Исследование открытых данных ФНС[1] России для прогнозирования финансового неблагополучия МСП[2]

1. ФНС – федеральная налоговая служба
2. МСП – малые и средние предприятия

Рекомендации бизнесу на основе изучения финансовых и организационных признаков МСП

Порфирьева О.В. porfirjeva.o@gmail.com
Факультет Аналитики Big Data
2023 г.

Цели проекта:

- собрать данные;
- изучить данные и обучить модель для предсказания;
- дать рекомендации;

Нестабильность экономических условий в первую очередь сказывается на малом и среднем бизнесе[1]. Узкий горизонт планирования затрудняет оценку инвестиционной привлекательности и возможности заключения контрактов и договоров.

Поэтому важно собирать и оценивать всю возможную информацию о второй стороне.

И именно здесь – в поиске и работе с информацией - полезен глубокий анализ данных и машинное обучение для выявления неявных закономерностей.

1. Например, статья на эту тему, Или ещё статья

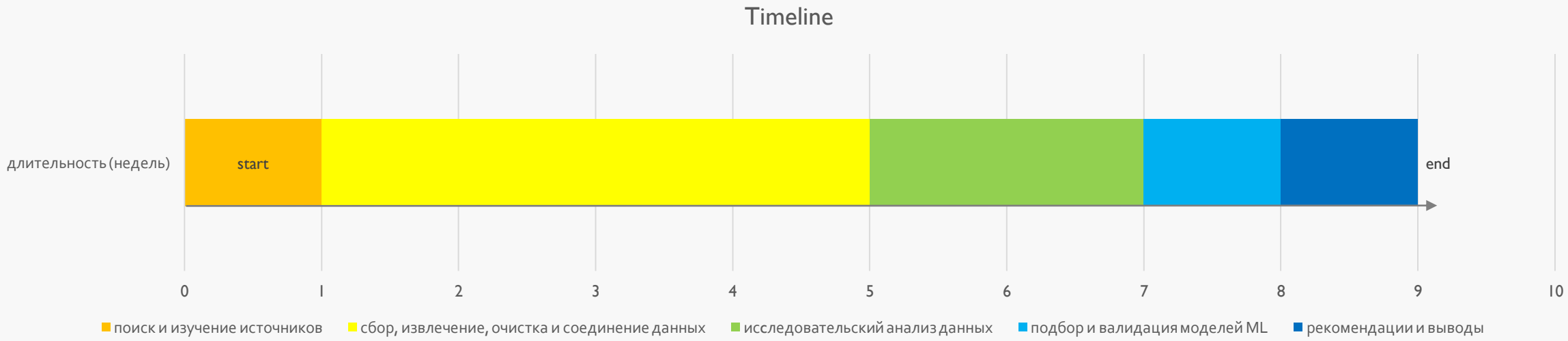
Задачи на пути к целям:

- Какие данные доступны? – Изучить открытые источники.
- В каком виде они доступны? – Изучить структуру данных, собрать, очистить и соединить.
- О чём говорят данные? – Провести исследовательский анализ данных, выбрать цель предсказания, создать новые признаке.
- Что неочевидно в данных? – Обучить модель и изучить вклад признаков
- Что можно получить из всего этого? – Сформулировать рекомендации.

> “Куда-нибудь ты обязательно попадешь, — сказал Кот. — Нужно только достаточно долго идти.”
>
> Л. Кэррол

Центр тяжести таймлайна:

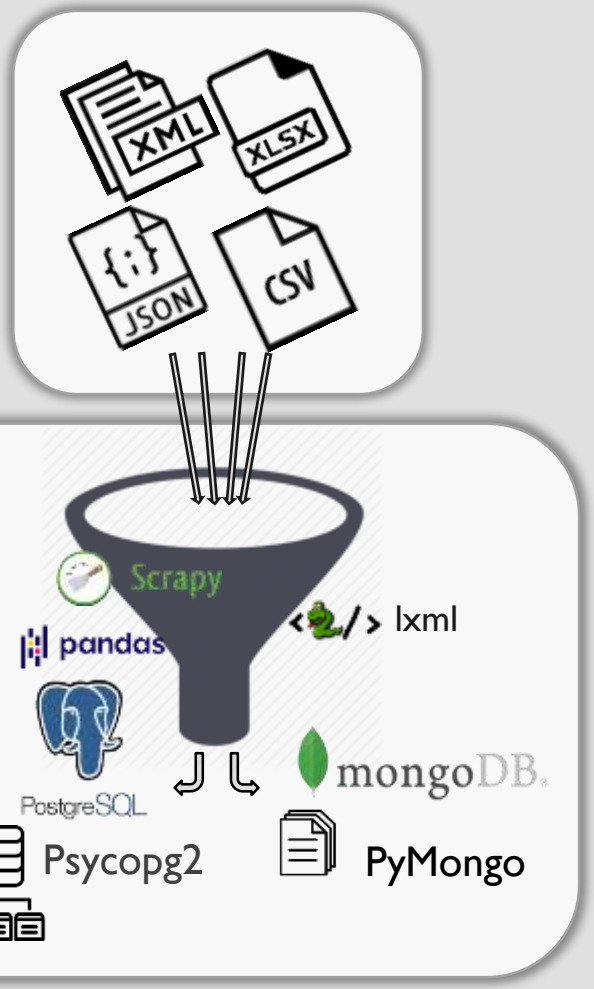
- сбор разнотипных данных из нескольких источников, их очистка и соединение.



Структура проекта

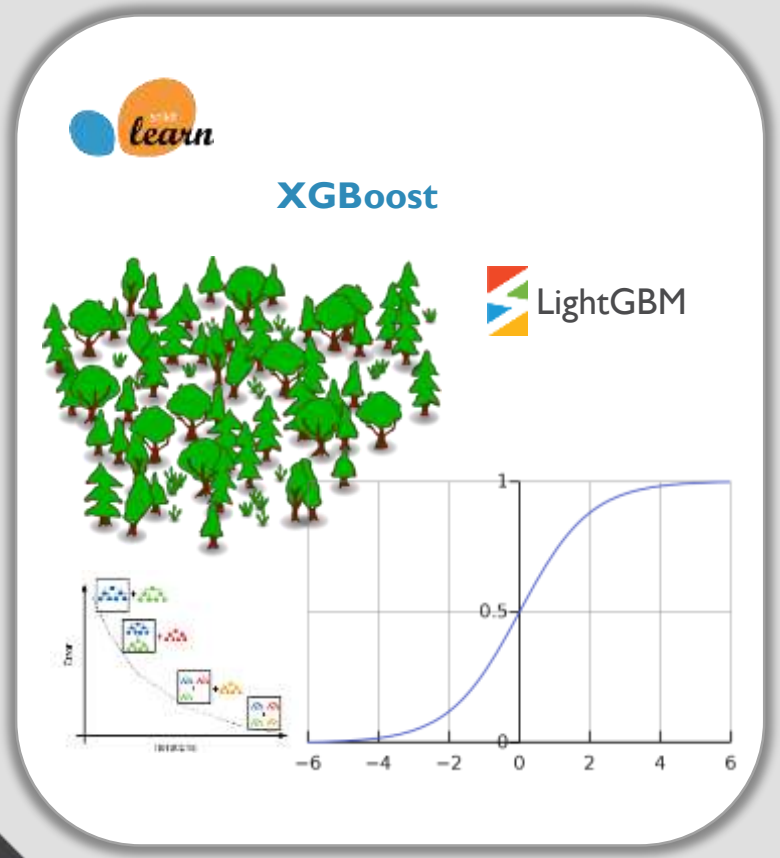
«Всему свое время, и время всякой вещи под небом…»
Экклезиаст

1. Собираем
~~камни~~ данные



2. Изучаем
~~камни~~ данные

3. ~~Разбрасываем~~
Предсказываем данные



jupyter

1. Получение данных

«Как закалялась сталь»
Н. Островский

Источник вдохновения – датасет

Company Bankruptcy Prediction

Основные источники данных:

