Санкт-Петербургский государственный университет

Информационно-аналитические системы

Группа 17Б.07-мм

ЯВИЧ Максим Ильич

Выпускная квалификационная работа

Использование алгоритмов градиентного бустинга для оценки объектов недвижимости

Уровень образования: бакалавриат

Направление *02.03.03 «Математическое обеспечение и администрирование информационных систем»*

Основная образовательная программа *СВ.5006.2017 «Математическое обеспечение и администрирование информационных систем»*

Научный руководитель:
к.ф.-м.н., доцент кафедры информационно-аналитических систем, Михайлова Е.Г.

Рецензент:
ООО «Ай Ти Сервис» Осипов Е.В.

Санкт-Петербург
2021

Saint Petersburg State University

Information Analytical Systems

Group 17B.07-mm

Iavich Maksim

Bachelor's Thesis

Using the gradient boosting algorithms for evaluating real estate objects

Education level: bachelor

Speciality *02.03.03 "Software and Administration of Information Systems"*

Programme *CB.5006.2017 "Software and Administration of Information Systems"*

Scientific supervisor:
Associate Professor, Ph.D. Mikhailova E. G.

Reviewer:
"IT Service" Osipov E.V.

Saint Petersburg
2021

Оглавление

1. Постановка задачи

Цель данной работы заключается в проектировании и тестировании модели машинного обучения для предсказания цен на рынке недвижимости г. Санкт-Петербург. Для этого были поставлены следующие задачи:

1. Собрать данные по продажам в г. Санкт-Петербург;
2. Получить из них нужные для моделей машинного обучения признаки;
3. Спроектировать модель используя алгоритмы градиентного бустинга;
4. Оценить качество модели используя подходящие метрики;

2. Введение

Рынок недвижимости является базовым элементом экономики любой страны [1]. В России инвестиции в недвижимость считаются надежными вложениями последние годы [2]. Количество сделок купли-продаж объектов жилой недвижимости в 2017 году составило 4 миллиона. При этом в статье [3] объясняется, что в большинстве случаев покупка производится не для улучшения жилищной ситуации, а в качестве инвестиции. Как и в других областях это порождает задачу точной оценки жилья, что является сложной, не линейной задачей. Решением этой задачи традиционно занимаются риелторы и оценщики недвижимости. Однако их оценки могут сильно разниться и достигать погрешности вплоть до 13%. Развитие машинного обучения показывает, что в большинстве случаев алгоритм решает эту задачу лучше человека при должном количестве входных данных. Поэтому сейчас особую нишу занимают агрегаторы объектов недвижимости, т.к. они предоставляют оценку онлайн по параметрам и часто гарантируют более высокую точность. Например, американский сайт Zillow [1] гарантирует медианную ошибку в 2% и считается лидером в вопросе оценивания. В российском сегменте тоже есть свои игроки, но они пока не добились таких результатов. Сайт Циан[2] дает коридор оценки в 10%. Важность точности оценки прежде всего заключается в том, что продавец может получить полезную корректировку и быстрее продать свой объект, а покупатель может понять, насколько хорошее предложение перед ним. Также оценка является крайне важна для банков, которые выдают кредиты на покупку жилья. Исследование, которое проводил Сбербанк было построено на кредитных заявках. Тема оценки объектов недвижимости актуальна не только для агентов, покупателей и продавцов, а также и для информационных ресурсов, различных агрегаторов, потому что они соревнуются в точности своих предсказаний. Отсюда формируется цель собрать и обработать дополнительные признаки, который описывают объект, для

[1] https://www.zillow.com/

[2] https://www.cian.ru/

уточнения текущих решений. В данной работе внимание было сосредоточено на точности цены в объявлении, для этого нужно было получить и использовать, как можно больше информации из объявлений. Об этом подробнее рассказано в главе 4.1

Задача предсказания цены представляет собой задачу регрессии. Некоторые методы и алгоритмы решения задачи регрессии рассмотрены в главе 4.3. Изучив работы, приведенные в главе 3 было установлено, что для решения этой задачи эффективнее всего будет использование алгоритмов, построенных на ансамблях деревьев.

3. Обзор

- Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboost Algorithm

  В статье [4] рассмотрены основные принципы использования градиентного бустинга и в частности XGBoost библиотеки. Результаты применения гридентного бустинга в реализации XGBoost сравнивали с многомерной линейной регрессией и деревьями выбора. Установлено, что XGBoost показывает наилучших результатов в области оценки вторичной недвижимости, достигая точности 0.9251 в R-квадрат метрике.

- Housing Price Prediction via Improved Machine Learning Techniques

  Сравнение различных реализаций методов, использующих градиентный бустинг, приводится в работе [5]. Кроме сравнения этих моделей между собой, строится стекинг модель которая показывает более высокую точность на тестовой выборке, хотя на тренировочных данных наблюдалось отставание от всех моделей в метрике RMSLE.

