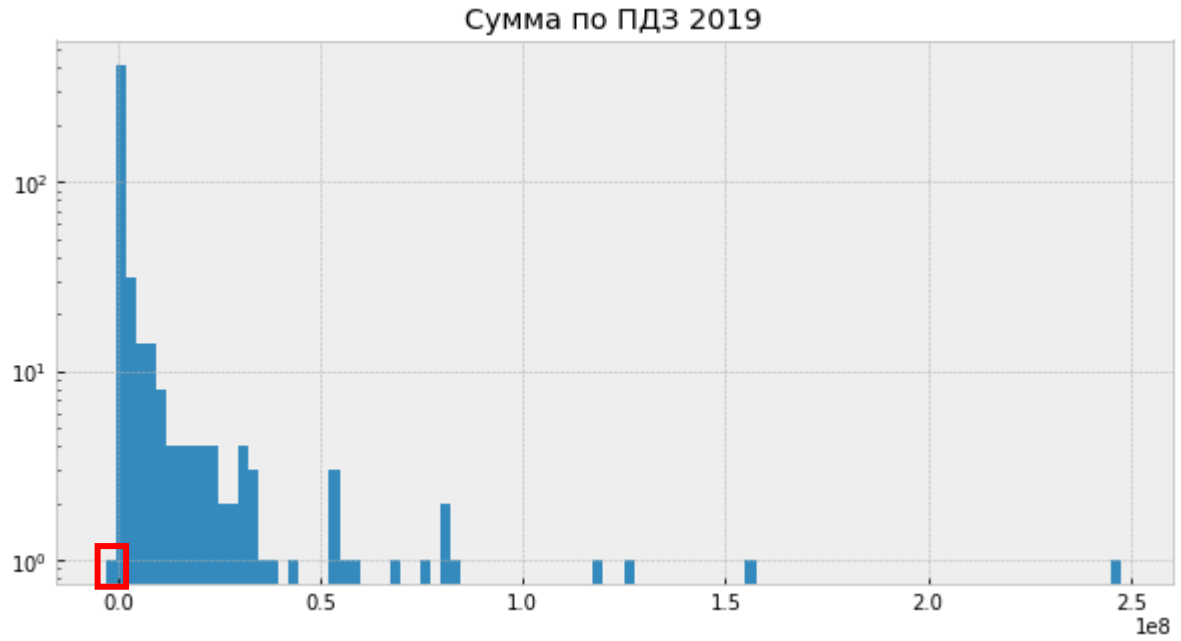
4. Анализ Контрагентов

Команда McSteel

1. Первичный анализ

Первичный анализ показал, что в данных нет явных корреляций с целевыми признаками, поэтому было принято решение о добавлении новых признаков и агрегация исходных.

Также были обнаружены строки, которые пометили как ошибки: отрицательное значение ПДЗ.



2. Дополнительные признаки

На каждый признак, который имеет историю (три года):

1. Добавляем категорию «нет сведений о признаке»
2. Добавляем значение динамики признака – получение тренда по историческим данным

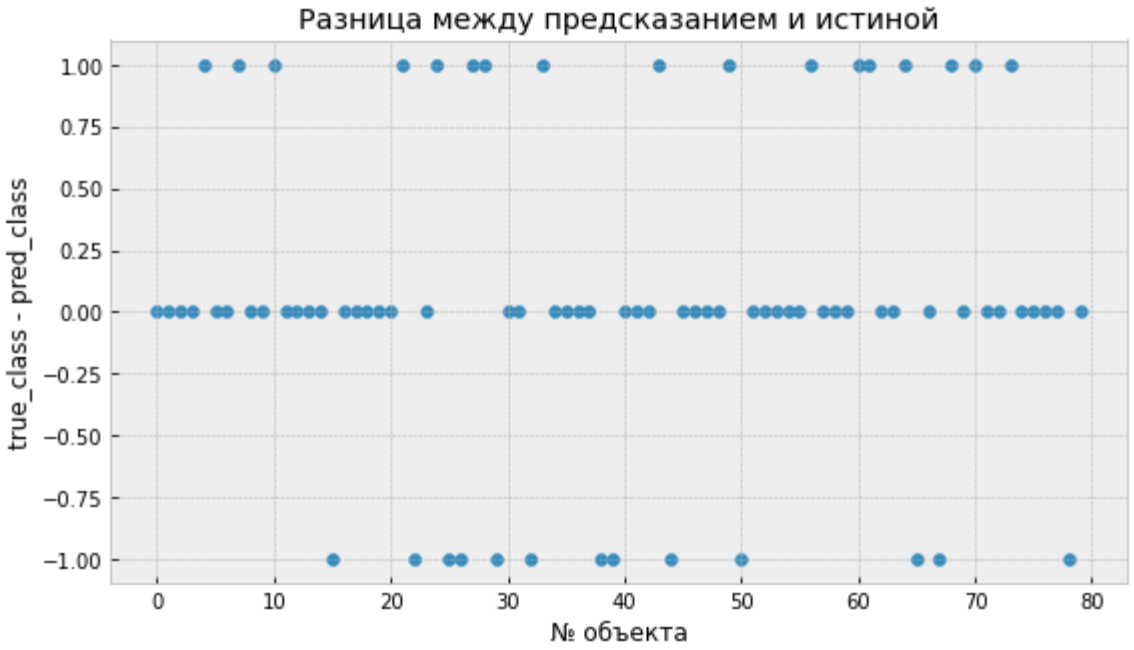
Признаки из внешних источников:

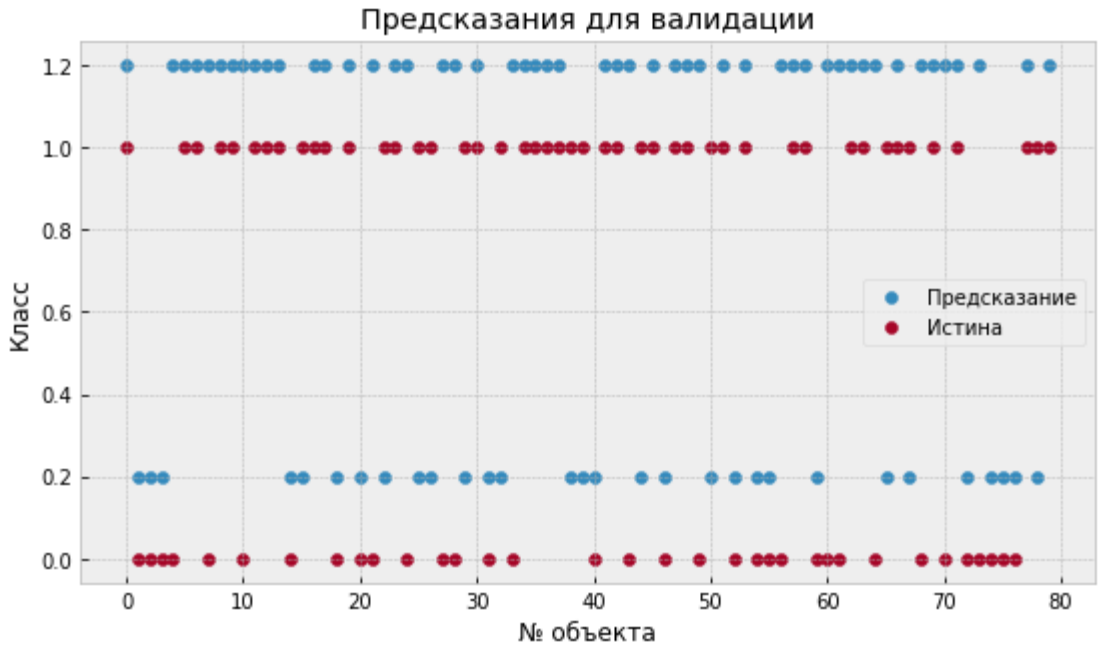
1. Индексы производства по отдельным видам экономической деятельности ОКВЭД2 (на 3 года назад: 2018, 2017, 2016)
    1. Производство металлургическое
    2. Производство стальных труб, полых профилей и фитингов
    3. Производство прочих стальных изделий первичной обработкой
    4. Производство готовых металлических изделий, кроме машин и оборудования
    5. Производство машин и оборудования, не включенных в другие группировки
2. Курс USD/RUB ЦБ РФ
    1. На 1 января каждого года
    2. Максимальный за год
    3. Минимальный за год

3. Обучение классификатора

- От задачи ранжирования переходим к задаче бинарной классификации - по признаку «Макс ПДЗ» устанавливаем порог, в решении были применены следующие пороги: 0, 5, 10.
- После этого был произведен поиск по сетке и построено 3 классификатора (CatBoost).
- Далее был проведен анализ значимости признаков для их дальнейшей интерпретации.

