# Credit Risk Similarity

Jonathan Bown, Christopher Harker, Travis, Tiner

## 1. - Introduction

From 2007 to 2009, the world economy experienced its worst decline since the Great Depression. Millions of Americans lost their jobs and their homes as the subprime mortgage bubble burst, and financial markets froze.

As a result of the economic collapse, the Dodd-Frank Act was enacted. This law requires the Federal Reserve and all large bank holding companies (50 billion in assets or more) to conduct an annual stress test which are then used to check whether a bank's system could theoretically handle a future economic shock. To prevent a repeat of taxpayer-financed bailouts, the Federal Reserve requires banks to maintain capital cushions that would allow them not only to stay afloat, but also to keep lending during periods of intense financial stress. It the results of these stress tests that show whether banks have enough capital to survive another recession.

A key component of these stress tests is the estimation of probability of default and its migration under hypothetical stress scenarios. This risk assessment involves studying customer relationships and behaviors, often grouping similar customers together based on the amount and magnitude of the risk they pose.

Using Fannie Mae single-family mortgage data, we attempt identify groups of customers with various degrees of risk by clustering obligor data. We attempt several clustering methods on the data, focusing mostly on K-means, and Gaussian Mixture Models. We compare these clustering results to the clusters given by performing dimensionality reduction before the clustering. The methods focused on in this paper produce clusterings that adequately group customers of various degrees of risk and whose characteristics align with those expected given the default rates in each cluster, allowing us to not only group together customers by the amount of risk they pose, but also allowing us to understand the characteristics underlying high risk loans.

## 2. - Experiments

In this section we describe our experiments and the analysis we conducted. We first describe the data set, including the variables used and how we derived certain variables, and the processing conducted on it, such as filling missing values and normalizing columns. We also describe the time analysis we conducted as well as a brief attempt at metric learning. Finally, we describe the clustering and the dimensionality techniques used and the metrics used to evaluate the performance.

## 2.1. - Data

### 2.1.1 - Data Source

The data set we explored contains single-family loan performance data from Fannie Mae. Each row represents a single-family mortgage along with its corresponding acquisition and performance attributes. The data is free, open to the public and can be obtained by creating an account on Fannie Mae's website and downloading the data quarter by quarter.

Originally, the acquisition data and loan performnce data were provided as two separate data sets. The acquisition data contains loan characteristics at the time of origination while the performance data describes loan performance over time. The first step in processing the data is the combining of these two data sets. To do this, we use a SAS script provided by Fannie Mae, which aggregated the performance data to the last point in time and joined this aggregated data to the origination characteristics from the acquisition set. After aggregating, we end up with a 25 GB file.

Due to the size of the data, we narrow our focus on Utah. After filtering out loans from Utah, we end up with a data set of 49 MB (498,578 observations).

### 2.1.2. - Loan Characteristics

The combined data set originally had 106 columns. Several of these columns describe costs and losses that occur after a loan has defaulted and will not be helpful in identifying risky customers. After examining each feature, and using our experience developing credit risk models, we decided on 11 variables that we felt would be the most significant in determining risky customers. These 11 variables are listed in the table below.

Table 1: Fannie Mae loan performance and aquisition variables used

| Variable | Type | Definition |
|---|---|---|
| orig_rt | Continuous | Interest Rate at Originiation |
| oltv | Continuous | Ratio of loan amount to value of home at origination |
| num_bo | Continuous | Number of Borrowers on Loan |
| dti | Continuous | Debt to income ratio |
| purpose | Categorical | Loan Purpose |
| cscore_b | Continuous | Borrower Credit Score At Origination |
| fthb_flg | Categorical | First Time Home Buyer Indicator |
| prop_type | Categorical | Property Type |
| num_unit | Categorical | Number of Units |
| occ_stat | Categorical | Property Type |
| mi_pct | Continuous | Primary Mortgage Insurance Percent |

Note that we do not consider using data more granular than the state level because the Fair Lending Act and Equal Credit Opportunity Act prohibit discrimination based on geography.

### 2.1.3. - Derived Variables

We also include some derived variables in our analysis. In addition to the variables described in Table 2 below, we also used one-hot encoding to create dummy variables for the categorical variables mentioned in Table 1 above. All in all, we used a total of 24 variables in our analysis. The Loan

Status variable was only used in evaluating the clusterings. Therefore, 23 variables were used in our experiments.

Table 2: Derived variables

| Derived Variable | Definition |
|---|---|
| coborrower_ind | Whether or not the borrower had a coborrower. Derived from the Coborrower Credit Score at origination, it has a value of 1 if true, otherwise 0. |
| loan_status | Defines the current status of the loan: 1 - Current, 2 - Late (Payment $>$ 30 days past due), 3 - Default (Payment $>$ 180 days past due). |

### 2.1.4. - Handling Missing Values

Before beginning our analysis, we examined the data for any missing values. Of the variables mentioned in Table 1, only three of them had missing values that we had to worry about. The number of missing values and how we handled them are described in Table 3 below. Given that the number of missing values for each variable is extremely small compared to the overall size of the data set ($\approx 1.7\%$ missing for DTI), how we handled them should have little impact on our overall results.

Table 3: Handling missing values

| Variable | No. Missing | Description |
|---|---|---|
| num_bo | 68 | Filled with the most common value, 2.0 |
| cscore_b | 1372 | We assumed a missing value indicated that the borrower did not have nor provide a credit score. Therefore, we assign missing values a credit score of 300. |
| dti | 8472 | Filled with the mean DTI over the entire dataset |

### 2.1.5. - Normalization

While most of the variables used in our experiments fall in the range of $[0, 1]$, there are a couple, specifically credit score and the number of borrowers, that do not. Therefore, we normalize our variables to ensure that all variables fall within this range.

We understand that since some of the variables have different units, such as FICO and the number of borrowers, the euclidean distance might not necessarily be the most appropriate distance for this data set. To address this, we attempted to use Multidimensional Scaling and the Mahalanobis distance.

In the case of Mahalonobis distance, we had difficulty preparing sets of known points that are similar and dissimilar to minimize and maximize this learned metric, given that determining which customers are similar to each other is the key focus of this project. We ran into memory issues with multidimensional scaling, even on subsets of 10,000 loans.

We applied distance metric learning to the data set and used the DML optimization algorithm from class. We ran into similar memory problems when trying to find the matrix M such that we

could us the Mahalanobis distance between points. To maximize the number of points and minimize the running time, we ran the DML algorithm 15 times with random samples of size 2,000 from each of the three classes. The data is normalized to prevent the final matrix from over compensating for variables with large scales. Then the DML algorithm finds the optimal M, this is repeated and the final matrix M is found by taking the mean of all 15 matrices. Below we show the points that were randomly chosen to compute the distance over.

