Ким Андрей ЭК-41

1. **Business Understanding**
   1. *Цель проекта*: создать модели машинного обучения для решения таких задач, как: регрессии – определить стоимость недвижимости (***price***); классификации – определить тип недвижимости (***room\_type***); кластеризации – поделить все недвижимости на группы по их качеству, используя оценки сделанные посетителями этих недвижимостей.
   2. *Поточная ситуация*: данные состоят из 3013 наблюдений и 45 характеристик. Все данные и созданные модели будут сберегаться на рабочем ноутбуке, модели планируются быть созданными на языке программирования Python. Возможные риски: срыв даты введения моделей в эксплуатацию.
   3. *Задачи, которые решаются с точки зрения разработчика*: определить наилучшие метрики качества всех моделей и обосновать их (для регрессии - MSE, RMSE, etc., Для классификации и кластеризации – recall, precision, F-мера, etc.)
   4. *План проекта*: написать план проекта.
2. **Data Understandung**
   1. *Сбор данных:* набор данных принадлежит Airbnb.
   2. *Описание данных:*
      1. Таблица – файл называется *train.csv*. Сам файл имеет объем 7.14 МБ;
      2. Ключ – атрибут id;
      3. Количество рядов – 3013 наблюдений;
      4. Количество столбцов – 45 столбцов. Большинство не будут нести какой-либо полезной информацией для моделей, поэтому их придется исключить из выборки;

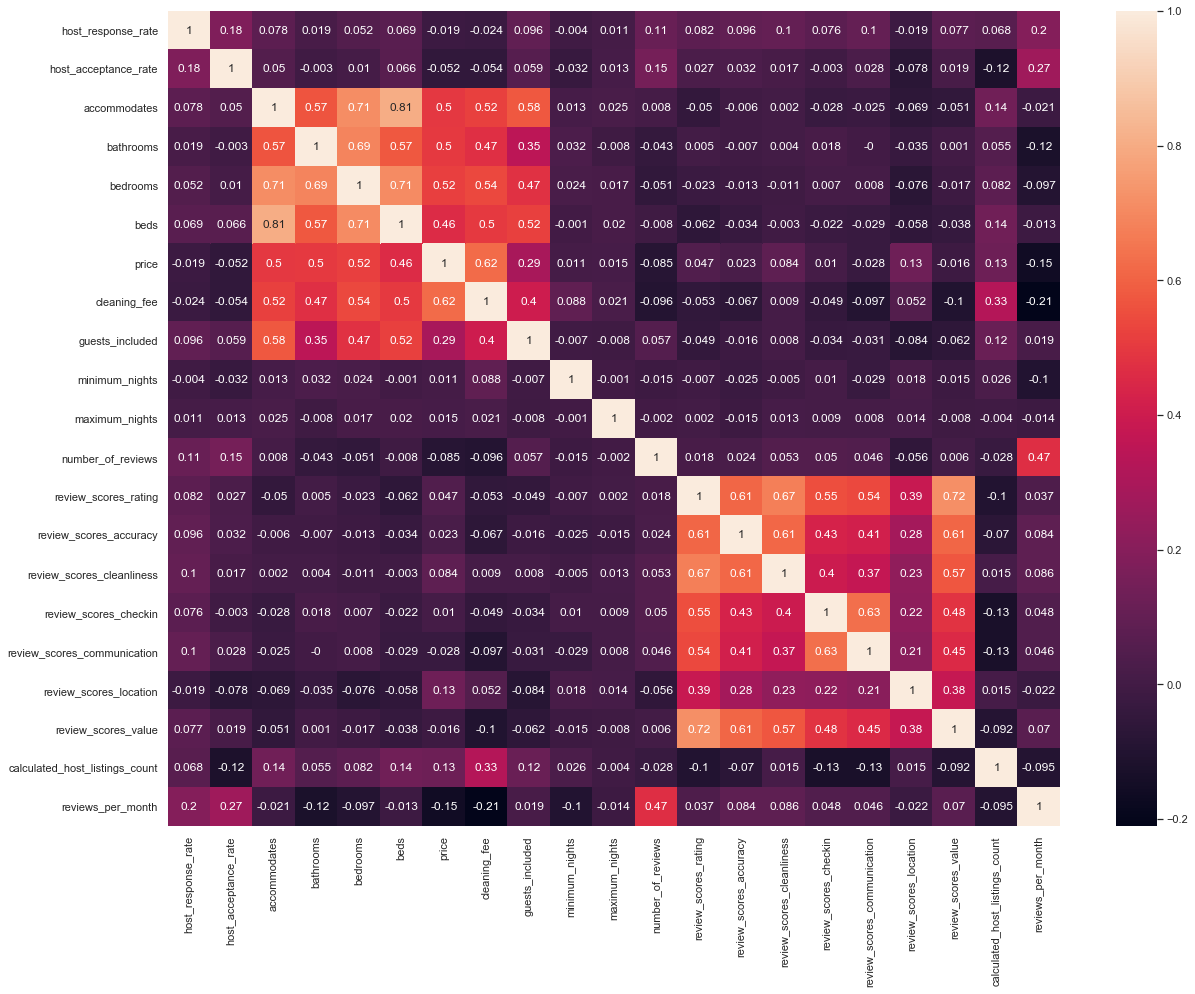
Рис. 1 Описательная статистика некоторых переменных



