Cairo university - Faculty of economics and political science - Statistics department - Arabic section

Multivariate Analysis of loan data

Authors

Saif El-Din Khaled - Alaa Abdullah - Doha Atef - Shaimaa Abdulrahim

Presented to

Dr. Nesma Saleh

Table of Contents

Introduction and Research Questions

In this project the data used is loan data which is a publicly available data from LendingClub.com and the data is published on kaggle.
Loan data will be used to classify which of the customers requesting a loan will pay back their loan in full. This classification using multivariate analysis techniques such as Factor analysis and Cluster analysis is beneficial for banking institutions and lending platforms to find out what are the factors affecting the ability of their customers to pay back their loans in full. The analysis is conducted in the R statistical package.

Research questions explored in this project are:

1. Investigating the interdependence between variables through discovering latent variables and finding indices using Factor analysis.

2. Identifying structures within the data that segments customers into two groups (Able to pay back the loan / Not able to pay back the loan) using Cluster Analysis.
3. Inference between two samples (loan fully paid - loan not fully paid) with the variables fico score and debt to income ratio.

Methodology and Analysis

Descriptive Statistics

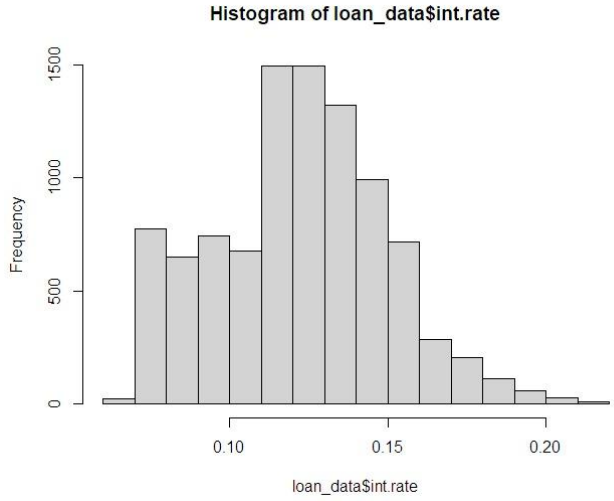

Figure 1

From Figure 1 it shows that most customers choose to have an interest rate on their loan between 0.10% and 0.15%. The mean interest rate is 12% the minimum is 6% and the maximum is 21%

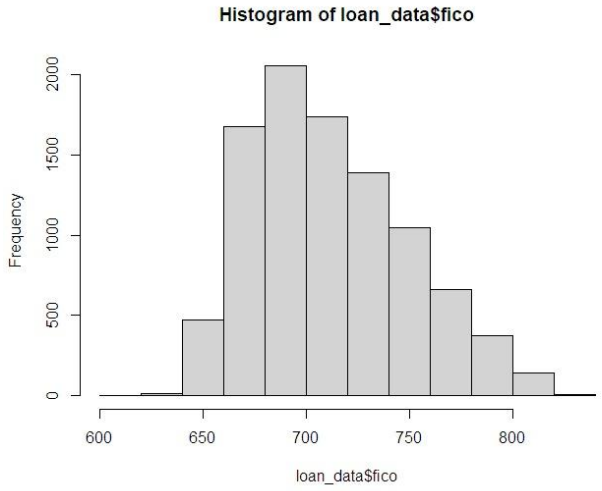

Figure 2

From Figure 2 it shows that most customers that request a loan have a fico score of 700, fico score ranges from 300 to 850, a fico score of 700 is generally considered good. The mean fico score is 710.8, the minimum in the data is 612 and the maximum is 827.

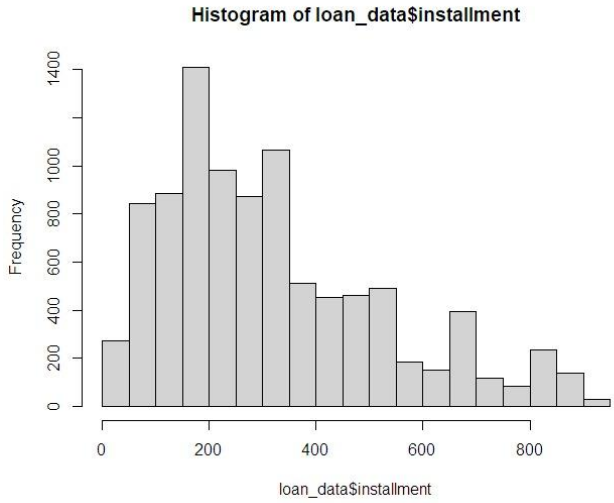

Figure 3

As shown in Figure 3, Most customers choose their monthly installment to range from 200 to 400 per month. The mean of monthly installments is 319

the minimum is 15.67 and the maximum is 940.14 there is a large spread in the monthly installment data.