Table 4: Loans chosen to compare distance

| Point No. | orig_dte | orig_rt | orig_amt | orig_trm | oltv | ocltv | num_bo | dti | cscore_b | mi_pct | default_st |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 07/01/2002 | 6.5 | 126000 | 360 | 95 | 95 | 2 | 33 | 720 | 30 | 3 |
| 2 | 03/01/2003 | 6 | 208000 | 360 | 95 | 95 | 2 | 33 | 608 | 25 | 3 |
| 3 | 03/01/2002 | 7.25 | 100000 | 360 | 77 | 77 | 1 | 35 | 641 | 0 | 2 |
| 4 | 11/01/2003 | 5.875 | 315000 | 360 | 73 | 73 | 1 | 54 | 768 | 0 | 2 |
| 5 | 10/01/2002 | 6.125 | 132000 | 360 | 80 | 80 | 2 | 19 | 758 | 0 | 1 |
| 6 | 02/01/2003 | 5.75 | 205000 | 360 | 80 | 80 | 2 | 25 | 756 | 0 | 1 |

The DML optimization algorithm results in a 9 by 9 matrix $M$ that is used to find the distance between loans with nine variables that were chosen from the original eleven. The resulting distances are shown below. For the sake of simplicity we only show a few of the computed distances.

Table 5: Mahalanobis Distance between loans

| Point No. Tuple | Distance ·100 |
|---|---|
| (1,2) | 0.0600623 |
| (1,5) | 0.002828574 |
| (2,3) | 0.01296889 |
| (3,4) | 0.1492825 |
| (3,5) | 0.04487866 |
| (5,6) | 0.02504693 |

From these computed distances, we indeed see that the most similar loans are number 1 and 5. Which can easily be seen by looking at the rate, amount, credit score, etc. However these distances don't seem to tell us anything about the resulting default status which is to be expected because similarity at origination doesn't imply similarity in eventual late or default status. If further research was to be conducted in this area, we would increase our sample size of each loan status and reevaluate the variables used in order to find the best set of variables for calculating loan similarity. Time and computational resource limits prevented us from using more variables or more loans in the sample.

For the above reasons, and provided that scaling to a range of $[0, 1]$ is an extremely common and accepted practice in credit risk modeling, we are comfortable proceeding using the data normalized to the range of $[0, 1]$.

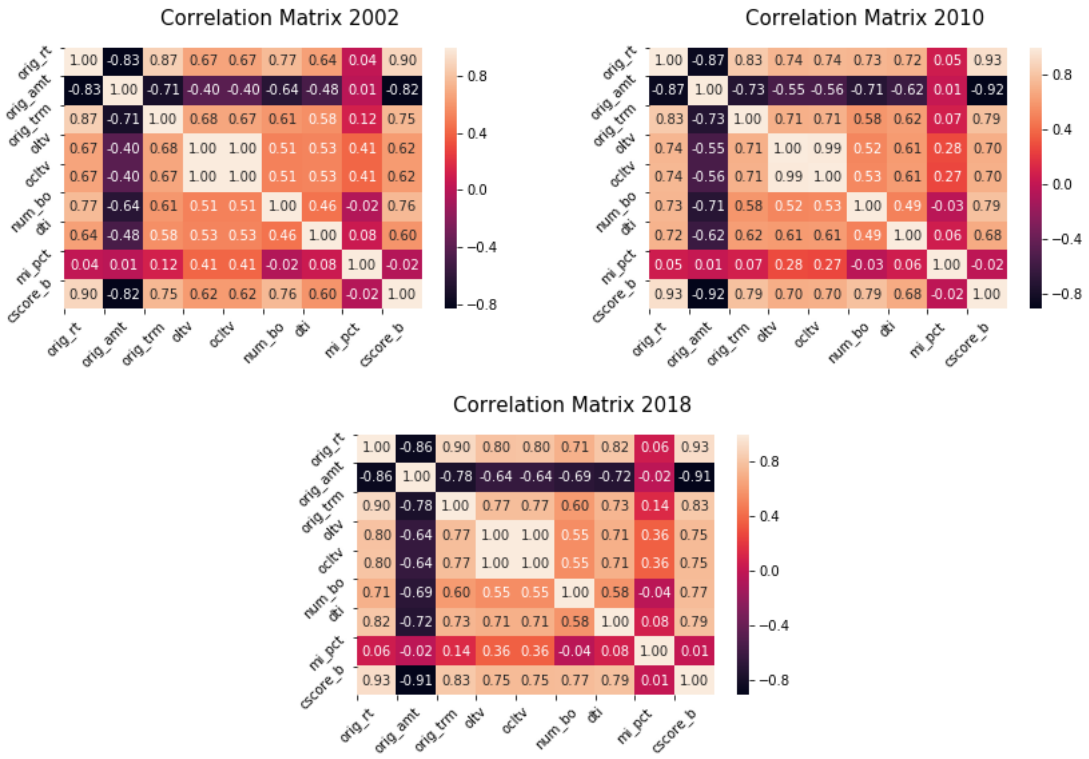
## 2.2 - Time Analysis

The performance data provided by Fannie Mae tracks the performance of loans over time. The fact that the time period provided includes the financial crisis implies that there might be some

difference in mortgage origination characteristics throughout time that can better separate groups or clusters. Our initial idea was to cluster loans in different sets of years to see what characteristics change and cause separation. We attempted several methods to evaluate the significance of the importance of this time component inclusing Singular Value Decomposition, feature importance, and variable correlation over time to see if there would be merit to including this in our clustering. Most of these experiments were exploratory in nature and didn't contribute to our conclusions. The details for these can be found in Appendix B.

The most reliable method chosen for this time component was the correlation matrices. We investigated the correlation matrices of a subset of variables contained in the aggregated data set over nine different periods in time. Three are shown below and the rest are located in Appendix B.2. For some variables, such as original LTV and DTI, the correlation as increased over the years covered in the data set. However, overall the matrices look fairly similar between time periods. We see this as evidence that behavior remains consistent over time and it is thus appropriate to cluster over the entire data set.

Figure 1: Variable correlation over time




## 2.3 - Clustering + PCA

### 2.3.1. - Clustering

As mentioned earlier, a key component of stress testing is the measurement of probability of default, which involves studying customer relationships and behaviors. This is often done grouping similar customers together based on the amount and magnitude of the risk they pose. Naturally, we decided to apply various clustering strategies on the data in order to find groups of similar customers.

We attempt several types of clustering, namely agglomerative clustering, spectral clustering, K-means clusters, and gaussian mixture models. Agglomerative and spectral clustering had memory issues, so we did not further pursue these methodologies. To determine the optimal number of clusters for K-means clustering, we calculated the sum of the squared distances to the closest center for various numbers of clusters and chose the "elbow" point. For the gaussian mixture model, we calculated the Bayesian Information Criterion (BIC) for the given clustering for various numbers of clusters. The gaussian mixture model also has a parameter describing the type of covariance parameters to use. Therefore, we also consider the four possible options available in the SkLearn library. These options are described in more detail in section 3.

