Department of Mechanical Engineering

ME 781: Engineering Data Mining and Applications

Assignment-1

With reference to the given data set: a1-data-set.csv

1. Visually inspect the data set (How? What will you do?) and list your general observations.
2. You will observe that there are some "missing values" in this data set:
    a) Can you afford to just ignore the missing data while creating a ML model using the data as a training set? Justify your answer.
3. Try the following strategies for imputing the missing values and create a new, complete data set (ie. with no missing values) in each case.
    a) Use the mean of the available data
    b) Use the median of the available data
    c) Use "Regression Tree" to estimate the missing values
    d) Use KNN regression to estimate the missing values
4. Using **each** of the above newly created data set (3a-3d) :
    a) Visualize the data set by creating a scatter plot of y.
    b) Fit an OLS model and note down / calculate parameters / quantities outlined in Table-1 and based on them provide your qualitative assessment of each case as shown in Table-2. (How do you interpret R, Adj. R, p-value, F-Stat, RMSE and MAE to assess the quality of your regression?)
    c) Fit an OLS model ignoring the data points with missing values and record your critical observations in Table-1 and Table-2. How does it compare with the OLS models fitted using the data sets of 3a … 3d? Explain your results.

Table-1

Case	R value	Adjusted R	p-value	F-Statistic	RMSE	MAE
3a						
3b						
…	…	…	…	…	…	…
4b						

Case 3a	Your assessments based on the above parameters
Case 3b	Your assessments based on the above parameters
…	…
Case 4b	

b) Create the following plots for each of the above cases, and record and explain your observations and interpretations.

i. Y-actual v/s y-predicted

ii. Residuals v/s serial number of the observation [Residuals = y_actual – y_predicted]

iii. Create a histogram of the residuals. Explain the shape of this histogram. Can its shape be different? What are the implications if the shape is different?

5. Which of the above OLS models most accurately represent the data? Justify your answer.

Note:

- Create a document containing all your Tables, graphs and conclusions. Number them based on the questions above.
- Save all your created data sets. Name the data sets 3a, 3b, …, 4b
- Place the document and data sets into a directory named Assignment-1
- Zip up the directory and submit it to the assignment submission point in Moodle
- The Test scheduled on Nov-2-2018 will assume you have completed this assignment.