2_Model_Selection_big

January 28, 2025

1 Exercise. Model Selection (part 2: big dataset)

Exercise 1: Building the Best Linear Model

This exercise involves working with a dataset called `model_selection_big.csv`. The dataset consists of 10 variables (features) and one independent variable (target). The goal is to explore the data and build the best possible linear model to predict the target variable. Follow the structured steps below to complete this task.

1.0.1 Task Overview:

1. **Data Exploration**
   - Analyze correlations between variables.
   - Perform descriptive statistics for each feature and the target variable.
   - Identify potential outliers using statistical methods or visualizations.
2. **Data Splitting**
   - Divide the data into training, validation, and test sets.
   - Use the training set to train models, the validation set to tune and compare models, and the test set only for final model evaluation.
   - Ensure the splitting process is randomized and stratified if necessary.
3. **Baseline Model Creation**
   - Create a baseline model using the simplest linear regression algorithm.
   - Compute metrics on the validation set, including Mean Squared Error (MSE), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and R-squared (R^2).
4. **Model Exploration with Quadratic Terms**
   - Investigate potential non-linearities in the data by including quadratic terms in the models.
   - Systematically test various model configurations by including or excluding certain variables, quadratic terms, and interactions.
   - Compare models using well-defined criteria, such as the lowest MSE, AIC, or BIC.
5. **Cross-Validation (Optional)**
   - Perform cross-validation to validate the robustness of the selected models.
   - Report metrics across folds to assess consistency and reliability.
6. **Final Model Selection and Testing**

- Choose the best-performing model based on validation metrics and cross-validation results.
- Evaluate the final model using the test set and report its performance metrics.

1.0.2 Expected Deliverables:

- A clear and concise report detailing each step of the process.
- Findings from data exploration.
- Splitting methodology and the resulting data partitions.
- Results and interpretation of the baseline model.
- Summary of the model exploration process and selected models.
- Cross-validation results, if performed.
- Final model performance on the test set.
- Well-commented code implementing the above steps.

1.0.3 Notes:

- Ensure all assumptions of linear regression are checked and reported, such as normality of residuals and homoscedasticity.
- Clearly justify decisions made at each step, such as the choice of metrics and model selection criteria.
- Use appropriate visualizations to support your analysis.

[ ]: