Sensor Data
User Behavior Analysis

Dan Ma

Capstone Project

McCormick
Northwestern Engineering

Department of Mechanical Engineering
Northwestern University

神策数据
SENSORS Data

产品特点 解决方案 用户中心 关于我们 数据驱动大会

体验 Demo

神策分析
驱动企业决策和产品智能

可以私有化部署的用户行为分析平台

体验 Demo

观看视频

Sign up Procedure

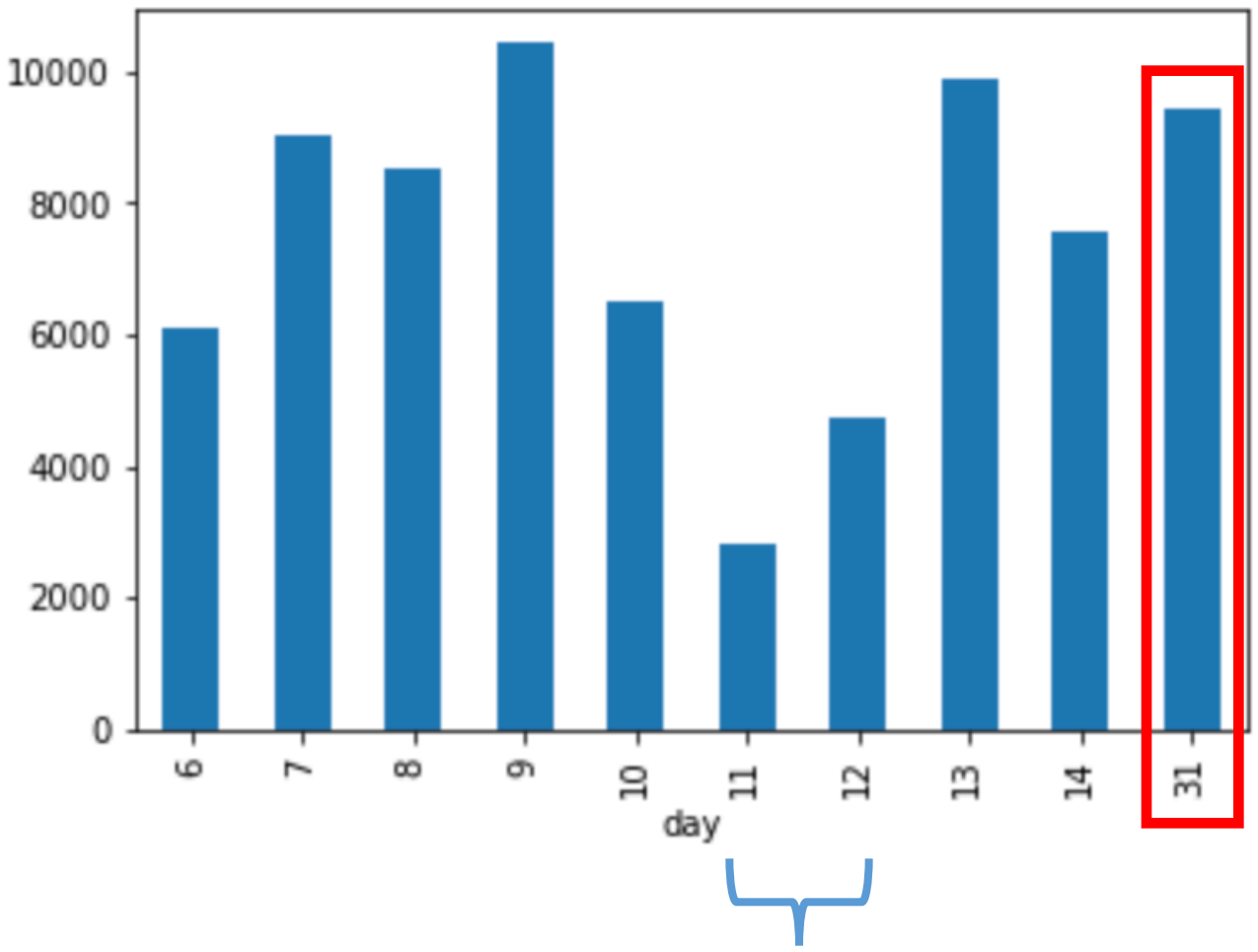
登录成功后，您将即刻体验到神策分析的强大分析功能。

企业服务类 Demo ，通过模拟企业服务类企业的典型应用场景，帮助处于不同发展阶段企业，实现潜客获取和管理、客户服务等流程的持续优化，从而实现整体营收的提升。在客户整个生命周期中，知道如何增强客户黏性、提升客户满意度，从而让保障客户续约率和提升 NPS 成为企业业绩可持续增长的终极奥义。Demo通过模拟数据可直观体验最佳行业实践范例。

