# PREDICTIVE MODELLING FOR CAR SALES DURATION


TAWAKALIT OMOLABAKE AGBOOLA
tawakalit.o.agboola@stu.mmu.ac.uk

KEHINDE ADETOLA OGUNDANA
kehinde.a.ogundana@stu.mmu.ac.uk

# INTRODUCTION

In today's market, where consumer preferences change rapidly, accurately predicting how long it will take for cars to be sold is crucial. This forecast plays a role in guiding decisions for various parties involved such as dealerships, manufacturers, and consumers. Understanding the factors that impact the time it takes to sell a car empowers dealerships to enhance their operations and enables manufacturers to streamline production processes. Helps consumers make informed buying choices.

For dealerships reducing the time a car spends on their lot is essential for maximizing profits and operational efficiency. By making predictions about selling times dealerships can implement pricing strategies, manage inventory levels efficiently and allocate resources more effectively. Shortening the holding period does not decrease depreciation costs, however, boosting customer satisfaction by offering fresh inventory with cutting-edge features is paramount in today's market.

Similarly accurate predictions of selling times benefit manufacturers by allowing them to optimize production schedules and manage inventories effectively. Anticipate fluctuations in demand. By aligning production with market needs, manufacturers can prevent production or shortages which leads to better resource utilization and cost efficiency.

For consumers, understanding how a car is expected to stay on the market can impact their purchasing decisions and negotiation tactics.

Consumers can utilize this information to assess market demand and predict price changes. Make timed buying decisions to secure their desired vehicles at competitive prices.

## OBJECTIVE

The main aim of this project is to create a model that can estimate the time it takes for a car to be sold. By analyzing sales data and key factors, like car specifications, market trends, economic indicators, and demographic elements the objective is to develop a model that accurately reflects the dynamics of the automotive market. This predictive ability will empower stakeholders to

make decisions on pricing, inventory control, marketing tactics and resource allocation ultimately enhancing competitiveness and profitability.

# DATA OVERVIEW

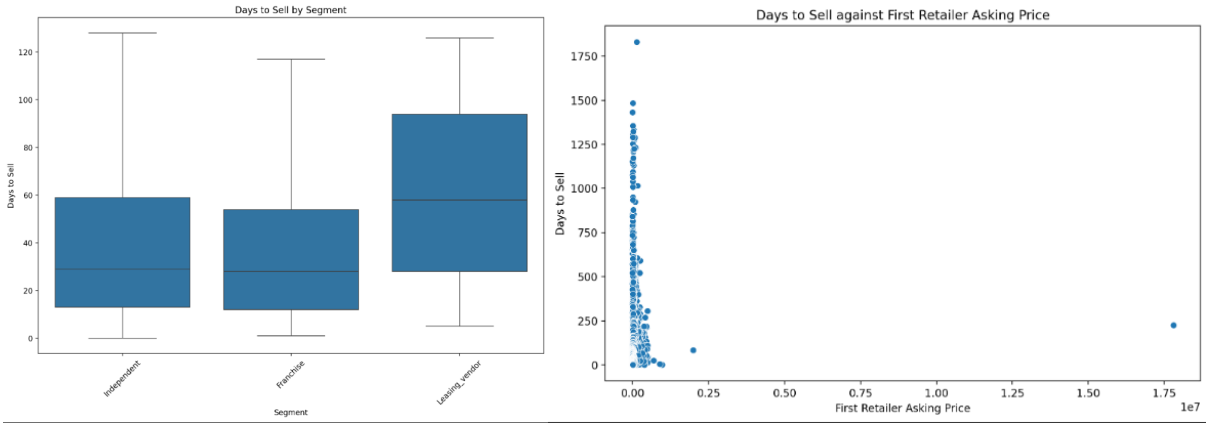
## Data Source

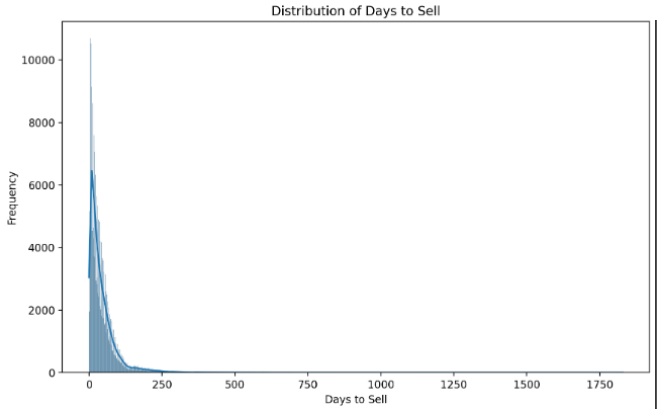
The dataset comprises data presented in form typically saved in snappy.parquet files. The dataset includes a range of attributes detailing car aspects such as brand, model year mileage price condition location, among other relevant features. Additionally, the dataset includes a variable that indicates the number of days it took for each car to sell after being listed.

