UNIVERSITÀ DEGLI STUDI DI MILANO

LA STATALE

Marketing Analytics Projects | Intarget Lab

Samaher Brahem | Franco Bonifacini | MohammedHadi ShahHosseini | Pooya Sabbagh

intarget:

June 2024 @ The University of Milan

What We Will Discuss

Your challenges, our solutions.

MARKETING ATTRIBUTION

The goal is to build a data driven attribution model to assign credits to Channels related to total conversions and to compare with the Last Click model

CUSTOMER SEGMENTATION

The goal is to cluster users based on the data provided in the table "orders"

CHURN RATE PREDICTION

The goal is to build a model to predict the Customer Churn probability for the customers based on the available data

1 MARKETING ATTRIBUTION

DATA PREPROCESSING

Handling Missing Values

We replaced missing values in the `transactionRevenue` and `transactionId` columns with appropriate default values.

Converting to Proper Types

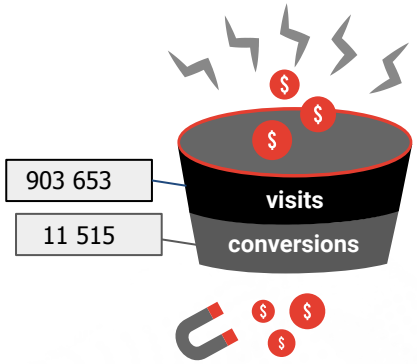
We converted the values in the `visitStartTime` column from Unix timestamps to a human-readable datetime format, allowing for easier manipulation and analysis of date and time information within the DataFrame.

Encoding conversions

We added a new column `conversion`, which indicates whether a conversion occurred for each entry. It sets the value to True if the `transactionRevenue` is greater than zero, indicating a revenue-generating transaction. If not, it sets the value to False, indicating no conversion.

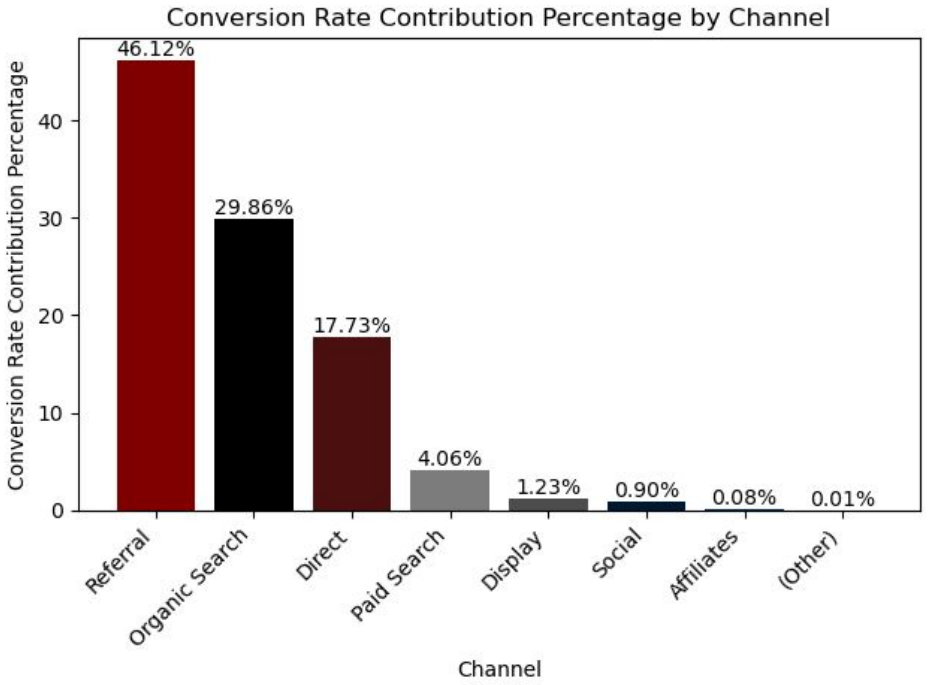
fullVisitorId	date	visitId	visitStartTime	channelGrouping	utm_source	utm_medium	transactionRevenue	transactionId	conversion
9416183380303809617	2016-12-17	1481990278	2016-12-17 15:57:58	Organic Search	ask	organic	0.0		False
342964634359205532	2016-12-17	1481999067	2016-12-17 18:24:27	Organic Search	ask	organic	0.0		False
294887852901730140	2016-12-17	1481997033	2016-12-17 17:50:33	Organic Search	ask	organic	0.0		False
2904105592463883270	2016-12-17	1482038661	2016-12-18 05:24:21	Display	dfa	cpm	0.0		False
8140805711484568839	2016-12-17	1482030051	2016-12-18 03:00:51	Display	dfa	cpm	0.0		False

