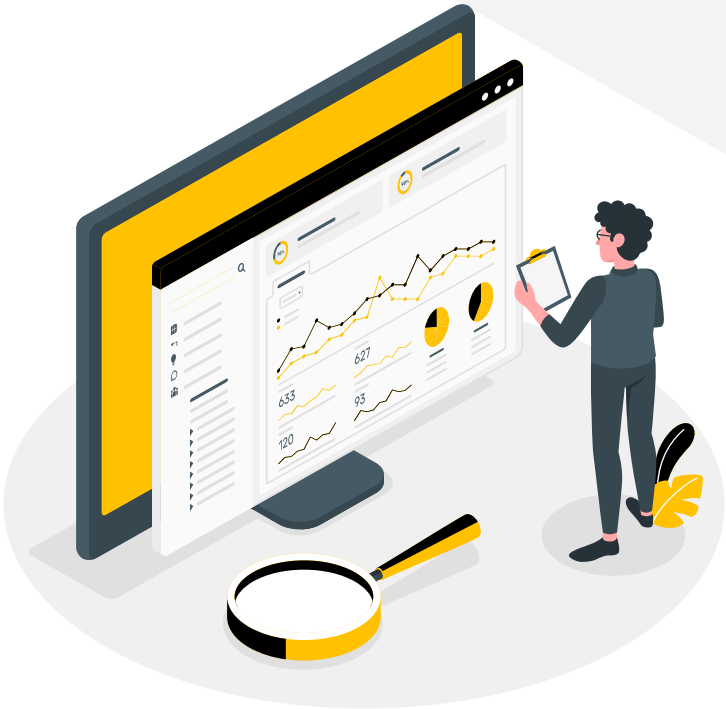
Click-Through Rate Prediction



Team 7:

Po-Han Yen, Shih-Siang Lin, Hsin-Yu Tsai, Yun-Jung Fan, Shu-Ping Chen

Agenda

01

Introduction

- Motivation
- Workflow

02

Data

- Description
- Challenges

03

Research Question

04

Methodology

05

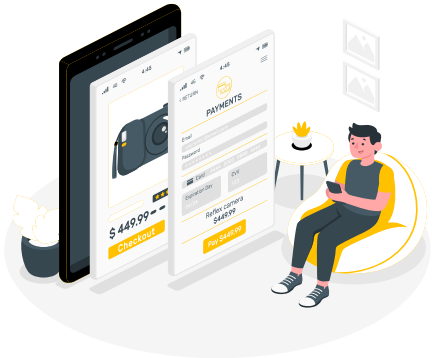
Results & Findings

06

Conclusion

Introduction - Motivation

- Online advertisements have a significant influence on the success of a business
- Click-through rate (CTR) is commonly used to evaluate ad performance



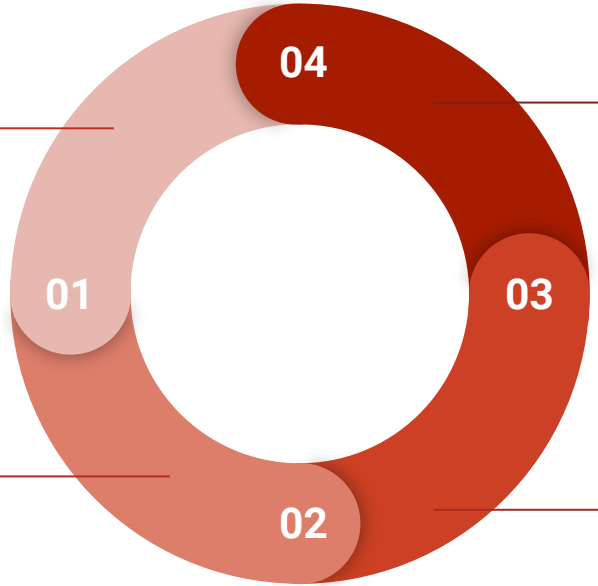
Introduction - Workflow

Gathering data

Kaggle

Pre-processing

Downsampling, Partitioning, Recategorizing



Evaluation

Accuracy, Sensitivity, Specificity, AUC

Training & testing models

Ensemble, Random Forest, Boosting, Naive Bayes, Prune classification tree, Logistic Regression, and KNN

