Fossil Age Prediction Using Linear Regression

Problem Statement:

Predict the age of fossils based on geological, chemical, and physical attributes using Linear Regression.

Dataset Overview:

The dataset consists of 4398 fossil samples with 13 features including:

- **uranium_lead_ratio:** Ratio of uranium to lead isotopes in the fossil sample.
- **carbon_14_ratio:** Ratio of carbon-14 isotopes present in the fossil sample.
- **radioactive_decay_series:** Measurement of the decay series from parent to daughter isotopes.
- **stratigraphic_layer_depth:** Depth of the fossil within the stratigraphic layer, in meters.
- **isotopic_composition:** Proportion of different isotopes within the fossil sample.
- **fossil_size:** Size of the fossil, in centimeters.
- **fossil_weight:** Weight of the fossil, in grams.
- **geological_period:** Geological period during which the fossil was formed.
- **surrounding_rock_type:** Type of rock surrounding the fossil.
- **paleomagnetic_data:** Paleomagnetic orientation data of the fossil site.
- **stratigraphic_position:** Position of the fossil within the stratigraphic column.
- **age:** Calculated age of the fossil based on various features, in years.

Data Pre-Processing:

Firstly, essential libraries like numpy, pandas for data manipulation, and matplotlib, seaborn for visualization were imported. The dataset was imported and copied to ensure original data integrity. Checks for data size, shape, and correctness of column names and types were conducted, with 'inclusion_of_other_fossils' converted to integer format for ease of analysis. No null values or duplicates were found. The dataset was then split into categorical and numeric variables for detailed univariate analysis.

Univariate Analysis:

Categorical Data Analysis

The initial phase of univariate analysis focused on examining categorical data. Using a structured approach, each categorical column was iterated through to identify subclasses and determine their distribution within the dataset. This process provided insights into the

diversity and frequency of subclasses present, essential for understanding the categorical variables' impact on the target variable. shown below.

```
geological_period
Cambrian          882
Triassic          676
Cretaceous        601
Devonian          498
Jurassic          490
Paleogene         405
Permian           365
Neogene           311
Ordovician        100
Carboniferous      52
Silurian           18
Name: geological_period, dtype: int64
paleomagnetic_data
Normal polarity      3160
Reversed polarity    1238
Name: paleomagnetic_data, dtype: int64
surrounding_rock_type
Sandstone        1497
Limestone        1166
Shale            1144
Conglomerate      591
Name: surrounding_rock_type, dtype: int64
stratigraphic_position
Bottom    2667
Middle    1267
Top        464
Name: stratigraphic_position, dtype: int64
```

Numeric Data Analysis

Following the categorical analysis, attention shifted to numeric columns. Utilizing box plots generated through a systematic loop, the distribution and presence of outliers in each numeric attribute were assessed. This step revealed outliers in certain columns, highlighting the need for outlier treatment to enhance the reliability of subsequent analyses and model performance.

Count Plot Analysis

Further exploration involved generating count plots for each categorical column. These plots visualized the spread and frequency of subclasses within each categorical variable. This methodical approach facilitated a deeper understanding of how categorical attributes contribute to the dataset's composition and variability.

