

Липецкий государственный технический университет

Институт компьютерных наук
Кафедра прикладной математики и системного анализа

Отчет по лабораторной работе № 1
по дисциплине «Интеллектуальные методы анализа данных»
на тему «Переработка данных»

Студент

Группа ПМ-23-1

Руководитель

к.т.н., доцент

учёная степень, учёное звание

подпись, дата

Габбасов.Д.В

фамилия, инициалы

Сысоев А.С.

фамилия, инициалы

Липецк 2025 г.

Постановка задачи

Цель работы

Исследование влияния различных методов обработки пропущенных значений на качество линейной регрессионной модели и сравнение эффективности методов предобработки данных.

Задачи

1. Реализовать функцию фильтрации данных методом Савицкого-Голея
2. Реализовать функцию нормализации данных методом IQR
3. Создать функцию для генерации пропущенных значений в датасете
4. Реализовать функцию восстановления пропусков двумя методами:
 - Заполнение средним значением по столбцу
 - Заполнение средним значением по скользящему окну
5. Провести сравнительный анализ качества моделей:
 - Модель, обученная на исходных данных без пропусков
 - Модель, обученная на данных с восстановлением средним
 - Модель, обученная на данных с восстановлением скользящим средним
6. Визуализировать результаты и сделать выводы

Оглавление

1. Введение	4
2. Описание данных и методов	5
2.1. Датасет California Houses	5
2.2. Методы предобработки	5
2.2.1. Создание и обработка пропусков	5
2.2.2. Методы восстановления пропусков	5
3. Экспериментальная часть	6
3.1. Визуализация пропусков и восстановленных данных	6
3.2. Сравнение распределений до и после восстановления	6
4. Результаты и анализ	7
4.1. Сравнительный анализ моделей	7
4.2. Важность признаков в модели	8
4.3. Общий вывод	8
Заключение	9
Список использованных источников	10

1. Введение

Надёжность и точность моделей машинного обучения критически зависят от качества исходных данных. Реальные наборы данных, такие как California Houses, часто содержат шумы, выбросы, пропущенные значения и признаки с различными масштабами, что негативно сказывается на работе алгоритмов.

Этап предобработки данных является необходимым и важным шагом в процессе анализа. В данной работе рассматриваются и применяются методы предобработки, включая сглаживание фильтром Савицкого-Голея для подавления шума и робастную нормализацию на основе межквартильного размаха (IQR) для приведения признаков к единому масштабу и уменьшения влияния выбросов.

Особое внимание уделяется проблеме пропущенных значений, которая часто встречается в реальных данных. Исследуется влияние различных методов импутации на качество линейной регрессионной модели.

Цель работы — исследование влияния методов обработки пропущенных значений на качество линейной регрессионной модели и сравнение эффективности методов предобработки данных.

Актуальность работы обусловлена необходимостью выбора оптимальных методов предобработки для конкретных типов данных и задач машинного обучения.

2. Описание данных и методов

2.1. Датасет California Houses

Для исследования использован датасет California Houses, содержащий информацию о жилых домах в Калифорнии. Датасет включает демографические, географические и структурные характеристики.

Целевая переменная:

- Median_House_Value — медианная стоимость дома

Факторы:

- Median_Income — медианный доход населения
- Median_Age — медианный возраст домов
- Tot_Rooms — общее количество комнат
- Tot_Bedrooms — общее количество спален
- Population — население
- Households — количество домохозяйств
- Latitude, Longitude — географические координаты
- Distance_to_coast, Distance_to_LA, Distance_to_SanDiego, Distance_to_SanJose, Distance_to_SanFrancisco — расстояния до побережья и крупных городов.

2.2. Методы предобработки

2.2.1. Создание и обработка пропусков

В данных были искусственно созданы пропуски (20% или 4128 значений) для каждого признака, чтобы оценить эффективность методов восстановления.

2.2.2. Методы восстановления пропусков

- Заполнение средним (Mean Imputation): Пропущенные значения в каждом столбце заменялись на среднее арифметическое имеющихся данных по этому столбцу.
- Скользящее среднее (Moving Average Imputation): Пропуски заполнялись с использованием скользящего среднего, что позволяет учесть локальные тенденции в данных.

