



Fake AD Click Detection

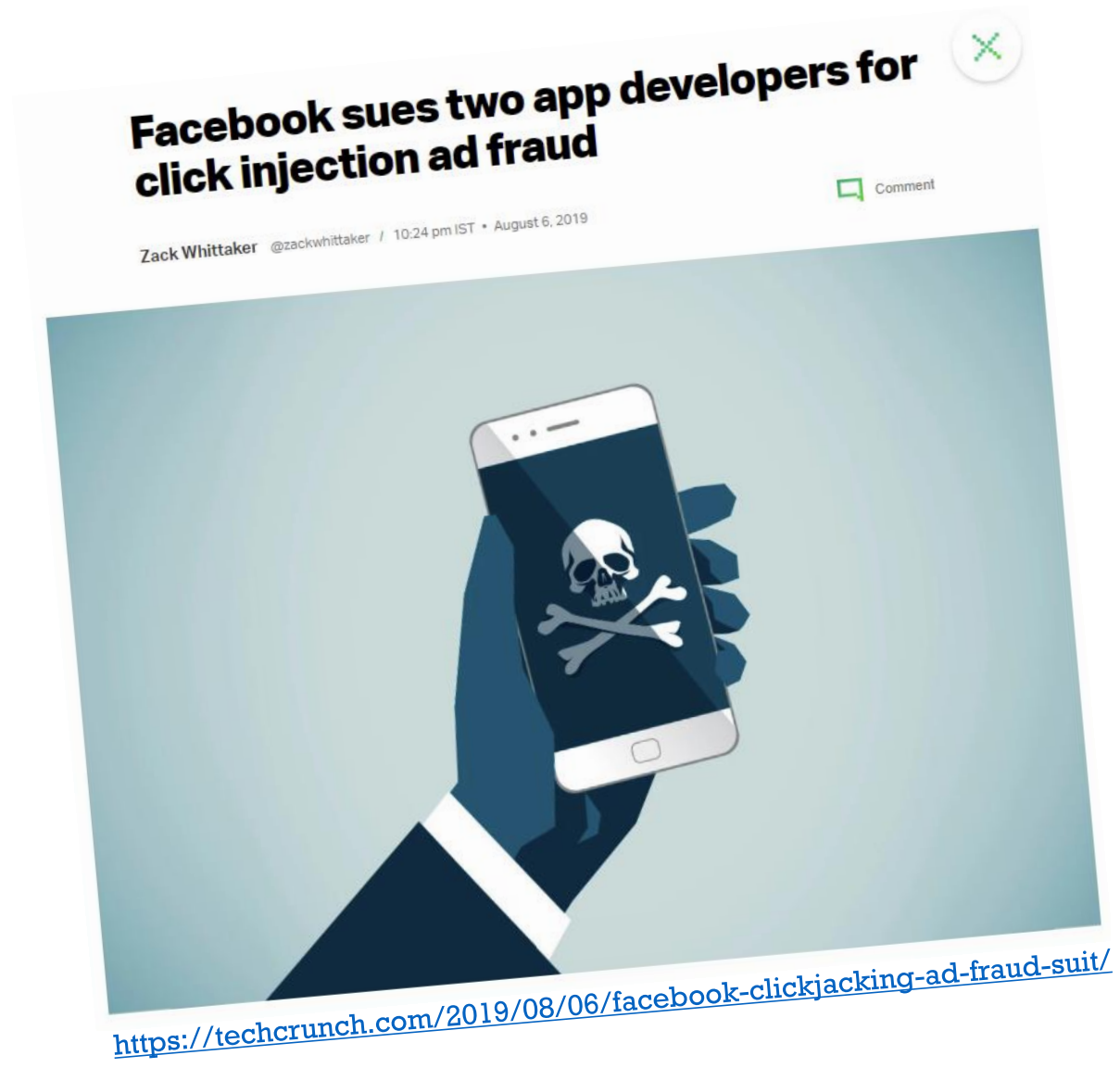
Gautam Shanbhag (18210455)
Supervisor : Prof. Mark Roantree
Dublin City University

Introduction to AD World



Pay-per-click (PPC), is an internet advertising model used to drive traffic to websites, in which an advertiser pays a **publisher** (typically a search engine, website owner, or a network of websites) when the ad is clicked.

