Analysis

13 October 2020 09:40

Jan 2004 – Dec 2016

Blakey : Feb 2005 — Sep 2014

we will do:

Jan 2005 —

Get all patients from Jan 2016 – Dec 2018

Now, get all data for these patients:

Codes Description

- Analysis_13Oct2020 : Extract all features / outcomes and select data
- Analysis_08Feb2021 : Bring all together

15th Oct 2020

Pat_id :: start from earliest,

pat-id date BTS step

BTS Steps :

① SABA (is it only inhaled?)
② ICS , or , LTRA
③ ICS_LABA , or, ICS (high dose only)
④ (ICS_LABA , and , LTRA) , or, (LABA, LTRA, ICS (high dose only))
⑤ OCS

02nd March 2021 :

Categorical features → ready for one-hot encoding

- Sex : male, female
- smoking: never, current, former
- PEF : not recorded, 60-80, . . .
- Eosinophil:
- Device Type:
-

} One-hot encoding

age, BMI } categorisation?

- average daily ICS dose
- prescribed daily ICS dose

} normalise

09th March 2021

discretize BMI → one-hot encoding?
→ ordinal coding ?

Data Description

Number of attacks:

Recode Outcomes

months			outcome
0-3	2017/3/31	2017/1/1	1
3-6	2017/6/30	2017/4/1	2
6-9	2017/9/30	2017/7/1	3
9-12	2017/12/31	2017/10/1	4
12-15	2018/3/31	2018/1/1	5
15-18	2018/6/30	2018/4/1	6
18-21	2018/9/30	2018/7/1	7
21-24	2018/12/31	2018/10/1	8

Total outcome counts: all in 3 month blocks

0-1, vs ≥2 : 212,441 vs 24,033

0,1,2,3 vs ≥4 : 232,571 vs 3,903

outcome period	number
0-3 months	15,888
3-6	11,083
6-9	11,320
9-12	16,202
12-15	16,277
15-18	11,153
18-21	10,375
21-24	14,030

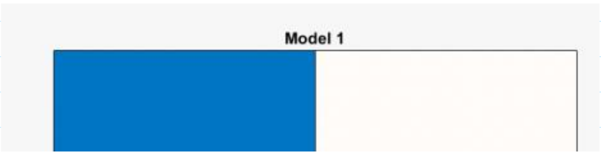
simple addition: 106,328

170,250 → no event
42,191 → 1 event
14,921 → 2 events
5,209 → 3 events

Vast majority is fine, but it's the small proportion where you have events...

19th March 2021

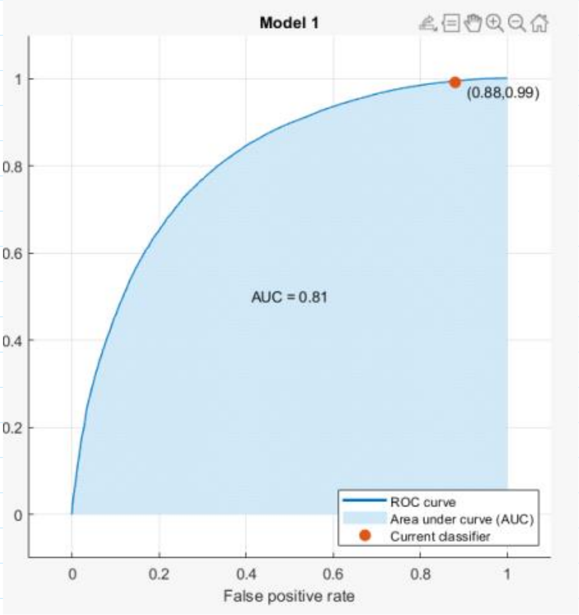
0 vs any:

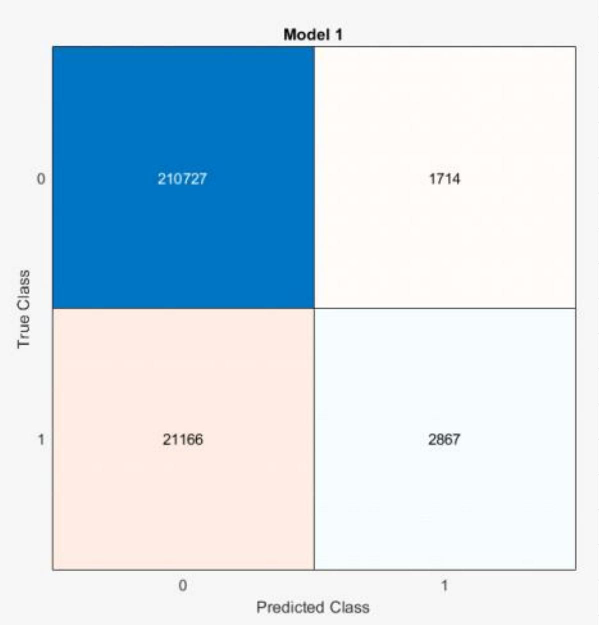


19 March 2021

0 vs any:

0,1 vs 2+ and using Logistic Regression



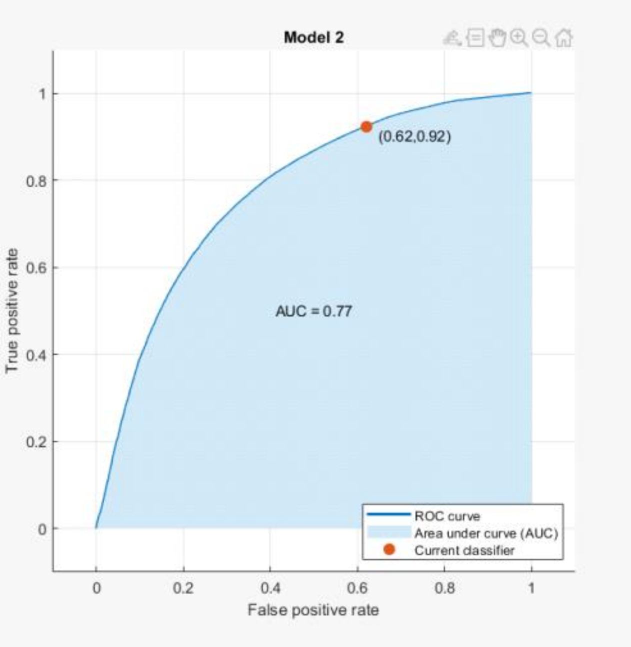


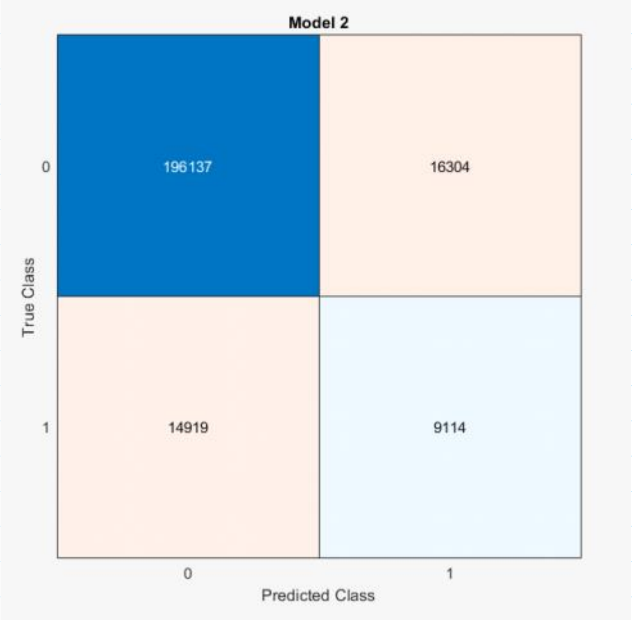
True Class \ Predicted Class	0	1
0	210727	1714
1	21166	2867

Screen clipping taken: 19/03/2021 16:09

This is a heavily unbalanced dataset though!

0,1 vs 2+ and using SVM





Screen clipping taken: 19/03/2021 16:14

Screen clipping taken: 19/03/2021 16:14

<u>0,1 vs any and using Logistic Regression</u>

<u>Two approaches, as classes are imbalanced:</u>

① Two-class classifier but resample:

② Novelty detection

① Two-class classifier:

26th May 2021

- Data Loading & Pre-processing

- Split the data into training, and testing

output measure: 0/1 vs 2 or more

- use k-fold cross validation and assess:
  - Logistic Regression
  - Novelty detection → Parzen, → Gm
  - NB
  - svm
  - DT

- Performance Evaluation

27th May 2021

feature (1) : 1, 2, 1, 5] mean rank
(2) :

<u>Feature importance</u>:

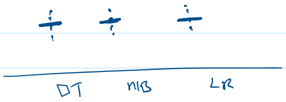
- numOCS
- BTS step

- number of asthma attacks
- number of PC consultations (general)
- age
- number of OCS with LRTI
- " " acute Resp events
- " " antibiotics with LRTI
- " " antibiotics (general)
-

Performance metric	0 vs 1/..	0/1 vs	2 or
AUC			
Accuracy			
P-Recall AUC			
Sensitivity			
Specificity			

NB, DT, LR

(A) AUC score :
(0-1 vs more)



(B) LR, Show table with metrics

(C) Feature importance

(D) Independent testing

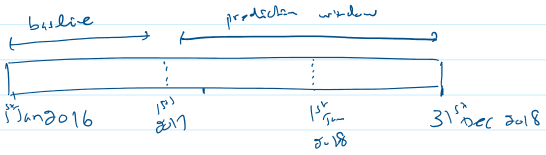
→ To prevent attacks.....
→ 0,1,2 vs more

- 2 or more may be relevant
-

→ Next one year

(A)
- External validation on CPRD . . .

04/06/2021 11:33



-Methods :

Dataset :
- OPCRD

- ~ xc million
  - ~ network of PC across UK

06th July 2021

- 5-12
- upto 18
- over 16

- 3 or more
- 4 or more } biologics
- 2 or more } steroids

- 65 thousand admissions in 4-5 million