EXPLORATORY DATA ANALYSIS



Conversion Rate: 1.27%

channelGrouping	total_conversions
Referral	5311
Organic Search	3438
Direct	2042
Paid Search	468
Display	142
Social	104
Affiliates	9
(Other)	1



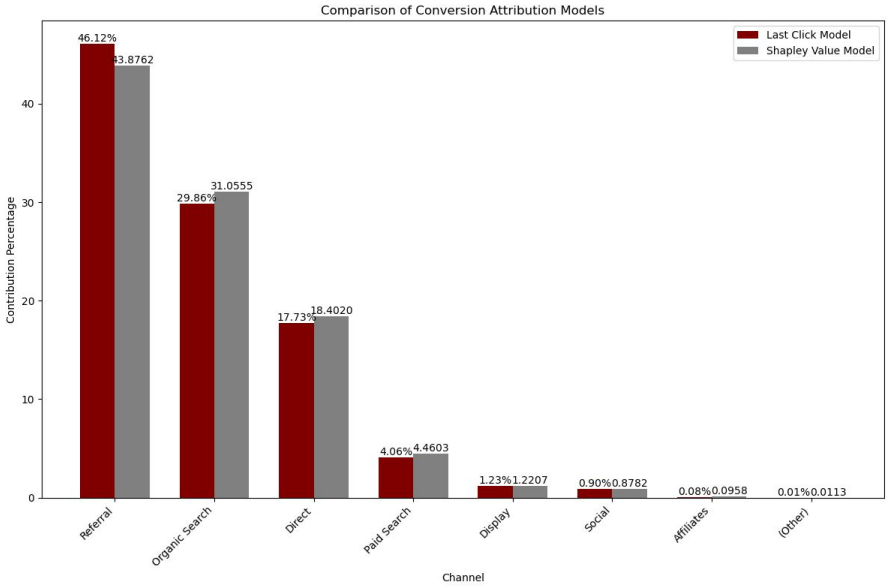
The **Referral** channel contributes the most to the total conversions, accounting for 46.12% of all conversions. **Organic Search** is the second largest contributor with 29.86%, followed by **Direct** with 17.73%. **Paid Search** channels contribute 4.06% to the total conversions. The **Display** channel accounts for 1.23%, and **Social** contributes 0.90%. **Affiliates** have a minimal contribution of 0.08%.

SHAPLEY VALUE MODEL

What it is?

The Shapley Value is a method from cooperative game theory used to fairly distribute the total gain (or payoff) among players based on their contributions. In marketing attribution, the "players" are different marketing channels or touchpoints that a user interacts with before converting.

fullVisitorId	channels	conversion	number of touchpoints
4823595352351	[Organic Search]	False	1
5103959234087	[Organic Search]	False	1
10278554503158	[Organic Search]	False	1
20424342248747	[Organic Search]	False	1
26722803385797	[Organic Search]	False	1
27376579751715	[Organic Search]	False	1
33471059618621	[Social]	False	1
35794135966385	[Direct]	False	1
39460501403861	[Social]	False	1
40862739425590	[Paid Search, Paid Search]	False	2



