# Project title

Car Dekho – Used Car

Price Prediction

Submitted by ,

R . Gayathri

**Problem Statement:**

**Objective:**

The aim is to enhance the customer experience and streamline the pricing process by leveraging machine learning. Need to create an accurate and user friendly streamlit tool that predicts the prices of used cars based on various features. This tool should be deployed as an interactive web application for both customers and sales representatives to use seamlessly.

**Project Scope:**

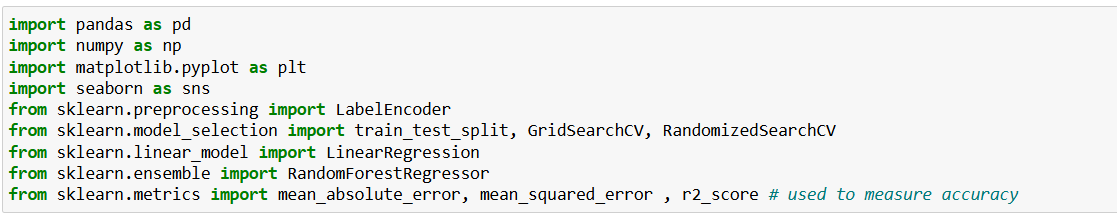
This project includes historical data on used car prices from CarDekho, including various features such as make, model, year, fuel type, transmission type and other relevant attributes from different cities. The task is to develop a machine learning model that can accurately predict the prices of used cars based on these features. The model should be integrated into a Streamlit – based web application to allow users to input car details and receive an estimated price instantly.

**Steps involved :**

1. **Data Processing**
   1. Import and Concatenate
   2. Handling missing values
   3. Standardising Data formats
   4. Encoding Categorial variables
   5. Normalizing numerical features
   6. Removing Outliers
2. **Exploratory Data Analysis(EDA)**
   1. Descriptive statistics
   2. Data Visualization
   3. Feature Selection
3. **Model Development**
   1. Train – Test Split
   2. Model Selection
   3. Model Training
   4. Hyperparameter Tuning
4. **Model Evaluation**
   1. Performance Metrics
   2. Model Comparison
5. **Optimization**
   1. Feature Engineering
   2. Regularization
6. **Deployment**
   1. Streamlit Application
   2. User Interface Design

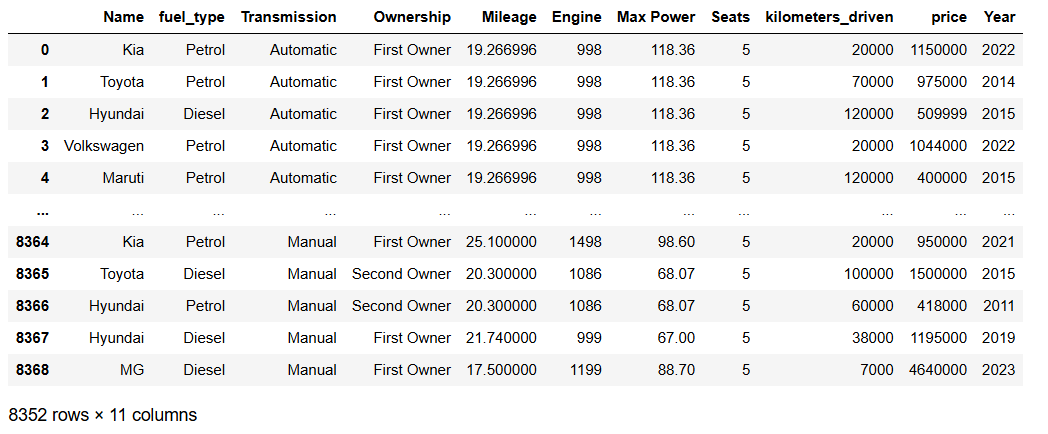
Steps involved in the model building process:

Packages used :



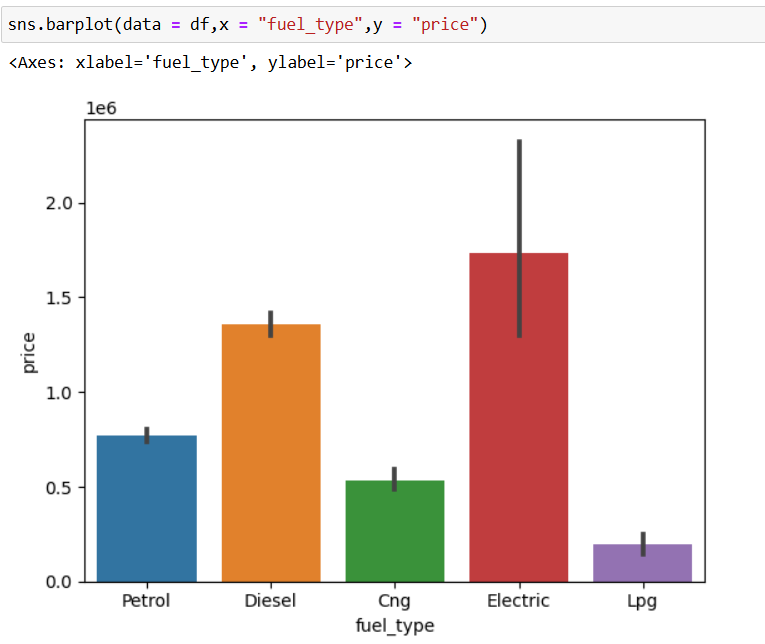
1. Pandas
2. Numpy
3. Matplotlib
4. Seaborn
5. Scikit – learn