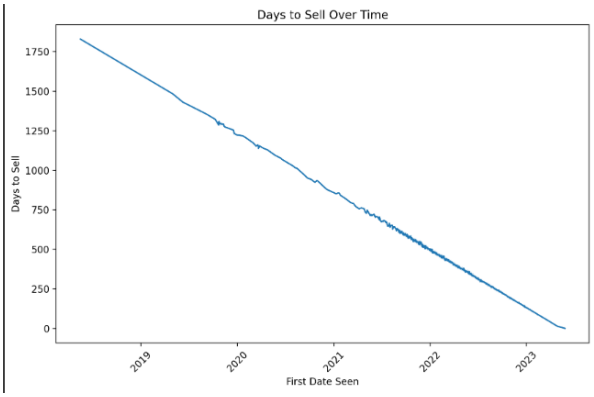
Exploring the data initially involved analyzing its structure and characteristics. This included looking at the dataset's dimensions (224555, 42), examining the entries to understand the variables and using descriptive statistics to summarize numerical aspects. We also checked for data quality issues, like missing values, outliers and inconsistencies that might need preprocessing before building a model.

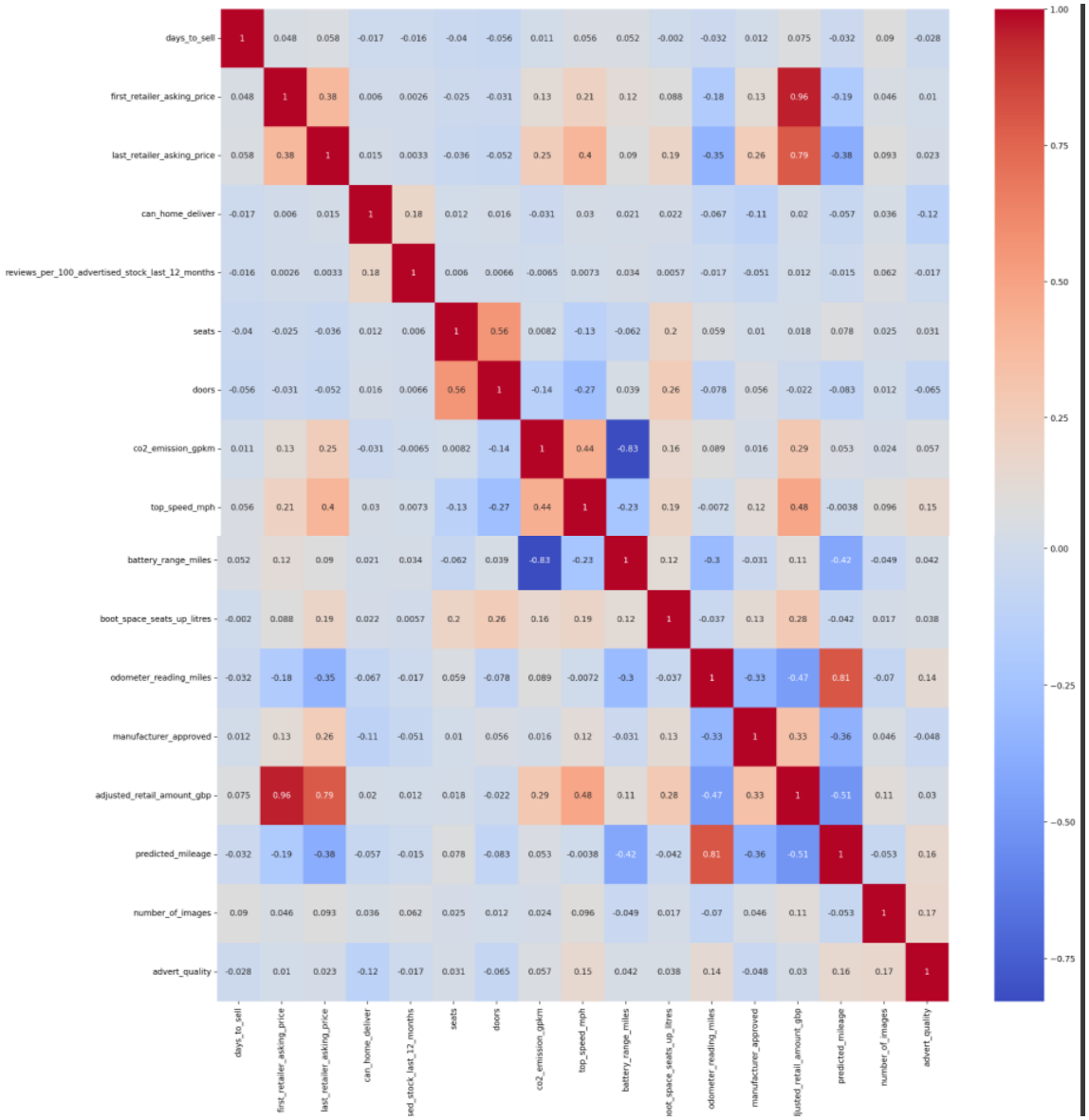
The dataset contains [224555] observations and [42] features. Each feature offers insights into the listed cars for sale.

