1. Employing Supervised & Unsupervised Learning

1.1. Introduction

Mock cleaned dataset consisting of second-hand car sales in UK has been provided, with a shape of (50000,7). Dataset consist of following features:

	Feature Name	Description of Feature	Feature Type
Independent Feature	Price	Price of the car in British Pounds (GBP)	Numerical
Dependent Feature	Manufacturer	Name of the manufacturer that produced the car	Categorical
	Model	Name of the model of the car	Categorical
	Engine Size	Size of the engine, in liters	Numerical
	Fuel Type	Type of fuel that the engine uses.	Categorical
	Year of Manufacture	Year of the car's manufacture	Numerical
	Mileage	Total distance the car has traveled, measured in miles	Numerical

Goal of this study is to do comparative analysis of accuracy of different models by predicting the price.

1.2. Supervised Learning

For evaluation, mean absolute error (MAE) (Bernico, n.d.) and Coefficient of Determination (R^2 score) (Anon., n.d.) is checked. A model is said to be robust if its Mean absolute error is low and Coefficient of Determination is closer to 1.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \qquad R^2 = 1 - \frac{RSS}{TSS}$$

1.2.1. Comparing linear & Non-linear model using single Numerical input feature

Dataset has 3 dependent numerical features, namely Engine size, Year of manufacture & Mileage. Simple linear regression and polynomial regression were applied on each set (Price vs. Mileage, Price vs. Engine Size & Price vs. Year of Manufacture) to predict price.

Linear Regression		
Features	**MAE**	**R^2**
Price vs Mileage	7964.78467	0.40131391
Price vs Engine Size	10817.49156	0.150625625
Price vs Year of Manufacture	7031.039209	0.511086524

Polynomial Regression (Degree 2)		
Features	**MAE**	**R^2**
Price vs Mileage	6409.911605	0.52235759
Price vs Engine Size	10807.26235	0.151263076
Price vs Year of Manufacture	5387.109075	0.609419402

From the result of both Linear Regression and Polynomial Regression (Degree 2), Price vs Year of Manufacture feature's MAE is lowest & R^2 score is highest. Therefore, we can say that **Year of Manufacture** variable is the best predictor for a car's price. Also, R^2 score of each feature in polynomial regression is better than Linear Regression thus we can conclude that **Polynomial Regression** is a better fit for price prediction.

1.2.2. Considering Multiple Numerical Variables

Numerical features like Engine size, year of manufacture & Mileage are all together used to predict the price. Both linear and polynomial model were applied.

Linear Regression		
Features	**MAE**	**R^2**
Price vs Mileage, Engine Size & Year of Manufacture	6091.458142	0.671456306

Polynomial Regression (Degree 2)		
Features	**MAE**	**R^2**
Price vs Mileage, Engine Size & Year of Manufacture	3196.824934	0.891991018

Inclusion of multiple numerical features significantly improves the accuracy of model. Previously for linear model highest R^2 score of 0.51 was recorded for the case of Year of Manufacture, now inclusion of all numerical feature improves the R^2 score to 0.67.

1.2.3. Considering All Input Features

In this case, in addition to Numerical features, Categorical features after label encoding are also considered to predict the price.

Linear Regression		
Features	**MAE**	**R^2**
Price vs All other features	6076.3458	0.671990366

Polynomial Regression (Degree 2)		
Features	**MAE**	**R^2**
Price vs All other features	2989.44386	0.906675288

Decision Tree Regression		
Features	**MAE**	**R^2**
Price vs All other features	486.1968	0.995799205

Random Forest Regression		
Features	**MAE**	**R^2**
Price vs All other features	332.2704	0.998246961

Slight improvement in accuracy is observed when including categorical features. Previously for linear case when only numerical features were considered R^2 score was 0.6714 which increases to 0.6719 when all features are included. For polynomial R^2 score increase from 0.891 to 0.906. Apart from linear and polynomial models, Decision Tree and Random Forest were also evaluated. Based on the results of all models, **Random Forest gives the best accuracy** with R^2 score of 0.99824.

1.2.4. Implementation of Artificial Neural Network

An Artificial Neural Network is made which takes 6 inputs both numerical and categorical features and give 1 output which is the price.

