**Data collection and wrangling summary**

Getting the data was not a tuff job as it was readily available in the database. We were just supposed to clean the data and make minor changes.

Source:

Dumped the data from a financial DB to a csv file and took it as an input.

Name of the raw file: dataset\_uncleaned\_28072020.csv

Set of columns taken under consideration:

[ 'industry', 'disbursal\_date', 'tnc\_amount\_first', 'tnc\_ir\_first', 'tnc\_tenure\_first',

'min\_cibil\_score', 'loan\_type', 'd\_date\_of\_birth', 'vintage',

'3m\_avg\_bal', '3m\_avg\_bal\_new', 'total\_liab\_new', '90adb',

'l\_created\_date', 'tenure', 'score', 'rounded\_score']

Sample:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **industry** | **disbursal\_date** | **tnc\_amount\_first** | **tnc\_ir\_first** | **tnc\_tenure\_first** | **min\_cibil\_score** | **loan\_type** | **d\_date\_of\_birth** | **vintage** | **3m\_avg\_bal** | **3m\_avg\_bal\_new** | **total\_liab\_new** | **90adb** | **l\_created\_date** | **tenure** | **score** | **rounded\_score** | **disbursal\_date** | **getdate** | **mob** |
| **Building supplies and accessories** | 2018-05-11 05:30:00 | 180000 | 1.8 | 24 | 725 | ecom | 1991-06-01 | 12 | 87212.7466666667 | 88753.76 | 0 | 88753.76 | 2018-05-03 05:42:12.601 | 19 | 1.47368421052632 | 1 | 2018-05-11 05:30:00 | 2020-07-28 16:49:28 | 1.10681700517722 |
| **Industrial Supplies & Solutions** | 2018-07-09 05:30:00 | 700000 | 1.8 | 24 | 644 | ecom | 1973-03-20 | 37 | 89512.4 | 87169.43 | 0 | 87169.43 | 2018-06-27 05:25:36.939 | 17 | 1.29411764705882 | 1 | 2018-07-09 05:30:00 | 2020-07-28 16:49:28 | 1.0261718438869 |
| **Computers,Mobile & related Accessories** | 2018-09-05 05:30:00 | 270000 | 2 | 24 | 741 | ecom | 1987-11-08 | 55 | 40480.3533333333 | 43564.94 | 350000 | 43564.94 | 2018-08-22 08:27:54.742 | 15 | 1.4 | 1 | 2018-09-05 05:30:00 | 2020-07-28 16:49:28 | 0.948214854639586 |
| **Other Service Activities** | 2018-07-27 05:30:00 | 50000 | 1.8 | 12 | 649 | ecom | 1983-09-26 | 12 | 12079.0066666667 | 11838.57 | 264313 | 11838.57 | 2018-06-04 05:13:18.035 | 12 | 1.25 | 1 | 2018-07-27 05:30:00 | 2020-07-28 16:49:28 | 2.0039565909996 |
| **Home, Furnishing and Decor** | 2018-07-24 05:30:00 | 70000 | 2.1 | 12 | 694 | ecom | 1987-12-31 | 22 | 8215.35333333333 | 8030.16 | 64959 | 8030.16 | 2018-06-13 10:30:35.362 | 12 | 1.41666666666667 | 1 | 2018-07-24 05:30:00 | 2020-07-28 16:49:28 | 2.01202110712863 |

Cleaning steps:

* **Inserted mean values in place of null/nan values**
* **Converted date columns to type datetime**
* **Chose numerical columns only for analysis (excluding categorical)**
* **new\_columns = ['disbursal\_date', 'tnc\_amount\_first',**

**'tnc\_ir\_first', 'tnc\_tenure\_first', 'min\_cibil\_score',**

**'d\_date\_of\_birth', 'vintage', '3m\_avg\_bal', '3m\_avg\_bal\_new',**

**'total\_liab\_new', '90adb', 'l\_created\_date', 'tenure']**

* **Normalized data via min-max**

Finally the cleaned data was saved to **‘cleaned\_dataset.csv’.**