CURACEL HEALTH

... Smart Practice

Spot A Bug

Automated Flagging of False Medical Insurance Claims

Consulting project for Curacel Systems

Shunling (Shirley) Guo, Ph.D.

20A Insight Health Data Science

INSIGHT

Catching false claims is costly

Curacel (Africa)

5% of total claims
$ 0.8 million

Process Claims
$ 0.17 million

40 +/- 28 days

USA

346,000 claim adjusters
(Bureau of Labor Statistics)
> $20 Billion/year

45 days

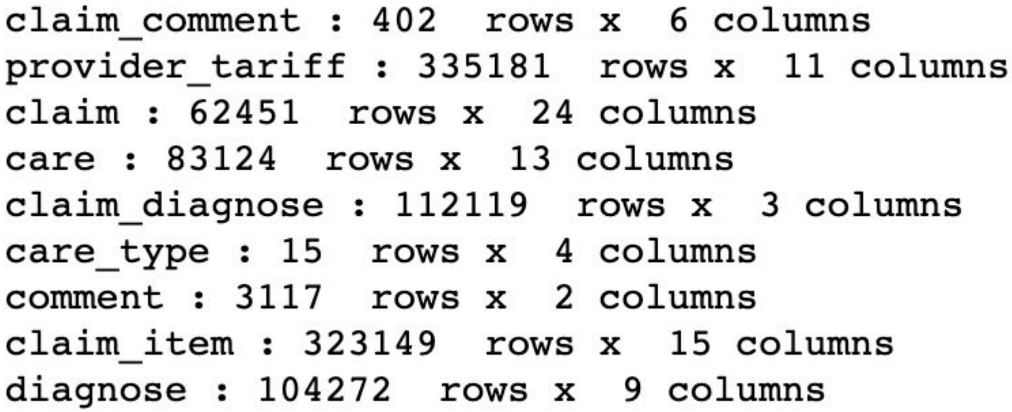
Premium

Data

9 Tables

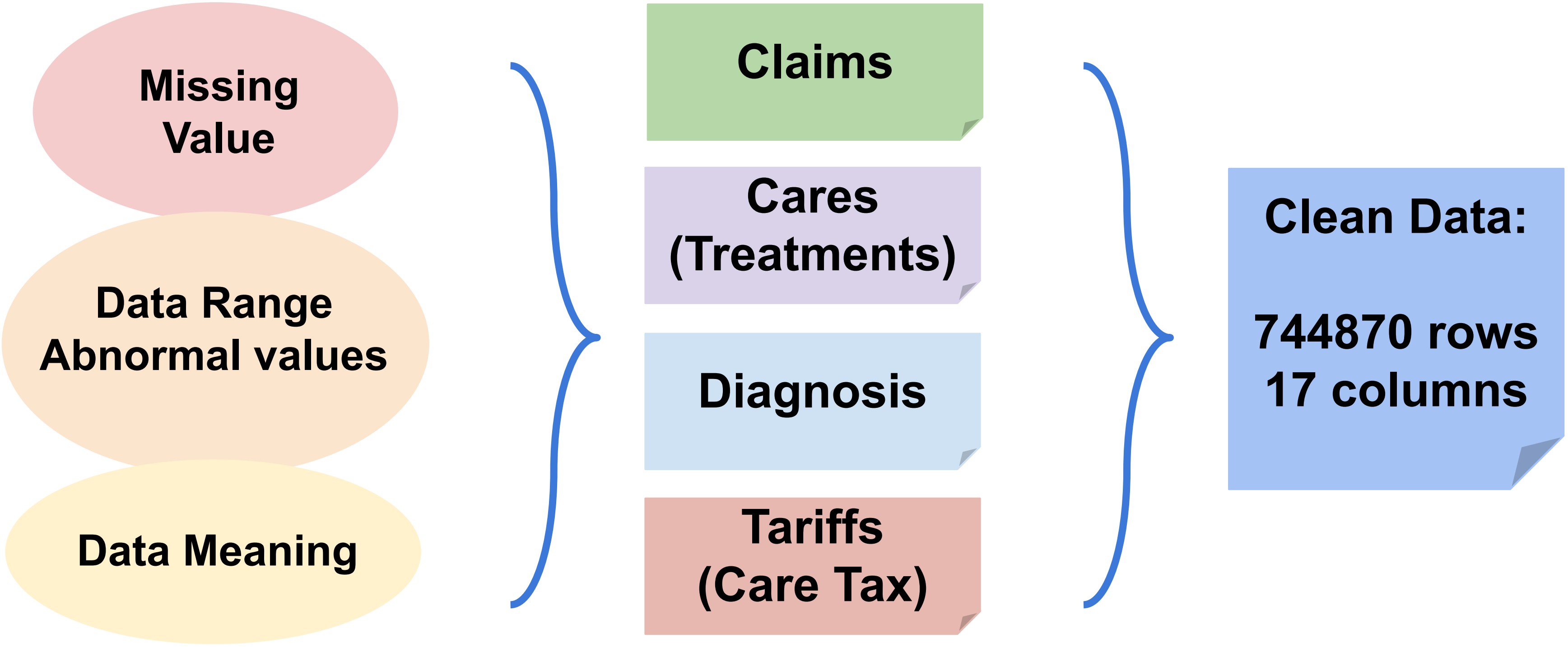
1+ Million rows

50+ attributes



```
