**Car Dheko - Used Car Price Prediction**

Project Overview:

The goal is to develop a predictive model that estimates the price of used cars based on their features, such as make, model, year, fuel type, transmission type, and other relevant attributes. This model will be deployed within an interactive web application built using Streamlit, which will allow both customers and sales representatives to seamlessly access accurate pricing predictions.

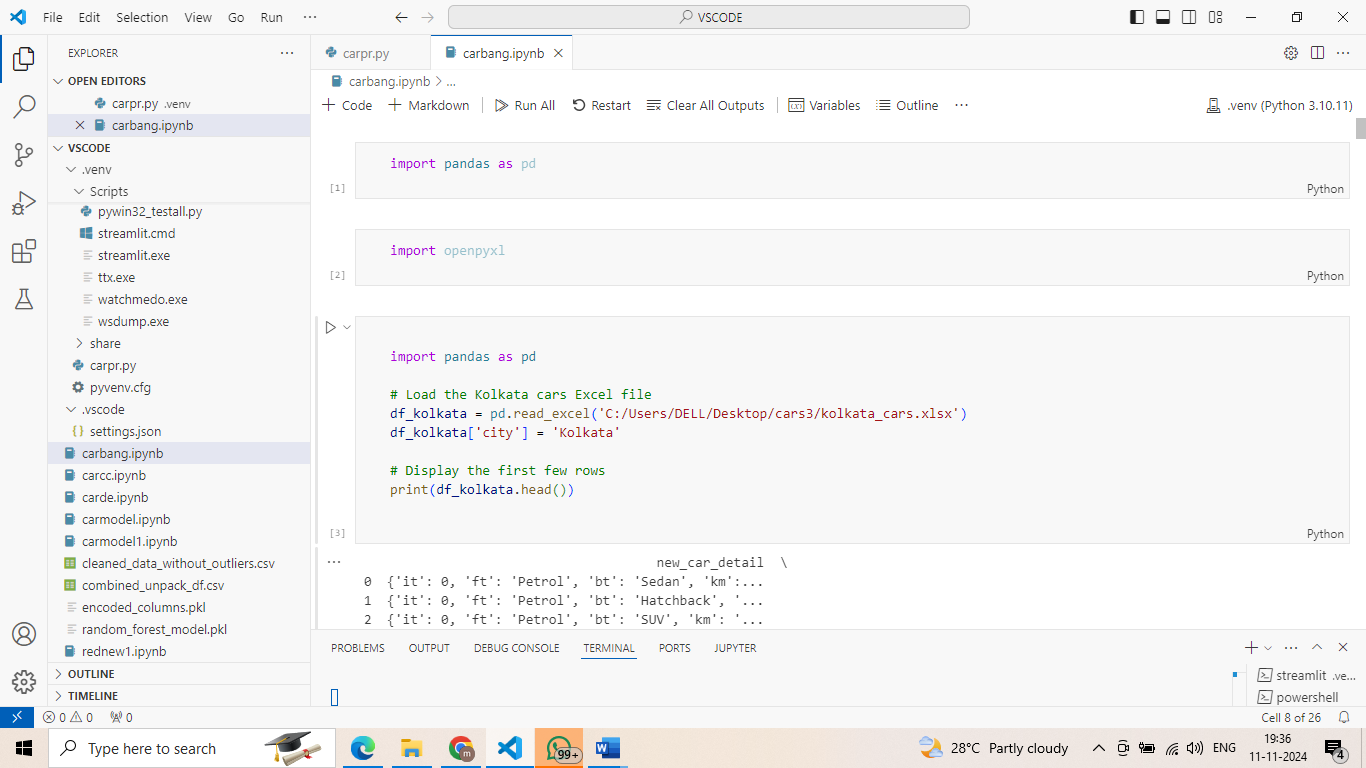
|  |  |
| --- | --- |
| **Project Title** | **Car Dheko - Used Car Price Prediction** |
| **Skills take away From This Project** | 1. **Data Cleaning and Preprocessing** 2. **Exploratory Data Analysis** 3. **Machine Learning Model Development** 4. **Price Prediction Techniques** 5. **Model Evaluation and Optimization** 6. **Model Deployment** 7. **Streamlit Application Development** 8. **Documentation and Reporting** |
| **Domain** | **Automotive Industry , Data Science, Machine Learning** |