1.2.4.1. ANN Architecture

Constructor Stage	Compilation Stage
➢ Input layer with 64 units and ReLU activation ➢ To prevent overfitting, a drop layer with a rate of 0.2 ➢ 1 Hidden layer with 64 units and ReLU activation ➢ Output layer with 1 unit and Linear activation	➢ Adam Optimizer with default learning rate ➢ Early stopping with patience of 20 ➢ Validation dataset split of 10% ➢ Epochs equal to 200

Artificial Neural Network		
Features	**MAE**	**R^2**
Price vs All other features	1460.894	0.972150981

With a simple Artificial Neural Network consisting of default settings, R^2 Score of 0.972 is obtained which is less than previously used Random Forest Regression model where obtained R^2 score was 0.99824.

1.2.4.2 Hyperparameter Tuning

To obtain the best performance, hyperparameters were tunned using Keras Turner Random Search. Parameters choices were given based on which Random Search gives the best possible combination of parameters that maximizes output.

Choices given to Random Search

- Choice of hidden layers between 1 to 3
- Vary neurons from 32 to 128 in each hidden layer with step size of 32
- Activation Function is set as ReLU
- Dropout choices: 10%, 20% or 30%
- Output layer with 1 unit and linear activation
- Learning rate choices: 0.01 or 0.001
- Epochs set as 50
- Validation split of 0.1
- Early stopping with patience of 20

Tunned parameters provided by Random Search

Constructor Stage	Compilation Stage
➢ Input layer with 128 units and ReLU activation ➢ To prevent overfitting, a drop layer with a rate of 0.1 ➢ 2 Hidden layers with 128 & 64 units respectively and ReLU activation ➢ Output layer with 1 unit and Linear activation	➢ Adam Optimizer with learning rate of 0.01 ➢ Early stopping with patience of 20 ➢ Validation dataset split of 10% ➢ Epochs equal to 200

Artificial Neural Network (Hyperparameter Tunning)		
Features	**MAE**	**R^2**
Price vs All other features	899.512	0.992128014

After hyperparameters tunning, R^2 Score of 0.9921 was obtained. As compared to the models used earlier, ANN performed well except it fall short of Random Forest Model, which delivered R^2 score of 0.9982.

1.2.5. Best Model

Figure 1 shows the R2 score of the tested models, based on the R2 scores, Random Forest performs the best predictions.

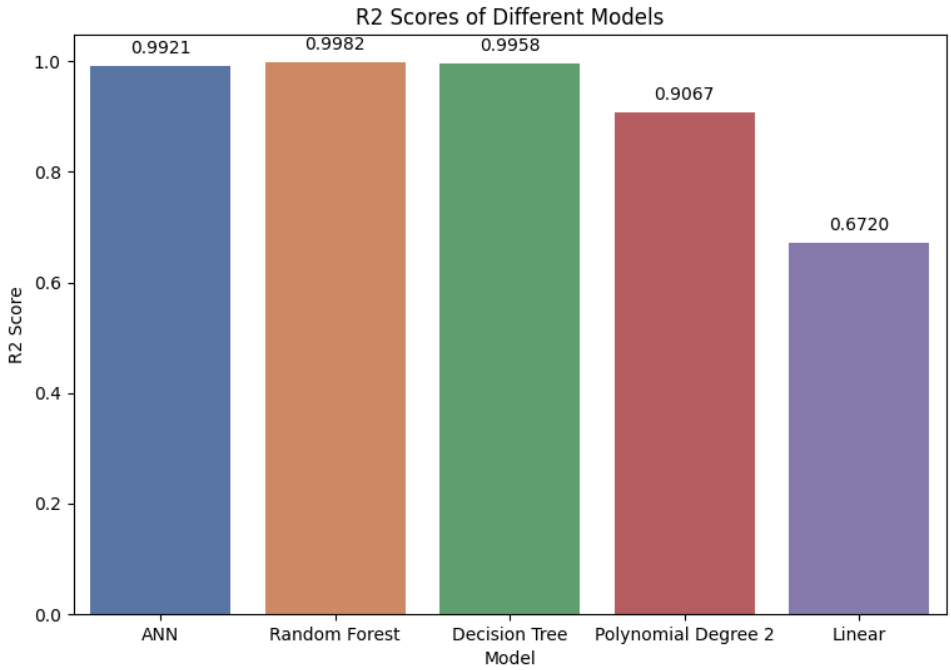

Figure 1, R2 score of different models

Figure 2 shows the graph of actual vs predicted price. Graph illustrates the wellness of model fit.

