

CMP6202

Artificial Intelligence and Machine Learning

2024–2025

Title: Car Price Prediction Using Machine Learning

Forename	Surname	Student Id	Module Leader	ML Model(s) developed
Talha Ahmed	Risath	22222895	Nouh Elmitwally	Car Price Prediction

Table of Contents

1. INTRODUCTION	4
1.1 DATASET IDENTIFICATION	4
1.2 SUPERVISED LEARNING TASK IDENTIFICATION	5
2. EXPLORATORY DATA ANALYSIS	5
2.1 QUESTION IDENTIFICATION	6
2.2 SPLITTING THE DATASET	6
2.3 EXPLORATORY DATA ANALYSIS PROCESS AND RESULTS	6
2.4 EDA CONCLUSIONS	12
3. EXPERIMENTAL DESIGN	13
3.1 IDENTIFICATION OF CHOSEN SUPERVISED LEARNING ALGORITHM.....	13
3.2 IDENTIFICATION OF APPROPRIATE EVALUATION TECHNIQUES.....	14
3.3 DATA CLEANING AND PRE-PROCESSING TRANSFORMATIONS	14
3.4 LIMITATIONS AND OPTIONS	15
4. PREDICTIVE MODELLING / MODEL DEVELOPMENT	16
4.1 THE PREDICTIVE MODELLING PROCESS.....	16
4.2 EVALUATION RESULTS ON “SEEN” DATA	16
5. EVALUATION AND FURTHER MODELLING IMPROVEMENTS	18
5.1 ALTERNATIVE MODEL IMPLEMENTATION: RANDOM FOREST	18
5.2 ALTERNATIVE MODEL IMPLEMENTATION: GRADIENT BOOSTING METHOD	18
5.3 COMPARISON OF MODELS.....	19
6. CONCLUSION	20
6.1 SUMMARY OF RESULTS.....	20
6.2 REFLECTION ON INDIVIDUAL LEARNING	20
6.3 PROJECT GAP AND FUTURE WORK.....	21
7. REFERENCES.....	22

List of Figures

Figure 1: Raw Data (from csv file)	5
Figure 2: Dataset Splitting	6
Figure 3: Bar graph showing different cars available in dataset	7
Figure 4: No. of Cars with their fuel type	7
Figure 5: No. of Cars with their transmission type (Manual vs Automatic)	7
Figure 6: Bar graph showing car condition (Old vs New)	8
Figure 7: Available Car Models	8
Figure 8: Histogram showing Engine Size	8
Figure 9: Density Plot for Mileage	9
Figure 10: Box Plot for price and Brand (prices in increasing range on x axis)	9
Figure 11: Box plot for price and model	9
Figure 12: Scatter Plot for price and engine size	10
Figure 13: Scatter plot for price and mileage	10
Figure 14: Correlation Heatmap	10
Figure 15: Scatter Plot for Price and Engine Size with Brand as Hue	11
Figure 16: Scatter Plot for Price and age with Brand as Hue	11
Figure 17: Scatter Plot for Price and Mileage with Brand as Hue	11
Figure 18: Handling Missing Values	15
Figure 19: Importing Libraries	16
Figure 20: Fitting linear regression model	16
Figure 21: Linear Regression Model Results	17
Figure 22: Correlation Heatmap using All Variables	17
Figure 23: Random forest Model Implementation	18
Figure 24: Evaluation of Random Forest Model	18
Figure 25: Gradient boost model implementation	19
Figure 26: Evaluation of Gradient boost Model	19
Figure 27: Cross Validation of all implemented models	19

1. Introduction

The exploration of machine learning technique application for predicting car price is done in the report using a dataset acquired from Kaggle that encompasses different car attributes. It is crucial to accurately predict car prices for dealers, manufacturers and consumers so that they make informed decisions. The focus of the report is on leveraging key features such as model, brand condition and engine size to develop predictive models. The report begins with the identification of dataset and its attributes followed by the detailed explanation of supervised learning tasks. The identification of relationships and patterns within the data is done using exploratory data analysis. The dataset is then split into training and testing sets for ensuring the validity of the model. The performance of different predictive models will be evaluated for determining the most effective approach in the report. The ML model learns about the relationship between the input features such as engine size, mileage, fuel type, etc and the target variable 'price' for predicting an exact price for a car which makes it a regression problem where the goal is to predict a continuous numerical value i.e., car price.

1.1 Dataset Identification

There are a total of 10 columns and 2,500 rows in the data set acquired from Kaggle. A balanced mix of numerical and categorical variables are provided in the dataset which is crucial for predictive modelling. The dataset represents car attributes for predicting their prices. The data provided in the CSV file is likely originated from industry databases or automotive listings which captures the key variables relevant to the price determination. The following columns represent key car attributes which include:

- Car ID: It is a unique identify for each car showing a categorical value
- Brand: The name of the manufacturer categorical value
- Year: Year of manufacture (numerical value)
- Model: Specifies version of car (categorical)
- Engine Size: Capacity of Engine in Litre (numerical)
- Fuel Type: Type of fuel used in car (petrol, diesel) (categorical value)
- Transmission: Gear system of car (automatic or manual) (categorical)
- Mileage: Car used in Kilometres (numerical)
- Condition: Shows if the car is used or new (categorical)
- Price: target variable in the dataset showing market price of car (numerical)

