DAV 6150 Module 2 Assignment

Cross Validation

We've learned how the concept of cross validation can be applied during model training to assess the performance of a model when it is applied to previously unseen data. For this assignment your primary task is to construct a cross validated linear regression model that predicts the energy production of a power plant. The data set you will be using is sourced from the UC Irvine machine learning archive:

- https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant#

The data set is comprised nearly 10,000 observations of 1 response/dependent variable (**net hourly electrical energy output)** and 4 explanatory/independent variables (temperature, ambient pressure, relative humidity, and exhaust vacuum). Please refer to the UCI web page for further details on these variables.

Once you are comfortable in your understanding of the various data attributes, get started on the assignment as follows:

1) Load the provided M2_Data.csv file to your DAV 6150 Github Repository.
2) Then, using a Jupyter Notebook, read the data set from your Github repository and load it into a Pandas dataframe.
3) Using your Python skills, perform some basic exploratory data analysis (EDA) to ensure you understand the nature of each of the variables (including the response variable). Your EDA writeup should include any insights you are able to derive from your statistical analysis of the attributes and the accompanying exploratory graphics you create (e.g., bar plots, box plots, histograms, line plots, etc.). You should also try to identify some preliminary predictive inferences, e.g., do any of the explanatory variables appear to be relatively more "predictive" of the response variable? There are a variety of ways you can potentially identify such relationships between the explanatory variables and the response variable. It is up to you as the data science practitioner to decide how you go about your EDA, including selecting appropriate statistical metrics to be calculated + which types of exploratory graphics to make use of. Your goal should be to provide an EDA that is thorough and succinct without it being so detailed that a reader will lose interest in it.
4) Using your Python skills, construct at least two different linear regression models that predict **net hourly electrical energy output** based on the provided explanatory variables and evaluate them using K-fold cross validation. Each of your models must include at least 2 explanatory variables. The explanatory variables you choose to include within each of your models should reflect some of the knowledge you have gained via your EDA work. The value(s) you choose for 'K' for your cross validation are up to you as a data science practitioner to select. After executing your K-fold cross validation for each of you models, select your preferred model based on the average accuracy scores you derived via the K-fold process.

Your first deliverable for this assignment is your Jupyter Notebook. It should contain a combination of Python code cells and explanatory narratives contained within properly formatted Markdown cells. The Notebook should contain (at a minimum) the following sections (including the relevant Python code for each section):

1) **Introduction (5 Points)**: Summarize the problem + explain the steps you plan to take to address the problem

2) **Exploratory Data Analysis (35 Points)**: Explain + present your EDA work including any conclusions you draw from your analysis including any preliminary predictive inferences. This section should include any Python code used for the EDA.

3) **Regression Model Construction & Evaluation (45 Points)**: Explain + present your linear regression models. This section should include an explanation of how you decided upon the explanatory variables to include in the models, how you implemented your K-fold cross validation, and any Python code used for construction + evaluation of the regression models.

4) **Conclusions (5 Points)**

Your Jupyter Notebook deliverable should be similar to that of a publication-quality / professional caliber document and should include clearly labeled graphics, high-quality formatting, clearly defined section and sub-section headers, and be free of spelling and grammar errors. Furthermore, your Pythion code should include succinct explanatory comments.

Upload your Jupyter Notebook to your online DAV6150 GitHub directory. Be sure to save your Notebook using the following nomenclature: **first initial_last name_M2_assn**" (e.g., J_Smith_M2_assn_). Then submit the resulting web link via Canvas within the Module 2 Assignment page.

Your second deliverable for this assignment (10 Points) is a short (approx. 5 minute) video presentation of your work. Your presentation should include a brief overview of your EDA findings, a high-level explanation of your regression models, and a summary of your cross validation findings + which model you prefer. Note that you do not need to appear on camera.

We recommend using Screencast-o-matic for its free cost (recordings up to 15 minutes), ease of use (including basic editing) and ability to save your recording as a link.

When complete, submit the link to your video presentation along with your GitHub link.