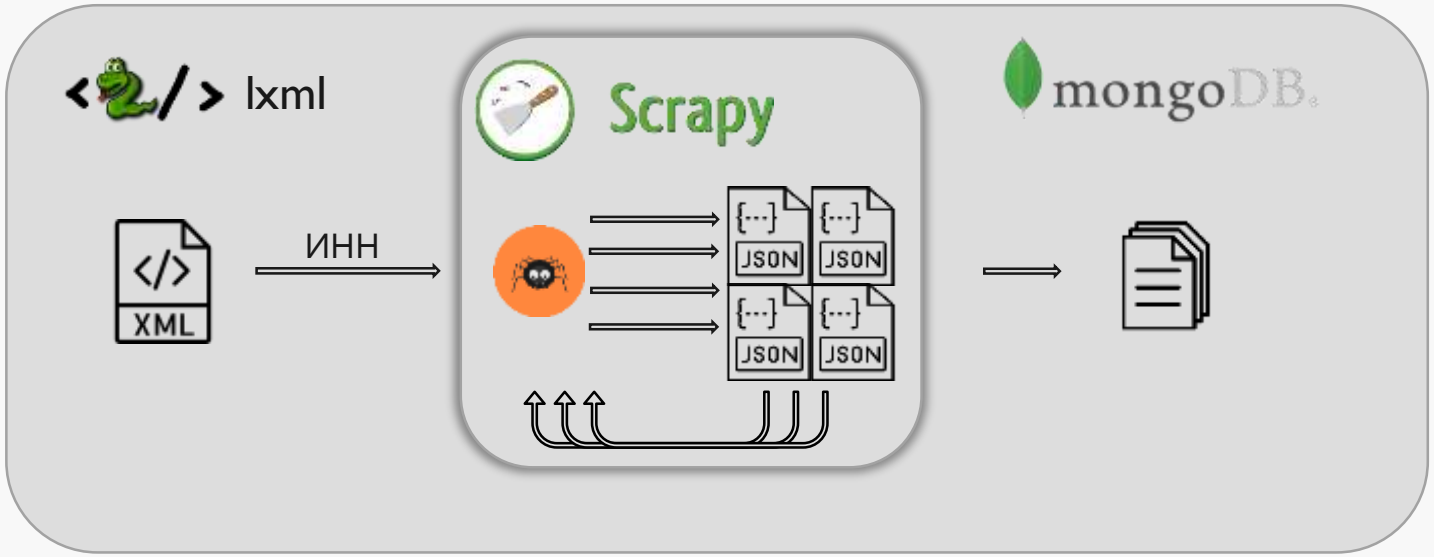
1. https://bo.nalog.ru/ : государственный информационный ресурс бухгалтерской (финансовой) отчётности.
2. https://www.nalog.gov.ru/opendata/: открытые государственные данные ФНС России

Дополнительно: https://rosstat.gov.ru/vpn_popul - итоги Всероссийской переписи населения

Особенности извлечения данных

«Не спи, не спи, работай,
Не прерывай труда»
Б.Пастернак

1. Источник: https://bo.nalog.ru/, json



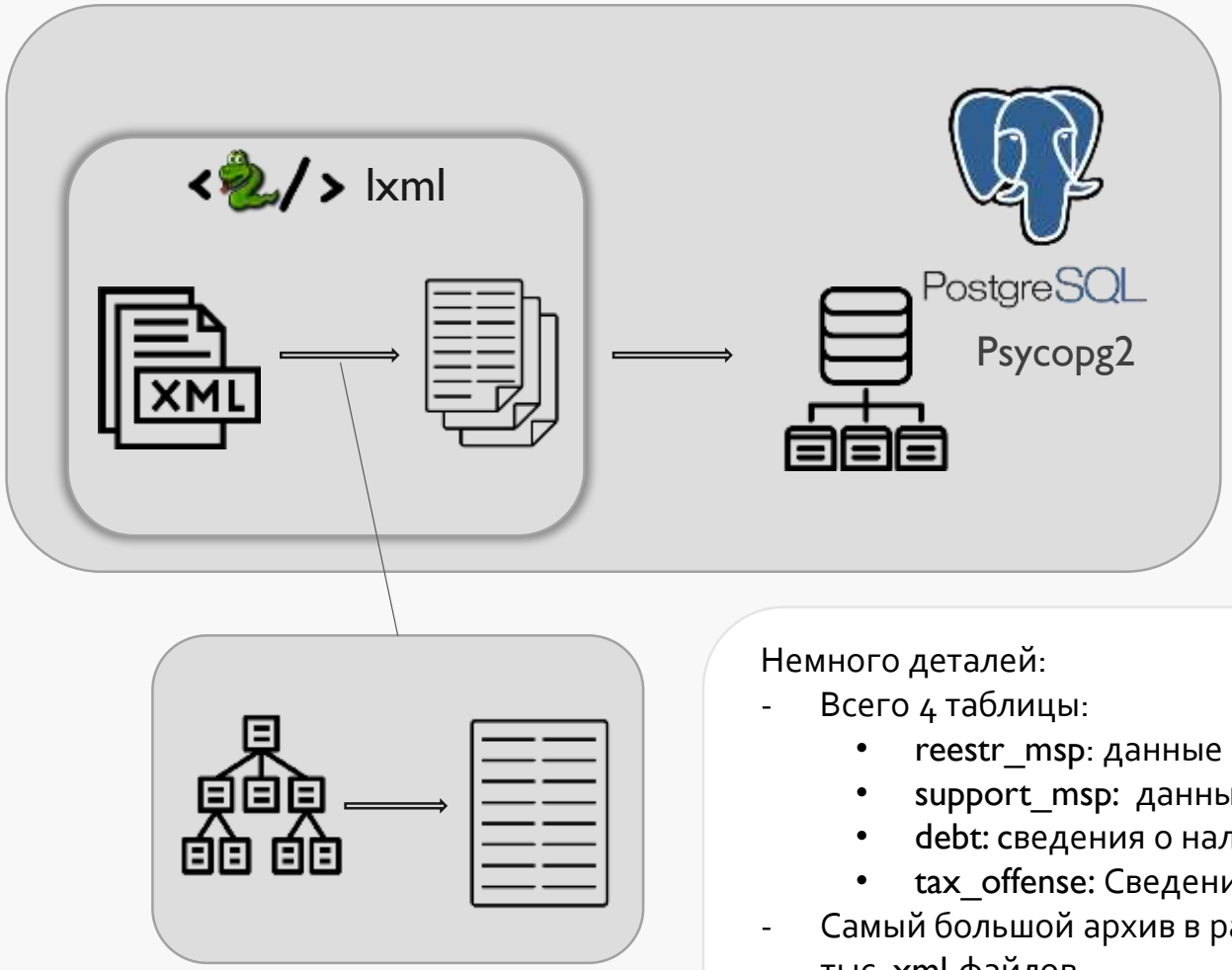
- Страница для каждой организации однозначно определяется номером ИНН. Список ИНН можно получить из реестра МСП (подробности в следующем слайде).
- Контент формируется динамически: чтобы собрать необходимый минимум информации по одному ИНН – требуется 4 последовательных запроса.
- Это наиболее продолжительный этап проекта.

Немного деталей:

- Всего в реестре около 2.3 млн. уникальных ИНН.
- В итоге получается около 9.2 млн. запросов (на самом деле меньше, так как не для всех ИНН есть отчёты в открытом доступе).
- Паук обрабатывает 50 тыс. ИНН чуть менее чем за 14 часов и получает из них около 40 тыс. документов для mongoDB.
- В результате получается около 1.9 млн. документов.
- Можно быстрее, если использовать платный API.
- Повторный проход для обновления или дополнения информации будет быстрее: в базе сохраняются необходимые для запросов id документов.
- Неплоские документы по одному за раз удобно записывать в NoSQL базу данных.

Особенности извлечения данных

2. Источник: https://www.nalog.gov.ru/opendata/, xml



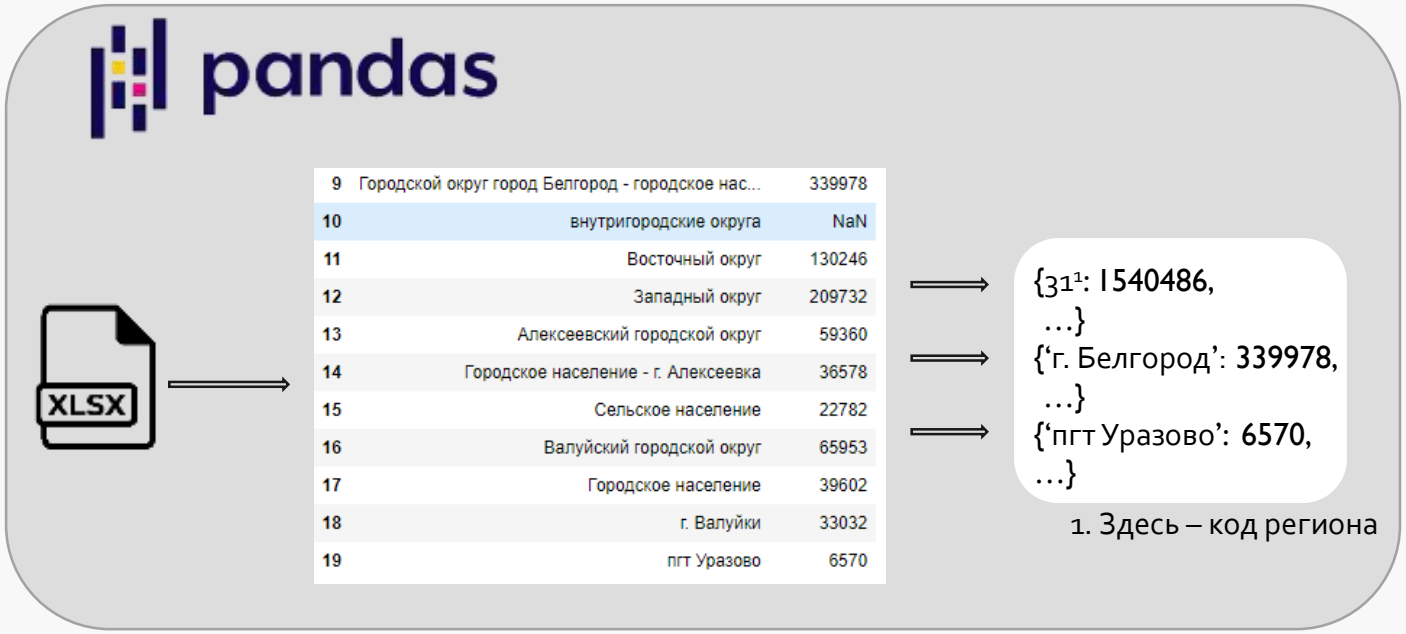
- Данные представляют собой архивы с множеством файлов xml.
- Данные имеют достаточно сложную вложенную структуру.
- Архивы можно сохранить локально, т.е. можно задать реляционную структуру и изменять её, если это необходимо.
- При пофайловой обработке удобно записывать данные пакетами.
- Можно установить связи между таблицами.

