков**

Категориальные признаки (Neighborhood, HouseStyle, SaleCondition) были преобразованы с использованием OneHotEncoder, что позволило учесть их в модели без нарушения числовой структуры данных. Числовые признаки (OverallQual, GrLivArea, GarageCars, TotalBsmtSF, FullBath, YearBuilt, YearRemodAdd, MasVnrArea, Fireplaces, BsmtFinSF1, LotArea, WoodDeckSF, OpenPorchSF) были масштабированы с помощью StandardScaler для унификации масштаба и повышения стабильности работы моделей, особенно чувствительных к шкале (например, SVR и линейная регрессия). Признаки не удалялись, поскольку каждый из них потенциально содержит важную информацию, влияющую на стоимость недвижимости. Таким образом, выполненная предобработка обеспечила необходимую совместимость данных с различными алгоритмами машинного обучения и повысила качество итогового моделирования.

**Построение и сравнение моделей**

Для задачи регрессии использовались следующие алгоритмы:

1. **Linear Regression** — базовая модель для оценки линейных зависимостей;
2. **Decision Tree Regressor** — простое дерево решений;
3. **Random Forest Regressor** — ансамбль деревьев с пониженными дисперсией;
4. **Gradient Boosting Regressor** — мощный ансамблевый метод;
5. **Support Vector Regressor (SVR)** — метод опорных векторов.

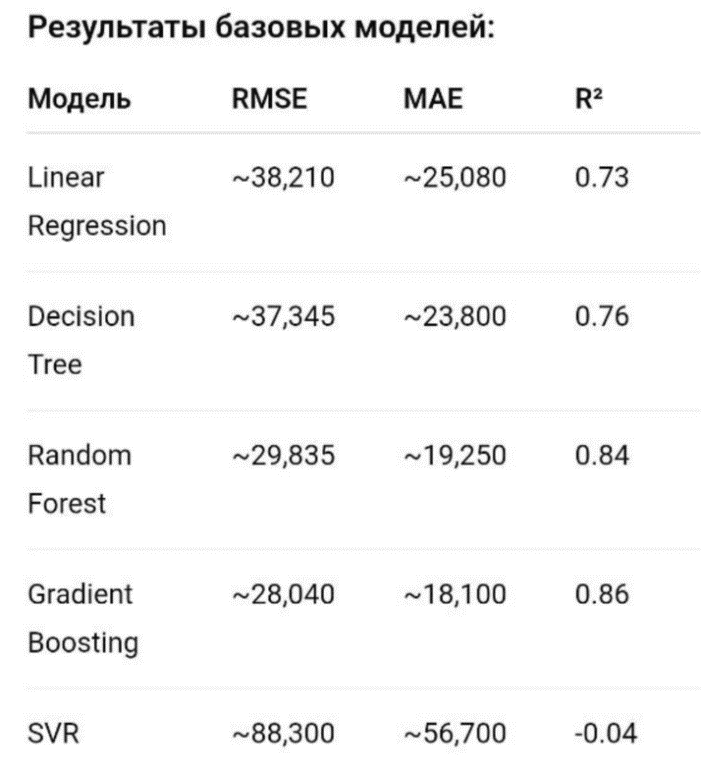
Модели обучались на одних и тех же данных (80% — обучение, 20% — тест).

