ПОЧТА РОССИИ

Разработка модели предсказания потери почтовых отправлений

logistics

Разработка модели предсказания потери почтовых отправлений.

Необходимо разработать модель предсказания потери почтовых отправлений на основе совокупности данных об отправлениях.

Создано в рамках «Цифровой прорыв 2022»

Автор: Аликин Геннадий Александрович

Данные

Для прогнозирования потери почтовых отправлений, представлены данные об посылках, атрибутах и месте операций.

Всего 28 признаков для каждой операции над посылкой и отметка о потере или не потере отправления.

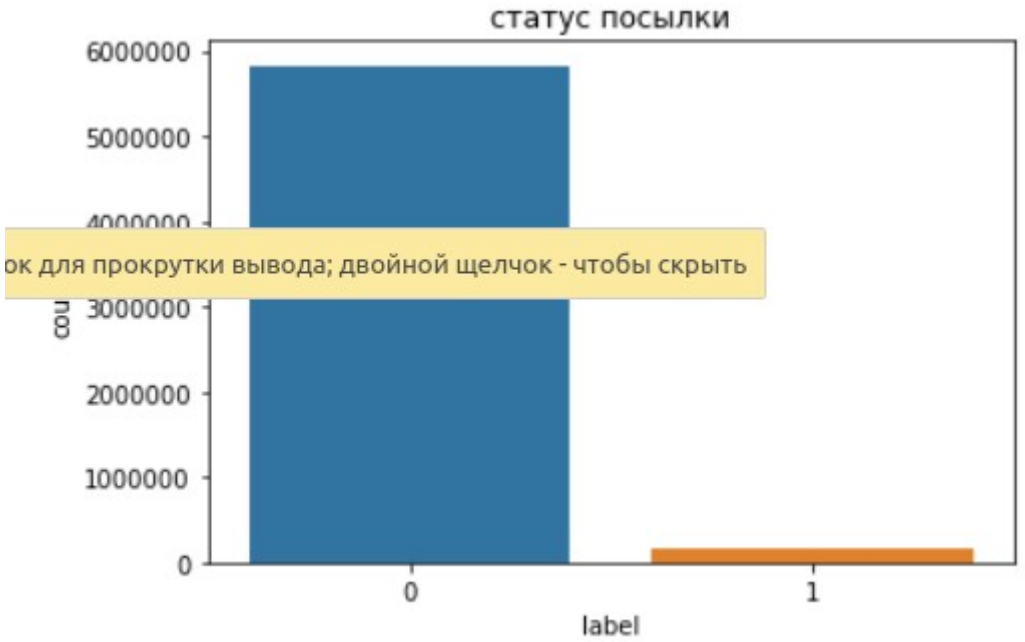
Обучающий сет

Представлены данные посылок, помечены пропавшие:

"**отправление было вручено или возвращено отправителю**": 0,

„**операция, на которой цифровой след оборвался, т. к. отправление потеряно**“: 1

В обучающем наборе присутствует значительный дисбаланс классов, необходима балансировка.



Формирование списка признаков для обучения модели

Из имеющихся данных о посылках сформирован датасет для обучения, где каждому действию добавлены:

- отдельный признак регион
- регионам общий объем посылок который через них проходит
- потерянные в каждом регионе посылки
- Процент потерянных в регионе посылок
- Добавлены характеристики операций над посылками с наибольшей вероятностью потери !

Финальный обучающий сет

Предсказания потери почтовых отправлений на основе отобранных из финального датасета одиннадцати признаков

Ввод [25]:
```
1 X
```

Out[25]:

	type	priority	class	dist_qty_oper_login_1	total_qty_oper_login_0	total_qty_over_index_and_type	is_wrong_phone_number	is_wrong	area	total_qty_
0	18	1	0.0	42.0	58950.0	779126.0	0	0	182	
1	4	1	0.0	914.0	83318932.0	132175590.0	0	0	3	
2	19	1	0.0	62.0	3233068.0	6479360.0	0	1	175	
3	19	1	0.0	55.0	653280.0	2714208.0	0	0	93	
4	18	1	0.0	16.0	27911.0	344830.0	0	0	183	
...	...	...	...	...	...	...	...	...	...	
5999995	4	1	0.0	1089.0	116432632.0	180702765.0	1	1	3	
5999996	19	1	0.0	31.0	144063.0	1911433.0	0	0	74	
5999997	19	3	0.0	186.0	10648.0	60624000.0	0	0	3	
5999998	19	3	0.0	105.0	4972424.0	20063762.0	0	1	3	
5999999	13	1	4.0	1.0	353.0	1894.0	0	0	22	

6000000 rows × 11 columns

Балансировка обучающего сета.

Проблема дисбаланса классов в обучающей выборке частично решена генерацией «синтетических» примеров операций (SMOTE). Таким образом удалось повысить точность модели на представленных в меньшинстве **операциях с потерей посылок.**

Выбор Модели

Бинарная классификация произведена **нейросетью** на базе библиотеки **keras**.

Технические особенности:

Python, keras, sklearn.

Уникальность: Используются стандартные подходы для бинарной классификации средствами библиотеки keras.

Метрики

На публичном датасете удалось добиться требуемой метрики.

0.825678

Что примерно соответствует представленным данным собственной валидации модели.

```
54648/54648 [==============================] - 40s 734us/step
              precision    recall  f1-score   support

           0       0.91      0.78      0.84    874353
           1       0.81      0.92      0.86    874353

    accuracy                           0.85   1748706
   macro avg       0.86      0.85      0.85   1748706
weighted avg       0.86      0.85      0.85   1748706
```

Контакты

Автор: Аликин Геннадий Александрович

Город Екатеринбург.

☎ +7(922)647-93-38

Email: genalll@mail.ru

Репозиторий решения:

https://github.com/genalll/mlsport3

Скринкаст:

https://disk.yandex.ru/i/2xpklCA4zcW-0g