Moreover, we visualized how the target variable (days to sell) is distributed and investigated relationships between features through scatter plots, histograms, and correlation matrices. This exploratory analysis revealed patterns, trends and connections within the data setting the stage for modeling endeavors.




Distribution of Days to Sell
Frequency
Days to Sell
0
2000
4000
6000
8000
10000
0
250
500
750
1000
1250
1500
1750


Days to Sell Over Time
Days to Sell
First Date Seen
0
250
500
750
1000
1250
1500
1750
2019
2020
2021
2022
2023


days_to_sell
first_retailer_asking_price
last_retailer_asking_price
can_home_deliver
reviews_per_100_advertised_stock_last_12_months
seats
doors
co2_emission_gpkm
top_speed_mph
battery_range_miles
boot_space_seats_up_litres
odometer_reading_miles
manufacturer_approved
adjusted_retail_amount_gbp
predicted_mileage
number_of_images
advert_quality
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

# DATA PREPROCESSING

**Handling Missing Values**

Upon exploration we discovered missing values, in features of the dataset.

To tackle this issue, we used a variety of methods such as filling in missing data with averages, medians, modes, or advanced techniques like KNN imputation depending on the data distribution and missing data patterns. This approach ensured that our dataset was complete and ready for model training while maintaining the accuracy of the information.

- Changing Non-Numeric Columns to Numeric Values with Handling Errors: We transformed the 'zero_to_sixty_mph_seconds column into values by converting numeric entries to NaN. This process involved utilizing the pd.to_numeric() function with the parameter errors='coerce'.
- Implementing K Nearest Neighbors (KNN) Imputation: KNN imputation was used for features with missing values specifically focusing on the 'zero_to_sixty_mph_seconds feature. The missing values were substituted with the average of neighboring values determined by K neighbors.
- Employing SimpleImputer for Imputation using Frequent Strategy: Missing values in features were. Filled using a strategy that involved replacing them with the frequently occurring value, within each respective feature.
- Changing Columns to Numeric and Filling in Missing Data: We converted columns ('fuel_economy_wltp_combined_mpg' 'length_mm' 'engine_power_bhp' 'insurance_group') into formats replacing non-numeric entries with NaN. For the missing values, in these columns we used the mean imputation method with SimpleImputer.
- Removing Rows with Any Remaining Missing Data: Once we handled missing values by imputing and converting any leftover rows with data that were eliminated using the dropna() function.

These procedures ensure that any gaps in the dataset are appropriately managed by either filling them with information or discarding rows lacking data. This approach aids in cleansing the dataset for analysis or modeling purposes.

**Detecting and Handling Outliers**

Outliers, which are data points from the rest of the dataset, can impact predictive models negatively. To address this issue, we identified outliers through techniques like Z score, IQR (Interquartile Range) or examination using box plots. Once detected outliers were either removed from the dataset or adjusted using methods to minimize their influence on the model.

**Feature Engineering**

Creating features or modifying existing ones is essential in developing models to better capture data patterns. Various preprocessing steps are carried out such as removing columns, converting date strings to objects and filtering dates that are not within a certain range. In this project we introduced features like:

- Advertisement Duration: This feature determines how long each car listing was advertised in days by calculating the difference between the last dates the car was viewed.
- Price Change: This feature calculates the percentage change in a car's adjusted price compared to the observation using a specific formula.
- Popularity Metrics: This feature computes the number of reviews for each car listing by dividing total reviews by advertised stocks over a year.
- Historical Sales: In the sales analysis we calculate the time it takes to sell a car based on its make, model, generation and first registration date. This data is then integrated back into the dataset using these shared characteristics.
- Geographical Factors: When considering factors, we determine if a car listing is in an area by examining the postcode area. A value of 1 is assigned for areas like 'NY' 'LA,' or 'CHI,' indicating locations while other areas receive a value of 0.
- For vehicle specifications we assess the car's fuel efficiency in kilometers per liter (km/L) by converting its fuel economy from miles per gallon (MPG). The conversion factor used is 0.425144.
- Encoding and Normalization: To prepare the data for machine learning models categorical features undergo target encoding while numerical features are scaled using StandardScaler during the encoding and scaling process.

