10/11/2024

EDA

TD Bank Credit Prediction Capstone

About the Datasets

Internal Dataset shape: (51336, 26)

External Dataset shape: (51336,62)

Unseen Data shape: (100, 42)

Features for Internal Banking Data

Variable Name	Description
Total_TL	Total trade lines/accounts in Bureau
Tot_Closed_TL	Total closed trade lines/accounts
Tot_Active_TL	Total active accounts

Features for External Cibil Data

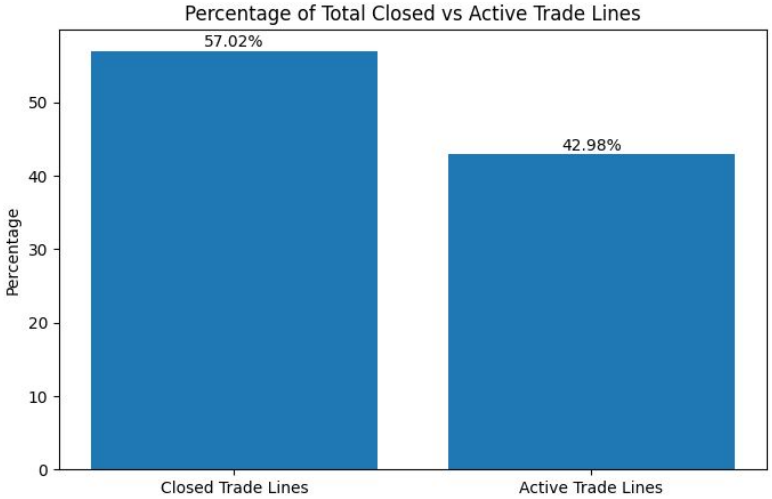
Variable Description

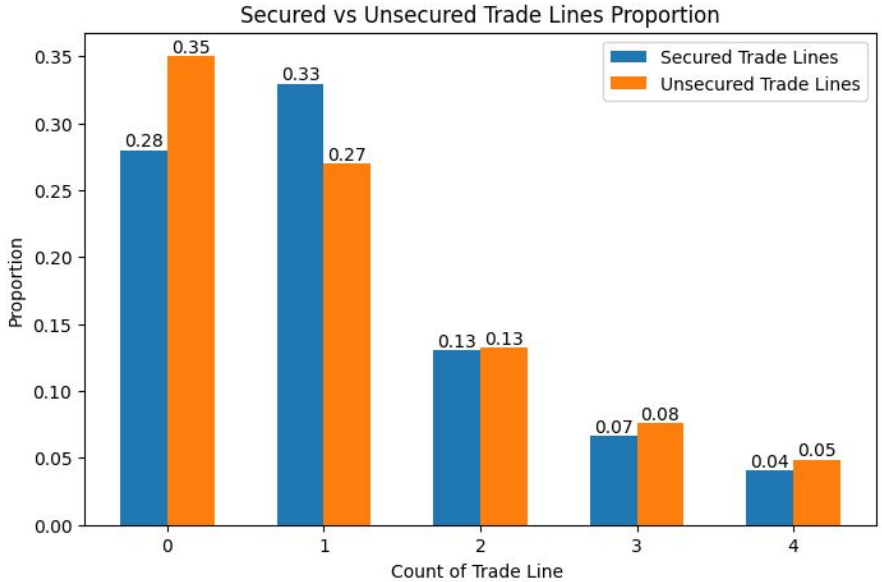
Variable	Description
time_since_recent_payment	Time Since recent Payment made
time_since_first_deliquency	Time since first Delinquency (missed payment)
time_since_recent_deliquency	Time Since recent Delinquency

