Week 3 Assignment (Individual) on CECL Data Merge, Cleaning and Exploration

Rohit Khurana

October 8, 2018

1. **Assignment Scope**

   The objective of this Assignment 3 is to address key topics of

   a. Data Loading, Merge and Cleaning

   b. Data Exploration

2. **Data Loading, Merge and Cleaning**

   2017Q2 data was downloaded as part of the last assignment and saved on the local machine in the .CSV format. This included two files

   - Acquisitions Data

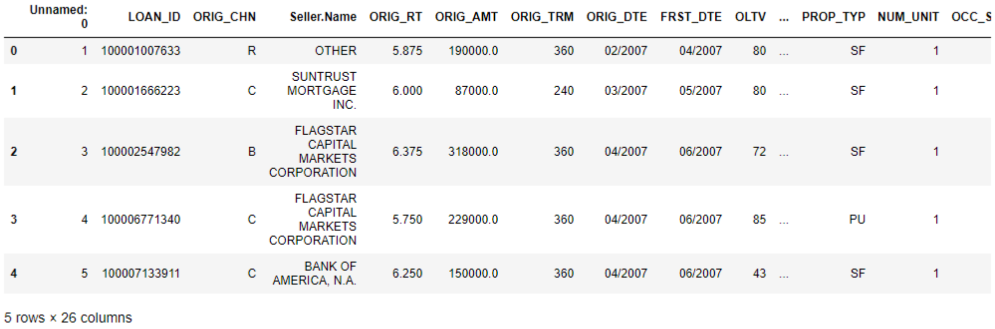
   - Performance Data

   For the purpose of analysis, we are interested in key elements from acquisition data, those are debt-to-income ratio, loan amount, credit score and interest rate, among other elements. From the Performance data, we are interested in loan payment history, delinquency, foreclosure date, etc. that will help us design the model of identifying the defaulting of loan.

Data Loading

Post reading the csv files into a data frames, following the two respective heads of the data frames.

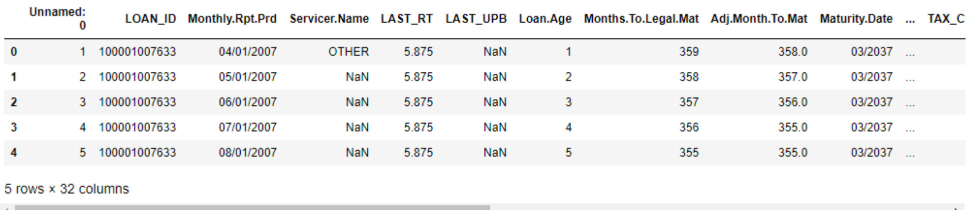
Acquisitions Data Frame



	Unnamed: 0	LOAN_ID	ORIG_CHN	Seller.Name	ORIG_RT	ORIG_AMT	ORIG_TRM	ORIG_DTE	FRST_DTE	OLTV	...	PROP_TYP	NUM_UNIT	OCC_S
0	1	100001007633	R	OTHER	5.875	190000.0	360	02/2007	04/2007	80	...	SF	1	
1	2	100001666223	C	SUNTRUST MORTGAGE INC.	6.000	87000.0	240	03/2007	05/2007	80	...	SF	1	
2	3	100002547982	B	FLAGSTAR CAPITAL MARKETS CORPORATION	6.375	318000.0	360	04/2007	06/2007	72	...	SF	1	
3	4	100006771340	C	FLAGSTAR CAPITAL MARKETS CORPORATION	5.750	229000.0	360	04/2007	06/2007	85	...	PU	1	
4	5	100007133911	C	BANK OF AMERICA, N.A.	6.250	150000.0	360	04/2007	06/2007	43	...	SF	1	

5 rows × 26 columns

Performance Data Frame



	Unnamed: 0	LOAN_ID	Monthly.Rpt.Prd	Servicer.Name	LAST_RT	LAST_UPB	Loan.Age	Months.To.Legal.Mat	Adj.Month.To.Mat	Maturity.Date	...	TAX_C
0	1	100001007633	04/01/2007	OTHER	5.875	NaN	1	359	358.0	03/2037	...	
1	2	100001007633	05/01/2007	NaN	5.875	NaN	2	358	357.0	03/2037	...	
2	3	100001007633	06/01/2007	NaN	5.875	NaN	3	357	356.0	03/2037	...	
3	4	100001007633	07/01/2007	NaN	5.875	NaN	4	356	355.0	03/2037	...	
4	5	100001007633	08/01/2007	NaN	5.875	NaN	5	355	355.0	03/2037	...	

