CHICAGO BOOTH

The University of Chicago Booth School of Business

Loan Default Analysis

BUSN 41201 (Big Data) - Group 20

May 29, 2023

Arthur Cheib, Yu-Wei Chen, Simone Zhang, Sheng-Hau Peng, Shu-Hsiang Wang

Content

Executive Summary

Our team has conducted an in-depth analysis of a loan default dataset of 30,000 observations and 25 variables/parameters. This somewhat popular online dataset (source: UIC) has now been downloaded more than 700,000 times. Our primary objective was to identify patterns among customers in terms of loan repayment and determine the key variables that contribute to predicting loan default.

Additionally, knowing that most financial institutions deal with a substantial amount of data, we assessed the efficiency of the different models we run, which could be extra information for decision-makers within the company. Also, given some findings, we aimed to provide insights into how financial institutions can effectively target customers.

We employed various modeling techniques to achieve these goals, including regression, clustering, and machine learning. We utilized logistic regression on default, LASSO regression, and principal components analysis (PCA) in the regression category. These models allowed us to understand the relationship between the input variables and the likelihood of loan default and identify significant predictors.

In the clustering category, we applied K-means clustering and hierarchical clustering algorithms. These techniques allowed us to group customers with similar characteristics, enabling financial institutions to segment their customer base better and tailor their strategies accordingly.

Finally, we employed decision trees, random forests, and neural networks in the machine learning category. These models provided more profound insights into complex relationships and patterns within the dataset, enhancing our ability to predict loan default and classify customers accurately.

Our Main Findings:

- The variables representing payment status in different months, demonstrate that information from the recent months carries more weight in determining default probabilities. This suggests that **a person's current payment behavior is a better indicator of their likelihood to default than their past payment history.**
- Our PCA analysis found higher values for its first component, which condense information on individuals who have consistently higher bill statements, suggesting a potential risk for default payment. Also, **higher frequency of lagged repayment over the six-month period seems to indicate a potential financial instability or difficulty in meeting payment obligations** (similar findings of those obtained by the Logistic regression).
- Our cluster analysis revealed two main groups: the first is a financially well-behaved group. Most of them pay the bill on time. They have the lowest default rate. The second is the group with a higher default rate. Given both clusters, **an excellent step for a company interested in dealing with loan defaults would be to obtain background data on clients from different sources and develop targeted, informative content emphasizing small resolutions to help customers attain financial responsibility**.
- About code effiency analysis revelaed that, of our models, the logistic regression model took approximately 1.20 seconds to run, indicating its relatively fast performance. The Lasso model exhibited a longer computation time of around 27.47 seconds and the tree models took longer than 30 seconds. **The clustering method has the been among the fastest one and with very accurate findings**.

Introduction to the Data

Source

Name: I-Cheng Yeh

Email addresses: (1) icyeh '@' chu.edu.tw (2) 140910 '@' mail.tku.edu.tw

Institutions: (1) Department of Information Management, Chung Hua University, Taiwan. (2) Department of Civil Engineering, Tamkang University, Taiwan. Other contact information: 886-2-26215656 ext. 3181

Parameters of the Data

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:

- X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- X2: Gender (1 = male; 2 = female).
- X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- X4: Marital status (1 = married; 2 = single; 3 = others).
- X5: Age (year).
- X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows:
  - X6 = the repayment status in September, 2005;
  - X7 = the repayment status in August, 2005;...;
  - X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months;...; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
- X12-X17: Amount of bill statement (NT dollar).
  - X12 = amount of bill statement in September, 2005;
  - X13 = amount of bill statement in August, 2005;...;
  - X17 = amount of bill statement in April, 2005.
- X18-X23: Amount of previous payment (NT dollar).
  - X18 = amount paid in September, 2005;
  - X19 = amount paid in August, 2005;...;
  - X23 = amount paid in April, 2005.

Exploratory Data Analysis

Data Cleaning

Before start the modelling process, we cleaned the dataset, which did not required a lot of work, because the data was very well prepared. However, to fit our purposes, here are the steps we did to clean the data:

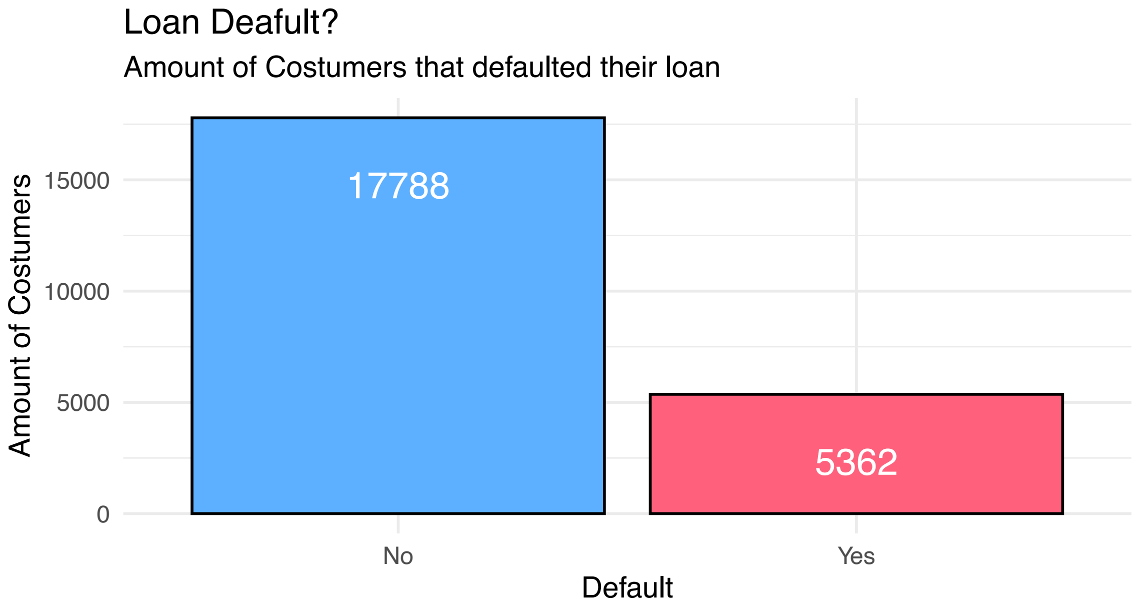
1) Removing Invalid Education Indices: we filtered out any education indices that are not defined or fall outside the valid range (only 1, 2 and 3 were valid).
2) Removing Invalid Marriage Indices: The next step removes rows where the "MARRIAGE" value is not equal to 0. We removed any marriage indices that were not defined in the data dictionary.
3) Handling Specific Data Values: we cleaned colums 6 to 11 (repayment status columns) by first removing rows where the value in a column is -2, because this value is not defined in the dataset. Secondly, we replace any repayment occurrences with 0.
4) Data Recoding: We then transform the columns "SEX," "EDUCATION," "MARRIAGE," and "default" into factors (categorical variables) for easier analysis.
5) Recoding Education and Marriage Categories: we recoded the "EDUCATION" column so that categories 0, 4, 5, and 6 are combined and labeled as "Other." The category 1 is labeled as "Grad.School," category 2 as "University," and category 3 as "High.School." Similarly, the "MARRIAGE" column had its categories 0 and 3 recoded as "Other," category 1 as "Married," and category 2 as "Single."
6) Adjusting the Order of Payment Categories: finally, we recoded the columns "PAY_0" to "PAY_6," which represent payment status for six months. These columns were transformed into ordered factors, where the levels are sequenced from 0 to 9.

```r
# Remove education indices which are not defined
data <- data %>% filter(EDUCATION<=4 & EDUCATION>0)
# Remove MARRIAGE indices which are not defined
data <- data %>% filter(MARRIAGE!=0)
for(i in 6:11){
  # -2 is not defined in the document + index pay duly as 0
  data <- data[data[,i]!=-2,]
  data[data[,i]==-1,i] <- 0
}
colnames(data)[24] <- 'default'

data <- data %>%
  mutate(SEX = factor(SEX, labels = c("Male", "Female")),
         EDUCATION = factor(EDUCATION),
         MARRIAGE = factor(MARRIAGE)) %>%
  mutate(EDUCATION = car::recode(EDUCATION, "c(0, 4, 5, 6) = 'Other';  1 = 'Grad.School'; 2 = 'Universi
         MARRIAGE = car::recode(MARRIAGE, "c(0, 3) = 'Other'; 1 = 'Married'; 2 = 'Single'"),
         default = factor(default, levels = c(0, 1), labels = c("No", "Yes")))

data <- data %>%
  mutate(PAY_0 = factor(PAY_0, order=TRUE,levels=c(seq(0,9, by = 1))),
         PAY_2 = factor(PAY_2, order=TRUE,levels=c(seq(0,9, by = 1))),
         PAY_3 = factor(PAY_3, order=TRUE,levels=c(seq(0,9, by = 1))),
         PAY_4 = factor(PAY_4, order=TRUE,levels=c(seq(0,9, by = 1))),
         PAY_5 = factor(PAY_5, order=TRUE,levels=c(seq(0,9, by = 1))),
         PAY_6 = factor(PAY_6, order=TRUE,levels=c(seq(0,9, by = 1))))
```

Data Visualization

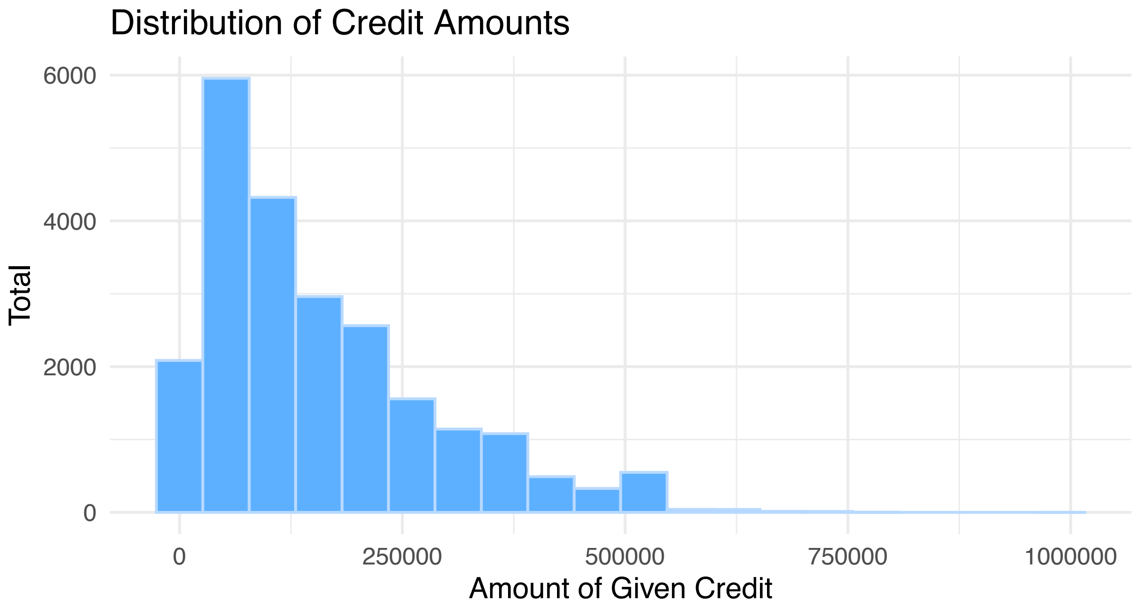
A. Default variable

The dataset consists of a total of 23150 individuals, out of which 17788 individuals do not default, and 5362 individuals default. Therefore, the overall default rate is 23.16%.