claim_comment : 402  rows x  6 columns
provider_tariff : 335181  rows x  11 columns
claim : 62451  rows x  24 columns
care : 83124  rows x  13 columns
claim_diagnose : 112119  rows x  3 columns
care_type : 15  rows x  4 columns
comment : 3117  rows x  2 columns
claim_item : 323149  rows x  15 columns
diagnose : 104272  rows x  9 columns
```

Data cleaning and wrangling



Missing Value

Data Range Abnormal values

Data Meaning

Claims

Cares (Treatments)

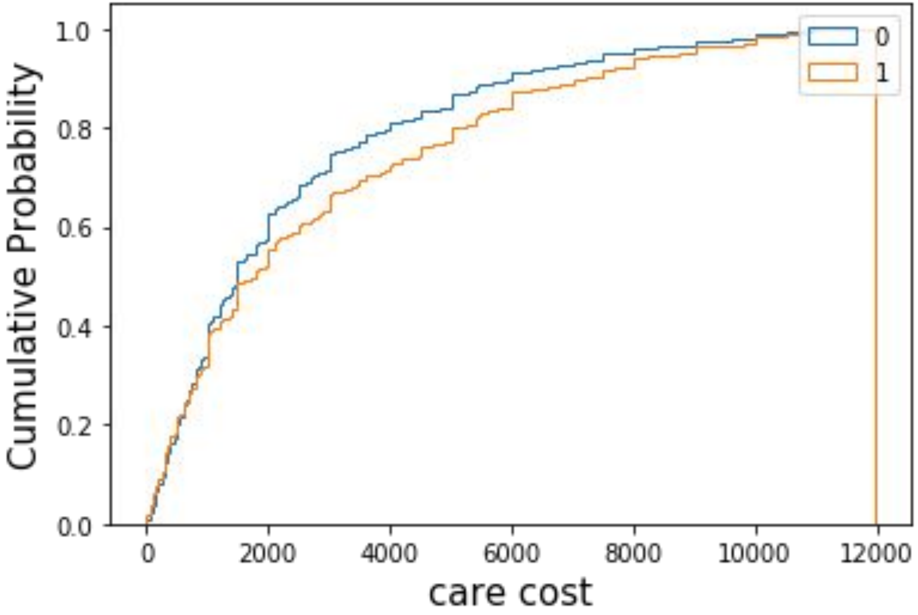
Diagnosis

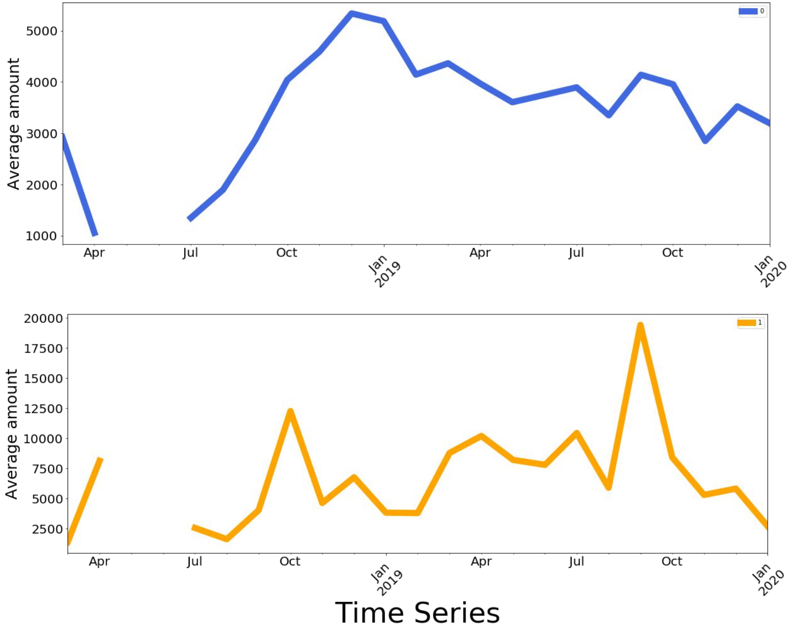
Tariffs (Care Tax)

Clean Data:

