Group 1:

1. Parham Faraji (230430778)
2. Deepak Srikanth (230324824)
3. Alim Rabbi (230410448)
4. Simran Kaur Saini (230425197)
5. Rajapriyan Ravi (240043117)

DATA MINING

GROUP ASSIGNMENT

## **Introduction**

* **What is the problem?**

The primary problem addressed in this report is the analysis and prediction of car prices based on various features. Understanding how different factors influence car prices can help stakeholders make informed decisions. We are using different machine learning algorithms on the historical dataset and showing the output using Python. This task is done in a group. To solve this problem, we are doing regression analysis on the dataset using different models.

* **Who are the stakeholders?**

1. Car Buyers

Car buyers are one of the primary stakeholders in this analysis. They are interested in getting the best value for their money when purchasing a vehicle. Understanding how different features affect the price of a car can help them make more informed decisions and choose a car that offers the best combination of features, condition, and price.

1. Car Sellers and Dealerships

Car sellers and dealerships want to price their cars competitively to attract buyers while maximizing their profit margins. By analyzing how various factors influence car prices, sellers can set prices that are attractive to buyers yet still profitable. Accurate pricing strategies can lead to increased sales and customer satisfaction.

1. Insurance Companies

Insurance companies use car price data to assess the risk associated with insuring different vehicles. Car value plays a significant role in determining insurance premiums. If an insurance company can accurately value a car, it can better assess the risk and set premiums that reflect the actual cost and risk of insuring the vehicle.

1. Financial Institutions

Financial institutions, such as banks and credit unions, provide car loans to buyers. To do this effectively, they need to evaluate car values accurately to determine the loan amounts they can safely offer. Proper car valuations help in making informed lending decisions, ensuring that the loans are adequately secured by the value of the vehicle.

* **Why does this problem matter to each stakeholder?**

1. For car buyers, knowing how features like model, year, condition, engine size, and safety ratings affect car prices can help them make better purchasing decisions. For instance, understanding that newer cars with better safety ratings tend to be more expensive allows buyers to prioritize which features they are willing to pay a premium for and which ones they can compromise on to stay within their budget.
2. Car Sellers and Dealerships

Accurate pricing strategies are crucial for car sellers and dealerships. By understanding the market dynamics and how different features impact prices, they can price their cars competitively, attract more buyers, and increase their sales volumes. This also helps in maintaining a good reputation in the market as a dealership that offers fair prices.

1. Insurance Companies

For insurance companies, accurate car valuations are essential for assessing risk and determining insurance premiums. If the value of a car is overestimated, the insurance company might charge higher premiums, which could deter customers. On the other hand, underestimating car values could lead to inadequate coverage and higher payouts in the event of a claim. Therefore, precise car valuations help insurance companies balance their risk and offer fair premiums to their customers.

1. Financial Institutions

Financial institutions rely on accurate car valuations to make informed lending decisions. If a car is overvalued, the institution might lend more money than the car is worth, increasing their risk. Conversely, undervaluing a car could lead to lower loan amounts, which might not be sufficient for the buyer. Accurate valuations ensure that the loans are properly secured by the value of the vehicle, reducing the risk for the financial institution and helping buyers get the necessary financing.

By addressing these concerns, the analysis of car prices based on various features becomes valuable for all stakeholders involved, helping them make informed and strategic decisions.

## **Data Set and Visualization**

* **Type of data set and dimensions**

Our data set is cross – sectional dataset and dimensions are 2231 rows and 32 columns.

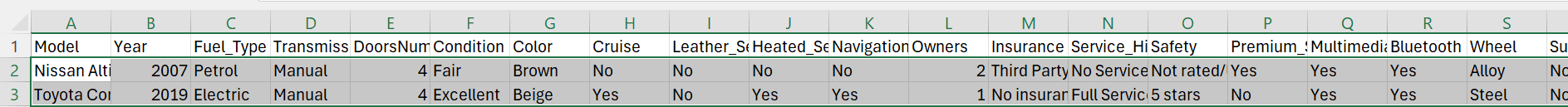
The 10 variables were chosen based on a combination of:

1. Correlation Analysis: Numeric features were selected based on their correlation with car prices. Strongly correlated features are more likely to be predictive.
2. Standard Deviation Calculation: For categorical features, those with higher standard deviations of average prices across different instances were selected, indicating they have a significant impact on price variations.
3. Data Cleaning: Categorical and Numerical features were cleaned and standardized to ensure consistent data representation, reducing noise and potential errors in the model.
4. Domain Knowledge: Features were selected based on domain knowledge about factors that typically affect car prices (e.g., model, age, condition, mileage).

