CS451 TIM & JERRY

# CLICK PREDICTION

REPO - - - -> BIT.LY/CS451CP

# 3. THE OUTLINE - -> 6 THINGS

- 1. The data
- 2. Experiment specs
- 3. problem description
- 4. data analysis
- 5. model accuracy
- 6. challenges

REPO - - - -> BIT.LY/CS451CP

# 4. THE DATA

- source: Kaggle
- size: ~4GB's
- format: ~11 .csv files
- rows: ~60 million
- predictors: ~10

Competition Details » Get the Data » Make a submission

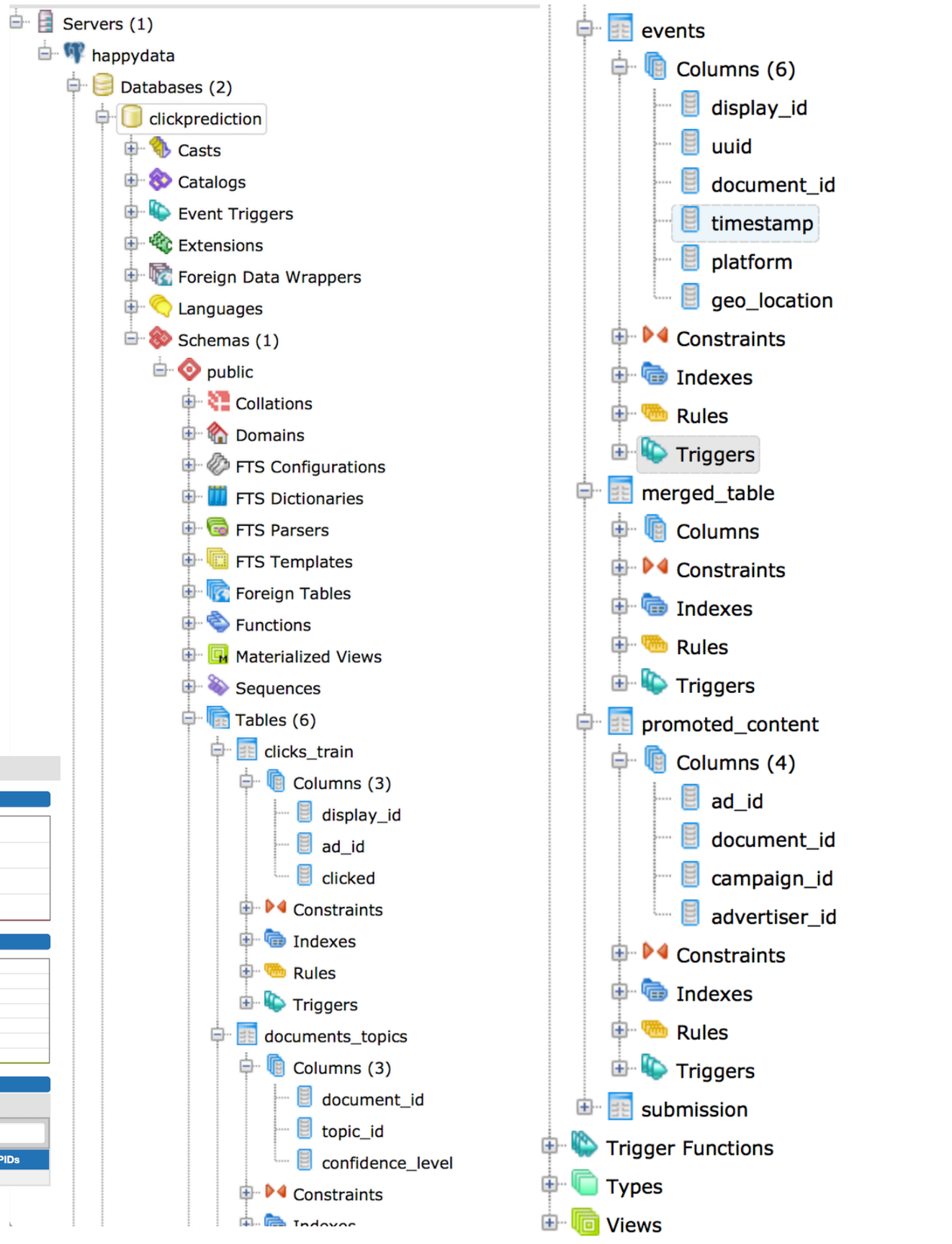
**Data Files**

| File Name | Available Formats |
|---|---|
| documents_categories.csv | .zip (32.34 mb) |
| clicks_test.csv | .zip (135.43 mb) |