- Housing Price Prediction Based on CNN

  Авторы статьи [6] использовали нейронные сети в качестве основного метода. Им удалось достигнуть результатов лучше, чем XGBoost модели. Однако они использовали градиентный бустинг для получение главных параметров для построения CNN сети.

- A Geographic Feature Integrated Multivariate Linear Regression Method for House Price Prediction

  Пример нестандартного использования гео-данных приведен в статье [7]. Обычно по гео-данных строится модель, результаты которой используются в основной модели, или географические признаким обрабатываются так, чтобы их можно было использовать в основной модели. Здесь же производится кластеризация по гео-

графическим признаки и по каждому кластеру строятся отдельные независимые модели.

- Identifying Real Estate Opportunities Using Machine Learning

  В статье [8] сравниваются не алгоритмы градиентного бустинга, а классические алгоритмы в том числе KNN модель, которая очень важна для области оценки недвижимости. В качестве представителей бустинг алгоритмов используются ансамбли деревьев. Был проведен анализ времени обучения моделей, что может служить важным параметров для использования в индустрии.

- Deep Learning with XGBoost for Real Estate Appraisal

  Авторам статьи [9] удалось использовать не только числовые признаки, но также и фотографии объектов, чтобы построить модель для оценки. Они использовали предобученные CNN сети с функцией активацией ReLu, чтобы получить уровень эстетичности, который впоследствии использовался наряду с другими числовыми показателями XGBoost модели. Однако не было предоставлено сравнения модели, использующей только числовые данные и с добавлением признаков, полученных из фото, поэтому тяжело судить о важности вклада этого параметра. В работе [10] авторам удалось собрать датасет из 5000 объектов с фотографиями. Они использовали только информацию о фото и расположение объекта для предсказания цены.

- Identifying Real Estate Opportunities Using Machine Learning

  Различные современные методы для предсказания цены на данных, собранных из открытых источников, были рассмотрены в статье [11]. Были проведены сравнения алгоритмов в разных метриках и условиях. Самые лучшие результаты показали модели использующие ансамбли деревьев решений. На тестовой выборке была получена MAPE 16.80% и MdApE 5.71%.

4. Сбор данных и построение модели

4.1. Сбор данных

Для построения модели необходим датасет с объектами на рынке недвижимости и их ценами. Для этого можно использовать уже собранные данные, часто предоставляемые различными компаниями, например для соревнований на Kaggle[1]. Однако, в данной работе было решено собрать данные посредством web scrapping [12] различных платформ. Такой подход более приближен к реальным условиям и позволяет работать со свежими данными без предварительной подготовки. Также аргументом в пользу второго способа является то, что для г. Санкт-Петербург такой датасет отсутствует, самый близкий набор данных в открытом в доступе к нашей задаче, это данные, выложенные Сбербанком[2]. Для сбора своего датасета было написано серверное приложение на языке python с использованием библиотек и методов, описанных в книге [12] работающее с агретаором Avito[3] результат работы, которого обновляющаяся база данных объявлений снятых с публикаций. Было принято, что последняя цена и будет являться искомой для построения модели. После работы такого приложения в течении полутора месяцев был собран датасет состоящий из 22134 объектов и хранящийся в виде, представленном в таблице 1.

[1]https://www.kaggle.com/

[2]https://www.kaggle.com/c/sberbank-russian-housing-market/data

[3]https://www.avito.ru/

Таблица 1: Первоначальный вид собранных данных.

Признак	Тип данных	Процент пропущенных значений
Город	object	0.0
Район	object	0.0022952458867375692
Улица	object	0.0022952458867375692
Номер дома	object	0.002363582645848958
Геолокация	object	0.002363582645848958
Тип сделки	object	0.0022952458867375692
Цена аренды	object	0.0022952458867375692
Цена	object	0.002523071327908351
Тип дома	object	0.002523071327908351
Количество комнат	object	0.002523071327908351
Общая площадь	object	0.002523071327908351
Количество этажей	object	0.002523071327908351
Материал стен	object	0.002523071327908351
Тип продажи	object	0.002523071327908351
Фото	object	0.002523071327908351
Описание	object	0.002523071327908351
Дата	object	0.0027510003637686432
Ссылка	object	0.0027510003637686432
Площадь кухни	object	0.0027510003637686432
Жилая площадь	object	0.0027510003637686432
Этаж	object	0.0027510003637686432

4.2. Обработка данных и выделение признаков

Было обнаружено, что пропущенные значения в основном встречаются в одних и тех же объектах, поэтому эти данные не были включены в датасет. Также с целью уточнения модели и более конкретной задачи были отобраны только квартиры в многоквартирных, домах выставленные на продажу. После этого в датасете осталось 16307 объектов.

