**Programming Language used : Python**

**Workflow Manager User: Luigi**

**Tasks**

1. **Downloading Data**
2. **Preprocessing Origination Data**
3. **Preprocessing Performance Data**
4. **Building prediction model**
5. **Downloading Data:**

File Location : Classes/Part1/Download\_sf\_loan.py

Task Requires no prior tasks to be completed.

Output of the task are all the sample origination and performance files.

Process:

* Asking user for username and password.
* Creating a browser agent (using the mechanicalsoup library) to store and pass the cookies
* Logging in with the user’s credentials.
* Checking if the user is succesfuly logged in or not.
* Landing to the page that contains the list of files and download links
* Asking user for the year and the quarter file to run prediction model.
* Putting the table of files in a dataframe
* Iterating through the rows in dataframe for the links that contain sample files and downloading them to a newly created (if it doesn’t already exist) “Downloads” directory
* The program also checks if the files are already present in the “Downloads” directory. It skips the downloading if the file already exists.
* Unzipping the downloaded file.

1. **Origination File Observation and Cleaning**
2. Credit Score: Deleted the rows that had missing credit score
   1. Cannot replace missing values as it is explicitly specified that credit score can be either less than 301 or greater than 850.
   2. Number of such instances is very less (0.002% in 2016 to 1.242% in 2000)
   3. Removing the rows that have blank values and nulls for credit score.

|  |  |  |
| --- | --- | --- |
| CREDIT SCORE | COUNT OF BLANKS | YEAR |
|  | 362 | 1999 |
|  | 621 | 2000 |
|  | 274 | 2001 |
|  | 201 | 2002 |
|  | 33 | 2003 |
|  | 42 | 2004 |
|  | 24 | 2005 |
|  | 39 | 2006 |
|  | 29 | 2007 |
|  | 29 | 2008 |
|  | 1 | 2009 |
|  | 1 | 2011 |
|  | 3 | 2013 |
|  | 2 | 2014 |
|  | 1 | 2016 |

1. FIRST PAYMENT DATE: No missing values in sample files
2. FIRST TIME HOMEBUYER FLAG:
   1. If blank it can be replaced by NA if Occupancy Status is either “I” or “S” (Investement property or Second Home)
   2. If blank it can replaced by NA if Loan Purpose is either “C” or “N” (Refinance)
   3. If blank, then replace it with NA
   4. Created three columns for – First Time HomeBuyer Flag YES (1,0) , NO(1,0) and NA(1,0)
3. MATURITY DATE:
   1. No missing values in the sample files
   2. Splitting Maturity year and month
4. METROPOLITAN STATISTICAL AREA(MSA) OR METROPOLITAN DIVISION:
   1. Replaced missing values with zero.
   2. Derived a new column for Metropolitan Area Flag, that had values in it
   3. Future Scope: Compare the values of zip codes, if the zip code belongs to a MSA or MD, then map the msa or md code in the data.

|  |  |
| --- | --- |
| YEAR | COUNT OF BLANKS |
| 1999 | 7640 |
| 2000 | 7542 |
| 2001 | 6978 |
| 2002 | 7309 |
| 2003 | 7182 |
| 2004 | 7844 |
| 2005 | 7913 |
| 2006 | 8209 |
| 2007 | 8671 |
| 2008 | 7729 |
| 2009 | 7528 |
| 2010 | 7022 |
| 2011 | 6944 |
| 2012 | 6593 |
| 2013 | 5475 |
| 2014 | 5030 |
| 2015 | 4845 |
| 2016 | 1184 |

