**Прогнозирование кредитного дефолта**

**Исходная задача и её актуальность**

Будь то в ресторане или покупая билеты на концерт, современная жизнь рассчитывает на удобство кредитной карты для совершения ежедневных покупок. Это избавляет нас от необходимости носить с собой большие суммы наличных, а также может ускорить полную покупку, которая может быть оплачена со временем. Как эмитенты карт узнают, что мы вернем то, что мы взимаем? Это сложная проблема со многими существующими решениями и еще большим количеством потенциальных улучшений.

Прогнозирование дефолта по кредиту занимает центральное место в управлении рисками в бизнесе потребительского кредитования. Кредиторы оптимизируют кредитные решения, что приводит к улучшению качества обслуживания клиентов и надежной экономики бизнеса.

Необходимо спрогнозировать дефолт по кредиту. В данной задаче используется набор данных промышленного масштаба для построения модели машинного обучения.

Главными задачами являются предсказание вероятности того, что клиент не вернет остаток на своей кредитной карте в будущем, на основе его ежемесячного профиля клиента, улучшение качества обслуживания клиентов для владельцев карт, упрощение процедуры получения одобрения кредитной карты.

**Используемые метрики**

Показатель оценки, *M* для этого соревнования является средним значением двух показателей рангового порядка: нормализованный коэффициент Джини, G, и показатель по умолчанию зафиксирован на уровне 4%, *D*.

*M=0.5(G+D)*

Показатель по умолчанию, зафиксированный на уровне 4%, представляет собой процент положительных меток (по умолчанию), зафиксированных в 4% прогнозов с наивысшим рейтингом, и представляет статистику чувствительности / отзыва.

Максимальное значение этого показателя равно 1.0.

**Анализ данных**

**Описание исходных данных**

Целевая двоичная переменная рассчитывается путем наблюдения за 18-месячным периодом выполнения после последней выписки по кредитной карте, и если клиент не выплачивает причитающуюся сумму в течение 120 дней после даты последней выписки, это считается событием по умолчанию.

Набор данных содержит агрегированные характеристики профиля для каждого клиента на каждую дату выписки. Функции анонимизированы и нормализованы и подразделяются на следующие общие категории:

* D\_\* = переменные просрочки
* S\_\* = Переменные расходов
* P\_\* = Переменные платежа
* B\_\* = переменные баланса
* R\_\* = переменные риска

при этом следующие функции являются категоричными:

['B\_30', 'B\_38', 'D\_114', 'D\_116', 'D\_117', 'D\_120', 'D\_126', 'D\_63', 'D\_64', 'D\_66', 'D\_68']

Задача - предсказать для каждого customer\_ID из них вероятность дефолта по платежу в будущем (target = 1).

**Первоначальный анализ данных**

Наборы данных для обучения, проверки и тестирования включают поведенческие данные временных рядов и анонимизированную информацию о профиле клиента.

Начнем с того, что на сайте Kaggle показана краткая характеристика для каждого признака (всего их 190). Первоначально отбросим те признаки, у которых пропусков больше 30 процентов, а это чуть больше 30 признаков из разных категорий. Затем отбросим признаки с малым отклонением это немного меньше 30. Первоначально остаётся примерно 130 признаков с которыми будем работать.

Следующим шагом отбросим данные с большим разбросом. Из тех данных, которые сильно похожи друг на друга выберем тот признак, который характеризует всю группу. Остаётся порядка 30 признаков с которыми мы в дальнейшем и будем работать. (нужно от чего – то отталкиваться, если результат будет плохим, необходимо будет вернуться к этому этапу)

В большинстве своём выбранные признаки имеют нормальное распределение, адекватные средние и отклонение, и малое количество пропусков.

**Анализ выбранных признаков**

Тренировочный файл состоит примерно из 5,5 миллионов строк и 190 столбцов (признаков), его вес составляет чуть больше 16 гигабайт с форматом csv.

Проблема заключается в том, что данный файл невозможно открыть с помощью Excel или различных программ для открытия csv файлов, также не годиться открывать их с помощью библиотеки pandas, поскольку время работы будет очень долгим (я работаю на слабеньком ноутбуке).

Первоначально я разделил файл на множество файлов равных по 15000 строк. Главная проблема, что такие файлы являются нерепрезентативными. И анализ данных файлов будет неправилен.

Решением данной проблемы является открыть этот файл как базу данных. Я пользовался программой DB2. Делаем select запрос для выбранных признаков, связываем её с файлом с тестовыми ответами по customer ID и импортируем их в csv файл для дальнейшего анализа в pandas.

1. Получим сводную статистику, а также для просмотра количества ненулевых данных
2. Из категориальных данных создаем фиктивные столбцы. И удаляем из них ненужные
3. Для определения группы с выбросами и пропусками используем кластеризацию и заменяем их средним значением
4. Проверяем данные на корреляцию
5. Строим нормальное и т – распределение
6. Выбираем 7 лучших признаков без учёта ID, дат и ответов

Делим данные на обучаемые и тестовые.

1. Проводим нормализацию данных
2. Выбираем модель для обучения
3. Проверяем её на недообучение и переобучение, график зависимости обучения
4. Метрики

**Тестовые данные**

Создаем тестовый файл с избранными признаками, обучаем модель и заливаем на сайт Kaggle. Подводим итог. Если оценка небольшая проводим анализ заново, выбираем другую модель и подводим результат.