|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **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)** |

ФАКУЛЬТЕТ \_\_Информатика и системы управления\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_Системы обработки информации и управления\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

***Модель машинного обучения для оценки\_\_\_\_\_\_\_\_\_\_\_ энергоэффективности зданий\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

Студент \_\_\_\_ИУ5-34м\_\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_ойкинция коммутатора на \_\_\_\_\_\_\_\_\_\_\_\_\_А.Е. Звонарев\_\_\_\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель курсового проекта **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_Ю.Е.Гапанюк\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2022 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)

« \_\_\_\_\_ » \_\_\_\_\_\_\_\_\_\_\_\_ 20 \_\_\_\_ г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме \_ Модель машинного обучения для оценки энергоэффективности зданий\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент группы \_\_\_\_ ИУ5-34М\_\_\_\_

\_\_\_Звонарев Алексей Евгеньевич\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

\_\_\_учебная\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Источник тематики (кафедра, предприятие, НИР) \_\_кафедра\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

График выполнения НИР: 25% к \_4\_ нед., 50% к \_8\_ нед., 75% к \_12\_ нед., 100% к \_17 нед.

***Техническое задание:*** *построить классифицирующую модель машинного обучения \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка на 19 листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата выдачи задания « \_1\_ » \_\_\_октября\_\_\_ 2022 г.

**Руководитель НИР**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ **\_Ю.Е. Гапанюк \_\_**

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ А.Е. Звонарев\_\_**

(Подпись, дата) (И.О.Фамилия)

Введение

Задачей научно-исследовательской работы работы являетя построение регрессионной модели машинного обучения. Модель должна решать задачу для оценки энергоэффективности зданий.

Задание буду выполнять на датасете "Energy and Water Data Disclosure for Local Law ". Датасет содержит значения различных характеристик энергопотребления зданий. В датасете есть как числовые и категориальные признаки, так и нестандартные признаки. Есть колонки с пропусками, ненормированными значениями.

Оглавление

[Введение 3](#_Toc121769809)

[1.Очистка данных 5](#_Toc121769810)

[1.1 Отсутствующие и аномальные данные 7](#_Toc121769811)

[1.2 Разведочный анализ данных 8](#_Toc121769812)

[1.3 Поиск взаимосвязей 9](#_Toc121769813)

[1.4 Двухпеременные графики 12](#_Toc121769814)

[1.5 Конструирование и выбор признаков 15](#_Toc121769815)

[2. Построение модели машинного обучения. 17](#_Toc121769816)

[2.1 Выбор признаков 17](#_Toc121769817)

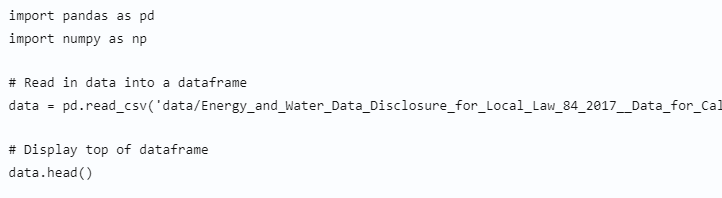
[2.2 Базовый уровень 18](#_Toc121769818)

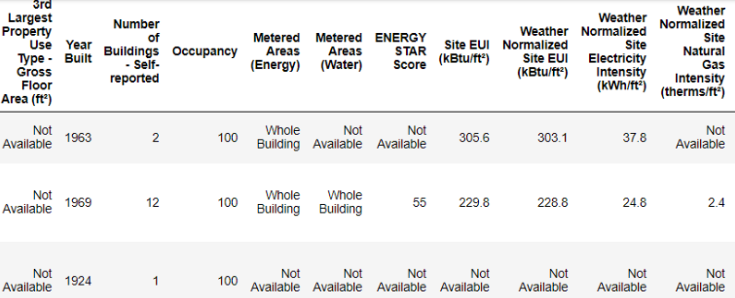
[3. Список литературы 19](#_Toc121769819)

1.Очистка данных

Далеко не каждый набор данных представляет собой идеально подобранное множество наблюдений, без аномалий и пропущенных значений (пример: mtcars и iris). В реальных данных мало порядка, так что прежде, чем приступить к анализу, их нужно очистить и привести к приемлемому формату. Очистка данных — обязательная процедура при решении большинства задач по анализу данных.