神策分析拥有八大分析模型，可实现深入洞察用户行为，数据驱动产品与运营优化。

User Activity Time Series

2017 March



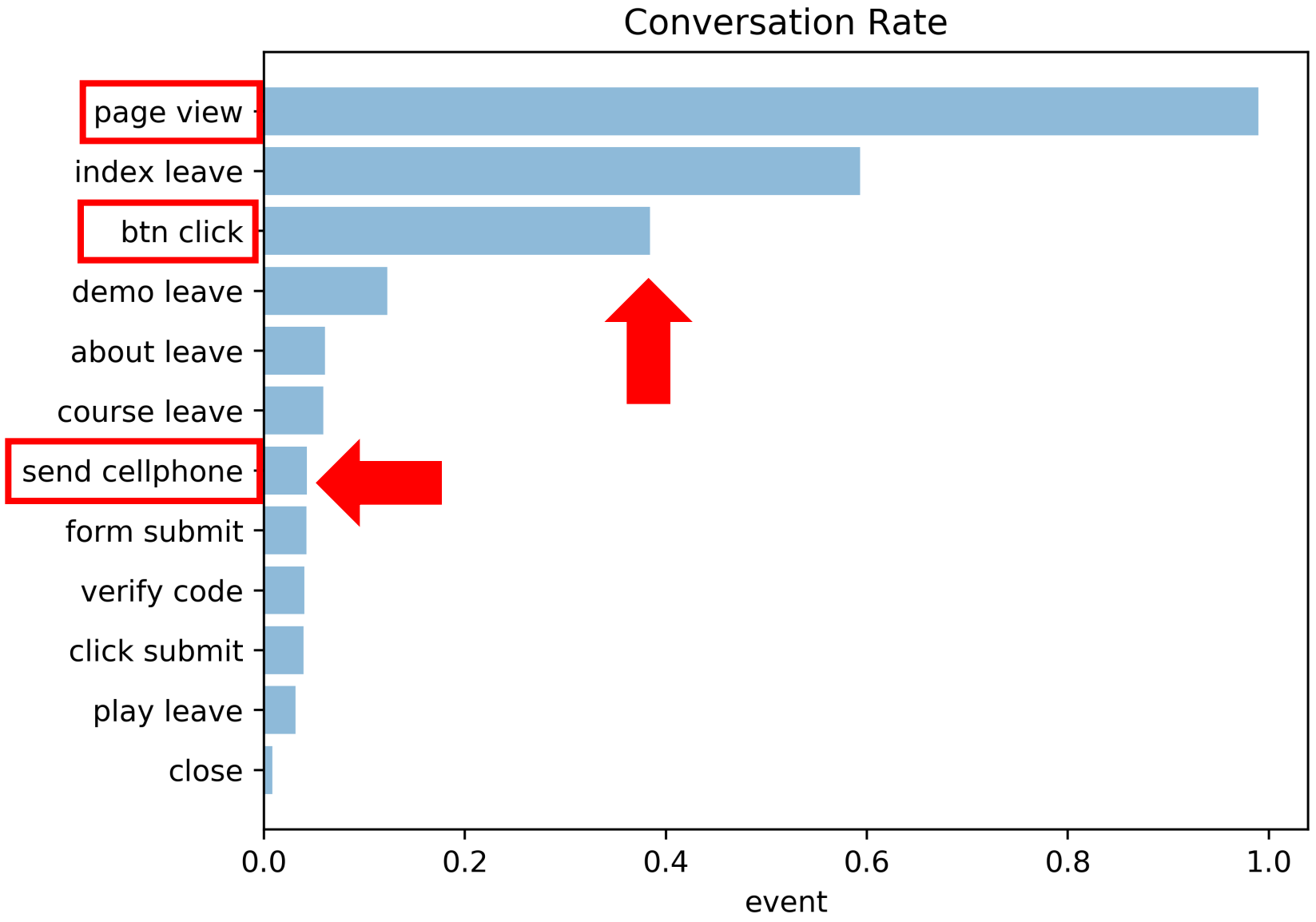
- The data set spans for 9 days (over a week)
- Weekends activity drop significantly: most users view this website due to work requirements/ interests

✓ Introduce a weekend or not feature

✓ Introduce a worktime or not feature (8AM to 5PM in Beijing time zone)

- 31st is an isolated day, probably contains wrong data, need to be excluded

Funnel Analysis



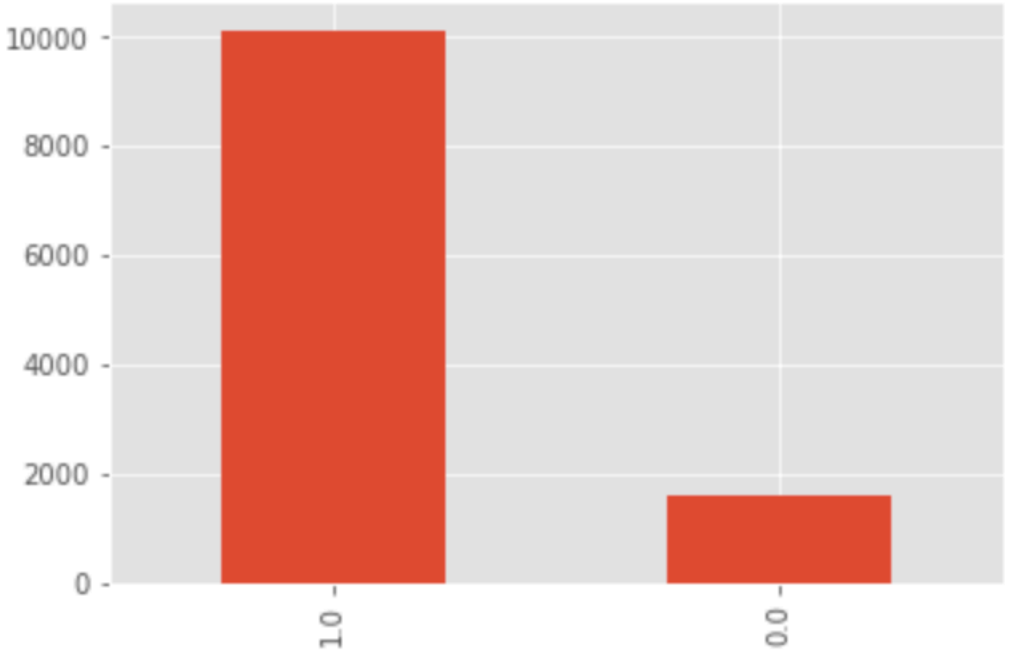
- Drop from page view to bottom click
- **Most users do not have the interest to click on pages**
- Page quality?

- Another sharp drop from bottom click to send cellphone verification code
- **Some interested users do not want to register with cellphone number**
- Privacy concerns?
- Do not have cellphone number from mainland China?