#### 2.3.2. - PCA

Dimensionality Reduction was also performed using Principal Component Analysis. While the number of dimensions in the original data set is not high (we only use 23 variables), we hope that we might be able to remove some of the noise by clustering on the first few components and that we will be able to better visualize the clusters. We specifically look at the first three components and the features that have the most influence over them. Also, as we will show below, clusters are more visible if viewed from the basis formed by the first three principal components. Therefore, we also perform K-means clustering and fit a gaussian mixture model on the first few components, hoping to improve the quality of the clusters.

We select the optimal number of clusters as before, except we have the added parameter of how many principal components to use. Therefore, for K-means, we examine the elbow plot for K-means performed using a various number of clusters and components. For the gaussian mixture model, we perform the selection similarly, except we use BIC.

#### 2.3.3. - Evaluating Quality of Clusters

To evaluate the quality of the clusters, we primarily use two methods. For the first method, we use silhouette analysis. Using the optimal number of clusters we found earlier, we calculate a silhouette score, calculated as the difference of the mean nearest-cluster distance and the mean intra-cluster distance divided by the maximum of the two. However, due to the large sample size, we have to use sub-samples of the data to calculate the score. Therefore, we take a sub-sample of 10,000 observations and calculate the silhouette score. After doing this 1,000 times, we use the mean as the final silhouette score.

The silhouette score can take on values in the range $[-1, 1]$. Negative values generally indicate that a sample has been assigned to an incorrect cluster, while a higher score generally indicates dense, well separated clusters. Scores around zero typically indicate overlapping clusters.

The second method used to evaluate the quality of the clusters involves examining the characteristics of the customers in each cluster, especially the mean default rate of the customers in each cluster. A good clustering will provide a good separation of customers with low default rates and high default rates.

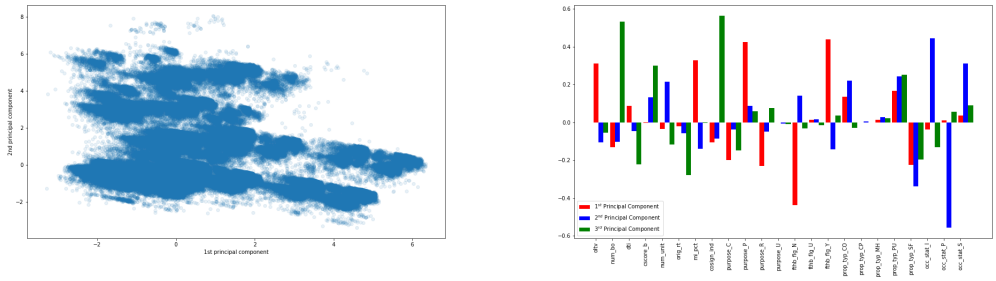
## 3 - Results and Discussion

### 3.1 - PCA

We ran Principal Component Analysis on the data set and examined the first two and three principal components. The plot below is a scatter plot of the first two principal components. As we can see,

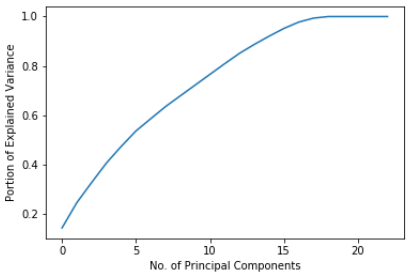
some clear clustering can be seen. We also examined the attributes that have the most influence over the first three principal components. Interestingly, the first time home buyer indicators and the original LTV had the most influence over the first principal component, while the occupancy status indicators had the most influence over the second principal component.

Figure 2: Initial PCA results




However, the first few components do not seem to explain a large portion of the variance within the data set. As seen below, the first five components only explain about 40% of the variance and we need 14 components to explain about 90%. Given that the data set only has 23 variables, we do not expect principal components to improve the clustering results much if at all.

Figure 3: Explained Variance

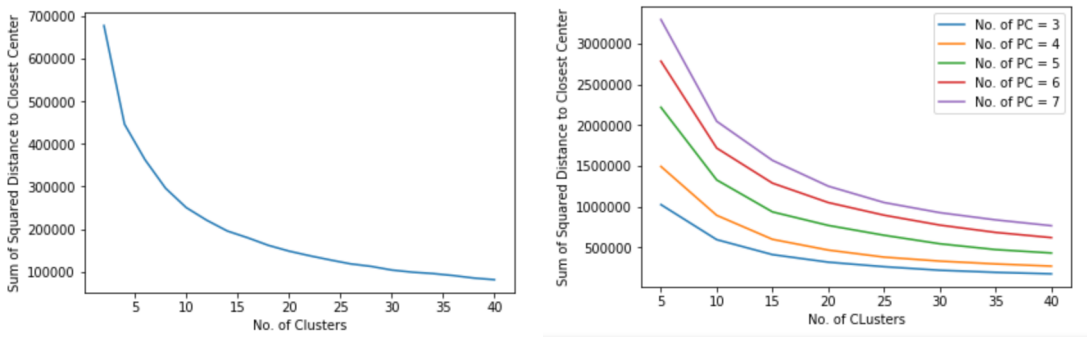



### 3.2 - Clustering Results

We performed K-means clustering before and after Principal Component Analysis. A key part of K-means is determining an appropriate number of clusters to use. We utilized elbow plots, where the sum of the squared distance to the closest center is plotted against the number of clusters. The number of clusters at the "elbow" is generally an appropriate number of clusters to use. However, when K-means is ran after performing principle component analysis, we also have to determine an appropriate number of principle components to cluster on. To accomplish this, we again utilize elbow plots, making a plot for $n = 3, ..., 7$ principal components.

We can see in Figure 4 below, that after around 10 clusters adding additional clusters does not drastically improve the sum of squared distances. Therefore, we choose 10 clusters as the optimal number of clusters. Also, when conducting PCA beforehand, we see that the more principle components we include in the K-means clustering, the larger the sum of squared distances generally is. For this reason, we choose to use three principle components in our clustering. Also after about 15 clusters, we see that the addition of clusters does not improve the sum of squared distances very much. Therefore, we choose to use 15 clusters.