Cleaned Dataset :



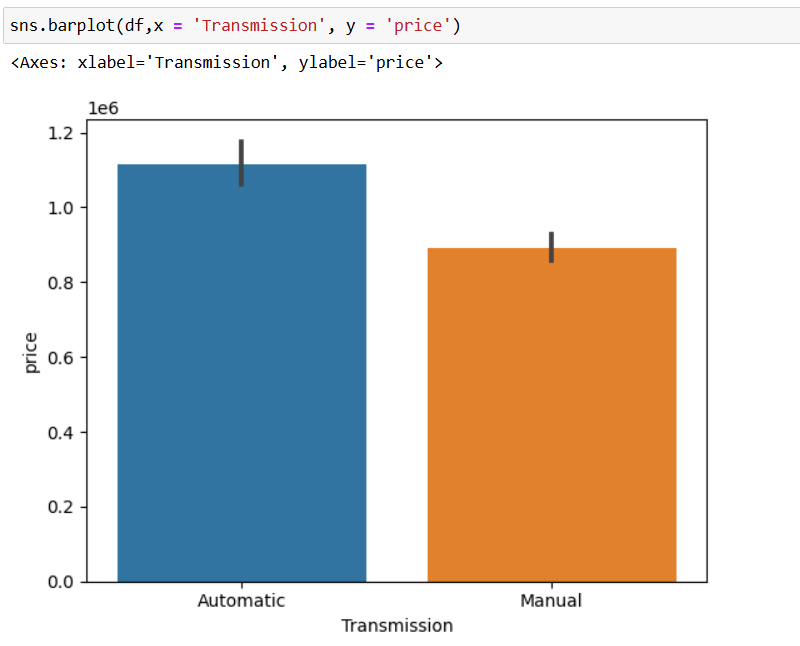
Exploratory Data Analysis:

Relation between Fuel type and price:



Based on the analysis, the price of Electric cars are higher than the others. Diesel is the Second High cost fuel type car .

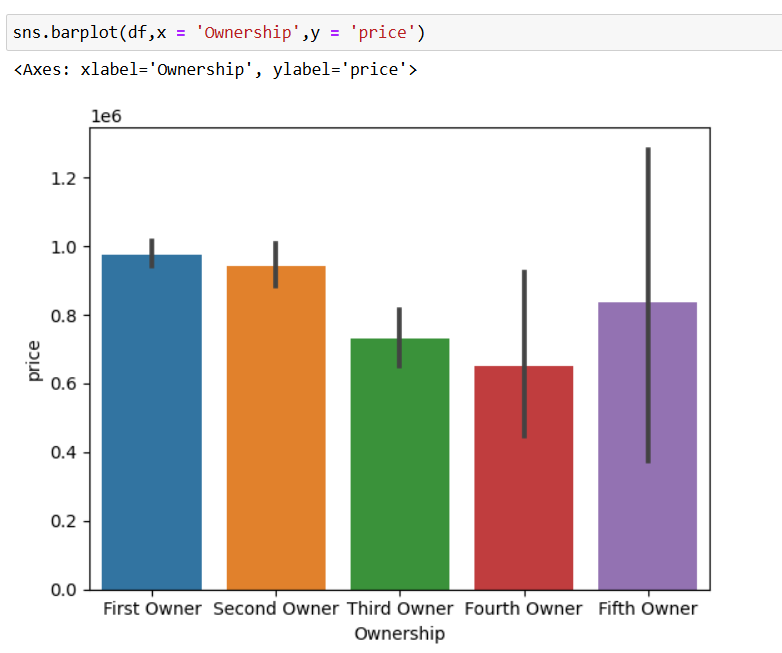
Relationship between Transmission and Price:



From the plot , we can see that Automatic transmission vehicles are high in cost .

