Transaction Fraud Detection

Илюткин Алексей Евгеньевич

Июнь’22

Проект ПЕРЕЗАПУСК

О себе

- ФИО: Илюткин Алексей Евгеньевич
- Образование: Высшее (ННГУ им.Н.И.Лобачевского, Менеджмент организации)
- Опыт в Сбере:
  - ПЦП ППиИК/Отдел по работе с операционными рисками
  - Должность: специалист
  - Основной функционал:
    - контроль и устранение ошибок в процессе мониторинга исполнения условий ипотечных сделок;
    - выгрузка данных по ипотечным кредитам из SQL;
    - сведение и анализ данных предоставляемых смежными подразделениями в Excel,
    - подготовка отчетности;
    - обнаружение и регистрация инцидентов операционного риска;
  - Системы и процессы: АС ЕКП/ЕКС, PL/SQL Developer, АС Transact SM, MS Excel, ипотечное кредитование
- Город: Нижний Новогород
- Контакты: +7 (920) 111-47-11, aeilyutkin@sberbank.ru

Описание проекта

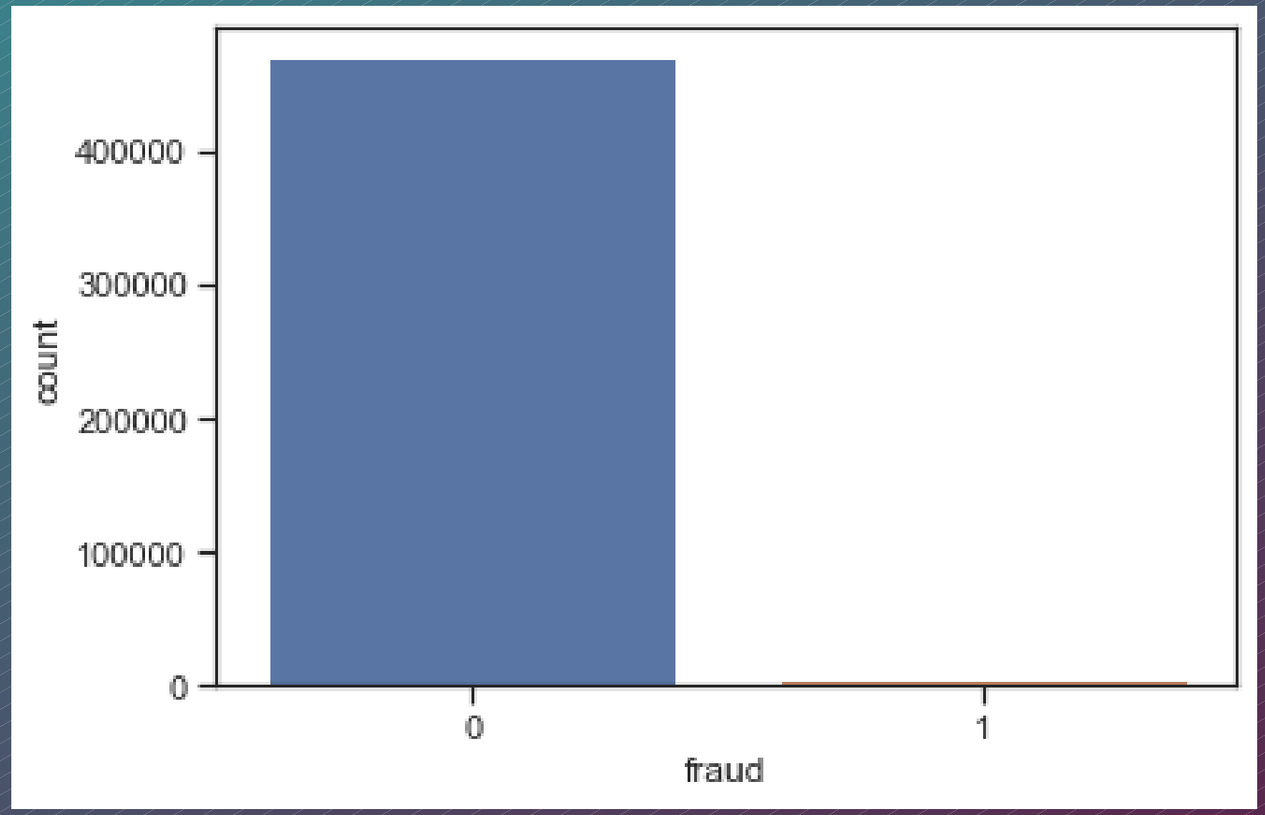
- *Выявление мошеннических операций различными методами (эвристическим и с помощью ML), улучшение качества модели с помощью SMOTE, визуализация результатов работы алгоритма*
- *Ссылка на репозиторий: https://github.com/Ilutkin/qsl_tasks_sber_da/blob/main/TFD_project.ipynb*