Немного деталей:

- Всего 4 таблицы:
  - reestr_msp: данные из единого реестра субъектов МСП
  - support_msp: данные о полученной субъектами поддержке
  - debt: сведения о налоговых правонарушениях
  - tax_offense: Сведения о суммах недоимки и задолженности по пеням и штрафам
- Самый большой архив в распакованном состоянии занимает около 20 Гб и включает около 7 тыс. xml файлов.
- Из одного файла можно забрать от 0 до чуть менее 1000 строк.
- Цикл обработки занимает 20-25 минут, конечная таблица занимает около 800 Мб.

Особенности извлечения данных

3. Дополнительные источники:
- https://www.nalog.gov.ru/opendata/, **csv**, данные о дисквалифицированных лицах
- https://rosstat.gov.ru/vpn_popul, **xlsx**, данные переписи населения

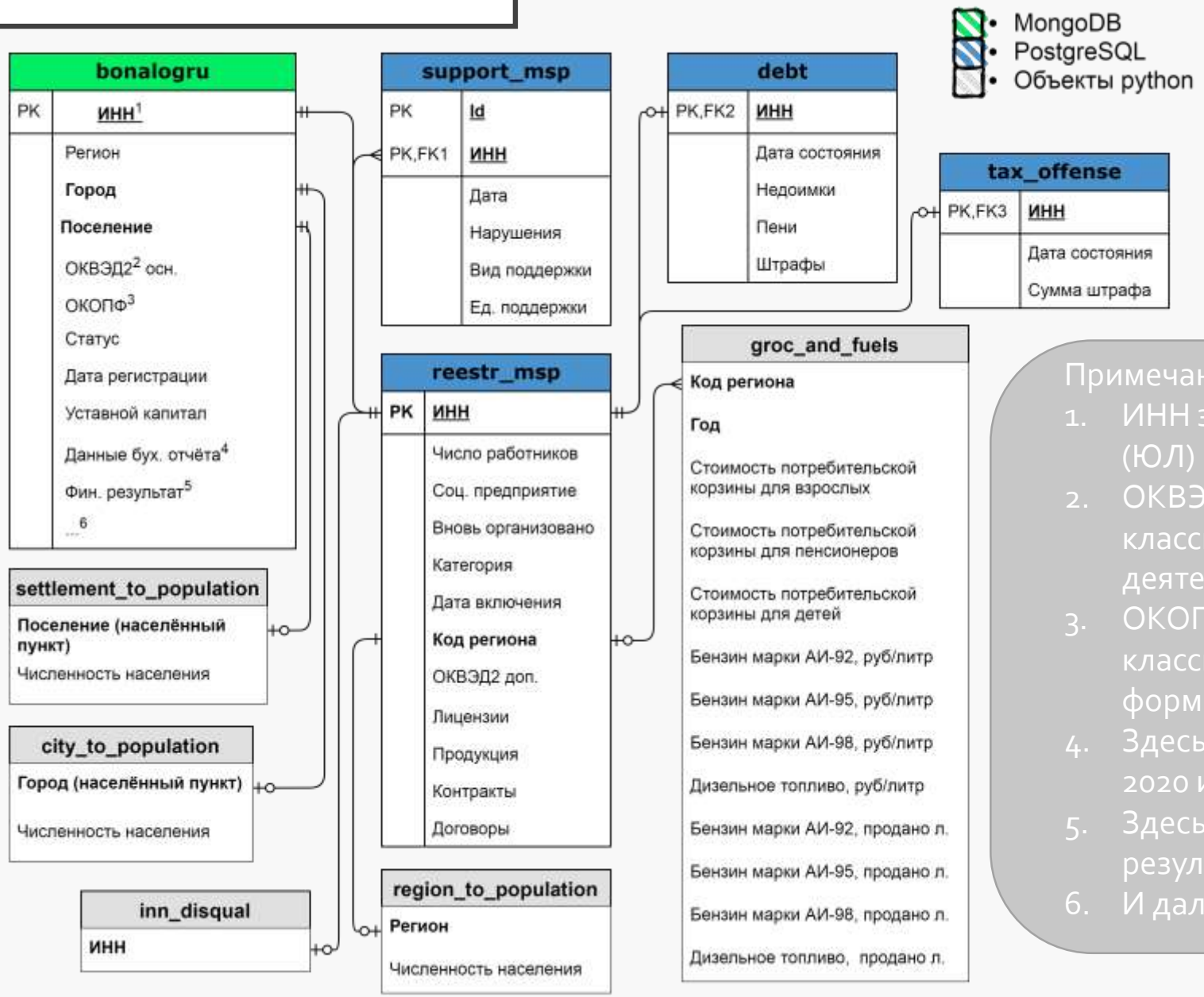


1. Здесь – код региона

- С извлечением данных из **csv** и **xlsx** без проблем справляется **pandas**.
- Сложности возникают на этапе соединения данных о переписи населения (**vpn_popul**) с основными данными.
- Таблица **vpn_popul** содержит строки разного уровня, а наименования населённых пунктов не унифицированы.
- Таким образом, приходится анализировать содержание строк в поисках нужного совпадения.
- Далее, в основных данных информация о населённых пунктах для некоторых строк указана некорректно: город в поле Регион, посёлок – в поле Город и т.д.
- Есть пропуски в данных и адреса, которые сложно интерпретировать с точки зрения численности населения ('23 Км Байкальского Тракта', 'ВНИИССОК').
- В таблице **vpn_popul** данные только для объектов с численностью от 3 000 ч-к.

- В итоге, придётся выбирать конечный населённый пункт и заполнять пропуски.
- На выходе – три словаря с численностью для региона, города и поселения
- Вопрос, насколько это будет полезно для модели, остаётся открытым.
- Наиболее чистая и полная информация – по численности для регионов.

Структура данных



«Если, согласно квантовой теории, наблюдатель создает или может частично создавать наблюдаемое, то мышь может переделать Вселенную, просто посмотрев на нее»

А. Энштейн

Примечания:

1. ИНН здесь и далее – ИНН юридических лиц (ЮЛ)
2. ОКВЭД – код по общероссийскому классификатору видов экономической деятельности
3. ОКОПФ – код по общероссийскому классификатору организационно-правовых форм собственности
4. Здесь данные бухгалтерского отчёта за 2021, 2020 и 2019 гг.
5. Здесь данные отчёта о финансовых результатах за 2021 и 2020 гг.
6. И далее – в совокупности более 150 полей

Что дальше?

«Стрижка только начата»
(«Почему у льва...»)

- на этапе соединения – 1.3 млн. сырых строк из MongoDB;
- паук всё ещё продолжает свою работу;
- итоговый датасет занимает около 0.6 Гб и содержит 124 столбца;
- можно сделать много интересных запросов к данным (по регионам, ОКВЭД, финансовым показателям...);
- а чтобы это было удобно делать – следовало бы собрать данные вместе в PostgreSQL, создать таблицы по сущностям (а не по файлам или источникам, как на этапе извлечения данных), настроить связи, создать справочники... - большой объём работы за рамками проекта.

Но чтобы показать, что собранные данные могут быть полезны – надо выбрать цель, отобрать данные и приступить к их исследованию.