Car ID	Brand	Year	Engine Size	Fuel Type	Transmission	Mileage	Condition	Price	Model
1	Tesla	2016	2.3	Petrol	Manual	114832	New	26613.92	Model X
2	BMW	2018	4.4	Electric	Manual	143190	Used	14679.61	5 Series
3	Audi	2013	4.5	Electric	Manual	181601	New	44402.61	A4
4	Tesla	2011	4.1	Diesel	Automatic	68682	New	86374.33	Model Y
5	Ford	2009	2.6	Diesel	Manual	223009	Like New	73577.1	Mustang
6	Audi	2019	2.4	Diesel	Automatic	246553	Like New	88969.76	Q7
7	Audi	2020	4	Electric	Automatic	135486	Used	63498.75	Q5
8	Tesla	2017	5.3	Hybrid	Automatic	83030	New	17381.19	Model Y
9	Honda	2023	5.7	Electric	Manual	120360	Like New	15905.62	Civic
10	Ford	2010	1.5	Electric	Automatic	135009	Like New	9560.22	Explorer
11	Tesla	2001	1.8	Diesel	Automatic	298875	Like New	58872.6	Model 3
12	Ford	2017	5.7	Electric	Automatic	169737	Used	28074.19	Mustang
13	Ford	2006	4.7	Petrol	Automatic	114360	New	74766.45	Fiesta
14	Audi	2023	5.4	Electric	Automatic	263894	Like New	70193.74	Q7
15	BMW	2014	2	Electric	Automatic	65018	New	35220.52	X3
16	Ford	2010	3.9	Electric	Automatic	240904	Used	21796.16	Mustang
17	Mercedes	2017	4.5	Electric	Automatic	136817	New	14728.03	GLA
18	Audi	2022	4.4	Hybrid	Automatic	192803	Like New	75044.95	A3
19	Honda	2011	3	Electric	Automatic	86984	New	47791.89	Civic
20	BMW	2005	1.1	Petrol	Automatic	290595	New	35735.34	X5
21	Mercedes	2019	3.9	Petrol	Automatic	192608	Used	86382.04	C-Class
22	Mercedes	2022	2.3	Electric	Manual	12150	Used	61393.26	E-Class
23	Honda	2012	3.3	Diesel	Manual	275550	New	54210.22	CR-V
24	BMW	2019	5.8	Hybrid	Automatic	150853	New	75621.02	X5
25	Audi	2015	3	Electric	Automatic	188489	New	82480.4	Q5
26	Toyota	2017	5.2	Electric	Automatic	18325	Used	70176.95	Camry
27	BMW	2009	1.9	Electric	Manual	199756	Used	46800.6	X3
28	Honda	2022	4.4	Diesel	Manual	204541	New	41033.39	Accord
29	Mercedes	2007	5.9	Diesel	Manual	17669	Used	78308.17	GLC
30	Audi	2017	1.5	Diesel	Automatic	207836	Like New	54201.18	A3

Figure 1: Raw Data (from csv file)

(Source: Khan, 2024)

1.2 Supervised Learning Task Identification

The aim of the supervised learning task is to predict car prices which is a regression problem using the acquired dataset. The determination of accurate car price will be done depending on features like model, brand, engine size, condition and mileage which is the predictive problem. The target variable in the dataset is “price” which is a numeric value representing the market value of the car. The ground truth is provided by “price” for model training and evaluation. Various machine learning models will be leveraged in the task for capturing the linear and non-linear relationships between the target variable and other features.

2. Exploratory Data Analysis

The key patterns, relationships and distributions are identified by exploratory data analysis in the dataset for informing the model development. The exploratory data analysis process includes several steps in which data cleaning is done initially for checking any inconsistent or missing values (Milo, 2022). The rows having incorrect mileage values or missing prices are identified in this stage and handled by their removal or imputation (Jones, 2015). The basic statistics are computed for numerical columns such as price and mileage. The visualisations are created in the next stage of EDA where correlation heatmap, box plots and histograms are used for analysing the distribution of numerical features such as engine size and year. The visualisation of relationship between numerical features such as price, year and mileage are done by correlation heatmap. The analysis of categorical variables is done for analysing its impact on price. The potential outliers are identified in features like price and mileage using scatterplot.

2.1 Question Identification

The exploratory data analysis process is guided by the following questions that help in uncovering relationships, trends and potential outliers in the data which guides the selection and engineering of features for predictive modelling:

- How is the distribution of car price done across different brands and condition?
- How do the numerical features like engine size and mileage correlate with the price?
- Is there any significant difference in price of the car based on the fuel type or its transmission?
- Are the newer car models based on manufacturing year have higher price as compared to older cars?

2.2 Splitting the Dataset

The dataset will be split into two subsets for preventing any data leakage and ensuring the validity of the model i.e., training and testing set (80% and 20% respectively). The first subset will be of training set in which the 80% of the dataset will be used for training. The model's training set will ensure that sufficient data is present for capturing the underlying relationships and patterns. The remaining 20% will be reserved for evaluating the performance of the model on unseen data. The 20% set of data will be the testing set through which generalisability of the model will be ensured to new data.

```
# Split into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Standardize the features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
pdf.chapter_body("Standardized the feature columns.")
```

Figure 2: Dataset Splitting

2.3 Exploratory Data Analysis Process and Results

The EDA analysis shows different graphs, charts and plots providing insights on how the variables of car price dataset are related to each other.