1. MORTGAGE INSURANCE PERCENTAGE (MI%):

|  |  |  |  |
| --- | --- | --- | --- |
| MORTGAGE INSURANCE PERCENTAGE (MI %) | COUNT | YEAR | Percentage |
|  | 9026 | 1999 | 18.052 |
| 0 | 21885 | 1999 | 43.77 |
|  | 44 | 2000 | 0.088 |
| 0 | 32764 | 2000 | 65.528 |
|  | 58 | 2001 | 0.116 |
| 0 | 36990 | 2001 | 73.98 |
|  | 11 | 2002 | 0.022 |
| 0 | 38304 | 2002 | 76.608 |
|  | 10 | 2003 | 0.02 |
| 0 | 40083 | 2003 | 80.166 |
|  | 9 | 2004 | 0.018 |
| 0 | 40437 | 2004 | 80.874 |
|  | 57 | 2005 | 0.114 |
| 0 | 43136 | 2005 | 86.272 |
| 0 | 43086 | 2006 | 86.172 |
| 0 | 39839 | 2007 | 79.678 |
| 0 | 40958 | 2008 | 81.916 |
| 0 | 46460 | 2009 | 92.92 |
| 0 | 46266 | 2010 | 92.532 |
| 0 | 44985 | 2011 | 89.97 |
| 0 | 43711 | 2012 | 87.422 |
| 0 | 40459 | 2013 | 80.918 |
| 0 | 36478 | 2014 | 72.956 |
| 0 | 37309 | 2015 | 74.618 |
| 0 | 9481 | 2016 | 18.962 |

* 1. Zero means No Mortgage insurance
  2. Blanks Means either less than 1% or greater than 55%, so the replacement cannot be generalized in this case. Also, such cases are ~18% in 1999 and ~0.01% in until 2005 and 0 in the later years.
  3. Deriving a new column for mortgage insurance flag is done, where the value is kept No if MI% is zero, otherwise it is made Yes

1. NUMBER OF UNITS:

|  |  |  |
| --- | --- | --- |
|  | 1 | 2000 |
|  | 7 | 2004 |

* 1. No missing values for most sample files. Only 1 in the year 2000 and 7 cases in 2004 where number of units is missing
  2. Replaced it with the mode OR Discard the row

1. OCCUPANCY STATUS:
   1. No missing values in the sample files.
   2. Handled the missing value by replacing it by mode or discarding the rows
2. ORIGINAL COMBINED LOAN-TO-VALUE(CLTV):

|  |  |  |
| --- | --- | --- |
| ORIGINAL COMBINED LOAN-TO-VALUE (CLTV) | COUNT | Year |
|  | 0 | 1999 |
|  | 3 | 2000 |
|  | 2 | 2001 |
|  | 4 | 2002 |
|  | 3 | 2003 |
|  | 3 | 2004 |
|  | 6 | 2005 |
|  | 1 | 2006 |
|  | 2 | 2007 |
|  | 0 | 2008 |
|  | 0 | 2009 |
|  | 1 | 2010 |
|  | 0 | 2011 |
|  | 2 | 2012 |
|  | 2 | 2013 |
|  | 1 | 2014 |
|  | 2 | 2015 |
|  | 0 | 2016 |

* 1. ~0.01% missing values in the sample files.
  2. If the LTV is less than 80 or greater than 200 or unknown, then this column is unknown. Also if CLTV is less than LTV then, CLTV is set to unknown.
  3. This value is dependent on each individual case, so may not be replaced by mean, median or mode.

1. ORIGINAL DEBT-TO-INCOME (DTI) RATIO:
   1. Ratio greater than 65% are represented as spaces. We replaced it by 70.
   2. Unknowns are represented by null, which we replaced by the median.
2. ORIGINAL UPB:
   1. No missing values in the sample files
   2. If value is missing then discard the rows.
3. ORIGINAL LOAN-TO-VALUE:
   1. Ratios below 6% and greater than 105% are unknown.

|  |  |  |
| --- | --- | --- |
| ORIGINAL LOAN-TO-VALUE (LTV) | COUNT | YEAR |
|  | 0 | 1999 |
|  | 2 | 2000 |
|  | 1 | 2001 |
|  | 1 | 2002 |
|  | 3 | 2003 |
|  | 3 | 2004 |
|  | 6 | 2005 |
|  | 1 | 2006 |
|  | 2 | 2007 |
|  | 0 | 2008 |
|  | 0 | 2009 |
|  | 1 | 2010 |
|  | 0 | 2011 |
|  | 2 | 2012 |
|  | 2 | 2013 |
|  | 1 | 2014 |
|  | 2 | 2015 |
|  | 0 | 2016 |

* 1. Close to zero percent of such occurrence. But, replacing of the values with mean/median cannot be justified as it is specifically said that these values are either less than 6 or greater than 105. So, discarding such rows.

