AdTracking Fraud Detection

Detect fraudulent click traffic for mobile app ads

Group 4:
Zhongxing Xue, Junkai Zhang, Xiaoxiao Guo, Sitong Chen, Wenshan Wang

CONTENTS

PART 1
Motivation and Goal

PART 2
EDA

PART 3
Model

PART 4
Evaluation

PART 1

Motivation and Goal

Motivation

Click Fraud

-- Occurs on the Internet in pay-per-click (PPC) online advertising.

-- Fraud occurs when a person, automated script or computer program imitates a legitimate user to click on ad.

-- According to TalkingData, 90% of clicks are potentially fraudulent.

Goal

Help app developers detecting and avoiding click fraud to save cost and obtain an accurate market feedback.

PART 2

EDA

is_attributed: the target that is to be predicted, indicating the app was downloaded

App :app id for marketing.

click_time: timestamp of click (UTC)

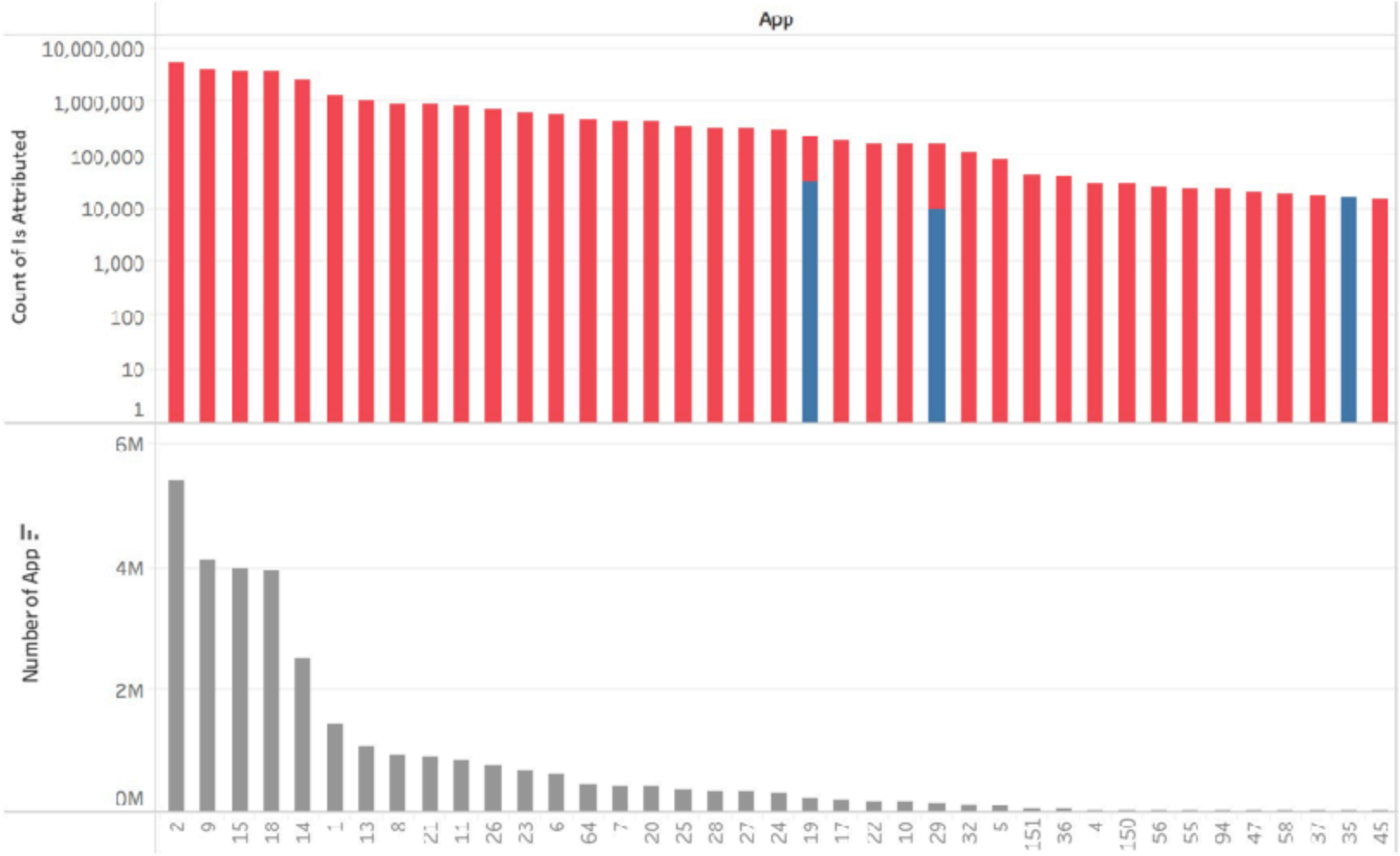
channel: channel id of mobile ad publisher

os: os version id of user mobile phone

device: device type id of user mobile phone (e.g., iphone 6 plus, iphone 7, huawei mate 7, etc.)

