Criteo

Summary	Here we will present the end-to-end project we have undertaken to analyse the attribution trends for Cretio
Category	Assignment 1
Status	Published
Author	Rishvita Reddy,, Jugal Sheth
Codelap Preview	https://codelabs-preview.appspot.com/?file_id=1PtcVieMyXwH4htg9uu3CA_s9FEqIyHFUc5j9DQG1OXE#0

Introduction

We are given the task of analysing the Data for Criteo's Banner Attribution data by the clicks of customers with time stamps, and have to create a Dashboard

criteo AI Lab

About Us Publications Blog Posts Datasets Events Progr

Criteo AI Lab > Machine Learning > Criteo Attribution Modeling for Bidding Dataset

Criteo Attribution Modeling for Bidding Dataset

By: Criteo AI Lab / 10 Oct 2017

We recently published a paper on Attribution Modelling for Bidding at TargetAd and AdKDD 2017. As usual when we publish research work on Criteo related data, we publicly release the dataset used for the experiments in the paper.

Compared to other online advertising datasets such as CTR prediction dataset or Conversion Modelling datasets, this one contains attribution data, i.e if the conversions are attributed to Criteo or not by the advertisers. This important information opens a new set of applications in the area of real-time bidding and conversion modeling in online advertising.

In our paper we show that using a simple attribution model in the bidder can significantly improve bidding performance, providing better ROI for the advertiser and reduced ad exposure for the user, compared to baseline last-click bidder.

Content of this dataset

This dataset includes following files:

- **README.md**
- **criteo_attribution_dataset.tsv.gz**: the dataset itself (623M compressed)
- **Experiments.ipynb**: ipython notebook with code and utilities to reproduce the results in the paper. Can also be used as a starting point for further research on this data. It requires python 3.* and standard scientific libraries such as pandas, numpy and sklearn.

Data description

This dataset represents a sample of 30 days of Criteo live traffic data. Each line corresponds to one impression (a banner) that was displayed to a user. For each banner we have detailed information about the context, if it was clicked, if it led to a conversion and if it led to a conversion that was attributed to Criteo or not. Data has been sub-sampled and anonymized so as not to disclose proprietary elements.

Dataset

- The dataset contains 30 days of Criteo traffic data.
- Each row represents one impression that was displayed to a user
- The dataset have information about the context,if the banner was clicked, if the click leads to the customer purchasing the product within 30 days
- The data was compressed in tar.gz format and by using sublime text and python we received the 2.5gb data.

Here is a detailed description of the fields (they are tab-separated in the file):

- **timestamp**: timestamp of the impression (starting from 0 for the first impression). The dataset is sorted according to timestamp.
- **uid** a unique user identifier
- **campaign** a unique identifier for the campaign
- **conversion** 1 if there was a conversion in the 30 days after the impression (independently of whether this impression was last click or not)
- **conversion_timestamp** the timestamp of the conversion or -1 if no conversion was observed
- **conversion_id** a unique identifier for each conversion (so that timelines can be reconstructed if needed). -1 if there was no conversion
- **attribution** 1 if the conversion was attributed to Criteo, 0 otherwise
- **click** 1 if the impression was clicked, 0 otherwise
- **click_pos** the position of the click before a conversion (0 for first-click)
- **click_nb** number of clicks. More than 1 if there was several clicks before a conversion
- **cost** the price paid by Criteo for this display (**disclaimer:** not the real price, only a transformed version of it)
- **cpo** the cost-per-order in case of attributed conversion (**disclaimer:** not the real price, only a transformed version of it)
- **time_since_last_click** the time since the last click (in s) for the given impression
- **cat[1-9]** contextual features associated to the display. Can be used to learn the click/conversion models. We do not disclose the meaning of these features but it is not relevant for this study. Each column is a categorical variable. In the experiments, they are mapped to a fixed dimensionality space using the Hashing Trick (see paper for reference).