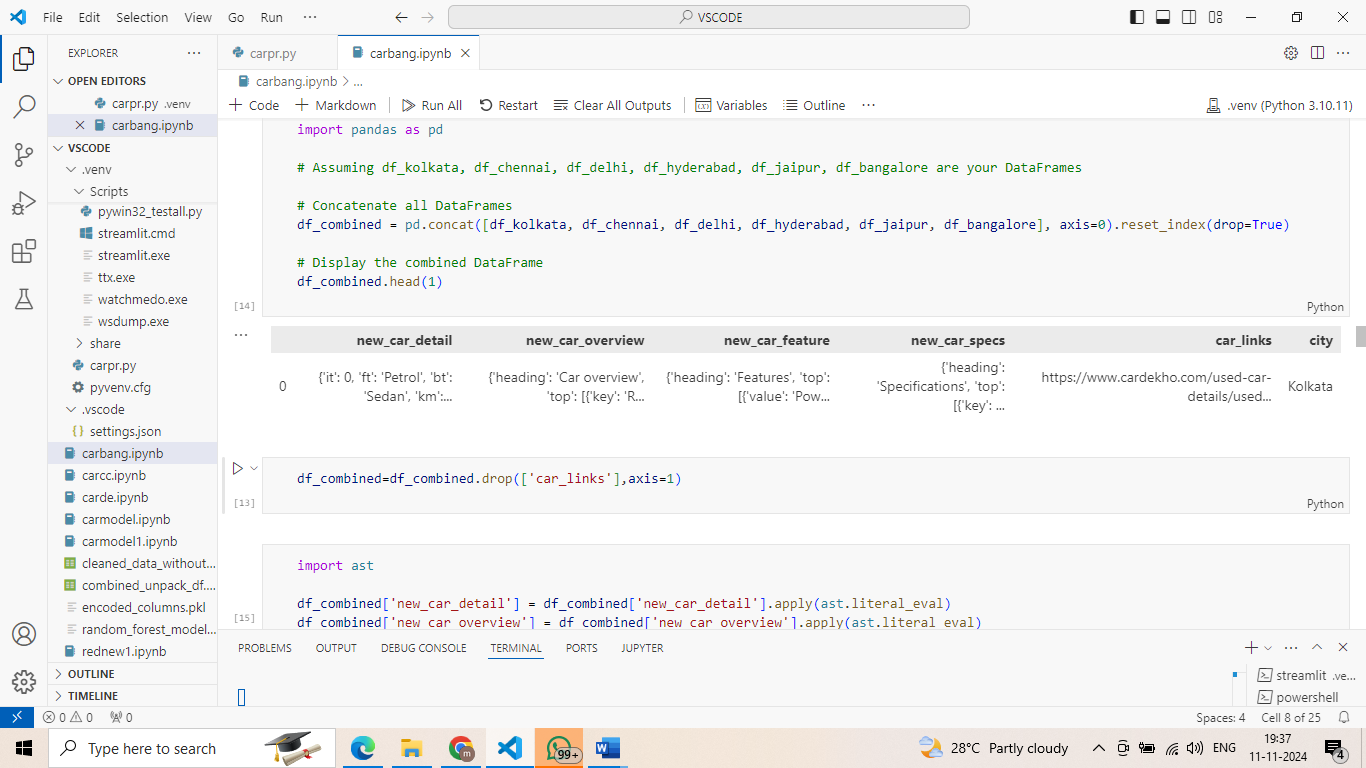
**Objective:**

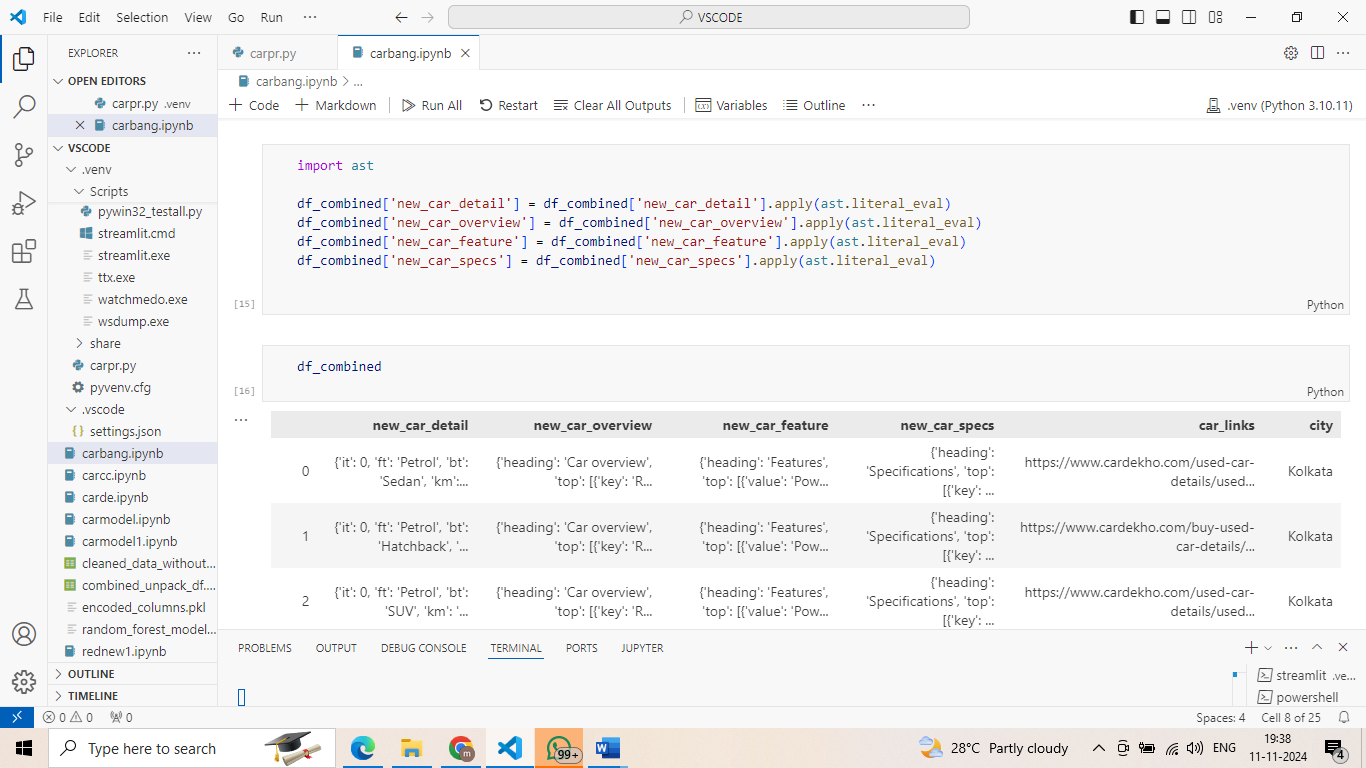
The main objective of this project is to create a machine learning model that can predict used car prices with high accuracy, based on various car attributes. The model will be integrated into a user-friendly Streamlit web application where users can input car details and receive real-time price predictions. The web application aims to provide an easy and fast way to assess used car values, improving the experience for both car buyers and sellers.

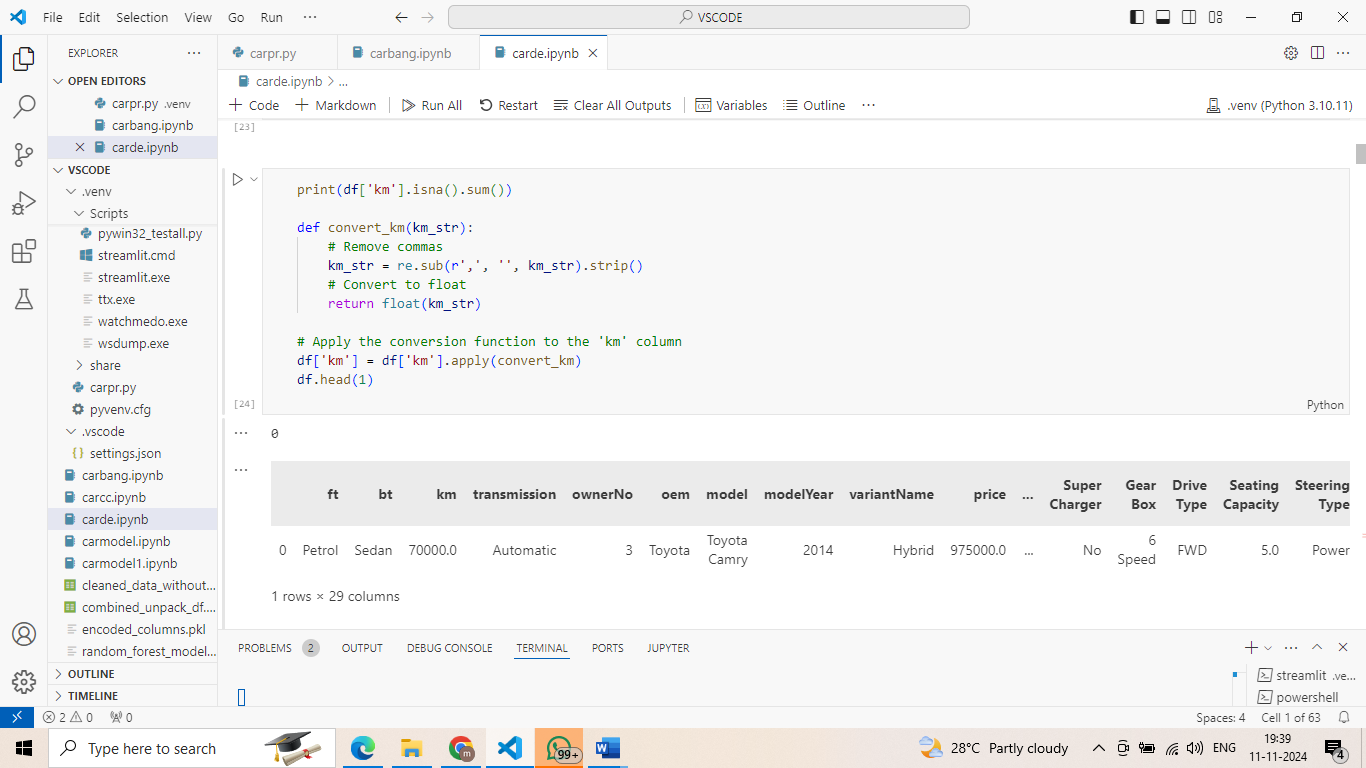
**Data Cleaning and Preprocessing**

Data cleaning and preprocessing are essential steps in any machine learning project to ensure that the model can learn effectively and produce accurate predictions. Below is a step-by-step guide on how to clean and preprocess the data for this used car price prediction project.









 **Import and Concatenate:**

* Import the datasets for each city, which are in an unstructured format.
