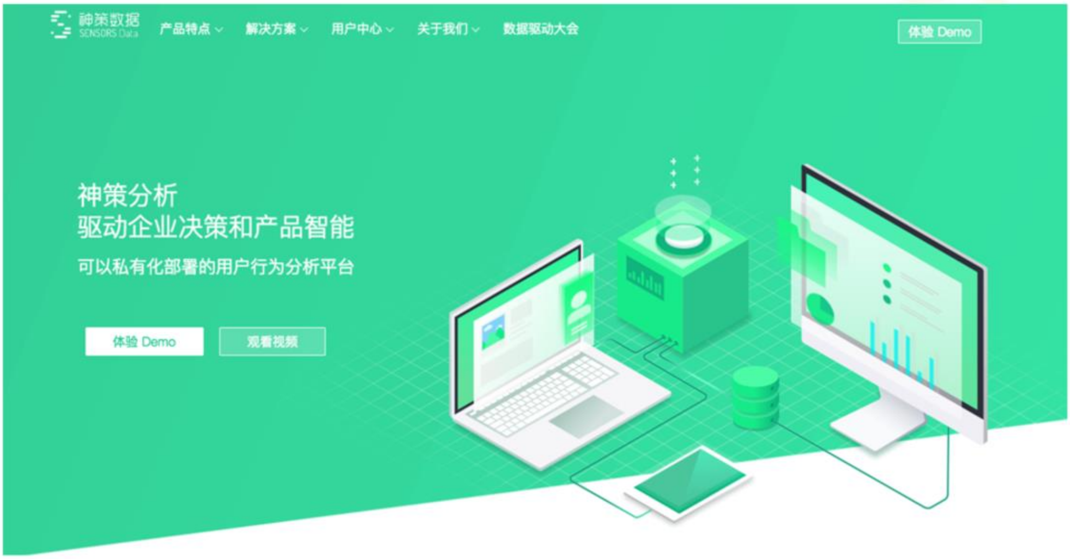
Sensors Data
User Behavior Analysis

Ella

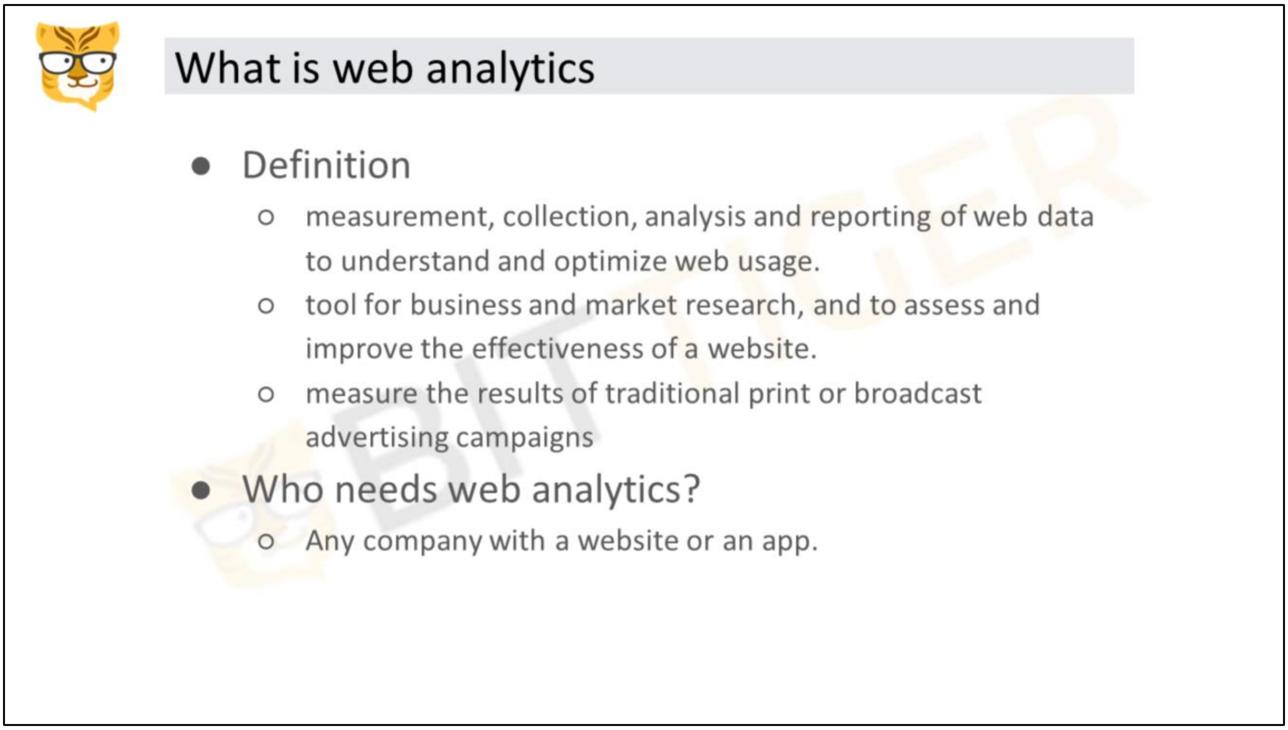
What is sensors data



神策分析
驱动企业决策和产品智能

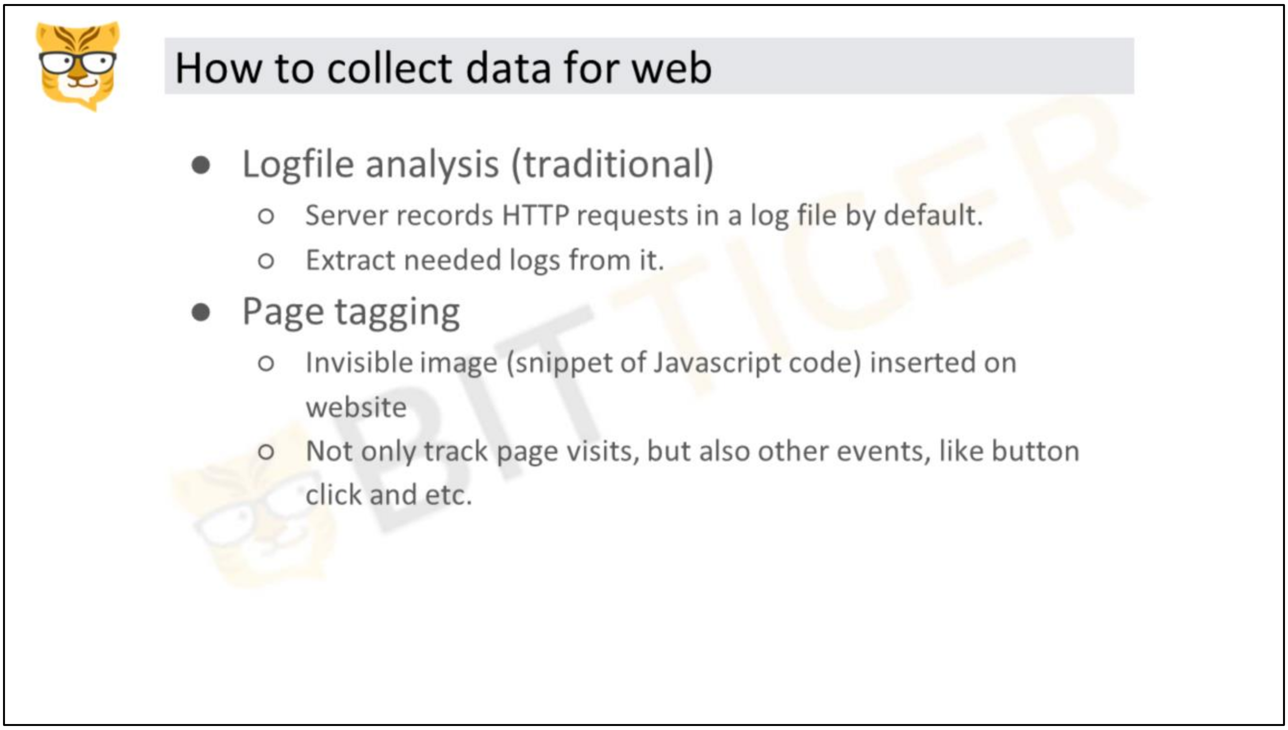
可以私有化部署的用户行为分析平台

https://www.sensorsdata.cn/



What is web analytics

- Definition
  - measurement, collection, analysis and reporting of web data to understand and optimize web usage.
  - tool for business and market research, and to assess and improve the effectiveness of a website.
  - measure the results of traditional print or broadcast advertising campaigns
- Who needs web analytics?
  - Any company with a website or an app.

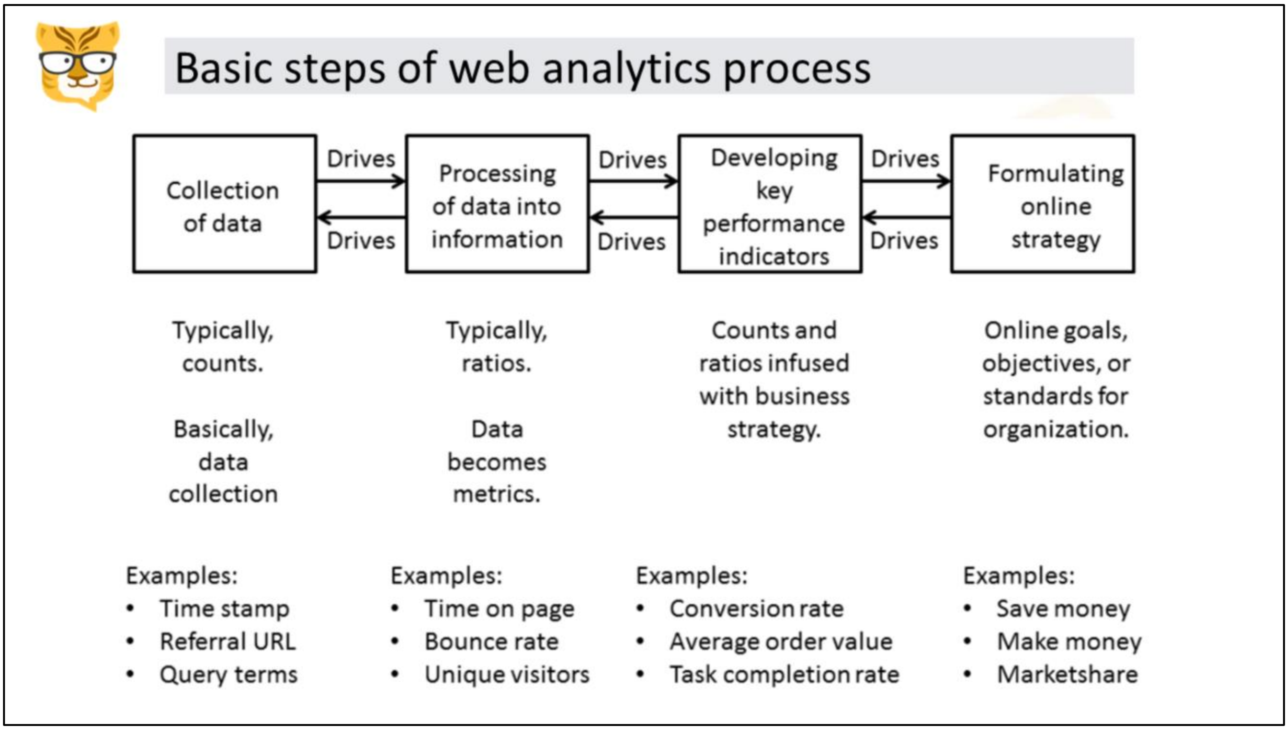


How to collect data for web

- Logfile analysis (traditional)
  - Server records HTTP requests in a log file by default.
  - Extract needed logs from it.
- Page tagging
  - Invisible image (snippet of Javascript code) inserted on website
  - Not only track page visits, but also other events, like button click and etc.

https://en.wikipedia.org/wiki/Web_analytics#Logfile_analysis_vs_page_tagging

Basic steps of web analytics process


Collection of data
Drives
Drives
Processing of data into information
Drives
Drives
Developing key performance indicators
Drives
Drives
Formulating online strategy

Typically, counts.

Basically, data collection

Examples:
- Time stamp
- Referral URL
- Query terms

Typically, ratios.

Data becomes metrics.