* **Variables, definitions, their types, and their roles**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable name | Definitions | Variable type | Role |
| Model | The brand and model of the car | Categorical | Predictors |
| Tage (Tyre Age) | The age of the car's tires | Categorical | Predictors |
| Condition | The overall condition of the car | Categorical | Predictors |
| Year | The manufacturing year of the car | Categorical | Predictors |
| Color | The color of the car | Categorical | Predictors |
| Safety | The safety rating of the car | Categorical | Predictors |
| Fuel\_Type | The type of fuel the car uses | Categorical | Predictors |
| Monthly\_mileage | The average monthly mileage of the car | Numerical | Predictors |
| Engine\_Size | The size of the car's engine | Numerical | Predictors |
| Owners | The number of previous owners of the car | Numerical | Predictors |
| Sales Price | Selling Price of the car | Numerical | Target Variable |

* **Verbal presentation of two records**

******

1. This record is about a Nissan Altima manufactured in 2007. It's a petrol-powered vehicle that's currently in fair condition. The car is brown in color and has an unrated or unknown safety rating. Its tires are 3 years old (TAge: 3). The car has had 2 previous owners and has an engine size of approximately 1.66 liters. The vehicle's length is 179.1 inches, and it's priced at about $8,005.
2. This record describes a Toyota Corolla from 2019. It's an electric vehicle in excellent condition. The car is beige and boasts a top safety rating of 5 stars. Its tires are relatively new, being only 1 year old (TAge: 1). This Corolla has had only one owner and features a larger engine size of about 2.70 liters. The car's length is 182.1 inches, and it's valued at approximately $25,385.

* **Level of the data: The data is at the individual vehicle level.**

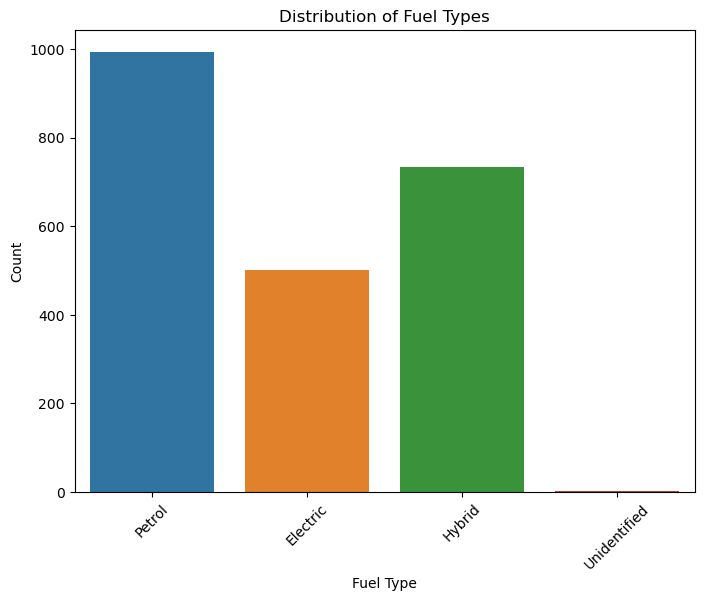
What each record is about: Each record in the dataset represents a specific car with its unique set of characteristics and attributes.

* **Univariate Analysis for Predictors**

A graph of different colored bars

Description automatically generated

Comment: The most common models are Toyota Camry and Chevrolet Impala. This distribution shows the mode of the dataset, with the mode being the most frequent model. There is a high spread, indicating variability in the types of models available.



Comment: The mode for fuel type is petrol, as it is the most common. The spread shows significant presence of hybrid and electric cars as well, indicating a variety in fuel types but with petrol being dominant.

