1

Advanced statistical process control forNew York City: Citywide Payroll Data– using PCA and Logistic regression

Submitted by **KEERTHANA SATISH**

Abstract—The salary details of the government employees have been always on the discussion. Citywide payroll data for New York City for the 2016 fiscal year was collected to analyse the details of how many people do the various city agencies employ, and how much does each department spend on salary in total. This data can be used to analyze how the city's financial resources are allocated and how much of the city's budget is being devoted to overtime. In this project, statistical process control using Principal Component Analysis and logistics regression Analysis is implemented to control the salary of the employees by analyzing principal components and latent factors.

**Index Terms: Principal components analysis (PCA), PCA Implementation,** **logistics regression analysis and logistics regression implementation**

* 1. INTRODUCTION

1. Factors influence on Employee salary

The data consists of employee salary details with 50 sample observations against 5 variables. This sample observations contains the salary, pay rate, and total compensation of every New York City employee. In this dataset this information is provided for the 2016, fiscal year and provides a list of who gets paid how much.

Note that fiscal years in the New York City budget cycle start on July 1st and end on June 30th.Each sample is being calculated against different parameters such as Base Salary, Regular Hours, Regular Gross Paid, OT Hours and Total Other Pay. The most important of these fields is "Regular Gross Pay", which provides that employee's total compensation.

The 50 observations represent the employees for which the inspections were carried out. The 5 variables have specific values.

X1= Base Salary

X2= Regular Hours,

X3= Regular Gross Paid.

X4= OT Hours

X5= Total Other Pay

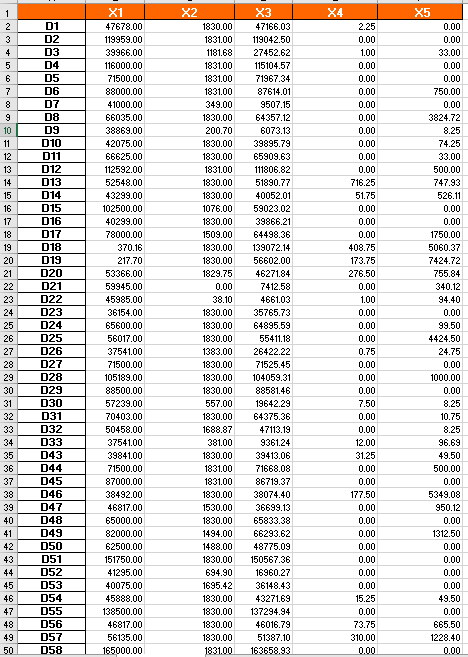


Fig. 1. rows of Citywide payroll data for New York City data.

Since different attributes have different measurements, we standardize the data before going into PCA and regression part. As shown in the Box plot for standardize data in figure 2, From the figure we can say that all the other variables have the outliers and X2 has the highest median and X5 has the lowest median X4 and X2 (in the same order) have more number of outliers among all the variables and for the remaining variables the median value is between 0 and -0.5. Based on this plot, X1 and X3 feature distribution seems to be the most similar to normal distribution. The first three features of the data set are plotted against each other in figure 3.

To have a better intuition into the relationship between the attributes of the data set, bi variate scatter plot is shown in figure 4. Figure 4 shows the correlation matrix of the data set attributes. This matrix shows how different features are related. For instance, X1 has negative correlation with X4 as well as X5, which means if one of them increases, the other one will decrease.

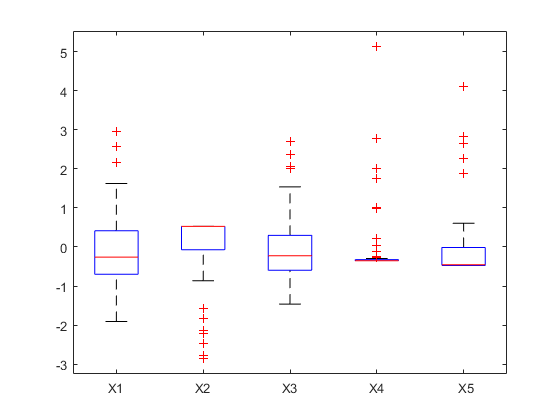


Fig. 2. Box plot of the New York City: Citywide Payroll Data

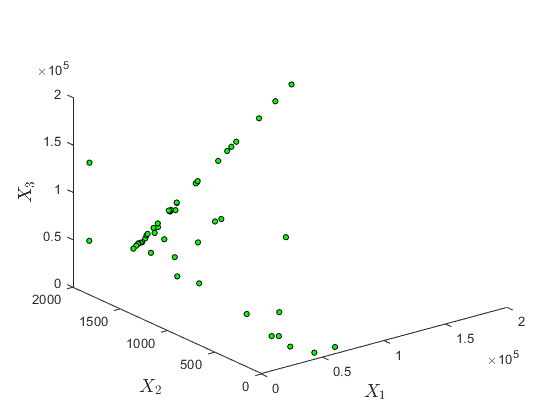


Fig.3. 3D visualization of data.