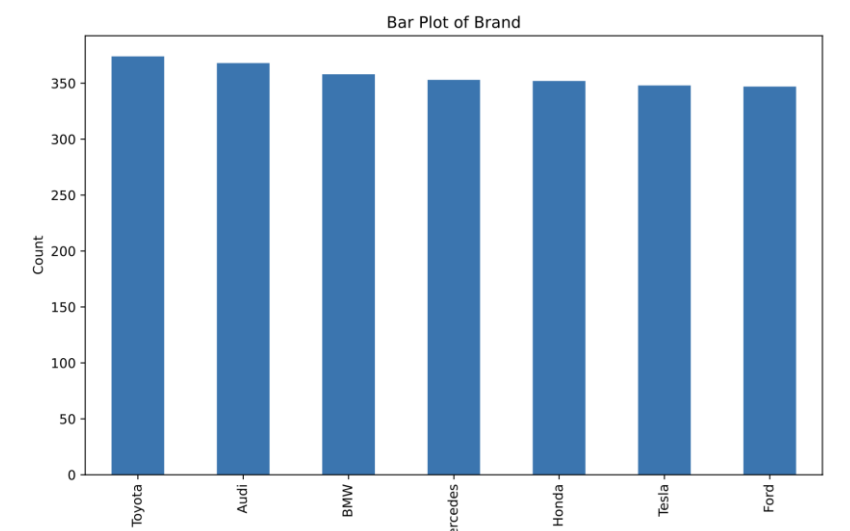


Figure 3: Bar graph showing different cars available in dataset

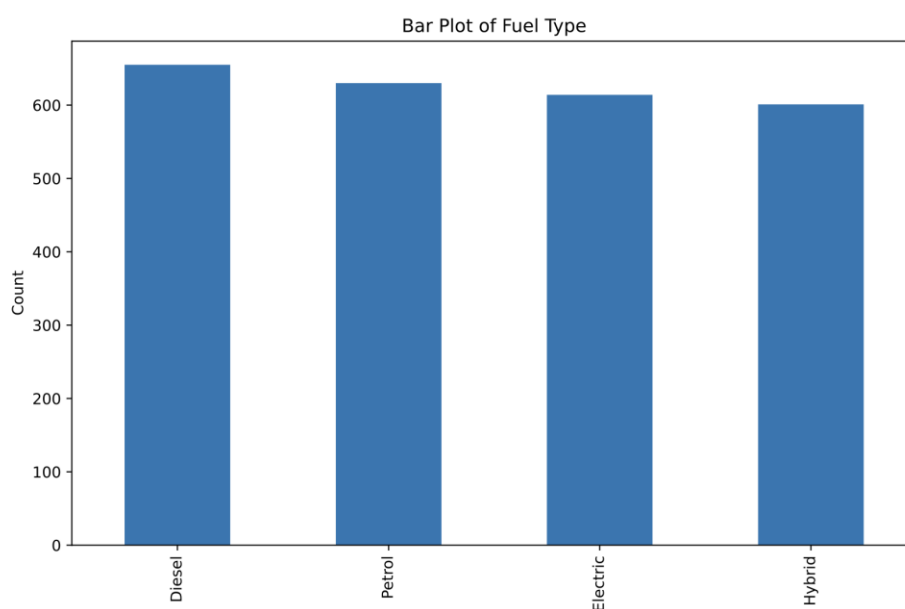


Figure 4: No. of Cars with their fuel type

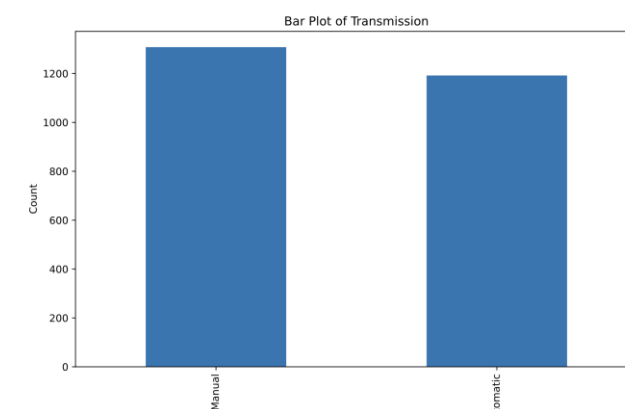


Figure 5: No. of Cars with their transmission type (Manual vs Automatic)

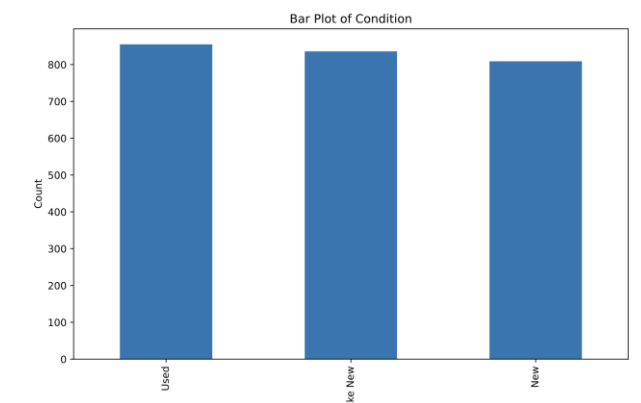


Figure 6: Bar graph showing car condition (Old vs New)

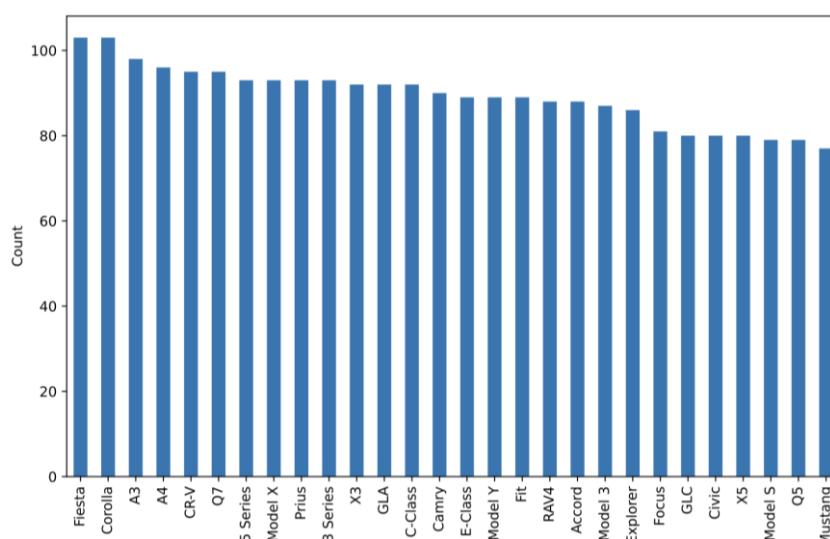


Figure 7: Available Car Models

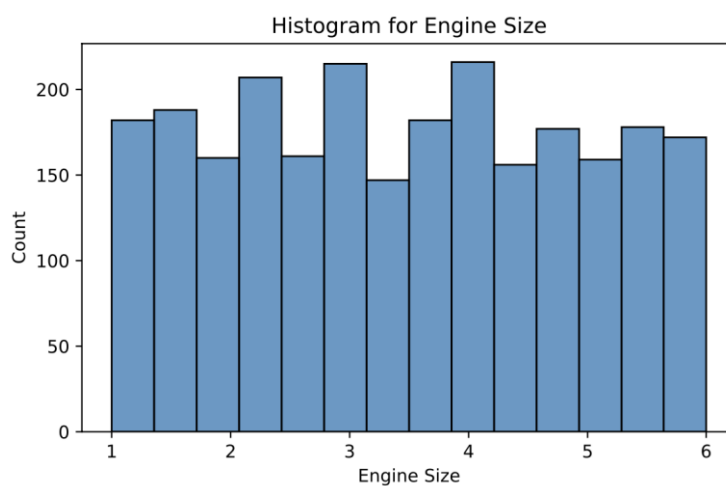


Figure 8: Histogram showing Engine Size

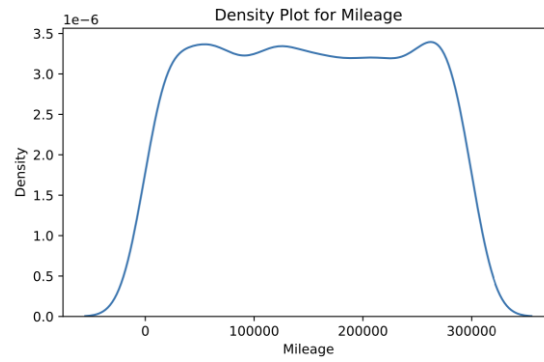


Figure 9: Density Plot for Mileage

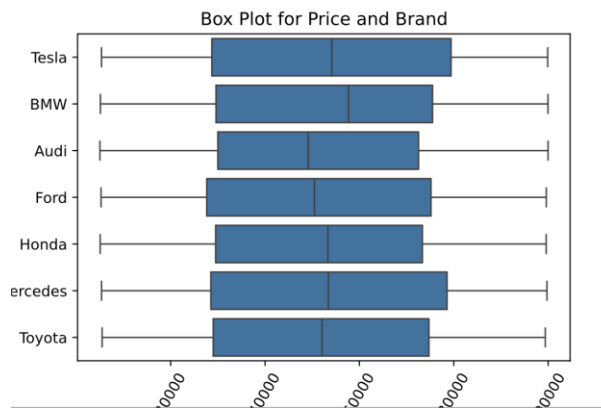


Figure 10: Box Plot for price and Brand (prices in increasing range on x axis)