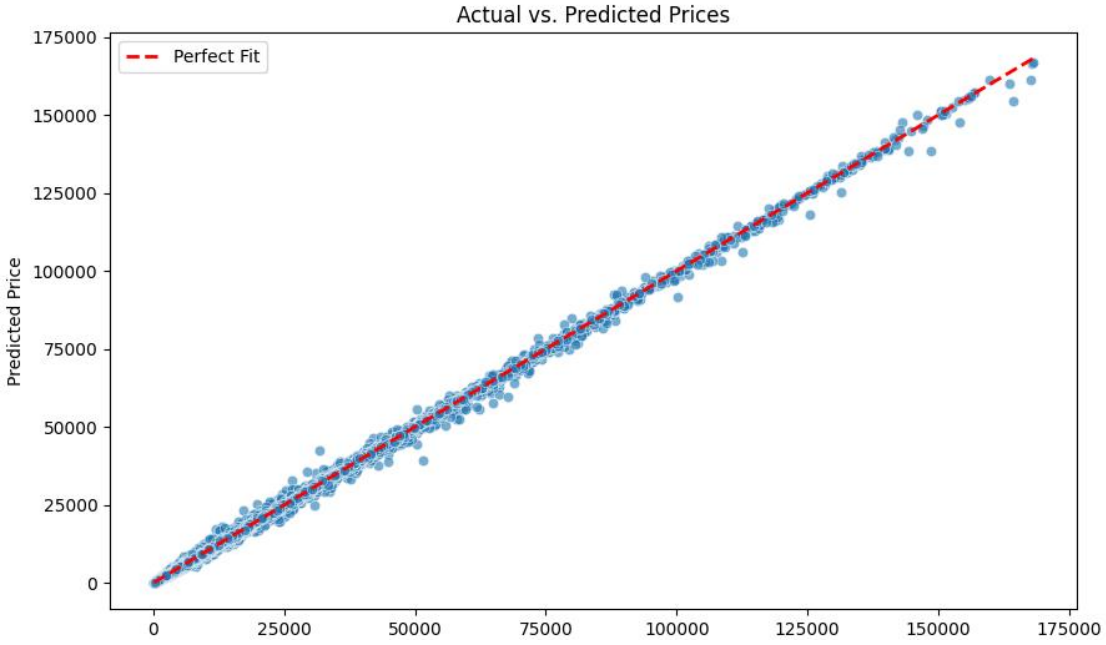

Figure 2, Random Forest, Actual Price vs. Predicted Price

Figure 3 shows the residuals plot, where it can be seen that residuals are randomly scattered around the horizontal axis without forming any pattern. This shows the model has captured all the underlying patterns in the data.

