NAME: CONSOLATA FAUSTIN MBILINYI

REG NO:220242406098

MODULE: DATA MINING

MODULE CODE:COU 08104

TASK: ASSIGNMENT 2

Qn1.1

Variables relationships

- Very strong positive linear relationship between s1 and s2(r=0.89663) and strong positive relationship betweebs4 and s2(r=0.659817)
- Strong negative relationship between s3 and s4(r=-0.7384) and moderate negative relationship between s3 and BMI(R=-0.366811)
- -Moderate positive relationship between BMI and BP(r=0.3954) and BMI and s5(r=0.0352)
- Weak relationship between age and sex and between s1 and sex

1.2 collinearlity

Collinearity refer to the situation in multiple regression model whwre two or more predictor variables are highly correlated with one another thus one variable can be accurately predicted by linaer combination of the others

Effects on the coefficients

It make the estimated coefficient values unstable and very sensitive to small changes in the data

1.3 significance values of variables and collinearity problem

Significance: not all variables are significant example age and serum often show p_value much higher than 0.05

Collinearity problem: yes there is , the presence of non significant variables despite a decent r2 value combined with the high correlation indicates that the model is struggling to distinguish the individual impacts of each predictor

1.4 Forward versus backward selection

Forward and backward selection are among the methods of attributes subset selection and here is how they differ

Forward selection procedure starts with an empty set of attributes as reduced set then best of original are determined and added to the set

WHILE

Backward selection procedure starts with full set of attribute at each step it remove the worst attribute of the remaining set

1.5 Stepwise approach in selection variables includes the set where either best variables are added into an empty set or worst attributes are reduced from the original or reduced set.

selected variables:['BMI', 'S5', 'BP', 'S1', 'SEX', 'S2']
the function work by identifying the smaller subset of the variables
stepwise MSE:2876.6833
stepwise R2: 0.5149

2.1 Differences between logistic and linear regression
logistic regression used for classification tasks where the dependent variable is categorical
WHILE
Linaer regression designed to predict the continuous, quantative output

2.4 parameter estimates
Significance : these parameters are statistically significant as their p_values are below 0.05 threshold

3.1 Qualitative Description of PCA:

Principal Component Analysis (PCA) is a dimensionality reduction linear technique that transforms a large set of correlated variables into a smaller set of uncorrelated variables called Principal Components.

Application of PCA in machine learning

- Feature extraction as create principal component(new feature) that summarize the original feature
- Noise reduction through sorting of data in descending order of significance thus those with low variance considered as noise
- Data visualization : high dimensional data can be projected for visualization
- Image and signal processing by compressing image data and extracting dominant patterns

PCA is useful in transform of explanatory variables because it simplifies complex datasets, removes noise, and resolves multicollinearity issues. By focusing on the components that capture the most variance, we can visualize high-dimensional data and improve the efficiency.

3.2 Mathematical Equations for PCA:

PCA transforms the raw input data matrix X into a new set of variables Z using the equation:

$$Z = XV$$

• X (Data Matrix): The standardized input data where each row is an observation and each column is a feature.

• V (Weight/Loading Matrix): The matrix of eigenvectors derived from the covariance or correlation matrix of X.

•Z (Principal Components): The newly transformed variables which are linear combinations of the original features, ordered by the amount of variance they explain.

3.3 PCA Weight Interpretation:

The First Principal Component (PC1) show consistent positive weights of stocks with highly similar magnitude between 0.15 and 0.20, this indicate that PC1 represents the “market factor” it captures the general trend where most stocks move in the same directions simultaneously .Because weights are relatively uniform PC1 can be thought as weighted average of entire Dow jones index.

The Second Principal Component (PC2) represent a differential factor by highlighting differences between sectors or identifies stocks that move contrary to the main market trend. Weight vary significantly with some stocks having strong positive example INTC, HD, GS and other having strong negative example AMGN, MCD HON

PC2 capture nuances since this stocks with positive PC2 weight move in one direction while those with negative weight moves in opposite direction

3.4 The principal component required to explain the 95% of the variance are 17

3.5 Identifying Distant Stocks:

Three most distant stocks: ['HD', 'INTC', 'AMGN']

Stocks that appear as outliers in the PCA scatter plot (the most distant from the average) are considered unusual because their price movements do not align with the broader Dow Jones index. These stocks likely belong to specific sectors that experienced unique volatility or independent growth trends during 2020-2021 period that were not shared by the rest of the market