Данные можно разделить на 4 типа: числовые, признаковые, геоло-

кация и текстовое описание. Признаковые данные, такие как материал стен, были перекодированы в числовые. Также, воспользовавшись данными с сайта Открытые данные г. Санкт-Петербург[3], были получены дополнительные признаки такие как район, наличие лифта, мусоропровода и перекодированы в числовые. Из этих же данных были добавлены дополнительные числовые данные, например возраст дома и время с последнего капитального ремонта. Из данных геолокации были получены расстояния [13] до ближайшего метро, центра города и количество зеленых зон в радиусе 1500 метров. После того как текстовые описания были почищены от цифр, знаков препинания и других шумов был произведен стемминг слов. Были убраны часто встречающиеся слова и были опробованы NLP[14] техники LDA[15], TF-IDF[16][17]. TF-IDF векторизация хорошо удобна для таких регрессионных моделей как XGBoost и может быть легко интерпретируема. Например, на Рис.2 показан вклад 1,2-грамм, полученных векторизацией описаний, где у каждого вектора 500 признаков, каждый из которого представляет слово или пару слов, в XGBoost модели. Таким образом все собранные данные были подготовлены для построения модели предсказания цены. В качестве целевой переменной для модели будет использоваться цена за квадратный метр, из которой легко получить конечную цену. На графике 1 можно посмотреть распределение логарифма цены и цены за квадратный метр в представленных данных.

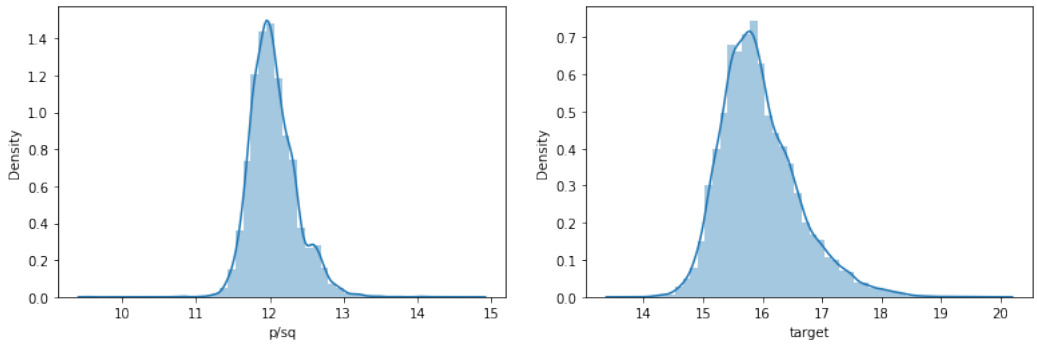

Рис. 1: Распределение логарифма цены за 1 квадратный метр и цены за всю квартиру

[3]https://data.gov.spb.ru/opendata/7840013199-passports_houses/

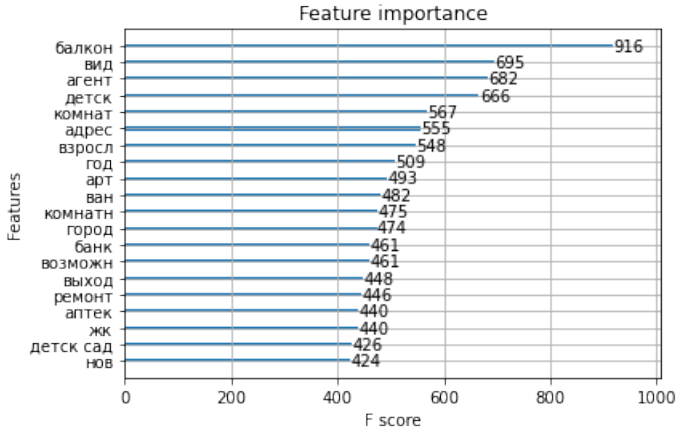

Рис. 2: Вклад слов в описание

4.3. XGBoost и базовые модели

- Линейная регрессия [18] - модель зависимости переменной y_p от одной или нескольких других переменных с линейной функцией зависимости

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\epsilon}_i \tag{1}$$

Коэффициенты уравнения линейной регрессии подбираются так, чтобы минимизировать сумму квадратов отклонения реальных точек данных от гиперплоскости, которую задает уравнение

- LASSO регрессия [19] - модификация линейной регрессии, путем введения дополнительного слагаемого регуляризации в функцию оптимизации модели. Тем самым условие минимизации отклонения при оценке параметров $\hat{\beta}$ принимает следующий вид

$$\hat{\beta} = \arg\min \left(\sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} \beta_j x_{ij})^2 + \lambda |\boldsymbol{\beta}| \right) \tag{2}$$

λ является параметром регуляризации, что позволяет снизить размерность гиперплоскости, тем самым отбрасывая параметры, которые имеют наименьший вклад в точность модели

- Elastic Net [20]- обобщение регрессии с регуляризацией. Эта модель устанавливает сразу два типа штрафных параметров - λ_1 и