744870 rows
17 columns

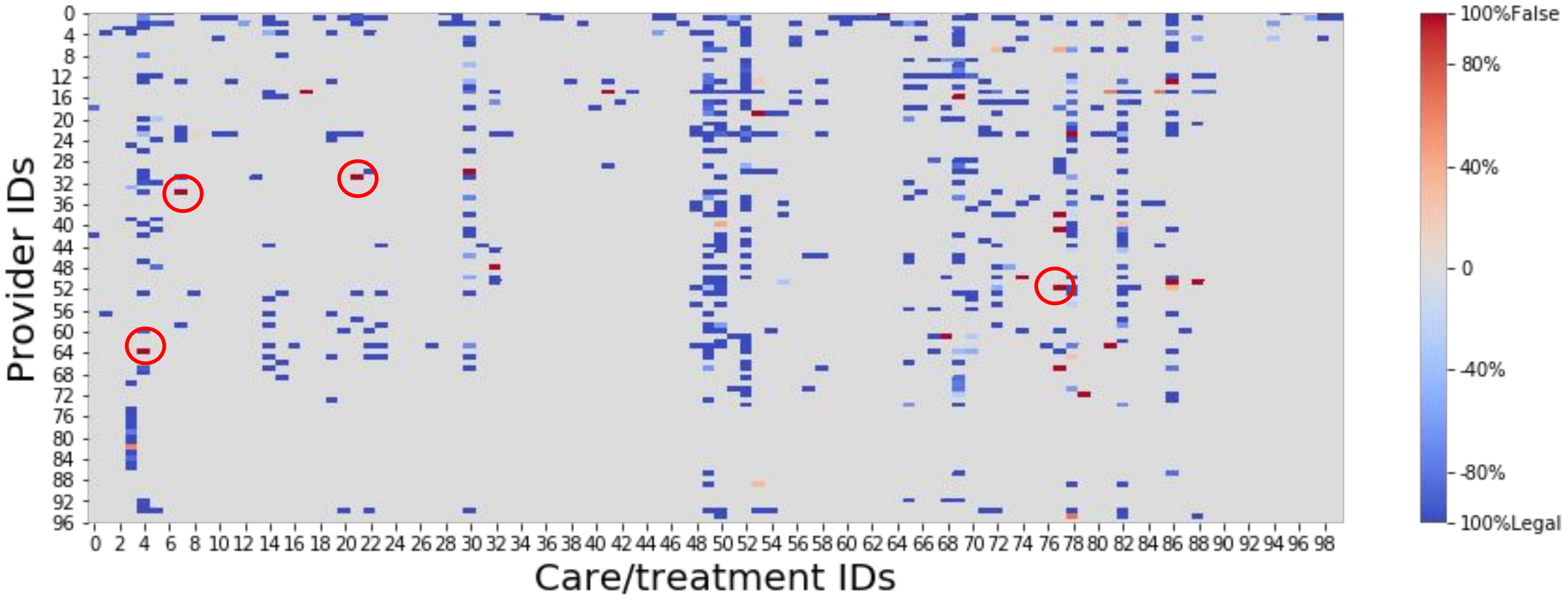
Example of features with distinct distribution patterns in two claim types





0: legal claim, 1: ‘problematic claim’

Example of features with specific value-pairs associated with higher probability of false claims



0: legal claim, 1: 'problematic claim'

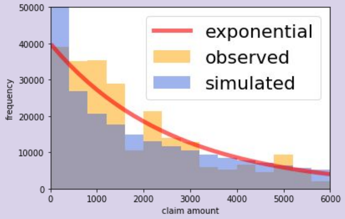
Model implementation: 11 total Features

Patients (11270):
Cum_claim_count
Cum_pay_amount

Provider (185)

Diagnosis (8901)

Claim amount



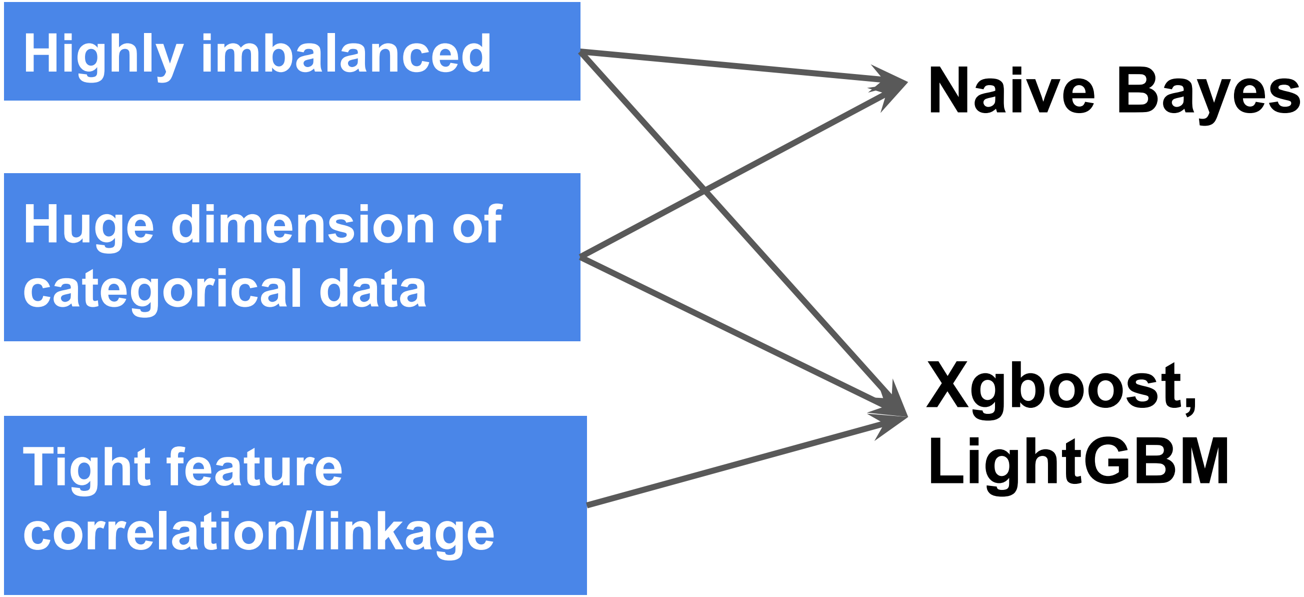
Care (12714):
Quantity
Unit_cost
ID
Tariffs
Created month