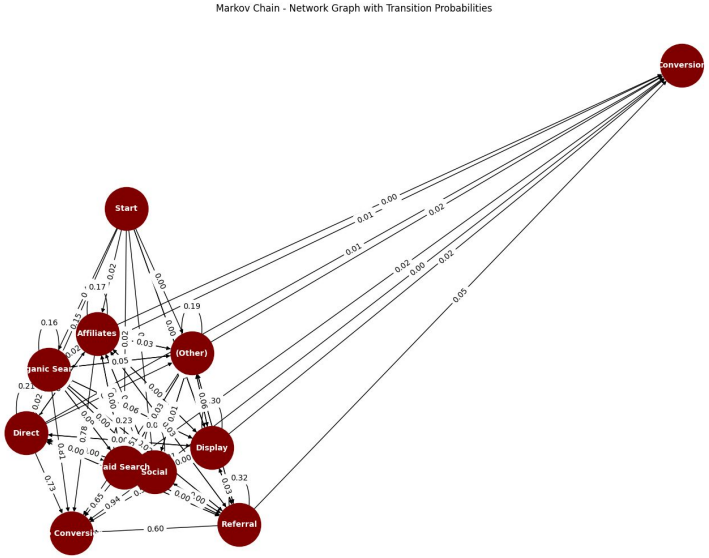
How it works?

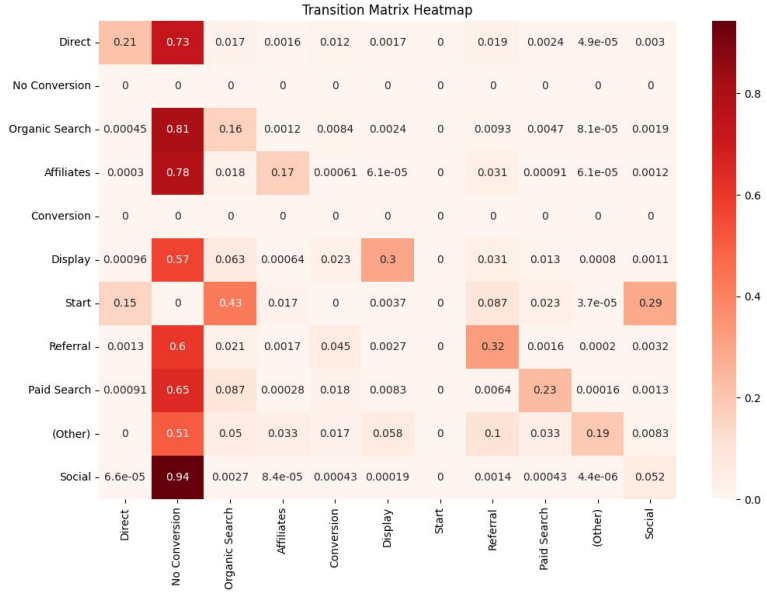
1. Identify All Touchpoints.
2. Calculate Marginal Contribution.
3. Calculate Average Contributions.
4. Distribute Credit.

DATA-DRIVEN (MARKOV CHAIN) MODEL

What it is?

Markov chain attribution models credit marketing touchpoints based on the likelihood of a customer moving from one touchpoint to another before converting. They analyze sequential interactions to evaluate each touchpoint's influence on conversion, offering insights into channel effectiveness.

