Контекст

Кредитные скоринговые карты являются распространенным методом контроля рисков в финансовой индустрии. Он использует личную информацию и данные, представленные заявителями на кредитные карты, для прогнозирования вероятности будущих дефолтов и заимствований по кредитным картам. Банк имеет право решать, выдавать ли кредитную карту заявителю. Кредитные баллы могут объективно количественно оценить величину риска.  
   
Вообще говоря, кредитные карты основаны на исторических данных. Однажды столкнувшись с большими экономическими колебаниями. Прошлые модели могут потерять свою первоначальную прогностическую силу. Логистическая модель является распространенным методом кредитного скоринга. Потому что Logistic подходит для задач бинарной классификации и умеет вычислять коэффициенты каждого признака. Чтобы облегчить понимание и работу, оценочная карта будет умножать коэффициент логистической регрессии на определенное значение (например, 100) и округлять его.  
   
В настоящее время с развитием алгоритмов машинного обучения. В скоринг кредитных карт были введены более прогностические методы, такие как бустинг, случайный лес и машины опорных векторов. Однако эти методы часто не обладают хорошей прозрачностью. Может быть сложно предоставить клиентам и регулирующим органам причину отклонения или принятия.

Задача

Создайте модель машинного обучения, чтобы предсказать, является ли кандидат «хорошим» или «плохим» клиентом, в отличие от других задач, определение «хорошего» или «плохого» не дается. Вы должны использовать некоторые методы, такие как [винтажный анализ](https://www.kaggle.com/rikdifos/eda-vintage-analysis) , чтобы создать свою этикетку. Кроме того, проблема дисбаланса данных является большой проблемой в этой задаче.

Содержание и объяснение

Две таблицы могут быть объединены ID:

| application\_record.csv |  |  |
| --- | --- | --- |
| Название функции | Объяснение | Примечания |
| ID | Номер клиента |  |
| CODE\_GENDER | Пол |  |
| FLAG\_OWN\_CAR | есть машина |  |
| FLAG\_OWN\_REALTY | Есть ли недвижимость |  |
| CNT\_CHILDREN | Число детей |  |
| AMT\_INCOME\_TOTAL | Годовой доход |  |
| NAME\_INCOME\_TYPE | Категория дохода |  |
| NAME\_EDUCATION\_TYPE | Уровень образования |  |
| NAME\_FAMILY\_STATUS | Семейное положение |  |
| NAME\_HOUSING\_TYPE | Стиль жизни |  |
| DAYS\_BIRTH | День рождения | Считать в обратном порядке от текущего дня  (0), -1 означает вчерашний день |
| DAYS\_EMPLOYED | Дата начала трудоустройства | Считать в обратном порядке от текущего дня (0).  Если положительный, это означает, что  человек в настоящее время безработный. |
| FLAG\_MOBIL | Есть ли мобильный телефон |  |
| FLAG\_WORK\_PHONE | Есть ли рабочий телефон |  |
| FLAG\_PHONE | Есть ли телефон? |  |
| FLAG\_EMAIL | Есть электронная почта |  |
| OCCUPATION\_TYPE | Занятие |  |
| CNT\_FAM\_MEMBERS | Семейный размер |  |

| кредит\_record.csv |  |  |
| --- | --- | --- |
| Название функции | Объяснение | Примечания |
| ID | Номер клиента |  |
| MONTHS\_BALANCE | Рекордный месяц | Месяц извлеченных данных является отправной  точкой, наоборот, 0 — текущий месяц,  -1 — предыдущий месяц и т. д. |
| STATUS | Статус | 0: Просрочка 1-29 дней 1: Просрочка 30-59 дней  2: Просрочка 60-89 дней  3: Просрочка 90-119 дней  4: Просрочка 120-149 дней  5: Просроченная или безнадежная задолженность,  списание более 150 дней C: погашен в этом месяце  X: нет кредита в  течение месяца |