1. ORIGINAL INTEREST RATE:
   1. No missing values
   2. If value is missing then replace by median
2. CHANNEL:
   1. No missing values in sample files
   2. If values are missing then replace by mode
3. PREPAYMENT PENALTY MORTGAGE (PPM) FLAG:

|  |  |  |
| --- | --- | --- |
| PREPAYMENT PENALTY MORTGAGE (PPM) FLAG | COUNT | YEAR |
|  | 1247 | 1999 |
|  | 236 | 2000 |
|  | 122 | 2001 |
|  | 171 | 2002 |
|  | 198 | 2003 |
|  | 73 | 2004 |
|  | 49 | 2005 |
|  | 65 | 2006 |
|  | 113 | 2007 |
|  | 1039 | 2008 |
|  | 317 | 2009 |
|  | 336 | 2010 |
|  | 580 | 2011 |
|  | 39 | 2012 |
|  | 4 | 2013 |
|  | 11 | 2014 |
|  | 41 | 2015 |
|  | 7 | 2016 |

* 1. Most number of blanks (unknown) in the year 1999 -> 2.49%, 2008 -> 2.078%

|  |  |
| --- | --- |
| **1999** | **48753** |
| N | 48491 |
| Y | 262 |
| **2000** | **49764** |
| N | 49737 |
| Y | 27 |
| **2001** | **49878** |
| N | 49867 |
| Y | 11 |
| **2002** | **49829** |
| N | 49784 |
| Y | 45 |
| **2003** | **49802** |
| N | 49652 |
| Y | 150 |
| **2004** | **49927** |
| N | 49752 |
| Y | 175 |

* 1. Maximum are “N” throughout the years. 97.5% in 1999, 99.5% in 2000…
  2. We are replacing unknown(blanks) values by mode as it wouldn’t affect the distribution.

1. PRODUCT TYPE:
   1. No missing values found in the observations
   2. If there are any missing values, then it is replaced with “FRM”
2. PROPERTY STATE:
   1. No missing values found in the observations
   2. If there are any missing values, then it is replaced with “Unknown”
3. PROPERTY TYPE:

|  |  |  |
| --- | --- | --- |
| PROPERTY TYPE | COUNT | YEAR |
|  | 8 | 2000 |
|  | 11 | 2001 |
|  | 3 | 2002 |
|  | 14 | 2004 |

|  |  |  |
| --- | --- | --- |
| PROPERTY TYPE | COUNT | YEAR |
|  | 8 | 2000 |
| CO | 4090 | 2000 |
| CP | 74 | 2000 |
| LH | 15 | 2000 |
| MH | 244 | 2000 |
| PU | 6531 | 2000 |
| SF | 39038 | 2000 |
|  | 11 | 2001 |
| CO | 3546 | 2001 |
| CP | 45 | 2001 |
| LH | 22 | 2001 |
| MH | 181 | 2001 |
| PU | 5470 | 2001 |
| SF | 40725 | 2001 |
|  | 3 | 2002 |
| CO | 3399 | 2002 |
| CP | 48 | 2002 |
| LH | 12 | 2002 |
| MH | 274 | 2002 |
| PU | 5053 | 2002 |
| SF | 41211 | 2002 |
|  | 14 | 2004 |
| CO | 3616 | 2004 |
| CP | 210 | 2004 |
| LH | 35 | 2004 |
| MH | 529 | 2004 |
| PU | 6829 | 2004 |
| SF | 38767 | 2004 |

* 1. No missing values for most of the years.
  2. Very few missing values observed for years 2000, 2001, 2002 and 2004.
  3. Replaced the missing values with the mode (“SF” as observed) because most number of records are categorized as Single Family Home (77% to 82%)

1. POSTAL CODE:

|  |  |  |
| --- | --- | --- |
| POSTAL CODE | COUNT | YEAR |
|  | 1 | 1999 |
|  | 72 | 2000 |
|  | 1 | 2001 |
|  | 1 | 2002 |
|  | 0 | 2003 |
|  | 0 | 2004 |
|  | 1 | 2005 |
|  | 0 | 2006 |
|  | 0 | 2007 |
|  | 0 | 2008 |
|  | 0 | 2009 |
|  | 0 | 2010 |
|  | 0 | 2011 |
|  | 0 | 2012 |
|  | 0 | 2013 |
|  | 0 | 2014 |
|  | 0 | 2015 |
|  | 0 | 2016 |