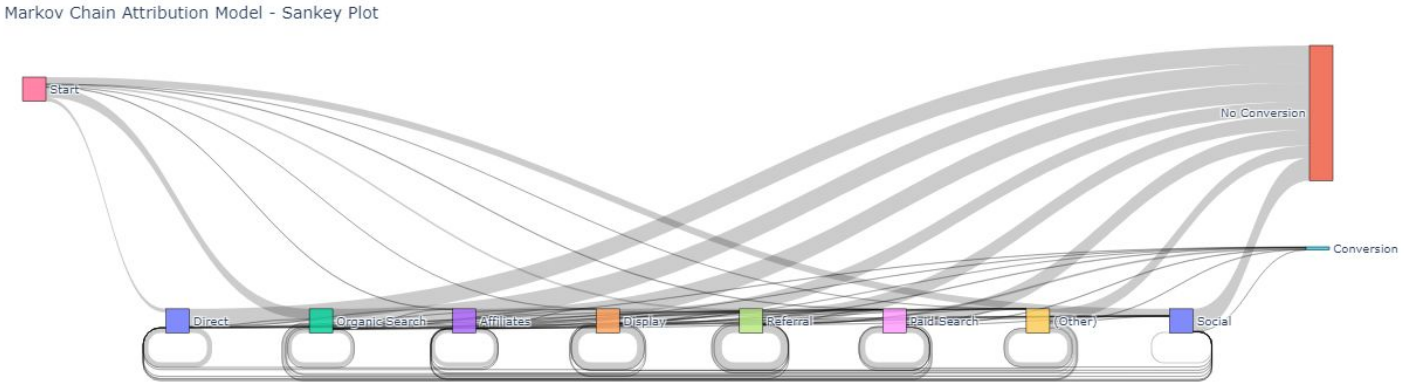


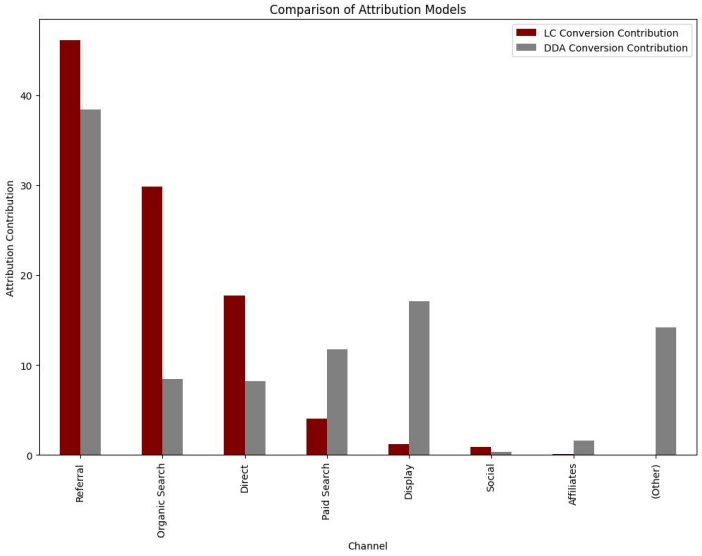


How it works?

1. Define touchpoint sequence
2. Calculate transition probabilities
3. Construct Markov chain model
4. Apply to customer journeys
5. Optimize marketing strategies

DATA-DRIVEN (MARKOV CHAIN) MODEL





Comparison

2 CUSTOMER SEGMENTATION

DATA PREPROCESSING

Final variables

q_purchases: number of purchases by customer

avg_units: average units per purchase

avg_value (monetary): average value per purchase

cheapest_unit: cheapest unit bought by customers

most_exp_unit: most expensive unit bought by customers

recency: last purchase counting from last day of March 2022

frequency: number of purchases in a time window of **3 months**

#_countries: number of countries where customers bought

Handling Missing Values

Cases where price per unit is 0: this cases were deleted for the project, nevertheless in a daily job is recommendable to check why (usually are cases with gifts or discounts)

	n_days_prepurchase
count	117606.00000
mean	117.53965
std	97.34904
min	1.00000
25%	39.00000
50%	95.00000
75%	172.00000
max	438.00000

	CustomerId	q_purchases	avg_units	avg_value	cheapest_unit	most_exp_unit	recency	frequency	#_countries
0	+++SJgx/2IJ+dXq7vF8COg==	1	3.0	172.21	10.48	35.95	123	0.0	1
1	+++aKiAiXhTfaqCLC/kyWA==	1	2.0	29.94	5.99	8.98	380	0.0	1
2	++/G67YHZTMKdpvANeYPLw==	1	46.0	884.61	1.35	22.50	441	0.0	2
3	++/GTDXvJzF11ZIUz81SPg==	1	3.0	59.91	5.99	20.97	216	0.0	1
4	++0Dxza60/nPDbfORBYuuA==	1	6.0	632.94	13.46	44.91	33	1.0	1

EXPLORATORY DATA ANALYSIS

Although most variables have **outliers**, we considered that they are **important if and only if the distance between other values is not significant**.