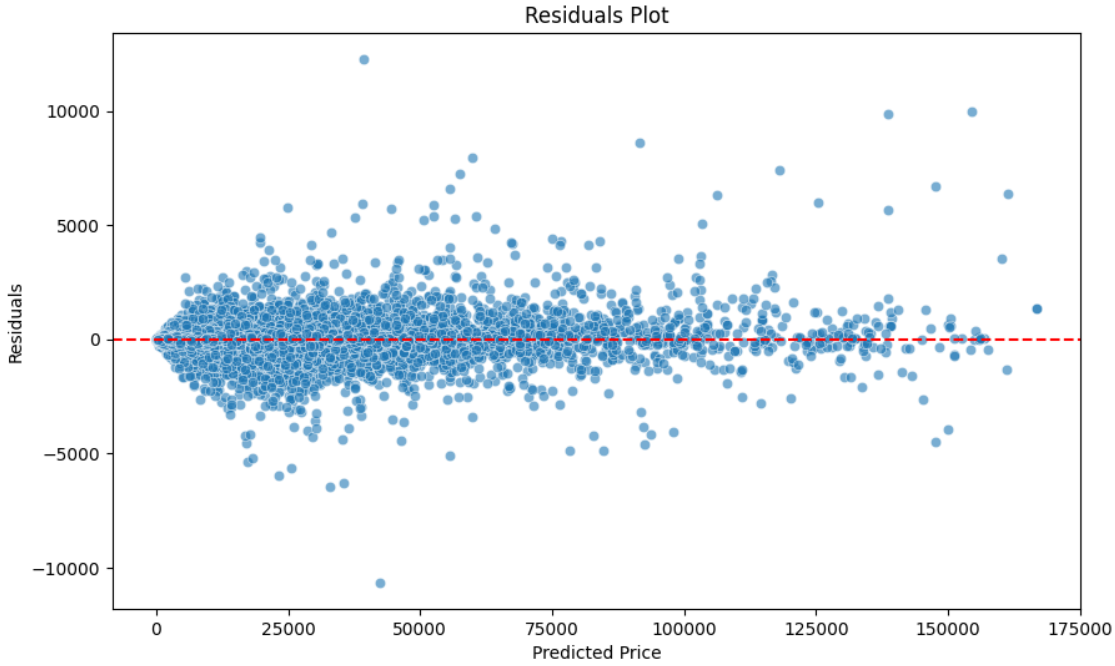

Figure 3, Random Forest, Residuals Plot

1.3. Unsupervised Learning

For accuracy evaluation of model, Davies-Bouldin (DB) Index and Silhouette Coefficient are used. A model is said to be robust if its Davies-Bouldin Index is small and Silhouette Coefficient (S) (Bonnin, n.d.) is close to +1.

$$DB = \frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j} R_{ij} \qquad s = \frac{b-a}{\max(a,b)}$$

1.3.1. K-Means Clustering

1.3.2. Optimal Number of Clusters (k)

Optimal number of clusters is determined using Elbow method & Silhouette coefficient. K is selected where minimum inertia and maximum Silhouette is obtained as shown in graphs below. This Point balances the trade-off between having too many clusters and too few clusters. Since the dataset is labelled so deliberately ignoring the price column to make dataset unlabeled.

Engine Size & Year of Manufacture (k = 3)

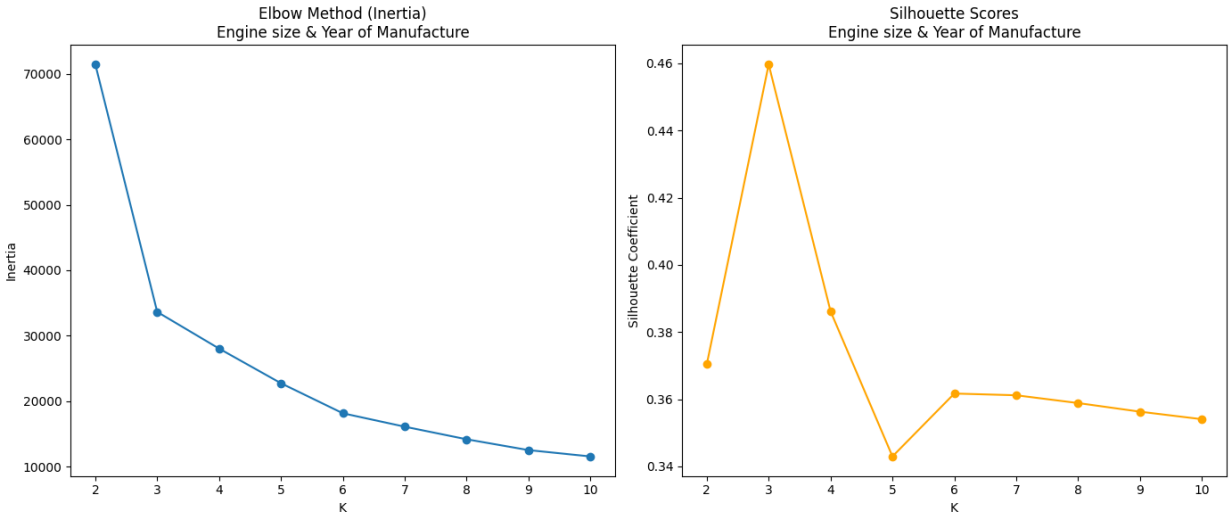

Figure 4: Elbow Method & Silhouette Score of Engine Size & Year of Manufacture

Year of Manufacturer & Mileage (k = 2)