5 rows × 32 columns

In the performance data, we are only interested in following columns, hence we can drop the rest:-

- Loan Identifier: LOAN ID

- Servicer Name: Servicer.Name

- Current Interest Rate: LAST RT

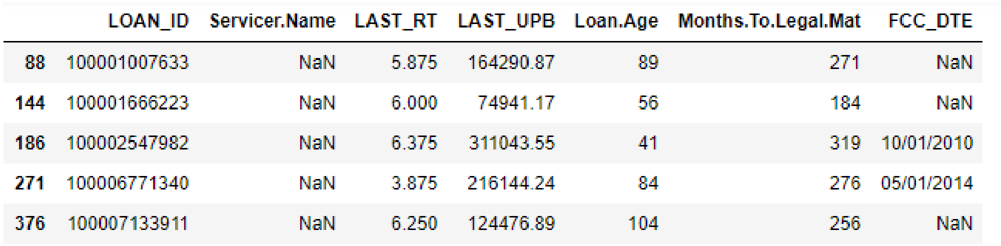
- Current Account Unpaid Balance: LAST UPB

- Loan Age: Loan.Age

- Months to Maturity: Months.To.Legal.Mat

- Foreclosure Date: FCC DTE

After dropping, revised header for performance as below looks more relevant.



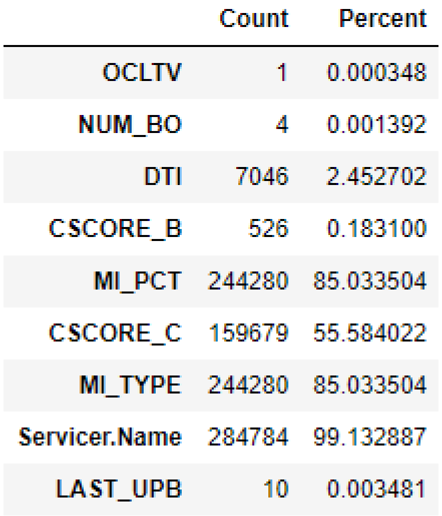
	LOAN_ID	Servicer.Name	LAST_RT	LAST_UPB	Loan.Age	Months.To.Legal.Mat	FCC_DTE
88	100001007633	NaN	5.875	164290.87	89	271	NaN
144	100001666223	NaN	6.000	74941.17	56	184	NaN
186	100002547982	NaN	6.375	311043.55	41	319	10/01/2010
271	100006771340	NaN	3.875	216144.24	84	276	05/01/2014
376	100007133911	NaN	6.250	124476.89	104	256	NaN

Data Merging

Joining the datasets using the LOAN ID column, and further dropping LOAN ID as it will not be of use for identifying defaults. Moreover we will rename the column FCC DTE to Default, as that will be our target variable.

Data Cleaning

The Default column, 1 is placed next to any borrower that has defaulted and 0 to those who has not defaulted. Also, Final dataframe, a merged one has many columns to dealt before we move further



	Count	Percent
OCLTV	1	0.000348
NUM_BO	4	0.001392
DTI	7046	2.452702
CSCORE_B	526	0.183100
MI_PCT	244280	85.033504
CSCORE_C	159679	55.584022
MI_TYPE	244280	85.033504
Servicer.Name	284784	99.132887
LAST_UPB	10	0.003481

The above table shows the number of null values in each column along with the percentages. Here the DTI columns contains 7046 null values and these null values makes upto 2.45 percentage of the entire column. In summary, dataset has nine columns with at least one null value and in the following case the percentage of nulls is large:-