По этим показателям мы видим, что есть пропущенные значения и подозрение на выбросы

* 1. *Исследование данных:*

Рис. 2 Корреляционная матрица

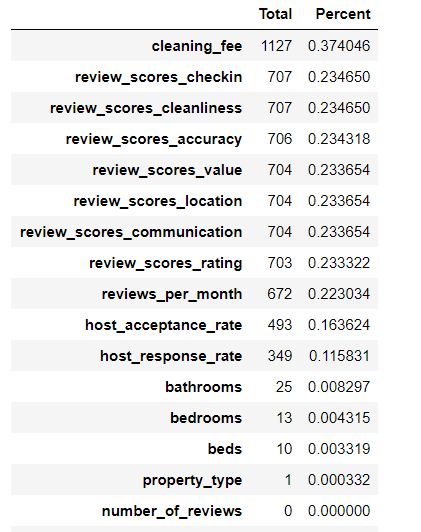


Как мы видим, далеко не все факторы сильно колеруют с ценой.

*2.4Качество данных.*

В наших данных помимо выбросов существуют пропущенные значения. На Рис. 3 можно увидеть количество таких пропущенных значений и их долю от всех наблюдений по каждому фактору.

Рис. 3 Пропущенные значения



Также, в столбцах *price* и *cleaning\_fee* есть символ *$*, а в *host\_response\_rate* и *host\_acceptance\_rate - %*. Поэтому для работы с этими переменными требуется удалить эти символы.

1. **Data Preparation**

*3.1 Отбор данных.*

Некоторые факторы не несут собой какую-либо значимую информацию для решений наших бизнес-задач, такие факторы следует убрать из набора данных. Среди них оказались столбцы: *name, summary, description,neighborhood\_overview, transit, host\_since, host\_location, host\_about, neighbourhood\_cleansed, city, state, market, smart\_location, country\_code, country, zipcode, amenities, host\_verifications, host\_neighbourhood*.

*3.2 Очищение данных.*

Как мы упоминали в пунктах *2.4 Качество данных* в наших данных существуют как выбросы, так и пропущенные значения. С выбросами было принято решения подставить вместо них крайние значения, пользуясь теоремой о 3-х сигмах. Для использование этой теоремой все числовые переменные прологарифмированы, т.к. в таком случае распределение схоже с нормальным, а после обработки выбросов им были возвращены к изначальным величинам.

С пропущенными значениями использовалось KNN-Imputation. Этот алгоритм позволял сохранить корреляцию между переменными.

*3.3 Генерация данных.*

Шкалирование (приведение переменных в нормализованным величинам) происходила перед решением задачи классификации.

*3.4 Интеграция данных.*

Все начальные данные находятся в одном файле.

*3.5 Формирование данных.*

Основная часть формирование данных была произведена на пункте *3.2 Очищение данных*, остальные манипуляции с данными буду проводиться перед, непосредственно, созданием моделей.

**Регрессия**

1. **Моделирование.**

*4.1*Выбор алгоритмов.