Сначала можно загрузить данные в виде кадра данных (dataframe) Pandas и изучить их:





Это фрагмент таблицы из 60 колонок. Даже здесь видно несколько проблем: нам нужно прогнозировать Energy Star Score, но мы не знаем, что означают все эти колонки. Хотя это не обязательно является проблемой, потому что зачастую можно создать точную модель, вообще ничего не зная о переменных. Но нам важна интерпретируемость, поэтому нужно выяснить значение как минимум нескольких колонок.

В решении данной задачи нет смысла изучать все колонки, но точно необходимо разобраться с Energy Star Score, которая описывается так:

Ранжирование по перцентили от 1 до 100, которая рассчитывается на основе самостоятельно заполняемых владельцами зданий отчётов об энергопотреблении за год. Energy Star Score — это относительный показатель, используемый для сравнения энергоэффективности зданий.

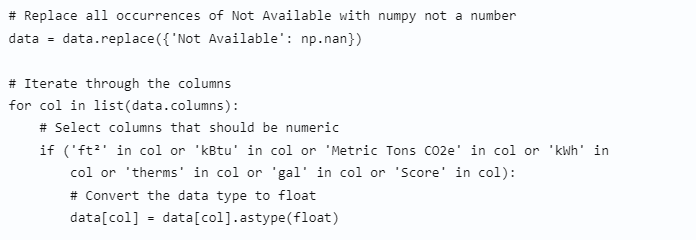


Первая проблема решилась, но осталась вторая — отсутствующие значения, помеченные как «Not Available». Это строковое значение в Python, которое означает, что даже строки с числами будут храниться как типы данных object, потому что, если в колонке есть какая-нибудь строковая, Pandas конвертирует её в колонку, полностью состоящую из строковых. Типы данных колонок можно узнать с помощью метода dataframe.info():



Некоторые колонки, которые явно содержат числа (например, ft²), сохранены как объекты. Невозможно применять числовой анализ к строковым значениям, так что конвертируем их в числовые типы данных (особенно float).

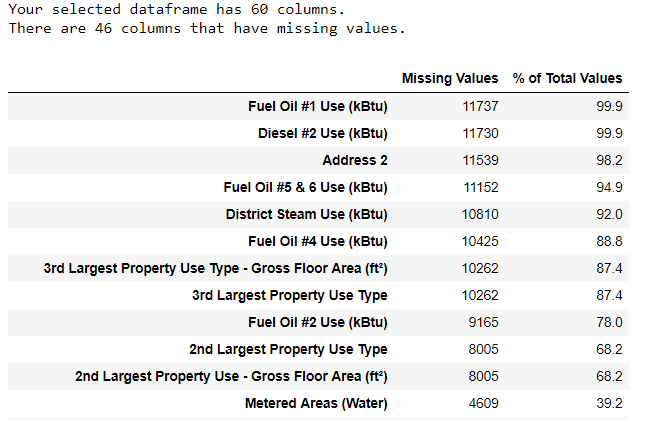
Данный код сначала заменяет все «Not Available» на not a number (np.nan), которые можно интерпретировать как числа, а затем конвертирует содержимое определённых колонок в тип float:



Когда значения в соответствующих колонках у нас станут числами, можно начинать исследовать данные.

1.1 Отсутствующие и аномальные данные

Наряду с некорректными типами данных одна из самых частых проблем — отсутствующие значения. Они могут отсутствовать по разным причинам, и перед обучением модели эти значения нужно либо заполнить, либо удалить. Сначала необходимо выяснить, сколько у нас не хватает значений в каждой колонке.



Убирать информацию всегда нужно с осторожностью, и если много значений в колонке отсутствует, то она, вероятно, не пойдёт на пользу нашей модели. Порог, после которого колонки лучше выкидывать, зависит от задачи, а в данном проекте будем удалять колонки, пустые более чем на половину.

Также на этом этапе лучше удалить аномальные значения. Они могут возникать из-за опечаток при вводе данных или из-за ошибок в единицах измерений, либо это могут быть корректные, но экстремальные значения. В данном случае удалим «лишние» значения, руководствуясь определением экстремальных аномалий:

* Ниже первого квартиля − 3 ∗ интерквартильный размах.
* Выше третьего квартиля + 3 ∗ интерквартильный размах.

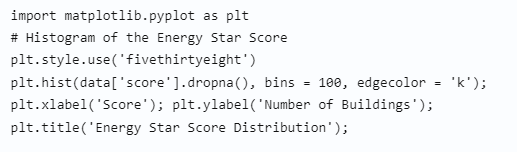
По завершении процесса очистки данных и удаления аномалий у нас осталось больше 11 000 зданий и 49 признаков.

