

PCA and LDA dimensionality reduction

Abigail Solomon - tsm190000, Subash Chandra - SXC200027, Aditi Chaudhari - APC180001 and Derrick

We will do PCA and LDA model on credits data set found in here

```
library(caret)

## Loading required package: ggplot2
## Loading required package: lattice
library(dplyr)

##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##   filter, lag
## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
library(ggplot2)
df <- read.csv('credit.csv')
```

Let's split the data set into Train and Test

```
set.seed(1234)
i <- sample(1:nrow(df), nrow(df)*0.8, replace=FALSE)
train <- df[i,]
test<-df[-i,]
```

Remove column 25, which is the classifying column, since PCA is unsupervised learning, we need unlabeled or unclassified data set.

```
pca_train <- train[,1:24]
pca_test <- test[,1:24]
```

Check PCA eligibility. Let's see if our variables are linearly correlated? We can check for correlation by creating a table with the `cor()` function. The average correlation value is approximately 0.2, so it shows that there is little correlation.

```
cor(pca_train)
```

```
##          ID  LIMIT_BAL      SEX  EDUCATION  MARRIAGE
## ID      1.000000000  0.02722999  0.020033812  0.038594391 -0.024946286
## LIMIT_BAL 0.027229988  1.000000000  0.019413122 -0.222787435 -0.106639521
## SEX      0.020033812  0.01941312  1.000000000  0.021252215 -0.030919784
## EDUCATION 0.038594391 -0.22278743  0.021252215  1.000000000 -0.145458030
## MARRIAGE -0.024946286 -0.10663952 -0.030919784 -0.145458030  1.000000000
## AGE      0.019001324  0.14643099 -0.088324775  0.176398011 -0.413589540
## PAY_1    -0.027278647 -0.27312113 -0.056848605  0.101516629  0.020024854
## PAY_2    -0.006330543 -0.29825455 -0.068826189  0.119715312  0.020060371
## PAY_3    -0.017184176 -0.28837524 -0.064693187  0.110370988  0.029280591
## PAY_4     0.001776109 -0.26992118 -0.052731539  0.107648189  0.031538067
## PAY_5    -0.017711653 -0.25184190 -0.049615053  0.093847552  0.036756628
## PAY_6    -0.015076729 -0.23743025 -0.037481987  0.079963509  0.034803996
## BILL_AMT1 0.021264651  0.28299928 -0.032032328  0.016107315 -0.026552972
## BILL_AMT2 0.019708142  0.27544980 -0.029661613  0.011309274 -0.024903906
## BILL_AMT3 0.027111681  0.27990007 -0.024205729  0.006984286 -0.028516466
## BILL_AMT4 0.042212370  0.29070446 -0.021060892 -0.005675557 -0.023889105
## BILL_AMT5 0.018641353  0.29206362 -0.015857380 -0.010956135 -0.025765046
## BILL_AMT6 0.018949610  0.28751960 -0.015653884 -0.013005116 -0.023095854
## PAY_AMT1  0.006752802  0.19451796 -0.003420698 -0.039915154 -0.008923307
## PAY_AMT2  0.006588766  0.17725173 -0.004584862 -0.031373728 -0.013004655
## PAY_AMT3  0.036602448  0.21177164 -0.009072133 -0.040215967 -0.003327253
## PAY_AMT4  0.006581535  0.20476583 -0.003084839 -0.038988771 -0.017832318
## PAY_AMT5  0.000118102  0.21724921 -0.005794148 -0.042513476 -0.008374648
## PAY_AMT6 -0.001822799  0.21805935 -0.002716164 -0.044694548 -0.005161726
##          AGE      PAY_1      PAY_2      PAY_3      PAY_4
## ID      0.01900132 -0.02727865 -0.006330543 -0.017184176  0.001776109
## LIMIT_BAL 0.14643099 -0.27312113 -0.298254546 -0.288375239 -0.269921183
## SEX     -0.08832477 -0.05684861 -0.068826189 -0.064693187 -0.052731539
## EDUCATION 0.17639801  0.10151663  0.119715312  0.110370988  0.107648189
## MARRIAGE -0.41358954  0.02002485  0.020060371  0.029280591  0.031538067
## AGE      1.00000000 -0.04234717 -0.049278207 -0.052188767 -0.049706511
## PAY_1    -0.04234717  1.00000000  0.676728158  0.581552138  0.542575677
## PAY_2    -0.04927821  0.67672816  1.000000000  0.770056955  0.664781623
## PAY_3    -0.05218877  0.58155214  0.770056955  1.000000000  0.782677110
## PAY_4    -0.04970651  0.54257568  0.664781623  0.782677110  1.000000000
## PAY_5    -0.05597050  0.50996701  0.622575604  0.689987410  0.820804840
## PAY_6    -0.05021041  0.47592087  0.575956997  0.636278586  0.716707823
## BILL_AMT1 0.05692233  0.18786609  0.236231649  0.210441866  0.202373561
## BILL_AMT2 0.05613573  0.19057613  0.236996322  0.239148906  0.224826422
## BILL_AMT3 0.05512568  0.17900463  0.224273596  0.228242161  0.243533181
## BILL_AMT4 0.04977059  0.18043868  0.223831034  0.229283614  0.246106846
## BILL_AMT5 0.04913813  0.18260102  0.223300062  0.227857376  0.243728613
## BILL_AMT6 0.04653135  0.17867922  0.220648123  0.224843066  0.240723923
## PAY_AMT1  0.03142215 -0.07750149 -0.079501608  0.001153613 -0.011071348
```

