Final Project Manual and Instructions

Introduction

As part of this course, students will be required to complete a regression analysis project. The objective of this project is to apply the regression models covered during the semester, evaluate their performance using appropriate evaluation metrics, and implement cross-validation to validate the model's robustness. Additionally, you will be required to properly handle train-test splits and perform necessary data preprocessing.

Dataset Selection

For this project, there is a **'data'** folder that contains **Eight different datasets** you can choose from. Additionally, there is a **'data_description.pdf'** file that provides information about each dataset, including its characteristics, variables, and potential use cases.

You are required to **choose (only) one of the datasets** from the 'data' folder for your regression analysis project. Review the descriptions in the 'data_description.pdf' file to understand the available datasets and choose the one that best fits the goals and requirements of your project.

Some key considerations when selecting the dataset:

1. Ensure that the dataset is appropriate for the application of the regression models covered in the course.
2. Choose a dataset that aligns with your interests and the insights you hope to gain from the analysis.
3. Consider the size and complexity of the dataset, as well as the number and types of variables, to ensure it is manageable for the scope of this project.
4. Evaluate the data quality, including the presence of missing values, outliers, or other potential issues that may require preprocessing.
5. Once you have selected the dataset, you should include it as part of your project deliverables, or provide clear instructions on how to access the dataset.

Remember, the dataset selection is an important step in the regression final project, as it will determine the focus and scope of your analysis. Choose wisely to ensure a successful and meaningful regression analysis.

Project Requirements

1. Jupyter Notebook File

In this file, you will implement the regression analysis using Python.

You should utilize the regression models taught during the semester, including multiple linear regression, ridge regression, and any other relevant techniques covered in the course.

1) Perform necessary **data preprocessing**, such as handling missing values, feature engineering, and scaling, as required by the **chosen dataset**
2) incorporating data visualization is a critical requirement for this regression final project. You should utilize a variety of visualization methods that have been covered during the semester and are applicable to your chosen dataset and regression analysis.
3) Split the data into training and testing sets appropriately.
4) Train the regression models using the training data and evaluate their performance using suitable evaluation metrics (e.g., RMSE, R-squared).
5) Implement cross-validation to assess the robustness of the models.
6) Use appropriate chart types, such as scatter plots, line plots, or bar charts, to visualize the relationships between variables.
7) Ensure that the charts are visually appealing, with thoughtful use of colors, font sizes, legends, and labels.

Document your code and provide explanations for the steps taken.

Some guidelines for the data visualizations:

1. Use clear and readable font sizes for labels, axis titles, and legends.
2. Choose a color palette that is visually appealing and helps distinguish different data elements.
3. Include informative legends and labels to make the charts self-explanatory.
4. Align and format the visualizations in a consistent manner throughout the project.
5. Consider using appropriate chart types based on the nature of your data and the insights you want to communicate.

2. Project Report (PDF file)

The regression final project should be presented in the form of a formal academic paper, akin to a journal or conference article. The project report should be structured with the following sections:

1. **Table of Contents**
2. **Table of Figures**
3. **Introduction**
   - Provide an overview of the problem statement and the objectives of your regression analysis project.
4. **Literature Review**
   - Highlight previous studies or findings that inform your approach to the current project.
5. **Data Description**
   - Discuss the key characteristics of the data, such as the data types, scales, and potential issues.
6. **Data Preprocessing**
   - Explain the steps taken to clean, transform, and prepare the data for analysis.
   - Address any data quality issues, such as missing values, outliers, or feature engineering.
7. **Exploratory Data Analysis (Data Visualization)**
   - Conduct a thorough exploratory data analysis, generating statistical summaries and visualizations to gain insights into the data.
8. **Regression Analysis**
   - Implement the regression models taught during the semester, including multiple linear regression, ridge regression, and any other relevant techniques.
   - Explain the rationale for selecting each regression model and the steps taken to train and evaluate them.
   - Assess the performance of the models using appropriate evaluation metrics, such as RMSE and R-squared.
   - Discuss the implementation of cross-validation to validate the robustness of the models.

9. **Results**
    - Present the predictions and performance metrics of the regression models.
    - Analyze and interpret the results, highlighting the strengths and limitations of the different models.
10. **Conclusion**
    - Summarize the key findings and insights gained from your regression analysis project.
11. **Bibliography**
    - Include all relevant citations and references used in your project report.

Good luck with your ISE 529 Final Project!