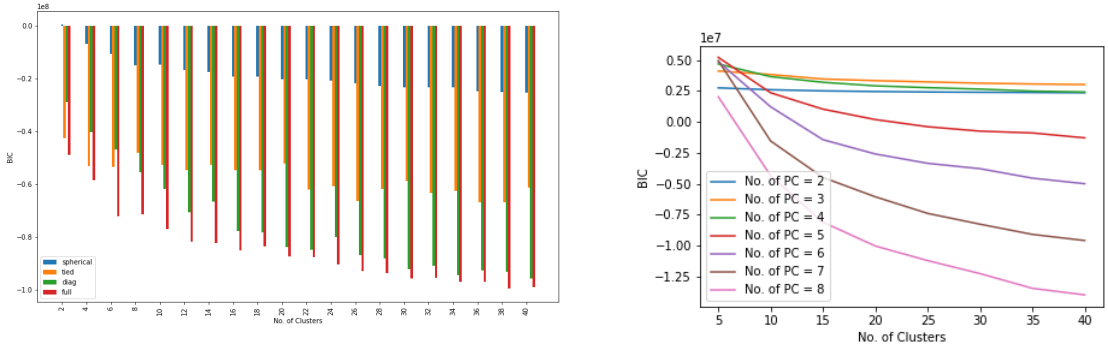
Figure 4: Elbow Plots




We perform a similar analysis with the gaussian mixture model. However, when choosing an appropriate number of clusters to use, we calculate the BIC for a particular clustering. We also need to consider the type of covariance parameters to use. Each component can have its own general covariance matrix (full), all components share the same general covariance matrix (tied), each component has its own diagonal covariance matrix (diag), or each component has its own single variance (spherical). The covariance parameter and the number of clusters that have the lowest BIC is considered an appropriate number of clusters to use. Also, when using GMM after performing PCA, we also have to determine an appropriate number of principle components to cluster on. We perform the same process as before, but using only the "Full" covariance type due to computational limitations, and repeat the processes for a various number of principal compoents.

If Figure 5, we can see that using the "Full" covarience type and 38 clusters gives the lowest BIC when performing only GMM. Interestingly, we see that BIC tends to keep decreasing as we increase the number of clusters. However, we did not explore more than 40 clusters due to computational constraints and to maintain some degree of interpretability. Also seen in figure 5, we see that the BIC also decreases with the number of principal components we use. Thus, for the case when PCA is performed before performing GMM, we use eight principal components with 40 clusters.

Figure 5: Add title




Below in Figure 6-7, we show the results of clustering with K-means and GMM after PCA. We only show clustering results in conjunction with PCA because it is easier to visualize. Also, plots in both 2-dimensions and 3-dimensions are shown because certain clusters are easier to see depending on the perspective. We can see that while some of the clusters are probably assigned appropriately,

there are some that might be overlapping with others. This fact is also supported by our silhouette analysis below.

Figure 6: Clustering with K-means and PCA

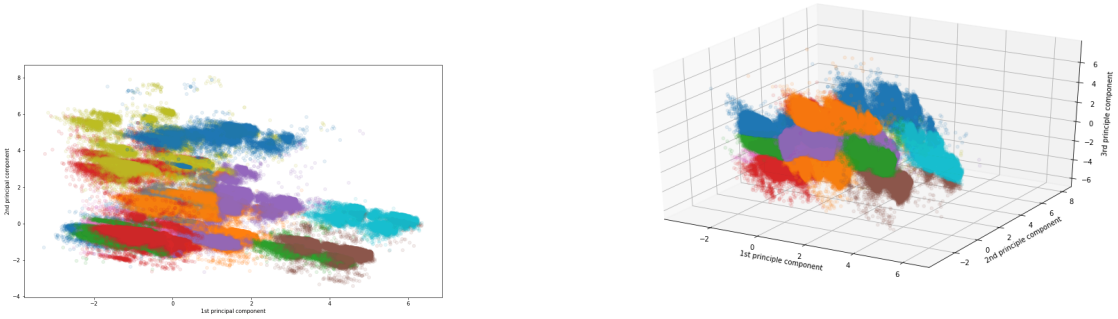
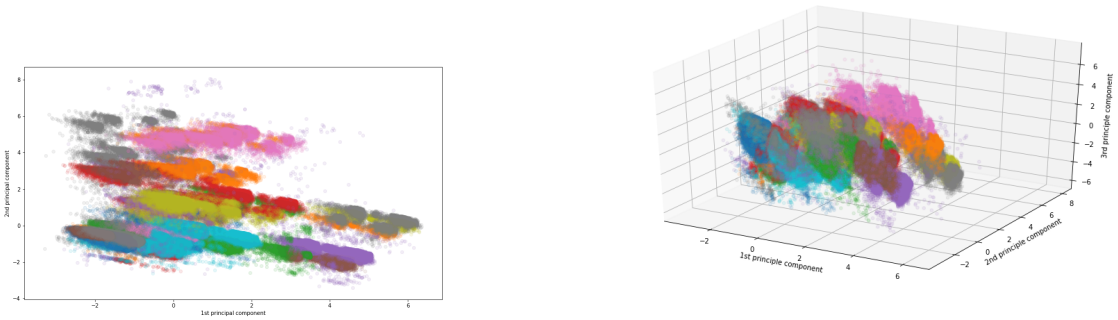



Figure 7: Clustering with GMM and PCA




### 3.3. - Evaluation

The first method used to evaluate the quality of the clusterings was silhouette analysis. Using the optimal number of clusters we found earlier, we calculated a silhouette score, calculated as the difference of the mean nearest-cluster distance and the mean intra-cluster distance divided by the maximum of the two. However, due to the large sample size, we have to use sub-samples of the data to calculate the score. Therefore, we take a sub-sample of 10,000 observations and calculate the silhouette score. After doing this 1,000 times, we use the mean as the final silhouette score.

Table 6 below shows the silhouette score for each of the four clustering methods. The gaussian mixture model had the highest score, followed closely by K-means clustering. These results are not suprising given that 14 components are needed in order to explain 90% of the variance in the dataset. Also, recall that K-means only uses 10 clusters while the Gaussian Mixture Model used 38. Given that the silhouette scores of the two methods are so similar, it seems that the added complexity and granularity of the GMM does not seem to drastically improve the performance when measured by a silhouette score.

Recall that the silhouette score is bounded between -1 for incorrect clustering and +1 for highly dense clustering. The higher scores for the k-means and GMM without PCA indicates that these

two methods produce clusters that are more dense and well separated than the other two methods. While the PCA methods are both positive, they are closer to zero, indicating that their clusters might overlap more than the other methods.

Table 6: Silhouette Analysis

| Method | Silhouette Score |
|---|---|
| K-means | 0.4181 |
| Gaussian Mixture Model | 0.4350 |
| PCA + K-means | 0.3585 |
| PCA + Gaussian Mixture Model | 0.2726 |

The second method used to evaluate the quality of a clustering involved examining how well the clusterings grouped low risk customers separately from high risk customers. In Table 7, we show the portion of the customers that are "Late" and the portion of customers that are in Default. Interestingly, while K-means clustering had the best silhouette score, it does not seem to separate low risk customers from high risk customers as well as the other methods, with the lowest default rate at 1.10% and the highest risk cluster of customers at 3.17%. The GMM performed slightly better, with the lowest risk cluster at 0.73% and the highest risk cluster at 4.7%.

Both k-means and GMM performed better, however, when PCA was performed first. PCA + K-means had a low risk cluster with a default rate of 0.16% and a high risk cluster at 5.32%, while PCA + GMM performed the best, with the low risk cluster at 0.007% and the high risk cluster at 6.28%.