Data - Description

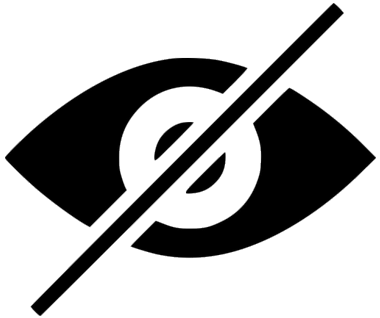
id	click	hour	C1	banner_pos	site_id	site_domain	site_category	app_id	app_domain	app_category	device_id	device_ip	device_mode	device_type	device_conn_	C14	C15	C16	C17	C18	C19	C20	C21
1E+18	0	14102100	1005	0	1fbe01fe	f3845767	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	ddd2926e	44956a24	1	2	15706	320	50	1722	0	35	-1	79
1E+19	0	14102100	1005	0	1fbe01fe	f3845767	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	96809ac8	711ee120	1	0	15704	320	50	1722	0	35	100084	79
1E+19	0	14102100	1005	0	1fbe01fe	f3845767	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	b3cf8def	8a4875bd	1	0	15704	320	50	1722	0	35	100084	79
1.0001E+19	0	14102100	1005	0	1fbe01fe	f3845767	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	e8275b8f	6332421a	1	0	15706	320	50	1722	0	35	100084	79
1.0001E+19	0	14102100	1005	1	fe8cc448	9166c161	0569f928	ecad2386	7801e8d9	07d7df22	a99f214a	9644d0bf	779d90c2	1	0	18993	320	50	2161	0	35	-1	157

- Data source: Kaggle CTR prediction contest
- Observations: 45 million records
- Variables: 23 in total
- dependent variable: click

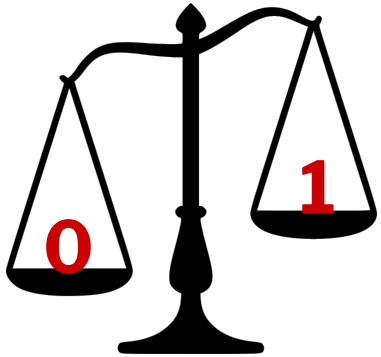
Data - Challenges

Big data

Anonymized features



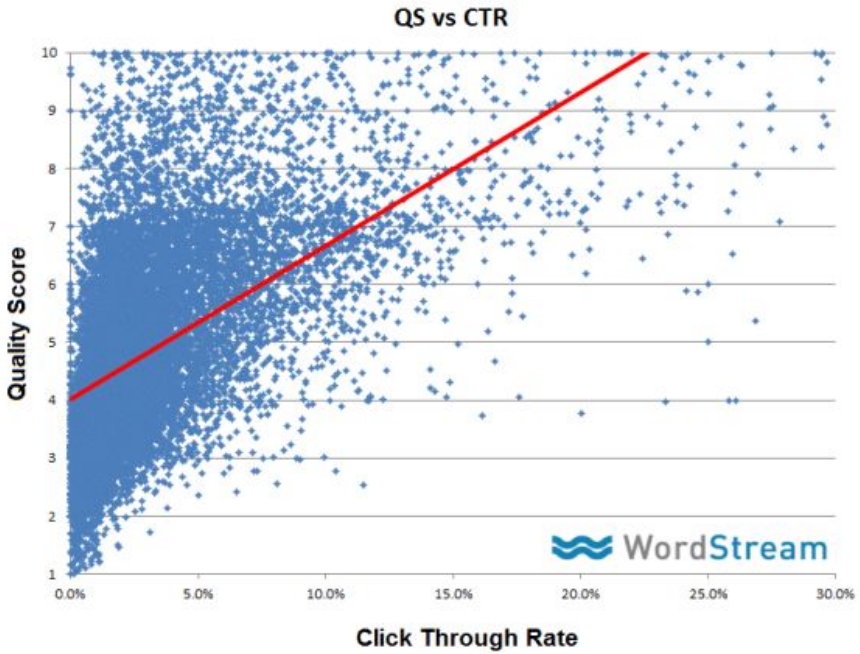
Imbalanced classes of the dependent variable



Research Question - 1

- There is a positive relationship between Click Through Rate and Quality Scores.