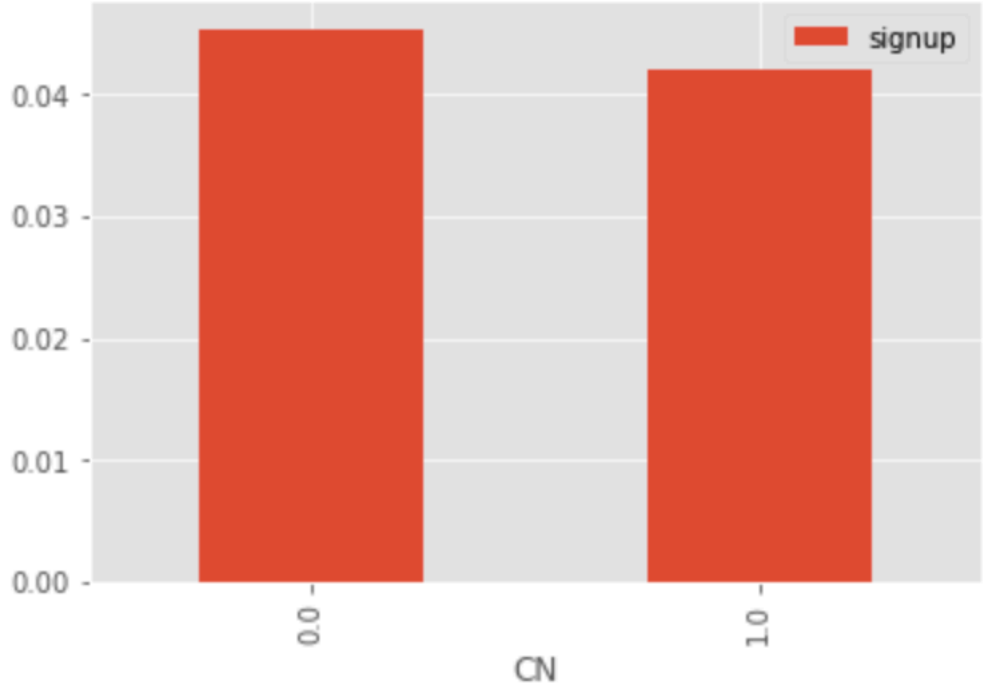
Overseas User Behavior

- Identification: IP address

# of users count in China and in abroad



Sign up rate for users in China and in abroad



- Here sign up rate considers form submit event only, other sign up related events also checked;
- Users in abroad still have high interest in sign up with cellphone verification

Utm Analysis

- Urchin Tracking Module (UTM) parameters are five variants of **URL** parameters used by marketers to track the effectiveness of online marketing campaigns across traffic sources and publishing media

Source

```
df.latest_utm_s.value_counts(dropna=False)
```

baidu	36085
NaN	25090
sogou	1943
sales4c	441
wechat	432
google	393
admin	374
sanjieke.cn	273
next.36kr.com	68

Medium

```
df.latest_utm_m.value_counts(dropna=False)
```

cpc	34623
NaN	25982
mcpc	3255
mfeed	934
default	538
answer	133
banner	67

cpc: cost per click

Value

```
df.latest_utm_t.value_counts(dropna=False)
```

NaN	26578
神策	7529
用户画像	5349
神策数据	3393
数据分析	1419
首页-通用词-三图-图1	934
大数据分析	813
用户分析	812
神策分析	677
电子商务数据	662
聚类分析	511
网站运营数据分析	506
网站数据统计	494

Campaign

```
df.latest_utm_campaign.value_counts(dropna=False)
```

NaN	25770
通用词	22180
品牌词	11929
S-通用词	1917
神策-移动推广	998
首页	934
G-通用词	391
用户行为	285

Content

```
df.latest_utm_content.value_counts(dropna=False)
```

NaN	26910
品牌-神策	11678
通用-用户画像	5529
通用-数据分析	3136
通用-数据分析-产品	1403
通用-数据分析-行业	1242
通用-数据分析-运营	1042
通用-用户分析	983

