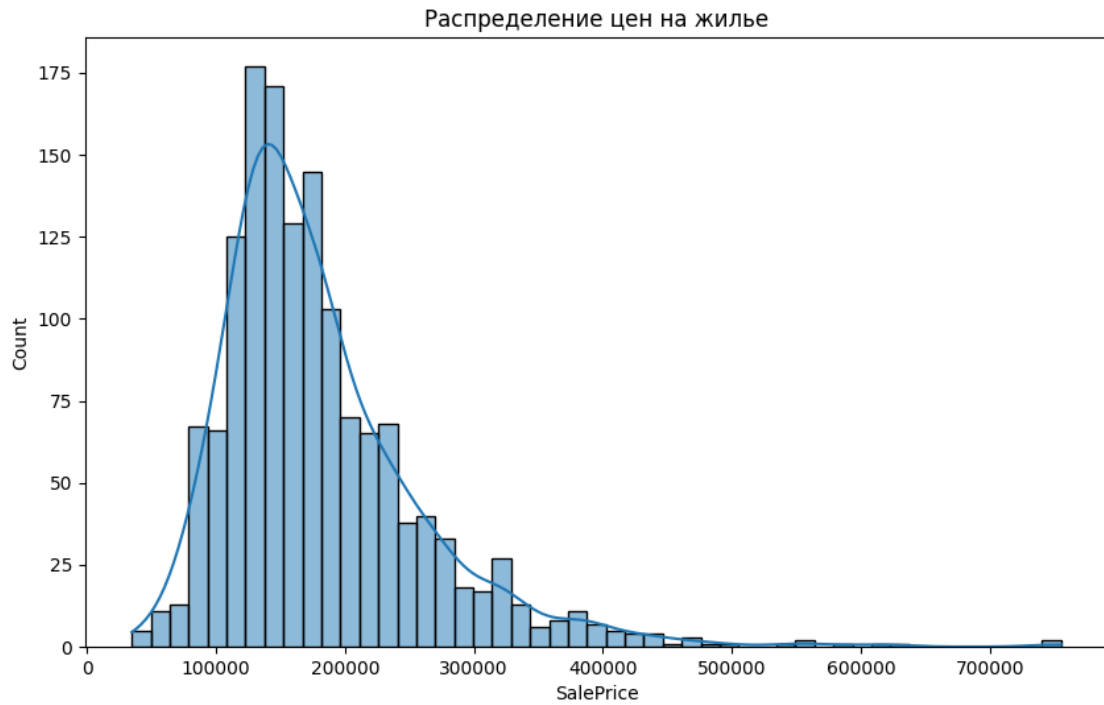


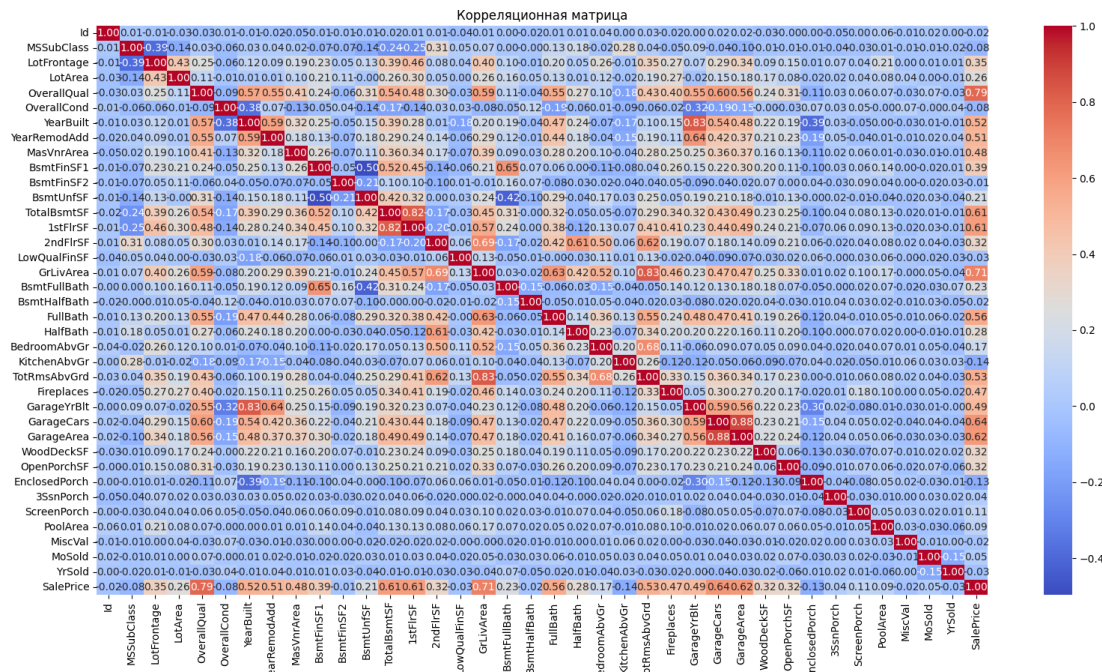
House Prices - Advanced Regression Techniques

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

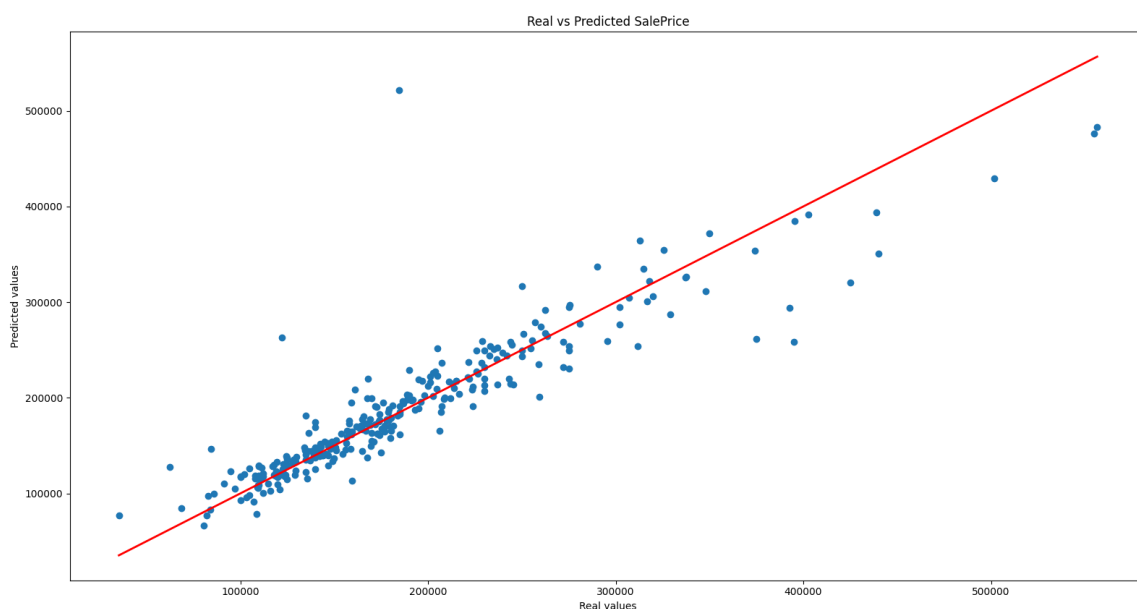
На этом этапе проводим предварительную обработку данных: заполняем пропущенные значения, производим one-hot encoding для категориальных данных, разделяем данные на признаки и целевую переменную, производим масштабирование признаков, разделяем данные на обучающую и тестовую выборки.

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 продемонстрировано 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)