Investigation of outliers

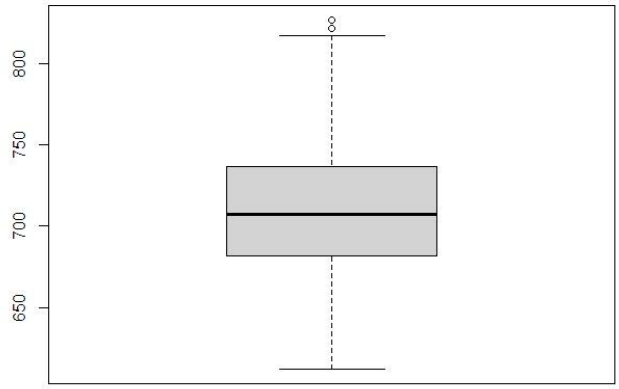

Figure 4(fico score outliers)

There are two outliers in the fico scores variable which are customers who has an excellent fico scores (above 800) these outlier values are not unreasonable as there are customers who maintain an excellent fico score in order to make it easy for them to request loans so there outliers are not omitted from the dataset.

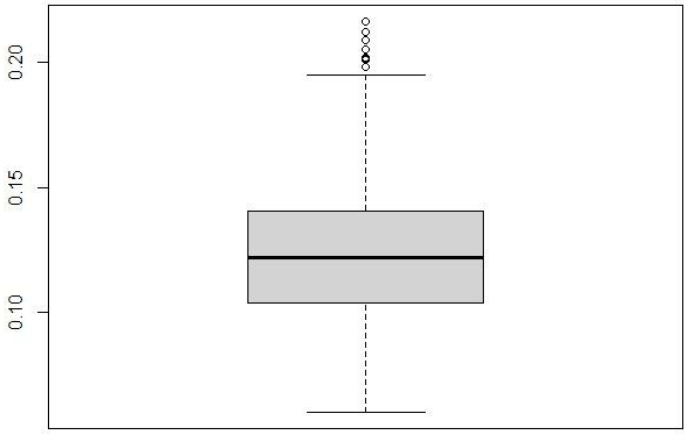

Figure 5(interest rate outliers)

As shown in figure 5 there are several outliers in the interest rate variable and these outliers are also not omitted from the data because it is acceptable for customers to request a higher interest rate on their loans.

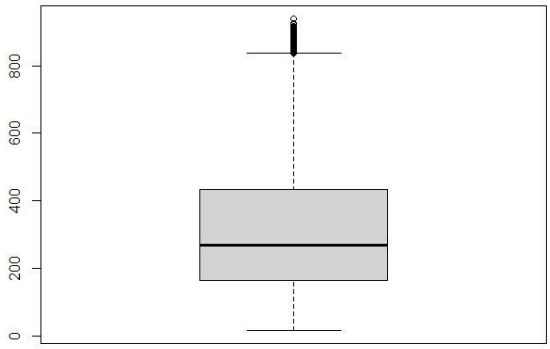

Figure 6 (Monthly installments outliers)

Shown in figure 6 are outlier values that represent customers who choose to pay a monthly installment more than 800 and again it is acceptable for

customers to choose their monthly installment plan so these values are not omitted from the data.