Policy (3)

{0}: 694362

{1}: 29386

Model selection:



Highly imbalanced

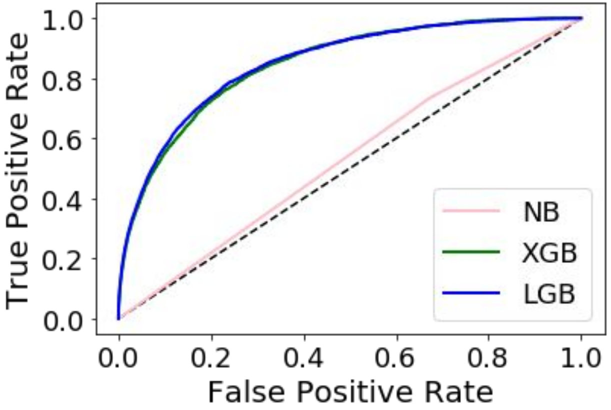
Huge dimension of categorical data

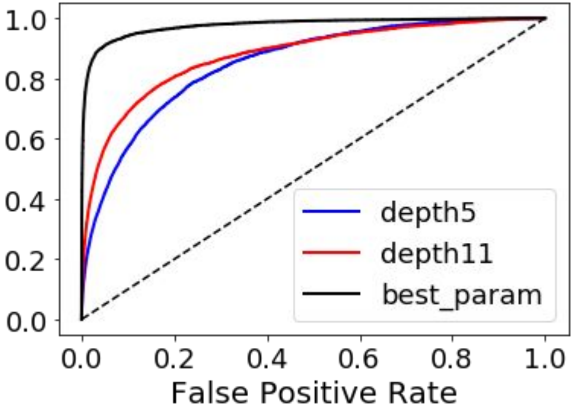
Tight feature correlation/linkage

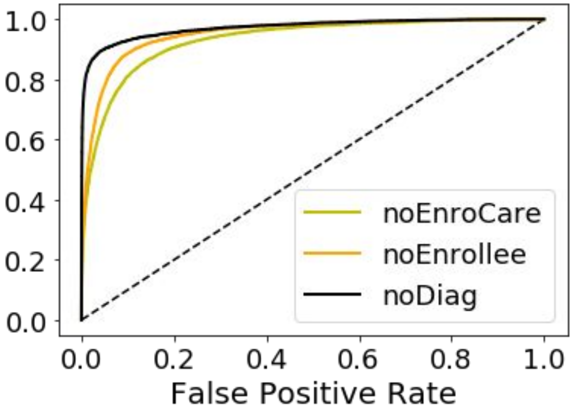
Naive Bayes

Xgboost, LightGBM

Evaluation by testing score







Best model:

recall/sensitivity/True Positive Rate (TPR): 0.893

specificity/True Negative Rate(TNR): 0.964

ROC_AUC score: 0.929

TNR: 0.964	**FPR: 0.036**
FNR: 0.107	**TPR: 0.893**

Without **enrollee** feature, roc_auc drop 10%:
TPR drop 17%, **FPR** increase 2.7%, **Loss ~$138K**

Without **diagnosis** feature, roc_auc drop 2%:
TPR drop 6.7%, **FPR** drop 2.4%, **Loss ~$50K**

Insights

TPR(sensitivity) vs FPR(1 - specificity)

Weight = 25

1% increase in TPR
Save $ 8000
~ 9.4 x $850

1% increase in FPR
Cost $ 850

Value provided:

