Решение тестового задания

НА СТАЖИРОВКУ ПО НАПРАВЛЕНИЮ "ИНТЕЛЛЕКТУАЛЬНЫЕ ТЕХНОЛОГИИ В ЗДРАВООХРАНЕНИИ" ИТМО НЦКР

Автор: Камышева Анна

Генерация признаков

Всего 11 признаков

+ Cases.csv

- **Готовые признаки**
  - *diagnosis*
  - *case_type*
  - *case_patient_condition*
- **Обработанные признаки**
  - *case_time*: производное от case_start и case_end - промежуток времени в днях между этими событиями

+ birthday_gender.csv

- *gender*
- *age* - возраст пациента на момент case_end на основе birthday_date

+ diaries.CSV

- *ibm* - ближайший замер к моменту case_end из cases,csv среди всех замеров для пациента
- *pressure_sys* - ближайший замер к моменту case_end из cases,csv среди замеров для данного случая
- *pressure_dias* - аналогично pressure_sys
- *relapses* - является ли случай рецидивом
- *relapses_num* - число записей со случаями в таблице cases, предшествующие данному для пациента

Предобработка данных

Приведение количественных переменных к типу float**:**

- Замена строк " " на NaN
- Замена неконвертируемых значений на NaN (колонка ibm)

Удаление строк с неизвестным значением таргетной переменной

Удаление выбросов:

Замена выбросов на пропущенные значения:

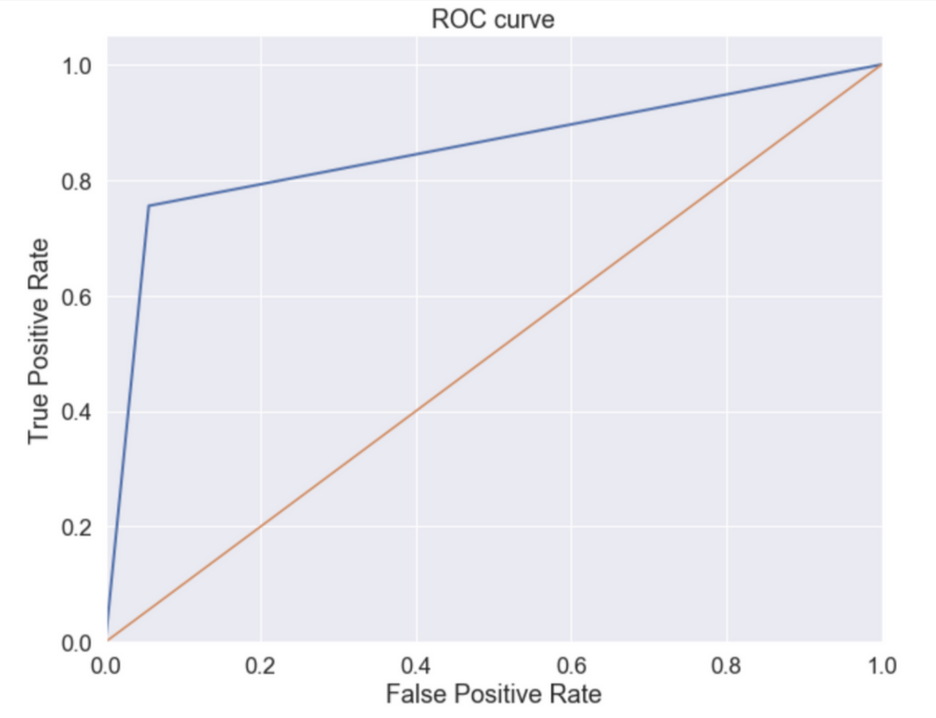
- case_time < 0
- ibm > 204
- pressure_dias > 220

Удаление околодубликатов в категориальных признаках:

case_patient_condition - слияние 10 вариантов "удовлетворительно" в 1

LightGBM

- Работает с категориальными признаками
- Работает с пропущенными значениями
- Быстрый



1

Step 1

Label-кодирование категориальных переменных

2

Step 2

Удаление признака relapses (дублирует relapses_num)