## PAY_AMT2	0.02557485	-0.07216395	-0.061868743	-0.068089287	-0.004645503
## PAY_AMT3	0.02709256	-0.06875714	-0.053752500	-0.051829679	-0.067756472
## PAY_AMT4	0.02551927	-0.06317477	-0.045542965	-0.044713065	-0.042522606
## PAY_AMT5	0.02670126	-0.05902910	-0.040199045	-0.037082959	-0.033186156
## PAY_AMT6	0.02125624	-0.06170637	-0.040308711	-0.037504607	-0.028545198
##	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3
## ID	-0.017711653	-0.015076729	0.02126465	0.01970814	0.027111681
## LIMIT_BAL	-0.251841900	-0.237430255	0.28299928	0.27544980	0.279900073
## SEX	-0.049615053	-0.037481987	-0.03203233	-0.02966161	-0.024205729
## EDUCATION	0.093847552	0.079963509	0.01610732	0.01130927	0.006984286
## MARRIAGE	0.036756628	0.034803996	-0.02655297	-0.02490391	-0.028516466
## AGE	-0.055970498	-0.050210415	0.05692233	0.05613573	0.055125683
## PAY_1	0.509967009	0.475920866	0.18786609	0.19057613	0.179004629
## PAY_2	0.622575604	0.575956997	0.23623165	0.23699632	0.224273596
## PAY_3	0.689987410	0.636278586	0.21044187	0.23914891	0.228242161
## PAY_4	0.820804840	0.716707823	0.20237356	0.22482642	0.243533181
## PAY_5	1.000000000	0.817125285	0.20749031	0.22728122	0.242748038
## PAY_6	0.817125285	1.000000000	0.20754735	0.22693197	0.240329478
## BILL_AMT1	0.207490306	0.207547345	1.000000000	0.95114455	0.889562606
## BILL_AMT2	0.227281218	0.226931967	0.95114455	1.000000000	0.924487451
## BILL_AMT3	0.242748038	0.240329478	0.88956261	0.92448745	1.000000000
## BILL_AMT4	0.273012237	0.267097246	0.85947992	0.89232691	0.922040453
## BILL_AMT5	0.271348325	0.291414305	0.82929774	0.86014088	0.883261312
## BILL_AMT6	0.263924876	0.285388860	0.80242457	0.83092902	0.853530392
## PAY_AMT1	-0.006737276	-0.001466827	0.13979311	0.27565834	0.237641275
## PAY_AMT2	-0.006727274	-0.007349921	0.09802975	0.08894766	0.324139251
## PAY_AMT3	0.008562841	0.006494514	0.15872741	0.15328393	0.133765723
## PAY_AMT4	-0.056241640	0.019713719	0.15485673	0.14490430	0.141681125
## PAY_AMT5	-0.034937825	-0.047912350	0.16448116	0.15270193	0.182859614
## PAY_AMT6	-0.022384532	-0.024277234	0.16476661	0.16130612	0.175164671
##	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2
## ID	0.042212370	0.01864135	0.01894961	0.006752802	0.006588766
## LIMIT_BAL	0.290704457	0.29206362	0.28751960	0.194517956	0.177251734
## SEX	-0.021060892	-0.01585738	-0.01565388	-0.003420698	-0.004584862
## EDUCATION	-0.005675557	-0.01095614	-0.01300512	-0.039915154	-0.031373728
## MARRIAGE	-0.023889105	-0.02576505	-0.02309585	-0.008923307	-0.013004655
## AGE	0.049770587	0.04913813	0.04653135	0.031422145	0.025574855
## PAY_1	0.180438683	0.18260102	0.17867922	-0.077501486	-0.072163946
## PAY_2	0.223831034	0.22330006	0.22064812	-0.079501608	-0.061868743
## PAY_3	0.229283614	0.22785738	0.22484307	0.001153613	-0.068089287
## PAY_4	