λ_2.

$$\hat{\beta} = \arg\min \left(\sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} \beta_j x_{ij})^2 + \lambda_1 (\boldsymbol{\beta})^2 + \lambda_2 |\boldsymbol{\beta}| \right) \tag{3}$$

- Дерево решений [21] - наглядная модель, которая строит дерево, в листьях которого оказываются значения, а в вершинах - условия разбиения. Очень популярная модель, особенно на признаковых характеристиках. Минимизация дисперсии позволяет достичь наилучшего разбиения. В каждой вершине выборка разбивается на соразмерные подвыборки.

$$D = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - \frac{1}{\ell} \sum_{i=1}^{\ell} y_i)^2, \tag{4}$$

  где ℓ– число объектов в листе, y_i – значения целевого признака

- Random forest [22] - метамодель, надстройка над деревьями решений. На каждой итерации работы алгоритма случайным образом выбирается подвыборка данных и поднабор признаков и в каждом таком пространстве строится дерево решений.

- kNN [23] - метод K ближайших соседей.
  Жадный метод, в принципе которого лежит построение метрического пространства признаков и в зависимости от выбранного K, для каждого объекта ищется ближайшие K объектов к нему. Задача регрессии решается нахождением среднего значения зависимой переменной.

- XGBoost [24] - реализация градиентного бустинга.
  Градиентный бустинг — это метод построения композиции алгоритмов, то есть этот алгоритм является оберткой над другими существующими алгоритмами. Он представляет собой итерационный процесс, на каждой итерации которого строится новый алгоритм, который исправляет ошибки предыдущего. Подробное описание принципа построение модели в работе [25]. Реализация

XGBoost позволяет использовать в качестве базовых моделей как линейный модели, так и деревья решений. Мы воспользуемся деревьями решений в силу результатов, изображенных на таблице 2, также градиентный бустинг над деревьями прост в использовании и хорошо показывает себя в работе с разнородными данными.

- Maximal Meta Ensemble - также является композиционным методом.

  Базовыми моделями могут выступать модели любой сложности. Множество признаков из тренировочной выборки разделяется на несколько возможно пересекающихся множеств. На каждом множестве тренируется одна из базовых моделей. Предсказания базовых моделей рассматриваются как новые признаки. Эти признаки, в сочетании с изначальным множеством, необходимы для тренировки старшей - ансамблевой - модели. Принцип был найден в работе [26].

4.4. Модель

После обработки собранных данных были получены три поднабора признаков: Числовые данные, описывающие объект; Координаты; Векторизованные описания объектов. Далее были построены базовые модели машинного обучения на основе каждого типа признаков.
По числовым признакам были построены: Линейная регрессия, LASSO регрессия, Elastic Net, Дерево решений, Random Forest. Построив эти модели, была проведена кросс-валидация. Лучшие результаты показали алгоритмы над деревьями, поэтому в качестве бустера в построение XGBoost регрессора будет выступать дерево решений. На таблице 2 приведены оценки точности построенных моделей в MAPE метрике.
Координаты были выделены в отдельный признак, чтобы построить по ним kNN, однако после ознакомления со статьями ([27][28][29]) стало понятно, что модель модифицируется после добавления дополнительных числовых и категориальных признаков в пространство, над которым строится kNN. На Рис. 5 наблюдается зависимость ошибки в MAPE