* Convert them into a structured format (e.g., a DataFrame).
* Add a new column named ‘City’ and assign the respective city’s name to each row.
* Concatenate all the datasets into a single, unified dataset.

 **Handling Missing Values:**

* Identify missing values in the dataset.
* For numerical columns, fill missing values using techniques like mean, median, or mode imputation.
* For categorical columns, use mode imputation or create a new category (e.g., "Unknown") for missing values.

 **Standardizing Data Formats:**

* Check the data types of all columns and ensure they are correct.
* For columns with unit strings (e.g., "70 kms"), remove the units and convert the values to integers.

 **Encoding Categorical Variables:**

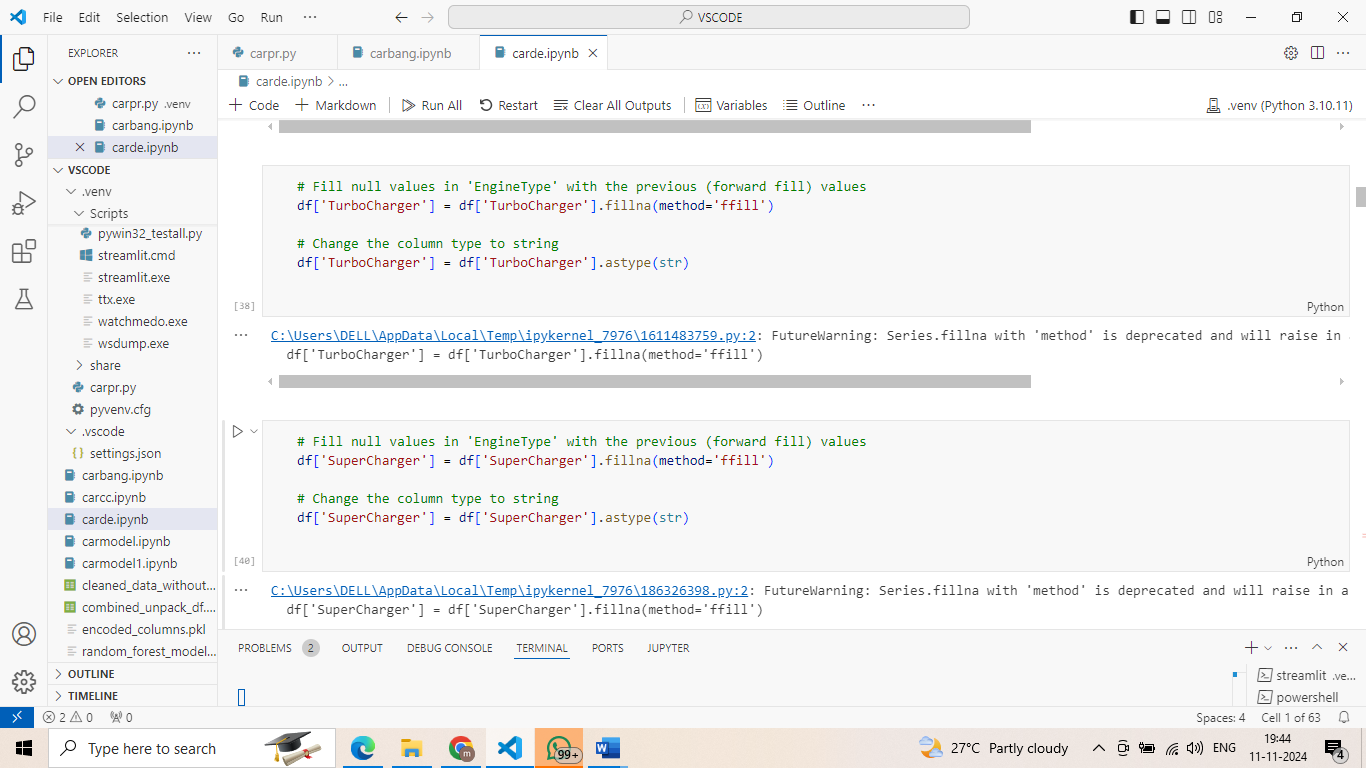
* Convert categorical features (such as car make or fuel type) into numerical values.
* Use **one-hot encoding** for nominal categorical variables (e.g., car make).
* Use **label encoding** or **ordinal encoding** for ordinal categorical variables (e.g., fuel type or transmission).