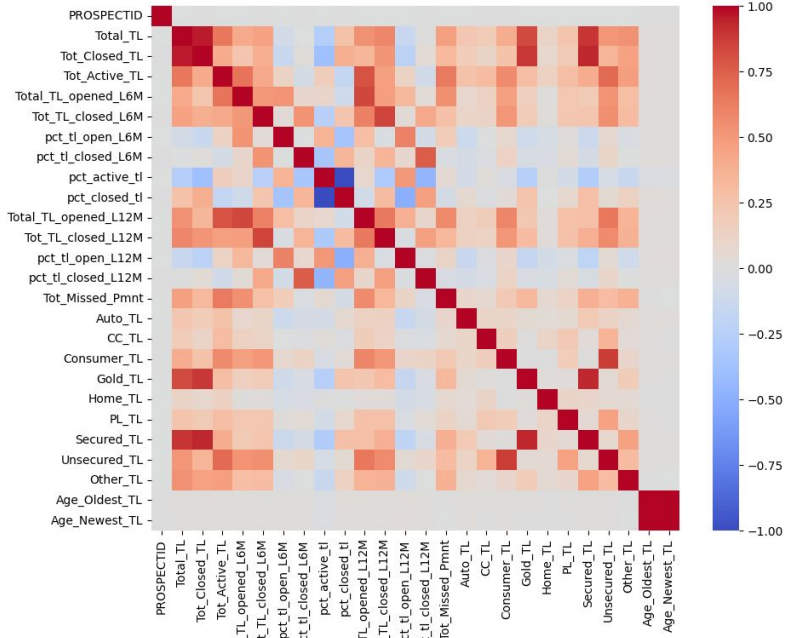
Comments:

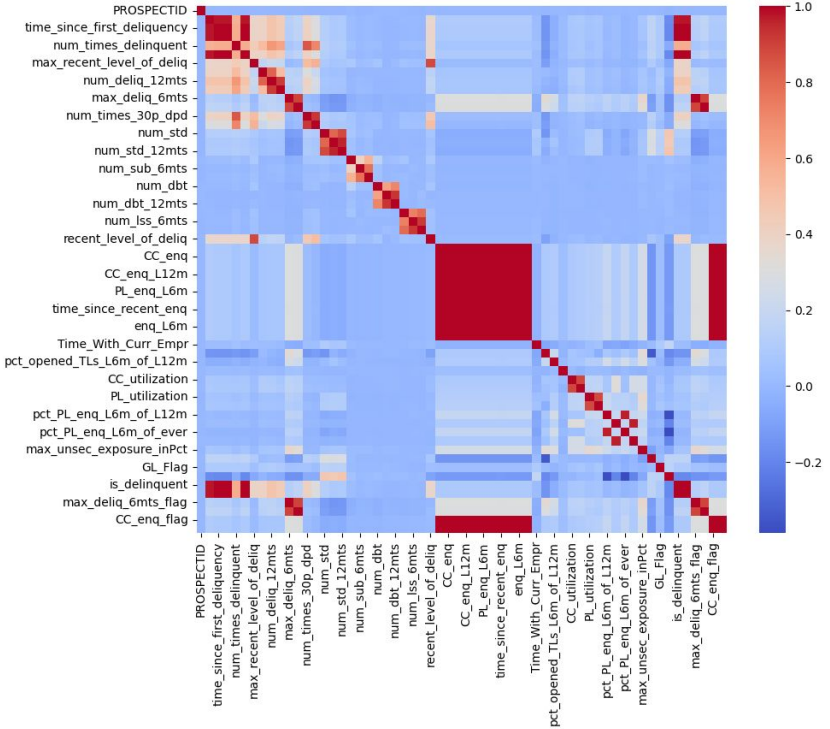
- For a single source of truth, Internal and External Datasets can be joined together using PROSPECTID (primary key) column

General Observations


Percentage of Total Closed vs Active Trade Lines
57.02%
42.98%
Percentage
0
10
20
30
40
50
Closed Trade Lines
Active Trade Lines


Secured vs Unsecured Trade Lines Proportion
Secured Trade Lines
Unsecured Trade Lines
Proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.28
0.35
0.33
0.27
0.13
0.13
0.07
0.08
0.04
0.05
0
1
2
3
4
Count of Trade Line





