1. load and check data

In [1]:

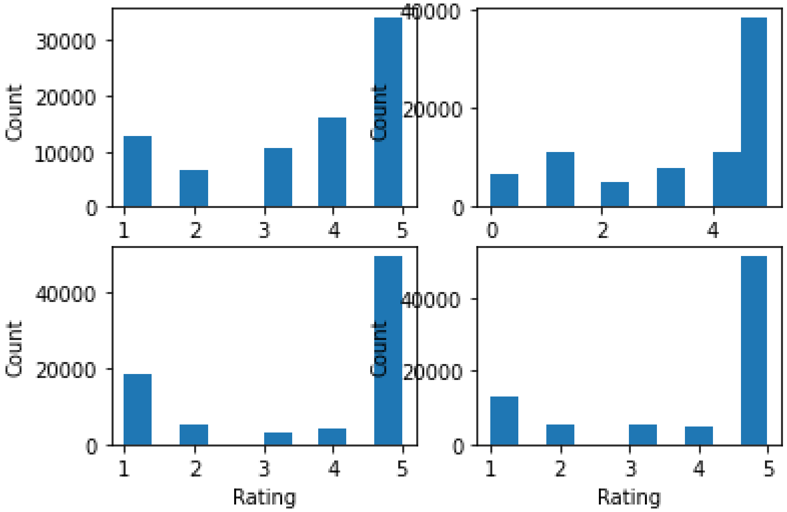
```
import pandas as pd


#load our data
# drRatings = pd.read_excel('./data/OBGYN_new_train_80000.xlsx',nrows=1000)
drRatings = pd.read_excel('./data/OBGYN_new_train_80000.xlsx')


#shuffle the data so that they are in random sequence
drRatings = drRatings.sample(frac=1)

import matplotlib.pyplot as plt
%matplotlib inline
```

In [2]:

```
#check the distribution of the four ratings.
plt.subplot(2, 2, 1)
plt.hist(drRatings['punctuality'])
plt.ylabel('Count')
plt.xlabel('Rating')
plt.subplot(2, 2, 2)
plt.hist(drRatings['staff'])
plt.ylabel('Count')
plt.xlabel('Rating')
plt.subplot(2, 2, 3)
plt.hist(drRatings['helpfulness'])
plt.ylabel('Count')
plt.xlabel('Rating')
plt.subplot(2, 2, 4)
plt.hist(drRatings['knowledge'])
plt.ylabel('Count')
plt.xlabel('Rating')
```

Out[2]:

```
Text(0.5, 0, 'Rating')
```



In [ ]:

```
drRatings.head(5)
```

Out[3]:

	reviewID	doctorID	doctorName	specialty	numReviews	city	state	doctorHomepage	averageRating	staff	punctuality
29734	29735	3298164	Dr. Okey Okoli	Gynecologist (OBGYN)	8	San Antonio	TX	/doctor-ratings/3298164/Dr-Okey-Okoli-San%2BAn...	4.50	4	4
75297	75298	162821	Dr. Michael J. Straker	Gynecologist (OBGYN)	18	Nutley	NJ	/doctor-ratings/162821/Dr-Michael%2BJ.-Straker...	5.00	5	5
5999	6000	3366471	Dr. Catalin G. Marinescu	Gynecologist (OBGYN)	25	Newport Beach	CA	/doctor-ratings/3366471/Dr-Catalin%2BG.-Marine...	4.50	5	3
48528	48529	50278	Dr. Craig L. Bissinger	Gynecologist (OBGYN)	32	Parsippany	NJ	/doctor-ratings/50278/Dr-Craig%2BL.-Bissinger-...	4.50	5	4
8338	8339	544050	Dr. Jung K. Choe	Gynecologist (OBGYN)	6	Marlton	NJ	/doctor-ratings/544050/Dr-Jung%2BK.-Choe-Marlt...	4.25	5	2

2. Prepare the input and output

In [3]:

```
#prepare input and output features
drRatings['highPunctuality'] = (drRatings['punctuality']>4).astype(int)
drRatings.head(5)
```

Out[3]:

	reviewID	doctorID	doctorName	specialty	numReviews	city	state	doctorHomepage	averageRating	staff	pu
54254	54255	887117	Dr. Xiorong Dai	Gynecologist (OBGYN)	3	West Palm Beach	FL	/doctor-ratings/887117/Dr-Xiorong-Dai-West%2BP...	2.00	5	
70949	70950	33769	Dr. Lisa Yang	Gynecologist (OBGYN)	13	Cincinnati	OH	/doctor-ratings/33769/Dr-Lisa-Yang-Cincinnati-...	4.75	4	
29367	29368	309188	Dr. Ernest H. Carlton	Gynecologist (OBGYN)	6	Macon	GA	/doctor-ratings/309188/Dr-Ernest%2BH.-Carlton-...	5.00	5	
71006	71007	113844	Dr. Michael J. Derosa	Gynecologist (OBGYN)	23	O%26apos%3BFallon	MO	/doctor-ratings/113844/Dr-Michael%2BJ.-Derosa-...	4.00	5	
57782	57783	28644	Dr. Dawn M. Hasson	Gynecologist (OBGYN)	2	Reading	PA	/doctor-ratings/28644/Dr-Dawn%2BM.-Hasson-Read...	4.75	4	

In [ ]:
```
drRatings['highPunctuality'].describe()
```

Out[5]:
```
count    80000.000000
mean         0.426500
std          0.494571
min          0.000000
25%          0.000000
50%          0.000000
75%          1.000000
max          1.000000
Name: highPunctuality, dtype: float64
```

In [4]:
```
temp = pd.get_dummies(drRatings['state'])
drRatings = pd.concat([drRatings,temp],axis=1)
drRatings.head()
```

Out[4]:

	reviewID	doctorID	doctorName	specialty	numReviews	city	state	doctorHomepage	averageRating	staff	...
54254	54255	887117	Dr. Xiorong Dai	Gynecologist (OBGYN)	3	West Palm Beach	FL	/doctor-ratings/887117/Dr-Xiorong-Dai-West%2BP...	2.00	5	...
70949	70950	33769	Dr. Lisa Yang	Gynecologist (OBGYN)	13	Cincinnati	OH	/doctor-ratings/33769/Dr-Lisa-Yang-Cincinnati-...	4.75	4	...
29367	29368	309188	Dr. Ernest H. Carlton	Gynecologist (OBGYN)	6	Macon	GA	/doctor-ratings/309188/Dr-Ernest%2BH.-Carlton-...	5.00	5	...
71006	71007	113844	Dr. Michael J. Derosa	Gynecologist (OBGYN)	23	O%26apos%3BFallon	MO	/doctor-ratings/113844/Dr-Michael%2BJ.-Derosa-...	4.00	5	...
57782	57783	28644	Dr. Dawn M. Hasson	Gynecologist (OBGYN)	2	Reading	PA	/doctor-ratings/28644/Dr-Dawn%2BM.-Hasson-Read...	4.75	4	...

5 rows × 70 columns

In [5]:

```
drRatings['postedTime']=pd.to_datetime(drRatings['postedTime'])
drRatings['year']=drRatings['postedTime'].dt.year
drRatings['hour']=drRatings['postedTime'].dt.hour
drRatings.head()
```

Out[5]:

	reviewID	doctorID	doctorName	specialty	numReviews	city	state	doctorHomepage	averageRating	staff	...
54254	54255	887117	Dr. Xiorong Dai	Gynecologist (OBGYN)	3	West Palm Beach	FL	/doctor-ratings/887117/Dr-Xiorong-Dai-West%2BP...	2.00	5	...
70949	70950	33769	Dr. Lisa Yang	Gynecologist (OBGYN)	13	Cincinnati	OH	/doctor-ratings/33769/Dr-Lisa-Yang-Cincinnati-...	4.75	4	...
29367	29368	309188	Dr. Ernest H. Carlton	Gynecologist (OBGYN)	6	Macon	GA	/doctor-ratings/309188/Dr-Ernest%2BH.-Carlton-...	5.00	5	...
71006	71007	113844	Dr. Michael J. Derosa	Gynecologist (OBGYN)	23	O%26apos%3BFallon	MO	/doctor-ratings/113844/Dr-Michael%2BJ.-Derosa-...	4.00	5	...
57782	57783	28644	Dr. Dawn M. Hasson	Gynecologist (OBGYN)	2	Reading	PA	/doctor-ratings/28644/Dr-Dawn%2BM.-Hasson-Read...	4.75	4	...