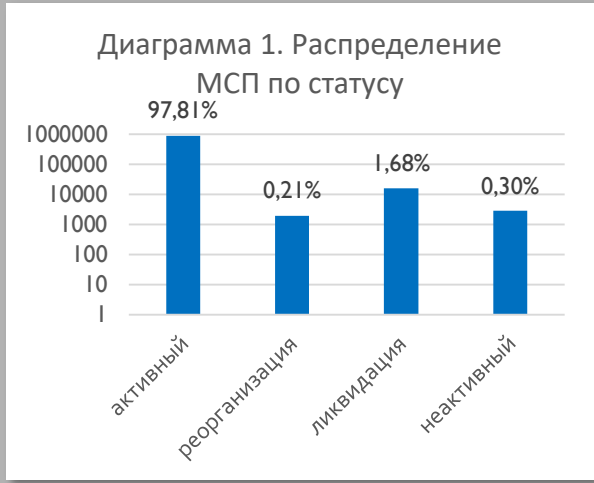
2. Анализ данных

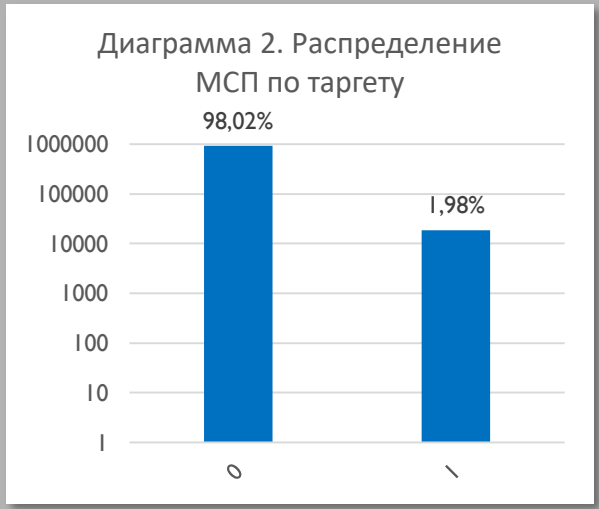
«Совы не то, чем они кажутся»
«Твин Пикс»

- О чём говорят данные?
- Насколько они точны и какие в них есть пропуски?
- Как они связаны между собой?
- Какие данные особенно важны и каких данных всё ещё не хватает?
- Какие выводы можно сделать, опираясь на данные и насколько далеко мы можем зайти в этих выводах?

Выбор цели

«В чём сила, брат?»
(«Брат»)





- Неожиданно интересное поле в данных – Статус организации - принимает одно из 4 значений:
  - активный
  - ликвидация
  - неактивный
  - реорганизация;
- Кажется, что это достаточно прямой показатель благополучия организации (помимо финансовых, где есть множество неочевидных тонкостей).
- Классы заметно несбалансированы (на Диагр. 1 логарифмическое отображение по шкале **y**);
- Но можно немного облегчить задачу предсказания, взяв за тарегт объединённые классы – см. Диагр. 2, где класс «0» объединяет «активный» и «реорганизация», а класс «1» – «неактивный» и «ликвидация»
- Учитывая особенности обновления данных – можно использовать данные за 20-21 гг., чтобы предсказать статус организации по состоянию на 22- начало 23 гг.

Очистка, исследование и генерация признаков

Бесконечно можно делать три вещи: смотреть на бегущую воду, на горящий огонь и вглядываться в данные.

Как извлечь максимальную пользу из данных?

- заполнение пропусков
- изучение распределения и статистик
- обрезка выбросов
- генерация новых признаков
- исследование связей с другими признаками и таргетом

Немного о том, как всё взаимосвязано

«И если ты долго смотришь в ~~данные~~ бездну, то бездна тоже смотрит в тебя» Ф. Ницше

Диаграмма 3. Вероятность статуса предприятия[1] в зависимости от изменения его категории

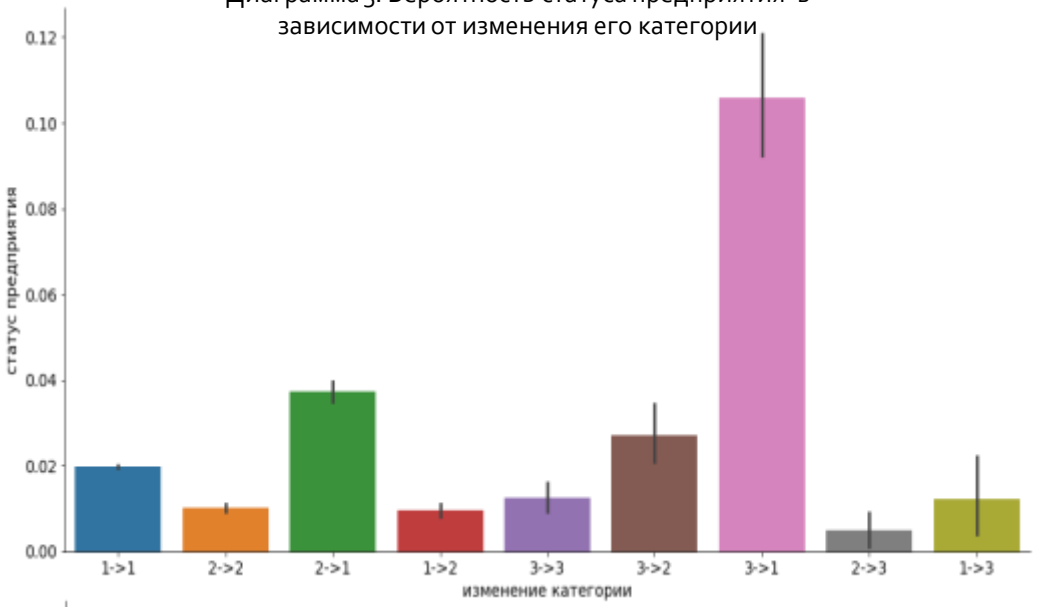
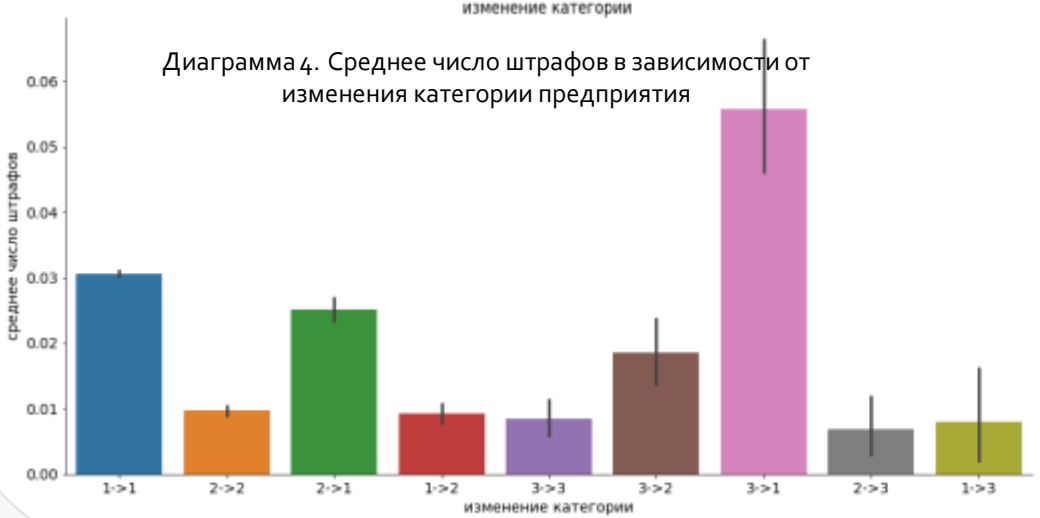


Диаграмма 4. Среднее число штрафов в зависимости от изменения категории предприятия



- Интересный для анализа признак – категория предприятия:
  - 1 - Микро: не более 15 человек, доход - не более 120 млн
  - 2 - Малое: не более 100, не более 800 млн
  - 3 - Среднее: не более 250, не более 2 млрд
- Но категория предприятия изменяется, только если в течение 3-х лет предприятие не соответствует критериям (и относительно вероятности таргета она не так показательна)
- При этом отдельно есть среднесписочная численность работников предприятия, которая подаётся за истекший год.
- На пересечении этих двух признаков возникает новая информация:
  - 1->1 – микропредприятия, сохранившие свою численность на прежнем уровне
  - 1->2 – микропредприятия, нарастившие численность до уровня малого… И т.д.
- И по этим группам наблюдается достаточно очевидная разница в среднем значении таргета (Диаграмма 3.):
  - Для средних предприятий, резко сокративших свою численность (3->1) – наибольшая опасность оказаться в неактивном статусе или статусе ликвидации.
  - На втором месте в группе риска - категория 2-> 1.
  - Т.е. в общем случае – снижение числа работников – угрожающий признак, увеличение или стабильность – неплохой знак.
  - Особняком стоит группа 1->1: микропредприятиям просто особо некуда дальше сокращаться, и их больше всего среди всех категорий.
- Схожим образом распределяется среднее число налоговых штрафов (Диаграмма 4.) и средняя сумма штрафов (за исключением группы «1->1» – наиболее низкие суммы штрафов, диаграмму можно увидеть в приложенном ноутбуке)