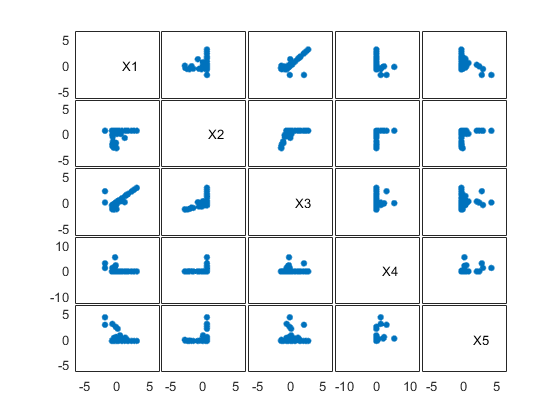


Fig. 3. bi variate scatter plot.

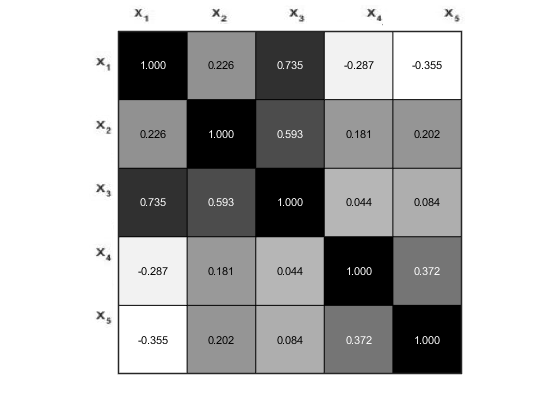


Fig. 4. Correlation matrix.

2

III. PRINCIPLE COMPONENT ANALYSIS

PCA is a multivariate technique to extract important features from a data set and represent them as orthogonal variables called principle components. This method is based on eigen decomposition for positive semi-definite matrices and singular value decomposition for rectangular matrices [4].

In general data matrix X, which is a n X p matrix will be pre- processed before the PCA analysis. First, it will be zero centered, which means the mean of each column will be equal to zero.

We have zero centered the data and calculated. The covariance matrix S, which is shown in Figure 6.

Next step in PCA analysis is to implement the eigen value decomposition of the covariance matrix S, as follows:

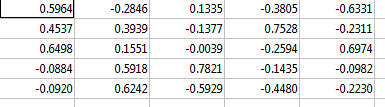
S = A**ʌ**; (1)

where A is a p X p matrix of eigenvectors, and **ʌ** is a diagonal matrix of eigenvalues. We have calculated the

eigenvalues and eigenvectors of the covariance matrix for the payroll data. The following matrix is the eigenvectors:



Fig. 5. Component correlation matrix.



0.5964 -0.2846 0.1335 -0.3805 -0.6331

0.4537 0.3939 -0.1377 0.7528 -0.2311

A= 0.6498 0.1551 -0.0039 -0.2594 0.6974

-0.0884 0.5918 0.7821 -0.1435 -0.0982

-0.0920 0.6242 -0.5929 -0.4480 -0.2230

The following vector is the eigenvalues. These values tell

us about the variance in particular dimensions.

2.0699

1.6626

λ= 0.6370

0.5266

0.1039

After determining the eigenvalues and eigenvectors, the

principle components can be calculated as follows:

Z = XA; (2)

where X is the zero-centered data, and the columns of Z are the principle components. For the payroll data, the following shows the first and second principle components:

Z1 = 0.5964 X1 +0.4537 X2+0.6498 X3 - 0.0884X4 -0.0920 X5

Z2 = -0.2846X1 +0.3939X2+0.1551X3 +0.5918X4 +0.6242X5

In the above-mentioned matrix, A table, the first column of the matrix gives the PC1🡪 Z1 and second column represents the PC2🡪Z2.

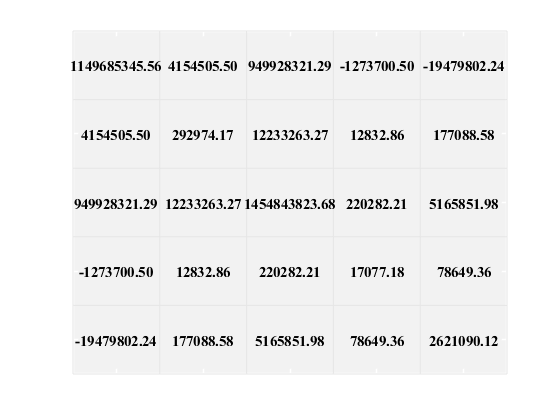


Fig. 6. Covariance matrix of zero centered data.

