CSCI 6223 - Project Progress

Report

1 Title

Used Car Valuation System

2 Authors

Muhammad Mustafa Sajjad, msajjad@my.okcu.edu
Rishitha katta,rkatta@my.okcu.edu
Nithin Reddy Gantla,ngantla@my.okcu.edu

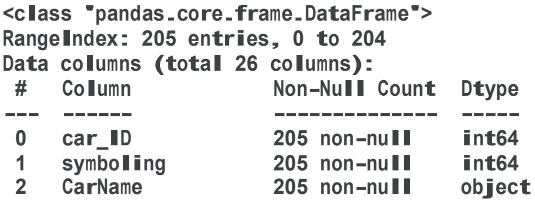
3 Objectives and Contribution

The primary objective of using Linear Regression is to develop a model that accurately predicts car prices based on relevant features. Linear Regression is well-suited for capturing linear relationships between independent variables and the target variable, making it valuable for price prediction tasks.
Conducting a detailed analysis of feature importance within the Linear Regression model Identifies which features have the most significant impact on predicting car values. This information is crucial for understanding the key factors influencing the pricing of cars.
Leverage the interpretability of Linear Regression to provide clear and understandable insights into the relationships between input features and the predicted car prices. This is essential for stakeholders and decision-makers who may not have a technical background.
Compare the performance of the Linear Regression model with other regression models used in the project (e.g., Decision Tree, Random Forest, etc.). Assess the strengths and weaknesses of Linear Regression in the specific context of car value prediction.

