Министерство науки и высшего образования Российской Федерации
Санкт-Петербургский политехнический университет Петра Великого
Высшая школа прикладной математики и вычислительной физики

Работа допущена к защите
Директор ВШПМиВФ
_____ Л.В. Уткин
«_____» _____ 2021 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
РАБОТА БАКАЛАВРА
ВЕКТОРНОЕ ПРЕДСТАВЛЕНИЕ СТРУКТУРИРОВАННЫХ ОБЪЕКТОВ

по направлению подготовки 01.03.02 Прикладная математика и информатика

Направленность (профиль) 01.03.02_03 Математическое и информационное обеспечение экономической деятельностью

Выполнил
студент гр. 3630102/70301 Д.М. Попеску

Руководитель
Доцент,
Кандидат физико-математических наук,
старший научный сотрудник С.Ю. Беляев

Консультант
Менеджер продукта В.А. Горовой

Консультант
по нормоконтролю Л.А. Арефьева

Санкт-Петербург
2021

РЕФЕРАТ

На 30 с., 36 рисунков,

КЛЮЧЕВЫЕ СЛОВА: ВЕКТОРНОЕ ПРЕДСТАВЛЕНИЕ, РЕКОМЕНДАЦИЯ, МЕТРИКА БЛИЗОСТИ, КЛАСТЕРИЗАЦИЯ.

Тема выпускной квалификационной работы: «Векторное представление структурированных объектов»

В данной работе изложена сущность подхода к получению векторного представления структурированных объектов. Это достигается путем применений моделей из сферы обработки естесственного языка. Приведен пример с предообработкой и анализом реальных данных. Описаны различные модели получения векторного представления слов. Разработана программаня реализация для предообработки данных, обучение модели и составление рекомендаций. Проведен анализ полученного векторного представления, а также кластеризация с целью интерпретации полученных результатов.

ABSTRACT

30 pages, 36 figures,

KEYWORDS: VECTOR REPRESENTATION, RECOMMENDATION, PROXIMITY METRIC, CLUSTERING.

The subject of the graduate qualification work is «Objects embeddings».

This paper outlines the essence of the approach to obtaining a representation of structured objects. This is achieved by applying models from the natural language processing realm. An example with preprocessing and analysis of real data is given. Various models of using word representation are described. A software implementation has been developed for data preprocessing, training the model and making recommendations. The analysis of the obtained representation, as well as clustering in order to interpret the results.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Мы повсеместно встречаемся с рекомендациями на различных сервисах. Это могут быть рекомендации фильмов, книг, музыки либо товаров. В вопросе рекомендаций остаются в выигрыше обе стороны взаимодействия – это компании предоставляющие свои товары или услуги, а также пользователи, которые пользуются товарами или услугами. От того насколько точны будут рекомендации, тем быстрее и качественнее пользователи буду удовлетворять свои потребности, тем самым они сэкономят себе время, и будут оставаться лояльны компаниям предоставляющие свои услуги.

В основе работы рассматриваемой модели лежит гипотеза об дистрибутивности, которая заключается в том, что объекты, встречающие в схожих контекстах, имеют близкое значение [**harris1954distributional**]. Самой популярной моделью, основанной на данной гипотезе, является модель word2vec, позволяющая представлять слова в векторном пространстве. В данной работе объекты (объявления) представляются в многомерном векторном пространстве.

В данной работе используются структурированные объекты – это такие объекты реального мира, которые описываются конечным множеством признаков в виде таблицы объект – признак. Главной проблемой исследования таких данных является высокая степень разреженности матрицы объект – признак. В данной работе также будут рассмотрены способы борьбы с пропусками.

Главное преимущество рассматриваемого метода перед остальными методами заключается в том, что данный метод придает семантический смысл используемым объектам. Это позволяет более точно предсказывать наиболее похожие объекты между собой.

Также особый интерес представляет устройство полученного векторного пространства, которое позволяет применять полностью весь математический аппарат для исследований и поиска закономерностей.

Стоит отметить также очень важный аспект данного исследования — это выявление семантических связей между объектами. Данное свойство имеет перспективы к дальнейшим исследованиям, направленных на выявление скрытых связей между объектами.

ГЛАВА 1. ПОСТАНОВКА ЗАДАЧИ

Перейдем к формальной постановке задаче.

В данном исследовании предоставлена выборка истории поведения обезличенных пользователей, а также информация об объектах, с которыми взаимодействовали пользователи. При каждом новом входе на веб-ресурс, пользователь начинает новую сессию, которая сохраняется в обезличенном варианте.

То есть для каждой пользовательской сессии известны идентификаторы объектов, для которых известны их признаковое описание.

Необходимо построить векторное представление объектов, которые помимо признакового описания хранили в себе еще и связь между историей взаимодействия пользователей.

Пусть H - история поведения пользователей на веб-ресурсе за все время, O - множество всех объектов присутствующих в базе данных веб-ресурса. U - множество всех пользователей посещавших веб-ресурс. Тогда поведения каждого пользователя $u \in U$ описывается следующим образом: $({o_1}_h^h \dots {o_k}_h^h)^u$, где $h \in H$ - интервал времени, k_h - длинна сессии пользователя за определенный интервал времени.

Задача построения векторного представления объектов заключается в сопоставление каждому объекту $o \in O$ вектора $\upsilon_o \in \mathbb{R}^m, m \ll |O|$. Такое отображение должно давать в результате такие вектора, чтобы похожие объекты были близки по расстоянию друг к другу.

Полученные векторные представления будут использоваться для:

– Анализа пользовательских сессий;

– Кластеризации пользователей исходя из их поведения на веб-ресурсе;

– Построение рекомендательной системы;

Опредение 1. Рекомендательная система - это подкласс систем фильтрации информации, которая стремится предсказать «рейтинг» или «предпочтение», которое пользователь дал бы объекту [**ricci2011introduction**].

Рекомендательная система представляет из себя задачу ранжирования. Определим формально понятие задачи ранжирования:

X – множество объектов

Имеется выборка, состоящая из n элементов:

$$X^n = x_1, \ldots, x_n \tag{1.1}$$

Данные объекты содержат в себе признаковое описание.

В задаче ранжирования целевой переменной является пара вида:

$$(i, j) : x_i < x_j \tag{1.2}$$

Необходимо построить ранжирующее отображение:

$$f : X \to \mathbb{R} \text{ такую, что } i < j \Rightarrow f(x_i) < f(x_j) \tag{1.3}$$

1.1. Цель работы

Целью данной работы является создание рекомендательной системы в основу которого ляжет векторное представление объектов.

1.2. Методы оценки качества полученных векторов

Суть векторного представления объектов в том, чтобы объекты находящиеся в одной сессии пользователя были близки по расстоянию друг к другу. Для проверки этого свойства можно воспользоваться методами из задачи близости [**rubenstein1965contextual**]. Но каждая задача близости привязана к конкретной выборке данных, поэтому данные методы не подходят для нашего исследования, так как исследуемые данные не являются публичными.

В статье [**mikolov2013efficient**] описывается метод оценки полученных векторов путем проведения алгебраических операций, то есть поиска аналогий для векторов. Пример:

$$\upsilon_{\text{царь}} - \upsilon_{\text{мальчик}} + \upsilon_{\text{девочка}} \approx \upsilon_{\text{царица}} \tag{1.4}$$

Известно, что для каждых 4 слов, первое находится в таком же семантическом отношении, как и третье с четвёртым (пример отношения: ягода - растение). Но даже в этом случае необходимо, чтобы ассесоры разметили данные и тогда можно было бы оценить качество. По этой причине придется оценивать релевантность рекомендаций путем перекрестного сравнения с уже имеющийся системой рекомендаций на веб-ресурсе.

ГЛАВА 2. ОБЗОР ЛИТЕРАТУРЫ

Исследуемая модель является Content-Based моделью рекомендаций, так как полученные вектора помимо своих характеристик хранят в себе и информацию о семантическом отношении в сессии пользователя. Поэтому в данном параграфе будут рассматриваться популярные модели получения векторов и методы рекомендации основанные на модели.

2.1. Модели получения векторного представления слова

2.1.1. Word2Vec (SGNS)

В основе работы данного алгоритма лежит идея о моделирование условного распределения вероятностей соседних слов. Также стоит отметить, что в отличие от других моделей дистрибутивной семантики (GloVe), Word2Vec работает с последовательностью слов, находящиеся от центрального слова на заданном расстояние - ширина окна.
В рассматриваемой модели хранятся и настраиваются два вектора для каждого слова. Первый вектор - является центральным представлением слова в рассматриваемом окне. Второй вектор - является контекстным представлением слова.
Для поиска оптимума в пространстве параметров данной модели используется градиентный спуск.

Skip Gram - предсказываем соседние слова по центральному слову[**word2vec**]:

$$W, D \in \mathbb{R}^{Vocab \times EmbSize} \sum_{CenterW_i} P(CtxW_{-2}, CtxW_{-1}, CtxW_{+1}, CtxW_{+2} | CenterW_i; W, D) \rightarrow \tag{2.1}$$

Стоит отметить, что сумма в вышеописанном выражение идет не по всем уникальным словам, а по всем возможным словоупотреблениям.

Мы предполагаем, что соседние слова условно независимы друг от друга, когда мы уже рассмотрели центральное слово.

$$P(CtxW_{-2}, CtxW_{-1}, CtxW_{+1}, CtxW_{+2}|CenterW_i; W, D) = \prod_j P(CtxW_j|CenterW_i; W, D) \tag{2.2}$$

Тогда наше распределение можно представить в виде произведения более простых распределений.

Каждое такое более простое распределение, будем моделировать при помощи softmax.

$$P(CtxW_j|CenterW_i; W, D) = \frac{e^{w_i \cdot d_j}}{\sum_{j=1}^{|V|} e^{w_i \cdot d_j}} = softmax \tag{2.3}$$

Из-за наличия в знаменателе суммы по всем объектам нашей выборки, каждый шаг градиентного спуска обходится вычислительно затратно.

