Classification with a Tabular Vector Borne Disease Dataset

Артемов Макар
Филиппенко Павел

Описания данных

В таблице данных соревнования содержатся сведения о **симптомах пациентов**.

В качестве **признаков** объектов используются **названия симптомов**. В каждой ячейке, соответствующей конкретному симптому находится значение **1.0 или 0.0**, что соответственно означает, **наблюдается ли этот признак у пациента или нет**.

Целевой меткой является категориальная величина – **название заболевания**. Всего представлено **11 различных целевых меток**. Таким образом, цель соревнования состоит в решении задачи **классификации на несколько**.

В качестве метрики в данном соревновании использовалась метрика mean average precision (MAP@k).

	id	sudden_fever	headache	mouth_bleed	nose_bleed	muscle_pain	joint_pain	vomiting	rash	diarrhea	hypotension
0	0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0
1	1	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0
2	2	0.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0
3	3	0.0	0.0	1.0	1.0	1.0	1.0	0.0	1.0	0.0	1.0
4	4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0

Список целевых меток:

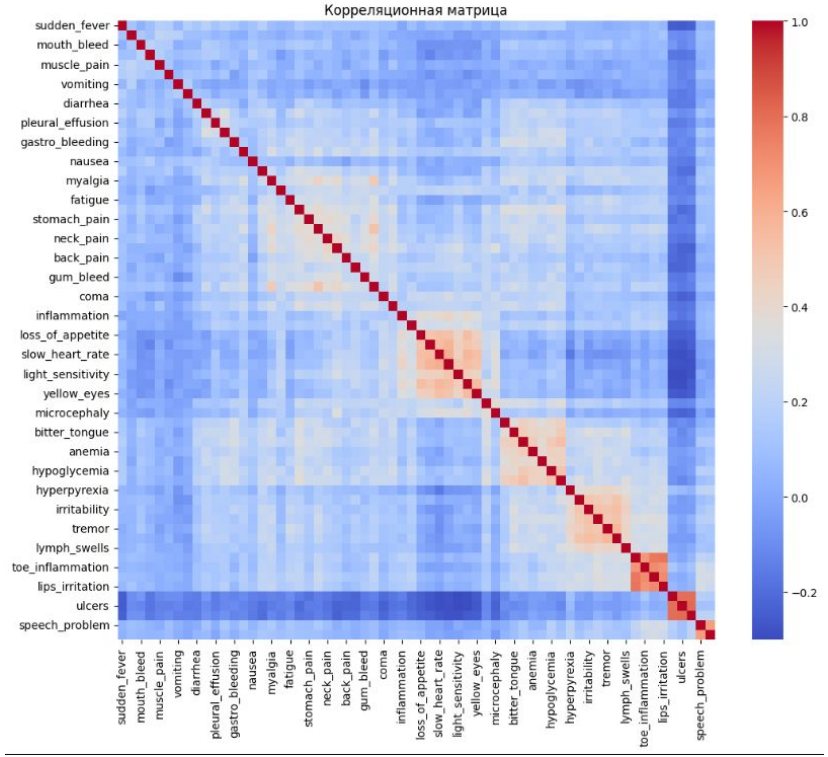
- *Chikungunya*
- *Dengue*
- *Japanese_encephalitis*
- *Lyme_disease*
- *Malaria*
- *Plague*
- *Rift_Valley_fever*
- *Tungiasis*
- *West_Nile_fever*
- *Yellow_Fever*
- *Zika*

Размеры тренировочного датасета (707, 66)
Таким образом, всего 707 тренировочных объектов и 65 признаков (последний столбец – целевая метка).

Комментарии по обработке датасета.

