

ОЛИМПИАДА ШКОЛЬНИКОВ «ШАГ В БУДУЩЕЕ»

НАУЧНО-ОБРАЗОВАТЕЛЬНОЕ СОРЕВНОВАНИЕ «ШАГ В БУДУЩЕЕ, МОСКВА»

10475

регистрационный номер

ИУ «Информатика и системы управления»

название факультета

ИУ7 «Программное обеспечение ЭВМ и информационные технологии»

название кафедры

Анализ отзывов клиентов на банковские каналы обслуживания, продукты/услуги. Сбор адресов отделений, построение моделей бинарной классификации, категоризация всех отзывов и реализация первых двух страниц сайта системы оценки банковских услуг

название работы

Автор:

Федченко Анастасия Кирилловна

фамилия, имя, отчество

ГБОУ Школа №1533 «ЛИТ», 11 класс

наименование учебного заведения, класс

Научный руководитель:

Чамров Михаил Валерьевич

фамилия, имя, отчество

ПАО Банк «ФК Открытие»

место работы

Вице-президент, Лидер трайба

необеспеченное кредитование

звание, должность

подпись научного руководителя

Анализ отзывов клиентов на банковские каналы обслуживания, продукты/услуги

Аннотация

Цифровые технологии открыли новые возможности и установили абсолютно новые правила игры для компаний и пользователей, в результате чего конкуренция сместилась от создания лучшего продукта или услуги к созданию лучшего клиентского опыта.

Актуальность нашего проекта заключается в отсутствии в открытом доступе программ или сайтов, которые занимаются сравнением клиентского опыта в банковской сфере РФ. *Целью нашей работы* является создание web-приложения, позволяющего пользователю сравнить по нескольким показателям любой банк с рынком в среднем и с лучшими игроками рынка, определить области улучшения и дальнейшего развития каналов обслуживания, продуктов и услуг по определенным критериям: удобство офиса, сеть банкоматов, уровень сервиса, персонал, продукты и услуги, дистанционные каналы обслуживания. *Целью моей работы* являлось выполнение нескольких частей: сбор адресов отделений, подготовка отзывов к обработке, ручная разметка отзывов, построение моделей бинарной классификации, категоризация всех отзывов, реализация первых двух страниц сайта.

Для создания веб-приложения мы воспользовались языком программирования Python, фреймворком для визуализации приложений с использованием машинного обучения streamlit, фреймворком scikit-learn для предиктивного анализа данных, библиотекой plotly для создания интерактивных диаграмм и модулем folium для создания карты банковских отделений.

Результатом моей работы являются две страницы сайта: “Средняя оценка по всем банкам” и “Средние оценки категорий по банкам и регионам”, предоставляющие визуализацию проведённого анализа на собранных данных.

Нами создан продукт, принятый в котором подход применим к оцениванию и сравнению различных видов обслуживания, например, приложения банков и доставки продуктов, сети продовольственных супермаркетов и пр. Для построения аналитики нужны лишь отзывы, а также их категоризация.

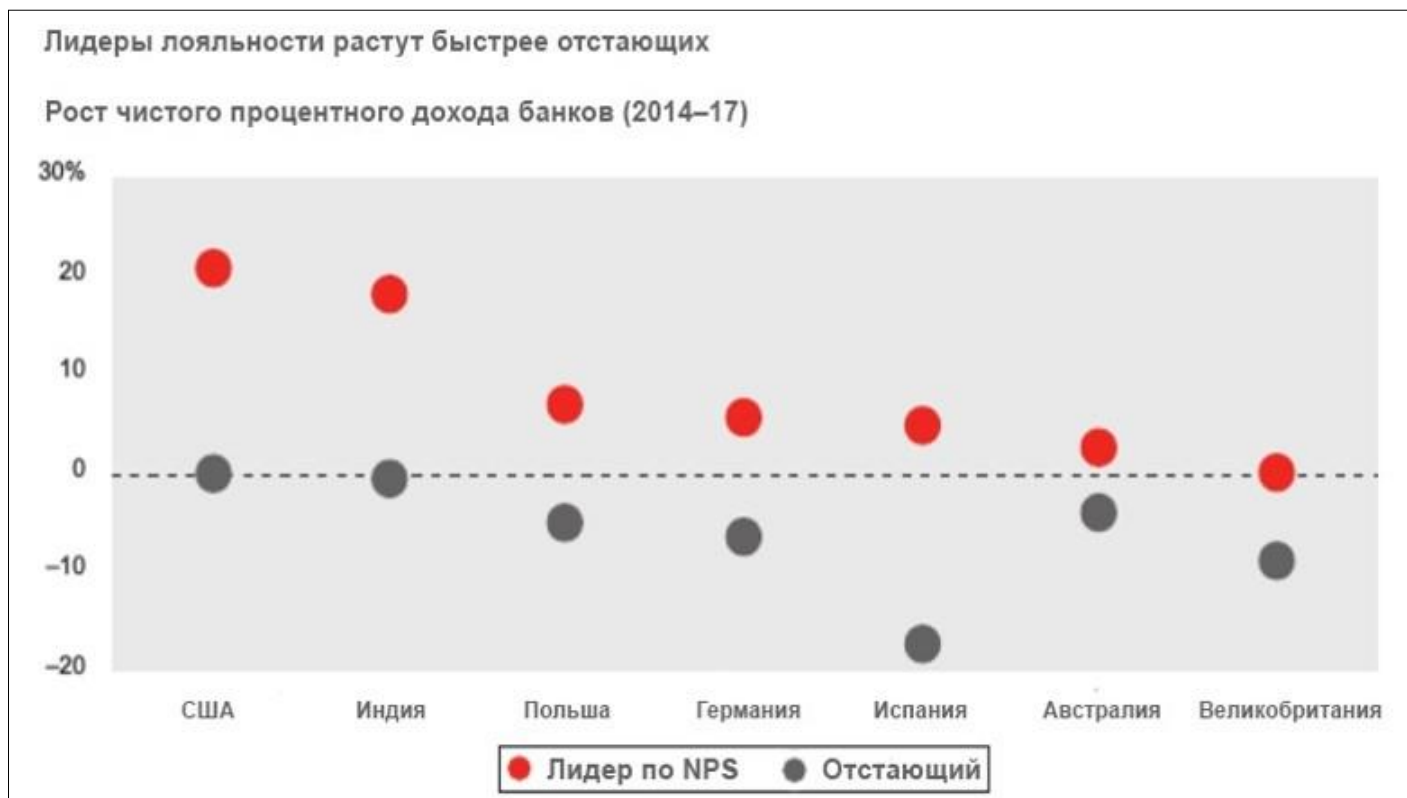
1. Содержание

1.	Содержание	3
2.	Введение	4
2.1	Анализ предметной области	4
2.2	Актуальность	5
2.3	Постановка задачи	6
2.3	Описание решаемых задач	8
2.4	Методы решения моей задачи	8
3.	Основная часть	10
3.1	Методы исследования	10
3.1.1	Критерий согласия Пирсона (χ^2 , хи-квадрат)	10
3.1.2	Метод опорных векторов	11
3.2	Программная реализация	12
3.2.1	Язык программирования, среда разработки, сторонние программы, API	12
3.2.2	Библиотеки Python	12
3.2.3	Структура базы данных	14
3.3	Практическое использование	16
3.4	Выводы	21
4.	Заключение	23
4.1	Результат	23
4.2	Перспективы дальнейшей разработки	23
5.	Список использованной литературы	24
6.	Приложения	25
1.	Sequence Diagram (диаграмма последовательностей) для сбора отделений с сайта https://1000bankov.ru	25
2.	Class Diagram (диаграмма классов)	26
3.	Activity diagram (диаграмма деятельности) для сбора отделений	27
4.	Component diagram (диаграмма компонентов)	28
5.	Графики precision-recall для моделей бинарной классификации	29
6.	ROC-AUC кривая для моделей бинарной классификации	30

2. Введение

2.1 Анализ предметной области

Цифровые технологии открыли новые возможности и установили абсолютно новые правила игры для компаний и пользователей, в результате чего конкуренция сместилась от создания лучшего продукта/услуги к созданию лучшего клиентского опыта. Качество взаимодействия бизнеса с людьми вычисляется с помощью индекса потребительской лояльности (NPS). NPS — индекс определения приверженности потребителей товару или компании (индекс готовности рекомендовать), используется для оценки готовности к повторным покупкам. На представленной диаграмме красными маркерами отмечены банки-лидеры по NPS; серыми маркерами отмечены отстающие по этому показателю. На графике мы можем увидеть, насколько изменился чистый процентный доход в зависимости от страны и NPS за три года. Обратим внимание на то, что у лидеров по NPS чистый процентный доход увеличился; у отстающих банков же он либо остался на прежнем уровне, либо уменьшился. Исходя из вышесказанного, клиентский опыт очень важен, и поэтому наш проект нацелен на его анализ в разных банках.