Перед созданием моделей, следует прологарифмировать целевую переменную *price* для того, чтобы её распределение было схожим к нормальному.

Были использованы такие методы для решения задачи регрессии, как:

1. Простая линейная регрессия;
2. Множественная регрессия;
3. Полиномиальная регрессия;
4. Дерево решений;
5. Случайный лес;
6. Бустинг.

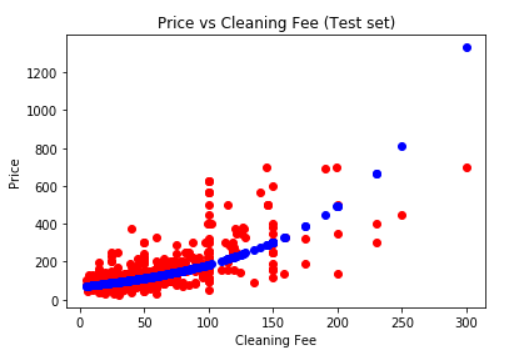
*4.2 Планирование тестирование.*

Изначальная выборка была поделена на обучающую и тестовую выборку с пропорцией 80/20.

*4.3 Обучение модели.*

1. Простая линейная регрессия. Исходя из Рис. 2 корреляционная матрица с целевой переменной наибольше всего коррелирует переменная *cleaning\_fee*, поэтому используем этот фактор для первой модели

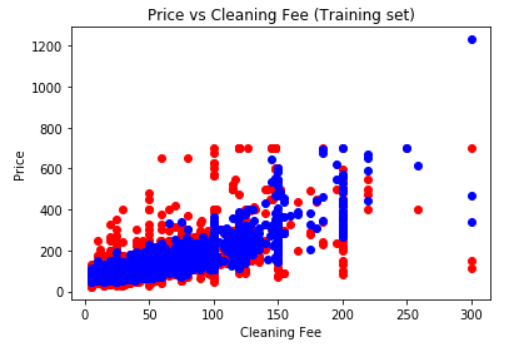
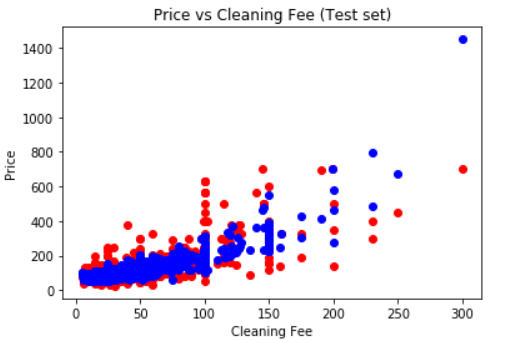
Рис. 4 Простая линейная регрессия.

Как видно по графикам, модель обнаружила некую тенденцию.

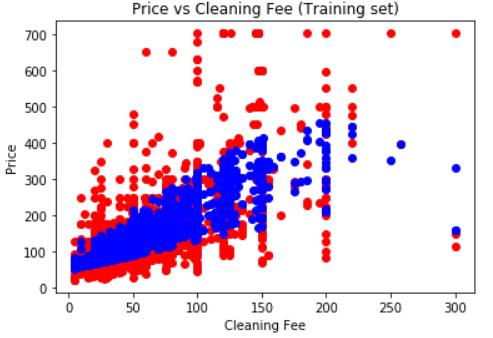
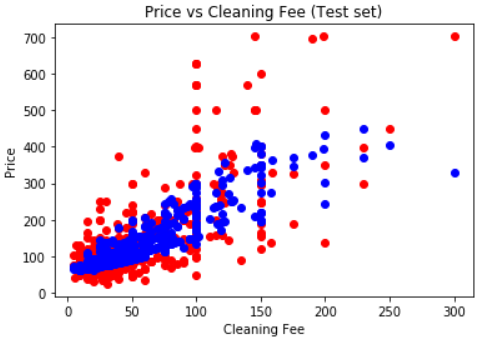
1. Множественная регрессия. Для этой модели были использованы все переменные. Это заметно отобразилось на качестве модели.

Рис. 5 Множественная регрессия.