Comprehensive Analysis

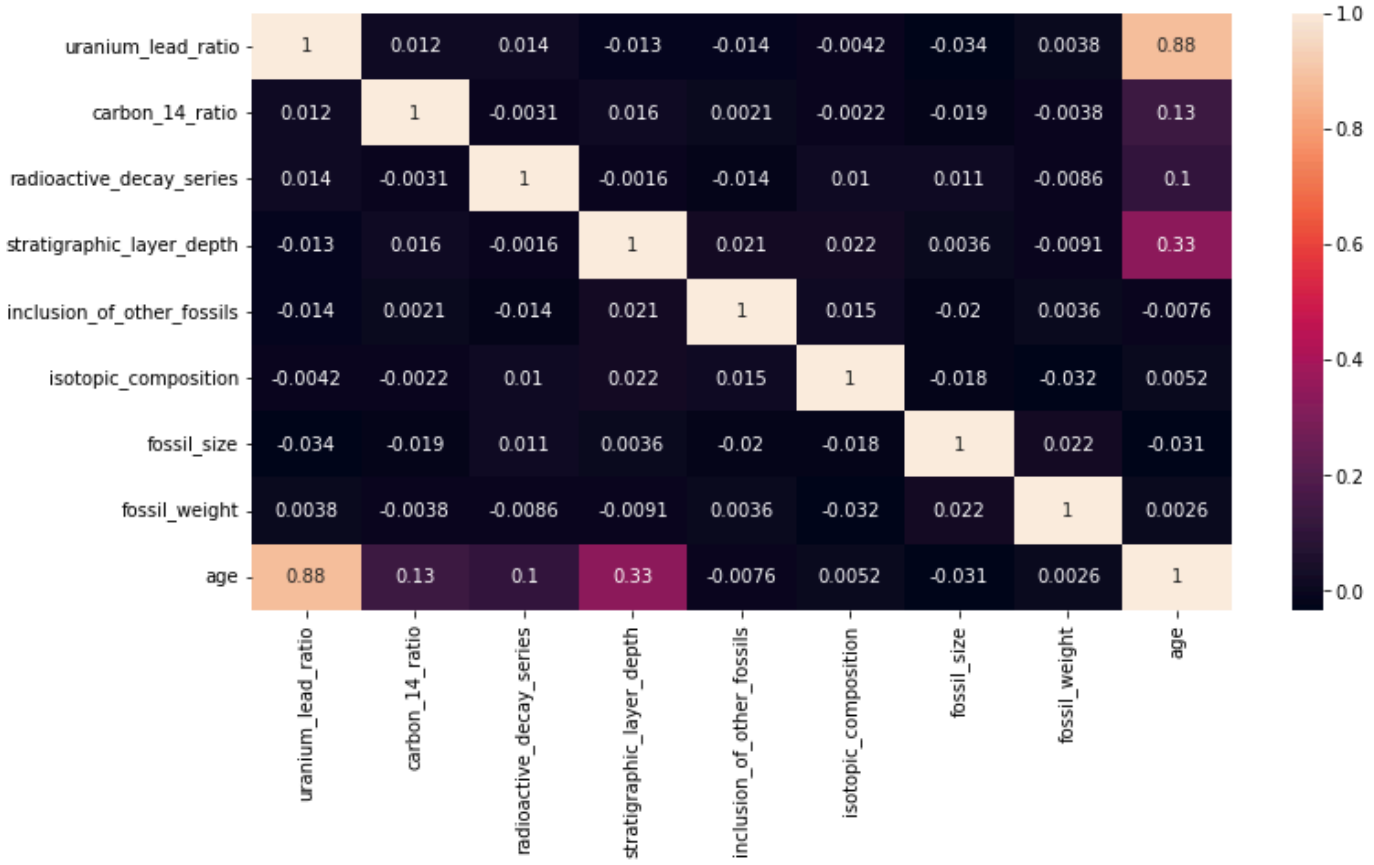
In concluding the univariate analysis phase, each column's characteristics, spread, and value distributions were thoroughly evaluated within the entire dataset context. While detailed visual representations are best explored in the accompanying notebook for clarity and

precision, this summary encapsulates the rigorous process undertaken to prepare the data for subsequent bivariate and multivariate analyses.

Bi-Variate Analysis:

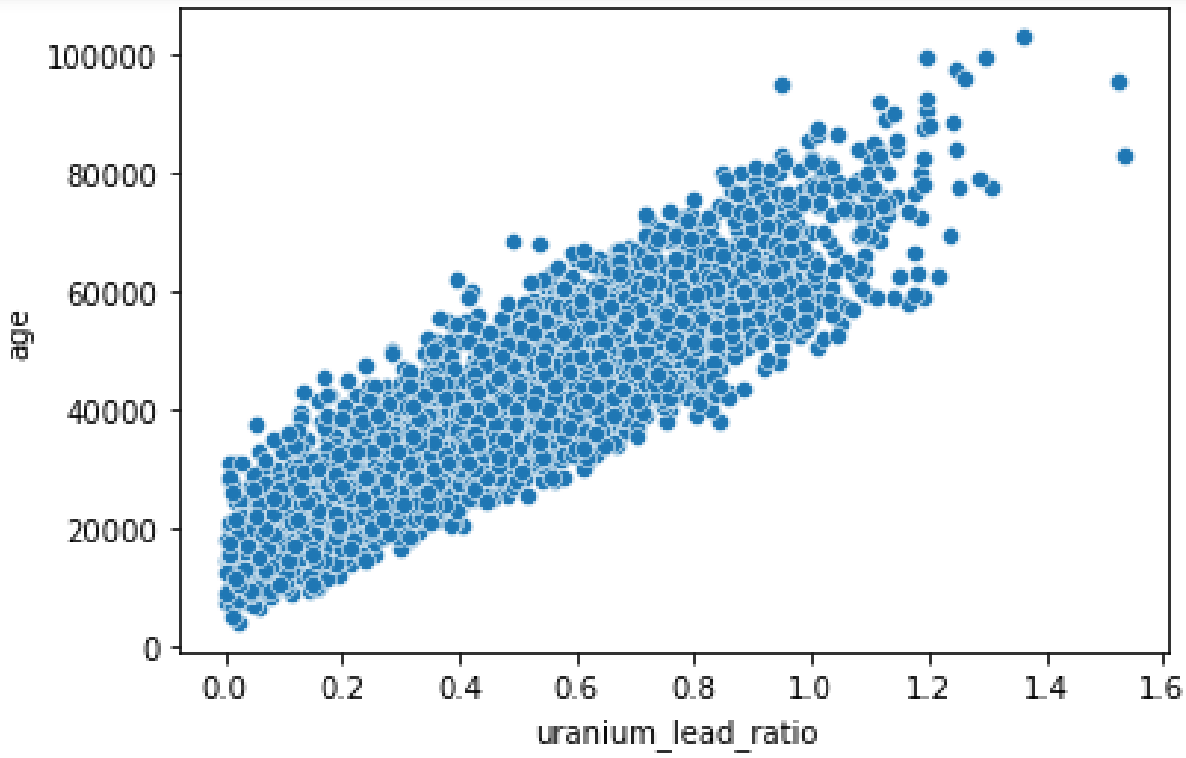
Heatmap of Correlation

The initial step in bivariate analysis involved generating a heatmap to visualize correlations between variables in the dataset. This heatmap provided a comprehensive overview of how each attribute relates to one another, particularly focusing on their impact on the target variable 'age'. Notably, the analysis revealed no significant multicollinearity issues among the predictors. Specifically, the 'uranium_lead_ratio' exhibited a notably strong positive correlation with 'age', suggesting it plays a pivotal role in predicting fossil ages.



Scatter Plots

Subsequent to the correlation analysis, scatter plots were created for each numeric variable plotted against the target variable 'age'. These plots were instrumental in identifying patterns and relationships between numeric predictors and the age of fossils. Notably, the 'uranium_lead_ratio' demonstrated a discernible positive correlation with age, corroborating findings from the correlation heatmap. This visual exploration provided insights into how changes in numeric attributes potentially influence the estimated age of fossils.



Box Plots for Categorical Variables

Further exploration involved generating box plots for each categorical variable against the target variable 'age'. These plots illustrated how different subclasses within each categorical attribute relate to fossil age. Such analysis was crucial in understanding the categorical variables' impact on age estimation, providing a deeper contextual understanding of their contributions to the predictive model.

Base Model Evaluation

Initial Linear Regression Model

The base model employed for initial evaluation was a Linear Regression model without any preprocessing steps such as data scaling or outlier removal. This approach aimed to establish a benchmark performance to assess subsequent improvements. The model yielded a high R-squared score of 97%, indicating a strong fit to the training data. However, the Mean Absolute Percentage Error (MAPE) was noted at 5.59%, suggesting potential overfitting due to the model's high performance on the training set.

Exploration of Data Preprocessing Techniques

To mitigate potential overfitting observed in the base model, various preprocessing techniques were explored. This included feature scaling and outlier removal to enhance the model's generalization capabilities. Despite these efforts, the R-squared score remained consistent or slightly improved, reinforcing the need for regularization techniques.

Introduction of Ridge and Elastic Net Regression

To address overfitting effectively, Ridge and Elastic Net Regression were implemented. These techniques introduce regularization penalties to the model, which help in reducing the complexity and variance of the model. Post application, the models exhibited an R-squared score of 85.89% and MAPE of approximately 13% for Elastic Net Regression. While not achieving perfection, these metrics indicate a balanced performance suitable for the dataset.

Selection of Final Model

After thorough evaluation, the Elastic Net Regression model with scaled input data was chosen as the final model. This decision was based on its ability to maintain strong predictive performance while mitigating overfitting issues observed in earlier stages. The model's performance metrics and stability make it suitable for reliable age prediction of fossils based on the dataset's attributes.

Cross-validation Insights

During cross-validation, R-squared scores consistently returned around 97%, indicating potential overfitting concerns. However, when evaluating using MAPE as the scoring metric, negative values were obtained, posing challenges in interpretation. This discrepancy suggests the need for further investigation into appropriate evaluation metrics that align with the dataset's characteristics and modeling objectives.

Conclusion:

The final document outlines a structured approach to predicting fossil ages using Linear Regression, emphasizing data exploration, model building, and evaluation. Further enhancements could involve refining feature selection or exploring alternative evaluation metrics for improved model interpretability and accuracy.