**Capstone Project 2 Project Report**

**Problem Statement:**

Predicting whether a mobile ad will be clicked or not. In online advertising, click-through rate (CTR) is a very important metric for evaluating ad performance. As a result, click prediction systems are essential and widely used for sponsored search and real-time bidding.

**Data fields:**

* id: ad identifier
* click: 0/1 for non-click/click
* hour: format is YYMMDDHH, so 14091123 means 23:00 on Sept. 11, 2014 UTC.
* C1 -- anonymized categorical variable
* banner\_pos
* site\_id
* site\_domain
* site\_category
* app\_id
* app\_domain
* app\_category
* device\_id
* device\_ip
* device\_model
* device\_type
* device\_conn\_type
* C14-C21 -- anonymized categorical variables

**Exploratory Data Analysis:**

* The dataset has about 40M records across 24 predictor variables
* A resampled data of about 2M records has been considered for the modeling efforts
* The data types are as below:

Data columns (total 24 columns):

# Column Dtype

--- ------ -----

0 id float64

1 click int64

2 hour int64

3 C1 int64

4 banner\_pos int64

5 site\_id object

6 site\_domain object

7 site\_category object

8 app\_id object

9 app\_domain object

10 app\_category object

11 device\_id object

12 device\_ip object

13 device\_model object

14 device\_type int64

15 device\_conn\_type int64

16 C14 int64

17 C15 int64

18 C16 int64

19 C17 int64

20 C18 int64

21 C19 int64

22 C20 int64

23 C21 int64

* The dataset does not contain any NaN/NULL values
* Number of unique items in the dataset are as below:

id 2021448

click 2

hour 240

C1 7

banner\_pos 7

site\_id 3042

site\_domain 3521

site\_category 22

app\_id 3929

app\_domain 242

app\_category 27

device\_id 281246

device\_ip 971277

device\_model 5790

device\_type 5

device\_conn\_type 4

C14 2366

C15 8

C16 9

C17 426

C18 4

C19 67

C20 161

C21 60

* The dataset is heavily imbalanced. With 0 clicks being more prevalent than 1’s.

Chart, bar chart

Description automatically generated

* Most of the ads were shown from device type 1

Chart, bar chart

Description automatically generated

**Data Preprocessing:**

* The dataset didn’t have any NaN/Null values; hence Imputation was not required.
* The hour field in the dataset which is of type timestamp is converted to multiple different datetime attributes; like year, month, day, int\_hour, is\_weekend, is\_weekday.
* Columns such as "app\_id","device\_id","device\_ip","dt\_hour","site\_id","hour" were dropped from the dataset, since they contained ids.
* The categorical columns were label encoded.
* The dataset was divided into test, train and val sets.

**Modeling:**

Baseline Evaluation: Logsitic Regression

* Using accuracy as the metric, LR model was 83% accurate on the imbalanced dataset. Since the 0 responses outweigh the 1’s, the accuracy is skewed.
* Upon treating the imbalance, the LR showed accuracy of about 57%.

LightGBM:

Feature Selection:

The LightGBM model produced the below list of features and their scores in descending order of their importance.

ImportanceScore

site\_domain 723

app\_domain 255

site\_category 242

C14 233

C21 225

app\_category 181

C18 158

C19 152

C17 141

C20 118

C16 102

device\_model 97

banner\_pos 92

device\_type 76

int\_hour 57

device\_conn\_type 54

C1 34

day 30

C15 21

is\_weekday 9

year 0

month 0

is\_weekend 0

* As per the LightGBM algorithm, site\_domain is of the highest importance when it comes to predict the click. Followed by app\_domain, C14, site\_category etc.
* Features with 0 importance score are dropped from the final dataset.

Hyperparameter Optimization:

RandomSearch algorithm was used to tune the hyperparameters of the model.

The parameter distributions are below:

params = {

'learning\_rate': [0.05,0.1],

'min\_child\_weight': [1,3,5,7],

'min\_split\_gain': [0.05,0.1,0.3,0.5,0.7,0.9,1.0,5.0],

'reg\_alpha': [0.01, 0.05, 0.1, 0.5, 1.0, 3.0, 10.0, 15.0],

'subsample': [0.6,0.7,0.8,0.9,1.0],

'colsample\_bytree': [0.6,0.7,0.8,0.9,1.0],

'reg\_lambda': [0.01,0.05,0.1,0.5,1.0,2.0,3.0],

'min\_data\_in\_leaf': [30, 45, 60],

'max\_depth': [3,5],

'num\_leaves': [7, 31],

'early\_stopping\_rounds':[5]

}

The accuracy score of the model is about 83%, which is better as compared to the LogisticRegression model.