5 rows × 72 columns

3. Train Logit Model (model1)

In [6]:

```python
#decide x and y
xcols = ['AK', 'AL', 'AR', 'AZ', 'CA',
       'CO', 'CT', 'DC', 'DE', 'FL', 'GA', 'HI', 'IA', 'ID', 'IL', 'IN', 'KS',
       'KY', 'LA', 'MA', 'MD', 'ME', 'MI', 'MN', 'MO', 'MS', 'NC', 'ND', 'NE',
       'NJ', 'NM', 'NV', 'NY', 'OH', 'OK', 'OR', 'PA', 'PR', 'RI', 'SC', 'SD',
       'TN', 'TX', 'UT', 'VA', 'WA', 'WI', 'WV', 'WY', 'year', 'hour','numReviews']

ycol = 'highPunctuality'

x = drRatings[xcols]
print(x.shape)

import statsmodels.api as sm
x = sm.add_constant(x)
y = drRatings[ycol]
print(x.shape,y.shape)
```

```
(80000, 52)

/usr/local/lib/python3.7/dist-packages/statsmodels/tools/_testing.py:19: FutureWarning: pandas.util.te
sting is deprecated. Use the functions in the public API at pandas.testing instead.
  import pandas.util.testing as tm

(80000, 53) (80000,)

/usr/local/lib/python3.7/dist-packages/statsmodels/tsa/tsatools.py:117: FutureWarning: In a future ver
sion of pandas all arguments of concat except for the argument 'objs' will be keyword-only
  x = pd.concat(x[::order], 1)
```

In [7]:

```
import statsmodels.api as sm
logit_model1 = sm.Logit(y, x)
logit_result = logit_model1.fit()
logit_result.summary2()
```

```
Optimization terminated successfully.
         Current function value: 0.674984
         Iterations 5
```

Out[7]:

Model:	Logit	Pseudo R-squared:	0.011
Dependent Variable:	highPunctuality	AIC:	108103.4113
Date:	2022-03-16 02:29	BIC:	108595.7697
No. Observations:	80000	Log-Likelihood:	-53999.
Df Model:	52	LL-Null:	-54584.
Df Residuals:	79947	LLR p-value:	5.1110e-211
Converged:	1.0000	Scale:	1.0000
No. Iterations:	5.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	-113.4387	5.5328	-20.5029	0.0000	-124.2828	-102.5946
AK	0.1092	0.1998	0.5467	0.5846	-0.2823	0.5007
AL	-0.0521	0.1146	-0.4546	0.6494	-0.2767	0.1725
AR	-0.0604	0.1281	-0.4712	0.6375	-0.3114	0.1907
AZ	-0.5092	0.1111	-4.5816	0.0000	-0.7270	-0.2914
CA	-0.5245	0.1043	-5.0282	0.0000	-0.7289	-0.3200
CO	-0.1732	0.1221	-1.4188	0.1560	-0.4125	0.0661
CT	-0.1520	0.1196	-1.2710	0.2037	-0.3864	0.0824
DC	-0.4453	0.1441	-3.0899	0.0020	-0.7278	-0.1628
DE	-0.6679	0.1613	-4.1402	0.0000	-0.9841	-0.3517
FL	-0.3097	0.1046	-2.9610	0.0031	-0.5147	-0.1047
GA	-0.2666	0.1100	-2.4247	0.0153	-0.4821	-0.0511

HI	-0.3483	0.1710	-2.0373	0.0416	-0.6835	-0.0132
IA	-0.1494	0.1477	-1.0117	0.3117	-0.4388	0.1400
ID	-0.2535	0.1502	-1.6878	0.0914	-0.5479	0.0409
IL	-0.3843	0.1083	-3.5475	0.0004	-0.5967	-0.1720
IN	-0.1337	0.1179	-1.1339	0.2568	-0.3648	0.0974
KS	-0.2142	0.1305	-1.6414	0.1007	-0.4699	0.0416
KY	-0.2517	0.1257	-2.0022	0.0453	-0.4981	-0.0053
LA	-0.1214	0.1183	-1.0264	0.3047	-0.3532	0.1104
MA	-0.4217	0.1151	-3.6653	0.0002	-0.6472	-0.1962
MD	-0.4424	0.1108	-3.9944	0.0001	-0.6595	-0.2253
ME	-0.0373	0.1960	-0.1904	0.8490	-0.4215	0.3469
MI	-0.3037	0.1095	-2.7737	0.0055	-0.5182	-0.0891
MN	-0.0700	0.1263	-0.5546	0.5792	-0.3175	0.1775
MO	-0.0367	0.1124	-0.3262	0.7443	-0.2570	0.1837
MS	0.1857	0.1434	1.2946	0.1955	-0.0954	0.4667
NC	-0.0562	0.1138	-0.4940	0.6213	-0.2792	0.1668
ND	-0.5078	0.2291	-2.2168	0.0266	-0.9567	-0.0588
NE	-0.0021	0.1424	-0.0146	0.9884	-0.2812	0.2770
NJ	-0.4000	0.1053	-3.7988	0.0001	-0.6064	-0.1936
NM	-0.4078	0.1591	-2.5628	0.0104	-0.7196	-0.0959
NV	-0.5368	0.1193	-4.4992	0.0000	-0.7706	-0.3029
NY	-0.5902	0.1044	-5.6527	0.0000	-0.7949	-0.3856
OH	-0.3712	0.1089	-3.4088	0.0007	-0.5846	-0.1578
OK	-0.3412	0.1149	-2.9701	0.0030	-0.5664	-0.1160
OR	-0.1915	0.1316	-1.4550	0.1457	-0.4494	0.0665
PA	-0.4238	0.1101	-3.8508	0.0001	-0.6395	-0.2081
PR	0.2510	0.2017	1.2448	0.2132	-0.1442	0.6463

RI	-0.0427	0.1581	-0.2703	0.7870	-0.3526	0.2671
SC	-0.2346	0.1218	-1.9263	0.0541	-0.4733	0.0041
SD	-0.0262	0.1899	-0.1378	0.8904	-0.3985	0.3461
TN	-0.0705	0.1124	-0.6267	0.5309	-0.2908	0.1499
TX	-0.3709	0.1042	-3.5596	0.0004	-0.5751	-0.1667
UT	-0.3752	0.1161	-3.2321	0.0012	-0.6027	-0.1477
VA	-0.4212	0.1083	-3.8886	0.0001	-0.6334	-0.2089
WA	-0.2617	0.1180	-2.2174	0.0266	-0.4931	-0.0304
WI	-0.0953	0.1218	-0.7826	0.4339	-0.3341	0.1434
WV	-0.2936	0.1401	-2.0949	0.0362	-0.5682	-0.0189
WY	-0.1841	0.2396	-0.7684	0.4423	-0.6537	0.2855
year	0.0565	0.0028	20.5362	0.0000	0.0511	0.0619
hour	-0.0016	0.0013	-1.2812	0.2001	-0.0041	0.0008
numReviews	-0.0060	0.0006	-9.8584	0.0000	-0.0071	-0.0048

In [9]:

```
#predict highPunctuality using the trained logit model
drRatings['highPunctuality_predictLogit'] = (logit_result.predict(x) >= 0.5).astype(int)
drRatings[['highPunctuality_predictLogit','highPunctuality']]
```

Out[9]:

	highPunctuality_predictLogit	highPunctuality
54254	0	0
70949	0	1
29367	0	1
71006	0	0
57782	0	1
...	...	...
64936	0	0
16701	0	1
24697	1	0
52354	0	0
14441	0	0