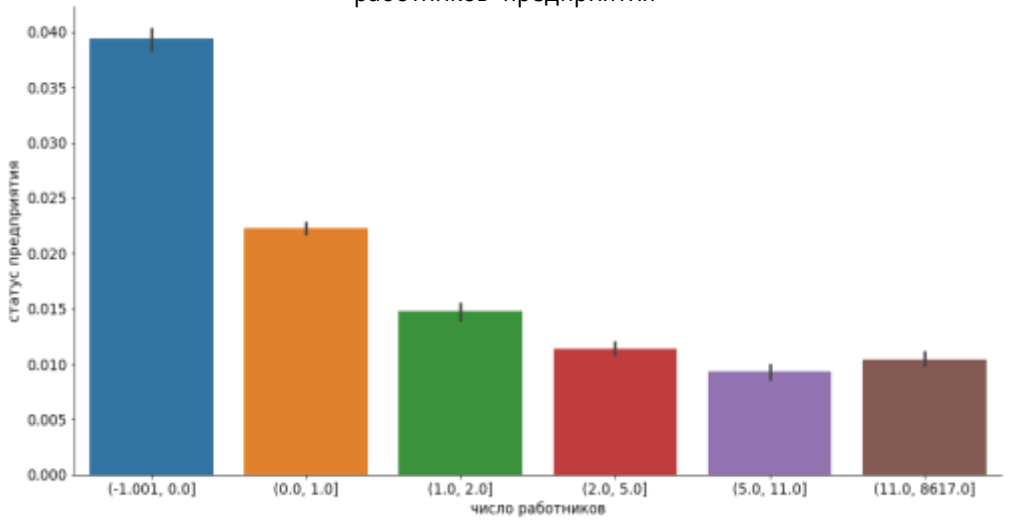
Примечания:

1. Статус предприятия: 1- неактивно или ликвидация, 0 – активно или реорганизация

Немного о том, как всё взаимосвязано. Численность

«Во всём мне хочется дойти до самой сути»
Б.Пастернак

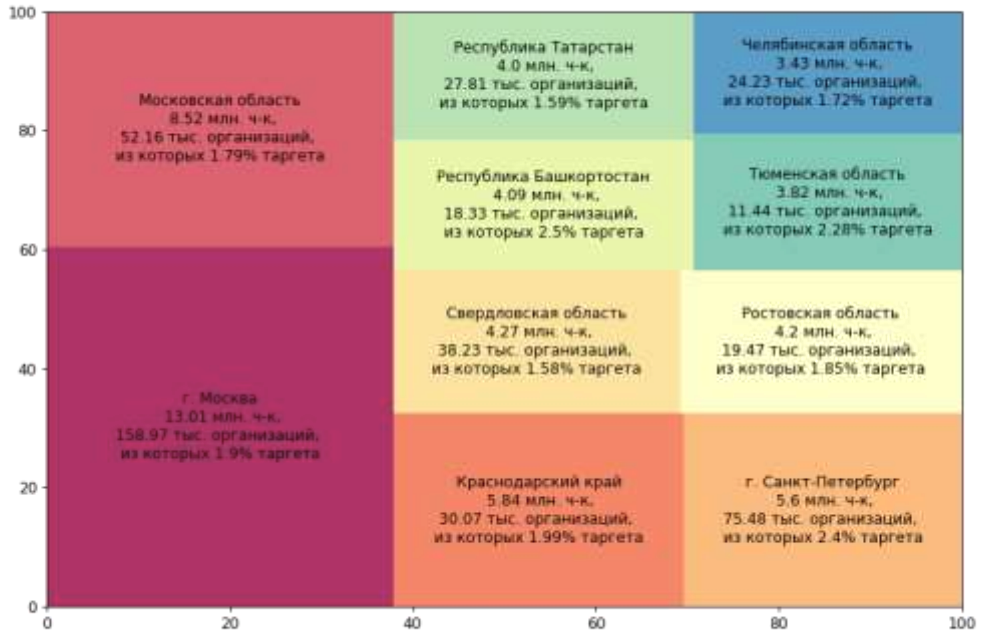
Диаграмма 5. Вероятность таргета в зависимости от числа работников[1] предприятия



Если пристальнее вглядеться в численность работников, можно увидеть:

- Наибольшая вероятность быть неактивными у предприятий, не подавших вовремя сведения о численности;
- Вероятность таргета постепенно снижается с увеличением числа работников;
- Оптимальная численность – от 5 до 11 человек, но в общем случае – не менее 5.
- Эта закономерность наблюдается прсктически для всех регионов – больше диаграмм можно увидеть в соответствующем ноутбуке.

Диаграмма 6. Соотношение численности населения и числа организаций для топ-10 регионов[2]



Ещё о численности:

- Площадь ячеек диаграммы отражает численность населения по региону
- Топ-10 регионов по численности населения практически совпадает с топ-10 регионов по численности организаций
- % таргета рассчитан относительно числа организаций в регионе
- Наибольшая концентрация неактивных предприятий – респ. Башкортостан, г. Санкт-Петербург, Тюменская обл.
- Т.е. помимо конкуренции (числа организаций в регионе) – есть другие факторы, влияющие на устойчивость компаний.

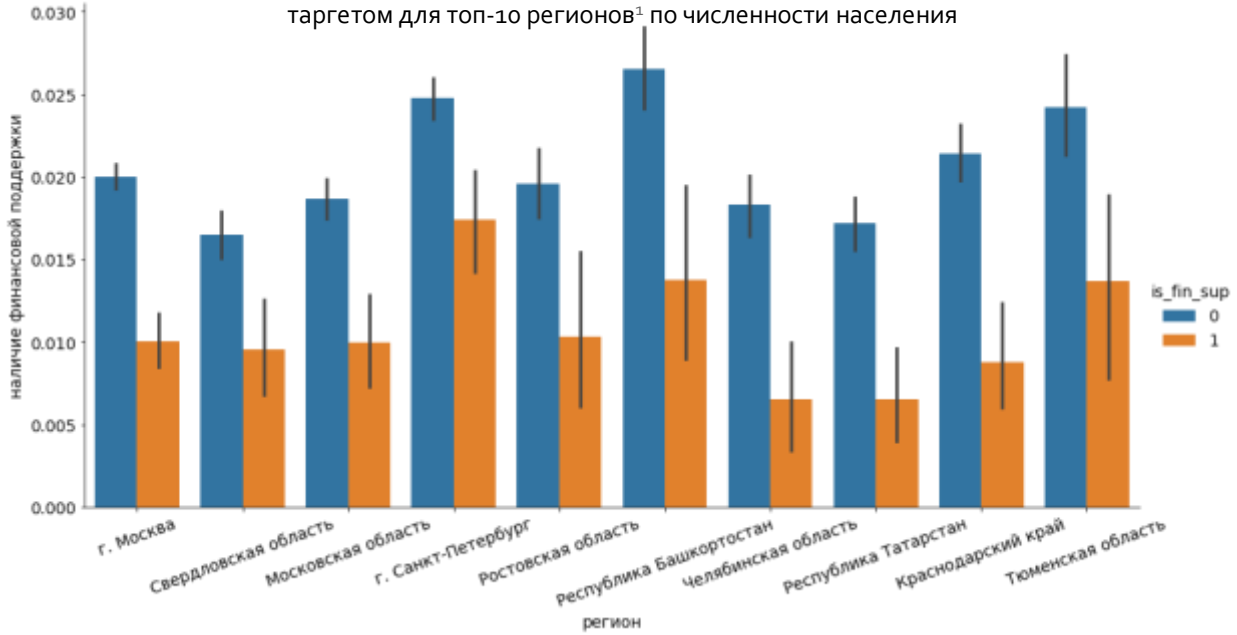
Примечания:

1. Числом -1 закодированы предприятия, вовремя не подавшие сведения о численности
2. Топ-10 по численности населения

Немного о том, как всё взаимосвязано. Финансовая поддержка

«Нужно бежать со всех ног, чтобы только оставаться на месте, а чтобы куда-то попасть, надо бежать как минимум вдвое быстрее!»
Л. Кэррол

Диаграмма 7. Взаимосвязь факта финансовой поддержки с таргетом для топ-10 регионов[1] по численности населения



