ChatGPT Review: SEMMA

Here's a detailed review of your notebook based on each SEMMA phase:

1. Sample Phase

- **Strengths**:
  - You properly loaded the dataset and reviewed the basic information, including data types and summary statistics.
  - Your approach to handling missing values was correct, as you examined and dropped/handled missing data appropriately.
- **Critique**:
  - Consider using stratified sampling if the target variable has imbalanced classes. This would ensure a better representation in smaller samples.
  - It would be beneficial to add exploratory insights about the sample distribution to confirm its representativeness visually.

2. Explore Phase

- **Strengths**:
  - Your visualizations covered the distributions of numerical features and relationships between variables effectively.
  - The correlation heat map was well-constructed, providing clear insights into feature relationships.
- **Critique**:
  - For categorical variables, more detailed plots (e.g., bar plots of counts) could help understand distributions better.
  - Residual and scatter plots for relationships between features and the target variable could further uncover important patterns or potential linear relationships.

3. Modify Phase

- **Strengths**:
  - You correctly handled encoding, scaling, and feature engineering, which improved data quality for modeling.
  - Feature selection methods like RFE were well-implemented.
- **Critique**:
  - When imputing missing values, provide rationale for the choice of imputation strategy, as this can affect model performance.
  - Some interactions or polynomial features might have improved model performance; consider adding such transformations.

4. Model Phase

- **Strengths**:
  - You used a variety of models to find the best performer, which is a solid strategy.
  - Hyperparameter tuning was conducted effectively, using GridSearchCV to optimize model performance.
- **Critique**:
  - Adding additional models like Gradient Boosting or Support Vector Regression could provide more comprehensive comparisons.
  - You could also include metrics like RMSE or Adjusted R^2 for regression models to better capture model performance.

5. Assess Phase

- **Strengths**:
  - You included evaluation metrics (MSE, MAE, and R^2) and visualizations like residual plots and actual vs. predicted plots, which provided a good performance overview.
  - Cross-validation was performed, which validated model consistency.
- **Critique**:
  - Consider adding a more in-depth error analysis, such as identifying specific areas (e.g., under/overestimation patterns) where the model struggles.
  - Include a brief explanation of findings and recommendations based on the assessment phase, highlighting potential deployment or model refinement steps.

Additional Comments

- Overall, your notebook is well-structured and follows the SEMMA methodology thoroughly.
- Improve documentation within the code cells by adding comments to explain why certain methods or techniques were chosen.
- Ensure consistency in data pre-processing steps (e.g., scaling and encoding) between training and testing datasets to avoid data leakage.