The National Bank of Fort Worth Report

A) Non-Technical Part:

in comparison with the performance of the two algorithms used.
There is no confidence difference between their performance, As the logistic regression metric was 76.7% and the LDA metric was 75.5%.
The logistic regression is slightly better than the LDA algorithm.

B) Technical Part:

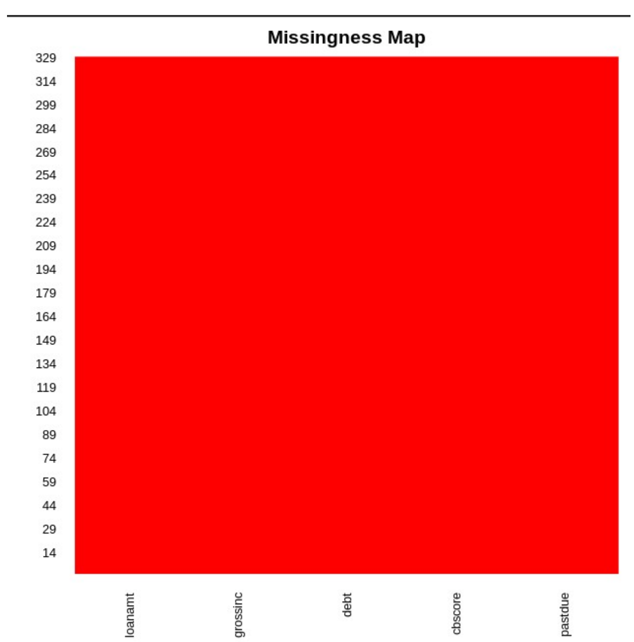
We have read the data into r data frame.
The data consists of 329 rows and 5 features, and here is the first 5 observations of the data.

```
> dim(bank)
[1] 329   5
> head(bank)
  pastdue cbscore debt grossinc loanamt
1       0     711   99   717.00     500
2       0     752   79  2417.00     500
3       1     654   63  3333.33    6547
4       0     650   62  2125.00    1200
5       0     605   57  2249.50   10000
6       1     774   56  4956.99   16000
```

Data has no missing values.

```
> sum(is.na(bank))
[1] 0
>
```

Assuring by plotting missing map:

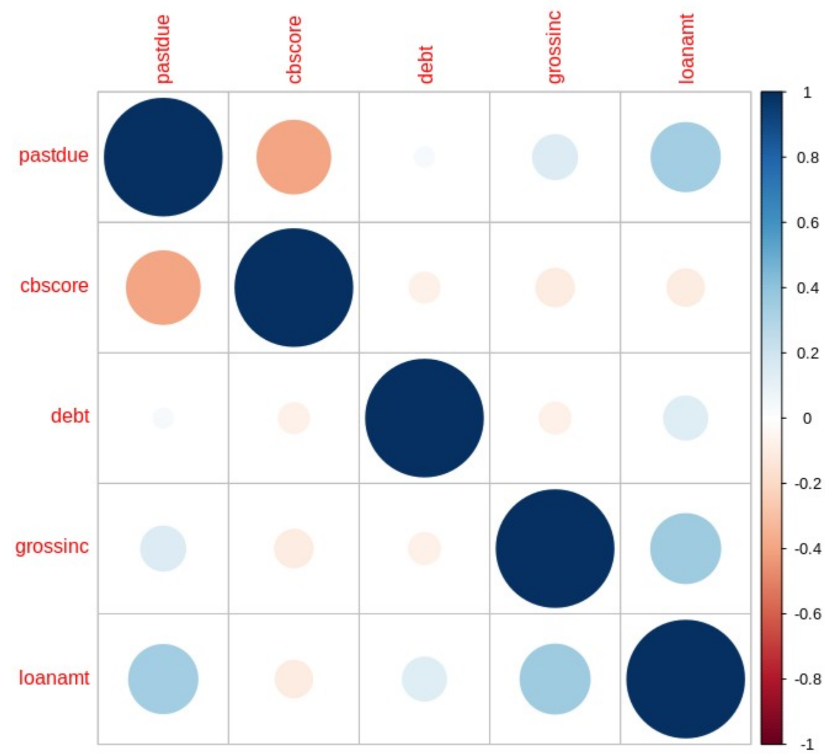


Summary of the data:

```
    pastdue          cbscore           debt           grossinc        loanamt     
 Min.   :0.0000   Min.   :508.0   Min.   : 0.00   Min.   : 509   Min.   :  200  
 1st Qu.:0.0000   1st Qu.:657.0   1st Qu.:19.00   1st Qu.:2247   1st Qu.: 2500  
 Median :0.0000   Median :696.0   Median :27.00   Median :3033   Median : 5000  
 Mean   :0.4286   Mean   :692.7   Mean   :26.78   Mean   :3330   Mean   : 5950  
 3rd Qu.:1.0000   3rd Qu.:726.0   3rd Qu.:35.00   3rd Qu.:4333   3rd Qu.:10000  
 Max.   :1.0000   Max.   :804.0   Max.   :99.00   Max.   :8292   Max.   :20000  
```

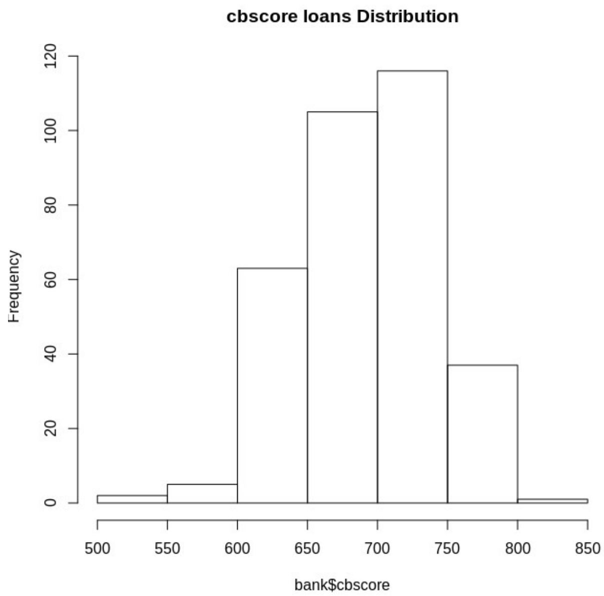
correlation matrix:

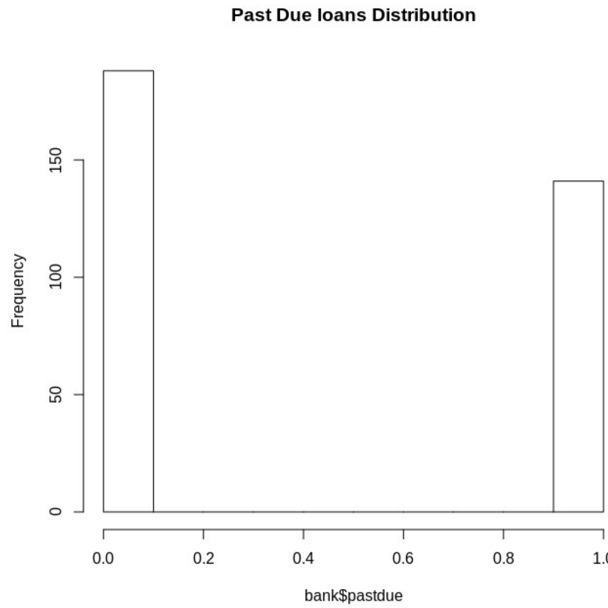
A dot-representation was used where blue represents positive correlation and red negative. The larger the dot the larger the correlation.
We can see that the loan amount has a weak positive correlation with past-due and Gross monthly income, and Score generated by the CSC has a weak negative correlation between past-due.



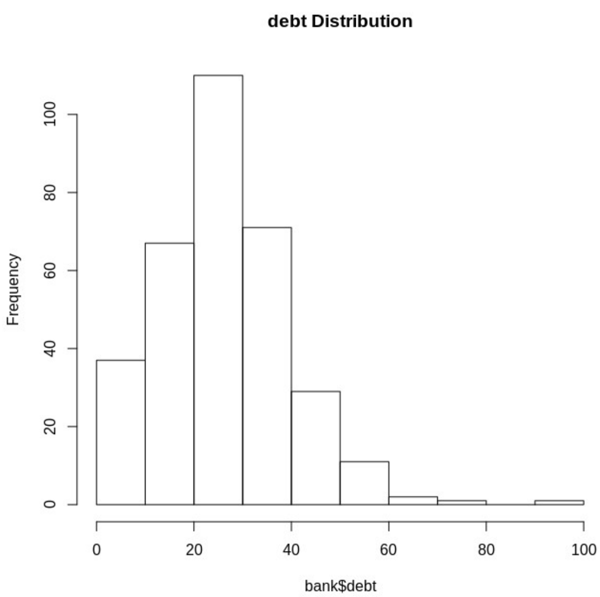
Investigate data features:

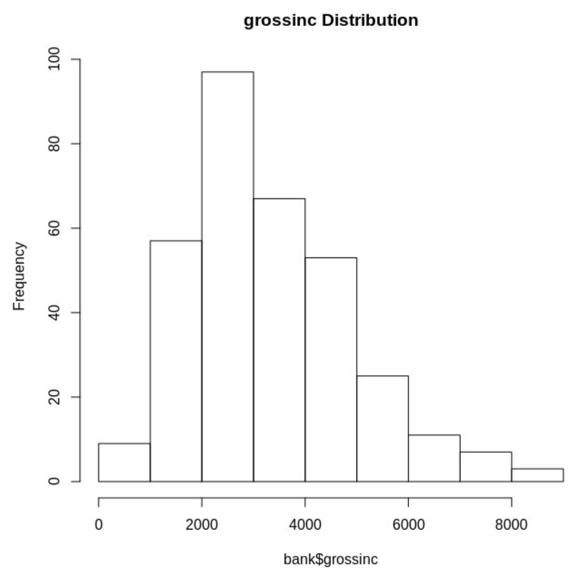
It's seen that the number of non past-due is greater than the past due loans.
Cbscore values limits from 500 till 850, and the majority value is set between 700 to 750.

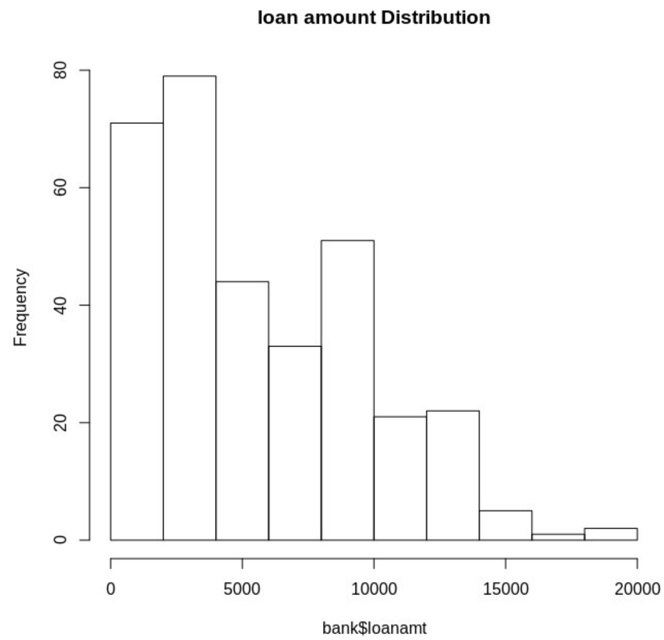




debt, loan amount, and grossinc features are right skewed.