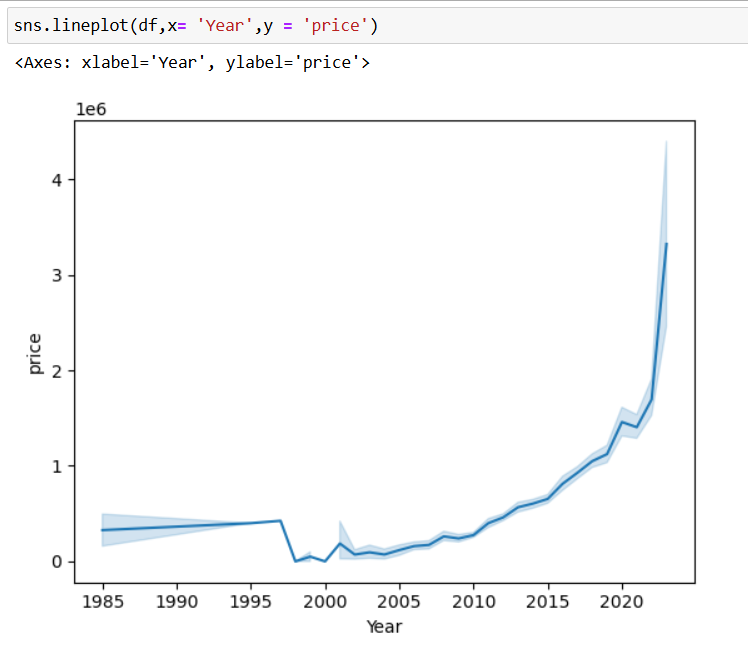
It has the highest price of above 10,00,000.

Relationship between Onwership and Price:



From the plot , The price of used cars bought from first owners has the highest selling price , the next one is from Second owner.

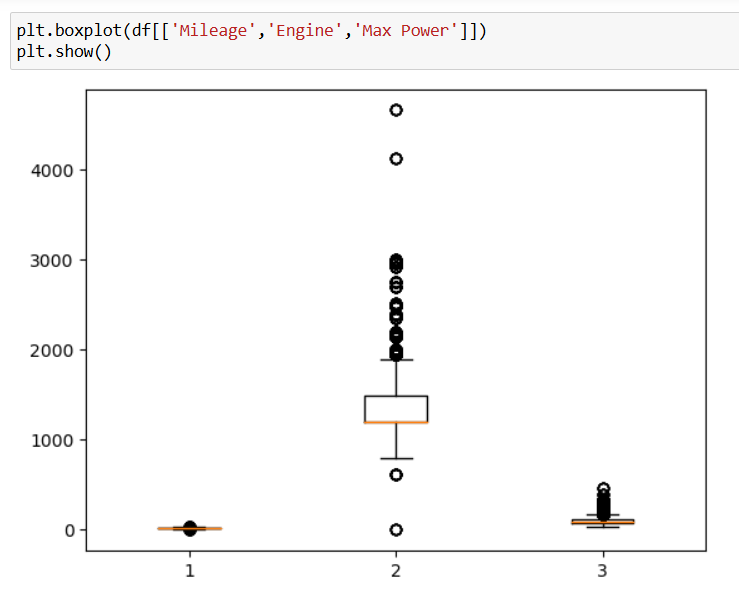
Trend of price :



From the plot , there is a trend which increases with the making year of car.

The cars which is made after 2020 has the highest price .

Outlier Detection:



Engine data has more outliers , it is removed using IQR (Inter Quartile Range) method.

**Model used :**

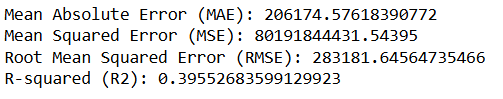
The models used in this project are

1. Linear Regression
2. Random Forest Regressor
3. XGBoost

**Linear Regression :**

1) **Overview**: Linear Regression is a simple and widely used algorithm for predicting a continuous target variable. It assumes a linear relationship between the input features and the target variable.

2) **How It Works**: The model finds the best-fitting line, represented by the equation y=mx+by = mx + by=mx+b (for one feature), where yyy is the target, mmm is the slope, xxx is the feature, and bbb is the intercept. For multiple features, it extends to y=w1x1+w2x2+⋯+wnxn+by = w\_1x\_1 + w\_2x\_2 + \dots + w\_nx\_n + by=w1​x1​+w2​x2​+⋯+wn​xn​+b.



The accuracy of linear regression model is too low.

It’s not a good model . So another model is used to improve the accuracy.

**Random Forest Regressor:**

1) **Overview**: Random Forest is an ensemble learning method that combines multiple decision trees to improve accuracy and stability. In regression tasks, it averages the predictions of individual trees to output a final prediction.

