

Task for Junior Quantitative Analyst in Credit Risk Model Validation

SEB

Mykhailo Naginailov

Second Question:

2. Provide opinion on data quality in terms of completeness, accuracy, consistency.

Definitions:

Completeness: Percentage of missing values to determine how much the model relies on estimated data rather than real facts.

Accuracy: Measures whether the data correctly reflects reality by checking for duplicate contracts and unrealistic outliers.

Consistency: Measures whether the data follows logical business rules (e.g., no defaults without past dues) and matches across different tables.

Short Answer for the question:

Completeness:

- Missing values for AGE, EDUCATION, DEBT_RATIO.
- The dataset relies heavily on a rule how to address missing values in key risk drivers (specifically Age and DPD), introducing values rather than purely observing data.

Accuracy:

- Data integrity was initially filled with over 2500 of duplicates and twin duplicates (having same AR_ID and YEAR but different values in other columns).
- Unrealistic outliers (i.e. DPD > 400)

Consistency:

- Internal Logic is clean (i. e. No DPD=0 Defaults found).

Long answer for the question:

1. Count of duplicate values is 2773

- Number of duplicates was 2773 after merging tables with the data on arrangements and the obligor demographics.

First I merged the data to have all variables in a single dataframe. The merge was done using values of AR_ID, IR_ID and YEAR.

Initial Row Count: 32224
Final Merged DataFrame head:

	YEAR	IP_ID	AR_ID	PD	PD_POOL	DFLT_FLAG	AGE	EDUCATION	DEBT_RATIO	DPD	M_LAST_DPD	grid
0	2022	CFG053	2795120	9.925260e-01	5.0	1	70.0	2.0	1.000000	1.0	6.0	
1	2022	NNA451	10210477	2.656939e-05	1.0	0	72.0	2.0	0.804767	NaN	NaN	
2	2022	VLZ931	24787848	3.514735e-06	1.0	0	44.0	2.0	0.136493	NaN	NaN	
3	2022	VVF064	35869938	5.868028e-06	1.0	0	63.0	1.0	0.522421	NaN	NaN	
4	2022	DXH176	71765042	8.378070e-07	1.0	0	20.0	1.0	0.295783	NaN	NaN	

Then I did an investigation on duplicates values and whether the merge was done correctly:

```
if len(dupes) > 0:
    print(f"\nFOUND {len(dupes)} DUPLICATE ROWS.")

# We check if the duplicates are EXACT copies of each other
perfect_copies = final_merged_df.duplicated(keep=False) # Checks ALL columns
# Checking for "Evil Twins" (Same ID, Different Data)
evil_twins = final_merged_df[final_merged_df.duplicated(subset=pk_cols, keep=False) & ~perfect_copies]

if len(evil_twins) > 0:
    print(f"\nCRITICAL WARNING: FOUND {len(evil_twins)} 'CONFLICTING' duplicates!")
    print("Displaying conflicts:")
    display(evil_twins.head(20))
    # Exploring what do they look like
```

I decided to check which kind of duplicates we have in the data: “Conflicting” duplicates which are having same AR_ID and YEAR but having different values in other columns or the duplicates with all same values in all columns.

The result:

Displaying conflicts:

	YEAR	IP_ID	AR_ID	PD	PD_POOL	DFLT_FLAG	AGE	EDUCATION	DEBT_RATIO	DPD	M_LAST_DPD
14	2022	CGC248	28580548	2.192613e-06	1.0	0	48.0	1.0	0.191020	NaN	NaN
15	2022	CGC248	28580548	2.192613e-06	1.0	0	48.0	1.0	NaN	NaN	NaN
49	2022	ZFG062	26749062	1.718946e-06	1.0	0	42.0	1.0	0.179365	NaN	NaN
50	2022	ZFG062	26749062	1.718946e-06	1.0	0	42.0	1.0	NaN	NaN	NaN
63	2022	NXT275	50254034	4.469723e-06	1.0	0	37.0	2.0	0.558583	NaN	NaN
64	2022	NXT275	50254034	4.469723e-06	1.0	0	37.0	2.0	NaN	NaN	NaN
77	2022	LCL666	26962273	5.860783e-06	1.0	0	36.0	2.0	0.808100	NaN	NaN
78	2022	LCL666	26962273	5.860783e-06	1.0	0	36.0	2.0	NaN	NaN	NaN
155	2022	OBR471	79359896	2.126369e-06	1.0	0	32.0	1.0	0.801096	NaN	NaN
156	2022	OBR471	79359896	2.126369e-06	1.0	0	32.0	1.0	NaN	NaN	NaN
233	2022	YDS333	13785325	8.169879e-07	1.0	0	75.0	3.0	0.255882	NaN	NaN
234	2022	YDS333	13785325	8.169879e-07	1.0	0	75.0	3.0	NaN	NaN	NaN
247	2022	UJQ208	73064678	8.908955e-01	5.0	1	36.0	2.0	0.319970	24.0	7.0
248	2022	UJQ208	73064678	8.908955e-01	5.0	1	36.0	2.0	NaN	NaN	NaN
262	2022	LNG232	69444207	5.228200e-06	1.0	0	62.0	1.0	0.651294	NaN	NaN
263	2022	LNG232	69444207	5.228200e-06	1.0	0	62.0	1.0	NaN	NaN	NaN
306	2022	EDD273	34666500	2.390969e-07	1.0	0	33.0	3.0	0.713276	NaN	NaN
307	2022	EDD273	34666500	2.390969e-07	1.0	0	33.0	3.0	NaN	NaN	NaN
451	2022	ZKP519	29581230	5.453794e-06	1.0	0	56.0	2.0	0.034554	NaN	NaN
452	2022	ZKP519	29581230	5.453794e-06	1.0	0	56.0	2.0	NaN	NaN	NaN