3

Step 3

Деление на тестовую и обучающую выборки (30/70)

4

Step 4

Подбор гиперпараметров на кроссвалидации

LightGBM

F1- мера - основная метрика

Гиперпараметры

- learning_rate=0.016
- max_depth=30
- n_estimators=200,
- num_leaves=140

Кросс-валидация

- f1 - 0.8302
- ROC AUC - 0.9190
- Accuracy - 0.8730

Тестовая выборка

- f1 : 0.8214
- ROC AUC: 0.8497
- accuracy : 0.8685
- precision : 0.9003
- recall : 0.7553

Random Forest

- Не работает с категориальными переменными
- Не работает с пропущенными значениями
- Позволяет удобно отбирать признаки

1

Step 1

One-hot-encoding
(997 признаков)

2

Step 2

Замена пропущенных значений на 0

3

Step 3

Выделение 100 наиболее значимых признаков

4

Step 4

Подбор гиперпараметров

Random Forest

Гиперпараметры

- bootstrap=False,
- max_depth=60
- min_samples_leaf=2
- min_samples_split=10
- n_estimators=1200,
- random_state=42

Кросс-валидация

- f1 - 0.82424
- ROC AUC - 0.9114
- Accuracy - 0.8699

Тестовая выборка

- f1 : 0.8185
- ROC AUC: 0.8474
- accuracy : 0.8667
- precision : 0.8993
- recall : 0.75096

XGBoost

1

Step 1

Использование выборок для LightGBM (дает лучший скор)

2

Step 2

Использование выборок для Random Forest с 10 самыми значимыми признаками

3

Step 3

Подбор параметров

XGBoost

Гиперпараметры

- colsample_bylevel=1.0
- colsample_bytree=0.419
- gamma=10.965
- learning_rate=0.2525
- max_depth=8
- min_child_weight=0.2271
- subsample=0.8

Кросс-валидация

- f1 - 0.8272
- ROC AUC - 0.9168
- Accuracy - 0.8717

Тестовая выборка

- f1 : 0.8202
- ROC AUC: 0.8487
- accuracy : 0.8680
- precision : 0.9022
- recall : 0.7517

Логистическая регрессия

- Не работает с категориальными признаками
- Не работает с пропущенными значениями
- Требует нормализации признаков
- Отбор признаков через L1-регуляризацию

1

Step 1

One-hot encoding

2

Step 2

Заполнение пропущенных значений через медиану значений рассматриваемой переменной, сгруппированной по значению переменной "АМБУЛАТОРНО" (самый значимый признак в других моделях)

3

Step 3

Нормализация количественных переменных

4

Step 4

Отбор значимых признаков через L1-регуляризацию

5

Step 4

Подбор гиперпараметров

Логистическая регрессия

Гиперпараметры

- C=0.616
- max_iter=1000
- penalty='l1'
- solver='saga'

Кросс-валидация

- f1 - 0.8217
- ROC AUC - 0.9029
- Accuracy - 0.8700

Тестовая выборка

- f1 : 0.8168
- ROC AUC: 0.8460
- accuracy : 0.8670
- precision : 0.9107
- recall : 0.7404

Выводы

- Таким способом удается решить задачу классификации с
  f1 - мерой = 0.82
  ROC AUC = 0.85
  accuracy = 0.87
- Лучший скор дает алгоритм LightGBM, хотя все результаты очень близки
- Наиболее важным для предсказания признаком оказался casy_type. В частности категория "АМБУЛАТОРНО". Также среди значимых признаков некоторые отдельные диагнозы, case_time, давление, возраст, ibm, relapses_num.

Что еще можно было бы сделать будь больше времени?

Solution 1

Выделить признаки из таблицы anamnesis.

Solution 2

Разобраться с индексацией пациентов (в ней содержатся символы обыкновенно не характерные ключам в таблицах, возможно это приводит к ошибкам).

Solution 3

Попробовать другие способы обработки пропущенных значений (через таргетную переменную) или формировать датасет для обучения без пропущенных значений.