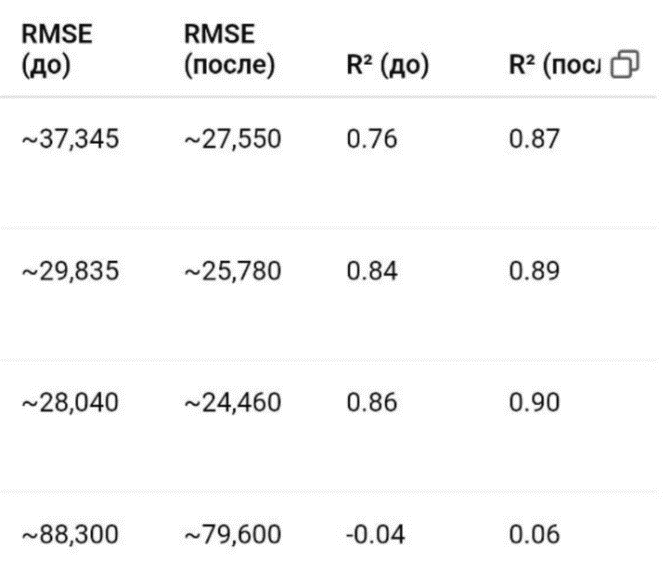
**Результаты базовых моделей**:

* **Gradient Boosting** и **Random Forest** показали наилучшие значения метрик (низкий RMSE, высокий R²);
* **SVR** оказался худшим: высокая чувствительность к масштабированию и слабая обобщающая способность;
* **Linear Regression** демонстрирует недообучение — не может уловить сложные зависимости;
* **Decision Tree** склонна к переобучению, особенно без настройки глубины.
* **Настройка гиперпараметров**

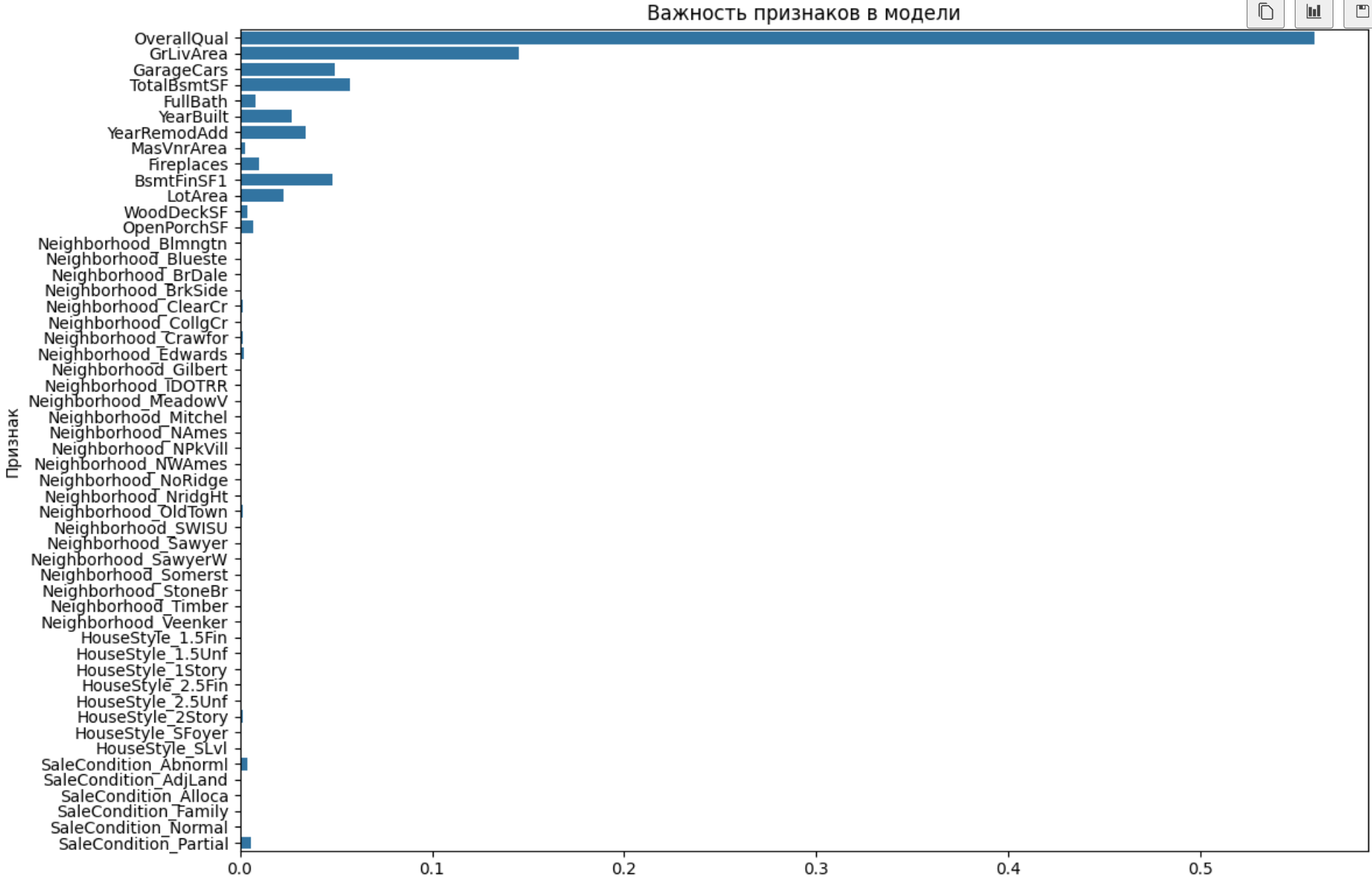
Для улучшения качества прогноза применялся метод GridSearchCV. Настроенные параметры дали ощутимый прирост точности:

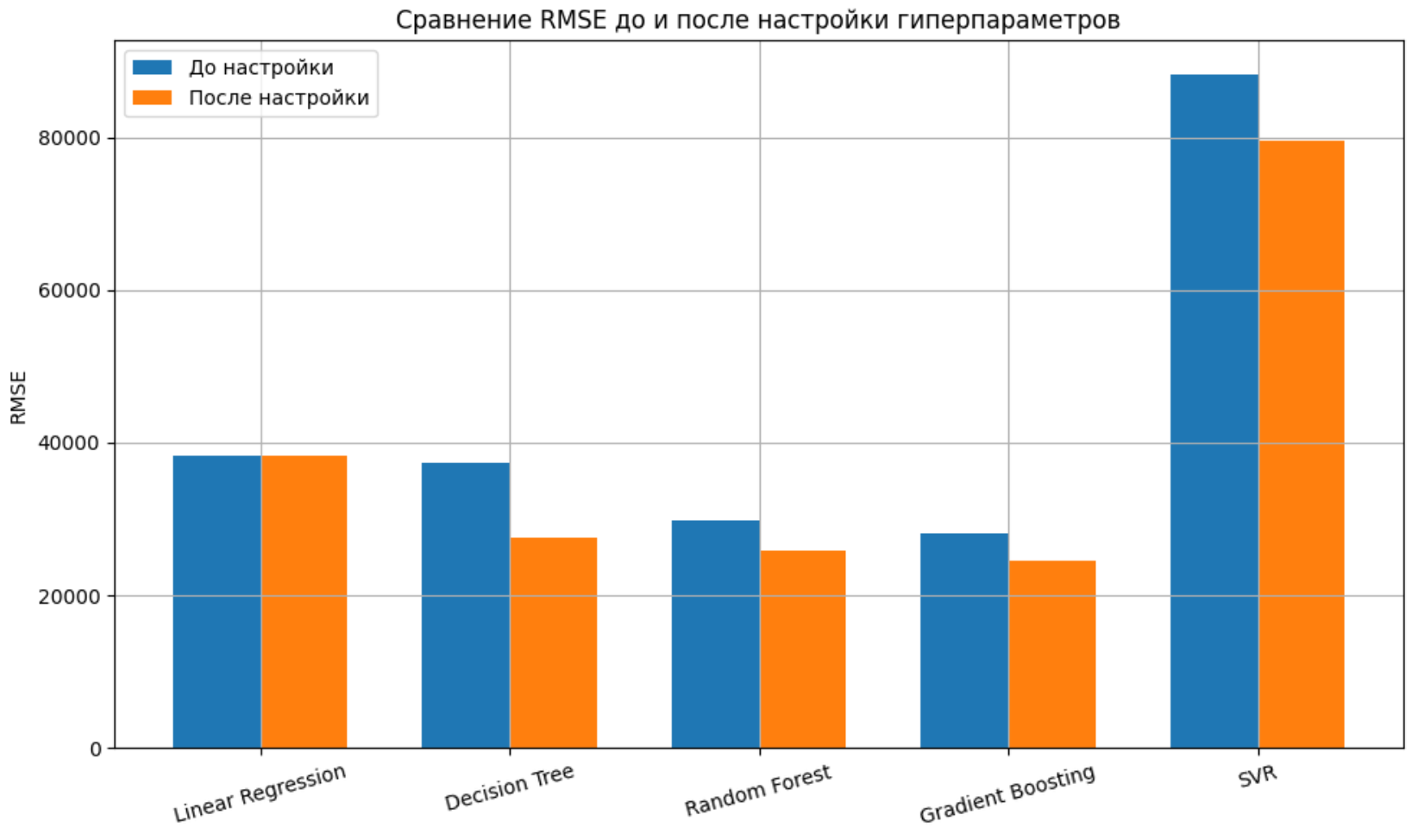
* **Gradient Boosting** достиг RMSE ≈ 28040 и R² ≈ 0.86;
* **Random Forest** — RMSE ≈ 29835, R² ≈ 0.84;
* **Decision Tree** улучшилась с RMSE ≈ 37345 до ≈ 27750;
* **SVR** остался худшим, несмотря на рост R² с -0.04 до 0.06.





