DATA QUALITY CHECK PROJECT

Retail Credit and Fraud Analytics Modelling team

Presented by:
Sarthak Dodeja

NEED FOR THE PROJECT

- When the banking Model runs, input variables are fed in it and the performance of the Model is dependent on these variables itself.
- If this data containing input variables have some irregularity, then it could lead to inappropriate predictions by the model.
- In the existing process, Model Monitoring is done quarterly, post model run, and by the time the problem is identified, its already too late as major business decisions would have already taken place.
- Hence quality check of Variables before the Model runs is essential.
- This report will help the Modelling Team to solve this problem as the Variables will be checked every time before they are fed into the Model.

MODELS COVERED IN THIS QUALITY CHECK

Model name	Base month	Execution
[redacted]	Sept 2023	Quarterly (On 24th)
[redacted]	October 2023	Monthly (On 23rd)
[redacted]	April 2024	Quarterly (On 1st)
[redacted]	April 2023	Monthly (On 24th)
[redacted]	March 2024	Monthly (On Last Day of the month)

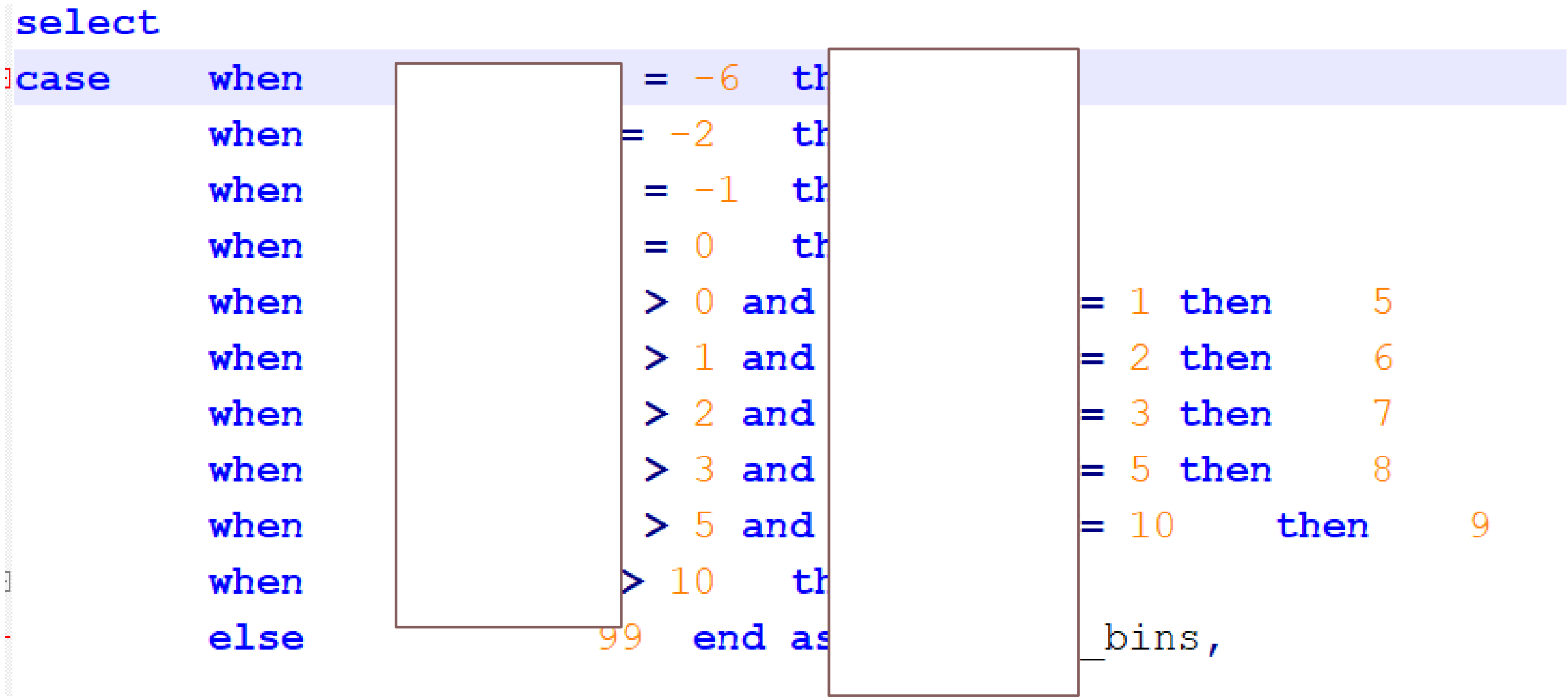
FINAL GOAL - TO CALCULATE CSI

- CSI is a metric that measures the stability of the relationship between two variables over time. It’s commonly used in credit Risk Modeling to monitor changes in the relationship between risk scores and loan defaults. By calculating the CSI between the current period and the baseline period, banks can identify significant changes in the relationship between the two variables that may require a model update.
- CSI is calculated by comparing the coefficients of a regression model fit on the baseline period with a model fit on the current period. CSI is calculated using the following formula:

