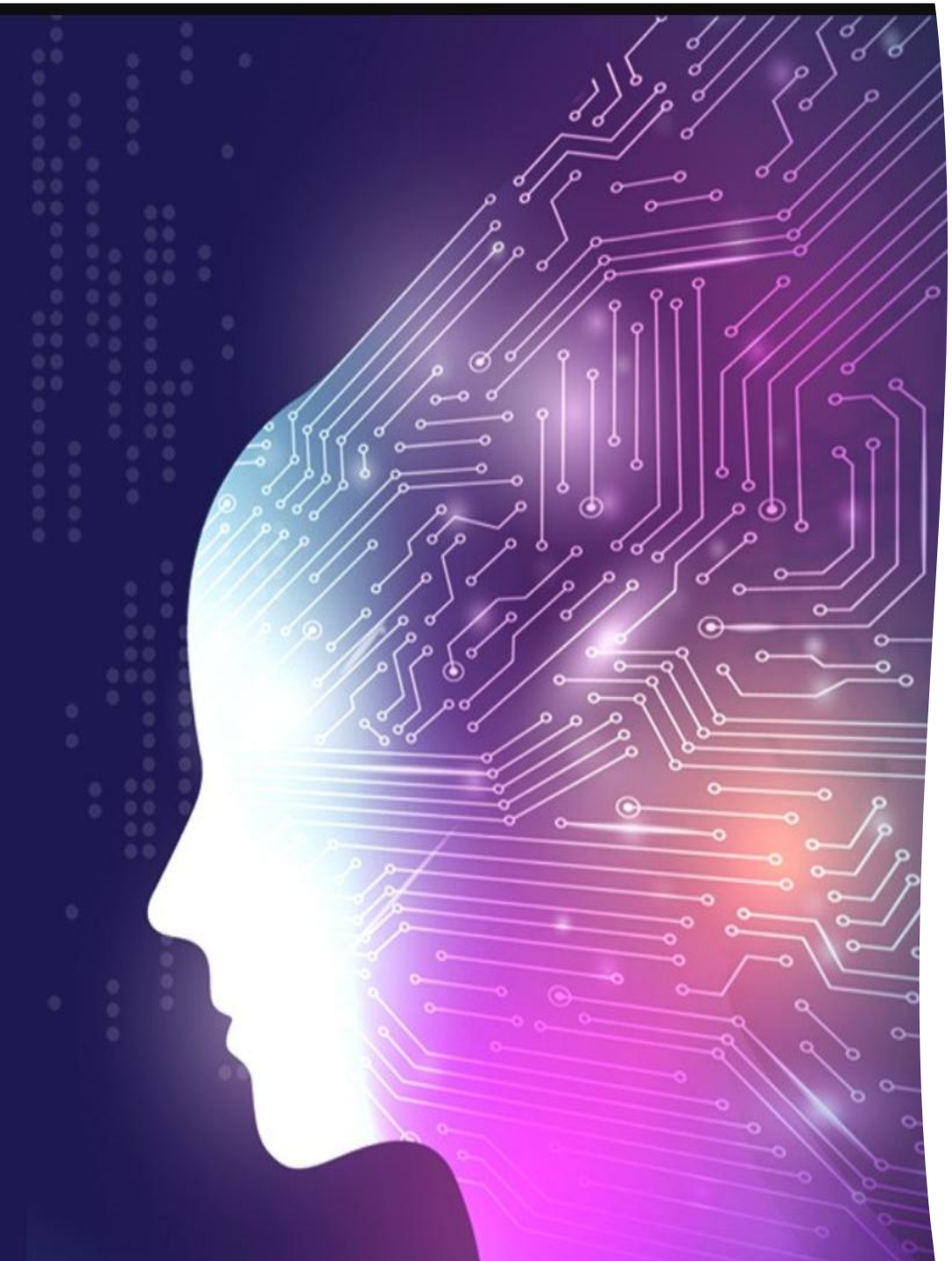
Case Study Sharing

Monthly Sales Prediction of Remy Products in JD Platform

Qin Lu

30th July 2021



Project Objective

- To use dynamic comment number data from the JD platform to predict and infer the monthly sales of all alcoholic items on the JD platform.

- This is akin to a task where our VIP client deliberately tested our capabilities—an endeavor they knew was practically infeasible but insisted on pursuing. Although it began as an exploration and experiment, the client ultimately ended up adopting this project, intending to use it as a reference to adjust their marketing strategies and inventory management.