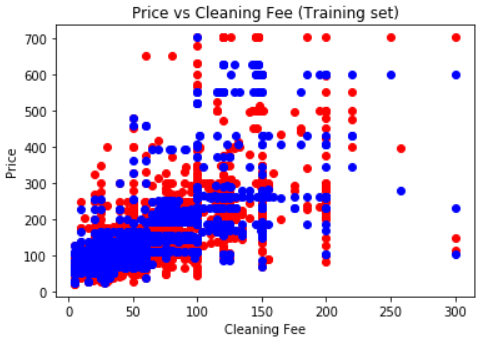
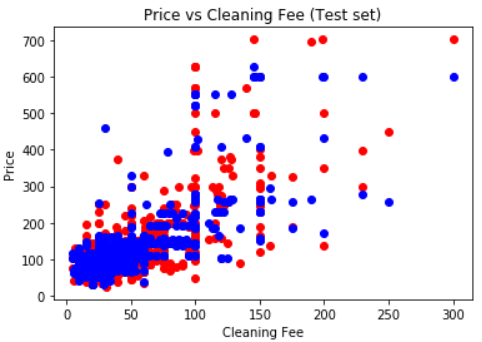
1. Полиномиальная регрессия. Для данной модели мы взяли факторы *accommodates, bedrooms, cleaning\_fee*. Здесь можно заметить, что прогнозные значения в среднем ниже, чем реальные.

Рис. 6 Полиномиальная регрессия.

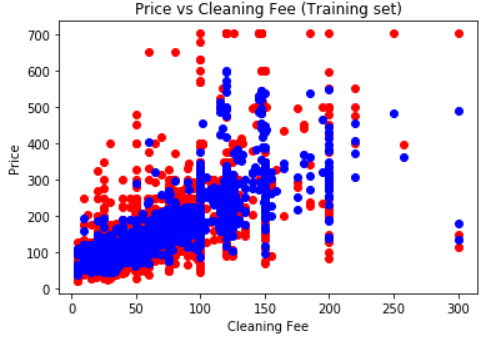
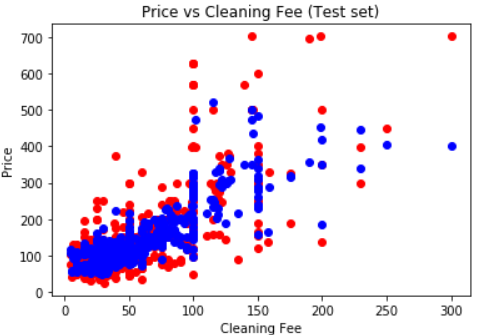
1. Дерево решений. На первый взгляд, ситуация здесь стоит намного лучше предыдущих моделей.

Рис. 7 Дерево решений.

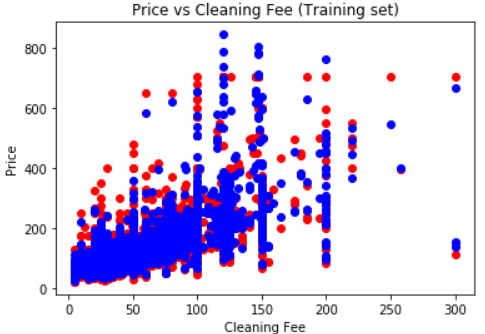
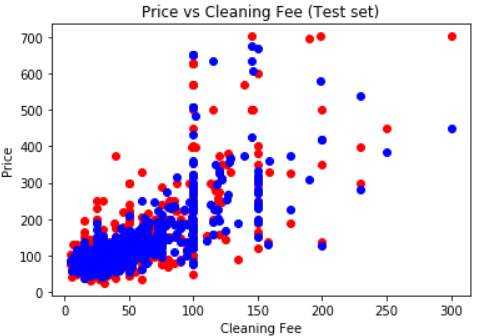
1. Случайный лес. Очень схожий результат с одним деревом решений.

Рис. 8 Случайный лес.

1. Бустинг. По графикам этот метод не уступает двум предыдущим.

Рис. 9 Бустинг.

*4.4 Оценка моделей*.