$$\left((\%Actual - \%Expected) \times \ln \frac{\%Actual}{\%Expected} \right)$$

STEP 1 - BINNING

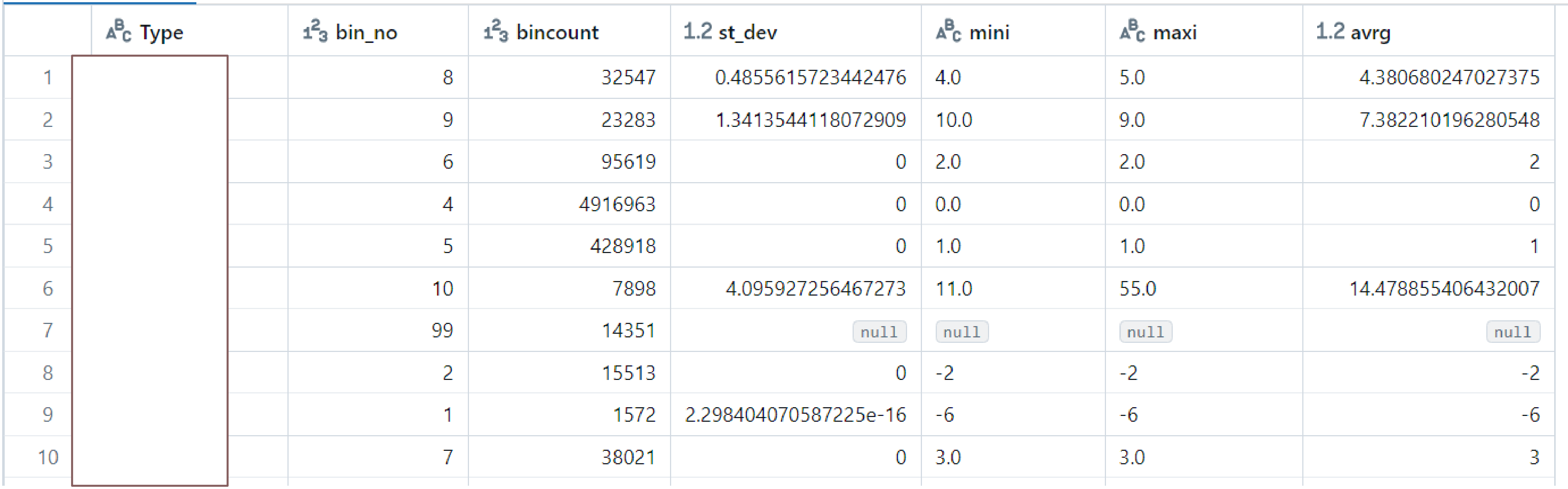
- Bins are made for each Variable based on its data distribution.
- Special provided by CIBIL are binned separately.



```
select
case        when  [redacted]  = -6   th[redacted]
            when  [redacted] = -2   th[redacted]
            when  [redacted]  = -1   th[redacted]
            when  [redacted]  = 0    th[redacted]
            when  [redacted]  > 0  and [redacted] = 1  then     5
            when  [redacted]  > 1  and [redacted] = 2  then     6
            when  [redacted]  > 2  and [redacted] = 3  then     7
            when  [redacted]  > 3  and [redacted] = 5  then     8
            when  [redacted]  > 5  and [redacted] = 10     then     9
            when  [redacted] > 10   th[redacted]
            else  [redacted]  99   end as[redacted]_bins,
```