Table 7: Portion of Loans that are Late or in Default by Cluster

| | K-means | | GMM | | PCA + K-means | | PCA + GMM | |
|---|---|---|---|---|---|---|---|---|
| Cluster No. | Avg. Late Ind. | Avg. Default Ind. | Avg. Late Ind. | Avg. Default Ind. | Avg. Late Ind. | Avg. Default Ind. | Avg. Late Ind. | Avg. Default Ind. |
| 0 | 0.068811 | 0.014936 | 0.078304 | 0.012987 | 0.038592 | 0.001555 | 0.090208 | 0.016604 |
| 1 | 0.084346 | 0.016784 | 0.072427 | 0.009739 | 0.112711 | 0.032576 | 0.045284 | 0.007005 |
| 2 | 0.105582 | 0.024678 | 0.121173 | 0.031240 | 0.051408 | 0.009602 | 0.077299 | 0.017424 |
| 3 | 0.080298 | 0.013947 | 0.108212 | 0.032310 | 0.090331 | 0.010418 | 0.106736 | 0.029880 |
| 4 | 0.075333 | 0.013415 | 0.142857 | 0.030612 | 0.072382 | 0.014346 | 0.127549 | 0.029434 |
| 5 | 0.112662 | 0.018421 | 0.113106 | 0.021544 | 0.092345 | 0.024121 | 0.054214 | 0.012940 |
| 6 | 0.068280 | 0.011044 | 0.083119 | 0.016647 | 0.085143 | 0.013462 | 0.079519 | 0.008879 |
| 7 | 0.101964 | 0.022232 | 0.054342 | 0.011391 | 0.082110 | 0.019624 | 0.103539 | 0.019864 |
| 8 | 0.06149 | 0.010990 | 0.122742 | 0.025753 | 0.135987 | 0.027954 | 0.058770 | 0.012167 |
| 9 | 0.118260 | 0.031704 | 0.077726 | 0.008276 | 0.057396 | 0.009066 | 0.127672 | 0.047605 |
| 10 | . | . | 0.076710 | 0.012440 | 0.081726 | 0.010277 | 0.077246 | 0.011195 |
| 11 | . | . | 0.111066 | 0.018913 | 0.046437 | 0.005117 | 0.088491 | 0.013841 |
| 12 | . | . | 0.073309 | 0.015931 | 0.108988 | 0.020643 | 0.121102 | 0.029029 |
| 13 | . | . | 0.087673 | 0.021213 | 0.171141 | 0.053211 | 0.048119 | 0.005837 |
| 14 | . | . | 0.081288 | 0.021758 | 0.090130 | 0.020671 | 0.085597 | 0.019417 |
| 15 | . | . | 0.114014 | 0.019002 | . | . | 0.092258 | 0.023064 |
| 16 | . | . | 0.061339 | 0.013693 | . | . | 0.061325 | 0.007953 |
| 17 | . | . | 0.085027 | 0.028915 | . | . | 0.123563 | 0.013471 |
| 18 | . | . | 0.094005 | 0.014688 | . | . | 0.059467 | 0.012928 |
| 19 | . | . | 0.108125 | 0.017308 | . | . | 0.097113 | 0.015748 |
| 20 | . | . | 0.147899 | 0.026891 | . | . | 0.167732 | 0.028754 |
| 21 | . | . | 0.100781 | 0.018630 | . | . | 0.101549 | 0.015018 |
| 22 | . | . | 0.105863 | 0.029625 | . | . | 0.076644 | 0.012236 |
| 23 | . | . | 0.079264 | 0.012682 | . | . | 0.096386 | 0.015014 |
| 24 | . | . | 0.070529 | 0.011083 | . | . | 0.068636 | 0.018788 |
| 25 | . | . | 0.052092 | 0.007808 | . | . | 0.125769 | 0.024036 |
| 26 | . | . | 0.086829 | 0.019051 | . | . | 0.110986 | 0.020282 |
| 27 | . | . | 0.098112 | 0.036212 | . | . | 0.077000 | 0.013224 |
| 28 | . | . | 0.110485 | 0.015264 | . | . | 0.102001 | 0.033919 |
| 29 | . | . | 0.046223 | 0.007704 | . | . | 0.130223 | 0.062797 |
| 30 | . | . | 0.099792 | 0.018433 | . | . | 0.072260 | 0.009662 |
| 31 | . | . | 0.111664 | 0.012058 | . | . | 0.099554 | 0.018733 |
| 32 | . | . | 0.066806 | 0.011830 | . | . | 0.091202 | 0.016631 |
| 33 | . | . | 0.050405 | 0.007261 | . | . | 0.086415 | 0.019544 |
| 34 | . | . | 0.066579 | 0.010870 | . | . | 0.119863 | 0.027397 |
| 35 | . | . | 0.115754 | 0.025786 | . | . | 0.036856 | 0.000686 |
| 36 | . | . | 0.127584 | 0.047598 | . | . | 0.084283 | 0.018556 |
| 37 | . | . | 0.097720 | 0.012622 | . | . | 0.106290 | 0.019735 |
| 38 | . | . | . | . | . | . | 0.068388 | 0.014556 |
| 39 | . | . | . | . | . | . | 0.106264 | 0.029941 |

# 4. - Discussion

All four clustering methods seem to perform well depending on the evaluation technique used. The silhouette score indicates that the k-means and Gaussian Mixture Models on their own seem to produced clustering that are reasonably dense and well separated while using PCA beforehand produces clusters that tend to overlap. This overlapping can especially be seen in the Figures 6-7. However, performing PCA beforehand seems to produce clusterings that do a more reasonable job of separating high risk customers from the low risk customers. Given that one purpose of credit risk management is to identify risky customers and behaviors, the results seem to suggest that one should lean towards performing PCA before clustering.

Table 8 below shows the mean value of each customer characteristic for the cluster that had the lowest average default rate (left column) and the cluster that had the highest default rate (right column) for each of the four methods. We can see that the cluster with the lower default rate generally has better credit scores, lower LTV, lower DTI, and lower interest rates than the cluster with the higher default rate, as expected. All four methods show these general characteristics that follow intuition. However, PCA used in conjuction with GMM produces clusters that reflect intuition at a larger magnitude than the others. Regardless, it appears that these methods can be used to group similar customers and the clusters' characteristics can be examined and used to inform capital allocation decisions and other credit risk forecasting models.