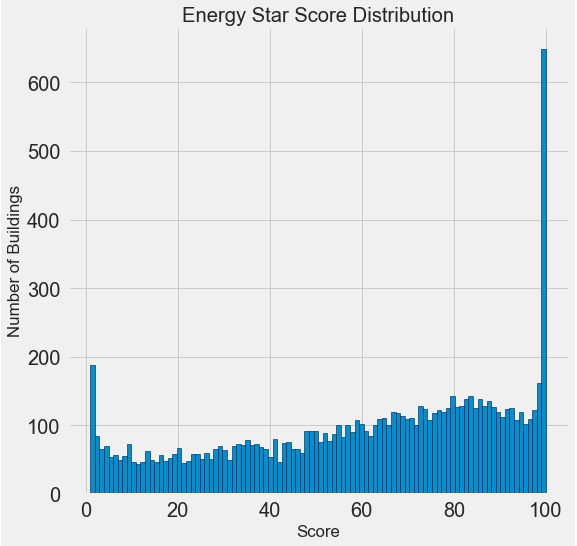
1.2 Разведочный анализ данных

Необходимый этап очистки данных закончен, можно перейти к исследованию! Разведочный анализ данных (РАД) — неограниченный по времени процесс, в ходе которого мы вычисляем статистику и ищем в данных тенденции, аномалии, шаблоны или взаимосвязи.

Коротко говоря, РАД — это попытка выяснить, что нам могут сказать данные. Обычно анализ начинается с поверхностного обзора, затем мы находим интересные фрагменты и анализируем их подробнее. Выводы могут быть интересными сами по себе, или они могут способствовать выбору модели, помогая решить, какие признаки мы будем использовать.

Цель проекта — прогнозировать значение Energy Star Score (в данных переименовано в score), так что имеет смысл начать с исследования распределения этой переменной. Гистограмма — простой, но эффективный способ визуализации распределения одиночной переменной, и её можно легко построить с помощью matplotlib.





Балл Energy Star Score является процентилем, значит следует ожидать единообразного распределения, когда каждый балл присваивается одному и тому же количеству зданий. Однако высший и низший результаты получило непропорционально большое количество зданий (для Energy Star Score чем больше, тем лучше).

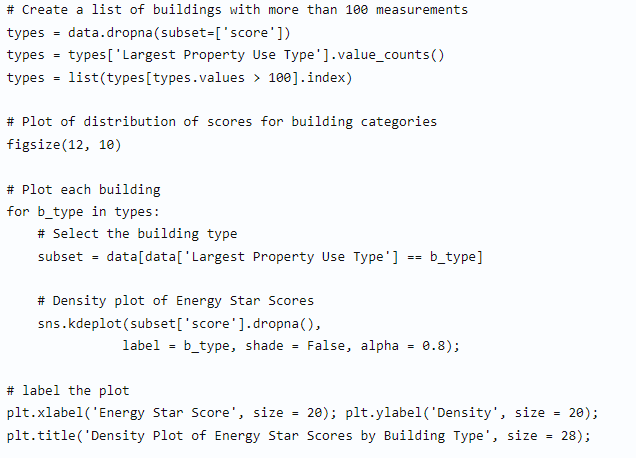
Если снова посмотреть на определение этого балла, то заметно, что он рассчитывается на основе «самостоятельно заполняемых владельцами зданий отчётов», что может объяснить избыток очень больших значений. Просить владельцев зданий сообщать о своём энергопотреблении, это как просить студентов сообщать о своих оценках на экзаменах. Так что это, пожалуй, не самый объективный критерий оценки энергоэффективности недвижимости.