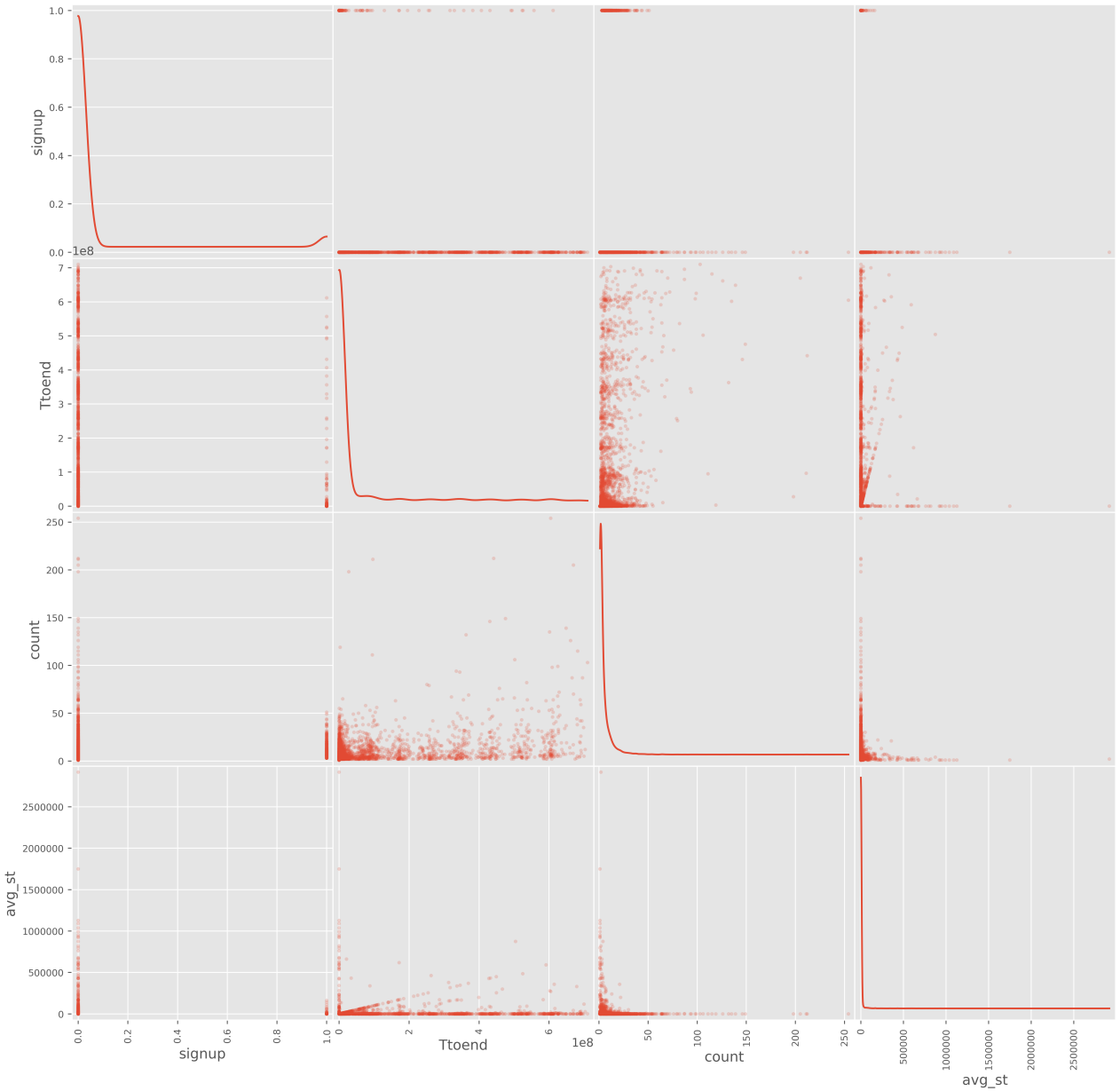
User activity Analysis

Sign up

First to register time period

Visit counts

Average page stay time



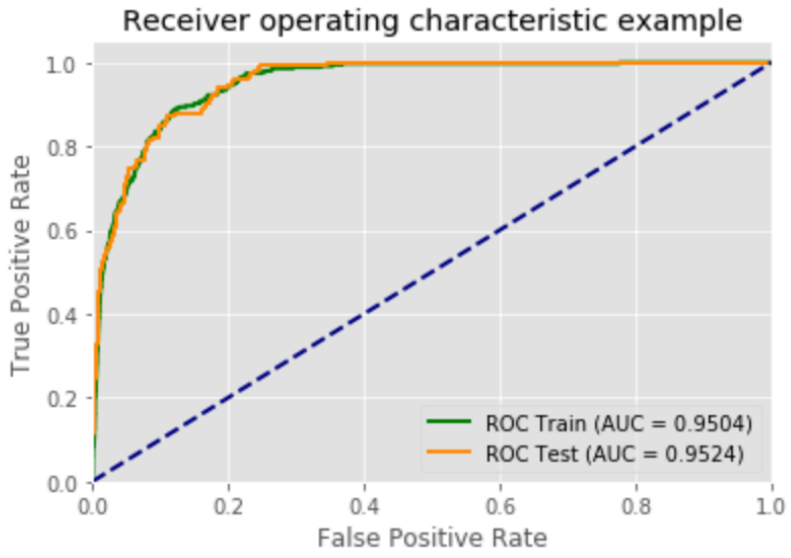
- Average stay time seems to have too broad distribution, use log function on it
- Visit counts highly correlated to average stay time on page

- Feature engineering and regularization are important, especially for logistic regression

Machine Learning Methods

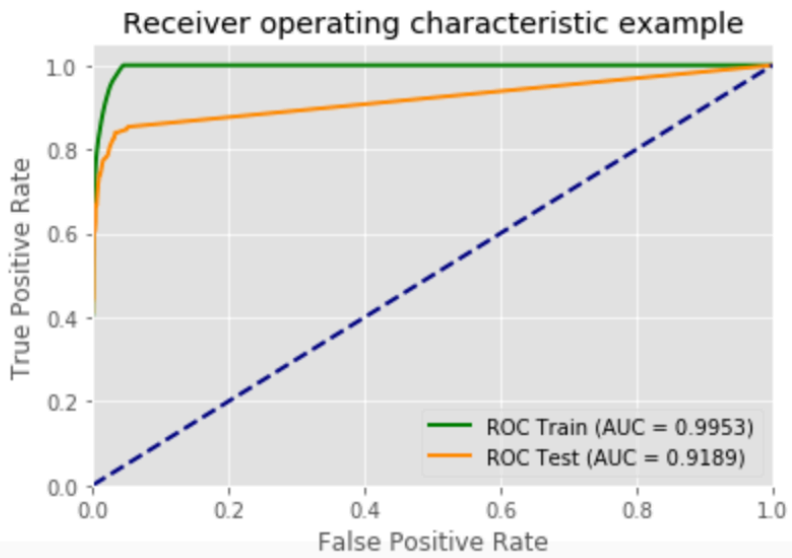
Logistic Regression

metrics	train	test
AUC	0.950415	0.952382
Accuracy	0.961928	0.963849
Precision	0.654867	0.755556
Recall	0.213256	0.226667
f1-score	0.321739	0.348718



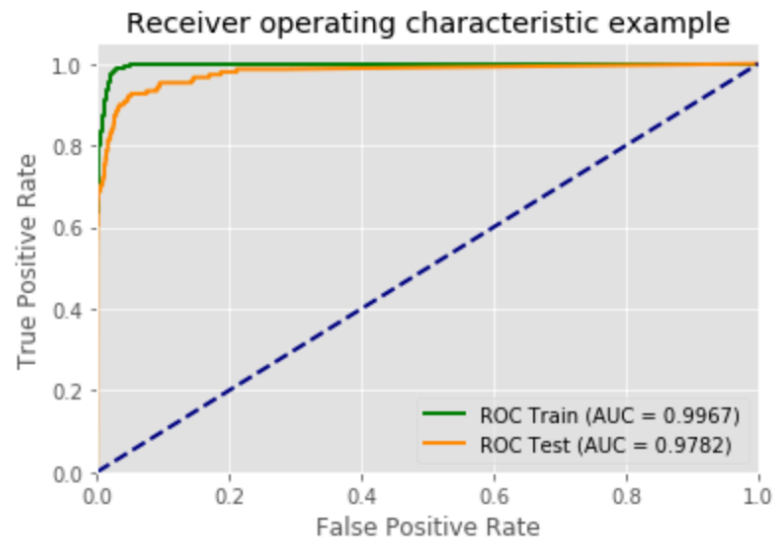
Single Tree

metrics	train	test
AUC	0.995323	0.918908
Accuracy	0.985479	0.980074
Precision	0.880000	0.822581
Recall	0.760807	0.680000
f1-score	0.816074	0.744526



Bagged Trees

metrics	train	test
AUC	0.996693	0.978180
Accuracy	0.986455	0.984344
Precision	0.940299	0.943925
Recall	0.726225	0.673333
f1-score	0.819512	0.785992