Бизнес логика

Модель данных

- *Данные представляют собой реляционную модель, т.е. представлены в виде таблицы:*
- *Информация о данных:*

	step	customer	age	gender	zipcodeOri	merchant	zipMerchant	category	amount	fraud
0	0	C1093826151	4	M	28007	M348934600	28007	es_transportation	4.55	0
1	0	C352968107	2	M	28007	M348934600	28007	es_transportation	39.68	0
2	0	C2054744914	4	F	28007	M1823072687	28007	es_transportation	26.89	0
3	0	C1760612790	3	M	28007	M348934600	28007	es_transportation	17.25	0
4	0	C757503768	5	M	28007	M348934600	28007	es_transportation	35.72	0

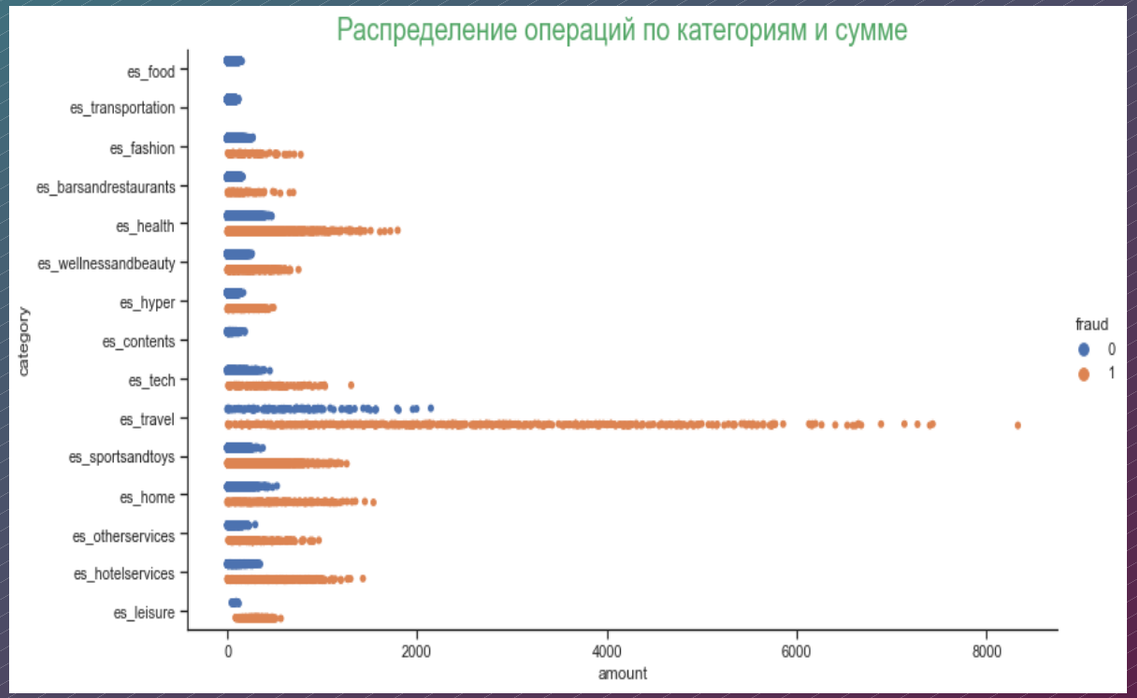


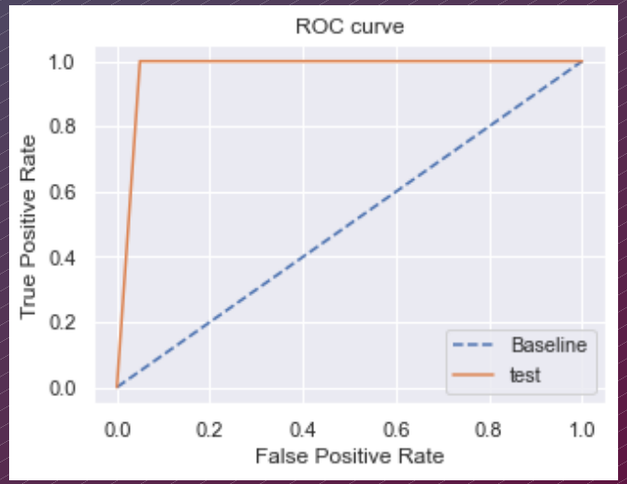

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 475714 entries, 357755 to 325735
Data columns (total 10 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   step         475714 non-null  int64
 1   customer     475714 non-null  object
 2   age          475714 non-null  object
 3   gender       475714 non-null  object
 4   zipcodeOri   475714 non-null  int64
 5   merchant     475714 non-null  object
 6   zipMerchant  475714 non-null  int64
 7   category     475714 non-null  object
 8   amount       475714 non-null  float64
 9   fraud        475714 non-null  int64
dtypes: float64(1), int64(4), object(5)
memory usage: 39.9+ MB
```

Задание №1

- Проверим распределение сумм операций по признаку фрода и построим диаграмму распределения операций по категориям и сумме
- Оперделим количество и процент операций по признаку фрод в групперовке по продавцу
- Для предсказания используем следующие условия, к фроду можно отнести:
  - Операций больше максимальной суммы не фрод операций в каждой из категорий покупок
  - Операции компаний соотношение фрода в кторых более 0 %
- Рассчитаем метрики для оценки алгоритма:

```
Gini_test: 0.9504719590770199
Accuracy: 0.9510716477898579
Precision: 0.1983744317399090 7
Recall: 1.0
F1: 0.3310725370732268
```



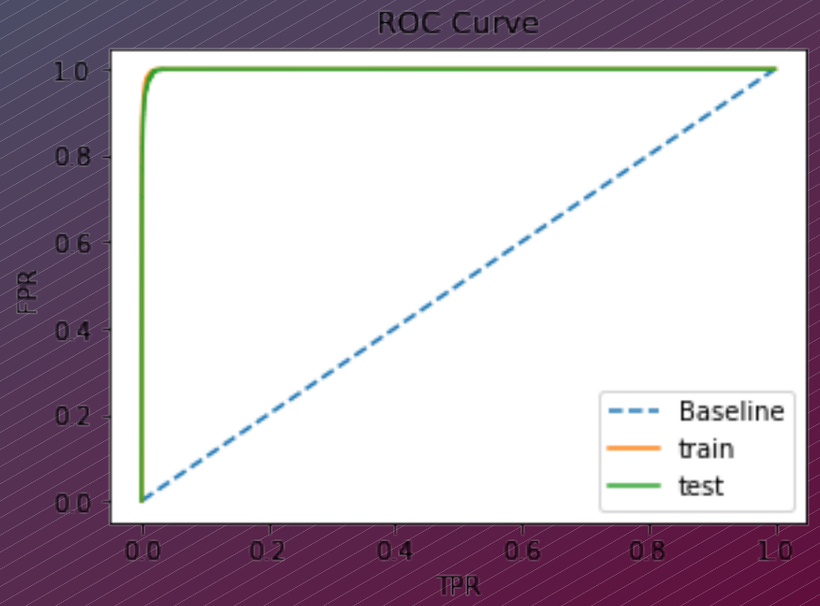


	merchant	0	1	percent
0	M1294758098	5.0	144.0	28.800000
1	M3697346	12.0	229.0	19.083333
2	M732195782	70.0	415.0	5.928571
3	M1873032707	33.0	177.0	5.363636
4	M1353266412	10.0	51.0	5.100000
5	M980657600	237.0	1182.0	4.987342
6	M2080407379	9.0	29.0	3.222222
7	M857378720	25.0	78.0	3.120000

Задание №2