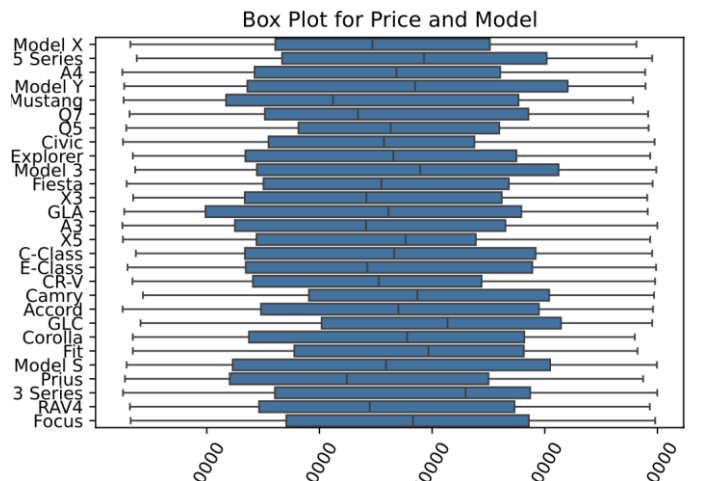


Figure 11: Box plot for price and model

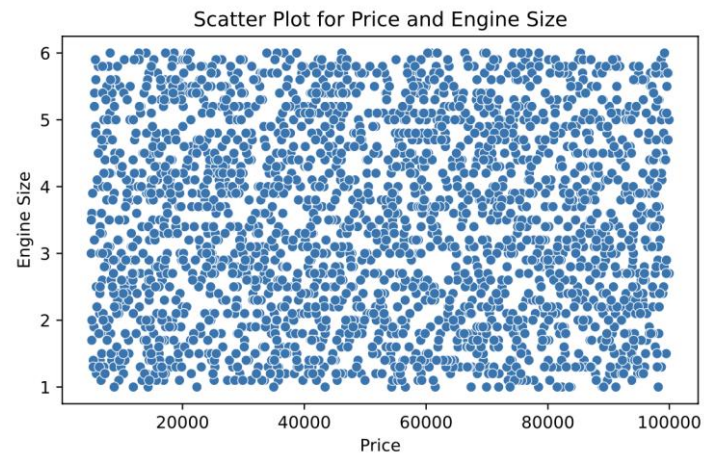


Figure 12: Scatter Plot for price and engine size

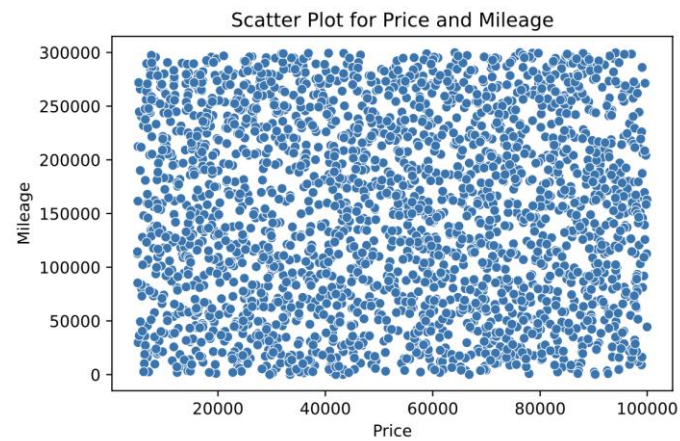


Figure 13: Scatter plot for price and mileage

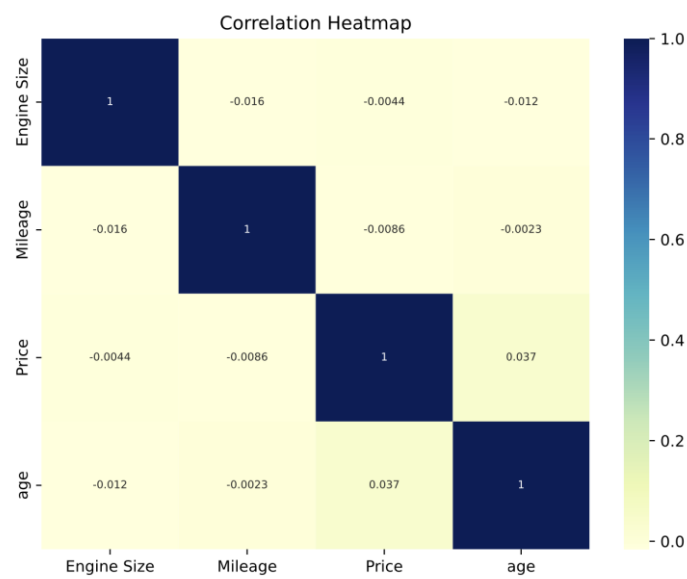


Figure 14: Correlation Heatmap

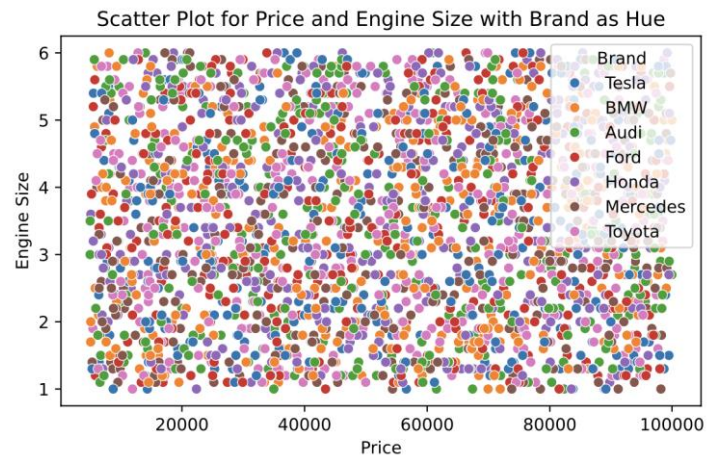


Figure 15: Scatter Plot for Price and Engine Size with Brand as Hue

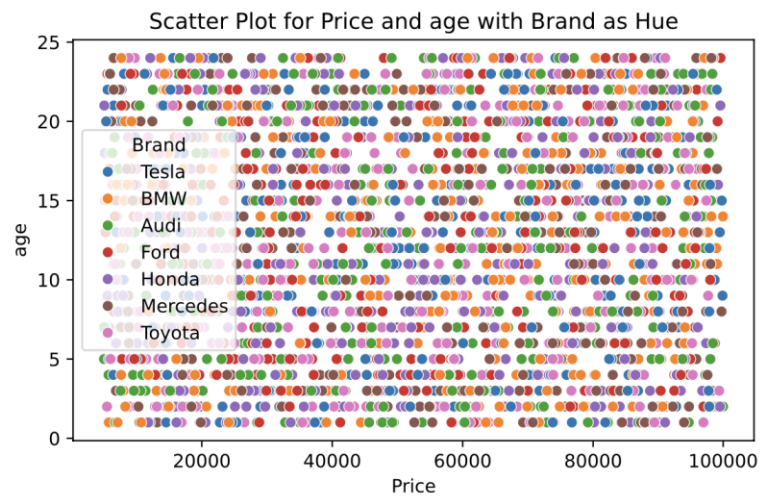


Figure 16: Scatter Plot for Price and age with Brand as Hue

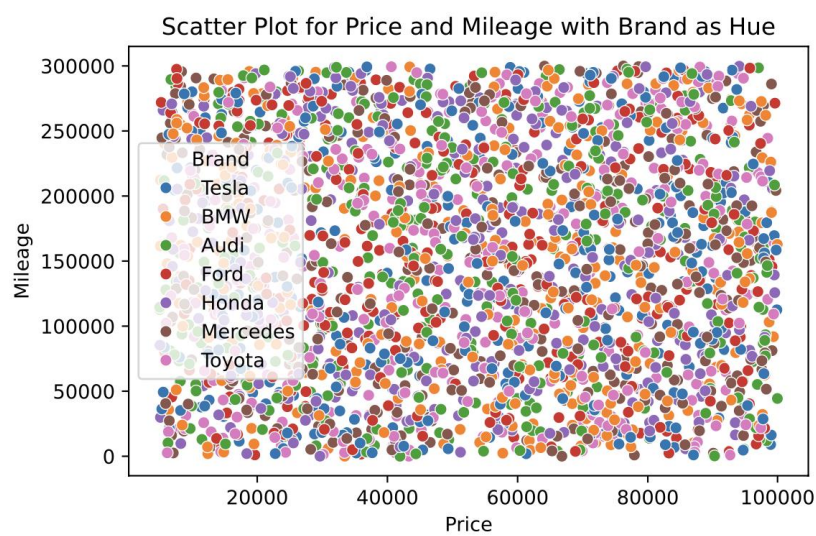


Figure 17: Scatter Plot for Price and Mileage with Brand as Hue

Valuable insights are provided by the EDA process into the car price dataset as it employs a variety of visualisations. The diverse visualisations are created which is one of the strengths of the analysis:

- The inclusion of bar graphs such as car condition and transmission type and histograms such as mileage and engine size provide clear distribution of numerical and categorical variable (Fig 3 to fig 8).
- The illustration of the key aspects like the predominance of fuel type as well as relative frequency of transmission type is done by these graphs.
- The categorical variables are presented effectively using the bar graphs showing fuel type and car condition revealing significant trends. For example, it is assessed that newer cars dominate the higher price bracket while the fuel type correlates with the specific price range (Fig 11).
- The relationship among the numerical variables is highlighted in the correlation heat map (Fig 14).
- The negative correlation between the price and mileage aligns with the expectation which is that higher mileage tends to decrease the car value (Fig 14).
- The brand specific price variations are provided in the box plot for price by model or brand which is important for predictive modelling (Fig 11).
- The outliers in the mileage and price are effectively flagged using box plot and scatter plot which is critical for understanding data quality as well as guiding decisions i.e., whether to remove or retain these values.

2.4 EDA Conclusions

Specific trends and patterns are analysed in the visualisations created using different variables of dataset. The brand impact is assessed using box plots which provide insights on luxury brands like Mercedes and BMW which have higher price range while the economy brands are categorised in lower price range. A layer of interpretability is added by the scatter plot with brand as a hue by combining numerical and categorical insights. There is a positive correlation identified between the price and engine size which is evident in the scatter plot (Fig 15 to Fig 17). The larger engines are often correspondent to premium model that justifies their higher cost. The scatter plot and density plot confirm that higher mileage cars have lower prices which is consistent with the depreciation due to the wear and tear.