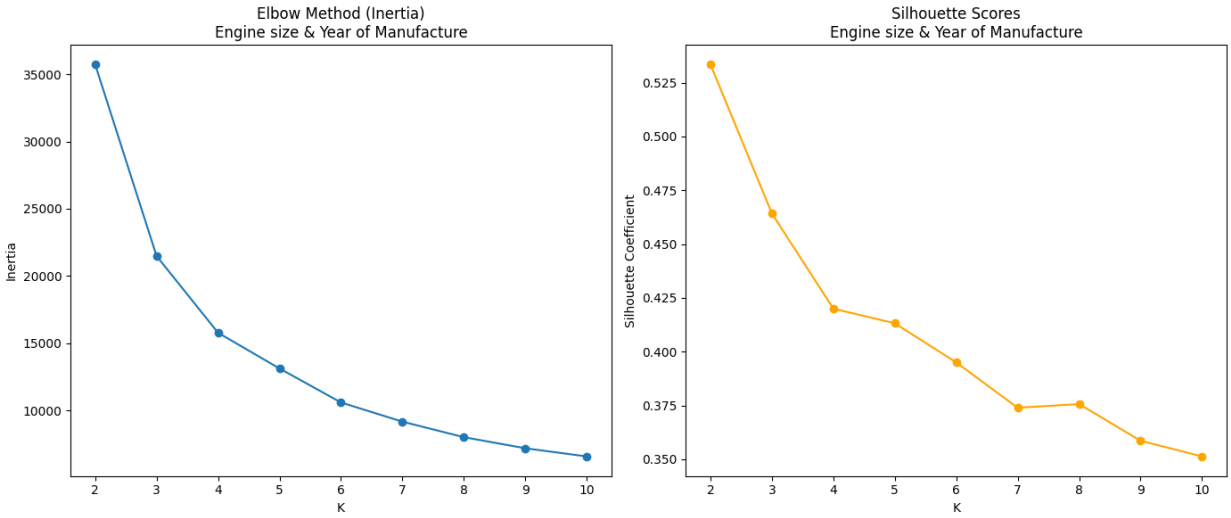

Figure 5: Elbow Method & Silhouette Score of Year of Manufacture & Mileage

Mileage & Engine Size (k = 3)

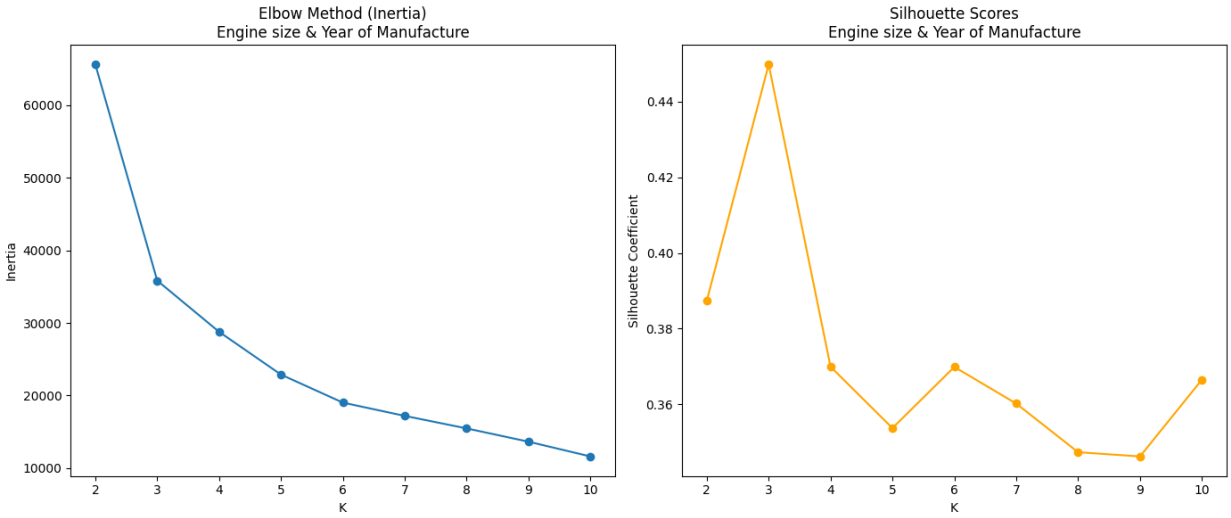

Figure 6: Elbow Method & Silhouette Score of Engine Size & Mileage

Engine Size, Year of Manufacture & Engine Size (k=3)

