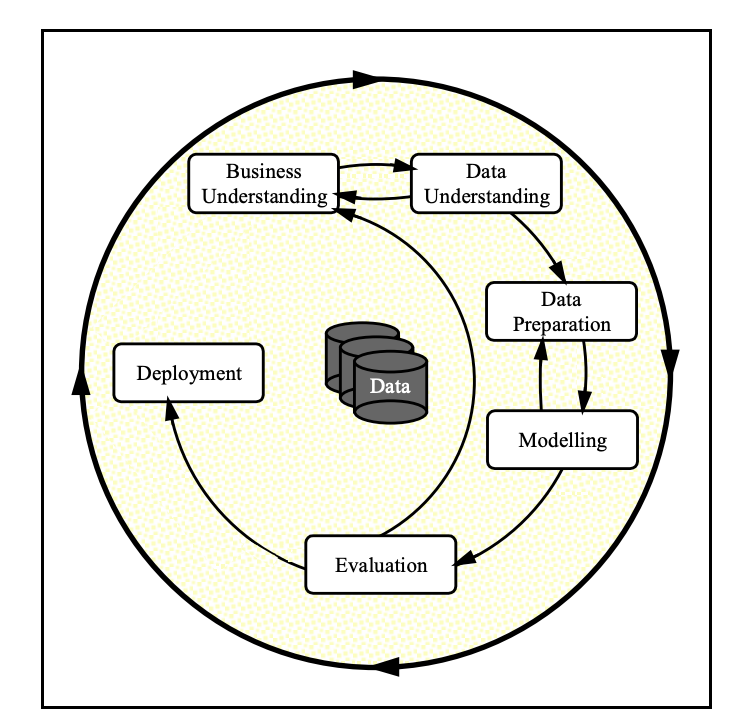
# **Used Car sales - what factors make a car more or less expensive**



**CRISP-DM Framework(Cross-Industry Standard Process for Data Mining): This application makes use of CRISP-DM framework.**

****

**To frame the task, throughout these practical applications, we will refer to a standard process in the industry for data projects called CRISP-DM. This process provides a framework for working through a data problem. Our first step in this application will be to read through a brief.**

### **1. Business Understanding/Project objective:**

**In this application, we explore a dataset from Kaggle. The original dataset contained information on 3 million used cars. The provided dataset contains information on 426K cars to ensure the speed of processing.**

**As a result of our analysis, we should provide clear recommendations to your client—a used car dealership—as to what consumers value in a used car.**

### **2. Data Understanding**

**The dataset contains detailed information about various vehicles available for sale across different regions, primarily in Alabama and Arizona. Each entry includes:**

* **ID: Unique identifier for each vehicle.**
* **Region: Location where the vehicle is being sold.**
* **Price: Listed price of the vehicle.**
* **Year: Manufacturing year of the vehicle.**
* **Manufacturer: Brand of the vehicle.**
* **Model: Specific model of the vehicle.**
* **Condition: Condition of the vehicle (e.g., good, excellent, fair).**
* **Cylinders: Number of engine cylinders.**
* **Fuel: Type of fuel used (e.g., gas, diesel).**
* **Odometer: Mileage of the vehicle.**
* **Title Status: Legal status of the vehicle's title (e.g., clean, rebuilt).**
* **Transmission: Type of transmission (e.g., automatic, manual).**
* **VIN: Vehicle Identification Number.**
* **Drive: Drivetrain type (e.g., 4WD, RWD).**
* **Size: Size category of the vehicle (e.g., full-size, mid-size).**
* **Type: Body type of the vehicle (e.g., SUV, truck).**
* **Paint Color: Color of the vehicle.**
* **State: State where the vehicle is located.**

**Earliest Year: The oldest vehicle in the dataset is from 1966. Latest Year: The most recent vehicle is from 2021.**

### **3. Data Preparation**

**Missing Values: Identify and handle missing values in columns such as manufacturer, model, condition, cylinders, fuel, odometer, title status, transmission, VIN, drive, size, type, and paint color. Data Types: Ensure all columns have appropriate data types (e.g., numerical for price and odometer, categorical for condition and type). Normalization: Normalize numerical values like price and odometer to ensure consistency. Categorical Encoding: Convert categorical variables into numerical values for modeling purposes.**

