CTR Prediction

By: Deepak Kumar Sisodia

Problem

Build a prediction model to predict whether a mobile ad will be clicked

CTR (click-through rate) Usage:

- Online Advertising
- Ad performance evaluation

Business Use Case :

- Sponsored search
- Real-time bidding

Approach

- **S** - Sample
- **E** - Explore
- **M** - Modify
- **M** - Model
- **A** - Assess

SEMMA - Sample

Original Data Set :

Rows – Over 40 million

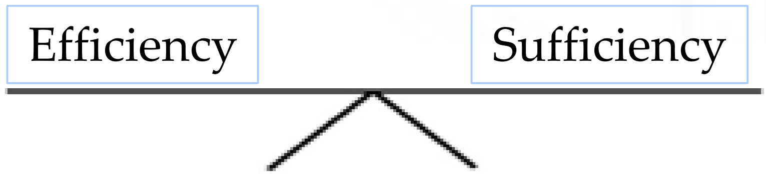
Columns – 24

Response Variable – Binary categorical

Class 0	Class 1
83%	17%

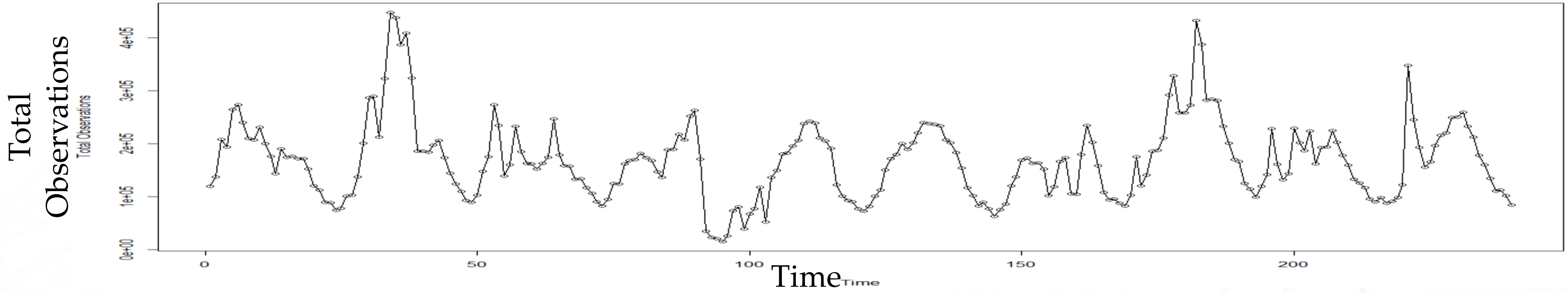
Data Set should be :

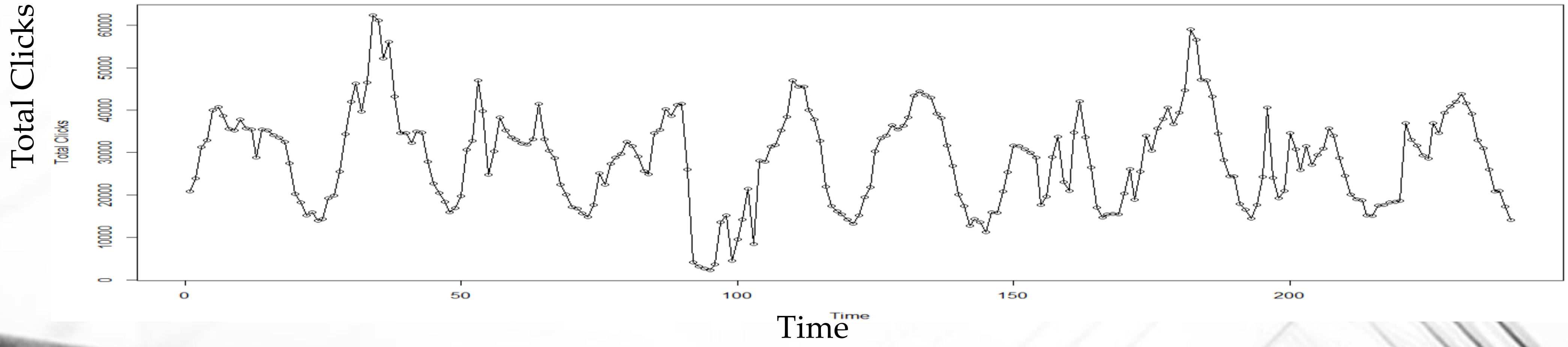
large enough – Sufficiency
small enough – Efficiency



SEMMA - Sample

10 days of data for each hour => Total 240 hours of data





SEMMA - Sample

On certain days at specific times, there are :