After looking on them - we can deduce the logic that the **first** from duplicate pair is the one with the value whereas the second is with missing values, so I decided to keep the first, the result:

=====

CLEANUP COMPLETE:

Rows Deleted: 1396

Final Dataset Size: 30828

2. For missing values the result is following:

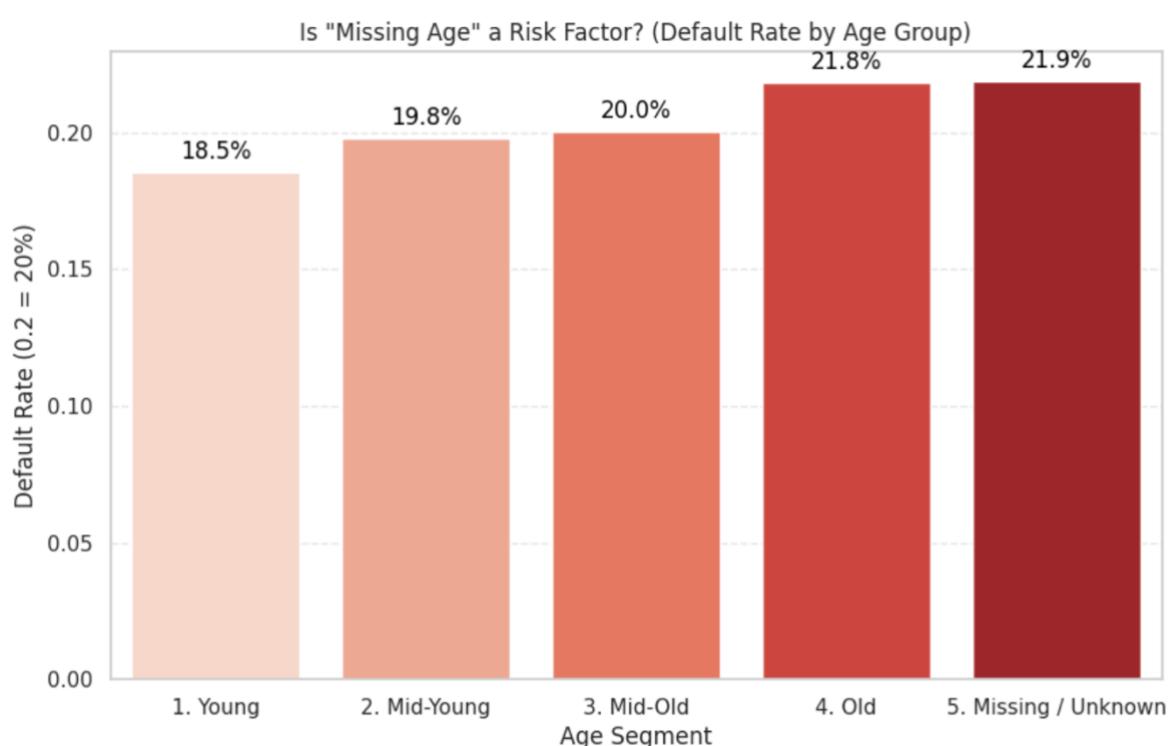
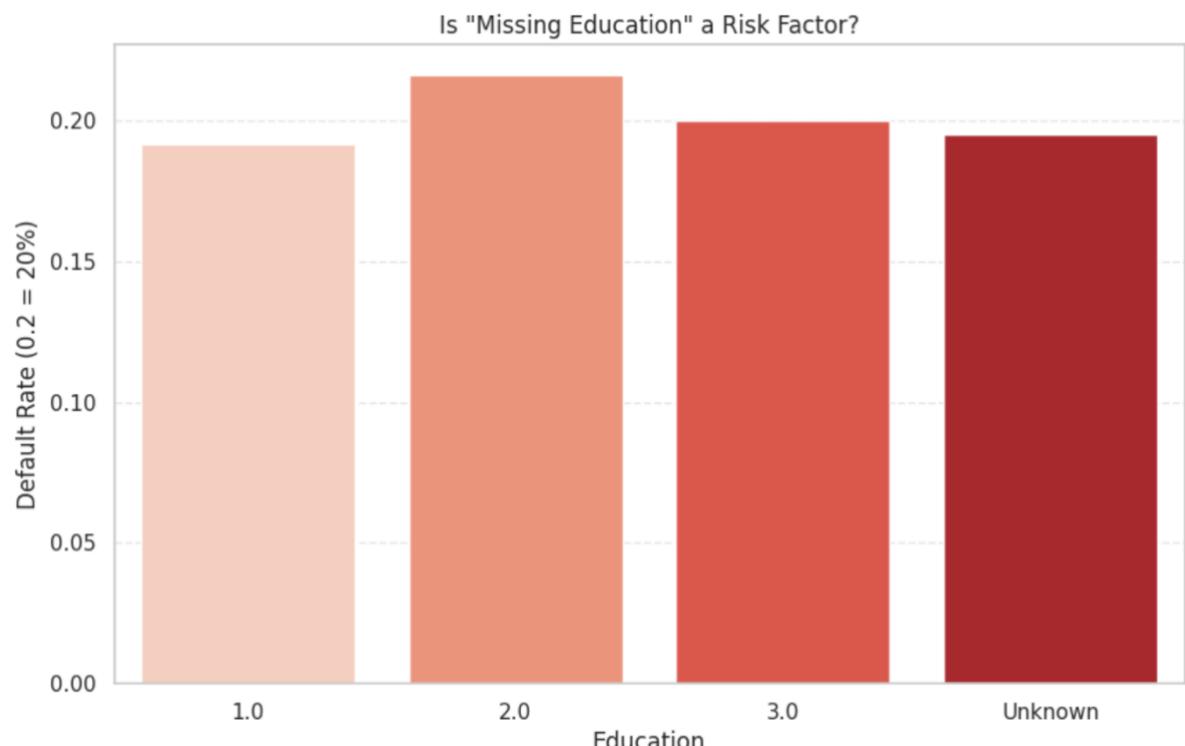
AGE - 645