0.246106846	0.24372861	0.24072392	-0.011071348	-0.004645503
## PAY_5	0.273012237	0.27134832	0.26392488	-0.006737276	-0.006727274
## PAY_6	0.267097246	0.29141430	0.28538886	-0.001466827	-0.007349921
## BILL_AMT1	0.859479918	0.82929774	0.80242457	0.139793109	0.098029749
## BILL_AMT2	0.892326915	0.86014088	0.83092902	0.275658339	0.088947657
## BILL_AMT3	0.922040453	0.88326131	0.85353039	0.237641275	0.324139251
## BILL_AMT4	1.000000000	0.94089881	0.90230117	0.230522687	0.206010816
## BILL_AMT5	0.940898808	1.000000000	0.94725446	0.215410623	0.179836252
## BILL_AMT6	0.902301174	0.94725446	1.000000000	0.199747245	0.176778661
## PAY_AMT1	0.230522687	0.21541062	0.19974724	1.000000000	0.264875571
## PAY_AMT2	0.206010816	0.17983625	0.17677866	0.264875571	1.000000000
## PAY_AMT3	0.298065418	0.24883727	0.23454305	0.259710307	0.258651947
## PAY_AMT4	0.131047724	0.28871014	0.24967904	0.208895009	0.192711371
## PAY_AMT5	0.160575515	0.14062534	0.30423755	0.154505402	0.202384416

```
## PAY_AMT6    0.173438013  0.15974236  0.11010206  0.185869971  0.169262638
##              PAY_AMT3    PAY_AMT4    PAY_AMT5    PAY_AMT6
## ID          0.036602448  0.006581535  0.000118102 -0.001822799
## LIMIT_BAL   0.211771641  0.204765826  0.217249211  0.218059348
## SEX         -0.009072133 -0.003084839 -0.005794148 -0.002716164
## EDUCATION   -0.040215967 -0.038988771 -0.042513476 -0.044694548
## MARRIAGE     -0.003327253 -0.017832318 -0.008374648 -0.005161726
## AGE          0.027092558  0.025519268  0.026701260  0.021256237
## PAY_1        -0.068757143 -0.063174771 -0.059029099 -0.061706369
## PAY_2        -0.053752500 -0.045542965 -0.040199045 -0.040308711
## PAY_3        -0.051829679 -0.044713065 -0.037082959 -0.037504607
## PAY_4        -0.067756472 -0.042522606 -0.033186156 -0.028545198
## PAY_5         0.008562841 -0.056241640 -0.034937825 -0.022384532
## PAY_6         0.006494514  0.019713719 -0.047912350 -0.024277234
## BILL_AMT1    0.158727411  0.154856730  0.164481162  0.164766605
## BILL_AMT2    0.153283931  0.144904304  0.152701933  0.161306119
## BILL_AMT3    0.133765723  0.141681125  0.182859614  0.175164671
## BILL_AMT4    0.298065418  0.131047724  0.160575515  0.173438013
## BILL_AMT5    0.248837272  0.288710138  0.140625337  0.159742364
## BILL_AMT6    0.234543053  0.249679044  0.304237551  0.110102062
## PAY_AMT1     0.259710307  0.208895009  0.154505402  0.185869971
## PAY_AMT2     0.258651947  0.192711371  0.202384416  0.169262638
## PAY_AMT3     1.000000000  0.234133517  0.166624206  0.165324233
## PAY_AMT4     0.234133517  1.000000000  0.154208364  0.157565455
## PAY_AMT5     0.166624206  0.154208364  1.000000000  0.156791454
## PAY_AMT6     0.165324233  0.157565455  0.156791454  1.000000000
```

```
mean(cor(pca_train))
```

```
## [1] 0.1828462
```