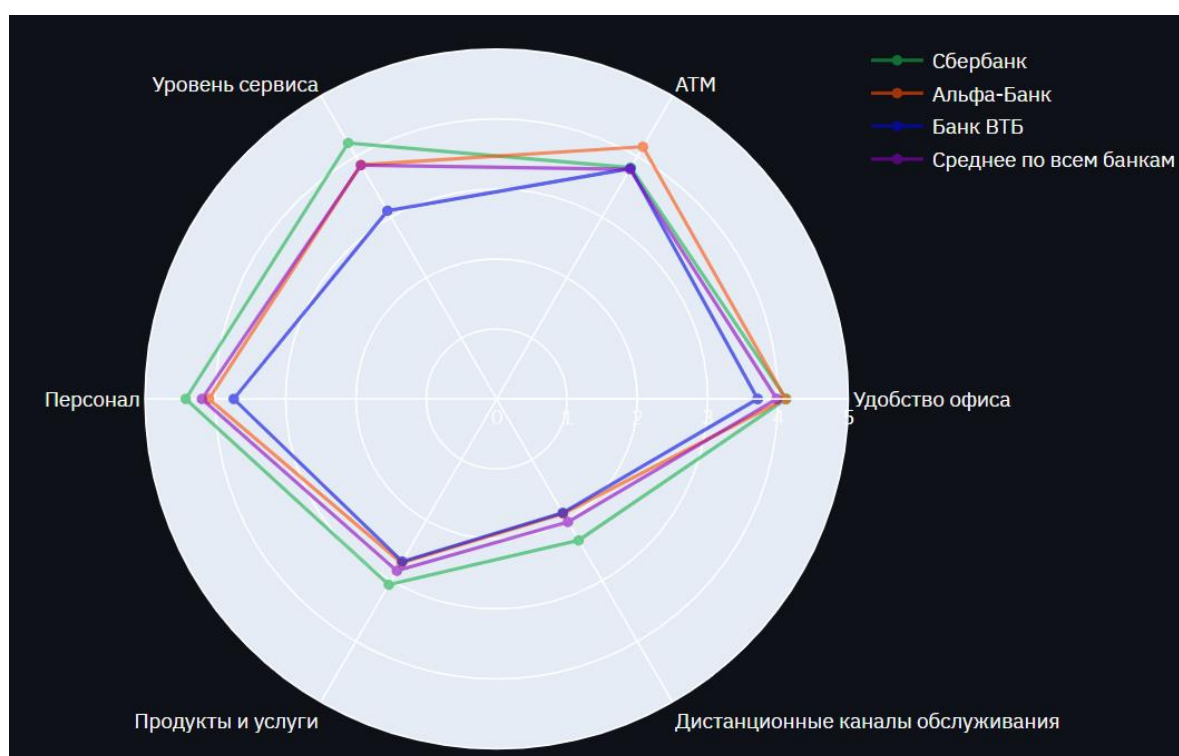
Источник: исследование Bain & Company “In search of customers who love their bank”

2.2 Актуальность

Лучший клиентский опыт включает в себя простые, умные и беспроblemные точки контакта пользователя с компанией, персонафикацию, простоту и понятность использования, позитивные эмоции от процесса использования продукта/услуги (не просто функциональные свойства) в условиях постоянно изменяющихся предпочтений пользователей и конкурентной среды.

Банки – традиционная и консервативная отрасль (в которой, конечно, есть и свои технологические лидеры), находящаяся в начале пути от предоставления традиционных услуг и создания экономически эффективных процессов к человеко-центричному дизайну.

Итак, причиной актуальности нашего продукта является отсутствие в открытом доступе программ/сайтов, занимающихся сравнением клиентского опыта в различных банках. Наша программа упрощает пользователю процесс сравнения нескольких банков по разным критериям. Вместо того чтобы искать информацию про каждый банк (т.е. открывать официальные сайты, форумы, отзывы на конкретные услуги), тратя большое количество времени, можно воспользоваться нашим сайтом и получить всю нужную информацию быстро и удобно, получив результат анализа в виде графика в формате “паутина” (пример на фотографии ниже).



2.3 Постановка задачи

Целью нашей работы является создание web-приложения, позволяющего пользователю сравнить любой банк с рынком в среднем и с лучшими игроками рынка, определить области улучшения и дальнейшего развития каналов обслуживания, продуктов и услуг по определенным критериям: удобство офиса, банкоматы, касса, уровень сервиса, персонал, продукты и услуги, дистанционное обслуживание. В качестве входных данных мы используем отзывы с картографического сервиса Яндекс.Карты (www.yandex.ru/maps) на банковские отделения. На основе полученных данных создаем модель оценки качества банковских каналов, продуктов и услуг с визуализацией результата.

Целью моей работы являлось выполнение нескольких частей:

- сбор адресов отделений,
- подготовка отзывов к обработке,
- ручная разметка отзывов,
- построение моделей бинарной классификации,
- категоризация всех отзывов,
- реализация первых двух страниц сайта.

Вместе с техническим заданием мы получили от научного руководителя список банков, отзывы на отделения которых нам необходимо проанализировать. Ниже приведена таблица, содержащая список с названиями 24 банков:

1	Сбербанк
2	Банк ВТБ
3	Альфа-Банк
4	Газпромбанк
5	Россельхозбанк
6	Почта Банк
7	Банк Открытие
8	Росбанк

9	Совкомбанк
10	Райффайзенбанк
11	Промсвязьбанк
12	Банк Хоум Кредит
13	Банк ДОМ.РФ
14	Уралсиб
15	ЮниКредит
16	Сетелем Банк
17	Русфинанс Банк
18	Ренессанс Кредит
19	Московский кредитный банк
20	Банк Санкт-Петербург
21	СМП Банк
22	Московский индустриальный банк
23	Уральский банк реконструкции и развития
24	Ситибанк

Для построения аналитики необходимо разделить отзывы пользователей на темы. Данная категоризация была предложена научным руководителем.

1. **Удобство офиса** (удобство расположения, транспортная доступность, наличие парковки, время работы, простота навигации внутри офиса, электронная очередь/возможность записаться в офис, комфортная зона ожидания, состояние и привлекательность интерьера офиса, чистота, наличие кулера с водой/туалета для клиентов, детского уголка, Wi-Fi и т.п.);

2. **Банкоматы** (их наличие, круглосуточная доступность, достаточность количества устройств, функциональность - выдача/прием наличных/платежи/переводы и платежи и т.п., работоспособность - в банкомате есть наличные/внесение наличных работает/у банкомата есть связь/он не висит и т.д.);

3. Уровень сервиса (время ожидания, скорость обслуживания, простота документов, доступность безбумажной/электронной подписи клиента, доступность нужной клиенту услуги, уполномоченность персонала решать вопросы по претензиям на месте без перенаправления в головной офис);

4. Персонал (вежливость, опрятность, клиентоориентированность/заинтересованность в решении вопроса клиента, компетентность, решение вопроса клиента без перенаправления к другому специалисту и т.п.);

5. Продукты и услуги (полнота и доступность, стоимость, прозрачность условий, удобство использования и др. характеристики стандартных продуктов и услуг банка);

6. Дистанционные каналы обслуживания (удобство интернет-банка (ИБ) и мобильного банка (МБ), простота подключения, доступность операций в ИБ/МБ без необходимости визита в офис, простота совершения операций в ИБ/МБ, ограничения по операциям в ИБ/МБ и пр.);

7. Прочее. В эту категорию следует относить отзывы, которые не принадлежат ни к одной категории выше.

2.3 Описание решаемых задач

1. упрощение процесса сравнения клиентского опыта в разных банках;
2. выделение наиболее важных критериев оценки по мнению клиентов;
3. определить области улучшения и дальнейшего развития каналов обслуживания, продуктов и услуг.

2.4 Методы решения моей задачи

Кратко опишу, как я решала мою задачу.

1. Сбор адресов отделений банков с сайта <https://1000bankov.ru>;
2. Подготовка отзывов к анализу: удаление стоп-слов, знаков препинания, приведение слов в начальную форму;
3. Ручной лейблинг отзывов на случайной выборке, состоящей из 1600 отзывов, на основе категоризации из «Постановки задачи»;

4. Обучение моделей бинарной классификации на сервисе Google Colab на основе пролейбленных отзывов из пункта 4 при помощи метода опорных векторов;
5. Категоризация всех отзывов при помощи обученных моделей;
6. Визуализация: создание двух первых страниц web-приложения (“Средняя оценка по всем банкам”, “Средние оценки категорий по банкам и регионам”), которые содержат горизонтальные гистограмм для отображения средних оценок, диаграммы в форматах «паутина» и «торнадо» (с применением критерия согласия Пирсона).

3. Основная часть

3.1 Методы исследования

3.1.1 Критерий согласия Пирсона (χ^2 , хи-квадрат)

Критерий согласия Пирсона проверяет значимость расхождения эмпирических (наблюдаемых) и теоретических (ожидаемых) частот. Он используется для проверки нулевой гипотезы о подчинении наблюдаемой случайной величины определенному теоретическому закону распределения. Нулевая гипотеза заключается в том, что частоты согласованы, то есть фактические данные не противоречат ожидаемым.

Критерий согласия Пирсона считается по данной формуле:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e},$$

где f_o и f_e — наблюдаемые и ожидаемые частоты соответственно. Суммирование производится по всем ячейкам таблицы.