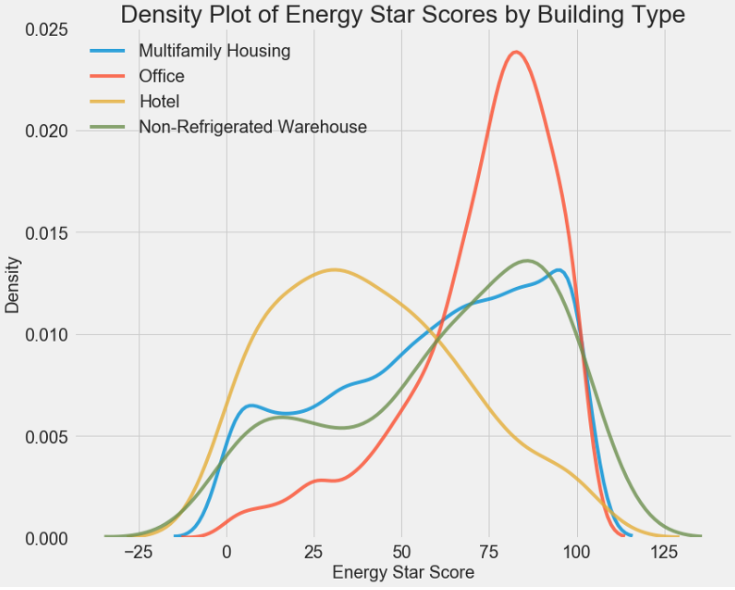
Если бы у нас был неограниченный запас времени, то можно было бы выяснить, почему так много зданий получили очень высокие и очень низкие баллы. Для этого нам пришлось бы выбрать соответствующие здания и внимательно их проанализировать. Но нам нужно только научиться прогнозировать баллы, а не разработать более точный метод оценки. Можно пометить себе, что у баллов подозрительное распределение, но мы сосредоточимся на прогнозировании.

1.3 Поиск взаимосвязей

Главная часть РАД — поиск взаимосвязей между признаками и нашей целью. Коррелирующие с ней переменные полезны для использования в модели, потому что их можно применять для прогнозирования. Один из способов изучения влияния категориальной переменной (которая принимает только ограниченный набор значений) на цель — это построить график плотности с помощью библиотеки Seaborn.

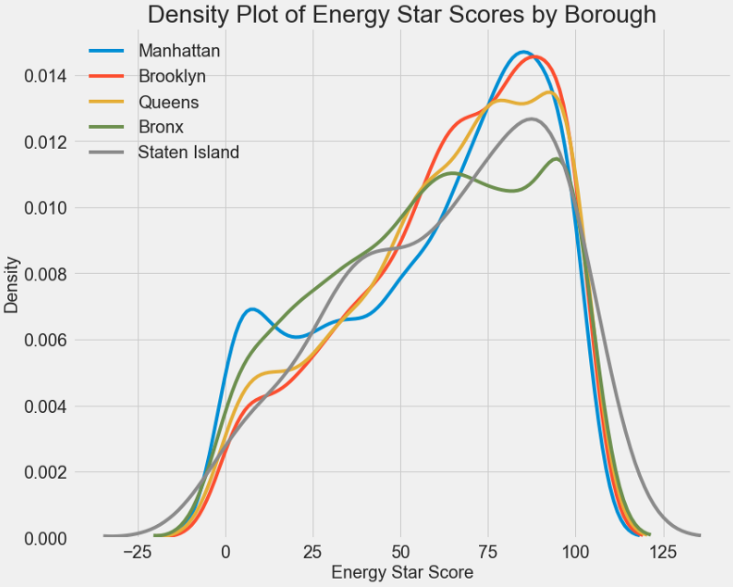
График плотности можно считать сглаженной гистограммой, потому что он показывает распределение одиночной переменной. Можно раскрасить отдельные классы на графике, чтобы посмотреть, как категориальная переменная меняет распределение. Этот код строит график плотности Energy Star Score, раскрашенный в зависимости от типа здания (для списка зданий сболее чем 100 измерениями):





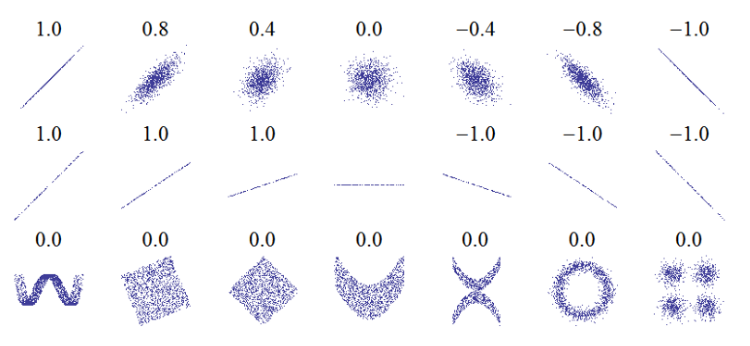
Как можно заметить, тип здания сильно влияет на количество баллов. Офисные здания обычно имеют более высокий балл, а отели более низкий. Значит нужно включить тип здания в модель, потому что этот признак влияет на нашу цель. В качестве категориальной переменной мы должны выполнить one-hot кодирование типа здания.