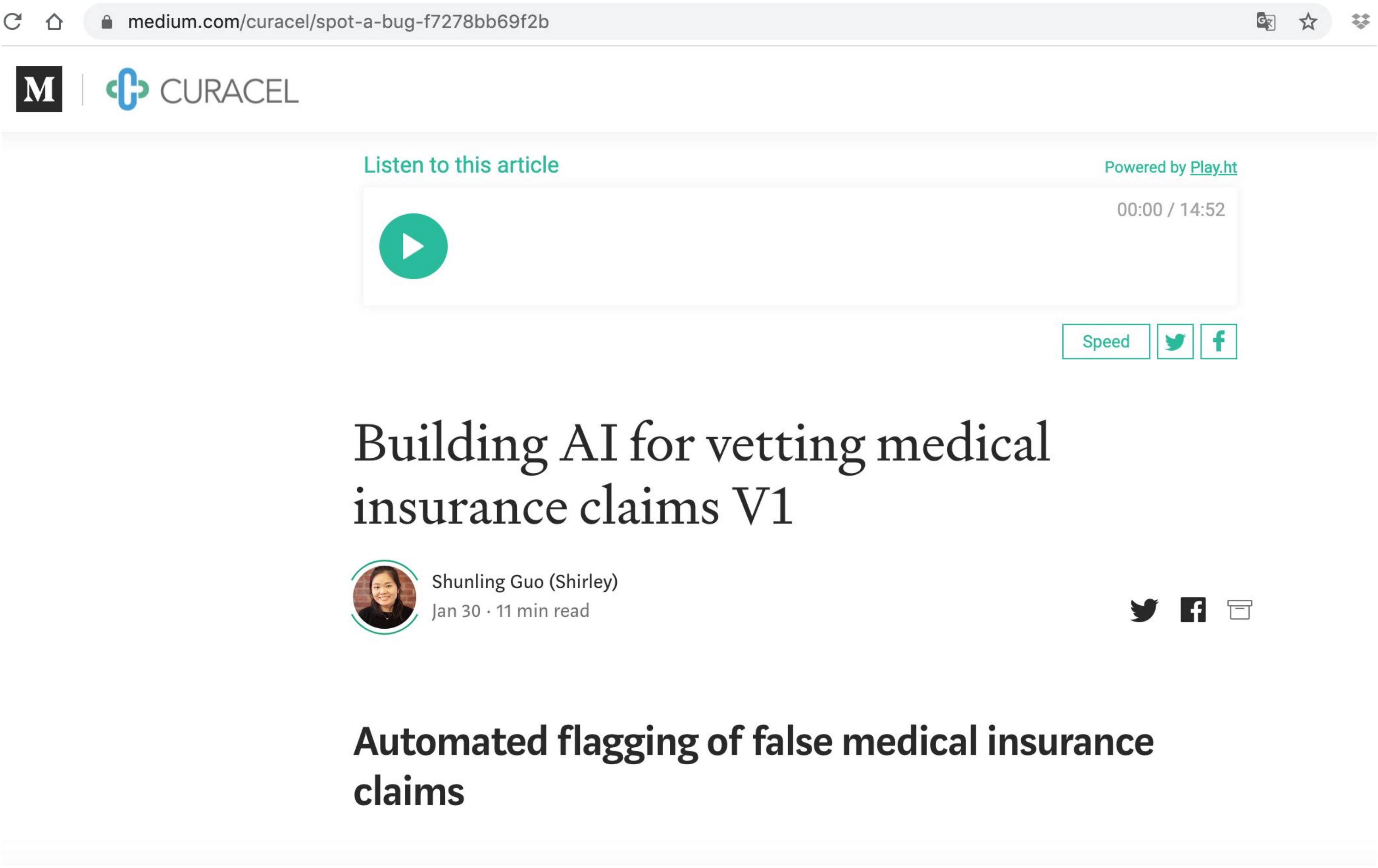
>$80K

40+ days -> instant

Value left to improve:

10.7% FN
$86K

3.6% FP
$3K



Building AI for vetting medical insurance claims V1

Shunling Guo (Shirley)
Jan 30 · 11 min read

Automated flagging of false medical insurance claims

Shunling (Shirley) Guo

INSIGHT

PhD in neuroscience
Chinese Academy of Sciences

Postdoc in Neurophysics
Stanford University

Scientist in Drug Discovery
Confometrx (GPCR structure)

Senior Scientist in Assay Development
Aromyx (Digitize Smell)