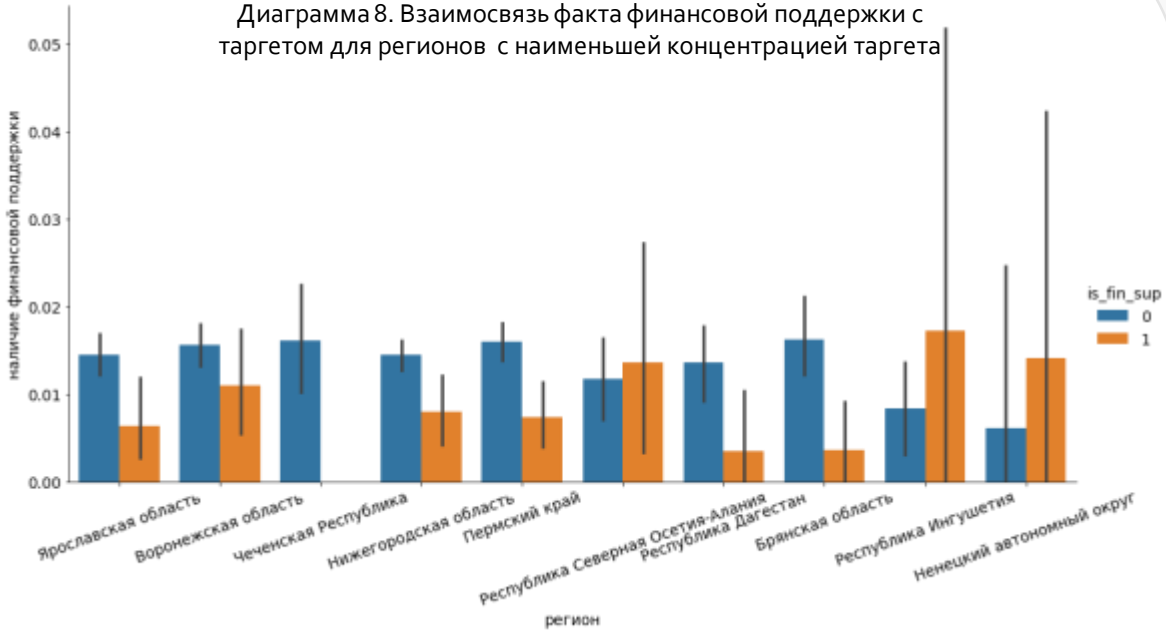
Финансовая поддержка – важный признак:

- Тенденция прослеживается достаточно отчётливо практически для всех регионов, не только для топ-10: наличие финансовой поддержки – благоприятный прогностический признак.
- Аналогичная взаимосвязь - для размера финансовой поддержки
- Интересно посмотреть, для каких регионов это не так – Диаграмма 8.

Некоторые взаимосвязи меняются со временем:

- Так, управленческие расходы организации в 2020 г. сильно коррелируют с чистой прибылью в 2020 же (отрицательная связь: -0.97)
- В 2021 г. такой связи нет (коэф-т 0.01) , но есть сильная положительная связь управленческих расходов с выручкой (коэф-т 0.96)
- А с таргетом из всех финансовых показателей наиболее тесно связаны данные за 2020 г. – прочие расходы (с коэффициентом 0.012).

Диаграмма 8. Взаимосвязь факта финансовой поддержки с таргетом для регионов с наименьшей концентрацией таргета



- Здесь отображены 10 регионов с наименьшей долей неактивных и ликвидируемых организацией относительно общего числа организаций региона;\
- Интересные регионы – респ. Ингушетия, Северная Осетия-Алания и Ненецкий автономный округ, в которых финансовая поддержка не так эффективна (хотя нельзя сделать однозначные выводы из-за малого числа примеров – полоса ошибка слишком велика)
- В целом для менее крупных регионов доля организаций, получающих поддержку, меньше.

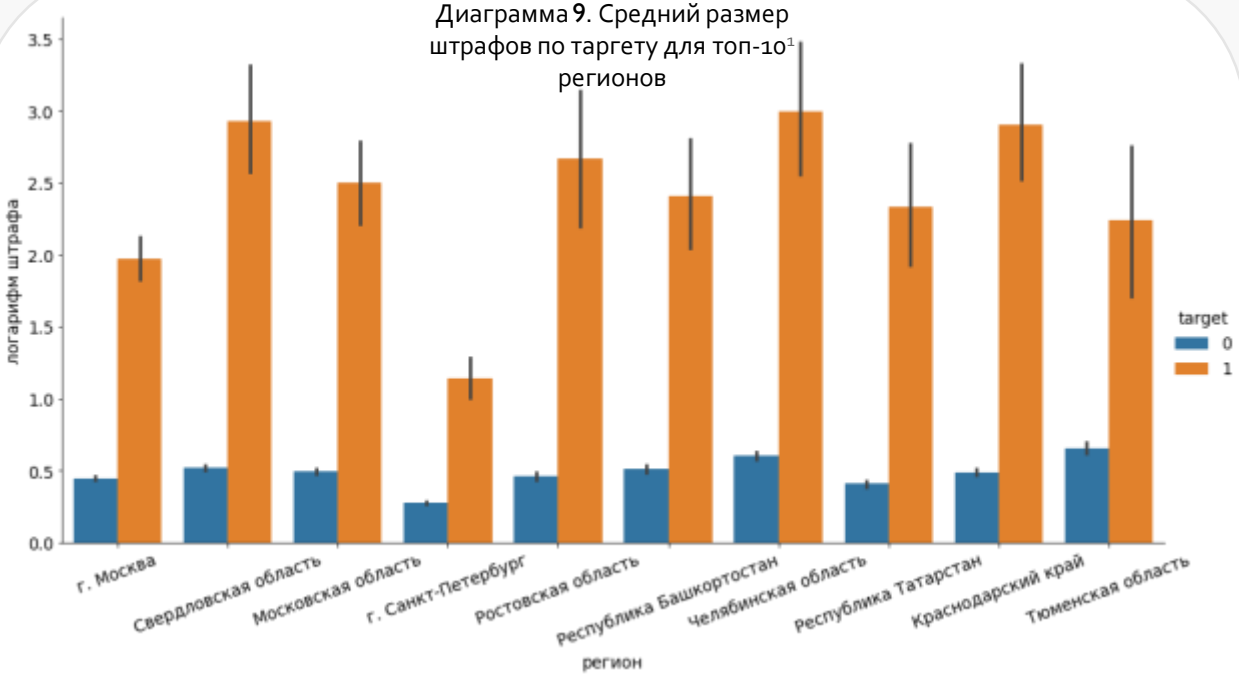
Примечания:

1. Похожие диаграммы по городам можно найти в соответствующих ноутбуках.

Немного о том, как всё взаимосвязано. Налоговые штрафы

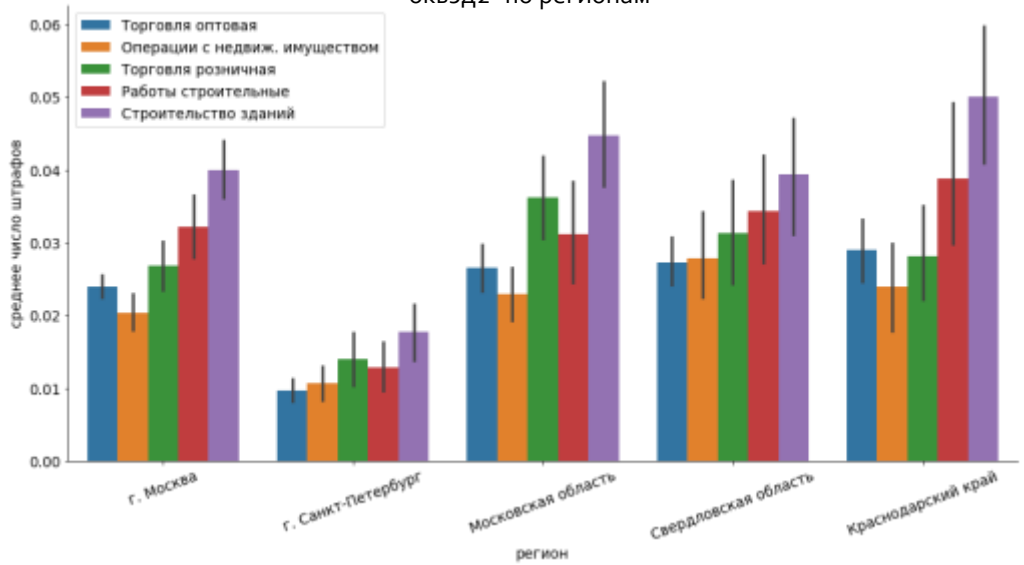
Наличие штрафов и задолженностей по налогам – один из самых сильных признаков финансового неблагополучия компании.
Но это достаточно «долгий» признак – у компании должна быть определённая история.