After calculating the principle components, they can be plotted against each other to have a better insight about their relationships. Figure 7 and 8 show these plots.

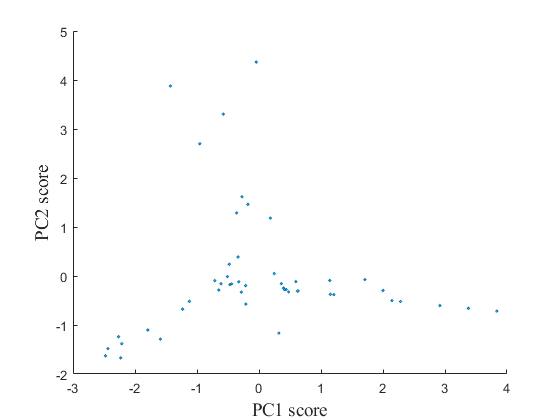


Fig. 7. scatter plot of PC2 score Vs. PC1 score.

3

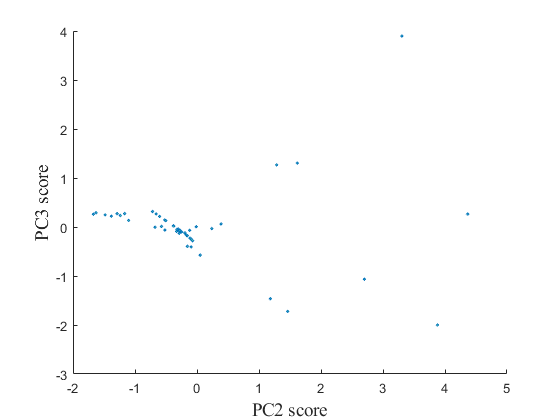


Fig. 8. scatter plot of PC3 score Vs. PC2 score.

Another important aspect of principle component analysis is that how much each attribute contributes to principle components.

For this purpose, PC coefficients are plotted against each other in Figure 9 and 10. The following items are some of the facts that can be inferred from these plots:

* X3 is the feature which contributes to PC1 more than others.
* X4 and X5 have similar coefficients for PC2.
* X1 is the feature which contributes to PC1 and PC2 less than others.
* X4 is the feature which contributes to PC3 more than others.
* X5 is the feature which contributes to PC3 less than others.

* Some features have positive coefficients like X5 for PC1 and some negative like X1 for PC2.

When the goal is to extract important information from the data matrix, the problem is to figure out how many components should be considered. There are different solutions to this problem. For instance, we can keep components whose eigen values are larger than the average eigen value. Another criterion which is used in this work is to calculate the explained variance for each component as follows:

= X 100%, j=1,……………………p (3)

Using the above equation, we have calculated the explained variance for the components in payroll data set, and the result is as follows:

l1 =41.3984% l2 =33.2518% l3 = 12.7395%

l4 = 10.5316% and l6 = 2.0787

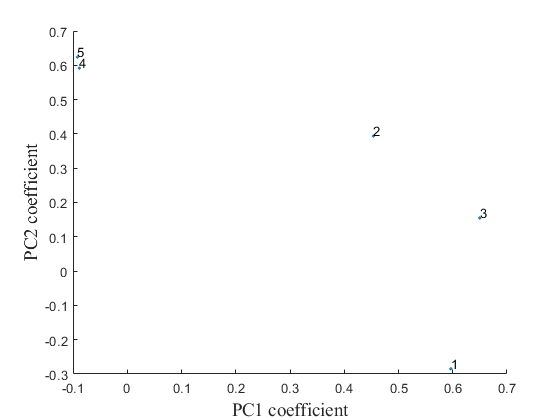


Fig. 9. scatter plot of PC2 coefficient Vs. PC1 coefficient.

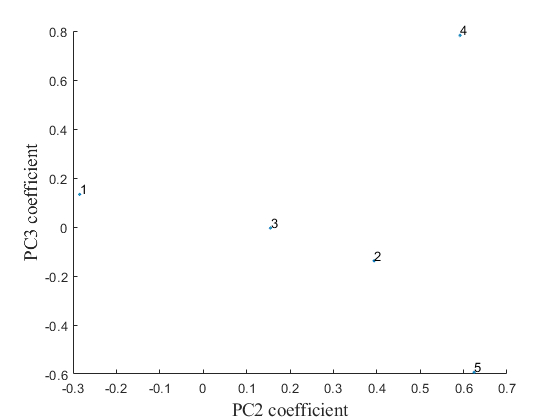


Fig. 10. scatter plot of PC3 coefficient Vs. PC2 coefficient.

