Close
Instructions
Overview

Final Project: Model Selection and Evaluation in Statistical Learning

Objective

As the culmination of our statistical learning class, you are tasked with a comprehensive project that allows you to apply the techniques learned throughout the course. Your challenge is to select a real-world dataset, perform exploratory data analysis, model the data using various statistical learning techniques, and identify the best performing model based on rigorous evaluation criteria.

Project Guidelines

. Dataset Selection:

. Choose a dataset that interests you. This can be from any domain, such as finance, healthcare, marketing, or environmental science.
. Ensure the dataset is complex enough to allow for meaningful analysis and modeling (e.g., a mix of categorical and numerical variables, a sizable number of observations).
. Data Exploration and Preprocessing:

. Conduct thorough exploratory data analysis to understand the characteristics and distributions of your data.
. Perform any necessary data cleaning and preprocessing steps, such as handling missing values, feature scaling, and encoding categorical variables.
. Modeling:

. Implement several statistical learning techniques discussed in class. This may include, but is not limited to, linear regression, logistic regression, decision trees, and support vector machines.
. For each model, explain your rationale for choosing it and how it fits with the nature of your data.
. Model Evaluation and Selection:

. Compare the models using appropriate evaluation metrics (e.g., accuracy, RMSE, AUC).
. Utilize a model validation technique such as validation sets, k-fold cross-validation, or bootstrapping to assess model performance and select hyperparameters.
. Discuss the strengths and limitations of each model and the reasons behind your choice of the best model.
. Conclusions:

. Draw conclusions based on your model evaluations. Discuss any insights gained from the data and any implications of your findings.
. Reflect on the performance of the models and any challenges you encountered during the project.

Deliverables

. Project Report:

. Provide a comprehensive write-up of your project. Your report should include:

. A description of the dataset and your exploratory data analysis findings.

. Details of the preprocessing steps undertaken.

. An explanation of each model implemented, including rationale and evaluation.

. A comparison of the models and justification for the model you determined to be the best.
. Conclusions drawn from your analysis.

. The report should be clear, concise, and well-structured. Technical details should be accurately explained but also accessible to readers not familiar with the specific techniques.
. If this project is coded in Jupyter Notebook, then the report should appear in that file as well. If it is coded as a .py file, the report should be written in word or similar editor and submitted as either a word document or a pdf.

. Code:

. Submit all code used for the project. Your code should be well-documented, with comments explaining key steps and decisions.

Evaluation Criteria
Data Analysis: Depth and thoroughness of the exploratory data analysis.

Model Implementation and Rationale: Correct implementation of various statistical learning techniques and the rationale for their use.
Model Evaluation and Selection: Appropriateness and thoroughness of model evaluation and selection process.
Report Quality: Clarity, structure, and completeness of the project report.

This project is an opportunity to demonstrate your understanding of statistical learning concepts and your ability to apply these methods to real-world data. We are excited to see your innovative approaches and analytical skills in action!