

Описание сервиса

Сервис позволяет анализировать цены за аренду жилья по его параметрам таким, как число комнат, географическое положение, число возможных гостей и т.д. и т.п. Данный модуль имеет возможность вычислить гистограммы распределения параметров жилого помещения для близких по реальной и прогнозной стоимостям, а также вычислять расстояние до некоторого числа ближайших соседей по всей базе данных. Данные подходы достаточно репрезентативно могут показать почему некоторое прогнозное значение сильно отличается от реальной стоимости, и, тем самым, может, например, намекнуть хозяину о завышении цены, или, наоборот, пользователю предложить поискать более разумный по отношению цена-качество вариант по сравнению с данным.

Интерфейс пользователя

Начальный вид сервиса представлен на рисунке:

Исследование модели

Выберите модель для анализа

model_1.model ▾

Выберите исследование модели:

☐ Вычислить shap-values

☒ Валидировать модель

Настройки выбранного исследования (Валидировать модель):

Выполнить

Исследование конкретных id

Выберите исследование id:

☒ Гистограммы (+ force plot)

☐ Расстояния

Модуль анализа модели

Проведите валидацию модели

Слева располагается панель выбора режимов исследования, пространство справа впоследствии будет содержать информацию результатов исследования.

Опишем функционал опций, представленных слева: в выпадающем списке выбора модели задаем уже посчитанную прогнозную модель, которая предварительно должна быть помещена в директорию проекта.

Для выбора исследования модели можно выбрать вычисление самых значимых признаков методом SHAP, для этого отмечаем “вычислить shap-values” и выбираем по желанию опции (можно пересчитать shap-values для модели, показать график, сохранить получившийся график):

Исследование модели

Выберите модель для анализа

model_1.model

Выберите исследование модели:

☒ Вычислить shap-values

☐ Валидировать модель

Настройки выбранного исследования (Вычислить shap-values):

Доля выборки для расчета shap-values

0.00

1.00

☒ Пересчитать shap-values для модели

☐ Сохранить график shap-values

☒ Показать график shap-values

Выполнить

Исследование конкретных id

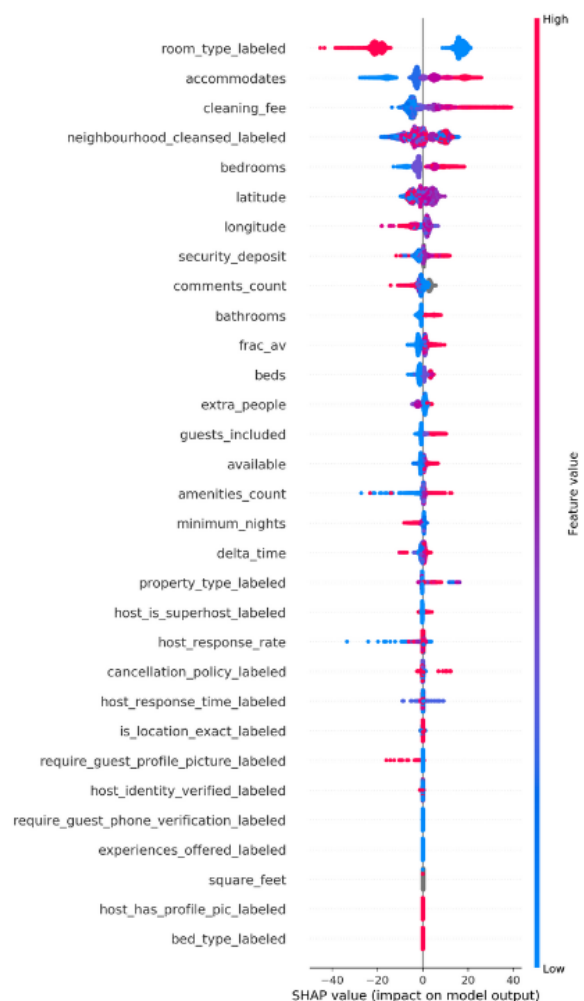
Выберите исследование id:

☒ Гистограммы (+ force plot)