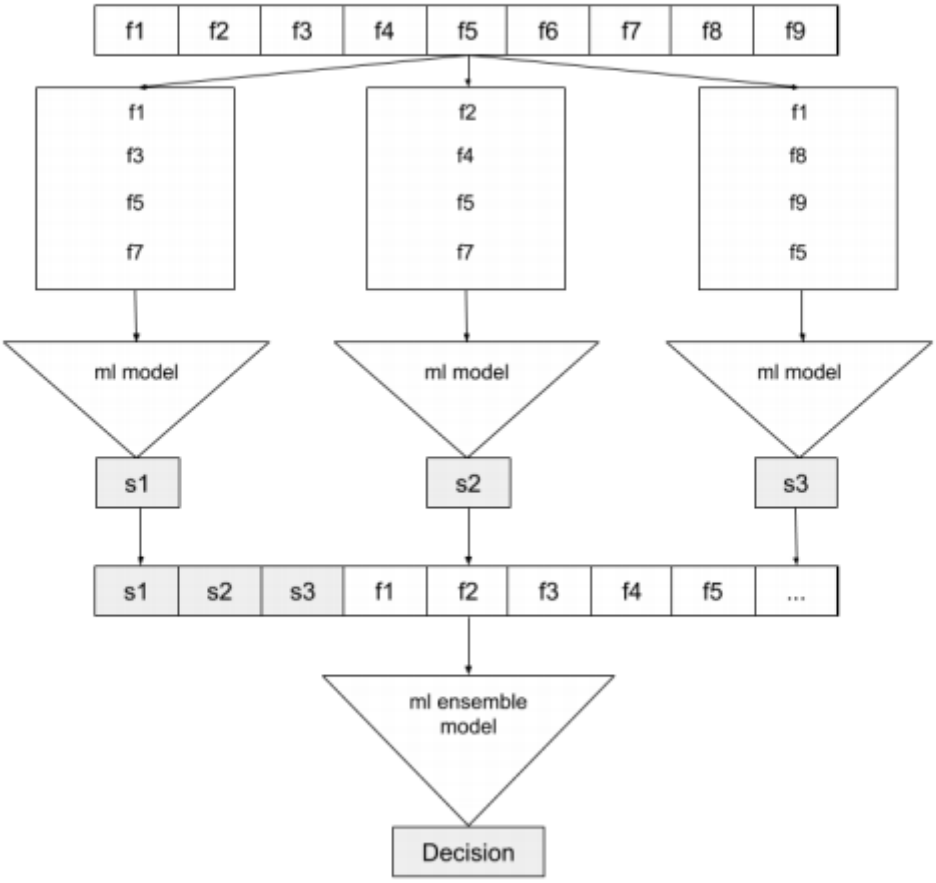

Рис. 3: Maximal Meta Ensemble

метрике от количества соседей. Синим цветом обозначены результаты невзвешенной регрессии, оранжевым цветом - взвешенной. На наших данных на пространстве с дополнительными признаками kNN модель также показала более точные предсказания. Для векторизации описаний было использовано ограничение в 500 самых встречаемых н-грамм. По этим данным была построена XGBoost регрессия. Архитектура конечного ансамбля моделей приведена на Рис 4

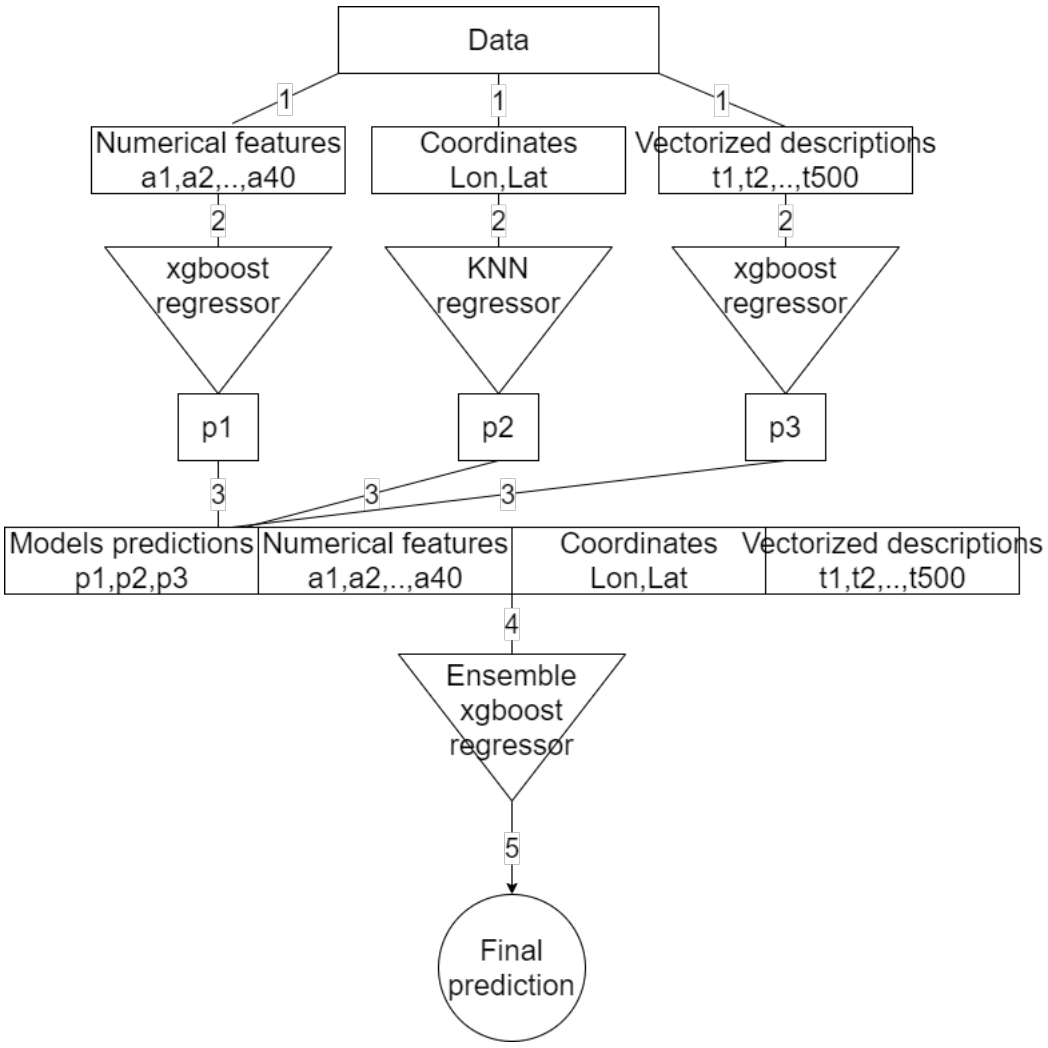

Рис. 4: Архитектура модели

5. Результаты

5.1. Исследовательские вопросы

- Получится ли использовать данные, содержащиеся в описаниях объектов для улучшения точности предсказания?
- Является ли принцип Maximal Meta Ensemble подходящим методом ансамбирования построенных моделей?