- *Для решения задачи используем LightAutoML со следующими параметрами:*
  - *Тип решаемой задачи: Binary*
  - *Целевой признак: fraud*
- *Определим основные фичи их значимость:*
- *Рассчитаем метрики для оценки алгоритма:*

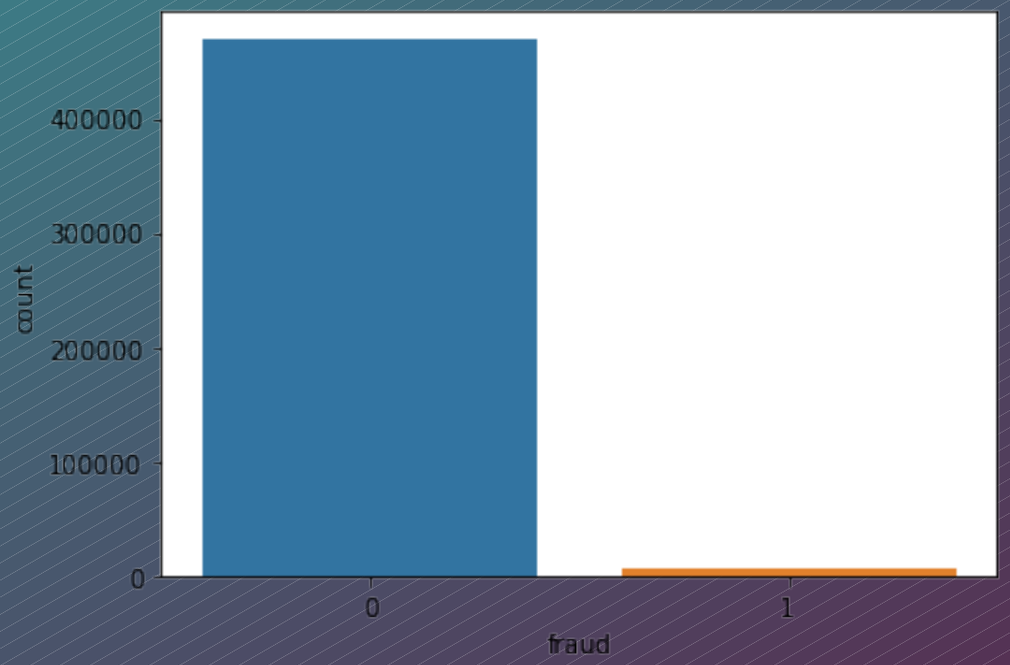
	Feature	Importance
0	amount	226278.858319
1	merchant	57230.306353
2	category	35959.464072
3	customer	15626.839403
4	step	12453.857836
5	age	2626.175654
6	gender	852.215438

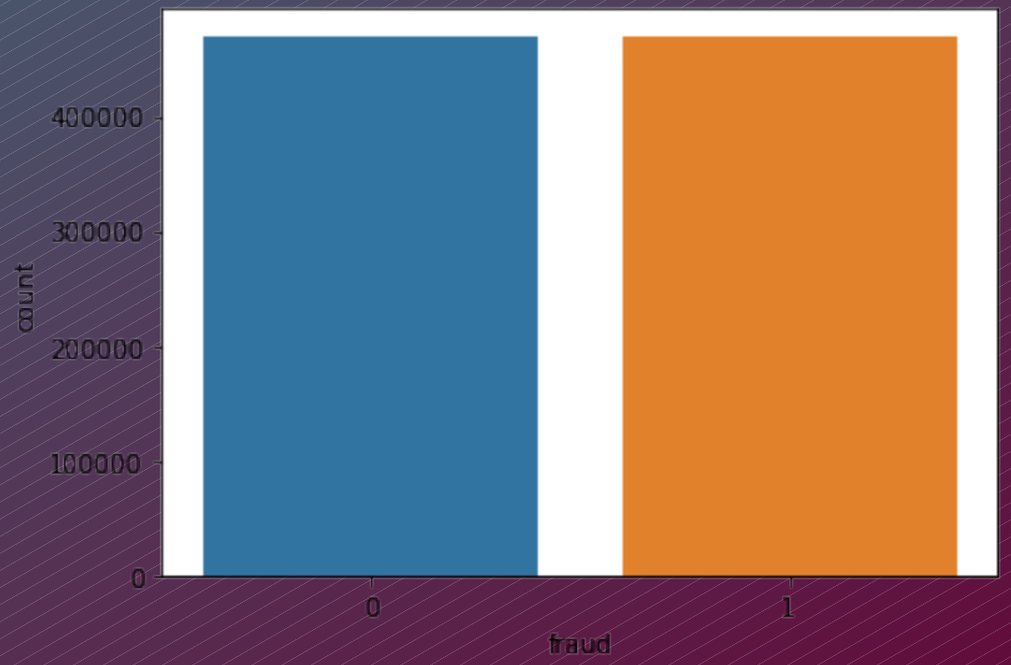