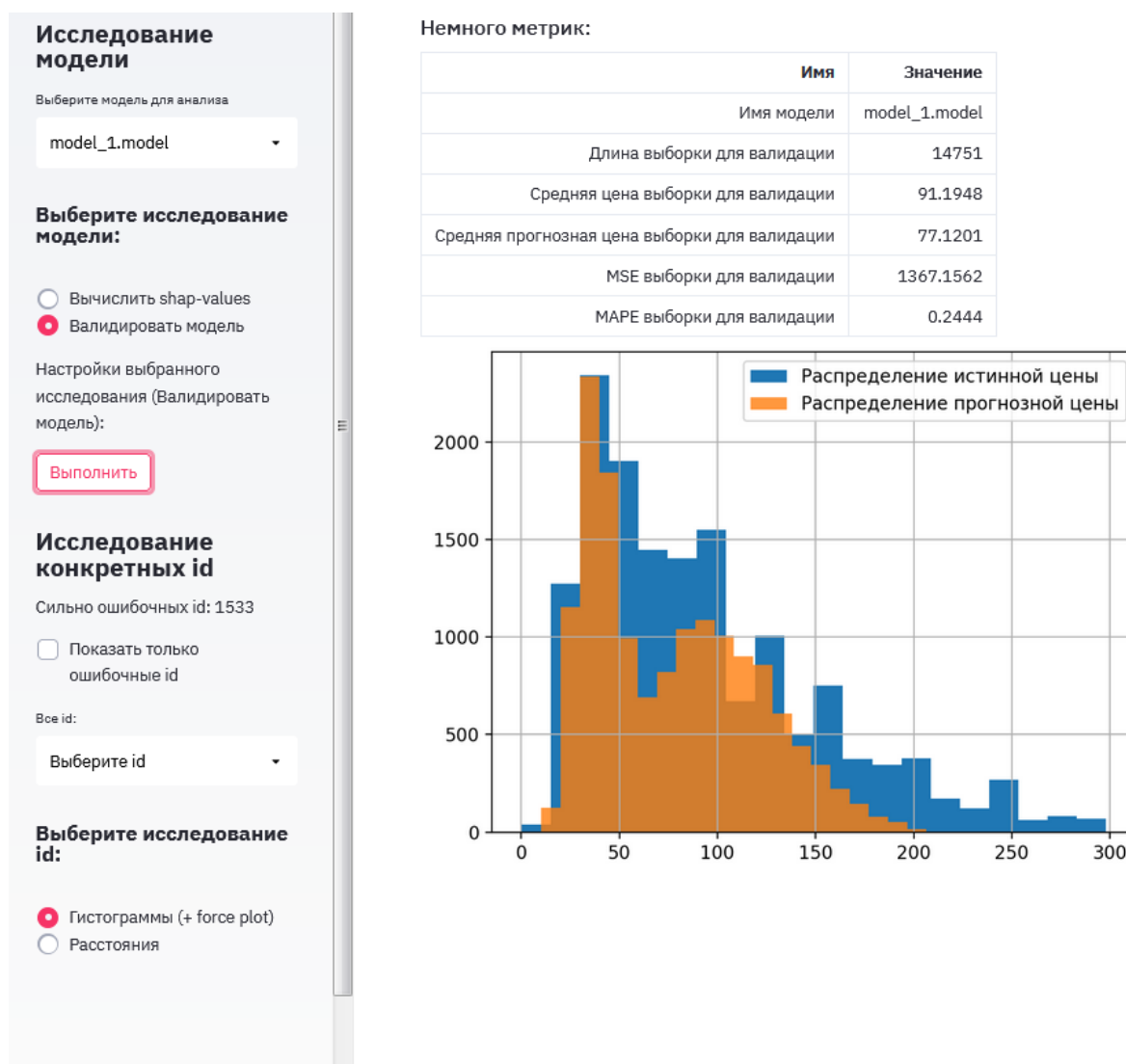
☐ Расстояния

Модуль анализа модели

Наиболее значимые признаки (SHAP-VALUES ГРАФИК):



Также можно (и даже нужно для последующего анализа) провалидировать модель, выбрав соответствующую опцию, и нажав кнопку ВЫПОЛНИТЬ:



