AUTO-DETECTION OF CLICK-FRAUDS USING MACHINE LEARNING

Anshuman Dash
MBA in Business Analytics
REVA University, Bengaluru

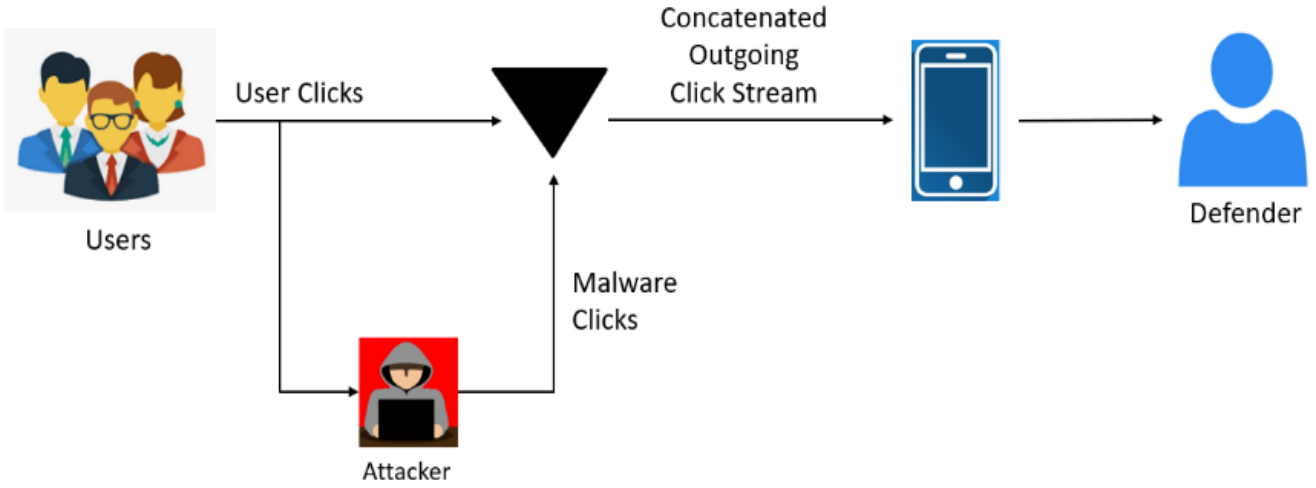
Satyajit Pal
MBA in Business Analytics
REVA University, Bengaluru

Agenda


Data Gathering & Understanding
Data Preparation
Problem Statement
Modeling
Business Case
Interpretation

Business Case

The click-fraud is considered one of the most critical issues in online advertising.



Introduction

While browsing online, number of advertisements are shown up and while clicking on them, the advertiser pay the publisher a fraction of money for every new user they bring in. This in simple terms is the Pay-Per-Click (PPC) revenue model of online advertising industry.

The current PPC revenue model is highly prone to click fraud spams which annually results in loss of billions of dollars from the pockets of the advertisers. Click-fraud is a major threat to the present online advertising ecosystem.

Hence, we have researched and come up with a machine learning approach which can effectively and accurately classify the clicks to be fraud or non-fraud with minimum number of false positives as possible.

Business Case (Continued…)

REVA UNIVERSITY

- Click fraud systems have been growing continuously in recent years.
- Existing detection approaches aim to classify click fraud behaviors from different perspectives, but each has its own limitations.
- Adding fuel to the fire is the lack of legislation or even resources to tackle this huge problem.
- The practice of click fraud is designed to negatively impact the advertisers advertising budget.
- Automated bots scraping information from various source would act as click frauds.

Challenges & Problem Statement

REVA UNIVERSITY

Challenges

- The click-fraud methods have been improving day by day.
- There is no legislation against these frauds.
- No industry standards for identifying and addressing these click-frauds.

Ad Fraud = ad impressions caused by bots, not seen by humans

Impression Fraud
(CPM) Fraud
(includes mobile display, video ads)

Click Fraud
(CPC) Fraud
(includes mobile search ads)

Problem Statement

- Pro-actively auto detect the click-spam using relative uncertainty between click-spam and valid clicks-streams.
- It does this by identifying repeated patterns from valid click-spam in the ad network.

Data Gathering

- A traffic monitoring on REVA University campus network gateways was set up to capture legal ad-click files.
  - Ad URL
  - Ad server IS,
  - Publisher page
  - Source IP address,
  - User agent string
  - Time stamp for every click.

- In total between August to November 2019, a total of **32,119 clicks** were recorded. Data were collected and all stored data are encrypted after proper process of obtaining ethical approval.