- Mortgage Interest Percentage: MI PCT

- Credit Scope of Co-Borrower: CSCORE C

- Mortgage Interest Type: MI TYPE

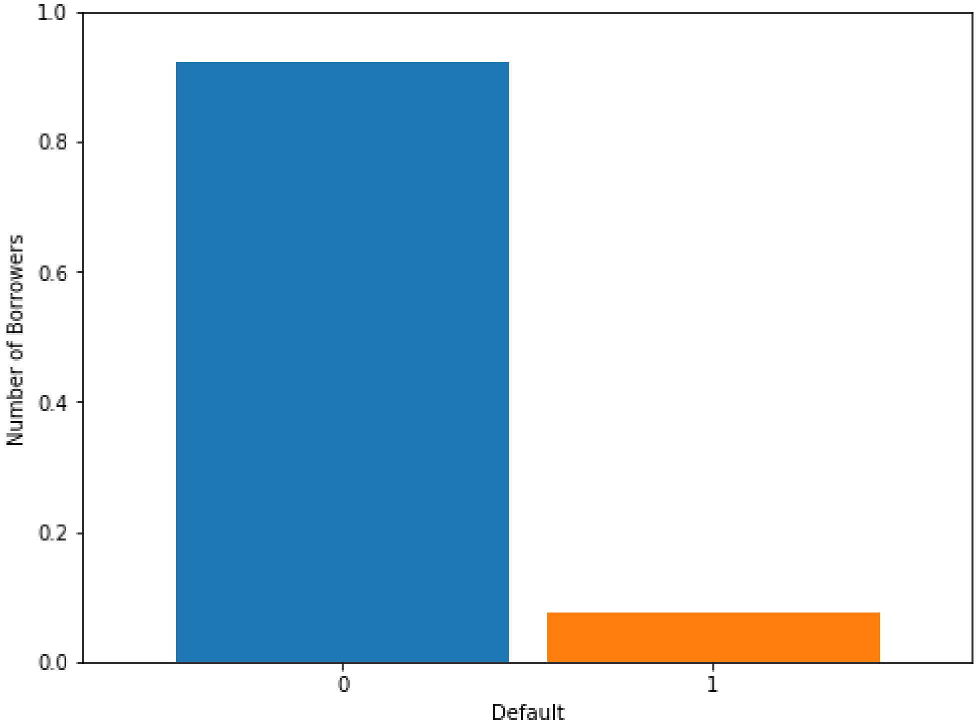
- Service Name: Service.Name

For the above dataset, percentage of nulls is greater than 10 percentage hence we can drop the null rows without much consequence. Also, product type column has only one variable, hence dropping that also.

3. **Data Exploration**

Here by we are further studying the quality of data and identifying the outliers.

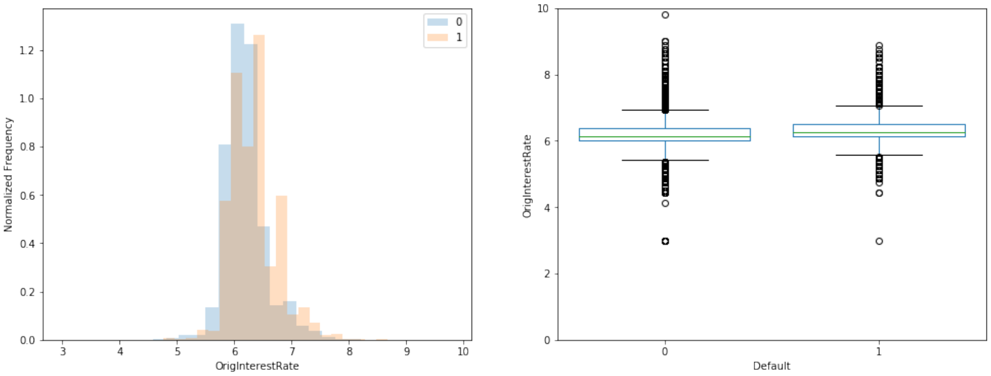
Default Column



The two classes (default = 1, non-default = 0) are extremely imbalanced here; defaulters make up only about 10 percent of all borrowers in this particular dataset. For very imbalanced datasets, it is often the case that machine learning algorithms will have a tendency to always predict the more dominant class when presented with new, unseen test data. To avoid an overabundance of false negatives, we can eventually balance the classes so that the dataframe contains equal numbers of defaulters and non-defaulters. However, lets continue looking at some more of the data first.

