Question 2 Regression

Title: Analysis of Housing Data using Regression Techniques

Introduction:

The given code snippet is an analysis of housing data using various regression techniques such as Linear Regression, Ridge Regression, Lasso Regression, Polynomial Regression, and Elastic Net Regression. The objective is to build models that can predict the median house value based on various features of the dataset.

Dataset:

The dataset contains information about various factors that influence the median house value. It has columns such as longitude, latitude, housing\_median\_age, total\_rooms, total\_bedrooms, population, households, median\_income, median\_house\_value, and ocean\_proximity.

Data Preparation:

The data is first loaded into a pandas DataFrame and then subjected to data exploration and cleaning. The 'ocean\_proximity' column is one-hot encoded, and missing values in the 'total\_bedrooms' column are filled with the median value.

Data Inspection:

The data is inspected for its shape, missing values, and data types of the columns. The scatter matrix plot is generated to visualize the relationship between the numerical columns and their correlations.

Model Building:

The dataset is split into training and testing sets, and several regression models are built using the training data, including Linear Regression, Ridge Regression, Lasso Regression, Polynomial Regression (degree 2 and 3), and Elastic Net Regression.

Model Evaluation:

The models are evaluated using R² score and Mean Squared Error (MSE). The results are as follows:

1. Linear Regression Model R² Score: 0.xxxxx
2. Ridge Regression Model R² Score: 0.xxxxx
3. Lasso Regression Model R² Score: 0.xxxxx
4. Polynomial Regression Model (Degree 2) R² Score: 0.xxxxx
5. Polynomial Regression Model (Degree 3) R² Score: 0.xxxxx
6. Elastic Net Regression Model R² Score: 0.xxxxx

The coefficients and intercepts for each model are also printed. These values are essential to understand the contribution of each feature in the prediction of the median house value.

Visualization:

The code provides various plots to visualize the performance of each model:

1. Scatter plots for true values vs. predicted values: These plots allow us to compare the actual and predicted median house values, providing a visual representation of the model's performance.
2. Residual plots: These plots show the residuals (difference between true and predicted values) against the true values. This visualization helps in understanding the distribution of errors in the predictions.
3. Regression plots: The code also plots the regression lines using a selected feature ('median\_income') for the Linear, Ridge, and Lasso regression models. This visualization helps in understanding how the model captures the relationship between the selected feature and the target variable.

Conclusion:

The code provided is an extensive analysis of the housing data using various regression techniques. It involves data preparation, model building, evaluation, and visualization. Based on the R² scores and visualizations, we can compare the performance of the different models and select the most appropriate one for predicting the median house value.

n conclusion, regression techniques are valuable tools for predicting numerical outcomes based on a set of input features. The analysis performed in the code snippet demonstrates the process of data preparation, model building, evaluation, and visualization, offering insights into the performance of various regression models on the housing dataset. This can serve as a starting point for further exploration and experimentation to identify the best model for predicting median house values.

Based on the provided results, the following conclusions can be drawn:

1. Linear Regression, Ridge Regression, and Lasso Regression models have similar R² scores, around 0.6388, which indicates that these models can explain approximately 63.88% of the variance in the target variable. This performance is considered moderate.
2. Polynomial Regression models with degrees 2 and 3 have very low and even negative R² scores, indicating that they perform poorly in explaining the variance in the target variable.
3. The Mean Squared Error (MSE) values for Linear, Ridge, and Lasso Regression models are similar, around 4.87 \* 10^9. This implies that the average squared difference between the predicted and actual values is relatively large.
4. Polynomial Regression models have significantly higher MSE values compared to the other models, further supporting their poor performance.
5. The coefficients and intercepts for Linear, Ridge, and Lasso Regression models are quite different, which might affect the model's interpretation and feature importance. Ridge and Lasso Regression models, in particular, are known for their regularization properties that help reduce the impact of less important features and prevent overfitting.
6. Overall, Linear, Ridge, and Lasso Regression models appear to be better choices for this dataset compared to Polynomial Regression models. However, further tuning and evaluation of these models might be required to improve their performance and achieve a higher R² score.

Based on the results you provided for the Elastic Net Regression model, we can now draw some conclusions.

1. Model performance: The R² score of 0.62475 indicates that the Elastic Net Regression model can explain about 62.47% of the variance in the target variable. While this is not an exceptionally high R² value, it does suggest that the model has some predictive power.
2. Interpretation of coefficients: The coefficients you provided are the estimated weights for each feature in your dataset. They represent how each feature impacts the target variable, with positive values indicating a positive relationship and negative values indicating a negative relationship. For example, the first coefficient of -1.71370589e+04 suggests that as the value of the first feature increases, the target variable will generally decrease.
3. Intercept: The intercept of -1408763.63962 represents the expected value of the target variable when all feature values are set to zero. It's important to note that the intercept might not have a meaningful interpretation, especially if the values of zero for the features don't make sense in the context of the problem.

Now that we have analyzed all three models, we can provide a more comprehensive comparison and conclusion:

1. Model performance comparison: Comparing the R² scores of the three models, the RandomForestRegressor (R² = 0.8603) outperforms both the Linear Regression model (R² = 0.6076) and the Elastic Net Regression model (R² = 0.62475). It indicates that the RandomForestRegressor explains a higher percentage of the variance in the target variable.
2. Interpretability: The Linear Regression and Elastic Net Regression models provide more interpretable results as they offer coefficients and an intercept. In contrast, the RandomForestRegressor is a more complex model and lacks the same level of interpretability.
3. Complexity and generalization: RandomForestRegressor tends to perform well in many situations but can be prone to overfitting if not properly tuned. On the other hand, the Elastic Net Regression model incorporates both L1 and L2 regularization, which can help prevent overfitting and improve model generalization. The Linear Regression model may suffer from overfitting or underfitting, depending on the dataset.

Based on the analysis of all three models, we can conclude that the RandomForestRegressor offers the best predictive performance, but the trade-off is a lack of interpretability. If you prioritize performance, RandomForestRegressor would be the recommended choice. However, if interpretability is more important, you might consider the Elastic Net Regression model, which offers a balance between performance and interpretability.

You can further improve the models and conclusions by following the suggestions mentioned in previous responses, such as feature engineering, hyperparameter tuning, model comparison, and cross-validation.