Для наилучшего представления качества созданных моделей является таблица их метрик

Таблица 1. Метрики моделей.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Простая лин регр.** | **Множ регр** | **Пол регр.** | **Дерево решений** | **Случайный лес** | **Бустинг** |
| **MSE (train)** | 8079.70 | 5360.79 | 5423.06 | 1722.53 | 1377.03 | 1366.44 |
| **MSE (test)** | 7209.08 | 5395.73 | 5005.85 | 3670.94 | 2456.20 | 2187.53 |
| **RMSE (train)** | 89.89 | 73.22 | 73.64 | 41.50 | 37.11 | 36.97 |
| **RMSE (test)** | 84.91 | 73.46 | 70.75 | 60.59 | 49.56 | 46.77 |
| **R^2 (train)** | 0.47 | 0.64 | 0.59 | 0.80 | 0.86 | 0.86 |
| **R^2 (test)** | 0.50 | 0.68 | 0.60 | 0.68 | 0.76 | 0.77 |

Исходя из этой таблицы сложно выбрать «подходящую» для нашей бизнес-задачи модель. У линейных моделей само по себе плохое качество, а деревья решений и их ансамбли с большой долей вероятности переобучены.

**Классификация**

1. **Моделирование.**

*4.1*Выбор алгоритмов.

Перед созданием моделей, следует прологарифмировать числовые переменны, чтобы их распределения было схожи к нормальному. Качественные переменные были закодированы. Затем все переменные прошкалировали.

Были использованы такие методы для решения задачи классификаци, как:

1. Логистическая регрессия;
2. KNN;
3. SVM с радиально-базисным ядром;
4. Наивный Байес;
5. Дерево решений;
6. Случайный лес.

*4.2 Планирование тестирование.*

Изначальная выборка была поделена на обучающую и тестовую выборку с пропорцией 80/20.

*4.3 Обучение модели.*

1. Логистическая регрессия. Данная модель была создана на всех факторах. Как мы видим, доля правильных ответов достаточно неплохая, однако, по остальным метрикам можно увидеть, что модель плохо обнаруживает класс Shared room, это нам говорит низкий показатель recall.

Таблица 2. Результаты Логистической регрессии.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | **Total** |
| Entire home/apt | 0.88 | 0.90 | 0.89 | 399 |
| Private room | 0.75 | 0.75 | 0.75 | 186 |
| Shared room | 0.50 | 0.24 | 0.32 | 17 |
| **Accuracy** |  |  | **0.84** | **602** |

1. KNN. Этот метод вообще не обнаружил последний класс.

Таблица 3. Результаты KNN.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | **Total** |
| Entire home/apt | 0.85 | 0.92 | 0.89 | 399 |
| Private room | 0.74 | 0.68 | 0.71 | 186 |
| Shared room | 0.00 | 0.00 | 0.00 | 17 |
| **Accuracy** |  |  | **0.82** | **602** |

1. SVM с радиально-базисным ядром. Алгоритм обнаружил какие-то уникальные характеристики присущие только последнему классу, однако, ими обладают лишь малая доля. Остальные очень схожи с другими классами.

Таблица 4. Результаты SVM.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | **Total** |
| Entire home/apt | 0.89 | 0.92 | 0.90 | 399 |
| Private room | 0.76 | 0.77 | 0.77 | 186 |
| Shared room | 1.00 | 0.12 | 0.21 | 17 |
| **Accuracy** |  |  | **0.85** | **602** |

1. Наивный Байес. Данный метод в целом имеет хуже результат, по сравнению с остальными.

Таблица 5. Результаты Наивный Байес.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | **Total** |
| Entire home/apt | 0.98 | 0.36 | 0.52 | 399 |
| Private room | 0.41 | 0.92 | 0.57 | 186 |
| Shared room | 0.25 | 0.53 | 0.34 | 17 |
| **Accuracy** |  |  | **0.54** | **602** |

1. Дерево решений. Модель намного лучше предыдущих, но это все равно ещё далеко от желаемого результата.

Таблица 6. Результаты Дерево решений.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | **Total** |
| Entire home/apt | 0.88 | 0.93 | 0.90 | 399 |
| Private room | 0.82 | 0.72 | 0.77 | 186 |
| Shared room | 0.47 | 0.47 | 0.47 | 17 |
| **Accuracy** |  |  | **0.85** | **602** |

1. Случайный лес. Этот алгоритм по результатам схож с методом SVM с радиально-базисным ядром.