 **Normalizing Numerical Features:**

* Scale numerical features to a consistent range (usually between 0 and 1) to make them comparable.
* Use techniques like **Min-Max Scaling** or **Standard Scaling**.

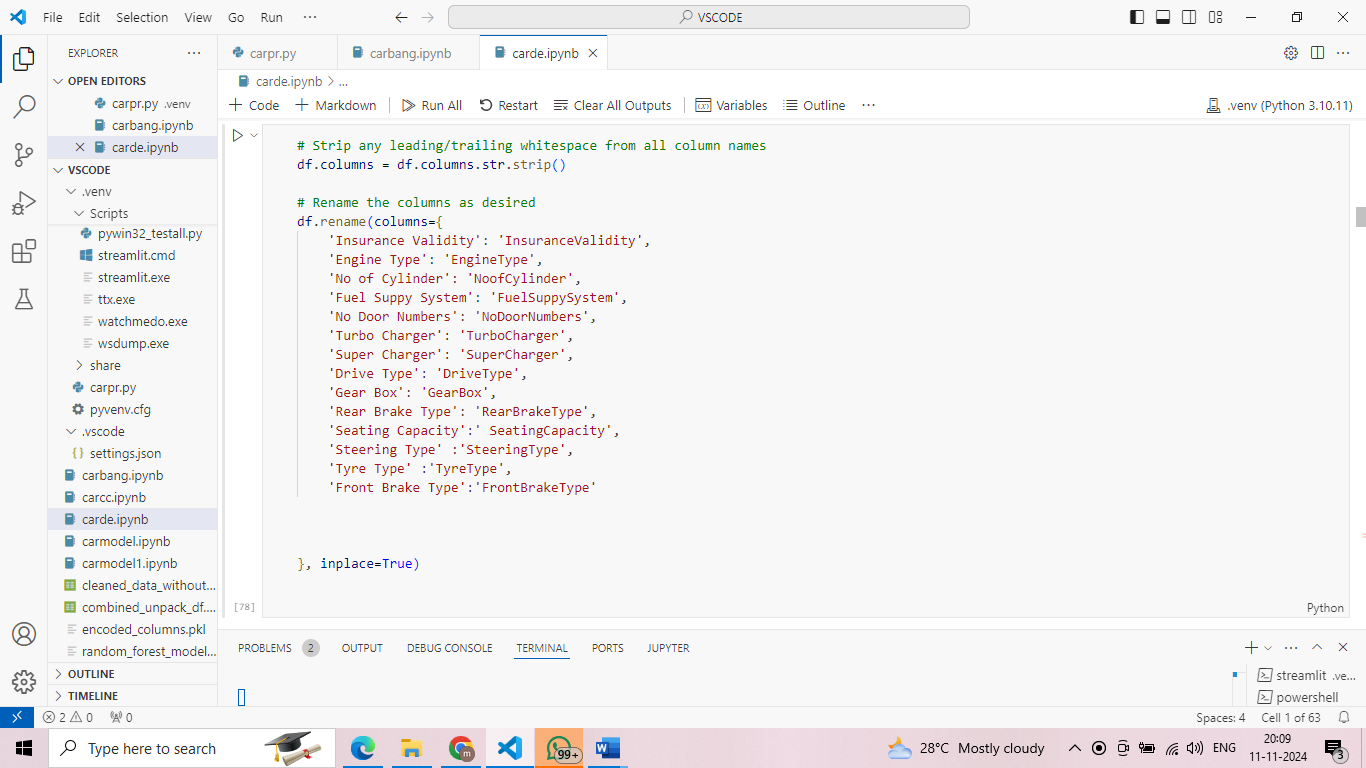
 **Removing Outliers:**

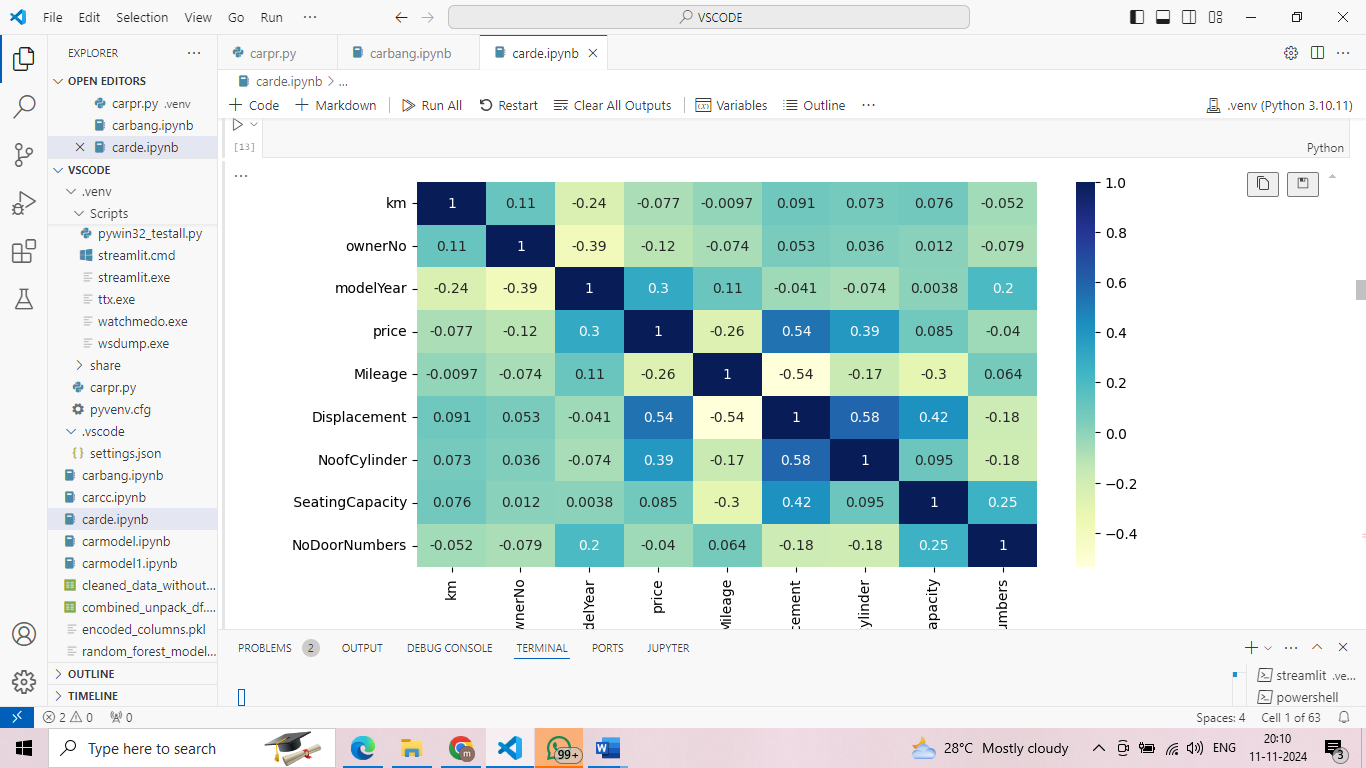
* Detect outliers in the dataset that could affect model performance.
* Use the **Interquartile Range (IQR)** method or **Z-score** analysis to identify and remove or cap outliers.