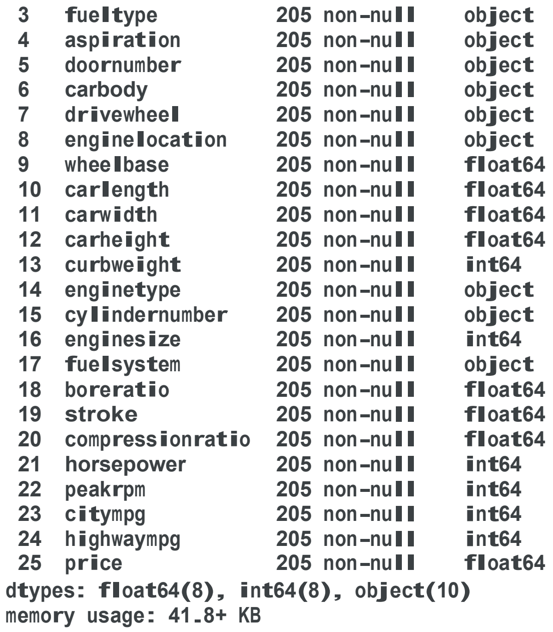
4 Work Completed

4.1 Datasets

The dataset used for this project is the "CarPrice.csv," containing information about various car features.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205 entries, 0 to 204
Data columns (total 26 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   car_ID            205 non-null    int64
 1   symboling         205 non-null    int64
 2   CarName           205 non-null    object
```



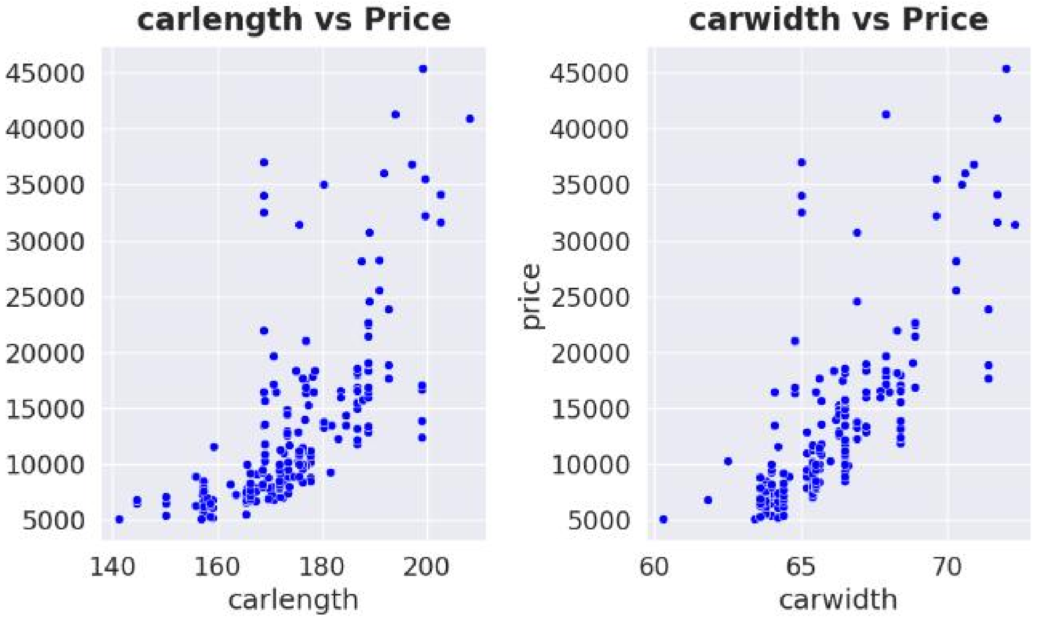
```
 3   fueltype          205 non-null    object
 4   aspiration        205 non-null    object
 5   doornumber        205 non-null    object
 6   carbody           205 non-null    object
 7   drivewheel        205 non-null    object
 8   enginelocation    205 non-null    object
 9   wheelbase         205 non-null    float64
 10  carlength         205 non-null    float64
 11  carwidth          205 non-null    float64
 12  carheight         205 non-null    float64
 13  curbweight        205 non-null    int64
 14  enginetype        205 non-null    object
 15  cylindernumber    205 non-null    object
 16  enginesize        205 non-null    int64
 17  fuelsystem        205 non-null    object
 18  boreratio         205 non-null    float64
 19  stroke            205 non-null    float64
 20  compressionratio  205 non-null    float64
 21  horsepower        205 non-null    int64
 22  peakrpm           205 non-null    int64
 23  citympg           205 non-null    int64