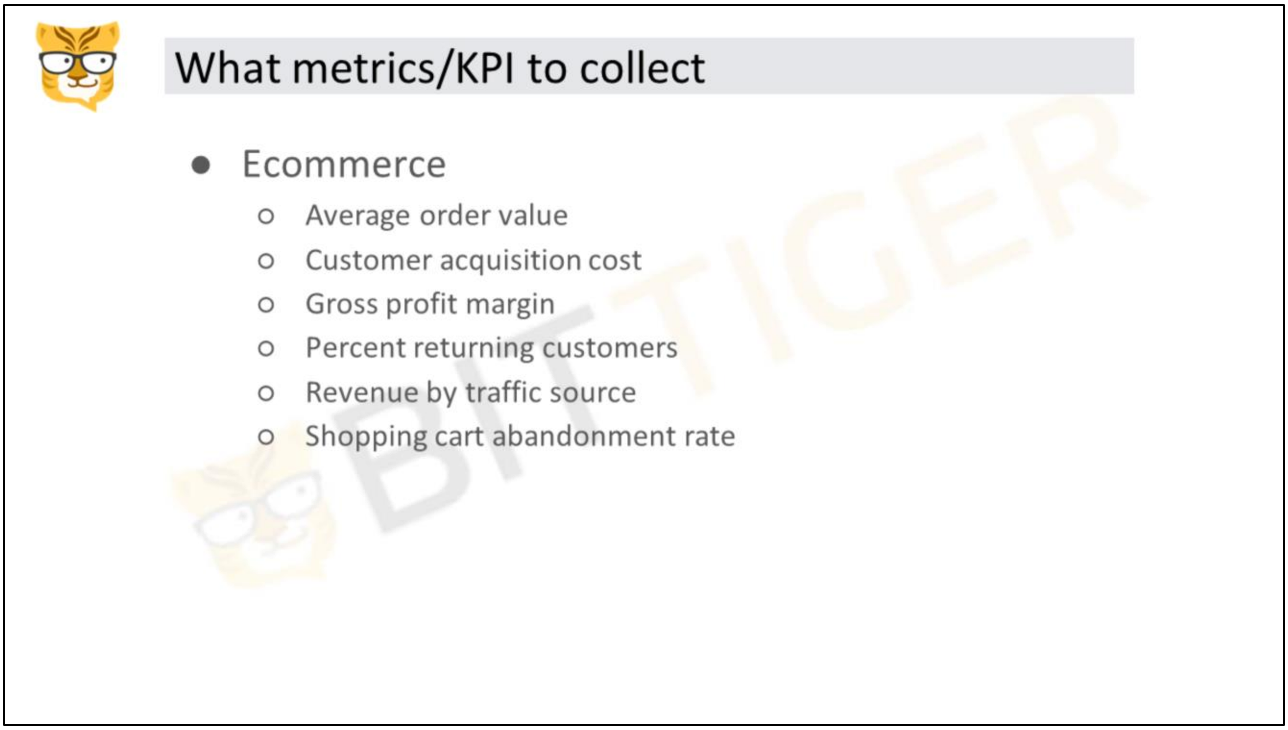
Examples:
- Time on page
- Bounce rate
- Unique visitors

Counts and ratios infused with business strategy.

Examples:
- Conversion rate
- Average order value
- Task completion rate

Online goals, objectives, or standards for organization.

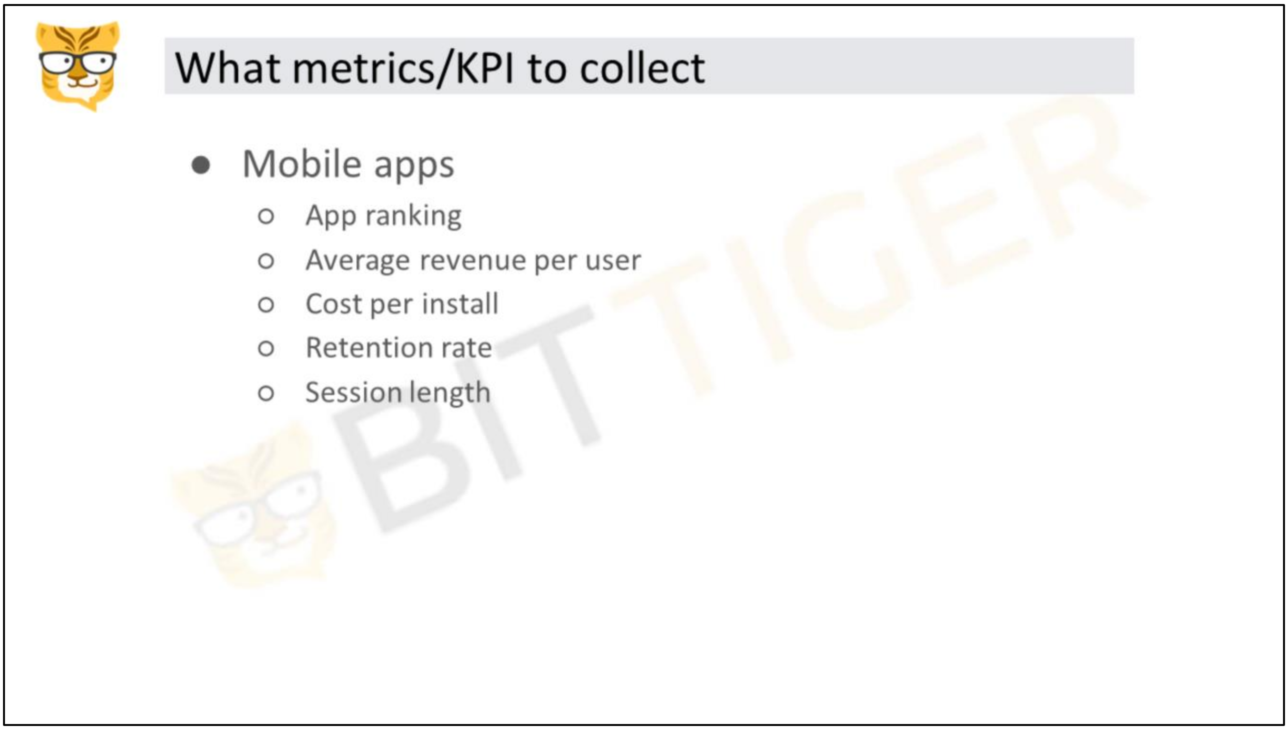
Examples:
- Save money
- Make money
- Marketshare