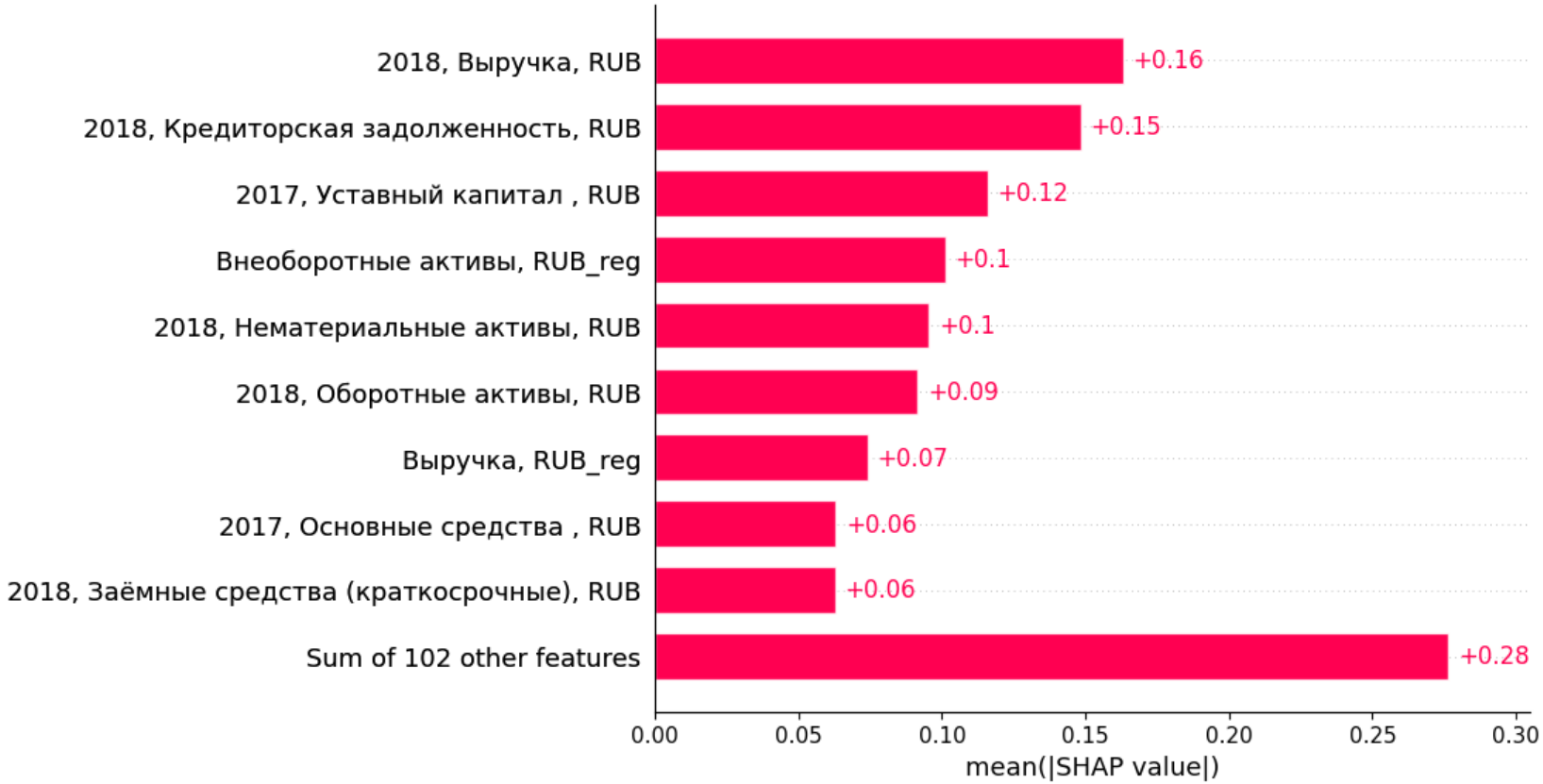
3. Обучение классификатора





На графиках приведены результаты работы классификатора на тестовой выборке (разбиение train/test 0,85).

3. Обучение классификатора



Были получены коэффициенты значимости Шапа, которые в последствии легли в основу системы выдачи информации по контрагенту. Коэффициенты считались на всех доступных данных (включая валидационные).

* Признаки _reg – тренд по историческим данным, _cat – наличие исторических данных

4. Рекомендательная система [основная идея]