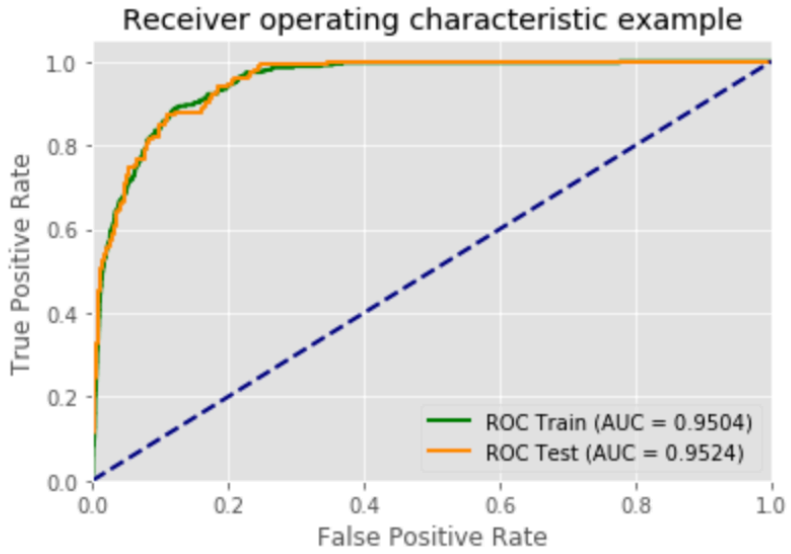


- Logistic regression is the simplest model, even though performance is not good, train and test has the most similar performance
- Bagging can make better performance

Why low Recall and f1 score

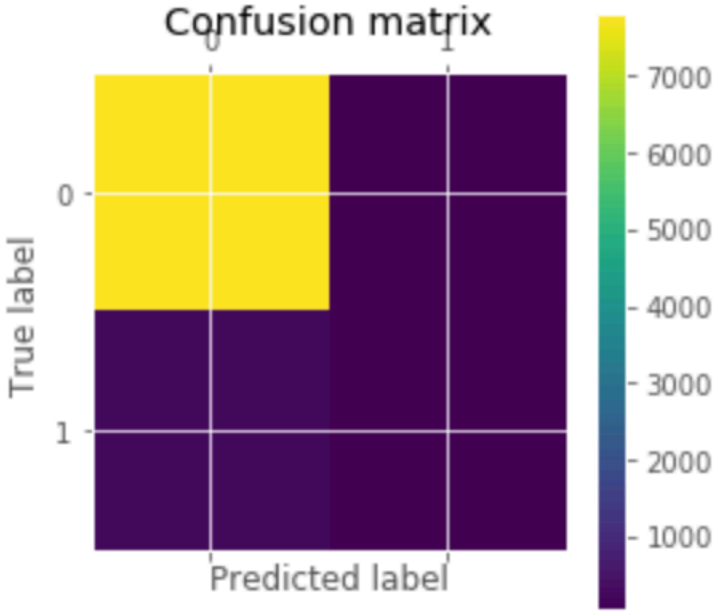
Logistic Regression

metrics	train	test
AUC	0.950415	0.952382
Accuracy	0.961928	0.963849
Precision	0.654867	0.755556
Recall	0.213256	0.226667
f1-score	0.321739	0.348718



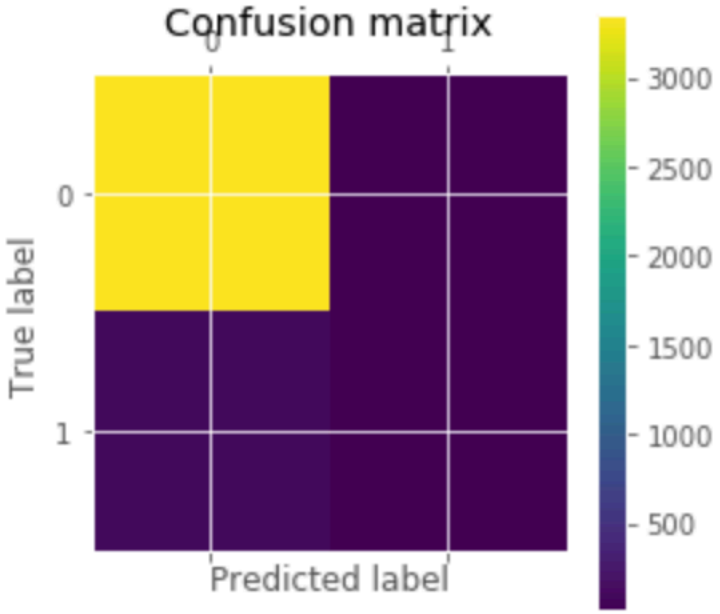
Train Set

```
[[7809   39]
 [ 273   74]]
```



Test Set

```
[[3352   11]
 [ 116   34]]
```

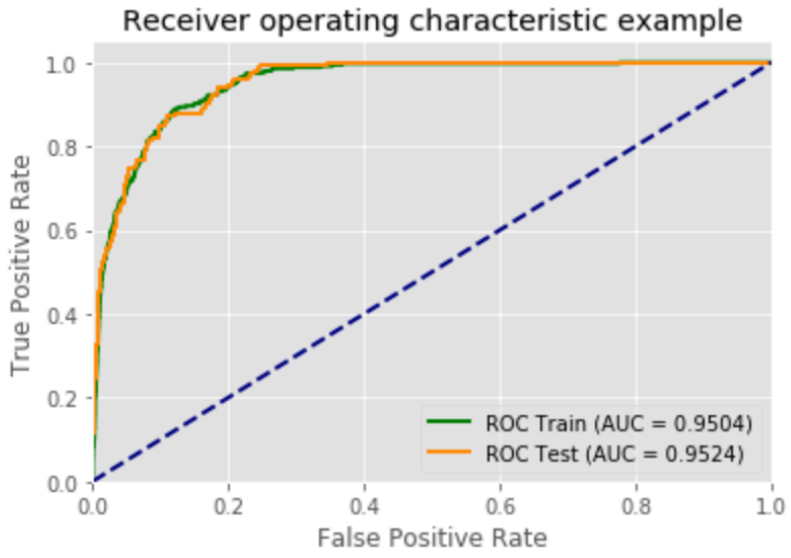


- recall=tp/(tp+fn): for this problem, sign up rate is low, thus tp and fp it is expected to be low