Аналогичный график можно использовать для оценки Energy Star Score по районам города:



Район не так сильно влияет на балл, как тип здания. Тем не менее включим его в модель, потому что между районами существует небольшая разница.

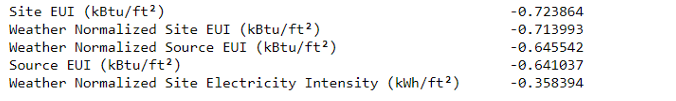
Чтобы посчитать взаимосвязи между переменными, можно использовать коэффициент корреляции Пирсона. Это мера интенсивности и направления линейной зависимости между двумя переменными. Значение +1 означает идеально линейную положительную зависимость, а -1 означает идеально линейную отрицательную зависимость. Вот несколько примеров значений коэффициента корреляции Пирсона:



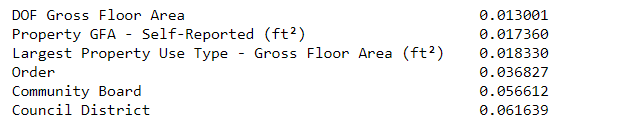
Хотя этот коэффициент не может отражать нелинейные зависимости, с него можно начать оценку взаимосвязей переменных. В Pandas можно легко вычислить корреляции между любыми колонками в кадре данных (dataframe):



Самые отрицательные корреляции с целью:



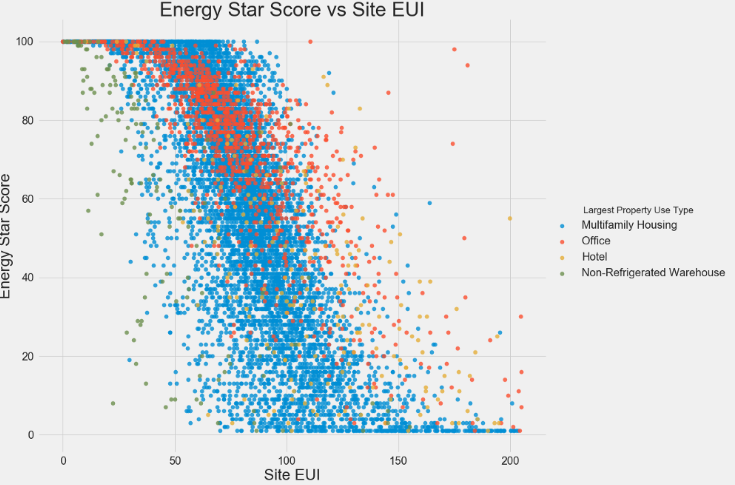
и самые положительные:



Есть несколько сильных отрицательных корреляций между признаками и целью, причём наибольшие из них относятся к разным категориям EUI (способы расчёта этих показателей слегка различаются). EUI (Energy Use Intensity, интенсивность использования энергии) — это количество энергии, потреблённой зданием, делённое на квадратный фут площади. Эта удельная величина используется для оценки энергоэффективности, и чем она меньше, тем лучше. Логика подсказывает, что эти корреляции оправданны: если EUI увеличивается, то Energy Star Score должен снижаться.