Challenges

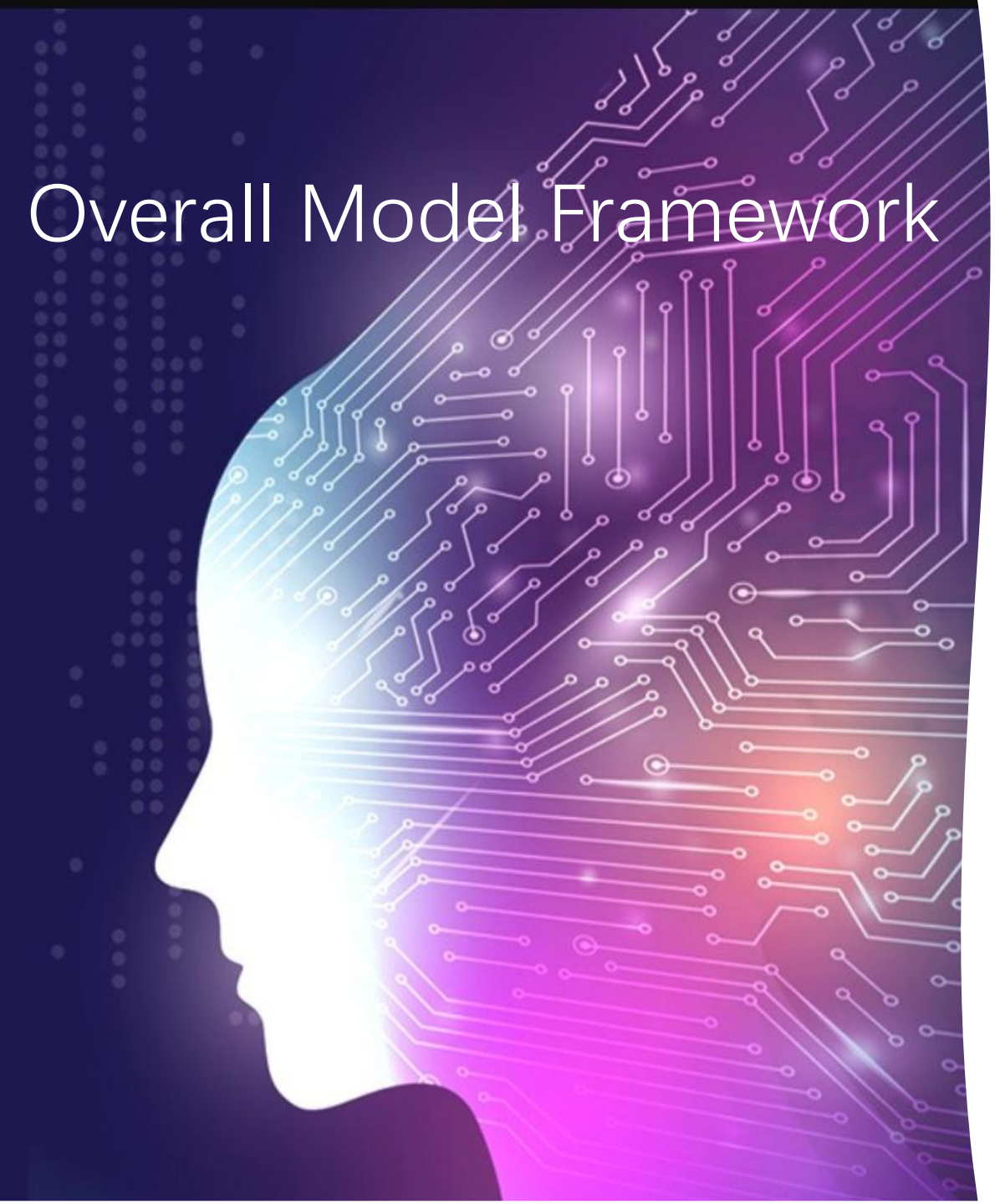
(1) The primary difficulty: to predict sales using comment data—it is uncertain whether the results are usable or meaningful.

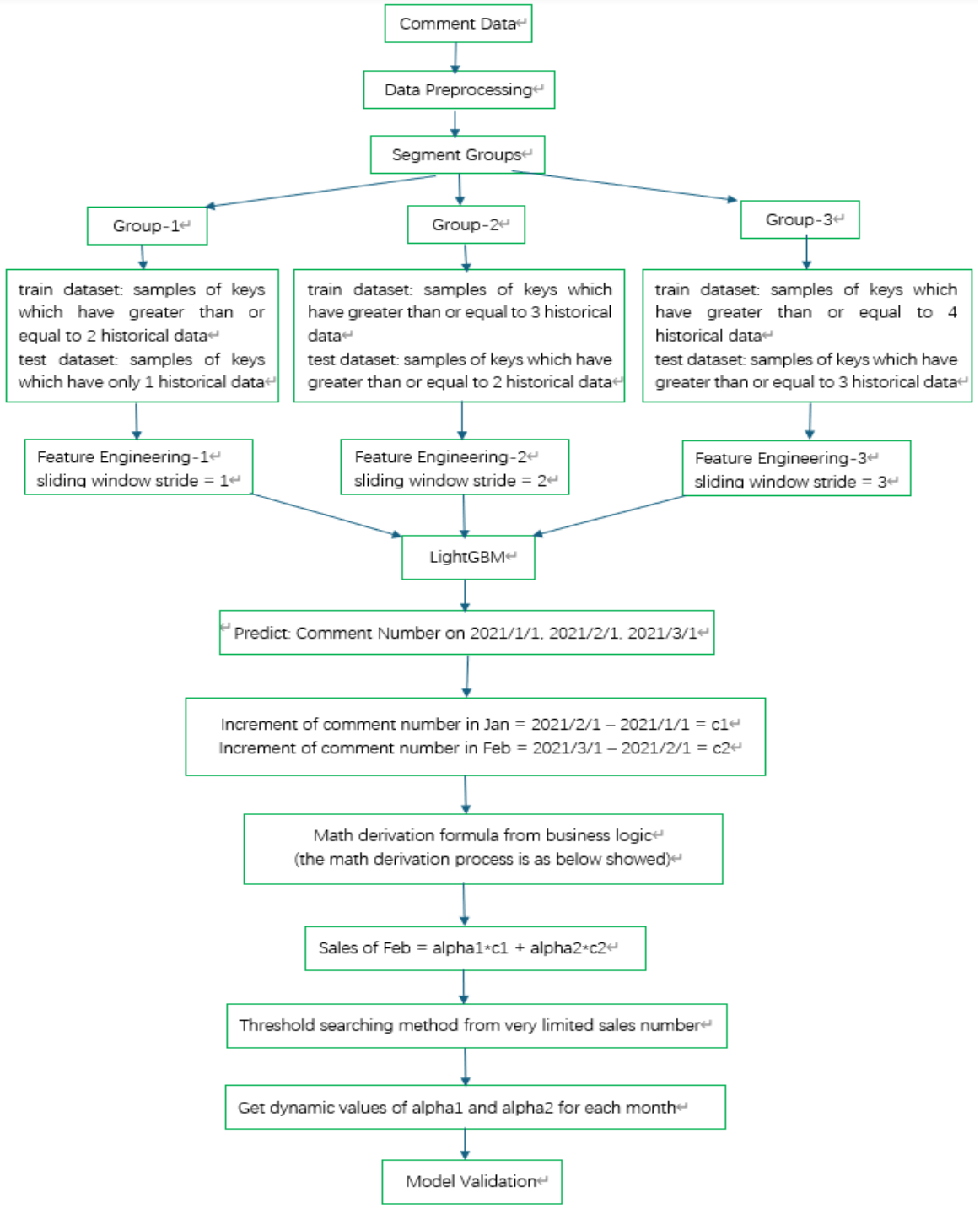
(2) Due to system maintenance or data migration, the comment dataset is relatively small and contains a high prevalence of dirty or noisy data.
For example, for a same key, only three samples might be extracted on 2025/1/1, 2025/1/5, and 2025/1/6.

(3) Review values exhibit a long-tail distribution, leading to highly imbalanced data.

(4) On the JD platform, keys corresponding to static product features may change dynamically;
For example, a same key representing a wine product could correspond to red wine for three months and later be replaced with white wine by the seller.