What metrics/KPI to collect

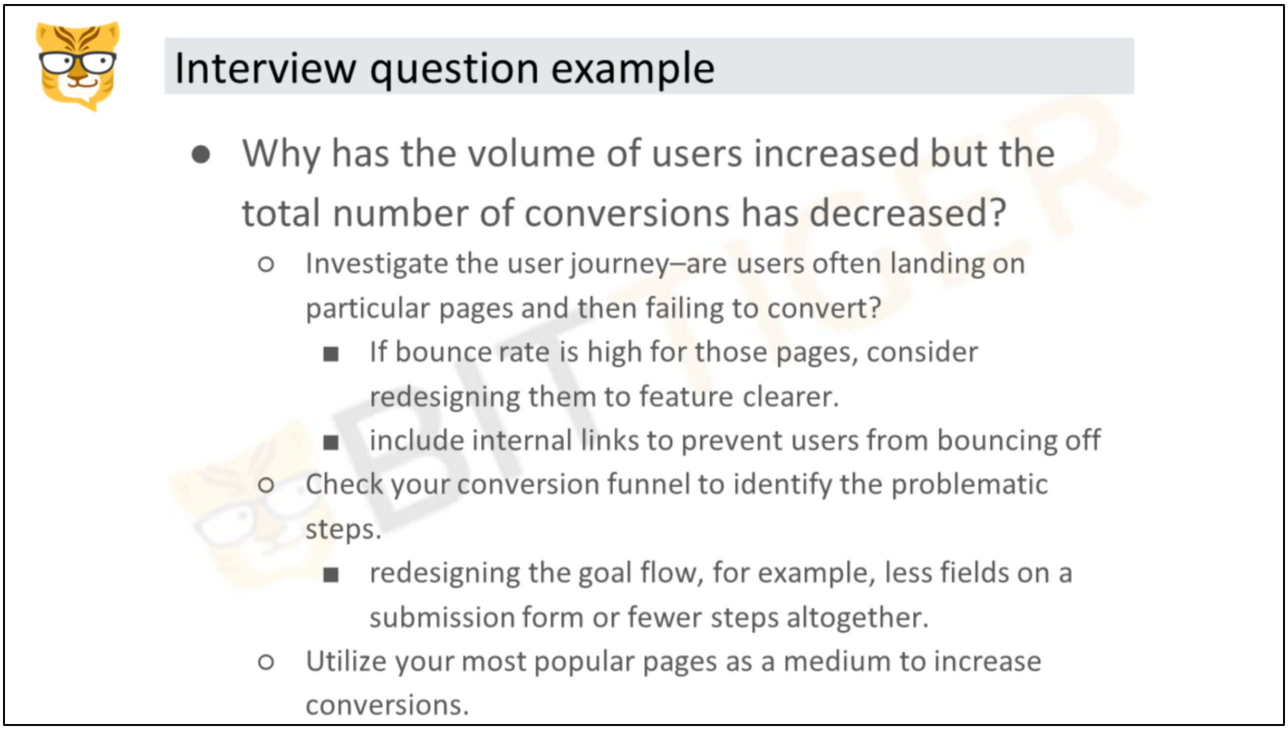
- Ecommerce
  - ○ Average order value
  - ○ Customer acquisition cost
  - ○ Gross profit margin
  - ○ Percent returning customers
  - ○ Revenue by traffic source
  - ○ Shopping cart abandonment rate

List of metrics https://www.geckoboard.com/learn/kpi-examples/#.WrgWZ5PwY1J



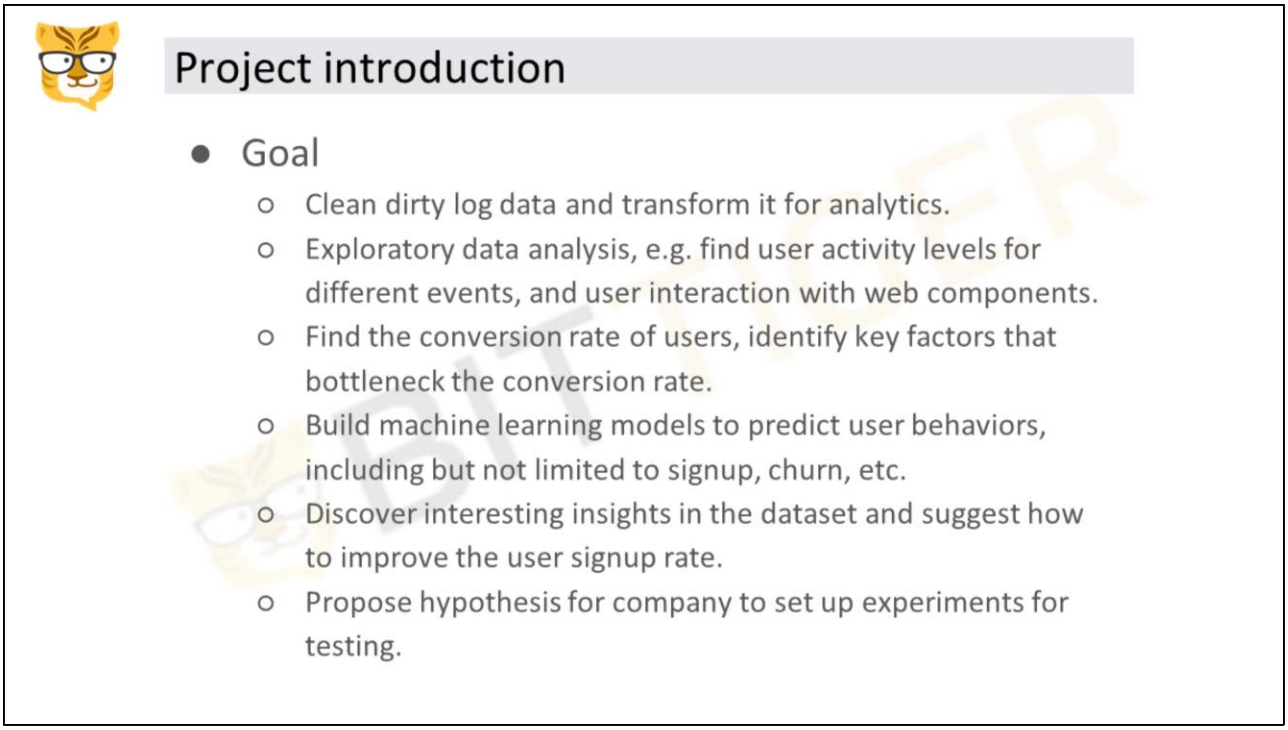
What metrics/KPI to collect

- Mobile apps
  - App ranking
  - Average revenue per user
  - Cost per install
  - Retention rate
  - Session length



Interview question example

- Why has the volume of users increased but the total number of conversions has decreased?
  - Investigate the user journey–are users often landing on particular pages and then failing to convert?
    - If bounce rate is high for those pages, consider redesigning them to feature clearer.
    - include internal links to prevent users from bouncing off
  - Check your conversion funnel to identify the problematic steps.
    - redesigning the goal flow, for example, less fields on a submission form or fewer steps altogether.
  - Utilize your most popular pages as a medium to increase conversions.