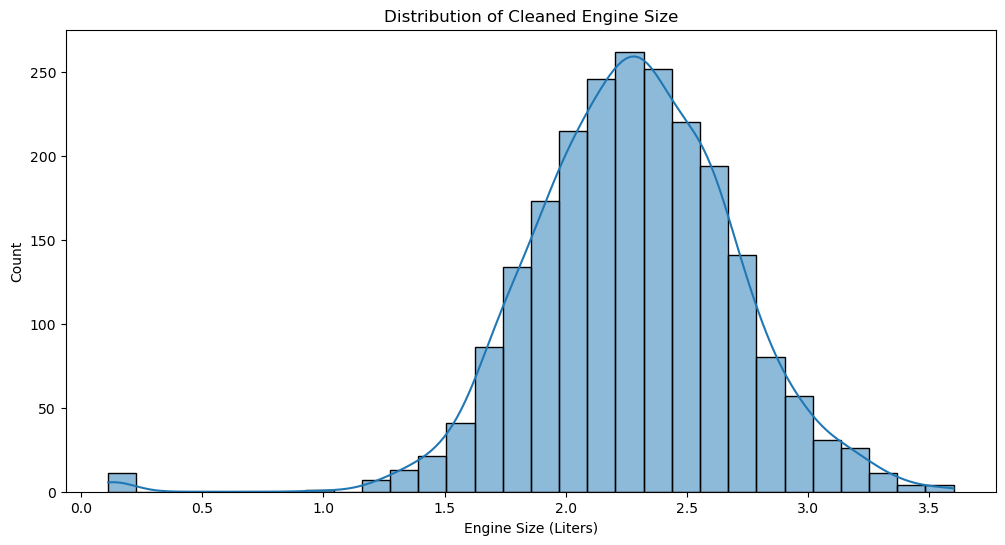
A graph of a distribution of car condition

Description automatically generated

Comment: The mode for car condition is "Good", as it is the most frequently occurring condition. The spread between "Fair", "Good", and "Excellent" conditions shows the distribution of the car's state, with a central tendency towards "Good".