3. Экспериментальная часть

3.1. Визуализация пропусков и восстановленных данных

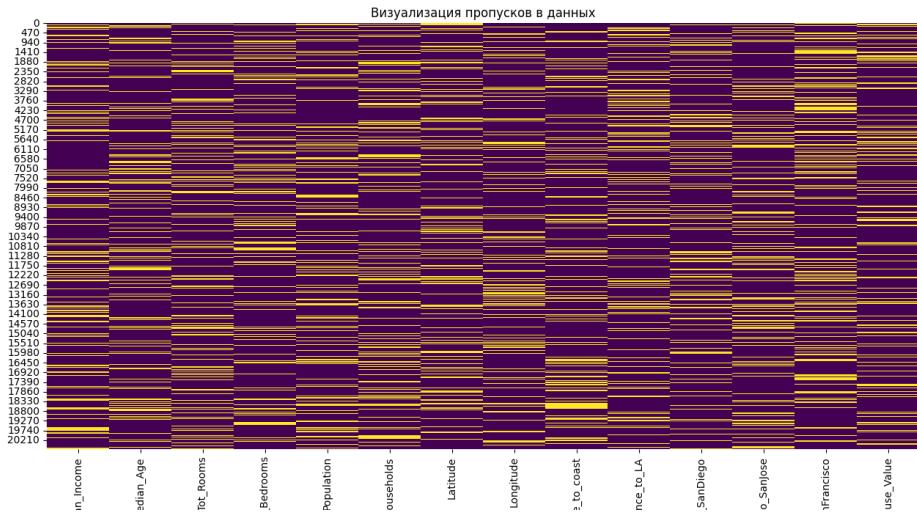


Рисунок 1 – Визуализация пропусков в данных (белые линии)

На Рис. 1 показана структура искусственно созданных пропусков. Пропущенные значения распределены случайным образом по всем признакам, что позволяет объективно сравнить методы восстановления.

3.2. Сравнение распределений до и после восстановления

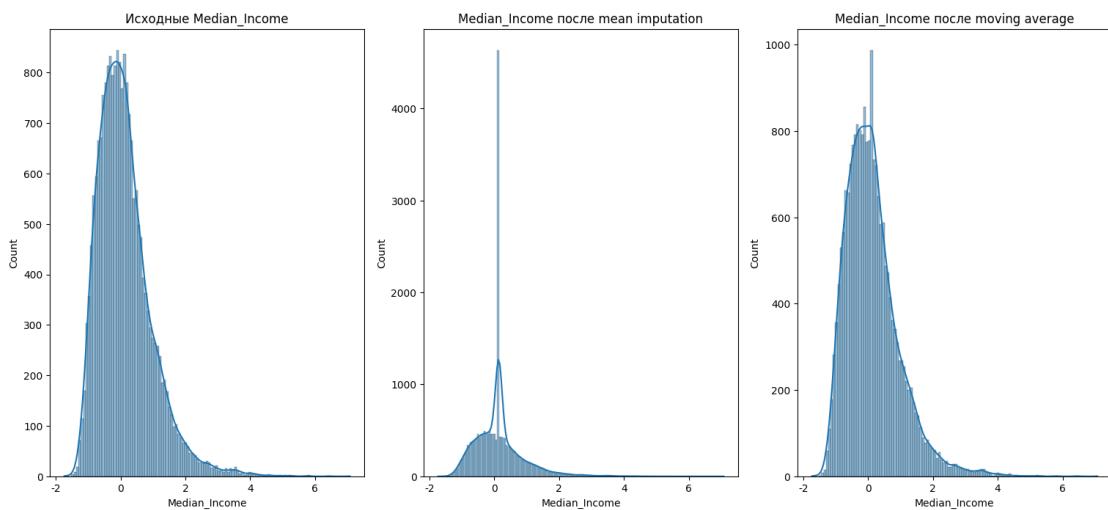


Рисунок 2 – Сравнение распределения признака Median_Income до и после восстановления пропусков

Анализ Рис. 2 позволяет оценить влияние методов восстановления на распределение данных:

- Исходное распределение (сверху) демонстрирует естественную форму.
- Заполнение средним (в центре) привело к резкому всплеску частоты вокруг среднего значения, что исказило исходное распределение.
- Скользящее среднее (снизу) сохранило форму распределения лучше, чем метод среднего, хотя и внесло некоторые искажения.