Overall Model Framework





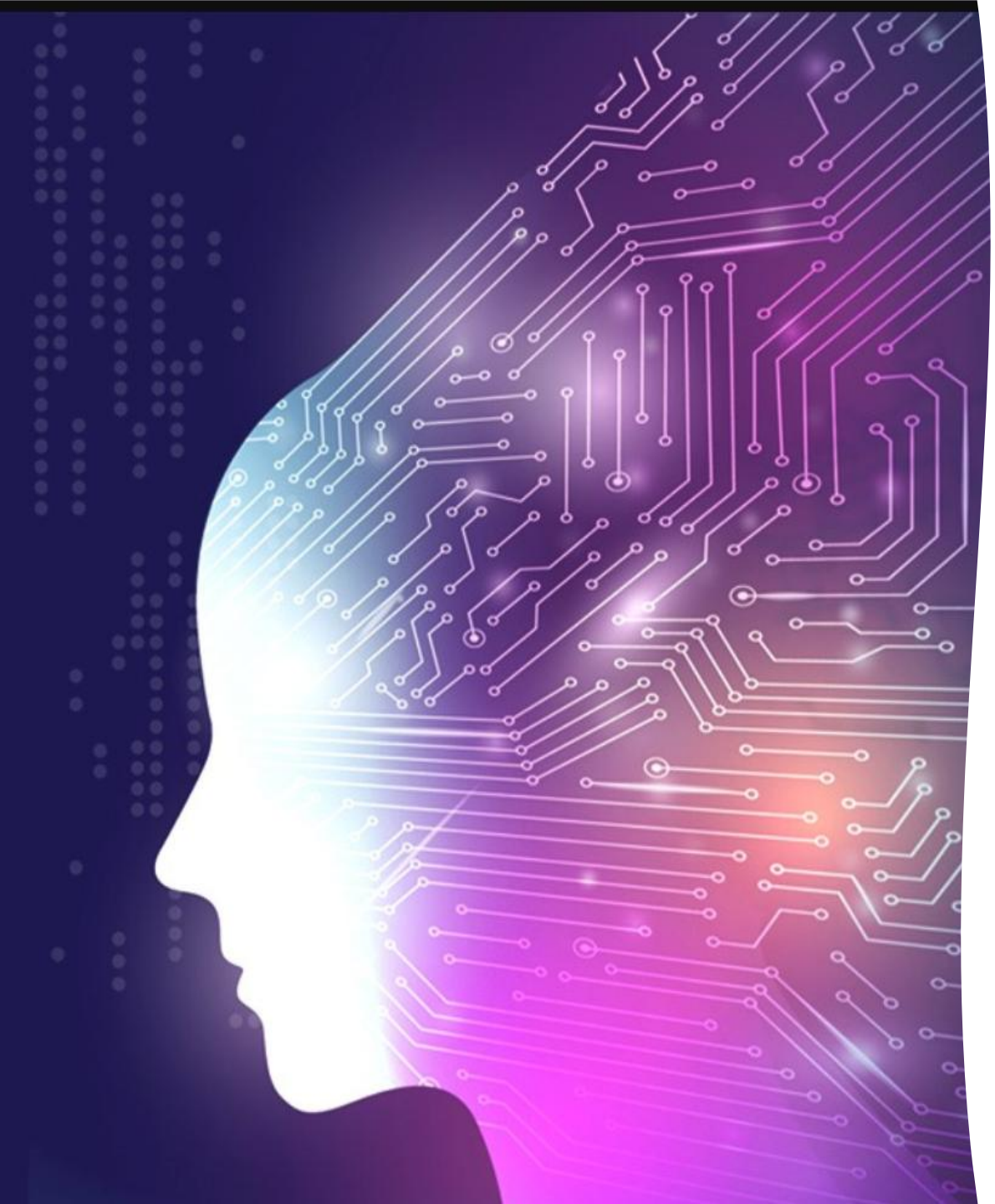
An example of sliding window method with stride=2

Original comment data samples:

KeyID	Date	Comment Number
key-1	2021/2/1	11
key-1	2021/1/22	10
key-1	2021/1/16	10
key-1	2021/1/2	9
key-2	2021/1/21	3
key-2	2021/1/10	1

After preprocessing:

Type of Dataset	KeyID	Date	Date-1	Date-2	Comment Number (y)
Train & Valid	Key-1	2021/2/1	2021/1/22	2021/1/16	11
	Key-1	2021/1/22	2021/1/16	2021/1/2	10
Test	Key-2	2021/2/1	2021/1/21	2021/1/10	?



How to get sales number from comment number?