Cases where the **distance with the next nearest value is considerably high**, then they were **dropped**.

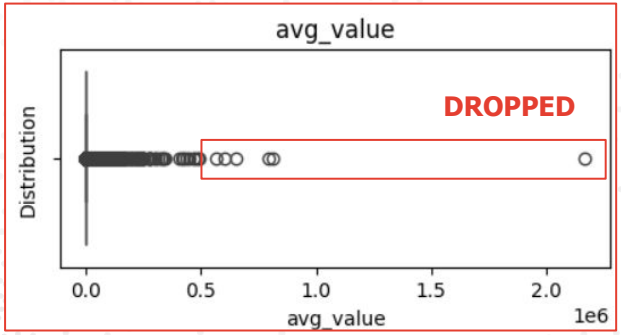
Nevertheless, in a daily job it will be highly recommended to analyze in detail this outliers as they may be interesting customers.

Finally, as **variables were not normally distributed**, for clustering we used **standard normalization**, as it outputted the **best results when clustering**.

The different normalization methods we used and compare were:

- MaxMin.
- **Standard (meand and std)**.
- Applying log to the variables.

>500: 11 records dropped.



>500.000: 6 records dropped.

RFM MODEL

RFM Quartiles

Cut the dataset into **quartiles** for each variable:

- **recency** (*lower values is preferred*)
- **frequency** (*higher values is preferred*)
- **monetary** (*higher values is preferred*)

RFM Score

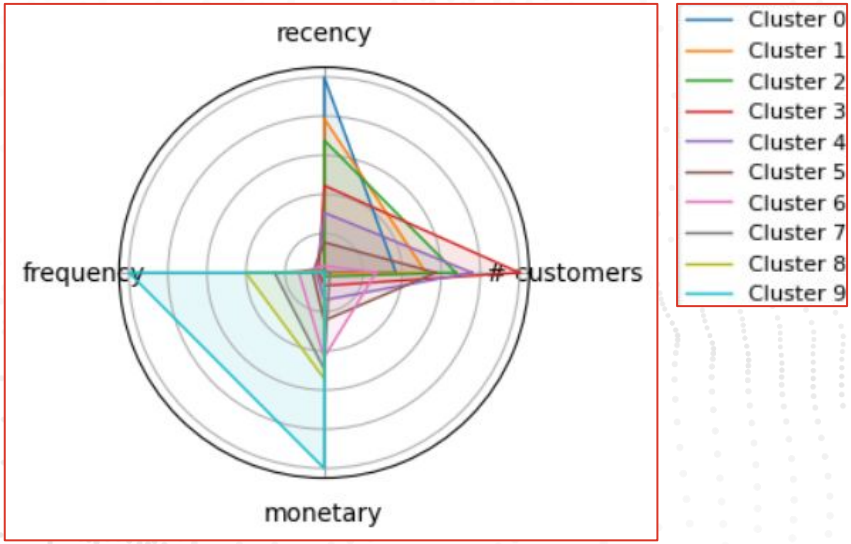
- Sum the three values obtained by customer.
- This is the **RFM Score**:
  - **Lower score** means a **valuable customer = retain**.
  - **Higher score** means **low-value customers = opportunity.**

Polar Plot

- Visual plot with the **resulting clusters**.
- Using a **polar plot** helps to **better differentiate the customers** according to their characteristics.

