Трек Хаб
Команда МАИ 2024

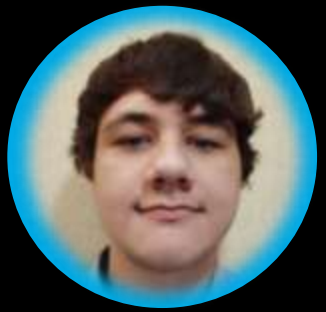
КОМАНДА

Артём Гаврилов
Капитан, разработчик

Михаил Фролов
Разработчик

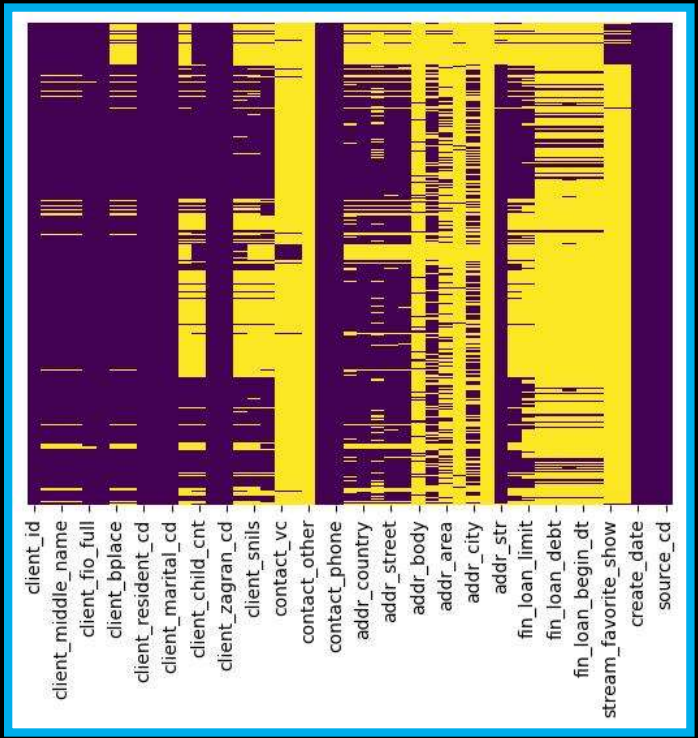
Дарья Фурлетова
Аналитик данных

Кира Калямина
Разработчик



Иван Мигурский
Разработчик

Анализ датасета




```
data.nunique()
```

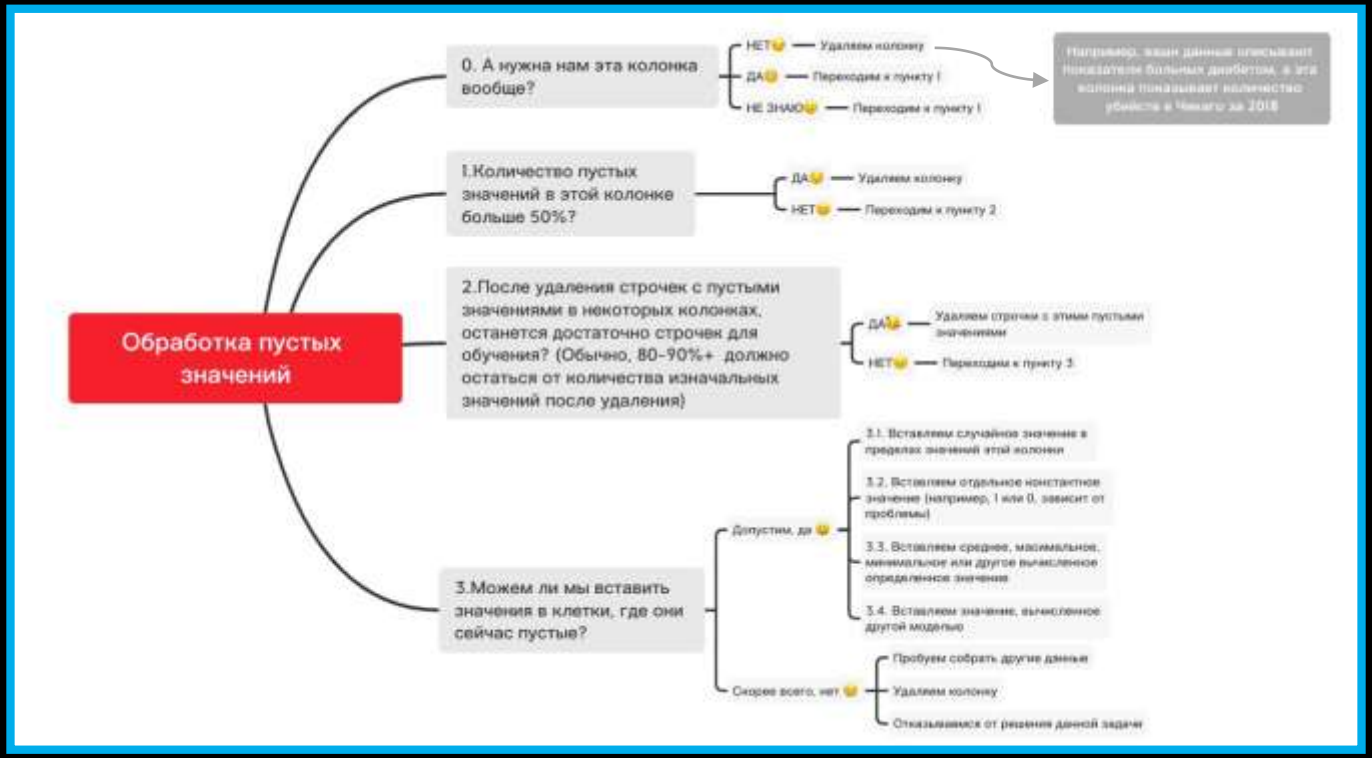
0.8s

```
client_id                985864
client_first_name         50093
client_middle_name        93756
client_last_name         418546
client_fio_full          473820
client_bday               31807
client_bplace            222881
client_cityzen              202
client_resident_cd            3
client_gender                 3
client_marital_cd             3
client_graduate               3
client_child_cnt              4
client_mil_cd                 3
client_zagran_cd              3
client_inn               244733
client_snils             243155
client_vip_cd               172
contact_vc                50911
contact_tg                50935
contact_other                 3
contact_email            258085
contact_phone            256731
addr_region                  96
addr_country                 60
...
stream_duration            1300
create_date              985864
update_date              985864
source_cd                     7
dtype: int64
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