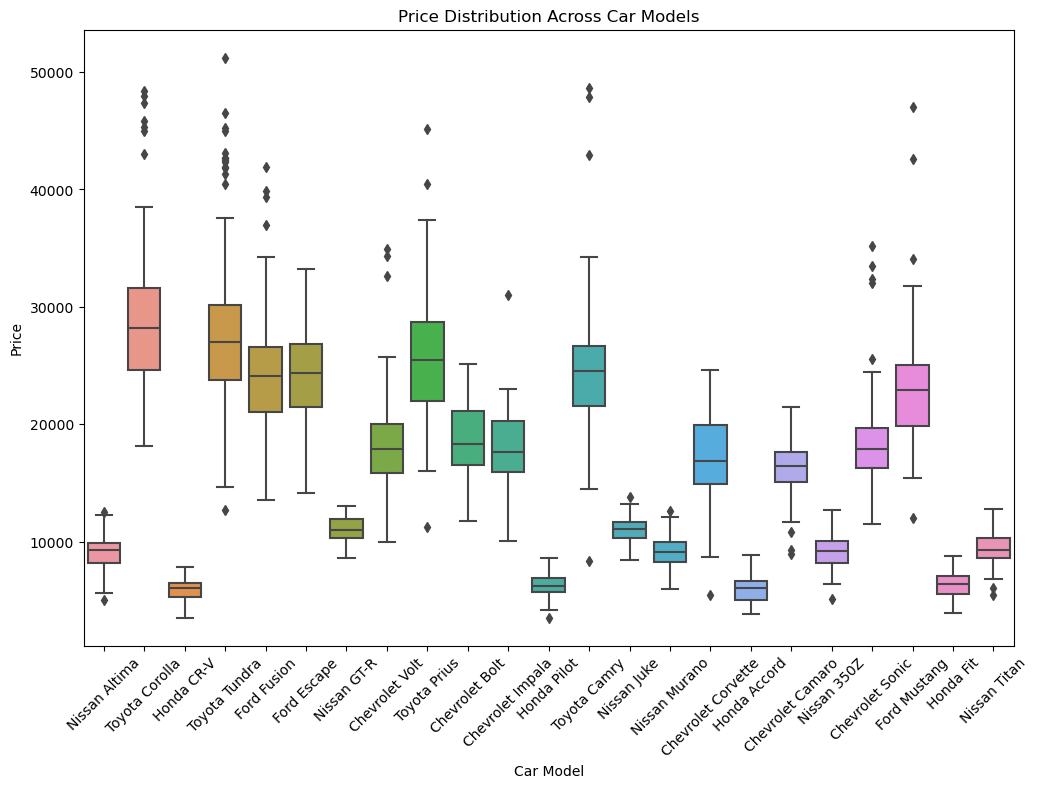
A graph with blue bars and red lines

Description automatically generated

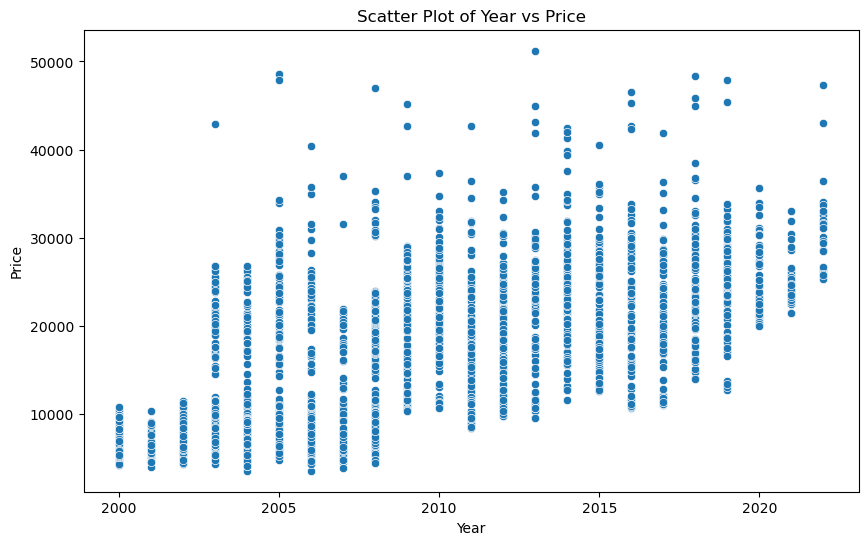
Comment: The mode is 2 previous owners, indicating that most cars have had this number of owners. The spread includes cars with up to 4 owners, with a few outliers having no previous owners. 

Comment: The mean engine size is around 2.0-2.5 liters, with the median close to this value as well, indicating a normal distribution. The spread (standard deviation) shows variability from around 1.0 to 3.5 liters, with most engines clustering around the central mean value.

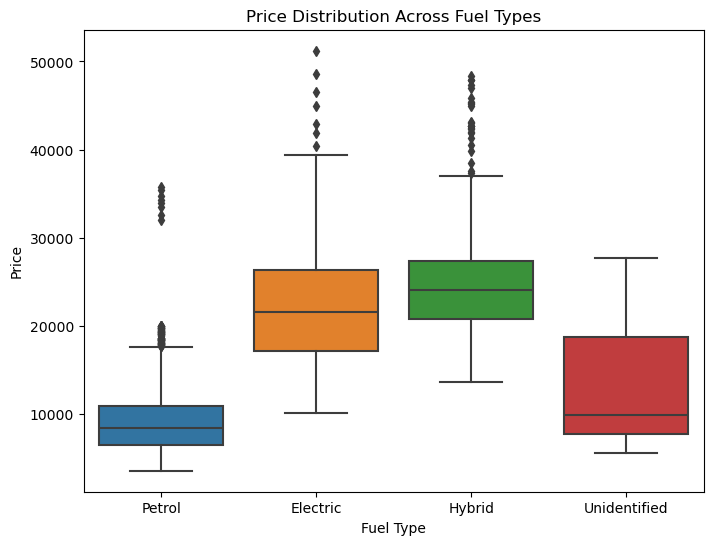
* **Bi-Variate Analysis for Predictors**