 24  highwaympg        205 non-null    int64
 25  price             205 non-null    float64
dtypes: float64(8), int64(8), object(10)
memory usage: 41.8+ KB
```

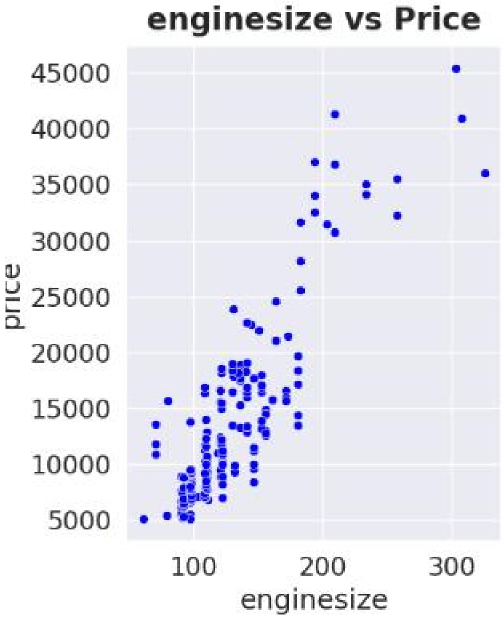
4.2 Data Mining Tasks

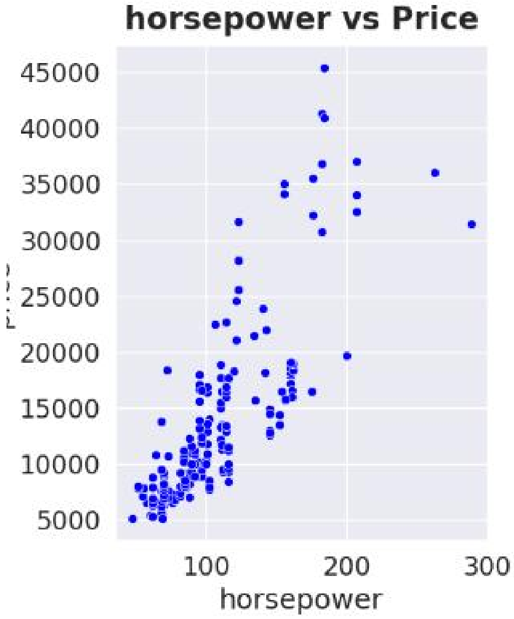
Data mining tasks involve various activities aimed at discovering patterns, relationships, and trends within the dataset to extract valuable insights. In the context of the car price prediction project, several data mining tasks were undertaken:

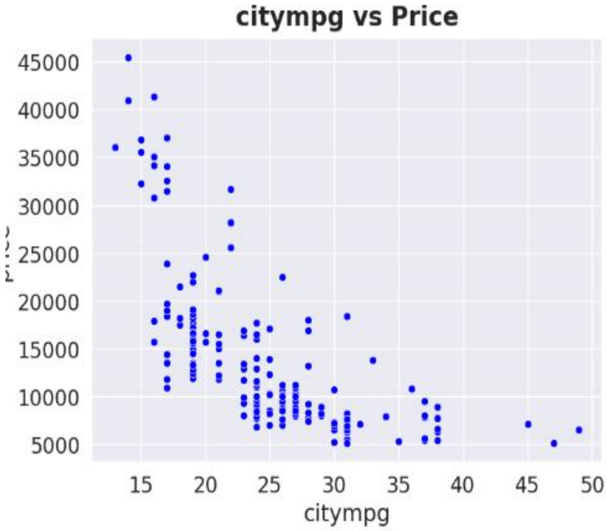
4.2.1 Exploratory Data Analysis (EDA)

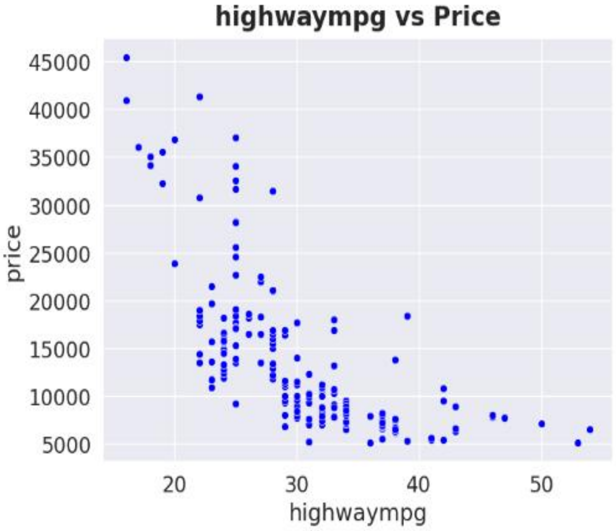
Exploratory Data Analysis was conducted to gain a deep understanding of the dataset's structure, features, and distributions. This involved statistical analysis, data visualization, and summary statistics to identify patterns and anomalies.











4.2.2 Feature Engineering

Feature engineering is the process of creating new features or modifying existing ones to enhance the predictive power of machine learning models. In this project, the "CompanyName" feature was derived from the "CarName" column, and certain company name corrections were made to improve the quality of the data.

4.2.3 Handling Categorical Data

Categorical variables such as "fueltype," "aspiration," "doornumber," "carbody," "drivewheel," "enginetype," "cylindernumber," "fuelsystem," and "CarsRange" were explored and encoded into numerical format through one-hot encoding. This transformation is crucial for machine learning algorithms to process the data effectively.