3. Experimental Design

3.1 Identification of Chosen Supervised Learning Algorithm

The selection of three supervised learning algorithms is done for predicting car prices which includes linear regression, random forest and gradient boosting. The selection of these algorithms is done due to their suitability for regression tasks, capacity of modelling both simple and complex relationships along with their ability of handling mixed data types.

The linear regression is a statistical method through which the relationship between the dependent variable and one or more independent variables is modelled (Maulud et al., 2020). The dependent variable in the dataset is 'price' and the independent variables include features like year, mileage and condition. A linear relationship is assumed by this statistical method between the predictor and the target variable (Kumari, 2018). The linear regression method is selected as it serves as a baseline model showing insights into the strengths of linear relationship in the data and feature importance. The non-linear interactions are not very well captured by linear regression, but a foundation is established for comparison with more complex models. The identification of primary contributor to car price is done in the prediction task which may be engine size and mileage.

The second chosen method is gradient boosting which is one of the ensemble methods which develops predictive models in a sequential manner where the focus of each model is on reducing the residual error of the previous model (Si and Du, 2020). The algorithms like LightGBM or XGboost implement the gradient boost technique in an efficient manner. The process of gradient boosting involves fitting a weak learner to data such as a decision tree, calculation of the residual, which is a difference between predicted and actual value, training next three models on the residual of previous ones for minimizing the loss function and aggregating the predictions of all the decision trees (Zhang et al., 2019). Complex and not linear relationships are captured effectively using gradient boosting method. The gradient boost method has the capacity of optimising loss functions that makes it highly accurate for the regression tasks such as this. For the prediction of car prices using the acquired dataset, the gradient boost model is well suited to find intricate patterns such as the relationships between mileage, brand and condition that provides superior predictive performance.