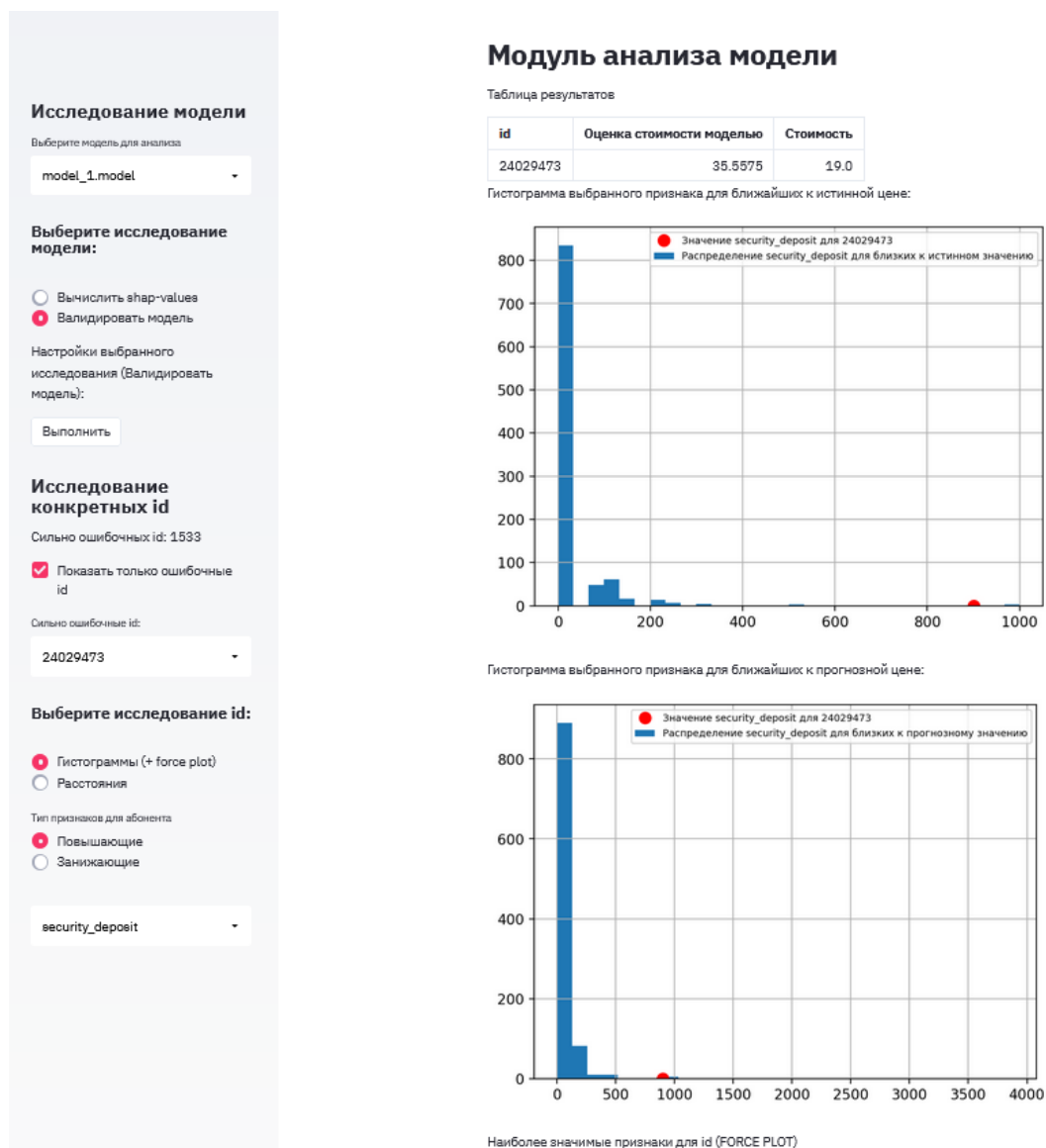
Справа появится информация о выборке и качестве модели, а также гистограмма распределения истинной цены и предсказанной моделью для валидационной выборки.

В качестве исследуемых объектов можно выбрать как все жилые помещения, так и только те, на которых прогноз дает большую ошибку. Выбор в пользу одного из этих наборов осуществляется установкой галочки около “Показать только ошибочные id”.

Из набора id выбираем один, интересующий пользователя, далее выбираем для него один из двух методов исследования — построение гистограммы выбранного признака для помещений, стоимость которых

близка к истинной стоимости выбранного пользователя, и гистограмма выбранного признака для, помещений, стоимость которых близка к прогнозной стоимости выбранного пользователя (что-то вроде байесовского подхода) и вычисление средней цены среди ближайших по признакам соседей (близко к kNN). Все параметры разделены на повышающие стоимость и понижающие стоимость и упорядочены по значению shar-values, и представлены в выпадающем списке.

Если выбрать вид исследования “гистограмма”, то справа появится следующий рисунок (как данные графики анализируются, посмотрим далее).



Для вычисления средней цены по ближайшим соседям (отмечаем кнопку “Расстояния”) нужно выбрать количество ближайших соседей и

число значимых признаков, по которым, собственно, и будут вычисляться ближайшие расстояния:

Исследование модели

Выберите модель для анализа

model_1.model

Выберите исследование модели:

☐ Вычислить shap-values

☒ Валидировать модель

Настройте выбранного исследования (Валидировать модель):

Выполнить

Исследование конкретных id

Сильно ошибочных id: 1533

☒ Показывать только ошибочные id

Сильно ошибочные id:

27712313

Выберите исследование id:

☐ Гистограммы (+ force plot)

☒ Расстояния

Значимые признаки:

