**Notes on Dataset**

*2019.08 | Ran Dou*

**CONTEXT FROM EMAILS**

**1. SOP** - From Josh to Andrea, Cody on Jun 30, 2019 at 11:34 PM

1. Long-term strategy of putting data to use – Risk Rating
2. FDIC – the effectiveness of consumer digital footprint as a proxy for default probability risk compared to a traditional credit report.
3. *Risk Rating Model, Predictive Default Probability Indicator*

**2. sample data** – From Peter to Andrea, Jason on May 06, 2019 at 5:29 PM

**Question decoding dataset** – From Peter to Andrea on Aug 9, 2019 at 8:27 AM

**Secret** – From Peter to Andrea, Jason, Max on May 6, 2019 5:43 PM

1. Sample spreads + six sample reports
2. Anything having a “CPR\_Group” of ‘COMMERCIAL BANKING (tab1, column c) is less likely to have an input error.

**3. Follow up** – From Steve to Andrea on June 26, 2019 at 2:17 PM

1. New designs for the industry insights view + formatted version of the financials tab
2. Maximum Debt Capacity Analysis – Projection

**4. Client Analytics** – From Steve to Many on Mar 4, 2019 at 6:06 PM

1. CFG generic pitch book – Business, Key teams, Industries, High-level product offering
2. Product List

**CONTEXT FORM DOCUMENTS**

**1. FDIC-CFR-On the Rise of the FinTechs-Credit Scoring using Digital Footprints**

**2.** **Sample pitch - Working Capital\_v4**

Working Capital Enhancement. 🡪 Identify problems in the DPO/CCC/etc. and ~~provide solutions~~

**3. Maximum Debt Capacity Analysis – Projection**

Assumptions:

1. Conservative growth rate of 2.0%
2. EBITDA Margin remains constant a xxx (in Citizens’ example: 10.6%)
3. Capital Expenditures remains consistent with historical years (in Citizens’ example: $4.0 million)

**4. CFG Commercial Banking Pitch Book FY2018**

Not helpful

**5. Ratings Summary Report Mapping**

Target columns for analysis

1. Raw Data: 1092x324 (14 object, 25 int64, 285 float64)

 111 features contain more than 1% NA value.

1. ﻿Dropped column Start with 'Cov' as they only contain 0 or NA (21 columns)
2. Also dropped columns with only one value (9 columns)

﻿*IndustryClass, rank\_by\_dt, rank\_by\_stmt\_id, ContextID, AcctsRecRelCo, CREATE\_BY\_NM, CREATE\_BY\_DT, LAST\_UPDATE\_BY\_NM, LAST\_UPDATE\_BY\_DT*

1092x294 (9 object, 21 int64, 264 float64)

1. Target ‘object’ features: CPR\_Group, Industry Type, AuditMethod, StatementType

1. ﻿'CPR $net\_inc\_YTD'

(All vs. Within 3std)

﻿count 426.000000

mean 6079.370663

std 24593.860080

min -178749.446140

25% 0.000000

50% 120.507499

75% 3242.455471

max 205897.888088

1. Focus ﻿on the variables in 'Rating Sammary Report Mapping.xlsx' (89 columns)

New subset: 92 features (89 + ﻿'AuditMethod', 'StatementType', 'StatementDate')

Drop columns with more than 50 NA records, then drop all the NA: 1070x85

1. ‘StatementDate’

 only 172 records for 2018

1. Among all numerical features, 34 have 0 for more than 50%

1. 31 columns relate to ratio, Many outliers

|  |  |
| --- | --- |
| ﻿LTD\_TO\_CAPITALIZATION\_RT | CASH\_MS\_TO\_TOTAL\_LIABILITIES\_RT |
| total records: 1029 | total records: 1026 |
| within 3std: 1026 | within 3std: 1019 |
| CHANGE\_IN\_LEVERAGE\_RT | LEVERAGE\_RT |
| total records: 922 | total records: 1029 |
| within 3std: 914 | within 3std: 1024 |
| DEBT\_TO\_EBITDA\_RT | EBITDA\_TO\_SALES\_RT |
| total records: 1026 | total records: 916 |
| within 3std: 1014 | within 3std: 906 |
| CASH\_TO\_TOTAL\_LIABILITIES\_RT | CASH\_TO\_TOTAL\_ASSETS\_RT |
| total records: 1026 | total records: 1030 |
| within 3std: 1019 | within 3std: 1001 |
| CHANGE\_IN\_WC\_TO\_SALES\_RT | TOT\_LIAB\_TO\_TOT\_ASSETS\_RT |
| total records: 922 | total records: 1030 |
| within 3std: 918 | within 3std: 1024 |
| ADJ\_DEBT\_TO\_EBITDA\_RT | ADJ\_EBITDA\_TO\_NET\_SALES\_RT |
| total records: 1026 | total records: 1025 |
| within 3std: 1016 | within 3std: 1023 |
| CASH\_TO\_SALES\_RT | MODEL\_DSCR\_PLUS\_NON\_CASH\_RT |
| total records: 1025 | total records: 904 |
| within 3std: 1016 | within 3std: 898 |
| OCF\_TO\_CPLTD\_IE\_RT | CASH\_AND\_LT\_INVEST\_TO\_TFD\_RT |
| total records: 823 | total records: 904 |
| within 3std: 816 | within 3std: 895 |
| EBIT\_INTEREST\_COVERAGE\_RT | MODEL\_RETURN\_ON\_ASSETS\_RT |
| total records: 901 | total records: 1030 |
| within 3std: 899 | within 3std: 1014 |
| ADJ\_FIX\_CHARGE\_COVERAGE\_RT | MODEL\_NET\_SALES\_GROWTH\_RT |
| total records: 913 | total records: 917 |
| within 3std: 910 | within 3std: 906 |
| TOT\_LIAB\_TO\_TANG\_NET\_WORTH\_RT | MODEL\_RETURN\_ON\_EQUITY\_RT |
| total records: 1028 | total records: 896 |
| within 3std: 1023 | within 3std: 890 |
| CASH\_MS\_TO\_SALES\_RT | MODEL\_OCF\_TO\_ADJ\_EBITDA\_RT |
| total records: 1025 | total records: 225 |
| within 3std: 1016 | within 3std: 223 |
| TOT\_DEBT\_TO\_ASSETS\_RT | MODEL\_OP\_PROFIT\_MARGIN\_RT |
| total records: 1030 | total records: 224 |
| within 3std: 1024 | within 3std: 223 |
| NPBT\_TO\_SALES\_RT | DSCR\_RT |
| total records: 1025 | total records: 205 |
| within 3std: 1022 | within 3std: 201 |
| CHANGE\_IN\_WC\_RT | QuickRatio |
| total records: 925 | total records: 1092 |
| within 3std: 924 | within 3std: 1072 |
|  | CurrentRatio |
|  | total records: 1092 |
|  | within 3std: 1071 |

1. Only 173 records with CPR\_Group of COMMERCIAL BANKIN

173X279

1. ‘Days’, for example, ﻿'NetAcctsRecDays', 'TotalInvDays', 'AcctsPayableDays' have negative value

In Progress:

1. Take inflation into consideration
2. If this is the only data we can have, clustering?