Key figures

- 2,4Gb uncompressed
- 16.5M impressions
- 45K conversions
- 700 campaigns

XSV:

Data Sampling and exploration

We have used XSV to sample and explore the data. The data had around 16 million rows and *2.5GB in size. By using XSV we randomly chose 20 campaigns and sampled the data to about 100k rows.

Data Exploration:Understanding the characteristics of data.

Total Counts and Headers:

Command Prompt

```
C:\Users\rishv\OneDrive\Northeastern\SEM3\Algorithmic Digital Marketing\Assignments\Assignment1-2\criteo_attribution_dat
aset>xsv count pcb_dataset_final.tsv
16468027

C:\Users\rishv\OneDrive\Northeastern\SEM3\Algorithmic Digital Marketing\Assignments\Assignment1-2\criteo_attribution_dat
aset>xsv headers pcb_dataset_final.tsv
1   timestamp
2   uid
3   campaign
4   conversion
5   conversion_timestamp
6   conversion_id
7   attribution
8   click
9   click_pos
10  click_nb
11  cost
12  cpo
13  time_since_last_click
14  cat1
15  cat2
16  cat3
17  cat4
18  cat5
19  cat6
20  cat7
21  cat8
22  cat9
```

Stats:

```
C:\Users\rishv\OneDrive\Northeastern\SEM3\Algorithmic Digital Marketing\Assignments\Assignment1-2\criteo_attribution_dat
aset>xsv stats pcb_dataset_final.tsv
field,type,sum,min,max,min_length,max_length,mean,stddev
timestamp,Integer,21662697362152,0,2671199,1,7,1315439.7525672794,769770.0361270809
uid,Integer,267401057636714,13,32458754,2,8,16237589.216772163,9373751.359085169
campaign,Integer,279692387314654,73322,32452111,5,8,16983964.58268292,9700052.225449245
conversion,Integer,806196,0,1,1,1,0.04895522699835389,0.2157744487836664
conversion_timestamp,Integer,1563478121339,-1,5262888,2,7,94940.22091042617,478966.63744865305
conversion_id,Integer,13073723398784,-1,32458519,2,8,793885.2297718121,4064784.1720765587
attribution,Integer,442424,0,1,1,1,0.026865634845022728,0.16169066920944608
click,Integer,5947563,0,1,1,1,0.3611582006758508,0.4803362934032628
click_pos,Integer,-13689309,-1,173,1,3,-0.8312658826708125,1.5322206203197763
click_nb,Integer,-10911742,-1,174,1,3,-0.6626016583527264,2.696254130340931
cost,Float,4829.340541025379,0.00001,0.0583448264308,5,17,0.000293255563702158,0.0008689670963295013
cpo,Float,3234792.6306751645,0.004,1.01631051174,5,16,0.19642866936420986,0.11863821555204436
time_since_last_click,Integer,4468753182520,-1,2592000,1,7,271359.35485891264,527310.8765171622
cat1,Integer,362745446968541,138937,30763035,6,8,22027256.02578344,12107310.172802933
cat2,Integer,241869634590874,138937,32440053,6,8,14687226.016261801,9122111.559270142
cat3,Integer,250858543819470,577,32457986,3,8,15233066.099508194,9847417.062123684
cat4,Integer,470410983488597,358249,32145478,6,8,28565108.831108402,2698653.3011883767
cat5,Integer,318482790236901,138937,32440053,6,8,19339462.47701044,11746115.865884133
cat6,Integer,248586757094861,138937,32440053,6,8,15095114.739296196,13406408.32272267
cat7,Integer,250864392790457,150,32458469,3,8,15233421.270832762,9002237.420802243
cat8,Integer,408663884955533,3225256,32440044,7,8,24815594.785912186,8254684.271293756
cat9,Integer,391463926601699,358246,32145483,6,8,23771149.18513154,7778014.745109545
```