В нашей задаче мы используем критерий согласия Пирсона при проверке расхождения между частотами в двух ситуациях:

1) Сравнение долей положительных и отрицательных отзывов в одном банке каждой категории и итоговых долей всех категорий:

Нулевая гипотеза такова: Распределение долей положительных и отрицательных отзывов в категории имеет то же распределение, что по всем отзывам банка в целом.

Пусть распределение долей положительных и отрицательных отзывов по всем отзывам банка — это ожидаемое распределение, а распределение долей положительных и отрицательных отзывов по отзывам конкретной категории — фактическое распределение.

Если вероятность p , которая посчитана при помощи функции `scipy.stats.chi2_contingency`, больше 0.05, то мы не показываем выбранную категорию, так как она с высокой вероятностью похожа на распределение долей по всем отзывам, то есть нулевая гипотеза оказалась правдивой.

Если вероятность $p \leq 0.05$, то распределения различаются, значит, доли выбранной категории стоит показать, то есть мы опровергли нулевую гипотезу.

2) Сравнение долей положительных и отрицательных отзывов каждой категории между банками.

Нулевая гипотеза такова: распределение долей положительных и отрицательных отзывов в пределах одной категории между всеми банками совпадают.

Мы попарно проверяем распределения между банками. Фактическим и ожидаемым распределениями мы ставим соответственно положительные и отрицательные доли отзывов на каждую пару банков. Если найдется хотя бы одно, вероятность p которого ≤ 0.05 , то мы показываем весь ряд, так как распределения с высокой вероятностью различны, то есть нулевая гипотеза опровергнута.

3.1.2 Метод опорных векторов

Метод опорных векторов — это алгоритм, позволяющий определить гиперплоскость (в частном случае — линию), которая распределяет данные на два класса. Разделяющей гиперплоскостью будет гиперплоскость, создающая наибольшее расстояние до двух параллельных гиперплоскостей.

Мы воспользовались классификатором на основе метода опорных векторов для создания двух типов моделей:

- определение отзыва к определенной категории;
- определение принадлежности слова к позитивным или негативным отзывам.

Для создания моделей первого типа мы "пролейблили" 1600 отзывов по категориям (см. «Постановка задачи»).

Входные данные: "чистый" отзыв и 0 либо 1 (0 — отзыв не принадлежит категории, 1 — отзыв принадлежит категории).

Каждая из шести моделей определяет принадлежность или непринадлежность отзыва к конкретной категории.

Для создания моделей второго типа мы в качестве входных данных использовали отзывы по каждой из категорий и 0 либо 1 (0 — оценка отзыва 1 или 2 (т.е. он негативный), 1 — оценка отзыва 4 или 5 (т.е. он положительный)).

Каждая из шести моделей может определить, в каких отзывах встречается слово: в положительных или отрицательных.

3.2 Программная реализация

3.2.1 Язык программирования, среда разработки, сторонние программы, API

В качестве языка программирования мы выбрали язык Python, так как на нем легко писать аналитику и визуализацию сайта.

Средой разработки мы выбрали PyCharm Community Edition. Она обладает большим функционалом, является простой в использовании и оснащена системой контроля версий Git.

Для работы с базой данных мы решили использовать программу SQLiteStudio. В ней мы можем смотреть записи из базы и делать тестовые запросы на языке SQL.

В качестве API мы использовали API Поиска по организациям от компании Яндекс для поиска координат отделения и его id в Яндекс.Картах. Также мы воспользовались Геокодером API Яндекс.Карт для нахождения субъекта Российской Федерации, в котором находится каждое отделение.

Для ручного «лейблинга» отзывов мы использовали Google Таблицы, чтобы параллельно друг с другом заниматься выделением тем в отзывах.

Для обучения моделей мы воспользовались сервисом Google Colab, так как он в разы ускоряет обучение моделей.

3.2.2 Библиотеки Python

В работе над проектом мы использовали большое количество библиотек. Ниже нам бы хотелось поподробнее рассказать про самые важные из них.

1. requests

Итак, первой библиотекой, которой мы воспользовались, является **requests**. Она нам необходима для обращения к сайтам в сети Интернет. При помощи метода `get()` мы получали HTML-структуру страницы сайта Яндекс.Карт и www.1000bankov.ru.

2. sqlite

Встроенный модуль языка Python позволил нам быстро и удобно работать с базой данных. Используя эту библиотеку, мы добавляли полученные с Яндекс.Карт отзывы и с www.1000bankov.ru адреса отделений, а также принадлежность отзывов к категориям и прочую необходимую информацию.

4. scikit-learn (sklearn)

scikit-learn – мощный инструмент для машинного обучения на языке Python. Мы воспользовались четырьмя классами из данной библиотеки:

1. **sklearn.svm.SVC** – классификатор, основанный на методе опорных векторов, мы использовали для создания моделей бинарной классификации;
2. **sklearn.feature_extraction.text.TfidfVectorizer** – трансформатор, отвечающий за перевод слов в цифры, а также оценку важности слова в контексте документа;
3. **sklearn.pipeline.Pipeline** – класс, который объединяет трансформаторы (в нашем случае это TfidfVectorizer) и модели (в нашем случае это SVC) для последовательной обработки данных и предсказания на обработанных данных;

В этой библиотеке также нашлось несколько полезных функций, которые мы использовали при работе:

1. **sklearn.model_selection.train_test_split()** – функция, которая делит датасет на тренировочные и тестовые данные.
2. **sklearn.metrics.roc_auc_score**, **sklearn.metrics.accuracy_score** – функции, показывающие оценку точности моделей.

5. nltk

nltk (Natural Language Toolkit) – пакет библиотек и программ для символьной и статистической обработки естественного языка. Из этой библиотеки мы импортировали список стоп-слов русского языка, которые нужно убрать из отзывов, а также функцию `nltk.word_tokenize(review)`, которая токенизирует предложения отзывов по словам.

Токенизация по словам – это процесс разделения предложений на слова-компоненты. Почти во всех языках пробел – это один из самых удобных разделителей слов, однако, могут возникнуть проблемы, если мы будем использовать только пробел – в русском языке составные существительные пишутся по-разному и иногда через пробел.

6. pymorphy2

pymorphy2 - морфологический анализатор для русского языка, написанный на языке Python и использующий словари из OpenCorpora.

Данный модуль необходим для приведения слов в начальную форму для корректной обработки «очищенных» отзывов.

7. plotly

Библиотека **plotly** позволяет строить красивые и удобные для просмотра диаграммы. С ее помощью мы построили три типа диаграмм:

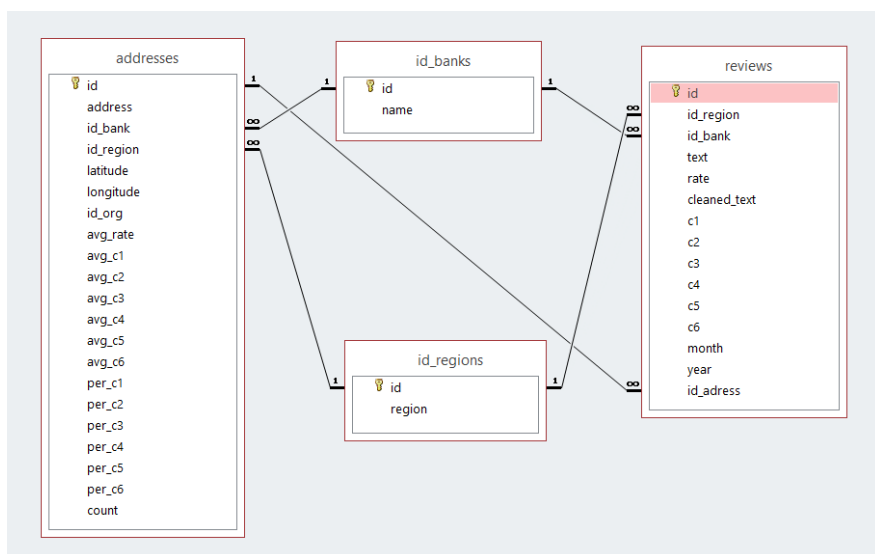
- горизонтальная гистограмма;
- диаграмма в формате «паутина» (Radar Chart);
- диаграмма в формате «торнадо» или «бабочка» (Tornado Chart).

8. streamlit

Streamlit – фреймворк, специально разработанный для визуализации приложений с использованием машинного обучения. Эта библиотека отлично работает с plotly, поэтому наш выбор пал именно на streamlit. Благодаря streamlit мы сохранили интерактивность и функциональность диаграмм.

3.2.3 Структура базы данных

На фотографии ниже приведена схема базы данных.



