Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

1.1 Постановка задачи

Сформулированная организаторами соревнования Kaggle задача является классической задачей классификации: нулевой класс соответствует беспроблемным кредитам, первый класс — кредитам, по которым заявитель допускал просрочки более некоторого количества дней.

В рамках этой работы поставлена задача:

- провести разведочный анализ данных
- при необходимости провести предобработку данных
- добавить в основной датасет агрегированные показатели, используя данные из бюро кредитных историй, данные кредитных карт и предыдущих кредитных заявок
- провести отбор наиболее информативных переменных-предикторов
- применить методы классического машинного обучения к задаче классификации, сравнить их эффективность
- разработать нейронную сеть для решения задачи классификации
- при построении моделей провести поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10.
- создать публичный репозиторий в GitHub и разместить там код и документацию исследования

Предоставленный датасет состоит из 7 основных файлов в табличном виде. Рассмотрим их по-порядку:

1) application_{train|test} – это основная таблица, разбитая на 2 файла — для обучения и тестирования, содержит статические данные для всех кредитных заявок. Одна строка представляет один кредит в выборке данных. В обучающей выборке содержится 307 тысяч заявок, в тестовой

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

— 48 тысяч заявок. Количество признаков (не считая целевой переменной и технического идентификатора заявки SK_ID_CURR) — 120.

2) bureau - все предыдущие кредиты клиента, предоставленные другими финансовыми учреждениями, о которых было сообщено в кредитное бюро (для клиентов, имеющих кредит в нашей выборке). Для каждого кредита в нашей выборке имеется столько строк, сколько кредитов было у клиента в кредитном бюро до даты подачи заявки. Содержит 1.7 млн записей и 15 признков (не считая ключей SK_ID_CURR и SK_BUREAU_ID, которые используются для джойна между таблицами).
3) bureau_balance - ежемесячные остатки по предыдущим кредитам в кредитном бюро. В этой таблице есть одна строка для каждого месяца истории каждого предыдущего кредита, сообщенного в кредитное бюро, т.е. в таблице есть строки (количество кредитов в выборке * количество относительных предыдущих кредитов * количество месяцев, в которых у нас есть некоторая история, наблюдаемая для предыдущих кредитов). Таблица содержит 27.3 млн записей и 2 признака.
4) POS_CASH_balance - ежемесячные справки о балансе предыдущих кредитов наличными, которые клиент имел в банке Home Credit. По одной записи для каждого кредита, каждый месяц. Таблица содержит 10.0 млн записей и 6 признаков.
5) credit_card_balance - ежемесячные балансы кредитных карт, которые заявитель имел в Home Credit. Таблица содержит 3.8 млн записей и 21 признак.
6) previous_application - все предыдущие заявки на кредиты Home Credit клиентов, имеющих кредиты в нашей выборке. Таблица содержит 1.6 млн записей и 35 признаков.
7) installments_payments - история погашения ранее выданных кредитов в Home Credit, относящихся к кредитам в нашей выборке. Имеется a) одна строка для каждого произведенного платежа плюс b) одна строка для

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

каждого пропущенного платежа. Таблица содержит 13.6 млн записей и 6 признаков.