ip: ip address of click

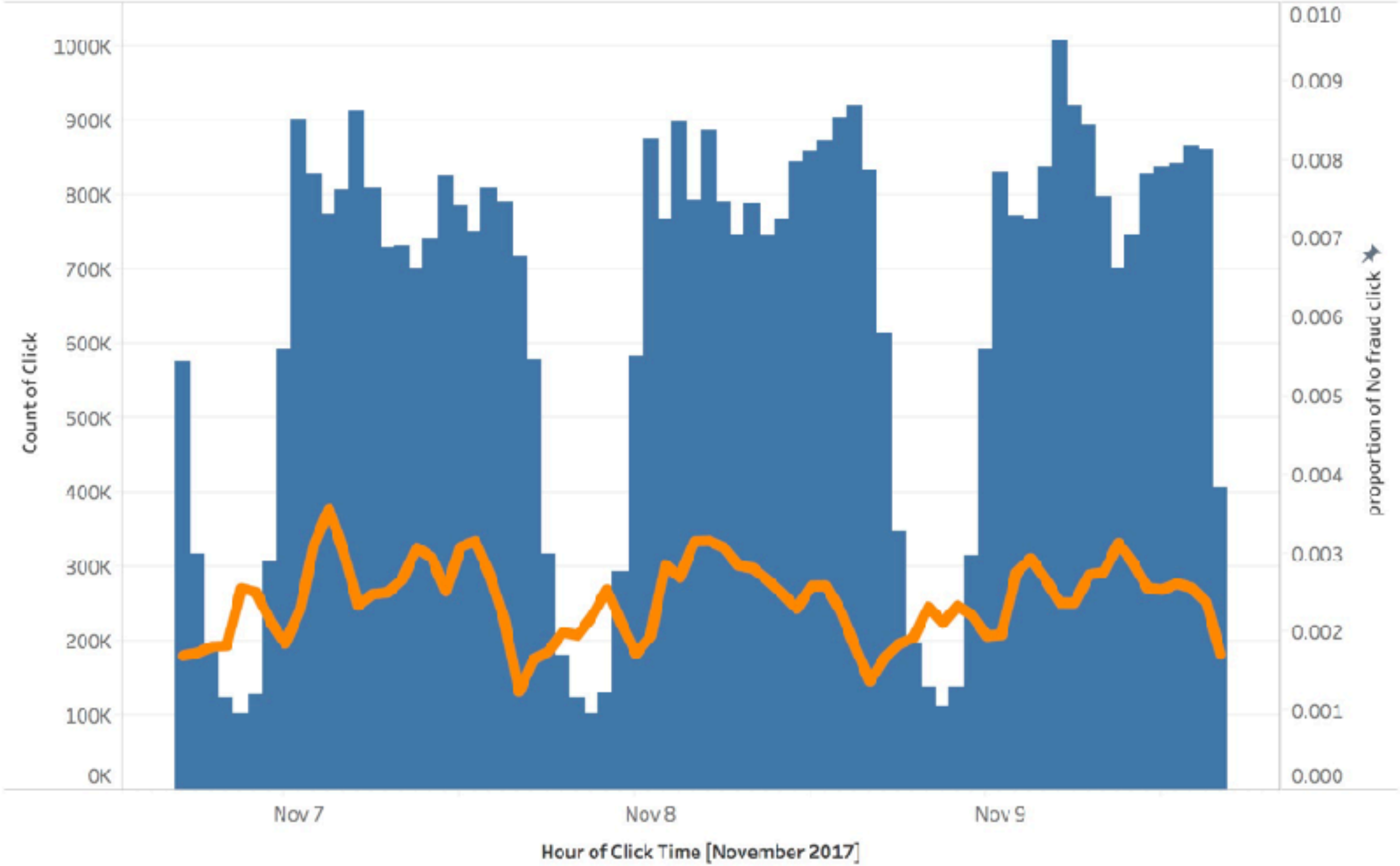
Apps which have large number of count



Is Attributed

- 0
- 1

Clicks and Proportion of "No fraud" per hour



PART 3

Model

0.2%

Is_attributed = 1
App was downloaded
Not fraud

Data is highly imbalanced.