Table 8: Customer characteristics by cluster

| | K-means | | GMM | | PCA + GMM | | PCA + K-means | |
|---|---|---|---|---|---|---|---|---|
| Variable | Cluster 8 | Cluster 9 | Cluster 33 | Cluster 36 | Cluster 35 | Cluster 29 | Cluster 0 | Cluster 13 |
| oltv | 70.781969 | 80.240792 | 66.155543 | 88.874891 | 66.191302 | 70.380879 | 44.252303 | 86.204060 |
| num_bo | 1.563844 | 1.161483 | 2.006702 | 1.209243 | 1.288732 | 1.152895 | 2.000000 | 1.795109 |
| dti | 33.510209 | 35.626107 | 30.988438 | 36.053785 | 31.687954 | 34.887106 | 26.517372 | 36.718518 |
| cscore_b | 752.050924 | 743.032403 | 764.034906 | 722.689447 | 748.751988 | 739.663865 | 784.883846 | 710.187085 |
| num_unit | 1.364076 | 1.009042 | 1.000140 | 1.000000 | 1.007668 | 1.543415 | 1.000000 | 1.000000 |
| orig_rt | 5.431242 | 5.365059 | 4.304174 | 5.466964 | 4.944771 | 5.625183 | 4.650456 | 5.881140 |
| mi_pct | 1.429775 | 10.585819 | 0.000000 | 21.291121 | 1.115600 | 1.332242 | 0.000000 | 16.581197 |
| cosign_ind | 0.497234 | 0.000000 | 1.000000 | 0.000000 | 0.057062 | 0.086274 | 1.000000 | 0.621439 |
| late_ind | 0.112662 | 0.100496 | 0.050405 | 0.127584 | 0.084729 | 0.141277 | 0.036856 | 0.130223 |
| default_ind | 0.018421 | 0.022433 | 0.007261 | 0.047598 | 0.013590 | 0.028428 | 0.000686 | 0.062797 |
| purpose_C | 0.260499 | 0.000000 | 0.000000 | 0.000000 | 0.360659 | 0.300538 | 1.000000 | 0.956553 |
| purpose_P | 0.418804 | 0.999549 | 0.000000 | 0.000000 | 0.028304 | 0.354675 | 0.000000 | 0.043447 |
| purpose_R | 0.320697 | 0.000000 | 1.000000 | 1.000000 | 0.610818 | 0.344787 | 0.000000 | 0.000000 |
| purpose_U | 0.000000 | 0.000451 | 0.000000 | 0.000000 | 0.000218 | 0.000000 | 0.000000 | 0.000000 |
| fthb_flg_N | 0.997674 | 0.999417 | 1.000000 | 0.999927 | 1.000000 | 0.999567 | 1.000000 | 1.000000 |
| fthb_flg_U | 0.002326 | 0.000583 | 0.000000 | 0.000073 | 0.000000 | 0.000433 | 0.000000 | 0.000000 |
| fthb_flg_Y | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| prop_typ_CO | 0.133566 | 0.104871 | 0.000000 | 0.000000 | 0.000520 | 0.031766 | 0.000000 | 0.000000 |
| prop_typ_CP | 0.000031 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| prop_typ_MH | 0.000943 | 0.006284 | 0.062971 | 0.000073 | 0.004144 | 0.000865 | 0.000000 | 0.000000 |
| prop_typ_PU | 0.019332 | 0.000000 | 0.937029 | 0.000000 | 0.000453 | 0.004511 | 0.000000 | 0.000000 |
| prop_typ_SF | 0.846127 | 0.888845 | 0.000000 | 0.999927 | 0.994883 | 0.962858 | 1.000000 | 1.000000 |
| occ_stat_I | 1.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.920771 | 0.000000 | 0.000000 |
| occ_stat_P | 0.000000 | 0.938165 | 1.000000 | 1.000000 | 0.999966 | 0.008158 | 1.000000 | 1.000000 |
| occ_stat_S | 0.000000 | 0.061835 | 0.000000 | 0.000000 | 0.000034 | 0.071071 | 0.000000 | 0.000000 |

# 5. - Conclusion

We applied clustering and dimensionality techniques to Fannie Mae mortgage performance data in an attempt to explore whether these tools can be used to identify high risk customers. K-means and Gaussian Mixture Models produce clusters that are reasonable well separated and do a descent job at grouping together customers with similar risk profiles. However, performing PCA before clustering, while producing overlapping clusters, outperforms k-means and Gaussian Mixture Models on their own when grouping together customers based on risk. All these methods do a reasonable job identifying customers with higher risk and could provide important insight when developing models used to forecast risk in stress testing and when making capital allocation decisions.

# Appendices

## A Contributions

Below is a table describing the contributions each team member made to the project. We would like to point out that we describe more analysis here than appears in the final report. We also like to point out that the contributions listed are not as clear cut as described and that each team member was very much involved in most aspects of the project.

Table 9: Team contributions

| Team member | Contribution |
|---|---|
| Jonathan Bown | • Gather and aggregated data.<br>• Time and correlation analysis.<br>• Singular Value Decomposition<br>• Metric Learning<br>• Linear Discriment Analysis |
| Christopher Harker | • Processing Data, including filling missing values and normalization.<br>• K-means clustering<br>• Principal Component Analysis<br>• K-means + PCA<br>• Silhouette analysis |
| Travis Tiner | • K-means clustering<br>• Gaussian Mixture Models<br>• PCA + Gaussian Mixture Models |

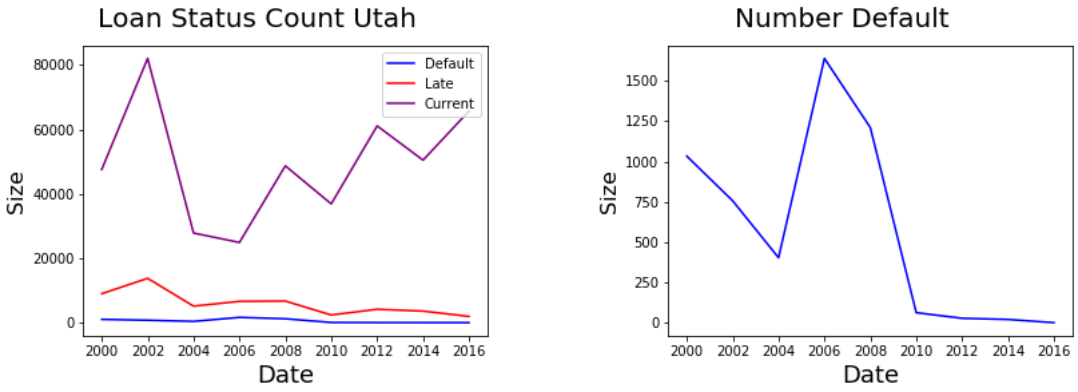
## B Time Analysis

### B.1 Composition over time

Below we show the composition of loans according to the derived loan_status variable described in table 2. Note the imbalance of those that are current to those that are late and in default. However,

these imbalances remain fairly consistent throughout the time span of this data set.