Problems of AD World



Problems of AD World

BRIEF

Mobile ad fraud rates double, with shopping apps hardest hit



Credit: Wikimedia Commons

<https://www.retaildive.com/news/mobile-ad-fraud-rates-double-with-shopping-apps-hardest-hit/523240/>

Problems of AD World

Mobile Ad Fraud Cost USD\$2.3bn (£1.89bn) in H1 2019; Sizmek Finalises Peer39 Sale

by Mathew Broughton on 2nd Aug 2019 in News



In this weekly segment, ExchangeWire sums up key industry updates on ad tech from around the European region – in this edition: Mobile ad fraud cost marketers over USD\$2.3bn (£1.89bn) in H1 2019; 79% of UK marketers acknowledge cross-channel video advertising wave; Centro enforces app-ads.txt to automatically target authorised and validated sellers; Teads offers 100% viewability on VCPM and CPCV buys; and Merkle launches data-driven planning methodology, Pando Planning, in the UK.

Mobile ad fraud cost marketers over
USD\$2.3bn (£1.89bn) in H1 2019

<https://www.exchangewire.com/blog/2019/08/02/mobile-ad-fraud-cost-usd2-3bn-1-89bn-in-h1-2019-sizmek-finalises-peer39-sale/>

Problems of AD World



HARVEST ACCESS RI

HOME · SUBSCRIPTIONS · RESEARCHSTORE · SERVICES · HARVEST · CONSULTANCY ·

Home > Press > Press releases > Advertising Fraud Losses to Reach \$42 Billion in 2019, Driven by Evolving Tactics by F

ADVERTISING FRAUD LOSSES TO REACH \$42 BILLION IN 2019, DRIVEN BY EVOLVING TACTICS BY FRAUDSTERS

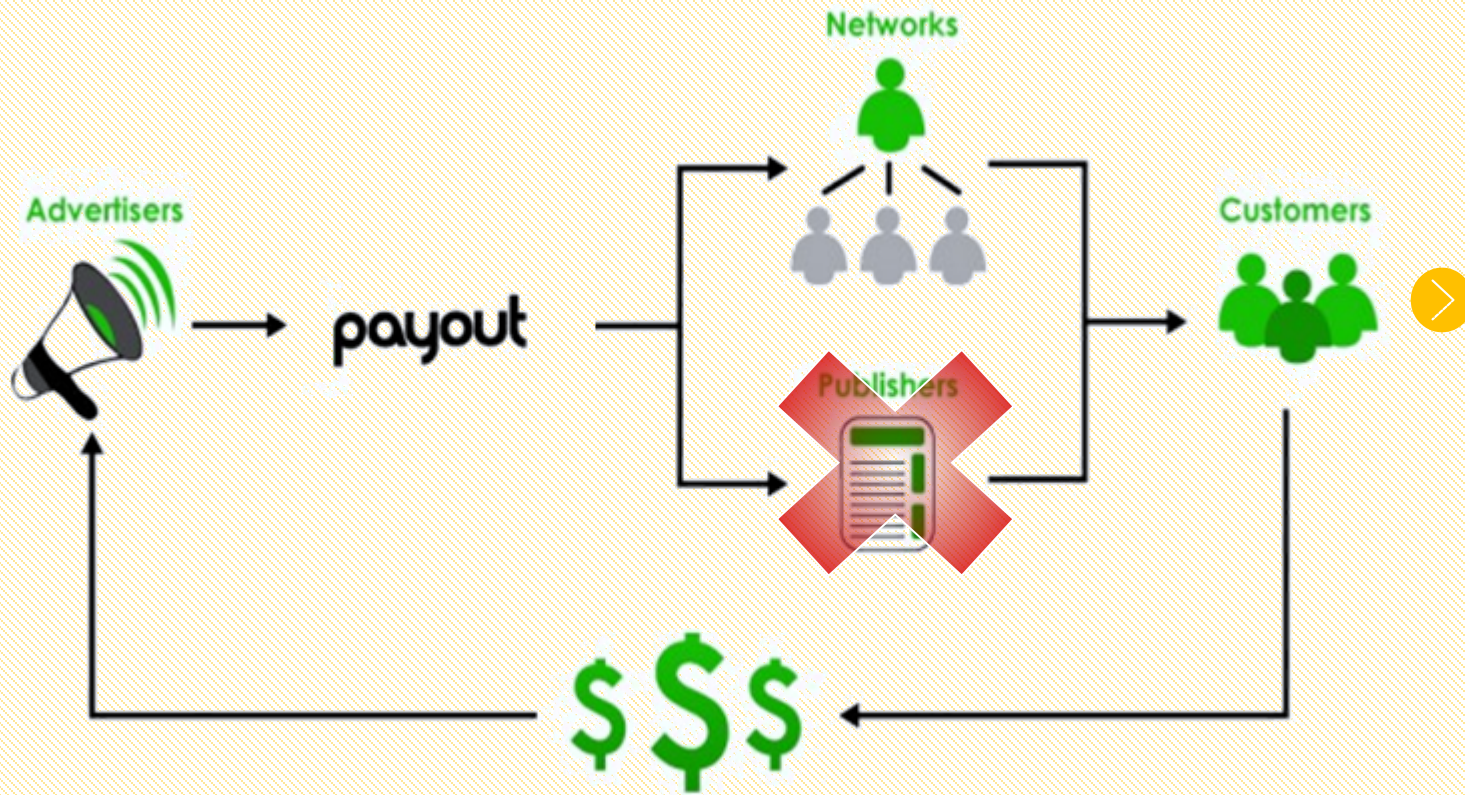
Hampshire, UK – 21st May 2019: A new study from **Juniper Research** forecasts that advertisers will lose \$42 billion of ad spend globally this year to fraudulent activities committed via online, mobile and in-app advertising. This is a 21% increase from the \$35 billion lost to advertising fraud in 2018. This growth will be driven by increasingly sophisticated techniques being implemented by fraudsters.