run PCA on the train data, out of the 25, only 16 components are needed to capture 95% of the variance

```
pca_out <- preProcess(pca_train, method=c("center", "scale", "pca"))
pca_out
```

```
## Created from 24000 samples and 24 variables
##
## Pre-processing:
##   - centered (24)
##   - ignored (0)
##   - principal component signal extraction (24)
##   - scaled (24)
##
## PCA needed 16 components to capture 95 percent of the variance
```

Reduced Data PCA

```
train_pc <- predict(pca_out, pca_train)
test_pc <- predict(pca_out, pca_test)
```

PCA reduced data in Classification, lets see if the reduced data can predict class

```
train_df <- train[,c(25)]
test_df <- test[,c(25)]

train_pc$dpnm <- train$dpnm
test_pc$dpnm <- test$dpnm

train_dfNew <- train_pc
test_dfNew <- test_pc
```

Check for missing values, We see that there are no NAs

```
sapply(df, function(x) sum(is.na(x)==TRUE))
```

```
##      ID LIMIT_BAL      SEX EDUCATION  MARRIAGE      AGE      PAY_1      PAY_2
##      0         0         0         0         0         0         0         0
##      PAY_3      PAY_4      PAY_5      PAY_6 BILL_AMT1 BILL_AMT2 BILL_AMT3 BILL_AMT4
##      0         0         0         0         0         0         0         0
## BILL_AMT5 BILL_AMT6 PAY_AMT1 PAY_AMT2 PAY_AMT3 PAY_AMT4 PAY_AMT5 PAY_AMT6
##      0         0         0         0         0         0         0         0
##      dpnm
##      0
```

Build a logistic regression model, almost half of the variables have a good P value.

```
glm1 <- glm(dpnm~., data=train_dfNew, family="binomial")
summary(glm1)
```

```
##
## Call:
## glm(formula = dpnm ~ ., family = "binomial", data = train_dfNew)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.1254  -0.6994  -0.5564  -0.2951   3.1630
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.450003   0.018325 -79.127 < 2e-16 ***
## PC1          0.077591   0.007755  10.006 < 2e-16 ***
## PC2         -0.396896   0.010667 -37.209 < 2e-16 ***
## PC3         -0.170534   0.023690  -7.199 6.08e-13 ***
## PC4          0.012293   0.019384   0.634  0.52597
## PC5         -0.133696   0.017142  -7.799 6.22e-15 ***
## PC6          0.002003   0.017086   0.117  0.90670
## PC7          0.119608   0.021279   5.621 1.90e-08 ***
## PC8         -0.040246   0.025155  -1.600  0.10961
## PC9         -0.076276   0.026473  -2.881  0.00396 **
```

```
## PC10      -0.066995   0.031440  -2.131   0.03310 *
## PC11      -0.071008   0.032449  -2.188   0.02865 *
## PC12       0.038320   0.039106   0.980   0.32713
## PC13       0.384037   0.021320  18.013 < 2e-16 ***
## PC14       0.049668   0.022258   2.231   0.02565 *
## PC15       0.035278   0.024481   1.441   0.14957
## PC16      -0.345859   0.025144 -13.755 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 25433  on 23999  degrees of freedom
## Residual deviance: 22419  on 23983  degrees of freedom
## AIC: 22453
##
## Number of Fisher Scoring iterations: 5
```

Evaluate on the test set. We got an accuracy of 0.81, TN=4584 and TP=297 . The accuracy is the same, it is a confirmation that PCA is capturing something important in the data.

```
probs <- predict(glm1, newdata=test_dfNew, type="response")
pred <- ifelse(probs>0.5, 1, 0)
acc <- mean(pred==test_dfNew$dpnm)
print(paste("accuracy = ", acc))
```