Поясним структуру каждой таблицы.

id_banks – таблица, которая хранит в себе пары id-название банка;

id_regions – таблица, которая хранит в себе пары id-регион РФ;

addresses – таблица, в которой хранятся адреса отделений банков:

1. id – уникальный идентификатор;
2. address – адрес отделения, взятого с сайта 1000bankov.ru;

3. `id_bank` – идентификатор банка из таблицы `id_banks`, отделением которого является адрес из предыдущего столбца `address`;
4. `id_region` – идентификатор региона, в котором находится отделение;
5. `latitude` и `longitude` – соответственно широта и долгота координат отделения.
Данные столбцы необходимы, чтобы избежать проблемы дубликации адресов с сайта 1000bankov.ru (про эту проблему рассказано в разделе 7.1.1 Выбор источника адресов отделений банков);
6. `id_org` – id организации, полученный из API Поиска по организациям Яндекса. Он нужен для получения отзывов.
7. `avg_rate` – средняя оценка отзывов, которые оставлены на конкретное отделение;
8. `avg_c1` – средняя оценка отзывов на отделение по категории «Удобство офиса»;
9. `avg_c2` – средняя оценка отзывов на отделение по категории «Банкоматы»;
10. `avg_c3` – средняя оценка отзывов на отделение по категории «Уровень сервиса»;
11. `avg_c4` – средняя оценка отзывов на отделение по категории «Персонал»;
12. `avg_c5` – средняя оценка отзывов на отделение по категории «Продукты и услуги»;
13. `avg_c6` – средняя оценка отзывов на отделение по категории «Дистанционные каналы обслуживания»;
14. `per_c1` – доля отзывов по категории «Удобство офиса»;
15. `per_c2` – доля отзывов по категории «Банкоматы»;
16. `per_c3` – доля отзывов по категории «Уровень сервиса»;
17. `per_c4` – доля отзывов по категории «Персонал»;
18. `per_c5` – доля отзывов по категории «Продукты и услуги»;
19. `per_c6` – доля отзывов по категории «Дистанционные каналы обслуживания»;
20. `count` – количество отзывов на отделение.

Столбцы 8-20 необходимы для быстрого построения интерактивной карты банковских отделений.

reviews – таблица, в которой хранятся отзывы:

1. id – уникальный идентификатор;
2. id_region – идентификатор региона, в котором находится отделение, на которое оставлен отзыв;
3. id_bank – идентификатор банка, на отделение которого оставлен отзыв;
4. text – текст отзыва;
5. rate – оценка, которую оставил пользователь;
6. cleaned_text – очищенный отзыв;
7. c1 – принадлежность отзыва к категории «Удобство офиса» (0 или 1);
8. c2 – принадлежность отзыва к категории «Банкоматы» (0 или 1);
9. c3 – принадлежность отзыва к категории «Уровень сервиса» (0 или 1);
- 10.c4 – принадлежность отзыва к категории «Персонал» (0 или 1);
- 11.c5 – принадлежность отзыва к категории «Продукты и услуги» (0 или 1);
- 12.c6 – принадлежность отзыва к категории «Дистанционные каналы обслуживания» (0 или 1);
- 13.month – месяц, в котором был оставлен отзыв на отделение;
- 14.year – год, в котором был оставлен отзыв на отделение;
- 15.id_address – идентификатор отделения, на который оставили конкретный отзыв (необходим для аналитики средних оценок по категориям для каждого отделения).

3.3 Практическое использование

Зайдём на сайт, пользователю предлагается выбрать язык (русский/английский) и одну из страниц: средняя оценка по всем банкам, средние оценки категорий по банкам и регионам, важные слова для различных категорий, интерактивная карта банковских отделений.

Выберите язык / Select language:

☒ Русский

☐ English

Меню

Выберите страницу:

☒ Главная

☐ Средняя оценка по всем банкам

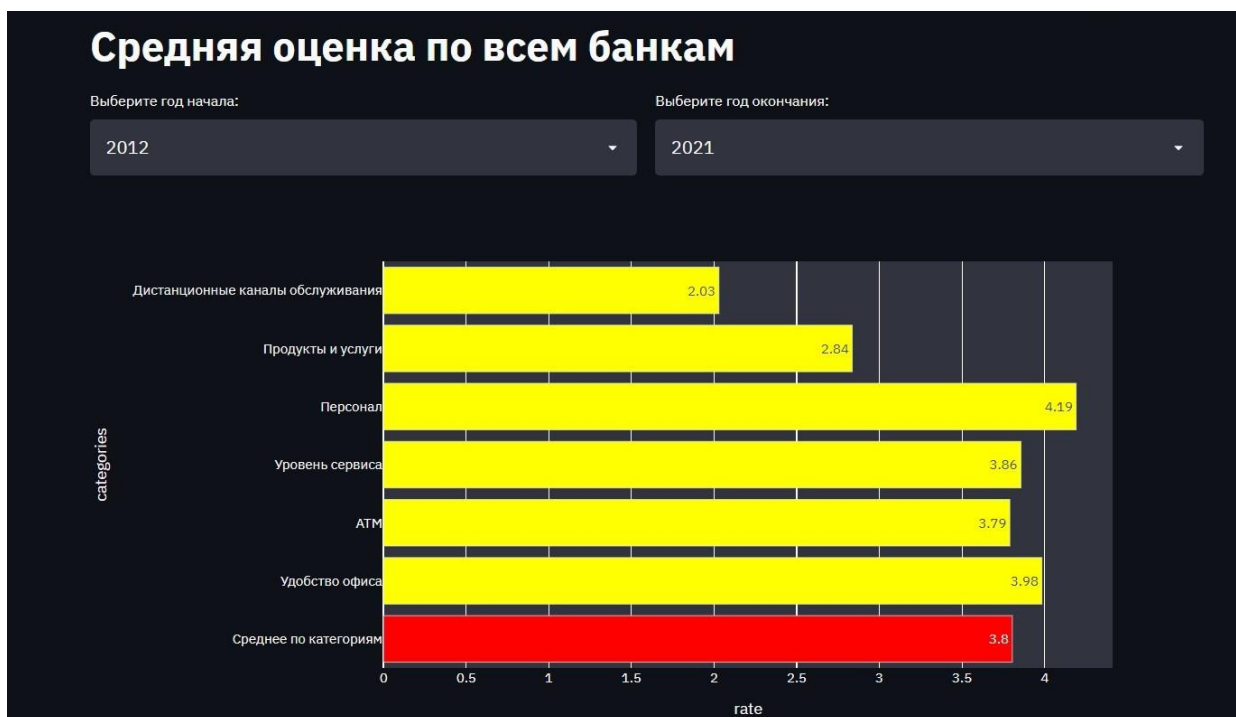
☐ Средние оценки категорий по банкам и регионам

☐ Важные слова для различных категорий

☐ Интерактивная карта банковских отделений

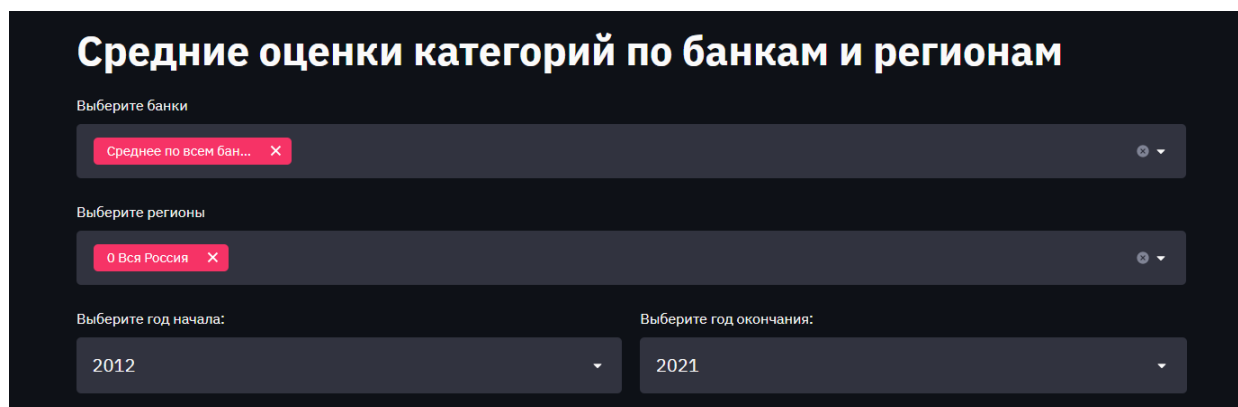
Средняя оценка по всем банкам

При переходе на первую страницу пользователь может увидеть диаграмму средней оценки по всем банкам. Имеется возможность выбора периода времени, за который будут выведены данные. Это диаграмма отлично подходит для анализа изменения клиентского опыта во всех банках в целом за определённый период времени.



Средние оценки категорий по банкам и регионам

На второй странице пользователю доступна диаграмма средних оценок категорий по банкам и регионам, представленная в формате “паутинки”. Такой формат позволяет легко сравнить между собой несколько банков. Для данной диаграмме также возможен выбор периода времени, банков (не больше 4) и регионов, которые будут соответственно распространяться и на другие диаграммы этой страницы.



The screenshot shows a web form titled "Средние оценки категорий по банкам и регионам" (Average category ratings by bank and region). The form is set against a dark background with light-colored text and input fields. It contains three main sections: 1. "Выберите банки" (Select banks) with a dropdown menu showing "Среднее по всем банкам..." (Average for all banks...). 2. "Выберите регионы" (Select regions) with a dropdown menu showing "0 Вся Россия" (0 All Russia). 3. Two date selection fields: "Выберите год начала:" (Select start year) with a dropdown showing "2012", and "Выберите год окончания:" (Select end year) with a dropdown showing "2021". Each dropdown menu has a small circular icon and a downward arrow on the right side.

Выше представлены начальные значения параметров.