5.2. Метрики

Наибольшей популярностью для оценки качества результатов используют метрики MAPE и RMSLE.

- MAPE $= \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{y_p - y_t}{y_t} \right|$
  средняя абсолютная ошибка, часто используется чтобы сравнить с оценкой, сделанной человеком. Также наглядно отражает точность модели.

- RMSLE - $\sqrt{\frac{1}{N} \sum_{t=1}^{N} (\log(y_p + 1) - log(y_t + 1))^2}$
  среднеквадратичная логарифмическая ошибка, часто используется для сравнения результатов моделей в Kaggle соревнованиях.

- MdAPE - медиана абсолютной ошибки является популярной метрикой для оценки качества на рынке, популярный американский агрегатор объектов недвижимости Zillow считается обладателем лучшей модели в этой области и гарантирует медиану абсолютной ошибки в размере 2%

5.3. Результаты

Была предложена новая модель оценки объектов недвижимости, не только по базовым характеристикам квартиры, дома и расположения, но и с учетом описания объекта, которое, как было показано, может нести полезную информацию. Установлено, что в конечной ансамблевой модели предсказание цены по текстовому описанию имеет большой вес (Рис. 6). Некоторые 1,2-граммы тоже занимают позиции в 20-ти наиболее важных признаках. Это позволяет утверждать, что наш подход оправдал себя и полученные из текстовых описаний признаки являются важной частью нашей модели. Также была получена MdAPE 3.47%, что говорит о том, что на половине тестовой выборке была достигнута ошибка меньше, чем 3.47%

model	cv1	cv2	cv3	cv4	cv5	mean
linear	18.77%	16.75%	18.45%	22.95%	17.99%	19%
lasso	18.81%	16.87%	18.46 %	22.69%	17.02%	18.63%
elastic_net	18.81%	16.87%	18.47%	22.69%	17.02%	18.63%
decision_tree	9.90%	5.59 %	11.54%	15.82%	7.31 %	10%
random forest	9.16%	5.66%	9.83%	12.45%	6.17 %	8.64%
xgboost	8.44%	4.88%	9.77%	11.80%	5.99%	8.23%

Таблица 2: Результаты простых моделей MAPE

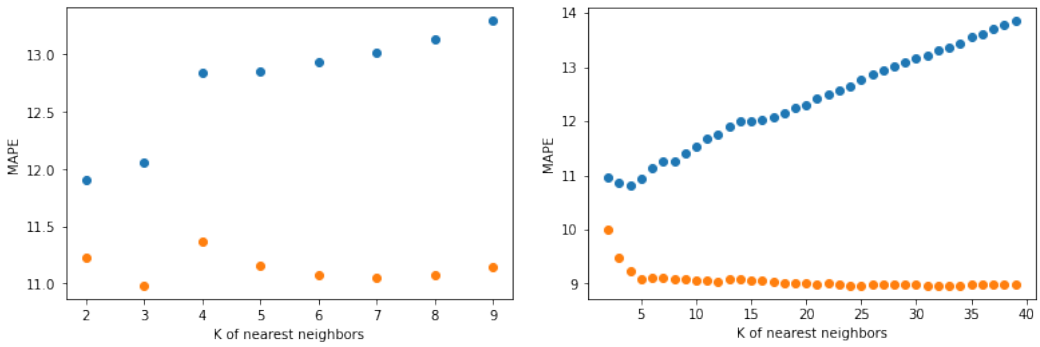

Рис. 5: kNN только по координатам и с добавлением категориальных признаков; Оранжевые точки - взвешенный способ; синие точки - невзвешенный способ

model name	xgboost	ensemble model	maximal meta ensemble
numerical	0.0322	-	-
text	0.0283	-	-
geo	0.0372	-	-
numerical+text	0.0217	0.0208	0.0204
numerical+geo	0.0247	0.0222	0.0215
text+geo	0.0280	0.0278	0.0277
text+geo+tex	0.0198	0.0196	0.0186