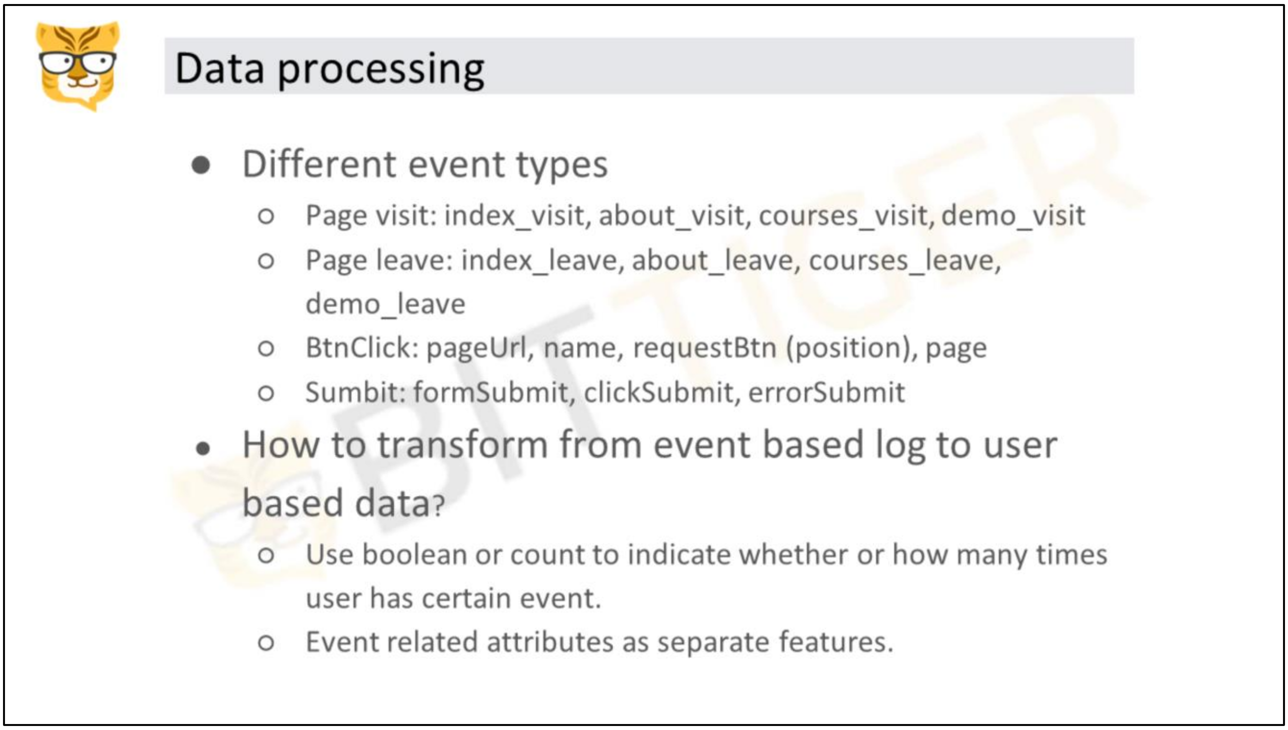


Project introduction

- Goal
  - Clean dirty log data and transform it for analytics.
  - Exploratory data analysis, e.g. find user activity levels for different events, and user interaction with web components.
  - Find the conversion rate of users, identify key factors that bottleneck the conversion rate.
  - Build machine learning models to predict user behaviors, including but not limited to signup, churn, etc.
  - Discover interesting insights in the dataset and suggest how to improve the user signup rate.
  - Propose hypothesis for company to set up experiments for testing.

Data example

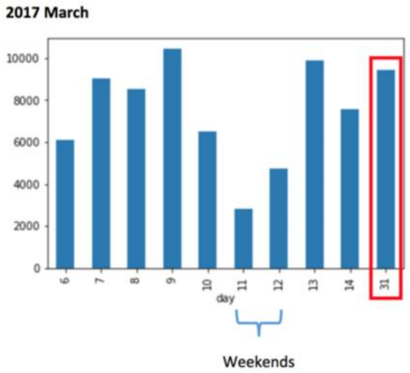
```
{“distinct_id":"595466e9a8e733434ce08de16e927d985e0b5d48",
“lib":{"$lib":"js","$lib_method":"code","$lib_version":"1.6.20"},
“properties”:{"$os":"windows","$model":"pc","$os_version":"6.1","$screen_height":800,"$
screen_width":1280,
“$lib":"js","$lib_version":"1.6.20","$browser":"chrome","$browser_version":"56","
$latest_referrer”:”","$latest_referrer_host":"","$latest_utm_source":"baidu","$latest_utm
_medium":"cpc","$latest_utm_campaign":"通用词","$latest_utm_content":"通用-用户画
像","$latest_utm_term":"用户画像
","_latest_ch":"demo","_session_referrer":"https://www.baidu.com/baidu.php","_session_
referrer_host":"www.baidu.com",
“session_page_url":"https://www.sensorsdata.cn/?utm_source=baidu&utm_medium=cpc&
utm_term=%E7%94%A&utm_content=%E9%80%9A%E7%&utm_campaign=%E9%80",
“pageUrl":"https://sensorsdata.cn/?ch=demo","pageStayTime":5.692,"pagePosition":2,"$is
_first_day":true,"$is_first_time":false,"$ip":"219.135.131.99"},
“type":"track","event":"index_leave","_nocache":"0654392402996","time":148879104795
3}
```