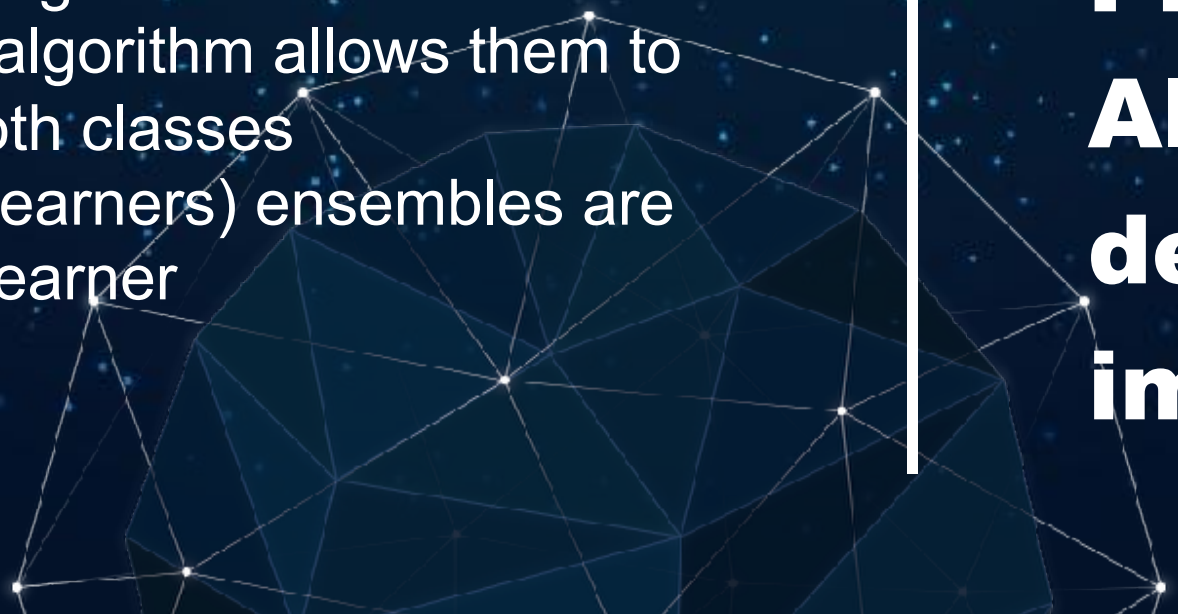
99.8%

Is_attributed = 0
App not downloaded
Fraud

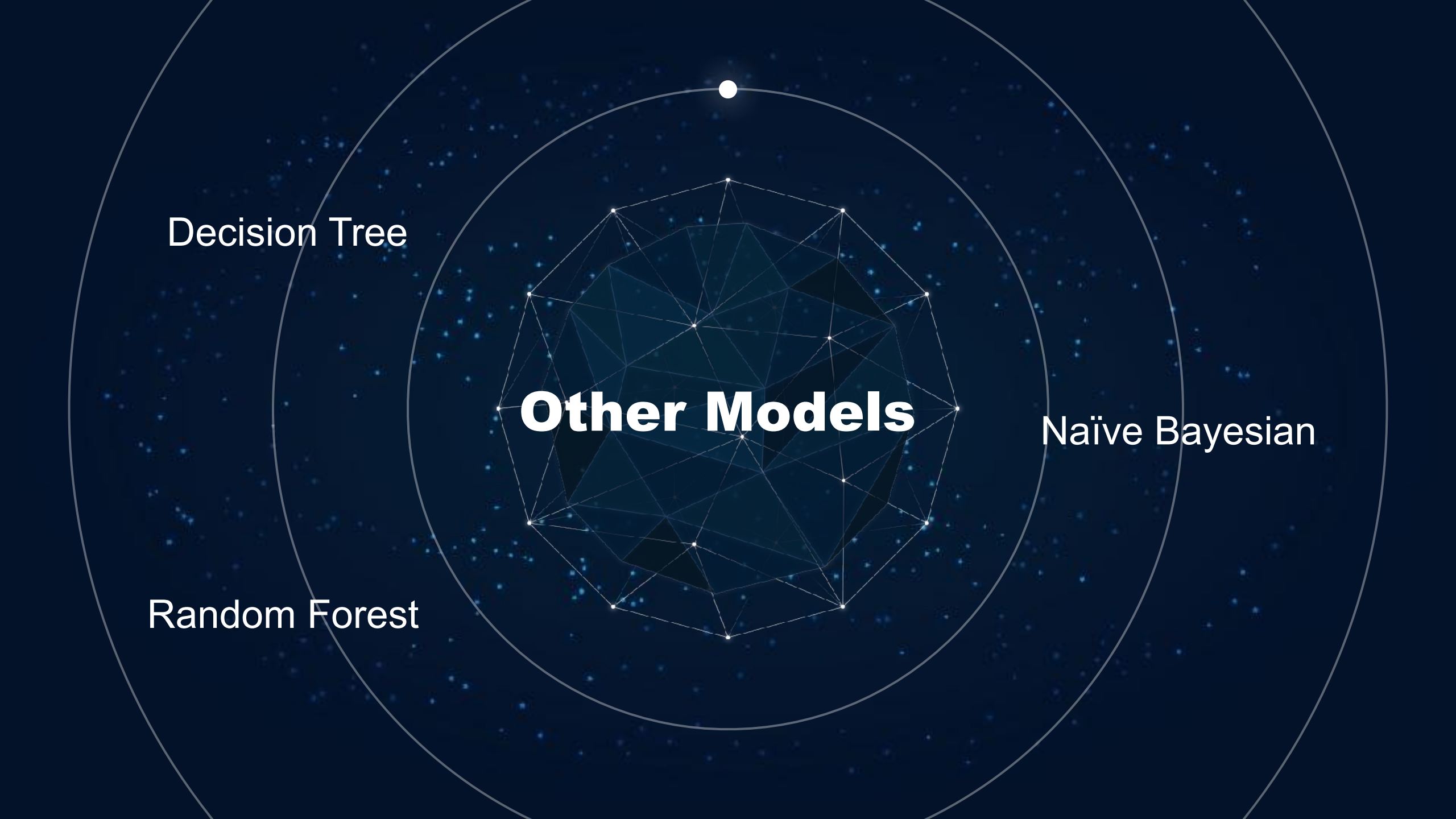
Procedure 1: Balance Data

-- Under-sampling the majority class or Oversampling the minority class
-- Smote algorithm (Synthetic Minority Over-sampling TEchnique)