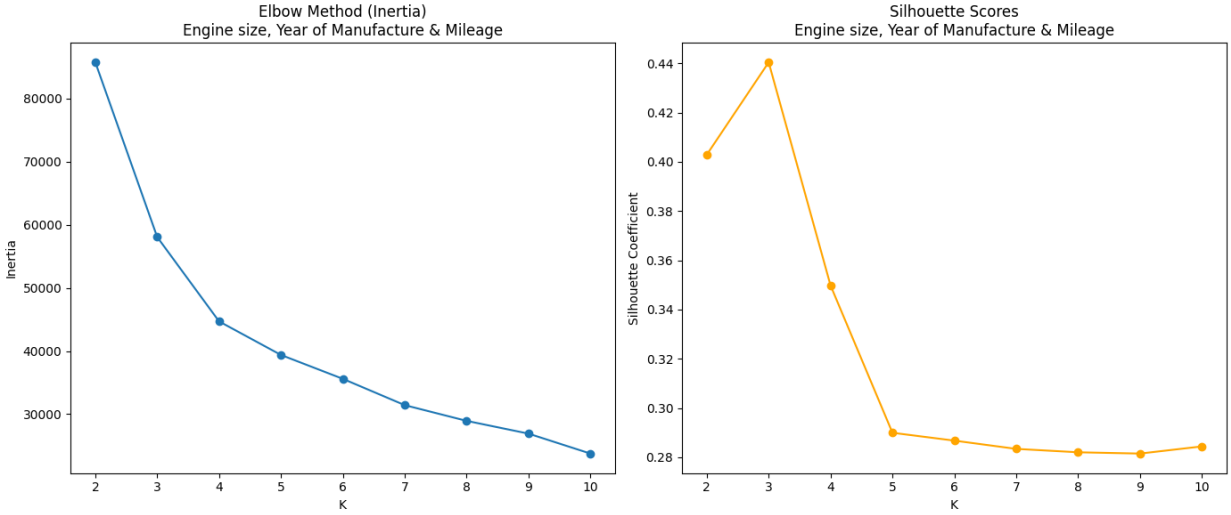

Figure 7: Elbow Method & Silhouette Score of Engine Size, Year of Manufacture, Mileage

Combination	Davies-Bouldin Index	Silhouette Coefficient
Engine size & Year of Manufacturer	0.752500164	0.459695553
Year of Manufacturer & Mileage	0.658826523	0.533497893
Mileage & Engine size	0.776741057	0.449839987
Engine Size, Mileage & Year of Manufacture	0.805472041	0.440399475

Combination of Year of Manufacturer & Mileage gives the best result as its DB score is lowest and Silhouette Coefficient is close to +1.

1.3.3. Gaussian Mixture Model

	K- Means		Gaussian Mixture	
Combination	**DB Index**	**Silhouette Coefficient**	**DB Index**	**Silhouette Coefficient**
Engine size & Year of Manufacturer	0.752500164	0.459695553	0.753737654	0.459851792
Year of Manufacturer & Mileage	0.658826523	0.533497893	0.708538567	0.474691579
Mileage & Engine size	0.776741057	0.449839987	0.784722482	0.446917018
Engine Size, Mileage & Year of Manufacture	0.805472041	0.440399475	0.877884627	0.390378180

For same number of clusters, DB index of K-Means clustering is lowest for all combination. Therefore, **K-Means produce best clustering**.

1.4. Conclusion

For labelled data set, supervised learning is usually applied for prediction purpose. Several Supervised Machine Learning algorithms were applied and appropriate evaluation metrics were used to evaluate the model performance. **Random Forest Regression model** turns out to be the best predictor with R^2 value of 0.9982. Clustering was also performed on the same dataset using different unsupervised algorithms by ignoring the price column and considering dataset as unlabeled. **K-Means** provides the best clustering in this case.