2) **How It Works**: Each tree in the forest is trained on a random subset of the data and features, leading to diverse predictions. For regression, the final prediction is the average of all tree predictions, which reduces variance and enhances generalization.

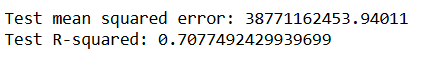
Accuracy :



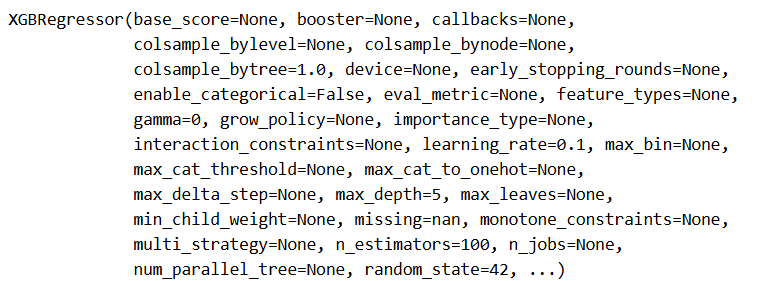
**XGBoost :**

1) **Overview**: XGBoost (eXtreme Gradient Boosting) is a powerful, efficient implementation of gradient boosting for supervised learning. It builds an ensemble of trees iteratively, where each tree corrects errors from the previous ones, enhancing model accuracy.

2) **How It Works**: Each new tree attempts to minimize the residual errors of the previous trees by assigning higher weights to harder-to-predict instances. This results in a series of trees, each improving on the last, leading to strong overall predictions.

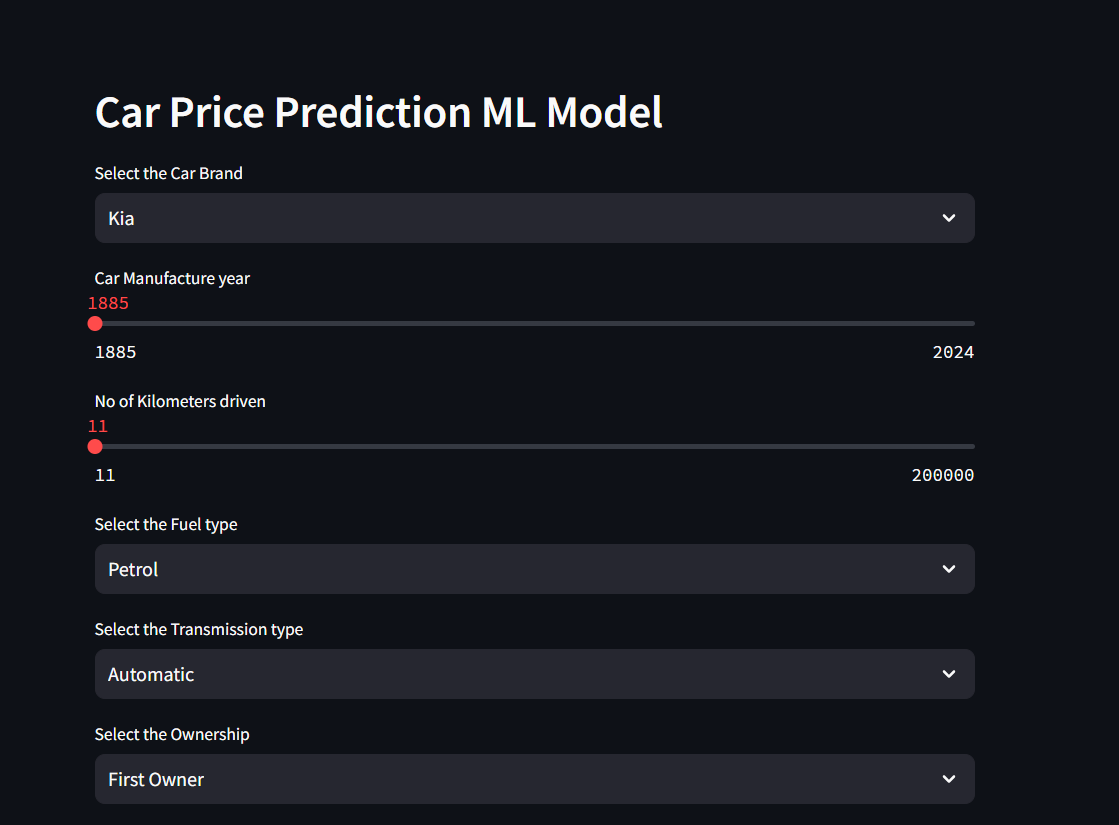


**Final model :**



**Deployment :**

The model is deployed using Visual Studio code editor.

Model link : <http://192.168.211.160:8505>

**Thank You**