-- introducing synthetic examples
-- k minority class nearest neighbors

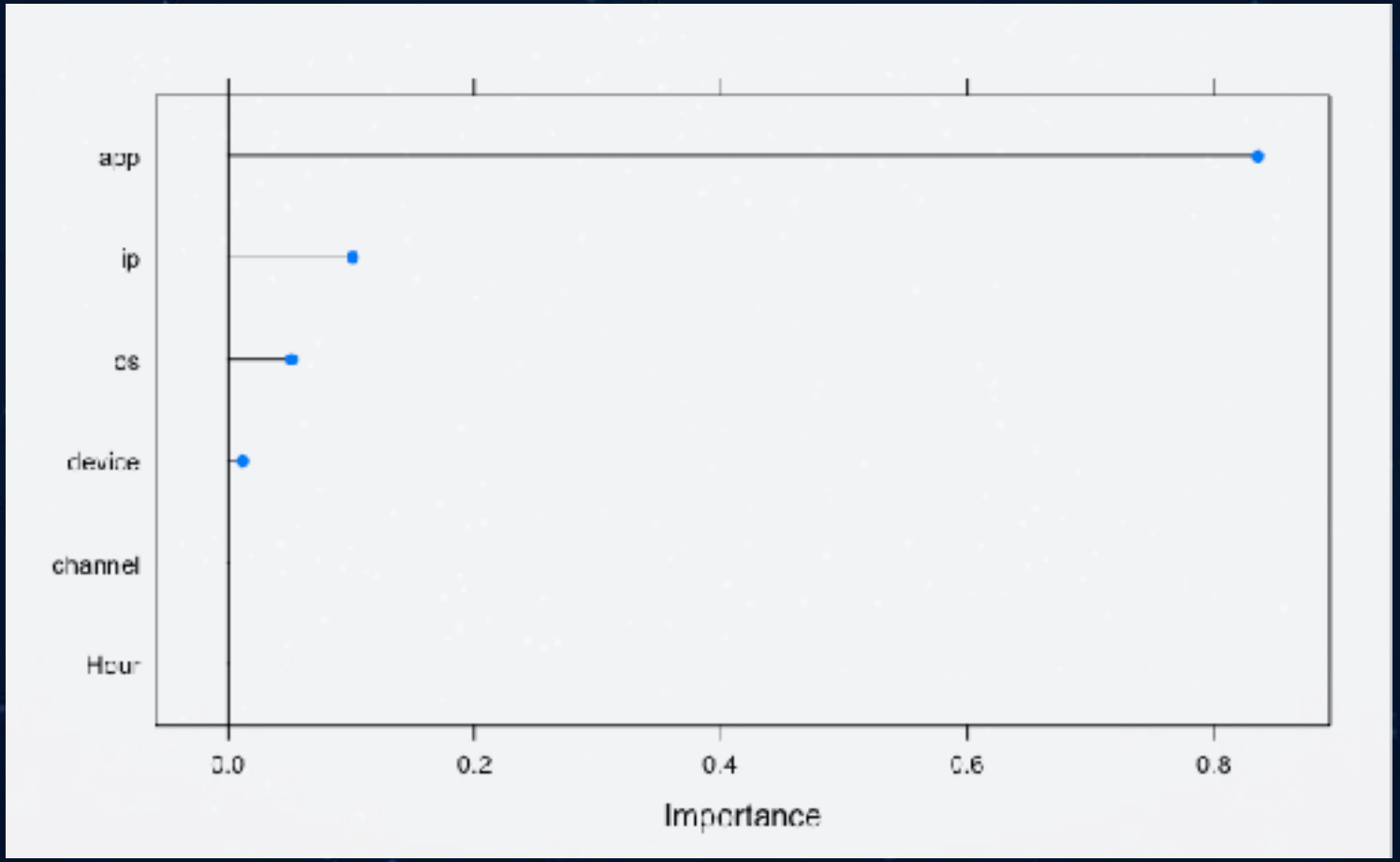
Gradient Tree Boosting
-- tree-based algorithm allows them to learn signals from both classes
-- tree (weak learners) ensembles are boosted into strong learner