```
Gini_train: 0.9984827593055206
Gini_test: 0.9975712265261714
f1_score 0.5436267071320182
precision_score 0.9951388888888889
recall_score 0.3739561586638831
```

Задание №3

- *Для балансировки датасета используем метод SMOTE*
  - *Размер выборки до балансировки: (475714, 10); после: (939908, 10)*
  - *Количество значений фрод (0/1) выравнялось*

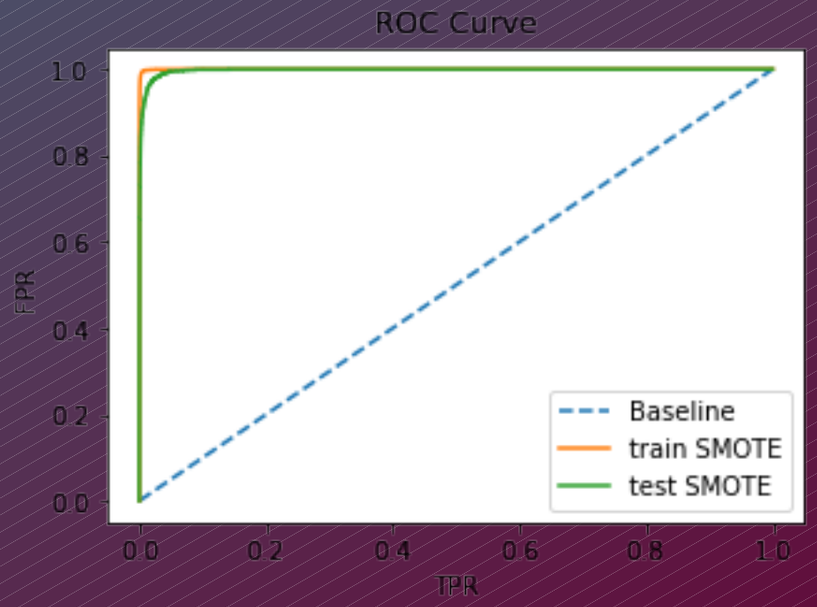




Задание №3

- *Для решения задачи используем LightAutoML на новых данных*
- *Определим основные фичи их значимость:*
- *Рассчитаем метрики для оценки алгоритма:*

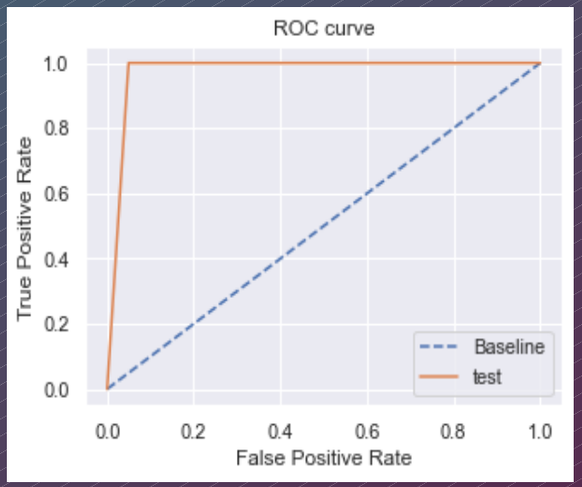
	Feature	Importance
0	amount	5190492.637061
1	category	1304650.320722
2	merchant	657248.485301
3	customer	132593.180845
4	step	81291.731516
5	gender	74064.964382
6	age	30750.614767



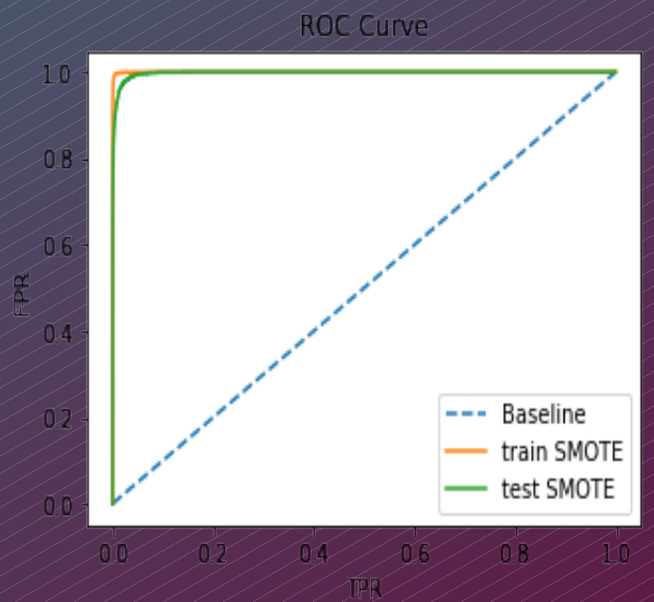