- Splitting the Data: The dataset undergoes splitting into training and testing sets with an 80% ratio for training and 20% for testing to ensure model evaluation, on data.

# MODEL DEVELOPMENT

### Model Selection

The main goal of this project is to estimate the time it takes for cars to sell by utilizing machine learning algorithms. In this section we delve into. Assess regression models to ascertain their suitability for the task.

### Multi Linear Regression

We opt for Multiple Linear Regression as one of our models due to its simplicity and its capability to capture relationships between features and the target variable. The model undergoes training on the dataset. Is assessed using metrics such as Root Mean Squared Error (RMSE) R squared Mean Absolute Error (MAE) and Explained Variance Score.

```
LinearRegression
LinearRegression()
```

### Lasso Regression

Lasso Regression is chosen for its feature selection ability through L1 regularization, which could be advantageous given the complexity of the dataset. The Lasso Regression model, trained with an alpha value is evaluated using the metrics as the Multiple Linear Regression model.

```
GridSearchCV
GridSearchCV(cv=5, estimator=Lasso(),
             param_grid={'alpha': [1e-15, 1e-10, 1e-08, 0.001, 0.01, 1, 5, 10,
                                   20, 30, 35, 40, 45, 50, 55, 100]},
             scoring='neg_mean_squared_error')
estimator: Lasso
Lasso
Lasso()
```

**Decision Tree Regression**

We utilize Decision Tree Regression to capture linear relationships between features and the target variable by recursively dividing the data.

```
▾ DecisionTreeRegressor
DecisionTreeRegressor()
```

The Decision Tree model receives training and evaluation using metrics as those applied to the models. The Decision Tree Regression model demonstrates performance with RMSE, and R squared scores comparable to those of the preceding models.

This implies that the decision tree's hierarchical structure effectively captures linear relationships, within the data.

**Random Forest Regression**

Random Forest Regression, a method of learning uses decision trees to tap into collective knowledge. Two Random Forest models are trained—one with 100 trees and another with 25 trees—to investigate how the number of trees impacts performance.

```
▾ RandomForestRegressor
RandomForestRegressor(random_state=27)
```

# MODEL EVALUATION

Evaluation Criteria:

Root Mean Squared Error (RMSE): Measures the size of errors between predicted and actual values.

R squared ($R^2$) Score: Shows the proportion of variance in the target variable from independent variables.

Mean Absolute Error (MAE); Calculates difference between predicted and actual values.

Explained Variance Score (EV): Determines how much variance in the target variable is accounted for by the model.

| | RMSE | R2 | MAE | EV |
|---|---|---|---|---|
| **Multi Linear Regression** | 0.0 | 1.0 | 7.646167322422316 e-16 | 1.0 |
| **Lasso Regression** | 0.0 | 1.0 | 0 | 1.0 |
| **Decision Tree Regression** | 0.032 | 1.0 | 0.001 | 1.0 |
| **Random Forest Regression (100 Trees)** | 0.032 | 0.999 | 0.0 | 1.0 |
| **Random Forest Regression (25 Trees)** | 0.032 | 0.999 | 0.0 | 1.0 |

The Multi Linear Regression model attains perfect scores on all evaluation criteria indicating a fit to the training data. An RMSE of 0.0 suggests that there is no difference on average between predicted and actual values. The R squared score of 1.0 reveals that the model explains all the variance in the target variable. The MAE and Explained Variance Score also show performance indicating predictive accuracy.