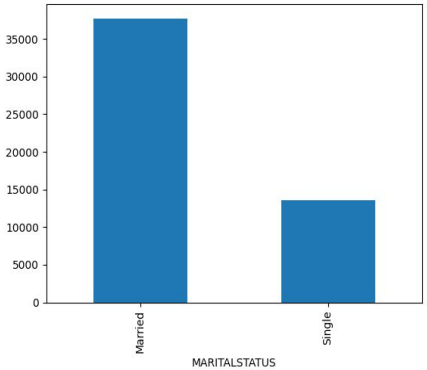
Internal Dataset Correlation

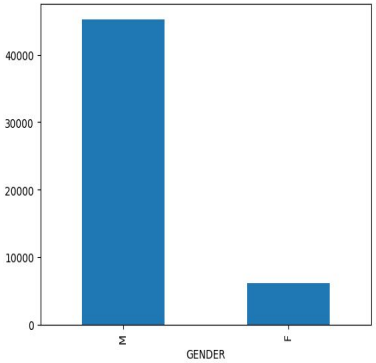
External Dataset Correlation

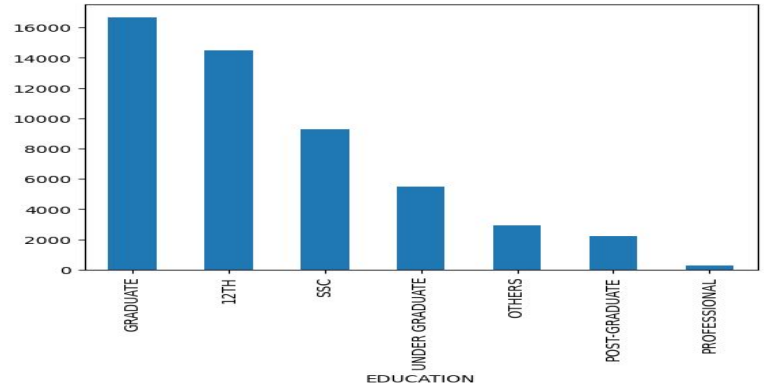
Interesting Observations

Categorical variables

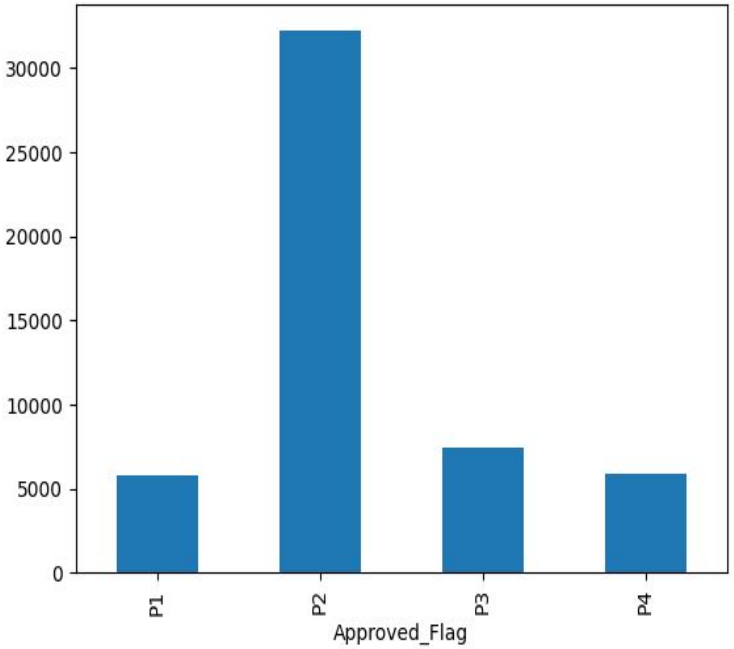
Predictors



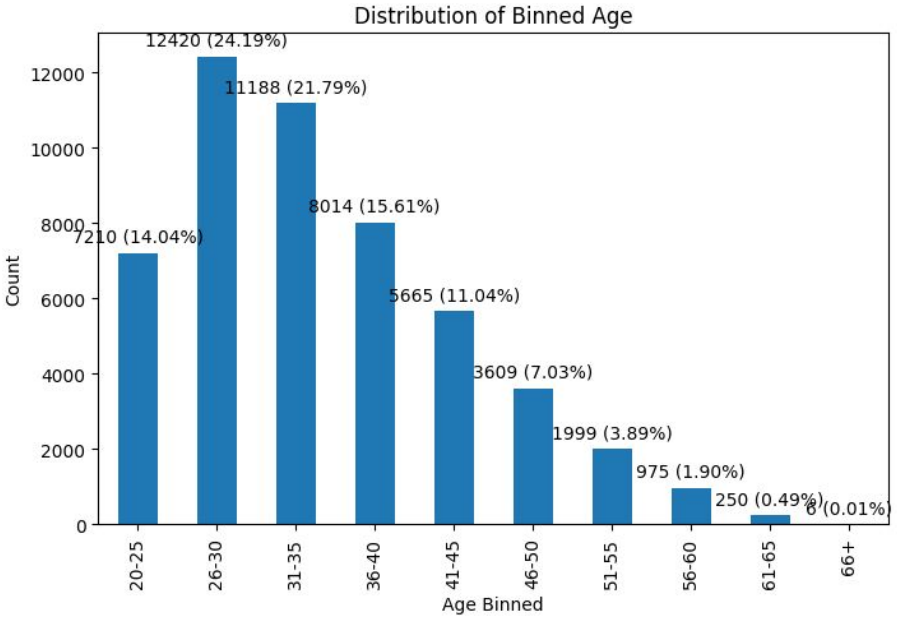


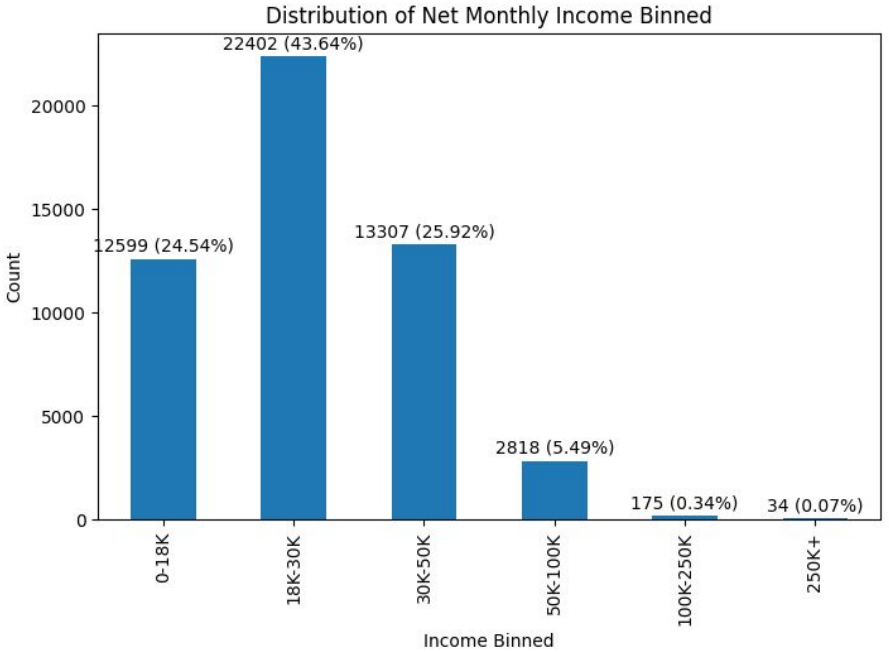


Target

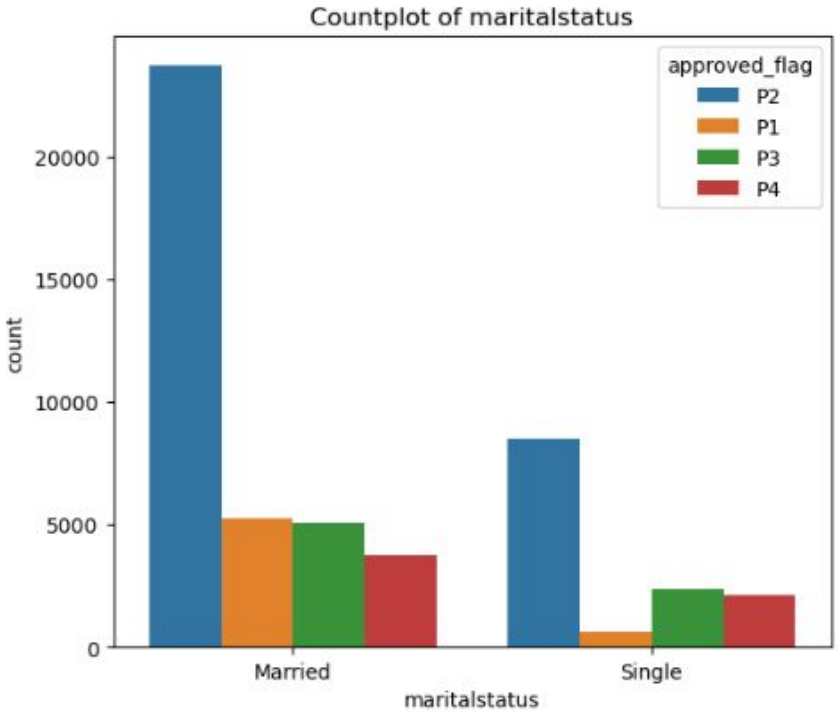


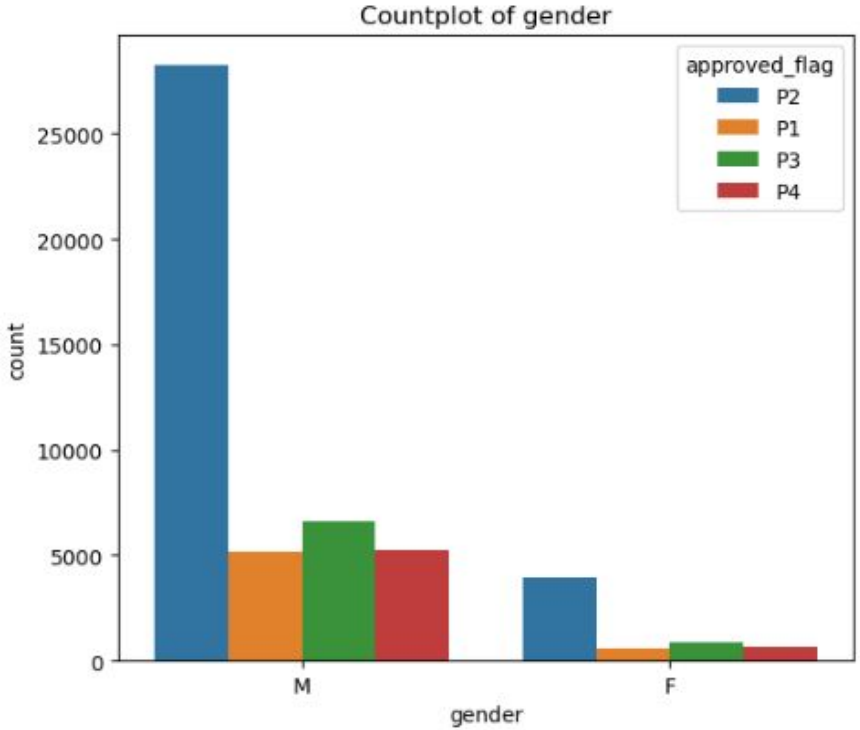
Distribution of numerical variables



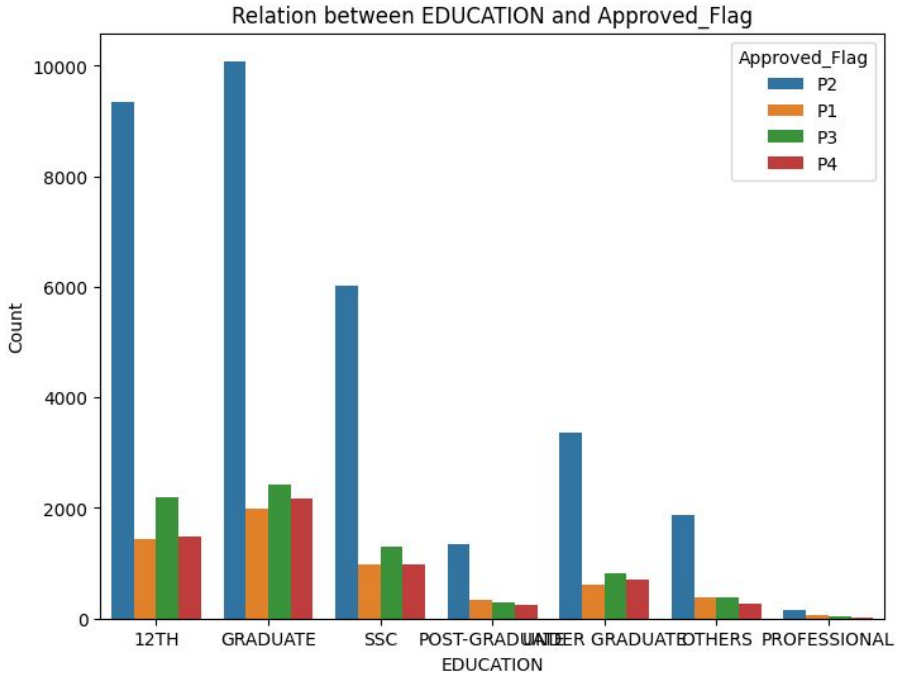


Categorical variables with The Target Variable





Categorical variables with The Target Variable



Challenges and Suggestions

