Постановка задачи:

Будем рассматривать задачу бинарной классификации (1 - квартира будет продана в течение месяца с заданной даты,0-квартира не будет продана). Для каждого класса рассчитаем вероятность принадлежности на основе набора признаков двух типов:

- Не зависящих от времени (размер, этаж, отделка,…),
- Зависящих от времени (спрос, акции, ценообразующие факторы,...).

Выбор модели.Формирование обучающей выборки:

Составление набора признаков зависит от способа формирования обучающей выборки. Рассмотрим три варианта:

1. Разделим весь рассматриваемый период времени на промежутки равной длительности (в зависимости, от длины интересующего нас интервала предсказания (в данном случае, 1 месяц), средней частоты происходящих событий, а также на основе величины ошибки по итогам тестирования различных вариантов).

Для каждого рассмотренного интервала сформируем набор признаков. Интервал, в котором произошла сделка-пометим единицей, остальные нулями.

Достоинства	Недостатки
• Большой размер обучающей выборки	• Несбалансированность обучающей выборки-количество негативных примеров

	значительно превышает количество позитивных. • Неопределенность в прогнозировании, в случае, если дата прогноза пересекает сразу несколько временных интервалов. • Сложность подбора оптимального размера интервала разделения.

2. Для каждой квартиры выделим ключевые даты изменения статуса: изменение цены, увеличение спроса, бронирование и т.д. Разобьём весь период времени на промежутки, ограниченные этими датами.

Достоинства	Недостатки
• Меньшая несбалансированность по сравнению с предыдущим способом и уменьшение количества незначимых для предсказания примеров.	• Неопределенность в определении ключевых дат

3. Рассмотрим все проданные квартиры как точки в пространстве признаков.

Разобьем все множество имеющихся примеров на кластеры, избавившись от аномалий. Для каждой непроданной квартиры рассчитаем вероятность продажи, как значение некоторой функцию от расстояния до ближайшего кластера.

Достоинства	Недостатки
• Меньшая несбалансированность по сравнению с предыдущим способом и уменьшение количества незначимых для предсказания примеров.	• Неопределенность в определении ключевых дат

4. Рассмотрим временной ряд спроса на квартиры разных типов (количество комнат, качество отделки и т.п.). Построим предсказание значение ряда на основе некоторой модели (например, ARIMA). Установим вероятность продажи квартиры как значение некоторой функции от спроса на квартиры данного типа.

Достоинства	Недостатки
• Большой набор инструментов работы с временными рядами (Facebook Prophet, Python StatsModels, и т.п.)	• Учитывается неполное множество признаков

Для данной задачи была выбрана вторая модель: для каждой квартиры были рассчитаны

значения признаков **включевые моменты времени**:

- Дата продажи квартиры (помечен меткой “1”, все остальные меткой “0”)
- Начало месяцев, в котором спрос на квартиры данного типа (в зависимости от количества комнат) стал выше среднего в k раз за весь период времени, k подобрано исходя из размера результирующей выборки.
- Начало месяцев, в котором отношение спроса на квартиры данного типа (в зависимости от количества комнат) к общему спросу стал выше среднего в k раз за весь период времени
- Начало месяцев, в которых спрос упал до 0
- Даты начала и продления бронирования
- Дата изменения цены

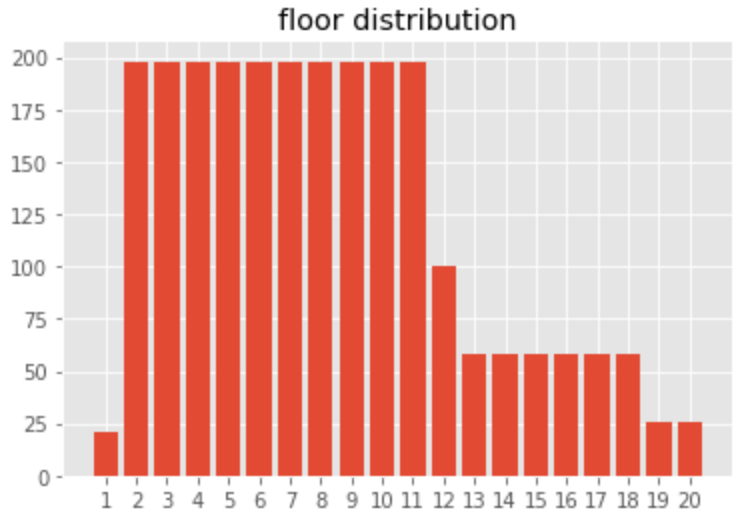
Выбор модели. Извлечение и преобразование признаков *(Feature extraction and transformation)*

Признаки, не зависящие от времени:

Квартиры, для которых не было информации в таблице *lead.csv* были исключены.

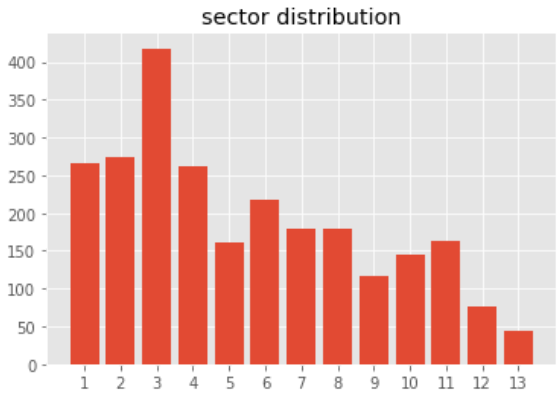
- Этаж:

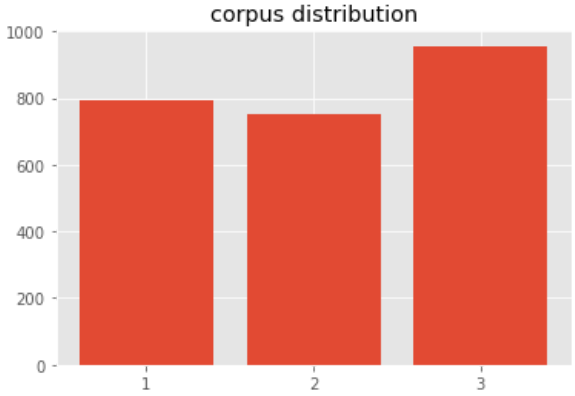
Векторизуем этот признак. Чтобы не увеличивать количество признаков значительно, разделим все допустимые значение параметра “Этаж” на интервалы: [1], [2-4], [5-8], [9-12], [13-18], [19-20] в соответствии с распределением рассматриваемых квартир по этажам:



- Сектор и корпус:

Признаки векторизованы, рассмотрены случаи исключения одного и двух признаков (см. раздел *Feature selection*):

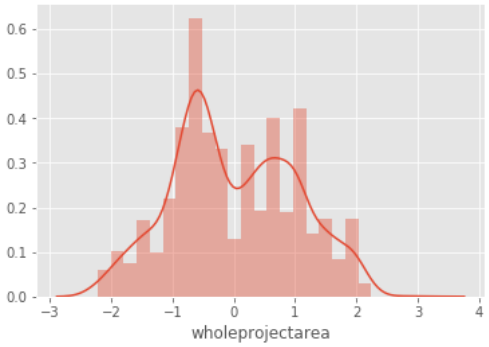




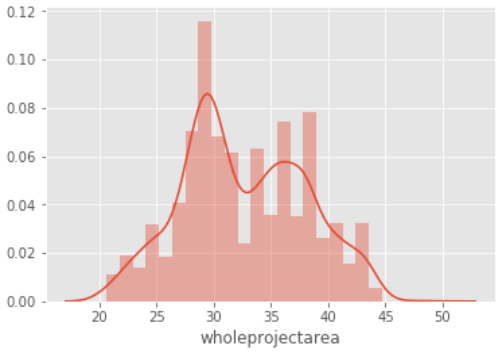
- Количество комнат:

Признак векторизован, в будущем рассматривается в совокупности с признаком “studio”.

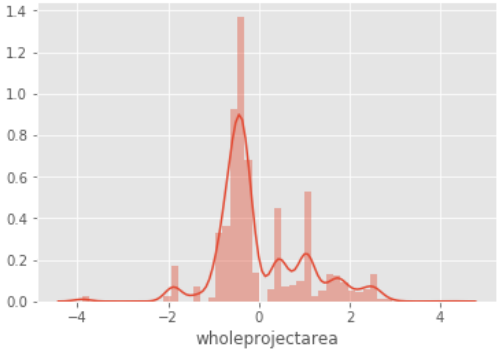
- Площадь:



Площадь квартиры заменена на площадь, приходящуюся на одну комнату.



Применен Standard Scaling:

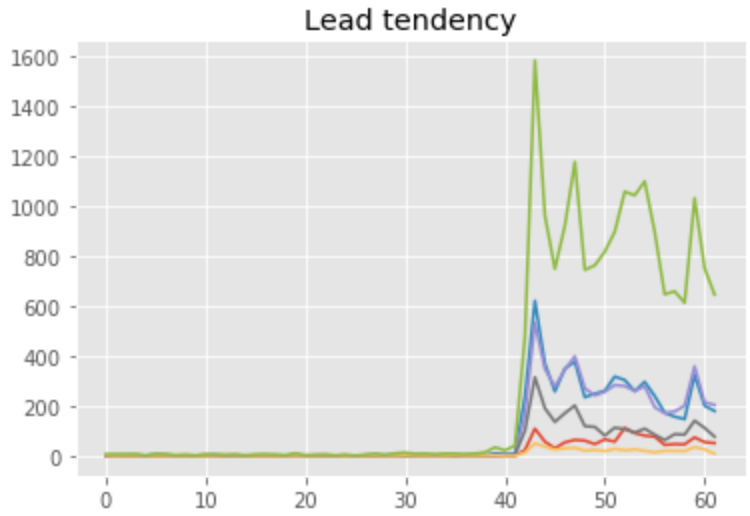


- Отделка *(finish)*

Признаки, зависящие от времени:

Чтобы устранить сезонные эффекты, вместо значений используются нормализованные соотношения.

Лиды:



На графике изображена зависимость количество лидов от номера месяца, начиная с самого первого месяца.

- *Demand_month* - отношение количества лидов на данный тип квартир к общему числу квартир за последний месяц
- *Demand* - экспоненциально взвешанное отношение количества лидов на данный тип квартир к общему числу квартир за весь период времени
- *Demand_up = 1,* если есть тенденция к повышению интереса к квартирам данного типа

Цены:

- *current_cost* - отношение текущей цены к количеству комнат
- *current_cost_up = 1,* если наблюдается тенденция к повышению цены
- *cost_dev* - отклонение цены от средней по данному типу квартир в данный момент времени

Бронирования:

- *last_resevation_duration* - длительность последнего бронирования
- *resevation_number* - количество бронирований на данный момент времени

- *resevation_duration* - Суммарная длительность бронирований на данный момент времени

Выбор модели. Отбор признаков *(Feature selection)*

- PCA
- Подбор

Выбор модели. Определение порогового значения разделения по классам

Были сравнены несколько видов классификаторов (kNN, decision tree, Random Forest, GBT). Лучший результат показал Random Forest. В качестве порогового значения для идентификации положительного прогноза выбрано 60%.