☒ Текущее shap-value

☐ Все shap-values

Число ближайших соседей

1

40

Число значимых признаков

1

40

Вычислить расстояние

Модуль анализа модели

Таблица результатов

id	Оценка стоимости моделью	Стоимость
27712313	101.7851	220.0

Важнейшие признаки id:

	room_type_labeled	accommodates	cleaning_fee	neighbourhood_cleansed_	bedro
89					

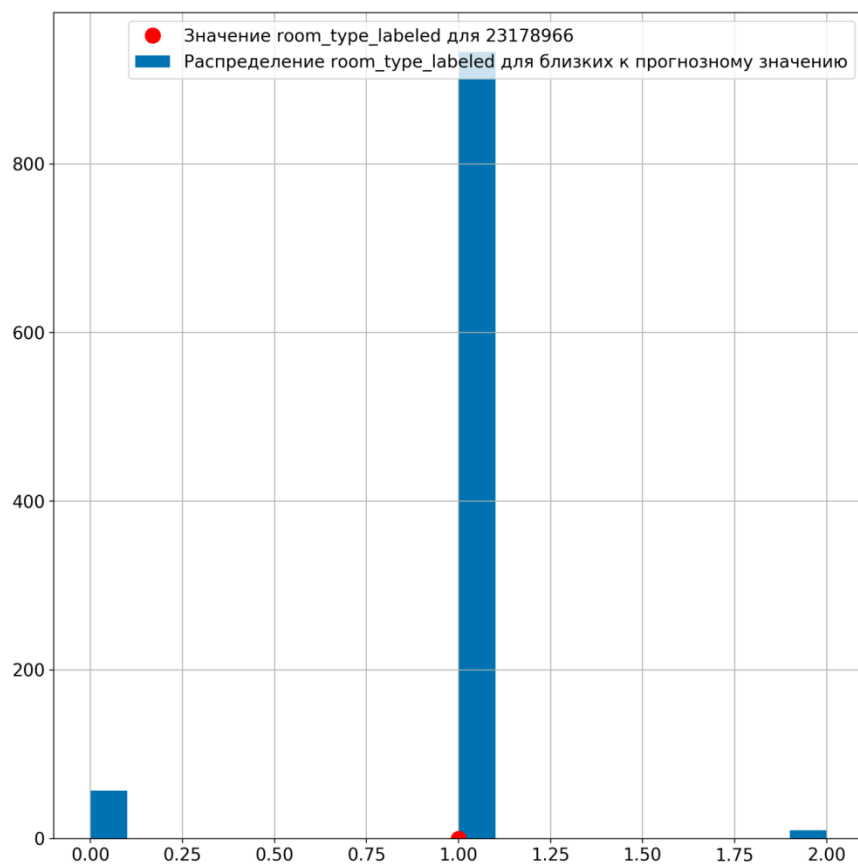
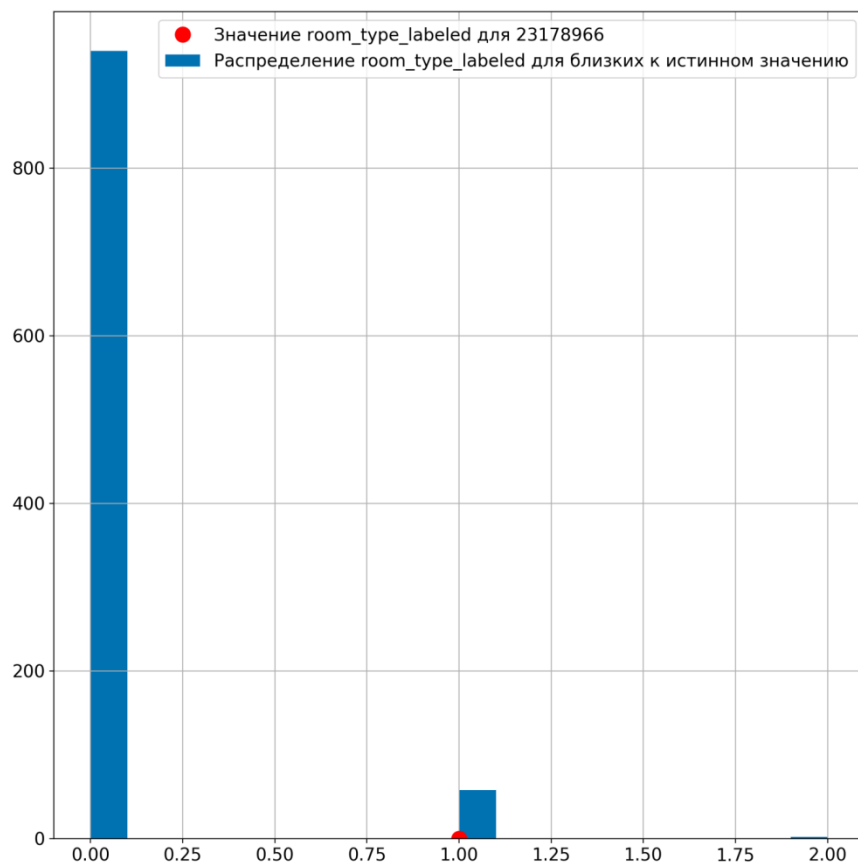
Важнейшие признаки ближайших соседей:

	room_type_labeled	accommodates	cleaning_fee	neighbourhood_cleansed_	bedro
18483	0	4	NaN		32
17245	0	4	NaN		31
21190	0	4	NaN		32
1296	0	4	NaN		32
12654	0	4	NaN		32
9681	0	4	NaN		31
28331	0	4	NaN		31
19069	0	4	0		32
29290	0	4	NaN		31
16329	0	4	NaN		32
13610					

Средняя цена ближайших соседей: 163.75

В качестве примеров покажем несколько кейсов такого анализа:

Арендуемое помещение с id=23178966 . Прогнозное значение 40.6372, реальная стоимость 125. Разница в цене почти трехкратная! Посмотрим на гистограмму распределения самого важного занижающего (по shap-values) признака room_type_labeled:



Ого! Признак `room_type_labeled`, равный 1, больше соответствует жилью за 40, нежели жилью за 120, так что, дорогой хозяин, будь так добр,

меняй цену, ну или сделай так чтобы room_type_label был равным нулю. Да и, вообще, посмотри на помещения, которые более остальных близки по признакам к твоей квартире:

Таблица результатов

id	Оценка стоимости моделью	Стоимость
23178966	40.6372	125.0

Признаки абонента:

room_type_labeled	accommodat...	cleaning_fee	neighbourhood_cleansed...	bec
884				

Признаки ближайших абонентов:

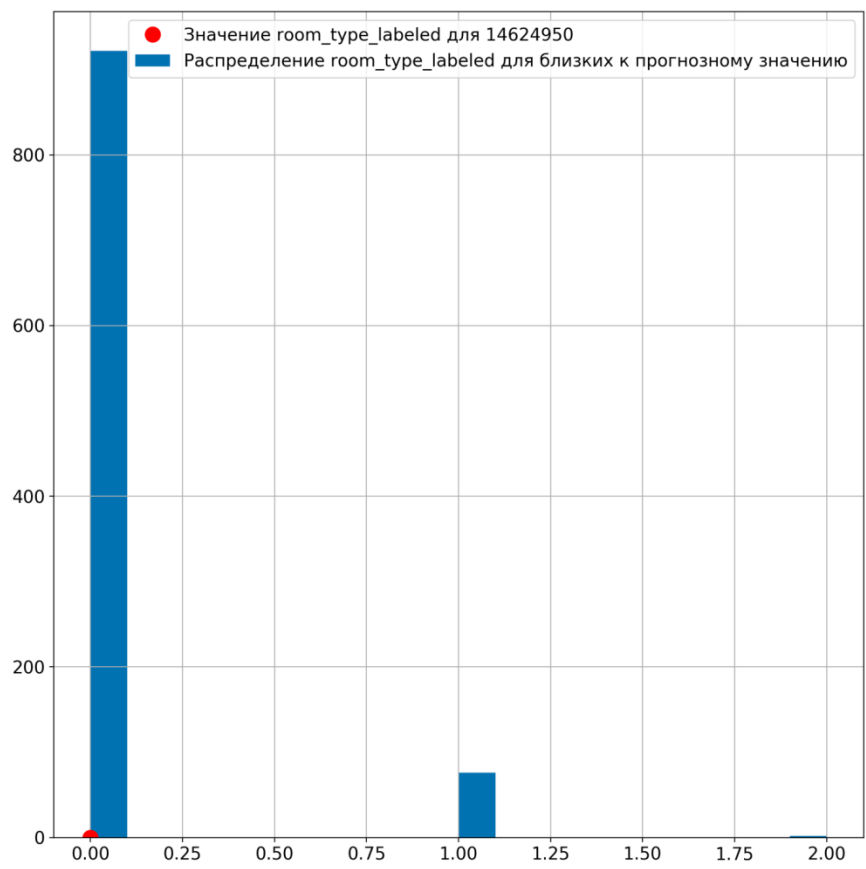
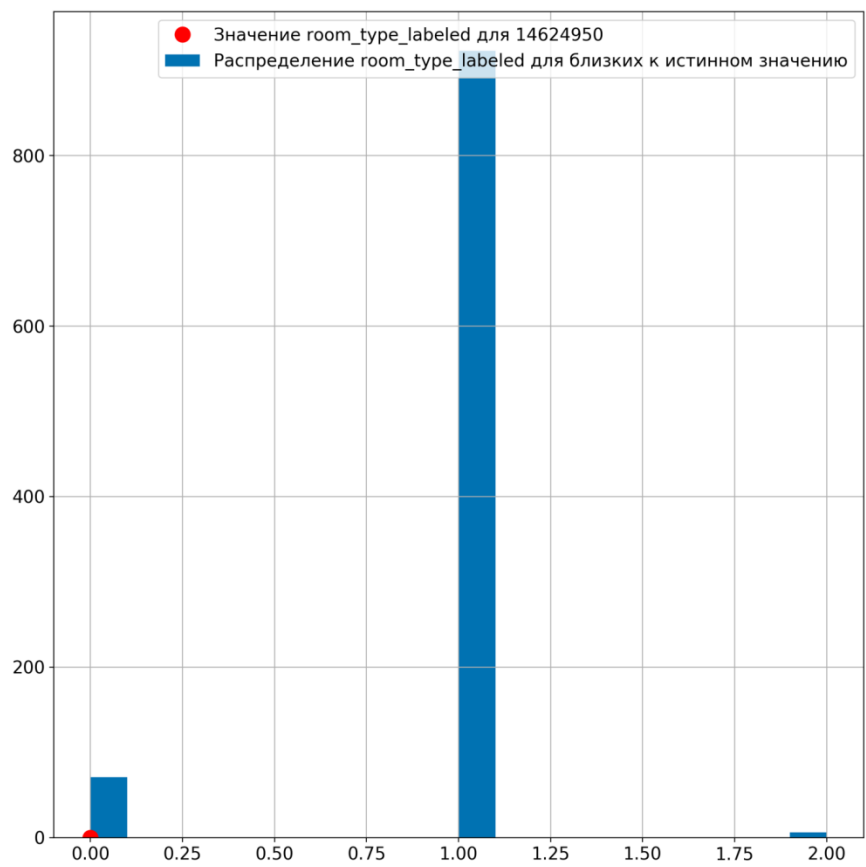
	room_type_labeled	accommodates	cleaning_fee	neighbourhood_cleansed...	t
2680	1	2	NaN	11	
6082	1	2	NaN	11	
3427	1	2	NaN	11	
18020	1	2	NaN	11	
6521	1	2	NaN	11	
8354	1	2	NaN	11	
33162	1	2	NaN	11	
27626	1	2	NaN	11	
22700	1	2	NaN	11	
25506	1	2	NaN	11	
3152					

