цифровой прорыв
сезон: ИИ

РОССИЯ – СТРАНА ВОЗМОЖНОСТЕЙ

Задача «Разработка модели предсказания потери почтовых отправлений»

Введение

Весь процесс доставки — от приема в отделении до вручения получателю — состоит из большого числа операций. Отправление запаковывают, перевозят на склад и транспортируют между сортировочными пунктами. Если доставка едет из-за границы, то дополнительно появляются операции на зарубежной и российской таможне.

Ускорить и удешевить доставку помогают крупные логистические хабы. Там мелкие грузы сортируются и отправляются в соседние регионы или собираются в новые контейнеры для отправки в другие макрорегионы. Таким образом, почтовые отправления путешествуют по сети сортировочных центров, как кровь по капиллярам, и в конце концов добираются в любые точки нашей страны.

Несмотря на высокий уровень системы безопасности, по-прежнему остается риск пропаж или порчи отправлений: перемещений и операций с посылками очень много, кроме того, в процессе может сыграть человеческий фактор. Точное предсказание пропаж и их локализация позволит повысить надежность системы — гарантировать доставку отправлений в срок и снизить расходы на транспортировку.

Условие задачи

Разработка модели предсказания потери почтовых отправлений.

Описание входных значений

- **train.csv** — файл, содержащий данные о посылках, включая статус отправления;
- **test.csv** — файл, содержащий данные о посылках для предсказания ;
- **sample_solution.csv** — пример файла для отправки;

Расшифровка столбцов

	Наименование поля	Описание данных
1	oper_type + oper_attr	Тип и атрибут операции
2	index_oper	Индекс места операции
3	type	Обозначение типа объекта почтовой связи
4	priority	Приоритет объекта
5	is_privatecategory	Y - является отделением закрытого типа
		N - иначе
6	class	Значение класса или категории объекта почтовой связи
7	is_in_yandex	Y - адрес отделения связи отображается в Яндекс-картах
		N - иначе
8	is_return	Y - Отправление движется в направлении возврата отправителю
		N - иначе
9	weight	Вес в граммах
10	mailtype	Код вида отправления
11	mailctg	Код категории почтового отправления
12	mailrank	Код разряда почтового отправления
13	directctg	Код классификации отправления
14	transport_pay	Общая сумма платы за пересылку в условной валюте
15	postmark	Код отметки
16	name_mfi	Наименование вложений (указано через запятую), указывается на бирке отправления
17	weight_mfi	Суммарная масса вложений
18	price_mfi	Суммарная стоимость вложений в условной валюте
19	dist_qty_oper_login_1	Количество уникальных имен операторов, задействованных в обработке данного типа отправлений (mailtype) на конкретном индексе, по которым возможно идентифицировать оператора
20	total_qty_oper_login_1	Количество отправлений с уникальным именем операторов, задействованных в обработке данного типа отправлений (mailtype) на конкретном индексе, по которым возможно идентифицировать оператора
21	total_qty_oper_login_0	Количество отправлений данного типа (mailtype), которые были обработаны неизвестным оператором на этом индексе

22	total_qty_over_index_and_type	Общее количество отправлений данного типа (mailtype), прошедших обработку на этом индексе
23	total_qty_over_index	Общее количество отправлений, прошедших обработку на этом индексе
24	is_wrong_sndr_name	Есть ли явные признаки, что имя отправителя введено некорректно? 1 - да, 0 - иначе
25	is_wrong_rcpn_name	Есть ли явные признаки, что имя получателя введено некорректно? 1 - да, 0 - иначе
26	is_wrong_phone_number	Есть ли явные признаки, что номер телефона получателя введен некорректно? 1 - да, 0 - иначе
27	is_wrong_address	Есть ли явные признаки, что адрес получателя введен некорректно? 1 - да, 0 - иначе
28	label	**СТРОКА ДЛЯ ПРЕДСКАЗАНИЯ** 1 - операция, на которой цифровой след оборвался, т.к. отправление потеряно 0 - отправление было вручено или возвращено отправителю

Метрика

В качестве метрики выступает комбинация двух значений:
Результирующее значение выглядит как:

$$Result = 0.1 * Recall + 0.9 * AUC\_ROC$$

Recall считается как:

$$recall = \frac{TP}{TP + FN}$$

AUC_ROC считается как:

$$AUC = \frac{1 + TPR - FPR}{2}$$

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

TP (True Positive) — количество верно угаданных значений того, что посылка пропала.

FN (False Negative) — количество значений, где было предсказано, что посылка пропадет, а она не пропала.

FP (False Positive) — количество значений, где было предсказано, что посылка не пропадет, а она пропала.

TN (False Negative) — количество верно угаданных значений того, что посылка не пропала.

Правила чемпионата:

1. С момента открытия датасета до момента завершения приема решений репозиторий участника, в котором он ведет разработку по задаче текущего чемпионата, должен оставаться закрытым.

2. Участник обязан открыть доступ к репозиторию на чтение по ссылке (которая была прикреплена в ЛК в поле «Ссылка на код (гитхаб)») не позднее чем в течение 12 часов с момента окончания дедлайна отправки решений на чемпионате.

3. Согласно п. 5.8 Положения в процессе верификации решений организаторы и технические эксперты, проверяющие решения участников, вправе назначить интервью с участниками чемпионата. Участник получит приглашение и ссылку на интервью не позднее чем за 12 часов до публикации итогового лидерборда. Пропуск интервью участником является поводом для дисквалификации.

4. Организаторы вправе исключить участника из призовых позиций лидерборда за непредоставление одного из артефактов решения задачи: тизера, скринкаста, презентации, ссылки на репозиторий.

5. Организаторы вправе дисквалифицировать участника в случае выявления плагиата кода или несоблюдения Положения проекта.

6. Участник, получивший 2 дисквалификации за сезон проекта, попадает в чёрный список с дальнейшим отстранением от участия в чемпионатах до конца сезона.