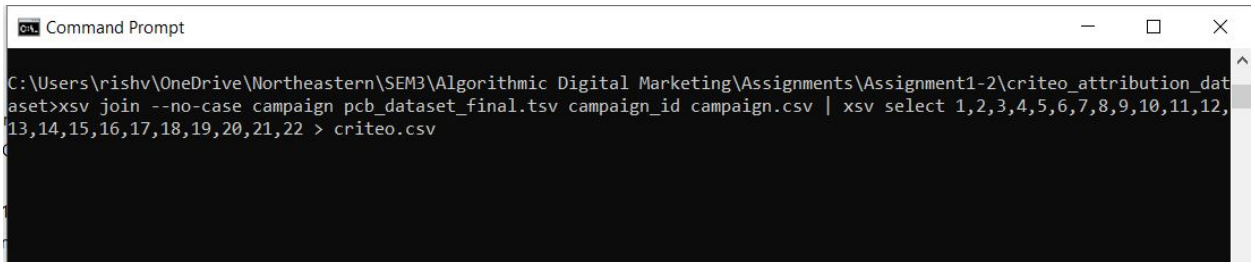
Sample Data:

```
C:\Users\rishv\OneDrive\Northeastern\SEM3\Algorithmic Digital Marketing\Assignments\Assignment1-2\criteo_attribution_dat
aset>xsv sample 10 pcb_dataset_final.tsv
timestamp,uid,campaign,conversion,conversion_timestamp,conversion_id,attribution,click,click_pos,click_nb,cost,cpo,time_
since_last_click,cat1,cat2,cat3,cat4,cat5,cat6,cat7,cat8,cat9
1529995,17721014,3073305,0,-1,-1,0,0,-1,-1,1e-05,0.004,-1,28928366,32440040,3781109,29196072,26611395,1973606,9312274,32
25256,29196072
244546,3999926,15184511,0,-1,-1,0,1,-1,-1,0.000397058830541,0.00619211307961,-1,25259032,28928366,10089519,29196072,5824
235,29196072,32081193,29196072,29196072
652368,18140180,28351001,0,-1,-1,0,1,-1,-1,0.00056232609117,0.0542864640501,255019,30763035,9312274,23322867,29196072,58
24236,3225256,16628728,29196072,21091108
668643,10950789,17686799,0,-1,-1,0,0,-1,-1,1.35598076742e-05,0.0699770487289,475434,30763035,9068207,4814486,29196072,32
440044,5824235,13595073,29196072,8661623
2462388,16993837,18498193,0,-1,-1,0,0,-1,-1,1e-05,0.318876550909,2579105,138937,9312274,146094,29196072,32440052,2892836
6,21280612,29196072,18291872
568185,4835629,5177431,0,-1,-1,0,0,-1,-1,0.000102434412521,0.0975390110284,-1,27093701,9312274,19228907,29196072,5824241
,1973606,18747137,26597096,29196072
1317239,1639136,7289590,0,-1,-1,0,1,-1,-1,9.86787589157e-05,0.193230268172,-1,138937,26597095,24485251,29196072,32440044
,29196072,1376925,29196072,29196072
403123,14919161,16823030,0,-1,-1,0,1,-1,-1,7.05882337163e-05,0.224521206906,-1,1973606,26597095,22173907,29196072,582423
7,1973606,20619782,9068204,29196072
2380940,23499696,18975813,0,-1,-1,0,0,-1,-1,3.24605681984e-05,0.149318964495,252489,30763035,9312274,22668889,29196072,5
824237,1973606,24731292,26597096,21091108
1675398,5851769,21257831,0,-1,-1,0,1,-1,-1,0.000251279619867,0.222847237151,251042,30763035,9312274,30356872,29196072,32
440047,29196072,9312274,29196072,21091108
```

Data Sampling:Extracting a subset of data

Step1: Extracted all the Campaigns to Excel sheet and selected 20 unique campaign id's to **campaign.csv**

Step2: Joined the original impressions data with multiple campaign id's to **campaign.csv** to extract all the impressions specific to these campaigns.