Диаграмма 9. Средний размер штрафов по таргету для топ-10[1] регионов



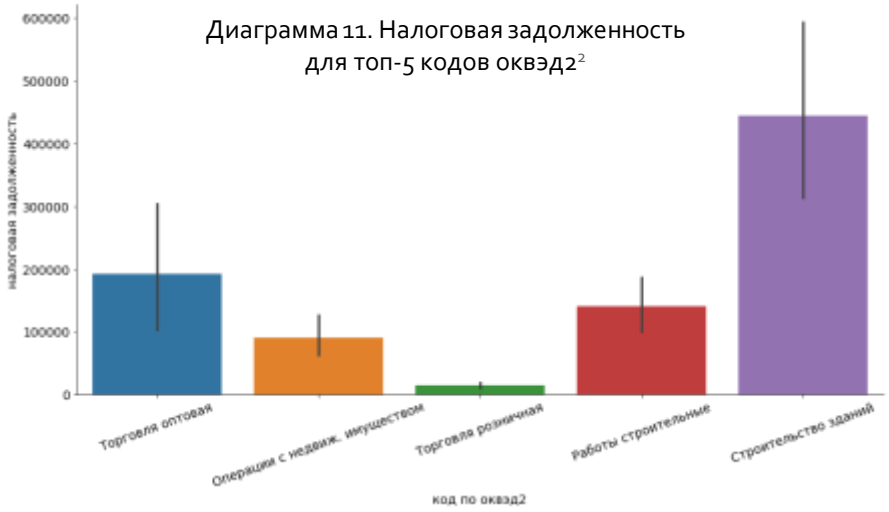
- В среднем у неактивных компаний заметно больший размер штрафов (шкала y отображена в логарифмическом масштабе).
- Схожим образом выглядят распределения для задолженностей и пени (распределения как сумм, так и факта наличия).

Диаграмма 10. Среднее число штрафов для топ-5 кодов по оквэд2[2] по регионам



- Но число и сумма штрафов и задолженностей различаются для организаций разных направленностей.
- Так, например, строительство зданий[3] – наиболее рискованный вид деятельности.
- Для других видов – может быть своя специфика по регионам и городам, но в целом тенденция повторяется и по суммам задолженностей и штрафов (за некоторым исключением – в розничной торговле наименьшие суммы штрафов) – Диаграмма 11.

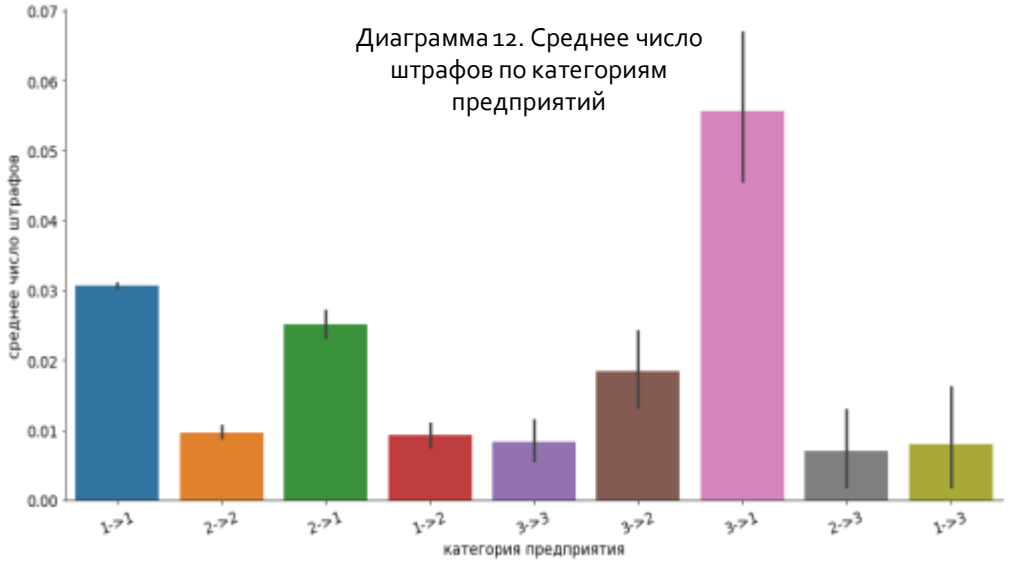
Диаграмма 11. Налоговая задолженность для топ-5 кодов оквэд2[2]



Примечания:
1. Топ-10 по численности населения
2. Топ-5 кодов по числу организаций
3. Больше диаграмм по сфере деятельности Строительство зданий – в соотв. ноутбуке

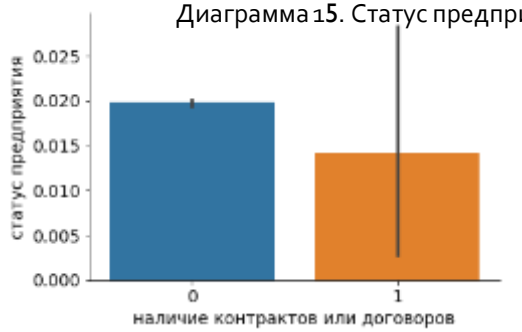
Немного о том, как всё взаимосвязано. Налоговые штрафы, топливо, контракты.

Диаграмма 12. Среднее число штрафов по категориям предприятий



- И замыкая круг – число штрафов у тех, кто сокращает численность работников, заметно выше. То же самое с суммами штрафов (за исключением того, что в группе микропредприятий самые низкие штрафы).

Диаграмма 15. Статус предприятия в зависимости от контрактов



Наличие муниципальных контрактов или договоров (по №44-ФЗ и №223-ФЗ) – скорее благоприятный (хоть и редкий) признак.

Диаграмма 13. Статус предприятия по нормированному объёму продаж дизельного топлива[2]

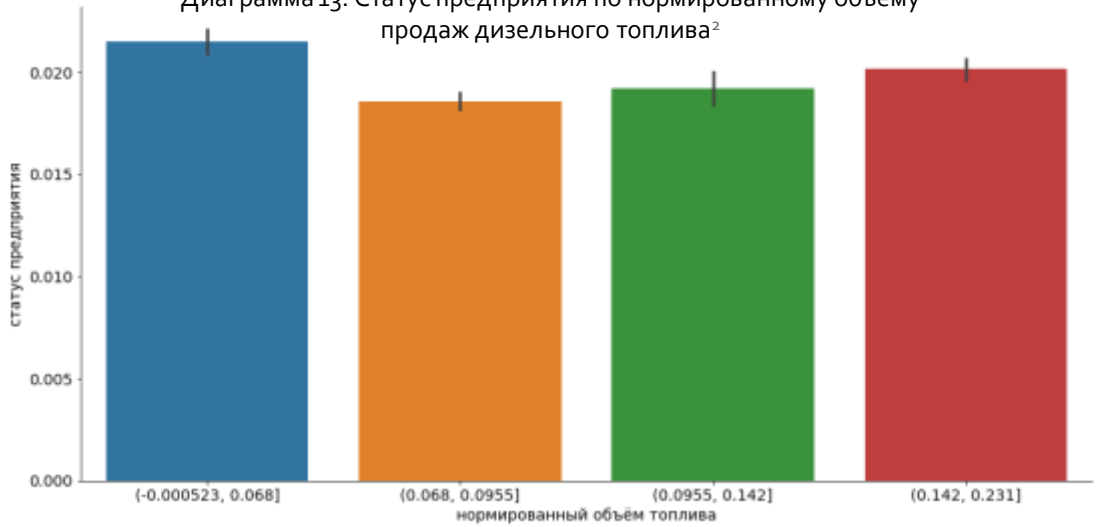
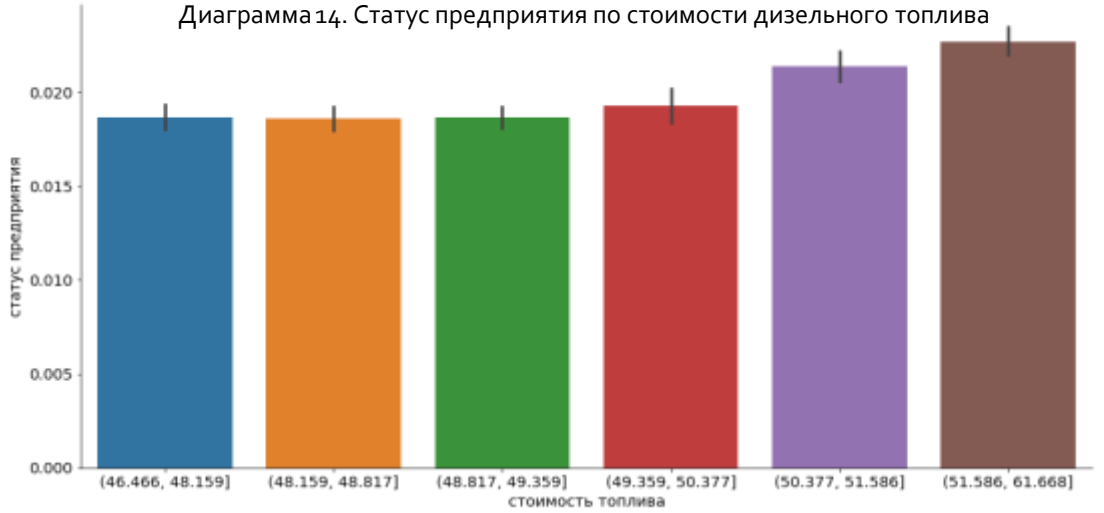


Диаграмма 14. Статус предприятия по стоимости дизельного топлива



Диаграммы 13 и 14 отражают слабую тенденцию (для других видов топлива – в целом аналогично):

- в регионах с наименьшими объёмами продаж дизельного топлива на душу населения – наибольшая вероятность таргета.
- В регионах с высокой средней стоимостью дизеля чуть выше вероятность таргета.