80000 rows × 2 columns

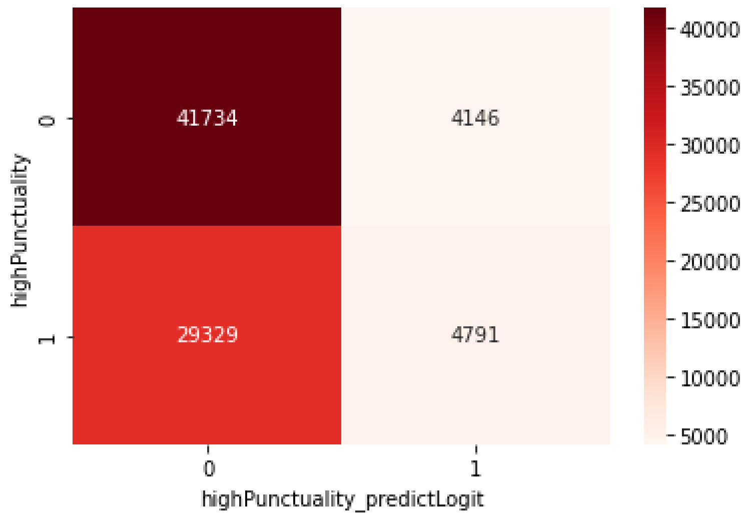
In [10]:

```
import sklearn.metrics as metrics
acc=metrics.accuracy_score(y_true=drRatings['highPunctuality'],y_pred=drRatings['highPunctuality_predict
print(acc)
confusion=metrics.confusion_matrix(y_true=drRatings['highPunctuality'],y_pred=drRatings['highPunctuality
print(confusion)
import seaborn as sn
sn.heatmap(confusion, annot=True, cmap='Reds', fmt='d')
plt.xlabel("highPunctuality_predictLogit")
plt.ylabel("highPunctuality")
```

```
0.5815625
[[41734  4146]
 [29329  4791]]
```

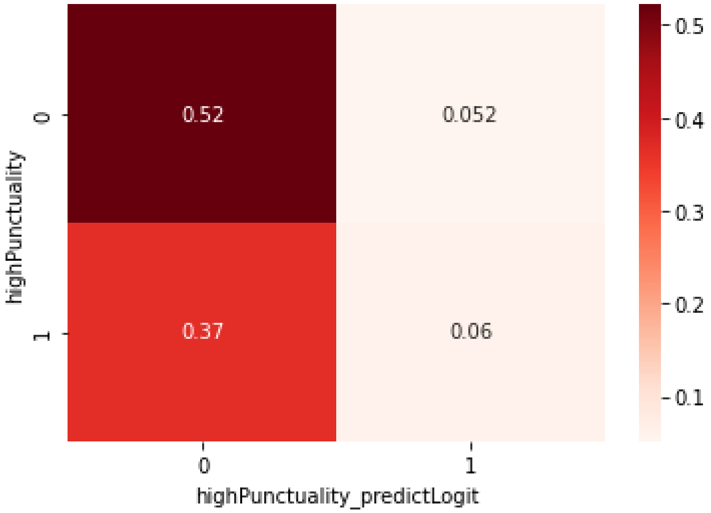
Out[10]:

```
Text(33.0, 0.5, 'highPunctuality')
```



In [11]:
```
import numpy as np
sn.heatmap(confusion/np.sum(confusion), annot=True, cmap='Reds', fmt='.2')
plt.xlabel("highPunctuality_predictLogit")
plt.ylabel("highPunctuality")
```

Out[11]:
```
Text(33.0, 0.5, 'highPunctuality')
```



Try a different model - Probit

In [12]:
```
probit_model2 = sm.Probit(y, x)
probit_result = probit_model2.fit()
probit_result.summary2()
```

```
Optimization terminated successfully.
         Current function value: 0.674993
         Iterations 5
```

Out[12]:

Model:	Probit	Pseudo R-squared:	0.011
Dependent Variable:	highPunctuality	AIC:	108104.8315
Date:	2022-03-16 02:31	BIC:	108597.1899
No. Observations:	80000	Log-Likelihood:	-53999.
Df Model:	52	LL-Null:	-54584.
Df Residuals:	79947	LLR p-value:	1.0087e-210
Converged:	1.0000	Scale:	1.0000
No. Iterations:	5.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	-70.5608	3.4290	-20.5778	0.0000	-77.2815	-63.8401
AK	0.0680	0.1247	0.5453	0.5855	-0.1764	0.3124
AL	-0.0332	0.0716	-0.4636	0.6429	-0.1736	0.1072
AR	-0.0382	0.0801	-0.4763	0.6338	-0.1952	0.1188
AZ	-0.3186	0.0693	-4.5942	0.0000	-0.4545	-0.1827
CA	-0.3278	0.0651	-5.0312	0.0000	-0.4555	-0.2001
CO	-0.1086	0.0763	-1.4225	0.1549	-0.2582	0.0410
CT	-0.0954	0.0747	-1.2762	0.2019	-0.2419	0.0511
DC	-0.2791	0.0896	-3.1135	0.0018	-0.4547	-0.1034
DE	-0.4161	0.0997	-4.1739	0.0000	-0.6115	-0.2207
FL	-0.1936	0.0654	-2.9625	0.0031	-0.3217	-0.0655
GA	-0.1669	0.0687	-2.4297	0.0151	-0.3016	-0.0323
HI	-0.2174	0.1067	-2.0377	0.0416	-0.4264	-0.0083

IA	-0.0938	0.0924	-1.0155	0.3099	-0.2749	0.0873
ID	-0.1587	0.0938	-1.6918	0.0907	-0.3426	0.0252
IL	-0.2404	0.0677	-3.5527	0.0004	-0.3730	-0.1078
IN	-0.0838	0.0737	-1.1367	0.2557	-0.2282	0.0607
KS	-0.1340	0.0815	-1.6444	0.1001	-0.2938	0.0257
KY	-0.1575	0.0786	-2.0045	0.0450	-0.3115	-0.0035
LA	-0.0761	0.0740	-1.0294	0.3033	-0.2211	0.0688
MA	-0.2634	0.0718	-3.6679	0.0002	-0.4042	-0.1227
MD	-0.2768	0.0691	-4.0030	0.0001	-0.4123	-0.1413
ME	-0.0236	0.1226	-0.1924	0.8475	-0.2639	0.2167
MI	-0.1903	0.0684	-2.7815	0.0054	-0.3243	-0.0562
MN	-0.0438	0.0789	-0.5550	0.5789	-0.1985	0.1109
MO	-0.0236	0.0703	-0.3360	0.7369	-0.1614	0.1141
MS	0.1166	0.0896	1.3019	0.1929	-0.0589	0.2922
NC	-0.0354	0.0711	-0.4974	0.6189	-0.1747	0.1040
ND	-0.3169	0.1424	-2.2246	0.0261	-0.5960	-0.0377
NE	-0.0019	0.0890	-0.0217	0.9827	-0.1764	0.1726
NJ	-0.2514	0.0658	-3.8223	0.0001	-0.3803	-0.1225
NM	-0.2547	0.0992	-2.5664	0.0103	-0.4491	-0.0602
NV	-0.3353	0.0743	-4.5126	0.0000	-0.4809	-0.1897
NY	-0.3684	0.0652	-5.6509	0.0000	-0.4962	-0.2406
OH	-0.2320	0.0680	-3.4118	0.0006	-0.3653	-0.0987
OK	-0.2138	0.0717	-2.9811	0.0029	-0.3544	-0.0733
OR	-0.1199	0.0823	-1.4573	0.1450	-0.2812	0.0414
PA	-0.2650	0.0687	-3.8568	0.0001	-0.3997	-0.1303
PR	0.1572	0.1252	1.2556	0.2093	-0.0882	0.4027
RI	-0.0275	0.0988	-0.2781	0.7809	-0.2211	0.1661

SC	-0.1470	0.0761	-1.9304	0.0536	-0.2962	0.0023
SD	-0.0164	0.1189	-0.1380	0.8902	-0.2495	0.2167
TN	-0.0444	0.0703	-0.6322	0.5273	-0.1822	0.0933
TX	-0.2324	0.0651	-3.5699	0.0004	-0.3600	-0.1048
UT	-0.2354	0.0725	-3.2488	0.0012	-0.3775	-0.0934
VA	-0.2640	0.0676	-3.9045	0.0001	-0.3965	-0.1315
WA	-0.1641	0.0738	-2.2246	0.0261	-0.3087	-0.0195
WI	-0.0596	0.0762	-0.7828	0.4337	-0.2090	0.0897
WV	-0.1842	0.0876	-2.1036	0.0354	-0.3559	-0.0126
WY	-0.1166	0.1499	-0.7781	0.4365	-0.4104	0.1771
year	0.0351	0.0017	20.6110	0.0000	0.0318	0.0385
hour	-0.0010	0.0008	-1.2858	0.1985	-0.0025	0.0005
numReviews	-0.0036	0.0004	-9.9667	0.0000	-0.0043	-0.0029