RFM CLUSTERS

	# customers	percentage	recency	frequency	monetary
0	46272	0.088111	404.475039	0.000000	43.639250
1	64967	0.123711	327.193637	0.000000	77.720261
2	83676	0.159336	286.287287	0.007374	164.741479
3	123275	0.234741	201.684916	0.235903	444.451197
4	93709	0.178441	150.897448	0.334525	882.784259
5	71164	0.135511	95.293660	0.497738	1511.553437
6	33950	0.064648	51.778203	1.109102	2635.808272
7	5063	0.009641	44.126802	2.175588	2985.793324
8	1431	0.002725	40.719776	3.433263	3282.149649
9	1646	0.003134	38.728433	8.554678	6019.418567



K-MEAN MODEL

Define variables

For this model some comparisons were done:

- **RFM:** using same variables as before, but RFM cluster provided a better result.
- **RFM+Variables**: add avg_units, #_countries.

K Definition

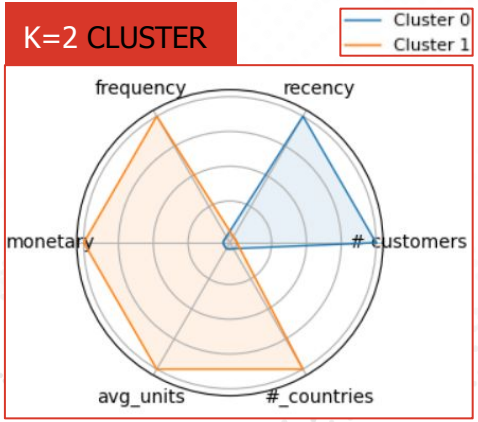
- To define number of clusters (k) an **elbow** and **silhouette method** were done.
- **cheapest_unit** and **most_exp_unit** were excluded because of low silhouette score.
- **Optimal K were 2 and 3** (same score).

Clustering Plot

When clustering, the final conclusion was:

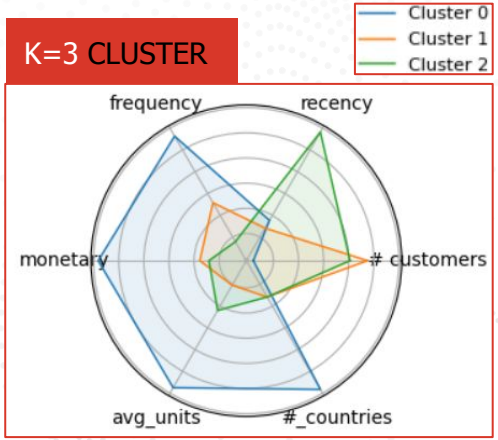
- For a **first marketing plan**, **K=2** is enough.
- **K=3** should be used for a **second step**.
- **RFM clustering** is recommended **for more detailed plans**.

K=2 CLUSTER



	# customers	percentage	recency	frequency	monetary	avg_units	#_countries
0	517867	0.986126	214.215362	0.295765	672.198526	5.286374	1.000000
1	7286	0.013874	120.903376	1.504666	5082.929347	7.940783	2.117897

K=3 CLUSTER



	# customers	percentage	recency	frequency	monetary	avg_units	#_countries
0	7267	0.013838	121.059172	1.484106	4850.905172	7.941986	2.12082
1	279866	0.532923	101.281685	0.547916	841.886616	4.947278	1.00000
2	238020	0.453239	346.991564	0.000008	480.113813	5.685260	1.00000