Columns with more than 20% of its values being -99999

Column	Error Percentage
cc_utilization	92.792582
pl_utilization	86.557192
time_since_recent_deliquency	70.026882
max_delinquency_level	70.026882
time_since_first_deliquency	70.026882
max_unsec_exposure_inpct	45.149603
max_deliq_6mts	25.109085
max_deliq_12mts	21.100203

-ve values are only -99999

For all features:

Binned counts for Age_Oldest_TL:

	Bins	Counts	Percentages (%)
0	-99999 to -99998	40	0.077918
1	-99998 to 0	0	0.000000
2	0 to 0.01	21	0.040907
3	0.01 to 43.56444444444444	30639	59.683263
4	43.56444444444444 to 87.11888888888889	13367	26.038258
5	87.11888888888889 to 130.67333333333332	4479	8.724871
6	130.67333333333332 to 174.22777777777776	2055	4.003039
7	174.22777777777776 to 217.7822222222222	594	1.157083
8	217.7822222222222 to 261.33666666666664	99	0.192847
9	261.33666666666664 to 304.89111111111106	27	0.052595
10	304.89111111111106 to 348.44555555555553	7	0.013636
11	348.44555555555553 to 392.0	8	0.015584

No other -ve values

-99999 can be missing data. Not 0, since one still can have Total Loans, Utilization, active CC account

```
Unique values for CC_TL: [0 1 2]
Unique values for CC_enq: [-99999]
Unique values for CC_enq_L6m: [-99999]
Unique values for CC_enq_L12m: [-99999]
Unique values for CC_utilization: [-9.9999e+04  0.0000e+00  5.3700e-01  1.0000e+00  9.9200e-01  8.2800e-01