Based on these values, which are also presented in Scree Plot Figure 11, we can infer that the first two components account for more than 82% of the variance in the banknote authentication data set. Thus, the minimum dimension to represent our data is d = 2, and we can omit the last two components for our classification phase, which will be described in the next section.

From the Pareto plot in figure 12, the cumulative explained variance for the components can be inferred. As it can be easily observed, the first 2 components account for most of the variance in data. Another useful plot in PCA is biplot. This plot shows two kinds of information at the same time: first it shows the PCA scores for observations, and second it shows how much each original variable contributes to the principle components.

Observations are presented as points in the plot, and variables are presented as vectors. The axis of bi-plot are the principle components. The 2D and 3D bi-plots for the bank data set is presented in Figures 13 and 14. The following information can be inferred from these plots:

4

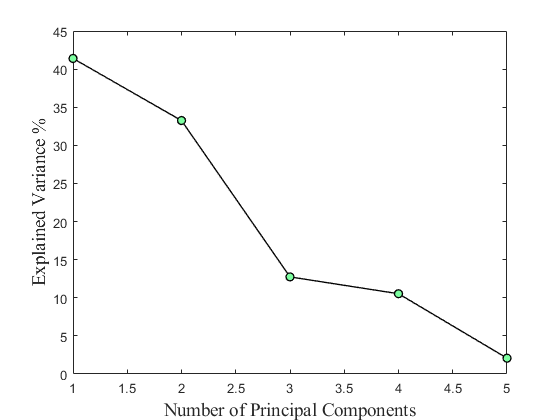


Fig. 11. Explained variance plot.

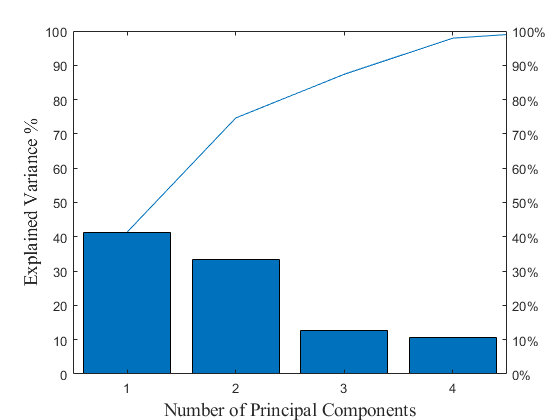


Fig. 12. Pareto plot.

* For PC1, X2, X3 and X1 have positive coefficients, but X4 and X5 have negative coefficients.
* For PC2, X2, X3, X4 and X5 have positive coefficients, but X1 has negative coefficient.
* Data points are distributed around the zero lines.
* X4 and X5 have almost the same coefficients for PC2.

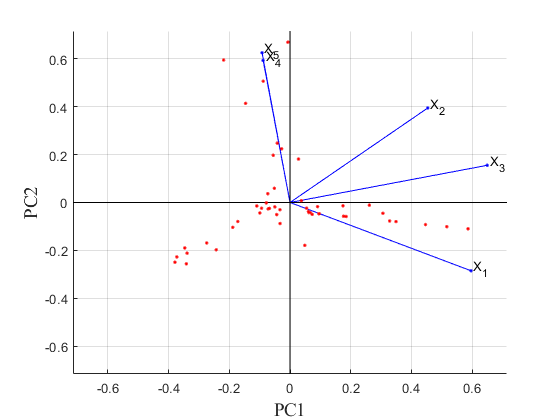


Fig. 13. 2D biplot.

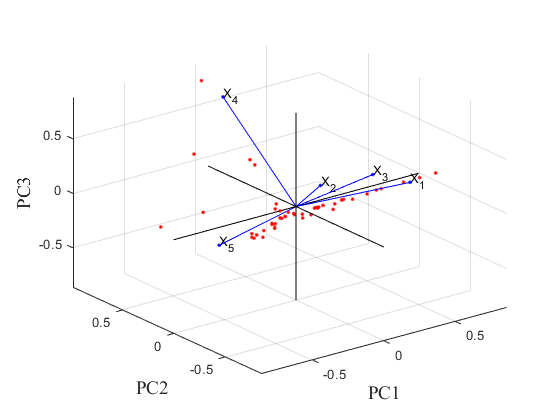


Fig. 14. 3D biplot.

Hotelling’s T2 is a measure of multivariate distance of observations from the center of data and is calculated as follows:

= (4)

Since we know that = A and =

= A = (5)

The following equations show the upper control limits for

phase one and phase two:

UCL = (6)

UCL = (7)

In these equations, (:) is the inverse CDF of Beta distribution and α is the significance level. We have calculated the Hotelling chart for the payroll data, and the result is shown in figure 15. As it can be inferred from this figure, several samples including 13, 18, 19, and 37 are out of control.