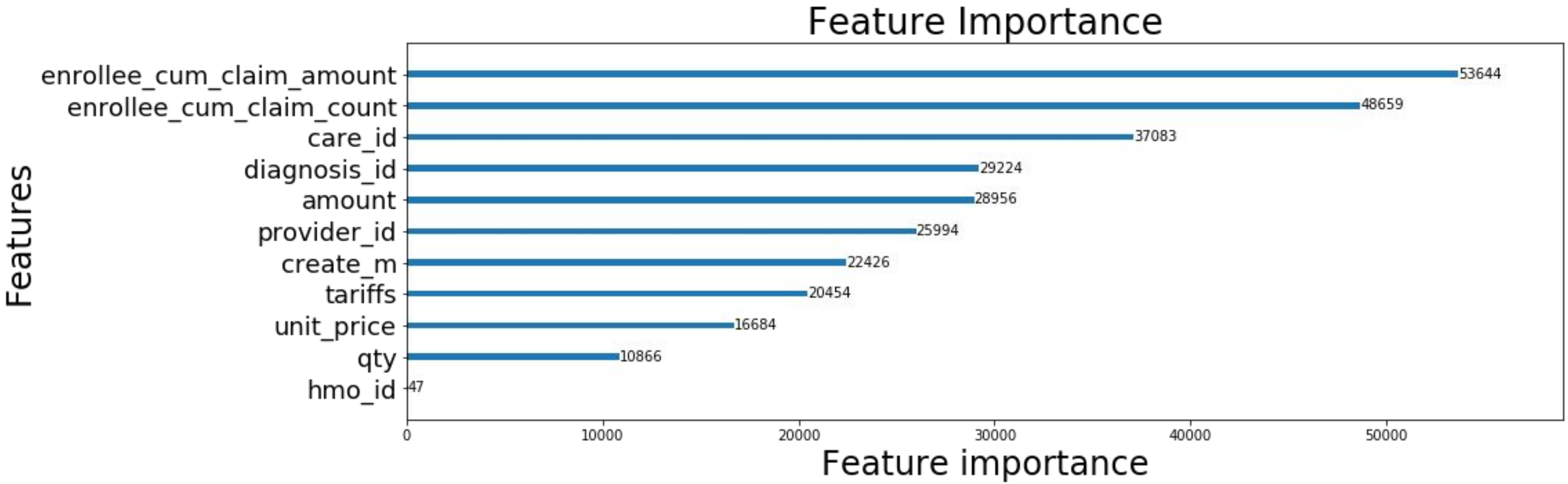
SpringBoard:
Data Science Career Track
Master Level Certificate

Udacity:
Artificial Intelligence
Nanodegree

github.com/Shunling

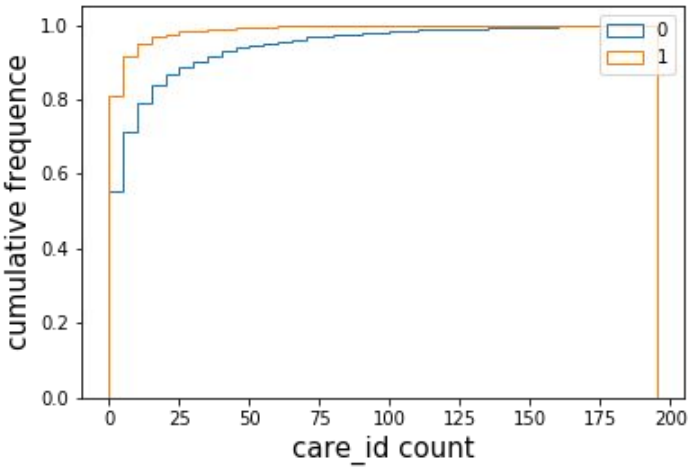
linkedin.com/in/ShunlingGuo

shirley.shunling@gmail.com

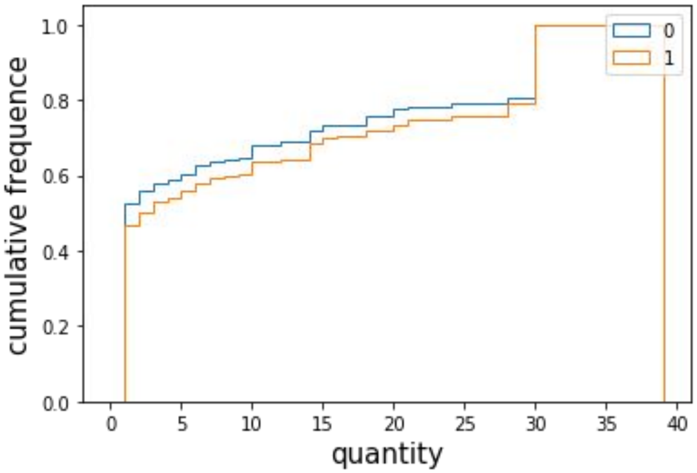


Confusion matrix for testing data:

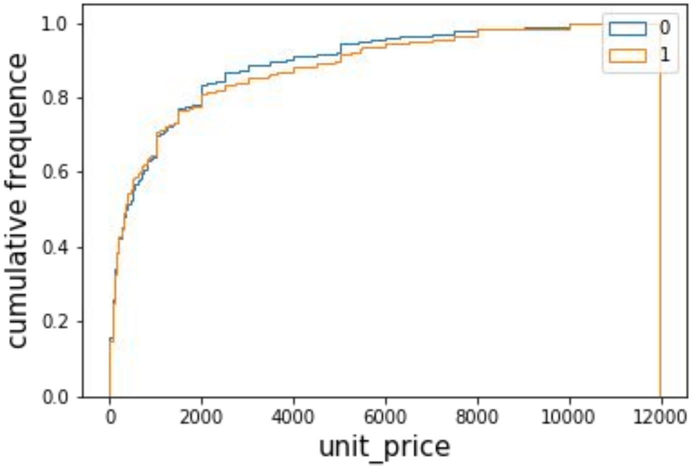
		Predicted	
		0	1
True	0	137582	5142
	1	667	5583