  9.1800e-01  2.7600e-01  5.9400e-01  6.3200e-01  9.0200e-01  9.4800e-01
  5.0000e-01  5.1800e-01  8.5400e-01  1.0020e+00  1.0000e-03  8.3300e-01
  2.0800e-01  5.1200e-01  9.3600e-01  2.3300e-01  8.4200e-01  2.6700e-01]
Unique values for CC_Flag: [0 1]
```

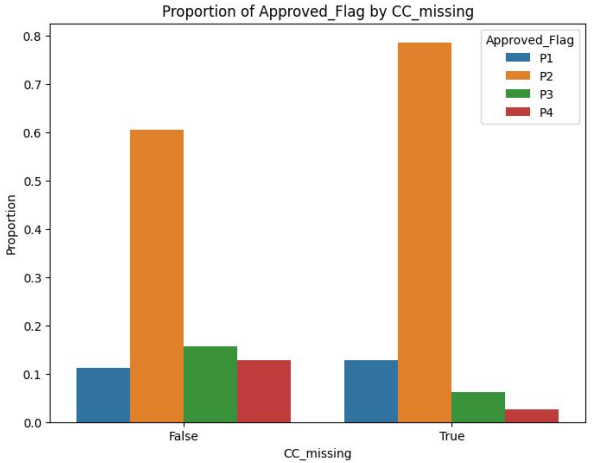
```
Unique values for CC_TL: [0 1 2 3 4]
Unique values for CC_enq: [0]
Unique values for CC_enq_L6m: [0]
Unique values for CC_enq_L12m: [0]
Unique values for CC_utilization: [-9.9999e+04  0.0000e+00  4.1400e-01  1.0000e-03  3.2800e-01  4.7600e-01
  6.7600e-01  9.5400e-01  1.0000e+00  1.1000e-01  2.0000e-03  8.3200e-01 ...
```

