Graded Project – Linear Regression

70 Marks

Problem Statement:

The data is taken from the comp-activ databases which is a collection of computer systems activity measures. The data was collected from a Sun Sparcstation 20/712 with 128 Mbytes of memory running in a multi-user university department. Users would typically be doing a large variety of tasks ranging from accessing the internet, editing files or running very cpu-bound programs. The data was collected continuously every 5 seconds. Using the various system programs which are running in the background for every task being performed by the user, Predict the percentage portion of time (out of 100), that cpu runs in user mode, and how does each system program affect the same.

Domain Name: Technology and IT

Data Set: Comp-active Data

The variables in the data are the key CPU parameters while measuring activities such as writing, reading, accessing some platform or a internal software. These are all numerical parameters. There is one categorical variable which suggests the dependency on the CPU

lread - Reads (transfers per second) between system memory and user memory

lwrite - writes (transfers per second) between system memory and user memory

scall - Number of system calls of all types per second

sread - Number of system read calls per second .

swrite - Number of system write calls per second .

fork - Number of system fork calls per second.

exec - Number of system exec calls per second.

rchar - Number of characters transferred per second by system read calls

wchar - Number of characters transfreed per second by system write calls

pgout - Number of page out requests per second

ppgout - Number of pages, paged out per second

pgfree - Number of pages per second placed on the free list.

pgscan - Number of pages checked if they can be freed per second

atch - Number of page attaches (satisfying a page fault by reclaiming a page in memory) per second

pgin - Number of page-in requests per second

ppgin - Number of pages paged in per second

pflt - Number of page faults caused by protection errors (copy-on-writes).

vflt - Number of page faults caused by address translation .

runqsz - Process run queue size (The number of kernel threads in memory that are waiting for a CPU to run. Typically, this value should be less than 2. Consistently higher values mean that the system might be CPU-bound.)

freemem - Number of memory pages available to user processes

freeswap - Number of disk blocks available for page swapping.

usr - Portion of time (%) that cpus run in user mode

Tasks to be performed:

1. Load data and describe data (5 Marks)
    a. Import necessary libraries & packages
    b. Load dataset
    c. Check necessary details about data like shape, data types of the variable, missing values etc
2. Perform EDA and data cleaning (20 Marks)
    a. Generate the summary statistics for each of the variables and write comments on your observations
    b. Working with Null/Missing values
        i. Check for Missing values and perform the necessary steps for data imputation and provide reasoning for your approach
        ii. Check for the Zero values and understand the importance of that data point. Please provide your comments on if we need to change them (impute) or drop them
    c. Working with Outliers
        i. Check for outliers and provide comments
        ii. Perform outlier treatment (only if required)
    d. Scaling the data
        i. Only if scaling is necessary, please perform the same and provide your reasoning
    e. Working with Categorical variables
        i. Identify the categorical data given as part of the data set?
        ii. Perform encoding and provide detailed comments and reasoning for the encoding approach
3. Perform univariate, bivariate & Multivariate analysis (20 Marks)
    a. Perform Univariate analysis for each variable and write comments
    b. Perform bivariate analysis and make necessary inference about the relation between the variables
    c. Perform Multivariate analysis and make necessary inferences about the relation between variables
    d. Check Covariance and Correlation and identify positively and negatively correlated variables
    e. Identify the variables which has multicollinearity. Check for multi collinearity and drop the variables

Hint: Use necessary plots such as Distribution plots, Box plots, Pair plots, heat maps, Histograms etc

4. Building a Linear Regression Model (10 Marks)
    a. Prepare data for model building
    b. Build linear regression model.
    c. Find the features that add value to the model. Identify the list of Variables which highly impact the prediction based on the correlation Metrix given for regression (target variable)
    d. Print test and train results with all variables and best fit line
5. Model Performance (10 Marks)
    a. Check for the performance measures for linear regression (Hint: RMSE, R-square, etc.) – 7 Marks
    b. Experiment with data transformation and suggest if we can improve the model performance. – 3 Marks
6. Business report, Inferences and observations for all the Tasks (5 Marks)
    a. Each step needs to be executed clearly separately gradually. Also ensure you provided proper comments at each step and provide respective inferences or observations
    b. Please note the hypothesis inferences, Observations and Conclusion should be available in both the workbook (.ipnyb) and the report clearly.
    c. All the answers to the questions to be submitted in a sequential manner as part of the business repot
    d. The Business report should include a detailed explanation of the approach used, insights, inferences, all outputs like graphs, tables, etc. The Level of detail and explanation in business report should be deeper
    e. Your report should not be filled with formulas. Only use important metrics or values or charts obtained from each step of analysis