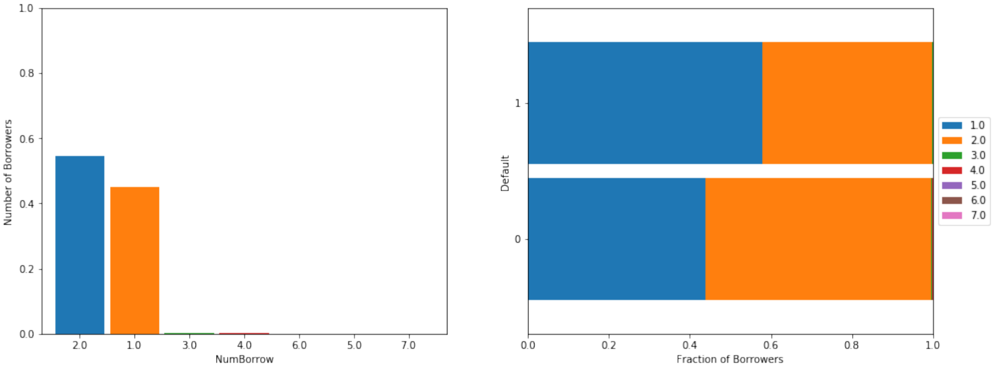
Original Interest Rate

The right-hand figure above shows original interest rate boxplots for both the default and non-default classes. It looks like the original loan interest rates for this time period sat around 6 %. The distributions for the two classes is similar, though interest rates for the default class were slightly higher.



Number of Borrowers

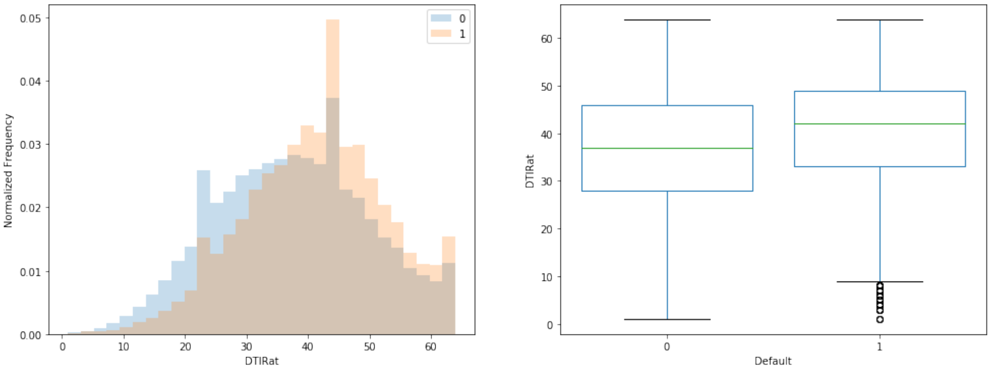
Further looking for relation between number of borrowers on a loan with default.



The vast majority of all loans (nearly 100 %) were for one or two individuals. Presumably the 2-borrower loans were usually for couples. Interestingly, more defaults occurred when only a single borrower was involved, perhaps because 2-borrower households had higher joint incomes and could more easily pay off the loan.