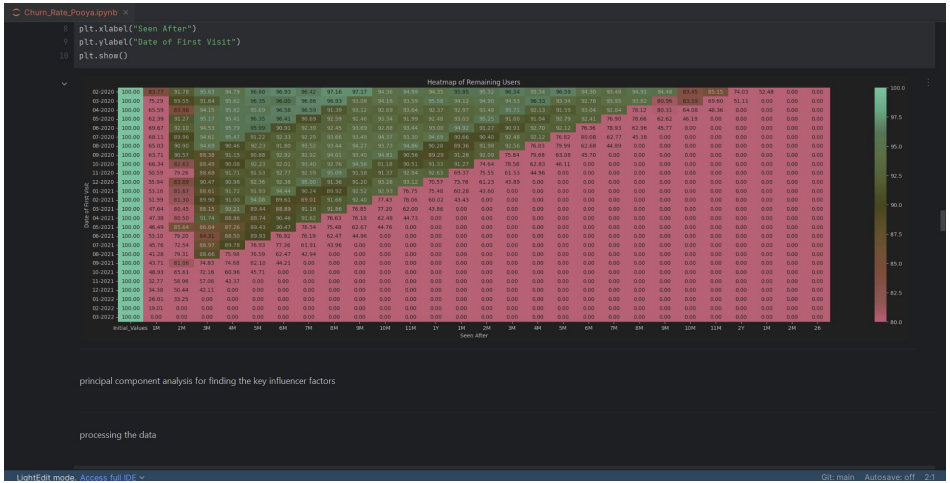
3 CHURN RATE PREDICTION

CHURN
MODEL

Final variables

	Initial_Values	1M	2M	3M	4M	5M	6M	7M	8M	9M	10M	11M	1Y	1M
02-2020	100	83.765228	91.782924	95.633835	94.786824	96.604485	96.932089	96.423135	97.163285	97.171254	94.364673	94.992183	94.349663	95.953253
03-2020	100	75.290823	89.553980	91.639890	95.615504	96.352704	95.998628	96.856918	96.926396	93.075767	94.160927	93.592638	95.083488	94.124390
04-2020	100	65.590056	83.981973	94.149340	95.823871	95.688082	96.584732	96.592177	91.389586	93.121478	92.692932	93.641366	92.374039	92.973856
05-2020	100	62.386845	91.274230	95.166110	95.414814	96.350332	96.412809	90.686179	92.589211	92.460053	93.339711	91.994153	92.482741	93.030925
06-2020	100	69.672933	92.096055	94.525533	95.793645	95.994563	90.914548	92.391948	92.450756	93.689426	92.877240	93.440233	92.996782	94.922807
07-2020	100	68.107444	89.958111	94.607230	95.469048	91.215128	92.328314	92.286101	93.660199	93.485807	94.370415	93.301602	94.694599	90.655078
08-2020	100	65.027275	90.898785	94.693411	90.458107	92.230041	91.796464	93.522499	93.441618	94.273532	93.727387	94.862848	90.276637	89.363731
09-2020	100	63.713934	90.576563	88.378050	91.169250	90.878607	92.926890	92.926422	94.010380	93.398117	94.809927	90.563218	89.287567	91.276337

Coding process and modeling (Notebook)



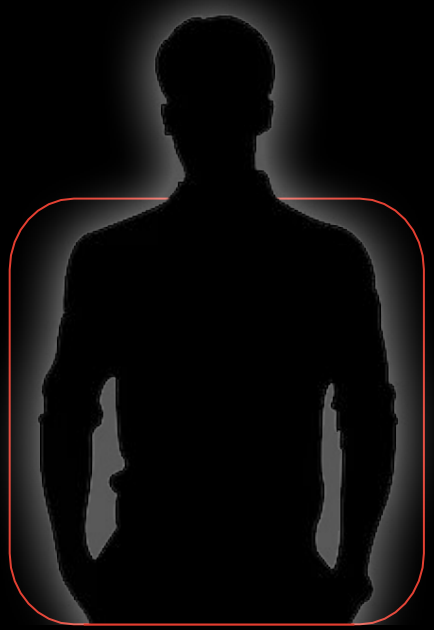
FOR MOUT OF THE BOX IDEAS, HERE'S OUOUT OF THE BOX TEAM

Quite Literally 👀

in/samaherbrahem
github.com/SamaherUNIMI
samaherbrahem.com

in/franco-bonifacini
github.com/Boni1995

in/mohammadhadishahhosseini
github.com/mohmmadhadi



Pooya Sabbagh