Средние оценки категорий по банкам и регионам

Выберите банки

Газпромбанк

Уралсиб

Выберите регионы

77 Москва

Выберите год начала:

2012

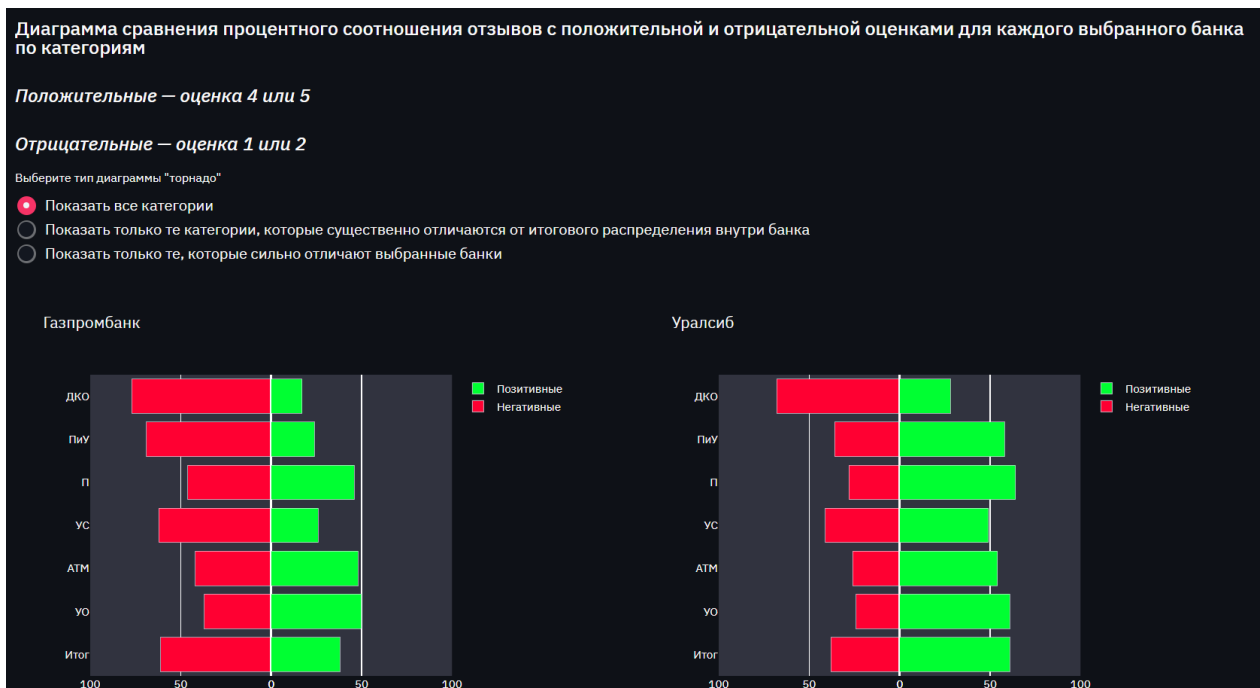
Выберите год окончания:

2021

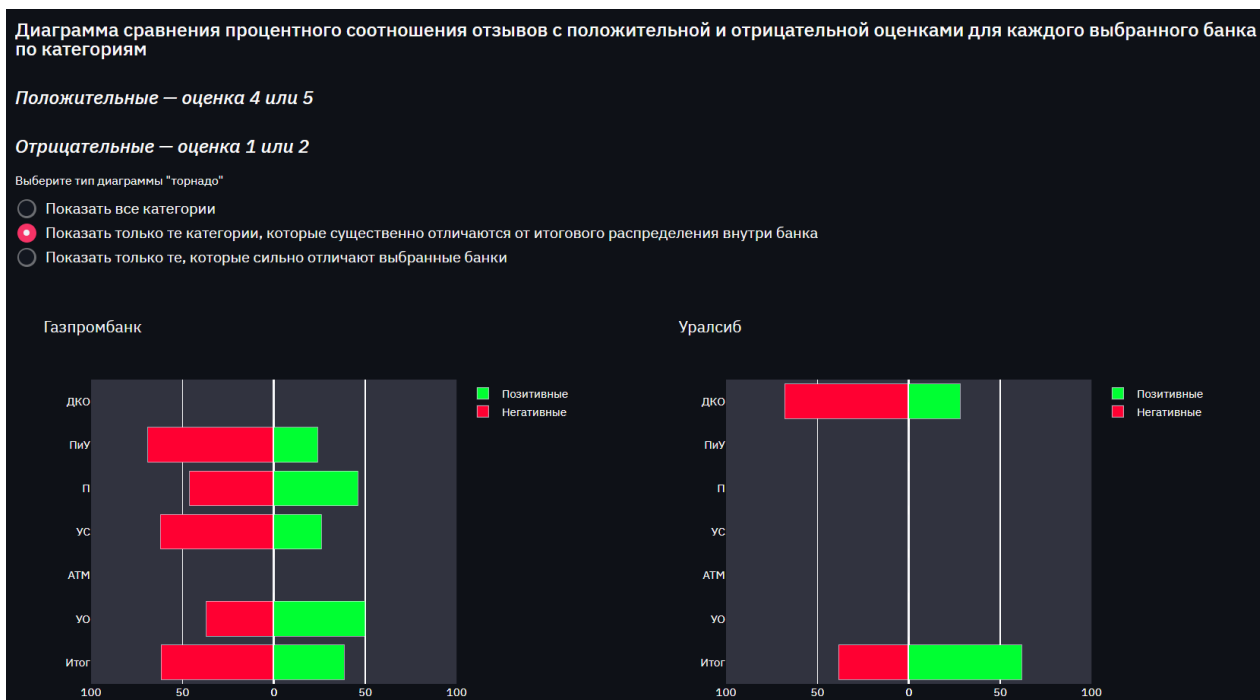


На этой картинке Вы можете увидеть пример выбора двух банков и региона.

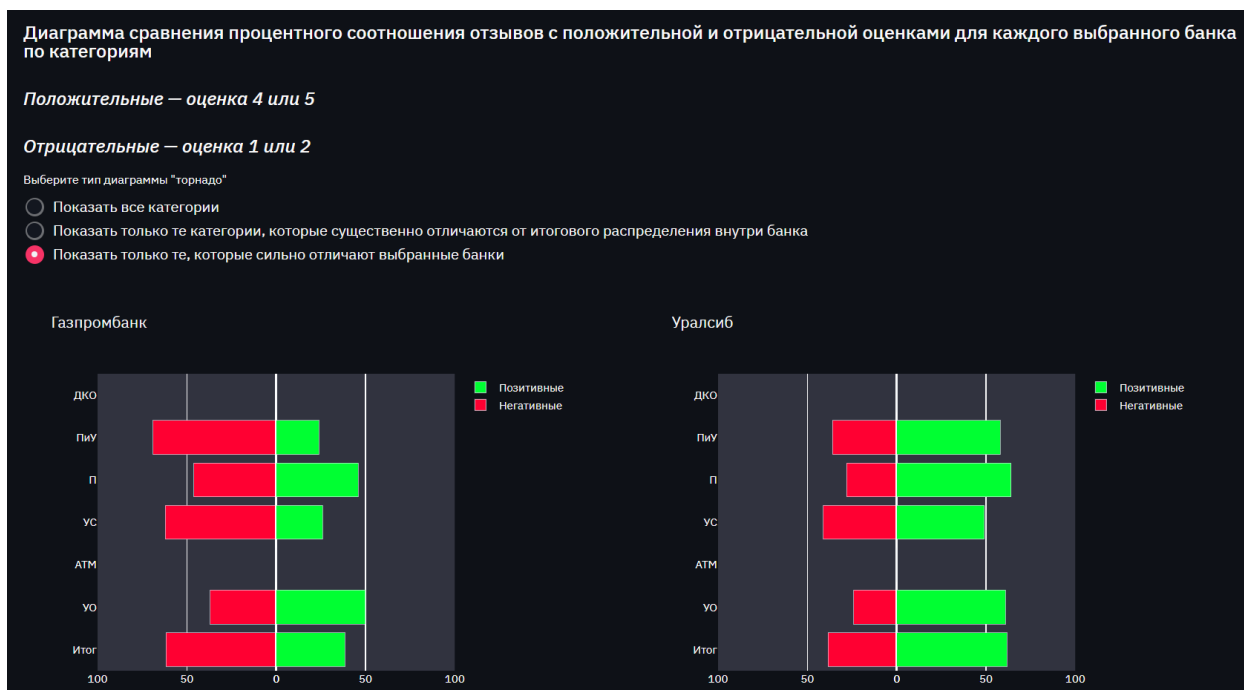
Ниже находится диаграмма сравнения процентного соотношения отзывов с положительной и отрицательными оценками для каждого выбранного банка по категориям в формате торнадо. Эта диаграмма представлена в трёх версиях. Первая показывает все распределения оценок, а остальные две показывают лишь те, которые существенно отличаются от итогового среднего в банке либо отличают выбранные банки по категориям между собой.



Пользователь может самостоятельно выбрать, какую именно версию смотреть.



При просмотре этой диаграммы, можем сделать вывод, что в Газпромбанке схожи с средним только две категории, а у Уралсиба, наоборот, отличается от итога только одна.



