データマイニングデーモ

ロリョウ

データセット収集

https://github.com/RoRyou/lupin4

特徴

ラベル

main ▾ lupin4 / lupin4 / data / demo.csv　　Go to file　…

RoRyou Add files via upload　　Latest commit 65ec38c 3 days ago　History

1 contributor

7862 lines (7862 sloc) | 161 KB　　Raw　Blame

Search this file...

	id	bad	wealth	max_unpay_day	score	age	period	education
2	1	0	3		280	26	6	4
3	2	0	4		316	35		4
4	3	0	6		361	29	4	4
5	4	0	4		290	23	6	
6	5	0	7		276	26	6	4
7	6	1	4		263	24	10	
8	7	0	8		257	25	6	4
9	8	0	6		294	24	6	3
10	9	0	2		196	22	6	4
11	10	0	4		300	26	10	1
12	11	0	5		214	25	6	3
13	12	0			0	36	4	4

データ整形、処理

- データの大体様子がわかる

	type	size	missing	unique	mean_or_top1	std_or_top2	min_or_top3	1%_or_top4	10%_or_top5	50%_or_bottom5	75%_or_bottom4	90%_or
id	int64	5502	0.0000	5502	3947.266630	2252.395671	2.0	87.03	820.1	3931.5	5889.25	
bad	int64	5502	0.0000	2	0.073246	0.260564	0.0	0.00	0.0	0.0	0.00	
score	int64	5502	0.0000	265	295.280625	66.243181	0.0	0.00	223.0	303.0	336.00	
age	float64	5502	0.0002	34	27.659880	4.770299	19.0	21.00	23.0	27.0	30.00	
wealth	float64	5502	0.0244	18	4.529806	1.823149	1.0	1.00	3.0	4.0	5.00	
education	float64	5502	0.1427	5	3.319483	1.005660	1.0	1.00	2.0	4.0	4.00	
period	float64	5502	0.1714	5	7.246326	1.982060	4.0	4.00	6.0	6.0	10.00	
max_unpay_day	float64	5502	0.9253	11	185.476886	22.339647	28.0	86.00	171.0	188.0	201.00	

データ処理

- 特徴のIVを計算
- 特徴のPSIを経時的に計算
- 特徴のVIFを計算

Information value

	iv	unique
score	0.758342	265.0
age	0.504588	35.0
wealth	0.275775	19.0
education	0.230553	6.0
max_unpay_day	0.170061	12.0
period	0.073716	6.0

PSI

```
score: 0.0037
age: 0.0017
wealth: 0.0034
education: 0.0006
max_unpay_day: 0.0010
id: 0.0113
period: 0.0035
```

VIF

```
wealth           1.107606
max_unpay_day    1.063695
score            0.875119
age              0.576168
period           1.110511
education        0.991708
dtype: float64
```

特徴フィルタリング

IV値　＞　０.０２
欠測値　＜　０.９５

- 特徴閾値を設定し、フィルタリング

```
11 train_selected, dropped = lupin4.select(train_df.drop(to_drop,axis=1),target = target, empty = 0.95,iv = 0.02, corr
12 print(dropped)
13 print(train_selected.shape)
14 train_selected
```

```
{'empty': array([], dtype=float64), 'iv': array([], dtype=object), 'corr': array([], dtype=object)}
(5502, 7)
```

