Project 3: Regression

Total Points: 100
Given Out: Week 7: Friday, Oct 4th, 2024
Due Date: Week 9: Friday, Oct 25th, 2024
Project is to be done in pairs.

Background & Introduction: In the vast landscape of machine learning and data mining, regression is a central and continuously explored task. Fundamentally, regression focuses on predicting a continuous value based on independent variables, rather than sorting data into categories like classification. Think of it as predicting a numeric score or outcome based on various influencing factors or features.

For this project, we're not just aiming to understand the basics of regression. We want to go deeper, looking at every step from selecting and preparing the data to training and evaluating the model. By working through this process, we'll get a hands-on understanding of how regression works. Through this project, our goal is to turn our theoretical knowledge into practical skills and get a clearer insight into the task of regression.

Keypoints: Just like previous projects, we would like to emphasize the importance of creativity in the project. Most of the questions are open-ended. The more innovative your explorations, the higher the scores you stand to achieve. Dive deep, and let your analytical creativity lead the way!

Datasets: **You will need to use a different type of data set than the one you chose in Project 2.**
For example, if you used tabular data in project 2, then we ask you to use other types of data such as image or text in project 3. You're encouraged to find datasets that are interesting and challenging to you! Make sure the data you choose is suited for regression, try to think of a regression task you are interested in solving over this dataset.
Dataset requirements: If you choose tabular data, it should contain at least 20 different features of different types and at least 200,000 records (rows). No particular requirements for textual/image/other data.

Submission and Deliverables: You will submit the project via Canvas. The submission includes two separate files (DO NOT compress your files to .zip): (1) A report that provides the tasks detailed below accompanied by explanations and (2) execution source codes that can reproduce the reported results.

1. Your report doc can be a .pdf file or .docx file. The file needs to contain the results (e.g., figures, tables) of each task and your analysis. If your answer contains figures/images, please take a screenshot or download them, and then insert them into your results doc.
2. Your code can be .py file or .ipynb file.

3. **Rename the files using your names** (student1's first name_student1's last name_student2's first name_student2's last name)**.** Eg., roee_shraga_yao_su.pdf roee_shraga_yao_su.ipynb

Using Generative-AI (e.g., ChatGPT): I encourage responsible and sensible use in assignment. Please report if, how, and what you used to complete the assignment. Explain how you validated the trustworthiness of the solution, which prompt/s you used,and you used the output of the model (basic code/documentation/etc.)

Demonstration: Once completed, several students may be asked to provide a brief demonstration of their results to your classmates. The instructor/TA may also request a one-on-one demonstration to support grading.

Taks 1 – BYOD (Bring Your Own Data) [10 points]

- Craft a compelling introduction for your project. Describe the dataset that you are going to use, introduce the background, explain its significance, and share your motivation for choosing it.
- Highlight the potential impact of a successful solution and anticipate any challenges ahead. This introduction should encapsulate the essence of your project, drawing readers in with clarity and conviction, while setting the stage for the details to follow.

Task 2 – EDA [10 points]

Perform exploratory analysis on the data. Research online for ideas, and then show analysis on **at least five** different aspects of the dataset. Analyze your findings.
Note: This part is open-ended, any valid analysis is fine as long as it useful for you to understand the data better! Use visualizations when necessary.

Task 3 – Problem Definition [10 points]

Based on Task 2, revisit Task 1 to write a compelling introduction for your project that also highlights the problem that you are going to solve and some observations to support the choice, motivation and significance based on your EDA.

Task 4 – Preprocessing [10 points]

Perform data processing on the data. Dive deep into data preparation, ensuring that the dataset is primed and ready for model training. Document all preprocessing steps and include them in the report, specify what happened to the data after the step, why you performed it and if you are going to use it moving forward.
For example:

1. **Data Cleaning:** Address any inconsistencies within the data. This includes dealing with missing values, anomalies, and duplicated entries.
2. **Data Transformation & Normalization:** Depending on the nature and distribution of your data, apply necessary transformations. Ensure features are on a similar scale, making them more amenable to analysis and model training.
3. **Specific Data Type Processing:**

a. Text Data: Implement tokenization to break down text into smaller chunks, remove stop words to filter out common but uninformative words, and utilize feature extraction techniques, like TF-IDF or word embeddings, to represent text in a form suitable for machine learning.
b. Image Data: Normalize image pixel values, ensuring they fall within a consistent range (typically 0 to 1). Consider image augmentation techniques to artificially enhance your dataset, creating varied representations of the same image to improve model robustness.

Task 5 – Model Selection, Training, and Optimization [20 points]

1. Identify and train **at least five** distinct machine learning models apt for your regression task.
2. try to boost the performance of the models using suitable techniques. For example, feature engineering, cross-validation, grid search, tuning model hyperparameters, etc. Try **at least three** techniques and compare the performance with the vanilla ones.

Task 6 – Model Evaluation [20 points]

1. Use **at least four** metrics to evaluate the performance of your selected models.
2. Analyze the experimental results and report your conclusion. The analysis should compare results (using the metrics you chose above) the different pre-processing steps (Task 4), the different models and optimization steps (Task 5) in a table/graph.
3. Report your observations discuss their implications (e.g., Model X with Optimization Y after preprocessing step Z was the most/least effective because ABC...).
4. In your report, you should refer explicitly to the nature of analysis each evaluation measure provides and the benefits from using it.

Task 7 – Explainability [20 points]

For this task, you will focus on making your regression model interpretable and understandable. In the world of machine learning, explainability is key for building trust in the model and understanding its decision-making process.

1. Use **at least two** techniques for explanations.
2. Explain Model Predictions: Use appropriate tools and techniques (e.g., SHAP, LIME, etc) to explain how your model is making predictions. Analyze the importance of each feature in influencing the model’s output. Highlight the key drivers behind the predictions.

3. Explore Feature Importance: Evaluate and rank the features based on their contribution to the prediction task. Provide a discussion on how different features impact the model’s performance and predictions. Focus on understanding the relationship between the features and the target variable.
4. Present Results in an Intuitive Way: Visualize the model's explainability using graphs and charts that clearly communicate the findings.
Some good examples for you: features_importance, LIME - Local Interpretable Model-Agnostic Explanations, An introduction to explainable AI with Shapley values.