Таблица 7. Результаты Случайный Лес.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | **Total** |
| Entire home/apt | 0.89 | 0.92 | 0.90 | 399 |
| Private room | 0.75 | 0.77 | 0.76 | 186 |
| Shared room | 1.00 | 0.06 | 0.11 | 17 |
| **Accuracy** |  |  | **0.85** | **602** |

*4.4 Оценка моделей*.

Все модели не смогли в полной мере найти какие-либо закономерности, связанные с последним классом. Это наталкивает на мысль проведение более глубокого анализа данных, а конкретно связанных с этим классом. Быть может, его лучше распределить по двум остальным классам.

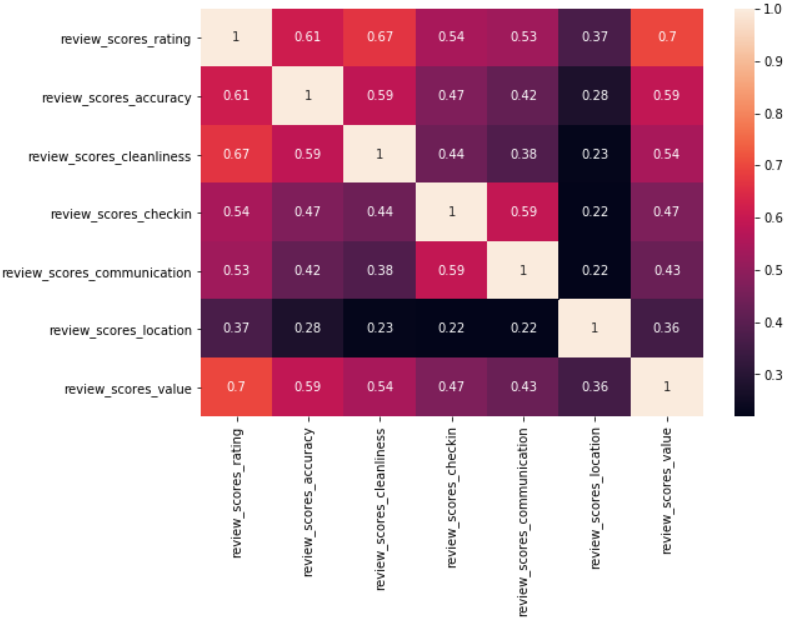
**Кластеризация**

1. **Моделирование.**

*4.1*Выбор алгоритмов.

Взятые мной факторы для кластеризации не требовали какого-либо дополнительного шкалирования. Эти факторы имеют неплохую корреляцию между собой.

Рис. 10 Матрица корреляций для кластеризации



Были использованы такие методы для решения задачи кластеризации, как:

1. Иерархическая, агломеративная кластеризация;
2. K-means.

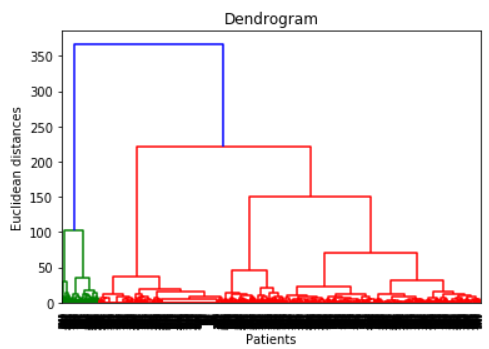
*4.2 Планирование тестирование.*

Для кластеризации была взята вся изначальная выборка.

*4.3 Обучение модели.*

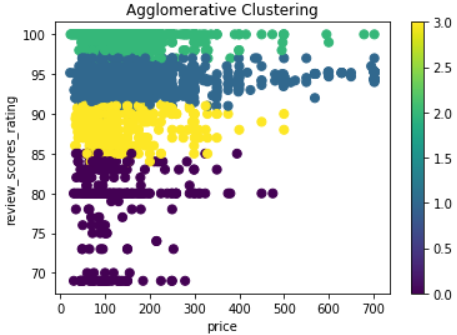
1. Иерархическая, агломеративная кластеризация. По имеющейся дендрограмме нельзя с уверенностью определить количество кластеров.