Data processing

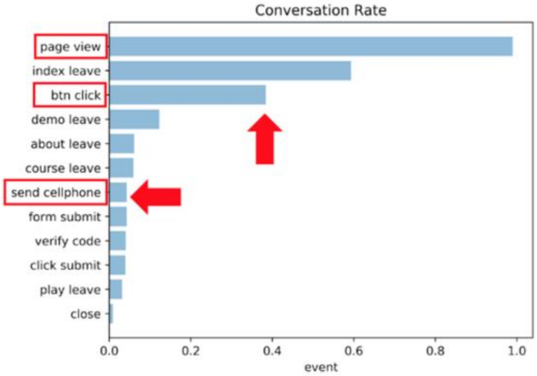
- Different event types
  - Page visit: index_visit, about_visit, courses_visit, demo_visit
  - Page leave: index_leave, about_leave, courses_leave, demo_leave
  - BtnClick: pageUrl, name, requestBtn (position), page
  - Sumbit: formSubmit, clickSubmit, errorSubmit
- How to transform from event based log to user based data?
  - Use boolean or count to indicate whether or how many times user has certain event.
  - Event related attributes as separate features.

Data exploration



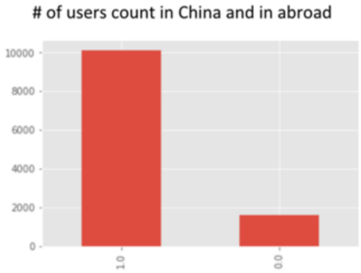
- User activity by day of week
  - Data set spans for 9 days (over a week)
  - Weekends activity drop significantly: most users view this website due to work requirements/ interests
  - Introduce a weekend or not feature
  - Introduce a work time or not feature (8AM to 5PM in Beijing time zone)
  - 31st is an isolated day, probably contains wrong data, need to be excluded

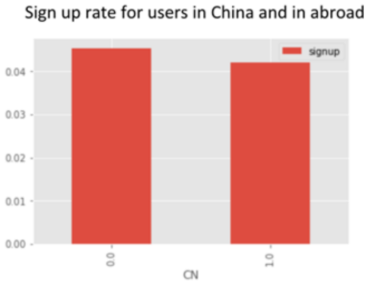
Funnel analysis



- Drop from page view to button click
  - Most users do not have the interest to click on pages
  - Improve page quality?
- Another sharp drop
  - From button click to send cell phone verification code
  - Some interested users do not want to register with cell phone number
  - Privacy concerns?
  - Do not have cellphone number from mainland China?

Compare user behavior in/out China



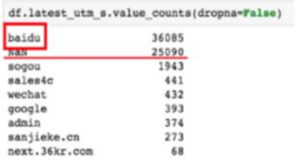


- Use IP address to identify
  - Users in abroad still have high interest in sign up with cell phone verification

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

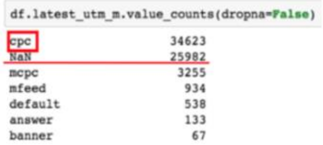
Source



```
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

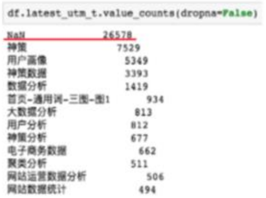


```
df.latest_utm_m.value_counts(dropna=False)
```

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

cpc: cost per click

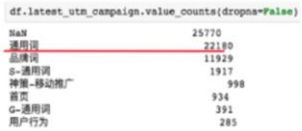
Value