B. Amount of given credit (X1)

From the plot and table, it can be observed that the distribution appears to be right-skewed, indicating that there are relatively more individuals with lower credit amounts compared to higher credit amounts. Additionally, the distribution has a relatively large range, suggesting that the credit amounts vary significantly among the individuals in the dataset. Finally, we observe that the third quartile is $220,000, which implies that only a few people could get large amount of credit.



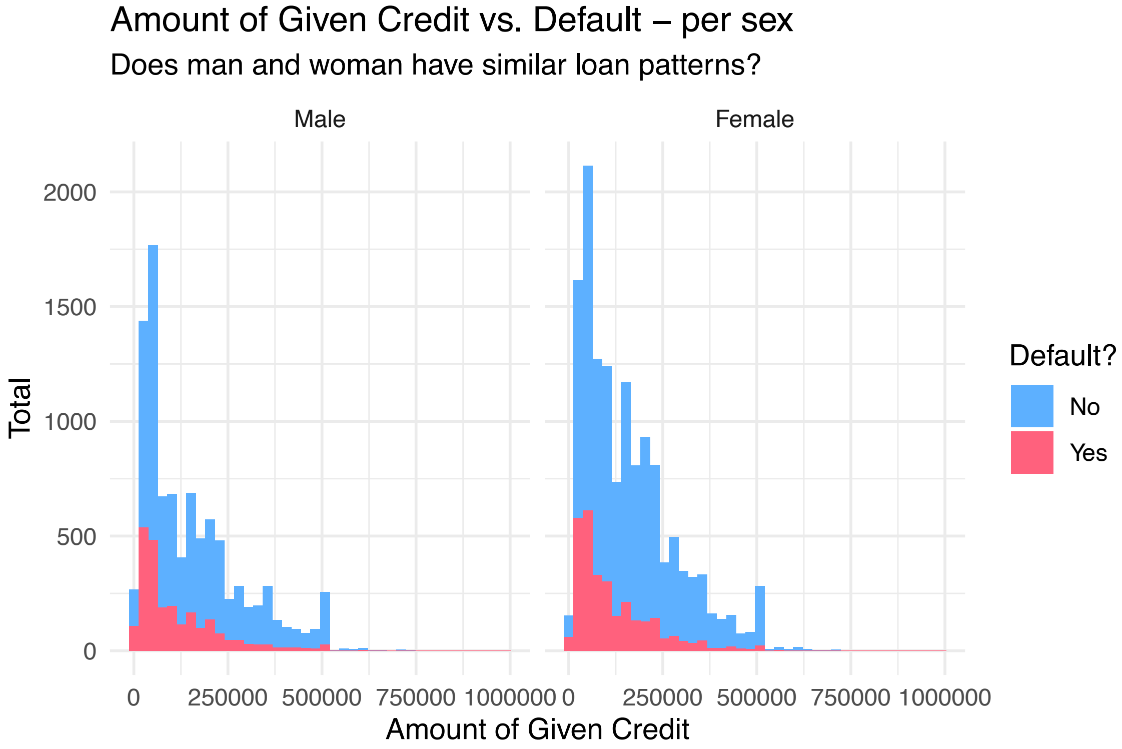
C. Gender

Among the male population of 9456 individuals, the default rate is 25.15%. In contrast, among the female population of 13694 individuals, the default rate is 21.81%.

Table 1: Loan Default Total among Gender

	Paid	Defaulted
Male	7078	2378
Female	10710	2984

From the figure, it can be observed that regardless of gender, the credit limits provided by the bank are concentrated between 100,000 and 200,000. Both male and female customers have their credit limits predominantly within this range. Also, the graph reveals that the default rate for males is slightly higher than that for females.



D. Education

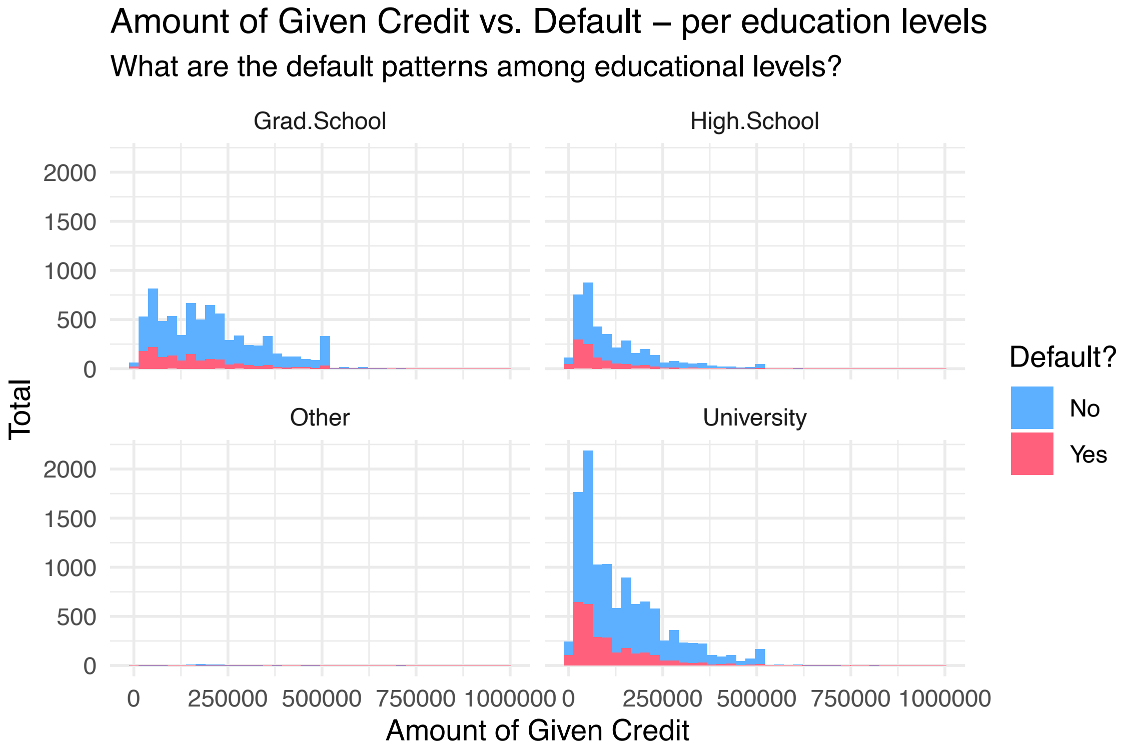
Among individuals with a high school education (3,995 people), the default rate is 25.86%. Among individuals with a university education (11,519 people), the default rate is 24.72%. Among individuals with a graduate school education (7,564 people), the default rate is 19.55%. Only 72 individuals fall under the "other" category, with a default rate of 4.17%.

Table 2: Loan Default Total among Educational Level

	Paid	Defaulted
Grad.School	6085	1479
High.School	2962	1033
Other	69	3
University	8672	2847

From the figure, it is evident that the majority of individuals have a university education, with a moderate default rate. The default rate is lowest for individuals with a graduate school education and highest for those with a high school education. This suggests a negative correlation between education level and default rates, indicating that individuals with higher education tend to have better financial capabilities, resulting in lower default rates. However, due to the small number of individuals in the "other" category and the uncertainty regarding their education level, no further explanation is provided for this group.

The figure reveals that customers with a high school or university education tend to be offered credit limits concentrated between 100,000 and 200,000. Specifically, there is a significant number of customers with a credit limit of 100,000. On the other hand, customers with a graduate school education have a relatively more evenly distributed credit limit range.



E. Marriage Status

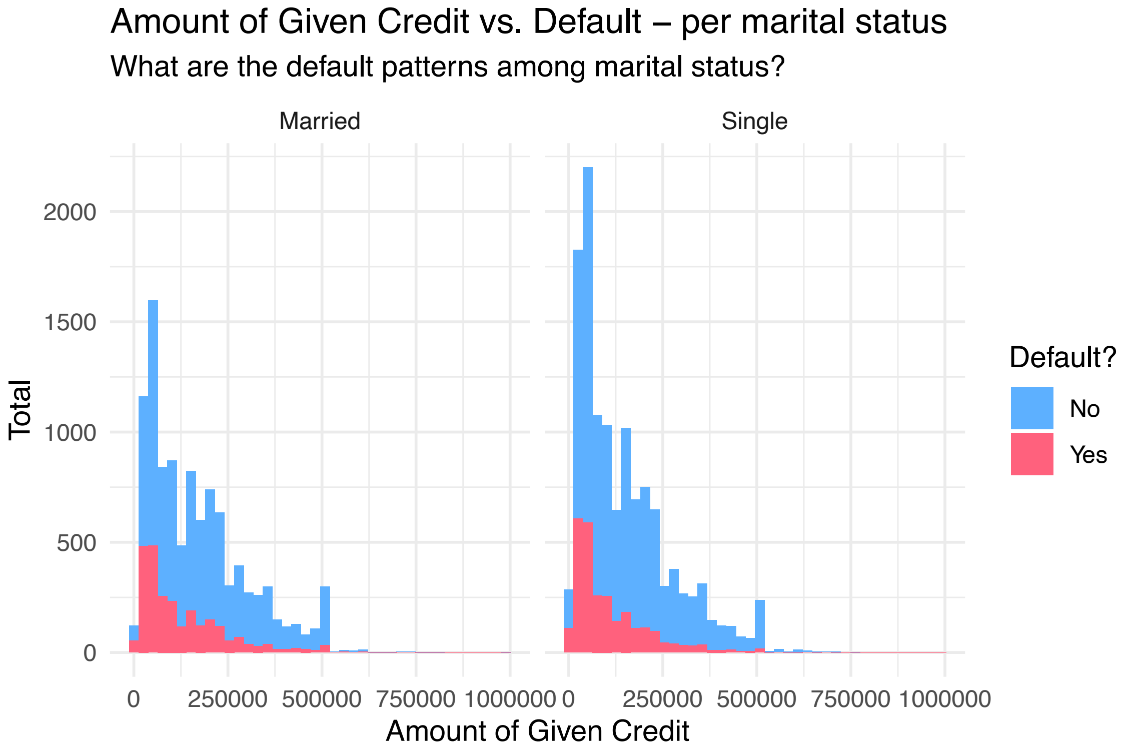
Among the married individuals, there are 10,354 people, with a default rate of 24.76%. Among the unmarried individuals, there are 12,527 people, with a default rate of 21.75%. There are 269 individuals classified under "other," with a default rate of 27.50%.

