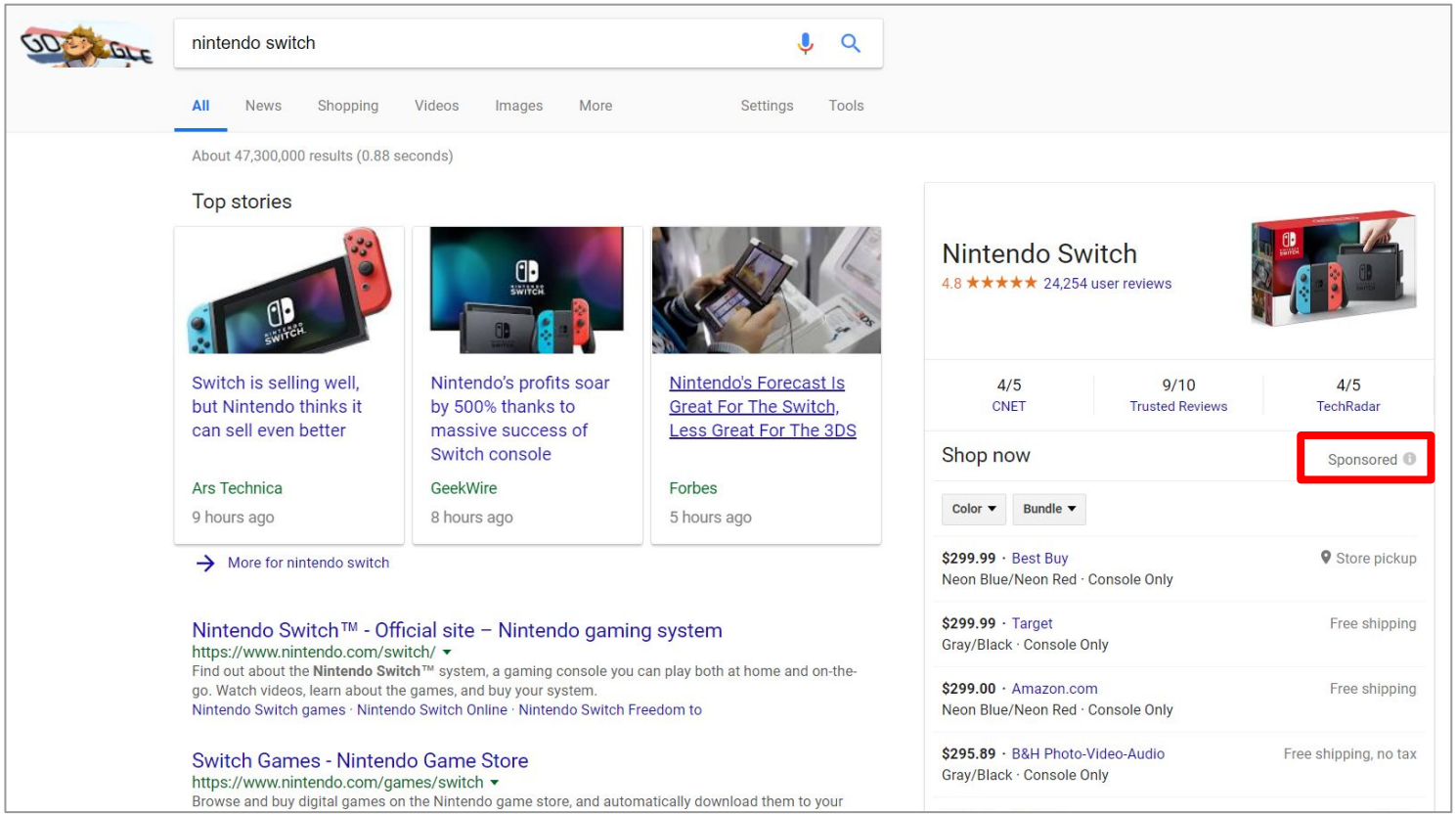

# CVR Prediction in Sponsored Search using Logistic Regression

Jiachao Fang, Yuchong Xiao

# Background




Sponsored Search

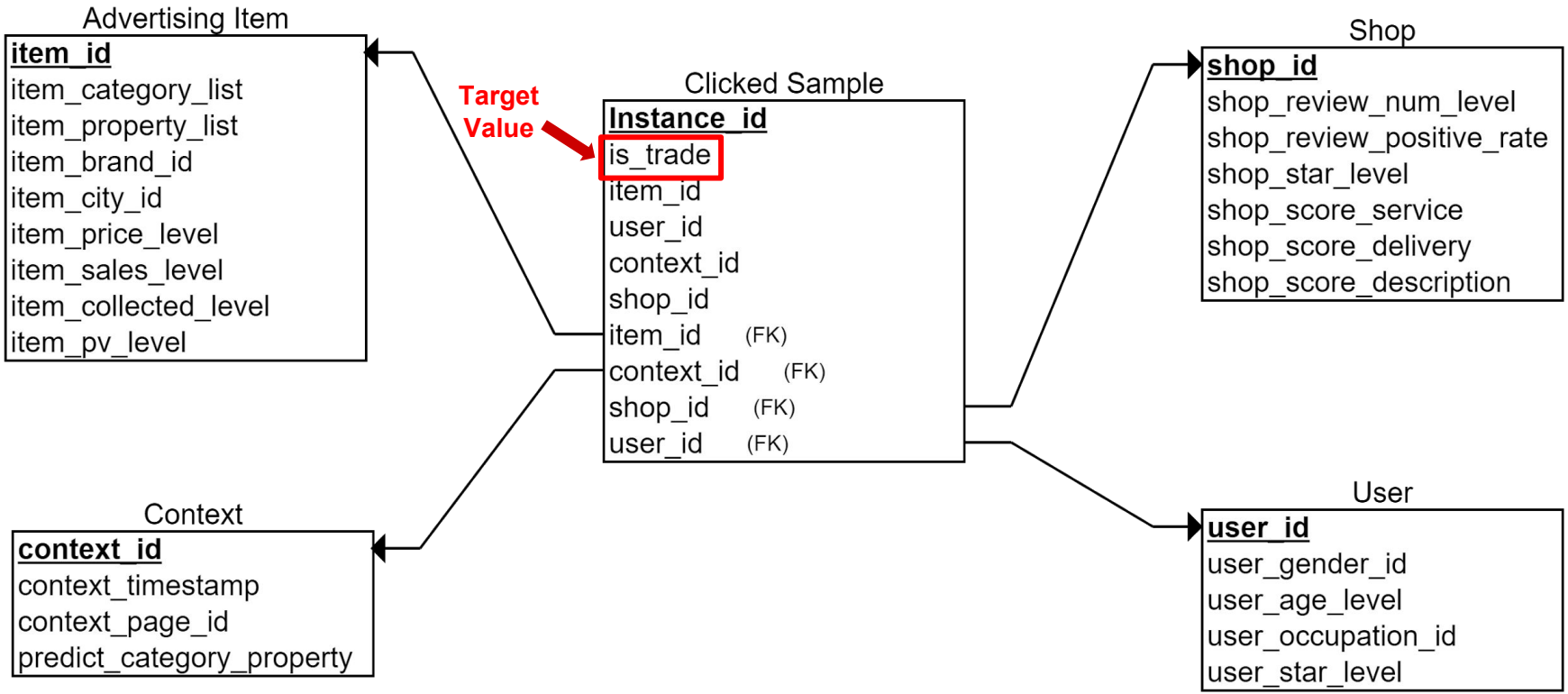
↓

Click-through rate

↓

Conversion Rate

# Dataset


Advertising Item
item_id
item_category_list
item_property_list
item_brand_id
item_city_id
item_price_level
item_sales_level
item_collected_level
item_pv_level
Target Value
Clicked Sample
Instance_id
is_trade
item_id
user_id
context_id
shop_id
item_id (FK)
context_id (FK)
shop_id (FK)
user_id (FK)
Shop
shop_id
shop_review_num_level
shop_review_positive_rate
shop_star_level
shop_score_service
shop_score_delivery
shop_score_description
Context
context_id
context_timestamp
context_page_id
predict_category_property
User
user_id
user_gender_id
user_age_level
user_occupation_id
user_star_level

