MSc. in Computing
Practicum Approval Form

Section 1: Student Details

Project Title:	Fake Ad Click Detection by identifying anomalies
Student ID:	18210455
Student name:	Gautam Shanbhag
Student email	gautam.shanbhag2@mail.dcu.ie
Chosen major:	MCM – Data Analytics
Supervisor	Mark Roantree
Date of Submission	20-Dec-2018

Section 2: About your Practicum

Click fraud is a black-hat technique of falsely inflating the number of clicks on a pay-per-click ad. Click fraud is usually driven by one of two incentives:

- Advertisers are trying to sabotage their competitors by driving up their costs and meeting their budget caps early on in the day
- Ad publishers are clicking on the ads displayed on their own sites to generate more revenue for themselves.

The topic of this practicum is **Fake Ad Click Detection by Identifying Anomalies.**

By Clustering, Classifying and Using ML Modelling, I intend to predict whether a user will download an app after clicking a mobile app advertisement by measuring the journey of a user's click across their portfolio, and to flag IP addresses who produce many clicks, but never end up installing apps. With this information, we can built an IP blacklist and device blacklist.

Section 3: Papers referred

[1] M. Taneja, K. Garg, A. Purwar and S. Sharma, "Prediction of click frauds in mobile advertising," 2015 Eighth International Conference on Contemporary Computing (IC3), Noida, 2015, pp. 162-166. doi: 10.1109/IC3.2015.7346672

[2] D. Antoniou et al., "Exposing click-fraud using a burst detection algorithm," 2011 IEEE Symposium on Computers and Communications (ISCC), Kerkyra, 2011, pp. 1111-1116. doi: 10.1109/ISCC.2011.5983854

[3] C. Walgampaya, M. Kantardzic, R. Yampolskiy, "Real Time Click Fraud Prevention using multi-level Data Fusion", WCECS 2010, October 20-22, 2010, San Francisco, USA

[4] R. Roy and K. T. George, "Detecting insurance claims fraud using machine learning techniques," 2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT), Kollam, 2017, pp. 1-6. doi: 10.1109/ICCPCT.2017.8074258

[5] S. Yaram, "Machine learning algorithms for document clustering and fraud detection," 2016 International Conference on Data Science and Engineering (ICDSE), Cochin, 2016, pp. 1-6. doi: 10.1109/ICDSE.2016.7823950

Section 4: Practicum Abstract

How does your proposal relate to existing work on this topic described in these papers? (200 words)

The paper [1] cited above is a similar use case where the author has used Recursive Feature Elimination technique for feature selection & Hellinger Distance Decision Tree for classification. Using Confusion matrix to evaluate the prediction. Antoniou in his paper [2] uses splay tree to evaluate click burst and detect fraudulent ip and click burst threshold. Walgampaya and his team [3] follow the Dempster-Shafer evidence theory to create a multi level data fusion mechanism to evaluate the frauds at real time using an online db. Papers [4] & [5] are reference papers used to identify the prediction techniques and methodologies in other domains.

What are the research questions that you will attempt to answer?

For this project, I intend to perform clustering analysis of the data set into informative groups depending upon various parametrical values present. Identifying anomalies to distinguish between users & bots, to identify regions, time of attack, devices used or os versions ideally preferred which are best suited for click frauds.

What software and programming environment will you use?

R & python libraries would be used for clustering, classification and machine learning. Open Refine would be used for data cleaning.

What coding/development will you do?

I intend to use Principal Component Analysis or Linear Discriminant Analysis for extracting relevant new features and then performing feature elimination based on weight to finalise the classifications. Then using various clustering algorithm like K-Means, Hierarchical Density Based, Gaussian Mixture to identify patterns and clusters dataset into informative groups and followed by Random Forest Algo for classification. I intend to train my ml with 70-30 dataset thus trying to optimise the efficiency of my program for better accuracy to detect fraud and real clicks.

What data will be used for your investigations?

The data set, which I will be using, is available on Kaggle by the name TalkingData AdTracking Fraud Detection Challenge (https://www.kaggle.com/c/talkingdata-adtracking-fraud-detection#Timeline).

How will the results be evaluated?

Analysis would be done using Python Language. I intend to use confusion matrix for evaluation. I have also read about Numenta Anamoly Benchmark technique which if feasible would be used to evaluate my ml algorithm.