Machine Learning Methods

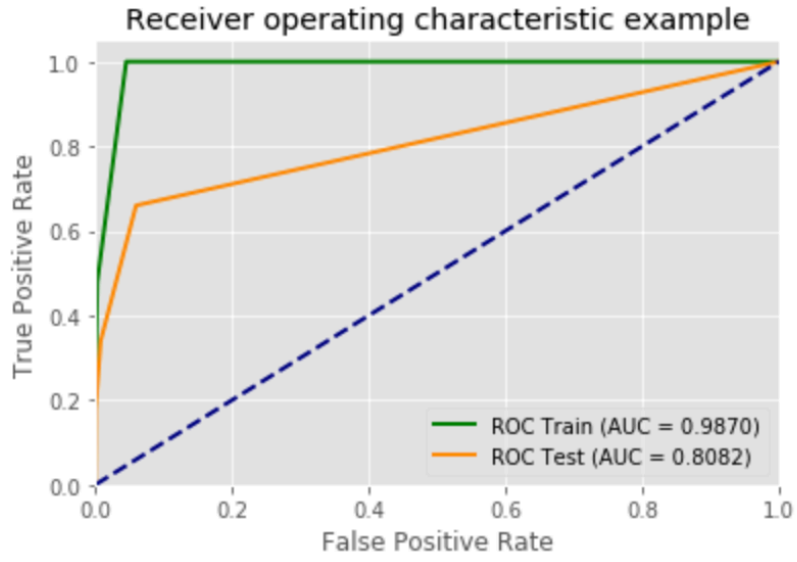
Logistic Regression

metrics	train	test
AUC	0.950415	0.952382
Accuracy	0.961928	0.963849
Precision	0.654867	0.755556
Recall	0.213256	0.226667
f1-score	0.321739	0.348718



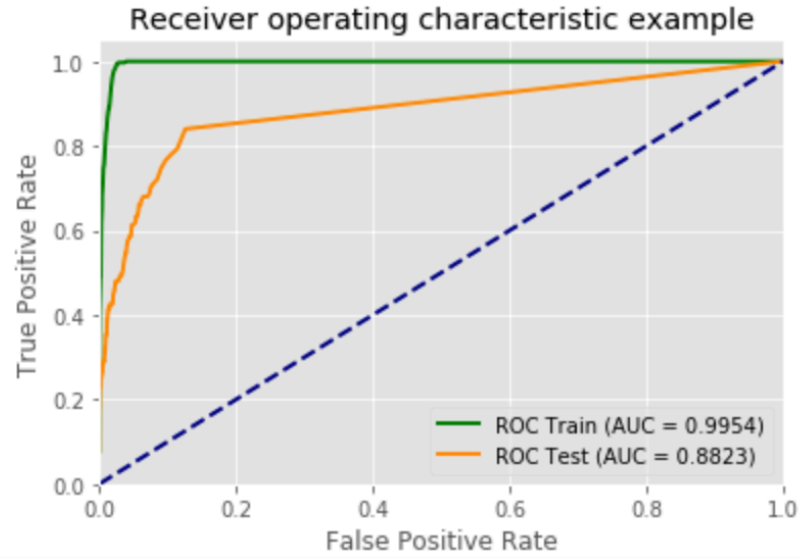
Single KNN

metrics	train	test
AUC	0.987025	0.808249
Accuracy	0.975107	0.964133
Precision	0.878307	0.653846
Recall	0.478386	0.340000
f1-score	0.619403	0.447368



Bagged KNN

metrics	train	test
AUC	0.995417	0.882304
Accuracy	0.976083	0.963279
Precision	0.931429	0.656716
Recall	0.469741	0.293333
f1-score	0.624521	0.405530

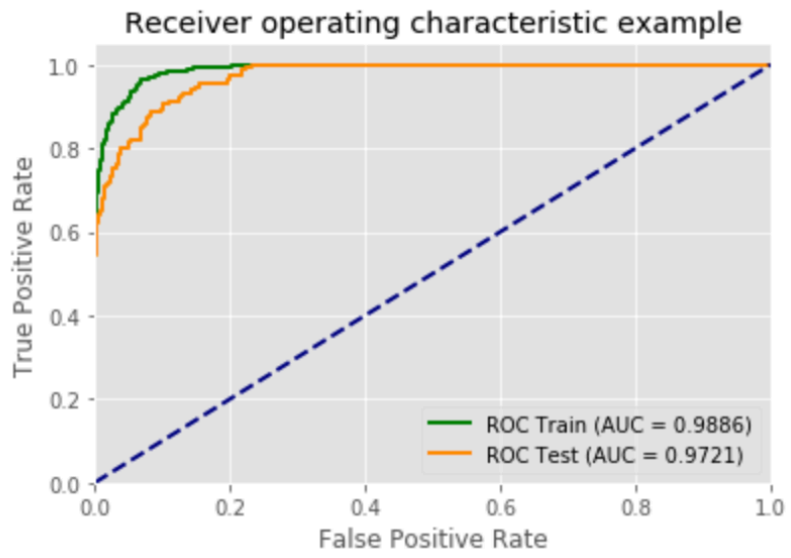


- KNN has improved recall compared to lR, but still bad performance

Machine Learning Methods

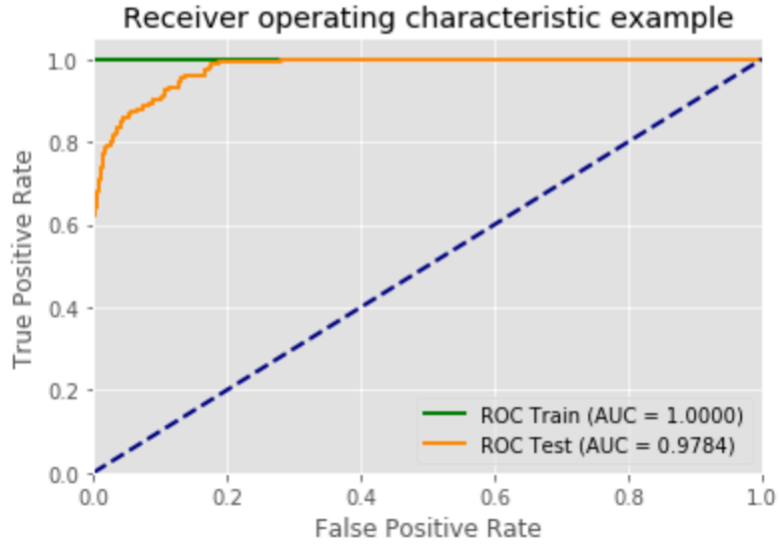
Random Forest

metrics	train	test
AUC	0.988645	0.972051
Accuracy	0.969494	0.968118
Precision	1.000000	1.000000
Recall	0.279539	0.253333
f1-score	0.436937	0.404255