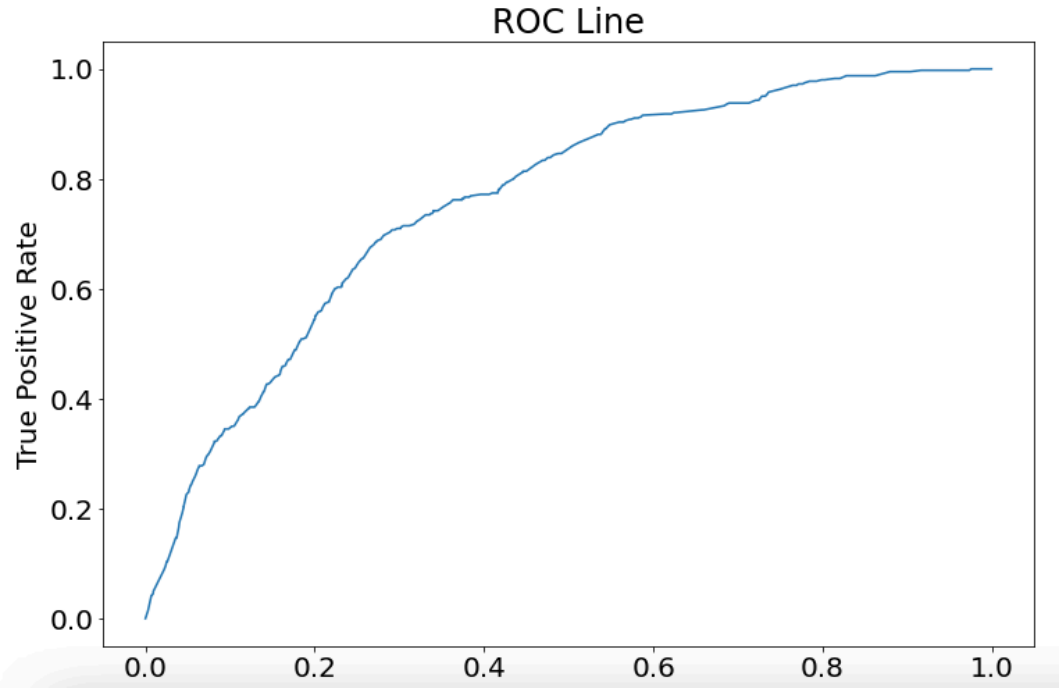
	bad	wealth	max_unpay_day	score	age	period	education
4168	0	4.0	NaN	288	23.0	6.0	4.0
605	0	4.0	NaN	216	32.0	6.0	4.0
3018	0	5.0	NaN	250	23.0	6.0	2.0
4586	0	7.0	171.0	413	31.0	NaN	2.0
1468	0	5.0	NaN	204	29.0	6.0	2.0
...	...	...	...	...	...	...	...
5226	0	4.0	171.0	346	23.0	NaN	3.0
5390	0	5.0	NaN	207	32.0	NaN	3.0
860	0	6.0	NaN	356	42.0	4.0	3.0
7603	0	3.0	NaN	323	34.0	NaN	3.0
7270	0	4.0	NaN	378	24.0	10.0	4.0