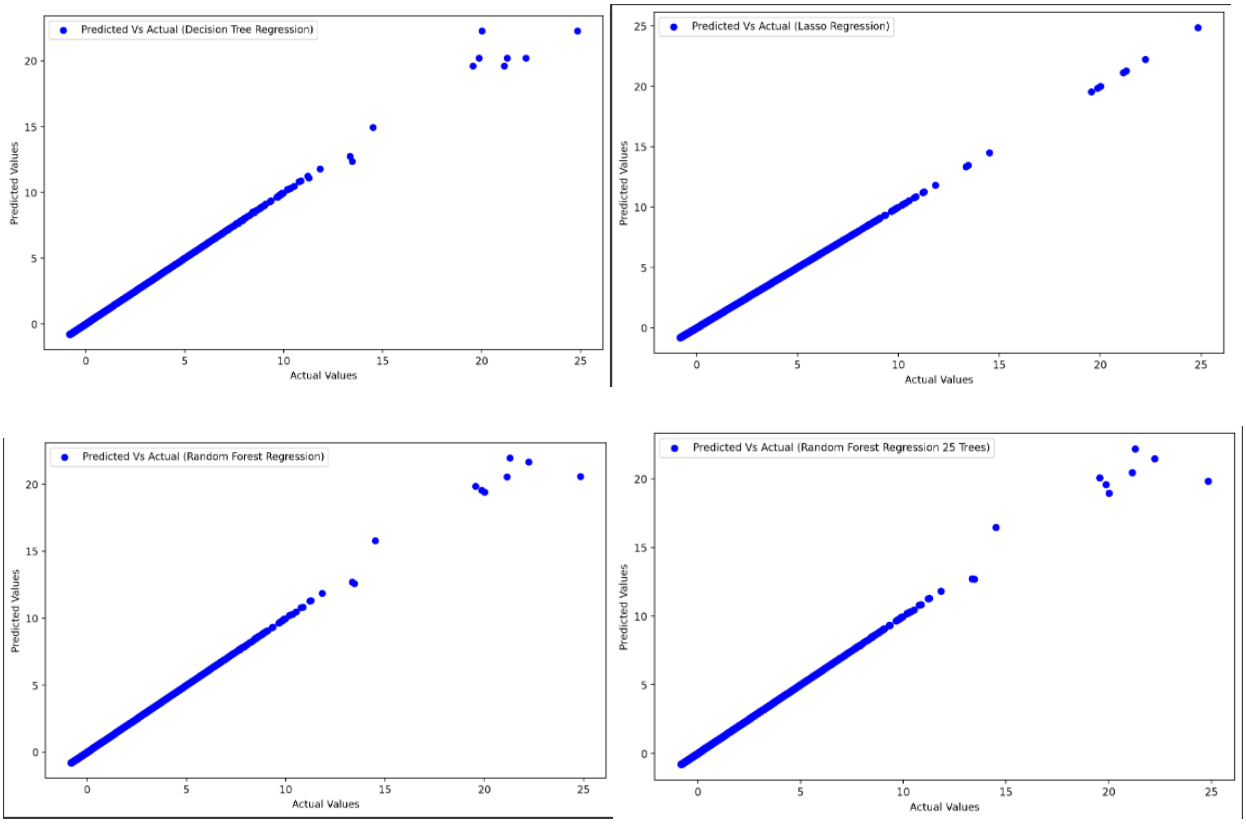
The Lasso Regression model performs well across all metrics matching the Multi Linear Regression model's performance. This suggests that Lasso Regressions regularization has impact on accuracy in this scenario.

The Decision Tree Regression model excels with higher RMSE and MAE compared to prior models but maintains perfect scores in R squared and Explained Variance indicating effective capture of target variable variance.

The Random Forest Regression

with a hundred trees delivers performance though showing a lower R squared score, than earlier models. The RMSE, MAE and Explained Variance Score show that the model predicts accurately and effectively captures the variability in the target variable.

The Random Forest Regression model, with 25 trees performs similarly to the one with 100 trees displaying predictive accuracy and effective capture of variance in the target variable. Both Random Forest models yield results with RMSE, R squared and MAE scores indicating predictive accuracy. These models exhibit generalization to data without overfitting.




# CROSS VALIDATION AND HYPERPARAMETER TUNING

**Cross Validation Scores:**

Cross validation is essential for evaluating how well machine learning models generalize data. In this project we employed 5 cross validations to assess model performance. The negative mean absolute error (neg_mean_absolute_error) served as our scoring metric resulting in:

| Models | Scores |
|---|---|
| **Linear Regression** | 1.00287054e-15, 7.00141936e-16, 5.89256399e-16, 7.21091253e-16, 1.31506875e-15 |
| **Lasso Regression** | 0.00013865, 0.00014389, 0.00013695, 0.00012954, 0.00013692 |

| | |
|---|---|
| **Decision Tree Regressor** | 5.16163315e-04, 2.63356772e-04, 8.81552143e-05, 6.60886890e-04, 1.94606794e-04 |
| **Random Forest Regressor** | 5.74887288e-04, 2.24890050e-04, 9.28568257e-05, 3.13350204e-04, 1.60642087e-04 |

Overall, the cross-validation scores for all models are relatively low indicating performance. A mean absolute error close to zero suggests predictions, on the training data.