```
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
```

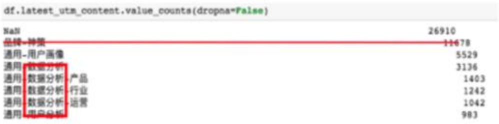
Campaign



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

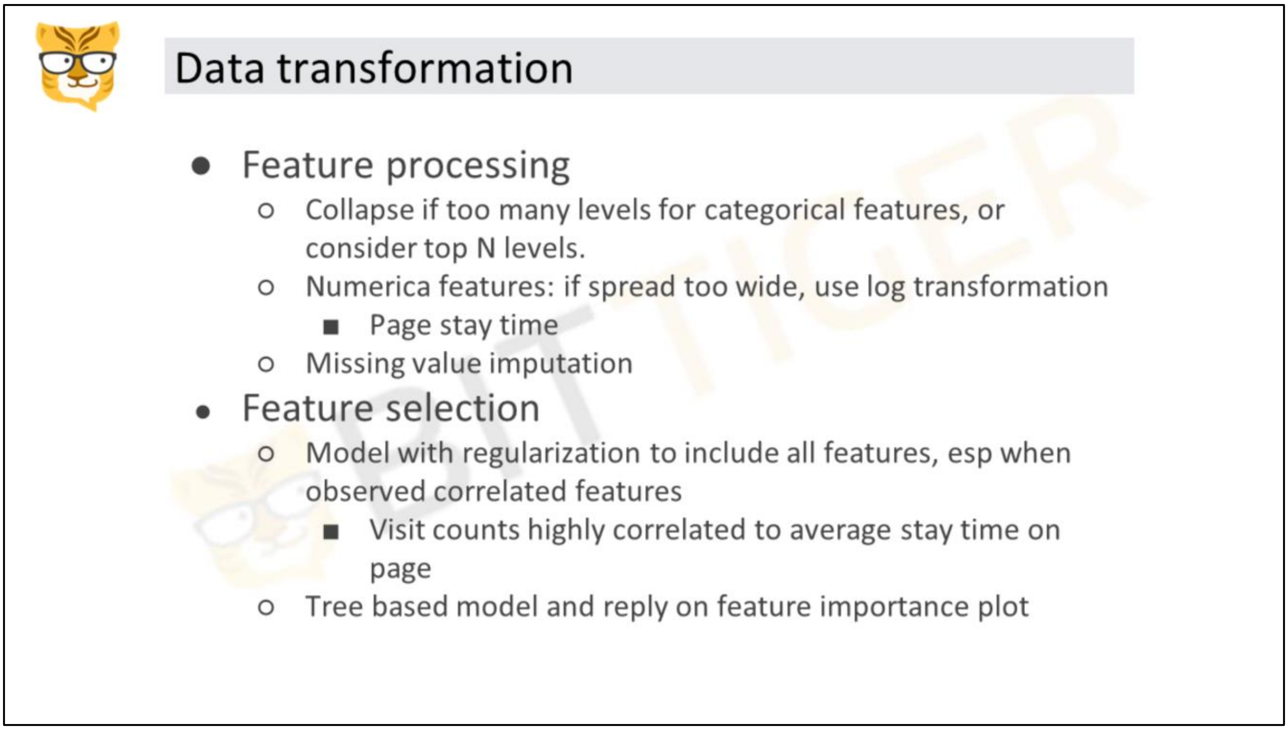
```
NaN          25770
通用词         22180
品牌词         11929
S-通用词        1917
神策-移动推广       998
首页            934
G-通用词         391
用户行为          285
```

Content



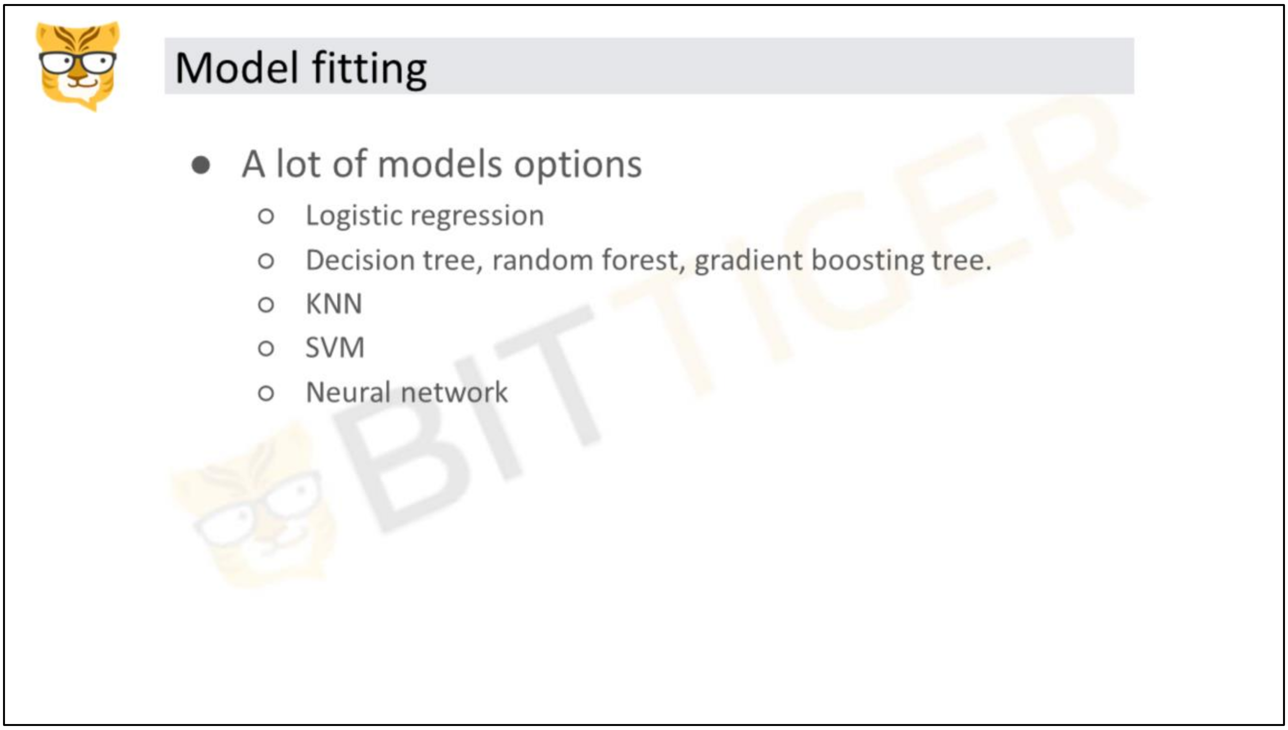
```
df.latest_utm_content.value_counts(dropna=False)
```

