В данной работе стояла задача создать модель предсказания выхода клиента банка в дефолт по кредиту. При оценке качества модели использовалась метрика ROC-AUC с порогом не ниже 75%.

С порядком выполнения работы можно ознакомиться в файле ReadMe.

Изначально было получено 12 файлов формата parquet и файл target\_train.csv с данными по дефолту для каждого клиента. Стоит отметить, что данные были обобщенные и не имели никаких отметок о том, по какому кредиту клиент совершил дефолт. В связи с этим можно по разному подойти к объединению датафреймов. Первый вариант размножить целевую переменную под все имеющиеся кредиты у клиента, второй вариант объединить каким-либо образом имеющиеся кредиты клиента. Подсказка от разработчиков говорит о том, что имея общую размерность всех файлов в 26млн строк, необходимо получить итоговый датафрейм в размере 3млн строк, поэтому был выбран второй вариант - группировка. Сами данные не содержали пропусков, неправильных значений, все были числового типа. Из за малого количества информации нельзя сказать о фальсификации данных, о их природе и качестве.

Недостаточно подробно были описаны данные, только название признаков, поэтому сложно изначально было оценивать их содержание. Можно ли придавать какой-то вес значениям или нет. С одной стороны, если люди, которые изначально кодировали эти значения, имеют возможность интерпретировать их обратно, значит связь между числами может существовать. В связи с этим был выбран вариант использования агрегации по среднему, так как мода при наличии двух кредитов у клиента ломала бы логику, медиана в некоторых значениях тоже, поэтому среднее значение, как статистический способ агрегации данных подходил наилучшим образом.

В дальнейшем были опробованы разные подходы к обработке данных, в том числе использование OneHotEncoding на явно закодированных данных, и затем кодирование всех данных таким образом кроме бинарных и агрегация по сумме всех данных, что является редукцией размерности, получая на выходе вместо множества бинарных столбцов, всего один с данными о активности клиента. Также были созданы 2 признака «Есть ли просрочка Да/Нет», «Общее количество просрочек».

Разные подходы к обработке данных были вызваны тем, что основываясь на экспериментах модели плохо прогрессируют и отзываются на них. Так между первым и вторым подходами к обработке разница составила менее 1%, 72.11 и 73 соответственно. При использовании моделей внутри на разных объемах данных, с использованием различных методов балансировки классов эксперименты не приводили к успеху, однако, стоит отметить, что при использовании RandomUnderSampling и тестировании на полноценных данных без балансировки, модели Бустинга справлялись лучше остальных и показывали самые высокие результаты. Можно дополнительно упомянуть метод опорных векторов, который был использован при тестировании и тоже показал близкие к максимальным результаты, но время его построения в десятки раз больше, поэтому в финальных экспериментах он участие не принимал.

Третий подход показал, что чем больше признаков, которые и появляются благодаря горячему кодированию, тем лучше модель способна предсказывать целевую метку.

По итогу получается на кросс-валидации модель CatboostClassifier с метрикой 76.17, при формировании пайплайна и выдаче предсказания на тестовых данных метрика 75.68.