Table 3: Loan Default among Marital Status (%)

Married	Other	Single
24.76	27.51	21.75

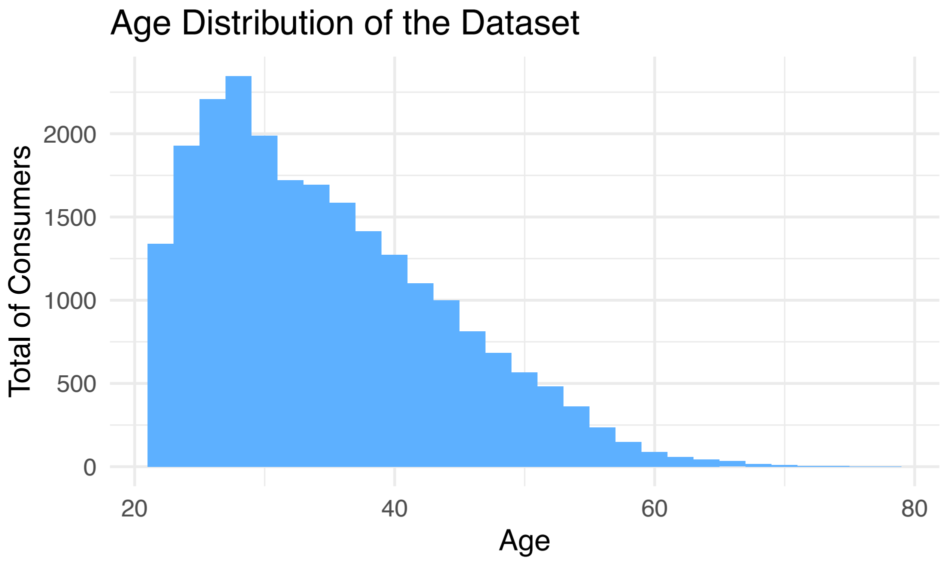
From the figure, it can be observed that the dataset contains a larger number of unmarried individuals, and their default rate is slightly lower compared to married individuals.

The figure indicates that regardless of marital status, the credit limits provided by the bank are concentrated between 100,000 and 200,000. Both married and unmarried individuals have their credit limits predominantly within this range.

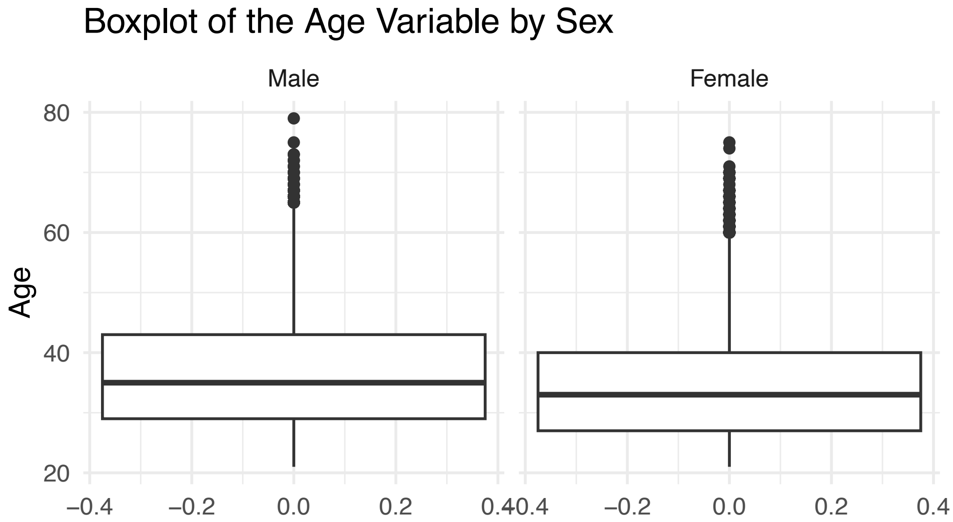


F. Age

From the figure, it is evident that the distribution exhibits a right-skewness. According to summary statistics of this variable, the youngest individual is only 21 years old, while the oldest individual is 79 years old. The average age is around 35 years, indicating that the majority of individuals fall into the youth and middle-aged categories



The boxplot can also helps us visualize the summary statistics of our data. It can be observed that the age distribution of males is slightly higher than that of females in the dataset.

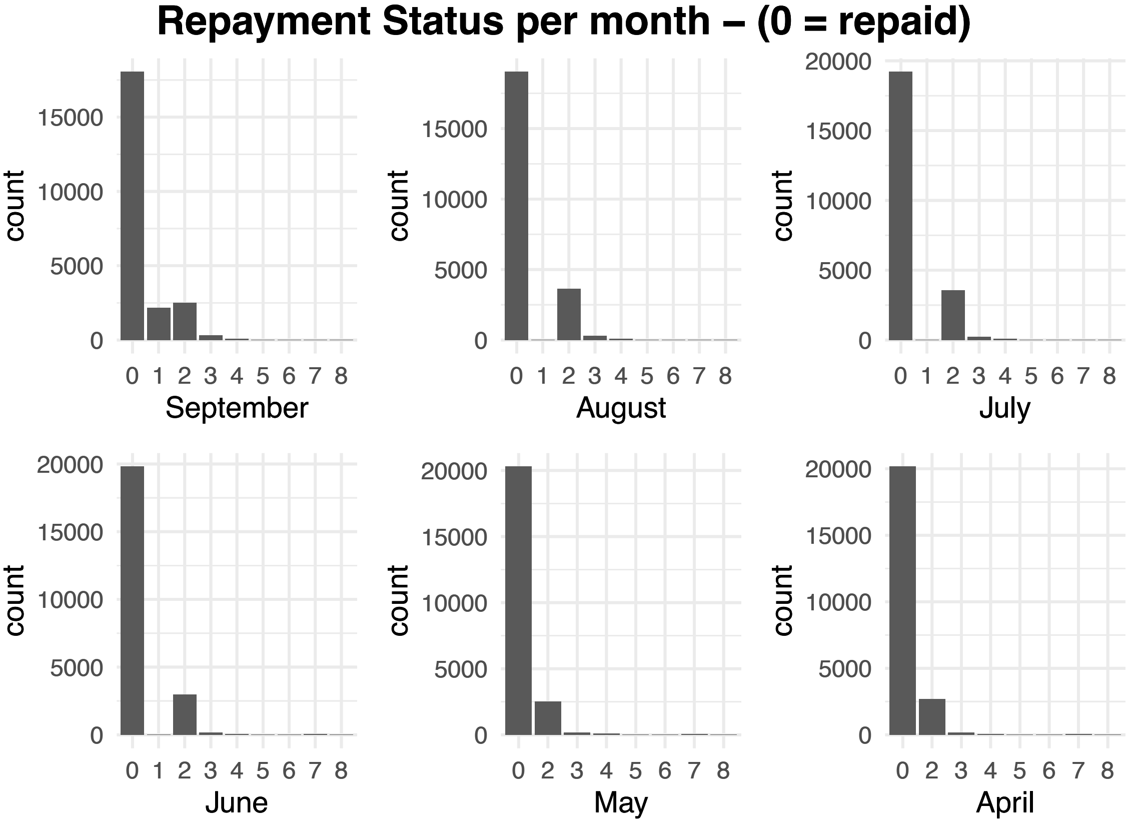


G. History of past-payments

From the figure, it can be observed that the majority of customers do not have any delayed payments. However, there is still a portion of customers who experience delayed payments, and it can be noted that these customers have a relatively higher default rate.

For customers who have no delayed payment for the past six months (PAY_0 ~ PAY_6 == 0), the default rate is 11.22%.

On the other hand, for customers who have delayed payments in every month for the past six months (PAY_0 ~ PAY_6 != 0), the default rate is 70.61%.



Model Building

Regression

Logistics Regression on Default

Our Logistic Regression model reveals some important insights. While Age doesn't seem to be a significant factor in predicting defaults, all the other coefficients play a crucial role. This implies that variables such as income, credit history, and payment behavior are more influential in determining whether an individual is likely to default on their payments.

Moreover, the results indicate that recent data has a more substantial impact on predicting defaults. The variables Pay_0 to Pay_6, representing payment status in different months, demonstrate that information from the recent months carries more weight in determining default probabilities. This suggests that a person's current payment behavior is a better indicator of their likelihood to default than their past payment history, which makes sense since the current financial situation of someone seems to answer more for one's need for a loan in the first place.

We observe a similar trend in the variables Bill_amt and Pay_amt. Categories closer to the current month are more statistically significant in predicting defaults, which shows that the current billing and payment amounts have a more significant influence on the prediction of defaults.

Table 4: Logistic Regression Results - Significant Variables

	coefficients	p_values
(Intercept)	-1.416	0.000
LIMIT_BAL	-0.000	0.000
SEX2	-0.152	0.000
EDUCATION4	-1.443	0.017
MARRIAGE2	-0.199	0.000
PAY_01	0.694	0.000
PAY_02	2.040	0.000
PAY_03	2.024	0.000
PAY_04	1.550	0.000
PAY_05	1.540	0.003
PAY_22	0.237	0.001
PAY_23	0.327	0.038
PAY_25	1.492	0.044
PAY_32	0.296	0.000
PAY_33	0.404	0.039
PAY_42	0.267	0.000
PAY_52	0.281	0.000
PAY_62	0.330	0.000
PAY_63	0.938	0.000
BILL_AMT1	-0.000	0.021
BILL_AMT2	0.000	0.036
PAY_AMT1	-0.000	0.000
PAY_AMT2	-0.000	0.001

Our findings emphasize the importance of considering two main factors when predicting the likelihood of defaults: (i) recent payment behavior and (ii) current financial status. Although this is an exciting finding, for a financial company, it is somewhat risky to have to wait to see if a loan receiver is likely to default on their payment. However, such results can incentivize the company to be more innovative when monitoring risky loans.