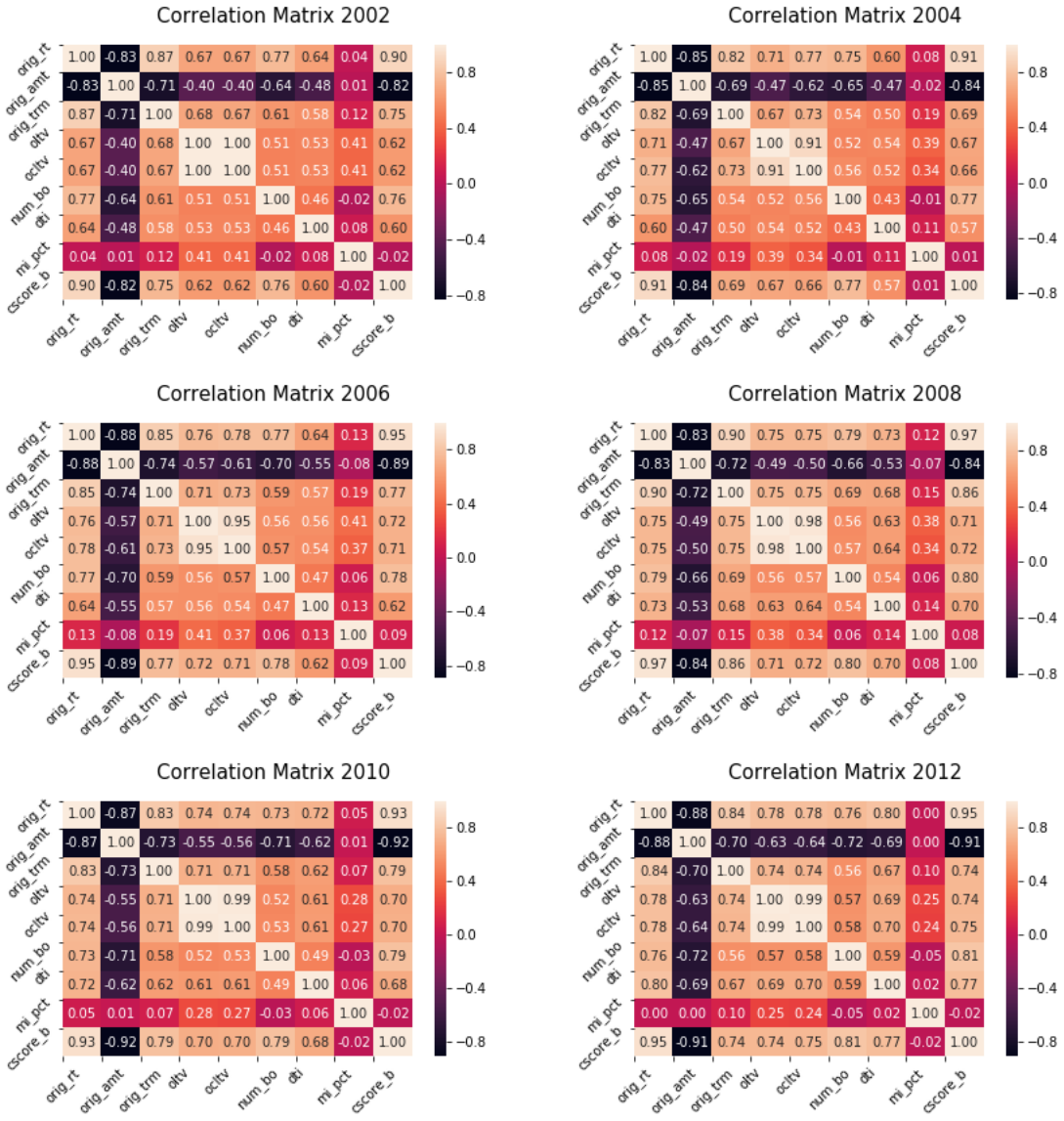
Figure 8: Loan status over time

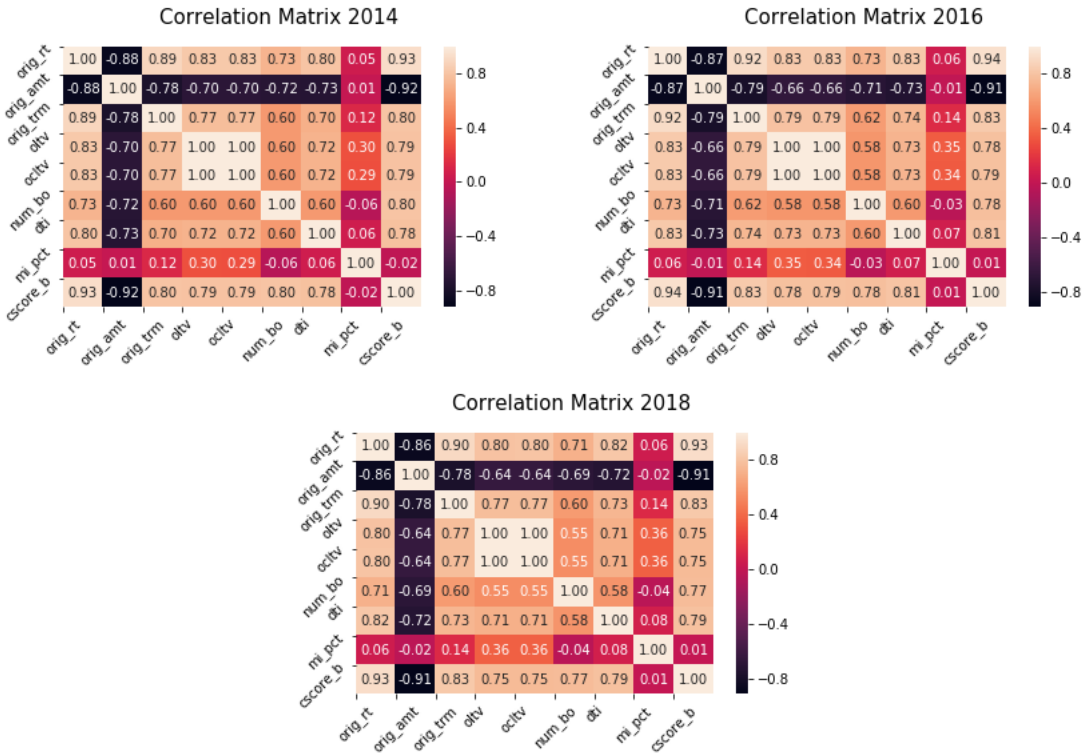



## B.2 Variable Correlation

We include all the correlation matrices found among the two year subsets below. Note that the year in the title corresponds to the loans included from the two years prior to the start of that year. For example, Correlation Matrix 2002 is calculating the correlation among variables for the years before 2002.