In [13]:

```
#predict highPunctuality using the trained probit model
drRatings['highPunctuality_predictProbit'] = (probit_result.predict(x) >= 0.5).astype(int)
drRatings[['highPunctuality_predictProbit','highPunctuality']]
```

Out[13]:

	highPunctuality_predictProbit	highPunctuality
54254	0	0
70949	0	1
29367	0	1
71006	0	0
57782	0	1
...	...	...
64936	0	0
16701	0	1
24697	1	0
52354	0	0
14441	0	0

80000 rows × 2 columns

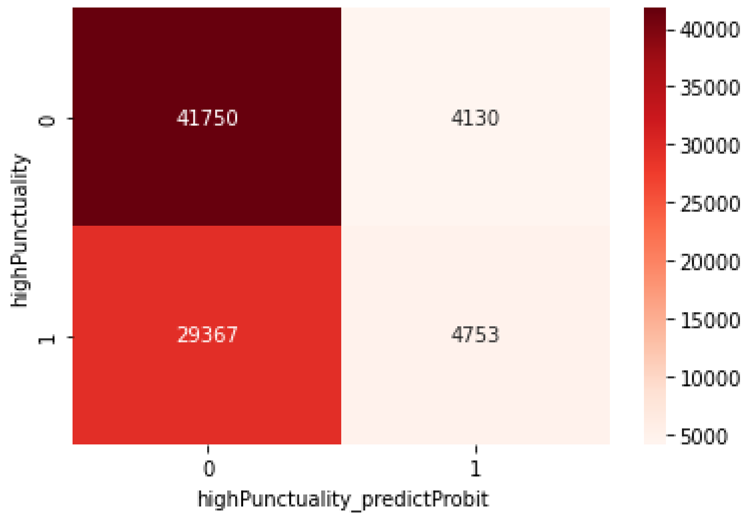
In [14]:

```
import sklearn.metrics as metrics
acc=metrics.accuracy_score(y_true=drRatings['highPunctuality'],y_pred=drRatings['highPunctuality_predict
print(acc)
confusion=metrics.confusion_matrix(y_true=drRatings['highPunctuality'],y_pred=drRatings['highPunctuality
print(confusion)
import seaborn as sn
sn.heatmap(confusion, annot=True, cmap='Reds', fmt='d')
plt.xlabel("highPunctuality_predictProbit")
plt.ylabel("highPunctuality")
```

```
0.5812875
[[41750  4130]
 [29367  4753]]
```

Out[14]:

```
Text(33.0, 0.5, 'highPunctuality')
```



In [ ]:

Try keywords for better predictions

In [15]:

```
drRatings['good']=drRatings['review'].apply(lambda x: int('good' in x))
drRatings['bad']=drRatings['review'].apply(lambda x: int('bad' in x))
drRatings.head()
```

Out[15]:

	reviewID	doctorID	doctorName	specialty	numReviews	city	state	doctorHomepage	averageRating	staff	...
54254	54255	887117	Dr. Xiorong Dai	Gynecologist (OBGYN)	3	West Palm Beach	FL	/doctor-ratings/887117/Dr-Xiorong-Dai-West%2BP...	2.00	5	...
70949	70950	33769	Dr. Lisa Yang	Gynecologist (OBGYN)	13	Cincinnati	OH	/doctor-ratings/33769/Dr-Lisa-Yang-Cincinnati-...	4.75	4	...
29367	29368	309188	Dr. Ernest H. Carlton	Gynecologist (OBGYN)	6	Macon	GA	/doctor-ratings/309188/Dr-Ernest%2BH.-Carlton-...	5.00	5	...
71006	71007	113844	Dr. Michael J. Derosa	Gynecologist (OBGYN)	23	O%26apos%3BFallon	MO	/doctor-ratings/113844/Dr-Michael%2BJ.-Derosa-...	4.00	5	...
57782	57783	28644	Dr. Dawn M. Hasson	Gynecologist (OBGYN)	2	Reading	PA	/doctor-ratings/28644/Dr-Dawn%2BM.-Hasson-Read...	4.75	4	...

5 rows × 76 columns

In [16]:

```
#decide x and y
xcols = ['AK', 'AL', 'AR', 'AZ', 'CA',
       'CO', 'CT', 'DC', 'DE', 'FL', 'GA', 'HI', 'IA', 'ID', 'IL', 'IN', 'KS',
       'KY', 'LA', 'MA', 'MD', 'ME', 'MI', 'MN', 'MO', 'MS', 'NC', 'ND', 'NE',
       'NJ', 'NM', 'NV', 'NY', 'OH', 'OK', 'OR', 'PA', 'PR', 'RI', 'SC', 'SD',
       'TN', 'TX', 'UT', 'VA', 'WA', 'WI', 'WV', 'WY', 'year', 'hour','numReviews','good','bad']

ycol = 'highPunctuality'

x = drRatings[xcols]
print(x.shape)
x = sm.add_constant(x)
y = drRatings[ycol]
print(x.shape,y.shape)
logit_model3 = sm.Logit(y, x)
logit_result = logit_model3.fit()
print(logit_result.summary2())
#predict highPunctuality using the trained logit model
drRatings['highPunctuality_predictLogit3'] = (logit_result.predict(x) >= 0.5).astype(int)
drRatings[['highPunctuality_predictLogit3','highPunctuality']]
acc=metrics.accuracy_score(y_true=drRatings['highPunctuality'],y_pred=drRatings['highPunctuality_predict
print(acc)
confusion=metrics.confusion_matrix(y_true=drRatings['highPunctuality'],y_pred=drRatings['highPunctuality
print(confusion)
sn.heatmap(confusion, annot=True, cmap='Reds', fmt='d')
plt.xlabel("highPunctuality_predictLogit3")
plt.ylabel("highPunctuality")
```