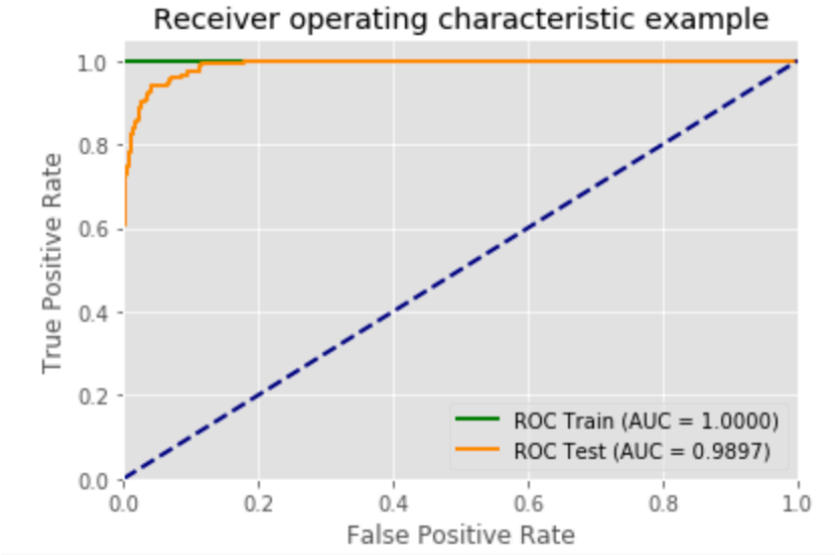
Random Forest Grid Search

metrics	train	test
AUC	0.999963	0.978440
Accuracy	0.986333	0.976373
Precision	1.000000	1.000000
Recall	0.677233	0.446667
f1-score	0.807560	0.617512



Gradient Boosting Tree

metrics	train	test
AUC	0.999997	0.989664
Accuracy	0.999268	0.984059
Precision	1.000000	0.879032
Recall	0.982709	0.726667
f1-score	0.991279	0.795620



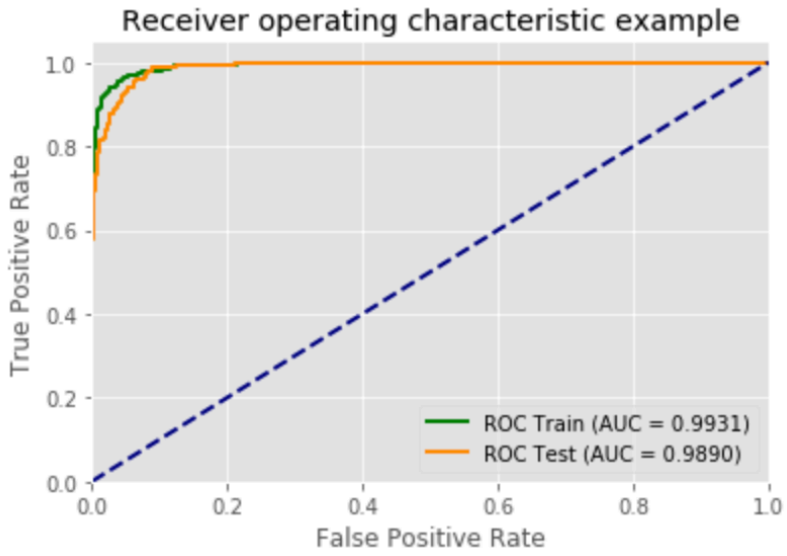
- Random forest need grid search to have better performance
- Gradient boosting tree has the best performance
- As expected, visit frequency, page stay time and time to register are important features

	feature	importance
31	log_Ttoend	0.263551
3	log_st	0.193434
28	log_npage	0.172168
0	count	0.153803

Machine Learning Methods

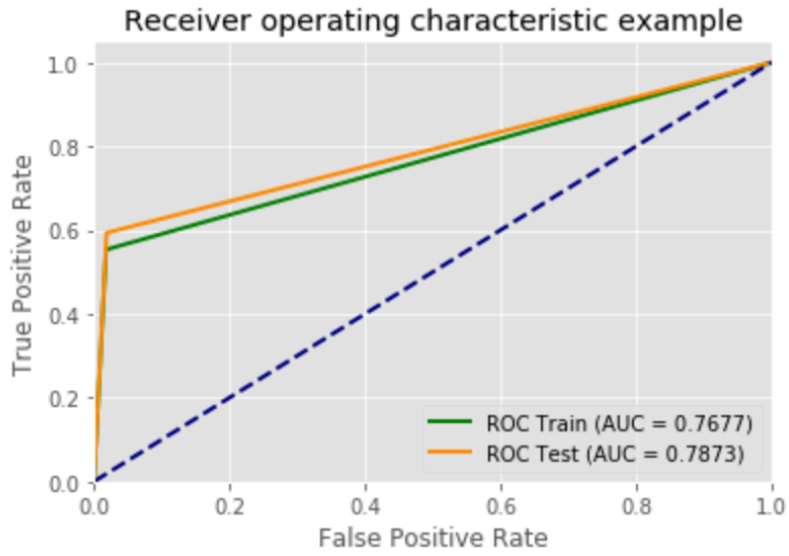
Neutral Network

metrics	train	test
AUC	0.993082	0.989034
Accuracy	0.986943	0.983205
Precision	0.931655	0.902655
Recall	0.746398	0.680000
f1-score	0.828800	0.775665