For more insights on the Digital Advertising market landscape, download our free whitepaper: The Impact of AI for Digital Advertisers.

Advertisers' Loss to Reach \$100 billion by 2023

<https://www.juniperresearch.com/press/press-releases/advertising-fraud-losses-to-reach-42-billion>

Introduction



CLICK FRAUD

black-hat technique of falsely inflating the number of clicks on a pay-per-click ad.

WHY

- ❖ Generate Revenue and profit for themselves
- ❖ Hurt a competitor by depleting their ad budget and improving their own ad position.
- ❖ Weaken companies by driving up their marketing cost or increase cost of acquisition

Previous Work

Dataset & Preprocessing



DATASET :

- ❖ TalkingData
- ❖ Avazu
- ❖ Buzzfeed

PREPROCESSING :

- ❖ Principal Component Analysis (PCA)
- ❖ Recursive Feature Elimination (RFE)
- ❖ SMOTE

Algorithms



- ❖ Hellinger Distance Decision Tree
- ❖ Bagging & Boosting Decision Tree
- ❖ Clustering Using K-Means
- ❖ Clustering Using Splay Tree Algorithm
- ❖ Dempster Shafer evidence theory

Evaluation



- ❖ Accuracy
- ❖ ROC-AUC

Dataset Description

Source : LARC SMU-CMU
Singapore Management
University provided by
BuzzCity



Dataset Snapshot

address	bankaccount	partnerid	status
tle0ao6u67qaiwgmek4817o3w	NaN	dv91f	OK
j8hl8uuip15ku56ere498tcwn	hlshfjmd9ftb7uf7wquuv9r3y	dv8sy	OK
rzqk95gpqy16bebgwo8znpbav	NaN	dv8sd	OK
tpsxjzmzmfjnk516gtfz28es	fx11691shvf7vevzvfe38cavz	dv8pz	OK
i82qiiadjfajw7aetq583gci4p	rjt5cv6fhxs4af54i22dolwh6	dv8pa	OK

agent	category	cid	cntr	id	iplong	partnerid	referrer	timeat
SonyEricsson_K70	ad	dsfag	us	9794476	1071324855	dv3va	NaN	2012-03-08 00:00:00.0
Samsung_S5233	mg	dswae	in	9794474	1000461055	dv4gs	riflql2a0yv8xoa9sq0recx4x	2012-03-08 00:00:00.0
Nokia_C3-00	co	dr75h	py	9794471	3386484265	duq7h	NaN	2012-03-08 00:00:00.0
Nokia_5233	es	ds3xq	vn	9794468	1907981997	dv6i3	gp53lqr9njqd6z2ap5d364sip	2012-03-08 00:00:00.0
MAUI	ad	dvb8g	in	9794467	1791989091	duxto	NaN	2012-03-08 00:00:00.0

