Направление Data Scientist, Alfa Digital

Аннотация

Добро пожаловать на виртуальную стажировку **Shift + Enter** от Alfa Digital — IT-подразделения Альфа-Банка, отвечающего за продуктовую разработку и запуск новых цифровых технологий.

Стажировка от Alfa Digital поможет тебе прокачать такие навыки, как:

- Написание SQL запроса, позволяющего объединять данные из разных таблиц.
- Исследование датасета в рамках разведочного анализа данных[1].
- Обоснование выбора скоринговой модели[2].

Развиваемые компетенции

По результатам выполнения заданий ты сможешь:

- Попрактиковаться в написании запроса с оператором JOIN.
- Узнать, почему при работе над скорингом существуют определенные требования к моделям.

А также будешь развивать аналитическое мышление, внимательность к деталям и поймешь, как Data Science используется в банковской сфере.

Описание подзадач

Команда дата сайентистов Alfa Digital приглашает тебя присоединиться к ним и помочь с автоматизацией процесса оценки кредитоспособности заемщиков на основе собранных данных, иными словами, кредитного скоринга.

Выполнение всего блока заданий займет у тебя не более 90-110 минут.

Рекомендуемый тайминг

1. 10-15 минут на первое задание.
2. 30-40 минут на второе задание.
3. 50-55 минут на третье задание.

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] Разведочный анализ данных (EDA, Exploratory Data Analysis) — предварительное исследование датасета с целью определения его основных характеристик и взаимосвязей между данными.

[2] Скоринговая модель — модель с определенным набором данных и процессами их обработки, задачей которой является оценка уровня риска заемщика.

Задание 1. Напиши SQL запрос для объединения данных

Твоим первым заданием станет написание запроса на языке SQL[3] для объединения двух таблиц с сохранением всех совпадающих записей. Эти данные послужат исходным датасетом для следующих задач в рамках твоей стажировки в Alfa Digital.

Мария[4], старший дата сайентист и по совместительству твой руководитель, прислала письмо с пояснениями к задаче.

Привет!

Тебе необходимо написать SQL запрос, чтобы объединить данные из двух таблиц, в которых содержится информация о наших заемщиках. В таблице Persons содержится информация о сумме предоставленного кредита (LIMIT_BAL) и возрасте заемщика, а в таблице Personal_information о его поле (SEX: 1 = мужской; 2 = женский), образовании (EDUCATION: 1 = аспирантура, 2 = университет, 3 = средняя школа, 4 = другое, 5 = нет образования, 6 = не хочу отвечать) и семейном положении (MARRIAGE: 1 = женат, 2 = холост, 3 = другое).

Persons

ID	LIMIT_BAL	AGE
1	20000	24
2	120000	26

Personal_information

Personal_ID	SEX	EDUCATION	MARRIAGE
1	2	2	1
2	2	2	2

При этом нужно сохранить только совпадающие по ID записи и отсортировать в порядке убывания результат по величине предоставленного кредита.

Hints. Чтобы твой запрос не был слишком громоздким, используй сокращения p и i для таблиц Persons и Personal_information соответственно.

С нетерпением жду твоего решения на почту.

Спасибо!

[3] SQL (от англ. Structured Query Language – «язык структурированных запросов») – это структурированный язык запросов, созданный для того, чтобы получать из базы данных необходимую информацию.
[4] Все имена и названия вымышленные, любые совпадения случайны. Данные заданий могут быть изменены в целях конфиденциальности

Полезные материалы

- Для проверки правильности запроса можно использовать сервис SQLite Online: в левом меню выбери PostgreSQL.
- Статья об операторе JOIN, который нужно использовать для объединения таблиц.

Формат конечного результата

Файл в формате .docx, содержащий SQL запрос.

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.

Задание 2. Проведи разведочный анализ данных

Ты отлично справился с первым заданием и готов поработать над вторым. Оно будет связано с анализом данных. Позже ты увидел письмо от Марии с пояснениями к задаче.