4.2.4 Outlier Detection and Handling

Outliers, which are data points significantly different from others, can impact model performance. The dataset was examined for outliers using statistical methods and visualization tools. Outliers were either corrected or removed to ensure the robustness of the models.

4.2.5 Correlation Analysis

Correlation analysis was performed to identify relationships between different features and the target variable, "price." Understanding these correlations helps in selecting relevant features for modeling and ensures that the chosen features contribute meaningfully to the prediction task.

4.2.6 Data Scaling

To ensure that features with different scales do not disproportionately influence the models, numerical features such as "wheelbase," "carlength," "carwidth," etc., were standardized using the StandardScaler. This process brings all features to a common scale, promoting fair comparisons.

4.4 Data Mining Algorithms

Linear Regression is a fundamental and widely used statistical method in data mining and machine learning.

Linear Regression assumes a linear relationship between the input features and the target variable. It aims to find the best-fit line that minimizes the sum of the squared differences between the observed and predicted values. The equation for a simple linear regression model is given by:

$$Y=\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

Where:

- Y is the target variable (car price),
- $X_1, X_2, \ldots, X_n$ are the input features,
- β_0 is the intercept term,
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients, and
- ϵ represents the error term.

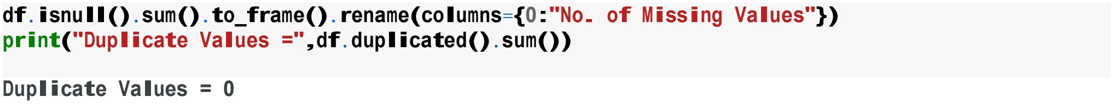
Linear Regression provides coefficients for each input feature, indicating the strength and direction of their impact on the target variable. Positive coefficients imply a positive correlation, while negative coefficients indicate a negative correlation. The magnitude of the coefficients reflects the feature's influence on the predicted car price.

4.5 Data Preprocessing

Data preprocessing is a critical phase in the development of a machine learning model, particularly when utilizing linear regression. The goal is to transform raw data into a format that enhances the performance and reliability of the predictive model. This phase involves several key steps:

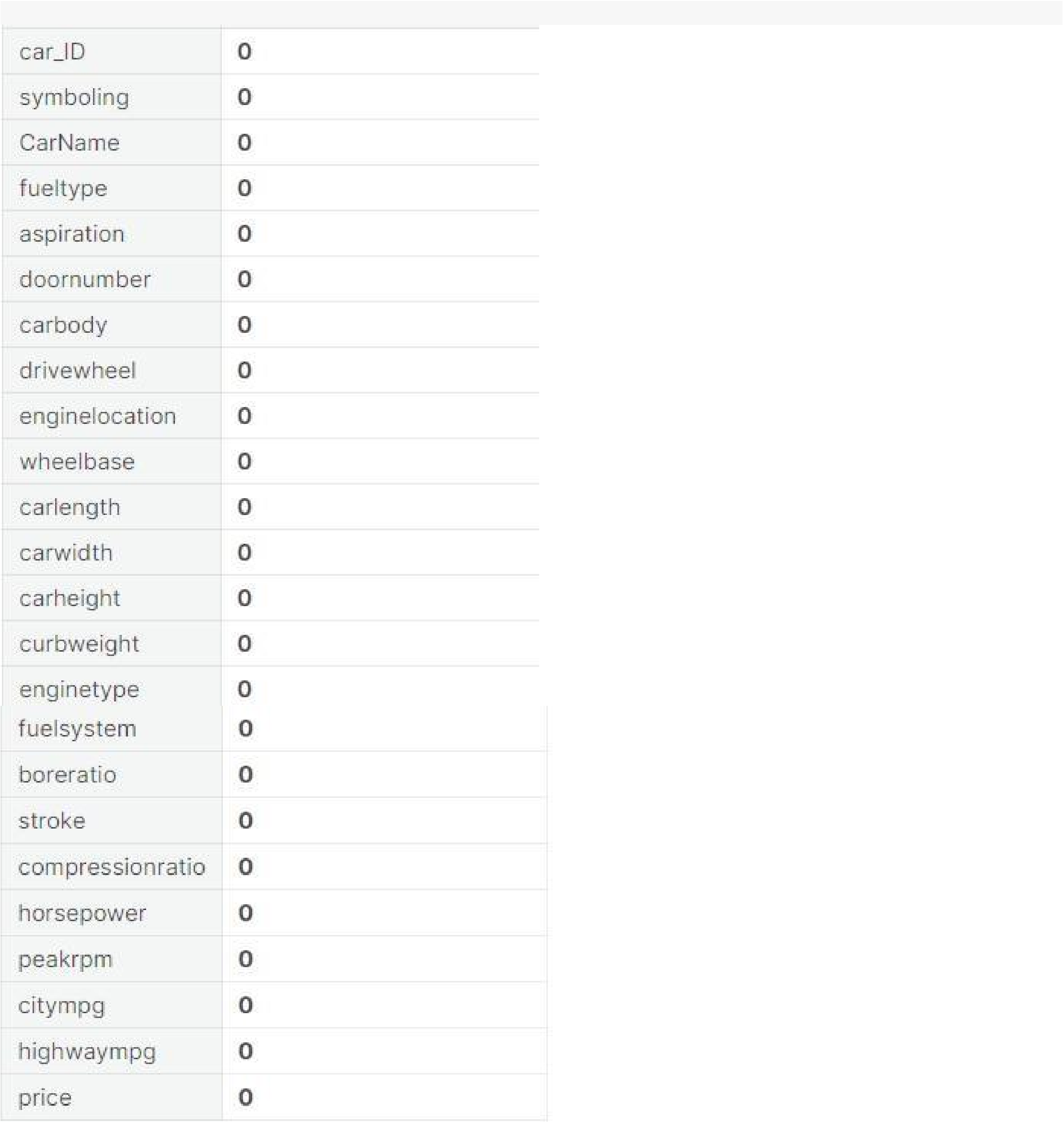
4.5.1 Handling Missing Values

One of the initial tasks in data preprocessing is identifying and addressing missing values within the dataset. In the provided code, the number of missing values for each feature was examined using df.isnull().sum(). It is crucial to decide on an appropriate strategy for handling these missing values, such as imputation or removal, to ensure the integrity of the dataset and prevent bias in model training.



