DAV 6150 Project 1 (Module 6)

Regression for Numeric Data

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

For your first Project of the course you will be building upon the work you completed for the Module 3 Assignment wherein you were tasked with cleaning a data set that contained a variety of data quality defects. As you will recall, the data set you were using for that Assignment is comprised of information on more than 12,700 wines, with many of the attributes being related to the chemical composition of the wines. A review of the data attributes is provided below:

Data Set Attribute	Description
INDEX	Unique ID
TARGET	Response Variable (indicates # of cases of wine sold)
AcidIndex	Measures total acidity of wine via a weighted average
Alcohol	Alcohol Content
Chlorides	Chloride content of the wine
CitricAcid	Citric Acid content of the wine
Density	Density of the wine
FixedAcidity	FixedAcidity of the wine
FreeSulfurDioxide	Sulfur Dioxide content of the wine
LabelAppeal	Subjective marketing score that indicates the appeal of the design of the label on the bottle
ResidualSugar	Residual sugar content of the wine
STARS	Wine rating as determined by experts (4 = excellent; 1 = Poor
Sulphates	Sulfate content of the wine
TotalSulfurDioxide	Total sulfur dioxide content of the wine
VolatileAcidity	Volatile acid content of the wine
pH	pH of the wine

Your objective for Project 1 is to apply the full data science project lifecycle to the implementation of a regression model. Your work should include EDA, data preparation (including transforms as needed), feature selection, and a thorough evaluation of model performance metrics. The response variable you will be modeling is the data set's "**TARGET**" attribute, which represents the number of cases of wine that were purchased by wine distributors subsequent to their sampling each of the wines. Sample cases of wine are used to provide tasting samples to wine shops and restaurants throughout the USA. You've been tasked by a large wine producer with the development of a model that can predict the number of wine cases ordered by distributors based on the various characteristics of the many wines represented in the data set. The wine producer is interested in understanding ways in which their own wine offerings can be adjusted to maximize wine sales.

Therefore, your task is to construct and compare/contrast a series of regression models (after completing the necessary EDA and data prep work) that predict the number of wine cases sold relative to certain properties/characteristics of the wine. It is up to you as the data science practitioner to determine which features should be included in these models. To get started on the Project:

1) Load the provided Project1_Eval.csv file to your DAV 6150 Github Repository. You will be using this data set to assess the effectiveness of your regression models.

2) Using a Jupyter Notebook, read the **M3_Data.csv** data set from your Github repository and load it into a Pandas dataframe. **The M3_Data.csv data set will serve as your model training data set for this Project.**

3) Perform EDA work as necessary (if you already have a high-quality EDA from the Module 3 Assignment, you may incorporate it here. If your M3 Assignment EDA was flawed, you should repeat the EDA work and address any shortfalls identified in you M3 Assignment EDA).

4) Perform any required data preparation work, including any feature engineering adjustments you deem necessary for your work.

5) Apply your knowledge of feature selection and/or dimensionality reduction techniques to identify explanatory variables for inclusion within your models. You may select the features manually via the application of domain knowledge, use forward or backward selection, or use a different feature selection method (e.g., decision trees, etc.). It is up to you as the data science practitioner to decide upon the most appropriate feature selection and/or dimensionality reduction techniques to be used with the data set.

6) Using the M3_Data.csv data set, construct at least two different Poisson regression models, at least two different negative binomial regression models, and at least two multiple linear regression models, using different explanatory variables (or the same variables if they have been transformed via different transformation methods). At times, Poisson and negative binomial models can produce identical results. Be sure to comment on that if it happens.

7) After training your various models, decide how you will select the "best" regression model from those you have constructed. For example, are you willing to select a model with slightly lower performance if it is easier to interpret or less complicated to implement? What metrics will you use to compare/contrast your models? Evaluate the performance of your models via cross validation using the training data set. Then apply your preferred model to the evaluation data set and assess how well it performs on that previously unseen data.

Your first deliverable for this Project 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 (10 Points)**: Explain + present your EDA work including any conclusions you draw from your analysis regarding the integrity + usability of the data in its raw state. This section should include any Python code used for the EDA

3) **Data Preparation (10 Points)**: Describe + show the steps you have taken to address the data integrity + usability issues you identified in your EDA, including any feature engineering you have applied to the data set. This section should include any Python code used for Data Preparation.

4) **Prepped Data Review (5 Points)**: Explain + present your post-Data Prep EDA analysis. This section should include any Python code used for re-running your EDA on the variables adjusted during your Data Preparation work.

5) **Regression Modeling (40 Points)**: Explain + present your regression modeling work, including your interpretation of the coefficients your models are generating. Do they make sense intuitively? If so, why? If not, why not? Comment on the magnitude and direction of the coefficients + whether they are similar from model to model, e.g., you might say something like "the fixed acidity variable had a noticeable positive effect in my Poisson model but had a minor negative effect in my linear regression model", etc.

6) **Select Models (15 Points)**: Explain how you selected your model selection criteria. Identify your preferred model. Discuss why you've selected that specific model as your preferred model. Apply your preferred model to the evaluation data set (Project1_Eval.csv) and discuss your results. Did your preferred model perform as well as expected?

7) **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_Project1**" (e.g., J_Smith_Project1_). Then submit the resulting web link via Canvas within the Project 1 Canvas page. ***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.***

Your second deliverable for this Project (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 data preparation + feature selection process + regression models, a summary of your model selection process, an explanation of why you chose your preferred model, and comments on the performance of your preferred model when applied to the evaluation data set. 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.