The third employed method is random forest which is another ensemble method that creates a forest of decision trees having each tree trained on the random set of data and features (Parmar et al., 2019). The aggregate of individual trees is done by random forest by averaging in case of regression task (Probst, 2018). The key characteristics of random forest include feature randomness and bagging. The random forest model combines predictions from different models for reducing variance and prevent the issue of overfitting. Variations are introduced by random forest by choosing random subset of features for each tree that

improves generalisation of model. A versatile and robust mechanism is followed by random forest that effectively handles mixed data types be it numerical or categorical. The issue of overfitting is mitigated by the ensemble approach used in the random forest algorithm that makes it ideal for the regression problems like car price prediction dataset (Bakır et al., 2024). The non-linear interactions are captured efficiently by random forest method such as influence of transmission, type or brand on price (target variable) while also maintaining interpretability through feature importance metrics.

Upon comparing, it is analysed that interpretability and simplicity is provided by linear regression that serves as the baseline model. High accuracy is offered by gradient boost method as it models intricate patterns which is highly suitable for competitive tasks such as this. The balance between robustness and performance is established by random forest that accommodates data heterogeneity. A spectrum of complexity is covered by these algorithms that enables comprehensive evaluation of predictive capability for car price prediction due to which they are selected for this regression task.

3.2 Identification of Appropriate Evaluation Techniques

A crucial role is played by evaluation techniques for assessing the performance of predictive model and ensuring their reliability for the prediction of car prices. The evaluation of model will be done by MAE, MSE, RMSE, MAPE, accuracy, recall, precision and F1 score which are chosen as the primary evaluation metric.

Mean Absolute Error (MAE) calculates the average magnitude of errors between predicted and actual values as per Frías-Paredes et al. (2018) which is the first metric chosen for evaluating model performance as it provides a straightforward measure of average prediction error. The calculation of average squared difference between actual and predicted values is done by Mean Squared Error (MSE) as per Chicco et al. (2021) which is significant when large errors are to be highlighted for e.g. outliers in car price predictions. An error measure on the same scale as the target variable provided by RMSE which is the square root of MSE. A balanced perspective is provided by RMSE between average and extreme prediction inaccuracies which will help in providing overall model reliability for car prices. The average percentage error between actual and predicted values is provided by MAPE i.e., Mean Absolute Percentage Error (Hodson, 2022).

3.3 Data Cleaning and Pre-processing Transformations

Minimal cleaning was required for the acquired dataset for the car price prediction task because of absence of missing values. However, several pre-processing techniques were applied for preparing the data for machine learning models to ensure that the dataset is properly optimised and formatted for performance. The data cleaning is performed by handling

missing values and validating the dataset for data integrity. The dataset was checked for any missing values using tools like `isnull ()` method and creating visualisations like heatmaps.

The `isnull ()` method confirmed that there were no missing values in the dataset that simplified the cleaning process. Additional checks were performed in the dataset for anomalies such as negative or zero values in numerical fields i.e., mileage and engine size. The inconsistencies were not identified in the dataset for potential corrections or exclusion of data. If there were missing values present in the dataset, the strategies like mean or median imputation would have been implemented for numerical features. However, for categorical features, the placeholder or most frequent categories such as 'unknown' would have been used in the dataset.

Handling Missing Values

Missing Values Before Handling:

Car ID	0
Brand	0
Year	0

Figure 18: Handling Missing Values

The data encoding is performed in the dataset where the target variable price was left as a numerical column because it directly shows the regression output. The encoding of features like model, brand, transmission, fuel type and condition was done using one hot encoding. The assignment of ordinal relationship to non-ordinal variable is prevented by this technique that maintains model accuracy. The feature engineering is performed in the dataset where the outlier detection in Price and Mileage is done by creating visualisations such as box plots.

3.4 Limitations and Options

The data set is comprehensive, but it presents limitations which includes potential bias in feature representation as well as insufficient data on external factors such as regional variations and market demand. The reliance is majorly on historical data that may overlook the emerging trends. The selected models may face challenges related to overfitting or multicollinearity especially for highly correlated features. The advanced techniques can be leveraged alternatively such as deep learning for capturing complex trends and patterns in the dataset related to external variables for enhancing model robustness. Feature engineering could be involved for improving the model performance and reducing redundancy.

4. Predictive Modelling / Model Development

4.1 The Predictive Modelling Process

The linear regression model is created in which the predictive modelling process starts with the data preparation and systematic steps are followed for developing a reliable model. The first stage was splitting the dataset in which the dataset was divided into training and testing subsets which is 80% and 20% respectively for ensuring that the model generalises effectively on the unseen data. In the next stage, features selection and engineering are done where relevant numerical features such as Price, mileage and engine size as well as categorical features such as brand and transmission are selected. The transformation of categorical variables is done using one-hot encoding to make them suitable for the regression model. The implementation of the linear regression model is done as the primary model for predicting the car prices. The training data was fitted into the model by optimising coefficient for minimising the sum of squared error.

```
"from sklearn.model_selection import train_test_split\n",
"from sklearn.linear_model import LinearRegression\n",
"from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor\n",
"from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score\n",
"from sklearn.preprocessing import StandardScaler"
```

Figure 19: Importing Libraries

```
"# Linear Regression Model\n",
"linear_model = LinearRegression()\n",
"linear_model.fit(X_train, y_train)\n",
"\n",
"# Predict using Linear Regression\n",
"y_pred_linear = linear_model.predict(X_test)\n",
"\n",
```

Figure 20: Fitting linear regression model

4.2 Evaluation Results on “Seen” Data

The linear regression model evaluation is done in which the metrics are evaluated which include mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE). The value of Mean Absolute Error is 23692.56. The value of Mean Squared Error is 759016533.57 and Root Mean Squared Error is 27550.25.

	Car ID	Brand	Year	Engine Size	Fuel Type	Transmission	Mileage	Condition	\
0	1	Tesla	2016	2.3	Petrol	Manual	114832	New	
1	2	BMW	2018	4.4	Electric	Manual	143190	Used	
2	3	Audi	2013	4.5	Electric	Manual	181601	New	
3	4	Tesla	2011	4.1	Diesel	Automatic	68682	New	
4	5	Ford	2009	2.6	Diesel	Manual	223009	Like New	
	Price	Model							
0	26613.92	Model X							
1	14679.61	5 Series							
2	44402.61	A4							
3	86374.33	Model Y							
4	73577.10	Mustang							
Linear Regression Model Evaluation:									
Mean Absolute Error: 23692.564222638546									
Mean Squared Error: 759016533.5678779									
Root Mean Squared Error: 27550.254691524686									

Figure 21: Linear Regression Model Results

The average magnitude of error is indicated by mean absolute error in the prediction by ignoring their direction. It is suggested by this value that on an average, the car price prediction of the model deviate by \$23,692.56 from the actual price. The higher value of mean squared error is suggesting that there may be some significant outliers present in the dataset which are causing larger prediction error. The root means squared error is signifying that the typical prediction errors are around \$27,550.25.