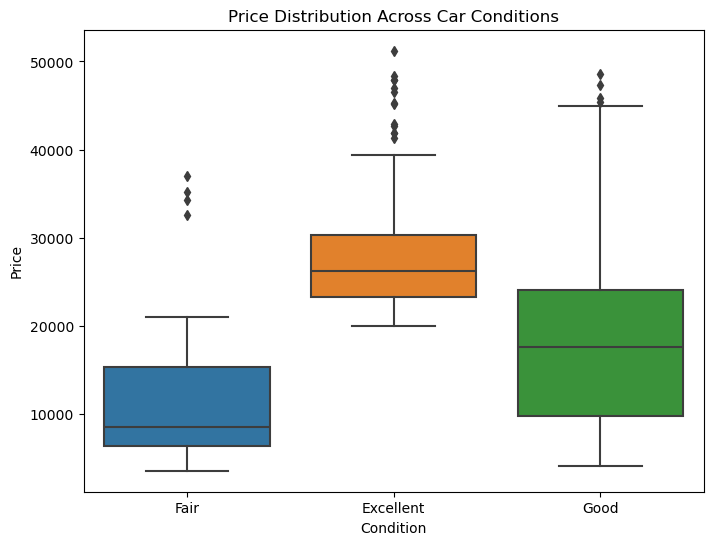
Comment: Significant variability in prices across models. Models like the Ford Escape and Chevrolet Impala show higher price ranges, indicating these models could be strong predictors of price.



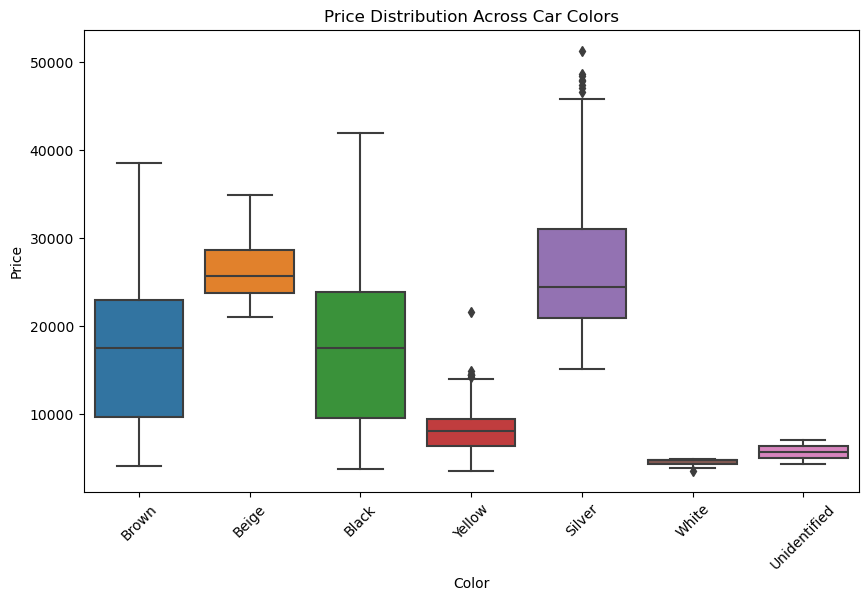
Comment: Newer cars generally have higher prices. The year is a strong predictor of price, with a clear upward trend.



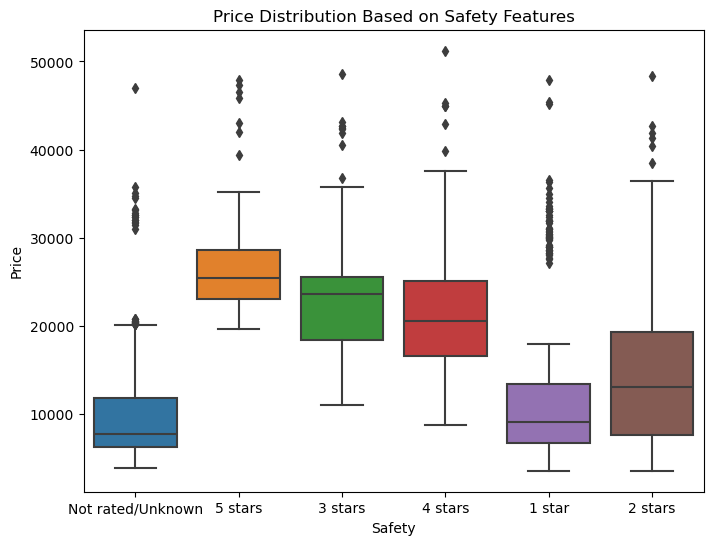
Comment: Electric and hybrid cars have higher median prices compared to petrol cars, suggesting fuel type is a good predictor of price.