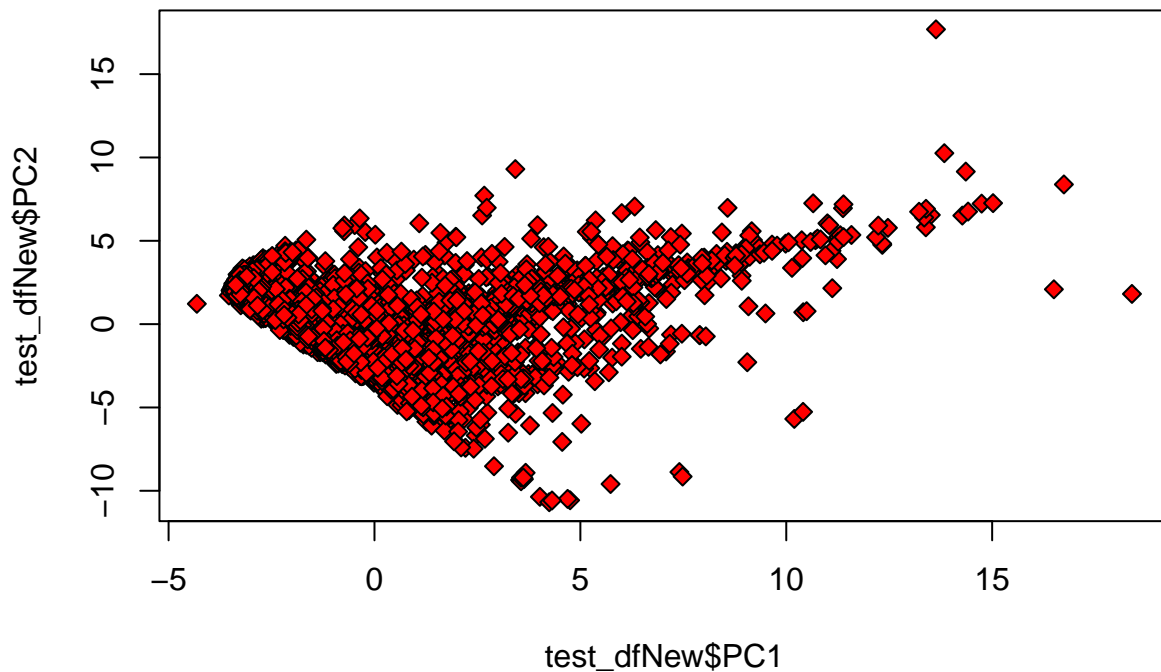
```
## [1] "accuracy = 0.8135"
```

```
table(pred, test_dfNew$dpnm)
```

```
##
## pred    0    1
##      0 4584 1003
##      1  116  297
```

PCA PLOT

```
plot(test_dfNew$PC1, test_dfNew$PC2, pch=c(23,21,22)[unclass(test_dfNew$dpnm)], bg=c("red","green","blue"))
```