Поэтому будем использовать аппроксимацию negative sampling (отрицательное сэмплирование)

$$P(CtxW_j|CenterW_i; W, D) \simeq \frac{e^{w_i \cdot d_j^+}}{\sum_{j=1}^{k} e^{w_i \cdot d_j^-}}, \quad k \ll |V| \tag{2.4}$$

Суть данной аппроксимации заключается в том, что мы будем считать скалярное произведение в знаменателе не по всей нашей выборки объектов, а лишь по некоторым случайно выбранным.

2.1.2. TF-IDF

Данный подкласс моделей еще называется "мешком слов". Главная идея таких алгоритмов это то, что тематика текста хорошо описывается не порядком слов в документе, а составом лексикона и частотой встречаемости слов.

Тогда каждый документ описывается разреженным вектором. Для того, чтобы модель адекватно описывала данные необходимо, чтобы у каждого слова был свой вес. Одним из методов подсчета веса слова и является метод TF-IDF.

Основная идея в том, что чем чаще слово встречается в документе, тем более характерно оно для этого документа. С другой стороны чем реже встречается слово в выборке документов, тем оно более специфично и информативно.

TF - term frequency - значимость слова в рамках документа [**tfidf**]

$$TF(w, d) = \frac{WordCount(w, d)}{Length(d)} \tag{2.5}$$

где $WordCount(w, d)$ - количество употреблений слова w в документе d, $Length(d)$ - длина документа d в словах.

IDF - inverse document frequency - специфичность слова [**tfidf**]

$$IDF(w, c) = \frac{Size(c)}{DocCount(w, c)} \tag{2.6}$$

где $DocCount(w, c)$ - количество документов в коллекции c, в которых встречается слово w,

$Size(c)$ - размер коллекции c в документах.

Тогда вес слова подсчитывается следующим образом:

$$TFIDF(w, d, c) = TF(w, d) * IDF(w, c) \tag{2.7}$$

2.2. Модель рекомендаций

Главная цель моделей рекомендаций - это моделирование отношения между поведением пользователя и товарами или услугами предоставляемыми сервисами.

Отношение между пользователем и объектами можно приблизить некоторыми числами, которые описывают параметры пользователя и параметры объектов. Таким образом образуются векторы в пространстве одной и той же размерности, при этом потребовав, чтобы скалярное произведение вектора, описывающего пользователя, и вектора, описывающий объект, хорошо приближала оценку отношений.

$$x_{ij} \approx \langle u_i, \nu_j \rangle \tag{2.8}$$

u_i - параметры пользователя

ν_j - параметры объектов

Таким образом мы перешли к оптимизационной задаче:

$$\sum (\langle u_i, \nu_j \rangle - x_{ij})^2 \rightarrow min \tag{2.9}$$

2.2.1. SVD (Singular Value Decomposition)

Пусть дана матрица пользователи - объекты, на пересечении которых стоят оценки пользователей.

Данная матрица имеет огромный размер (количество пользователей интернет ресурса может достигать нескольких миллионов, как и количество объектов предоставляемых веб-ресурсом)

Для любой вещественной $(n \times n)$ – матрицы существуют две вещественные ортогональные матрицы U и V такие, что

$$U^T AV = \Lambda \tag{2.10}$$

[**svd**]

Используя SVD-разложение матрицы пользователи - объекты, мы получим 2 матрицы: $U\ n \times k$ и $V\ m \times k$, где n - число пользователей, m - число объектов, k - набор факторов. Данные факторы и являются характеристикой вкусов и предпочтений пользователей.

ГЛАВА 3. ПРЕДООБРАБОТКА И АНАЛИЗ ДАННЫХ

В данной главе будет описана обработка реальных данных, и методы борьбы с пропусками.

3.1. Исходные данные

Для проведения экспериментов были предоставлены обезличенные данные о поведение пользователей на интернет-ресурсе «Яндекс.Недвижимость». Также была представлена вся информация об имеющийся на определённый момент времени объявлениях недвижимости.
За представленный период 1.3 миллиона уникальных пользователей совершили 17 миллионов «кликов» (Клик – это взаимодействие пользователя с объектом интернет-ресурса, где объекты представляют из себя объявления о продаже/аренде недвижимости) по 0.4 миллионам объектам.

3.2. Анализ данных

Для исследования были предоставлены табулированные данные.

A. Данная таблица несет в себе информацию об истории активности пользователей на интернет-ресурсе, в частности их клики на объявления. Эта информация содержала в себе, уникальный идентификатор пользователя, уникальный идентификатор объявления, с которым взаимодействовал пользователь, а также временная отметка данного действия.

	offer_id	user_id	event_time
0	4803055886953936205	1	4290716
1	4750707305963955879	1	4290735
2	6504772239483419419	1	4290786
3	6504772239427956238	1	4290898
4	4833302310798700801	1	4291745

Рис.3.1. Структура таблицы истории пользователей

B. Также была предоставлена таблица в виде матрицы объект-признак. В ней перечисленны все характеристики объявлений за определенный промежуток времени. Она содержит в себе 380036 уникальных объявлений, описываемых 36 признаками

	agent_fee	category_type	create_time_millis	first_met_date_millis	flat_type	floors	floors_total	geo_text	is_grandmother_renovation	is_internal	...
100009511263526656	NaN	APARTMENT	1519027563392	1577375640356	SECONDARY	3.0	5.0	Россия, Санкт-Петербург, проспект Юрия Гагарин...	None	True	...
1000163439824483072	NaN	APARTMENT	1580723624187	1580723624187	UNKNOWN	18.0	23.0	Россия, Санкт-Петербург, Русановская улица, 16к3	False	True	...
1000192056170345729	NaN	LOT	1559060128980	1559060128980	UNKNOWN	NaN	NaN	Россия, Ленинградская область, Гатчинский райо...	False	True	...
1000270208695790848	NaN	APARTMENT	1576171290432	1576171290432	SECONDARY	5.0	12.0	Россия, Санкт-Петербург, улица Кораблестроител...	True	True	...
1000284101403898880	50.0	APARTMENT	1569565896483	1577014613724	UNKNOWN	5.0	12.0	Россия, Санкт-Петербург, проспект Королёва, 49	True	True	...

Рис.3.2. Информация об объявлениях

В практических задачах, реальные данные не очень хорошие, так как они сильно разреженны. С этой проблемой необходимо бороться.

Обработка пропусков в данных - это отдельная обширная область в анализе данных.

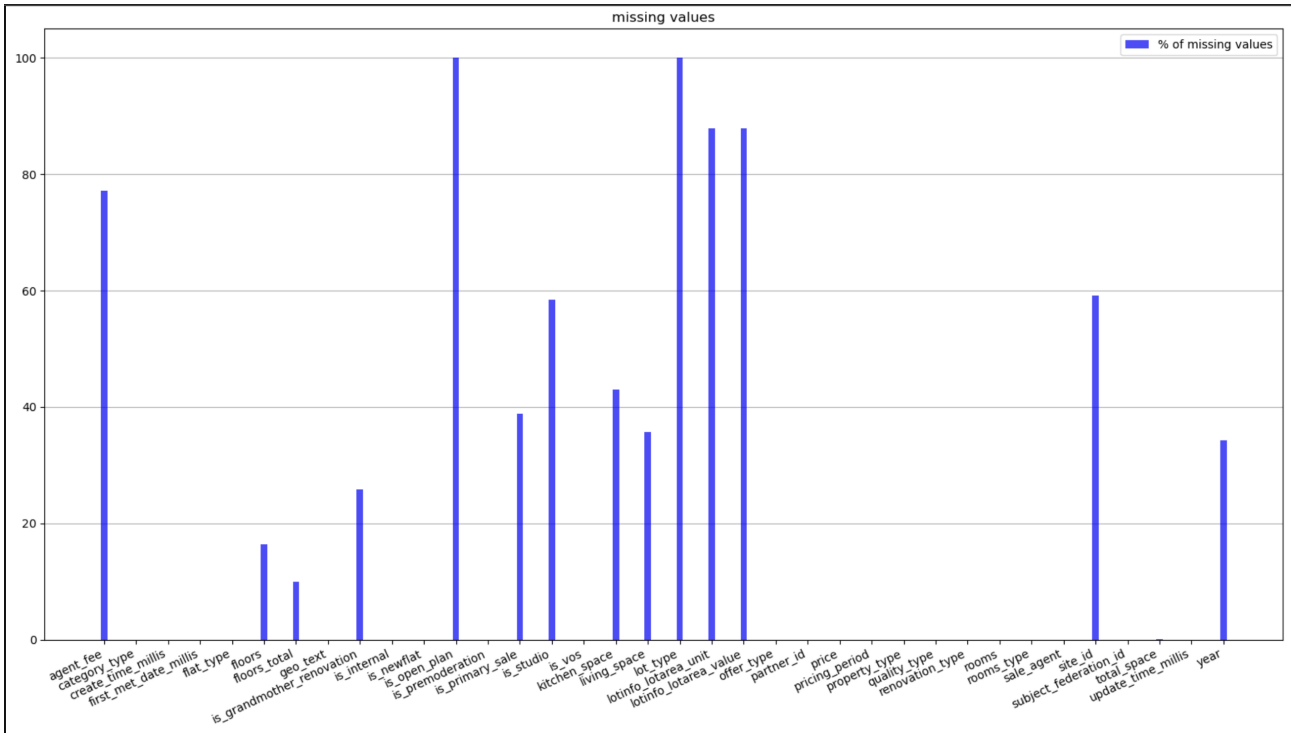

Рис.3.3. Процентное соотношения пропусков

Существует множество методов, которые позволяют бороться с пропусками в данных, одним из них является построение решающих функций, которые будут предсказывать пропущенные значения. Но данное решение весьма трудоемко и не вкладывается в общий ход решения поставленной задачи. Поэтому приходилось каждую характеристику объявления обрабатывать отдельно.