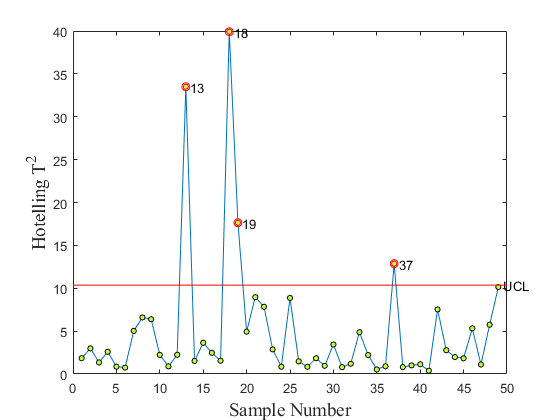


Fig. 15. Hotelling control chart.

Another useful control chart is the PC chart which can be

determined using the following equations:

5

UCL = +3

CL = 0

LCL = -3 (8)

This chart is provided in figure 16. According to this chart,

all samples are between the control limits, which means they are in control.

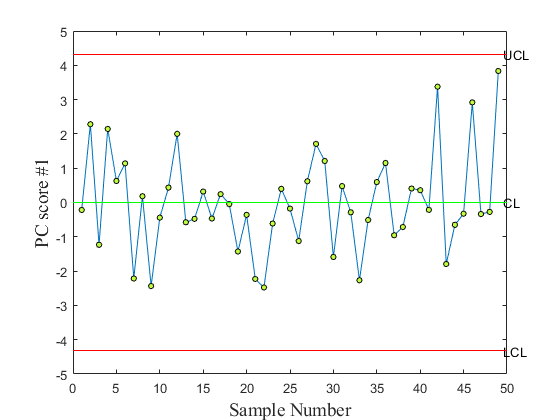


Fig. 16. PC control chart.

IV. LOGISTIC REGRESSION

The goal of a supervised learning algorithm is to train a model using n observations and try to distinguish between m different classes. logistic regression is used for analyzing a data set in which there are one or more independent variables that determine the outcome. In this logistic regression we took a new data set and analyzed the data .

The input of the model is a vector of length d as follows:

x = [x1, ..., xd]| ∈ Rd. (9)

Class label y is defined as follows:

y = [y(1), y(2), ..., y(m)]|, (10)

such that y(i) = 1 if x belongs to i and y(i) = 0

otherwise. Thus, the n training samples are presented as:

D = (x1, y1), ..., (xn, yn). (11)

One of the well-known supervised learning methods is regression. Regression analysis helps to understand the relationship between a dependent and several independent variables. In other words, it shows how much the dependent variable changes according to the variations in the independent variables.

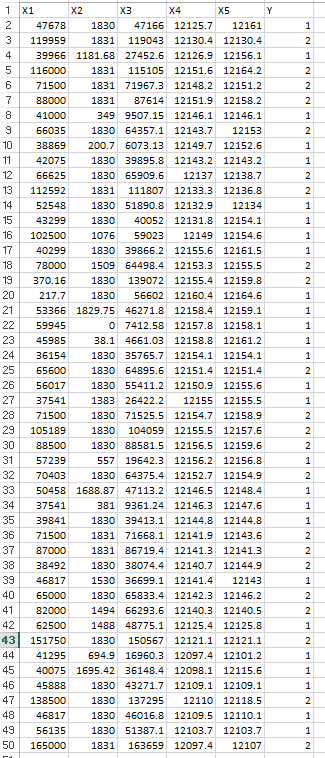


Fig. 1. Modified rows of Citywide payroll data for New York City data.

The general regression model is applicable when the class attribute is continuing, however, in the case of binary classes, we can use logistic regression.

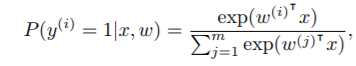
Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome.

The outcome is measured with a binary variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the binary characteristic of interest (dependent variable) and a set of independent variables.

Multinomial logistic regression is a method that generalizes

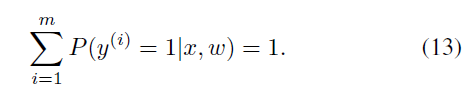
logistic regression to multi-class problems. In other words, it predicts the probabilities of different outcomes, and it is applicable when the outcomes are nominal [5].

Based on a multinomial logistic regression model, the probability that x belongs to class i is described as:

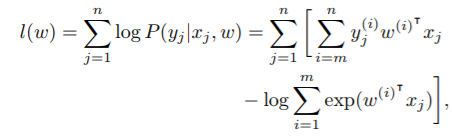


for i ∈ {1, ..., m}. In this equation, w(i) is the weight vector corresponding to class i. For binary problems, this equation is known as logistic regression.