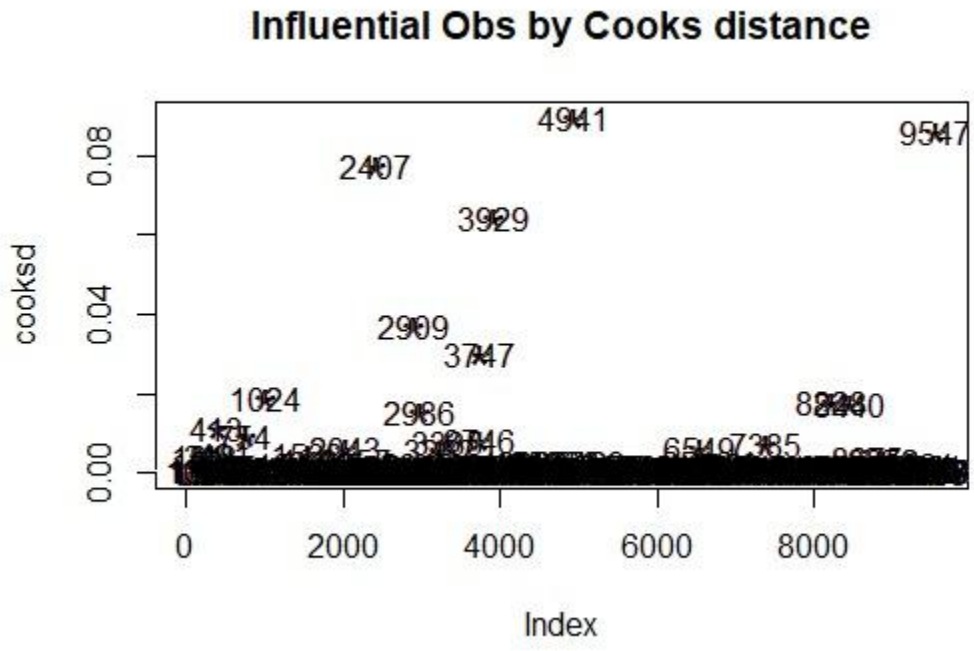

Figure 7

Figure 7 shows all the outliers in the data, these outliers will not be omitted because they are useful in our analysis.

There are no missing values in the data, the data is clean and ready for further analysis.

Factor Analysis

In order to to apply factor analysis, the following steps will be taken:

First: Determining the factors to retain using scree plot that illustrates if applying factor analysis is appropriate and then determining the number of factors to retain.

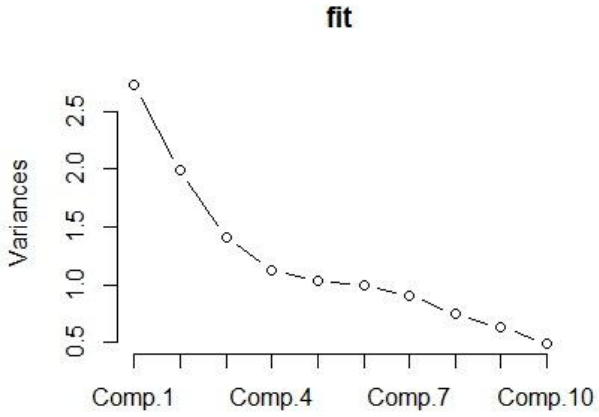

Figure 8 (scree plot)

As shown in figure 8 the scree plot illustrates two findings. First, that factor analysis is appropriate to use, and that the elbow shape of the scree plot breaks at the fourth point so the first three factors will be retained.

Using the proportion of common variance explained technique 5 factors will be retained.

Applying Kaiser-Guttman rule, components with eigenvalues(Variation) greater than 1 will be retained so 5 factors have variation more than 1 and they will be retained.

We will retain three factors according to the scree plot as the results will be more logical.

Principal factor method

Non-rotation

```
> fa <- fa(loan_data, nfactors = 3, rotate = 'none', fm = 'pa')
> fa
Factor Analysis using method =  pa
Call: fa(r = loan_data, nfactors = 3, rotate = "none", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
                    PA1   PA2   PA3    h2   u2 com
credit.policy     -0.53  0.08  0.58 0.625 0.38 2.0
int.rate           0.77  0.11  0.14 0.626 0.37 1.1
installment        0.09  0.56  0.04 0.322 0.68 1.1
log.annual.inc     0.00  0.74 -0.09 0.549 0.45 1.0
dti                0.30  0.09  0.14 0.118 0.88 1.6
fico              -0.90  0.21 -0.23 0.899 0.10 1.2
days.with.cr.line -0.15  0.44 -0.05 0.224 0.78 1.3
revol.bal          0.17  0.49 -0.08 0.274 0.73 1.3
revol.util         0.56  0.14  0.33 0.444 0.56 1.8
inq.last.6mths     0.35 -0.06 -0.59 0.475 0.53 1.7
delinq.2yrs        0.14  0.00 -0.01 0.019 0.98 1.0
pub.rec            0.13 -0.01 -0.01 0.016 0.98 1.0
not.fully.paid     0.21  0.00 -0.10 0.054 0.95 1.5

                       PA1  PA2  PA3
SS loadings           2.34 1.39 0.92
Proportion Var        0.18 0.11 0.07
Cumulative Var        0.18 0.29 0.36
Proportion Explained  0.50 0.30 0.20
Cumulative Proportion 0.50 0.80 1.00
```

Figure 9

As shown in figure 9 and based on the unrotated loadings, the first factor can be interpreted using the variables with loadings >0.6 which are interest rate and fico score.