| documents_meta.csv | .zip (15.51 mb) |
| documents_entities.csv | .zip (125.67 mb) |
| promoted_content.csv | .zip (2.52 mb) |
| sample_submission.csv | .zip (99.57 mb) |
| documents_topics.csv | .zip (120.91 mb) |
| clicks_train.csv | .zip (389.75 mb) |
| events.csv | .zip (477.74 mb) |
| page_views.csv | .zip (29.71 gb) |
| page_views_sample.csv | .zip (148.51 mb) |

REPO - - - -> BIT.LY/CS451CP

# 5. EXPERIMENT SPECS

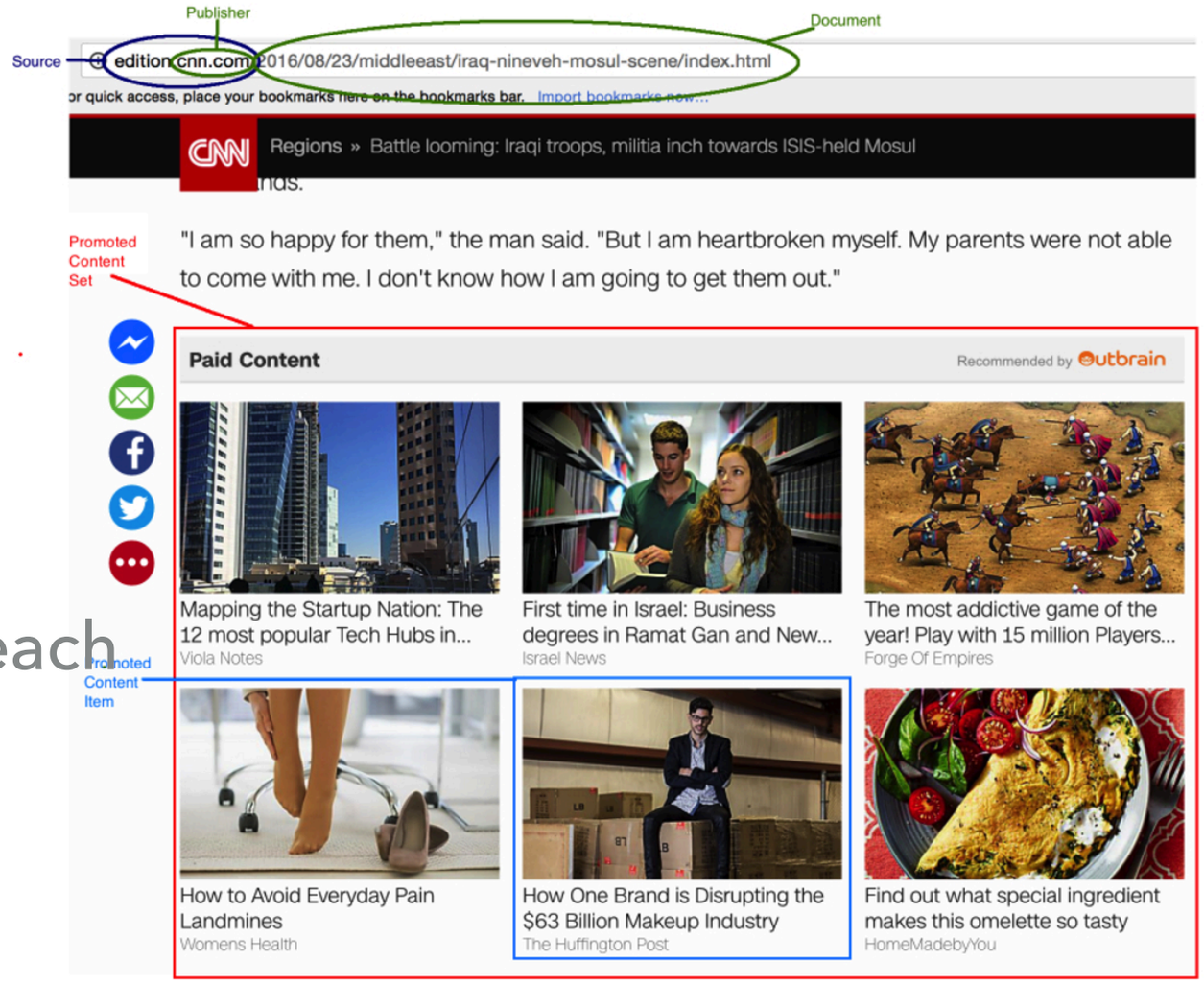
- database server: postgres
- data integration: custom table
- libraries: RPostgreSQL, tree, randomForest, e1071, knitr, markdown