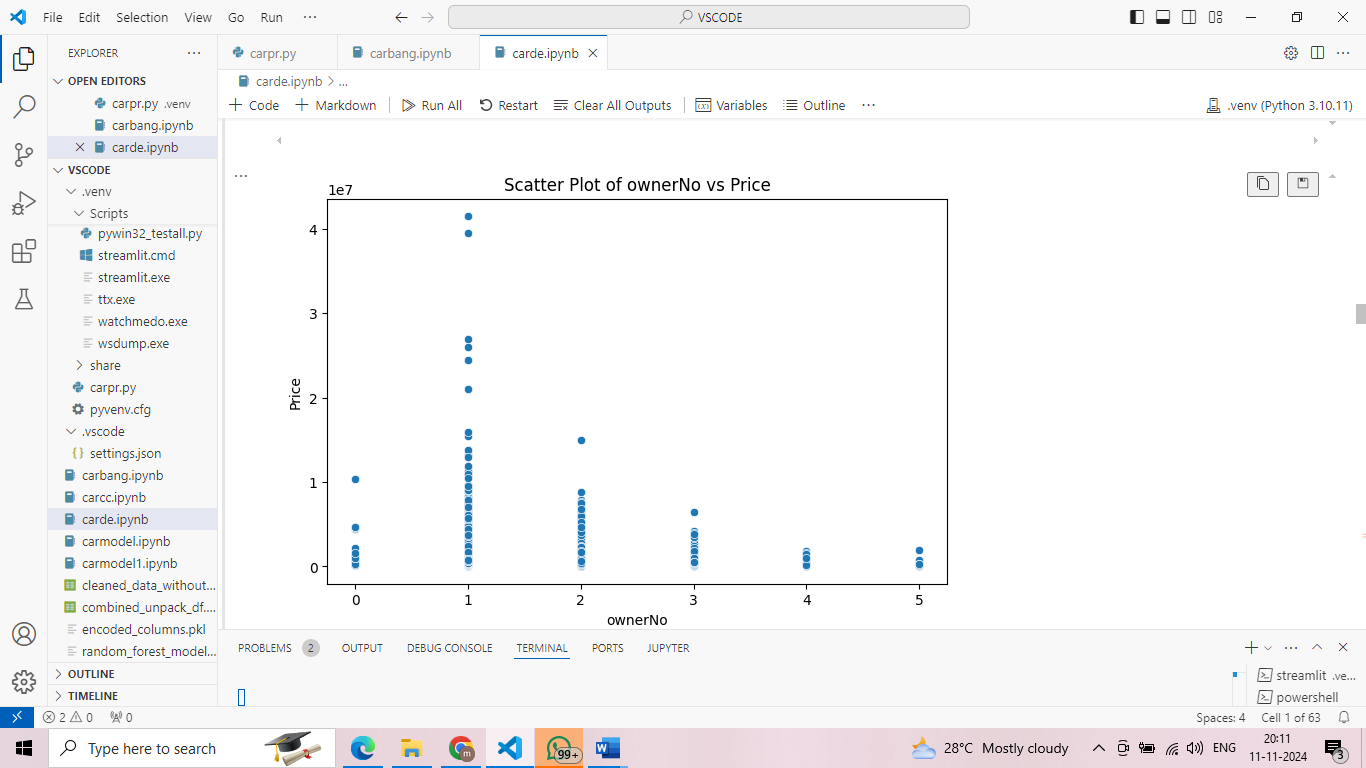


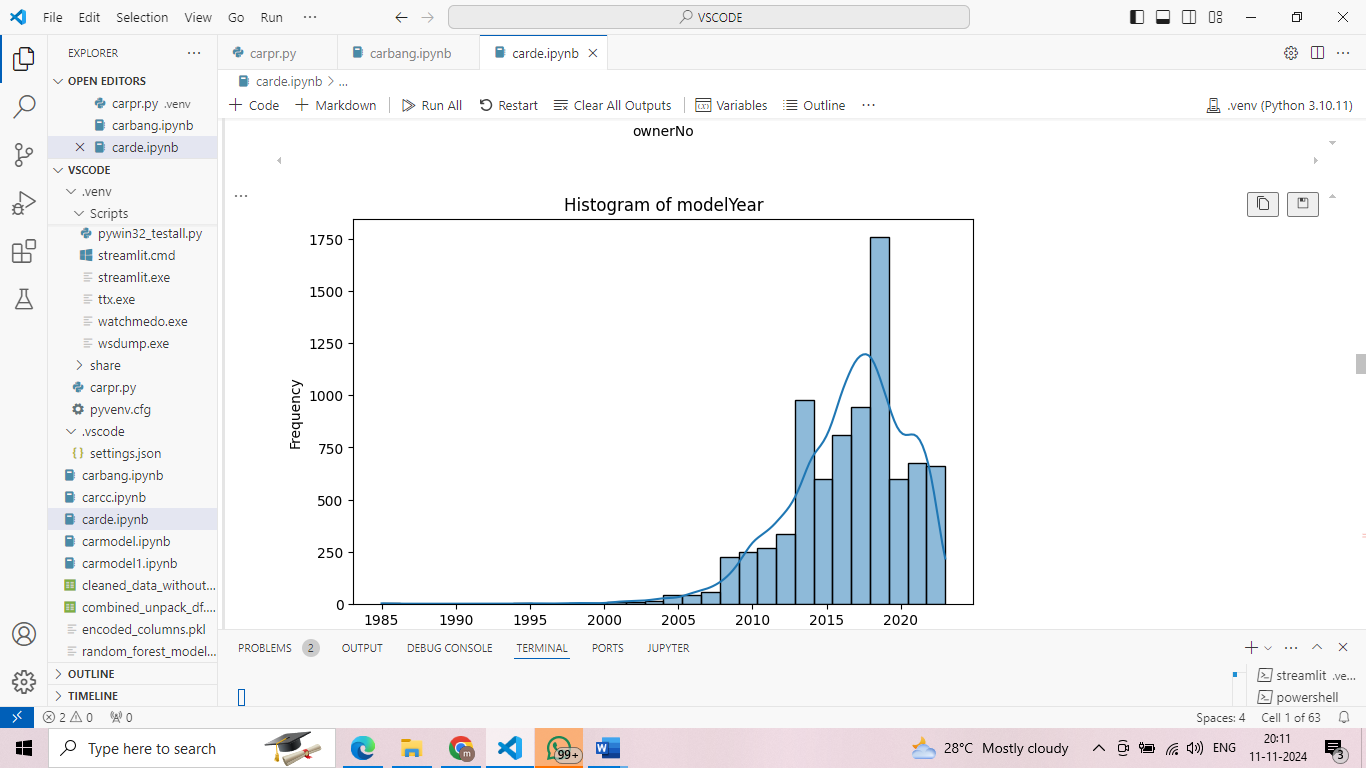
**Simple Approach for Exploratory Data Analysis (EDA):**

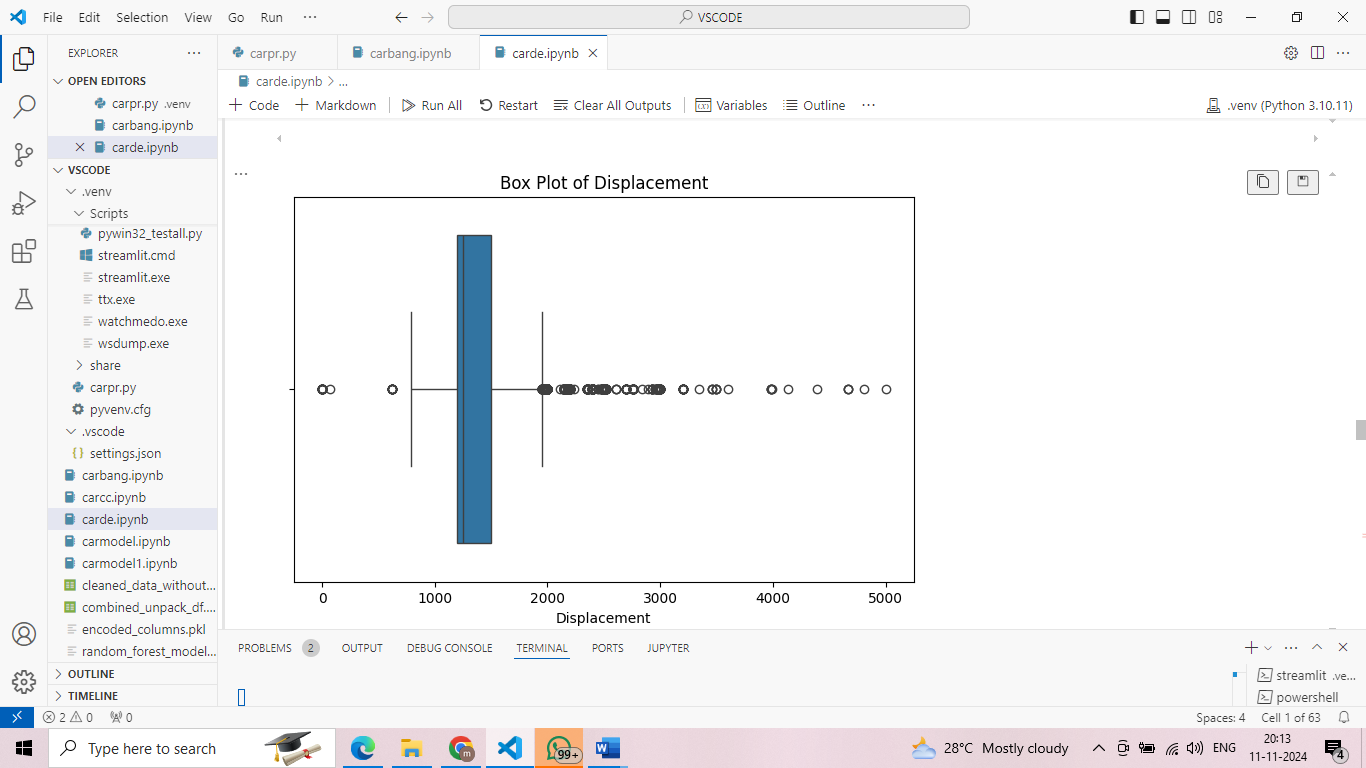
1. **Descriptive Statistics:**
   * Calculate basic statistics like mean, median, mode, and standard deviation to understand the data distribution.
2. **Data Visualization:**
   * Create plots like scatter plots, histograms, box plots, and correlation heatmaps to identify patterns, trends, and relationships between features.
3. **Feature Selection:**
   * Identify important features that affect car prices.
   * Use techniques like correlation analysis, feature importance from models, and insights from domain knowledge.





