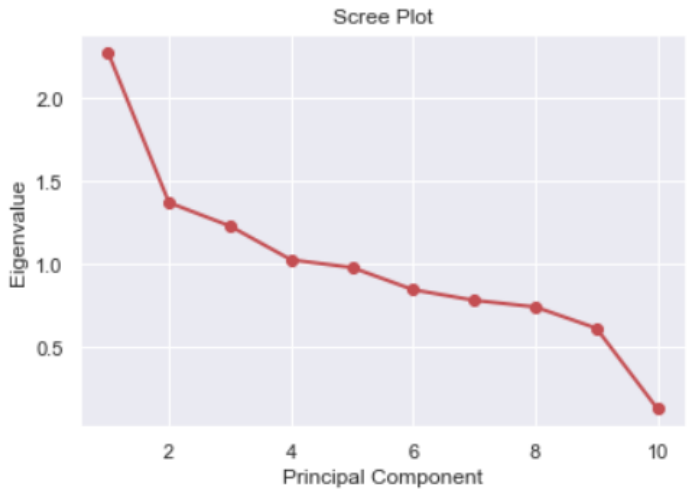
MSDS-422-SEC56 Quick Assignment PCA Daniel Zaremba

PCA Data Observations



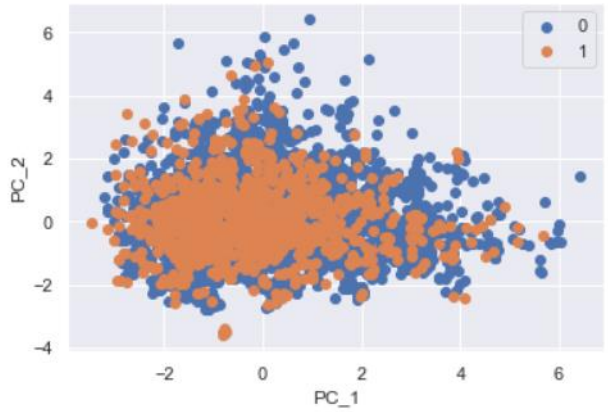

```
2.28 variation= 22 %  total= 22 %
1.37 variation= 13 %  total= 36 %
1.23 variation= 12 %  total= 48 %
1.03 variation= 10 %  total= 59 %
0.98 variation= 9 %  total= 68 %
0.85 variation= 8 %  total= 77 %
0.78 variation= 7 %  total= 85 %
0.74 variation= 7 %  total= 92 %
0.61 variation= 6 %  total= 98 %
0.13 variation= 1 %  total= 100 %
```

How to know how many Principal Components to use, using a Scree Plot look for where the line is under 1. Looking at the Scree Plot above the line is under 1 at PC5.

Out[44]:

	0	1	2	3	4
PC_1	-2.65841	-1.42711	-2.97271	-0.987707	-0.337499
PC_2	0.228203	0.928828	0.318572	-0.338656	-0.0345762
PC_3	-0.60107	0.32643	-0.569613	-0.385871	-1.85886
PC_4	-0.28366	2.3236	-0.336563	0.573674	1.15291
PC_5	-0.861568	-1.37411	-1.17241	-1.37372	-1.05699
TARGET_BAD_FLAG	1	1	1	1	0
TARGET_LOSS_AMT	641	1109	767	1425	NaN
IMP_REASON	HomeImp	HomeImp	HomeImp	MISSING	HomeImp
IMP_JOB	Other	Other	Other	MISSING	Office

The above screenshot shows that the target variables and the categorical variables were appended back to the dataset.



Notes & Observations

PCA is method of using mathematical formulas to combine multiple variables into one variable the new variable will be called a Principal Component or PC

Benefit 1: It helps with dimensionality reduction, which makes things faster by reducing the size of dataset to be stored and processed

Benefit 2: Removes Correlated Features, PCA does this for you efficiently. After implementing the PCA on the dataset, all the Principal Components are independent of one another. There is no correlation among them. Improves Algorithm Performance. Improves Visualization: It is very hard to visualize and understand the data in high dimensions. PCA transforms a high dimensional data to low dimensional data (2 dimension) so that it can be visualized easily. We can use 2D Scree Plot to see which Principal Components result in high variance and have more impact as compared to other Principal Components.

PCA is method of using mathematical formulas to combine multiple variables into one variable the new variable will be called a Principal Component or PC

For this assignment it was imperative to handle the outliers prior to normalizing the variables, in this exercise we dropped the outliers that were 3+ standard deviations.