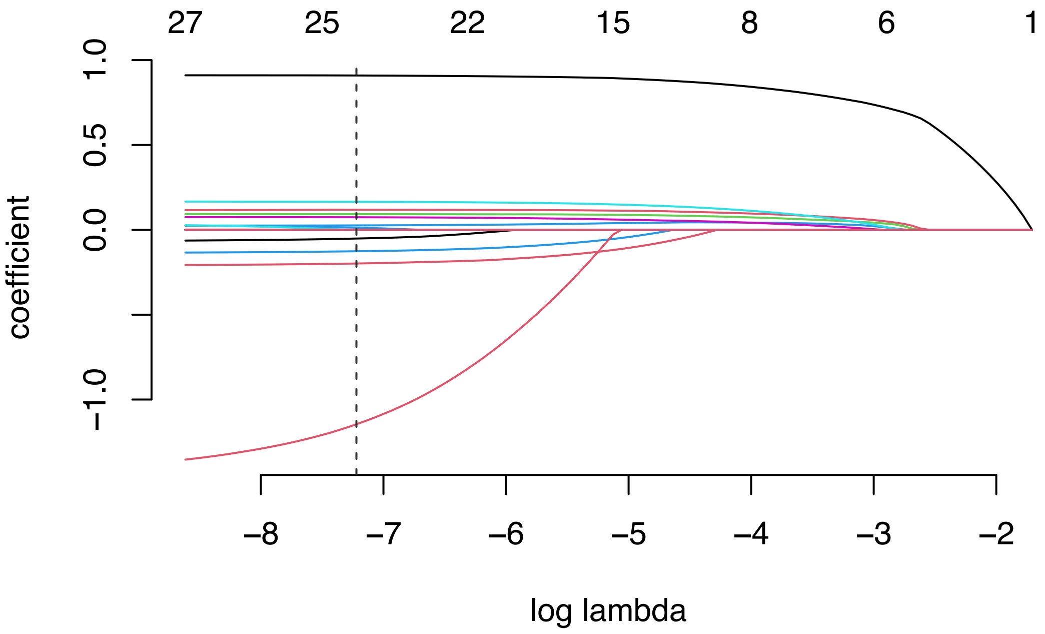
Two suggestions could be:

- Send customers automated alerts that can be generated when significant deviations from standard payment patterns or certain predefined thresholds are crossed.
- Include analyzing non-traditional data points such as utility bill payments, rental payment history, online purchase behavior, or even social media activity to assess borrowers' creditworthiness (any indication of a recent default history).

Lasso

For the Lasso model's penalty parameter (λ), we determined an optimal value of approximately -7 based on AICc, which is a reliable criterion for model selection in generalized linear models. This choice of λ yielded favorable results across various glm models. The Lasso results revealed that categories such as PAY_0-6, with closer proximity to the current month, have a more pronounced impact on predicting defaults. Additionally, we observed the significance of education status in predicting defaults, which aligns with our expectations. Typically, individuals with higher education levels tend to have higher salaries, providing them with greater ability to pay their bills. Hence, it is sensible that education status plays a crucial role in our predictive model.

The results obtained from Lasso, both with and without cross-validation, indicate poor performance. In the Lasso model without cross-validation, we obtained an R^2 value of 0.1826, which suggests limited interpretability of the model for our dataset. Similarly, in the Lasso model with cross-validation, the mean squared error (MSE) value of 0.8876 is quite high considering our response variable takes binary values of 0 or 1. This indicates that the Lasso model is not a suitable choice for accurately predicting defaults in our dataset.

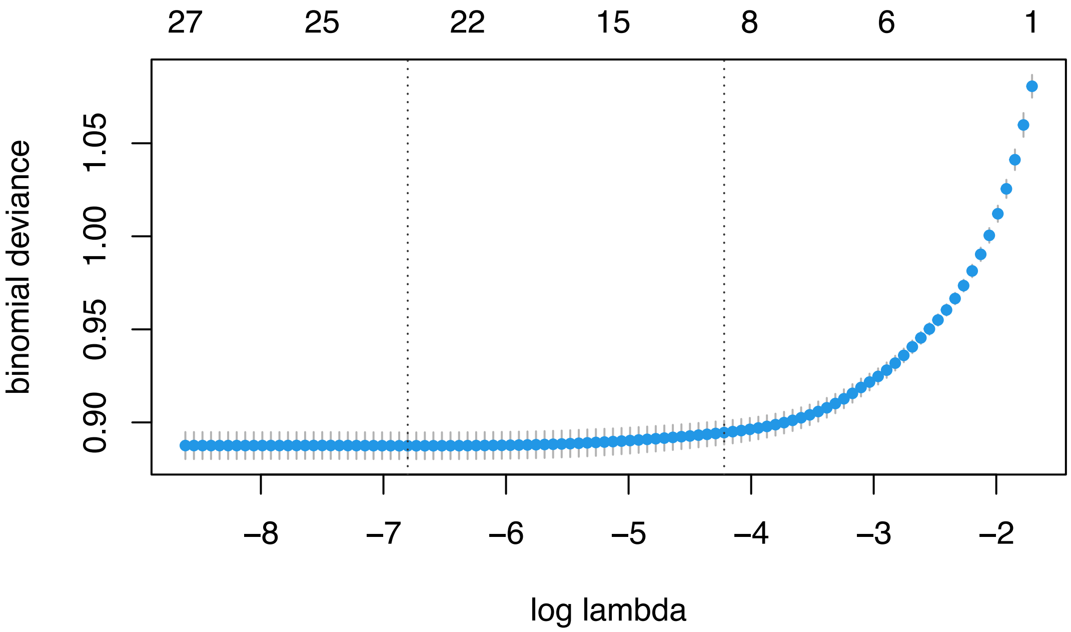


	coef.Lasso_model.1..o..1....
EDUCATION_Other	-1.146

	coef.Lasso_model.1..o..1....
PAY_0	0.9095
MARRIAGE_Single	-0.1984
PAY_6	0.165
SEX	-0.1263
PAY_3	0.1184
PAY_5	0.09245
PAY_4	0.07408
EDUCATION_High.School	-0.05126
PAY_2	0.02618
EDUCATION_Grad.School	0.0129
AGE	0.001024
PAY_AMT1	-6.849e-06
PAY_AMT2	-6.677e-06
LIMIT_BAL	-2.172e-06
PAY_AMT5	-1.93e-06
PAY_AMT4	-1.477e-06
BILL_AMT3	9.636e-07
PAY_AMT3	-5.84e-07
BILL_AMT2	3.894e-07
BILL_AMT1	0
BILL_AMT4	0
BILL_AMT5	0
BILL_AMT6	0
PAY_AMT6	0
EDUCATION	0
MARRIAGE	0
SEX_Male	0
SEX_Female	0
EDUCATION_University	0
MARRIAGE_Married	0
MARRIAGE_Other	0

Lasso 2 - using CV

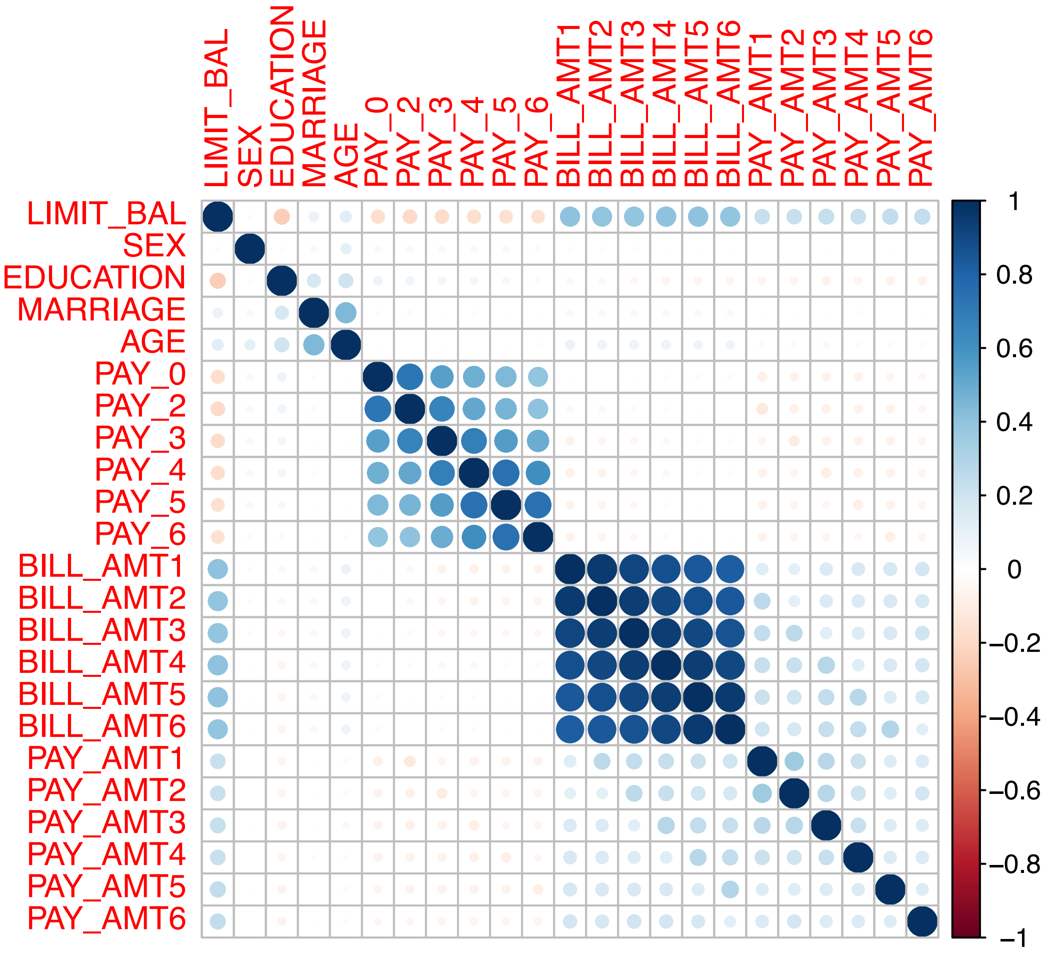
We would choose λ = -8.4066597.


27
25
22
15
8
6
1
binomial deviance
0.90
0.95
1.00
1.05
−8
−7
−6
−5
−4
−3
−2
log lambda

Principle Components Analysis (PCA)

Examining the dataset, we noticed that certain variables capture similar information but at different points in time, so we consider Principal Component Analysis (PCA) as a practical option to reduce the dimensionality of these variables. By applying PCA, we aim to simplify the analysis process and identify the primary factors that influence the response variable, which, in our case, is the default payment behavior.