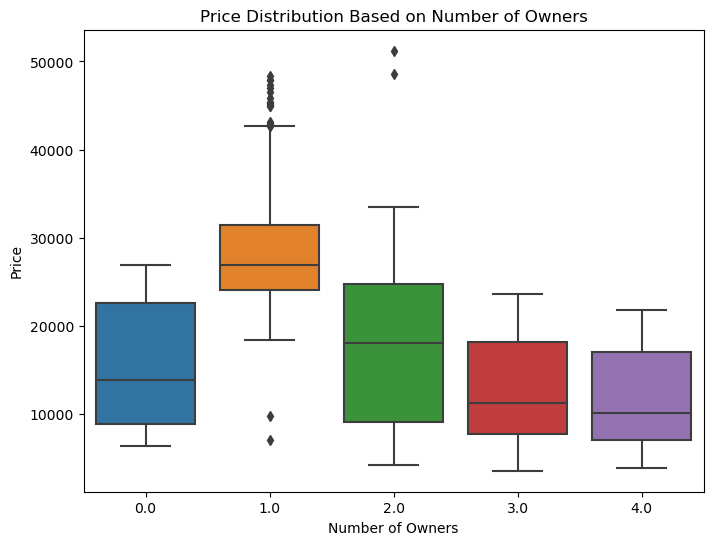
Comment: Cars in excellent condition have significantly higher prices. Condition is a strong predictor of price.



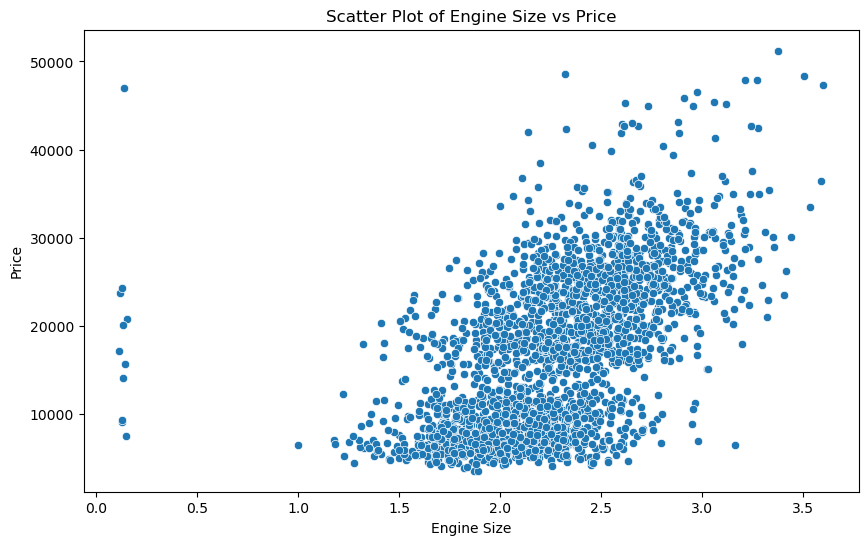
Comment: Some colors, like silver and black, have higher median prices, but overall, color is a less significant predictor of price.



Comment: Higher safety ratings correlate with higher prices, making safety features a good predictor of price.



Comment: Fewer previous owners correlate with higher prices. The number of owners is a moderate predictor of price.



Comment: Larger engines generally command higher prices. Engine size is a strong predictor of price, showing a positive correlation.

* **Data Quality Assessment & Treatment**

**Outliers:** Data points significantly different from others in the dataset, which can be due to variability or errors.

Extremes: Specific type of outlier that lies at the far end of the distribution

Manually identified based on domain knowledge and visual inspections, reduced extreme values to more plausible figures (e.g., changing 44 owners to 4).

Standardized values for consistency (e.g., converting large engine sizes from cc to liters).

**Missing Values:** Occur when no data is available for a variable in an observation, possibly due to data entry errors or non-responses.

For Categorical Data: Converted missing values to a placeholder like 'Unidentified' to make them explicit.