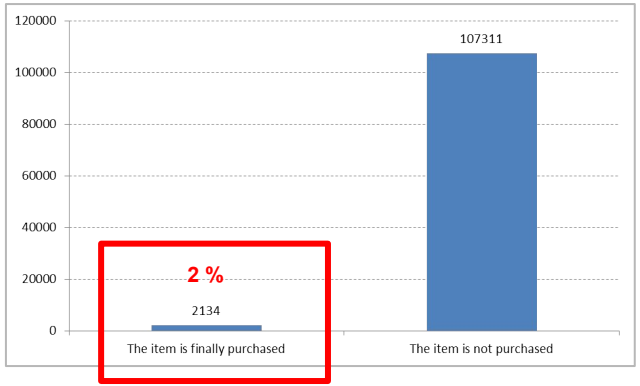
# Data Preparation

478,138
Transaction Data

Raw Data

80% (368,693 records)
with any missing values

Delete
Missing Data




Total: 109,445

Data Processing

Numerical
Variables

Normalization

Variables
(Range 0-1)

Standardization

Standardized
Variables

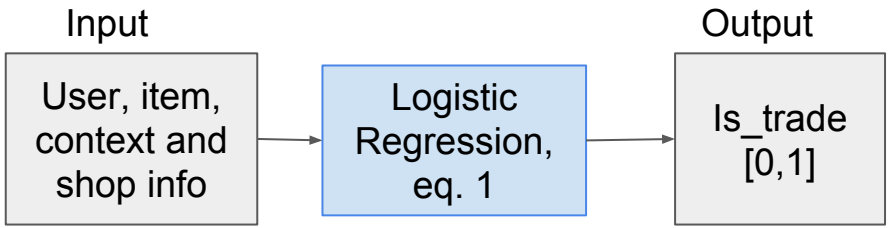
Categorical
Variables

One Hot
Encoder

Dummy
Variables

Others

# Experiment Design




$$Ln[Y/(1-Y)] = \beta_0+\beta_1\times X1+\beta_2\times X2+\beta_3\times X3... \quad (1)$$

Evaluation: Cross-entropy Loss, eq. 2, 3, 4, 5

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} Cost(h_\theta(x^i), y^{(i)}) \quad (2)$$

$$Cost(h_\theta(x), y) = -log(h_\theta(x)) if y = 1 \quad (3)$$

$$Cost(h_\theta(x), y) = -log(1 - h_\theta(x)) if y = 0 \quad (4)$$

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}log(h_\theta(x^{(i)}))+(1-y^{(i)})log(1-h_\theta(x^i))] \quad (5)$$

```python
from sklearn.model_selection import train_test_split
from sklearn.metrics import log_loss
from sklearn.linear_model import LogisticRegression

def model_log_loss(model):
    X = train[select_cols]
    Y = train['is_trade']
    X_train, X_test, y_train, y_test = train_test_split
    (X, Y, test_size=0.3, random_state=0)

    print("Training...")
    model.fit(X_train, y_train)
    print("Predicting...")
    y_prediction = model.predict_proba(X_test)
    test_pred = y_prediction[:, 1]
    print('log_loss ', log_loss(y_test, test_pred))

def result(model):
    X = train[select_cols]
    Y = train['is_trade']
    model.fit(X, Y)
    y_pred = model.predict_proba(test[select_cols])[:, 1]
    result = pd.DataFrame({'instance_id':test['instance_id'],
                           'predicted_score':y_pred})
    result.to_csv('result.txt', sep=" ", index=False)

if __name__ == "__main__":
    result(LogisticRegression(C=100, n_jobs=-1))
    model_log_loss(LogisticRegression(C=100, n_jobs=-1))
```