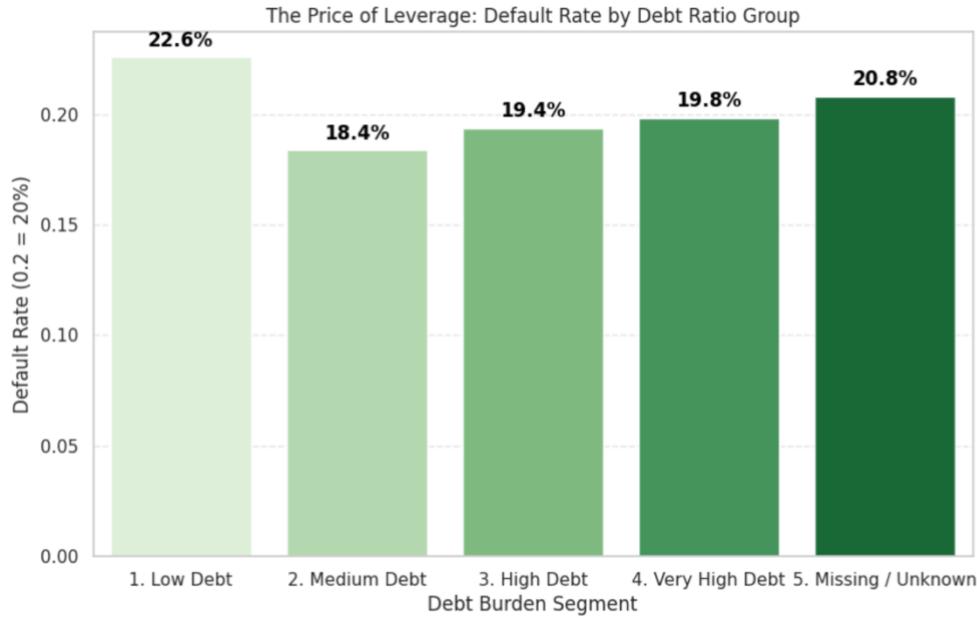
EDUCATION - 780

DEBT_RATIO - 418
 DPD - 26844
 M_LAST_DPD - 26844

But after taking a look on these graphs below, we can say that these missing value are not a big deal and doesn't distort our data significantly. Because there is no unpredictable distribution of the data (the share of missing values is almost same from year to year), and the behaviour is kind of aggregating all of other columns for Education, Age and Exposure to the debt.





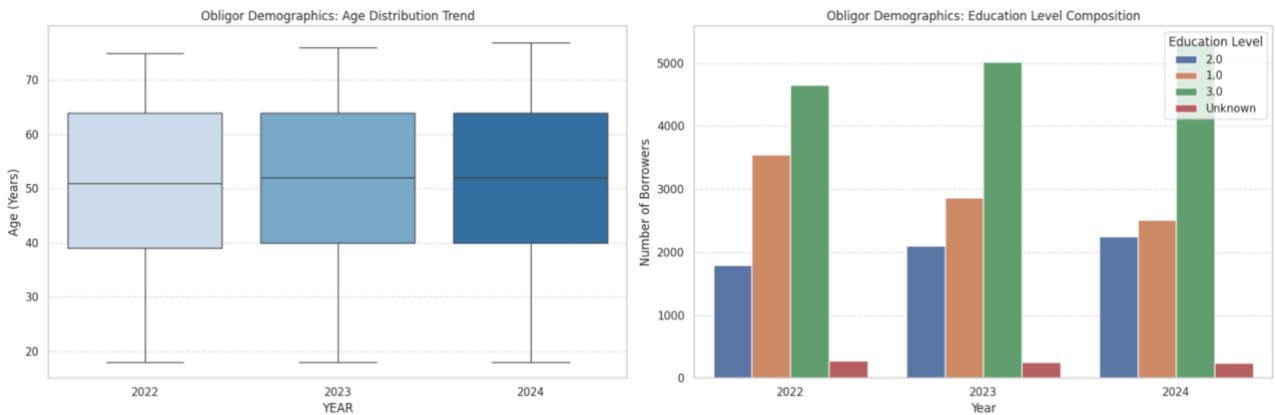


3. The descriptive statistics overall is fine. **But** the average amount of defaulters looks scary (0.200467 - 20%), however, it is more about the portfolio side of the model than the data quality side.

