EXPERIMENT: 4

Aim: Feature Reduction using Principal Component Analysis (PCA)

Theory:

PRINCIPAL COMPONENT ANALYSIS (PCA):

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is

often used to reduce the dimensionality of large data sets, by transforming a large set of

variables into a smaller one that still contains most of the information in the large set.

Reducing the number of variables of a data set naturally comes at the expense of

accuracy, but the trick in dimensionality reduction is to trade a little accuracy for

simplicity. Because smaller data sets are easier to explore and visualize and make

analyzing data much easier and faster for machine learning algorithms without

extraneous variables to process.

So to sum up, the idea of PCA is simple — reduce the number of variables of a data set,

while preserving as much information as possible.

TEPS TO PERFORM PCA USING XLMINER IN EXCEL:

Step 1: Select a cell within the data set, then on the Data Mining ribbon, select

Transform - Principal Components to open the Principal Components Analysis.

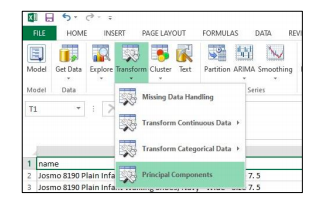


Fig. 1 Selecting Principal Components from Data Mining Ribbon

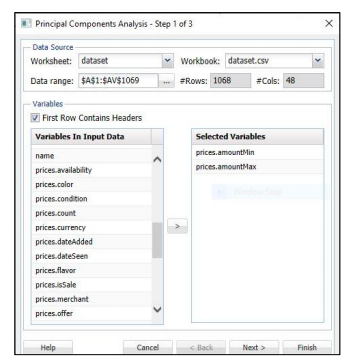


Fig. 2 Selecting variables (columns) to perform PCA

From the Variables In Input Data list, select variables, then click the > button to

move them to the Selected Variables list, and click Next to open the Principal

Components Analysis - Step 2 of 3 dialog.

Step 2: XLMiner provides two routines for specifying the number of principal

components: Fixed #components and Smallest # components explaining. Use the

Fixed # components method to specify a fixed number of components or variables

to be included in the reduced model. Use the Smallest #components explaining

method allows the user to specify a percentage of the variance.

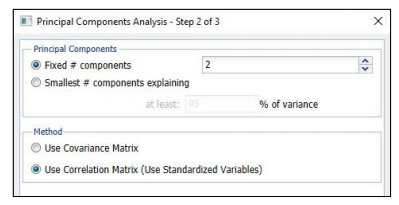


Fig. 3 Selecting Fixed Components Option

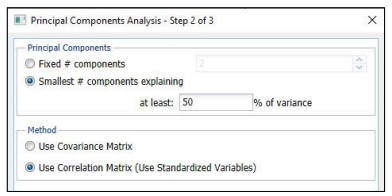


Fig. 4 Selecting Smallest Components Option

XLMiner provides two methods for calculating the principal components: using the

covariance, or the correlation matrix. When using the correlation matrix method, the

data is normalized first before the method is applied (i.e., the data set is normalized

by dividing each variable by its standard deviation). Normalizing gives all variables

equal importance in terms of variability.

Select Use Correlation Matrix (Use Standardized Variables), then click Next to open

the Principal Components - Step 3 of 3 dialog.

Step 3: Confirm Show principal components score is selected, then click Finish.

This option displays an output matrix where the columns are the principal

components, the rows are the individual data records, and the value in each cell is

the calculated score for that record on the relevant principal component.

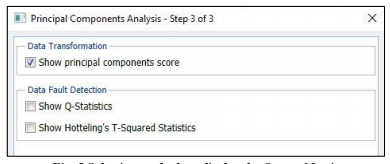


Fig. 5 Selecting methods to display the Output Matrix

Output:

Two worksheets are inserted after the Description worksheet: PCA\_Output, and

PCA\_Scores. The output is displayed below:

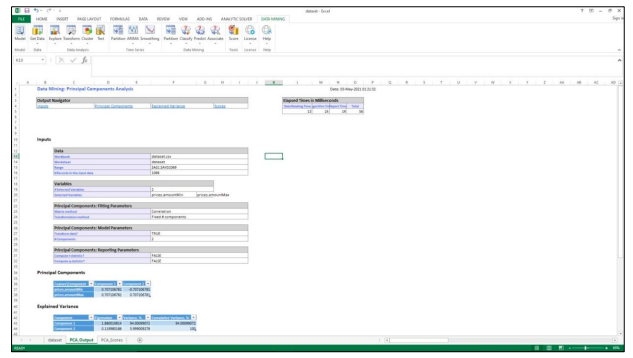


Fig. 6 PCA\_Output for Fixed Components

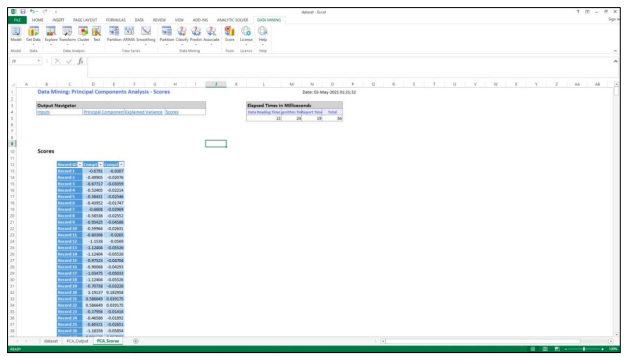


Fig. 7 PCA\_Scores for Fixed Components

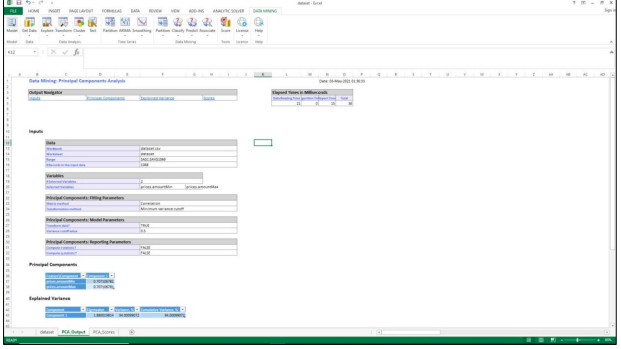


Fig. 8 PCA\_Output for Smallest Components

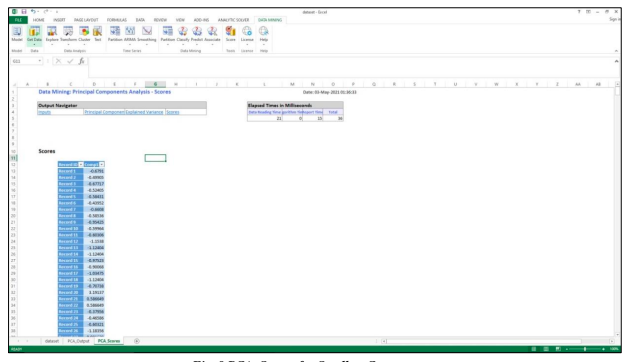


Fig. 9 PCA\_Scores for Smallest Components