Trifacta:

Data Wrangling :

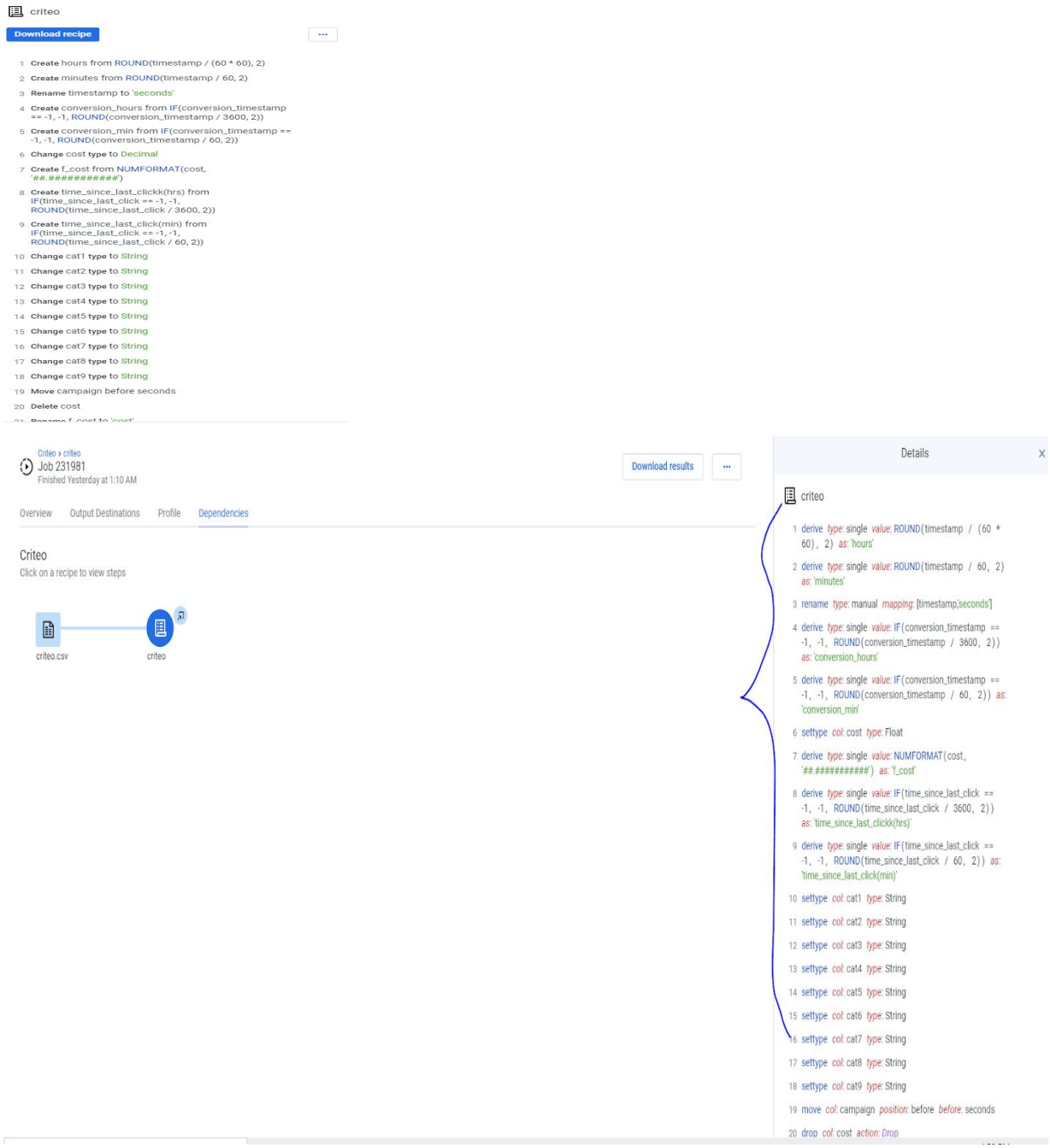
Transforming and mapping raw data.

