Hw4 Discussion of Visualization and Process

Ruijing Chen

1. Essential Questions

The dataset used in this Shiny app is from the Spring 2023 STAT course, provided by the instructor as part of a project. The visualizations aim to answer two key questions: What are the most important features influencing the prediction of final grades in the dataset? How do selected features individually affect predicted grades, as visualized through Partial Dependence Plots?

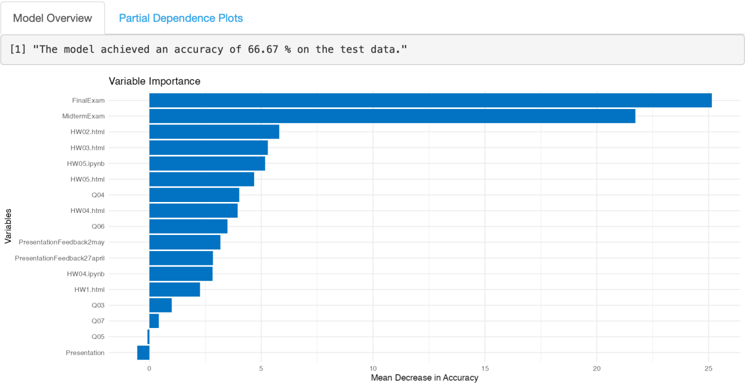
2. Design Choices and Tradeoffs

To address these questions, the app integrates several interactive elements: The Variable Importance Plot highlights the key features influencing prediction accuracy. This design allows users to easily interpret feature relevance. Partial Dependence Plots (PDPs) help users understand the relationship between individual features and predicted outcomes.

Tradeoffs include: Simplifying PDPs to focus on one variable at a time (univariate plots) to ensure interpretability for nontechnical users, though this limits insights into interactions between variables. Using a static grid resolution of 50 points for PDPs to balance computation time and plot granularity.

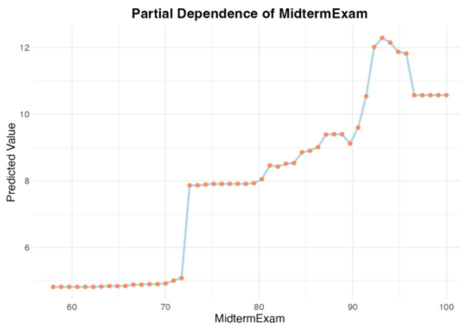
3. Key Findings

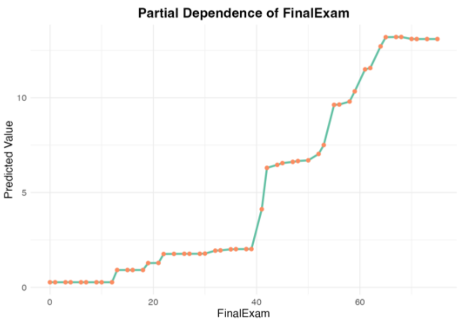
The model has the highest accuracy (66.67%) and uses the fewest variables when the variables in the chart below are selected.



```
[1] "The model achieved an accuracy of 66.67 % on the test data."
```

The app reveals that MidtermExam and FinalExam consistently show high importance in the model, indicating a strong influence on grade prediction. PDPs visually confirm trends in the data, where variables such as exam scores positively correlate with final grades. These findings align with prior expectations about the dataset while also providing new insights into feature interactions and their predictive power.





4. Creation Process

Data Preparation: The dataset was read and preprocessed to ensure the target variable ("Grade") and potential predictors were correctly formatted. Factorization of "Grade" was necessary for Random Forest modeling. Selected features were automatically extracted based on the dataset structure, with options dynamically updated in the user interface.

5. Visualization Development:

The Variable Importance Plot was created using the `ggplot2` package to display the Mean Decrease in Accuracy for each variable. PDPs were generated using the `pdp` package, which calculates predictions across a range of values for the selected feature. Interactive elements such as dropdown menus allow users to control which feature to visualize. This approach ensures that the visualizations are dynamically generated based on user inputs and model outputs, making the app adaptable to different datasets and exploration needs.

Github Link: https://github.com/18930415187/stat436-hw4