ИМПУЛЬС ТІ

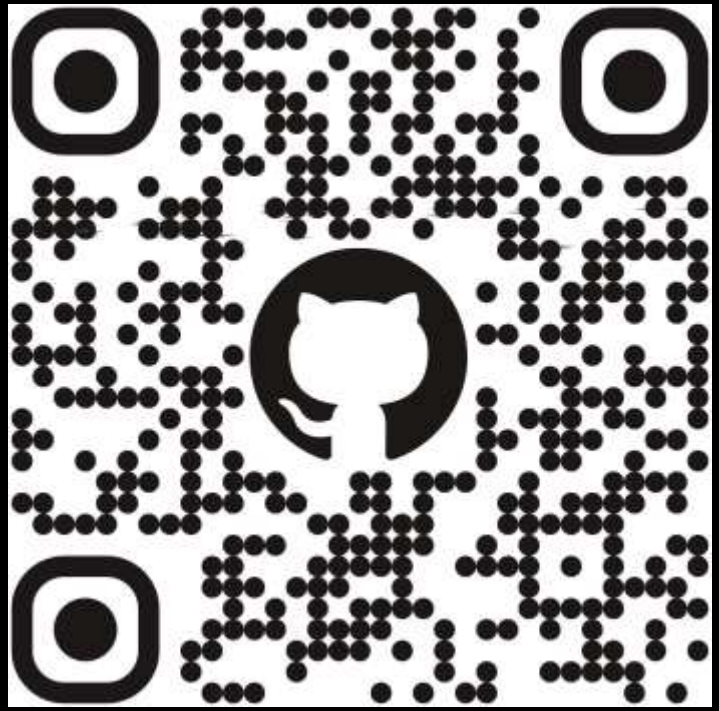
Идея предобработки датасета



Колонки для группировки


```
[
    'client_fio_full’,
    'client_yo’,      -  заменена дата рождения клиента на его
возраст, скорректированы выбросы
    'client_cityzen’,
    'addr_region’,
    'addr_country’,
    'addr_city’,
    'fin_rating’,
    'fin_loan_percent’,
    'stream_favorite_show’,
    'source_cd’
]
```

Ссылка на репозиторий



Оценка результата

- Создана программа с корректно работающими: загрузкой датасета, предобработкой датасета, кластеризацией, выделением «золотой записи» для каждого кластера, выдачей результата в виде scv-файла.
- Обработка данных происходит с учётом их актуальности, частоты и полноты. Методология работы с данными описана в текстовом файле, находящемся в github-репозитории проекта.
- Использована актуальная версия python 3.12.1 и современные библиотеки для работы с данными и машинного обучения (Pandas, Scikit-learn). Нет зависимости от внешних сервисов, т.к. использованы только open-source библиотеки.
- Масштабируемость достигается разделением предобработки данных и группировки уже предобработанных данных по кластерам. Кроме того возможно использование библиотек, позволяющих параллельно обрабатывать данные для увеличения скорости работы.

Спасибо за внимание

ИМПУЛЬС TI 2024

Инженеры больших идей

TI