For Numerical Data: Missing or placeholder values were either corrected or replaced with appropriate values (e.g., non-numeric values in Weight were replaced with 0).

## **Model and result**

* **Predictive Modelling Formulation**

In this assignment, we aim to predict the price of cars using a cleaned dataset. This involves formulating a predictive model, determining the type of problem, partitioning the data, selecting performance metrics, setting up a baseline model, performing feature engineering, and improving the model's performance.

* **Type of Problem**

To find the sale price based on different variables, we use the regression method because our target variable (price) is continuous.

* **Partitioning**

Data partitioning is essential to evaluate the performance of the predictive model effectively. By splitting the dataset into training and testing sets, we can train the model on one subset and evaluate its performance on another, ensuring that the model generalizes well to unseen data.

* **Performance Metrics**

We use regression analysis, and the outputs of our analysis are:

• **MAE (Mean Absolute Error): Calculates the average magnitude of error.**

**• RMSE (Root Mean Squared Error): Measures the square root of the average of squared errors.**

**• MSE (Mean Squared Error): Measures the average of squared errors.**

We take MAE as our primary performance metric.

* **Baseline Model**

We take one model as our baseline model to compare other models against it. Using the Decision Tree, we calculate our baseline model. We run our dataset only on testing data. The performance metrics for the Decision Tree model are:

• **MAE:** 2392.3247

• **MSE:** 11448996.7983

• **RMSE:** 3383.6366

These values serve as our benchmark for comparison with other models.

* **Further Understanding**

To further understand the data, we have done analysis using three models – Support Vector Machine (SVM), Random Forest, and Decision Tree.

**1. Decision Tree**

The Decision Tree model is our baseline model. Decision trees are commonly used for regression analysis because they are simple to understand and interpret. The performance metrics for the Decision Tree model are:

• **MAE:** 2392.3247

• **MSE:** 11448996.7983

• **RMSE:** 3383.6366

**2. Support Vector Machine (SVM)**

Support Vector Machines are supervised learning models that can be used for regression by minimizing the error between predicted and actual values. The performance metrics for the SVM model are:

• MAE: 6894.5803

• MSE: 63247833.0477

• RMSE: 7952.8506

Compared to the Decision Tree, the SVM model has higher errors across all metrics, indicating that it may not be as effective for this particular dataset.

**3. Random Forest**

Random Forest is an ensemble learning method that creates multiple decision trees to make a forest. It helps in improving the accuracy of the model. The performance metrics for the Random Forest model are:

• MAE: 1918.3656

• MSE: 7407168.8633

• RMSE: 2721.6114

The Random Forest model outperforms the Decision Tree in all metrics, indicating a better fit for the data.

**Model Parameters Used**

The following table summarizes the parameters used for each model:

Model Parameters Used

| **Model** | **Parameters Used** |
| --- | --- |
| Support Vector Machine | Default: kernel = ‘rbf’, C = 10, degree = 3, epsilon = 0<br>Changed: kernel = ‘linear’, C = 200, degree = 4, epsilon = 0 |
| Random Forest | Default: Criterion = ‘absolute\_error’, min\_samples\_leaf = 1, n\_estimators = 100, max\_depth = none<br>Changed: Criterion = ‘absolute\_error’, min\_samples\_leaf = 15, n\_estimators = 100, max\_depth = 20 |
| Decision Tree | Default: Criterion = ‘squared\_error’, max\_leaf\_nodes = none, min\_samples\_leaf = none, max\_depth = 20<br>Changed: Criterion = ‘squared\_error’, max\_leaf\_nodes = 29, min\_samples\_leaf = 29, max\_depth = 20 |

**Model Improvement**

The following table summarizes the improvement in performance metrics after parameter tuning for each model:

Model Training Data (Before) Testing Data (Before) Training Data (After) Testing Data (After)

Support Vector Machine MAE: 85885.96 MAE: 76780.42 MAE: 28564.06 MAE: 26157.12

Random Forest MAE: 28663.427 MAE: 25208.02 MAE: 18582.52 MAE: 18195

Decision Tree MAE: 18638.89 MAE: 17233.937 MAE: 26047.28 MAE: 27820.53

**Error Cost Analysis**

In error cost analysis, we check the error in our model prediction to determine whether it is cost-effective or not. It consists of two methods: overestimation and underestimation.