First Question:

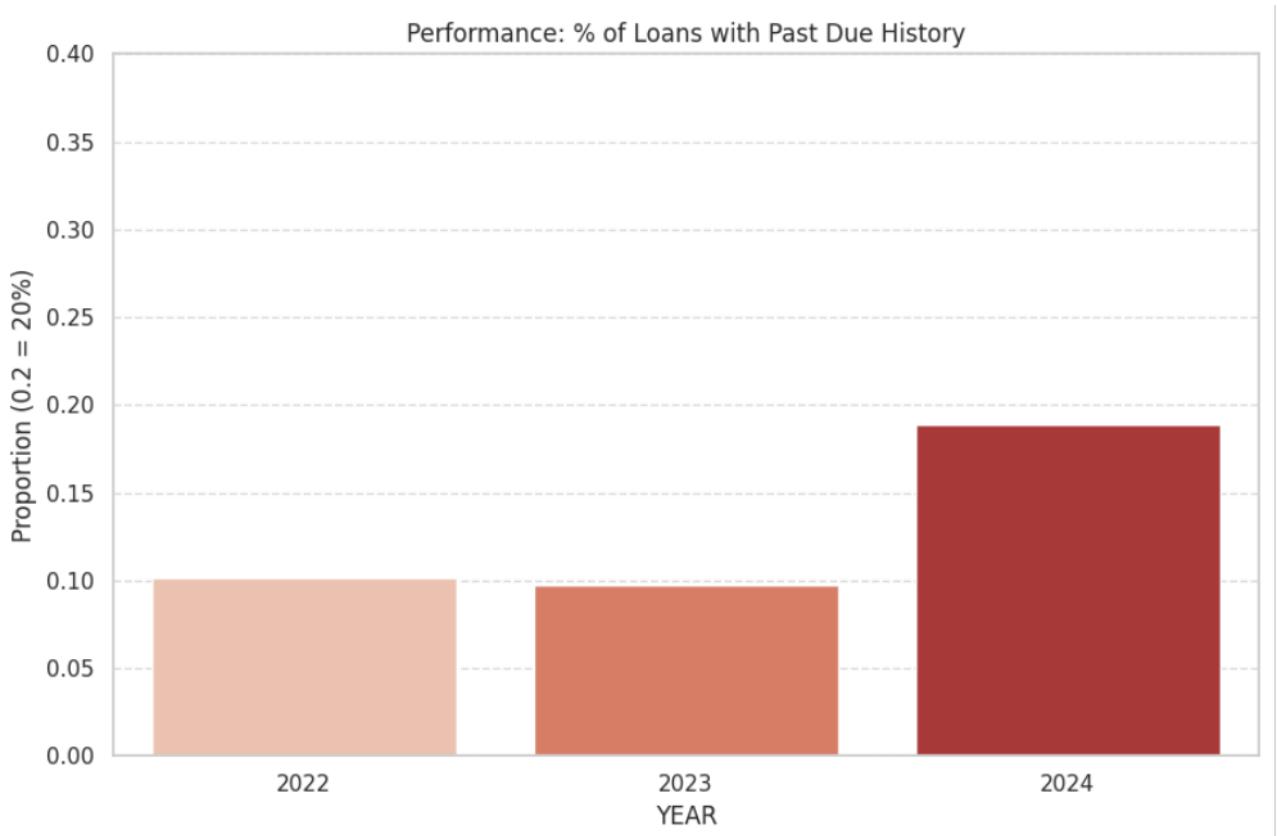
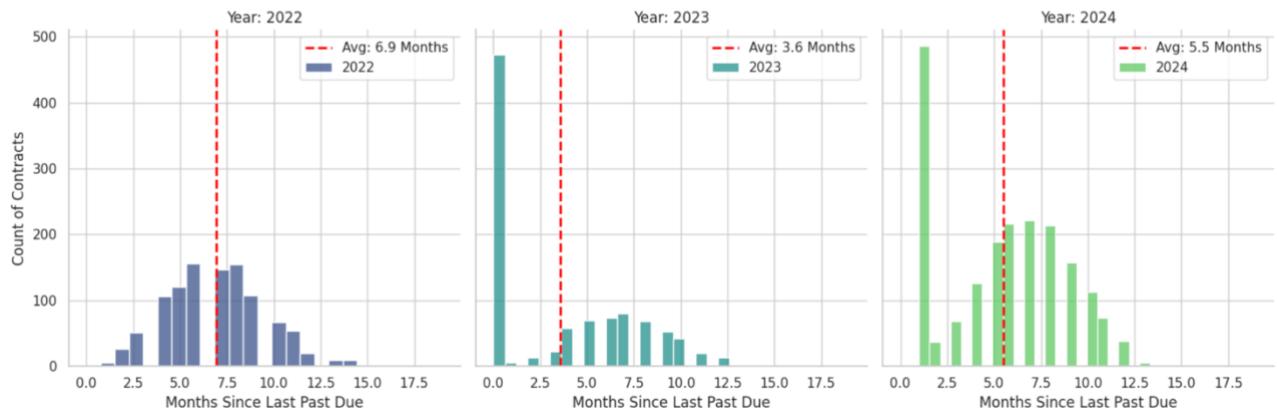
3. **Visualize and provide a brief description on qualitative portfolio structure in terms of obligor, arrangement and scoring features available, portfolio composition and performance. Assess consistency over time.**

a) Obligor demographics:



The provided graphs indicate standard benchmarks for a healthy bank: the average client age is approximately 50 years, and there is a preference for establishing contracts with individuals who have a higher level of education.

```
--- ANALYZING FEATURE: M_LAST_DPD (By Year) ---
   count      mean       std    min   25%   50%   75%   max
YEAR
2022  1041.0  6.947166  2.591423  0.0  5.0  7.0  9.0  19.0
2023   997.0  3.557673  3.830486  0.0  0.0  3.0  7.0  15.0
2024  1946.0  5.480987  3.313119  1.0  2.0  6.0  8.0  15.0
```



Graphs above give us insight that bank was in “steady-state” or in some sort of equilibrium before the 2023 (in 2023 the average dropped from 6.9 months since the last missed payment to the 3.6). However starting from 2023 some change happened, in particular, people start behaving more as risky borrowers in general.

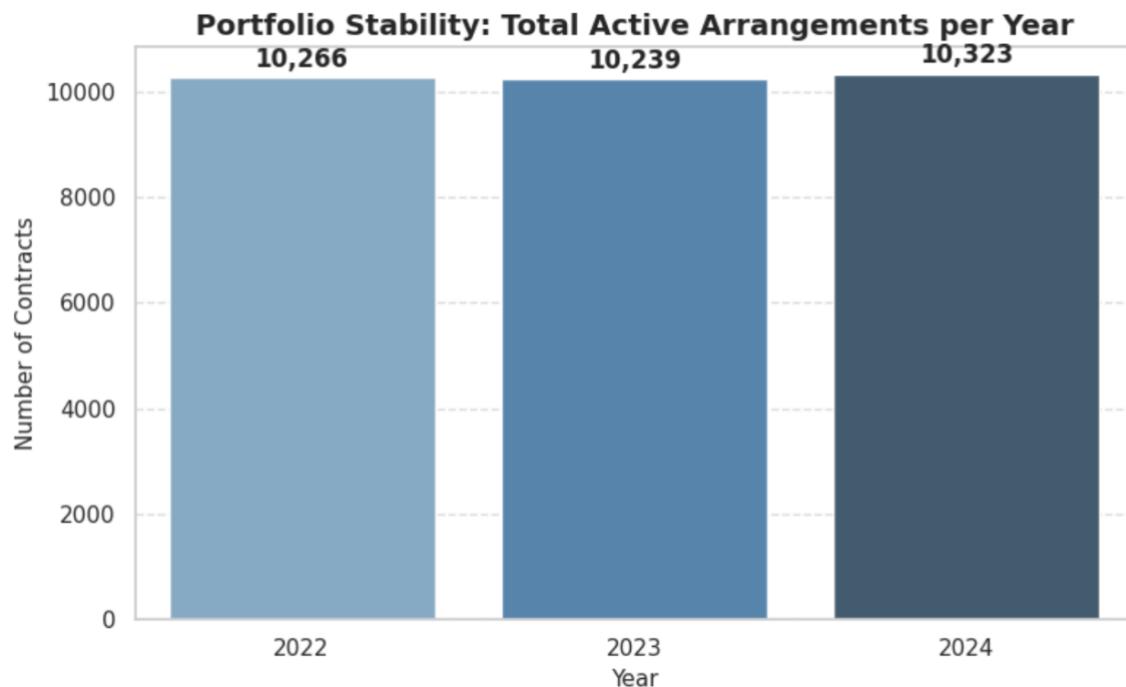
b) Portfolio Composition:



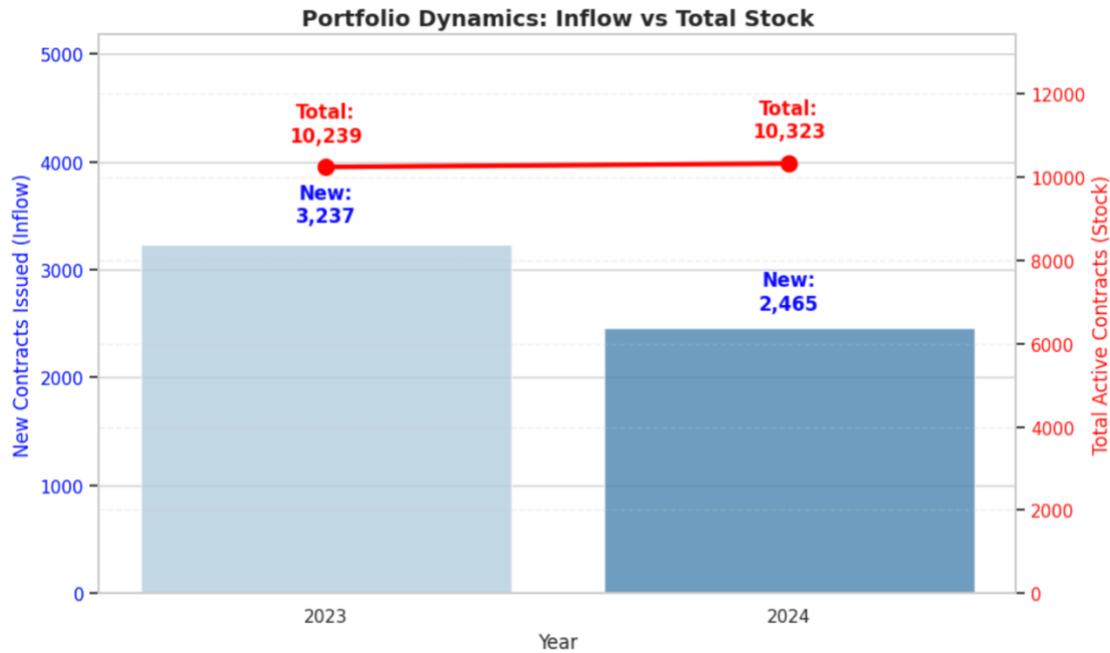
Combining violin plot of arrangement structure and the education level composition (where in 2024 the amount of uneducated people fell, while on contrast the amount of people with higher education increased), the story is following: The bank is closing up shop. Or in other words: It is collecting money from old loans (the bottom bulge) but they are refusing to issue new loans (the decreasing top bulge) except to a tiny group of highly educated people (university degree).

In banking, this is called a "Run-Off Portfolio."

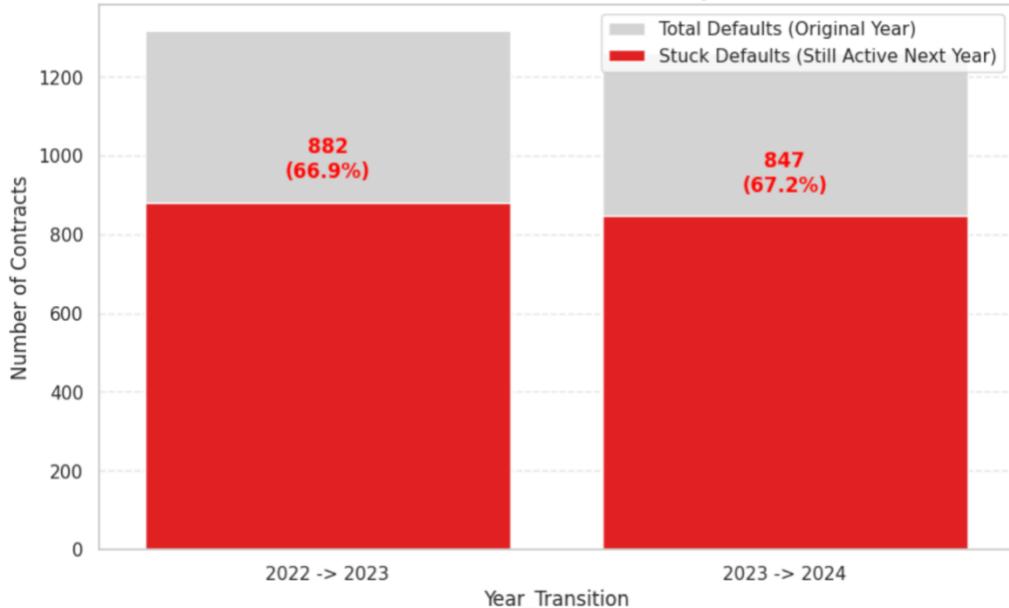
However the total amount of loans is about to be same for our years in our data:



Digging deeper to the total amount of arrangements leads to the picture where the total amount of loans increase but the amount of new arrangements (unique AR_ID) is decreasing. Hence, fewer amount of loans are paid off fully or fewer amount of loans are written off from the books.