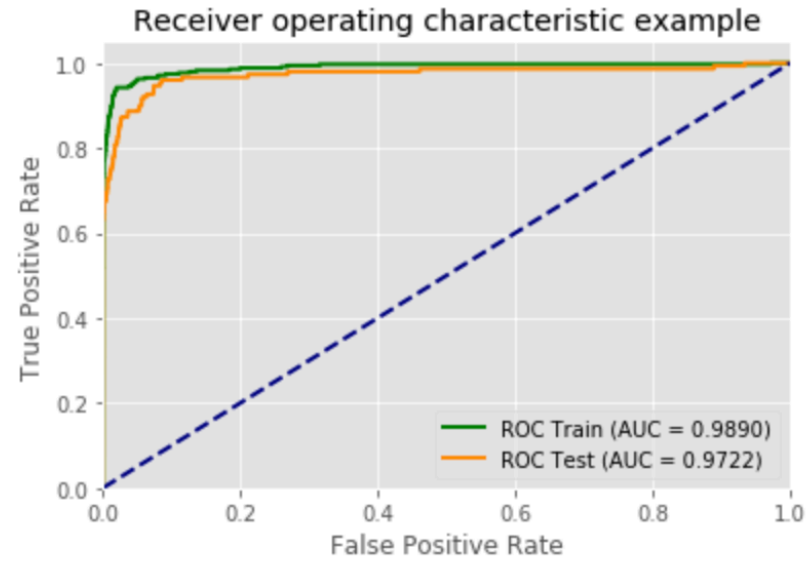
Linear SVM

metrics	train	test
AUC	0.767674	0.787300
Accuracy	0.963880	0.964703
Precision	0.576577	0.585526
Recall	0.553314	0.593333
f1-score	0.564706	0.589404



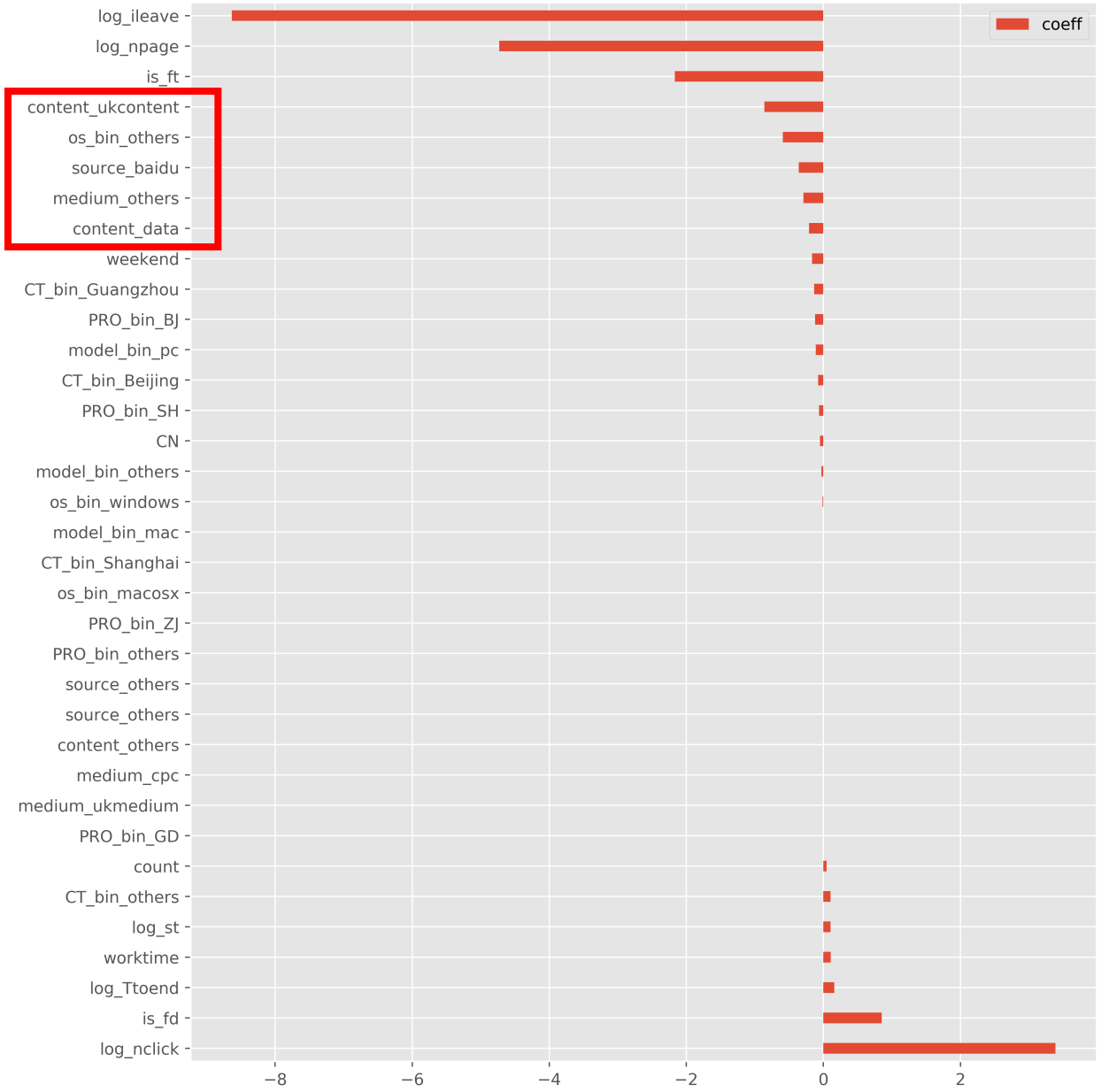
Nonlinear SVM

metrics	train	test
AUC	0.988964	0.972228
Accuracy	0.988041	0.981782
Precision	0.952727	0.870690
Recall	0.755043	0.673333
f1-score	0.842444	0.759398



- SVM is also a simple model, close train/test performance, but much better performance than IG
- Neutral network also has great performance and close train/test performance, but might take time if more layers are employed

Logistic Regression feature Coefficient



- Feature coefficient after L1 regularization
- Negative index leave/Negative page view: users trying to find other pages to check demo or more content without cellphone registration? Or users did not understand how to register
- Bottom click reflect users' interest to the website
- Highly interested users will come to register the other day or another time
- Medium or campaign have no positive or even negative effects

Summary

- According to Funnel Analysis: **page quality** and **cellphone privacy concern** might be key factors that bottleneck sign up rate
- Product promotion or strategic campaign have no significant effect
- Suggestions on sign up rate improvement:

✓Provide one or two simple free registration demo to attract new registration

✓Hire Web UX designer

✓Invest on media promotion and marketing campaign

- Top choice of models: Non-linear SVM, Gradient Boosting tree, Neutral Networks

Next Step

- Work on missing data imputation, may have a better interpretation on utm related features
- Which page, and on what aspect the interface need to be improved, for example: work on content targeted ads
- Detailed analysis on medium effect, campaign effect, and find out the things need to be invest on

- Questions?