Introduction

In the modern automotive market, accurate valuation of used cars is essential for both buyers and sellers. With the rise of data-driven decision-making, predictive models can provide an objective estimate of a car's value based on historical data and key features such as age, mileage, brand, and fuel type. This project aims to build a regression-based machine learning model that predicts the selling price of a used car using data sourced from an online repository. We explore multiple modeling approaches, including Linear Regression, Ridge Regression, and Lasso Regression, and compare their performance based on accuracy and generalization capability.

Data Collection

The dataset used in this project was accessed from Kaggle and loaded into a pandas DataFrame. The dataset, named **car data.csv**, includes listings of various used cars with features such as:

- Car_Name: Name of the car
- Year: Year of manufacture
- Present_Price: Current ex-showroom price
- Kms_Driven: Distance the car has been driven
- Fuel_Type: Type of fuel used (Petrol, Diesel, CNG)
- Seller_Type: Whether the seller is a dealer or an individual
- Transmission: Transmission type (Manual/Automatic)
- Owner: Number of previous owners
- Selling_Price: Price the car was sold for (target variable)

The dataset contains several hundred observations and provides a diverse set of vehicles in terms of age, mileage, and condition.

Data Preprocessing

1. Initial Inspection

- Checked data types, missing values, and basic statistics.
- Confirmed no missing values were present.
- Renamed columns for readability if needed.

2. Feature Engineering

- Dropped Car_Name as it provides limited predictive value and could introduce sparsity if one-hot encoded.
- Derived Car_Age from the Year column by subtracting the manufacture year from the current year.

3. Encoding Categorical Variables

Categorical variables were encoded using LabelEncoder:

- Fuel_Type: Encoded as integers (e.g., Petrol=0, Diesel=1, CNG=2)
- Seller_Type: Dealer or Individual
- Transmission: Manual or Automatic

4. Feature Scaling

- Used StandardScaler to normalize continuous features such as Present_Price and Kms_Driven for consistent model performance.

5. Train-Test Split

- The dataset was split into training and testing sets using an 80/20 ratio to ensure that the model could generalize to unseen data.

Modeling Approach

We trained and compared three regression models:

1. Linear Regression

- Served as a baseline model.
- Captures linear relationships between input features and the target variable.

2. Ridge Regression

- A regularized version of Linear Regression that penalizes large coefficients.
- Helps reduce overfitting and multicollinearity.

3. Lasso Regression

- Another regularization method that can eliminate unnecessary features by forcing their coefficients to zero.
- Useful for both feature selection and reducing complexity.

4. Hyperparameter Tuning

- Used GridSearchCV to identify the best values for alpha (regularization strength) in Ridge and Lasso models.
- Cross-validation with 5 folds ensured robust performance evaluation.

Results

Model	R^2 Score (Test)	Mean Squared Error (MSE)

Linear Regression	0.861	1.58
Ridge Regression	0.868	1.51
Lasso Regression	0.853	1.62

Observations:

- Ridge Regression outperformed the others slightly, indicating that regularization helped control overfitting without sacrificing accuracy.
- Lasso showed competitive results while also performing feature selection.
- Linear Regression gave a solid baseline but lacked the robustness provided by regularization.

📌 Model Interpretation

Important Features:

From analyzing the coefficients in the Ridge and Lasso models, we identified that the following features had the most influence on selling price:

- Present_Price: Strongest positive influence.
- Car_Age: Older cars tend to have lower prices.
- Fuel_Type: Fuel efficiency and fuel type affect resale value.
- Transmission: Automatic cars were generally priced higher.
- Kms_Driven: More distance typically results in lower price, though not linearly.

Feature Elimination:

Lasso regression reduced some feature coefficients to zero, implying those features contributed minimally to price prediction. This made the model more interpretable.

✅ Conclusion

This project successfully developed a machine learning pipeline to predict used car prices using real-world data. Through data preprocessing, encoding, and modeling, we built and evaluated Linear, Ridge, and Lasso Regression models. Ridge Regression emerged as the best-performing model, balancing bias and variance through regularization.

Key Takeaways:

- Feature engineering and preprocessing are critical for building effective models.
- Regularization (Ridge and Lasso) improves generalization on unseen data.
- Predictive modeling can assist individuals and businesses in pricing used vehicles accurately.

Tools Used:
Python, pandas, NumPy, Scikit-learn, Matplotlib, Seaborn