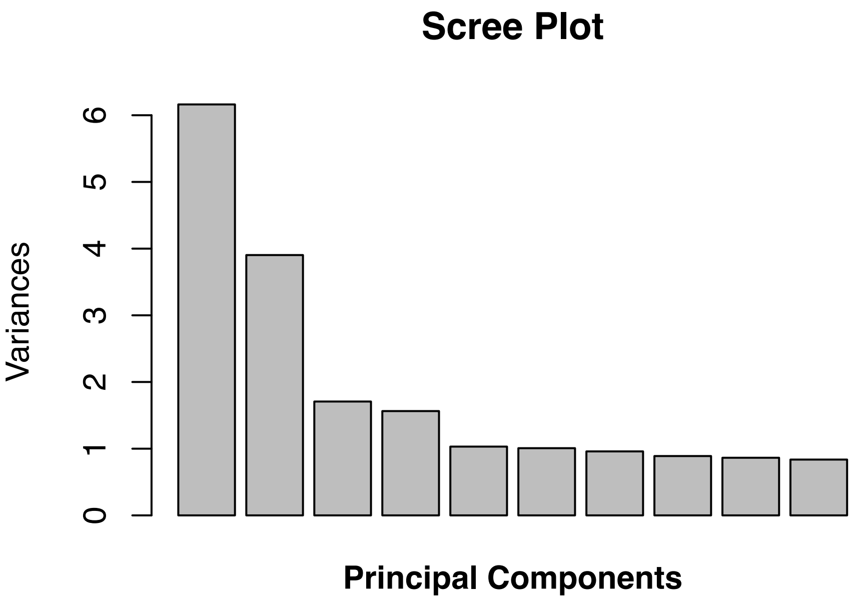
Looking at the correlation plot, we can see some of our suspicions confirmed: the variables that measure the same phenomenon show a certain degree of correlation. Regarding the demographic factors, age, and marriage display a positive correlation. Specifically, when MARRIAGE=1, indicating being married, and MARRIAGE=0, indicating being single, the correlation with age is expected since the likelihood of marriage tends to increase. However, the correlations between other pairs of demographic factors were weak.



From the output below, it seems that the first 4 principal components explains most the variation in the dataset and:

- PC1 seems to capture the people who have large amount of bill statement for the six month;
- PC2 seems to capture the people who have more lagged repayment for the six months;
- PC3 seems to capture the people who have large amount of previous payment;
- PC4 seems to capture people who are older and married which makes sense since this two covariates correlates positively as older people are more likely to be married.

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.4824 1.9757 1.30667 1.25051 1.01559 1.00375 0.97948
## Proportion of Variance 0.2679 0.1697 0.07423 0.06799 0.04484 0.04381 0.04171
## Cumulative Proportion  0.2679 0.4376 0.51187 0.57986 0.62470 0.66851 0.71022
##                            PC8     PC9    PC10    PC11   PC12    PC13    PC14
## Standard deviation     0.94317 0.92961 0.91470 0.87182 0.8111 0.74410 0.71500
## Proportion of Variance 0.03868 0.03757 0.03638 0.03305 0.0286 0.02407 0.02223
## Cumulative Proportion  0.74890 0.78647 0.82285 0.85589 0.8845 0.90857 0.93080
##                           PC15    PC16    PC17    PC18    PC19   PC20    PC21
## Standard deviation     0.67937 0.56715 0.49156 0.48418 0.43224 0.2583 0.18443
## Proportion of Variance 0.02007 0.01399 0.01051 0.01019 0.00812 0.0029 0.00148
## Cumulative Proportion  0.95086 0.96485 0.97535 0.98555 0.99367 0.9966 0.99805
##                           PC22   PC23
## Standard deviation     0.15546 0.1439
## Proportion of Variance 0.00105 0.0009
## Cumulative Proportion  0.99910 1.0000
```



AICC and Lasso selection

Here we decided to use two techniques to select which principal components are most important in explaining the variation in default payment and could be later used in Principal Components Regression.

Glm on First K

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.2316	0.002526	91.7	0
PC1	-0.01734	0.001017	-17.04	1.021e-64
PC2	-0.08518	0.001278	-66.62	0
PC3	0.006911	0.001933	3.575	0.0003508
PC4	-0.006614	0.00202	-3.275	0.001059

(Dispersion parameter for gaussian family taken to be 0.1476846)

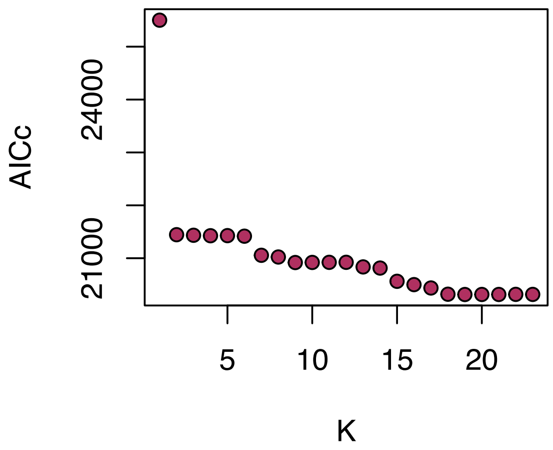
Null deviance:	4120 on 23149 degrees of freedom
Residual deviance:	3418 on 23145 degrees of freedom

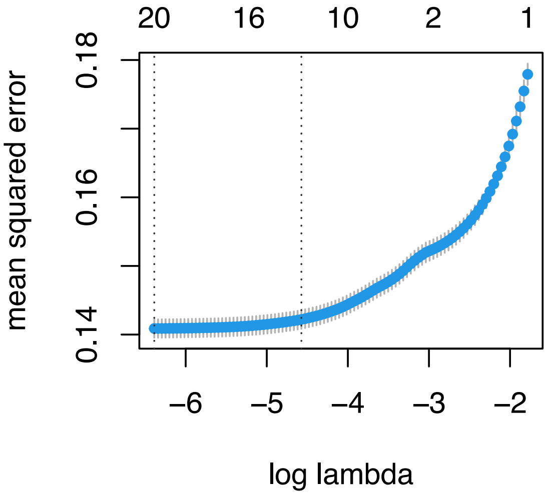
Glm on selected factors (AICc) AICc had its lowest value for the first nineteen factors, a somewhat suspicious result for the amount of variables selected. But we will comeback to this later.

Using 19 Principal components selected by AIC the goodness of fit is about 21% which is not a good enough result. Let's then take a look at LASSO.

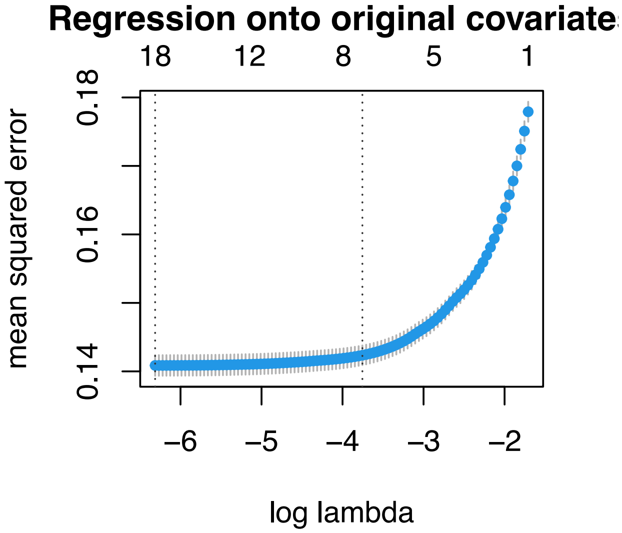
We can see that LASSO elinimates PC3-6, 10-12 and 19 that AIC selects.

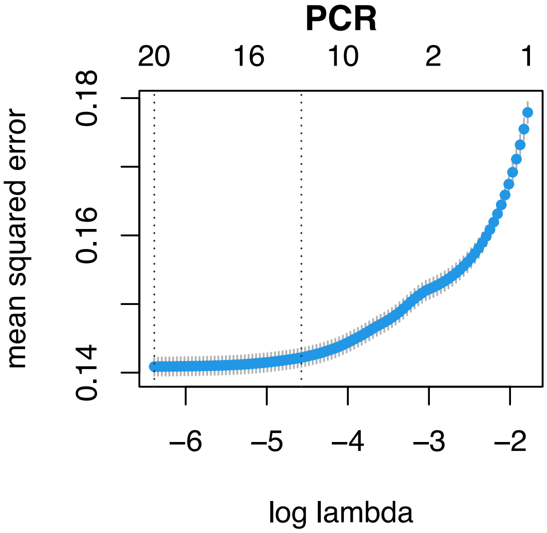
```
## 24 x 1 sparse Matrix of class "dgCMatrix"
##                  seg61
## intercept  0.231619870
## PC1       -0.013179172
## PC2       -0.079949068
## PC3        .
## PC4        .
## PC5        .
## PC6        .
## PC7        0.038412458
## PC8        0.004585184
## PC9       -0.016659692
## PC10       .
## PC11       .
## PC12       .
## PC13       0.017814892
## PC14      -0.002381513
## PC15      -0.042895700
## PC16      -0.016972917
## PC17      -0.019826620
## PC18      -0.034372362
## PC19       .
## PC20       .
## PC21       .
## PC22       .
## PC23       .
```





We then run the regression on the initial covariates and later the PCR model


Regression onto original covariate


PCR

We can see that the minimum MSE of using principal components to do LASSO regression and using original covariates to do LASSO differs very little. It may suggest that doing PCA this way is not very effective. Looking at the AIC plot above, we see that there are several flat stages meaning that the PC's of that stage didn't contribute much in improving the model. We suspect that it might be caused by the redundancy in the covariates. Thus we decided to **collapse the columns that measure the same thing and do PCA again.**

Feature engineering and running PCA again

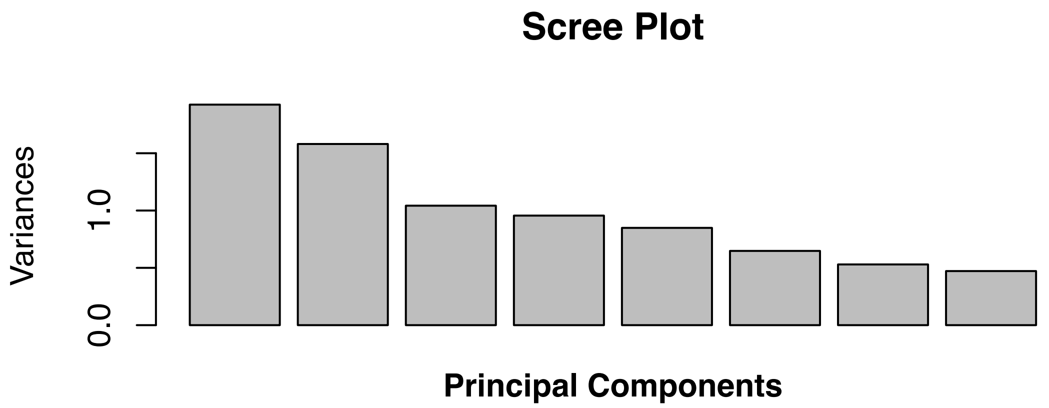
After collapsing the columns that measure similar phenomenons in different moments in time, we run our PCA again and these were our main findings.

1. PC1 seems to represent the young educated adults that have good credit history as they spend more and pay on time;

2. PC2 seems to represent married adults who does not really use credit card;
3. PC3 seems to represent young male who likes to delay payments;
4. PC4 seems to represent young adults that spend more and always delay payments.

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5     PC6     PC7
## Standard deviation     1.3870 1.2571 1.0209 0.9776 0.9212 0.80479 0.72789
## Proportion of Variance 0.2405 0.1975 0.1303 0.1195 0.1061 0.08096 0.06623
## Cumulative Proportion  0.2405 0.4380 0.5683 0.6878 0.7938 0.87480 0.94103
##                            PC8
## Standard deviation     0.68684
## Proportion of Variance 0.05897
## Cumulative Proportion  1.00000
```

```
##              PC1    PC2    PC3    PC4    PC5    PC6
## LIMIT_BAL  0.592 -0.047  0.027 -0.031  0.164 -0.164
## SEX        0.002  0.074 -0.855 -0.442 -0.145 -0.013
## EDUCATION -0.196  0.453  0.115  0.177 -0.767 -0.025
## MARRIAGE   0.122  0.612  0.164 -0.033  0.358  0.086
## AGE        0.163  0.629 -0.071 -0.159  0.154 -0.002
## BILL_AMT   0.495  0.008 -0.167  0.407 -0.218 -0.566
## PAY_AMT    0.510 -0.074 -0.056  0.139 -0.251  0.787
## PAY       -0.253  0.105 -0.438  0.748  0.317  0.156
```



Now we do AIC and LASSO selection again on the collapsed data and we see AIC choose first 6 and LASSO choose first 5 which is pretty similar. And look at the AIC plot we don't observe the flat stage anymore, indicating that the PC's are now more representive.