И здесь мы можем увидеть, соотношение отзывов в выбранных нами банках совпадает только в двух из шести категориях.

3.4 Выводы

За последние три года уровень предоставления банковских услуг в отделениях увеличился (средний показатель вырос на 0.5 балла из 5).

	Средняя оценка за 2017-2018	Доля категории во всех отзывах (2017-2018)	Средняя оценка за 2020-2021	Доля категории во всех отзывах (2020-2021)	Изменение средней оценки	Изменение доли категории
Удобство офиса	3.59	16.8%	4.02	17.6%	+0.43	+0.8%
Банкоматы	3.26	15.1%	3.91	9.8%	+0.65	-5.3%
Уровень сервиса	3.3	35.6%	3.87	30%	+0.57	-5.6%
Персонал	3.63	24%	4.22	28.4%	+0.59	+4.4%
Продукты и услуги	2.38	6.9%	2.77	6.8%	+0.39	-0.1%
Дист. каналы обслуживания	1.63	2.2%	2.03	1.5%	+0.4	-0.7%

Рост средней оценки вызван огромным ростом средних оценок в каждой из категорий.

За три года предпочтения клиентов банковских отделений изменились: пользователи стали чаще писать про персонал, но реже про банкоматы и уровень сервиса.

Изменения средних оценок вызваны резким скачком улучшения предоставления услуг у главного банка России – Сбербанка (занимает ~50% банковского рынка РФ). У остальных банков из топ-10 списка, который приведен в «Постановке задачи», особых изменений в оценках нет. Сбербанк смело можно назвать драйвером изменений на всем рынке за три года.

Можно сказать, что Сбербанк и Тинькофф, который нацелен на дистанционное обслуживание, задали новый тренд в дистанционных каналах обслуживания (средняя оценка за 2017-2018 у Сбербанка 1.79, 2020-2021 – 2.86). Несмотря на уменьшение доли данной категории во всех отзывах люди стали чаще оценивать приложения (скорее всего, отзывы на приложения клиенты оставляют в магазинах Apple Store и Play Market) и колл-центры банков; у многих банков за этот период дистанционные каналы появились. Клиенты банков оставляют отзывы на данную категорию лишь в случае, когда они посещают отделение банка по причине плохой работы приложения или колл-центра, именно поэтому эта категория имеет самый низкий средний балл (т.е. люди жалуются).

Клиенты банков стали гораздо чаще стали оставлять отзывы на картографических сервисах (например, Яндекс.Карты, отзывы с которого мы обзоредали). За 2017-2018 годы пользователи написали около 25 тысяч отзывов, а за 2020-2021 – уже 380 тысяч. Как мы можем заметить, количество отзывов за последний период увеличилось более чем в 15 раз.

4. Заключение

4.1 Результат

Результатом моей работы являются две страницы сайта, предоставляющие визуализацию проведённого анализа на собранных данных.

На этих страницах пользователю доступно три типа диаграмм:

1. горизонтальная гистограмма (для отображения средних оценок);
2. диаграмма в формате “паутинка” (для удобного отображения средних оценок по категориям);
3. диаграмма в формате “торнадо” (для отображения всех распределений положительных и отрицательных отзывов по категориям и в среднем, а также только тех распределений, которые сильно отличаются от среднего распределения и от распределения по категориям между банками);

Нами создан продукт, подход которого применим к оцениванию и сравнению различных видов обслуживания, например, приложения банков и доставки продуктов, сети продовольственных супермаркетов и пр. Для построения аналитики нужны лишь отзывы, а также их категоризация (т.е. про что пользователи пишут).

4.2 Перспективы дальнейшей разработки

1. Анализ отзывов пользователей на мобильные приложения банков из магазина приложений Google Play;
2. Анализ развернутых отзывов клиентов об услугах банков в «Народном рейтинге» на сайте banki.ru.

Эта задача является более сложной, так как на «Народном рейтинге» пользователи в основном пишут отзывы не о конкретных плюсах (компетентный персонал, отсутствие очередей и пр.) или минусах отделения (неудобные часы работы, неработающие банкоматы и пр.), а о банковском обслуживании в целом.