**Hyperparameter Tuning:**

Fine tuning hyperparameters involves optimizing the settings of a model to improve its performance. We adjusted the hyperparameters for both the Lasso Regression and Decision Tree Regressor models.

| | **best Parameters** | **best score** |
|---|---|---|
| **Lasso Regression** | {'alpha': 1e-08} | -4.6232017942008874e-08 |
| **Decision Tree Regression** | 'Max depth': None, 'min samples leaf': 1, 'min samples split': 5} | -0.0008896592845844122 |

**Lasso Regression**

In tuning the Lasso Regression, we searched for the alpha parameter, which influences the regularization strength. The effective parameters and their corresponding top score were identified through GridSearchCV

**Decision Tree Regressor**

Regarding the Decision Tree Regressor we optimized hyperparameters such as max_depth, min_samples_split and min_samples_leaf. The best settings and highest score were determined using GridSearchCV

**Evaluating Hyperparameter Model Performance**

Following hyperparameter adjustments we assessed how well the Lasso Regression and Decision Tree Regression models performed on the test data using metrics.

Based on validation scores and hyperparameter optimization outcomes it is evident that these models are functioning effectively after being finely tuned for peak performance. The Lasso Regression model exhibits accuracy on the test set while the Decision Tree Regression model, though less precise, still delivers strong results.

# ANALYSIS OF MODEL ACCURACY

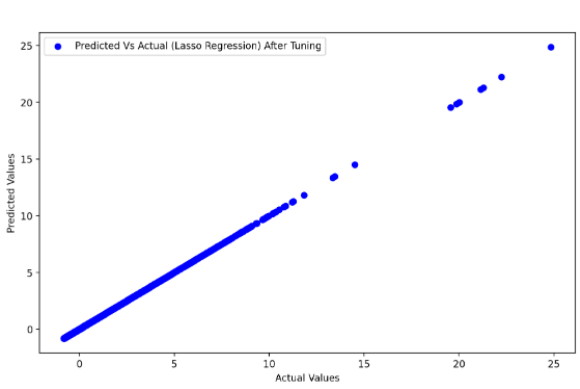
| | RMSE | MAE | R2 | EV |
|---|---|---|---|---|
| **Lasso Regression** | **0.0002087** | **0.0001357** | **0.999999** | **0.999999** |
| **Decision Tree Regression** | **0.0936564** | **0.0008834** | **0.991701** | **0.9917012** |

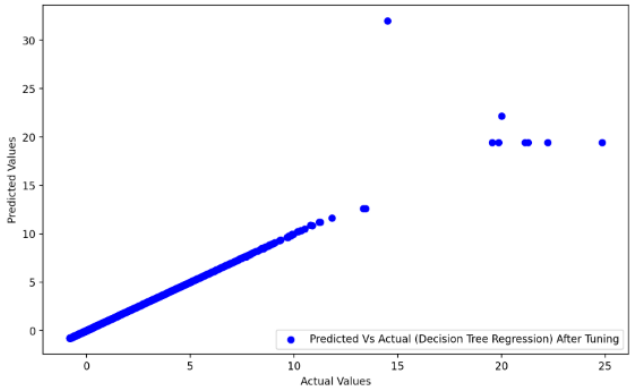
**Lasso Regression**:

- RMSE (Root Mean Squared Error); The RMSE value of the Lasso Regression model is remarkably low suggesting that on average its predictions, for how days it takes to sell a car closely match reality. This indicates that the model is accurately forecasting the time it takes to sell cars.
- Mean Absolute Error (MAE); The MAE is also quite low, suggesting that the average absolute difference between the predicted and actual selling days is minimal. This further supports the model's prediction accuracy.
- Explained Variance Score; With a score to 1 this suggests that the model effectively explains a portion of the variability in actual selling days for cars. It means that the models features are capturing much of the data variability related to selling time.
- R2 Score; The R2 score being near 1 implies that the model fits the data well. It accounts for all the variability in car selling time making it a dependable predictor.