Так например если не указана стоимость комиссия агента по продажам, то мы заполняли данный пропуск, как отсутствие комиссии. Также характеристику о том, что данное объявление выставляется впервые, в случае пропуска заполнялось как

истина.

Остальные пропуски учитывались как самостоятельный элемент, эта эвристика выходит из принципа максимального правдоподобия. Также одной из важных подзадач в анализе данных было определение оптимальной максимальной длинны пользовательской сессии. Этот фактор имел несколько предпосылок:

A. Работа алгоритма word2vec на длинных сессиях было бы вычислительна затратная
B. Длинные сессии могли быть сгенерированы ботами, которые обрабатывают интернет-ресурс, их активность не имеет интереса для нашей задачи.

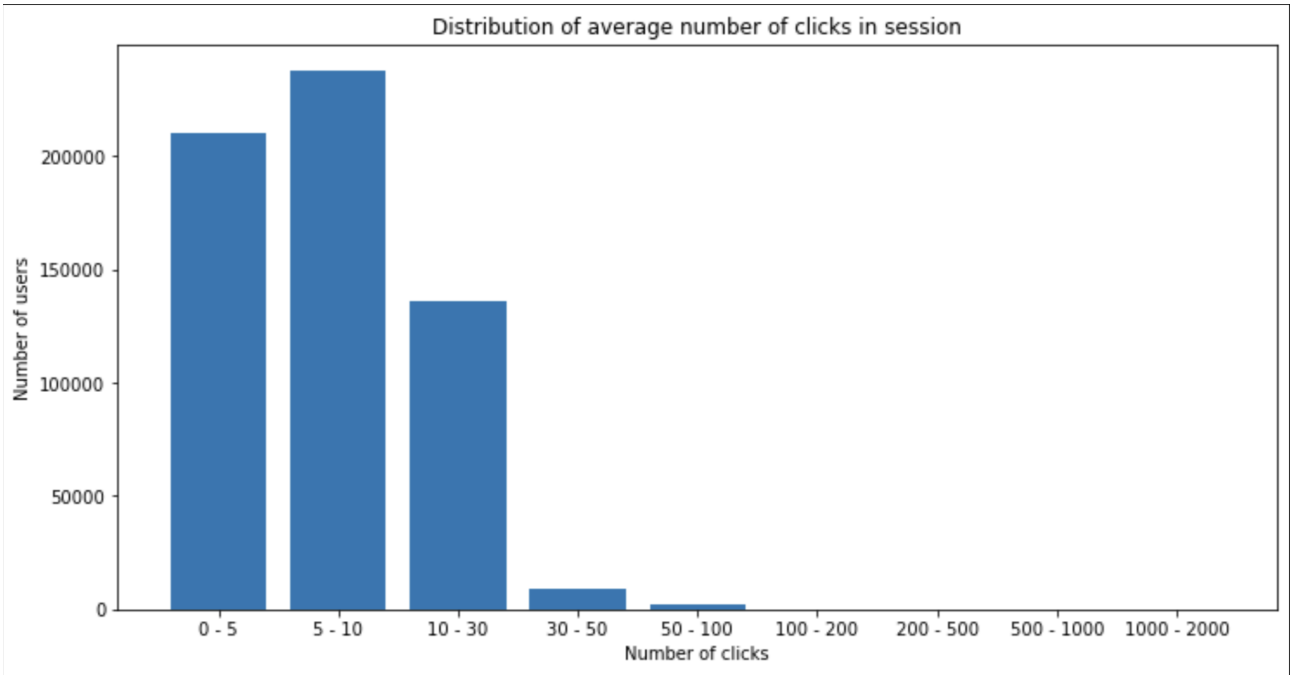

Рис.3.4. Распределение среднего числа кликов пользователей в сессии

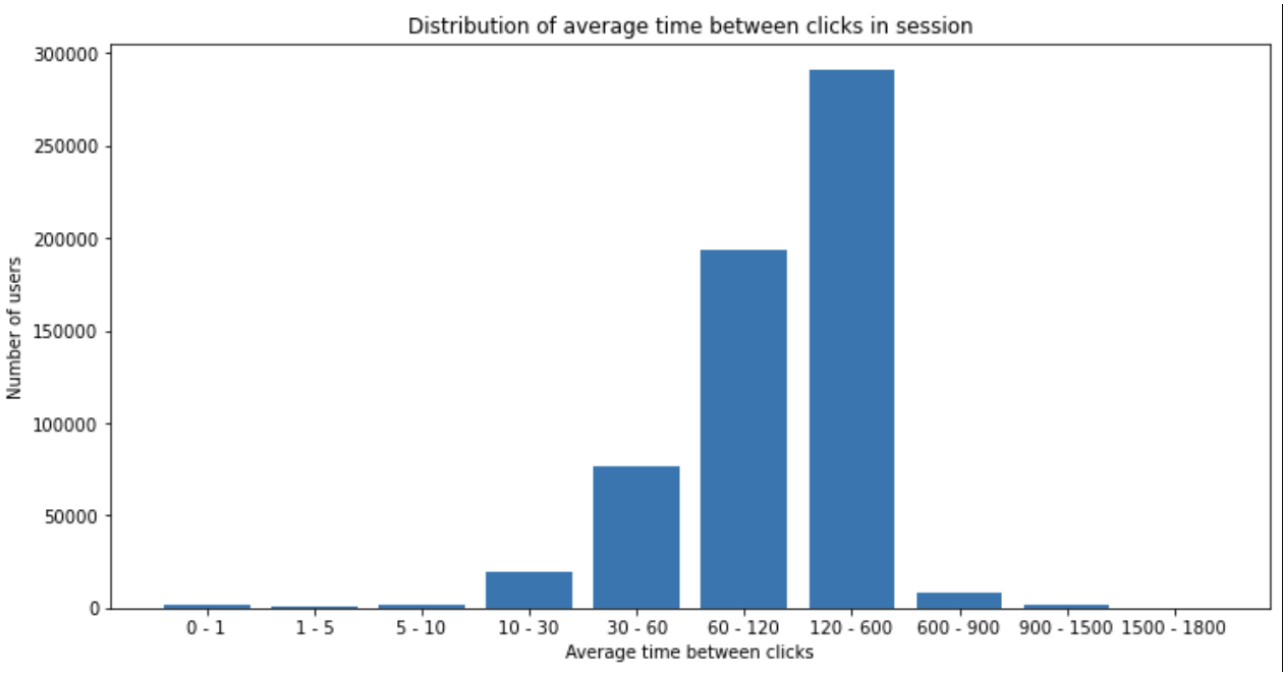

Рис.3.5. Распределение средней разности времени между кликами пользователей в сессии

ГЛАВА 4. МОДЕЛИ ИССЛЕДОВАНИЯ И РЕЗУЛЬТАТЫ

В данной главе будут рассмотрены исследуемые модели машинного обучения. Результаты их работы и сравнение рекомендаций полученных в результате работы рассмотренных моделей.

4.1. Предообучение признаков объекта

В данной модели моделируется условное распределение вероятностей признаков, находящихся в соседних объектах пользовательских сессий.

Признаки, имеющие шкалу абсолютных величин, разбивались на перцентили, данная эвристика исходила из того, что в задачах рекомендации пользователи в общей совокупности делятся на некоторые подмножества, которые обуславливаются общими характеристиками по некоторым признакам. То есть множества сущностей всех признаков мы разбивали на категории.

Каждый признак имел векторное представление размером $min(n/2 + 1, 50)$, n - мощность множества сущностей признака объекта. Данная эмпирическая закономерность была выведена путем поиска по сетке, описанная в [**sizeEmbedding**].

После чего полученные вектора сущностей признаков конкатенировались в соответствие с матрицей объект-признак. Тем самым мы получем векторное представление для каждого объекта.

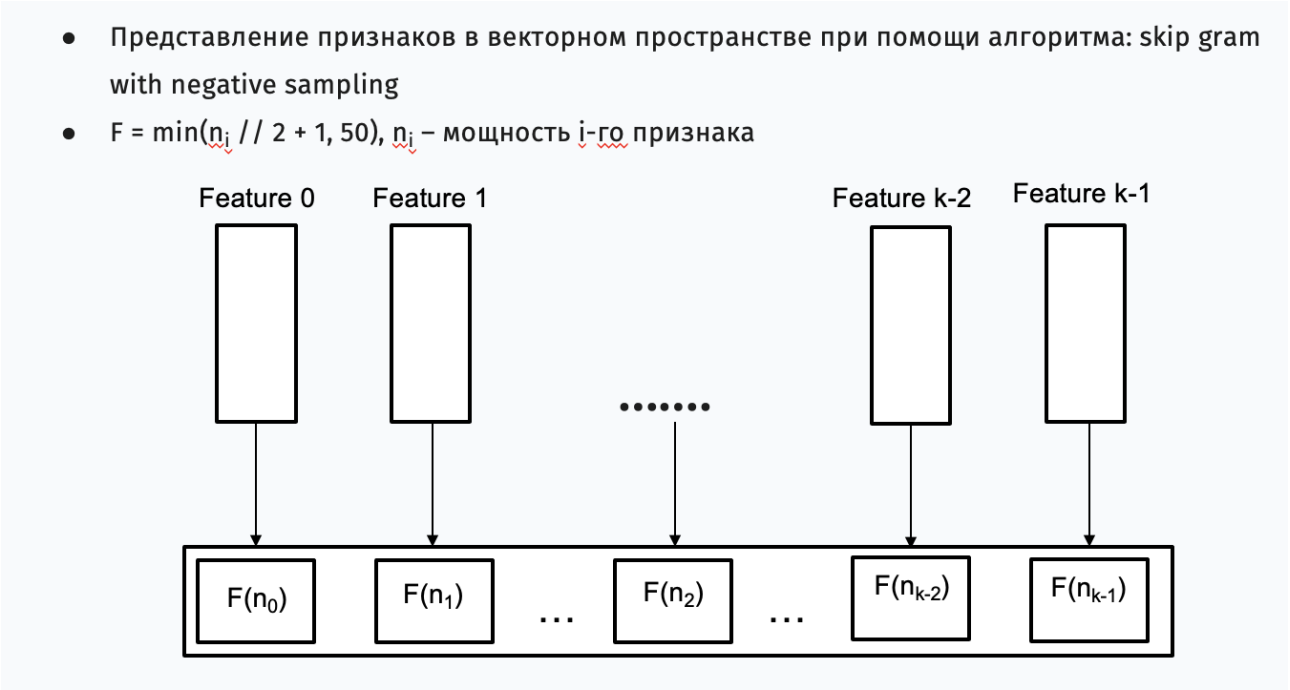

Рис.4.1. Архитектура модели предобучения признаков

Полученные вектора признаков будут использоваться как новые признаки для следующих моделей.