Таблица 3: Результаты ансамблей моделей RMSLE

model_name	xgboost	ensemble model	maximal meta ensemble
numerical	8.49%	-	-
text	11.89%	-	-
geo	9.59%	-	-
numerical+text	8.37%	8.24 %	8.19%
numerical+geo	8.61%	8.31%	8.02%
text+geo	10.74%	9.24%	9.06%
text+geo+tex	8.51%	8.17%	7.77%

Таблица 4: Результаты ансамблей моделей MAPE

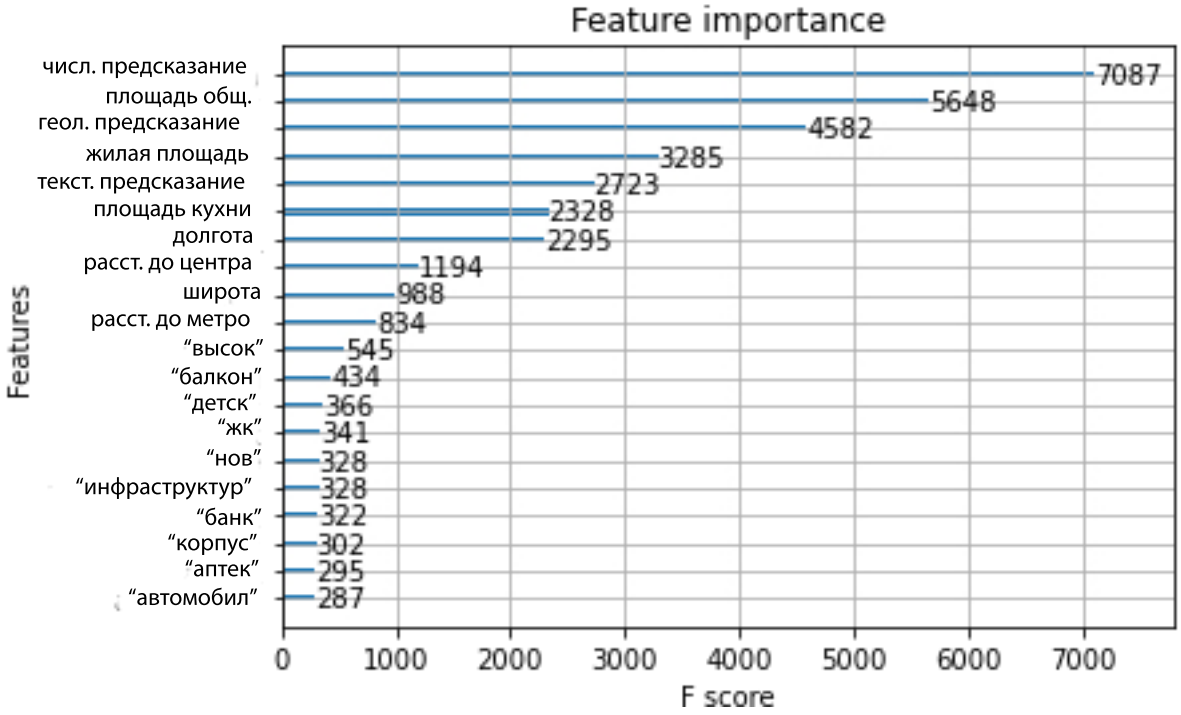

Рис. 6: Вклад признаков в конечную модель

5.3.1. ИВ1

На результирующей таблице показано изменение точности модели при добавлении текстовых данных. Улучшение предсказаний при добавлении результатов модели, обученной на текстах описаний, было экспериментально доказано.

5.3.2. ИВ2

Принцип Maximal Meta Ensemble продемонстрировал лучшее ансамблирование моделей.

6. Применение того, что сделано на практике

В большинстве агрегаторов присутствующих на рынке существуют свои модели предсказания насколько точно пользователь указал цену размещенного объекта. Это делается для того, чтобы подсказать пользователям релевантность цены. Однако в этих моделях для предсказания не рассматриваются описания объектов, которые в свою очередь занимают большую часть объявления и как было показано в работе несут полезную информации. Применением результатов этой работы может быть добавление обработки текстовой информации в существующие агрегаторы для уменьшения ошибки предсказания. Также одно из возможных приложений этого алгоритма может быть анализ разницы предсказаний, чтобы определить не является ли объявление мошенническим, однако это требует дополнительной проверки, в том числе сбор мошеннических объявлений. Протестировать работу полученного алгоритма можно в приложении, которое выложено на GitHub.[30]

7. Заключение

1. Был написан парсер на языке python использующий принципы Web scrapping описанные в [12] для сбора данных с популярных агрегаторов объявлений;

2. Из собранных данных был получен датасет числовых признаковых, которые описывают каждый объект. Для перевода тектовых описаний и выделения из них числовых признаков был применен TF-IDF метод [16][17];

3. Была спроектирована архитектура ансамблевой модели, в основе которой лежит алгоритм градиентного бустинга в реализации XGBoost[24]. Архитектура построенной модели приведена на Рис. 4;