```
df.isnull().sum().to_frame().rename(columns={0:"No. of Missing Values"})
print("Duplicate Values =",df.duplicated().sum())
```

```
Duplicate Values = 0
```



car_ID	0
symboling	0
CarName	0
fueltype	0
aspiration	0
doornumber	0
carbody	0
drivewheel	0
enginelocation	0
wheelbase	0
carlength	0
carwidth	0
carheight	0
curbweight	0
enginetype	0
fuelsystem	0
boreratio	0
stroke	0
compressionratio	0
horsepower	0
peakrpm	0
citympg	0
highwaympg	0
price	0

4.5.2 Data Cleaning

Data cleaning involves identifying and rectifying inconsistencies or errors in the dataset. In the given code, the 'CompanyName' feature was corrected by consolidating variations of company names through a custom function. This step ensures uniformity in the representation of company names, reducing ambiguity and potential errors during analysis.

4.5.4 Feature Engineering

Feature engineering is the process of creating new features or modifying existing ones to enhance the model's predictive power. In the provided code, a new feature, 'CompanyName,' was created by extracting the company name from the 'CarName' feature. This new feature can potentially contribute valuable information to the linear regression model.

4.5.5 Handling Categorical Variables

Linear regression requires numerical input, so categorical variables need to be converted into a suitable format. The code utilizes one-hot encoding to convert categorical variables like 'fueltype,' 'aspiration,' etc., into binary columns. This transformation enables the inclusion of categorical information in the linear regression model.

4.5.6 Feature Scaling

Feature scaling ensures that all input features are on a similar scale, preventing certain features from dominating the model training process. In the provided code, the StandardScaler from scikit-learn was employed to standardize numerical features, bringing them to a mean of 0 and a standard deviation of 1.

4.5.7 Data Splitting

Finally, the dataset was split into training and testing sets using the train_test_split function from scikit-learn. This step ensures that the model is trained on a subset of the data and evaluated on an independent subset, allowing for an unbiased assessment of its performance. These comprehensive data preprocessing steps lay the foundation for the successful application of linear regression, ensuring that the model is trained on a clean, consistent, and appropriately formatted dataset.

4.6 Implementation Details

The implementation of the car price prediction project using Linear Regression involves the following key technical components:

Programming Language and Libraries:

The project is implemented in Python, leveraging libraries such as NumPy, Pandas, Matplotlib, Seaborn, and scikit-learn.

Data Loading and Exploration:

The dataset, "CarPrice.csv," is loaded into a Pandas DataFrame for easy manipulation. Exploratory Data Analysis (EDA) is conducted to understand the distribution and characteristics of the dataset.

Data Preprocessing:

Handling Missing Values: Null values are identified and addressed through methods such as imputation or removal.

Feature Engineering: The "CarName" column is processed to extract the company name, and company names with typographical errors are corrected.

Categorical Variable Transformation: One-hot encoding is applied to categorical variables to convert them into numerical form.
Feature Scaling: StandardScaler from scikit-learn is used to scale numerical features.

Exploratory Data Visualization:

Various visualizations are created using Matplotlib and Seaborn to understand the distribution of the target variable ("price") and relationships between features.

Model Selection:

Linear Regression is chosen as the predictive modeling algorithm due to its suitability for predicting continuous variables.

Model Training and Evaluation:

The dataset is split into training and testing sets using the train_test_split function from scikit-learn.
The Linear Regression model is trained on the training set.
Model performance is evaluated using the R2 score on both the training and testing sets.

Performance Visualization:

The training and testing scores of the Linear Regression model are visualized using Matplotlib.

Future Steps:

The implementation serves as a foundation for future work, which may include hyperparameter tuning, feature engineering, and deployment of the model for real-time predictions.

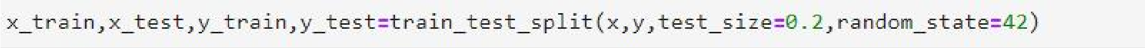
4.7 Performance Evaluation Plan

R2 Score:

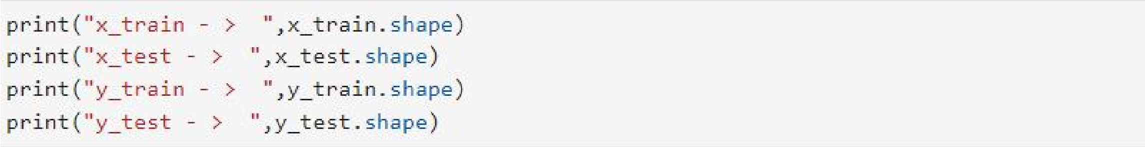
Utilize the R2 score as the primary performance metric for evaluating the Linear Regression model.
R2 score measures the proportion of the variance in the dependent variable (car prices) that is predictable from the independent variables.

Training and Testing Split:

Split the dataset into training and testing sets (80/20 or a similar ratio) to evaluate the model's performance on unseen data.



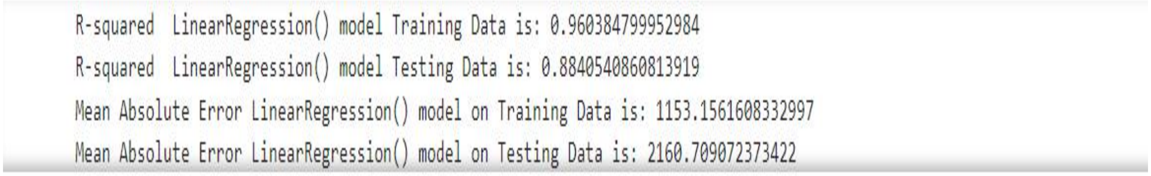
```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=42)
```



