Анализ контрагентов.

Для решения этой задачи было предоставлено 3 набора данных. Данные agents2021.csv включают в себя данные с agents2019.csv и agents2020.csv. То для рассмотрения была взята только база по контрагентам agents2021.csv.

Объем объектов совсем небольшой – всего 325 контрагентов.

Признаки – данные финансовой отчетности за 2016-2020 года, факты под номерами, и данные о задолженности 2019, 2020, 2021 (если я правильно поняла последние колонки).

Требуется: построить модель ML, которая должна рассчитывать возможность просрочки оплаты контрагентом , основываясь на показатели прошлых периодов. И если просрочка возможна, то на какой период.

Признаки, которые были добавлены:

На основании "Методоглогических указаний по проведению анализа финансового состояния организаций" (приказ ФСФО от 23 января 2001г. № 16) в качестве признаков рассчитаны коэффициенты платежеспособности и финансовой устойчивости.

Платежеспособность характеризует возможность организации своевременно расплачиваться по своим обязательствам.

1. Общие показатели:
К_1 - среднемесячная выручка;

2. Расчет показателей платежеспособности и финансовой устойчивости:
К_4 - степень платежеспособности общая;
К_5 - коэффициент задолженности по кредитам банков и займам;
К_6 - коэффициент задолженности другим организациям;
К_9 - степень платежеспособности по текущим обязательствам;
К_10 - коэффициент покрытия текущих обязательств оборотными активами;
К_11 - собственный капитал в обороте;
К_12 - доля собственного капитала в оборотных средствах (коэффициент обеспеченности собственными средствами);
К_13 - коэффициент автономии (финансовой независимости).

3. Показатели эффективности использования капитала
К_14 - коэффициент обеспеченности оборотными средствами;
К_17 - рентабельность оборотного капитала;
К_18 - рентабельность продаж;
К_20 - эффективность внеоборотного капитала (фондоотдача);

По всем коэффициентам есть значения сильно отклоняющиеся от медианного, произошедшие в результате резких скачков по показателям участвующими в расчете или их нулевым значениям. Необходимо отобрать этих контрагентов и изучить их более

внимательно, чтоб откорректировать эти значение для машинного обучения. Т.к. в текущем варианте это выглядит как выбросы. Просто удалить нельзя – данных очень мало. На это не хватило времени.

Для отслеживания динамики изменения показателей финансовой отчетности, в качестве признаков были добавлены показатели:

- темп роста,
- темп прироста
- ускорение.

В ходе анализа не выявлено никаких явных признаков, которые бы способствовали более четкому разграничению - заплатит контрагент во время или нет.

Создание модели:

Для машинного обучение сформировано 2 варианта выборок по признакам.

Так как данных мало, то выборку смысла дробить нет.

Модель должна рассчитывать возможность просрочки основываясь на показателях прошлых периодов.

Для target_2019 возьмем года 2016-2018 - Проверять будем на target_2020 и годах 2017-2019

Для target_2020 возьмем года 2017-2019 - Проверять будем на target_2021 и годах 2018-2020

Т.е. у нас будет 3 обучающие выборки и 3 с верными значениями в нескольких вариациях

df_2019	Года для обучения 2016-2018	target_2019_01
df_2020	Года для обучения 2017-2019	target_2020_01
df_2021	Года для обучения 2018-2020	target_2021_01

В результате было испробовано много разных вариантов моделей машинного обучения и вариантов комбинации признаков.

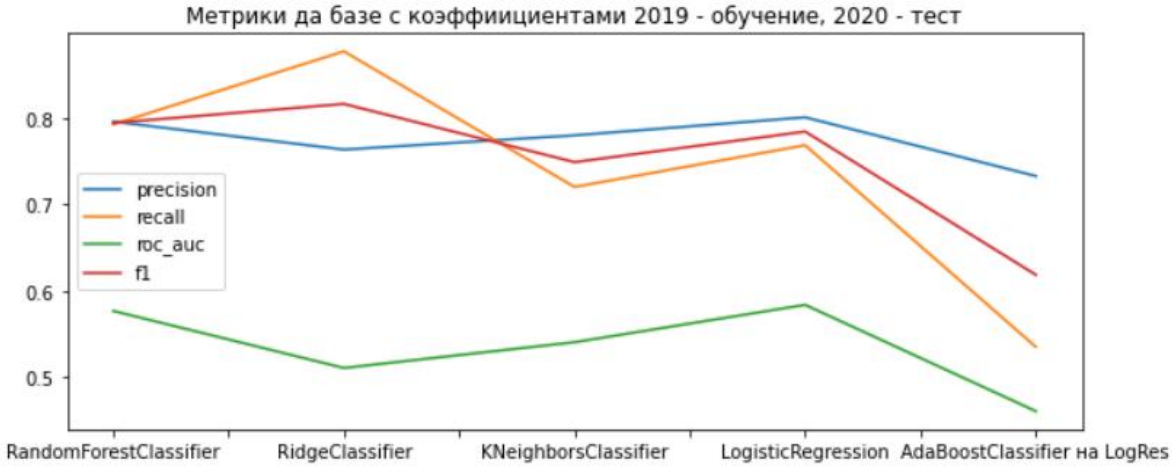
Модели, которые давали боле менее лучший результат:

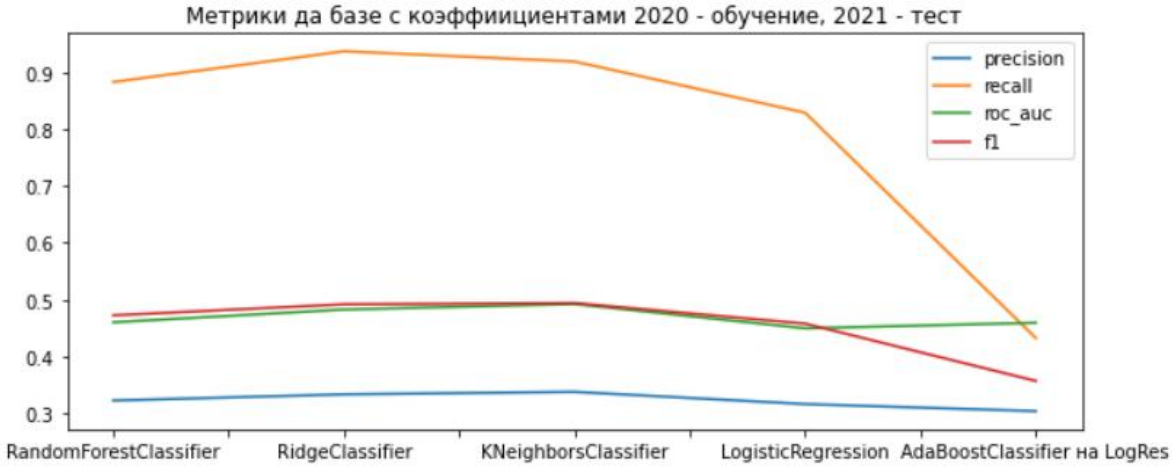
обучение на df_2019, прогноз на df_2020

	conf_matrix	precision	recall	roc_auc	f1
RandomForestClassifier	[[28, 50], [51, 196]]	0.80	0.79	0.58	0.80
RidgeClassifier	[[11, 67], [30, 217]]	0.76	0.88	0.51	0.82
KNeighborsClassifier	[[28, 50], [69, 178]]	0.78	0.72	0.54	0.75
LogisticRegression	[[31, 47], [57, 190]]	0.80	0.77	0.58	0.79
AdaBoostClassifier на LogRes	[[30, 48], [115, 132]]	0.73	0.53	0.46	0.62

обучение на df_2020, прогноз на df_2021

	conf_matrix	precision	recall	roc_auc	f1
RandomForestClassifier	[[8, 206], [13, 98]]	0.32	0.88	0.46	0.47
RidgeClassifier	[[6, 208], [7, 104]]	0.33	0.94	0.48	0.49
KNeighborsClassifier	[[14, 200], [9, 102]]	0.34	0.92	0.49	0.49
LogisticRegression	[[15, 199], [19, 92]]	0.32	0.83	0.45	0.46
AdaBoostClassifier на LogRes	[[104, 110], [63, 48]]	0.30	0.43	0.46	0.36





Метрики на базах обучении 2020 – тест 2021 гораздо хуже, чем на обучении 2019 – тест 2020, что очень хорошо отражено на графиках

Данные, которые использовались для предсказания df_2021 включает в себя года 2019 – 2020, которые пришлись на пандемию короновируса и многие организации столкнулись разными проблемами и ограничениями, что отразилось на финансово-хозяйственной деятельности и соответственно данных финансовой отчетности.

Так как в коэффициентах есть выбросы, а обработать их не было времени. Просто округлить, уменьшить арифметически – не вариант. Надо смотреть причину. И исходя из этого принимать решение.