Средняя цена ближайших соседей: 51.8

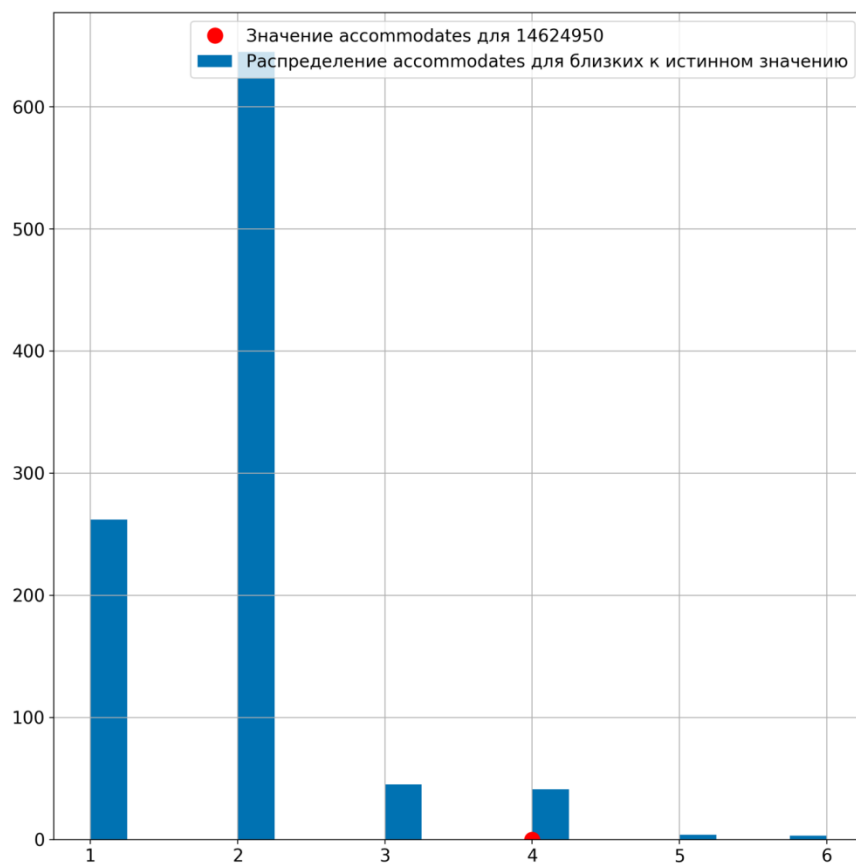
Да ты совсем с ума сошел, все пользователи же уйдут к твоим конкурентам: за меньшие деньги они получают тот же пакет удобств!

