DAV 6150 Project 3 (Module 12)

Gradient Descent + Gradient Boosting

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

Gradient descent algorithms lie at the heart of a wide variety of machine learning models, and a variety of enhanced gradient descent algorithms are available for our use for both classification and regression problems. One nagging question we have yet to properly address is: just how well do algorithms that are based on gradient descent concepts perform relative to both each other and other types of models? This assignment provides you with an opportunity to gauge the effectiveness of gradient descent-based models firsthand. Your task for **Project 3** is to construct a series of different models for a provided data set and compare/contrast the performance of the varying models against one another. Specifically, you will be constructing a decision tree, a random forest, a gradient boosting classifier, a stochastic gradient descent classifier, and an XG Boost classifier. The data set you will be using for this Project is a well-known set of attributes that describe both the physical characteristics and prices of nearly 54,000 diamonds. A description of the attributes represented within the data set can be found here:

- https://ggplot2.tidyverse.org/reference/diamonds.html.

This data set has been widely used for purposes of demonstrating machine learning algorithms that predict the price of a diamond. However, for this assignment the **cut** attribute (a categorical variable) will serve as the response variable for your models. As such, your machine learning models should be designed for purposes of predicting which of the five **cut** values is most likely to apply to a given observation.

Get started on the Assignment as follows:

1) Load the provided **Project3_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. Ensure your data attributes are properly labeled within the data frame.
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 provided explanatory variables to identify variables that you believe will prove to be relatively useful within your models. 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 accuracy (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) After splitting the data into training and testing subsets, use the training subset to construct each of the following models:
    - Decision Tree
    - Random Forest
    - Gradient Boosting Classifier
    - Stochastic Gradient Descent Classifier
    - XG Boost Classifier

Your models must each include at least four (4) explanatory variables. Be sure to make use of the same four explanatory variables for each of the models. This will allow you to compare the performance of the various models in a much more effective and understandable manner.

6) After training your various models, decide how you will select the "best" classification 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 testing subset 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 (15 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) **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 techniques 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) **Classifier Modeling (35 Points)**: Explain + present your classifier modeling work, including your feature selection / dimensionality reduction decisions and the process by which you selected the hyperparameters for your models. This section should include any Python code used for feature selection, dimensionality reduction, and model building.
6) **Select Models (15 Points)**: Explain your model selection criteria. Identify your preferred model. Compare / contrast its performance with that of your other models. Discuss why you've selected that

specific model as your preferred model. Apply your preferred model to the testing subset and discuss your results. Did your preferred model perform as well as expected? Be sure include any Python code used as part of your model selection work and to frame your discussion within the context of the classification performance metrics you have derived from the models.

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_Project3**" (e.g., J_Smith_Project3). Then submit the resulting web link via Canvas within the Project 3 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. 10 minute) video presentation of your work. Your presentation should include a brief overview of your EDA work, a high-level explanation of your data preparation + feature selection process, a discussion of your models including the hyperparameter values you selected for each, 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 testing 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 10 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.