Примечания:

1. 1 – микро, не более 15 ч-к; 2 - малое: не более 100 ч-к; 3 – среднее, не более 250 ч-к
2. Объём, нормированный на численность по региону (л/ч-к)

Уже сейчас можно сделать первые выводы

«Связывать реальность с закономерностью – значит определять её довольно-таки произвольным образом»

П. Фейерабенд

В зоне риска:

- численность работников менее 4, особенно резко сократившее численность;
- с налоговыми задолженностями или нарушениями;
- не получает финансовую поддержку;
- не имеет муниципальных контрактов или договоров;
- занимается строительством зданий;
- в регионе с дорогим топливом и низким уровнем потребления топлива на человека.

Идеальный вариант:

- численность работников не менее 4, растущее (или, по крайней мере, стабильное) по численности);
- без налоговых задолженностей и нарушений;
- получает финансовую поддержку;
- имеет контракты или договоры;
- не строительное.
- расположено в регионе с не очень дорогим топливом и достаточно высоким уровнем потребления топлива на человека;

Вопросы, которые стоит задать для оценки перспектив:

- Какова численность работников на предприятии?
- Как менялась численность за последний год?
- Есть ли налоговые нарушения или задолженности
- Есть ли финансовая поддержка или муниципальные контракты и договоры?
- Как чувствуют себя схожие (по экономической форме деятельности, форме собственности, размеру и т.д.) предприятия в данном регионе (городе)?
- Каково экономическое состояние региона?

- Отдельно стоит уделить больше внимания финансовым признакам (здесь они практически не рассмотрены).
- Важно учитывать локальные и глобальные исторические события, влияние которых может перекрывать все найденные тенденции.

3. Модели ML

«Я познание сделал своим ремеслом»
О.Хайям

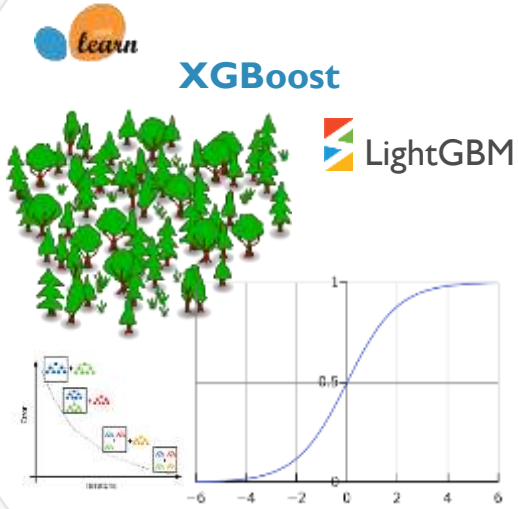
- Какой прогноз мы можем сделать?
- Насколько точен он может быть?
- Как и для чего можно использовать предсказание?
- Каков возможный потенциал у моделей ML?

Постановка задачи и выбор моделей.

«Земную жизнь пройдя до половины, Я очутился в сумрачном лесу»

Данте Алигьери

- Решаем задачу классификации.
- Учитывая значительный дисбаланс классов и особенности собранной информации , вряд ли удастся обучить модель достаточно чувствительную, чтобы отловить существенную часть целевого класса.
- А, значит, предсказание модели можно будет использовать как ещё один фактор в комплексной оценке перспектив



- Как отправную точку и для отбора признаков – возьмём **Linear Regression** и **Random Forest** (достаточно простые и быстро отрабатывающие модели);
- Для валидации и подбора гиперапараметров используем градиентный бустинг (**XGBoost** и **LightGBM** – всё ещё достаточно быстрые модели) .

- После очистки данных и генерации новых признаков – около 370 полей в датасете;
- Значительная часть из них тесно связана и должна быть исключена на этапе отбора признаков;
- После нескольких итераций отбора – в датасете 128 полей.

Обучение и метрики.

«Мене, текел, фарес»
Ветхий Завет

Отправная точка:

Таблица 1. Метрики стартовых моделей до и после отбора признаков

Модель	ROC-AUC	Точность	Полнота	F-score
LR до отбора	0.761	0.374	0.181	0.244
RF (oob[1])	0.708	0.431	0.196	0.269
LR после отбора	0.764	0.356	0.184	0.243
RF (oob) после отбора	0.707	0.479	0.187	0.269

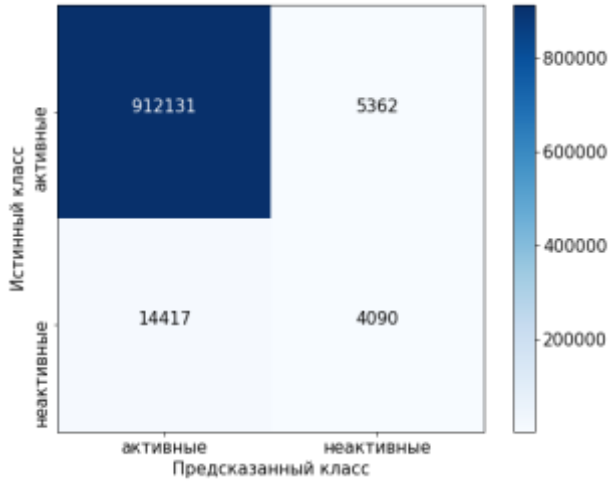
- Определённый потенциал у моделей есть, но требуется больше внимания к признакам;
- Модели недостаточно чувствительны к 1 классу (неактивные предприятия и предприятия в стадии ликвидации), но относительно неплохо отбирают класс 0.

Итоговая точка:

Таблица 2. Итоговые метрики после подбора параметров

Модель	ROC-AUC	Точность	Полнота	F-score
LGBM train	0.835	0.433	0.221	0.293
LGBM валидация	0.798	0.050	0.697	0.093
XGBoost train	0.826	0.502	0.232	0.317
XGBoost валидация	0.793	0.789	0.130	0.223

Диаграмма 15. Матрица ошибок для LGBM train



- Модель верно угадывает лишь около 1/5 неактивных предприятий (4 тыс. из 18.5 тыс.)
- Ситуацию можно немного улучшить, если подобрать соответствующий ß-коэффициент и максимизировать метрики по f1-score
- Но как предварительный фильтр модель может быть использована даже в таком виде

- LGBM чуть лучше по ROC-AUC (и на трейне и на валидации) и заметно быстрее;
- XGBoost немного переобучается на стандартных параметрах, чуть лучше по полноте и точности

Примечания:
1. out of bag, так как Random Forest сильно переобучается на тренировочной выборке.

4. Выводы

«Мы строили, строили и, наконец, построили! Ура!»
«Чебурашка и крокодил Гена»

1. Основная цель проекта - сбор разнотипных данных из нескольких источников, их очистка и соединение достигнута:

- Для датасета отобрано около 1.3 млн строк и 124 столбца (но в сырых данных их больше)
- Собранные данные позволяют делать интересующие выборки по регионам, ОКВЭД, финансовым показателям и т.д.
- Но чтобы это удобнее было делать - следовало бы собрать данные вместе в PostgreSQL, создать таблицы по сущностям, настроить связи, создать справочники

2. Анализ данных позволяет выявить некоторые тенденции:

- Характеристики распределения признаков могут колебаться от региона к региону, в рамках вида деятельности или категории предприятия, а замеченные тенденции могут быть слабыми – но большое число наблюдений позволяет говорить о них с достаточной долей уверенности.
- Среди важных для оценки будущего статуса предприятия признаков – численность работников и её динамика, наличие налоговых нарушений, задолженностей и штрафов, наличие финансовой поддержки и муниципальных контрактов и договоров.
- Важно также сравнивать конкретное предприятие со схожими по региону (городу).
- Вероятно, также было бы полезным большее внимание к финансовым показателям.

3. Использование ML имеет некоторый потенциал:

- Модели недостаточно чувствительны к 1 классу (неактивные предприятия и предприятия в стадии ликвидации), но могут быть использованы для предварительного отбора или в качестве ещё одного фактора при принятии решений
- Вероятно, можно улучшить качество предсказаний, если уделить большее внимание признакам и использовать более сложные модели

4. Предсказание статуса предприятия – не единственная возможная цель:

- Можно задаться вопросом: а кому дают финансовую поддержку?
- Или ставить задачу регрессии и прогнозировать финансовый рост/падение.
- Ещё одно интересное поле – задолженности по кредитам.
- Датасет (объём данных в том числе) позволяет также отобрать компании только по интересующему направлению деятельности или размеру и делать более прицельные выводы и предсказания.