However, when we have more than two class labels it is called multinomial logistic regression. Due to the normalization problem:



According to this equation, the weight vector for one of the classes need not to be calculated, because it can have estimated using other weight vectors. The training data D is used to estimate the components of w, using the maximum likelihood estimator (MLE), as follows:



which is done using Newton’s method or other maximization methods.

We are going to use both the features extracted using PCA, and all the features and compare them in terms of accuracy and calculation time.

The data set consists of 50 observations. These observations are divided into the training and testing data set. Training part includes 80% of observations. The rest of the observations are used in testing part. The results we report in this section are obtained from the test data set.

First step in using logistic regression is to fit the model.

The output of fitting logistic regression is a matrix of weights. Each column of this matrix corresponds to one class, however the number of columns is the number of classes minus one, because the last class can be calculated using previous ones.

In the problem we have two labels, therefore the weight matrix has one column. The first element is the bias term. We have fitted the logistic regression model using all the features and the training observations. The following is the weight matrix:

1.2295

-8.2859

B = -1.9167

-0.6121

Also,

0.8978

-8.0019

B2= -1.8100

Second step is to use this matrix to predict the labels. First element of this matrix is the bias term, and the other elements are the coefficients of features. As a result, the following formula can be used to calculate the probability of the classes:

P (y(i) = 1) = −1.2295X1−8.2859X2−1.9167X3 -0.6121X4 -0.3094X5

P (y(i) = 2) = 1 − P (y(i) = 1)

Finally, the label with the bigger probability will be assigned

to the observation.

We have applied the principle component analysis to our data, and adopted the first two components, which corresponds to more than 80% of variability in the data set. Figure 17 and 18 show the histogram of these two components. Figure 19 show the histogram of labels in the training data set.

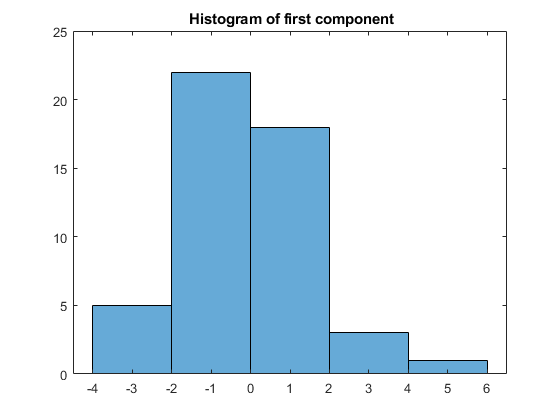


Fig. 17. Histogram of first component.

Then, we have fitted the logistic regression model, and predicted the labels of testing data. Figure 20 shows the boundaries obtained from this model, and how much it has

predicted the correct labels. Blue points that are located in

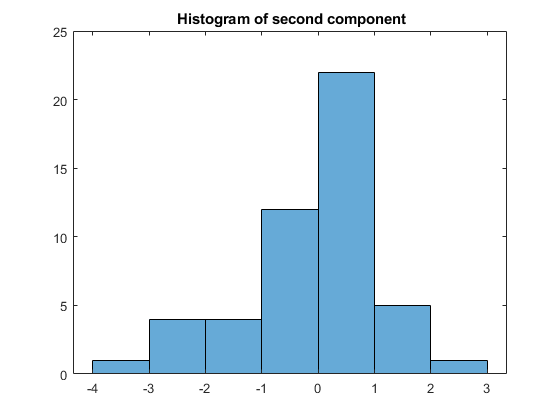


Fig. 18. Histogram of second component

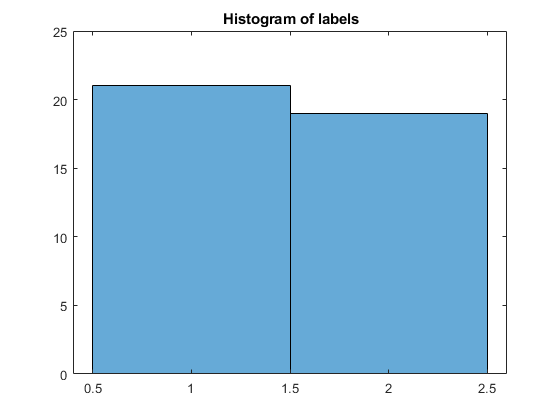


Fig. 19. Histogram of labels in the training data set

Table2: COMPARING RESULTS OF USING ALL THE FEATURES AND USING THE FIRST TWO COMPONENTS.

|  |  |  |
| --- | --- | --- |
| Number of features | Accuracy | Time |
| All the features | 0.7778 | 0.4158 |
| First two components | 0.8889 | 0.0161 |

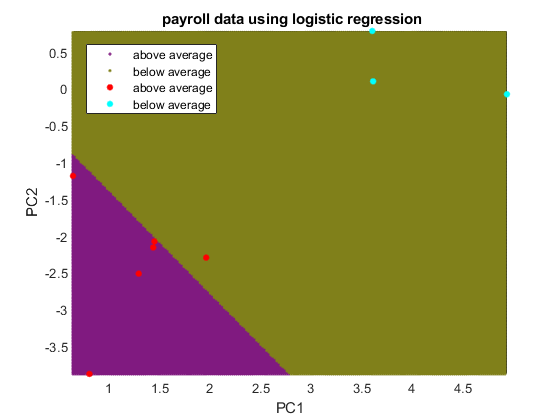


Fig. 20. Boundaries obtained from logistic regression.