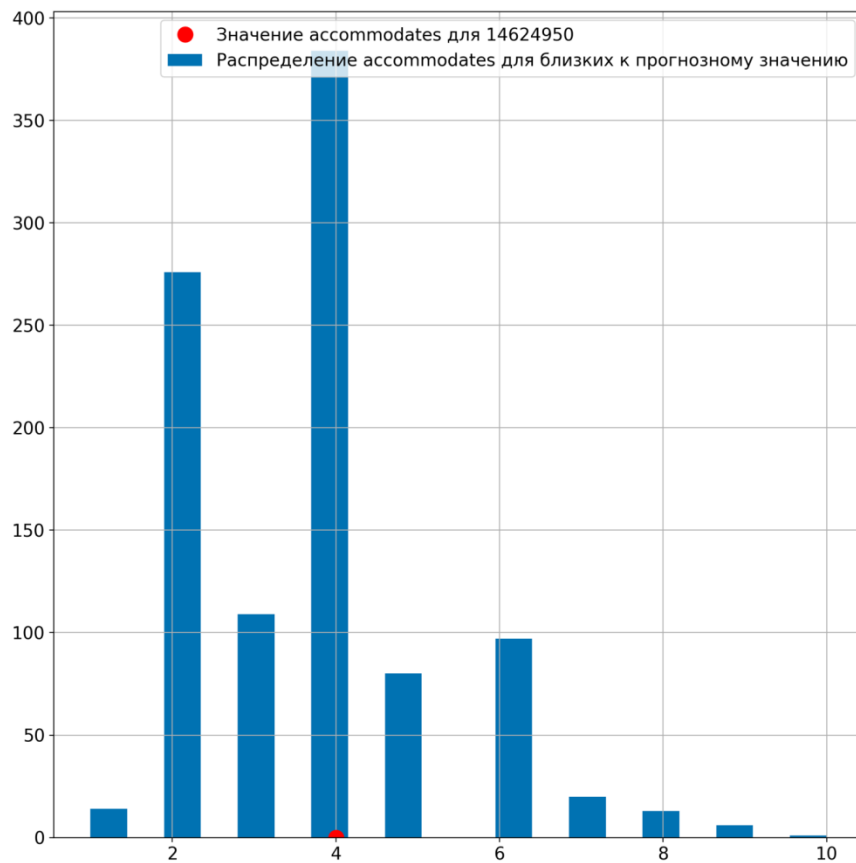
Или же диаметрально противоположный случай:

Арендуемое помещение с id=14624950 . Прогнозное значение 105, реальная стоимость 41. Разница в цене почти трехкратная? Но уже в другую сторону! Посмотрим на гистограмму распределения самого важного повышающего признака room_type_labeled:



Здесь также видно, что данный `room_type_labeled` нехарактерен для данной цены – у тебя такая хата, чувак, а ты ее за бесценок отдаешь, подними хоть в два раза – будут как горячие пирожки разбирать. А посмотрим-ка мы на вместительность:





У тебя же огромная квартира, если не неудобно , когда тебе кладут много денег сразу – можешь разделить свои апартаменты на две равные квартирki и с каждой иметь столько сколько сейчас с одной большой.

Ты только посмотри, почем твои соседи по Барвихе особняки продают:

Таблица результатов

id	Оценка стоимости моделью	Стоимость
14624950	105.0548	41.0

Признаки абонента:

	room_type_labeled	accommodates	cleaning_fee	neighbourhood_cleansed_	be
2862					

Признаки ближайших абонентов:

	room_type_labeled	accommodates	cleaning_fee	neighbourhood_cleansed_	t
30624	0	4	0		18
4500	0	5	NaN		13
11227	0	5	NaN		13
3542	0	5	NaN		11
24768	0	5	NaN		11
13430	0	5	NaN		11
15688	0	5	30		18
3490	0	5	NaN		11
3669	0	5	NaN		13
22779	0	5	10		11
28909					

Средняя цена ближайших соседей: 126.65

Ты не хуже, давай к ним!