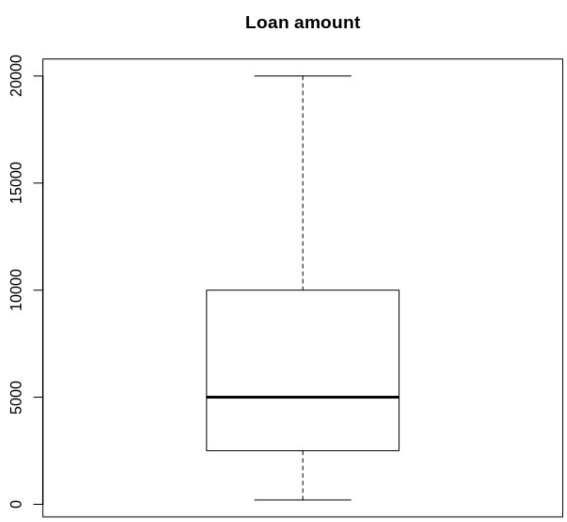


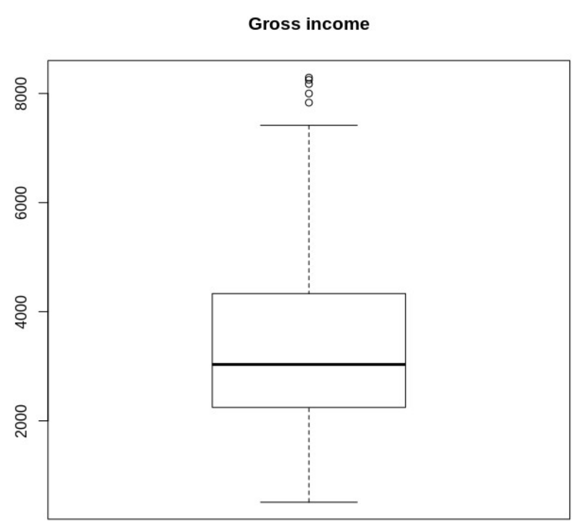




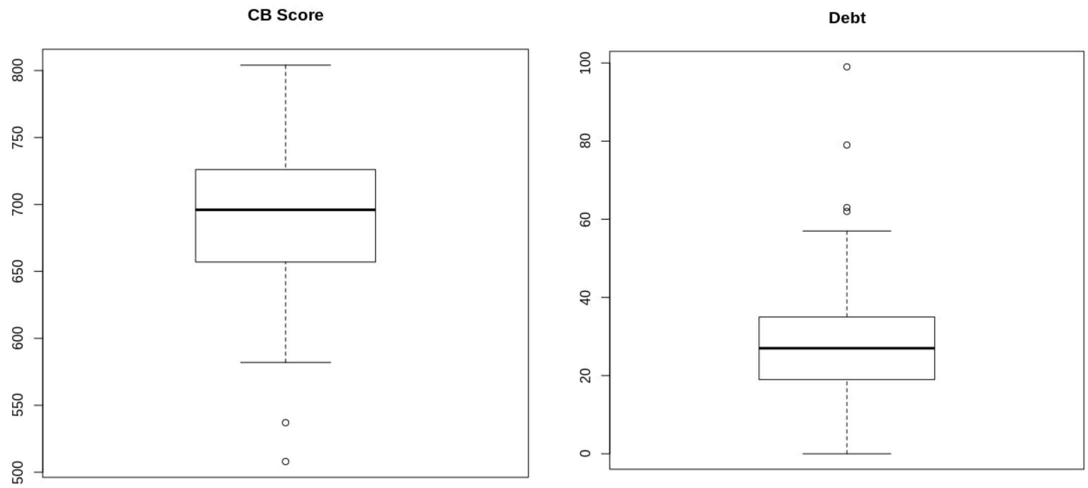
Box plotting:

We can see that the loan amount feature has no outliers, while the outliers for the gross income feature are starting from 7 K value.





The CB score feature has few outliers and starting to appear below the 550 value, while the debt feature outliers are starting to appear for the values above the 60 value.



Logistic regression summary:

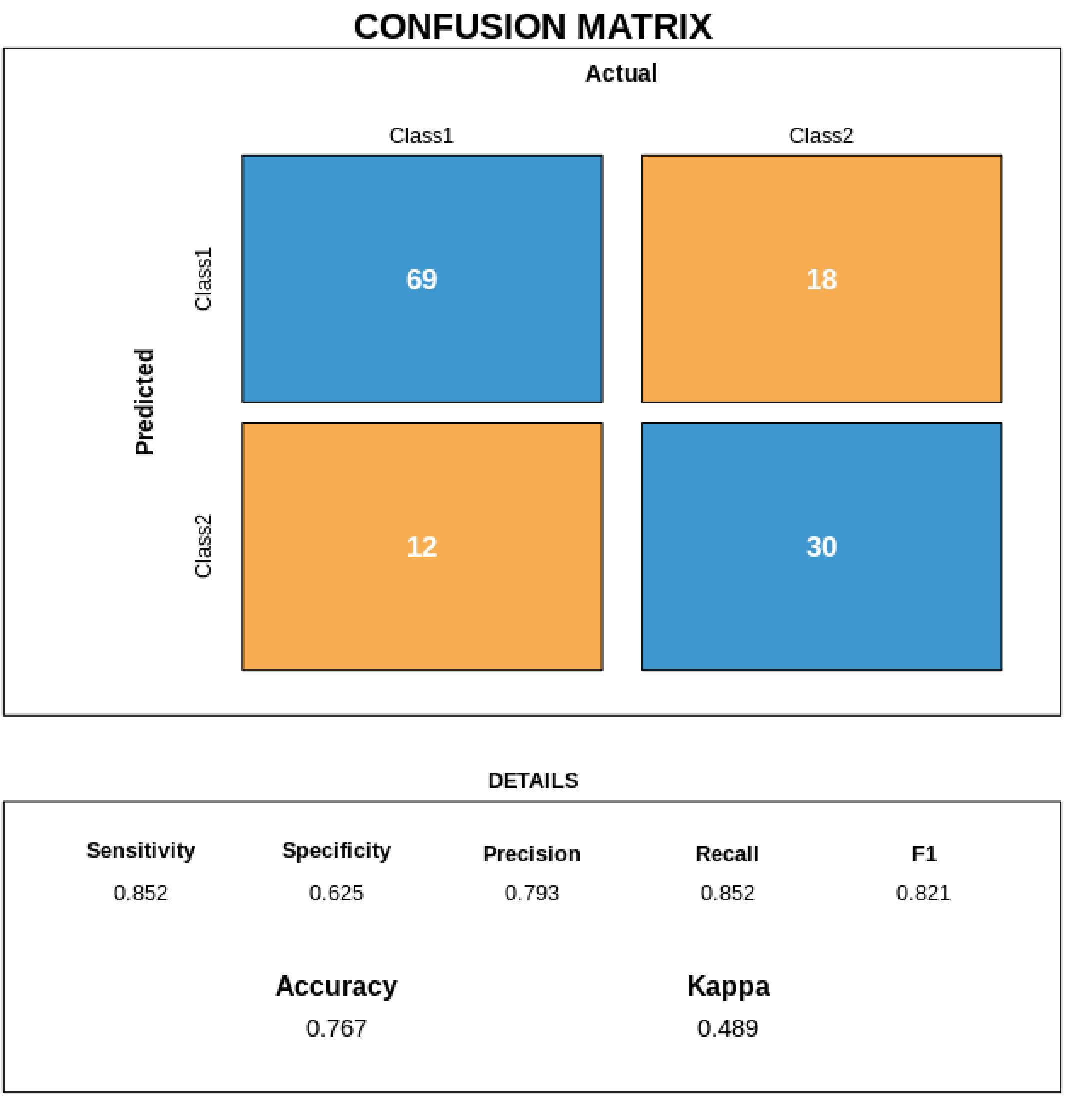
We can see here the standard error is very low for all features used.
It’s seen that all features are significant except debt and grossinc features.
The difference between deviance is not low as it’s more than 100.

```
> summary(glm.fit)

Call:
glm(formula = pastdue ~ cbscore + debt + grossinc + loanamt,
    family = binomial, data = bank)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0887  -0.8776  -0.4492   0.9275   2.2856

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.289e+01  2.201e+00   5.855 4.76e-09 ***
cbscore     -2.047e-02  3.142e-03  -6.514 7.33e-11 ***
debt        -7.546e-03  1.040e-02  -0.725    0.468
grossinc    -2.542e-05  8.732e-05  -0.291    0.771
loanamt      2.045e-04  3.737e-05   5.473 4.41e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 449.35  on 328  degrees of freedom
Residual deviance: 355.72  on 324  degrees of freedom
AIC: 365.72

Number of Fisher Scoring iterations: 4

> |
```

Evaluate logistic regression:

Here is the confusion matrix of the logistic regression model.
The false positive for past-due is 12 observations and for non past-due loan is 30 observations.
The model accuracy is 0.76.

```
glm.pred  0  1
       0 69 18
       1 12 30
```



LDA:

We have fitted a lda object to the train split, and here is the lda object printed.

```
> lda
Call:
lda(pastdue ~ ., data = train)

Prior probabilities of groups:
    0     1
0.535 0.465

Group means:
   cbscore     debt grossinc  loanamt
0 701.3738 36.07477 3028.900 4837.554
1 676.7849 33.35484 3579.633 8625.184

Coefficients of linear discriminants:
                   LD1
cbscore  -1.112225e-02
debt     -2.168952e-02
grossinc  4.080381e-05
loanamt   2.033883e-04
>
```

We called the prediction function , and here is the difference between the actual values and predicted values for the first 5 observations in the test split.

```
> data.frame(original = test$pastdue, pred = pred_lda$class)
    original pred
1          0    0
2          1    0
3          1    1
4          0    0
5          1    1
```

LDA Evaluation:

the lda model resulted in an area under the curve of 0.775 which is fair.
An AUC of 0.75 means that if we take two data points from two different classes, there is a 75% chance that the model will correctly rank order them, hence the positive class has a greater prediction probability than the negative class.