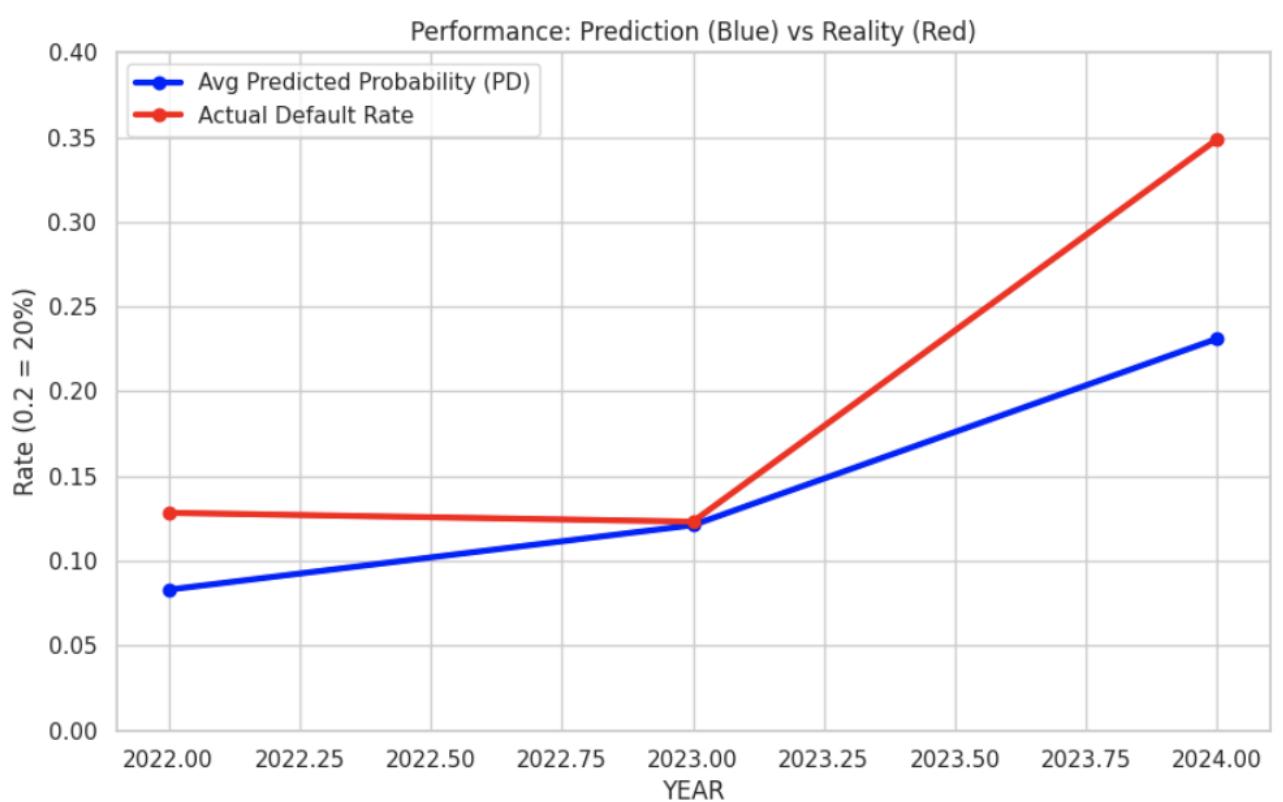


I decided to investigate whether all the defaulters is gone from the bank's books or not, and the story is next: Around **800** of defaulted loans stay in total amount of arrangements.



c) Performance

On the graph below we see two lines: the reality and the prediction. The blue line which is the prediction is almost always below the red one (actual default rate) which means that model underestimate the risk that bank faces on average.



Regarding the pools (graph below) we see that the system of having those pools is actually new to the company, in 2022 it was only very good borrowers and very bad ones. There is a clear switch to the new system in 2023.

Another addition to the theory that the model underestimates the risk – the red area in 2024 is not bigger than 1/3 of our portoflio pillar, which means that it failed to classify the 35% default rate in 2024.



Below I checked whether the total amount of arrangements assigned to pool tiers of level 2,3,4 is equal to 0:

```
checking_pools = final_merged_df.groupby(['YEAR', 'PD_POOL']).count()
display(checking_pools)
```

		IP_ID	AR_ID	PD	DFLT_FLAG	AGE	EDUCATION	DEBT_RATIO	DPD	M_LAST_DPD	HAS_PAST_DUE	
YEAR	PD_POOL											
2022	1.0	8986	8986	8986	8986	8816	8744	8873	3	3	8986	
	2.0	6	6	6	6	6	6	6	6	6	6	
	3.0	6	6	6	6	6	6	6	6	6	6	
	4.0	53	53	53	53	52	53	53	44	44	53	
	5.0	1215	1215	1215	1215	1190	1180	1200	982	982	1215	
2023	1.0	2969	2969	2969	2969	2969	2969	2960	0	0	2969	
	2.0	2442	2442	2442	2442	2360	2376	2396	0	0	2442	
	3.0	2228	2228	2228	2228	2198	2108	2199	0	0	2228	
	4.0	1375	1375	1375	1375	1315	1340	1334	0	0	1375	
	5.0	1225	1225	1225	1225	1189	1187	1207	997	997	1225	
2024	1.0	2742	2742	2742	2742	2742	2742	2731	0	0	2742	
	2.0	2084	2084	2084	2084	1997	2032	2045	0	0	2084	
	3.0	1937	1937	1937	1937	1900	1824	1916	0	0	1937	
	4.0	1188	1188	1188	1188	1126	1160	1157	0	0	1188	
	5.0	2372	2372	2372	2372	2317	2321	2327	1946	1946	2372	

We see that total amount assigned for pools 2,3,4 is not zero.

Two reasons are possible:

- Human behaviour (somebody put some arrangements manually to the system).
- Or this these 65 people assigned were likely specific branches or specific product types where the risk team was "beta testing" the math.

