Assignment - 2

1. The data on features given in the table below are collected to estimate the published relative performance (PRP) of a centralised processing unit. The data is given in the CPU\_Data file.

|  |  |  |
| --- | --- | --- |
| Feature Name | Description | Unit |
| MYCT | Machine Cycle Time | Nanoseconds |
| MMIN | Minimum Main Memory | Kilobytes |
| MMAX | Maximum Main Memory | Kilobytes |
| CACH | Cache Memory | Kilobytes |
| CHMIN | Minimum Channels | Channels |
| CHMAX | Maximum Channels | Channels |

1. Provide a descriptive summary of fields in the dataset
2. Explore the relationship between the response variable PRP and explanatory features using scatter plots and correlation matrix. Give your interpretation of the relationship between response and explanatory features.
3. Develop a model to predict the PRP using explanatory features. Provide the model coefficient table and interpret. Are all features have a significant impact on the response? Provide the mathematical expression of your model.
4. Provide F statistic value and corresponding p\_value? Give your comments on model significance?
5. Provide R2 and adjusted R2 values. Comment on model accuracy.
6. Comment on model adequacy based on residual analysis and plots. Provide normal probability plot of residuals and normality test result (test statistic and p\_value) and comment on the normality of residuals.
7. Provide Residuals versus Predicted values plot and given your interpretation.
8. Perform k fold (k = 10) cross-validation. Provide the mean square error and root mean square error obtained for original data and during cross-validation. Give your comments on model generalizability.
9. The data has collected on health profile parameters of people shown symptoms of heart disease and their diagnostic results are given in the Heart\_Disease\_Data file. The list of health profile features on which data is collected is given in the table below.

|  |  |  |
| --- | --- | --- |
| SL No | Feature Name | Description |
| 1 | Age | Age |
| 2 | Sex | Sex |
| 3 | CP | Chest pain type |
| 4 | RestBP | Resting blood pressure |
| 5 | Cholesterol | Serum cholesterol in mg/dl |
| 6 | FBP | Fasting blood sugar > 120 mg/dl |
| 7 | RestECG | Resting electrocardiographic results |
| 8 | Max\_HR | Maximum heart rate achieved |
| 9 | ExAngina | Exercise-induced angina |
| 10 | Oldpeak | ST depression induced by exercise relative to rest |
| 11 | Slope | The slope of the peak exercise ST segment |
| 12 | CA | Number of major vessels (0-3) colored by flourosopy |
| 13 | Thal | 3 = normal; 6 = fixed defect; 7 = reversible defect |

1. Provide a descriptive summary of the features
2. Explore the relationship between response “Result” and the features using graphically.
3. Develop a logistic regression model to classify whether a patient is having heart disease or not using the feature values. Give the model coefficient table and the mathematical expression of the model.
4. Compare the model with a null model and provide corresponding p\_value. Is the model significant?
5. Compute and give actual versus predicted matrix. Compute accuracy & misclassification %. Comment on model accuracy.
6. Check model generalizability using k -fold (k = 10) cross-validation. Provide accuracy & misclassification % obtained for data used for developing the model and during cross-validation. Is there a deterioration in model performance? Comment on model generalizability.
7. Compute sensitivity, specificity, precision & f-measure. Comment on the aforementioned performance measures. Is the model equally good at predicting having heart disease (positive) and not having heart disease (negative) cases correctly?