Procedure 2: Algorithm deal with imbalance


Decision Tree
Other Models
Naïve Bayesian
Random Forest

Variable Importance



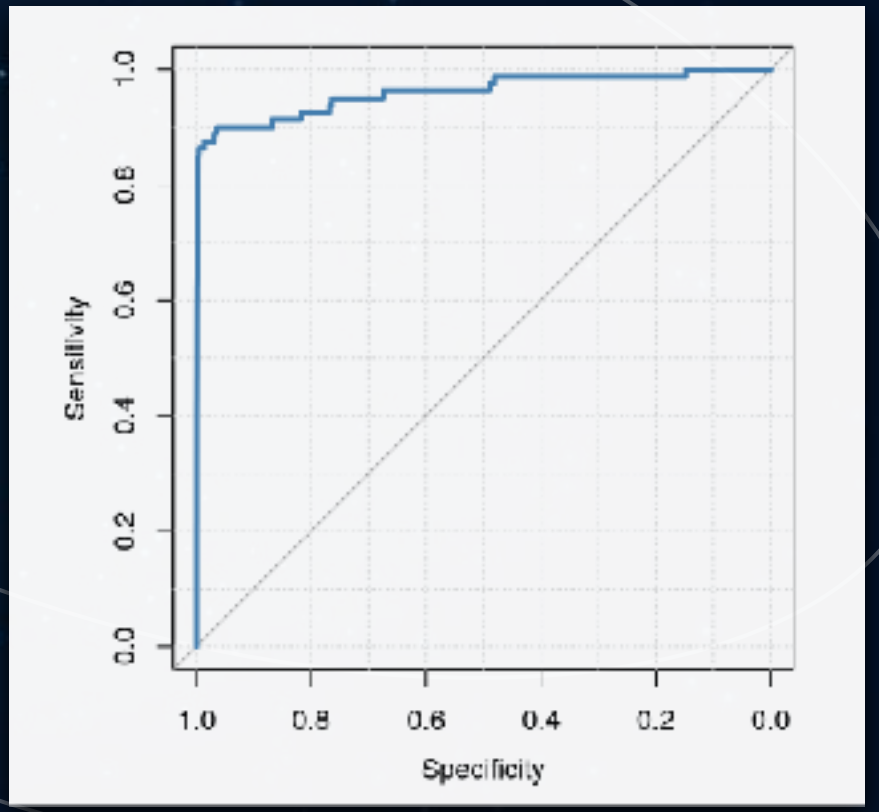
PART 4

Evaluation

AUC
(Area Under Curve)

-- Shows the tradeoff between sensitivity (TP) and specificity (FP)

-- The closer the curve follows the left-hand border and top border, the more accurate the test



Evaluation Result

	XGBoost	Decision Tree	Random Forest	Naïve Bayesian
AUC	0.9530	0.8944	0.8857	0.9560

THANKS