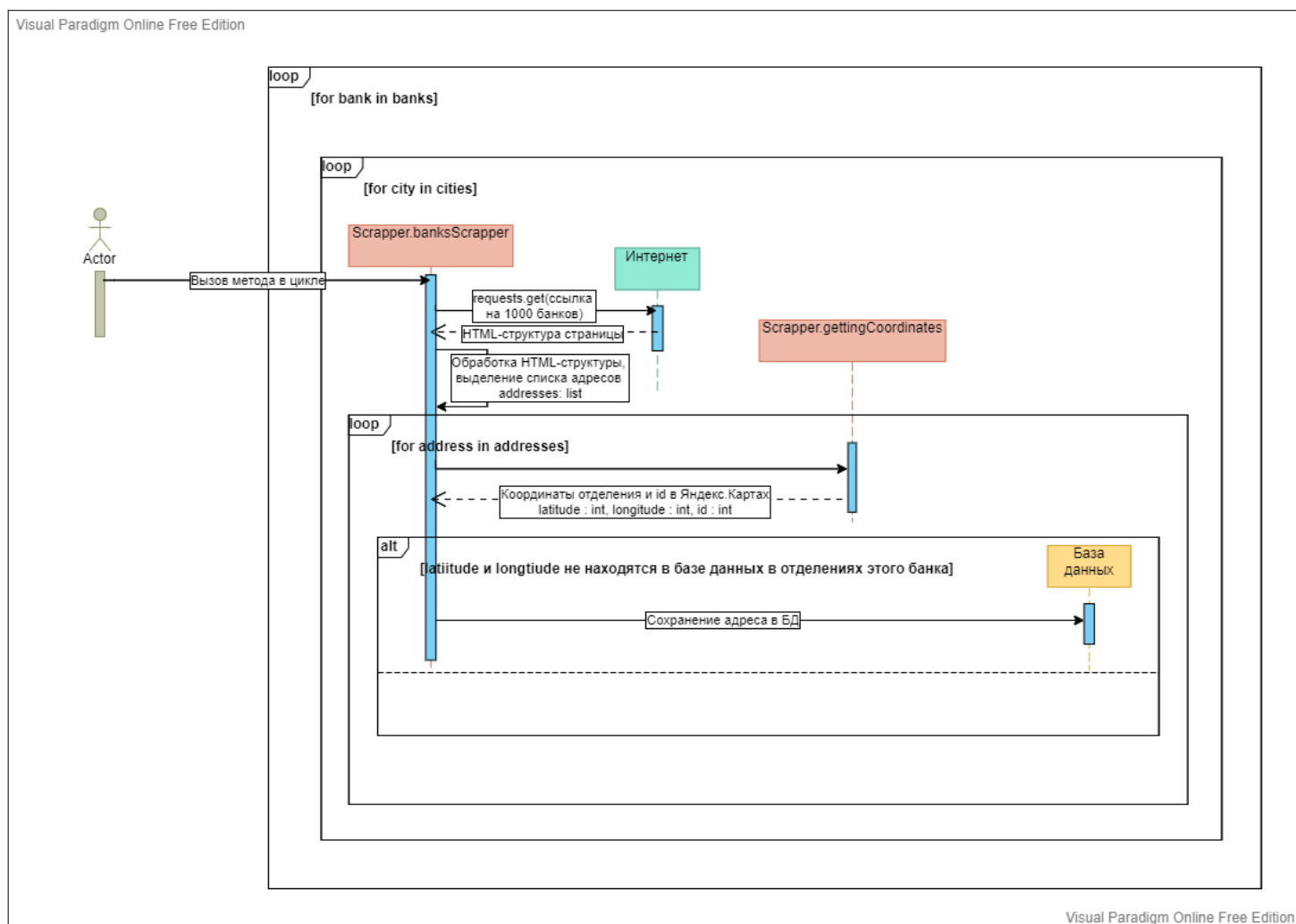
5. Список использованной литературы

1. ООО «Яндекс»: О сервисе. API поиск по организациям [Электронный ресурс]: 2020 — Режим доступа: <https://yandex.ru/dev/maps/geosearch/doc/concepts/about.html?from=geosearch>
2. Voximplant: Основы Natural Language Processing для текста / Блог компании Voximplant / Хабр [Электронный ресурс]: 15.04.2019 — Режим доступа: <https://habr.com/ru/company/Voximplant/blog/446738/>
3. Mikhail Korobov Revision: Документация — Морфологический анализатор pymorphy2 [Электронный ресурс]: 2013-2020 — Режим доступа: <https://pymorphy2.readthedocs.io/en/latest/user/index.html>
4. Список регионов (субъектов, областей) России 2021 РФ с кодами согласно данным ФНС по алфавиту [Электронный ресурс]: 2020 — Режим доступа: <https://www.sites.google.com/site/ruregdatav1/spisok-regionov-rossii-s-kodamy>
5. Udey: Бесплатное учебное руководство по теме “Обработка и анализ данных” [Электронный ресурс]: 2021 — Режим доступа: <https://www.udemy.com/course/knime-bootcamp/>
6. Working With Text Data — scikit-learn 0.24.2 documentation [Электронный ресурс]: 2020 — Режим доступа: https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html
7. 3.1 Cross-validation: evaluating estimator performance — scikit-learn 0.24.2 documentation [Электронный ресурс]: 2020 — Режим доступа: https://scikit-learn.org/stable/modules/cross_validation.html
8. Receiver Operating Characteristic (ROC) with cross validation — scikit-learn 0.24.2 documentation [Электронный ресурс]: 2020 — Режим доступа: https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc_crossval.html
9. Краткий обзор алгоритма машинного обучения Метод Опорных Векторов (SVM) [Электронный ресурс]: 01.11. 2018 — Режим доступа: <https://habr.com/ru/post/428503/>
10. Welcome to Streamlit — Streamlit 0.82.0 documentation [Электронный ресурс]: 2021 — Режим доступа: <https://docs.streamlit.io/en/stable/>
11. Plotly Python Graphic Library | Python | Plotly [Электронный ресурс]: 2021 — Режим доступа: <https://plotly.com/python/>
12. Классические методы статистики: критерий хи-квадрат [Электронный ресурс]: 05.08.2012 — Режим доступа: <https://r-analytics.blogspot.com/2012/08/blog-post.html?m=1>
13. Running Chi-Square Tests with Die Roll Data in Python | By Jake Huneycutt | Towards Data Science [Электронный ресурс]: 07.05.2018 — Режим доступа: <https://towardsdatascience.com/running-chi-square-tests-in-python-with-die-roll-data-b9903817c51b>
14. Критерий согласия Пирсона χ^2 (Хи-квадрат) | statanaliz.info [Электронный ресурс]: 07.10.2019 — Режим доступа: <https://statanaliz.info/statistica/proverka-gipotez/kriterij-soglasiya-pirsona-khi-kvadrat/>
15. Random Forest Feature Importance Computed in 3 Ways with Python | MLJAR Automated Machine Learning [Электронный ресурс]: 2021 — Режим доступа: <https://mljar.com/blog/feature-importance-in-random-forest/>
16. Градиентный бустинг — просто о сложном [Электронный ресурс]: 27.11.2018 — Режим доступа: <https://neurohive.io/ru/osnovy-data-science/gradientyi-busting/>

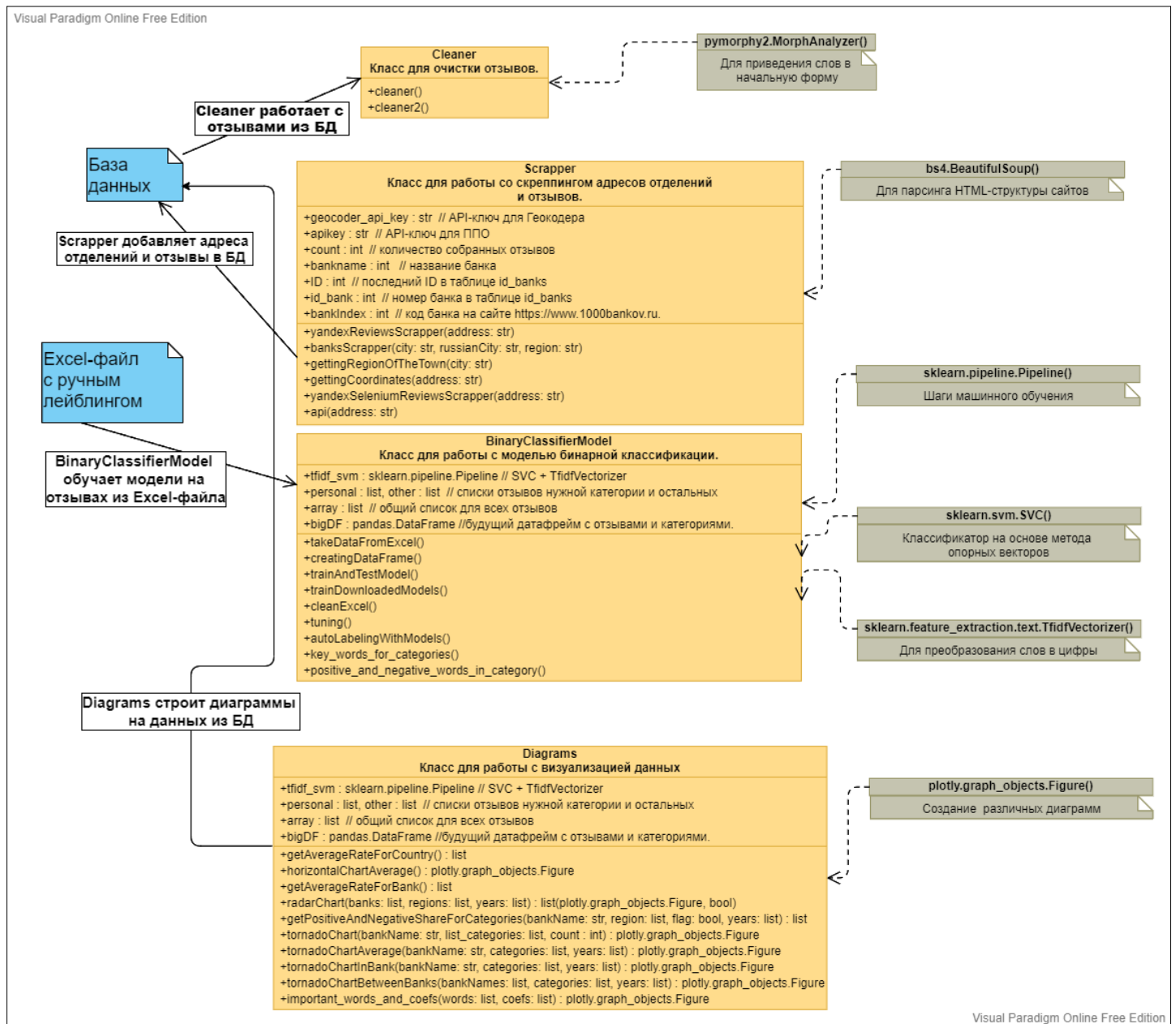
17. Статистический анализ гипотез в Python с корреляцией Anova, Хи-квадрат и Пирсона - pythobyte.com [Электронный ресурс]: 2021 — Режим доступа:
<https://pythobyte.com/statistical-hypothesis-analysis-in-python-with-anovas-chi-square-and-pearson-correlation-be15ad06/>
18. Folium — Folium 0.12.1 documentation [Электронный ресурс]: 2021 — Режим доступа:
<https://python-visualization.github.io/folium/>

6. Приложения

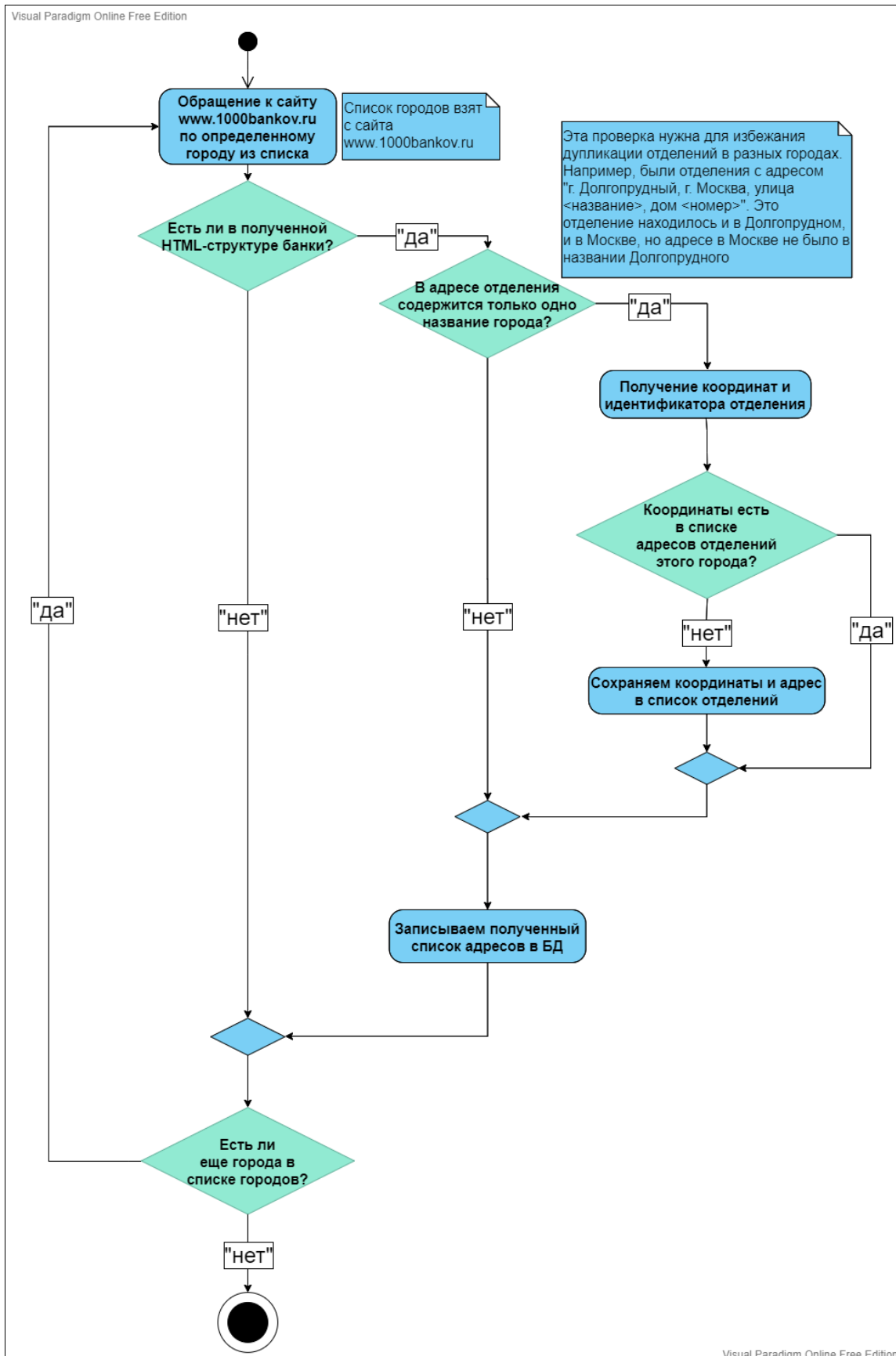
1. Sequence Diagram (диаграмма последовательностей) для сбора отделений с сайта <https://1000bankov.ru>