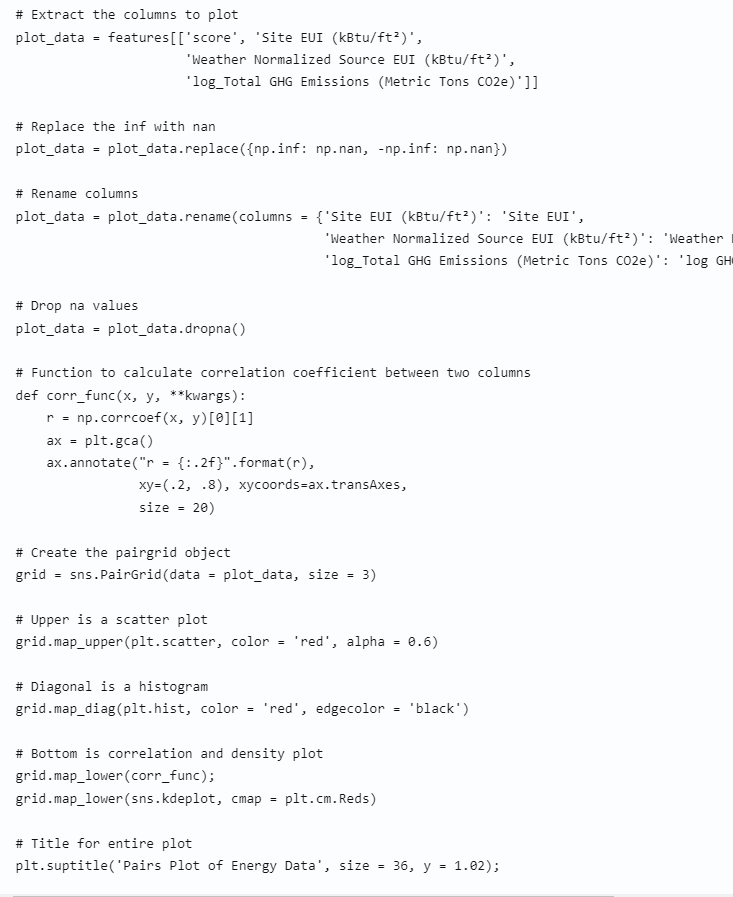
1.4 Двухпеременные графики

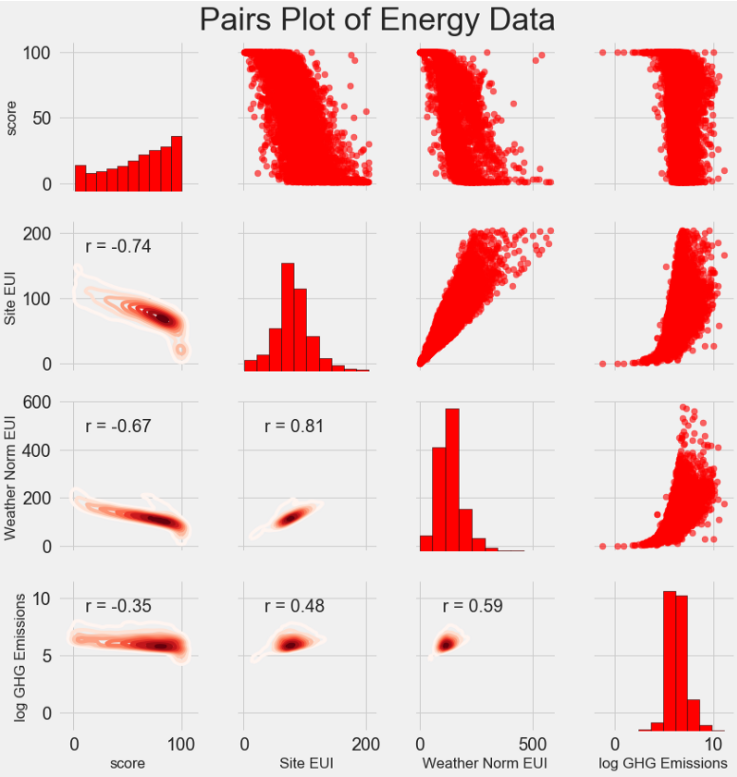
Воспользуемся диаграммами рассеивания для визуализации взаимосвязей между двумя непрерывными переменными. К цветам точек можно добавить дополнительную информацию, например, категориальную переменную. Ниже показана взаимосвязь Energy Star Score и EUI, цветом обозначены разные типы зданий:



Этот график позволяет визуализировать коэффициент корреляции -0,7. По мере уменьшения EUI увеличивается Energy Star Score, эта взаимосвязь наблюдается у зданий разных типов.

Последний исследовательский график называется Pairs Plot (парный график). Это прекрасный инструмент, позволяющий увидеть взаимосвязи между различными парами переменных и распределение одиночных переменных. Воспользуемся библиотекой Seaborn и функцией PairGrid для создания парного графика с диаграммой рассеивания в верхнем треугольнике, с гистограммой по диагонали, двухмерной диаграммой плотности ядра и коэффициентов корреляции в нижнем треугольнике.





Чтобы увидеть взаимосвязи переменных, поищем пересечения рядов и колонок. Допустим, нужно посмотреть корреляцию Weather Norm EUI и score, тогда мы ищем ряд Weather Norm EUI и колонку score, на пересечении которых стоит коэффициент корреляции -0,67.