- We used Trifacta to explore, transform, and enrich raw data into clean and structured formats

Key Transformations:

- Splitting of timestamp columns to hours,minutes and seconds
- Rounding off column values after decimals
- Changing the data type
- Conditional statements (IF) while converting
- Deleting columns that we do need for analysis

The overview of the recipe is shown below:

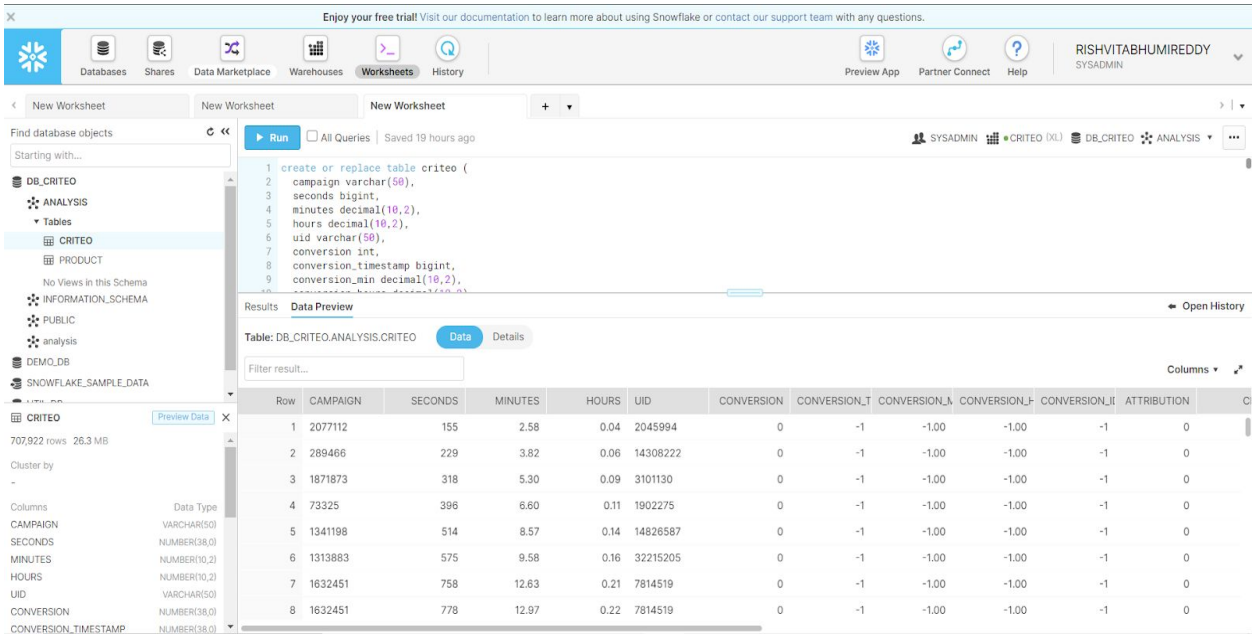


Snowflake:

Staging

- After wrangling the data we downloaded the data in Csv format.
- Created a table named Criteo in snowflake
- Imported the data into snowflake table

- The staging area can be used for further editing, storage and connecting to other cloud softwares we used Snowflake



Here we have staged the data by creating a new data warehouse(CRITEO) → data base(DB_CRITEO)→ Schema (Analysis)→Table(CRITEO)