График функции потерь при обучении модели для 4 признаков:

4.2. Использование полносвязных слоев

Данная модель по своей структуре является классификатором.

Его идея заключается в том, чтобы найти взаимосвязь между предобученными признаками объекта и представить в более меньшей размерности.

4.2.1. Метод обучения

Для обучения модели необходимо выбрать пару из нашей выборки, которая входит в сессию пользователя. К данной паре необходимо подобрать некоторое количество негативных примеров. Это случайные объекты из нашей выборки.

Затем прогоняем через сеть пару объектов, а также негативные примеры. После чего считаем скалярное произведение между парой и негативными элементами. Выбираем 100 элементов с самым большим значением скалярного произведения из множества негативных примеров. И считаем функцию потерь - cross entropy loss.

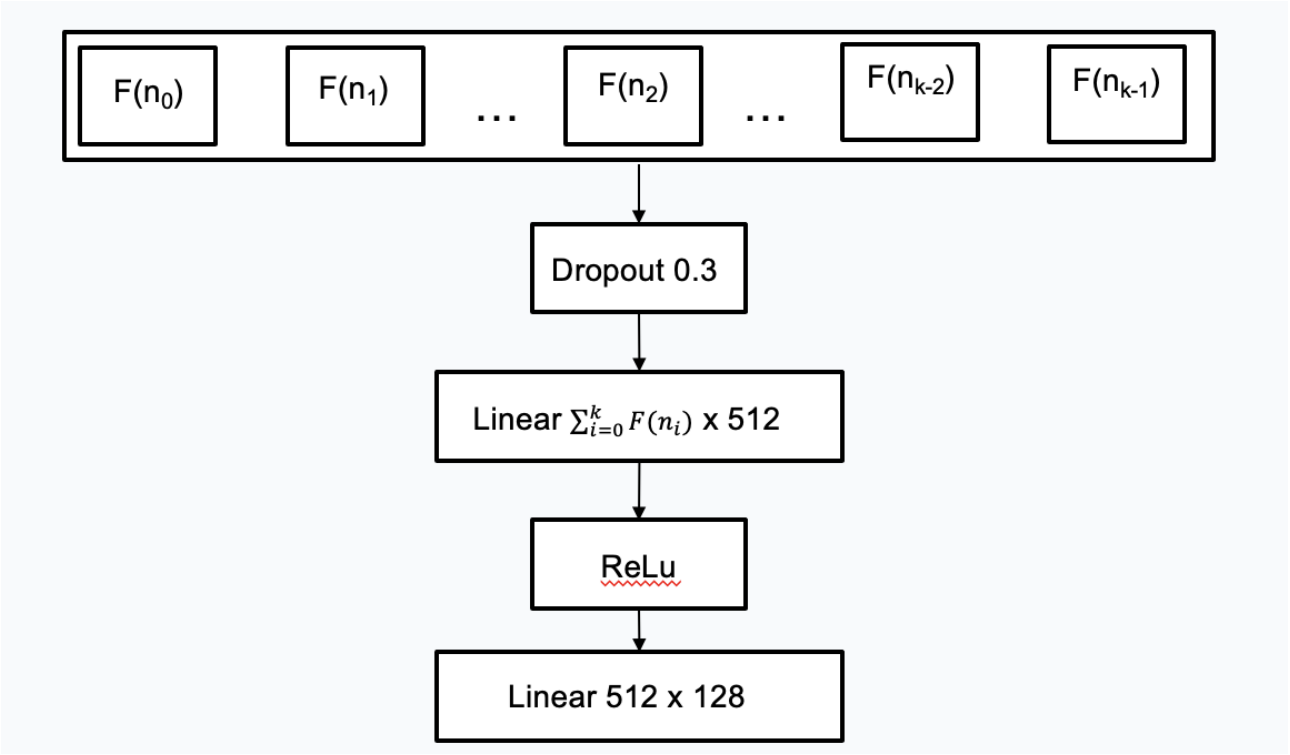

Рис.4.2. Архитектура нейронной сети

$$loss(x, class) = -\log(\frac{\exp(x[class])}{\sum(\exp(x[j]))}) = -x[class] + \log(\sum(exp(x[j]))) \quad (4.1)$$

4.3. Модифицированная модель с полносвязными слоями

В данной модели усовершенствование работы алгоритма заключается из того, что для полученных вектора уже хранят в себе семантический смысл, поэтому для каждой сессии пользователя, используются алгебраические операции над векторами объектов. То есть теперь данными являются не пара похожих объектов, а пара: среднее векторов между последовательностью объектов входящих в одну сессию и следующим за ними объектом из этой же сессии.

4.4. Ансамблирование алгоритмов обучения

В данной модели моделируется условное распределение вероятностей объектов, находящихся в соседних объектах пользовательских сессий. При этом имея в наличии уже предобученные вектора признаков.

Данное распределение моделировалось на таком же алгоритме word2vec, как и для обучения признаков объектов. Отличие заключается лишь в том, что изна-

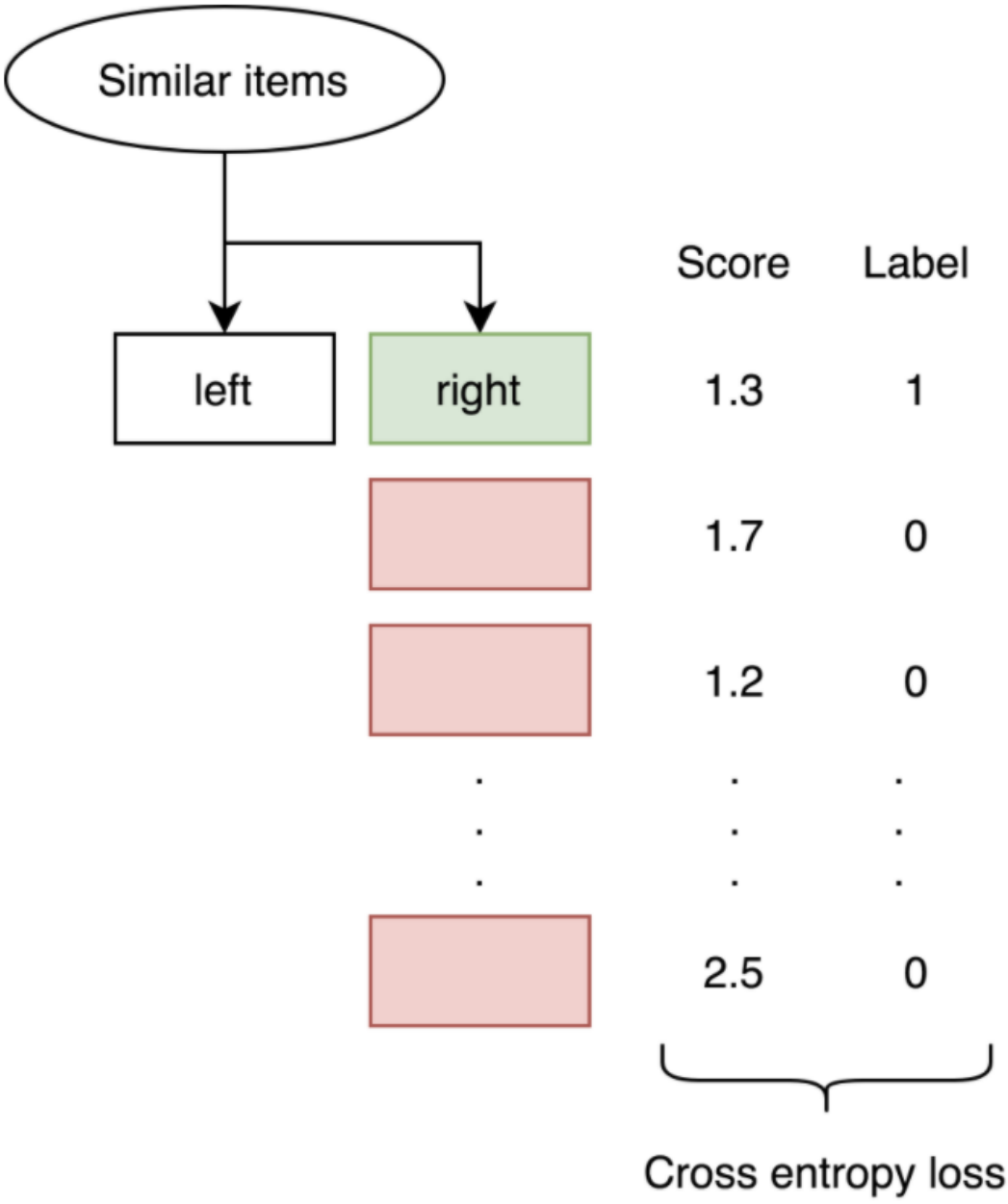

Рис.4.3. Структура тренировки модели

чальные вектора были получены в результате конкатенации векторов-признаков, а не инициализации случайным шумом.