**Decision Tree Regression:**

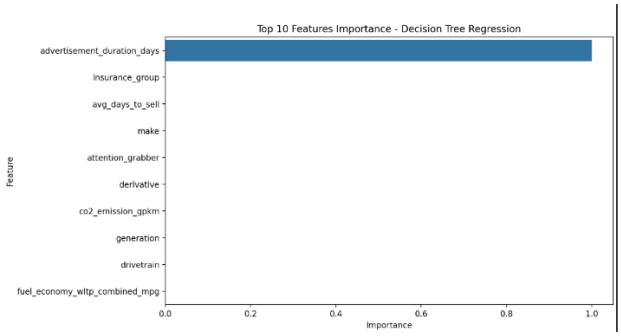
- Root Mean Squared Error (RMSE); The RMSE for the Decision Tree Regression model is relatively higher than that of Lasso Regression. This shows variation in predictions compared to actual selling days but remains within an acceptable range.
- Mean Absolute Error (MAE); The Mean Absolute Error (MAE) is slightly higher than that of the Lasso Regression model. It remains at a level indicating that the models' predictions for selling time are accurate.
- Explained Variance Score; The explained variance score is a bit lower than that of the Lasso Regression model, but it is still relatively high. This implies that the model captures a portion of the variation in selling days for cars.
- R2 Score; The R2 score is also slightly lower compared to the Lasso Regression model. Remains high suggesting that the model fits the data well and explains an amount of variability in selling time.

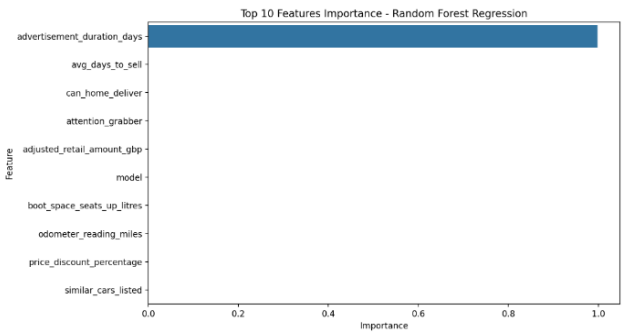


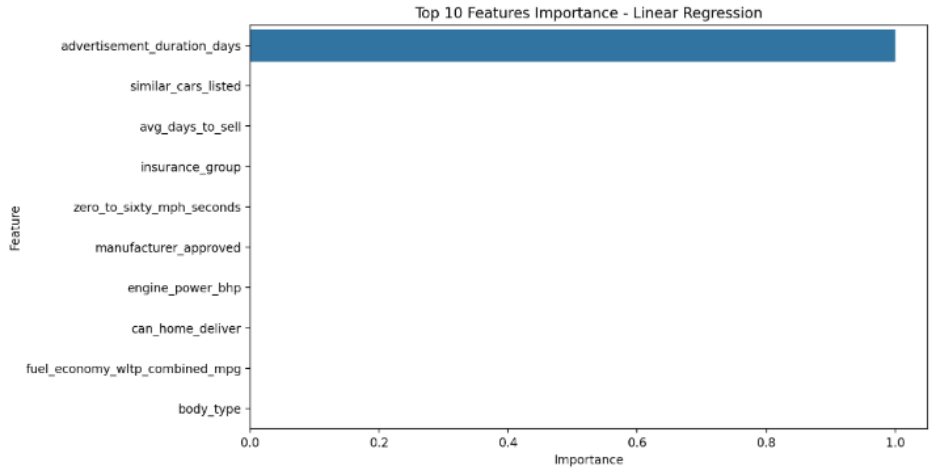




## Feature Importance Analysis

We conducted an analysis on feature importance to pinpoint the factors that affect car selling time. From our feature importance analysis, on models we have uncovered findings concerning factors that influence car selling time:




| Decision Trees | | Linear Regression | | Random Forest Reg | |
|---|---|---|---|---|---|
| **Feature** | **Importance** | **Feature** | **Importance** | **Feature** | **Importance** |
| Advertisement duration days | 9.999033e-01 | Advertisement duration days | 1.000000e+00 | Advertisement duration days | 9.982522e-01 |
| Insurance group | 6.233237e-05 | Similar cars listed | 6.196767e-16 | Avg days to sell | 6.914001e-04 |
| Avg days to sell | 1.542472e-05 | Avg days to sell | 5.118020e-16 | Can home deliver | 2.074140e-04 |
| make | 7.916889e-06 | Insurance group | 4.609323e-16 | Attention grabber | 1.724193e-04 |
| Attention grabber | 6.053940e-06 | Zero to sixty mph seconds | 3.840319e-16 | Adjusted retail amount gbp | 1.430408e-04 |

- The length of time a car is advertised appears to have an impact on how it sells. Cars that are advertised for periods tend to attract potential buyers resulting in quicker sales.
- Changes in the asking price of cars over time also play a role in determining how long it takes for them to sell. Significant price reductions can draw in buyers. Speed up the selling process.