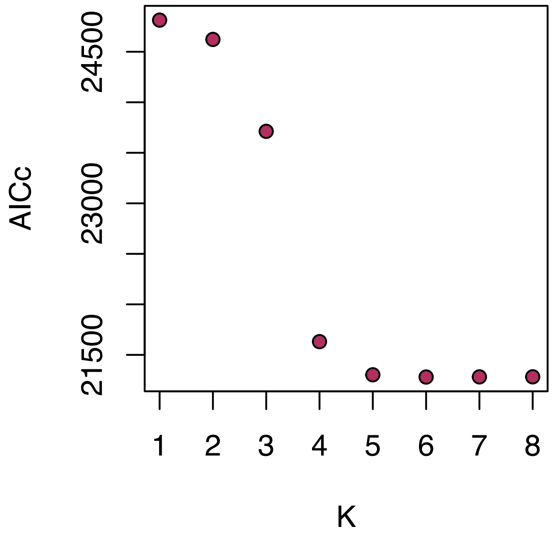
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.2316	0.002518	91.99	0
PC1	-0.06035	0.001815	-33.24	9.499e-237
PC2	0.02997	0.002003	14.96	2.18e-50
PC3	-0.07927	0.002466	-32.14	8.817e-222
PC4	0.1212	0.002575	47.04	0
PC5	0.04988	0.002733	18.25	7.145e-74
PC6	0.01532	0.003129	4.897	9.814e-07

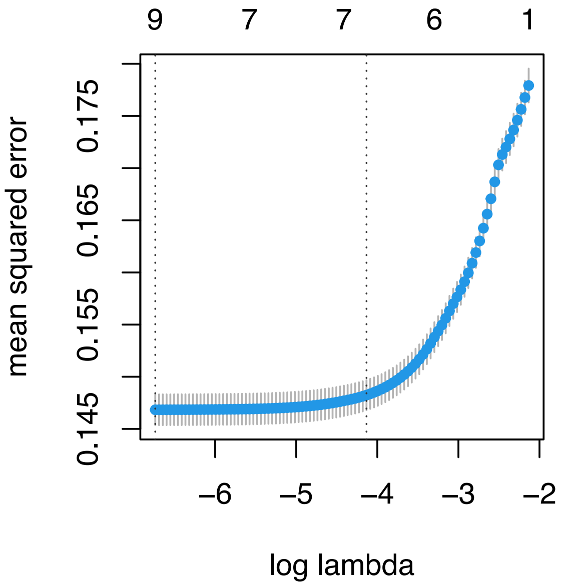
(Dispersion parameter for gaussian family taken to be 0.1467512)

Null deviance:	4120 on 23149 degrees of freedom
Residual deviance:	3396 on 23143 degrees of freedom

AIC and Lasso selection using our Featured engineered

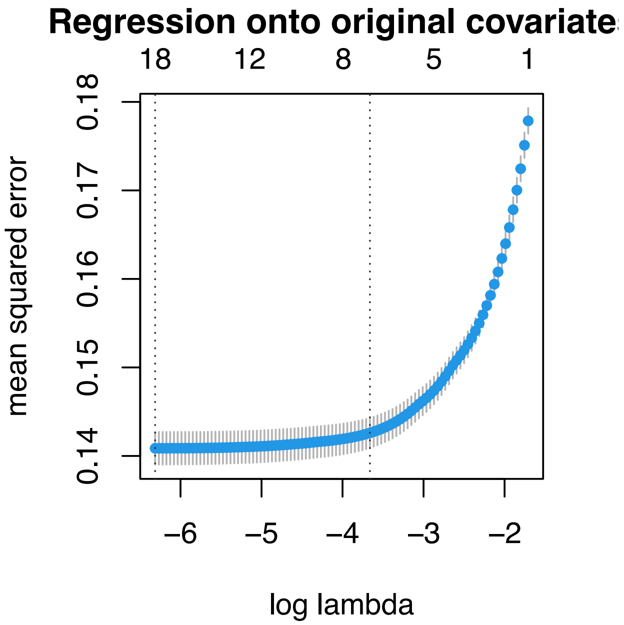
```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                 seg44
## intercept  0.23161987
## PC1       -0.04879070
## PC2        0.01722170
## PC3       -0.06356780
## PC4        0.10476404
## PC5        0.03247732
## PC6        .
## PC7        .
## PC8        .
```

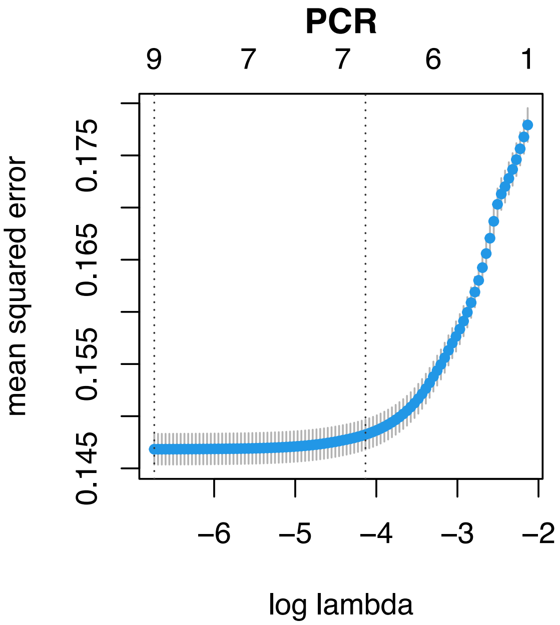




LASSO PCR and simple LASSO

Now we compare the LASSO PCR using collapsed data (featured engineered) and LASSO using original 23 covariates and now the minimum MSE of LASSO PCR improves more than not using collapsed data. This makes sense since by adding the adding up the Repayment status, Amount of bill statement and Amount of previous payment up across month, we are still able to get the idea of the spending and payment behavior of each individual. Therefore, since we're interested in the overall behavior of each individual in explaining and predicting the default payment rather than monthly behavior, it makes sense to combine some covariates as it simplifies analysis and makes PCA more meaningful.





Clustering

Cluster analysis is a valuable technique used in various fields to gain insights from data by identifying groups or clusters of similar objects or observations. For the Loan Default data, we have two main goals in clustering our data:

1. Segmentation and targeting: because clustering is popular in market segmentation and customer profiling, we want to identify the specific customer segment that are more likely to default their loans.
2. Anomaly detection: because clustering is a very good technique to identify outliers or anomalies within the data, this will be especially relevant in our case. For financial firms such ours (loan business) we need to identify the outlier that might have higher probability to get default, causing financial losses.

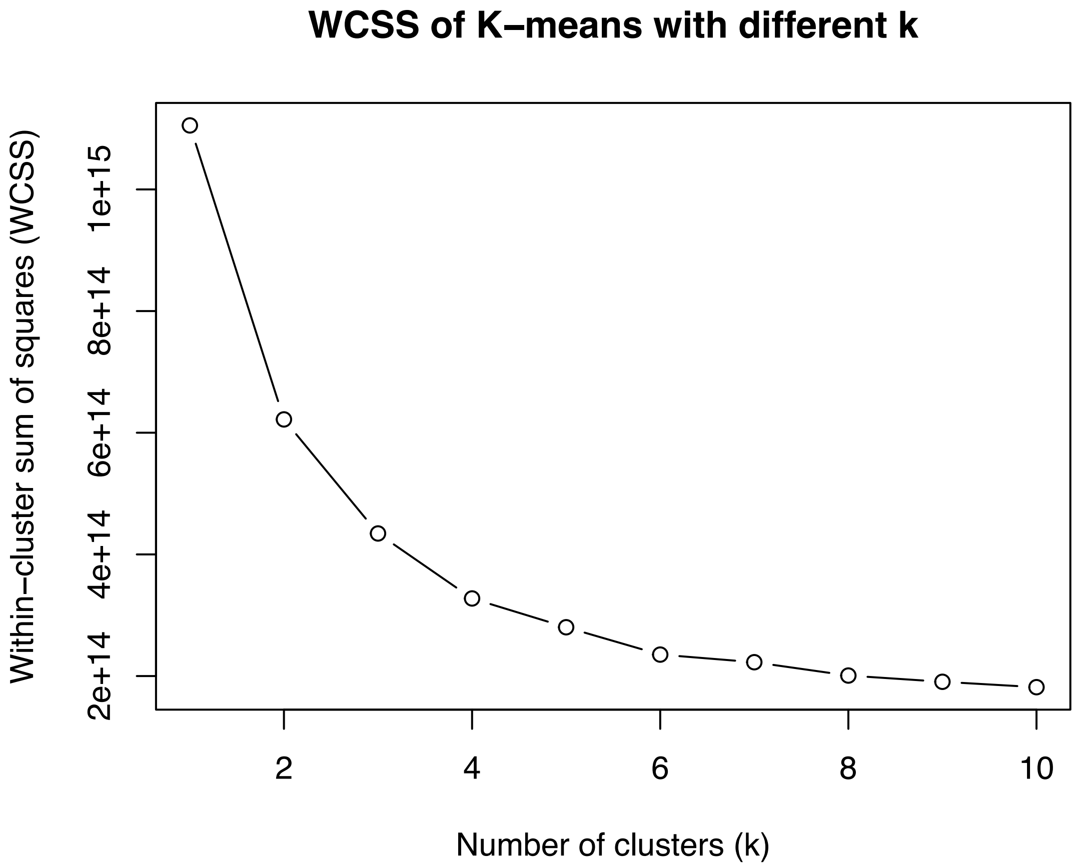
Following, we will perform three different clustering methods (K-means, Hierarchial clustering and Gaussian Mixture Models - GMM). Some methods are model-based, and some are not.