* **vehicles.csv: Raw data.**
* **vehicles\_no\_null.csv : Removed nan and missing values filled with IterativeImputer.**
* **vehicles\_pre\_processed.csv : Removed outliers and data is promed for modeling.**

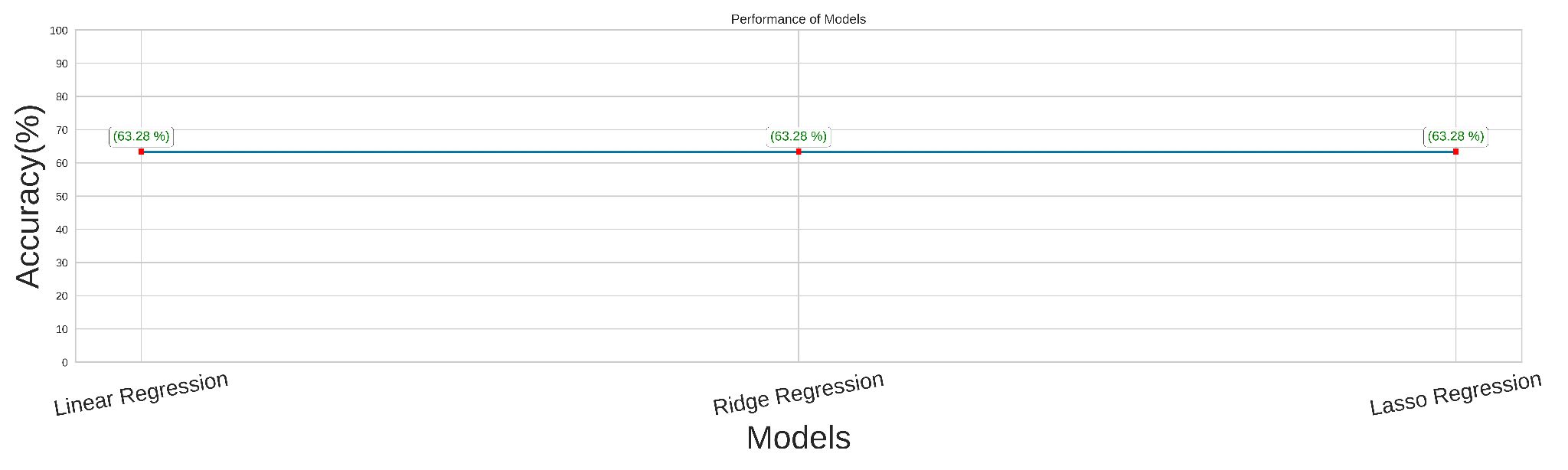
### **4. Modeling**

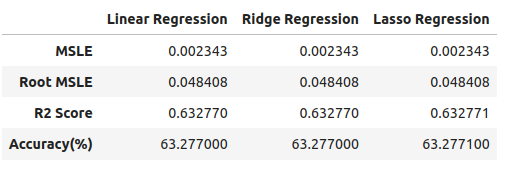
**Models used to predict vehicle prices based on features like year, manufacturer, model, condition, mileage, and location were Linear Regression, Ridge Regression and Lasso Regression.**

### **5. Evaluation**

**Used metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) for regression models, and accuracy, precision, recall for classification models. Feature Importance: Analyze feature importance to understand which attributes most significantly impact vehicle prices and conditions.**

## **Score Table**



****

**Linear Regression has the lowest MSLE (0.0023433223), indicating slightly better performance in predicting the target variable compared to Ridge and Lasso. Linear Regression has the smallest value (0.0484078745), suggesting it is marginally better at minimizing logarithmic errors. Lasso Regression has the highest R2 (63.2771%), but the difference is negligible (0.0001%) compared to the other models. Linear Regression is the better model because: It has the lowest MSLE and Root MSLE, indicating better overall predictive accuracy. The R2 Score for all models is nearly identical, with no practical advantage in using Ridge or Lasso in this case.**

## **Conclusion:**

**All models suggest that "Year" and "Odometer" are the features that contributes to the model. The used car dealer should take these two factors into consideration for selling the cars.**