- Is it possible to predict whether an ad would be clicked by viewers or not based on historical data? Furthermore, is it possible to successfully identify both clicked ads and non-clicked ads?



Research Question - 2

- The challenge of the humongous data sets :

  Is it possible to drive values from the data in an efficient way under constraints?

Methodology

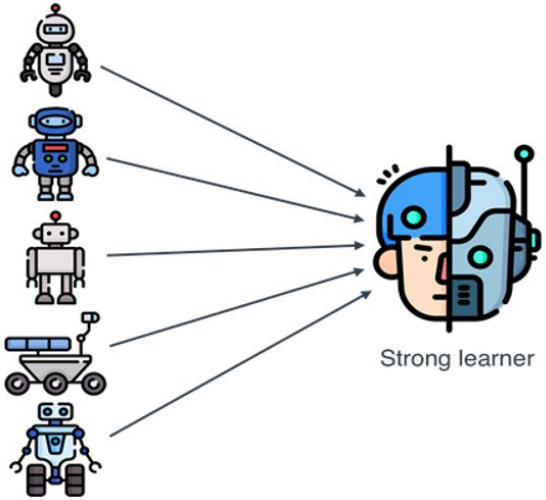
Anonymized features

Feature Category

Group Mean CTR vs Total Mean CTR

Very Low, Low, Median, High, Very High

Big data



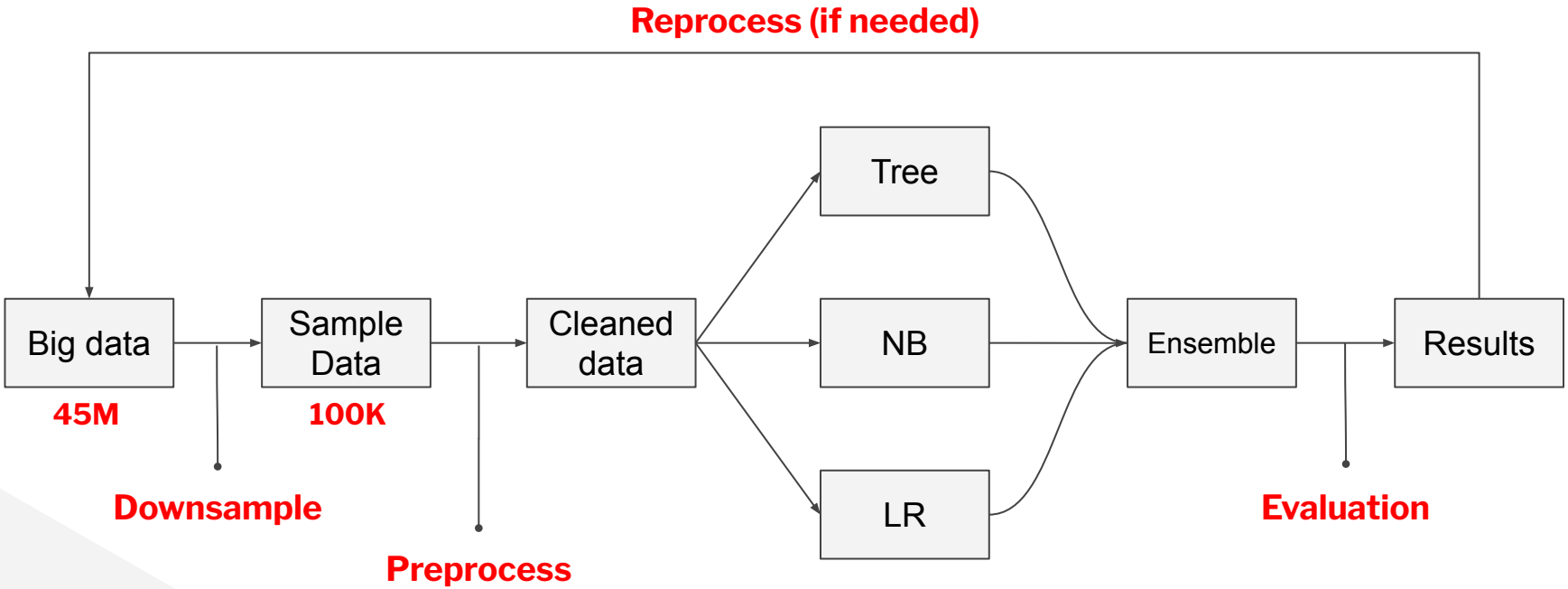
https://livebook.manning.com/concept/machine-learning/ensemble-method

Imbalanced classes

- Weights
- Synthetic

✅ Downsample

Auto modeling



Our Ensemble Algorithm

- Randomly choose a subset of data & 2-6 features

- Build NaiveBayes, Logistic Regression or Classification Tree

- Choose proper prediction method based on input model

- Average the predictive probabilities to get final result

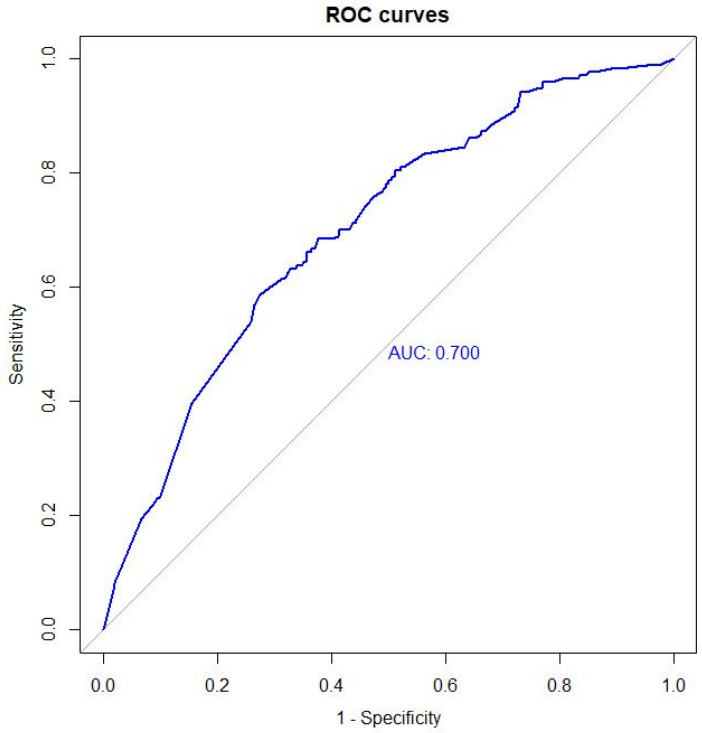
Results & Findings

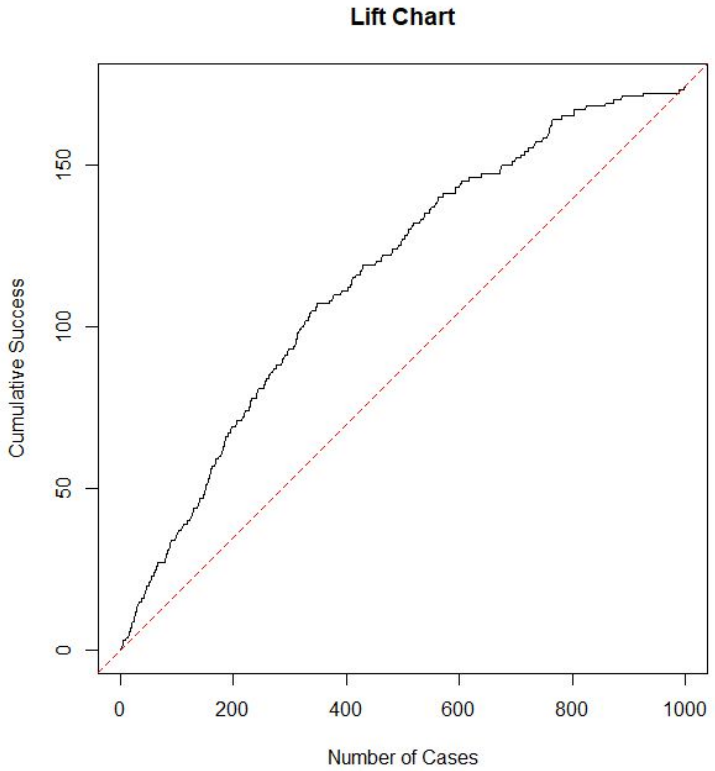
Stage I	Accuracy	Sensitivity	Specificity
NaiveBayes	0.637	0.271	0.894
KNN	0.636	0.609	0.641
Prune Tree	0.608	0.267	0.907
Bagging	0.591	0.666	0.575
Gradient Boosting	0.562	0.216	0.863
XGB	0.635	0.402	0.684
Random Forest	0.627	0.775	0.595