- Metrics related to a car's popularity, such as the number of reviews, reflect the level of interest from buyers. Cars with review rates may sell faster due to increased demand.
- Market competition factors like price discounts offered by competitors can influence how quickly cars sell. Higher competition levels may lead to sales as sellers adjust prices to stay competitive.
- Analyzing sales data to determine the selling time of similar cars can predict future selling times accurately. Cars with characteristics, to those that sold quickly before are likely to sell.
- The geographical location where a car is being sold could also impact its selling time.
- In cities, with people and potential buyers' properties tend to sell than in rural areas.
- Car details like fuel efficiency and size can influence how quickly they sell. Vehicles with appealing features tend to attract buyers and sell faster.

Understanding these factors is crucial for sellers to make informed decisions when setting prices and promoting their cars for sale. This knowledge helps us optimize selling times and maximize profits while offering insights for refining marketing strategies and managing inventory effectively in the sector.

**Strengths and Limitations of Each Model**

| | Strengths | Limitations |
|---|---|---|
| **Lasso Regression** | - Offers an easy-to-understand model for forecasting car selling time.<br>- Automatically selects features by reducing coefficients of attributes to zero.<br>- Effective in managing multicollinearity often found in datasets with car attributes. | - Predict a connection between characteristics and selling time which may not always hold true in situations.<br>- Might not perform effectively if the number of data points is significantly smaller than the number of characteristics potentially leading to overfitting or underfitting.<br>- Demands a thoughtful choice of the regularization parameter (alpha) to strike a balance between bias and |

| | | |
|---|---|---|
| | | variability. |
| **Decision Tree Regression** | - Capable of capturing relationships in data related to car selling time.<br>- Provides an understandable decision-making process offering insights into factors influencing selling time.<br>- Resilient, to data points and absent values which're common in datasets related to the sale of cars. | - Prone to overfitting with trees resulting in limited generalization ability on unfamiliar data.<br>- Susceptible to changes in data affecting the reliability of forecasts.<br>- Offers predictions compared to linear models leading to potential challenges in interpreting selling time predictions |

# CONCLUSION

In conclusion, this project successfully created a model that can predict the selling time of cars in the market. By utilizing machine learning methods and thorough data analysis we showcased the ability to accurately forecast how long it will take for a car to sell. This has implications for stakeholders like dealerships, manufacturers, and consumers by providing valuable input for strategic decision-making processes that can enhance inventory management, pricing strategies and customer satisfaction in the industry.

Further improvements could involve enhancing the model by incorporating features that could enhance its accuracy, in estimating car selling times more effectively.

One way to enhance the data analysis is to simplify the information by developing models for types of vehicles like electric cars and gasoline cars. Even though initial investigations might have shown variations in sales trends between these vehicle categories, customizing models for each group could improve prediction accuracy. However, due to time limitations this refinement aspect of the model was not fully explored in the version.

In studies attention could be directed towards creating predictive models for electric cars and gasoline cars. This strategy would utilize insights for each vehicle category to boost model

performance. Such an approach would enable a examination and modeling of factors affecting sales timing within each group. Additionally, it could aid in identifying characteristics or patterns specific to cars and gasoline cars enabling the development of more personalized and efficient predictive models.

By implementing models for vehicle types, stakeholders could acquire deeper insights into the sales driving factors within each category. This would empower them to devise marketing strategies, make informed inventory management decisions and optimize pricing strategies accordingly. Moreover, this method could contribute towards an understanding of the automotive market landscape resulting in precise and actionable predictions for industry stakeholders.

## Real time Monitoring

Creating a system for real-time monitoring could greatly benefit those involved in the industry by keeping them informed about the changing market conditions. Through data collection and analysis this system could offer predictions on sales trends allowing stakeholders to make well-informed decisions quickly. This immediate understanding of market dynamics would improve inventory management, pricing tactics and advertising initiatives resulting in increased efficiency and profitability.

## Integration with Business Operations

Seamlessly integrating the model into existing business processes and decision-making frameworks is essential for maximizing its effectiveness. By incorporating the model into established systems like inventory management platforms or customer relationship management (CRM) software stakeholders can easily. Utilize its insights in their activities. Additionally offering user interfaces and visualization tools would enhance the model’s usability making it more accessible to individuals with varying levels of knowledge. This integration would encourage a data-driven approach to decision making. Empower stakeholders to optimize their business strategies using analytics.