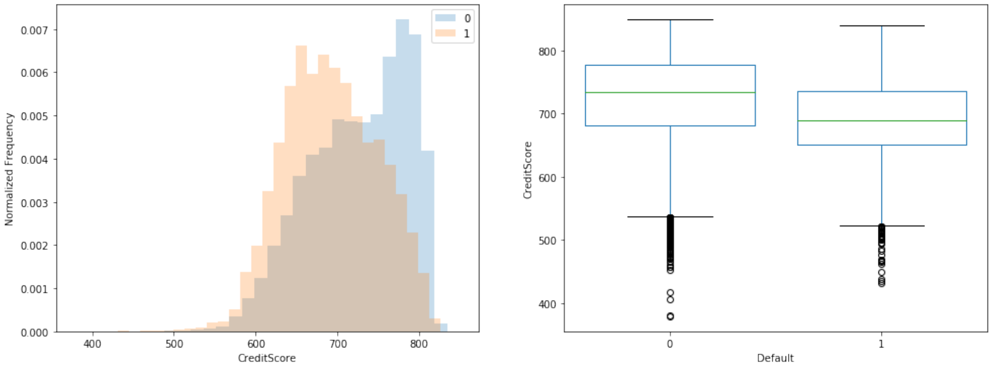
Debt to Income Ratio

Relation between debt to income ration vis--vis default has been explored below. As expected, defaulters typically had higher debt-to-income ratios than non-defaulters.



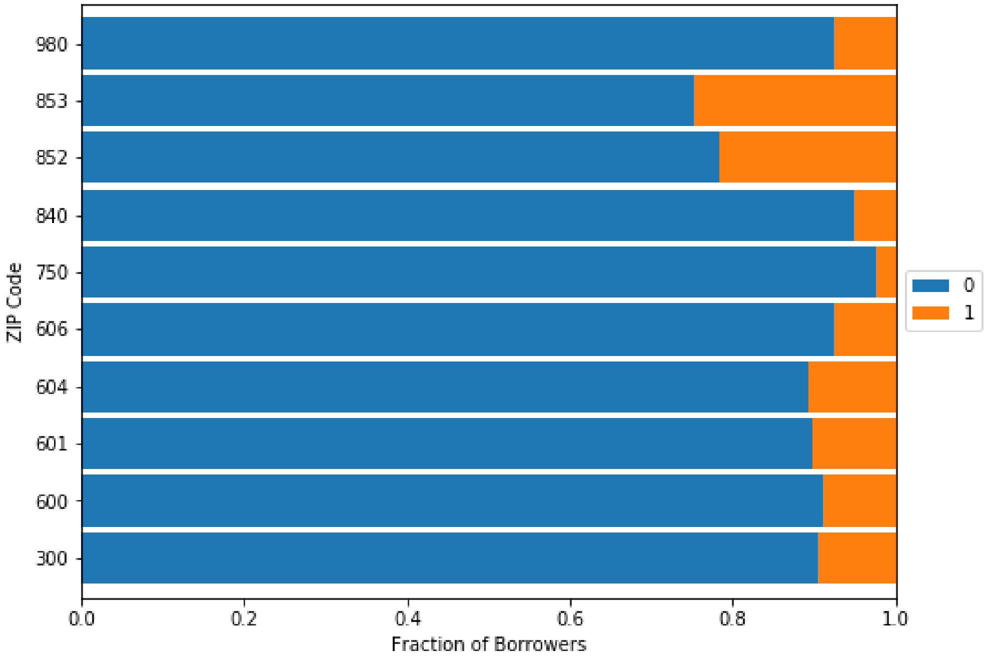
Credit Score

Vast differences can be observed between defaulters and non-defaulters, hence those defaulting had credit score less than 700 and non-defaulters are having more than 700 of credit score, there average around 750 as evident from the attached image:-