1.5 Конструирование и выбор признаков

Конструирование и выбор признаков зачастую приносит наибольшую отдачу с точки зрения времени, потраченного на машинное обучение. Сначала дадим определения:

* Конструирование признаков: процесс извлечения или создания новых признаков из сырых данных. Чтобы использовать переменные в модели, возможно, их придётся преобразовывать, скажем, брать натуральный логарифм, или извлекать квадратный корень, или применять one-hot кодирование категориальных переменных. Конструирование признаков можно рассматривать как создание дополнительных признаков из сырых данных.
* Выбор признаков: процесс выбора из данных самых релевантных признаков, в ходе которого мы удаляем часть признаков, чтобы помочь модели лучше обобщать новые данные ради получения более интерпретируемой модели. Выбор признаков можно рассматривать как удаление «лишнего», чтобы осталось только самое важное.

Модель машинного обучения может учиться только на предоставленных нами данных, поэтому крайне важно удостовериться, что мы включили всю релевантную для нашей задачи информацию.

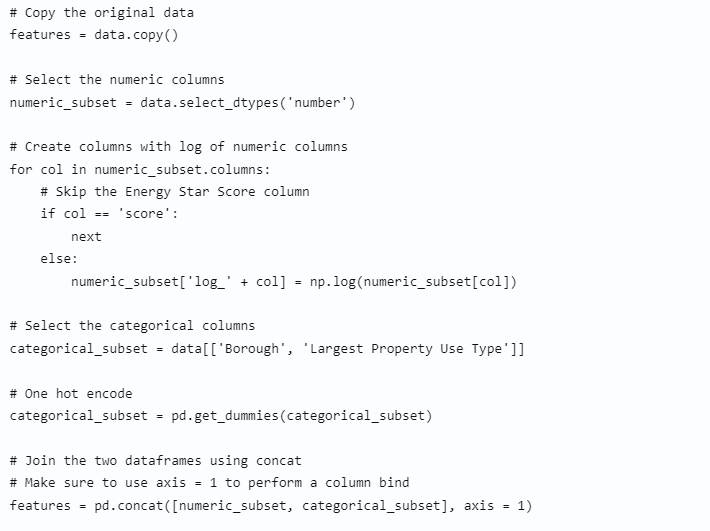
Сделаем следующее:

* Применим к категориальным переменным (квартал и тип собственности) one-hot кодирование.
* Добавим взятие натурального логарифма от всех числовых переменных.

One-hot кодирование необходимо для того, чтобы включить в модель категориальные переменные. Алгоритм машинного обучения не сможет понять тип «офис», так что если здание офисное, мы присвоим ему признак 1, а если не офисное, то 0.

Добавление преобразованных признаков поможет модели узнать о нелинейных взаимосвязях внутри данных. В анализе данных является нормальной практикой извлекать квадратные корни, брать натуральные логарифмы или ещё как-то преобразовывать признаки, это зависит от конкретной задачи или вашего знания лучших методик. В данном случае добавим натуральный логарифм всех числовых признаков.

Этот код выбирает числовые признаки, вычисляет их логарифмы, выбирает два категориальных признака, применяет к ним one-hot кодирование и объединяет оба множества в одно.



Теперь у нас есть больше 11 000 наблюдений (зданий) со 110 колонками (признаками). Не все признаки будут полезны для прогнозирования Energy Star Score, поэтому займёмся выбором признаков и удалим часть переменных.

2. Построение модели машинного обучения.

2.1 Выбор признаков

Многие из имеющихся 110 признаков избыточны, потому что сильно коррелируют друг с другом. К примеру, вот график EUI и Weather Normalized Site EUI, у которых коэффициент корреляции равен 0,997.

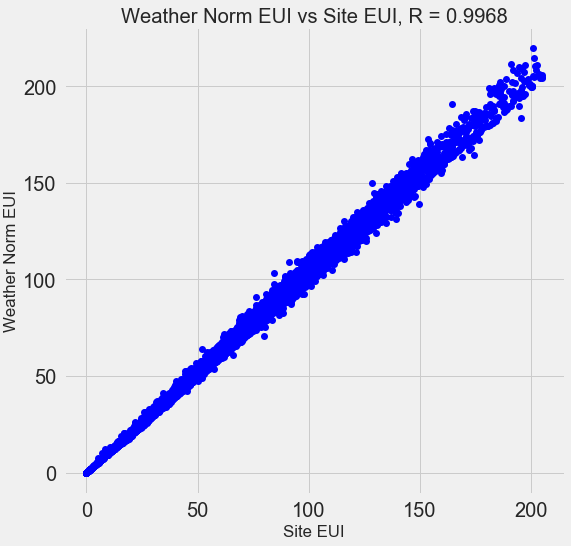


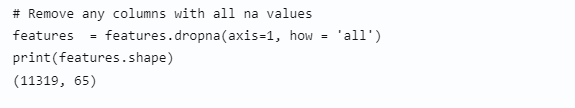
Рисунок 18. Матрица корреляций.