A high FICO score indicates a low credit risk. It indicates the customer is less likely to send a late payment. Low credit risk helps the customer qualify for a lower interest rate. We can say there is a relationship between fico score and interest rate, as the fico score increases the interest rate that the customer can qualify for decreases. So, the first factor can be called “Credit risk and Interest”.
The second factor reflects the log of annual income.

Varimax rotation

```
Fit based upon off diagonal values = 0.81> fa.v <- fa(loan_data, nfactors = 3, rotate = 'varimax', fm = 'p
a')
> fa.v
Factor Analysis using method =  pa
Call: fa(r = loan_data, nfactors = 3, rotate = "varimax", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
                    PA1   PA2   PA3    h2   u2 com
credit.policy     -0.18  0.02 -0.77 0.625 0.38 1.1
int.rate           0.75  0.09  0.24 0.626 0.37 1.2
installment        0.12  0.55 -0.04 0.322 0.68 1.1
log.annual.inc     0.00  0.74  0.01 0.549 0.45 1.0
dti                0.34  0.08  0.01 0.118 0.88 1.1
fico              -0.88  0.24 -0.26 0.899 0.10 1.3
days.with.cr.line -0.14  0.45 -0.07 0.224 0.78 1.2
revol.bal          0.14  0.49  0.11 0.274 0.73 1.3
revol.util         0.66  0.10 -0.03 0.444 0.56 1.1
inq.last.6mths     0.02  0.01  0.69 0.475 0.53 1.0
delinq.2yrs        0.11  0.00  0.07 0.019 0.98 1.7
pub.rec            0.10 -0.01  0.07 0.016 0.98 1.8
not.fully.paid     0.13  0.01  0.19 0.054 0.95 1.8

                       PA1  PA2  PA3
SS loadings           2.01 1.38 1.26
Proportion Var        0.15 0.11 0.10
Cumulative Var        0.15 0.26 0.36
Proportion Explained  0.43 0.30 0.27
Cumulative Proportion 0.43 0.73 1.00
```

Figure 10

Varimax is an orthogonal rotation method that maximizes the variance of the factor loadings on each factor, while keeping the factors orthogonal. maximizing the variance means that some of the loadings will be close to zero while others will be large. Variables with loadings >0.6 on each factor are used to interpret that factor.

As shown in figure 10 and based on the varimax rotation, the first factor can be interpreted using the variables with loadings >0.6 which are interest rate, fico score and the borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available). In other words, it's how much the customer currently owes divided by their credit limit. The lower the rate, the better. and it is generally advised to have a rate below 30%. Since the revolving line utilization rate also falls in the category of credit risk as well as fico score, The first factor can be called “Credit risk and Interest”.
The second factor can be interpreted using the variables with loadings >0.6 which is the log of annual income.
The third factor can be interpreted using the credit policy and credit inquiries in the last 6 months, too much credit inquiries in a short period of time can negatively affect fico score. This factor can be called “Credit behavior”

Since the “revolving line utilization rate” in the first factor and the “credit inquiries in the last 6 months” in the third factor in the varimax rotated loadings is taken into consideration while calculating the customers fico scores, it would make more sense to take the non-rotated loadings.
In general it is better to take the method that captures more variables, but in our case the variables are already calculated within the fico score variable.

Oblimin rotation

```
> fa.o <- fa(loan_data, nfactors = 3, rotate = 'oblimin', fm = 'pa')
Loading required namespace: GPArotation
> fa.o
Factor Analysis using method =  pa
Call: fa(r = loan_data, nfactors = 3, rotate = "oblimin", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
                    PA1   PA2   PA3    h2   u2 com
credit.policy     -0.08 -0.01 -0.76 0.625 0.38 1.0
int.rate           0.75  0.15  0.09 0.626 0.37 1.1
installment        0.10  0.56 -0.05 0.322 0.68 1.1
log.annual.inc    -0.06  0.74  0.03 0.549 0.45 1.0
dti                0.34  0.10 -0.06 0.118 0.88 1.2
fico              -0.91  0.17 -0.07 0.899 0.10 1.1
days.with.cr.line -0.16  0.43 -0.03 0.224 0.78 1.3
revol.bal          0.09  0.51  0.10 0.274 0.73 1.1
revol.util         0.69  0.15 -0.17 0.444 0.56 1.2
inq.last.6mths    -0.07  0.02  0.71 0.475 0.53 1.0
delinq.2yrs        0.11  0.01  0.05 0.019 0.98 1.4
pub.rec            0.10  0.00  0.05 0.016 0.98 1.6
not.fully.paid     0.11  0.02  0.17 0.054 0.95 1.8

                       PA1  PA2  PA3
SS loadings           2.07 1.39 1.18
Proportion Var        0.16 0.11 0.09
Cumulative Var        0.16 0.27 0.36
Proportion Explained  0.45 0.30 0.25
Cumulative Proportion 0.45 0.75 1.00

 With factor correlations of
      PA1   PA2   PA3
PA1  1.00 -0.02  0.33
PA2 -0.02  1.00 -0.04
PA3  0.33 -0.04  1.00
```