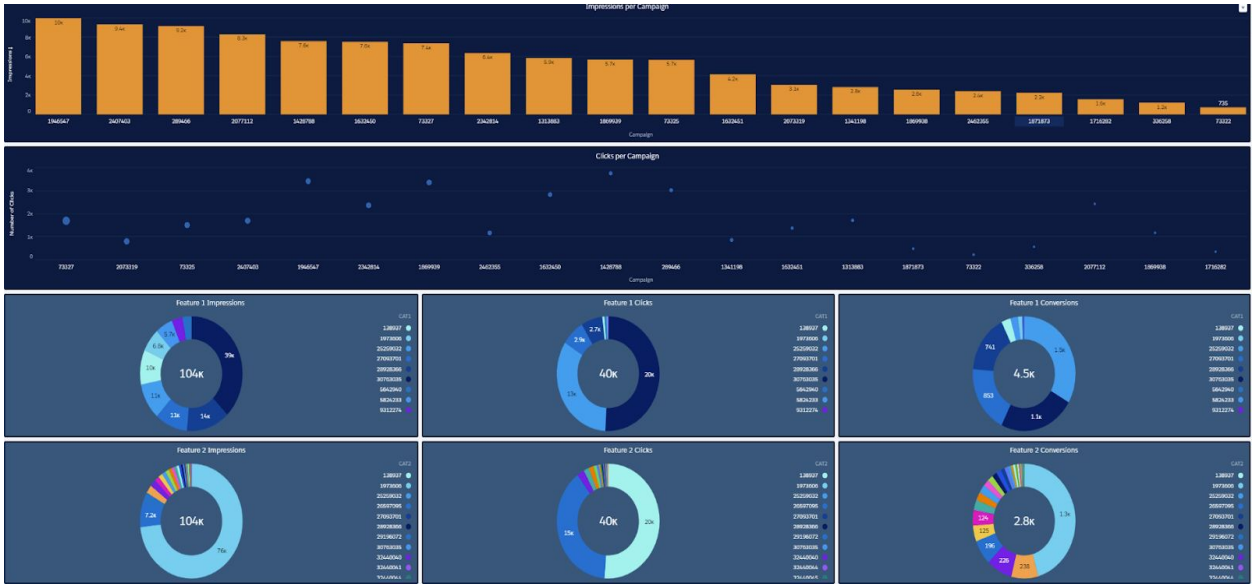
Einstein Analytics:

- The dashboards provide an at a glance summary of how different campaigns are performing.
- The dashboard helps in gaining insights on marketing analytics,KPI's and revenue metrics using visualization.

Dashboards:

RECOMMENDATION:

- This dashboard gives details about how various campaigns have performed with respect to views,clicks and conversions and thus helps us to draft the most ideal campaigns.It also gives insights on how each contextual feature attribute towards the decision making thus suggesting the ideal features to consider for each campaign.



[Reccomendation]

PRICING:

- This dashboard helps Criteo to develop comprehensive pricing strategies.
- It gives an overview of total cost for each campaign.
- The cost per 1000 impressions for each campaign.
- The dashboard gives the details of revenue earned by each campaign.

[Pricing]

ATTRIBUTION ANALYSIS:

- The dashboard highlights the attribution merits of different campaigns .
- The dashboard helps in understanding the conversion rates.
- The conversions that are attributed to Criteo based on each Campaign.
- The attribution ratio gives the details about total conversions that are attributed to Criteo with respect to the total number of clicks.

[Attribution]

DIGITAL MARKETING TRENDS:

The trends dashboard shows metrics of how each campaign has performed and varied over time, facilitating data-driven decision making.

The dashboard has viewers counts ,click counts, conversion counts with respect to campaigns.

Analysing various click positions.

On an average the hours taken since last click and last conversion.

Trends

KPI:

This dashboard gives an overview of KPI's :

Total Cost: The cost incurred by Criteo for running these ad campaigns.

CPM:It gives publishers the cost for every 1000 views (impressions) an advertisement receives.
Formula: **(Cost to the Advertiser / No. of Impressions) x 1000**

ROI: Return on Investment (ROI) is a performance measure used to evaluate the efficiency of an investment.

Formula: **(Total ad Revenue – Total ad campaign cost) / Total ad campaign cost**

CTR:Click-through rate is the ratio of users who click on a specific link to the number of total users who view an impression.