4. Результаты и анализ

4.1. Сравнительный анализ моделей

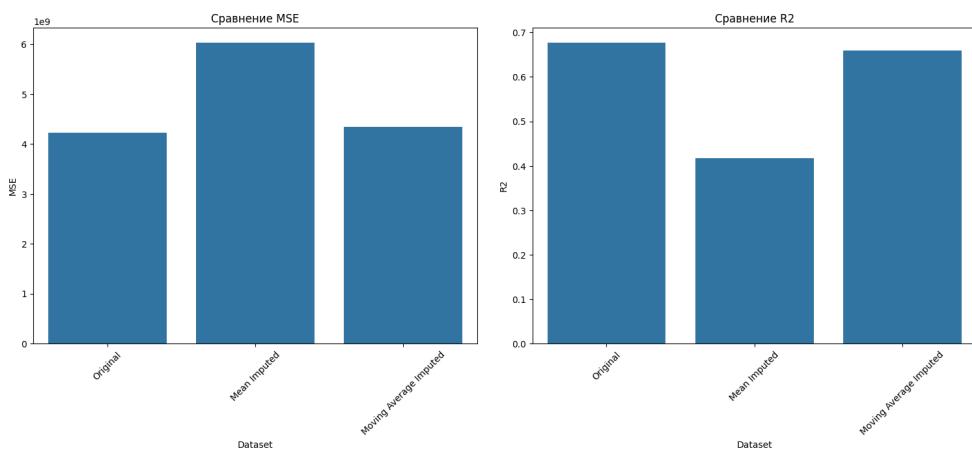


Рисунок 3 – Сравнение моделей по метрикам MSE и R^2

Метод обработки	MSE	R^2
Исходные данные	4,223,569,271	0.678
Скользящее среднее	4,339,766,846	0.660
Заполнение средним	6,035,201,337	0.418

Таблица 1. Сравнение метрик качества регрессионных моделей

По результатам, представленным на Рис. 3 и в Таблице 1, можно сделать следующие выводы:

- Модель, обученная на исходных данных (без пропусков), показывает наилучшее качество ($R^2 = 0.678$).
- Метод скользящего среднего для восстановления пропусков показал результат, близкий к исходному ($R^2 = 0.660$), что свидетельствует о его эффективности.
- Заполнение средним значительно ухудшило качество модели ($R^2 = 0.418$), что согласуется с сильным искажением распределений признаков, видимым на Рис. 2.

