DAV 6150 Module 4 Assignment

Feature Selection & Dimensionality Reduction

***** You may work in small groups of no more than three (3) people for this Assignment *****

When the number of explanatory variables is relatively large with respect to the number of observations contained within a data set, data science practitioners need to know how to effectively reduce the number of explanatory variables required for the intended model. For this assignment your primary task is to apply feature selection and/or dimensionality reduction techniques to identify the explanatory variables to be included within a linear regression model that predicts **the number of times an online news article will be shared**. The data set you will be using is sourced from the UC Irvine machine learning archive:

- https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

The data set is comprised of 39,797 observations and 61 attributes. Please refer to the UCI web page for further details on these variables. The **shares** variable will serve as the response variable for your regression model. As such, you are to apply your feature selection / dimensionality reduction expertise to the remaining 60 attributes for purposes of identifying the explanatory variables that you believe will be most useful when included in a linear regression model that estimates **shares**.

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

1) Load the provided M4_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, apply your knowledge of feature selection and dimensionality reduction to the 60 candidate explanatory variables to identify variables that you believe will prove to be relatively useful within the required linear regression model. Your work here should reflect some of the knowledge you have gained via your EDA work. While selecting your features, be sure to consider the tradeoff between model performance and model simplification, e.g., if you are reducing the complexity of your model, are you sacrificing too much in the way of Adjusted R^2 (or some other performance measure)? The ways in which you implement your feature selection and/or dimensionality reduction decisions are up to you as a data science practitioner to determine: will you use filtering methods?

PCA? Stepwise search? etc. It is up to you to decide upon your own preferred approach. Be sure to include an explanatory narrative that justifies your decision making process.

5) Train/cross validate your model and report on its performance.

Your 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 (30 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) **Feature Selection / Dimensionality Reduction (45 Points)**: Explain + present your feature selection / dimensionality work, including any Python code used as part of that process.
4) **Regression Model Evaluation (15)**: Explain + present your linear regression model and discuss its accuracy. This section should include any Python code used to construction + evaluate your regression model.
5) **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 within the provided M4 Assignment Canvas submission portal. Be sure to save your Notebook using the following nomenclature: **first initial_last name_M4_assn**" (e.g., J_Smith_M4_assn_). ***Small groups should identity all group members at the start of the Jupyter Notebook and each team member should submit their own copy of the team's work within Canvas.***