Same for PL_enq

Type of columns with -99999

- Time-based:
  - Ex) time_since_first_deliquency
    - If no delinquency, then -99999
      => Create another dummy variable

- Quantity-based:
  - "PL_enq" (The # of personal loan queries)
    - 0 values are already there
    - Does not seem to be random
      => Dummy variable or imputation



Percent based columns with max value > 100

Binned counts for pct_currentBal_all_TL:

	Bins	Counts	Percentages (%)
0	-99999 to -99998	72	0.140252
1	-99998 to 0	0	0.000000
2	0 to 0.01	11946	23.270220
3	0.01 to 703.0644444444445	39314	76.581736
4	703.0644444444445 to 1406.118888888889	1	0.001948
5	1406.118888888889 to 2109.1733333333336	1	0.001948
6	2109.1733333333336 to 2812.227777777778	0	0.000000
7	2812.227777777778 to 3515.2822222222226	0	0.000000
8	3515.2822222222226 to 4218.336666666667	0	0.000000
9	4218.336666666667 to 4921.391111111112	0	0.000000
10	4921.391111111112 to 5624.445555555556	1	0.001948
11	5624.445555555556 to 6327.5	1	0.001948

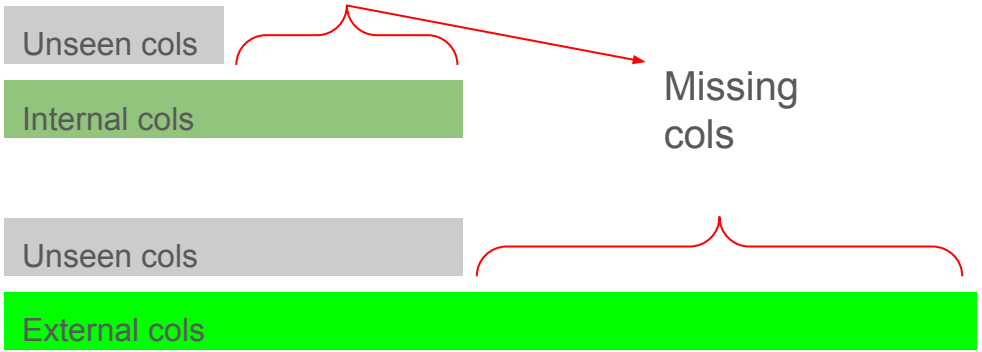
=> Drop these rows

Binned counts for max_unsec_exposure_inPct:

	Bins	Counts	Percentages (%)
0	-99999 to -99998	23178	45.149603
1	-99998 to 0	0	0.000000
2	0 to 0.01	891	1.735624
3	0.01 to 19311.12	27255	53.091398
4	19311.12 to 38622.23	5	0.009740
5	38622.23 to 57933.340000000004	2	0.003896
6	57933.340000000004 to 77244.45	2	0.003896
7	77244.45 to 96555.56	2	0.003896
8	96555.56 to 115866.67	0	0.000000
9	115866.67 to 135177.78000000003	0	0.000000
10	135177.78000000003 to 154488.89	0	0.000000
11	154488.89 to 173800.0	1	0.001948

=> Drop these rows

Mismatch in Unseen dataset against In/External



Number of columns in Unseen that match with Internal: 13 of 26

Number of columns in Unseen that match with External: 29 of 62

Number of columns in Unseen that do NOT match with either Internal or External: 0 of 100

=> Cannot use Unseen dataset (features in In/External datasets are missing in Unseen)

=> Carve out our own “Unseen” data from In/External