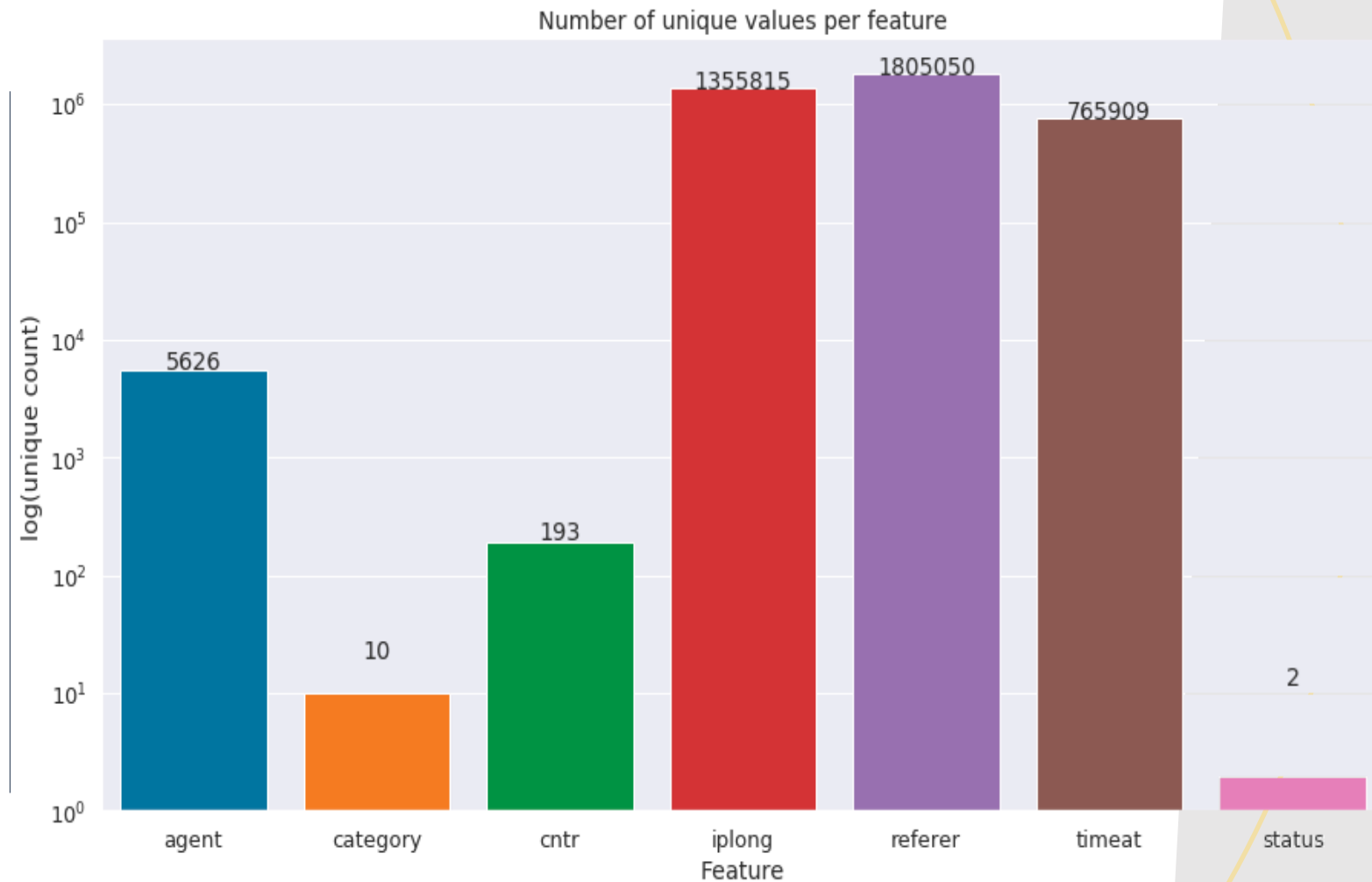
Understanding Dataset



Size : 69,38,523 ie 7 Million

Fraud Count – 4,44,439

Ok Count – 64,94,084



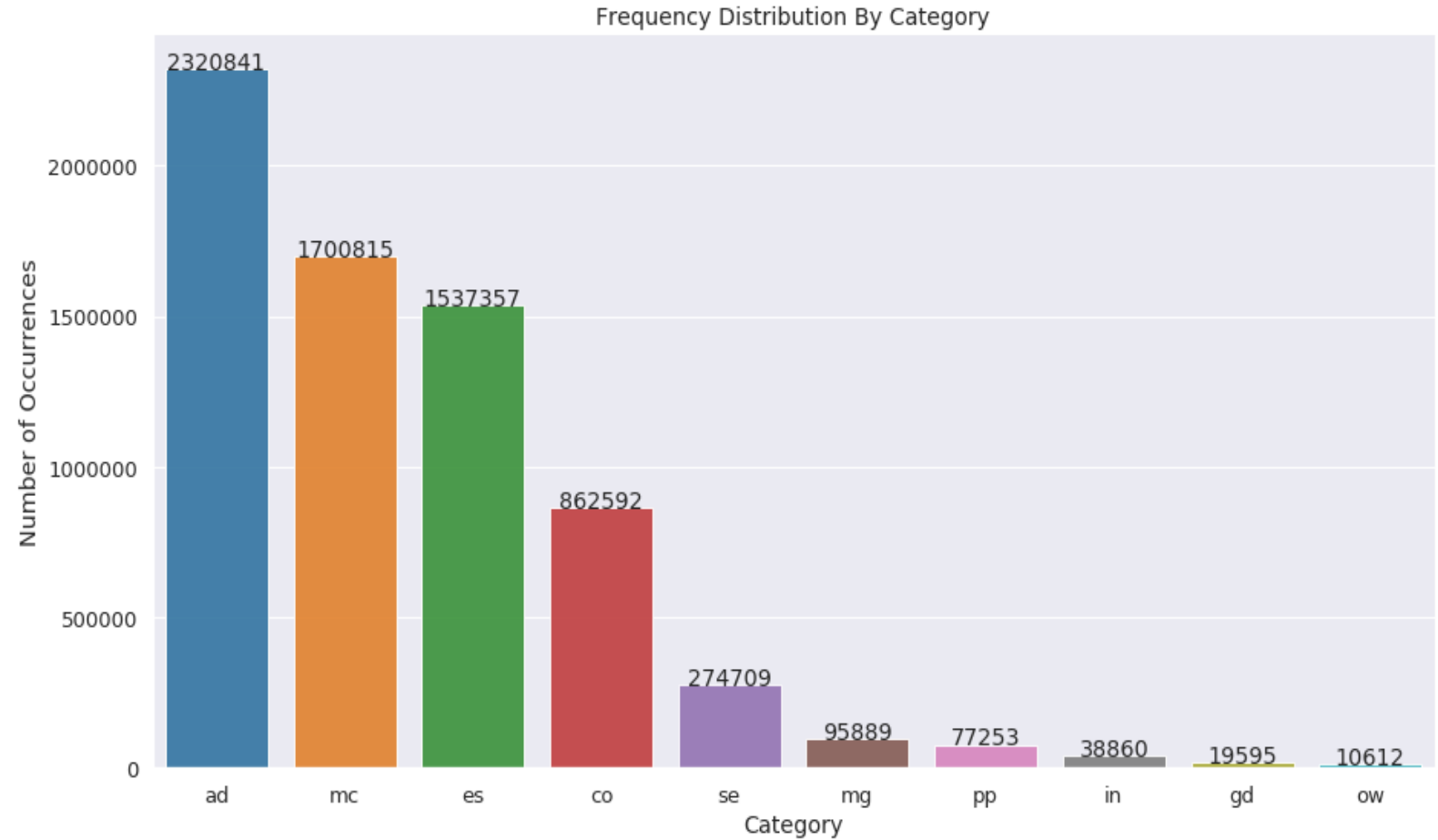
Understanding Dataset



Distribution of Categories for Mobile Data

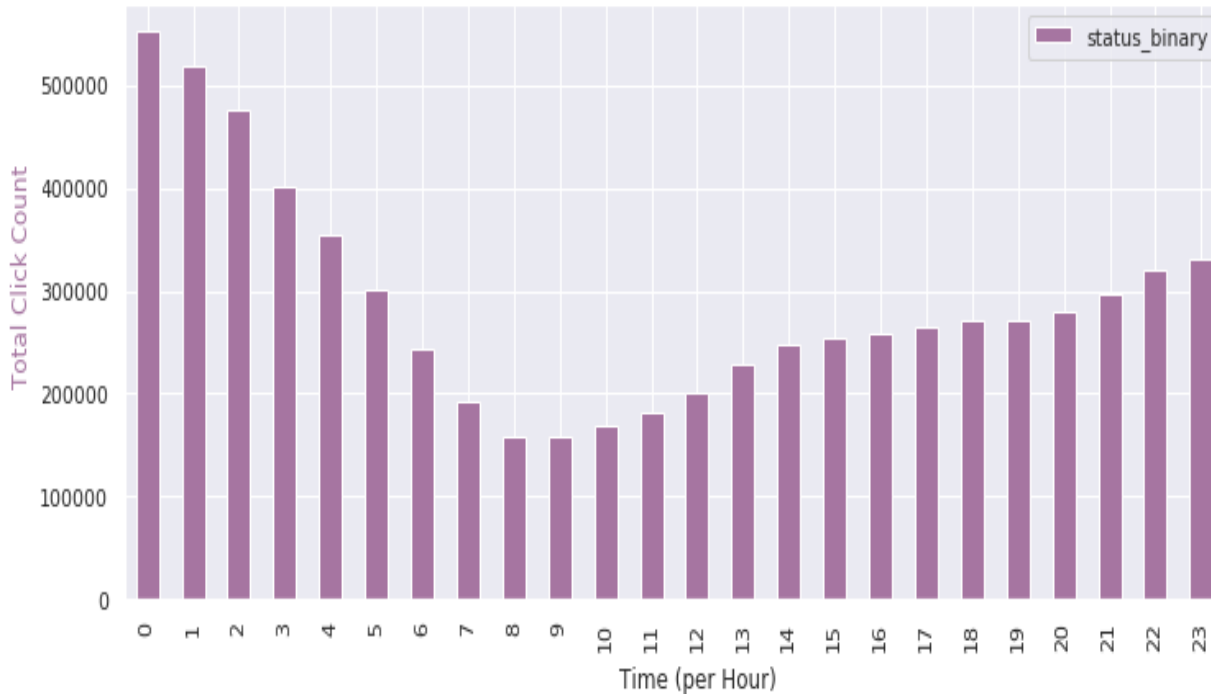
Top 5 Categories which attract clicks

1. Adult Sites
2. Mobile Content
3. Entertainment & Lifestyles
4. Community
5. Search Portal Services



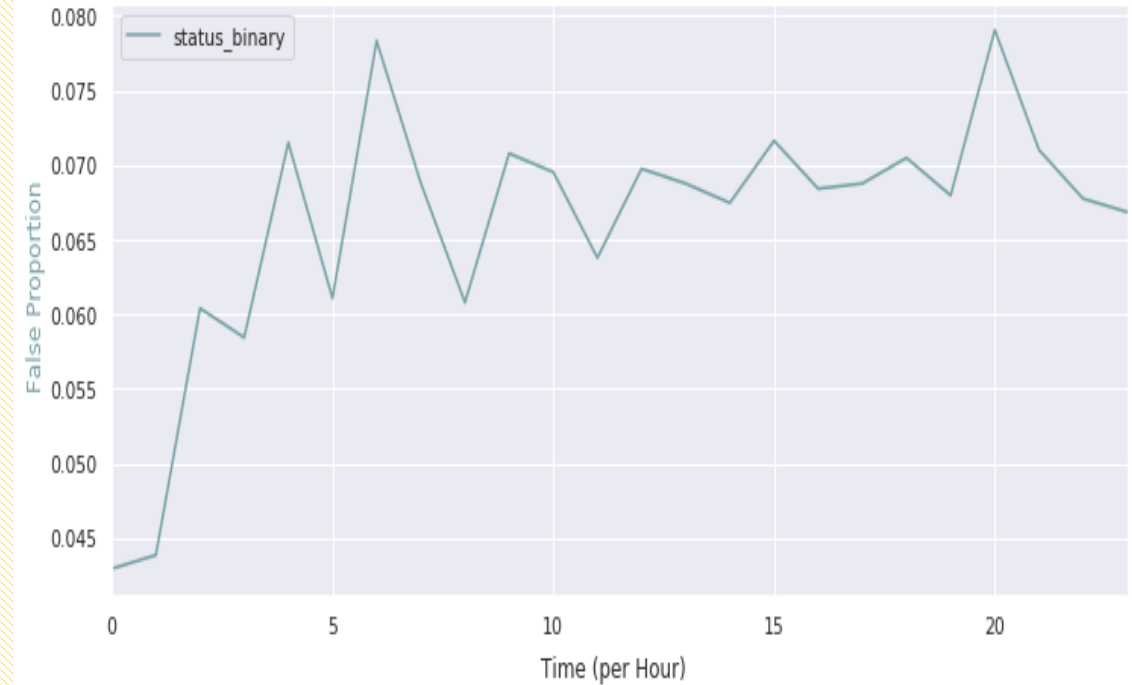
Analytics

HOURLY CLICK FREQUENCY Barplot



Count of clicks increases after 8 am till midnight and then falls


HOURLY CONVERSION RATIO Lineplot

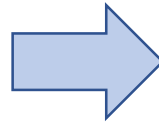


False Count Proportion to the total count received per hour

Feature Extraction



One Hot  Encoding



Curse of
Dimensionality

Feature Extraction

- ❖ Agent
- ❖ Category
- ❖ Country
- ❖ Ip
- ❖ Referer
- ❖ Timeat