- Because the JD platform allows users to leave a comment within 60 days after purchasing a product.

- Therefore, from the business logical, we got the math formula:

Sale Volume of April = w1*c1 + w2*c2 + w3*c3 (A)

* **C1** = incremental number of comments in April
* **W1** = The ratio/weight of users who both left a comment and purchased the Remy product in April
* **C2** = incremental number of comments in March
* **W2** = The ratio/weight of users who both left a comment in April and purchased the Remy product in March
* **C3** = incremental number of comments in February
* **W3** = The ratio/weight of users who both left a comment in April and purchased the Remy product in February

- Above is the first step for a mathematical derivation formula to show how from comment number prediction to sales number.

- But it is still too complicated because we need to predict the comment number on 2021/5/1, 2021/4/1, 2021/3/1, 2021/2/1 if we need to get sales of April.

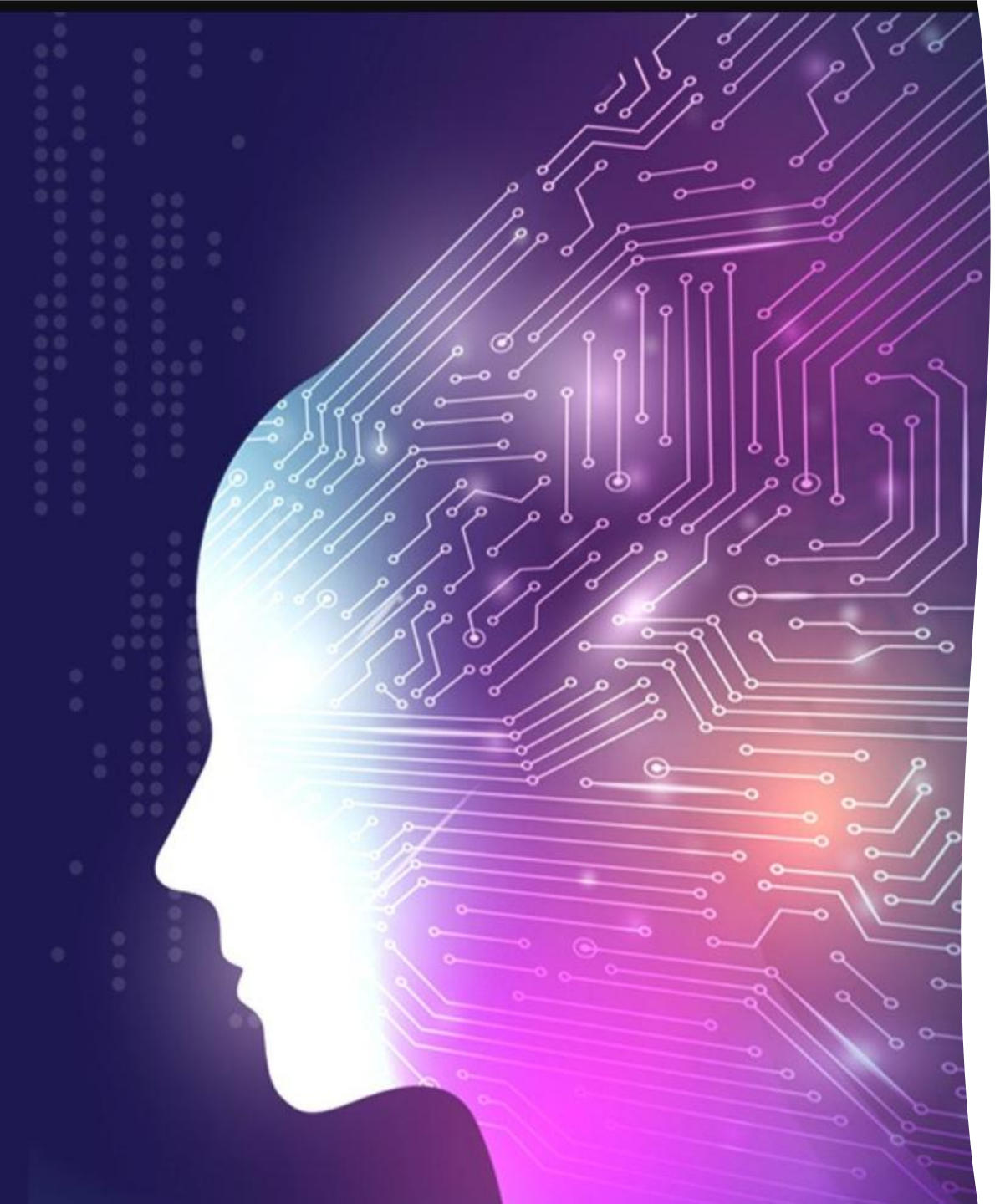
- Then, we derive formula A continuously to try to find easier formula:

Sale Volume of April
= w1*c1 + w2*c2 + w3*c3
= w1*c1 + w2*c2 + w3*constant*c1
= (w1+w3*constant)*c1 + w2*c2
= alpha1*c1 + alpha2*c2

- Therefore, we define the formula:

Sale Volume of April = alpha1*c1 + alpha2*c2 (B)

- It means that we just predict comment numbers of 2021/5/1, 2021/4/1, 2021/3/1 to derive sales volume of April.



- Also, we could get further from formula (B):

Sale Volume of April

= w1*c1 + w2*c2 + w3*c3

= w1*c1 + w2*ratio1*c1 + w3*ratio2*c1

= (w1 + w2*ratio1 + w3*ratio2) * c1

= alpha*c1

- Therefore, we define the formula:

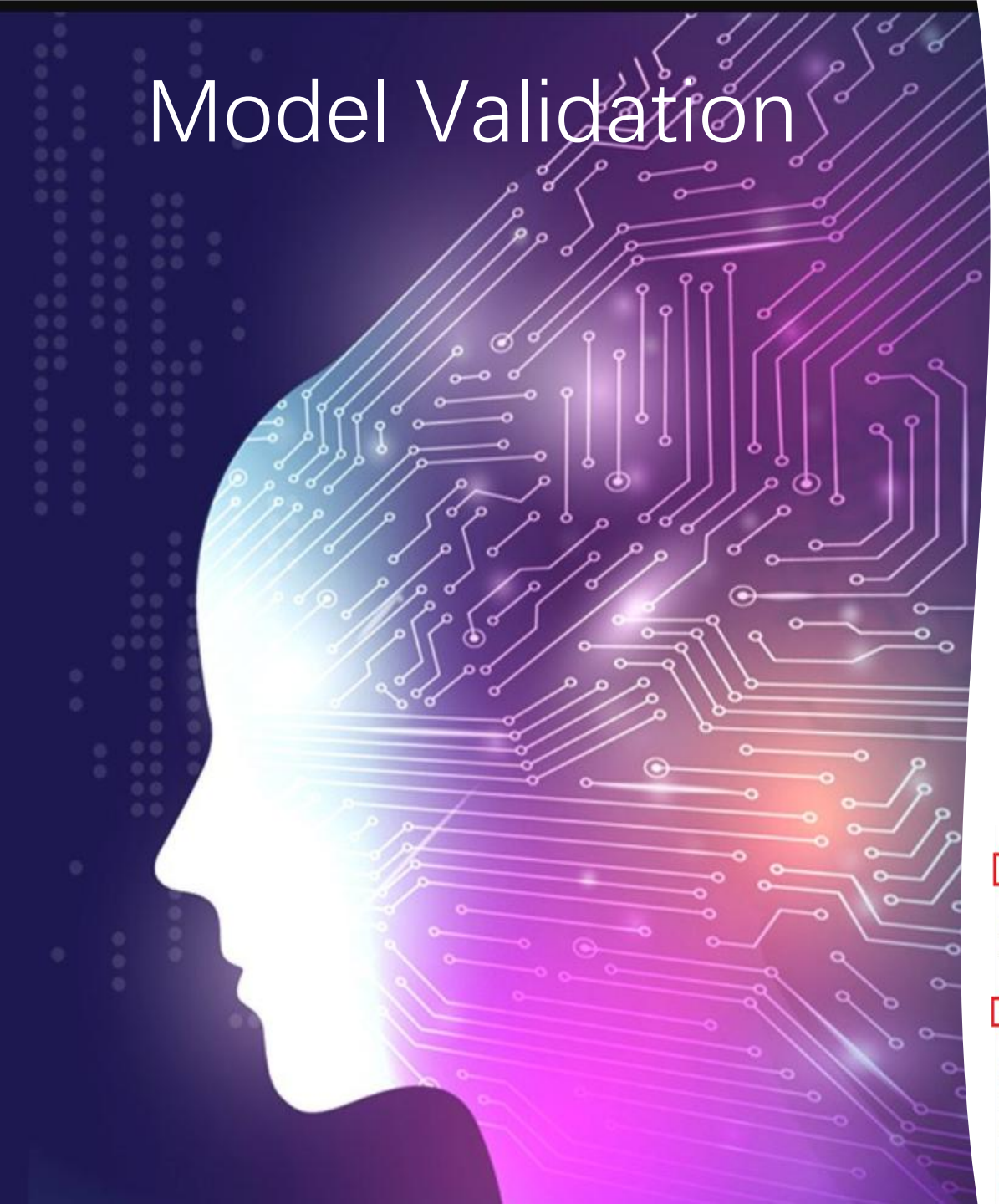
Sale Volume of April = alpha*c1 (C)