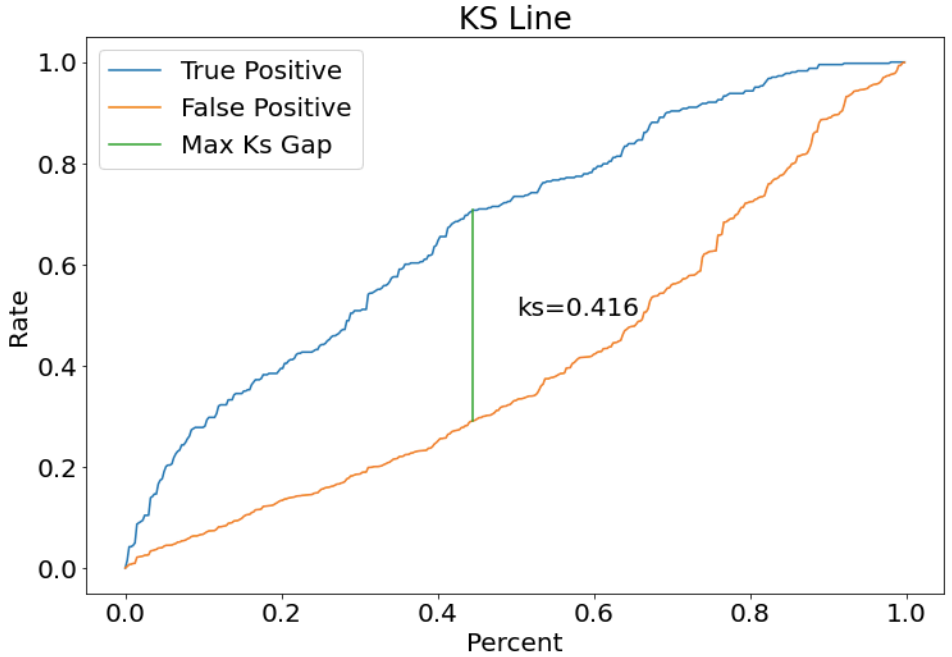
5502 rows × 7 columns

モデル効果可視化

```
KS: 0.4160
AUC: 0.7602
```

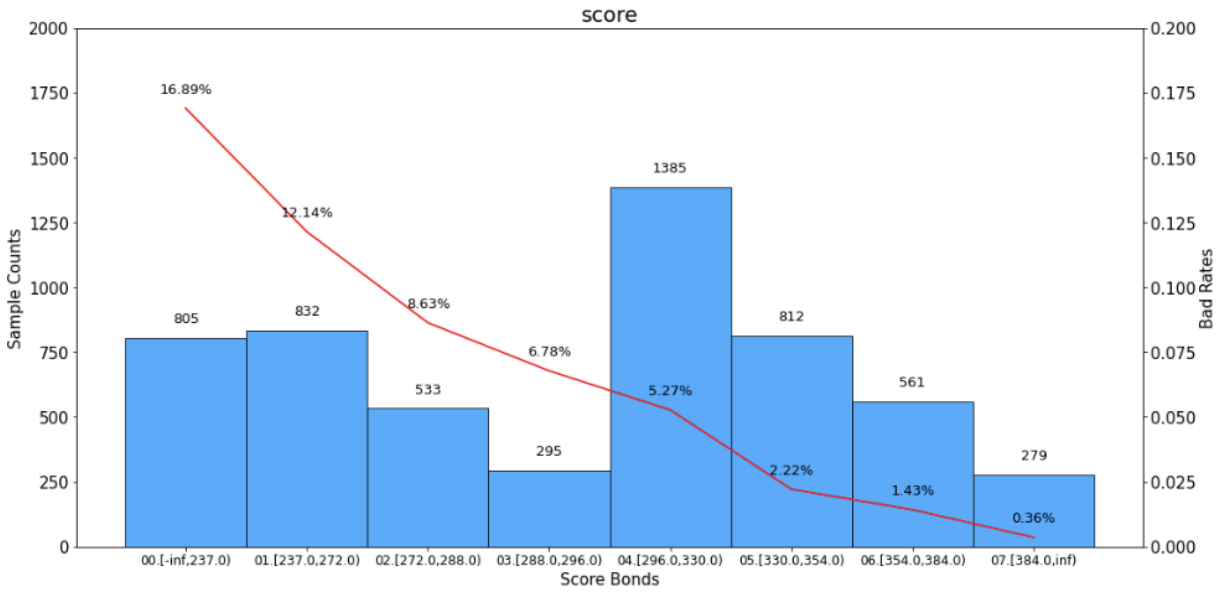
```
findfont: Font family ['sans-serif'] not found. Falling back to DejaVu Sans.
findfont: Generic family 'sans-serif' not found because none of the following families wer
```

