Report on Regression Model Analysis

Name: CAO Xinyang
ID: 321793

1. Data Summary and Preprocessing

Dataset: The Auto MPG dataset from the UCI repository, which was used to predict miles per gallon (mpg) based on various vehicle features.

Preprocessing:

Handling Missing Values: Missing values were filled using the mean.

Feature Selection: The *car_name* column was excluded as it contained non-numeric, irrelevant information.

Train-Test Split: The data was split into training and testing sets (80% training, 20% testing) to ensure the models could generalize to new data.

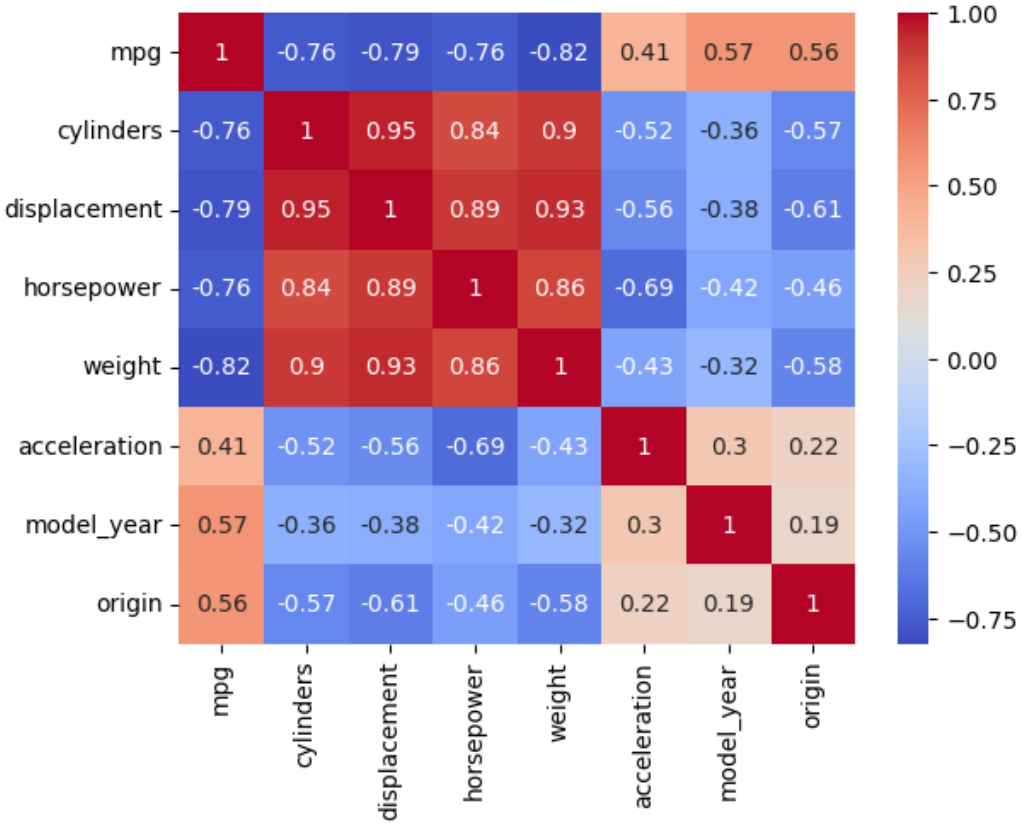
	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin
0	18.0	8.0	307.0	130.0	3504.0	12.0	70.0	1.0
1	15.0	8.0	350.0	165.0	3693.0	11.5	70.0	1.0
2	18.0	8.0	318.0	150.0	3436.0	11.0	70.0	1.0
3	16.0	8.0	304.0	150.0	3433.0	12.0	70.0	1.0
4	17.0	8.0	302.0	140.0	3449.0	10.5	70.0	1.0

2. Exploratory Data Analysis

Correlation Analysis:

Features most correlated with mpg include weight (-0.82), displacement (-0.79), and horsepower (-0.76), indicating that heavier vehicles with larger engines tend to have lower fuel efficiency.

Positive correlations with mpg were observed with model_year (0.57) and origin (0.56), suggesting that newer vehicles and certain geographic origins have better fuel efficiency.



Target Variable (mpg):

mpg has a wide range of values, indicating substantial variability in fuel efficiency across different vehicles, which may allow the models to capture diverse patterns.

3. Model Selection and Tuning

Five models were selected and tuned for comparison:

Linear Regression: Baseline model with no hyperparameters to tune.

Ridge Regression: Tuned for regularization strength (alpha). The parameter grid included alpha values of [0.01, 0.1, 1, 10, 100].

Lasso Regression: Also tuned for alpha, with the parameter grid consisting of [0.01, 0.1, 1, 10, 100].

k-Nearest Neighbors (kNN) Regression: Optimized for three hyperparameters:

n_neighbors: Number of neighbors considered in predictions, with values [3, 5, 7, 9, 11, 13, 15].

weights: Weighting function used in prediction, with options ['uniform', 'distance'].

p: Power parameter for the Minkowski distance, with values [1, 2].

Decision Tree Regression: Tuned for five hyperparameters:

criterion: The function to measure the quality of a split, with options ['squared_error', 'absolute_error', 'friedman_mse'].

max_depth: Maximum depth of the tree, with values [3, 5, 10, 15, None].

max_features: Number of features to consider for the best split, with options ['sqrt', 'log2'].

min_samples_leaf: Minimum number of samples required to be at a leaf node, with values [1, 2, 4, 6].
min_samples_split: Minimum number of samples required to split an internal node, with values [2, 5, 10, 15].

```
models = {
    'LinearRegression': (LinearRegression(), {}),
    'Ridge': (Ridge(), {'alpha': [0.01, 0.1, 1, 10, 100]}),
    'Lasso': (Lasso(), {'alpha': [0.01, 0.1, 1, 10, 100]}),
    'KNN': (KNeighborsRegressor(), {
        'n_neighbors': [3, 5, 7, 9, 11, 13, 15],
        'weights': ['uniform', 'distance'],
        'p': [1, 2]
    }),
    'DecisionTree': (DecisionTreeRegressor(), {
        'max_depth': [3, 5, 10, 15, None],
        'min_samples_split': [2, 5, 10, 15],
        'min_samples_leaf': [1, 2, 4, 6],
        'max_features': ['sqrt', 'log2'],
        'criterion': ['squared_error', 'absolute_error', 'friedman_mse']
    })
}
```

4. Model Evaluation and Comparison

The following models were evaluated based on their prediction accuracy for the target variable (miles per gallon), comparing the predicted values versus actual values on the test data. A red dashed line indicates the ideal "perfect prediction" line where predicted values exactly match actual values.

Linear Regression:

Best Parameters: Default parameters
R^2 Score: 0.7796
Performance: The predictions closely follow the actual values, especially along the middle range. However, some variance is observed at the higher end, with slight deviations from the ideal line.

Ridge Regression:

Best Parameters: alpha=10
R^2 Score: 0.7813
Performance: Ridge regression shows a similar pattern to linear regression, with predictions closely following the actual values. Minor deviations are present, but the regularization helps control overfitting.

Lasso Regression:

Best Parameters: alpha=0.01

R^2 Score: 0.7805

Performance: Lasso regression performs comparably to linear and ridge regression, though it slightly underpredicts some high values. The model provides a balanced fit overall.

K-Nearest Neighbors (KNN) Regression:

Best Parameters: n_neighbors=15, p=1, weights='distance'

R^2 Score: 0.6939

Performance: The KNN model has a wider spread around the ideal line, with larger deviations for high and low values, suggesting a lower ability to generalize compared to the linear models. This could indicate sensitivity to local variations.mpg

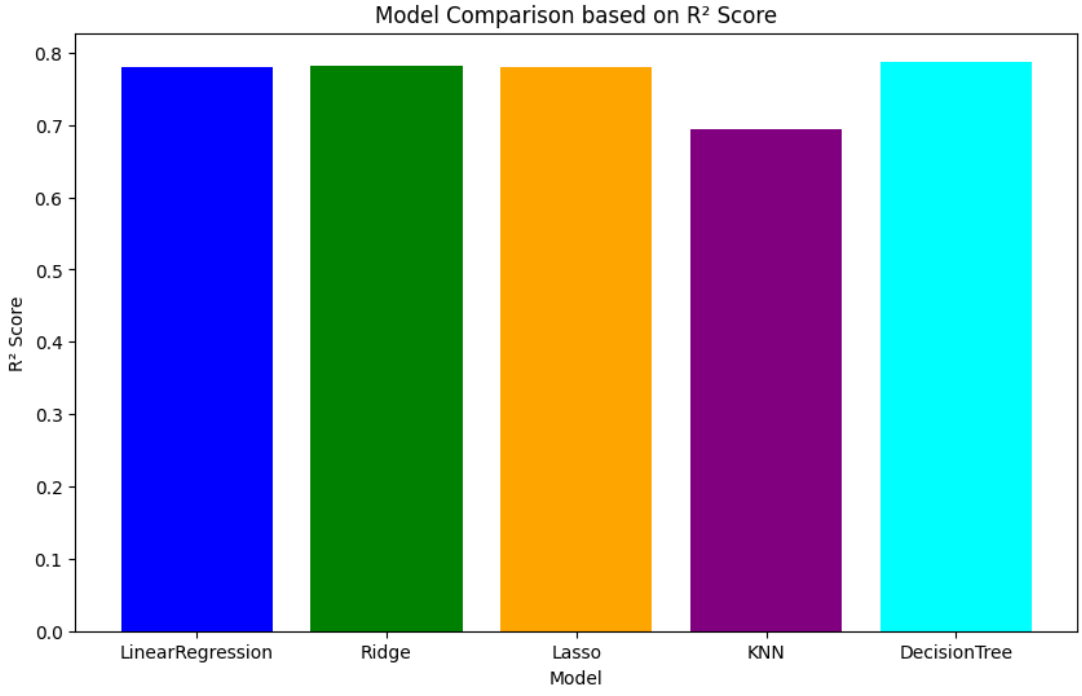
Decision Tree Regression:

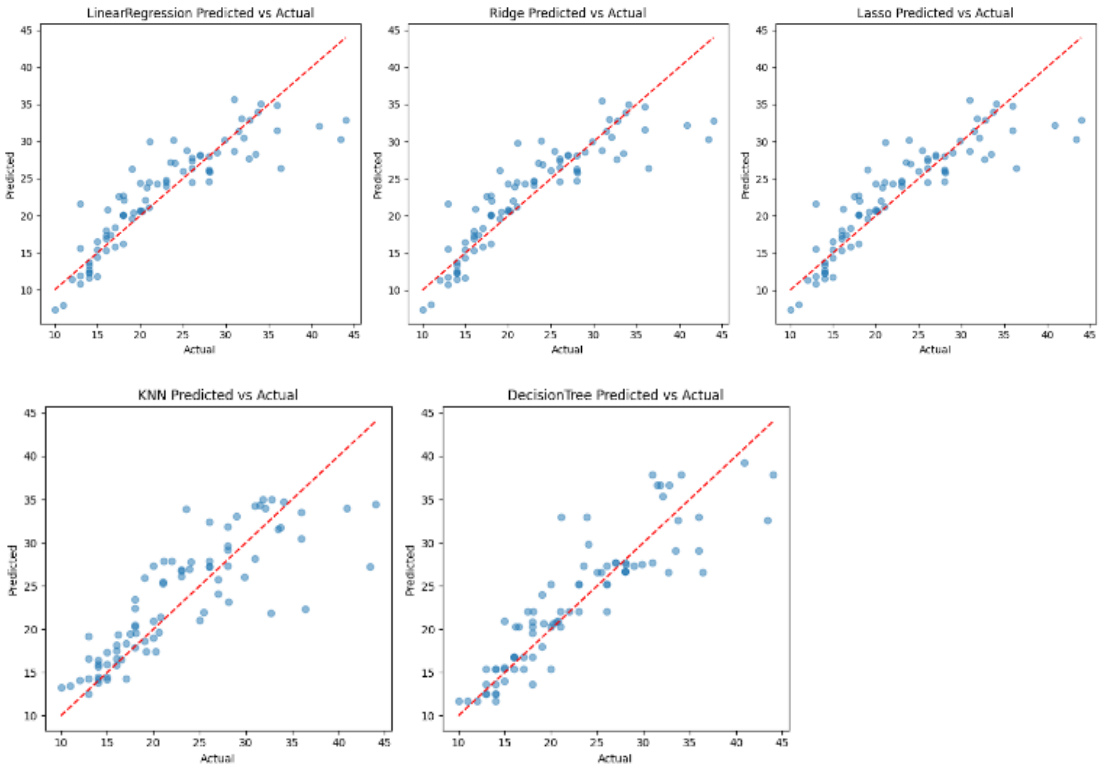
Best Parameters: criterion='friedman_mse', max_depth=10, max_features='sqrt', min_samples_leaf=6, min_samples_split=5

R^2 Score: 0.7877

Performance: The decision tree model shows more noticeable variance around the ideal line, especially at extreme values, where it sometimes overpredicts or underpredicts. This is common for decision trees due to their tendency to create discrete "steps" in predictions.

```
LinearRegression best params: {}
Ridge best params: {'alpha': 10}
Lasso best params: {'alpha': 0.01}
KNN best params: {'n_neighbors': 15, 'p': 1, 'weights': 'distance'}
DecisionTree best params: {'criterion': 'friedman_mse', 'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 6, 'min_samples_split': 5}
```





5. Conclusions

Model Performance:

Overall, Decision Tree Regression achieved the highest R^2 score among the models, but Ridge Regression and Lasso Regression provide stable and well-balanced predictions without large deviations. KNN Regression showed the largest scatter and had the lowest R^2 score, indicating it may not be as suitable for this dataset.

Recommendations:

Decision Tree Regression is recommended if accuracy is the priority, as it captured non-linear patterns better.

For a simpler model with interpretability, Ridge Regression is recommended due to its straightforward and relatively high accuracy.