**Overestimation**

In overestimation, our predicted values are higher than the actual values in our dataset.

**Underestimation**

In underestimation, our predicted values are lower than the actual values in our dataset.

For our selected model, which is Random Forest, the small difference between our training and testing data errors indicates minimal overestimation and underestimation.

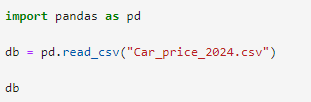
## **Conclusion**

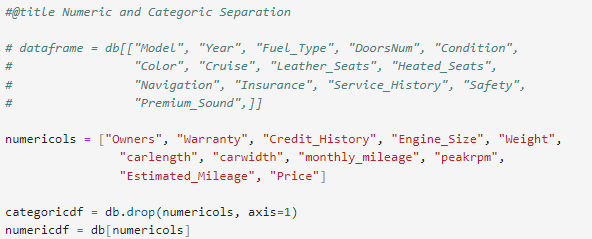
During our analysis, we ran our dataset on different models. Using regression techniques, our optimum model was Random Forest, with a mean absolute error of 1918.3656. This indicates that our model still has errors, but it performs significantly better than the baseline Decision Tree model and the SVM model. To improve the model further, we can use techniques such as feature engineering or look for more precise data.

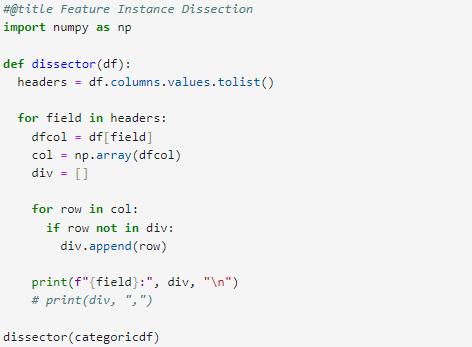
## **Bibliography**

* IBM. (n.d.). Random Forest. Retrieved from https://www.ibm.com/uk-en/topics/random-forest
* NIST. (n.d.). Engineering Statistics Handbook: Section 1. Retrieved from https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm
* Techtarget. (n.d.). Support Vector Machine (SVM). Retrieved from https://www.techtarget.com/whatis/definition/support-vector-machine-SVM#:~:text=A%20support%20vector%20machine%20(SVM,which%20are%20labeled%20for%20classification
* Towardsdata science. (n.d.). How to Do Error Analysis to Make All of Your Models Better. Retrieved from https://towardsdatascience.com/how-to-do-error-analysis-to-make-all-of-your-models-better-a13c4ca643a
* Wikipedia contributors. (2023, July 2). Outlier. In Wikipedia, The Free Encyclopedia. Retrieved from https://en.wikipedia.org/w/index.php?title=Outlier&oldid=1136047256
* Wikipedia contributors. (2023, July 2). Missing data. In Wikipedia, The Free Encyclopedia. Retrieved from <https://en.wikipedia.org/w/index.php?title=Missing_data&oldid=1054697233>

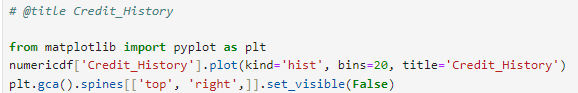
## **Appendix**

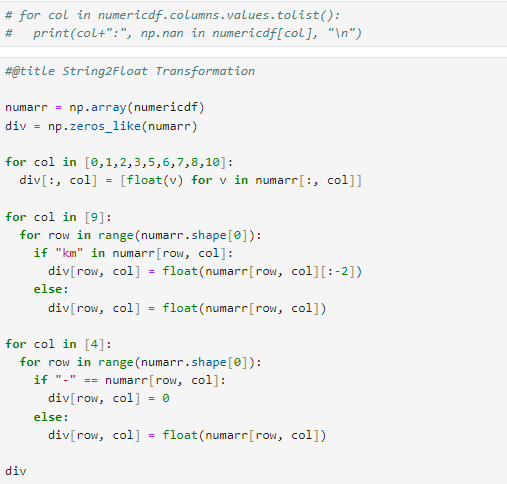












\