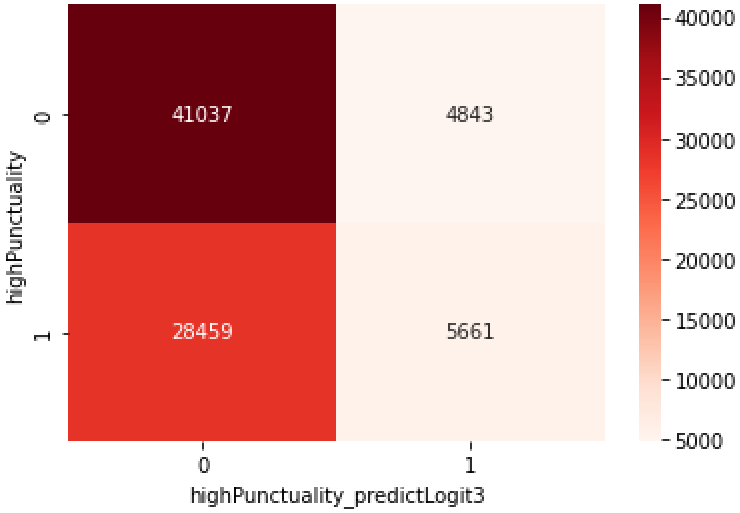
```
(80000, 54)
(80000, 55) (80000,)

/usr/local/lib/python3.7/dist-packages/statsmodels/tsa/tsatools.py:117: FutureWarning: In a future ver
sion of pandas all arguments of concat except for the argument 'objs' will be keyword-only
  x = pd.concat(x[::order], 1)

Optimization terminated successfully.
         Current function value: 0.672216
         Iterations 5
                          Results: Logit
==================================================================
Model:              Logit            Pseudo R-squared: 0.015
Dependent Variable: highPunctuality  AIC:              107664.5465
```

```
Date:               2022-03-16 02:33 BIC:              108175.4845
No. Observations:   80000            Log-Likelihood:   -53777.
Df Model:           54               LL-Null:          -54584.
Df Residuals:       79945            LLR p-value:      3.2451e-302
Converged:          1.0000           Scale:            1.0000
No. Iterations:     5.0000
------------------------------------------------------------------
              Coef.   Std.Err.    z     P>|z|    [0.025    0.975]
------------------------------------------------------------------
const       -115.9477   5.5469 -20.9030 0.0000 -126.8195 -105.0758
AK             0.1025   0.2002   0.5120 0.6086   -0.2899    0.4949
AL            -0.0548   0.1149  -0.4768 0.6335   -0.2800    0.1704
AR            -0.0659   0.1284  -0.5131 0.6079   -0.3176    0.1858
AZ            -0.5036   0.1114  -4.5187 0.0000   -0.7220   -0.2851
CA            -0.5180   0.1046  -4.9526 0.0000   -0.7230   -0.3130
CO            -0.1648   0.1224  -1.3460 0.1783   -0.4047    0.0752
CT            -0.1500   0.1199  -1.2510 0.2109   -0.3850    0.0850
DC            -0.4399   0.1445  -3.0436 0.0023   -0.7231   -0.1566
DE            -0.6652   0.1618  -4.1123 0.0000   -0.9822   -0.3482
FL            -0.3088   0.1049  -2.9446 0.0032   -0.5144   -0.1033
GA            -0.2630   0.1103  -2.3849 0.0171   -0.4791   -0.0469
HI            -0.3452   0.1715  -2.0132 0.0441   -0.6812   -0.0091
IA            -0.1539   0.1480  -1.0398 0.2984   -0.4440    0.1362
ID            -0.2576   0.1506  -1.7102 0.0872   -0.5528    0.0376
IL            -0.3805   0.1086  -3.5030 0.0005   -0.5934   -0.1676
IN            -0.1312   0.1182  -1.1097 0.2671   -0.3629    0.1005
KS            -0.2167   0.1308  -1.6561 0.0977   -0.4731    0.0398
KY            -0.2606   0.1260  -2.0678 0.0387   -0.5075   -0.0136
LA            -0.1202   0.1186  -1.0135 0.3108   -0.3526    0.1122
MA            -0.4192   0.1154  -3.6340 0.0003   -0.6453   -0.1931
MD            -0.4464   0.1111  -4.0196 0.0001   -0.6641   -0.2287
ME            -0.0525   0.1963  -0.2676 0.7890   -0.4373    0.3322
MI            -0.3034   0.1098  -2.7642 0.0057   -0.5186   -0.0883
MN            -0.0773   0.1266  -0.6107 0.5414   -0.3254    0.1708
MO            -0.0423   0.1127  -0.3755 0.7073   -0.2633    0.1786
MS             0.1941   0.1438   1.3495 0.1772   -0.0878    0.4760
NC            -0.0629   0.1141  -0.5513 0.5814   -0.2865    0.1607
ND            -0.5129   0.2296  -2.2338 0.0255   -0.9628   -0.0629
NE            -0.0037   0.1428  -0.0263 0.9791   -0.2836    0.2761
NJ            -0.3973   0.1056  -3.7633 0.0002   -0.6043   -0.1904
NM            -0.4006   0.1595  -2.5108 0.0120   -0.7132   -0.0879
NV            -0.5251   0.1196  -4.3886 0.0000   -0.7596   -0.2906
NY            -0.5900   0.1047  -5.6353 0.0000   -0.7952   -0.3848
```

```
OH             -0.3683   0.1092  -3.3729 0.0007   -0.5823   -0.1543
OK             -0.3388   0.1152  -2.9409 0.0033   -0.5646   -0.1130
OR             -0.1956   0.1319  -1.4831 0.1381   -0.4542    0.0629
PA             -0.4216   0.1104  -3.8209 0.0001   -0.6379   -0.2054
PR              0.2229   0.2018   1.1045 0.2694   -0.1727    0.6186
RI             -0.0430   0.1585  -0.2712 0.7863   -0.3536    0.2677
SC             -0.2319   0.1221  -1.8988 0.0576   -0.4713    0.0075
SD             -0.0253   0.1905  -0.1326 0.8945   -0.3986    0.3481
TN             -0.0713   0.1127  -0.6321 0.5273   -0.2922    0.1497
TX             -0.3648   0.1045  -3.4912 0.0005   -0.5696   -0.1600
UT             -0.3686   0.1164  -3.1663 0.0015   -0.5967   -0.1404
VA             -0.4140   0.1086  -3.8118 0.0001   -0.6269   -0.2011
WA             -0.2562   0.1184  -2.1643 0.0304   -0.4882   -0.0242
WI             -0.0924   0.1222  -0.7564 0.4494   -0.3318    0.1470
WV             -0.2901   0.1405  -2.0642 0.0390   -0.5655   -0.0146
WY             -0.1800   0.2404  -0.7489 0.4539   -0.6511    0.2911
year            0.0578   0.0028  20.9413 0.0000    0.0523    0.0632
hour           -0.0015   0.0013  -1.1625 0.2450   -0.0039    0.0010
numReviews     -0.0056   0.0006  -9.2382 0.0000   -0.0068   -0.0044
good           -0.0742   0.0235  -3.1501 0.0016   -0.1203   -0.0280
bad            -0.7937   0.0405 -19.6019 0.0000   -0.8730   -0.7143
===================================================================

0.583725
[[41037  4843]
 [28459  5661]]
```

Out[16]:

```
Text(33.0, 0.5, 'highPunctuality')
```



Search for a good set of keywords

In [17]:

```
keywordlist = ['good','bad','time','wait','no']
for keyword in keywordlist:
  drRatings[keyword]=drRatings['review'].apply(lambda x: int(keyword in x))

drRatings.head()
```

Out[17]:

	reviewID	doctorID	doctorName	specialty	numReviews	city	state	doctorHomepage	averageRating	staff	...
54254	54255	887117	Dr. Xiorong Dai	Gynecologist (OBGYN)	3	West Palm Beach	FL	/doctor-ratings/887117/Dr-Xiorong-Dai-West%2BP...	2.00	5	...
70949	70950	33769	Dr. Lisa Yang	Gynecologist (OBGYN)	13	Cincinnati	OH	/doctor-ratings/33769/Dr-Lisa-Yang-Cincinnati-...	4.75	4	...
29367	29368	309188	Dr. Ernest H. Carlton	Gynecologist (OBGYN)	6	Macon	GA	/doctor-ratings/309188/Dr-Ernest%2BH.-Carlton-...	5.00	5	...
71006	71007	113844	Dr. Michael J. Derosa	Gynecologist (OBGYN)	23	O%26apos%3BFallon	MO	/doctor-ratings/113844/Dr-Michael%2BJ.-Derosa-...	4.00	5	...
57782	57783	28644	Dr. Dawn M. Hasson	Gynecologist (OBGYN)	2	Reading	PA	/doctor-ratings/28644/Dr-Dawn%2BM.-Hasson-Read...	4.75	4	...

5 rows × 80 columns

In [18]:

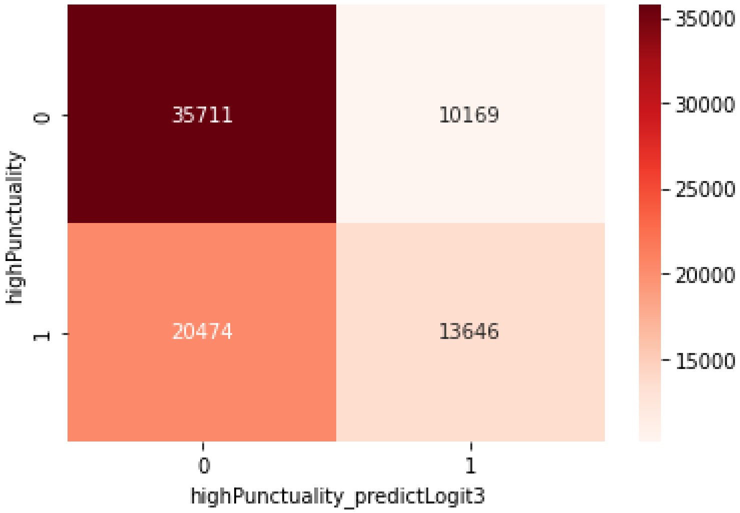
```
xcols = ['AK', 'AL', 'AR', 'AZ', 'CA',
       'CO', 'CT', 'DC', 'DE', 'FL', 'GA', 'HI', 'IA', 'ID', 'IL', 'IN', 'KS',
       'KY', 'LA', 'MA', 'MD', 'ME', 'MI', 'MN', 'MO', 'MS', 'NC', 'ND', 'NE',
       'NJ', 'NM', 'NV', 'NY', 'OH', 'OK', 'OR', 'PA', 'PR', 'RI', 'SC', 'SD',
       'TN', 'TX', 'UT', 'VA', 'WA', 'WI', 'WV', 'WY', 'year', 'hour','numReviews','good','bad','time','

ycol = 'highPunctuality'

x = drRatings[xcols]
print(x.shape)
x = sm.add_constant(x)
y = drRatings[ycol]
print(x.shape,y.shape)
logit_model3 = sm.Logit(y, x)
logit_result = logit_model3.fit()
print(logit_result.summary2())
#predict highPunctuality using the trained logit model
drRatings['highPunctuality_predictLogit3'] = (logit_result.predict(x) >= 0.5).astype(int)
drRatings[['highPunctuality_predictLogit3','highPunctuality']]
acc=metrics.accuracy_score(y_true=drRatings['highPunctuality'],y_pred=drRatings['highPunctuality_predict
print(acc)
confusion=metrics.confusion_matrix(y_true=drRatings['highPunctuality'],y_pred=drRatings['highPunctuality
print(confusion)
sn.heatmap(confusion, annot=True, cmap='Reds', fmt='d')
plt.xlabel("highPunctuality_predictLogit3")
plt.ylabel("highPunctuality")
```

```
(80000, 57)
(80000, 58) (80000,)

/usr/local/lib/python3.7/dist-packages/statsmodels/tsa/tsatools.py:117: FutureWarning: In a future ver
sion of pandas all arguments of concat except for the argument 'objs' will be keyword-only
  x = pd.concat(x[::order], 1)

Optimization terminated successfully.
         Current function value: 0.646484
         Iterations 6
                          Results: Logit
==================================================================
Model:              Logit            Pseudo R-squared: 0.052
Dependent Variable: highPunctuality  AIC:              103553.4742
Date:               2022-03-16 02:34 BIC:              104092.2816
No. Observations:   80000            Log-Likelihood:   -51719.
```

```
Df Model:           57               LL-Null:           -54584.
Df Residuals:       79942            LLR p-value:       0.0000
Converged:          1.0000           Scale:             1.0000
No. Iterations:     6.0000
-------------------------------------------------------------------
              Coef.   Std.Err.    z      P>|z|    [0.025    0.975]
-------------------------------------------------------------------
const       -128.7661   5.6814 -22.6647 0.0000 -139.9014 -117.6309
AK             0.0674   0.2031   0.3320 0.7399   -0.3307    0.4655
AL            -0.0338   0.1170  -0.2885 0.7729   -0.2631    0.1956
AR            -0.0638   0.1310  -0.4873 0.6261   -0.3206    0.1929
AZ            -0.4174   0.1135  -3.6784 0.0002   -0.6398   -0.1950
CA            -0.4350   0.1064  -4.0877 0.0000   -0.6436   -0.2264
CO            -0.1279   0.1248  -1.0245 0.3056   -0.3725    0.1168
CT            -0.1265   0.1221  -1.0362 0.3001   -0.3658    0.1128
DC            -0.3564   0.1473  -2.4196 0.0155   -0.6452   -0.0677
DE            -0.6047   0.1651  -3.6624 0.0002   -0.9282   -0.2811
FL            -0.2476   0.1067  -2.3198 0.0203   -0.4568   -0.0384
GA            -0.2098   0.1123  -1.8690 0.0616   -0.4298    0.0102
HI            -0.2975   0.1752  -1.6974 0.0896   -0.6409    0.0460
IA            -0.1182   0.1508  -0.7833 0.4334   -0.4138    0.1775
ID            -0.2216   0.1534  -1.4440 0.1487   -0.5223    0.0792
IL            -0.3132   0.1106  -2.8322 0.0046   -0.5299   -0.0965
IN            -0.0840   0.1204  -0.6971 0.4857   -0.3200    0.1521
KS            -0.1594   0.1335  -1.1933 0.2328   -0.4211    0.1024
KY            -0.2090   0.1285  -1.6259 0.1040   -0.4608    0.0429
LA            -0.0651   0.1210  -0.5384 0.5903   -0.3022    0.1720
MA            -0.3791   0.1175  -3.2274 0.0012   -0.6094   -0.1489
MD            -0.3930   0.1130  -3.4772 0.0005   -0.6146   -0.1715
ME            -0.0392   0.1994  -0.1967 0.8441   -0.4301    0.3517
MI            -0.2711   0.1117  -2.4262 0.0153   -0.4901   -0.0521
MN            -0.0667   0.1288  -0.5174 0.6049   -0.3192    0.1859
MO            -0.0087   0.1148  -0.0761 0.9393   -0.2337    0.2163
MS             0.1806   0.1472   1.2268 0.2199   -0.1079    0.4691
NC            -0.0265   0.1161  -0.2285 0.8193   -0.2542    0.2011
ND            -0.4701   0.2349  -2.0011 0.0454   -0.9305   -0.0097
NE            -0.0219   0.1456  -0.1504 0.8805   -0.3073    0.2635
NJ            -0.3246   0.1074  -3.0211 0.0025   -0.5351   -0.1140
NM            -0.3736   0.1627  -2.2961 0.0217   -0.6925   -0.0547
NV            -0.4224   0.1221  -3.4593 0.0005   -0.6617   -0.1831
NY            -0.4973   0.1065  -4.6683 0.0000   -0.7061   -0.2885
OH            -0.3280   0.1111  -2.9520 0.0032   -0.5458   -0.1102
OK            -0.2762   0.1174  -2.3530 0.0186   -0.5063   -0.0461
```

```
OR             -0.1865   0.1343  -1.3884 0.1650   -0.4497    0.0768
PA             -0.3960   0.1123  -3.5260 0.0004   -0.6161   -0.1759
PR              0.1360   0.2055   0.6616 0.5082   -0.2668    0.5387
RI             -0.0546   0.1615  -0.3378 0.7355   -0.3712    0.2620
SC             -0.1645   0.1246  -1.3200 0.1868   -0.4087    0.0797
SD             -0.0786   0.1936  -0.4061 0.6847   -0.4581    0.3008
TN             -0.0014   0.1149  -0.0119 0.9905   -0.2266    0.2238
TX             -0.2775   0.1063  -2.6096 0.0091   -0.4859   -0.0691
UT             -0.2863   0.1187  -2.4124 0.0158   -0.5188   -0.0537
VA             -0.3514   0.1105  -3.1793 0.0015   -0.5680   -0.1348
WA             -0.1994   0.1206  -1.6532 0.0983   -0.4358    0.0370
WI             -0.0637   0.1243  -0.5126 0.6082   -0.3074    0.1799
WV             -0.2099   0.1440  -1.4577 0.1449   -0.4922    0.0723
WY             -0.1825   0.2437  -0.7489 0.4539   -0.6602    0.2952
year            0.0643   0.0028  22.7692 0.0000    0.0588    0.0699
hour           -0.0014   0.0013  -1.0627 0.2879   -0.0039    0.0012
numReviews     -0.0023   0.0006  -3.7134 0.0002   -0.0035   -0.0011
good            0.0274   0.0242   1.1315 0.2578   -0.0201    0.0750
bad            -0.6622   0.0413 -16.0189 0.0000   -0.7432   -0.5812
time           -0.0869   0.0165  -5.2747 0.0000   -0.1192   -0.0546
wait           -1.1322   0.0263 -43.0783 0.0000   -1.1838   -1.0807
no             -0.5650   0.0155 -36.3841 0.0000   -0.5954   -0.5346
==================================================================

0.6169625
[[35711 10169]
 [20474 13646]]
```

```
Out[18]: Text(33.0, 0.5, 'highPunctuality')
```


35711
10169
20474
13646
highPunctuality
0
1
highPunctuality_predictLogit3
35000
30000
25000
20000
15000

In [19]:

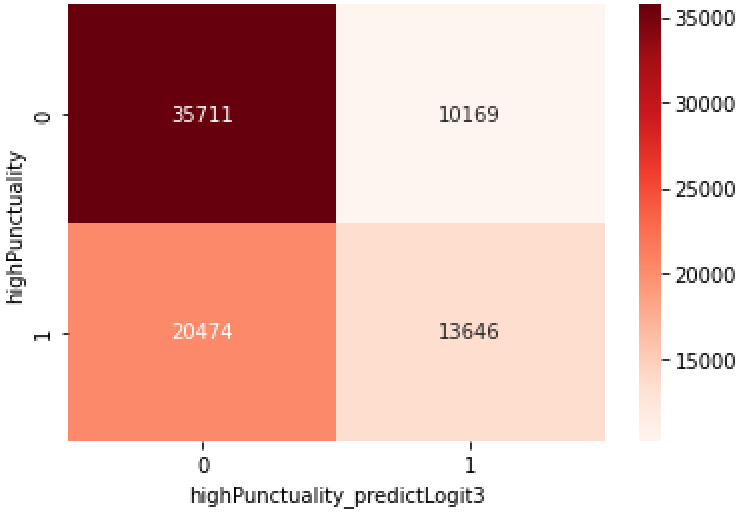
```
def search4keywords(keywordlist):
  for keyword in keywordlist:
    drRatings[keyword]=drRatings['review'].apply(lambda x: int(keyword in x))

  drRatings.head()
  xcols = ['AK', 'AL', 'AR', 'AZ', 'CA',
       'CO', 'CT', 'DC', 'DE', 'FL', 'GA', 'HI', 'IA', 'ID', 'IL', 'IN', 'KS',
       'KY', 'LA', 'MA', 'MD', 'ME', 'MI', 'MN', 'MO', 'MS', 'NC', 'ND', 'NE',
       'NJ', 'NM', 'NV', 'NY', 'OH', 'OK', 'OR', 'PA', 'PR', 'RI', 'SC', 'SD',
       'TN', 'TX', 'UT', 'VA', 'WA', 'WI', 'WV', 'WY', 'year', 'hour','numReviews']+keywordlist

  ycol = 'highPunctuality'

  x = drRatings[xcols]
  print(x.shape)
  x = sm.add_constant(x)
  y = drRatings[ycol]
  print(x.shape,y.shape)
  logit_model3 = sm.Logit(y, x)
  logit_result = logit_model3.fit()
  print(logit_result.summary2())
  #predict highPunctuality using the trained logit model
  drRatings['highPunctuality_predictLogit3'] = (logit_result.predict(x) >= 0.5).astype(int)
  drRatings[['highPunctuality_predictLogit3','highPunctuality']]
  acc=metrics.accuracy_score(y_true=drRatings['highPunctuality'],y_pred=drRatings['highPunctuality_predi
  print(acc)
  confusion=metrics.confusion_matrix(y_true=drRatings['highPunctuality'],y_pred=drRatings['highPunctuali
  print(confusion)
  sn.heatmap(confusion, annot=True, cmap='Reds', fmt='d')
  plt.xlabel("highPunctuality_predictLogit3")
  plt.ylabel("highPunctuality")
  return(acc)
```

In [20]:

```python
keywordlist = ['good','bad','time','wait','no']
acc=search4keywords(keywordlist)
```

```
(80000, 57)
(80000, 58) (80000,)

/usr/local/lib/python3.7/dist-packages/statsmodels/tsa/tsatools.py:117: FutureWarning: In a future ver
sion of pandas all arguments of concat except for the argument 'objs' will be keyword-only
  x = pd.concat(x[::order], 1)

Optimization terminated successfully.
         Current function value: 0.646484
         Iterations 6
                          Results: Logit
==================================================================
Model:              Logit            Pseudo R-squared: 0.052
Dependent Variable: highPunctuality  AIC:              103553.4742
Date:               2022-03-16 02:35 BIC:              104092.2816
No. Observations:   80000            Log-Likelihood:   -51719.
Df Model:           57               LL-Null:          -54584.
Df Residuals:       79942            LLR p-value:      0.0000
Converged:          1.0000           Scale:            1.0000
No. Iterations:     6.0000
------------------------------------------------------------------
              Coef.   Std.Err.    z     P>|z|    [0.025    0.975]
------------------------------------------------------------------
const       -128.7661   5.6814 -22.6647 0.0000 -139.9014 -117.6309
AK             0.0674   0.2031   0.3320 0.7399   -0.3307    0.4655
AL            -0.0338   0.1170  -0.2885 0.7729   -0.2631    0.1956
AR            -0.0638   0.1310  -0.4873 0.6261   -0.3206    0.1929
AZ            -0.4174   0.1135  -3.6784 0.0002   -0.6398   -0.1950
CA            -0.4350   0.1064  -4.0877 0.0000   -0.6436   -0.2264
CO            -0.1279   0.1248  -1.0245 0.3056   -0.3725    0.1168
CT            -0.1265   0.1221  -1.0362 0.3001   -0.3658    0.1128
DC            -0.3564   0.1473  -2.4196 0.0155   -0.6452   -0.0677
DE            -0.6047   0.1651  -3.6624 0.0002   -0.9282   -0.2811
FL            -0.2476   0.1067  -2.3198 0.0203   -0.4568   -0.0384
GA            -0.2098   0.1123  -1.8690 0.0616   -0.4298    0.0102
HI            -0.2975   0.1752  -1.6974 0.0896   -0.6409    0.0460
IA            -0.1182   0.1508  -0.7833 0.4334   -0.4138    0.1775
ID            -0.2216   0.1534  -1.4440 0.1487   -0.5223    0.0792
IL            -0.3132   0.1106  -2.8322 0.0046   -0.5299   -0.0965
IN            -0.0840   0.1204  -0.6971 0.4857   -0.3200    0.1521
```

```
KS            -0.1594   0.1335  -1.1933 0.2328   -0.4211    0.1024
KY            -0.2090   0.1285  -1.6259 0.1040   -0.4608    0.0429
LA            -0.0651   0.1210  -0.5384 0.5903   -0.3022    0.1720
MA            -0.3791   0.1175  -3.2274 0.0012   -0.6094   -0.1489
MD            -0.3930   0.1130  -3.4772 0.0005   -0.6146   -0.1715
ME            -0.0392   0.1994  -0.1967 0.8441   -0.4301    0.3517
MI            -0.2711   0.1117  -2.4262 0.0153   -0.4901   -0.0521
MN            -0.0667   0.1288  -0.5174 0.6049   -0.3192    0.1859
MO            -0.0087   0.1148  -0.0761 0.9393   -0.2337    0.2163
MS             0.1806   0.1472   1.2268 0.2199   -0.1079    0.4691
NC            -0.0265   0.1161  -0.2285 0.8193   -0.2542    0.2011
ND            -0.4701   0.2349  -2.0011 0.0454   -0.9305   -0.0097
NE            -0.0219   0.1456  -0.1504 0.8805   -0.3073    0.2635
NJ            -0.3246   0.1074  -3.0211 0.0025   -0.5351   -0.1140
NM            -0.3736   0.1627  -2.2961 0.0217   -0.6925   -0.0547
NV            -0.4224   0.1221  -3.4593 0.0005   -0.6617   -0.1831
NY            -0.4973   0.1065  -4.6683 0.0000   -0.7061   -0.2885
OH            -0.3280   0.1111  -2.9520 0.0032   -0.5458   -0.1102
OK            -0.2762   0.1174  -2.3530 0.0186   -0.5063   -0.0461
OR            -0.1865   0.1343  -1.3884 0.1650   -0.4497    0.0768
PA            -0.3960   0.1123  -3.5260 0.0004   -0.6161   -0.1759
PR             0.1360   0.2055   0.6616 0.5082   -0.2668    0.5387
RI            -0.0546   0.1615  -0.3378 0.7355   -0.3712    0.2620
SC            -0.1645   0.1246  -1.3200 0.1868   -0.4087    0.0797
SD            -0.0786   0.1936  -0.4061 0.6847   -0.4581    0.3008
TN            -0.0014   0.1149  -0.0119 0.9905   -0.2266    0.2238
TX            -0.2775   0.1063  -2.6096 0.0091   -0.4859   -0.0691
UT            -0.2863   0.1187  -2.4124 0.0158   -0.5188   -0.0537
VA            -0.3514   0.1105  -3.1793 0.0015   -0.5680   -0.1348
WA            -0.1994   0.1206  -1.6532 0.0983   -0.4358    0.0370
WI            -0.0637   0.1243  -0.5126 0.6082   -0.3074    0.1799
WV            -0.2099   0.1440  -1.4577 0.1449   -0.4922    0.0723
WY            -0.1825   0.2437  -0.7489 0.4539   -0.6602    0.2952
year           0.0643   0.0028  22.7692 0.0000    0.0588    0.0699
hour          -0.0014   0.0013  -1.0627 0.2879   -0.0039    0.0012
numReviews    -0.0023   0.0006  -3.7134 0.0002   -0.0035   -0.0011
good           0.0274   0.0242   1.1315 0.2578   -0.0201    0.0750
bad           -0.6622   0.0413 -16.0189 0.0000   -0.7432   -0.5812
time          -0.0869   0.0165  -5.2747 0.0000   -0.1192   -0.0546
wait          -1.1322   0.0263 -43.0783 0.0000   -1.1838   -1.0807
no            -0.5650   0.0155 -36.3841 0.0000   -0.5954   -0.5346
==================================================================
0.6169625
[[35711 10169]
 [20474 13646]]
```