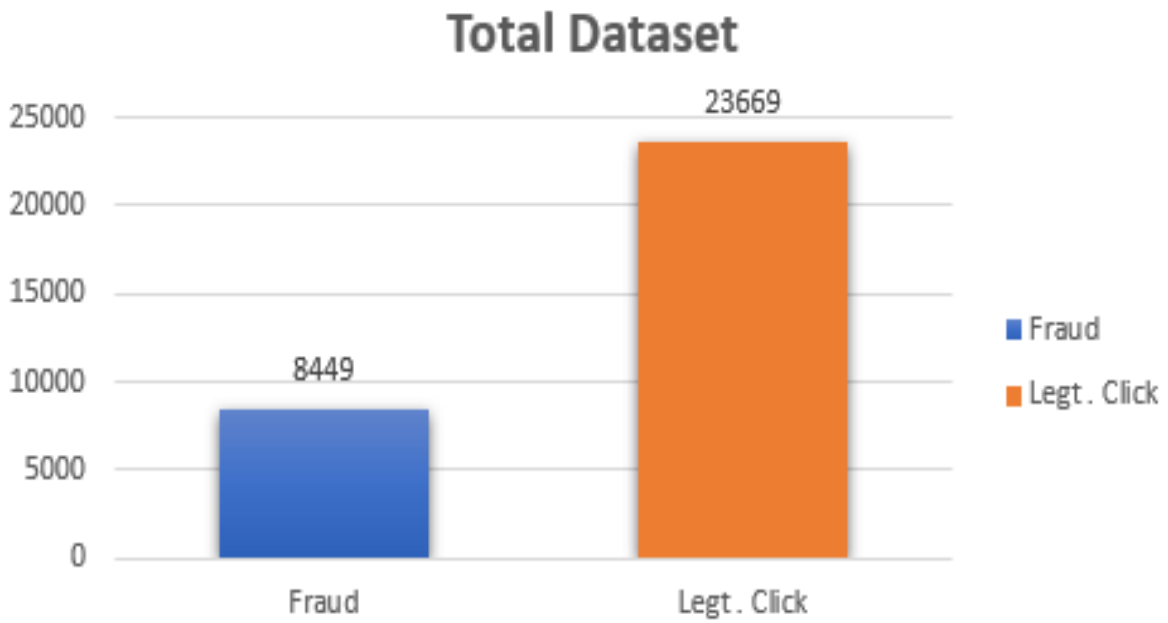
Data Gathering (Continued…)

REVA UNIVERSITY

Dataset Source: REVA University Click Data

Time Period: August 2019 to October 2019

Variable	Definition
IP	IP address of the click
APP	app id for marketing
DEVICE	workstation type
OS	OS version of the workstation
CHANNEL	CHANNEL version of the workstation
CLICK_TIME	CLICK_TIME version of the workstation
ATTRIBUTED TIME	User clicking the ad time
ID	target ID
TIMESTAMP	target ID prediction time
STATE	Types of clicks distribution



Data Understanding (Continued..)