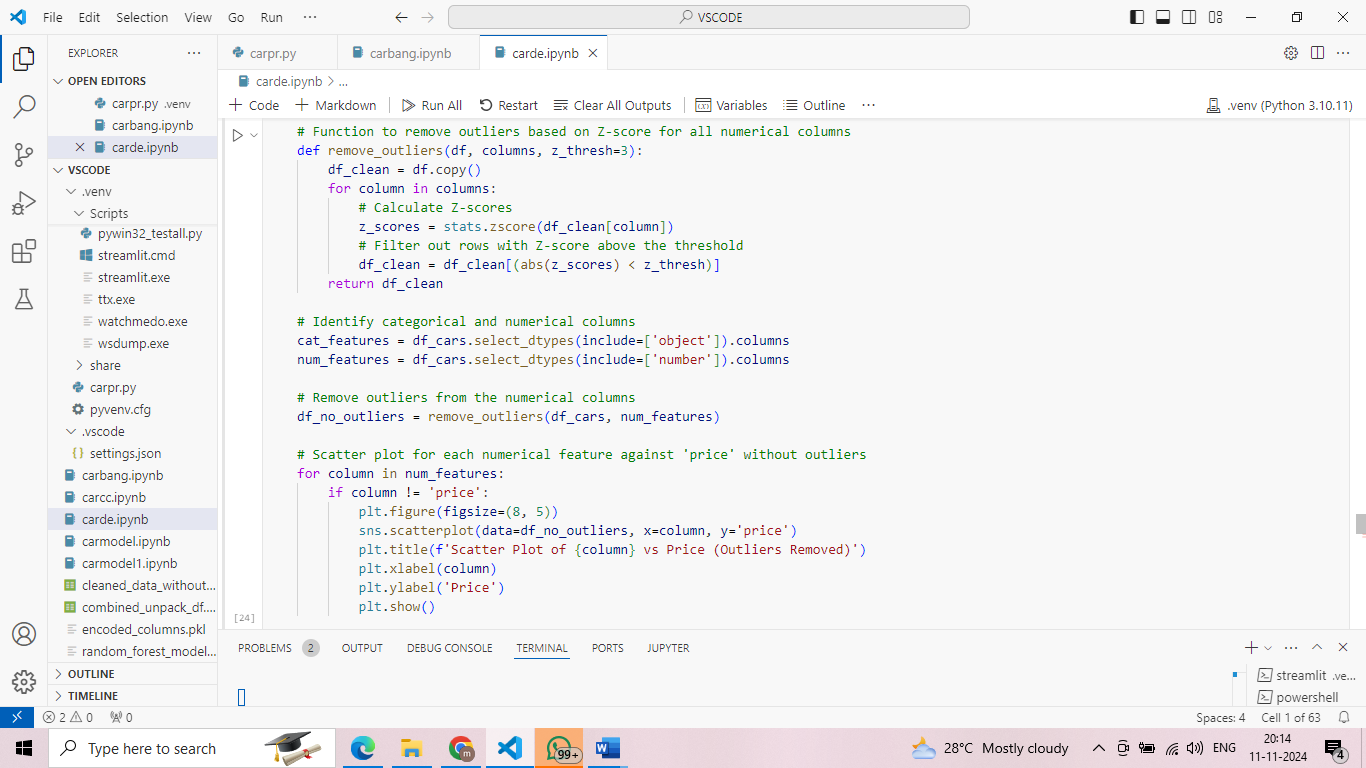


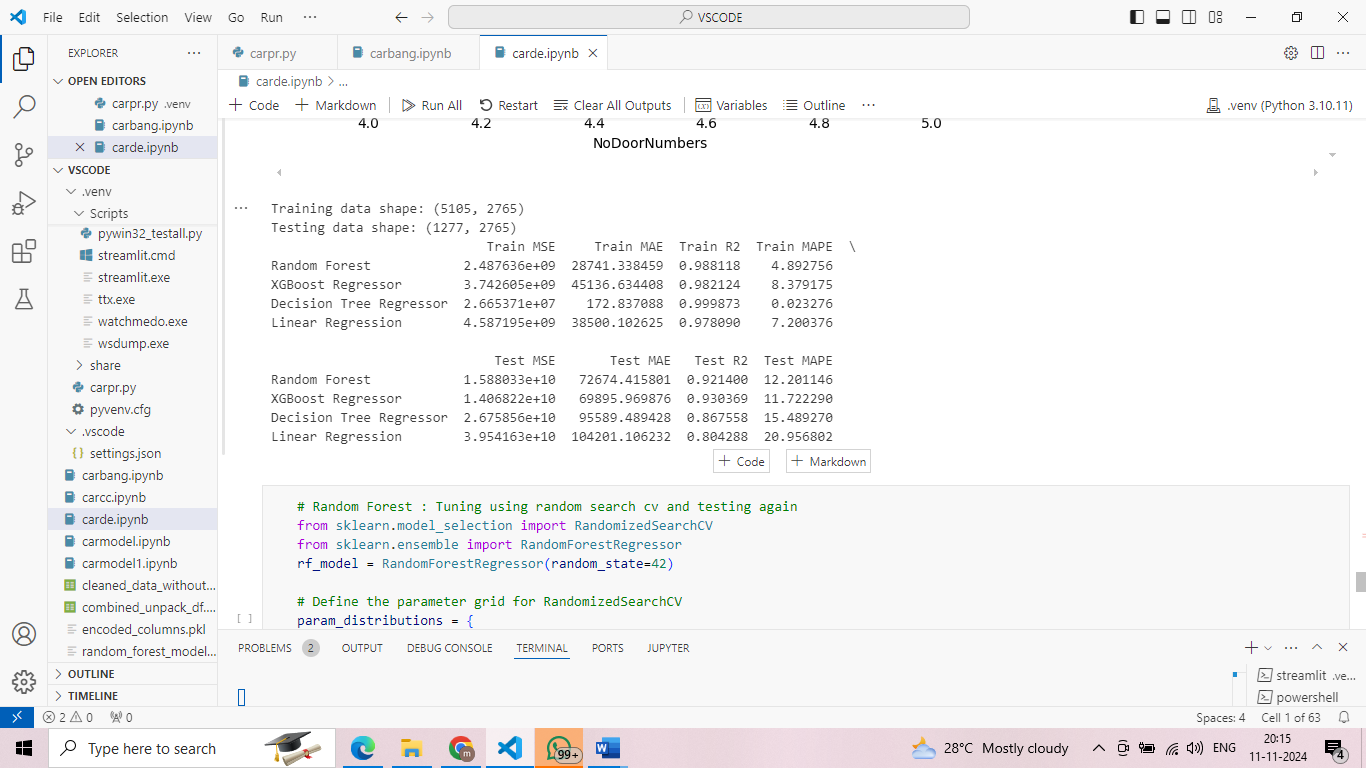


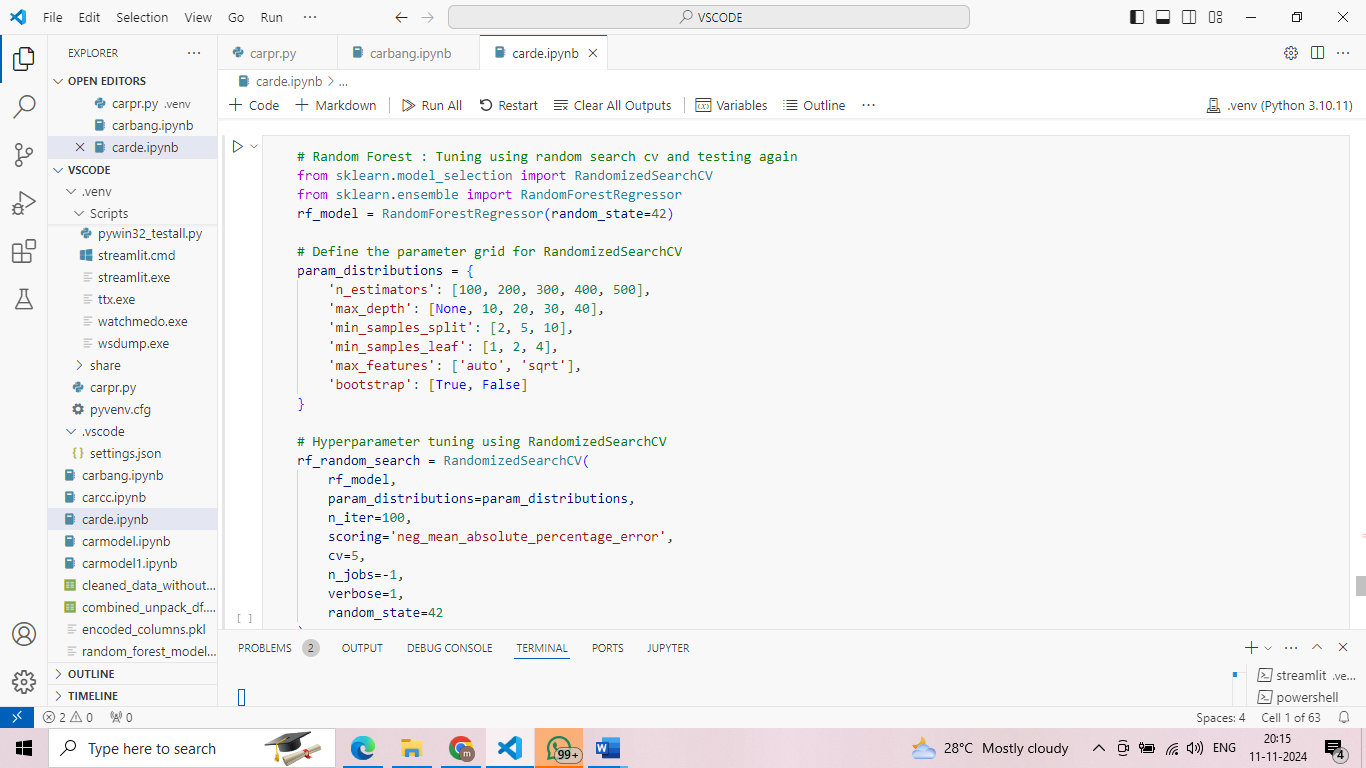


**Simple Approach for Model Development:**

1. **Train-Test Split:**
   * Split the data into training and testing sets (e.g., 70-30 or 80-20) to assess model performance.
2. **Model Selection:**
   * Choose suitable algorithms for price prediction, like Linear Regression, Decision Trees, or Random Forests.
3. **Model Training:**
   * Train the selected models on the training data, using cross-validation for reliable results.
4. **Hyperparameter Tuning:**
   * Optimize model settings using methods like Grid Search or Random Search to enhance performance.