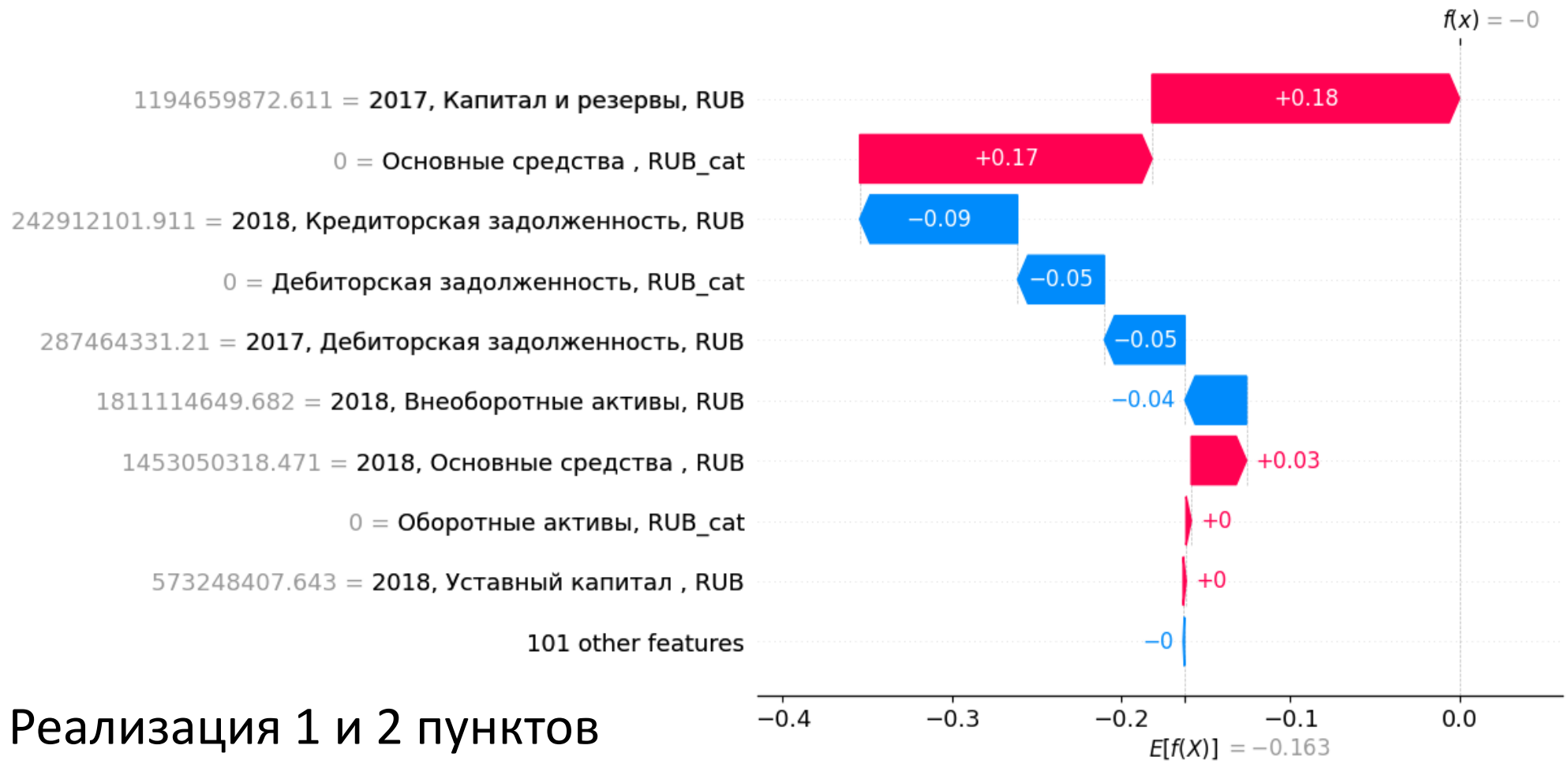
Любая аналитика необходима в конечном счете для лица, которое на ее основе принимает решение. Таким образом, необходимо сделать понятную неспециалисту систему, которая бы не была перегружена техническими подробностями, а давала понятную информацию.

В нашем случае она объединяет 3 сущности:

1. Предсказание классификаторов – дает в процентах оценку надежности контрагента.
2. Графическое представление наиболее значимых признаков, повлиявших на данную оценку.
3. Сводку по уже имеющимся данным, в разрезе признаков из п. 2.

4. Рекомендательная система

```
Введите индекс строки для построения отчета: 500
Контрагент 500 классифицирован как 'контрагент с высоким риском' [84.4%]
Наиболее значимые параметры контрагента, повлиявшие на решение:
```



Реализация 1 и 2 пунктов

4. Рекомендательная система

```
===== Сводка по наиболее значимым признакам=====
[предполагается наличие исторических данных]
['Аналогичные' контрагенты - процент, допустивших ПДЗ по данному критерию (выше/ниже) в зависимости от 'Вклада']
```

	Признак	Вклад	Аналогичные
0	2018, Кредиторская задолженность, RUB	-0.093322	21.8
1	Дебиторская задолженность, RUB_cat	-0.051280	1.1
2	2017, Дебиторская задолженность, RUB	-0.047656	16.2
3	2018, Внеоборотные активы, RUB	-0.036394	9.4
4	2018, Оборотные активы, RUB	-0.000654	19.4

Реализация 3 пункта

В приведенной тетрадке данная идея обернута в класс и можно интерактивно с ней повзаимодействовать.

Заключение

Не успели детально изучить данные по другим годам, однако в построенный пайплайн они довольно легко впишутся.

Значимые признаки не показались необычными – например, достаточно логично влияние кредиторской задолженности на вероятность ПДЗ.

При дальнейшей работе, скорее всего, основная идея не изменится, так как самое важное – это получить на выходе аналитику, понятную человеку, для которого она предназначена.