Привет,

Ты, наверное, слышал, что существуют определенные требования регулятора к интерпретируемости скоринговых моделей. Именно поэтому мы используем простые и понятные модели, например решающие деревья или логистическую регрессию.

Первым этапом при создании модели является анализ исходных данных. Сейчас мы активно работаем над автоматизацией этого процесса, поэтому просим тебя помочь провести EDA.

В процессе EDA мы считаем основные статистики и рисуем графики, чтобы найти тренды и связи внутри данных.

Для этого предлагаю тебе следующий план работы:

- Собрать Pandas DataFrame и проверить, есть ли пропуски в данных.
- Посчитать распределение целевой переменной в датасете.
- С помощью стандартных библиотек Python (Matplotlib, seaborn) построить графики распределения заемщиков по:
  а) возрасту,
  б) полу,
  б) образованию,
  в) семейному положению.
- Ответить на вопросы, проанализировав результаты EDA:
  а) Кого среди заемщиков больше: тех, у кого высока вероятность наступления дефолта в следующем месяце или тех, у кого такая вероятность равна нулю.
  б) Каков средний возраст заемщиков?
  в) Среди заемщиков преобладают мужчины или женщины?
  г) Верна ли гипотеза о том, что люди с высшим образованием склонны чаще брать кредиты?
  д) Различается ли распределение возрастов для мужчин и женщин в выборке?

Пожалуйста, пришли свой Jupyter Notebook с пояснениями и ответами на вопросы сегодня до конца рабочего дня.

Спасибо за помощь!

Полезные материалы

- Исходный датасет[5] с описанием данных.
- Статья про то, как проводить EDA.

Формат конечного результата

Файл Jupyter Notebook (.ipynb) с пояснениями.

[5] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients]. Irvine, CA: University of California, School of Information and Computer Science.

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.

Задание 3. Используй модель решающего дерева для скоринга и визуализируй своё дерево решений

У тебя за плечами появился успешный опыт работы над двумя заданиями, которые обычно выполняют дата сайентисты в Alfa Digital. Пришло время перейти к следующему. Открыв почту, ты внимательно изучаешь детали финального задания стажировки.

Привет!

Проблема кредитного скоринга является важнейшей составляющей процесса кредитования в банковской сфере. На основе результатов моделей кредитного скоринга, среди прочего, рассчитывается средний уровень вероятности дефолта (Probability of Default - PD). Дефолт возникает, когда заемщик прекращает вносить необходимые платежи по долгу.

Это может подвергнуть заемщика судебным искам и ограничить доступ к кредитам в будущем. Модель будет использоваться для оценки вероятности дефолта заемщика в процессе жизни кредита.

В качестве модели предлагаю использовать модель решающего дерева из-за простоты интерпретируемости результатов (особенно, если имеем дело с неглубокими деревьями).

Тебе необходимо:

- Разделить датасет на обучающую и тестовую выборки в соотношении 80:20 с использованием train_test_split.
- Используя обучающую выборку датасета, обучить модель, которая будет использоваться для оценки вероятности дефолта заемщика в процессе жизни кредита. Акцентируй внимание на том, какие гиперпараметры были подобраны для модели.
- Проверить работу модели с использованием тестовой выборки:
  - Посмотреть метрику ROC AUC на обучающей и тестовой выборках. Сильно ли меняется её качество?
  - Какие признаки получились наиболее важными (feature importance)?
  - Можно ли объяснить именно такое распределение признаков по важности?
- Визуализировать дерево в Jupyter.
  Hints.
  - Для построения модели используй библиотеку Scikit-learn.
  - Для отображения дерева в блокноте Jupyter используй функцию export_graphviz Scikit-learn, но если хочешь, то можно использовать и другие. Для построения дерева также необходимо установить graphviz и pydotplus.

Как обычно, жду файл с кодом на почту.

Спасибо!

Полезные материалы

Статья про то, как можно решить задачу кредитного скоринга.

Формат конечного результата

Файл Jupyter Notebook (.ipynb) 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.