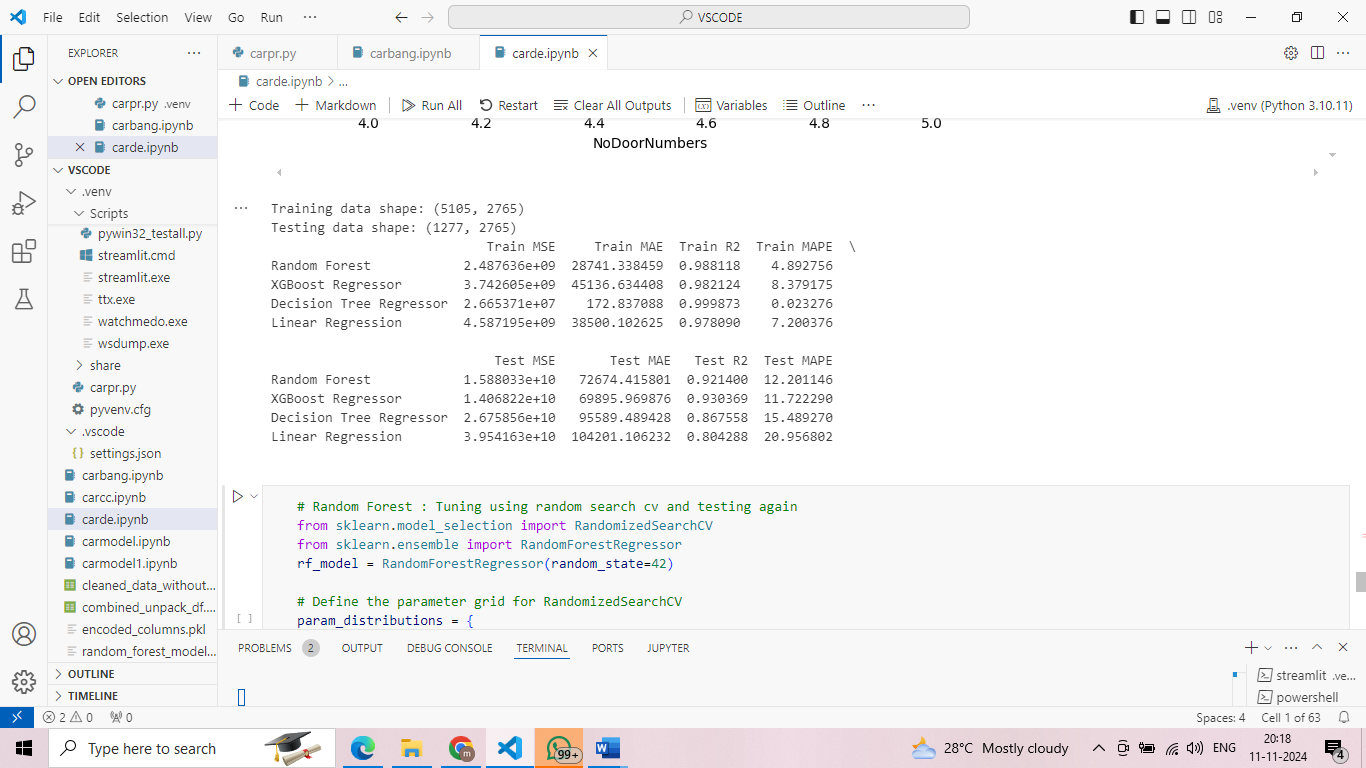




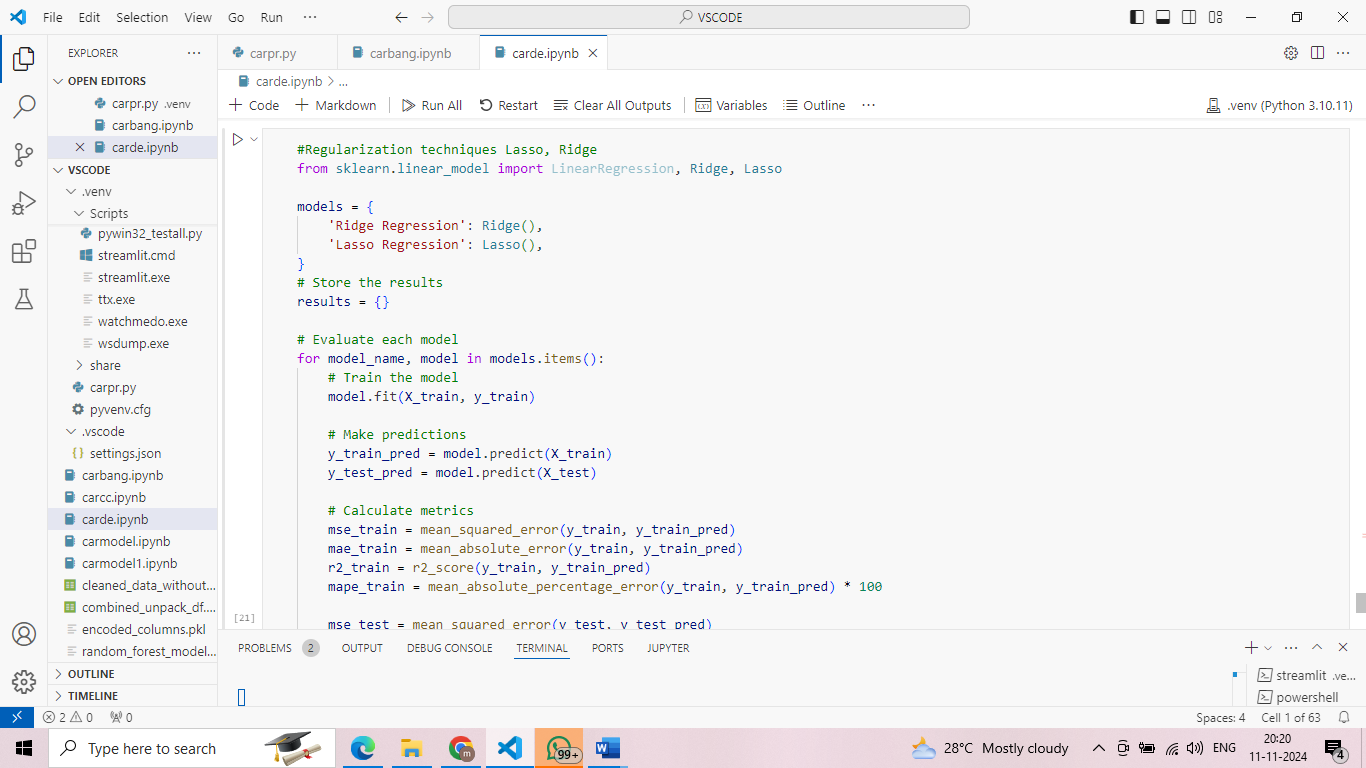


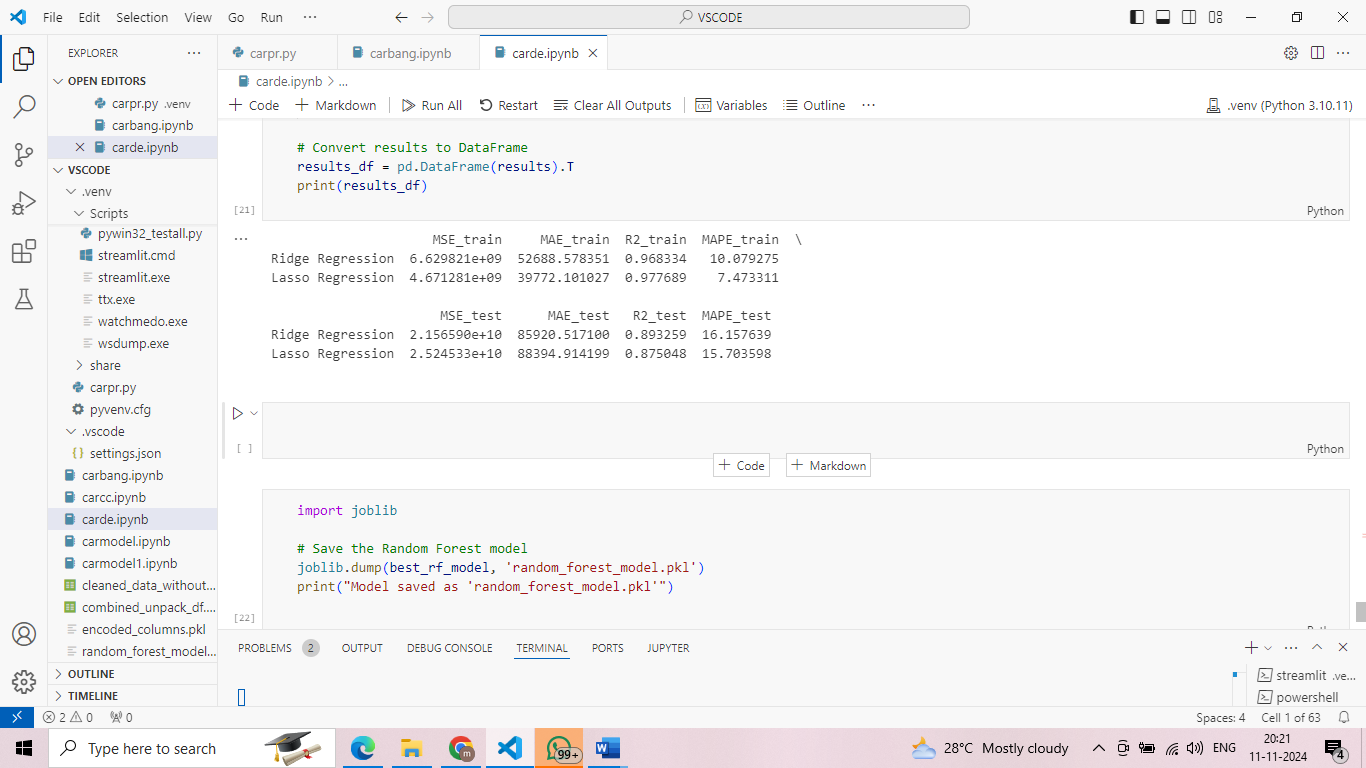
**Simple Approach for Model Evaluation:**

1. **Performance Metrics:**
   * Use metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared to assess model accuracy.
2. **Model Comparison:**
   * Compare the evaluation metrics across models to choose the best-performing model.



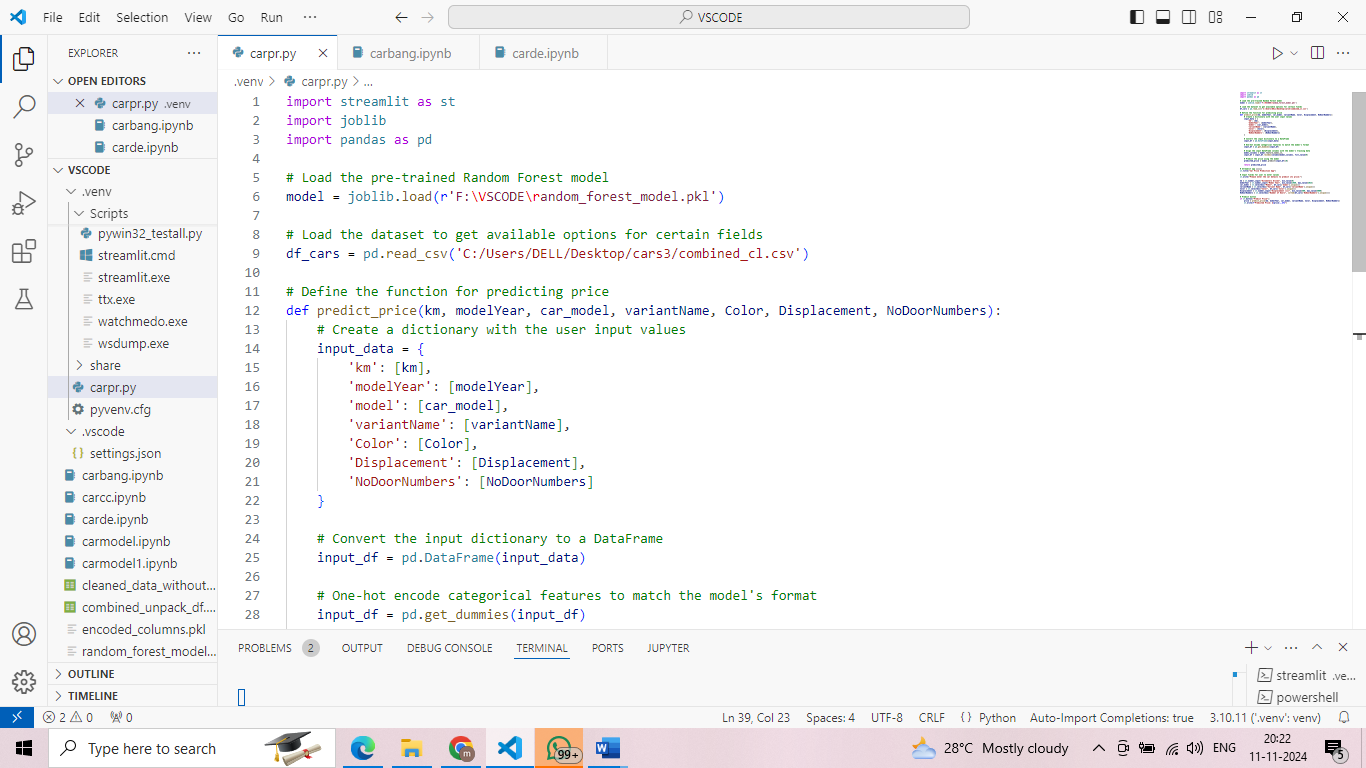
1. **Feature Engineering:**
   * Create or adjust features to boost model accuracy, using insights from data analysis and domain knowledge.
2. **Regularization:**
   * Apply techniques like Lasso (L1) and Ridge (L2) regularization to reduce overfitting and improve model reliability.