- More Observations
- More ad clicks

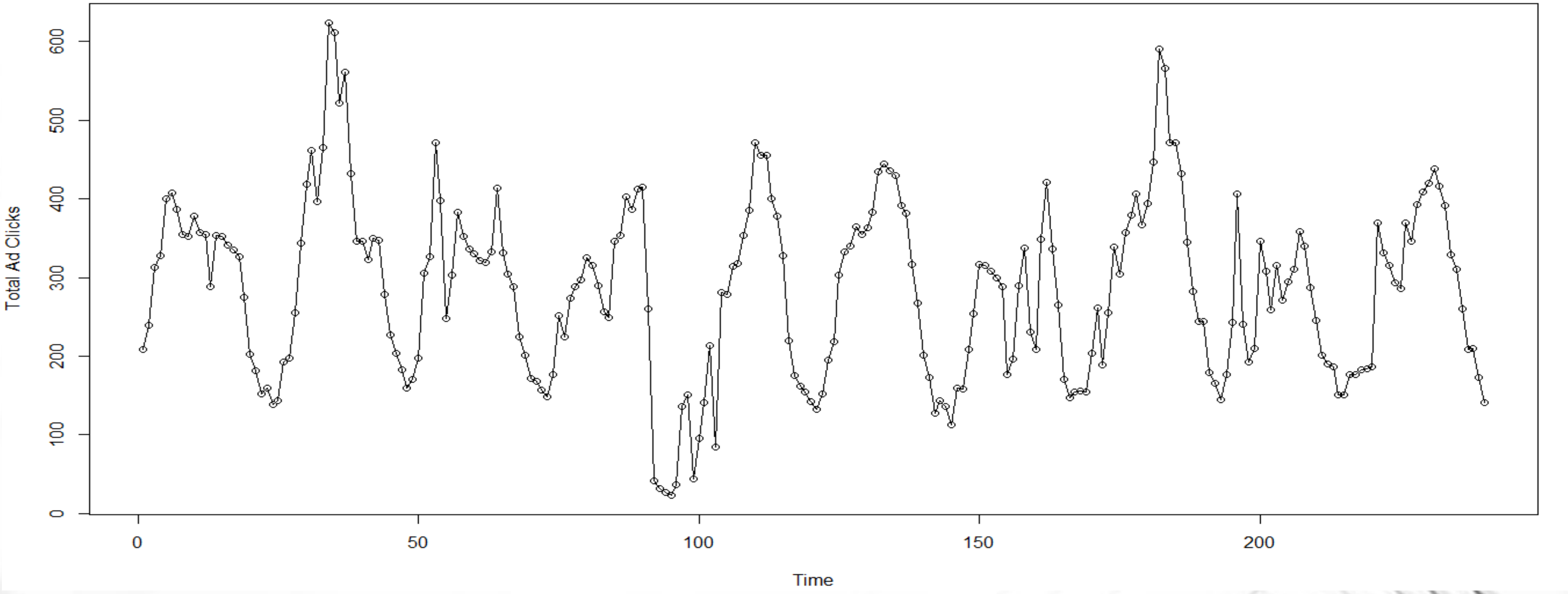
Sampling Strategy: Take a 1% stratified sample from each of the 240 hours of data

Total sample size = 404299

Motivation: Preserve the proportion of total observations as well as ad clicks across different hours.

SEMMA - Explore

- Time is an important variable based on which ad clicks are varying.



SEMMA - Explore

- All independent variables are Categorical

Independent Variable	# Levels
C1	7
Banner_pos	7
site_id	**2184**
site_domain	**2146**
site_category	21
app_id	**2299**
app_domain	**142**
app_category	26
device_id	**64709**

Independent Variable	# Levels
device_ip	**262641**
device_model	**4351**
device_type	4
device_con_type	4
C14	**2070**
C15	8
C16	9
C17	**413**
C18	4

Independent Variable	# Levels
C19	66
C20	**161**
C21	60

10 variables have more than 100 levels

SEMMA - Modify

Handling Categorical Levels:

1. One hot Encoding: This will create dummy variables.
   Huge no. of dimensions will be created
   Not a good option here because no. of levels are very high

2. Impact Coding: Uses naïve Bayes
   Example: Suppose a categorical input variable has 3 levels A,B,and C.

For each level calculate the conditional probability of output=1
For level A of an input variable calculate:
$P(y=1|A)= P(A|y=1) * P(y=1) / P(A)$
$P(y=1|A)= 1/4 * 4/6 / (2/6) = ½$

i/p variable	Response
A	1
A	0
B	0
B	1
B	1
C	1

SEMMA - Modify

i/p variable	Response
A	1
A	0
B	0
B	1
B	1
C	1

Impact Coding

Modified i/p Variable	Response
0.5	1
0.5	0
2/3	0
2/3	1
2/3	1
1	1

Modified all ten variables with more than 100 levels by using impact coding.
Variables modified : site_id, site_domain, app_id, app_domain, device_id, device_ip, device_model, C14, C17, C20

SEMMA - Modify

Input variable :

hour: format is YYMMDDHH, 14091123 means 23:00 on Sept. 11, 2014

Created 2 new categorical variables from the hour variable:

day_of_week : categorical with 7 levels
1 is Monday,…, 7 is Sunday

hour_of_day: categorical with 24 levels
00,01,02,…,23

Motivation: Capture seasonality present in Days of a week and Time of the day.

SEMMA - Modify

Another Technique to deal with large no. of categorical variables is **Hashing**.

Hashing uses less memory and requires little pre-processing. It is a fast and space-efficient way of vectorizing features.

Hash Size is a critical parameter.

Large Hash size - Will handle more variables (i.e. unique values).
Smaller Hash size - Risk having memory collisions and loss of data.

Example: Suppose we choose a Hash size of 4

Var1	Var2
abcd	xyzw

Hash Function $h_1(x)$

123456 → 123456 % 4 = 0

Hash Function $h_2(x)$

679043 → 679043 % 4 = 3

Hashed Output

H_0	H_1	H_2	H_3
1	0	0	1

SEMMA - Model

Binary Classification Problem

Models Used:

a. Logistic Regression
b. Logistic Regression with Hashing
c. Random Forest
d. Gradient Boosting

SEMMA - Assess

Created a 70:30 stratified split to create Training and Validation sets

Evaluation Metric: Logloss

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1-y_i)\log(1-p_i))$$

Model	Logloss on Validation set
Logistic	0.61
Logistic with Hashing	0.413
Random Forest	0.53
GBM	0.43
GBM + Hashing	0.410

Out of all tried models, logistic and GBM models with hashing technique gave the least logloss error

Thank You