Признаки, которые сильно коррелируют друг с другом, называются коллинеарными. Удаление одной переменной в таких парах признаков часто помогает модели обобщать и быть более интерпретируемой.

Существует ряд методов вычисления коллинеарности признаков, и один из самых популярных — фактор увеличения дисперсии (variance inflation factor). Мы для поиска и удаления коллинеарных признаков воспользуемся коэффициентом В-корреляции (thebcorrelation coefficient). Отбросим одну пару признаков, если коэффициент корреляции между ними больше 0,6. Код приведён в блокноте.

Машинное обучение эмпирично, и часто приходится экспериментировать, чтобы найти лучшее решение. После выбора у нас осталось 64 признака и одна цель.

Видно, что признаки T\_bc и T коррелируют друг сдругом и ни один из них не является целевым. Стоит исключить один из них.



2.2 Базовый уровень

Мы очистили данные, провели разведочный анализ и сконструировали признаки.

И, прежде чем перейти к созданию модели, нужно выбрать исходный базовый уровень (naive baseline) — некое предположение, с которым мы будем сравнивать результаты работы моделей. Если они окажутся ниже базового уровня, мы будем считать, что машинное обучение неприменимо для решения этой задачи, или что нужно попробовать иной подход.

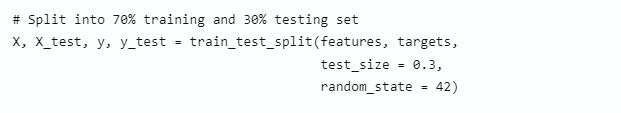
Для регрессионных задач в качестве базового уровня разумно угадывать медианное значение цели на обучающем наборе для всех примеров в тестовом наборе. Эти наборы задают барьер, относительно низкий для любой модели.

В качестве метрики возьмём среднюю абсолютную ошибку (mae) в прогнозах. Для регрессий есть много других метрик, но мне нравится совет выбирать какую-то одну метрику и с её помощью оценивать модели. А среднюю абсолютную ошибку легко вычислить и интерпретировать.

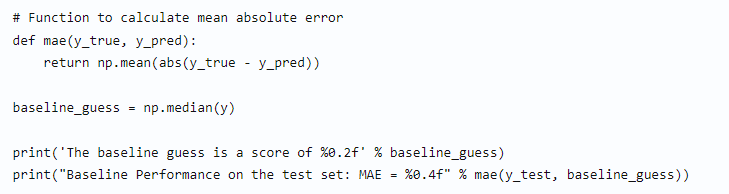
Прежде чем вычислять базовый уровень, нужно разбить данные на обучающий и тестовый наборы:

1. Обучающий набор признаков — то, что мы предоставляем нашей модели вместе с ответами в ходе обучения. Модель должна выучить соответствие признаков цели.
2. Тестовый набор признаков используется для оценки обученной модели. Когда она обрабатывает тестовый набор, то не видит правильных ответов и должна прогнозировать, опираясь только на доступные признаки. Мы знаем ответы для тестовых данных и можем сравнить с ними результаты прогнозирования.

Для обучения используем 70 % данных, а для тестирования — 30 %:



Теперь вычислим показатель для исходного базового уровня:



В результате получаем:

The baseline guess is a score of 66.00

Baseline Performance on the test set: MAE = 24.5164

Средняя абсолютная ошибка на тестовом наборе составила около 25 пунктов. Поскольку мы оцениваем в диапазоне от 1 до 100, то ошибка составляет 25 %.

3. Список литературы

1. К. Элбон. Машинное обучение с использованием Python. Сборник рецептов— Санкт-Петербург, Вильямс, 2019 – 200 с.

2. Ю.Е. Гапанюк. Обработка пропусков в данных, кодирование категориальных признаков, масштабирование данных. — [Электронный ресурс] — Режим доступа. — URL: https://nbviewer.org/github/ugapanyuk/ml\_course\_2020/blob/master/common/notebooks/missing/handling\_missing\_norm.ipynb (Дата обращения: 15.12.2021).

3. А. Мюллер, С. Гвидо. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. — Санкт-Петербург, Вильямс, 2020 – 480 с.