Рис. 11 Дендрограмма.



Сначала, я решил посмотреть, как изменятся цена в зависимости от общего рейтинга недвижимости, если разбить выборку на 4 кластера.

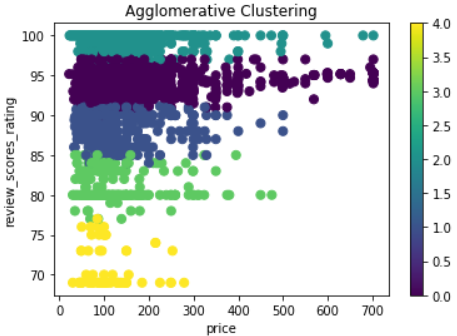
Рис. 12 4-кластера



Заметим, что чем выше рейтинг, тем возможнее дороже окажется недвижимость.

Теперь посмотрим, насколько измениться картина, если мы поделим на 5 кластеров.

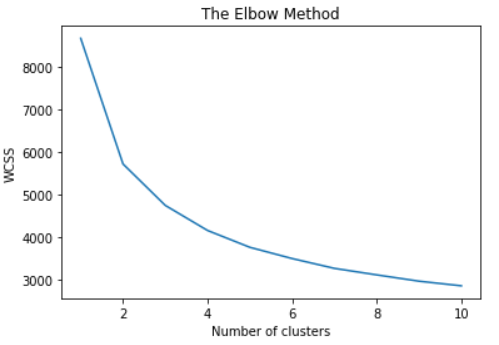
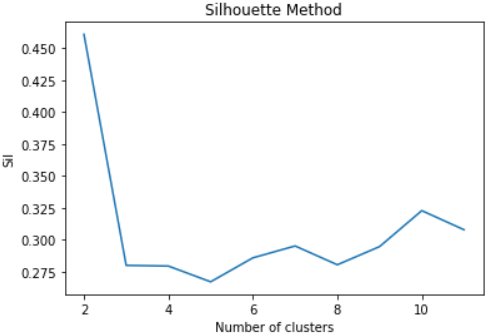
Рис. 13 5-кластера



Как видим, вывод, сделанный для 4 кластеров актуален и для 5.

1. K-means. Чтобы определить оптимальное число кластеров для алгоритма K-means следует использовать метод локтя и силуэта.

Рис. 14 Метод локтя и метод силуэта.

По методу локтя начиная со 2 кластера внутригрупповой разбросы приблизительно с одинаковой скоростью уменьшаются. А методу силуэта лучше всего оставить 2 кластера.

Однако, для сравнения с результатами иерархической кластеризации возьмем тоже 4 и 5 кластеров.

*4.4 Оценка моделей.*

В итоге, кластеры между собой пересекаются примерно на 78%, что является неплохом показателем правильности разбиение на такое количество кластеров.

1. **Оценка результата**
   1. *Оценка результатов моделирования.*

Результаты задач регрессии и классификации получились неудовлетворительными, что оговаривалось в пунктах 4.4 Оценка моделей. В то же время, с помощью кластеризации по оставленным оценкам от арендаторов, недвижимости можно выделить на категории по рейтингам «Низкий», «Нормальный», «Хороший», «Очень хороший», «Отличный», благодаря чему арендодатели смогут оценивать и анализировать свое положение на рынке.

* 1. *Анализ процесса выполнения проекта.*

Как и ожидалось, самое время затратное и трудоемкое оказалось подготовка данных. Для улучшения результатов стоит вернуться к этому этапу и провести более глубокий анализ, а именно:

* Обработать фактор «удобства» для его использования в моделировании;
* Подбор наилучшего метода обработки выбросов.

Также, было обнаружено, что мощность рабочего компьютера оказалось недостаточной для выполнения задания.

* 1. *Принятие решения.*

Весь имеющийся материал и записи остаются для анализа и усовершенствования в будущем. Возможно, стоит обдумать проект по созданию рекомендательной системы недвижимости на основе результатов кластерного анализа, где целевой аудиторией будут выступать арендаторы и арендодатели.