It is interpreted from the evaluation metrics that some of the variability in the car prices is captured by the linear regression model but it may not be completely accurate. The relatively high MAE and RMSE values indicate that the predictions of the model could be improved. The reasons behind these predictions may be due to potential multicollinearity present among features like fuel type and engine size. There may be outliers in features or data with non-linear relationship relationships.

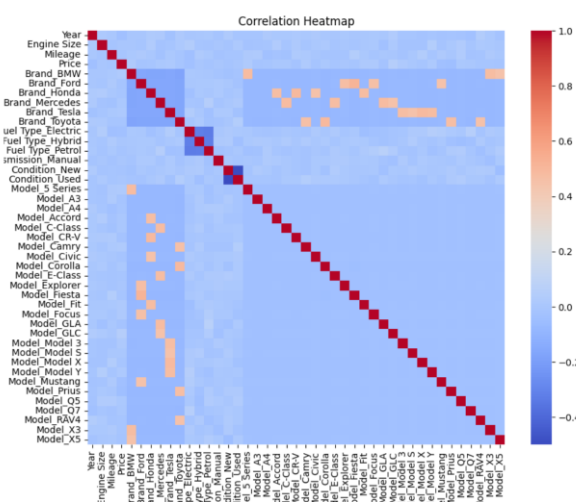


Figure 22: Correlation Heatmap using All Variables

5. Evaluation and Further Modelling Improvements

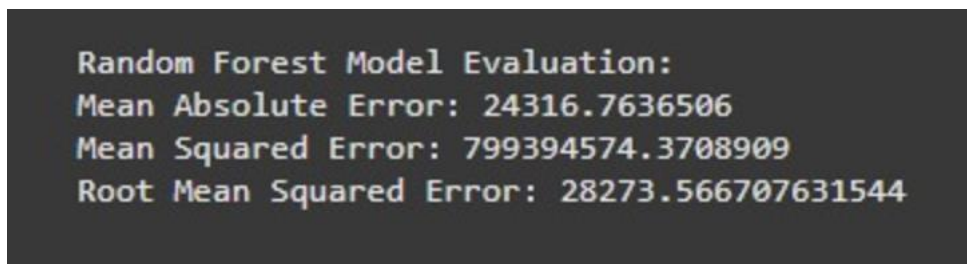
5.1 Alternative Model Implementation: Random Forest

The alternative models are implemented with anticipation of improved prediction accuracy. The implementation of random forest model is done in which the performance of the model is checked on the test set i.e., unseen data accounting for 20%.

```
"# Random Forest Model\n",
"rf_model = RandomForestRegressor(n_estimators=100, random_state=42)\n",
"rf_model.fit(X_train, y_train)\n",
"\n",
"# Predict using Random Forest\n",
"y_pred_rf = rf_model.predict(X_test)\n",
"\n",
```

Figure 23: Random forest Model Implementation

The random forest model evaluation is done in which the mean absolute error is 24316.76 while the mean squared error is 799394574.37. The Root Mean Squared error is calculated to be 28273.56.



```
Random Forest Model Evaluation:
Mean Absolute Error: 24316.7636506
Mean Squared Error: 799394574.3708909
Root Mean Squared Error: 28273.566707631544
```

Figure 24: Evaluation of Random Forest Model

It is highlighted by the mean absolute error that on an average, the predicted car prices are deviating from the actual price by approximately \$24,316.76. Reasonably large errors are suggested by this value possibly due to the feature selection or inherent data variability. A high value of mean square error is indicated which suggests that there may be some significant errors in the model possibly due to the outliers in the dataset. The typical prediction error is suggested by the root mean squared error which is around 28273.56 which is significant depending on the average car price in the dataset. However, slightly better results are showed by random forest model as compared to the linear regression model in terms of error reduction. However, the values of MAE and RMSE are relatively high suggesting the room for improvement.

5.2 Alternative Model Implementation: Gradient Boosting Method

The implementation of Gradient Boosting method is done for improving the model performance than random forest and linear regression. The steps for implementing the model

are similar as in linear regression where python libraries are loaded, and dataset is cleaned and split into two subsets: training and testing. The model is fit into training subset and trained based on which it performs on unseen data whose evaluation is provided below:

```
'# Gradient Boosting Model\n",
'gb_model = GradientBoostingRegressor(n_estimators=100, random_state=42)\n",
'gb_model.fit(X_train, y_train)\n",
'\n",
'# Predict using Gradient Boosting\n",
'y_pred_gb = gb_model.predict(X_test)\n",
'\n".
```

Figure 25: Gradient boost model implementation

The calculation of Mean Absolute Error is done which is 23692.49 while the Mean Squared Error is 770497442.16. The Root Mean Squared Error is computed with gradient boost which is 27757.83.

```
Gradient Boosting Model Evaluation:
Mean Absolute Error: 23692.49040846372
Mean Squared Error: 770497442.1678281
Root Mean Squared Error: 27757.8356895459
```

Figure 26: Evaluation of Gradient boost Model

From the analysis, it is seen that on an average, the predictions of gradient boosting model are deviating from the actual car price by \$23,692.49 which shows a slightly lower error value as compared to the random forest model highlighting improved performance. The mean squared error value is slightly lower as compared to the random forest model which suggests that the significant prediction errors are reduced more effectively by gradient boosting model. Upon comparison, the root mean square error is also lower of gradient boosting model than random forest and linear regression that shows an overall improvement in the prediction error.

```
Cross-validation scores (R-squared) for Random Forest:
[-0.07680381 -0.09211411 -0.04126547 -0.03782143 -0.09710084]
Cross-validation scores (R-squared) for Gradient Boosting:
[-0.05419946 -0.04512603 -0.00954059 -0.0144158 -0.03465471]
Cross-validation scores (R-squared) for Linear Regression:
[ 0.00617265 -0.00193943 -0.00071524 -0.01607991 -0.00640036]
```

Figure 27: Cross Validation of all implemented models

5.3 Comparison of Models

Upon comparison of models, it is analysed that the gradient boosting method is showing slight improvement over random forest and linear regression models that makes it the best

performing model among all three. There is reduction in the error particularly in MAE and RMSE which demonstrates that the patterns are captured more effectively in the data by gradient boosting method. It is indicated by the low score of MAE and RMSE in the gradient boosting method that more accurate price predictions are delivered by this model on unseen data, although incremental improvements are required in the model. It is highlighted by the differences in the models that the dataset will benefit from further outlier handling, feature engineering and data pre-processing for significantly reducing the errors. Hyperparameter adjustment may also result in improved model performance which includes tree depth, number of boosting stages and learning rate for further improving the gradient boosting model performance.