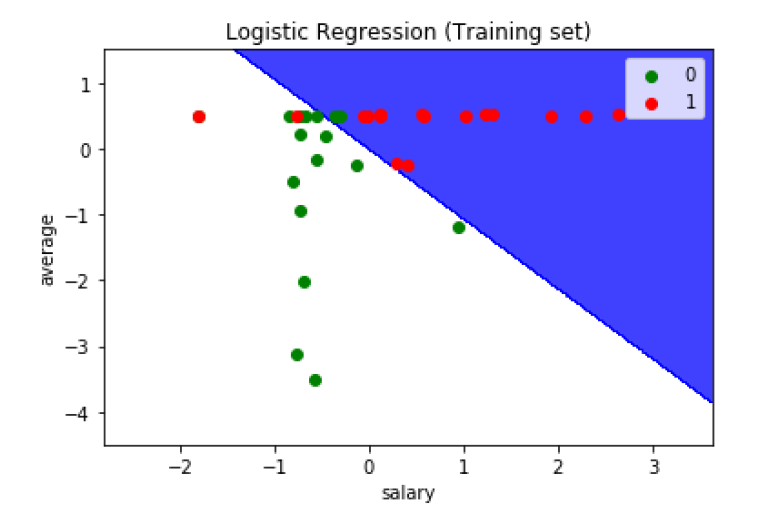


Fig. 21. Boundaries obtained from logistic regression for training test.

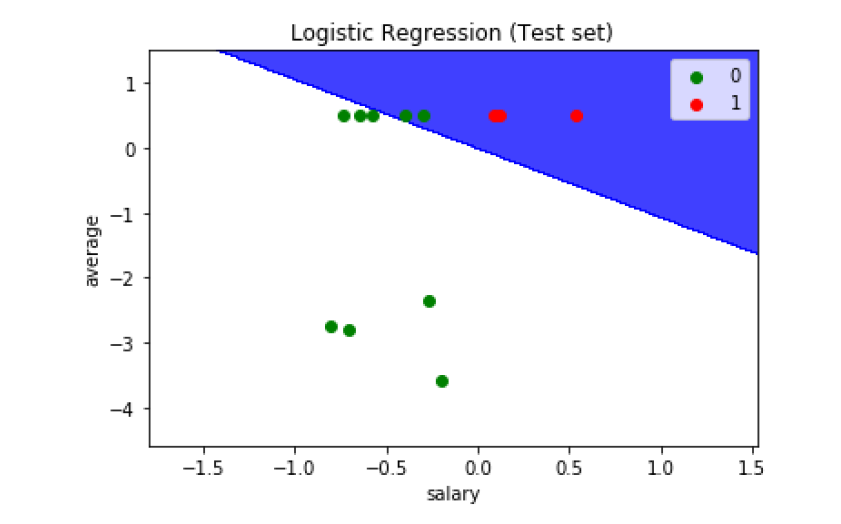


Fig. 22. Boundaries obtained from logistic regression for training test.

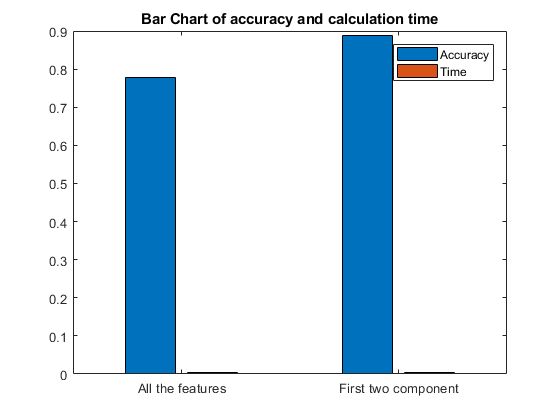


Fig. 23. Bar chart for comparing accuracy and calculation time of logistic

regression using all the features and the first two components.

Another useful criteria to compare two models, is using the confusion matrix, which allows the visualization of the

performance of an algorithm. Columns of this matrix are

the predicted classes, and rows the are actual classes [6].

Confusion matrices calculated for the two cases are shown in figure 24 and 25.

Predicted class

|  |  |
| --- | --- |
| 4 | 2 |
| 0 | 3 |

Target

Fig. 24. Confusion matrix in the case of using all the features.