4.5. Выбор метрики

В результате применение вышеописанного алгоритма, мы сможем получить векторное представление объектов более низкой размерности, при этом данный алгоритм смог сохранить семантику объектов в векторном представление. Эта семантика выражается через отношение близости объектов. Наиболее схожие объекты между собой находятся близко в построенном векторном пространстве.

Эта идея является ключевой для построения рекомендательной системы.
В статье [**dist**] описываются сравниваются различные оценки близости такие как: евклидово расстояние, косинусная близость, метрика Манхэттена, расстояние Бхаттачарья, расстояние Хеллингера, дивергенция Кульбака-Лейблера. В результате экспериментов косинусная близость показала наилучший результат.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

4.6. Результаты моделей

4.6.1. Графики функции потерь

A. Векторное представление признаков объекта.

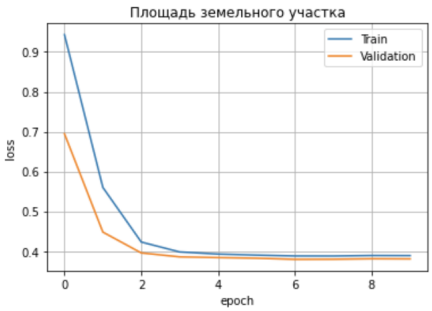

Рис.4.4. Функция потерь для площади земельного участка

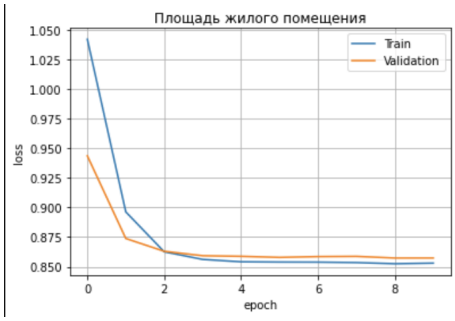

Рис.4.5. Функция потерь для жилплощади

Можно увидеть, что обучение на всех признаков достигает некоторого оптимума.

B. Модель с полносвязными слоями
C. Модифицированная модель с полносвязными слоями

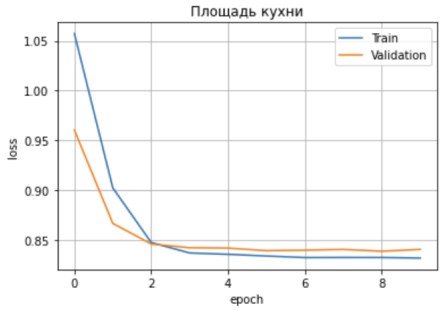

Рис.4.6. Функция потерь для площади кухни

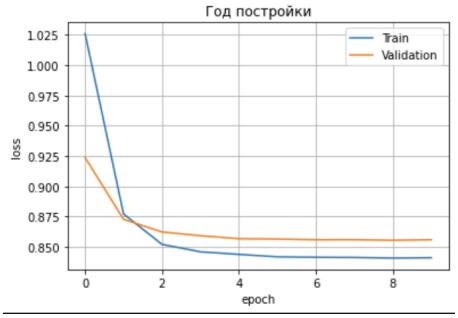

Рис.4.7. Функция потерь для года постройки

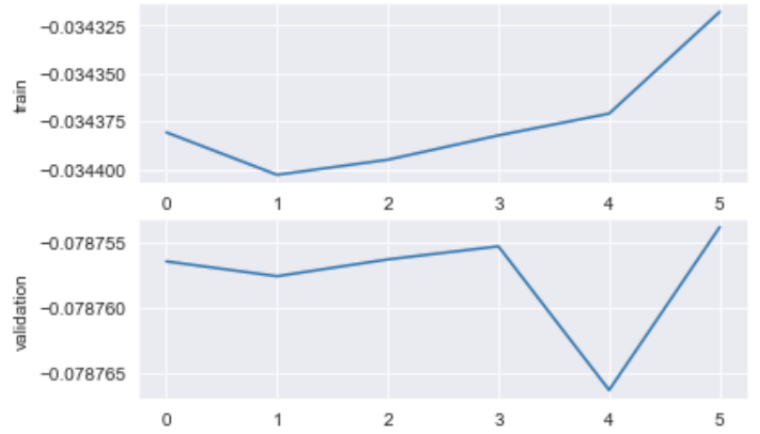

Рис.4.8. Функция потерь модели с полносвязными слоями

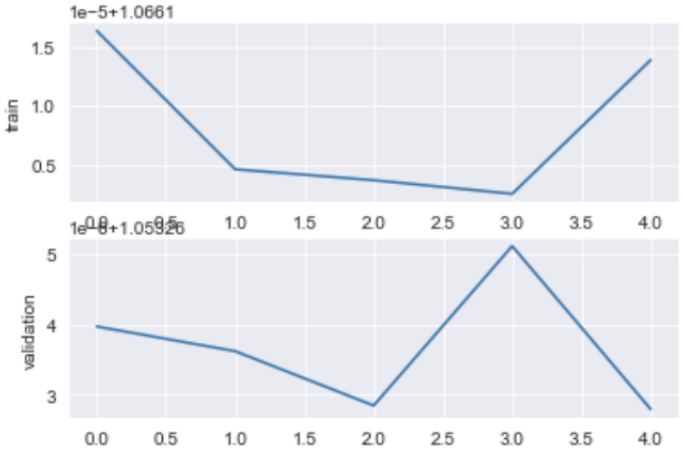

Рис.4.9. Функция потерь модели с полносвязными слоями

D. Ансамблирование алгоритмов обучения

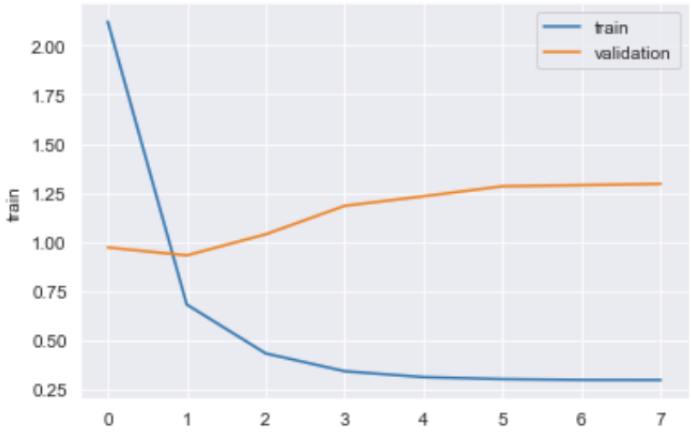

Рис.4.10. Функция потерь для ансамбля моделей

Как можно заметить модель быстро переобучается и теряет "знания"о предобученных признаках.

4.6.2. Поиск похожих объектов

В данном исследовании будет рассмотрено, как алгоритм хорошо представляет объекты в векторном пространстве, то есть располагает похожие объекты близко. Для этого будем строить рекомендации для некоторых случайно выбранных объектов, которые относится к разным категориям, таким как:

- Продажа квартиры
- Аренда квартиры
- Аренда дома

parameter	100493917162964737	4265194567116977665	8252048366830426881	1670582614124560897	4927840560490826753	2685624910493028608
agent_fee	0	0	0	0	0	0
category_type	APARTMENT	APARTMENT	APARTMENT	APARTMENT	APARTMENT	APARTMENT
flat_type	NEW_SECONDARY	NEW_SECONDARY	NEW_SECONDARY	NEW_SECONDARY	NEW_SECONDARY	NEW_SECONDARY
floors	2	3	8	1	10	3
floors_total	18	18	18	12	18	18
is_grandmother_renovation	False	False	False	False	False	False
is_primary_sale	False	False	False	False	False	False
is_studio	True	True	True	True	True	True
is_vos	True	True	True	True	True	True
kitchen_space	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
living_space	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
lotinfo_lotarea_value	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
offer_type	True	True	True	True	True	True
price	2200000	2600000	2490000	2600000	2690000	2650000
pricing_period	WHOLE_LIFE	WHOLE_LIFE	WHOLE_LIFE	WHOLE_LIFE	WHOLE_LIFE	WHOLE_LIFE
quality_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
renovation_type	EURO	EURO	EURO	COSMETIC_DONE	EURO	EURO
rooms	0	0	0	0	0	0
sale_agent	OWNER	OWNER	OWNER	OWNER	PRIVATE_AGENT	PRIVATE_AGENT
total_space	24	24	26	22	22	25
year	2017	2019	2018	2017	2017	2018
distance	0.000	1.726	1.726	0.000	1.221	1.726

Рис.4.11. Рекомендация векторного представления признаков для продажи квартиры

parameter	7971252646849736960	4651016148161830400	5873663946069185536	3476717717078956288	8853842806943915009	2783043973510204928
agent_fee	0	0	0	0	0	0
category_type	COMMERCIAL	COMMERCIAL	COMMERCIAL	COMMERCIAL	COMMERCIAL	COMMERCIAL
flat_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
floors	3	5	5	3	2	2
floors_total	4	6	6	3	6	3
is_grandmother_renovation	False	False	False	False	False	False
is_primary_sale	True	True	True	True	True	True
is_studio	True	True	True	True	True	True
is_vos	True	True	True	True	True	True
kitchen_space	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
living_space	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
lotinfo_lotarea_value	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
offer_type	False	False	False	False	False	False
price	24000	22500	22500	25200	23400	30000
pricing_period	PER_MONTH	PER_MONTH	PER_MONTH	PER_MONTH	PER_MONTH	PER_MONTH
quality_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
renovation_type	COSMETIC_DONE	COSMETIC_DONE	COSMETIC_DONE	UNKNOWN	EURO	COSMETIC_DONE
rooms	1	1	1	1	1	1
sale_agent	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT
total_space	24	30	30	25	18	25
year	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
distance	0.000	0.000	0.000	0.000	0.000	0.000