K Means

K-means is a popular clustering algorithm. Here, we aim to partition our dataset into a predetermined number of clusters. In the code bellow, we first choose the number of clusters (k) and randomly select k initial centroids. We then proceed to calculate the distance between each data point and the centroids. Assign each data point to the nearest centroid. Finally, we recalculate the centroids based on the data points assigned to each cluster, and repeat steps two and three until the clusters stabilize (convergence).

Once convergence is reached, the we then check the final cluster assignments and the coordinates of the centroid for each cluster, which will then be used for further analysis and interpretation.

We compare their within-cluster sum of square (WCSS) and choose the appropriate number k. Noted that K-means only works on numerical variables.



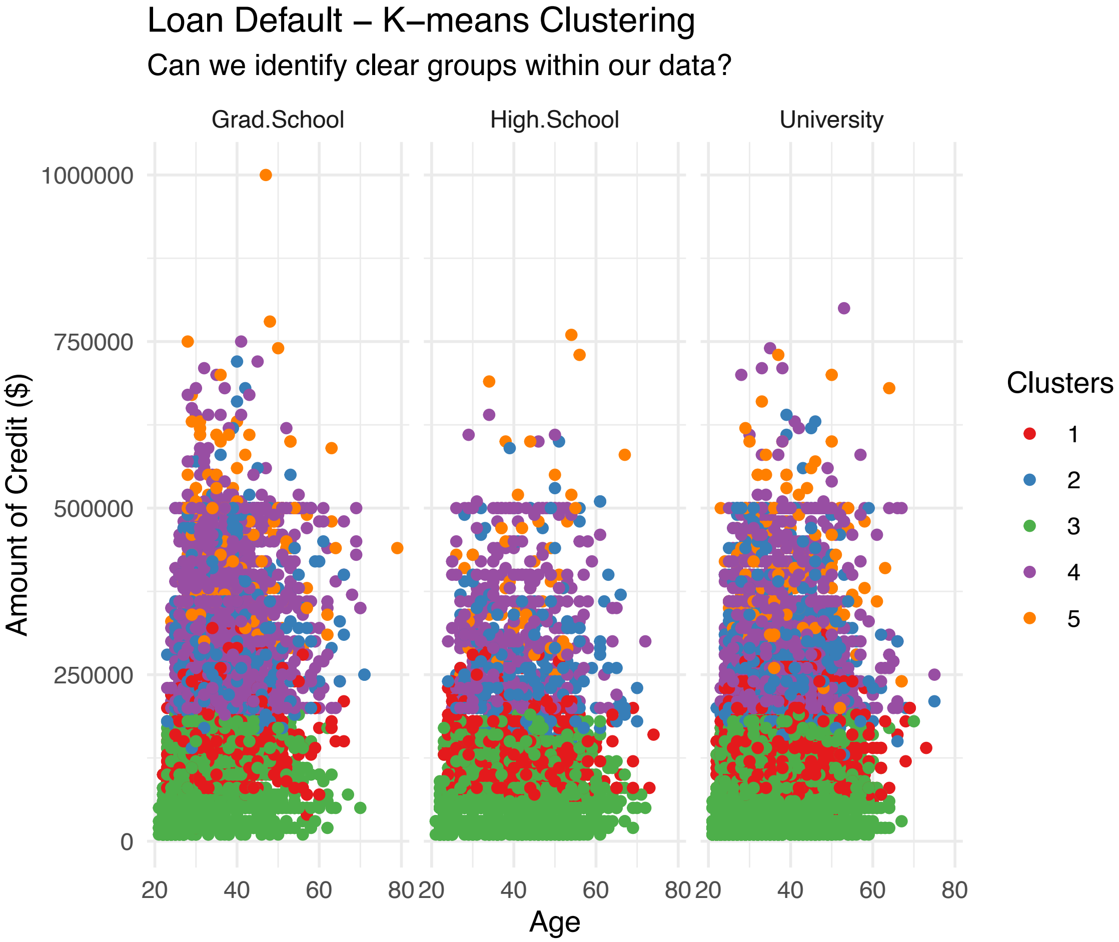
```
# find elbow, k=4 is appropriate
kmeans_result = kmeans(data_kmeans, centers = 5)
kmeans_result$centers
```

```
##   LIMIT_BAL      AGE BILL_AMT1 BILL_AMT2 BILL_AMT3 BILL_AMT4 BILL_AMT5
## 1 142554.41 35.22602  96795.80  94517.04  90067.02  83504.08  78541.07
## 2 275288.99 36.73119 190158.89 185627.41 178770.43 164922.87 152850.48
## 3  66732.18 34.35366  25222.09  24543.12  23313.48  21062.54  19541.61
## 4 313580.49 36.45758  23754.41  22092.16  21941.01  22108.82  21317.07
## 5 413509.33 38.54595 354819.22 350109.68 340912.72 320856.35 299168.37
##   BILL_AMT6  PAY_AMT1  PAY_AMT2  PAY_AMT3  PAY_AMT4  PAY_AMT5  PAY_AMT6
## 1  76110.90  6936.033  6403.116  5986.669  5594.856  5332.253  5173.778
## 2 145766.82 13893.089 14568.792 11346.958 10676.325 10142.867 10630.837
## 3  18839.88  2990.791  2864.303  2550.303  2410.688  2385.467  2364.063
## 4  20883.65  7741.274  8336.017  8413.509  7486.566  7802.061  8839.270
## 5 285641.60 20635.517 21725.942 20286.105 16612.510 17015.166 20258.841
```

When plotting the age as our x-axis and the amount of credit given as our y-axis, and breaking the plot by Education (into three categories), we have an interesting visualization of the job done by our clustering algorithm.

We can see that clusters three (3) and four (4) discovered a somewhat of a niche costumer. Cluster three seems to be formed predominantly by costumers for which a low limit was given, almost similar across all ages. Cluster four (4), however, seems a costumer segment for which a higher amount of credit was given, which seems to be more common for people with grad school level of education.

Another relevant takeaways from our clusters is that, it seems that clusters one (1) and five (5) share a common characteristics across all level of education for a range of amount of credit given. This might indicate that, for these variables, this is a grey area and that we might need other models to help us identify which other variables can help us disentangle this information.



Hierarchical Clustering

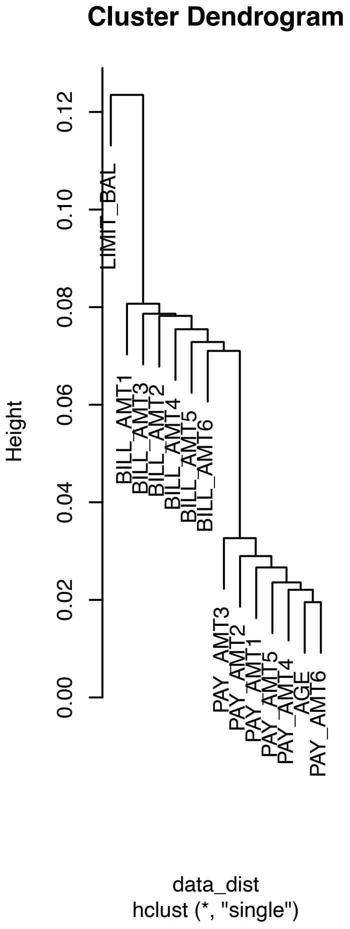
Hierarchical clustering is a clustering algorithm that aims to create a hierarchical structure of clusters based on the similarity or dissimilarity between data points.

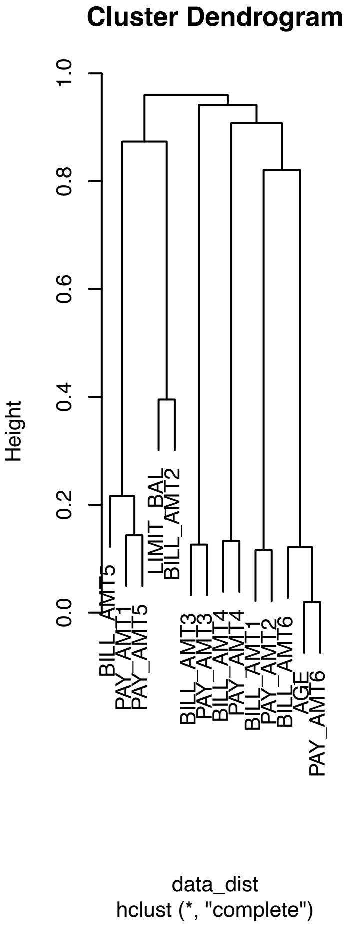
Hierarchical clustering use different linkage criteria to define the similarity or distance between clusters and guide the merging process. In our model, we will run our hierarchical clustering for the three following linkage criteria:

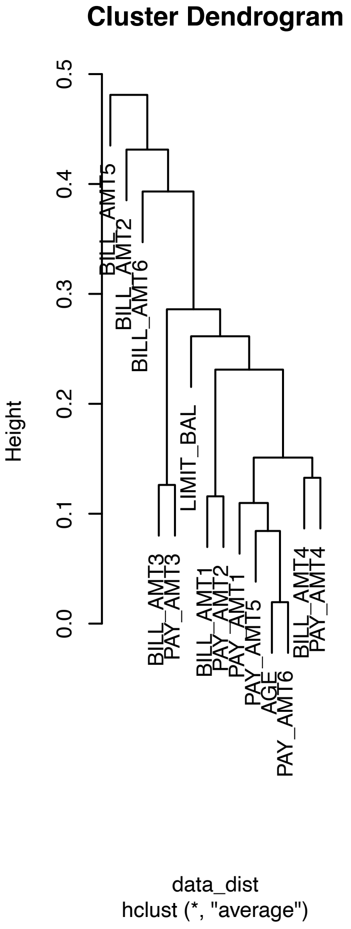
- Single linkage: The similarity between two clusters is defined as the minimum distance between any two data points, one from each cluster.
- Complete linkage: The similarity between two clusters is defined as the maximum distance between any two data points, one from each cluster.
- Average linkage: The similarity between two clusters is defined as the average distance between all pairs of data points, one from each cluster.

For each one of them, our findings can be summarized as follows:

1. Single Linkage: We can see that `PAY` and `BILL` two different categorical variables are clustering together. And, `Age` is the most closed to the most far payment `PAY_AMT6`.
2. Complete Linkage: It prefers to cluster some period of `PAY` and `BILL` together.
3. Average Linkage: The pattern is unclear. But, we still have that `Age` is the most closed to the most far payment `PAY_AMT6`.