- датасет не имеет пропущенных значений
- нет смысла исследовать данные на выбросы
- целевая метка имеет категориальный тип, было принято решение использовать позиционное кодирование целевой метки
- исследование распределений признаков выявило, что некоторые из них распределены неравномерно (есть очень редкие симптомы)
- исследование распределения целевой переменной показало, что имеется незначительное смещение в сторону некоторых болезней

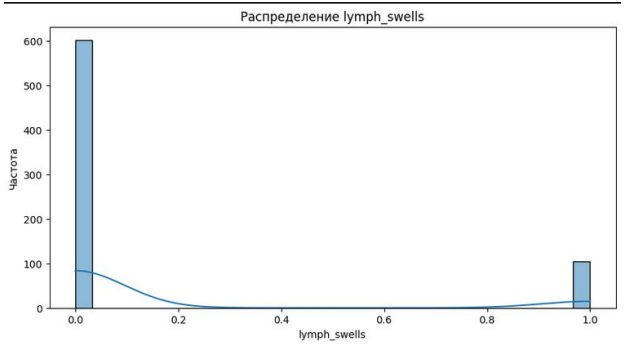
Матрица корреляции признаков

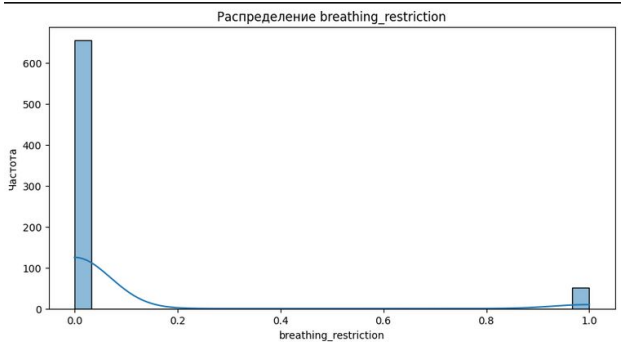


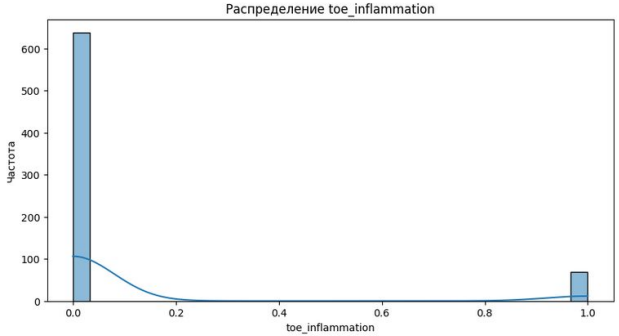
Пары признаков с наибольшей корреляцией:

(finger_inflammation, breathing_restriction) – 0,7684
(lips_irritation, breathing_restriction) – 0,7783
(ulcers, itchiness) – 0,7945
(toenail_loss, itchiness) – 0,7974

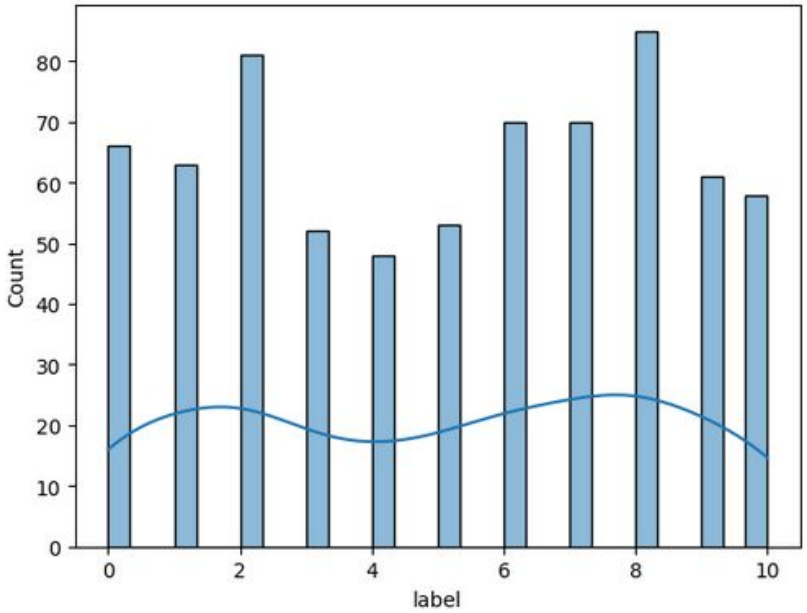
Примеры неравномерного распределения признаков







Распределение целевой переменной



где:

0: 'Chikungunya',
1: 'Dengue',
2: 'Japanese_encephalitis',
3: 'Lyme_disease',
4: 'Malaria',
5: 'Plague',
6: 'Rift_Valley_fever',
7: 'Tungiasis',
8: 'West_Nile_fever',
9: 'Yellow_Fever',
10: 'Zika'

Генерация признаков

По итогу исследования различных способов генерации признаков значительный прирост дало объединение (кластеризация) определенных комбинаций симптомов. Новые признаки дали значительный прирост в оценке модели.