# Result

| Testing Global Null Hypothesis: BETA=0 | | | |
|---|---|---|---|
| Test | Chi-Square | DF | Pr > ChiSq |
| Likelihood Ratio | 2772.7245 | 16 | <.0001 |
| Score | 2860.6242 | 16 | <.0001 |
| Wald | 2821.1625 | 16 | <.0001 |

The model as a whole fits significantly better than an empty model.

| Type 3 Analysis of Effects | | | |
|---|---|---|---|
| Effect | DF | Wald Chi-Square | Pr > ChiSq |
| item_price_level | 1 | 617.7158 | <.0001 |
| item_sales_level | 1 | 858.9710 | <.0001 |
| item_collected_level | 1 | 149.7540 | <.0001 |
| item_pv_level | 1 | 271.1387 | <.0001 |
| user_gender_id | 3 | 32.9101 | <.0001 |
| user_age_level | 1 | 47.5474 | <.0001 |
| user_star_level | 1 | 47.2900 | <.0001 |
| context_page_id | 1 | 39.5969 | <.0001 |
| shop_review_num_leve | 1 | 10.0710 | 0.0015 |
| shop_review_positive | 1 | 0.0139 | 0.9062 |
| shop_star_level | 1 | 2.2379 | 0.1347 |
| shop_score_service | 1 | 6.1350 | 0.0133 |
| shop_score_delivery | 1 | 16.9826 | <.0001 |
| shop_score_descripti | 1 | 11.6373 | 0.0006 |

Statistically significant at the 5% significance level

# Result

| Analysis of Maximum Likelihood Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Parameter | | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | | 1 | -120.9 | 130.5 | 0.8580 | 0.3543 |
| item_price_level | | 1 | -0.2589 | 0.0104 | 617.7158 | <.0001 |
| item_sales_level | | 1 | 0.3009 | 0.0103 | 858.9710 | <.0001 |
| item_collected_level | | 1 | -0.1144 | 0.00935 | 149.7540 | <.0001 |
| item_pv_level | | 1 | -0.1283 | 0.00779 | 271.1387 | <.0001 |
| user_gender_id | 0 | 1 | -0.1147 | 0.0260 | 19.4051 | <.0001 |
| user_gender_id | 2 | 1 | -0.0855 | 0.0762 | 1.2591 | 0.2618 |
| user_gender_id | -1 | 1 | -0.3976 | 0.0855 | 21.6423 | <.0001 |
| user_age_level | | 1 | 0.0513 | 0.00744 | 47.5474 | <.0001 |
| user_star_level | | 1 | -0.0171 | 0.00248 | 47.2900 | <.0001 |
| context_page_id | | 1 | -0.0185 | 0.00294 | 39.5969 | <.0001 |
| shop_review_num_leve | | 1 | -0.0758 | 0.0239 | 10.0710 | 0.0015 |
| shop_review_positive | | 1 | -0.0643 | 0.5452 | 0.0139 | 0.9062 |
| shop_star_level | | 1 | 0.0389 | 0.0260 | 2.2379 | 0.1347 |
| shop_score_service | | 1 | 10.9562 | 4.4233 | 6.1350 | 0.0133 |
| shop_score_delivery | | 1 | -17.1848 | 4.1701 | 16.9826 | <.0001 |
| shop_score_descripti | | 1 | 5.1509 | 1.5099 | 11.6373 | 0.0006 |

Variables increasing odds:
shop_score_service,
shop_score_description,
item_sales_level,
user_age_level

Variables decrease odds:
Item_price_level,
Item_collected_level,
Item_pv_level,
User_star_level,
Context_page_id,
Shop_review_num_level,
Shop_score_delivery,

Categorical variable:
User_gender_id

# Conclusion


***Shop***
- Better service
- Better shop description
- Reasonable delivery speed

***User***
- Senior customers
- Male customers

***Item***
- Higher sales, more popular
- Cheaper

***Advertisement***
- Put on first few pages


THANK YOU