ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

Кредитная индукция

В этой задаче тебе нужно проанализировать кредитный датасет и найти в нем закономерность, которая позволяет предсказывать надежность заемщика. Оцениваться будет не только качество найденной закономерности, но и описание процесса исследования.

Описание задачи

Методы искусственного интеллекта (ИИ) часто используются, чтобы найти закономерности в данных. Например, банки используют ИИ, чтобы понять, как разные признаки человека (карьерный путь, образование, ...) связаны с вероятностью возврата кредита. В этой задаче мы предлагаем тебе попробовать себя в роли исследователя и найти закономерности в данных об образовательных кредитах.

Данные В этой задаче ты будешь работать с датасетом CREDIT, в котором собрана статистика по *образовательным кредитам*. Данные сгенерированы специально для этой задачи и не соответствуют реальности. Любые совпадения случайны.

Всего в файле 7 976 строк и 21 столбец. Каждая строка соответствует заявке на кредит, а каждый столбец содержит характеристику заявки. Названия столбцов соответствуют реальным названиям в базе данных крупного банка.

Столбец	Описание
`name`	ФИО заемщика
`create_dttm`	дата создания заявки
`pid`	id заявки
`gender_cd`	пол заемщика
`age`	возраст заемщика
`short_nm`	университет
`education_level_code`	ступень образования
`specialty`	специальность
`semestr_cnt`	число семестров до конца обучения
`subside_rate`	ставка субсидирования
`semester_cost_amt`	стоимость обучения за семестр
`initial_approved_amt`	начальная сумма одобренного кредита
`initial_term`	срок кредитования (в годах)
`utm_source`	рекламный источник
`marketing_flag`	сталкивался ли заемщик с рекламой университета
`approve_dttm`	дата согласования заявки
`approve_flg`	флаг согласования заявки
`reject_reason`	причина отказа
`util_dttm`	дата перевода денежных средств
`util_flg`	флаг перевода денежных средств
`score`	оценка плетёжеспособности: число на отрезке [0; 1]

Важный нюанс

К сожалению, имена столбцов в датасете были утеряны. Тебе предстоит самому догадаться, какое имя какому столбцу соответствует.

Подсказки:

- флаг — это бинарный признак (0 или 1), который отмечает, состоялось ли какое-то событие;
- в учебном году обычно 2 семестра.

Как работать с данными Ты можешь использовать любой удобный инструмент, в том числе:

- `Python` и библиотеку `pandas` (рекомендовано)
- Excel
- Google Sheets

Задача

В датасете есть столбец `score`.

Твоя цель — придумать формулу, по которой значение `score` можно *оценить* из значений других столбцов.

Мы сгенерировали данные так, чтобы существовала формула, которая для каждого заемщика точно вычисляет значение `score` из значений других столбцов. Найти точную формулу может быть непросто, поэтому тебе нужно подобрать формулу, которая оценивает `score` максимально точно.

Твоя формула может использовать значения любых из 20 столбцов, описывающих заемщика: от `name` до `util_flg`. На выходе формула должна возвращать одно число для каждого заемщика: оценку `score`. Пример формулы в `pandas`:

```
result = (df['education_level_code'].map({'SPECIALTY': 0.2, 'BACHELOR': 0.3, 'MASTER': 0.5})
          + 1
          - df['approve_flg']) / 2
```

Здесь уровень обучения кодируется вещественным числом, затем к нему прибавляется единица и вычитается значение флага согласования. Итоговый результат делится на два.

Что можно и нельзя использовать В своем решении ты можешь использовать:

- любые элементарные функции и функции округления
- операции сложения, умножения, вычитания и деления, поэлементно примененные к столбцам
- подстановки: замены символьных выражений числами, например «MASTER» $\to$ 0.3
- функции подсчета свойств: число символов в строке, день недели для даты, ...
- если решаешь на `Python`, любые функции `pandas` и `numpy`

При этом тебе нужно соблюдать некоторые ограничения:

- ты можешь использовать алгоритмы машинного обучения для исследования датасета, но твоя итоговая формула должна зависеть только от значений исходных признаков и использовать только описанные выше функции
- твоя формула должна работать только со значениями двадцати столбцов, описывающих заемщика: нельзя ссылаться на номер строки, других заемщиков, ...

Важно В формуле все столбцы должны называться настоящими именами (`name`, `create_dttm` и т.д.).

Оценка решения Твое решение будет оцениваться по двум критериям:

- *Точность формулы.* Мы будем использовать метрику MAE (Mean Absolute Error, см. например `https://wiki.loginom.ru/articles/mae.html`). Все решения будут упорядочены по возрастанию ошибки — чем выше место участника в рейтинге (то есть, чем меньше ошибка), тем больше баллов он(а) получит.
- *Качество отчета.* Мы оценим полноту и ясность отчёта, а также оригинальность применённых методов. Кроме того, оценивается компактность и простота формулы — старайся найти простое и красивое решение.

Оформление решения

Для отправки решения тебе нужно сделать три шага.

Шаг 1. Подсчитай ошибку своей формулы и округли ее по правилам математики до третьего знака после запятой. Полученное значение укажи в первой строке текстового поля формы решения.

Шаг 2. Запиши свою итоговую формулу в txt файл и приложи его к решению. Если ты решаешь задачу в `pandas`, скопируй итоговую формулу и вставить ее в файл. Если ты решаешь задачу в Excel или Google Sheets, опиши формулу псевдокодом.

Шаг 3. Оформи отчет с точным описанием итоговой формулы и решением: логикой сопоставления имен столбцов, этапами поиска формулы, перечнем рассмотренных вариантов решений. В отчете можно приводить фрагменты кода. Если ты использовал(а) какие-то источники информации, перечисли их. Отчет должен быть оформлен в один файл формата pdf без ссылок на дополнительные части. Объем файла: не более 5 страниц A4.

Важные замечания

- Мы вручную проверим точность твоей формулы. Если значение, которое мы получим, будет расходиться со значением, которые ты укажешь на шаге 1, ты получишь штраф.
- Мы проверим точность твоей формулы не только на данных, с которыми ты работаешь, но и на закрытых тестовых данных. Тестовые данные сгенерированы по тому же принципу, что и основной датасет, поэтому твоя формула должна показывать похожий результат. Если точность твоей формулы резко ухудшится, ты получишь штраф.
- Мы будем применять твою формулу к датасету, в котором все столбцы названы правильно. Если ты неправильно определишь названия столбцов, твоя формула получит низкий балл.

Загрузи свое решение в личный кабинет до **5 ноября 11:59 по московскому времени** по ссылке:

`https://edu.tinkoff.ru/all-activities/courses/ced0e9d5-fbb4-451a-b423-9da0a4ffc3d6`

По любым вопросам пиши нам в Телеграм: @cu_reshis.

Удачи!