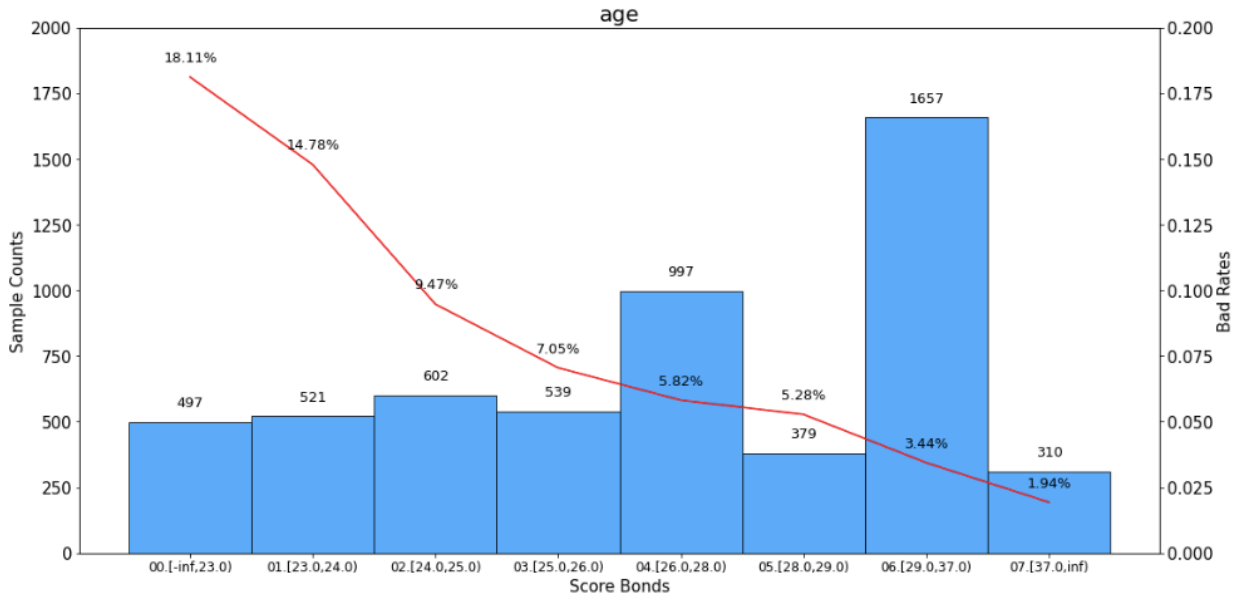




特徴ビニングと単変量解析

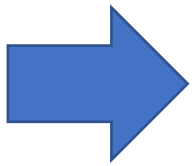
```
{'wealth': [3.0, 4.0, 5.0, 7.0],
 'max_unpay_day': [171.0],
 'score': [237.0, 272.0, 288.0, 296.0, 330.0, 354.0, 384.0],
 'age': [23.0, 24.0, 25.0, 26.0, 28.0, 29.0, 37.0],
 'period': [6.0, 10.0],
 'education': [3.0, 4.0]}
```