REVA UNIVERSITY

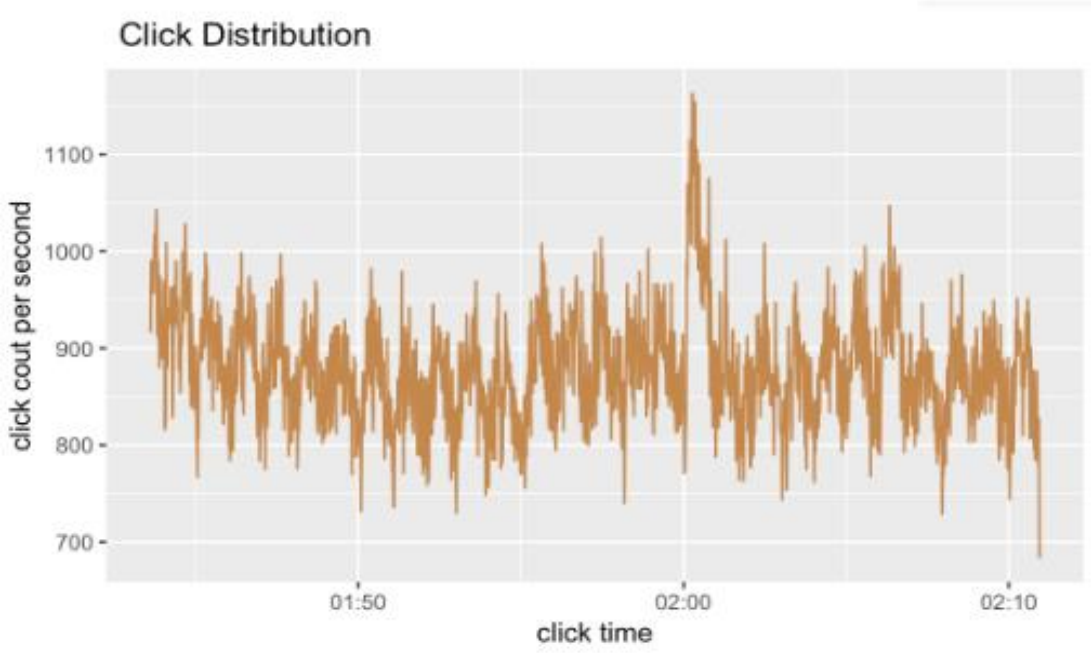
Summarizing Shape & Spread Of Data

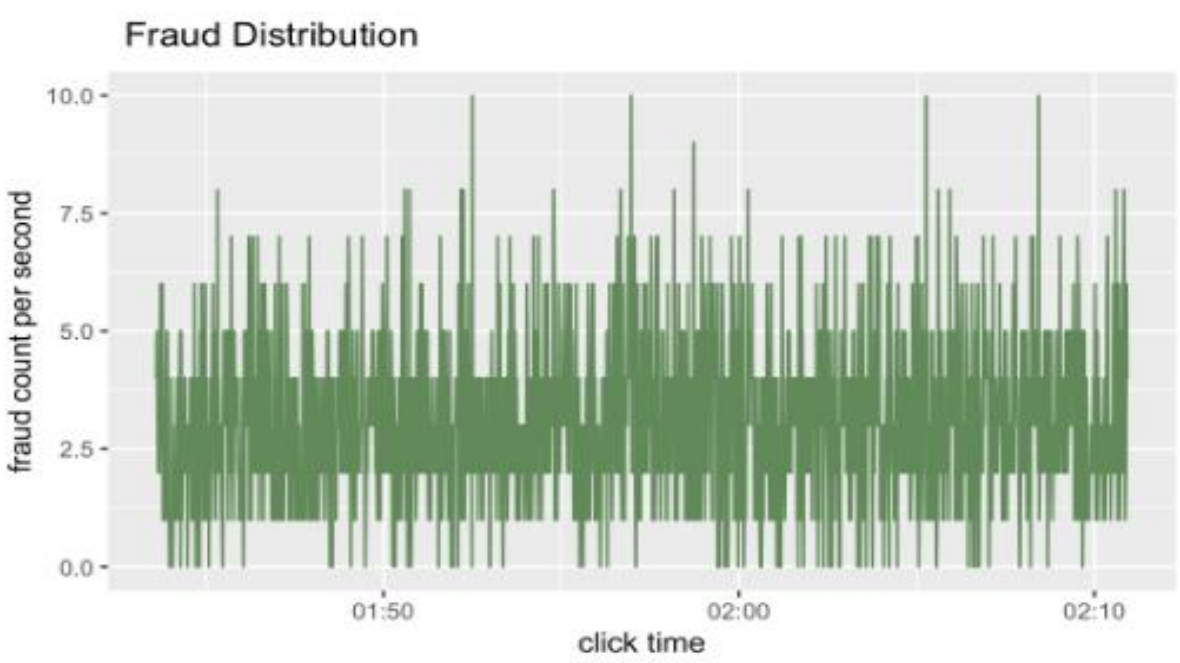
High Level Summary

Dimension	: **32119 Observations**
Count of Unique Features	: **9 Input + Fraud data**
Unique Count of Users	: **1283 Students**
Types of Click-Streams	: **2 (Click-Distributions, Fraud-Distributions)**