6. Conclusion

6.1 Summary of Results

In conclusion, successful demonstration of the use of machine learning models is done for predicting car prices using linear regression, random forest and gradient boosting algorithms. The prediction of car price is done using these machine learning models with reasonable accuracy whose exploration is shown in the report. In comparison of all the implemented models, the gradient boosting method emerged as the slightly better in performance showing lowest MAE and RMSE score which indicates that it is superior as compared to random forest and linear regression for capturing complex relationships in the data. Valuable insights are provided by the analysis into the key factors which influence car prices. The distribution of car price across different brands and conditions reveals that the well-maintained cars and luxury brands command higher price. It is also analysed in the report that the numerical features like mileage and engine size show significant correlation with price where the larger engine size typically increase the price while the high mileage reduces it. In addition to that, the categorical features such as transmission and fuel type are found to be influencing prices of the car where the manual transmission and electric cars often reflect distinct pricing pattern.

6.2 Reflection on Individual Learning

I experienced significant growth in both analytical and technical skills throughout this project. I faced challenges initially in understanding the intricacies of data pre-processing and the selection of appropriate machine learning models. I cleared my understanding on data pre-processing using my course concepts and project examples provided on machine learning problems (especially regression tasks) where I made a considerable effort to handle duplicate and missing values in the dataset. I encoded the categorical variables and normalised the features in an effective manner.

The challenges pushed me to enhance my understanding on data manipulation techniques so that I would not face difficulty in the later stages of model development. After implementing the random forest, linear regression and gradient boosting model, I was exposed to the importance of model evaluation metrics like MAE, MSE and RMSE. With the slightly better performance, the gradient boosting method was the best performing model among all three, but I learnt that with more effort put into dataset optimisation and cleaning, I could have achieved enhanced performance of the model.

Each of the implemented model has unique insights but upon their performance comparison, I learnt the value of iterative experimentation for achieving accurate predictions. It is important that I understand the feature importance where I believe my understanding and knowledge is limited (in practical aspect) and I lack the most. The performance of the models could be improved with Hyperparameter Tuning, Feature Importance Analysis and Residual Analysis. One of the key learnings from this project for me was importance of residual analysis for identifying systematic errors that will guide me to refine the models further. Overall, my ability of approaching machine learning tasks is enhanced through this project with structured thinking because I feel confident that I can approach similar machine learning tasks in the future effectively.

6.3 Project Gap and Future Work

Several limitations are encountered in the project which impacted the overall effectiveness of the model and its analysis. First comes the size of the data set where only 1257 rows were present in the data which may not have been sufficiently diverse or large to capture all variations in the car prices that led to potential underrepresentation or overfitting of certain factors. In addition to that, the certain critical features were lacking from the dataset such as regional market variation, brand reputation or historical price trend that would have improved the accuracy of the prediction. There is the potential presence of outliers in the dataset as well such as cars showing extremely low or high price that affected the ability of the model to generalise effectively. Even though the steps of pre-processing were applied in the dataset like encoding and normalisation, there is need of further refinement such as implementing advanced featuring engineering or outlier removal that will minimise the influence of outliers and improve model performance.

7. References

- Bakır, R., Orak, C. and Yüksel, A., 2024. Optimizing hydrogen evolution prediction: A unified approach using random forests, lightGBM, and Bagging Regressor ensemble model. *International Journal of Hydrogen Energy*, 67, pp.101-110.
<https://www.sciencedirect.com/science/article/pii/S0360319924014460>
- Chicco, D., Warrens, M.J. and Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peerj computer science*, 7, p.e623. <https://peerj.com/articles/cs-623/>
- Frías-Paredes, L., Mallor, F., Gastón-Romeo, M. and León, T., 2018. Dynamic mean absolute error as new measure for assessing forecasting errors. *Energy conversion and management*, 162, pp.176-188.
<https://www.sciencedirect.com/science/article/pii/S0196890418301341>
- Hodson, T.O., 2022. Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development Discussions*, 2022, pp.1-10.
<https://gmd.copernicus.org/preprints/gmd-2022-64/>
- Jones, Z. and Linder, F., 2015, April. Exploratory data analysis using random forests. In *Prepared for the 73rd annual MPSA conference* (pp. 1-31).
<https://pdfs.semanticscholar.org/e7b7/3565b07a7f1369a20b1055f222423f0feb34.pdf490>
- Khan, Z. (2024). Car Price Prediction. <https://www.kaggle.com/datasets/zafarali27/car-price-prediction>
- Kumari, K. and Yadav, S., 2018. Linear regression analysis study. *Journal of the practice of Cardiovascular Sciences*, 4(1), pp.33-36.
- Maulud, D. and Abdulazeez, A.M., 2020. A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, 1(2), pp.140-147.
<https://jastt.org/index.php/jasttpath/article/view/57>
- Milo, T. and Somech, A., 2020, June. Automating exploratory data analysis via machine learning: An overview. In *Proceedings of the 2020 ACM SIGMOD international conference on management of data* (pp. 2617-2622).
<https://dl.acm.org/doi/abs/10.1145/3318464.3383126>
- Parmar, A., Katariya, R. and Patel, V., 2019. A review on random forest: An ensemble classifier. In *International conference on intelligent data communication technologies and*

internet of things (ICICI) 2018 (pp. 758-763). Springer International Publishing.

https://link.springer.com/chapter/10.1007/978-3-030-03146-6_86

Probst, P. and Boulesteix, A.L., 2018. To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, 18(181), pp.1-18.

Si, M. and Du, K., 2020. Development of a predictive emissions model using a gradient boosting machine learning method. *Environmental Technology & Innovation*, 20, p.101028.

<https://www.sciencedirect.com/science/article/pii/S2352186420313286>

Zhang, C., Zhang, Y., Shi, X., Almpandis, G., Fan, G. and Shen, X., 2019. On incremental learning for gradient boosting decision trees. *Neural Processing Letters*, 50, pp.957-987.

<https://link.springer.com/article/10.1007/s11063-019-09999-3>