Но, к сожалению, и сервис, бывает, сбоит:

Арендуемое помещение с id=16700630 . Прогнозное значение 33, реальная стоимость 65. Посмотрим на гистограмму распределения самого

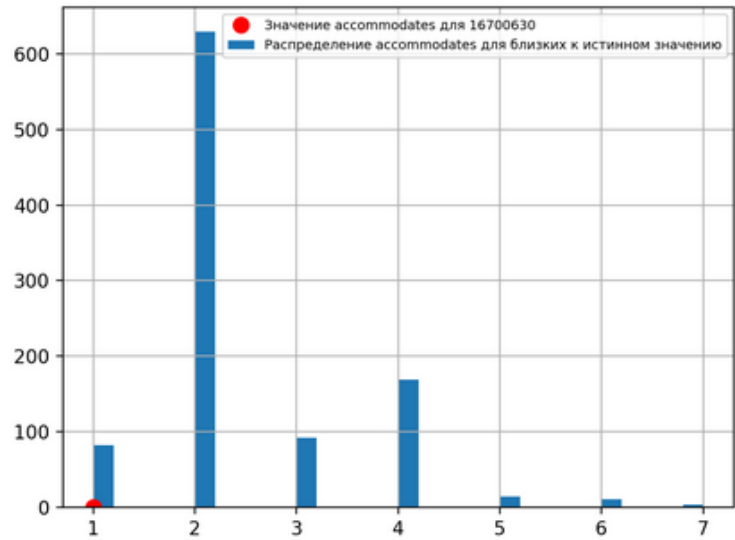
важного занижающего признака accomodates:

Модуль анализа модели

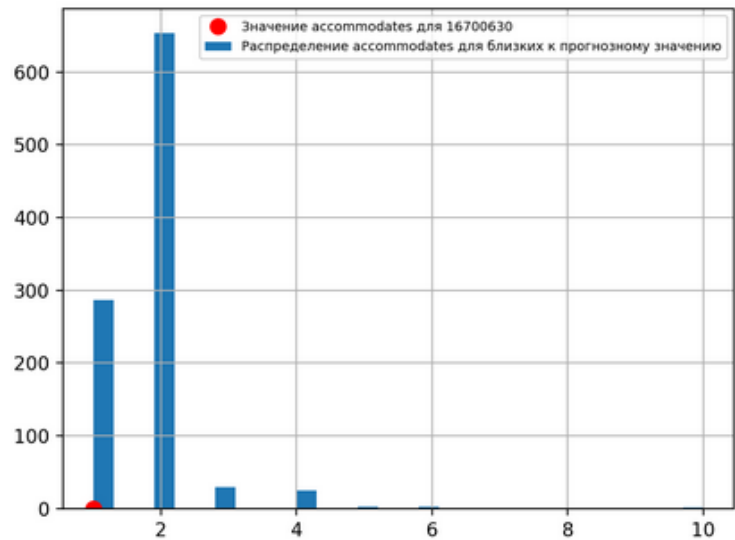
Таблица результатов

id	Оценка стоимости моделью	Стоимость
16700630	33.2163	65.0

Гистограмма выбранного признака для ближайших к истинной цене:



Гистограмма выбранного признака для ближайших к прогнозной цене:



Наиболее значимые признаки для id (FORCE PLOT)

Похоже, понятно, почему модель выбрала именно эту цену. А вот если посмотреть на ближайших соседей, то нас ждет сюрприз:

Модуль анализа модели

Таблица результатов

id	Оценка стоимости моделью	Стоимость
16700630	33.2163	65.0

Важнейшие признаки id:

	room_type_labeled	accommodates	cleaning_fee	neighbourhood_cleansed..	bedro
57	III				

Важнейшие признаки ближайших соседей:

	room_type_labeled	accommodates	cleaning_fee	neighbourhood_cleansed..	be
30195	1	1	10	32	
24544	1	1	10	32	
15676	1	1	20	32	
3420	1	1	30	32	
27560	1	1	30	32	
34235	1	1	10	32	
21868	1	2	30	32	
28906	1	1	15	32	
5306	1	1	15	32	
16103	1	2	20	32	
20076	III				

Средняя цена ближайших соседей: 58.85

У ближайших соседей средняя стоимость как раз близка к реальной стоимости, а модель выдает цену в 2 раза меньшую, возможно, благодаря влиянию вместительности жилого помещения, соответствующего жилью классом ниже.

Дальнейшее развитие.

В перспективе сервис позволит дать рекомендации не только нетипичным с точки зрения прогноза пользователям, но и тем, кто полностью соответствует рынку. Сделать это можно, например, благодаря графику FORCEPLOT, который отображает влияние признаков на формирование цены для конкретного пользователя (в отличие от графика shar-, который показывает усредненную важность признаков по всей выборке). Так, имея подобный график, можно попытаться скорректировать признаки, наиболее сильно формирующие цену.