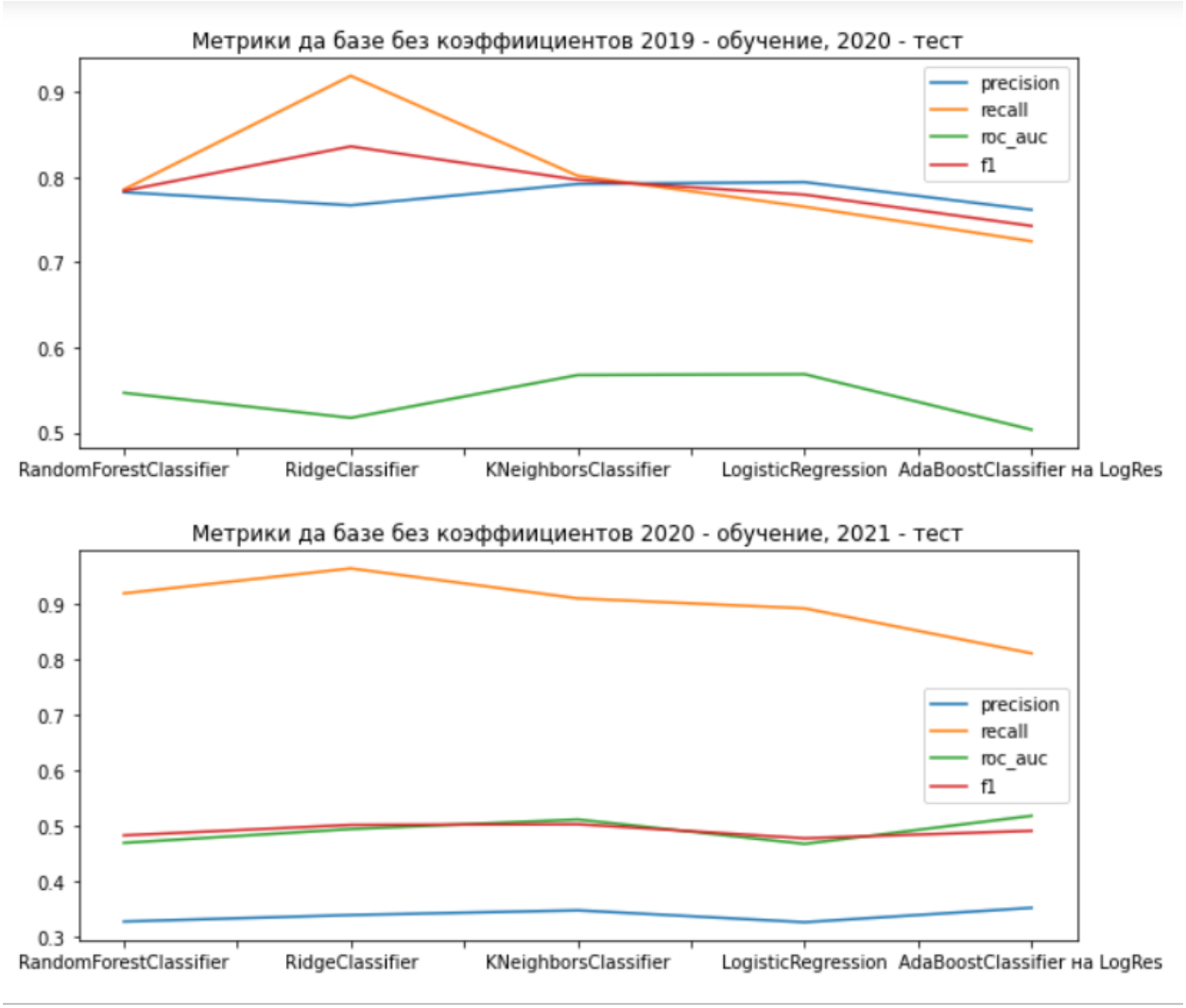
Сделала выборки без учета этих коэффициентов.

обучение на df_2019, прогноз на df_2020 без коэффициентов

	conf_matrix	precision	recall	roc_auc	f1
RandomForestClassifier	[[24, 54], [53, 194]]	0.78	0.79	0.55	0.78
RidgeClassifier	[[9, 69], [20, 227]]	0.77	0.92	0.52	0.84
KNeighborsClassifier	[[26, 52], [49, 198]]	0.79	0.80	0.57	0.80
LogisticRegression	[[29, 49], [58, 189]]	0.79	0.77	0.57	0.78
AdaBoostClassifier на LogRes	[[22, 56], [68, 179]]	0.76	0.72	0.50	0.74

обучение на df_2020, прогноз на df_2021 без коэффициентов

	conf_matrix	precision	recall	roc_auc	f1
RandomForestClassifier	[[4, 210], [9, 102]]	0.33	0.92	0.47	0.48
RidgeClassifier	[[5, 209], [4, 107]]	0.34	0.96	0.49	0.50
KNeighborsClassifier	[[24, 190], [10, 101]]	0.35	0.91	0.51	0.50
LogisticRegression	[[9, 205], [12, 99]]	0.33	0.89	0.47	0.48
AdaBoostClassifier на LogRes	[[48, 166], [21, 90]]	0.35	0.81	0.52	0.49



Картина та же.

Работа по данной задаче не завершена. Требуются дополнительные данные которые могут оказать существенное влияние на результат.

Подбор гиперпараметров для моделей не изменил ситуацию.

Скорее всего, необходимо:

- информация о наличие судебных процессов у этого контрагента по вопросам взыскания долгов.
- вид экономической деятельности
- является ли Северсталь основным поставщиком этого контрагента
- суммы контрактов, регулярность поставок
-возможно регион осуществления деятельности.