Рис.4.12. Рекомендация векторного представления признаков для аренды квартиры

parameter	8877539465255334401	4773698031729608193	2012428460346647808	2440192171730253312	4516451544737701889	9168731149268184064
agent_fee	0	0	0	0	0	0
category_type	HOUSE	HOUSE	HOUSE	HOUSE	HOUSE	HOUSE
flat_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
floors	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
floors_total	2	2	1	1	1	1
is_grandmother_renovation	False	False	False	False	False	False
is_primary_sale	True	True	True	True	True	True
is_studio	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
is_vos	True	True	True	True	True	True
kitchen_space	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
living_space	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
lotinfo_lotarea_value	<NA>	<NA>	800	1200	1200	400
offer_type	False	False	False	False	False	False
price	1200	4500	1900	1750	1200	2000
pricing_period	PER_DAY	PER_DAY	PER_DAY	PER_DAY	PER_DAY	PER_DAY
quality_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
renovation_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
rooms	0	0	0	0	0	0
sale_agent	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT
total_space	35	50	40	45	30	45
year	1980	<NA>	<NA>	<NA>	<NA>	1917
distance	0.000	47.871	62.511	62.511	63.152	205.327

Рис.4.13. Рекомендация векторного представления признаков для аренды загородного дома

parameter	100493917162964737	4265194567116977665	5732924502375163136	8252048366830426881	6555443516320912640	6075245557494210817
agent_fee	0	0	0	0	0	0
category_type	APARTMENT	APARTMENT	APARTMENT	APARTMENT	APARTMENT	APARTMENT
flat_type	NEW_SECONDARY	NEW_SECONDARY	NEW_SECONDARY	NEW_SECONDARY	NEW_SECONDARY	NEW_SECONDARY
floors	2	3	2	8	16	6
floors_total	18	18	18	18	18	16
is_grandmother_renovation	False	False	True	False	False	False
is_primary_sale	False	False	False	False	False	False
is_studio	True	True	True	True	True	True
is_vos	True	True	True	True	True	True
kitchen_space	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
living_space	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
lotinfo_lotarea_value	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
offer_type	True	True	True	True	True	True
price	2200000	2600000	2300000	2490000	2680000	3050000
pricing_period	WHOLE_LIFE	WHOLE_LIFE	WHOLE_LIFE	WHOLE_LIFE	WHOLE_LIFE	WHOLE_LIFE
quality_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
renovation_type	EURO	EURO	COSMETIC_DONE	EURO	EURO	EURO
rooms	0	0	0	0	0	0
sale_agent	OWNER	OWNER	AGENCY	OWNER	PRIVATE_AGENT	OWNER
total_space	24	24	23	26	24	24
year	2017	2019	2017	2018	2016	2017
distance	0.000	1.726	0.000	1.726	1.221	1.221

Рис.4.14. Рекомендация модели с полносвязными слоями для продажи квартиры

parameter	7971252646849736960	4651016148161830400	5873663946069185536	3476717717078956288	8853842806943915009	8763171408389065984
agent_fee	0	0	0	0	0	0
category_type	COMMERCIAL	COMMERCIAL	COMMERCIAL	COMMERCIAL	COMMERCIAL	COMMERCIAL
flat_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
floors	3	5	5	3	2	4
floors_total	4	6	6	3	6	6
is_grandmother_renovation	False	False	False	False	False	False
is_primary_sale	True	True	True	True	True	True
is_studio	True	True	True	True	True	True
is_vos	True	True	True	True	True	True
kitchen_space	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
living_space	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
lotinfo_lotarea_value	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
offer_type	False	False	False	False	False	False
price	24000	22500	22500	25200	23400	27750
pricing_period	PER_MONTH	PER_MONTH	PER_MONTH	PER_MONTH	PER_MONTH	PER_MONTH
quality_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
renovation_type	COSMETIC_DONE	COSMETIC_DONE	COSMETIC_DONE	UNKNOWN	EURO	COSMETIC_DONE
rooms	1	1	1	1	1	1
sale_agent	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT
total_space	24	30	30	25	18	37
year	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
distance	0.000	0.000	0.000	0.000	0.000	0.000

Рис.4.15. Рекомендация модели с полносвязными слоями для аренды квартиры

parameter	8877539465255334401	2012428460346647808	4773698031729608193	6728846443168103169	5608653784653049345	8693575741440514560
agent_fee	0	0	0	0	0	0
category_type	HOUSE	HOUSE	HOUSE	APARTMENT	HOUSE	HOUSE
flat_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
floors	<NA>	<NA>	<NA>	4	<NA>	<NA>
floors_total	2	1	2	5	2	2
is_grandmother_renovation	False	False	False	False	False	False
is_primary_sale	True	True	True	True	True	True
is_studio	<NA>	<NA>	<NA>	False	<NA>	<NA>
is_vos	True	True	True	True	True	True
kitchen_space	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
living_space	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
lotinfo_lotarea_value	<NA>	800	<NA>	<NA>	<NA>	<NA>
offer_type	False	False	False	False	False	False
price	1200	1900	4500	1800	6500	12046
pricing_period	PER_DAY	PER_DAY	PER_DAY	PER_DAY	PER_DAY	PER_DAY
quality_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
renovation_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
rooms	0	0	0	1	0	0
sale_agent	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT	OWNER	AGENT
total_space	35	40	50	40	80	200
year	1980	<NA>	<NA>	1979	1954	1985
distance	0.000	62.511	47.871	81.765	102.801	21.367

Рис.4.16. Рекомендация модели с полносвязными слоями для аренды загородного дома

parameter	100493917162964737	4265194567116977665	8252048366830426881	5188591056744850944	7358685847518509569	5279241030997223168
agent_fee	0	0	0	0	0	0
category_type	APARTMENT	APARTMENT	APARTMENT	APARTMENT	APARTMENT	APARTMENT
flat_type	NEW_SECONDARY	NEW_SECONDARY	NEW_SECONDARY	NEW_SECONDARY	NEW_SECONDARY	NEW_SECONDARY
floors	2	3	8	16	14	2
floors_total	18	18	18	19	17	17
is_grandmother_renovation	False	False	False	False	False	False
is_primary_sale	False	False	False	False	False	False
is_studio	True	True	True	True	True	True
is_vos	True	True	True	True	True	True
kitchen_space	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
living_space	<NA>	<NA>	<NA>	<NA>	<NA>	19
lotinfo_lotarea_value	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
offer_type	True	True	True	True	True	True
price	2200000	2600000	2490000	2530000	2290000	2200000
pricing_period	WHOLE_LIFE	WHOLE_LIFE	WHOLE_LIFE	WHOLE_LIFE	WHOLE_LIFE	WHOLE_LIFE
quality_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
renovation_type	EURO	EURO	EURO	EURO	UNKNOWN	NEEDS_RENOVATION
rooms	0	0	0	0	0	0
sale_agent	OWNER	OWNER	OWNER	PRIVATE_AGENT	PRIVATE_AGENT	OWNER
total_space	24	24	26	24	25	26
year	2017	2019	2018	2019	2017	2019
distance	0.000	1.726	1.726	0.000	1.726	1.221

Рис.4.17. Рекомендация модифицированной модели с полносвязными слоями для продажи квартиры

parameter	7971252646849736960	7933577522989033728	4651016148161830400	5873663946069185536	3476717717078956288	1874015770311020801
agent_fee	0	0	0	0	0	0
category_type	COMMERCIAL	COMMERCIAL	COMMERCIAL	COMMERCIAL	COMMERCIAL	COMMERCIAL
flat_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
floors	3	5	5	5	3	3
floors_total	4	5	6	6	3	4
is_grandmother_renovation	False	False	False	False	False	False
is_primary_sale	True	True	True	True	True	True
is_studio	True	<NA>	True	True	True	True
is_vos	True	True	True	True	True	True
kitchen_space	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
living_space	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
lotinfo_lotarea_value	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
offer_type	False	False	False	False	False	False
price	24000	25600	22500	22500	25200	19990
pricing_period	PER_MONTH	PER_MONTH	PER_MONTH	PER_MONTH	PER_MONTH	PER_MONTH
quality_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
renovation_type	COSMETIC_DONE	COSMETIC_DONE	COSMETIC_DONE	COSMETIC_DONE	UNKNOWN	COSMETIC_DONE
rooms	1	0	1	1	1	1
sale_agent	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT	AGENT
total_space	24	25	30	30	25	30
year	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
distance	0.000	1.224	0.000	0.000	0.000	1.733

Рис.4.18. Рекомендация модифицированной модели с полносвязными слоями для аренды квартиры

parameter	8877539465255334401	4773698031729608193	9168731149268184064	2440192171730253312	2012428460346647808	5220037825451960064
agent_fee	0	0	0	0	0	0
category_type	HOUSE	HOUSE	HOUSE	HOUSE	HOUSE	HOUSE
flat_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
floors	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
floors_total	2	2	1	1	1	1
is_grandmother_renovation	False	False	False	False	False	False
is_primary_sale	True	True	True	True	True	True
is_studio	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
is_vos	True	True	True	True	True	True
kitchen_space	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
living_space	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
lotinfo_lotarea_value	<NA>	<NA>	400	1200	800	<NA>
offer_type	False	False	False	False	False	False
price	1200	4500	2000	1750	1900	4000
pricing_period	PER_DAY	PER_DAY	PER_DAY	PER_DAY	PER_DAY	PER_DAY
quality_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
renovation_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
rooms	0	0	0	0	0	0
sale_agent	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT	PRIVATE_AGENT	AGENCY
total_space	35	50	45	45	40	80
year	1980	<NA>	1917	<NA>	<NA>	<NA>
distance	0.000	47.871	205.327	62.511	62.511	72.779