```
Gini_train: 0.9995275520169113
Gini_test: 0.9952660698259224
f1_score 0.527563499529633
precision_score 0.9736111111111111
recall_score 0.3618064516129032
```

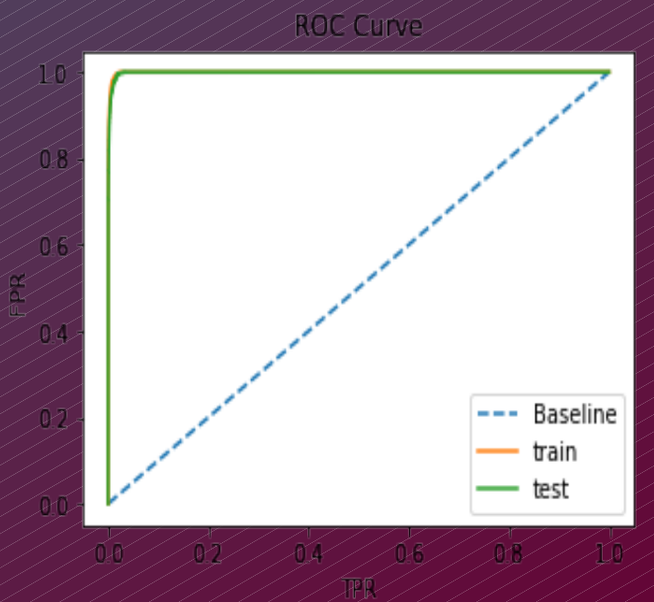
Задание №4

- *Т.к. при решении задачи Класс 1 (мошенничество) в целом скорее важнее, чем класс 0, а дисбаланс классов очень большой основной метрикой для определения точности предсказаний будет являться ROC-кривая*
- *Сравним метрики из 3-х задач*
- *Наибольшее значение Gini получено в Задаче № 2*



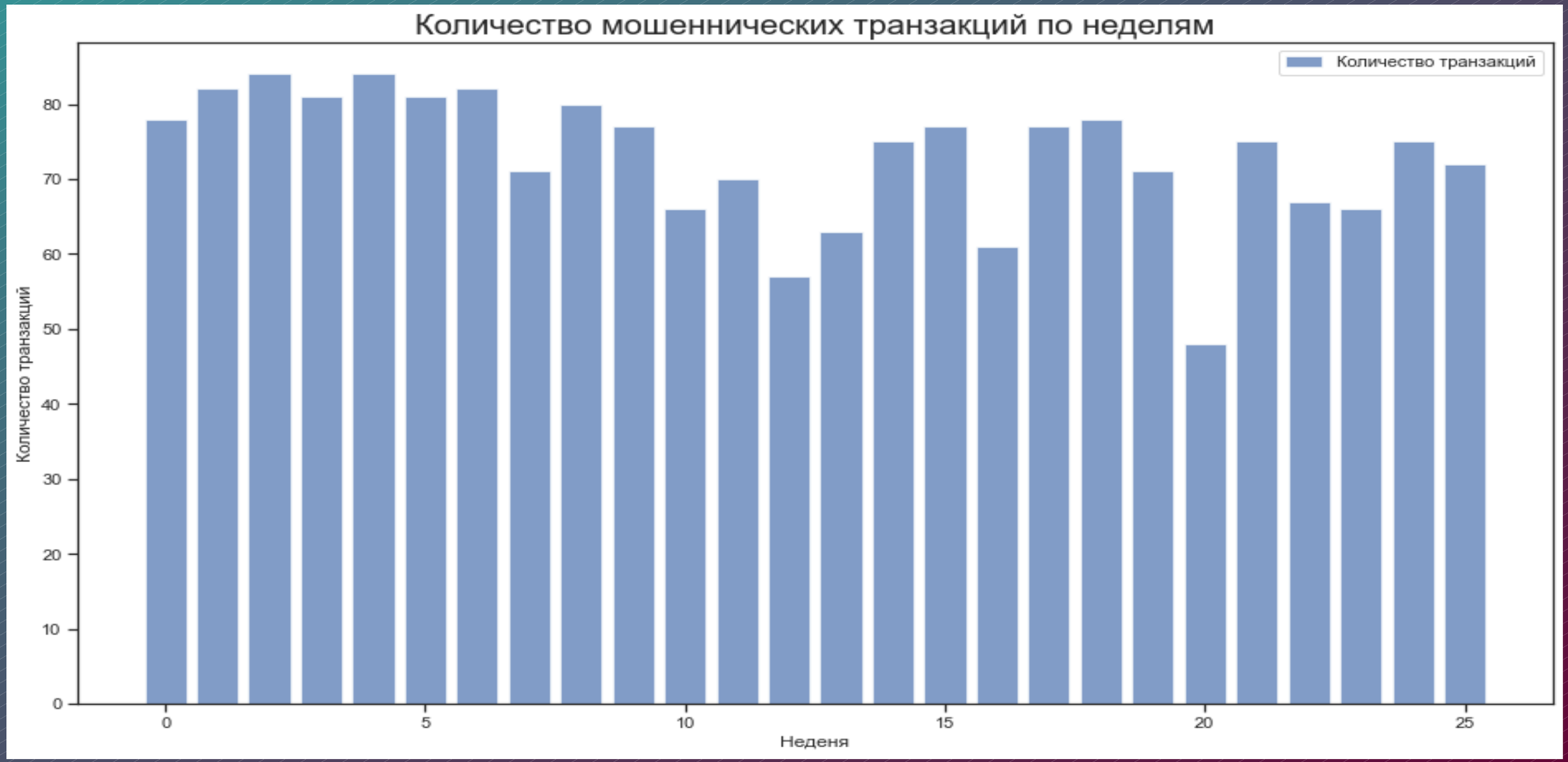
	Gini