Results & Findings

Stage II	Accuracy	Sensitivity	Specificity
125 Tree	0.651	0.638	0.653
125 NB	0.598	0.706	0.575
125 LR	0.609	0.695	0.59
125 Tree + 125 NB	0.623	0.666	0.613
125 Tree + 125 LR	0.64	0.666	0.634
125 NB + 125 LR	0.604	0.683	0.587
125 Tree + 125 NB + 125 LR	0.628	0.666	0.619
375 Tree	0.652	0.638	0.655

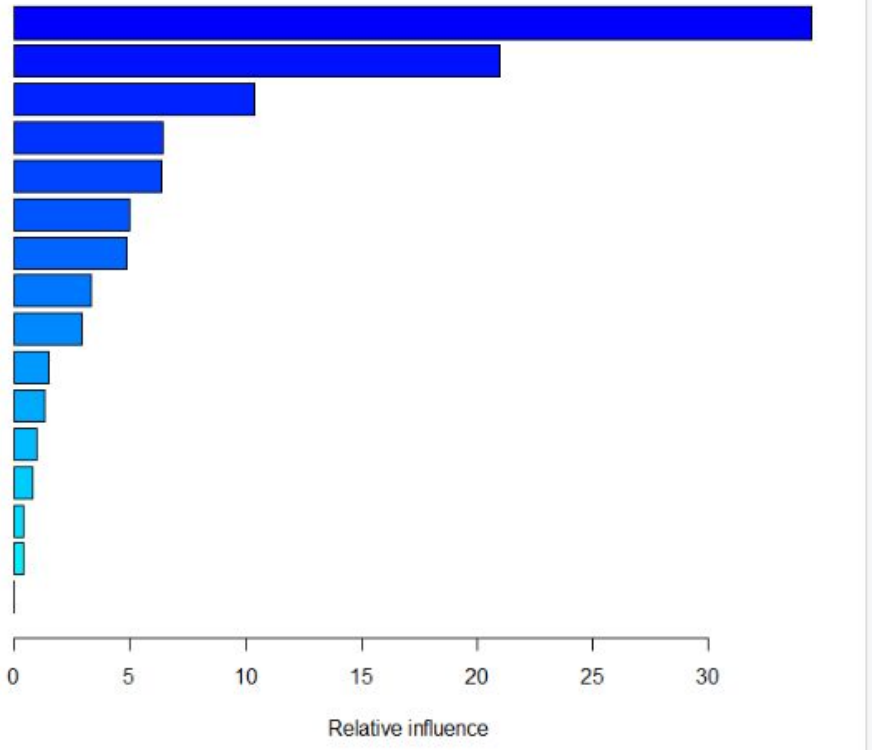
Results & Findings





Conclusion

var	rel.inf
hour	34.4686489
weekday	20.9877846
device_model	10.3461106
site_domain	6.4150408
site_id	6.3270323
C18	4.9969144
site_category	4.8590808
app_id	3.3239721
app_domain	2.9384772
banner_pos	1.4918899
app_category	1.3075987
device_conn_type	0.9726429
C1	0.7914517
C16	0.4034392
C15	0.3699158
device_type	0.0000000



Conclusion

Recommended prediction model:

- Self-build ensemble tree-based model
  - Accuracy: 65.2%
  - Sensitivity: 63.79%
  - Specificity: 65.5%
  - AUC: 70%

Future works:

- More data and models
- Deep learning algorithms
- Process data on big data platforms
  - Hadoop and Spark

Thank You!

Team 7