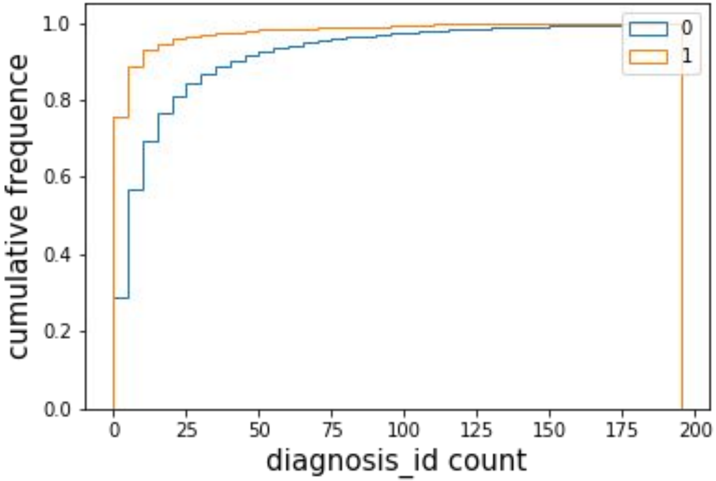
cumulative frequence
0.0
0.2
0.4
0.6
0.8
1.0
0
25
50
75
100
125
150
175
200
care_id count
0
1



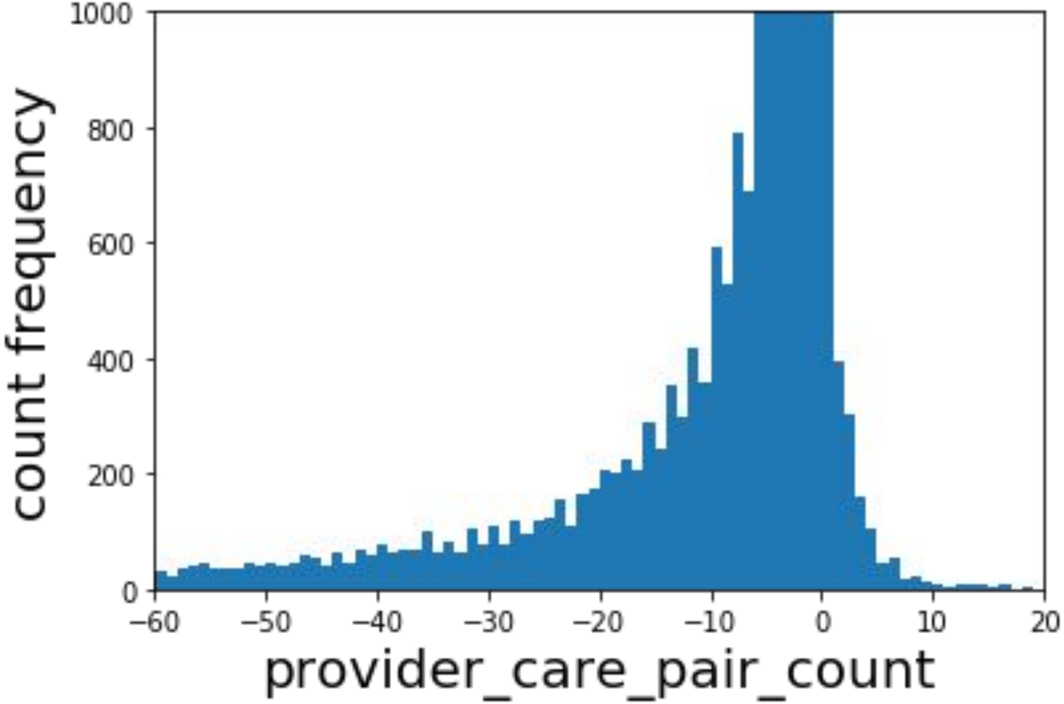
cumulative frequence
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
35
40
quantity
0
1