Задача 1	0.9504719590770199
Задача 2	0.9975712265261714
Задача 3	0.9952660698259224





Задание №4

- *Информация о количестве фрод-операций за неделю*



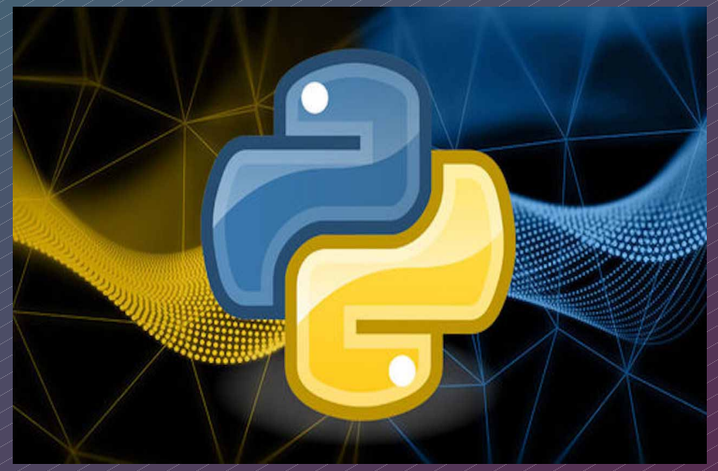
Задание №4

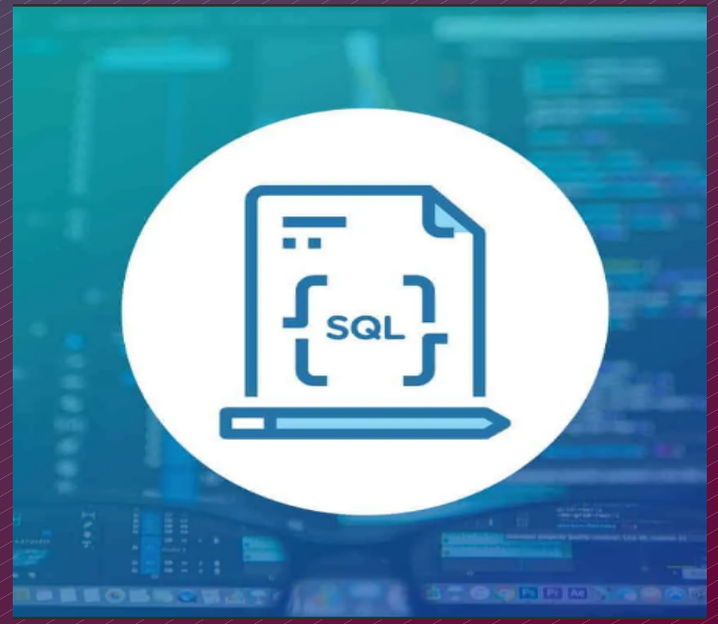
- *Информация о последних 10 фрод транзакциях:*

	Идентификатор покупателя	Возрастная группа	Пол	Идентификатор продавца	Категория покупки	Объем транзакции
0	C2113737065	2	F	M480139044	es_health	466.35
1	C2078888187	3	F	M980657600	es_sportsandtoys	167.90
2	C1193034305	4	F	M2011752106	es_hotelservices	491.56
3	C1459810780	2	F	M1873032707	es_hotelservices	291.93
4	C1847495093	3	M	M980657600	es_sportsandtoys	396.66
5	C2113737065	2	F	M732195782	es_travel	3631.60
6	C1886871597	4	F	M480139044	es_health	487.12
7	C910454738	4	F	M980657600	es_sportsandtoys	203.18
8	C76313164	3	F	M732195782	es_travel	2761.61
9	C2080088206	5	F	M3697346	es_leisure	213.62

Используемые технологии

- *Язык программирования Python*
- *Язык запросов к базам данных SQL*
- *Фреймворк LightAutoML (LAMA) для функций ML*
- *Технология оверсемплинга SMOTE*





СПАСИБО ЗА ВНИМАНИЕ