Figure 11

The pattern of relationship between the factors and the variables are not different from those obtained using the orthogonal rotation(which gives the same confidence in the results) As shown in figure 11 the correlations between the second factor and the first factor is very weak (-0.02) and the correlation between the third factor and the first factor is also weak(0.33) then there is no difference between the varimax and oblimin rotations.

Principal component method

Non-rotation

```
> pa.pc <- principal(loan_data, nfactors = 3, rotate = 'none')
> pa.pc
Principal Components Analysis
Call: principal(r = loan_data, nfactors = 3, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
                    PC1   PC2   PC3    h2   u2 com
credit.policy     -0.57  0.13 -0.60 0.698 0.30 2.1
int.rate           0.82  0.05 -0.14 0.699 0.30 1.1
installment        0.14  0.68 -0.01 0.483 0.52 1.1
log.annual.inc     0.03  0.79  0.15 0.647 0.35 1.1
dti                0.42  0.12 -0.35 0.312 0.69 2.1
fico              -0.84  0.25  0.20 0.810 0.19 1.3
days.with.cr.line -0.16  0.62  0.11 0.423 0.58 1.2
revol.bal          0.25  0.62  0.08 0.458 0.54 1.3
revol.util         0.65  0.12 -0.48 0.671 0.33 1.9
inq.last.6mths     0.41 -0.11  0.72 0.703 0.30 1.6
delinq.2yrs        0.18 -0.02  0.09 0.041 0.96 1.5
pub.rec            0.18 -0.02  0.07 0.036 0.96 1.4
not.fully.paid     0.30 -0.03  0.24 0.146 0.85 1.9

                       PC1  PC2  PC3
SS loadings           2.73 1.99 1.41
Proportion Var        0.21 0.15 0.11
Cumulative Var        0.21 0.36 0.47
Proportion Explained  0.45 0.32 0.23
Cumulative Proportion 0.45 0.77 1.00
```

Figure 12

As shown in figure 12 and based on the unrotated loadings, the first factor can be interpreted using the variables with loadings >0.6 which are interest rate, fico score, and revolving line utilization rate. The first factor can be called “Credit risk and Interest”

The second factor reflects installment, log of annual income, days with credit line, and the borrower's revolving balance (amount unpaid at the end of the credit card billing cycle). The second factor can be called “payment ability”

The third factor can be interpreted using the credit policy and credit inquiries in the last 6 months, too much credit inquiries in a short period of time can negatively affect fico score. This factor can be called “Credit behavior”.

Varimax rotation

```
> pa.pc.v <- principal(loan_data, nfactors = 3, rotate = 'varimax')
> pa.pc.v
Principal Components Analysis
Call: principal(r = loan_data, nfactors = 3, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
                    RC1   RC2   RC3    h2   u2 com
credit.policy     -0.05  0.02 -0.83 0.698 0.30 1.0
int.rate           0.73  0.07  0.41 0.699 0.30 1.6
installment        0.15  0.68 -0.02 0.483 0.52 1.1
log.annual.inc    -0.03  0.80  0.03 0.647 0.35 1.0
dti                0.55  0.09 -0.02 0.312 0.69 1.1
fico              -0.76  0.24 -0.42 0.810 0.19 1.8
days.with.cr.line -0.16  0.62 -0.11 0.423 0.58 1.2
revol.bal          0.17  0.64  0.13 0.458 0.54 1.2
revol.util         0.81  0.09  0.03 0.671 0.33 1.0
inq.last.6mths    -0.15  0.02  0.82 0.703 0.30 1.1
delinq.2yrs        0.08  0.00  0.19 0.041 0.96 1.3
pub.rec            0.09 -0.01  0.17 0.036 0.96 1.5
not.fully.paid     0.07  0.02  0.37 0.146 0.85 1.1

                       RC1  RC2  RC3
SS loadings           2.20 1.98 1.95
Proportion Var        0.17 0.15 0.15
Cumulative Var        0.17 0.32 0.47
Proportion Explained  0.36 0.32 0.32
Cumulative Proportion 0.36 0.68 1.00
```

Figure 13

As shown in figure 13 and based on the varimax rotation, the first factor can be interpreted using the variables with loadings >0.6 which are interest rate, fico score and the borrower's revolving line utilization rate. The first factor can be called “Credit risk and Interest”.