* 1. 72 of 50000 unknowns in 2000, 1 row each in 1999, 2001,2002 and 2005 of unknowns
  2. Replaced the blanks with 99999 as unknown value
  3. Future Scope: Get a complete dictionary of Metropolitan Statistical Area or Metropolitan Division codes and map the MSA or MD for the row to the dictionary to find the missing postal code

1. LOAN SEQUQENCE NUMBER:
   1. Unique Identifier Column.
   2. No missing values. If the value is missing for a row, then replace by random Loan sequence number the complete row or generating a unique identifier UUID
   3. Derived two new columns for origination year and origination quarter
2. LOAN PURPOSE:
   1. No missing values in the sample files.
   2. If the values are missing then, loan purpose is unknown. Assuming that the percentage of such occurrence in the yearly data would be close (if not equal to) 0%, and it wouldn’t affect the distribution of the data, we replaced it by the mode of the column
3. ORIGINAL LOAN TERM:
   1. No missing values observed.
4. NUMBER OF BORROWERS:

|  |  |  |
| --- | --- | --- |
| NUMBER OF BORROWERS | COUNT | YEAR |
|  | 30 | 1999 |
|  | 20 | 2000 |
|  | 11 | 2001 |
|  | 9 | 2002 |
|  | 7 | 2003 |
|  | 14 | 2004 |
|  | 17 | 2005 |
|  | 17 | 2006 |
|  | 23 | 2007 |
|  | 19 | 2008 |
|  | 6 | 2009 |
|  | 0 | 2010 |
|  | 0 | 2011 |
|  | 0 | 2012 |
|  | 0 | 2013 |
|  | 0 | 2014 |

* 1. 0% to 0.6% Missing values found.
  2. Replacing missing values with the mode.

1. SELLER NAME:
   1. No missing values found in the sample files.
   2. Replacing missing values by “Unknown”
2. SERVICES NAME:
   1. No missing values found in the sample files.
   2. Replacing missing values by “Unknown”
3. SUPER CONFORMING FLAG:

|  |  |  |
| --- | --- | --- |
| SUPER CONFORMING FLAG | COUNT | YEAR |
| Y | 80 | 2008 |
| Y | 1236 | 2009 |
| Y | 1364 | 2010 |
| Y | 1967 | 2011 |
| Y | 2189 | 2012 |
| Y | 1718 | 2013 |
| Y | 1995 | 2014 |
| Y | 2223 | 2015 |
| Y | 492 | 2016 |

1. Per the data dictionary, all the missing values are Not super conforming, so replaced the missing values by “N”
2. **PERFORMANCE FILE**
3. LOAN SEQUENCE NUMBER:
   1. Derived two new columns for origination year and origination quarter
4. MONTHLY REPORTING PERIOD:
   1. Derived two new columns for monthly reporting period year and month
5. CURRENT ACTUAL UPB:
6. CURRENT LOAN DELINQUENCY STATUS:
   1. No Missing values observed in the sample files.
   2. Replacing missing values with “XX” which is also used for unknown.
7. LOAN AGE:
   1. No missing values observed.
8. REMAINING MONTHS TO LEGAL MATURITY:
   1. No missing values found.
9. REPURCHASE FLAG:
   1. This field is only populated at loan termination. For all others the value is not applicable.
   2. Replacing nulls with NA.
10. MODIFICATION FLAG:
    1. Replacing nulls with “NO” (Not modified)
11. ZERO BALANCE CODE:
    1. Replacing nulls and spaces with “NA” as it is not applicable if the balance is not reduced to zero.
12. ZERO BALANCE EFFECTIVE DATE:
    1. Replacing missing values with 999999, which will denote not applicable.
    2. Deriving 2 new columns for zero balance effective year and month.
13. CURRENT INTEREST RATE:
    1. Replacing empty values with 0.
14. DUE DATE OF LAST PAID INSTALLMENT:
    1. Replacing missing values with 999999.
    2. Deriving 2 new columns for due year and month of last paid installment.
15. Replacing missing values with 0 for the following columns
    1. MI RECOVERIES
    2. NET SALES PROCEEDS
    3. NON MI RECOVERIES
    4. EXPENSES
    5. LEGAL COSTS
    6. MAINTENANCE AND PRESERVATION COSTS:
    7. TAXES AND INSURANCE:
    8. MISCELLENEOUS EXPENSES:
    9. ACTUAL LOSS CALCULATION:
    10. MODIFICATION COST
    11. CURRENT DEFERRED UPB