**Вывод**: Лучшая модель по всем метрикам — Gradient Boosting. Она сочетает высокую точность с устойчивостью к переобучению.



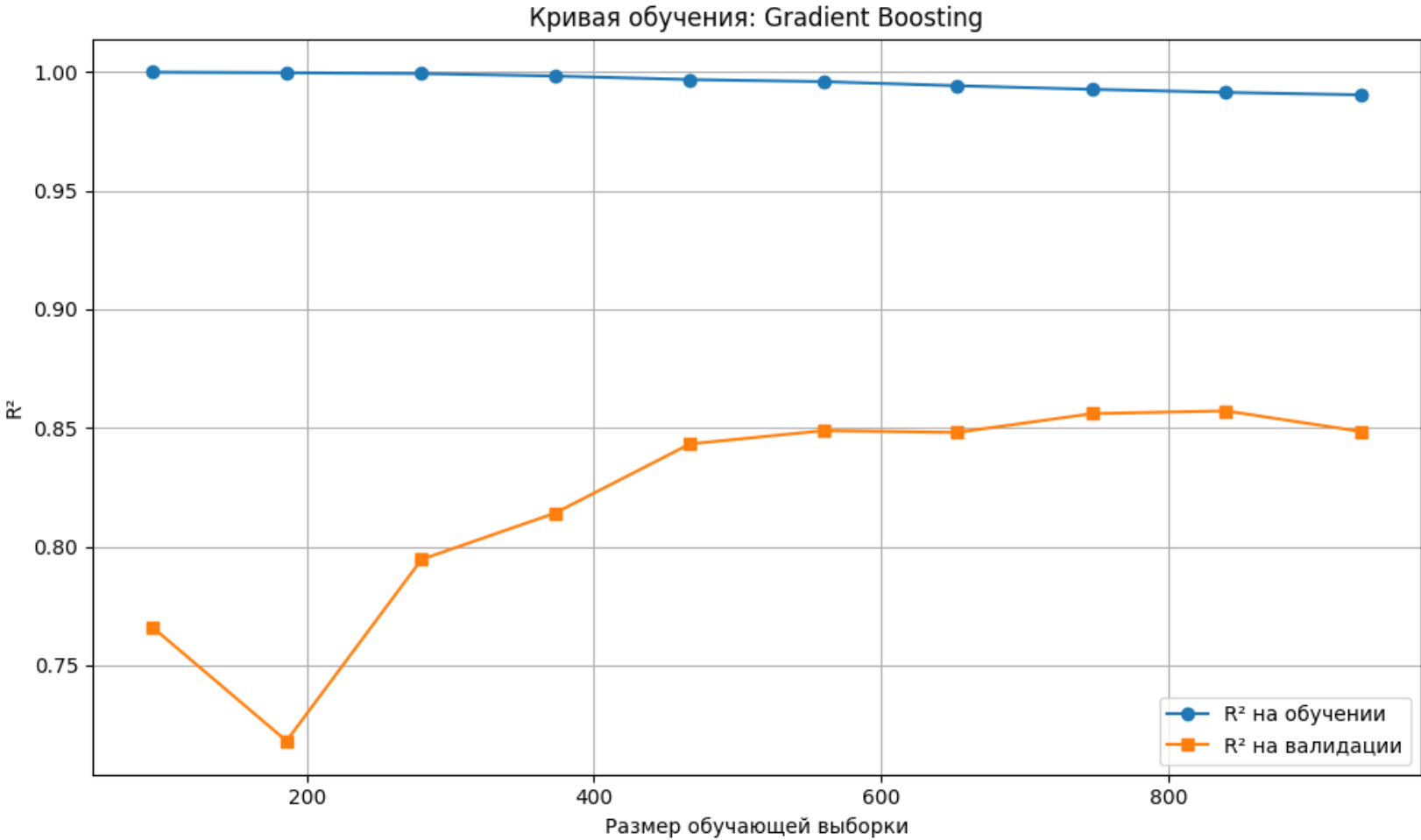


**Рисунок 5 – Сравнение моделей по RMSE-мере (базовая настройка)**

Визуальное сравнение показало:

* Gradient Boosting и Random Forest после настройки значительно улучшили RMSE;
* SVR остался аутсайдером даже после оптимизации;
* Decision Tree показала значительное улучшение, но уступает ансамблевым методам.

**Анализ кривой обучения Gradient Boosting**

****

**Рисунок 6 – Кривая обучения Gradient Boosting**

* Ось X: Объём обучающей выборки;
* Ось Y: R² (отрицательные значения свидетельствуют о слабом обобщении);
* Синий график — производительность на обучающей выборке;
* Оранжевый — на валидационной.

Анализ показал:

На малых объёмах обучающей выборки наблюдается переобучение (R² ≈ 1 на обучении и низкий на валидации);

При увеличении объёма выборки разрыв между криваями сокращается;

Валидационный R² стабилизируется на уровне ≈0.88–0.90, что говорит о хорошем обобщающем поведении модели.

**Заключение**

Результаты исследования показали, что для задачи регрессии по датасету House Prices наиболее эффективны ансамблевые модели, особенно Gradient Boosting:

Высокая точность и устойчивость к переобучению;

Значительное улучшение метрик после настройки гиперпараметров;

Подтверждение зависимости цены от качества, площади, года постройки и района.

Ключевые факторы, повлиявшие на результат:

Грамотная предобработка и выбор признаков;

Использование кросс-валидации;

Подбор гиперпараметров с помощью GridSearchCV.

**Список использованных источников**

1. Kaggle: Heart Failure Prediction Dataset — https://www.kaggle.com/datasets/mirichoi0218/insurance
2. Scikit-learn documentation — https://scikit-learn.org/
3. Streamlit documentation — [https://docs.streamlit.io/](https://scikit-learn.org/)
4. Материалы курса "Машинное обучение", OpenAI, Stepik, Coursera
5. Seaborn, Matplotlib — [https://geeksforgeeks.org/](https://www.python.org/)
6. Визуализация и EDA: <https://seaborn.pydata.org/>, <https://matplotlib.org/>