The second factor reflects installment, log of annual income, days with credit line, and the borrower's revolving balance (amount unpaid at the end of the credit card billing cycle). The second factor can be called “payment ability”
The third factor can be interpreted using the credit policy and credit inquiries in the last 6 months, too much credit inquiries in a short period of time can negatively affect fico score. This factor can be called “Credit behavior”.
The results with the varimax rotation are similar to the results with no rotation. Then the rotation is of no use.

Oblimin rotation

```
> pa.pc.o <- principal(loan_data, nfactors = 3, rotate = 'oblimin')
> pa.pc.o
Principal Components Analysis
Call: principal(r = loan_data, nfactors = 3, rotate = "oblimin")
Standardized loadings (pattern matrix) based upon correlation matrix
                    TC1   TC2   TC3    h2   u2 com
credit.policy     -0.13 -0.01 -0.80 0.698 0.30 1.0
int.rate           0.77  0.09  0.18 0.699 0.30 1.1
installment        0.13  0.68 -0.04 0.483 0.52 1.1
log.annual.inc    -0.06  0.80  0.06 0.647 0.35 1.0
dti                0.56  0.10 -0.19 0.312 0.69 1.3
fico              -0.82  0.21 -0.18 0.810 0.19 1.2
days.with.cr.line -0.20  0.62 -0.04 0.423 0.58 1.2
revol.bal          0.16  0.65  0.09 0.458 0.54 1.2
revol.util         0.83  0.10 -0.21 0.671 0.33 1.2
inq.last.6mths    -0.08  0.04  0.85 0.703 0.30 1.0
delinq.2yrs        0.10  0.01  0.16 0.041 0.96 1.6
pub.rec            0.11  0.00  0.14 0.036 0.96 1.9
not.fully.paid     0.11  0.03  0.34 0.146 0.85 1.2

                       TC1  TC2  TC3
SS loadings           2.43 1.98 1.71
Proportion Var        0.19 0.15 0.13
Cumulative Var        0.19 0.34 0.47
Proportion Explained  0.40 0.32 0.28
Cumulative Proportion 0.40 0.72 1.00

 With component correlations of
     TC1   TC2   TC3
TC1 1.00  0.01  0.21
TC2 0.01  1.00 -0.05
TC3 0.21 -0.05  1.00
```

Figure 14

The pattern of relationship between the factors and the variables are not different from those obtained using the orthogonal rotation(which gives the same confidence in the results) As shown in figure 14 the correlations between the second factor and the first factor is very weak (0.01) and the correlation between the third factor and the first factor is also weak(0.21) then there is no difference between the varimax and oblimin rotations. And both rotations results are similar to the non-rotation results, so the rotations are of no use.

According to the analysis above, the decision will be to depend on the non rotated principal component method. With the factors described as follows:

First factor “Credit risk and Interest”: interest rate, fico score, and revolving line utilization rate, the first factor explains 21% of the variance.

Second factor "payment ability": installment, log of annual income, days with credit line, and the borrower's revolving balance, the second factor explains 15% of the variance.

Third factor "Credit behavior": credit policy and credit inquiries in the last 6 months, the third factor explains 11% of the variance.

The three factors explain 47% of the variance which is an acceptable percentage because finance is a social science domain.

Cluster Analysis

We use cluster analysis to split customers requesting a loan into 2 groups based on their eligibility and ability to pay back the loan. and based on this clustering the decision will be made to either give the customer the requested loan or not.

We will group the customers in the sample into two groups

1. Eligible to request a loan
2. Not eligible to request a loan

We start by random centroids.
In order to understand how the two formed clusters differ we will calculate the means of the different variables in each of the two clusters.

```
K-means clustering with 2 clusters of sizes 4198, 5380

Cluster means:
  credit.policy   int.rate installment log.annual.inc        dti       fico days.with.cr.line
1     0.4008194 -0.7811954 -0.08898297    0.002867837 -0.3744089  0.8276845         0.1557604
2    -0.3127583  0.6095647  0.06943318   -0.002237765  0.2921503 -0.6458401        -0.1215394
   revol.bal revol.util inq.last.6mths delinq.2yrs    pub.rec not.fully.paid
1 -0.1639220 -0.6984828     -0.2362567  -0.1653932 -0.1452065     -0.2292646
2  0.1279079  0.5450243      0.1843505   0.1290559  0.1133043      0.1788946
```

Figure 15

The first cluster is characterized by a high fico score and low debt to income ratio and also a high annual income, generally the results of the first cluster indicate a customer that shows a high ability in paying back their requested loan.