Third Question:

3. Assess correctness of PD estimate calculation.

1. **Predictive Power.** I measured the **Expected-to-Observed (E/O) Ratio**, where a value of 1.0 indicates perfect accuracy.

```
--- 1. CALIBRATION ACCURACY (Right Graph Analysis) ---
YEAR Avg_Predicted_PD (Blue Line) Actual_Default_Rate (Red Line) \
0 2022 0.083014 0.128385
1 2023 0.121085 0.123157
2 2024 0.231150 0.348833
```

	Prediction_Error	E0_Ratio (Target = 1.0)
0	-0.045371	0.646606
1	-0.002072	0.983177
2	-0.117683	0.662639

Results:

- 2022 & 2023 - The "Good" Year, E/O of 0.98
- 2024 -The "Crash" - The model significantly **underestimates risk**. The E/O Ratio dropped to **0.66**, meaning the model predicted a 23% default rate, but the reality was a catastrophic 35%.

Conclusion: The PD estimate is **statistically incorrect for the current portfolio**, failing to capture the systemic risk shift in 2024. The model is "optimistic bias," which is dangerous for capital adequacy planning.

2. Risk Differentiation (Ranking Logic)

--- 2. RISK TIER LOGIC CHECK (Left Graph Analysis) ---					
Do the values increase from Left (Pool 1) to Right (Pool 5)?					
PD_POOL	1.0	2.0	3.0	4.0	5.0
YEAR					
2022	0.005119	0.500000	1.000000	0.962264	0.997531
2023	0.002695	0.004095	0.008079	0.019636	0.977959
2024	0.154267	0.156910	0.150749	0.169192	0.994098

I analyzed whether higher PD Pools correspond to higher actual default rates (Monotonicity).

Results:

- a) 2023 (2022 do not really use pools): Demonstrates robust logic: the default rate increases monotonically from Pool 1 (0.2%) to Pool 5 (97.8%).
- b) 2024: Loss of differentiation power. The distinction between Pool 1 (15.4%), Pool 2 (15.7%), and Pool 3 (15.1%) is negligible.

Conclusion: In the high-stress environment of 2024, the model's ability to distinguish "Safe" from "Risky" clients has collapsed. High-quality borrowers (Pool 1) are defaulting at rates previously associated with bad borrowers.

Fourth Question:

4. Using quantitative and/or statistical tools, measure PD model performance from at least one of the following perspectives:

- (a) PD model performance in terms of PD estimate sufficiency and accuracy
 - 2023 (Optimally Calibrated): The E/O ratio hovered near **1.0**, indicating the model accurately predicted the aggregate default rate during stable period.
 - 2024 (Underestimation): The ratio dropped significantly to 0.66. This indicates a severe "optimistic bias," where the model predicted a 23% default rate while the actual realized rate was 35%.
- (b) PD model performance in terms of risk differentiation on risk attribute raw PD estimate levels
 - 2022: AUC = 0.9899 (Gini = 0.9799)
 - 2023: AUC = 0.9851 (Gini = 0.9702)
 - 2024: AUC = 0.8298 (Gini = 0.6596)

I measured the model's ability to distinguish between high and low-risk obligors using the **AUC (Area Under Curve)** metric.

- 2022: AUC = 0.9899 (Gini = 0.9799)
- 2023: AUC = 0.9851 (Gini = 0.9702)
- 2024: AUC = 0.8298 (Gini = 0.6596)

Results:

- **Performance Decay:** The AUC deteriorated from a robust **[0.98]** in 2023 to **[0.83]** in 2024.
- **Meaning:** While an AUC of 0.83 is mathematically acceptable, the sharp drop (-15 points) confirms that the risk factors (like Age/Education) lost significant predictive power in the 2024 environment. So, new variables should be considered for the future estimation of risk.

(c) PD model stability over time in terms of model coefficients and model result.

I attempted to replicate the bank's model results by strictly following their specified methodology, including missing value treatment, variable capping (minimum and maximum), and the chosen standardization technique.

I initially estimated a simple logistic regression model; however, for the 2024 period, the coefficients were excessively high or low across certain variables. To address this, I used L2 regularization to penalize the model and constrain the magnitude of the coefficients.

Regression outputs:

... --- STABILITY CHECK (Yearly vs Pooled) ---						
YEAR	ORIGINAL_TABLE_4	2022	2023	2024	POOLED_ESTIMATE	
AGE	0.5910	0.1550	0.1470	0.0580	0.1090	
EDUCATION_1	2.0370	0.3020	0.2370	-0.0320	-0.0870	
EDUCATION_2	2.8570	0.2720	0.4710	-0.0110	0.1140	
DEBT_RATIO	0.4260	0.0740	0.0640	0.0010	-0.0320	
MAX_DPD	0.8250	0.3250	0.5320	0.3480	0.2640	
MIN_M_LAST_DPD	-2.2370	-1.4360	-1.4740	-1.6660	-1.3760	
INTERCEPT	7.2570	5.7070	5.9420	9.4380	6.5210	

Results:

- a) Financial Health Stopped Mattering: In the past, factors like a person's EDUCATION level and DEBT_RATIO helped predict if they would default. During the 2024 crisis, these became almost useless as predictors. This suggests that even people who were financially well-off were defaulting.
- b) Only Late Payments Counted: The only reliable sign of risk during the crisis was a history of MAX_DPD (maximum days past due). The model basically said, "If they're currently late on payments, they're a risk; nothing else matters."
- c) Everyone Became Riskier: The overall Intercept (the baseline risk for an average person) shot way up. This means the general chance of anyone defaulting increased dramatically, regardless of their personal details.