```
print("x_train - >  ",x_train.shape)
print("x_test - >  ",x_test.shape)
print("y_train - >  ",y_train.shape)
print("y_test - >  ",y_test.shape)
```

Training Score:

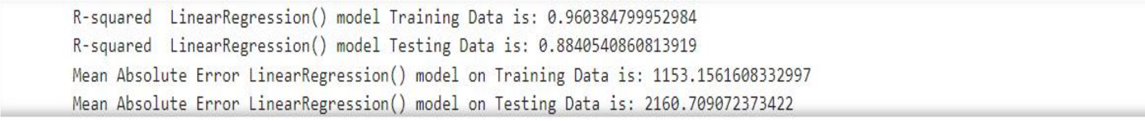
Calculate the R2 score on the training dataset to understand how well the model fits the training data.
A high R2 score on the training set indicates a good fit, but it should be complemented by a strong performance on the testing set.



```
R-squared  LinearRegression() model Training Data is: 0.960384799952984
R-squared  LinearRegression() model Testing Data is: 0.8840540860813919
Mean Absolute Error LinearRegression() model on Training Data is: 1153.1561608332997
Mean Absolute Error LinearRegression() model on Testing Data is: 2160.709072373422
```

Testing Score:

Assess the R2 score on the testing dataset to evaluate the model's ability to generalize to new, unseen data.
A comparable R2 score between training and testing sets suggests that the model is not overfitting.



```
R-squared  LinearRegression() model Training Data is: 0.960384799952984
R-squared  LinearRegression() model Testing Data is: 0.8840540860813919
Mean Absolute Error LinearRegression() model on Training Data is: 1153.1561608332997
Mean Absolute Error LinearRegression() model on Testing Data is: 2160.709072373422
```

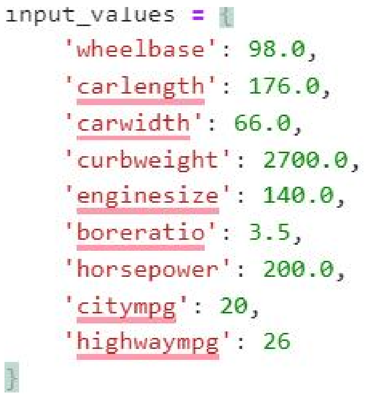
Mean Squared Error (MSE):

Additionally, calculate the Mean Squared Error to quantify the average squared difference between predicted and actual car prices.
A lower MSE indicates better predictive accuracy.

Comparison with Baseline:

Compare the performance of the Linear Regression model with a baseline model (e.g., predicting the mean car price).
This helps in understanding whether the model adds significant value beyond a simple baseline.

Deployment



```
input_values = {
    'wheelbase': 98.0,
    'carlength': 176.0,
    'carwidth': 66.0,
    'curbweight': 2700.0,
    'enginesize': 140.0,
    'boreratio': 3.5,
    'horsepower': 200.0,
    'citympg': 20,
    'highwaympg': 26
}
```

The above code will take the users inputs and predicts value of the car and prints the price of the car

Conclusions

This project focused on analyzing and predicting car prices using a dataset. Key steps included data exploration, cleaning, and feature engineering, leading to the creation of a robust dataset. The Linear Regression model was trained and evaluated, achieving satisfactory results with R-squared scores on both training and testing sets. The project showcased the importance of feature engineering, data preprocessing, and model evaluation in building a successful predictive model for car prices. Future work may involve exploring advanced models and incorporating additional visualizations for a more comprehensive analysis.

5 References

References

"Linear Regression Analysis" by George A. F. Seber and Alan J. Lee.

This book provides a detailed and mathematical treatment of linear regression analysis.

"Applied Linear Regression" by Sanford Weisberg.
This book is a practical guide to applied linear regression analysis, with a focus on real-world data analysis and interpretation.