- It means that we just predict comment numbers on 2021/5/1, 2021/4/1 to derive sales volume of April.

- Compared with formula (C), we found formula (B) is better after testing.

Then, how to find best alpha1 and alpha2 for formula (B): Sale Volume of April = alpha1*c1 + alpha2*c2 ?

- We leveraged the limited available sales data combined with a parameter search approach to identify the optimal dynamic alpha1 and alpha2 for each month
- Then, alpha1 and alpha2 vary across different months, but remain consistent within the same month.



Group-1

	error_num	cnt	ratio
0	0-5	227348	0.746524
1	10以上	56097	0.184201
2	5-10	21097	0.069275

	mape_per	cnt	ratio
0	0-0.01	183367	0.602107
1	0.01-0.02	47948	0.157443
2	0.02-0.03	21941	0.072046
3	0.03-0.04	12755	0.041883
4	0.04-0.05	7701	0.025287
5	0.05以上	30599	0.100475
6	999999	231	0.000759

Group-2

	error_num	cnt	ratio
0	0-5	433013	0.784006
1	10以上	80993	0.146645
2	5-10	38302	0.069349

	mape_per	cnt	ratio
0	0-0.01	373652	0.676528
1	0.01-0.02	67994	0.123109
2	0.02-0.03	32414	0.058688
3	0.03-0.04	20546	0.037200
4	0.04-0.05	13309	0.024097
5	0.05以上	44306	0.080220
6	999999	87	0.000158

Group-3

	error_num	cnt	ratio
0	0-5	255195	0.768606
1	10以上	51016	0.153652
2	5-10	25812	0.077742

	mape_per	cnt	ratio
0	0-0.01	230817	0.695184
1	0.01-0.02	45608	0.137364
2	0.02-0.03	19997	0.060228
3	0.03-0.04	11757	0.035410
4	0.04-0.05	7092	0.021360
5	0.05以上	16749	0.050445
6	999999	3	0.000009

Sales Validation from Comment data

求的月份	使用月份	alpha1	alpha2	alpha3	r2_score	mse	mae	有效key的数量
求202101	用连续三个月	0.9	0.9	nan	0.951892	1146.916	16.96154	65
求202101	用单月202101	1.1	nan	nan	0.932527	1721.517	14.95692	65

求的月份	使用月份	alpha1	alpha2	alpha3	r2_score	mse	mae	有效key的数量
求202012	用连续三个月	0.3	0.3	0	0.723879	1474.163	16.596	50
求202012	用单月202012	0.8	nan	nan	0.645365	1762.438	18.236	50
求202012	用单月202101	0.5	nan	nan	0.642826	1810.368	18.16176	68
求202012	用单月202102	2	nan	nan	0.346736	3218.882	28.91176	68

求的月份	使用月份	alpha1	alpha2	alpha3	r2_score	mse	mae	有效key的数量
求202011	用连续三个月	1.7	0.4	0	0.905237	605.4082	12.53725	51
求202011	用单月202012	1	nan	nan	0.835313	1238.192	18.38462	52
求202011	用单月202011	2	nan	nan	0.662888	1025.843	12.27451	51
求202011	用单月202101	0.7	nan	nan	0.583656	4176.831	26.63043	69

Results

- We believed that this exploration had no reference value and explicitly informed the client at the outset that such predictions were not advisable.

- However, the client insisted that we carry out the project and ultimately chose to adopt the model as a reference for their marketing strategies and inventory management.

- Therefore, the above outlines our overall approach to tackling this seemingly impossible task.


Thank You