By clustering variables, we can identify groups or clusters of variables that are related to each other. Variables within the same cluster often have similar behavior or share common underlying factors. This grouping can help in understanding the structure and organization of the dataset, revealing hidden patterns or relationships.

Next, we want to cluster by observations, but our observations is too large, that we need to reach from other approaches. we choose to randomly sample 500 observations from our dataset, and trying to find the pattern in the sample.

Hierarchical clustering allows for a bottom-up approach, starting with individual data points and interactively merging clusters based on similarity or distance. The resulting dendrogram provides a visual representation of the cluster hierarchy, and by cutting the dendrogram at an appropriate level, you can obtain clusters at different granularities.

This algorithm is flexible and can handle various types of data and distance/similarity measures, making it widely used in exploratory data analysis and visualization.

Gaussian Mixture Models (GMM)

Gaussian Mixture Models (GMM) is a probabilistic clustering algorithm that assumes the underlying data distribution is a mixture of Gaussian distributions. GMM clustering aims to model the data as a collection of Gaussian components, each representing a cluster.

But now we can use package `mclust` to avoid probability calculation. We also found three clusters in this method.

As a summary, here are the three main conclusions from our GMM model.

1. First cluster is a well-behaved group. Most of them pay the bill on time. They have the lowest default rate. We can proactively communicate with them, keeping customers informed about changes, updates, and new offerings. For this group, one major suggestion would be to increase the amount of relevant information shared, such as interest rate changes or new products/services. Being proactive in the our company's communication will likely demonstrate a transparency that keeps customers engaged.

2. Second cluster is more likely to be a group has higher default rate. We can develop informative content that emphasizes financial responsibility and highlights the importance of meeting financial obligations. Share tips on budgeting, managing debt, and improving credit scores. As a company trying to target loan default, providing a trusted advisor and resource for responsible financial practices can be useful for costumers identified within this group.

3. Third cluster includes some extreme outliers that has the highest payment. And, overall they have a little higher credit. We can offer exclusive benefits, discounts, or cashback programs that provide tangible value to your customers. For this group, establishing a formal channel of regular communications of rewards and benefits they can enjoy, is likely to reinforce their loyalty to our brand.

Machine Learning

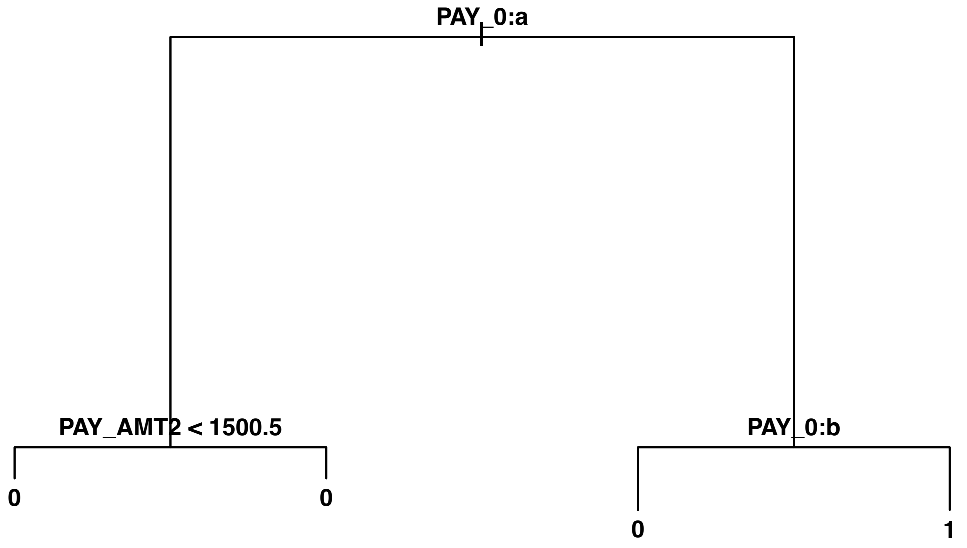
Decision Trees

By using cross-validation, we have determined that utilizing a decision tree with a size of 4 would yield a better result, as evidenced by a lower deviance.

	4	3	2	1
deviance	20674	20903	21444	25062

In the initial split, the decision tree divides the group into two subsets based on whether individuals paid their September's bill duly or not. In the subsequent split, despite considering the amount of money paid for the bill in the past two months, the tree finds that both subsets still belong to the same group. This phenomenon may occur due to the decision tree independently considering multiple features within a group and creating separate splits based on their respective conditions. Moving forward, the third split occurs by examining whether a person's September's bill payment was delayed by one month or more.

The model achieves an accuracy of approximately 82.01%, which is a promising outcome considering the simplicity of the decision tree approach. The efficiency of this model makes it suitable for prediction tasks. However, it is crucial to acknowledge that decision tree models lack robustness. They are sensitive to minor variations in the training data, leading to significant differences in tree structures and predictions. This lack of robustness reduces the reliability of decision trees in scenarios with high data variability or when encountering slight changes in the input data.



Random Forest

For the Random Forest model, we sacrifice the interpretability of a single tree. However, in this algorithm, we essentially aggregate the predictions of multiple bootstrapped trees, which helps to smooth out the individual tree predictions. As a result, we typically achieve better overall results compared to using a single tree.

Using the random forest model on this dataset resulted in an accuracy of approximately 92.65%. Random forests are known for their better robustness and ability to reduce overfitting compared to individual decision trees. However, when evaluating the models on a hold-out set, the decision tree model actually outperformed the random forest model, achieving an accuracy of 82.31% compared to 82.09% for the random forest. We believe this difference in performance may be attributed to the hold-out set exhibiting relatively straightforward patterns that a single decision tree can capture effectively. Considering the significant increase in time required to build the random forest model (approximately 20 times longer), the decision tree model appears to be more suitable for this dataset.

Artificial Neural Network

The neural network (NN) model is one of the machine learning models that operates as a black box, meaning we don't have complete visibility into how it works internally. However, conceptually, the NN model involves performing various operations such as logistic regression, ReLU, and other models. These operations help to project the input data points onto different spaces, enabling the model to make better predictions.

Using the neural network model with 5 hidden layers, we achieved an accuracy of 76.67%. Considering that the model building process took only about 2 seconds, this is a decent result. If we were to add more hidden layers, it could increase the accuracy further. Nowadays, it is common for people to use models with even 50 hidden layers, as deeper architectures can capture more intricate patterns in the data. However, due to our computational limitations, we opted for 5 hidden layers, which still provided a satisfactory accuracy while requiring less time to train. Therefore, we believe the neural network model is a viable and effective choice for our dataset.

```
start.time = Sys.time()

data_matrix = model.matrix(
  ~ LIMIT_BAL + SEX + EDUCATION + MARRIAGE + AGE + PAY_0 + PAY_2 + PAY_3 + PAY_4 + PAY_5 + PAY_6 + BILL_
  data = data2)

train_data_matrix = data_matrix[train_idx, ]
test_data_matrix = data_matrix[-train_idx, ]

n = colnames(train_data_matrix[,-1])
f = as.formula(paste("factor(default) ~", paste(n[!n %in% "default"], collapse = " + ")))

nn_model = neuralnet(f, data = train_data_matrix,
                     hidden = 4,
                     threshold = 0.01,
                     linear.output = T)

# Typically, increasing the number of hidden layers in a nn leads to improved accuracy. However, in thi

nn_model.pred = predict(nn_model, test_data_matrix)
nn_model.pred_res = pred_res_func(nn_model.pred)
nn_model.accuracy = 1 - (length(which(nn_model.pred_res != test_data$default))/nrow(test_data_matrix))

end.time = Sys.time()
nn.time = end.time - start.time
```

Extra Analysis - model efficiency

The efficiency analysis of the different models reveals varying computational times. Given financial companies normally due with huge volumes of data, model efficiency can sooner become a very relevant topic. Thus, we displayed this table below to show a comparison of the time needed - in seconds - for each model to run.

The logistic regression model took approximately 1.20 seconds to run, indicating its relatively fast performance. The Lasso model exhibited a longer computation time of around 27.47 seconds, suggesting a higher complexity due to its regularization component. The tree model without cross-validation required about 7.13 seconds, while the tree model with cross-validation and the random forest model took around 95.63 seconds each, indicating a significantly longer computation time compared to the other models.Lastly, the neural network model also took approximately 95.63 seconds to complete, suggesting a longer computational duration due to its complex architecture.

Table 11: Table continues below

Logistic Regression	Lasso(without cv)	Lasso(with cv)	Tree(without cv)	Tree(with cv)	Random Forest	Neural Network
1.411	0.692	3.683	0.1047	0.9757	2.524	66.8

Conclusion

Our Loan Default Analysis focused on answering four questions:

1. What are the key variables that contribute to predicting loan default?
2. How can financial institutions effectively target customers based on our findings?
3. Given most financial institutions deal with large amounts of data, what is the efficiency of the different models we used?
4. How can financial institutions proactively address loan defaults?

Our primary objective was to identify the variables that have the most significant impact on predicting loan default. By utilizing various modeling techniques such as logistic regression, LASSO regression, and principal components analysis (PCA), we aimed to uncover the key factors influencing the likelihood of loan default. Our study showed that recent data has a more substantial impact on predicting defaults. For instance, the variables representing payment status in other months demonstrate that information from recent months carries more weight in determining default probabilities.

Another key objective was to provide insights into how financial institutions can leverage the findings from our analysis to target customers better. By understanding the patterns and characteristics of customers who are more likely to default on their loans, financial institutions can develop targeted strategies to minimize default rates and optimize loan repayment rates.

Given that financial institutions typically deal with large volumes of data, we needed to assess the efficiency of the different models we employed. By analyzing the computation time of each model, we provided insights into their efficiency in handling the dataset. Efficient models can significantly enhance the speed and scalability of data analysis processes, allowing institutions to use their data more efficiently. Our analysis suggested that the tree models and the Logistic Regression are good options to balance accuracy and efficiency.

Finally, based on our main findings, we aimed to provide recommendations on how financial institutions can proactively address loan defaults. For example, since we discovered that a person's current payment behavior carries more weight in determining default probabilities than in their past payment history. This insight suggests that financial institutions should focus on monitoring and analyzing customers' recent payment behaviors to identify potential default risks and take appropriate actions promptly.

Additionally, the clustering analysis helped identify distinct groups of customers with different default rates, what could help institutions to gather background data and develop targeted strategies to assist customers in achieving financial responsibility.

In conclusion, our analysis aimed to answer these critical questions by applying various modeling techniques to the loan default dataset. By identifying key predictors, assessing model efficiency, and suggesting proactive measures, our findings offer valuable guidance for financial institutions to manage loan default risks and optimize their lending strategies.