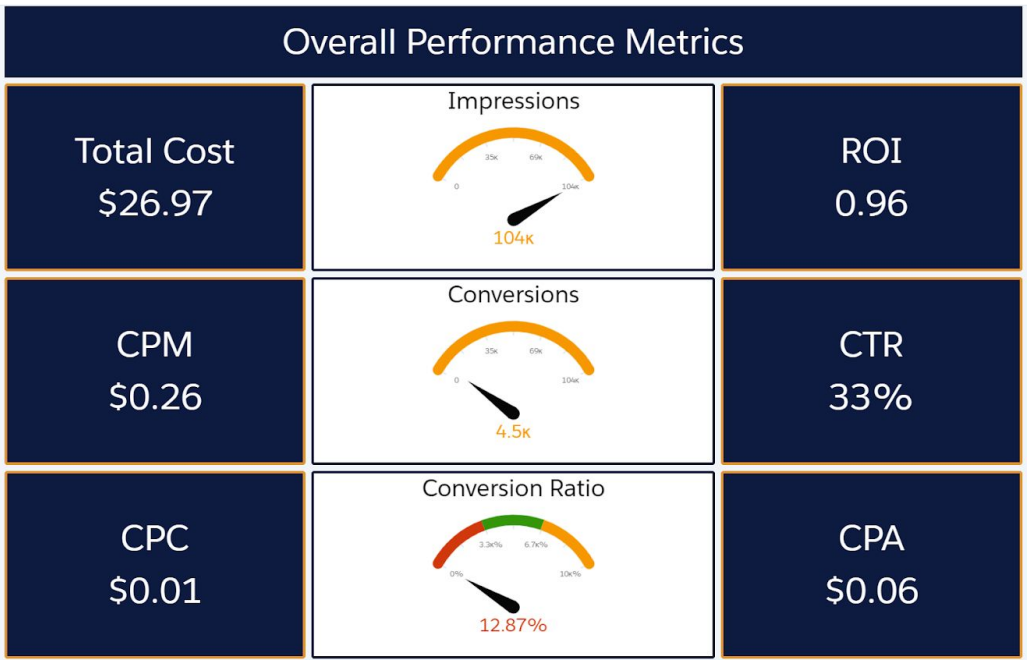
Formula: **(Total number of clicks/total number of clicks)*100**

CPC: It is calculated by dividing the cost to the advertiser by the number of clicks received on the ad

Formula:**Cost to the Advertiser / Number of Clicks**

CPA: It is calculated by dividing the cost to the advertiser with the number of actions received on the ad.

Formula: **Cost to the Advertiser / Number of Conversions**



KPI

Strengths and Weaknesses of Wrangling tools:

Trifacta:

Advantages:

- The GUI is very interactive and easy to understand and work.
- It offers suggestions on data transformations
- It helps us understand the column data patterns like min,max,unique values
- Flexibility in terms of importing and exporting data

Limitations:

- Upload limit for a single file is 100 MB.
- Results can be written to CSV or JSON format only.
- Integration with backend data storage is not supported. All files must be uploaded and downloaded from the application.
- Sharing and scheduling is not possible.

XSV:

Advantages:

- XSV is good with handling large datasets.
- It is much quicker compared to other platforms.

Limitations:

- We can only perform limited operations like join, count, stats etc
- Looking and working on multiple files is not easy
- GUI is not intuitive.

Questions to consider:

1. **Which columns are dimensions, which columns are measures?**

   Dimensions:
   - Uid
   - Campaign
   - conversion_id
   - click_pos
   - cat[1-9]

   Measures:
   - Timestamp
   - conversion
   - conversion_timestamp
   - attribution
   - click
   - click_nb
   - cost
   - cpo
   - time_since_last_click

2. **How would you generate new dimensions? What will you do to summarize measures?**
   - We can have additional description dimensions on various categorical variables as that would help us to understand the various features. We can use Trifacta, Pandas to generate new dimensions.
   - We can summarize the measure by SUM, COUNT, AVG

3. **Who would use this dashboard?**
    - Business Analysts: to understand the revenues and profits with respect to each campaign.
    - Digital Marketing Analysts: Analysing statistics and looking for ways that company can improve its online marketing.
    - Marketing analysts to compare the performance of each campaign.

4. **What value would be generated using this dashboard ?**
    - The number of clicks to campaigns and how many conversions were achieved.
    - The impressions that have been attributed to Criteo.
    - The revenue generated
    - Analysing various contextual features for targeted advertisements.
    - Conversions which could be attributed to Criteo and the company's investment to return on investment can be observed.