

Модель кредитного риск-менеджмента

Автор решения: Павел Грешников

Описание задачи

Требуется обучить модель кредитного риск-менеджмента для банка.

Данные содержат информацию о различных атрибутах заёмщиков и кредитных продуктов: о клиентах, которые уже имеют кредиты, их кредитной истории и финансовых показателях. Каждая запись в датасете представляет один конкретный кредитный продукт, выданный конкретному заёмщику.

Еще немного о данных

Область значений поля разбивается на N непересекающихся промежутков. Каждому промежутку случайным образом назначается уникальный номер от 0 до $N-1$, а значение поля заменяется номером промежутка, которому оно принадлежит.

Feature Engineering

- Удалены фичи с высокой корреляцией
- Добавлены новые фичи, такие как:
 - общее число просрочек по платежам
 - среднее количество дней с открытия
 - общее количество периодов без просрочек
 - соотношение невыплаченной суммы и лимита по кредиту
 - и некоторые другие.

Что дальше делать с данными?

Данные были предоставлены в виде нескольких строк для каждого клиента. Был сделан OneHotEncoding всех фич, так как номера интервалам, которым принадлежат значения фичей не несут в себе информации об отношении порядка этих номеров интервалов. Далее были сгруппированы эти номера для всех записей в рамках одного клиента.

Модели, использованные в решении:

	Logistic Regression	XGBoost
roc_auc на тестовой выборке	0.741	0.754

Что дальше?

Далее, был построен пайплайн, состоящий из предобработки данных и модели XGBoost. Пайплайн был обучен на всех имеющихся данных и сохранен в бинарном виде для дальнейшего использования.

Вопросы?