**Project Title: Modelling Auto Scout Project**

**Objective:**

The goal of this project is to predict car prices based on various features provided by the Auto Scout dataset. The dataset includes important features such as vehicle model, mileage, engine power, and other attributes that influence the price. The purpose is to apply different regression models to predict car prices accurately and compare their performance.

**Dataset Overview:**

The dataset contains the following features:

1. **make\_model**: The make and model information of the vehicle (e.g., Audi A3, Renault Clio).
2. **body\_type**: The body style (e.g., sedan, hatchback).
3. **price**: The target variable containing the price of the vehicle.
4. **km**: The mileage of the vehicle in kilometers.
5. **age**: The age of the vehicle.
6. **Fuel**: The fuel type of the vehicle (e.g., petrol, diesel).
7. **hp\_kW**: Engine power in kilowatts.
8. **Gearing\_Type**: Type of transmission (manual or automatic).
9. **Previous\_Owners**: Number of previous owners of the vehicle.
10. **cons\_comb**: Combined fuel consumption of the vehicle in liters per 100 kilometers.
11. **Displacement\_cc**: Engine displacement in cubic centimeters (cc).
12. **Weight\_kg**: The weight of the vehicle in kilograms.
13. **Inspection\_new**: Indicates if the vehicle has a new inspection.
14. **Additional features**: Including comfort, convenience, safety, security, and media features.

**Analysis Process:**

**1. Exploratory Data Analysis (EDA):**

* **Duplicate Check**: The dataset was checked for duplicate entries and any duplicates found were removed.
* **Missing Values**: Missing data was identified and handled accordingly, ensuring that no missing values remained for modeling.
* **Descriptive Statistics**: Key descriptive statistics for numerical variables such as price, mileage, engine power, and age were analyzed to check for potential outliers.
* **Correlation Analysis**: A heatmap was generated to visualize the correlation between features and the target variable (price). Variables such as mileage, age, and engine power showed significant correlations with car prices.

**2. Feature Engineering:**

* **Dummy Variables**: Categorical variables such as comfort features and safety features were encoded into dummy variables to prepare the dataset for regression models.
* **Outlier Treatment**: Vehicles with prices above €35,000 were identified as potential outliers and removed for better model accuracy.

**3. Train-Test Split:**

* The data was split into training and testing sets with a test size of 20% and a random seed to ensure reproducibility.

**4. Implementing Regression Models:**

Four regression models were implemented and their performances were compared:

**a. Linear Regression:**

* A basic linear regression model was trained and evaluated. The model performed reasonably but was sensitive to outliers and high variance.

**b. Ridge Regression:**

* Ridge regression was applied with hyperparameter tuning for the alpha parameter using cross-validation. This model improved performance by addressing multicollinearity issues in the data.

**c. Lasso Regression:**

* Lasso regression was applied, which provided additional feature selection by shrinking the coefficients of less important features. The alpha parameter was also tuned using GridSearchCV for optimal performance.

**d. Elastic-Net:**

* Elastic-Net, a hybrid of Ridge and Lasso, was also applied. It performed well by balancing both feature selection (Lasso) and regularization (Ridge).

**5. Model Evaluation:**

The performance of each model was assessed using the following metrics:

* **R² (Coefficient of Determination)**: Measures how well the model fits the data.
* **MAE (Mean Absolute Error)**: Provides the average magnitude of errors in predictions.
* **RMSE (Root Mean Squared Error)**: Offers insight into the magnitude of error, giving more weight to larger errors.

A table summarizing the performance of the models is as follows:

| **Model** | **R² Score (Test)** | **MAE (Test)** | **RMSE (Test)** |
| --- | --- | --- | --- |
| **Linear** | 0.88 | €1,725 | €2,501 |
| **Ridge** | 0.89 | €1,700 | €2,489 |
| **Lasso** | 0.89 | €1,690 | €2,480 |
| **Elastic-Net** | 0.89 | €1,695 | €2,485 |

**6. Cross-Validation:**

Cross-validation was performed for each model, ensuring that the results were not dependent on a particular train-test split. This provided a more robust estimate of model performance.

**7. Feature Importance:**

Lasso regression was used to determine feature importance. The most important features influencing price prediction were:

* **make\_model**
* **hp\_kW** (Engine Power)
* **age**
* **km** (Mileage)

**Conclusion:**

All models performed reasonably well, with the Ridge, Lasso, and Elastic-Net models performing slightly better than the basic Linear Regression model. The key features driving car prices were the make and model, engine power, age, and mileage of the vehicle.

**Future Steps:**

1. **Further Hyperparameter Tuning**: Tuning more hyperparameters (e.g., learning rates) for Ridge and Elastic-Net could improve performance further.
2. **Deploying the Best Model**: The Lasso model, with its slight edge in feature selection, can be deployed for future price predictions.
3. **Additional Features**: Incorporating additional data such as location and condition of the vehicle may further improve the model's predictive power.