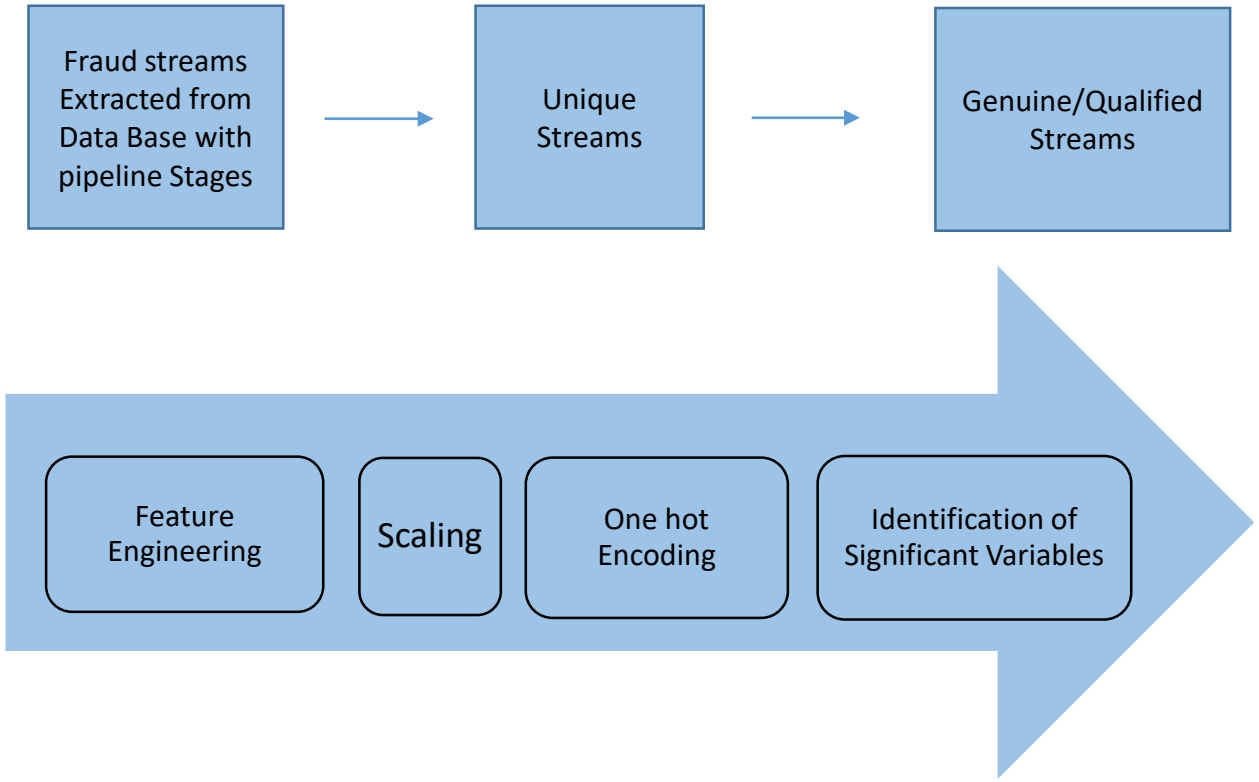
Target Variable: State

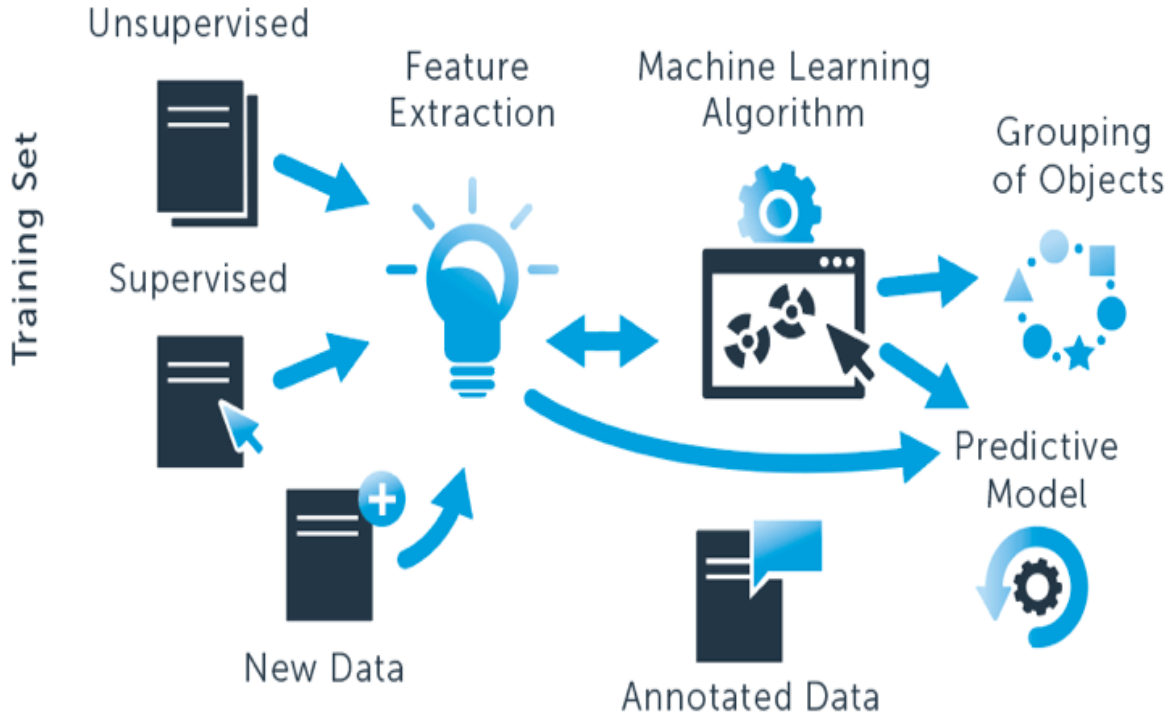
Class: Legt. Click & Fraud Click





Data Preparation





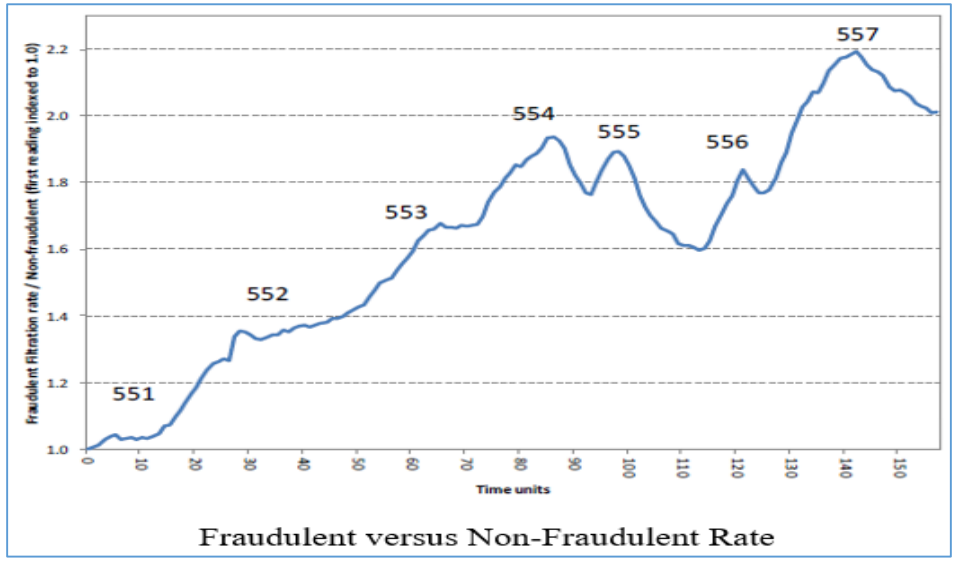
The data is prepared for effective analyses after data collection. The data set obtained consists of several attributes which are not required, so the data was prepared according to the requirements so that the algorithm produces accurate results.

Modelling

We used various classification algorithms to come up with the best model for this problem. The ultimate purpose was to build a model that classify the frauds best and hence accuracy was used as the chief parameter.

Along with accuracy, precision, recall and F1 scores have been considered to identify best model.

	TPR	FPR	TNR	FNR	ACCURACY
Random Forest	95.40%	14.30%	85.70%	4.60%	89.40%
Classification Trees	94.40%	7.60%	92.40%	5.60%	93.20%
Support Vector Machines	95.40%	9.10%	90.90%	4.60%	92.70%
knn Classification	69.00%	69.00%	31.00%	31.00%	45.70%
Gradient Tree Boosting	2.80%	76.90%	23.10%	15.20%	97.20%



Fraudulent versus Non-Fraudulent Rate

Interpretation

REVA UNIVERSITY

- The best tuned model is Gradient Tree Boosting with 12 most important attributes including numerical click count variables and transformed categorical features. The F1 score on test dataset reached 0.92, while the overfitting was avoided.

- As per the research, **fraudulent ad publishers** frequently engage in fraud clicks on certain channels rather than decentralize them in order to demand higher prices for a small number of clicks.

	Accuracy	Precision	Recall	F1 Score
Random Forest	89.4%	0.79	0.89	0.84
Clasificaiton Trees	93.2%	0.85	0.94	0.89
Support Vector Machine	92.7%	0.83	0.91	0.87
KNN	45.7%	0.48	0.55	0.51
Gradient Tree Boosting	97.2%	0.89	0.95	0.92

Conclusion

- This model can be used to mitigate the threat of click-fraud spam.
- Further modifications can be made to the model so that it can perform well when trained over the new data.
- The classifier can be embedded in a software which can be used on the webserver to run automatically while importing the click-logs.


Q&A
questions
answers
ask
help
investigation
inquiry
confused
problem
info
query