REPO - - - -> BIT.LY/CS451CP

# 6. PROBLEM DESCRIPTION

- click prediction
- calculate probabilities
- sort list by most probable, for each display_id

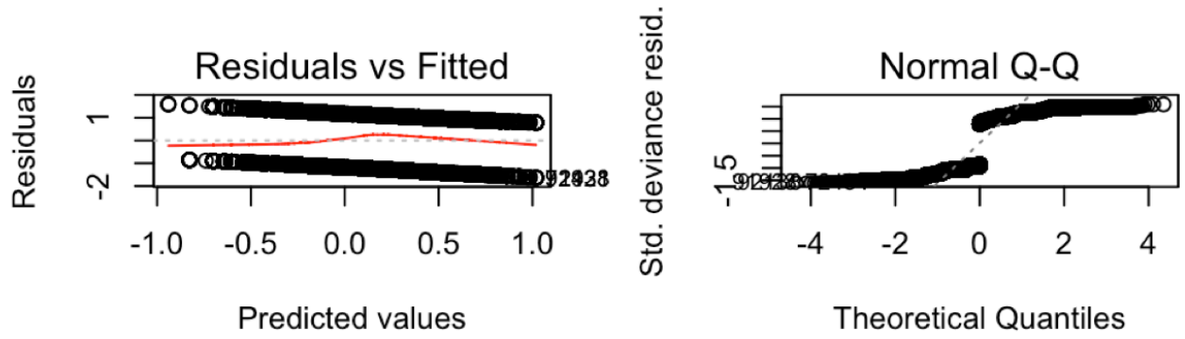


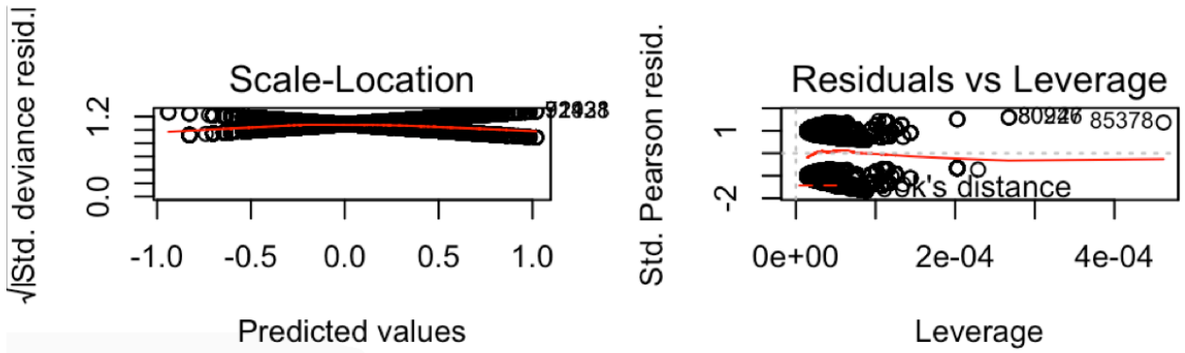

```
######## KAGGLE SAMPLE SUBMISSION FORMAT #########
# "display_id",  "ad_id"
# 16874594,      "170392 172888 162754 150083 66758 180797"
# 16874595,      "8846 143982 30609"
```