4. Для оценки модели были выбраны MAPE и RMSLE метрики. Также была посчитана MdAPE для сравнения с популярными на рынке решениями. Результаты получилось, уточнить используя текстовые данные;

Алгоритмы градиентного бустинга показали самые точные предсказания среди рассмотренных методов на числовых данных. В будущем возможно добавление дополнительных признаков для построения модели. Из неиспользуемой информации есть фотографии, по которым тоже можно построить модель классификации или регрессии и добавить в конечный ансамбль. Также было создано консольное приложения с реализацией алгоритма. [30]

Список литературы

[1] А. М. Королева, "Роль Рынка Недвижимости В Экономике Государства," *Общество: политика, экономика, право*, no. 6, pp. 71–73, 2016.

[2] Н. А. Сучкова, "Инвестиции в недвижимость-надежный и доходный способ вложения денежных средств," *Научные записки Орел-ГИЭТ*, no. 1, p. 5, 2012.

[3] N. Kosareva and T. Polidi, "Housing affordability in russia," *Housing Policy Debate*, vol. 31, no. 2, pp. 214–238, 2021.

[4] Z. Peng, Q. Huang, and Y. Han, "Model research on forecast of second-hand house price in chengdu based on xgboost algorithm," pp. 168–172, Oct 2019.

[5] Q. Truong, M. Nguyen, H. Dang, and B. Mei, "Housing price prediction via improved machine learning techniques," *Procedia Computer Science*, vol. 174, pp. 433–442, 2020. 2019 International Conference on Identification, Information and Knowledge in the Internet of Things.

[6] Y. Piao, A. Chen, and Z. Shang, "Housing price prediction based on cnn," pp. 491–495, 2019.

[7] Y. Mao and R. Yao, "A geographic feature integrated multivariate linear regression method for house price prediction," pp. 347–351, 2020.

[8] A. Baldominos, I. Blanco, A. Moreno, R. Iturrarte, ☒. Bernárdez, and C. Afonso, "Identifying real estate opportunities using machine learning," *Applied Sciences*, vol. 8, p. 2321, 11 2018.

[9] Y. Zhao, G. Chetty, and D. Tran, "Deep learning with xgboost for real estate appraisal," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1396–1401, 2019.

[10] Q. You, R. Pang, L. Cao, and J. Luo, "Image-based appraisal of real estate properties," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2751–2759, 2017.

[11] A. Baldominos, I. Blanco, A. Moreno, R. Iturrarte, ☒. Bernárdez, and C. Afonso, "Identifying real estate opportunities using machine learning," *Applied Sciences*, vol. 8, p. 2321, 11 2018.

[12] R. Mitchell, *Web scraping with Python: Collecting more data from the modern web.* " O'Reilly Media, Inc.", 2018.

[13] C. F. F. Karney, "Algorithms for geodesics," *Journal of Geodesy*, vol. 87, pp. 43–55, Jan 2013.

[14] R. Feldman, "Techniques and applications for sentiment analysis," *Commun. ACM*, vol. 56, p. 82–89, Apr. 2013.

[15] A. Onan, S. Korukoglu, and H. Bulut, "Lda-based topic modelling in text sentiment classification: An empirical analysis.," *Int. J. Comput. Linguistics Appl.*, vol. 7, no. 1, pp. 101–119, 2016.

[16] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM Journal of Research and Development*, vol. 1, no. 4, pp. 309–317, 1957.

[17] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, 1972.

[18] M. Kutner, C. Nachtsheim, and J. Neter, *Applied Linear Regression Model*, vol. 26. 01 2004.

[19] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[20] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[21] L. Breiman, *Classification and regression trees*. 01 1984.

[22] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[23] E. Fix, *Discriminatory analysis: nonparametric discrimination, consistency properties*, vol. 1. USAF school of Aviation Medicine, 1985.

[24] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[25] J. H. Friedman, "Stochastic gradient boosting," *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[26] M. Olden, "Predicting stocks with machine learning. stacked classifiers and other learners applied to the oslo stock exchange," Master's thesis, 2016.

[27] M. F. Mukhlishin, R. Saputra, and A. Wibowo, "Predicting house sale price using fuzzy logic, artificial neural network and k-nearest neighbor," in *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*, pp. 171–176, 2017.

[28] J. W. Weikun Zhao, Cao Sun, "The Research on Price Prediction of Second-hand houses based on KNN and Stimulated Annealing Algorithm," *International Journal of Smart Home*, vol. 8, pp. 191–200, 2014.

[29] S. Borde, A. Rane, G. Shende, and S. Shetty, "Real estate investment advising using machine learning," *Int. Res. J. Eng. Technol*, vol. 4, no. 3, pp. 1821–1825, 2017.

[30] Y. Maxim, "yavichmaxim/app_for_testing," May 2021. Available at https://github.com/YavichMaxim/app_for_testing, version 1.0.