Рис.4.19. Рекомендация модифицированной модели с полносвязными слоями для аренды загородного дома

parameter	100493917162964737	456482203167953152	1428604561695214592	1383660821553482496	2801590516726016512	1527768702518715393
agent_fee	0	70	0	0	0	0
category_type	APARTMENT	APARTMENT	APARTMENT	COMMERCIAL	APARTMENT	APARTMENT
flat_type	NEW_SECONDARY	UNKNOWN	UNKNOWN	UNKNOWN	SECONDARY	NEW_SECONDARY
floors	2	23	19	1	11	1
floors_total	18	25	22	26	17	7
is_grandmother_renovation	False	False	False	False	False	False
is_primary_sale	False	True	True	True	False	False
is_studio	True	False	False	False	False	False
is_vos	True	True	True	True	True	True
kitchen_space	<NA>	14	25	<NA>	10	10
living_space	<NA>	<NA>	<NA>	<NA>	22	17
lotinfo_lotarea_value	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
offer_type	True	False	False	True	True	True
price	2200000	60000	39000	33375000	4925000	3845640
pricing_period	WHOLE_LIFE	PER_MONTH	PER_MONTH	WHOLE_LIFE	WHOLE_LIFE	WHOLE_LIFE
quality_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
renovation_type	EURO	DESIGNER_RENOVATION	EURO	NEEDS_RENOVATION	COSMETIC_DONE	NEEDS_RENOVATION
rooms	0	2	3	4	1	1
sale_agent	OWNER	PRIVATE_AGENT	PRIVATE_AGENT	AGENCY	OWNER	AGENCY
total_space	24	70	72	222	48	43
year	2017	2007	2005	<NA>	2008	2019
distance	0.000	13.212	25.757	1.726	10.919	54.079

Рис.4.20. Рекомендация ансамбля моделей для продажи квартиры

parameter	7971252646849736960	2287670986313528832	7056913162025300224	5667574266998424833	3167052337943843584	9166458779125870593
agent_fee	0	50	0	50	0	0
category_type	COMMERCIAL	ROOMS	APARTMENT	ROOMS	APARTMENT	ROOMS
flat_type	UNKNOWN	UNKNOWN	SECONDARY	UNKNOWN	SECONDARY	UNKNOWN
floors	3	5	2	16	2	3
floors_total	4	10	5	16	4	3
is_grandmother_renovation	False	False	False	True	True	False
is_primary_sale	True	True	False	True	False	True
is_studio	True	False	False	False	True	False
is_vos	True	True	True	True	True	True
kitchen_space	<NA>	9	17	11	<NA>	20
living_space	<NA>	20	80	12	<NA>	16
lotinfo_lotarea_value	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
offer_type	False	False	True	False	True	True
price	24000	13000	12300000	10000	2950000	1050000
pricing_period	PER_MONTH	PER_MONTH	WHOLE_LIFE	PER_MONTH	WHOLE_LIFE	WHOLE_LIFE
quality_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
renovation_type	COSMETIC_DONE	COSMETIC_DONE	COSMETIC_DONE	EURO	UNKNOWN	COSMETIC_DONE
rooms	1	2	4	2	0	7
sale_agent	PRIVATE_AGENT	PRIVATE_AGENT	OWNER	PRIVATE_AGENT	AGENCY	PRIVATE_AGENT
total_space	24	50	135	60	15	100
year	<NA>	1998	<NA>	2005	<NA>	1917
distance	0.000	27.641	2.738	8.670	0.000	23.659

Рис.4.21. Рекомендация ансамбля моделей для аренды квартиры

parameter	8877539465255334401	1558680857804242688	8391680986576270081	9187507115879021569	7567893589656635989	7152714844283064577
agent_fee	0	0	0	50	0	0
category_type	HOUSE	APARTMENT	COMMERCIAL	APARTMENT	APARTMENT	APARTMENT
flat_type	UNKNOWN	SECONDARY	UNKNOWN	UNKNOWN	UNKNOWN	SECONDARY
floors	<NA>	7	2	3	4	1
floors_total	2	14	5	9	6	6
is_grandmother_renovation	False	False	False	False	False	False
is_primary_sale	True	False	True	True	True	False
is_studio	<NA>	False	False	False	False	False
is_vos	True	True	True	False	False	True
kitchen_space	<NA>	11	<NA>	12	12	12
living_space	<NA>	18	<NA>	36	101	41
lotinfo_lotarea_value	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>
offer_type	False	True	False	False	False	True
price	1200	6700000	70050	50000	75000	8300000
pricing_period	PER_DAY	WHOLE_LIFE	PER_MONTH	PER_MONTH	PER_MONTH	WHOLE_LIFE
quality_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
renovation_type	UNKNOWN	DESIGNER_RENOVATION	EURO	UNKNOWN	UNKNOWN	EURO
rooms	0	1	2	2	4	2
sale_agent	PRIVATE_AGENT	OWNER	PRIVATE_AGENT	AGENCY	AGENCY	PRIVATE_AGENT
total_space	35	44	46	84	150	63
year	1980	2001	<NA>	2011	1906	1912
distance	0.000	60.827	70.524	61.904	69.585	66.852

Рис.4.22. Рекомендация ансамбля моделей для аренды загородного дома

4.6.3. Кластеризация

В данной секции экспериментов было исследовано выделение кластеров из построенного векторного пространства. Было выбрано первые 1000 пользовательских взаимодействий (объектов), и на векторном представлении данных объектов выполнялся алгоритм кластеризации - K-Means[**likas2003global**] Так как алгоритм кластеризации запускался с фиксированными параметрами, но с разными векторными пространствами, то далее для интерпретации будет использоваться кластер под номером 5.

Чтобы оценивать качество кластеризации при одних и тех же параметров, я использовал Силуэт (англ. Silhouette) [**sil**]

$$Sil() = \frac{1}{N} \sum_{c_k \in C} \sum_{x_i \in c_k} \frac{b(x_i, c_k) - a(x_i, c_k)}{max\{a(x_i, c_k), b(x_i, c_k)\}} \tag{4.2}$$

Чем ближе данная оценка к 1, тем лучше.

Предобученные признаки Элементы кластера являются интерпретируемыми. Можно сказать, что в данный кластер попали те объекты, которые находятся в исторических районах города и имеют ремонт жилища.

$Sil(C) = 0.065$

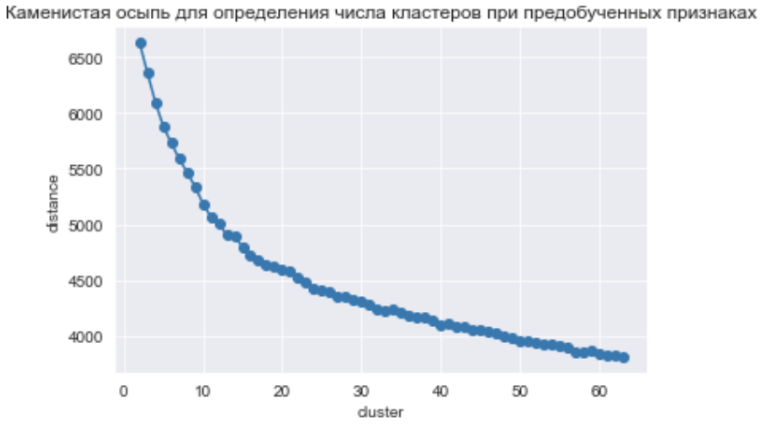

Рис.4.23. График каменистой осыпи для определения числа кластеров

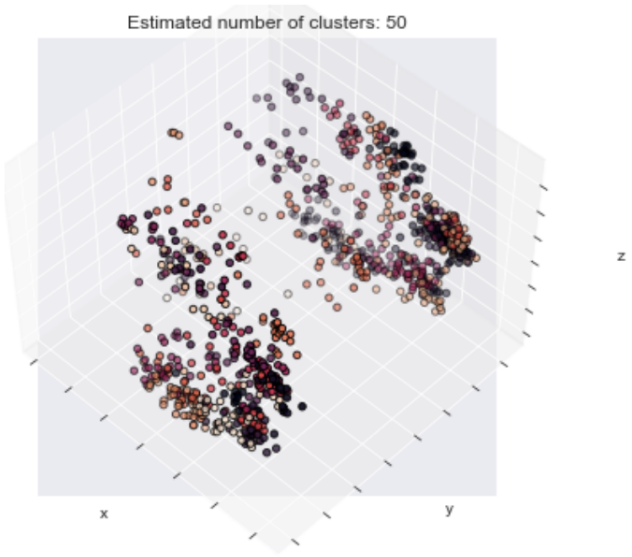

Рис.4.24. Представление кластеризации 1000 объектов при помощи PCA[**pca**]

parameter	2126020278928819713	5830883082242449467	2418201858031258624	1197296442389053745	1093030360823106816
agent_fee	0	50	0	0	35
category_type	ROOMS	ROOMS	ROOMS	ROOMS	ROOMS
flat_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
floors	3	3	4	5	3
floors_total	6	4	4	5	4
is_grandmother_renovation	True	False	True	True	False
is_primary_sale	True	True	True	True	True
is_studio	False	False	False	False	False
is_vos	True	False	True	False	True
kitchen_space	10	9	15	29	20
living_space	15	15	22	23	18
lotinfo_lotarea_value	<NA>	<NA>	<NA>	<NA>	<NA>
offer_type	False	False	False	False	False
price	15000	11000	14000	8000	14000
pricing_period	PER_MONTH	PER_MONTH	PER_MONTH	PER_MONTH	PER_MONTH
quality_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
renovation_type	EURO	COSMETIC_DONE	COSMETIC_DONE	UNKNOWN	EURO
rooms	2	5	7	20	5
sale_agent	OWNER	AGENCY	OWNER	AGENCY	PRIVATE_AGENT
total_space	60	115	110	490	100
year	1902	2005	1900	1951	1802
distance	0.000	8.918	2.449	5.487	3.465