STEP2 – BIN WISE INFORMATION

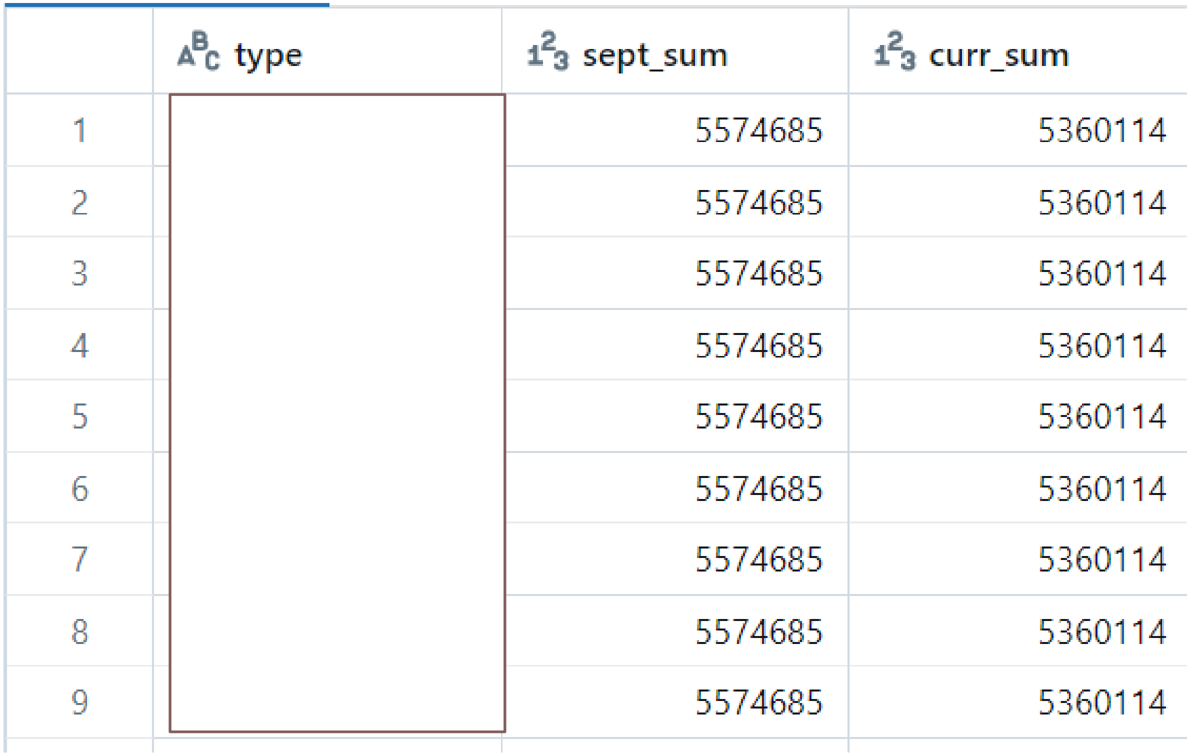
- Count of bins are calculated to check the distribution of Variable.
- Standard deviation , min, max and average are also calculated to understand distribution of these Variables.



	Type	bin_no	bincount	st_dev	mini	maxi	avrg
1		8	32547	0.4855615723442476	4.0	5.0	4.380680247027375
2		9	23283	1.3413544118072909	10.0	9.0	7.382210196280548
3		6	95619	0	2.0	2.0	2
4		4	4916963	0	0.0	0.0	0
5		5	428918	0	1.0	1.0	1
6		10	7898	4.095927256467273	11.0	55.0	14.478855406432007
7		99	14351	null	null	null	null
8		2	15513	0	-2	-2	-2
9		1	1572	2.298404070587225e-16	-6	-6	-6
10		7	38021	0	3.0	3.0	3

STEP3 – COMPARISON OF BASE DATA WITH LATEST DATA

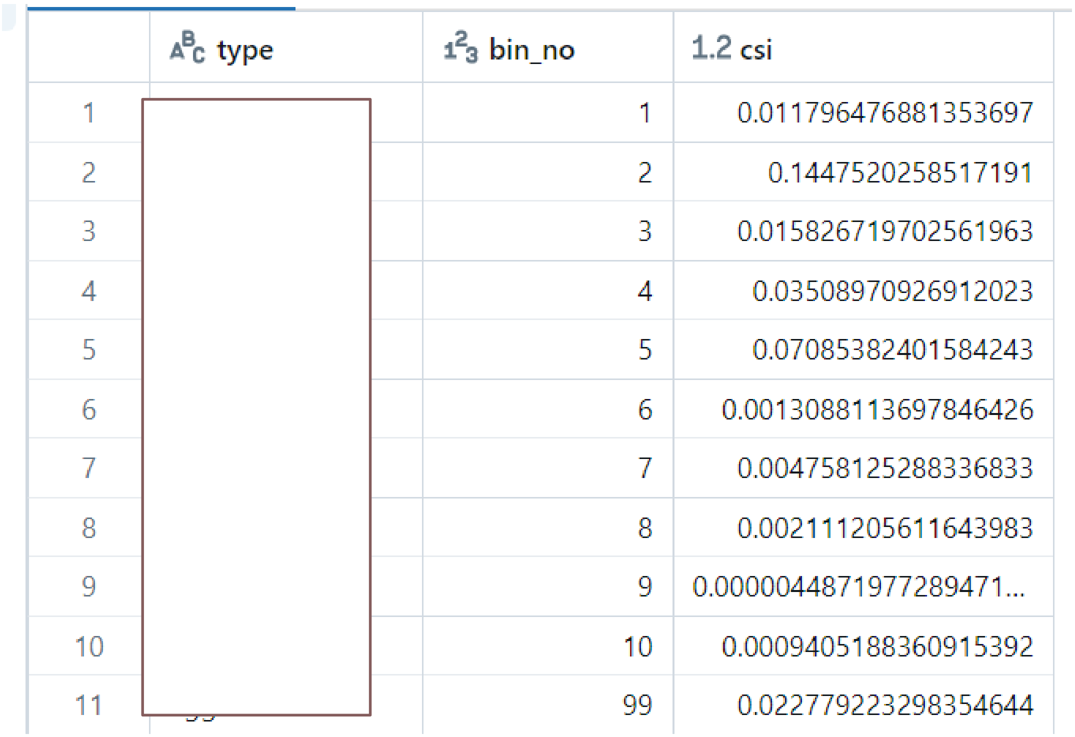
- For reference comparison of CV Variable distribution between September 2023 (Base) and Latest Month is carried out, for this count across bins is published below:



	type	sept_sum	curr_sum
1		5574685	5360114
2		5574685	5360114
3		5574685	5360114
4		5574685	5360114
5		5574685	5360114
6		5574685	5360114
7		5574685	5360114
8		5574685	5360114
9		5574685	5360114

STEP 4 - CALCULATION OF VARIABLE BIN WISE CSI

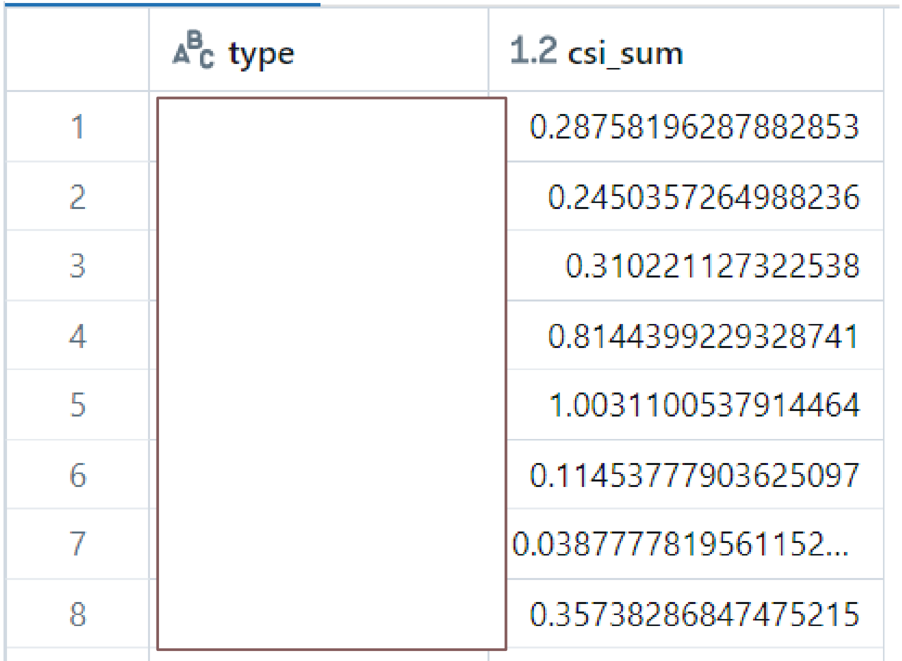
- Bin wise CSI is calculated so that , if the total CSI exceeds the threshold limit we can check individual bins for the irregularity.



	type	bin_no	csi
1	[redacted]	1	0.011796476881353697
2	[redacted]	2	0.1447520258517191
3	[redacted]	3	0.015826719702561963
4	[redacted]	4	0.03508970926912023
5	[redacted]	5	0.07085382401584243
6	[redacted]	6	0.0013088113697846426
7	[redacted]	7	0.004758125288336833
8	[redacted]	8	0.002111205611643983
9	[redacted]	9	0.0000044871977289471...
10	[redacted]	10	0.0009405188360915392
11	[redacted]	99	0.022779223298354644

STEP 5 – CALCULATION OF VARIABLE WISE CSI

- Finally we get the consolidated CSI (sum of all bin's CSI of a variable)
- This is done so as to get the final value of CSI that can be compared with the threshold value.



	ABC type	1.2 csi_sum
1	[redacted]	0.28758196287882853
2	[redacted]	0.2450357264988236
3	[redacted]	0.310221127322538
4	[redacted]	0.8144399229328741
5	[redacted]	1.0031100537914464
6	[redacted]	0.11453777903625097
7	[redacted]	0.0387777819561152...
8	[redacted]	0.35738286847475215

WORKING OF THE PROJECT

- The Base data table will be fixed (Sep'23 in the example demonstrated here).
- Creation of Table containing Latest Data will be scheduled on their respective dates and this table will be automatically updated.
- The same process will be repeated for all models giving us the final CSI for all the required variables.
- Variables where CSI exceeds the threshold limit of 10% will be highlighted.
- The bins of these high CSI variables will be analysed to check for the irregularities.
- Moving forward this process will be automated to run on scheduled dates.