スコア転換

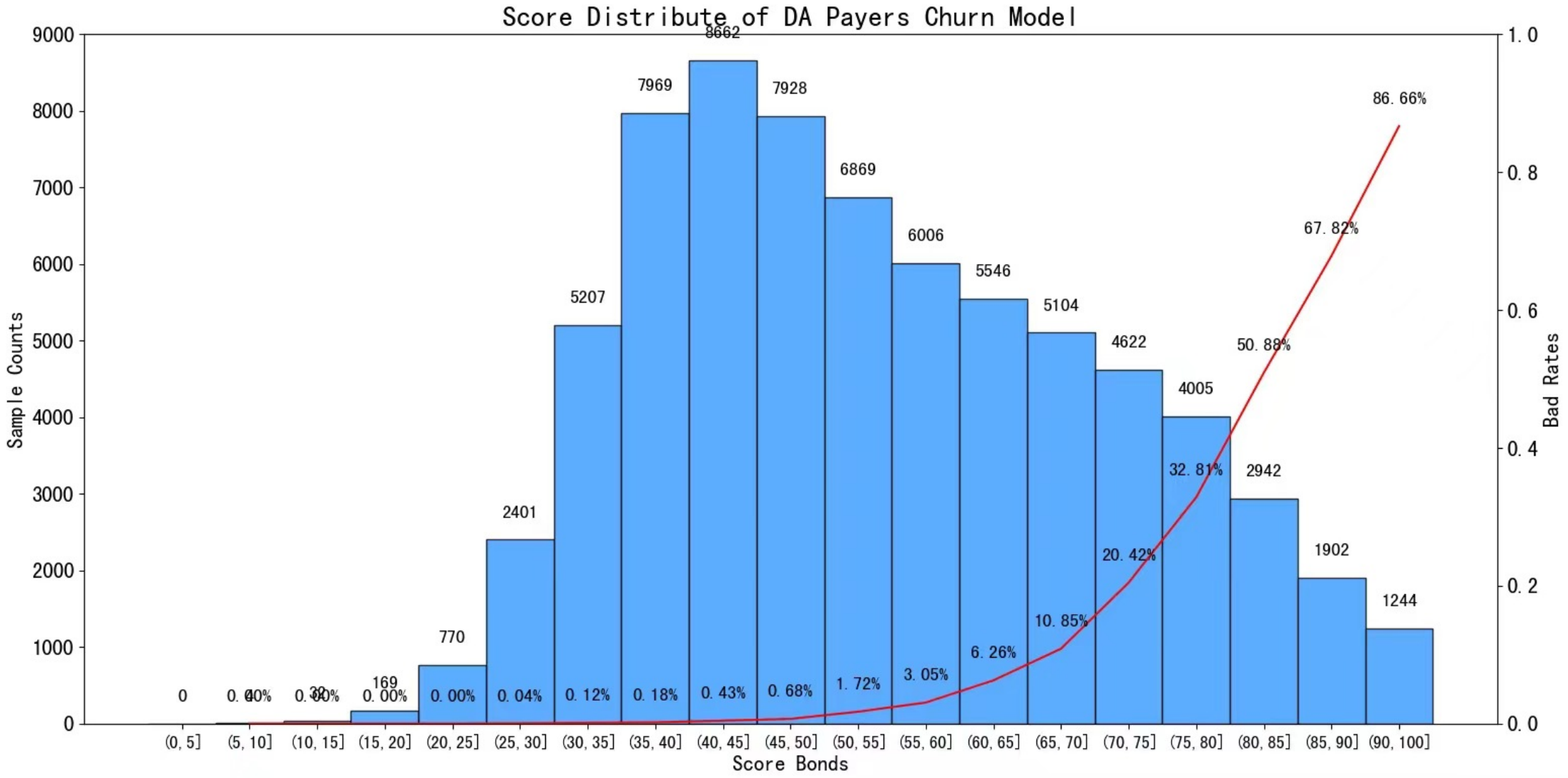
```
{'intercept': {'[-inf,inf)': 509.19},
 'wealth': {'[-inf,3.0)': -18.75,
  '[3.0,4.0)': -1.45,
  '[4.0,5.0)': 4.07,
  '[5.0,7.0)': 4.92,
  '[7.0,inf)': 11.37},
 'max_unpay_day': {'[-inf,171.0)': 2.64, '[171.0,inf)': -20.45},
 'score': {'[-inf,237.0)': -37.23,
  '[237.0,272.0)': -22.01,
  '[272.0,288.0)': -7.02,
  '[288.0,296.0)': 3.28,
  '[296.0,330.0)': 13.83,
  '[330.0,354.0)': 49.22,
  '[354.0,384.0)': 66.92,
  '[384.0,inf)': 121.77},
 'age': {'[-inf,23.0)': -39.69,
  '[23.0,24.0)': -30.31,
  '[24.0,25.0)': -10.81,
  '[25.0,26.0)': 1.59,
  '[26.0,28.0)': 9.51,
  '[28.0,29.0)': 13.49,
  '[29.0,37.0)': 30.74,
  '[37.0,inf)': 53.52}}
```



	bad	wealth	max_unpay_day	score	age
4168	0	-0.219698	-0.132726	-0.083176	0.785844
605	0	-0.219698	-0.132726	0.944734	-0.796844
3018	0	-0.265803	-0.132726	0.558570	0.785844
4586	0	-0.614215	1.026204	-3.089758	-0.796844
1468	0	-0.265803	-0.132726	0.944734	-0.796844
...	...	...	...	...	...
5226	0	-0.219698	1.026204	-1.248849	0.785844
5390	0	-0.265803	-0.132726	0.944734	-0.796844
860	0	-0.265803	-0.132726	-1.698053	-1.387405
7603	0	0.078071	-0.132726	-0.350985	-0.796844
7270	0	-0.219698	-0.132726	-1.698053	0.280129

5502 rows × 5 columns

スコアカードモデル可視化（客の角度）



因果分析

1. 側面からユーザーを評価
2. ABtestで、因果関係のある特徴がわかる