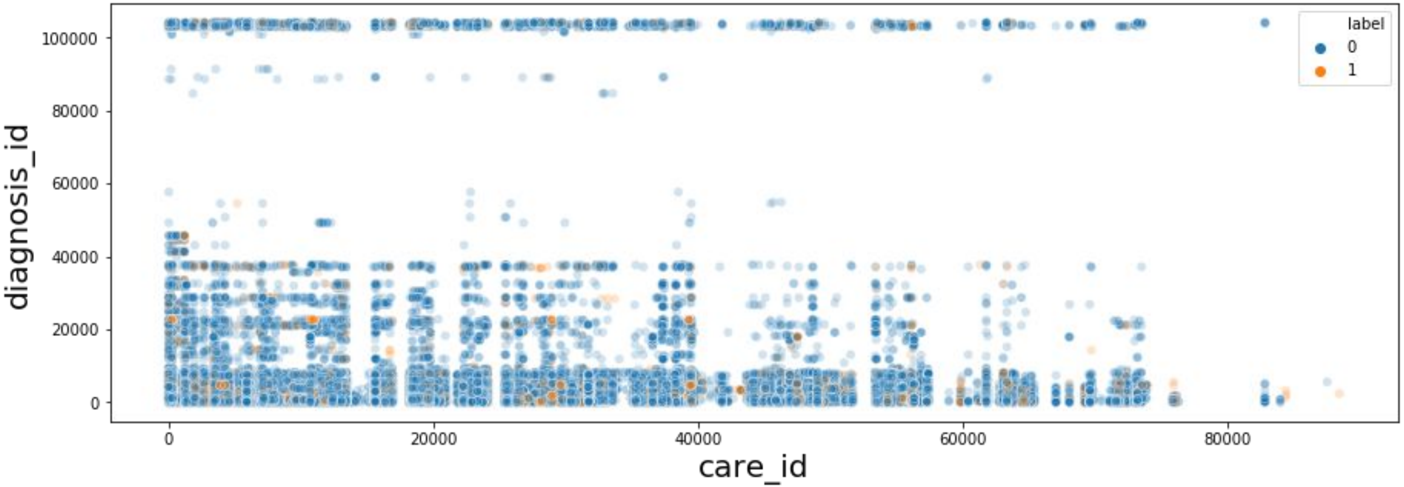


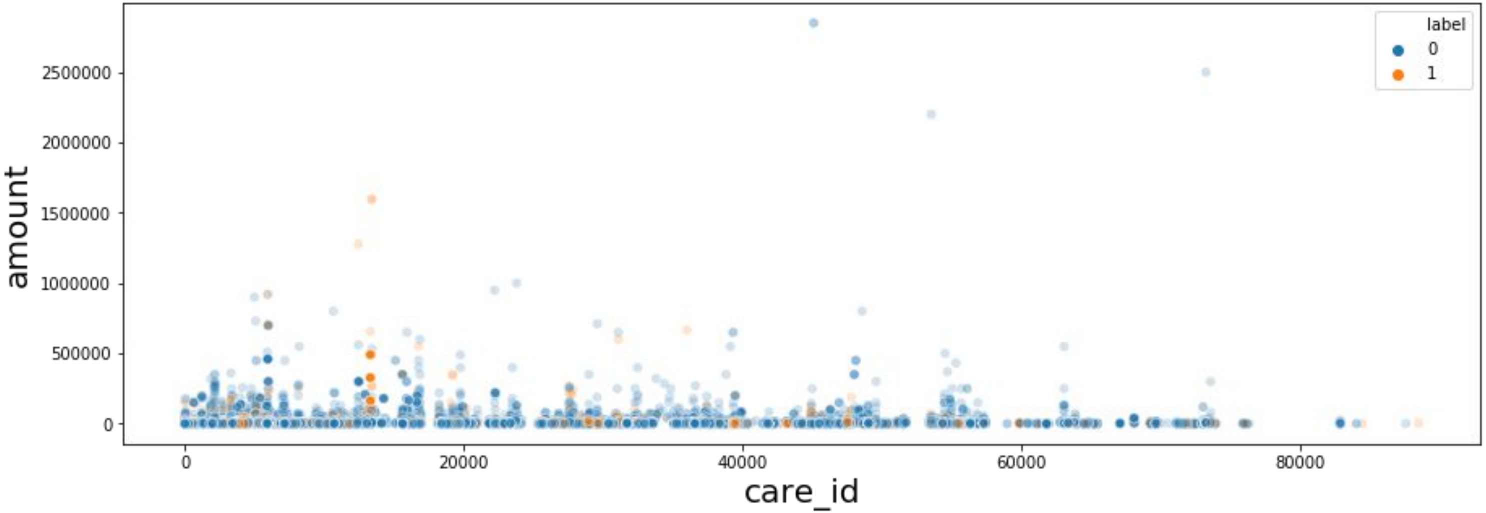
cumulative frequence
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
12000
unit_price
0
1



cumulative frequence
0.0
0.2
0.4
0.6
0.8
1.0
0
25
50
75
100
125
150
175
200
diagnosis_id count
0
1


1000
800
600
400
200
0
count frequency
−60
−50
−40
−30
−20
−10
0
10
20
provider_care_pair_count


diagnosis_id
100000
80000
60000
40000
20000
0
0
20000
40000
60000
80000
care_id
label
● 0
● 1


amount
2500000
2000000
1500000
1000000
500000
0
0
20000
40000
60000
80000
care_id
label
● 0
● 1