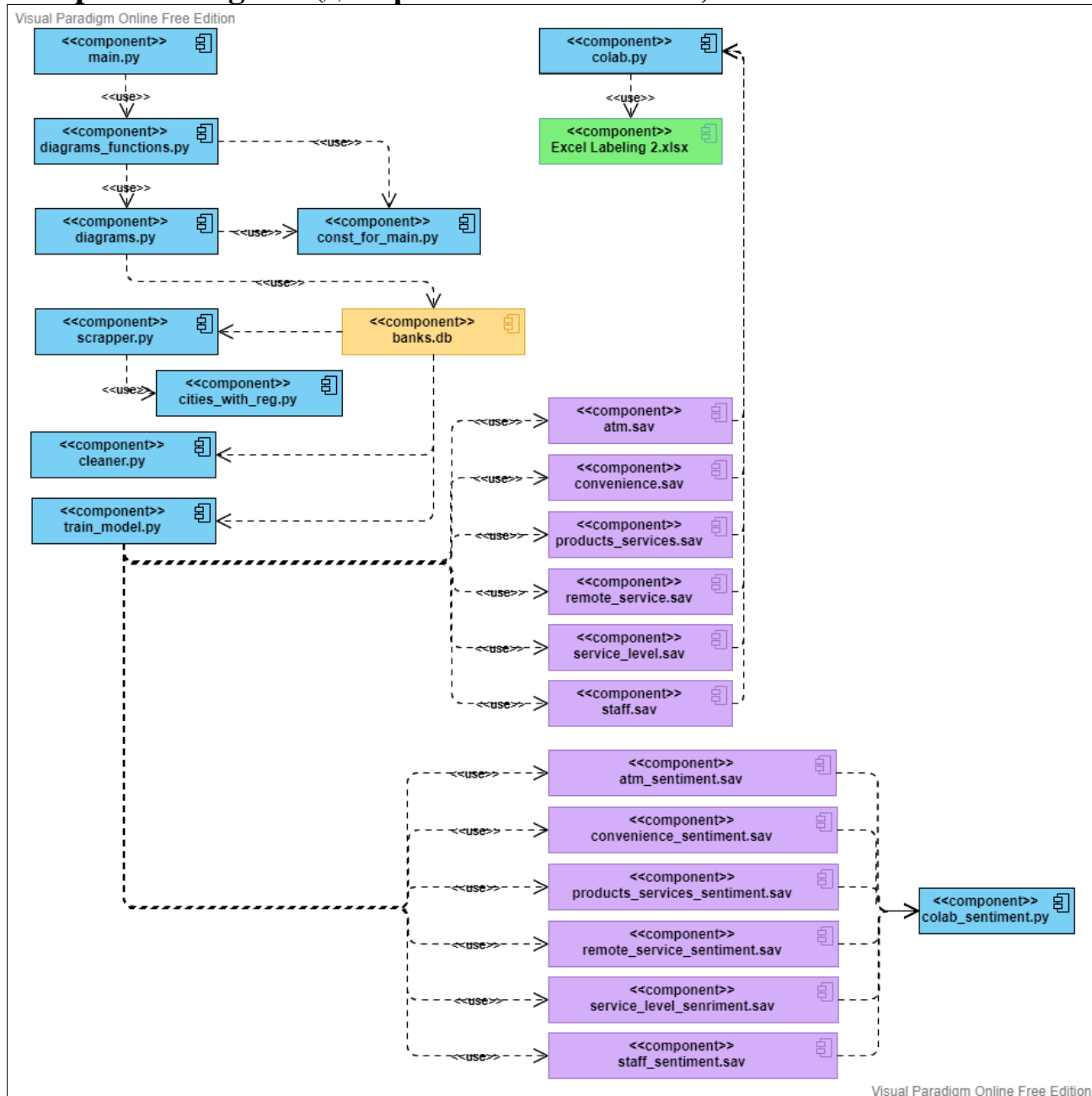
2. Class Diagram (диаграмма классов)



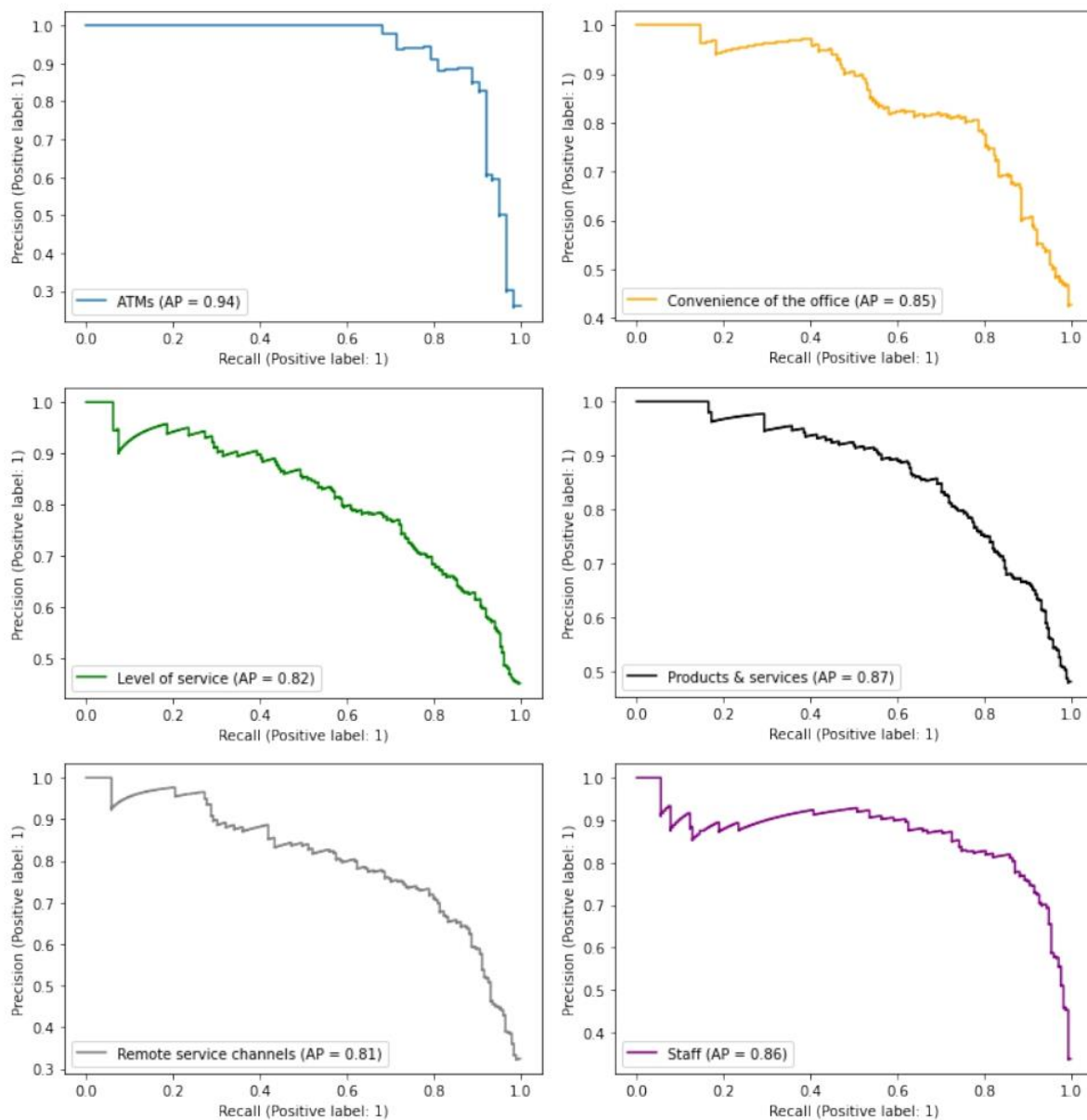
3. Activity diagram (диаграмма деятельности) для сбора отделений



4. Component diagram (диаграмма компонентов)



5. Графики precision-recall для моделей бинарной классификации



6. ROC-AUC кривая для моделей бинарной классификации