Build a logistic regression model on the original data, more variables have got good P values compared to the PCA variables.

```
glm1 <- glm(dpnm~., data=train, family="binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(glm1)
```

```
##
```

```
## Call:
```

```
## glm(formula = dpnm ~ ., family = "binomial", data = train)
```

```
##
```

```
## Deviance Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -3.1457  -0.7012  -0.5476  -0.2903   3.2510
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.457e-01  1.351e-01  -4.778 1.77e-06 ***
## ID          -2.257e-06  1.955e-06  -1.154 0.248422
## LIMIT_BAL   -7.583e-07  1.755e-07  -4.320 1.56e-05 ***
## SEX         -9.614e-02  3.431e-02  -2.802 0.005079 **
## EDUCATION   -9.785e-02  2.341e-02  -4.180 2.92e-05 ***
## MARRIAGE    -1.706e-01  3.536e-02  -4.826 1.39e-06 ***
```

```
## AGE          7.352e-03  1.991e-03   3.693 0.000222 ***
## PAY_1        5.707e-01  1.977e-02 28.870 < 2e-16 ***
## PAY_2        8.252e-02  2.268e-02   3.639 0.000274 ***
## PAY_3        6.331e-02  2.552e-02   2.481 0.013118 *
## PAY_4        2.417e-02  2.806e-02   0.861 0.389146
## PAY_5        3.545e-02  2.995e-02   1.184 0.236466
## PAY_6        2.365e-02  2.461e-02   0.961 0.336589
## BILL_AMT1    -6.774e-06  1.300e-06  -5.213 1.86e-07 ***
## BILL_AMT2     3.505e-06  1.668e-06   2.101 0.035623 *
## BILL_AMT3     1.836e-06  1.471e-06   1.248 0.211998
## BILL_AMT4    -3.100e-07  1.502e-06  -0.206 0.836538
## BILL_AMT5    -1.611e-07  1.700e-06  -0.095 0.924497
## BILL_AMT6     9.640e-07  1.329e-06   0.726 0.468132
## PAY_AMT1     -1.640e-05  2.728e-06  -6.010 1.86e-09 ***
## PAY_AMT2     -8.025e-06  2.209e-06  -3.633 0.000280 ***
## PAY_AMT3     -2.969e-06  1.984e-06  -1.496 0.134640
## PAY_AMT4     -4.528e-06  2.033e-06  -2.228 0.025913 *
## PAY_AMT5     -1.623e-06  1.838e-06  -0.883 0.377329
## PAY_AMT6     -1.478e-06  1.447e-06  -1.022 0.307007
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 25433  on 23999  degrees of freedom
## Residual deviance: 22348  on 23975  degrees of freedom
## AIC: 22398
##
## Number of Fisher Scoring iterations: 6
```

Evaluate on the test set of the original data. We got an accuracy of 0.81, TN=4585 and TP=282

```
probs <- predict(glm1, newdata=test, type="response")
pred <- ifelse(probs>0.5, 1, 0)
acc <- mean(pred==test$dpnm)
print(paste("accuracy = ", acc))
```

```
## [1] "accuracy = 0.811166666666667"
```

```
table(pred, test$dpnm)
```

```
##
## pred    0    1
##      0 4585 1018
##      1  115  282
```


LDA considers the class `dpnm`, it tries to find a linear combination of predictors that maximizes the separation of the classes, while minimizing within-class standard deviation.

```
library(MASS)

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##      select

lda1 <- lda(dpnm~., data=train)
```

LDA analysis has identified means for all variables by class

```
lda1$means

##          ID LIMIT_BAL      SEX EDUCATION MARRIAGE      AGE      PAY_1      PAY_2
## 0 15062.68 177889.6 1.615463 1.840495 1.560169 35.38561 -0.2086905 -0.3039006
## 1 14734.91 129815.9 1.571589 1.894865 1.526799 35.68722 0.6752249 0.4636432
##          PAY_3      PAY_4      PAY_5      PAY_6 BILL_AMT1 BILL_AMT2 BILL_AMT3
## 0 -0.3181526 -0.3572653 -0.3908594 -0.4080583 51902.78 49602.30 47420.19
## 1 0.3697526 0.2655547 0.1772864 0.1240630 48612.40 47516.98 45471.96
##          BILL_AMT4 BILL_AMT5 BILL_AMT6 PAY_AMT1 PAY_AMT2 PAY_AMT3 PAY_AMT4 PAY_AMT5
## 0 43525.39 40461.48 38921.08 6274.820 6629.747 5740.429 5268.031 5196.229
## 1 42290.20 39795.40 38708.72 3351.269 3468.872 3337.650 3110.524 3379.658
##          PAY_AMT6
## 0 5643.074
## 1 3445.893
```

Use LDA model for prediction, again we get an accuracy of approximately 0.81. Even though data is reduced, the model still keeps the accuracy of the original data

[illegible]

[illegible]

[illegible]


```
## .. ..$ : chr [1:6000] "5" "8" "9" "11" ...
## .. ..$ : chr [1:2] "0" "1"
## $ x      : num [1:6000, 1] -0.1066 -0.214 0.2758 0.0955 -0.7717 ...
## ..- attr(*, "dimnames")=List of 2
## .. ..$ : chr [1:6000] "5" "8" "9" "11" ...
## .. ..$ : chr "LD1"
```

Plot to see the separation of classes by taking two predictors generated by the LDA model

```
#plot(lda_pred$x[,1], lda_pred$x[,2], pch=c(23,21,22)[unclass(lda_pred$class)], bg=c("red","green","blue")
```