Linear Regression Homework

CMPE 188

Instructor: Jahan Ghofraniha

1. Perform full EDA on Salary data and determine the candidate features for a multiple linear regression model.(hint: you may want to use log transformation on the salary column and use it instead of Salary provided in the data). Use the log salary as the output and the rest of the variables as inputs. (you can ignore the categorical data columns).
2. Build a multiple linear regression model using the RFE and the stepwise methods.
3. For feature selection start with a large number of features and monitor the performance measures. Pick the number of features based on the performance measure when there is a significant change and stop when you do not see a major improvement.
4. Standardize the data by removing the mean and making the standard deviation equal to one (use from sklearn.preprocessing import StandardScaler, , and look at an example on how it is used)
5. Normalize the features by scaling them to a range between 0 and 1. Use the normalize object in scikit learn library to perform normalization on the data. Read the documentation from the preprocessing library documentation and look at the sample code given in the documentation as a guide on how to perform normalization.
6. Perform steps 1-3 for the standardized data and compare the results with the original analysis. What are the differences and how do you interpret the impact of standardization?
7. Perform steps 1-3 for the normalized data and compare the results with the original analysis. What are the differences and how do you interpret the impact of normalization?
8. Include your code, the results and explanation of the results either as a Jupyter notebook file (.ipynb) or a (.py) plus the output/results as comments in your code or as screenshots if it involves plots/graphics and upload to Moodle.