Kaggle: Classification with a Tabular Vector Borne Disease Dataset

Андрей Ткачик
Гор Аперян

Кафедра БИТ СберТех, МФТИ, 2025

Задача и данные

Тип задачи: Мультиклассовая классификация с ранжированием (MAP@3)
Данные:

- 11 различных заболеваний
- 64 бинарных признака (симптомы)
- Необходимо предсказать топ-3 наиболее вероятных диагноза

Метрика оценки: Mean Average Precision at K (MAP@3)

Анализ данных

- Анализ распределения классов
- Визуализация: график распределения показал относительно сбалансированные классы
- Heatmap корреляции симптомов и заболеваний
- Ключевой инсайт: разные заболевания имеют уникальные паттерны симптомов

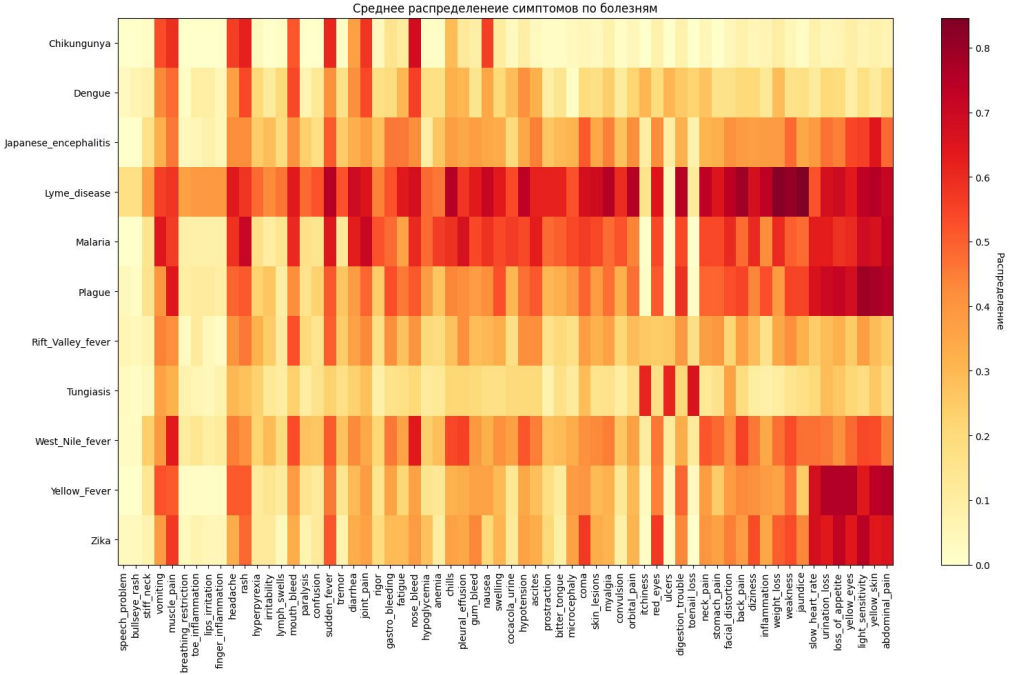
Baseline

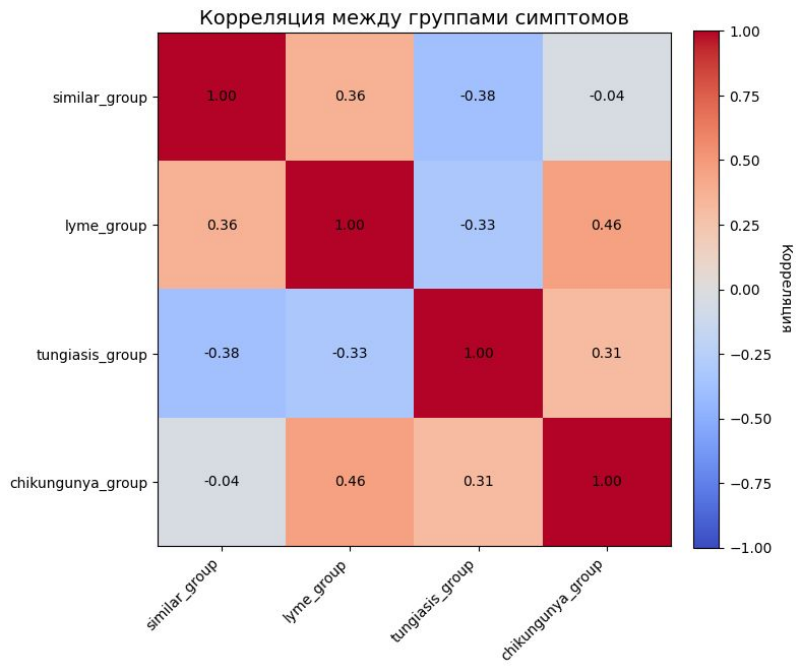
- Модель: Random Forest Classifier
- Оптимизация гиперпараметров через Optuna (30 trials)
- Лучшие параметры baseline:
  - n_estimators: 2500
  - max_depth: 30
  - min_samples_split: 7
  - min_samples_leaf: 8
  - bootstrap: False
- Результат baseline: MAP@3 ≈ 0.49

Feature Engineering

- Создание групповых признаков на основе медицинских знаний:
  - **similar_group**: общие симптомы (потеря аппетита, желтуха и др.)
  - **lyme_group**: симптомы болезни Лайма (лихорадка, боли, воспаление)
  - **tungiasis_group**: специфичные симптомы тунгиоза (язвы, зуд)
  - **chikungunya_group**: очень редкие симптомы чикунгуньи (судороги, воспаления)
- Анализ корреляции между группами симптомов. Группировка неплохо уменьшила корреляцию между признаками.

Feature Engineering





Улучшенная модель

- Повторная оптимизация с новыми признаками
- Новые лучшие параметры:
  - n_estimators: 5000
  - max_depth: 24
  - min_samples_leaf: 16
  - min_samples_split: 15
  - bootstrap: True
- Финальный результат: MAP@3 ≈ 0.50

Submission and Description	Private Score	Public Score
submission (5).csv Complete (after deadline) · 2h ago	**0.50000**	**0.37527**
submission_baseline (2).csv Complete (after deadline) · 2h ago	**0.48793**	**0.36754**

Спасибо за внимание!