The second cluster is characterized by a low fico score and high debt to income ratio and also a low annual income, generally the results of the second cluster indicate a customer that shows a low ability in paying back their requested loan.
We will use visualization in order to better grasp how our clusters are formed.

Figure 16

Color coded red is group 1, these are customers that are eligible to take the loan and demonstrate a high ability in paying it back.
Color coded blue is group 2, these are customers that are not eligible to take the loan and demonstrate a low ability in paying it back.
As it shows in figure 16 there is a higher number of non-eligible customers requesting loans.

We then tried grouping customers into three clusters (eligible - moderately eligible - not eligible) and the results were as follows:

```
K-means clustering with 3 clusters of sizes 1793, 3368, 4417

Cluster means:
  credit.policy   int.rate installment log.annual.inc        dti       fico days.with.cr.line
1    -2.0314965  0.6759771 -0.09804288   -0.055980704  0.2309321 -0.7738768       -0.18755598
2     0.4359979 -0.9219069 -0.12858669    0.016969104 -0.4553733  1.0103284        0.20305160
3     0.4921966  0.4285614  0.13784715    0.009785252  0.2534833 -0.4562429       -0.07869366
    revol.bal revol.util inq.last.6mths delinq.2yrs     pub.rec not.fully.paid
1  0.37574448  0.2738107      1.1385431  0.17304166  0.12471709     0.34682929
2 -0.18602596 -0.8334162     -0.2411354 -0.18447524 -0.15430325    -0.22596855
3 -0.01068019  0.5243385     -0.2783028  0.07042086  0.06703093     0.03151396
```

Figure 17

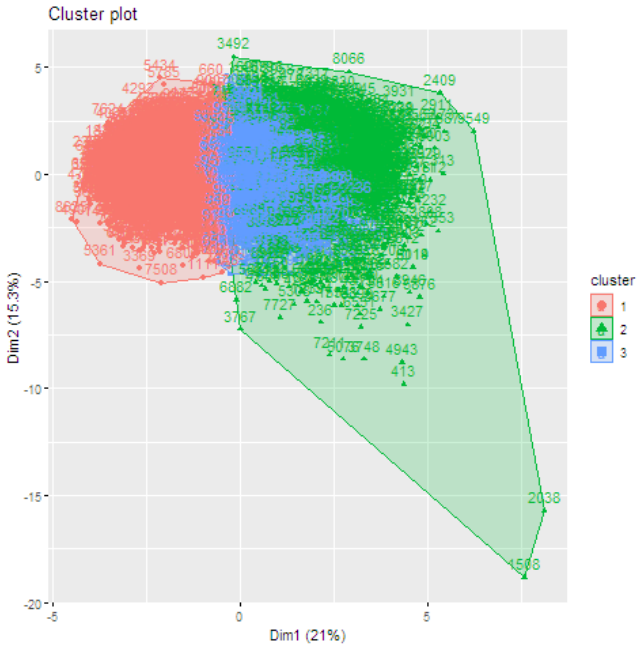

Figure 18

We found out that the three way grouping isn't beneficial and it is better to group customers into two clusters (eligible/not eligible).

Inference two samples

We are going to compare the two means of two independent samples (loan fully paid - loan not fully paid) based on the variables (fico score - debt to income ratio).
First we used Box's M-test for homogeneity of covariance matrices and the results were as follows:

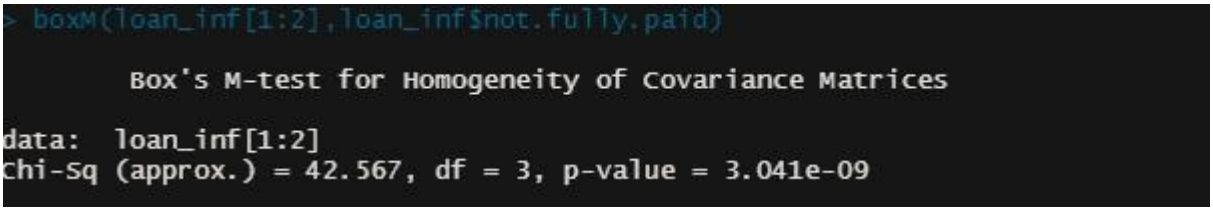

Figure 19

H0: Covariances are equal
H1: Covariances are not equal

According to figure 19 the p-value is < 0.05 that means that we reject H0 that means that the covariances in the two samples are not equal.
Since the sample is large, the central limit theorem is applied.