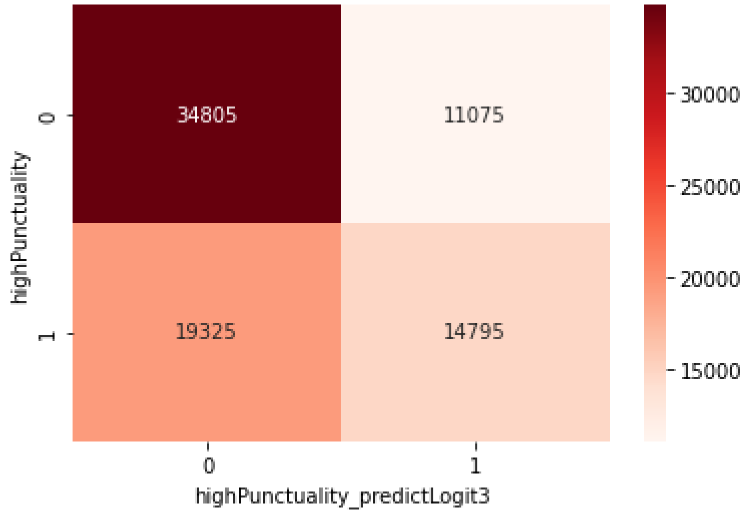


In [21]:
```
keywordlist = ['good','bad','time','wait','no','early','late','minute']
acc=search4keywords(keywordlist)
```

```
(80000, 60)
(80000, 61) (80000,)
Optimization terminated successfully.
         Current function value: 0.640691
         Iterations 6
                          Results: Logit
==================================================================
Model:              Logit            Pseudo R-squared: 0.061
Dependent Variable: highPunctuality  AIC:              102632.6143
Date:               2022-03-16 02:36 BIC:              103199.2910
No. Observations:   80000            Log-Likelihood:   -51255.
Df Model:           60               LL-Null:          -54584.
Df Residuals:       79939            LLR p-value:      0.0000
Converged:          1.0000           Scale:            1.0000
No. Iterations:     6.0000
------------------------------------------------------------------
              Coef.   Std.Err.    z     P>|z|    [0.025    0.975]
------------------------------------------------------------------
const       -132.6808   5.7127 -23.2257 0.0000 -143.8774 -121.4841
AK             0.0831   0.2049   0.4055 0.6851   -0.3185    0.4847
AL            -0.0498   0.1177  -0.4229 0.6724   -0.2805    0.1809
AR            -0.0782   0.1318  -0.5935 0.5528   -0.3365    0.1800
AZ            -0.4059   0.1142  -3.5556 0.0004   -0.6297   -0.1822
CA            -0.4282   0.1071  -4.0001 0.0001   -0.6381   -0.2184
CO            -0.1298   0.1256  -1.0336 0.3013   -0.3759    0.1163
CT            -0.1401   0.1228  -1.1414 0.2537   -0.3808    0.1005
DC            -0.3478   0.1482  -2.3462 0.0190   -0.6383   -0.0573
DE            -0.6071   0.1662  -3.6539 0.0003   -0.9328   -0.2815
FL            -0.2604   0.1074  -2.4252 0.0153   -0.4708   -0.0499
GA            -0.2227   0.1129  -1.9724 0.0486   -0.4440   -0.0014
HI            -0.3244   0.1760  -1.8427 0.0654   -0.6694    0.0206
IA            -0.1365   0.1515  -0.9012 0.3675   -0.4335    0.1604
ID            -0.2225   0.1542  -1.4427 0.1491   -0.5247    0.0798
IL            -0.3162   0.1112  -2.8423 0.0045   -0.5342   -0.0981
IN            -0.0913   0.1211  -0.7536 0.4511   -0.3287    0.1461
KS            -0.1566   0.1344  -1.1647 0.2441   -0.4200    0.1069
KY            -0.2207   0.1292  -1.7082 0.0876   -0.4740    0.0325
LA            -0.0809   0.1216  -0.6653 0.5058   -0.3193    0.1574
MA            -0.3928   0.1181  -3.3256 0.0009   -0.6243   -0.1613
MD            -0.3965   0.1137  -3.4874 0.0005   -0.6193   -0.1736
```

```
ME            -0.0562   0.2002  -0.2808 0.7789   -0.4485    0.3361
MI            -0.2763   0.1124  -2.4587 0.0139   -0.4966   -0.0561
MN            -0.0501   0.1297  -0.3859 0.6995   -0.3043    0.2042
MO            -0.0085   0.1155  -0.0734 0.9415   -0.2349    0.2179
MS             0.1526   0.1478   1.0320 0.3021   -0.1372    0.4423
NC            -0.0425   0.1168  -0.3636 0.7162   -0.2714    0.1864
ND            -0.4906   0.2357  -2.0817 0.0374   -0.9526   -0.0287
NE            -0.0369   0.1463  -0.2523 0.8008   -0.3237    0.2499
NJ            -0.3401   0.1081  -3.1472 0.0016   -0.5519   -0.1283
NM            -0.3676   0.1637  -2.2454 0.0247   -0.6885   -0.0467
NV            -0.4218   0.1228  -3.4346 0.0006   -0.6624   -0.1811
NY            -0.5070   0.1071  -4.7320 0.0000   -0.7170   -0.2970
OH            -0.3358   0.1118  -3.0047 0.0027   -0.5549   -0.1168
OK            -0.2830   0.1181  -2.3966 0.0165   -0.5144   -0.0515
OR            -0.1948   0.1351  -1.4425 0.1492   -0.4596    0.0699
PA            -0.4099   0.1129  -3.6295 0.0003   -0.6313   -0.1886
PR             0.0944   0.2059   0.4586 0.6465   -0.3091    0.4979
RI            -0.0767   0.1623  -0.4724 0.6366   -0.3947    0.2414
SC            -0.1690   0.1253  -1.3480 0.1777   -0.4146    0.0767
SD            -0.0565   0.1951  -0.2898 0.7720   -0.4390    0.3259
TN            -0.0132   0.1156  -0.1145 0.9088   -0.2397    0.2133
TX            -0.2819   0.1070  -2.6354 0.0084   -0.4915   -0.0722
UT            -0.2763   0.1194  -2.3139 0.0207   -0.5104   -0.0423
VA            -0.3435   0.1112  -3.0887 0.0020   -0.5614   -0.1255
WA            -0.2006   0.1214  -1.6526 0.0984   -0.4385    0.0373
WI            -0.0655   0.1251  -0.5234 0.6007   -0.3107    0.1797
WV            -0.2331   0.1446  -1.6117 0.1070   -0.5166    0.0504
WY            -0.2007   0.2449  -0.8198 0.4123   -0.6807    0.2792
year           0.0663   0.0028  23.3341 0.0000    0.0607    0.0718
hour          -0.0013   0.0013  -0.9909 0.3217   -0.0038    0.0013
numReviews    -0.0020   0.0006  -3.2744 0.0011   -0.0032   -0.0008
good           0.0277   0.0244   1.1365 0.2557   -0.0201    0.0755
bad           -0.6420   0.0416 -15.4373 0.0000   -0.7235   -0.5605
time          -0.0527   0.0166  -3.1715 0.0015   -0.0853   -0.0201
wait          -1.0417   0.0270 -38.5760 0.0000   -1.0946   -0.9887
no            -0.5116   0.0157 -32.6330 0.0000   -0.5424   -0.4809
early         -0.1866   0.0416  -4.4807 0.0000   -0.2683   -0.1050
late          -0.8440   0.0327 -25.7937 0.0000   -0.9082   -0.7799
minute        -0.4127   0.0415  -9.9504 0.0000   -0.4940   -0.3314
==================================================================

0.62
```

```
[[34805 11075]
 [19325 14795]]
```


highPunctuality
highPunctuality_predictLogit3