Рис.4.25. Элементы кластера

Модель с полносвязными слоями Представление пространства при помощи PCA при использование данной модели, дает более равномерное покрытие пространства точками. Интерпретация кластера в данном случае является не столь тривиальной, но логика все равно прослеживается. Можно сказать, что в данном кластере сосредоточены аренда квартир в определенной ценовой категории.

$$Sil(C) = 0.053$$

Рис.4.26. График каменистой осыпи для определения числа кластеров

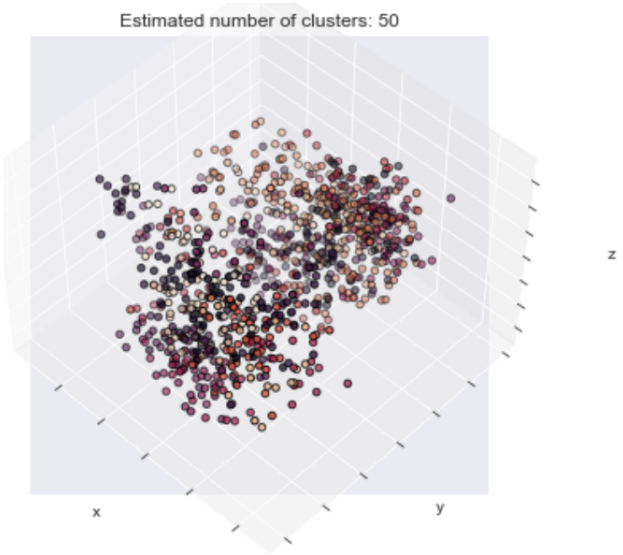

Рис.4.27. Представление кластеризации 1000 объектов при помощи PCA[**pca**]

parameter	6504772238540423604	6504772238533985087	6504772238562469096	5141797489725351349	6504772238539589606
agent_fee	60	60	60	50	60
category_type	APARTMENT	APARTMENT	APARTMENT	ROOMS	APARTMENT
flat_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
floors	2	2	3	2	5
floors_total	5	4	5	5	5
is_grandmother_renovation	True	False	False	False	False
is_primary_sale	True	True	True	True	True
is_studio	False	False	False	False	False
is_vos	False	False	False	False	False
kitchen_space	6	3	5	8	6
living_space	16	14	16	14	20
lotinfo_lotarea_value	<NA>	<NA>	<NA>	<NA>	<NA>
offer_type	False	False	False	False	False
price	20000	22000	24500	15000	24000
pricing_period	PER_MONTH	PER_MONTH	PER_MONTH	PER_MONTH	PER_MONTH
quality_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
renovation_type	RENOVATED	DESIGNER_RENOVATION	DESIGNER_RENOVATION	UNKNOWN	COSMETIC_DONE
rooms	1	1	1	10	1
sale_agent	AGENCY	AGENCY	AGENCY	AGENCY	AGENCY
total_space	30	17	25	130	32
year	1961	1897	1917	1926	1961
distance	0.000	9.893	11.570	11.309	12.511

Рис.4.28. Элементы кластера

Модифицированная модель с полносвязными слоями В данном случае, полученное векторное представление не такое равномерное как у обычной модели с полносвязными слоями, оно похоже на пространство векторов предобученных

признаков объекта. Кластер также вызывает трудности при интерпретации, но общее сходство у объектов найти можно.

$$Sil(C) = 0.056$$

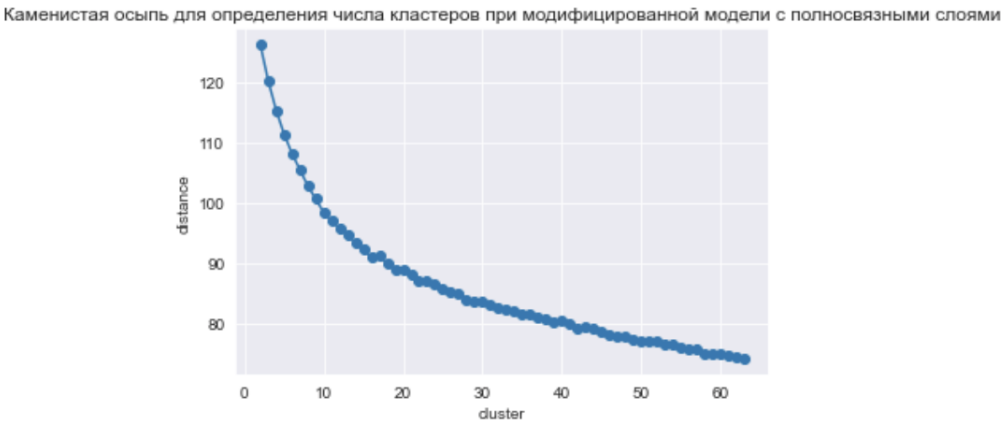

Рис.4.29. График каменистой осыпи для определения числа кластеров

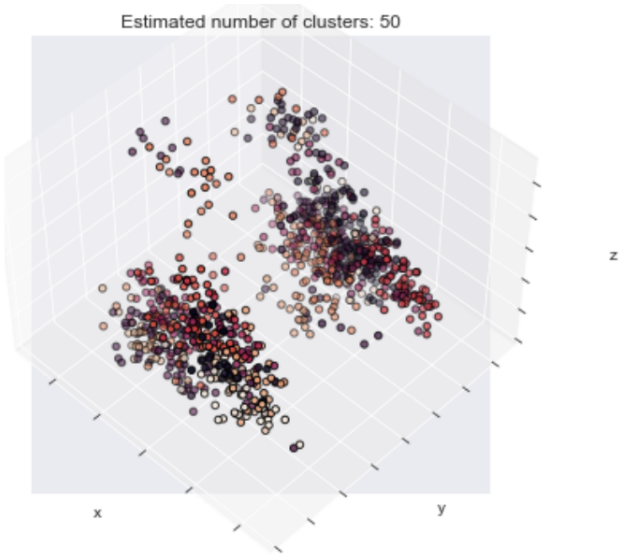

Рис.4.30. Представление кластеризации 1000 объектов при помощи PCA[**pca**]

parameter	5202971678865596616	3488971020659994625	8311717175058678272	6289570373522399489	2084590294170045952
agent_fee	50	50	50	0	50
category_type	APARTMENT	APARTMENT	APARTMENT	APARTMENT	APARTMENT
flat_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
floors	5	4	1	5	1
floors_total	10	5	5	9	5
is_grandmother_renovation	False	False	True	True	False
is_primary_sale	True	True	True	True	True
is_studio	False	False	False	False	False
is_vos	False	True	True	True	True
kitchen_space	9	<NA>	6	6	6
living_space	82	32	43	43	47
lotinfo_lotarea_value	<NA>	<NA>	<NA>	<NA>	<NA>
offer_type	False	False	False	False	False
price	100000	26000	25000	18000	20000
pricing_period	PER_MONTH	PER_MONTH	PER_MONTH	PER_MONTH	PER_MONTH
quality_type	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
renovation_type	UNKNOWN	COSMETIC_DONE	COSMETIC_DONE	EURO	COSMETIC_DONE
rooms	4	3	3	3	3
sale_agent	AGENCY	AGENCY	PRIVATE_AGENT	OWNER	PRIVATE_AGENT
total_space	112	48	58	60	57
year	<NA>	1965	1962	1979	1969
distance	0.000	10.557	2.738	33.266	22.567

Рис.4.31. Элементы кластера

4.7. Выводы

Исходя из проделанных экспериментов можно сделать вывод о том, что рассмотренные модели, кроме модели, использующая ансамбль, адекватны данным. Косинусная мера в построенных векторных пространствах дает интерпретируемые рекомендации.

Модель с полносвязными слоями исходя из перекрестной проверки дает лучшие рекомендации, а также более равномерно заполняет векторное пространство объектами. Модификация этой модели не привнесла никаких изменений, а лишь усложнила интерпретацию алгоритмов кластеризации.

Модель основанная на предобученных признаках также дает, адекватные рекомендации объектов.

Модель, основанная на ансамбле моделей, показала наихудший результат. Исходя из графика функции потерь, было видно, как модель быстро переобучается и теряет связь с изначальными характеристиками объектов, и выучивает лишь их ранги в пользовательской сессии.

ЗАКЛЮЧЕНИЕ

В результате выполнения выпускной квалификационной работы удалось рассмотреть актуальные методы векторизации слов, применить эти методы к структурированным объектам, и исследовать модели на реальной выборке данных.

В ходе исследования моделей векторизации слов было выяснено, что получение векторного представление любого структурированного объекта - осуществима, необходимо лишь иметь данные об историческом взаимодействии с рассматриваемыми объектами.

Построенная модель на основе полносвязных слоев дает релевантные рекомендации, оцененные при помощи перекрестной проверки. Также предобученные признаки объектов можно использовать на вход более мощным моделям построения рекомендаций.

Программная реализация была выполнена на языке Python, она состояла из pipeline обработки данных, обучения модели и построения рекомендаций для изначально выбранных объектов.

Проведенную работу можно считать успешной, так как выполнена основная задача работы - это построение рекомендательной системы, основанной на векторном представлении структурированных объектов.

Дальнейшие исследования заключается в проведение A/B тестирования на реальных пользователях веб-ресурса и изучение поведения модели при большем объеме данных.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