Predicted class

|  |  |
| --- | --- |
| 5 | 1 |
| 0 | 3 |

Target

Fig. 25. Confusion matrix in the case of using the first two components.

Another very useful tool to compare two models is the Receiver operating characteristic (ROC) curve, which describes the diagnostic ability of a binary classification model [7]. This plot is created by plotting true positive rate against the false positive rate. True positive rate which is also called sensitivity is defined as follows:

Sensitivity = (15)

where TP is the number of samples which are correctly classified as positive, and P is the total number of samples with positive label.

False positive rate which is also called fall-out is defined as follows:

Fallout = 1 - (16)

where TN is the number of samples which are correctly predicted as negative, and N is the total number of samples with negative label.

Different models can be plotted in a single ROC curve, and the model which has the bigger area under its curve is the more preferred model. We determined this plot for the two cases, and this plot is shown in figure 26. According to this figure, using all the features leads to a better ROC curve.

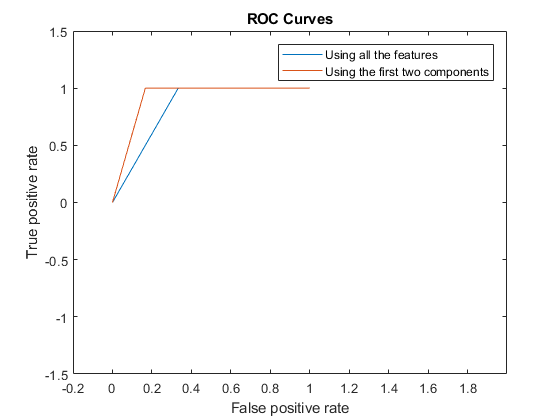
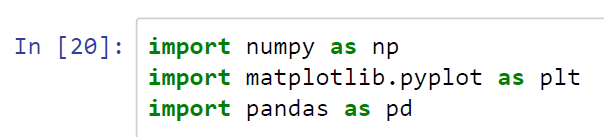
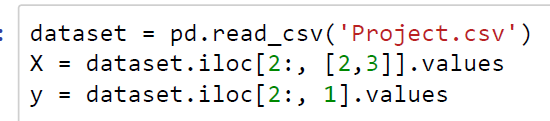


Fig. 26. ROC curve for the two cases.

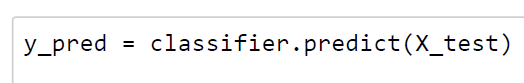
* initially we have to import data and mat plot



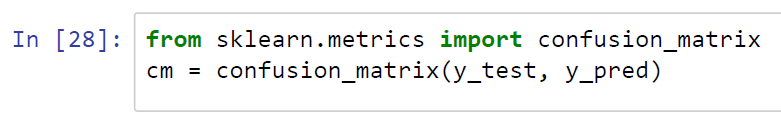
* and then assigning values for x and y



* then we have to predict the classifier



* and the last step we have to insert confusion matrix



In the analysis of the graph we plotted the graph for salary and average and the binary value there shows red denotes 1 green denotes 0, initially here we have to import the data and convert it into .csv file and upload the data here x label is salary and y label is average.

The above-mentioned graph distinguishes the data for average salary and individual salary and here they have been scattered and plotted in the in the graph.

V. CONCLUSION

In this work we have considered the Citywide payroll data for New York City for the 2016 fiscal year. Our work has two phases. In the first phase, we have adopted PCA to extract important features from the data set. Second, we have used logistic regression for classifying the observations. We have examined the performance of the logistic regression by considering two

cases.

The first case is to use all features in the data set, and the second case is two use the first two components obtained from PCA. Our results show that using all the features lead to a higher accuracy. however, it is more time consuming. It can be concluded that by analysing either the first three principal components, rather than analyzing the whole data which can be cumbersome task. As it is proved that the first three principal components contains more than 80% of variability will be enough to predict the status.

Therefore, there should be a trade-of between accuracy and time, based on the goal

* 1. BIBLIOGRAPHY

[1] “Kaggle Repository,” https://www.kaggle.com/new-york-city/nyc-citywide-payroll-data

[2] Principal Component Analysis, Wikipedia

Internet: http://en.wikipedia.org/wiki/Principal\_component\_analysis

[3] “Logistic regression,” https://en.wikipedia.org/wiki/Logistic\_regression

[4] Herve Abdi , Lynne J. Williams, “Principal Component Analysis,” Wiley

Interdisciplinary Reviews: Computational Statistics, 2010.

[5] “Confusion matrix,” https://en.wikipedia.org/wiki/Confusion\_matrix

[6] “Receiver operating characteristic,” https://en.wikipedia.org/wiki/

Receiver\_operating\_characteristic

[7] INSE 6220 - Lecture Notes.