Sr. No.	Type Of Features	Features Extracted
1	Time Features :	click_month
		click_day
		click_hour
		click_min
		click_sec
2	Count Features :	clicks_by_ip
		clicks_by_cntr
		clicks_by_referer
		clicks_by_category
		clicks_by_agent
3	Count feature By time :	clicks_per_minute
		clicks_per_hour
4	Count feature By Grouping :	ip_ref_cat_per_min
		ip_ref_cat_per_hr
		ip_cat_per_min
		ip_cat_per_hr
		ip_min_count
		agent_ip_count
		agent_cntr_count
5	Avg feature By Grouping :	ip_min_avg
		ip_ref_cat_min_avg
		ip_cat_min_avg

Pre Processing



Recursive Feature Elimination
on Decision Tree



UnderSampling after train test split



Scaling : Min-Max
Scaler between 0-1

	clicks_by_ip	clicks_by_cntr	clicks_by_referer	clicks_by_category	clicks_per_minute	clicks_per_hour	ip_cat_min_avg	clicks_by_agent	agent_ip_count	agent_cntr_count
clicks_by_ip	1	0.03	0.055	0.087	0.086	0.083	0.18	-0.016	0.38	0.044
clicks_by_cntr	0.03	1	-0.023	-0.07	-0.0092	0.013	0.16	-0.039	0.15	-0.11
clicks_by_referer	0.055	-0.023	1	-0.035	0.062	0.038	0.025	-0.0021	0.038	0.028
clicks_by_category	0.087	-0.07	-0.035	1	0.0079	-0.014	0.047	0.0092	0.097	0.055
clicks_per_minute	0.086	-0.0092	0.062	0.0079	1	0.44	0.084	0.0014	-0.00042	0.034
clicks_per_hour	0.083	0.013	0.038	-0.014	0.44	1	0.076	-0.0042	0.021	0.0092
ip_cat_min_avg	0.18	0.16	0.025	0.047	0.084	0.076	1	-0.012	0.47	-0.028
clicks_by_agent	-0.016	-0.039	-0.0021	0.0092	0.0014	-0.0042	-0.012	1	0.056	0.22
agent_ip_count	0.38	0.15	0.038	0.097	-0.00042	0.021	0.47	0.056	1	0.18
agent_cntr_count	0.044	-0.11	0.028	0.055	0.034	0.0092	-0.028	0.22	0.18	1

Correlation Matrix

Algorithms



Logistic Regression

- Base Model
- Simple Classification Model
- Fast Computation



CatBoost

- Gradient Boosting on Decision Trees
- Works directly on Categorical Data without encoding



ADA Boost with Decision Tree Classifier

- Boosting - weak learners are tweaked
- Flowchart-like tree structure
- Easy to understand and interpret










Random Forest Classifier

- Bagging Ensemble Learning
- Works on categorical & continuous data
- Not sensitive to overfitting
- Provides feature importance

Evaluation

Algorithm	Recall	Precision	Accuracy	AUC-ROC
Random Forest	94	54	94.44	94.14
CatBoost	88	57	95.04	91.84
ADA Boost + Decision Tree	92	36	89.22	90.64
Logistic Regression	65	12	66.67	65.95

Conclusion & Future Work

- 
- **Top Impacting Features - IP, Category, Referrer** 
 - **Sampling is important due to Bias in Data** 
 - **Feature Extraction and Selection Increased Accuracy** 
 - **Decision Trees works best for Fraud Detection** 
 - **More Parameter Tuning** 
 - **Model Combination and Ensemble Approach** 



THANK YOU