Направление Risk Data Scientist, ЮниКредит Банк

Аннотация

Добро пожаловать на виртуальную стажировку Shift + Enter от ЮниКредит Банка! Ты сможешь пройти все ее этапы, начиная с отбора и заканчивая итоговым проектом. Готов стать Юникумом и помочь департаменту стратегических, кредитных и интегрированных рисков?

Стажировка от ЮниКредит Банка поможет тебе освоить такие навыки, как:

- Написание скрипта для ранжирования переменных при составлении кредитного скоринга.
- Реализация собственного алгоритма «жадного отбора».

Развиваемые компетенции

По результатам выполнения заданий ты сможешь:

- Узнать больше о том, чем занимаются Risk DS в банковской сфере.
- Изучить профильные статистические методы и получить опыт их применения.
- Прокачать навыки программирования на Python.

Описание подзадач

Подразделение Data Science департамента приглашает тебя присоединиться к созданию пайплайна[1] по разработке моделей оценки кредитного риска на основе алгоритмов машинного обучения.

Выполнение всего блока заданий (без финального проекта) займет у тебя не более 45-60 минут. В конце стажировки тебя ждет настоящий челлендж в виде написания сложного алгоритма «жадного отбора», который ты сможешь сравнить с вариантом от экспертов ЮниКредит Банка.

Рекомендуемый тайминг

1. 15-20 минут на первое задание.
2. 30-40 минут на второе задание.

Работа над финальным проектом может потребовать несколько часов, но оно того стоит! Ведь **ты сможешь использовать это в своих дальнейших проектах**, например при участии в хакатонах или соревнованиях на Kaggle. :) По итогам выполнения задания эксперты ЮниКредит Банка поделятся с тобой своим вариантом реализации этого метода.

Информация о загрузке решения

Стажировка содержит несколько подзадач. Можно загрузить файл, содержащий решение части заданий, но по возможности постарайся сделать их все. Желаем удачи!

[1] Пайплайн (от английского pipeline — «трубопровод») устанавливает порядок действий, обозначает инструменты и сроки для каждой задачи. В итоге каждый знает, когда и как именно он взаимодействует с другими участниками процесса.

Этап 1. Отбор на стажировку

Две недели назад ты подал заявку на виртуальную стажировку от ЮниКредит Банка. Открыв сегодня электронную почту, ты с радостью обнаружил, что тебя пригласили решить отборочное задание.

Привет!

В качестве тестового задания мы предлагаем тебе проверить знания в области теории вероятностей и статистики.

Задача 1. **Какое из константных предсказаний в задаче регрессии с метрикой MSE[2] является наилучшим?**

Задача 2. **Допустим, у нас обучен некоторый бинарный классификатор. Сначала мы считаем метрику Джини[3] (GINI) для пары «целевые значения – предсказания» (target vs prediction), затем считаем для target vs prediction[3] (то есть для куба предсказания). Как изменится GINI?**
Пояснение: GINI = 2 × ROC AUC – 1

Пожалуйста, пришли свое решение в течение ближайшего получаса, и мы обсудим его на видеоинтервью.

Спасибо!

Полезные материалы

Статья про метрику Джини (GINI).

Формат конечного результата

Файл в файле формата .docx.

Форма загрузки результата

Пожалуйста, загрузи свой вариант ответа в формате zip-архива, используя инструмент «Загрузить решение». Необходимо сформировать единый zip-архив, содержащий решение одного или всех заданий по выбранной специальности.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания от эксперта после отправки собственной версии.

[2] Mean Squared Error (MSE) измеряет среднюю сумму квадратной разности между фактическим значением и прогнозируемым значением для всех точек данных.

[3] Коэффициент или метрика Джини (GINI) – интегральный показатель качества ранжирующей способности модели, используемый в задачах кредитного скоринга.

Этап 2. Добро пожаловать!

Поздравляем, ты прошел отбор и готов приступить к стажировке. Твое первое задание связано с анализом данных и подготовкой переменных для моделирования. Позже ты увидел письмо от Дмитрия[4], твоего руководителя, с пояснениями к задаче.

Привет!

Я хочу, чтобы ты подключился к работе над проектом, связанным со сравнительным анализом моделей бинарной классификации для кредитного скоринга.

Проблема кредитного скоринга является важнейшей составляющей процесса кредитования в банковской сфере. Кредитный скоринг — это методология оценки потенциальных и действующих клиентов, в основе которой лежит анализ статистических данных. Мы, как специалисты в области моделирования кредитных рисков, отвечаем за то, чтобы заранее предсказать, сможет ли тот или иной наш клиент выплатить кредит.