Figure 9: Feature correlation



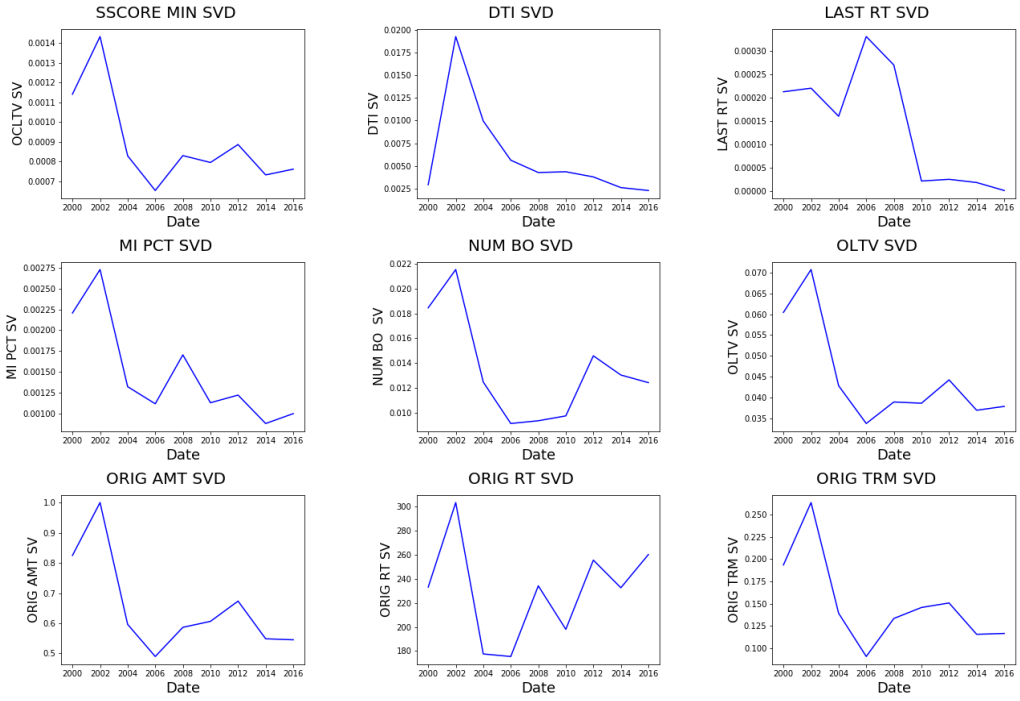



## B.3 SVD

Our first experiment was to take the set of explanatory variables and take the singular value decomposition for each two year group. To remove any influence from the number of loans in each group of years we randomly sampled 10,000 loans and then normalize the values of each variable. The resulting matrix is then run through the singular value decomposition and the singular values are recorded. What we can see in the plots below is that there is some variation over time in the singular value, but the scale of the singular value remains fairly consistent. We weren't surprised that the Original Rate variable is the most influential in creating a best subspace for dimensionality reduction. The rate given on a loan contains a lot of information that is explained by credit score, loan to value, and debt to income just to name a few.

For the SVD it is well known that this decomposition is somewhat unstable. A small change in the data can possibly lead to fairly significant changes in the subspace that is generated or the singular values themselves. This is ultimately why we decided to choose other methods of evaluating the importance of the time component.

Figure 10: Singular values over time

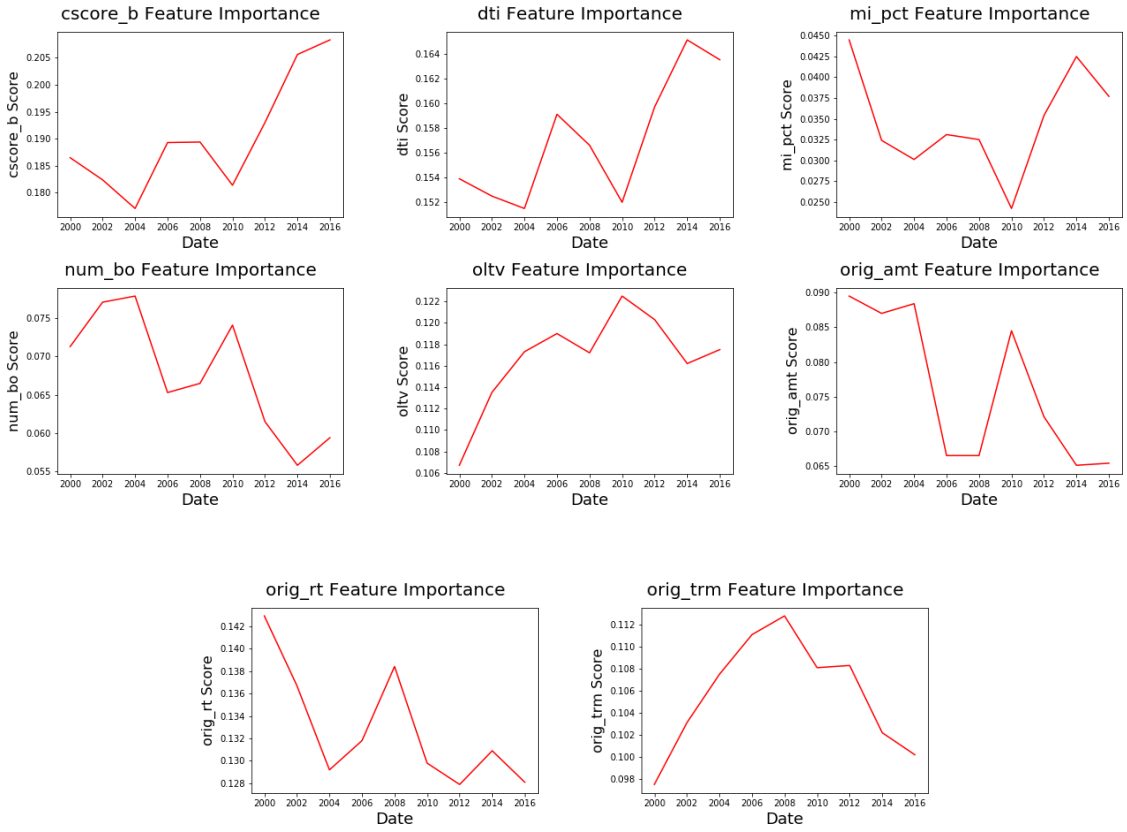



## B.4 Feature Importance

To further investigate if there are variables that change importance over time we take two year subsets and fit a random forest with 100 decision trees. This allows us to get a feature importance score for each variable. We then plot the feature importance score of a certain variable across these nine groups. The plots below indeed show that there is some clear variation over time. As expected the features vary in importance for these types of classification but the changes and overall importance of each variable did not seem to be influenced by the difference in time of origination.

Notice below that the origination characteristics before the rate is set such as credit score, debt to income, and original loan to value have become more influential in predicting loan status than the rate, amount, and percent mortgage insurance. This is one indication that origination characteristics outside the banks control have become more understood in the risk assessment process.

Figure 11: Feature importance over time




## B.5 Linear Discriminant Analysis

Linear Discriminate Analysis was considered as a method to incorporate into our overall goal of separating loans that default and those that don't. However, as can be seen by the histograms below, the assumptions of normality required for this method are violated. Variables such as credit score, loan to value, last rate, and original amount don't have a distribution shape that would suggest normality. Despite these violations, LDA was attempted to build some sort of boundary between defaults and non defaults. As can be seen in Figure 8 this is already tough to do because of the low proportion of defaults. Results weren't promising and much wasn't done further with this method in order to prioritize other techniques.

Figure 12: Variable histograms