Using Hotelling's T to compare the two means in the two samples, that is testing if the mean fico score and mean debt to income ratio differs between customers who fully pay back the loan and customers who don't fully pay back the loan.

H0: $\mu 1 - \mu 2 = 0$
H1: $\mu 1 - \mu 2 \neq 0$

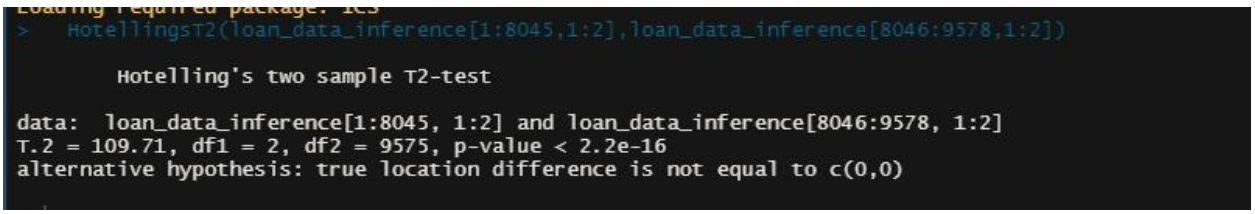

Figure 20

As shown in figure 20 the p-value <0.05 indicates that we reject the null hypothesis and that the two means are not equal. Which is supported by the data since by looking at the fico scores and debt to income ratios of customers that fully pay back the loan and customers that don't fully pay back the loan we find the customers which fully pay back the loan have fico scores much higher than customers that don't fully pay back the loan giving us an early indicator that the means of the two groups are going to be different.

Conclusions

Using factor analysis we have successfully reduced the variables into 3 factors (Credit risk and Interest, Payment Ability, and Credit Behavior)

Using cluster analysis we have successfully grouped customers into two groups according to their eligibility for requesting a loan (eligible/not eligible).

Using inference techniques we have compared the means between the two groups (loan fully paid - loan not fully paid) and found that the means are not equal in the two samples.
The conclusions driven from this report can have two uses:

1. It can be used by banks and lending platforms to aid their decision making in finding out which customers to give loans to.
2. It can be used by individuals seeking to request a loan in order to find out how to better qualify so that their request could be accepted.

References

https://www.kaggle.com/itssuru/loan-data?select=loan_data.csv

Appendix

Variables description:

Here are what the columns represent:

credit.policy: 1 if the customer meets the credit underwriting criteria, and 0 otherwise.
int.rate: The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Borrowers judged to be more risky are assigned higher interest rates.
installment: The monthly installments owed by the borrower if the loan is funded.
log.annual.inc: The natural log of the self-reported annual income of the borrower.
dti: The debt-to-income ratio of the borrower (amount of debt divided by annual income).
fico: The FICO credit score of the borrower.
days.with.cr.line: The number of days the borrower has had a credit line.
revol.bal: The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle).
revol.util: The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available).
inq.last.6mths: The borrower's number of inquiries by creditors in the last 6 months.
delinq.2yrs: The number of times the borrower had been 30+ days past due on a payment in the past 2 years.

pub.rec: The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments).

Codes:

```
fa <- fa(loan_data, nfactors = 3, rotate = 'none', fm = 'pa')
fa
pa.pc <- principal(loan_data, nfactors = 3, rotate = 'none')
pa.pc
fa.v <- fa(loan_data, nfactors = 3, rotate = 'varimax', fm = 'pa')
fa.v
fa.o <- fa(loan_data, nfactors = 3, rotate = 'oblimin', fm = 'pa')
fa.o
pa.pc.v <- principal(loan_data, nfactors = 3, rotate = 'varimax')
pa.pc.v
pa.pc.o <- principal(loan_data, nfactors = 3, rotate = 'oblimin')
pa.pc.o
```

```r
install.packages("factoextra")
library(factoextra)

# Loading dataset
df <- loan_data

# Omitting any NA values
df <- na.omit(df)

# Scaling dataset
df <- scale(df)

# output to be present as PNG file
png(file = "KMeansExample.png")

km.m <- kmeans(df, centers = 2, nstart = 25)
km.m
# Visualize the clusters
s=fviz_cluster(km, data = df)
s
# saving the file
dev.off()

# output to be present as PNG file
png(file = "KMeansExample2.png")

km.alolaaa <- kmeans(df, centers = 3, nstart = 25)
km
# Visualize the clusters
fviz_cluster(km, data = df)

# saving the file
dev.off()
```