Процесс разработки модели для целей кредитного скоринга состоит из нескольких этапов:

1. **Подготовка, очистка и предобработка данных, их первичный анализ (Exploratory Data Analysis).**
2. **Инженерия переменных, в том числе их биннинг при необходимости. Присвоение WoE-значений (Weight of Evidence) построенным бинам.**
3. **Выбор алгоритма машинного обучения, отбор переменных и тюнинг параметров.**
4. **Построение итоговой скоринговой модели.**

Сейчас мы работаем над автоматизацией первого и второго пунктов, поэтому просим тебя написать скрипт, который поможет нам в этом. Ты должен:

- **Проанализировать данные. Прежде чем работать с реальными данными наших пользователей, предлагаем потренироваться на данных из открытых источников (https://www.kaggle.com/competitions/tabular-playground-series-may-2022).**
- **Сделать WoE-биннинг и вывести список лучших WoE-переменных с индивидуальным коэффициентом GINI (метрика GINI для target vs feature).**

Hints. Для простоты закодированную текстовую категориальную переменную можно проигнорировать.

Прежде чем приступать к задаче, советую изучить полезные материалы.

Пришли свое решение в течение дня, и мы сможем обсудить его завтра на встрече с коллегами. Спасибо за помощь!

Полезные материалы

- Еще одна ссылка на данные, которые нужно использовать при работе над скриптом.
- Статья о WoE-преобразовании.
- Документация python-библиотеки для построения WoE-биннинга.

Формат конечного результата

Файл Jupyter Notebook (.ipynb) с пояснениями.

[4] Все имена и названия вымышленные, любые совпадения случайны. Данные заданий могут быть изменены в целях конфиденциальности.

Форма загрузки результата

Пожалуйста, загрузи свой вариант ответа в формате zip-архива, используя инструмент «Загрузить решение». Необходимо сформировать единый zip-архив, содержащий решение одного или всех заданий по выбранной специальности.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания от эксперта после отправки собственной версии.

Этап 3. Итоговый проект

Кредитный скоринг — это важная часть большого кросс-функционального проекта ЮниКредит Банка. Прелесть таких проектов в том, что в них вовлечены специалисты из разных подразделений. Это дает возможность узнать больше о других ролях в банке и понять, в каких направлениях тебе будет интересно развиваться. Ну и конечно, ты можешь общаться с драйвовыми коллегами из других команд - с ними не бывает скучно! :)

После выполнения различных заданий ты отлично прокачал свои навыки. Готов перейти на новый уровень мастерства? Твое финальное задание стажировки посвящено работе над алгоритмом отбора переменных.

Твой старший коллега Иван обещал прислать тебе больше информации о том, как работают жадные алгоритмы. Открыв почту, ты внимательно изучаешь детали проекта.

```
Привет!

Молодец, что обратился ко мне, всегда рад помочь нашим стажерам. Два года назад я
сам начинал со стажировки, и тогда мне тоже помогали более опытные коллеги.

Уверен, что через пару лет ты тоже сможешь быть наставником для новых стажеров банка
;)

Смотри, после подготовки данных тебе нужно написать другой скрипт, реализующий
алгоритм отбора переменных. Предполагается, что он войдет в пайплайн моделирования
нашего подразделения, что позволит быстрее разрабатывать модели и автоматически
обновлять их. Так ты облегчишь жизнь коллегам и повысишь продуктивность всего
департамента. Звучит круто, правда? :)

Твоя задача – написать скрипт для автоматического «жадного отбора» признаков в
рамках задачи бинарной классификации.

На первом шаге берется переменная с наибольшим индивидуальным коэффициентом GINI по
кросс-валидации (в среднем по фолдам). Потом рассматриваются всевозможные наборы из
двух переменных (с зафиксированной на предыдущем шаге первой переменной) и
выбирается та пара, на которой GINI по кросс-валидации максимален.

Далее рассматриваются всевозможные наборы из трех переменных (с зафиксированной на
предыдущем шаге двойкой переменных) и выбирается та тройка, на которой GINI по
кросс-валидации имеет наибольшее значение. И так далее, пока не упремся в некоторый
критерий останова. Его можно задать самостоятельно, например прирост качества при
добавлении очередной переменной к уже отобранным ниже порогового.

Hints. По желанию можешь прикрутить к данному алгоритму отбора дополнительный
функционал, мы только за.

Как обычно, жду файл с решением на почту. Будут вопросы - пиши. Спасибо!
```

Формат конечного результата

Файл Jupyter Notebook (.ipynb) или Python (.py) с пояснениями.

Форма загрузки результата

Пожалуйста, загрузи свой вариант ответа в формате zip-архива, используя инструмент «Загрузить решение». Необходимо сформировать единый zip-архив, содержащий решение одного или всех заданий по выбранной специальности.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания от эксперта после отправки собственной версии.