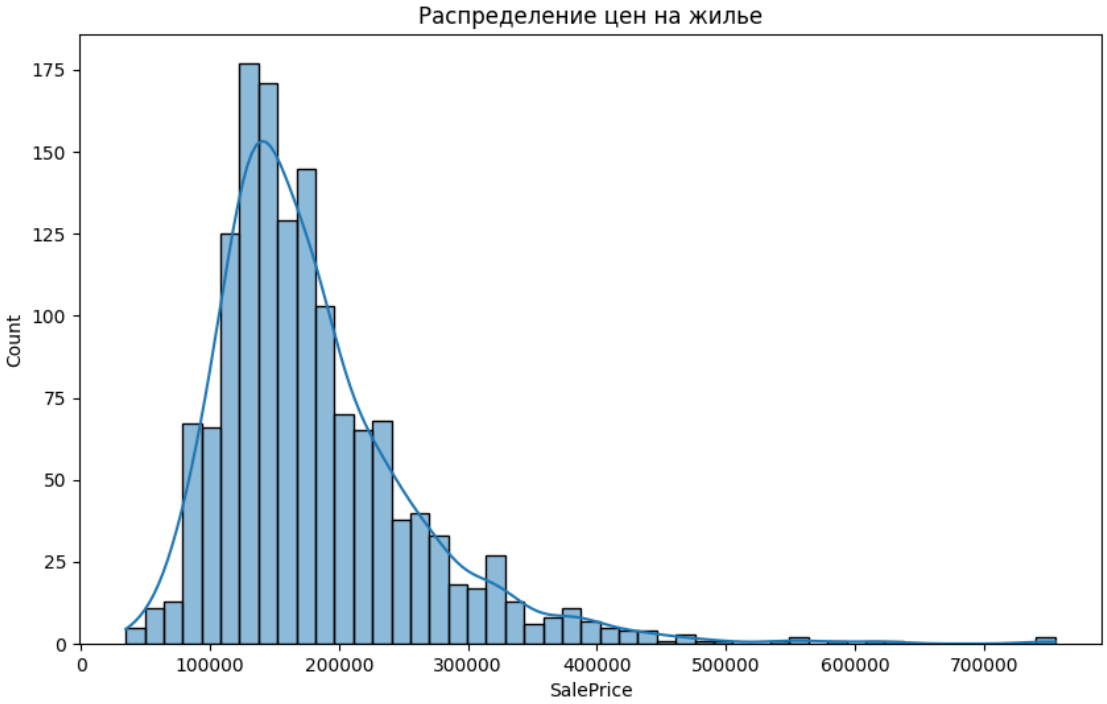
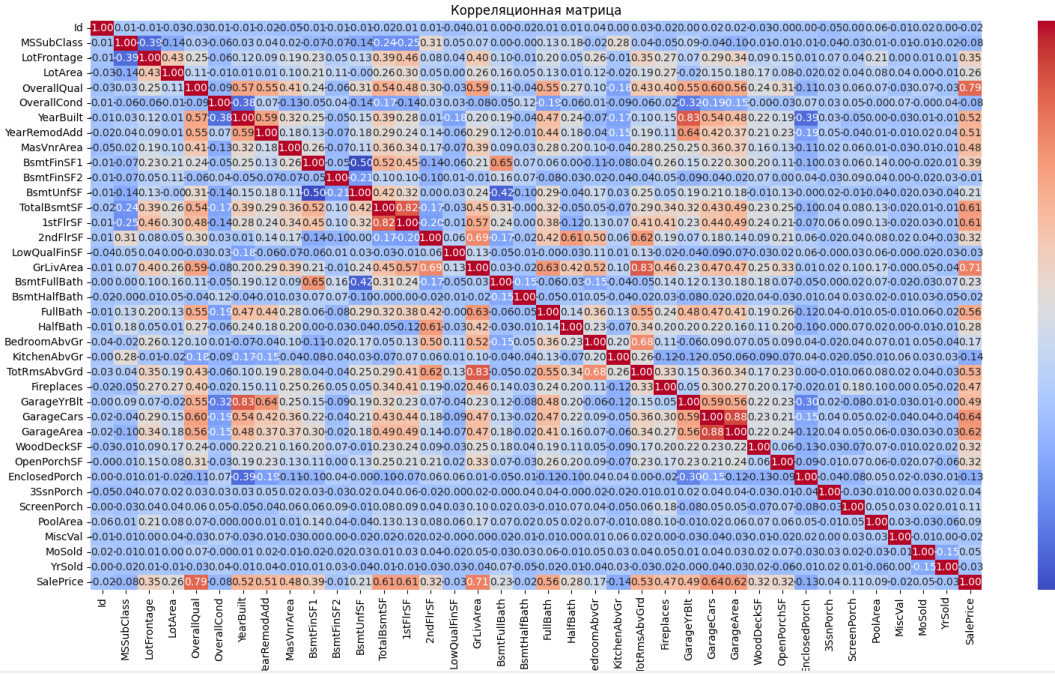
House Prices - Advanced Regression Techniques

Исследовательский анализ данных (EDA)

На этом этапе импортируем необходимые библиотеки, загружаем данные и проводим их исследовательский анализ. Здесь представлен анализ целевой переменной в виде гистограммы распределения:



А также корреляционная матрица числовых признаков:



Предварительная обработка данных

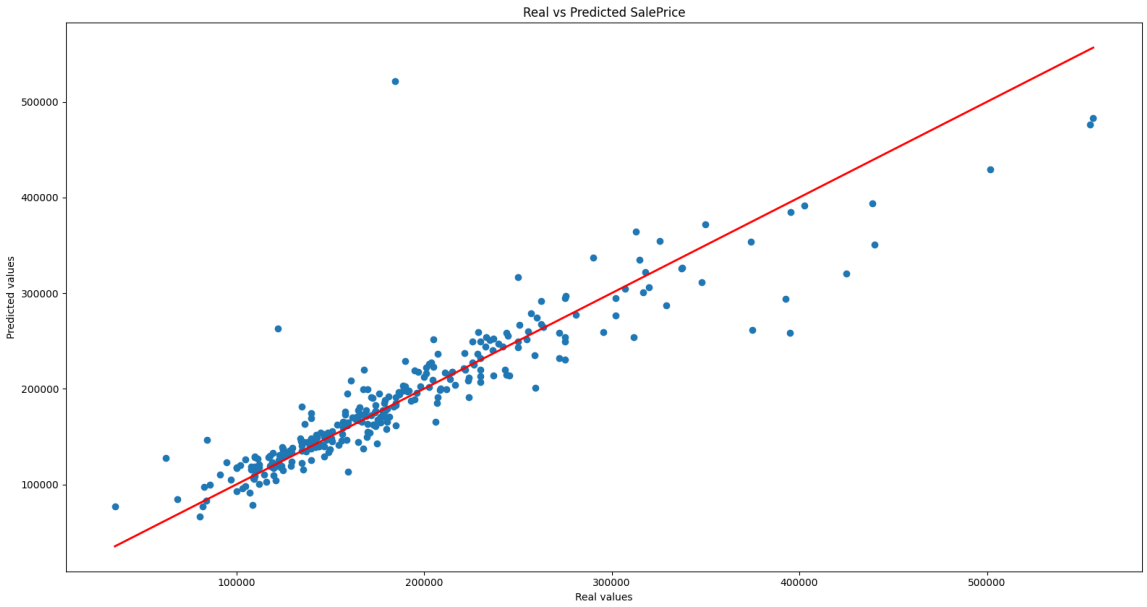
На этом этапе проводим предварительную обработку данных: заполняем пропущенные значения, производим one-hot encoding для категориальных данных, разделяем данные на признаки и целевую переменную, производим масштабирование признаков, разделяем данные на обучающую и тестовую выборки.

Построение и оценка модели

На этом этапе инициализируем модель, обучаем ее, делаем прогноз по тестовым данным, оцениваем результат по метрике R^2:

```
R²: 0.8383249640464783
```

А также строим график для визуальной оценки:



Красная линия соответствует уравнению Predicted = Real, синие точки – предсказания модели. Если точки выше прямой – стоимость была завышена моделью, ниже – стоимость занижена моделью.

Значение $R^2 > 0.8$ считается достаточно хорошим. Также из графика видно, что модель довольно неплохо справляется с предсказанием цен на жилье с реальной ценой в диапазоне 100000–300000, что неудивительно, т. к. в этом диапазоне находится большинство данных, что было продемонстрированно на графике распределения цен. Для жилья с низкими и высокими ценами было предоставлено немного данных, что вызвало слабую предсказательную силу для такого жилья. Забавно, что цены для всех дорогих объектов были занижены моделью, что, при наличии доверия к модели, можно интерпретировать как неоправданно завышенную стоимость этого жилья)

Теоретическая информация о XGBoost

XGBoost (Extreme Gradient Boosting) — это эффективная библиотека для реализации градиентного бустинга, которая используется для решения задач классификации, регрессии и других. Она является одним из самых популярных методов машинного обучения благодаря своей скорости и высокой точности.

XGBoost изначально начинался как исследовательский проект Чэн Тяньци как часть группы Distributed (Deep) Machine Learning Community (DMLC). Изначально она начиналась как консольная программа, которую можно было настроить с помощью конфигурационного файла libsvm. XGBoost стал широко известен в кругах участников соревнований по машинному обучению после его использования в решении победителя конкурса Higgs Machine Learning Challenge. Вскоре после этого были созданы пакеты для Python и R, и теперь XGBoost имеет реализации пакетов для Java, Scala, Julia, Perl и других языков. Это позволило привлечь к библиотеке больше разработчиков и способствовало ее популярности среди сообщества Kaggle, где она использовалась для проведения большого количества соревнований.

Основные особенности XGBoost, отличающие его от других алгоритмов градиентного бустинга, включают:

- Умная штрафовка деревьев
- Пропорциональное уменьшение узлов листьев
- Метод Ньютона в оптимизации
- Дополнительный параметр рандомизации
- Реализация на одиночных, распределенных системах и out-of-core вычислениях
- Автоматический отбор признаков

XGBoost использует Метод Ньютона-Рафсона в пространстве функций, в отличие от градиентного бустинга, который работает как градиентный спуск в пространстве функций, в функции потерь используется ряд Тейлора второго порядка для связи с методом Ньютона-Рафсона.