REPO - - - -> BIT.LY/CS451CP

# 7. DATA ANALYSIS

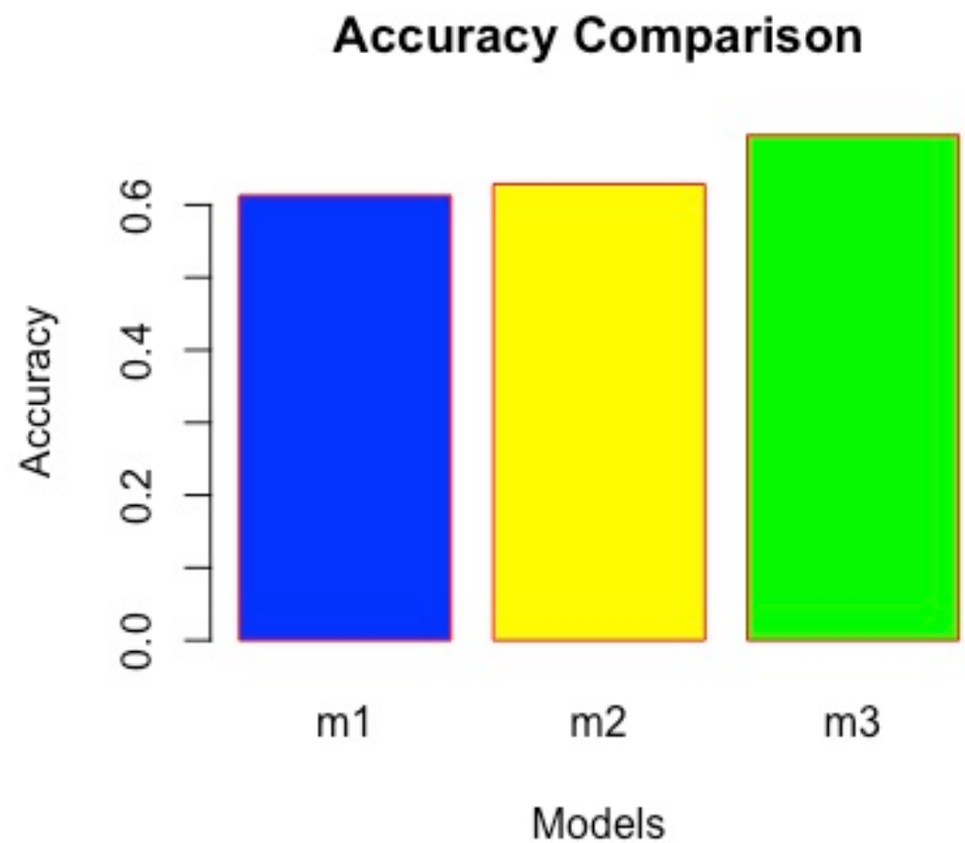
- significant features: topic_id, document_id, confidence_level
- Balanced the data set 50/50 of clicked/not clicked
- Worked with increments of 500K in memory from the database
- Created new table
- Omitted empty or null values
- Train @ 75% - - Test @ 25%



# 8. MODELS & ACCURACY

- m1:glm, predictors: topic_id, confidence_level, accuracy: 61%
- m2: glm, predictors: topic_id, confidence_level, document_id, topic_id, accuracy: 62%
- m3: random forest, predictors: topic_id, confidence_level, accuracy: 69%



# 9. CHALLENGES !

- time: lost 50% of time on time series prediction
- data: big time sink setting up the database server
- tools: RStudio is lackluster for any trivial sized data set

- Proposal: http://bit.ly/2gcCLQ4
- Kaggle: http://bit.ly/2gMVpPG
- Github: http://bit.ly/2gZoTwy
- Data set: http://bit.ly/2fQ0LHW

# SOURCE URLS