```
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
```



Data transformation

- Feature processing
  - Collapse if too many levels for categorical features, or consider top N levels.
  - Numerica features: if spread too wide, use log transformation
    - Page stay time
  - Missing value imputation
- Feature selection
  - Model with regularization to include all features, esp when observed correlated features
    - Visit counts highly correlated to average stay time on page
  - Tree based model and reply on feature importance plot

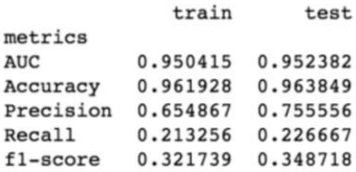


Model fitting

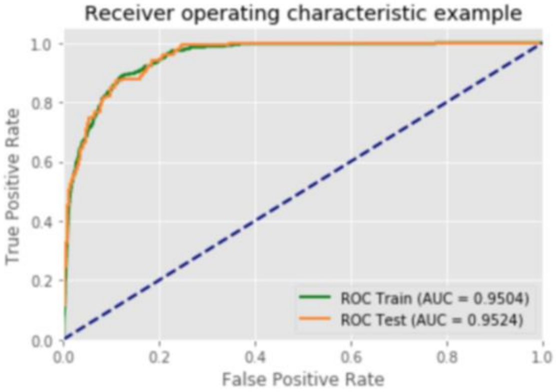
- A lot of models options
  - Logistic regression
  - Decision tree, random forest, gradient boosting tree.
  - KNN
  - SVM
  - Neural network

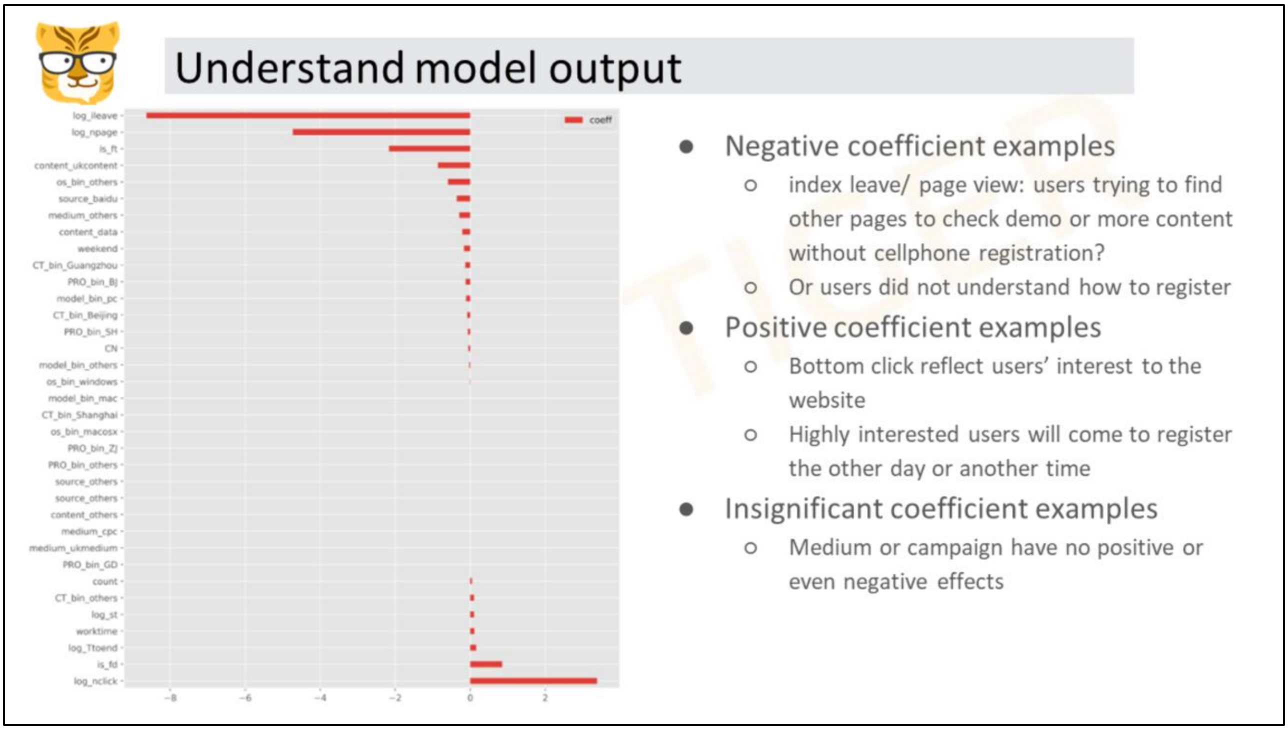
Logistic regression

- How model performs
  - AUC, Precision, recall, F1 score
- Model comparison
  - Why outperform/ underperform



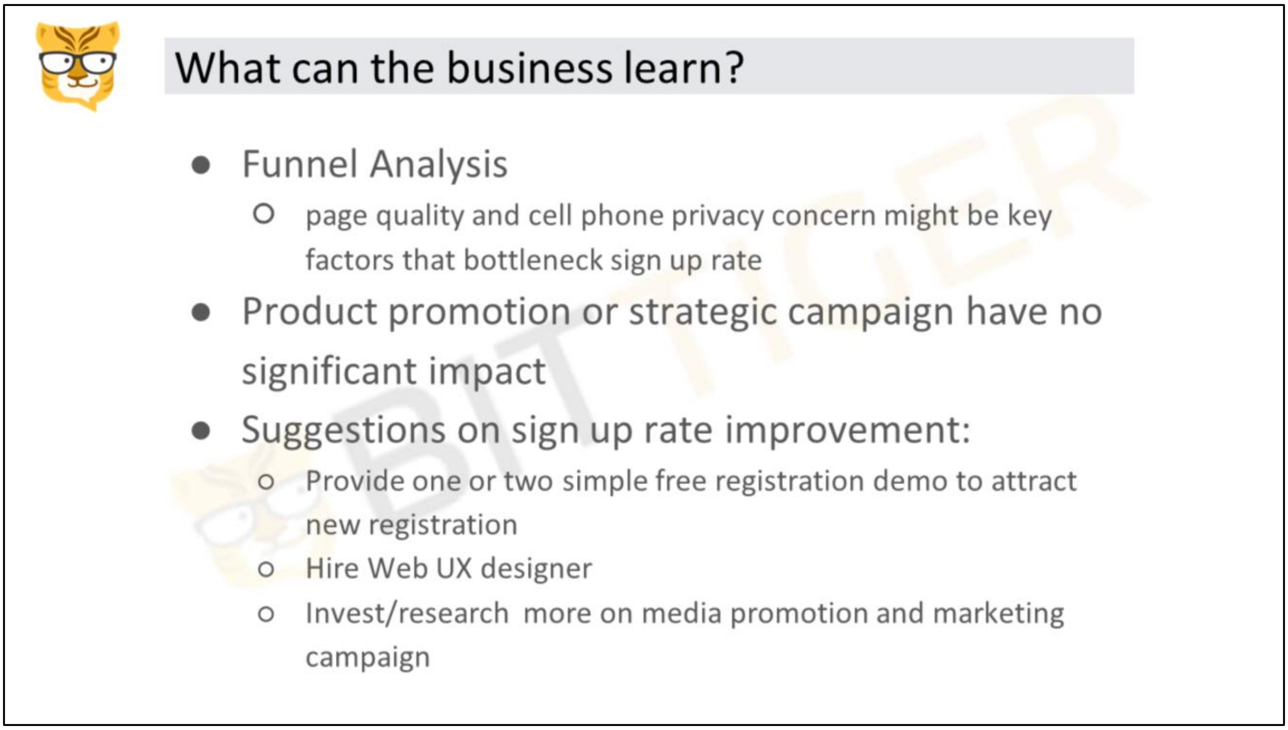
```
                train      test
metrics
AUC          0.950415  0.952382
Accuracy     0.961928  0.963849
Precision    0.654867  0.755556
Recall       0.213256  0.226667
f1-score     0.321739  0.348718
```





Understand model output

- Negative coefficient examples
  - index leave/ 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
- Positive coefficient examples
  - Bottom click reflect users' interest to the website
  - Highly interested users will come to register the other day or another time
- Insignificant coefficient examples
  - Medium or campaign have no positive or even negative effects



What can the business learn?

- Funnel Analysis
  - page quality and cell phone privacy concern might be key factors that bottleneck sign up rate
- Product promotion or strategic campaign have no significant impact
- Suggestions on sign up rate improvement:
  - Provide one or two simple free registration demo to attract new registration
  - Hire Web UX designer
  - Invest/research more on media promotion and marketing campaign