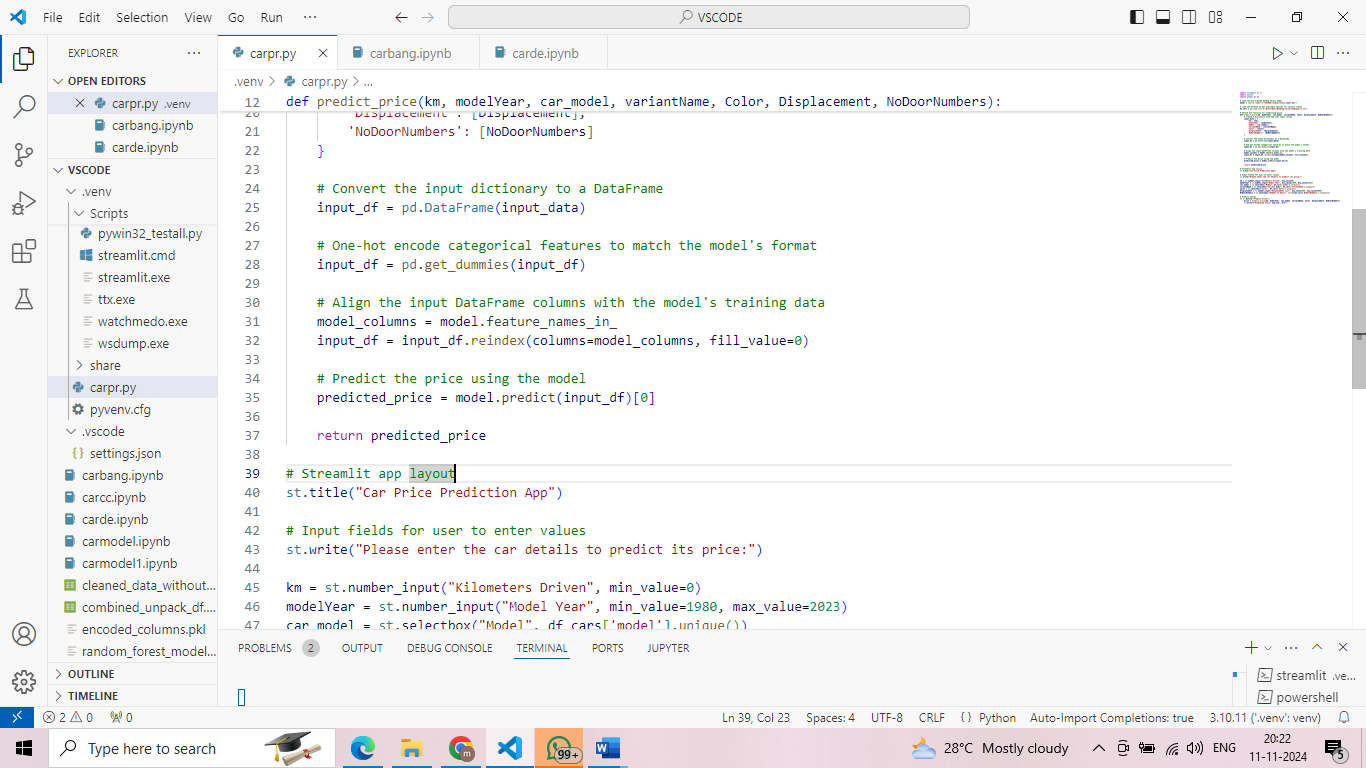


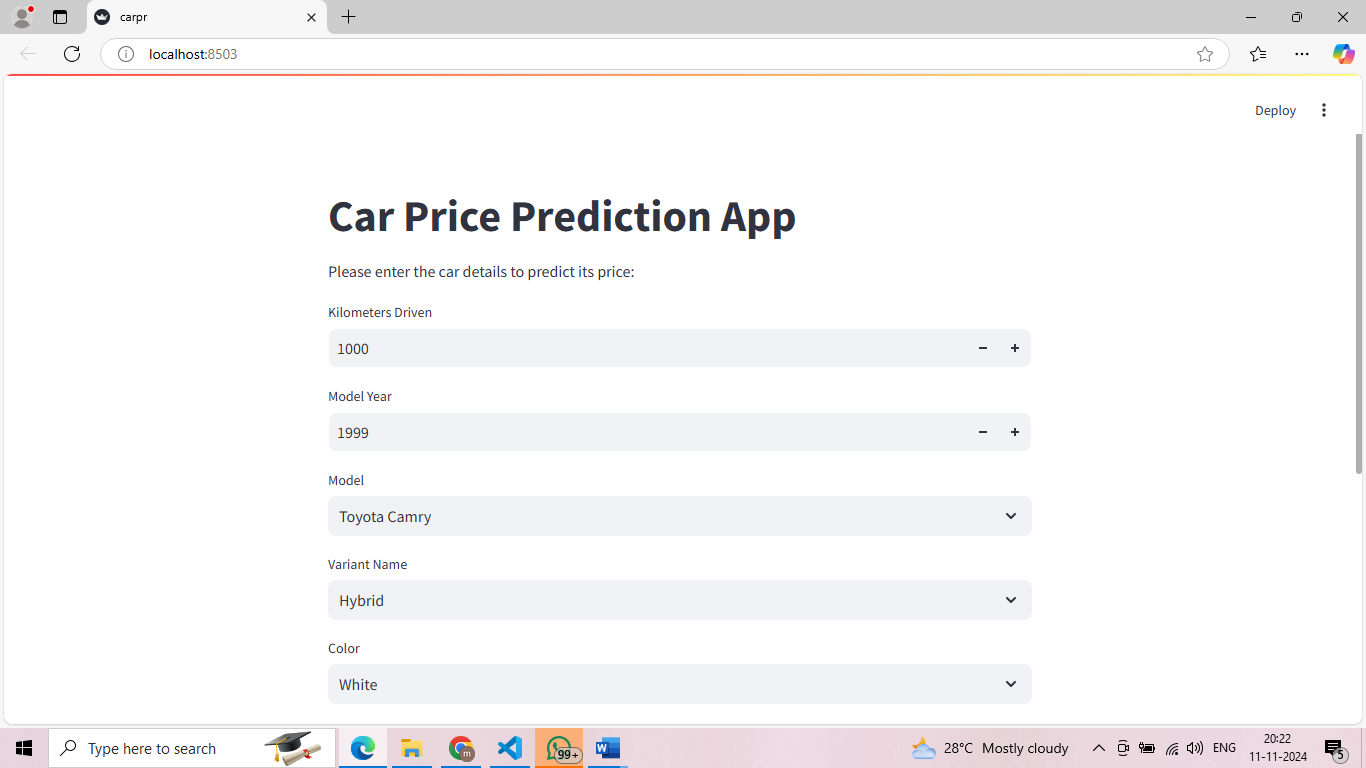


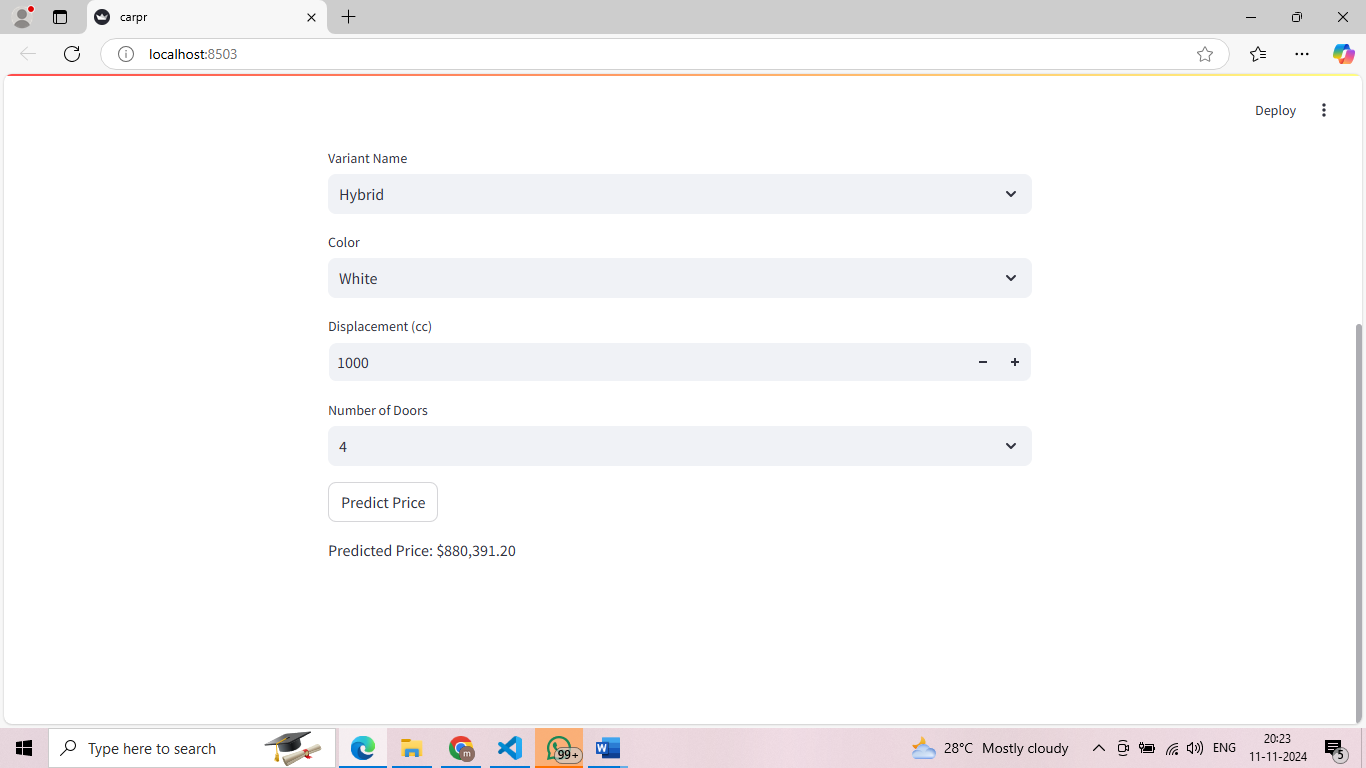
**Deployment**

* 1. **Streamlit Application:** Deploy the final model using Streamlit to create an interactive web application.
     1. Allow users to input car features and get real-time price predictions.
  2. **User Interface Design**: Ensure the application is user-friendly and intuitive.
     1. Provide clear instructions and error handling.







  
**Conclusion**

This project successfully developed a predictive model for used car prices, supported by a structured approach from data preprocessing to model deployment. The following deliverables were achieved:

1. **Source Code for Data Processing and Model Development:** Comprehensive scripts were created to handle data cleaning, preprocessing, feature engineering, and model training, ensuring a smooth and reproducible workflow.
2. **Detailed Documentation:** The methodology, model selection rationale, and evaluation results were thoroughly documented to provide transparency and support further development.
3. **Exploratory Data Analysis (EDA) Reports:** Visualizations and descriptive statistics were used to identify key trends and relationships within the data, informing feature selection and aiding in model refinement.
4. **Final Predictive Model and User Guide:** The final model, optimized through hyperparameter tuning and validated against appropriate metrics, was selected based on performance, simplicity, and interpretability. A user guide was created to assist users in understanding and using the model effectively.
5. **Deployed Streamlit Application:** The model was integrated into a Streamlit web application for seamless, real-time price prediction, enhancing user accessibility and customer experience.

The chosen approach and models were justified based on accuracy, robustness, and ease of use, ensuring that the model performs well across diverse car features and customer inputs. This systematic approach addresses the project’s goal to streamline the used car pricing process and provides a valuable tool for both customers and sales representatives.