• Нелинейность: Многие экономические явления не подчиняются линейным закономерностям. Модели машинного обучения, такие как нейронные сети, способны улавливать нелинейные зависимости.

• Обработка больших данных: Алгоритмы машинного обучения могут обрабатывать и анализировать огромные объемы данных — больше, чем это возможно с помощью традиционных статистических методов.

• Высокая точность прогнозов: Исследования показывают, что модели машинного обучения часто превосходят традиционные модели в точности прогнозирования банкротства.

Например, Ф. Барбоза, Х. Кимура, Э. Альтман выявили, что модели машинного обучения, такие как метод опорных векторов, бэггинг, бустинг и случайный лес, в среднем демонстрировали примерно на 10% более высокую точность по сравнению с традиционными моделями, такими как дискриминантный анализ, логистическая регрессия.

Набор данных представляет из себя две выгрузки с платформы Spark: арбитражные дела и строительные компании малого и среднего размера (определяются по выручке за 2019 отчетный год от 120 млн до 2 млрд рублей), которые были объединены в один файл посредством мэтчинга и добавлением столбца “Дата иска”.

Выборка взята из Spark за период с 2008 года по 2022 размером (14 лет) в 9121 строку. На начальном этапе были выбраны следующие переменные

Сравниваем столбцы “Ответчик” (арбитражные дела) и “Наименование” (списки компаний), если значения совпадают, то берем в нашу выборку. Далее данные из трех файлов объединяются в один по столбцам, получается общий dataframe с данными по годам. После этого я выбираю все столбцы, которые не относятся к нашим показателям (не числа) и транспонирую их. Считаю количество пропущенных значений в столбцах и убираю те, где процент превышает 30, после этого избавляюсь от компаний с пропущенными значениями. Создаю категориальную переменную, основанную на дате иска. True - компания является банкротом, False - не является. В отфильтрованные по категориальной переменной данных вычисляем год банкротства (год иска): если до апреля - 2, если в апреле или позже, то -1 (март - месяц отчетности компаний). Оставляю записи, где “Год” совпадает с “Годом иска”. Как итог получаю выборку с обанкротившимися компаниями. Далее делю выборку на тестовую и тренировочную с соотношением тестовых и тренировочных 0.3 (70% тренировочных данных и 30% тестовых от общих 1260) и с соотношением True False 14% - 86% для обеих выборок соответственно.

Тк у нас в работе присутствуют модели, связанные с деревьями, два основным гиперпараметра это: глубина деревьев и количество, L1 L2 регуляризация (для предотвращения переобучения), а также несколько специфичных для каждой модели.

Наша целевая переменная является категориальной, где True – обанкротившаяся компания, False – здоровая компания.

Удаление выбросов с помощью Isolation Forest (удалили 3% от выборки)

В целом, на основе представленных данных можно заключить, что тюнинг гиперпараметров приносит ожидаемое улучшение в производительности моделей для обоих алгоритмов. Также заметно, что LightGBM слегка превосходит XGBoost под каждым аспектом при отсутствии тюнинга гиперпараметров. При настройке гиперпараметров оба алгоритма показывают схожие результаты, но всё же LGBM имеет небольшое преимущество в R², это означает лучшую объяснительную способность данной модели. XGoost проще запустить в работу на видеокарте.