В итоге были использованы следующие объединения (кластеры):

- Кластер по схожести симптомов: объединить те симптомы, которые имеют схожие наиболее частые заболевания, целевой прогноз - малярия, чума, желтая лихорадка, вирус Зика
- Кластер Лайма: объединить те симптомы, которые чаще всего встречаются при болезни Лайма
- Кластер Тунгиоза: объединить те симптомы, которые чаще всего встречаются при Тунгиозе
- Кластер Чикунгунья: объединить те симптомы, которые НИКОГДА не встречаются при заболевании Чикунгунья.

Генерация признаков

```
similar_columns = ['loss_of_appetite', 'urination_loss',
                   'slow_heart_rate', 'abdominal_pain',
                   'light_sensitivity', 'yellow_skin', 'yellow_eyes']
X['similar_cluster'] = X[similar_columns].sum(axis=1)

lyme_columns = ['jaundice', 'weight_loss', 'weakness',
                'back_pain', 'sudden_fever', 'myalgia',
                'chills', 'orbital_pain', 'digestion_trouble']
X['lyme_cluster'] = X[lyme_columns].sum(axis=1)

tungiasis_columns = ['ulcers', 'toenail_loss', 'itchiness']
X['tungiasis_cluster'] = X[tungiasis_columns].sum(axis=1)

chikungunya_columns = ['convulsion', 'finger_inflammation', 'speech_problem',
                       'toenail_loss', 'ulcers', 'itchiness',
                       'lips_irritation', 'breathing_restriction', 'toe_inflammation',
                       'paralysis', 'stomach_pain', 'confusion',
                       'irritability', 'bullseye_rash']
X['chikungunya_cluster'] = X[chikungunya_columns].sum(axis=1)

columns = [col for col in X if col != 'prognosis']
X[columns] = X[columns].astype(int)
```

Построение моделей

В ходе решения задач были исследованы следующие модели и методы:

- SVM
- Gradient boosting
- Random forest
- Stacking

На каждой модели проводился подбор наилучших гиперпараметров.

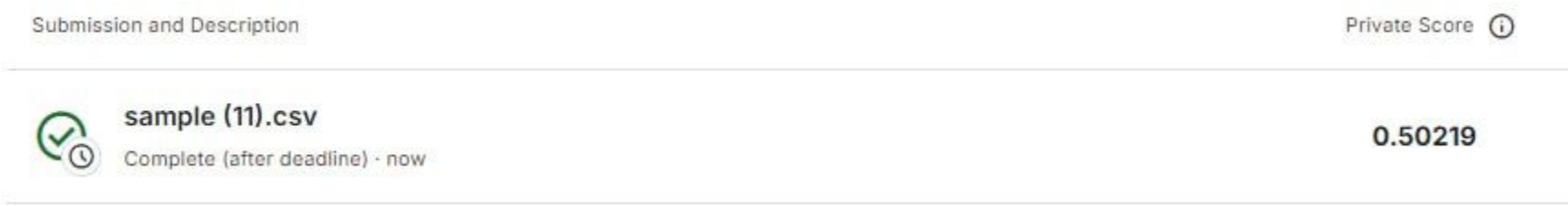
В качестве baseline модели использовался random forest на 100 деревьях.

Построение модели

По итогу исследования наилучшим образом себя показала модель random forest с гиперпараметрами + сгенерированные признаки:

- n_estimators = 9000
- min_samples_split = 6
- min_samples_leaf = 11

Итоговая наилучшая оценка MAP@k: 0.50219



Baseline

Submission and Description	Private Score
sample (8).csv Complete (after deadline) · 2d ago	0.47039

Random forest + feature engineering

Submission and Description	Private Score
sample (11).csv Complete (after deadline) · now	0.50219