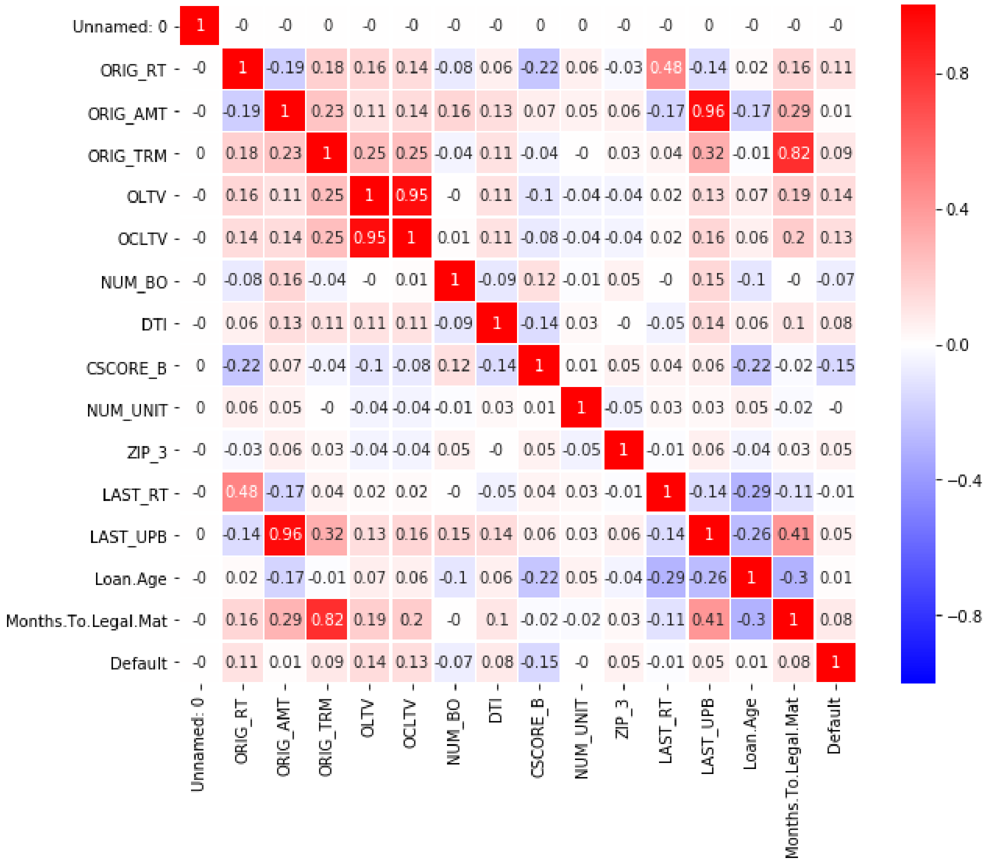
Borrowers Location

Further checking for borrowers location (ZIP Code), figure below shows the fraction of people that have defaulted from the ten most common ZIP Codes. Comparing certain locations (for example, ZIP Code 853 vs. 750), there are significant differences in the fraction of borrowers that defaulted.

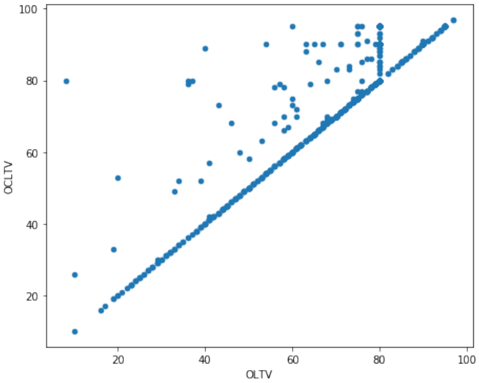


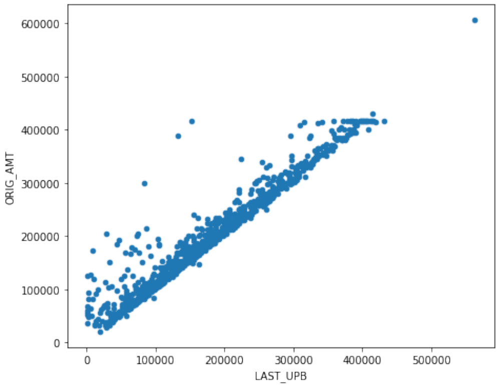
Correlation Features

Checking if any of the elements of our dataset are highly correlated. Highly correlated features can often lead to unstable models with poor interpretability.



We can see that the correlation between Current Account Unpaid Balance (LAST UPB) and Original Unpaid Balance (LAST UPB) is very high 0.96. Further, the correlation between Original CLTV(COLTV) and Original LTV (OLTV) is high as well 0.95. Let's plot these features against one another and see what we find.





It looks like for any value of OLTV or LAST UPB, OCLTV and ORIG AMT show more variability., hence we can drop these two features from the dataset.