4.2. Важность признаков в модели

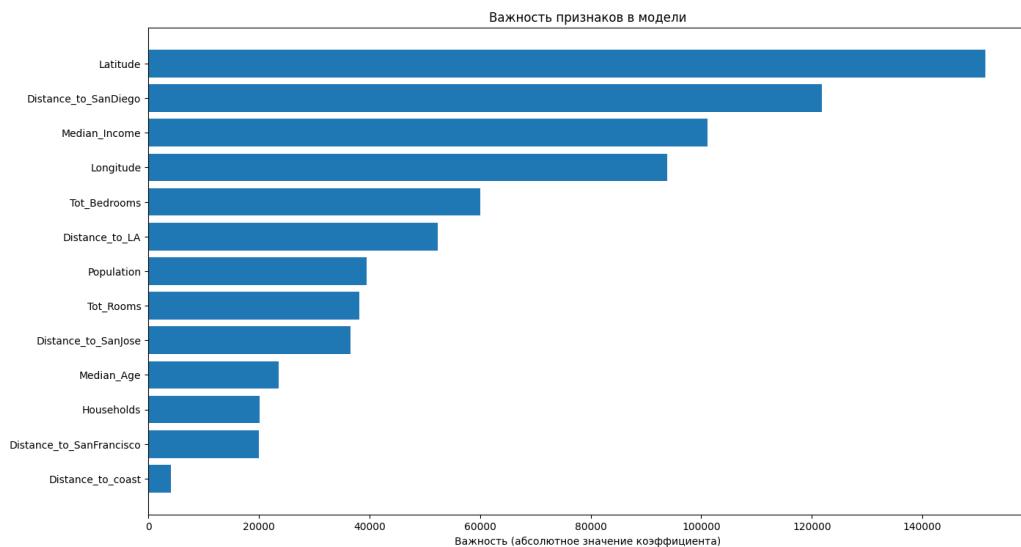


Рисунок 4 – Важность признаков в регрессионной модели (на основе абсолютных значений коэффициентов)

Анализ важности признаков (Рис. 4) для лучшей модели (на исходных данных) показывает:

- Median_Income является самым важным фактором, предсказывающим стоимость дома, что интуитивно понятно и соответствует экономической логике.
- Географические факторы (Latitude, Longitude, Distance_to_SanDiego) также имеют высокую важность, подчеркивая влияние местоположения на стоимость недвижимости.
- Такие признаки, как Distance_to_SanFrancisco и Distance_to_coast, оказались менее значимыми в данной конкретной модели.

4.3. Общий вывод

Эксперимент показал, что метод восстановления пропусков может оказывать существенное влияние на качество прогнозной модели. В данном случае скользящее среднее оказалось более предпочтительным, чем простое заполнение средним, так как оно лучше сохраняет структуру данных и позволяет достичь результатов, близких к модели на исходных данных без пропусков.

Заключение

В ходе выполнения работы был проведен комплексный анализ методов обработки пропущенных данных в контексте построения прогнозных моделей для оценки стоимости недвижимости. Экспериментальное исследование позволило сделать следующие ключевые выводы:

1. Критическое влияние методов импутации на качество моделей. Сравнительный анализ продемонстрировал существенное различие в эффективности методов восстановления пропущенных данных. Модель, обученная на исходных данных без пропусков, показала наивысшее качество ($R^2 = 0.678$), что подтверждает важность сохранения первоначальной структуры данных.
2. Преимущество локальных методов восстановления. Метод скользящего среднего ($R^2 = 0.660$) значительно превзошел простое заполнение средним значением ($R^2 = 0.418$), что объясняется его способностью учитывать локальные тенденции и временные зависимости в данных.
3. Визуальное подтверждение искажений распределений. Анализ гистограмм распределения признака Median_Income наглядно продемонстрировал, что заполнение средним значением создает искусственный пик в области среднего, существенно искажая исходное распределение, в то время как скользящее среднее лучше сохраняет форму распределения.
4. Экономическая интерпретируемость результатов. Анализ важности признаков подтвердил ожидаемые экономические закономерности: медианный доход населения оказался наиболее значимым фактором стоимости жилья, а географическое расположение — вторым по важности предиктором.
5. Практическая значимость исследования. Разработанный инструментарий и полученные результаты имеют практическую ценность для аналитиков недвижимости и специалистов по обработке данных, демонстрируя необходимость тщательного выбора методов обработки пропусков в зависимости от природы данных и требований к качеству прогнозов.

Список использованных источников

- [1] Savitzky A., Golay M.J.E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures // Analytical Chemistry. 1964. Vol. 36, No. 8. P. 1627–1639.
- [2] Tukey J.W. Exploratory Data Analysis. Reading, MA: Addison-Wesley, 1977. 688 p.
- [3] Little R.J.A., Rubin D.B. Statistical Analysis with Missing Data. 3rd ed. Hoboken, NJ: Wiley, 2019. 458 p.
- [4] Van Rossum G., Drake F.L. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace, 2009. 242 p.
- [5] McKinney W. Data Structures for Statistical Computing in Python // Proceedings of the 9th Python in Science Conference. 2010. P. 56–61.
- [6] Pedregosa F. et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2011. Vol. 12. P. 2825–2830.
- [7] Hunter J.D. Matplotlib: A 2D Graphics Environment // Computing in Science & Engineering. 2007. Vol. 9, No. 3. P. 90–95.