KASDI Merbah University - Ouargla Faculty of New Technologies of Information and Communication Department of Computer Science and Information Technology

Machine Learning 2: Evaluating and Tunning Machine Learning Models for Real-World Applications

By: Moncef DJEZZAR, and Abdessamad SADOUDI
Under the supervision of: Dr. Belal KHALDI

Abstract: This workshop focuses on evaluating and tuning machine learning models for predicting car sales and resale value. The dataset includes features such as price, size, and other attributes of vehicles. It demonstrates practical steps for model evaluation, feature selection, and hyperparameter optimization using tools such as SHAP and LIME for model interpretability.

Keywords: Machine Learning, Model Evaluation, Hyperparameter Tuning, Explainable AI (SHAP, LIME)

1. Introduction

The growing demand for accurate predictions of car sales volumes and resale values highlights the importance of advanced machine learning techniques. Traditional methods often rely on oversimplified assumptions, while machine learning captures complex patterns in large datasets.

This study evaluates three models: **Linear Regression, Decision Tree, and Random** Forest. Using a dataset with features such as price, engine size, and fuel efficiency, we optimize these models through performance evaluation, hyperparameter tuning, and interpretability tools like **SHAP, PDP** and **LIME** to ensure accuracy and transparency in predicting sales volumes and resale values.

2. Exploratory Data Analysis (EDA)

2.1. Missing Values Analysis:

- We found many missing values in the `__year_resale_value` column
- We filled these missing values with 0, since they likely meant no resale occurred
- We removed the few rows that had missing values in other columns

2.2. Data Quality Checks:

- We checked for duplicates and found none
- We looked for outliers using histograms and box plots
- While we found some outliers, we kept them since they were all reasonable values

2.3. Descriptive Statistics:

	count	mean	std	min	25%	50%	75%	max
Sales_in_thousands	152	53.359072	68.93838	0.11	13.714	29.213	68.06975	540.561
__year_resale_value	152	13.879539	12.707497	0	7.66875	12.8925	17.80625	67.55
Price_in_thousands	152	27.331822	14.418669	9.235	17.88875	22.747	31.93875	85.5
Engine_size	152	3.049342	1.049818	1	2.3	3	3.575	8
Horsepower	152	184.809211	56.823152	55	147.5	175	211.25	450
Wheelbase	152	107.413816	7.717839	92.6	102.9	107	112.2	138.7
Width	152	71.088816	3.464666	62.6	68.375	70.4	73.1	79.9
Length	152	187.059211	13.471247	149.4	177.475	186.65	195.125	224.5
Curb_weight	152	3.376184	0.636593	1.895	2.96475	3.336	3.8215	5.572
Fuel_capacity	152	17.959211	3.937582	10.3	15.775	17.2	19.8	32
Fuel_efficiency	152	23.842105	4.304788	15	21	24	26	45
Power_perf_factor	152	76.704153	25.180983	23.276272	59.755537	71.514623	89.408406	188.144323

2.4. Relationship Between Performance Factor and Price:

- We noticed a positive relationship between a car's performance factor (Power_perf_factor) and its price (Price_in_thousands), where higher performance factors correspond to higher prices.

2.5. Relationship Between Resale Value and Price:

- We observed a positive correlation between a car's resale value (__year_resale_value) and its price. This relationship became clearer when we excluded cars that did not sell.

2.6. Relationship Between Sales Volume and Price:

- We found that cars priced over 50 thousand dollars tend to have lower sales volumes, while cheaper cars do not necessarily achieve higher sales.

2.7. Correlation Analysis:

- From the correlation heatmap, we identified several significant relationships among numeric variables:
- We found that higher Power_perf_factor values strongly correlate with higher prices, larger engines, and lower fuel efficiency.
- We observed that higher fuel efficiency is associated with smaller car sizes and lower horsepower.
- We noticed a strong correlation between wheelbase and both car sales and car length.

2.8. Car Brand Analysis:

- **Sales**: We analyzed sales volumes across brands and noted significant variations.
- **Resale Value**: We identified brands with higher average resale values.
- **Pricing**: We observed that certain brands consistently have higher average prices, distinguishing premium brands from economy options.

2.9. Car Type Analysis:

- **Resale Value**: We found that some vehicle types, such as SUVs, have higher average resale values.
- **Sales**: We noticed variations in sales volumes across vehicle types.
- **Pricing**: We observed that luxury vehicle types tend to have significantly higher average prices compared to other types.

2.10. Data Preparation

We prepared the dataset for modeling by performing the following steps:

- **Feature and Target Separation**: Removed irrelevant columns (Sales_in_thousands and __year_resale_value) and separated the target variable from the feature set.
- **Numerical Feature Normalization**: Applied Z-score normalization to numerical features using StandardScaler, ensuring consistent scaling with a mean of 0 and a standard deviation of 1.
- **Categorical Variable Encoding**: Converted categorical variables into binary indicators using one-hot encoding, dropping the first category to avoid redundancy.
- **Target Normalization**: Normalized numerical target variables using Z-score normalization for consistency with the feature scaling.

3. Model Selection and Implementation

We selected the following machine learning models to predict **__year_resale_value** and **Sales_in_thousands**:

1. **Linear Regression**: A baseline model to identify linear relationships between predictors and target variables.
2. **Decision Tree Regressor**: Captures non-linear relationships by recursively partitioning the dataset based on feature values.
3. **Random Forest Regressor**: An ensemble method that averages multiple decision trees to improve performance and reduce overfitting.
4. **Least-Squares Linear Regression**: A statistical approach minimizing the sum of squared residuals to achieve the best fit.

3.1. Model Evaluation

- **TSS, RSS, ESS**: Analyzed variance decomposition to evaluate explained and unexplained variability in the target variable.
- **R^2**: Measured the proportion of variance in the target variable explained by the model.
- **Adjusted R^2**: Accounted for the number of predictors to ensure fair model comparison.
- **Mallows' Cp**: Balanced goodness-of-fit with model simplicity to avoid overfitting.
- **AIC**: Penalized overfitting by combining model complexity and fit.
- **BIC**: Penalized complex models, especially effective for large datasets.

3.2. Cross-Validation for Model Evaluation

We implemented **5-fold cross-validation** to evaluate the performance of the selected models—**Linear Regression, Decision Tree, and Random Forest**.

3.3. Results

3.3.1. Results for __year_resale_value:

Model	TSS (±std)	RSS (±std)	ESS (±std)	R^2 (±std)	Adjusted R^2 (±std)	Mallows' Cp(±std)	AIC (±std)	BIC (±std)
Linear Regression	29.434 (±13.954)	22.071 (±8.772)	13.489 (±5.828)	0.008 (±0.815)	-0.026 (±0.843)	0.819 (±0.320)	80.247 (±11.983)	84.490 (±11.995)
Decision Tree	29.434 (±13.954)	39.031 (±12.839)	28.705 (±6.644)	-0.493 (±0.452)	-0.545 (±0.467)	1.454 (±0.497)	98.475 (±7.897)	102.718 (±7.873)
Random Forest	29.434 (±13.954)	23.959 (±6.603)	9.888 (±3.872)	0.041 (±0.436)	0.008 (±0.451)	0.892 (±0.253)	83.835 (±8.241)	88.077 (±8.232)

3.3.2. Result Analysis for __year_resale_value:

Decision Tree

- **R^2**: The negative average value (-0.493) indicates that the model performs worse than predicting the mean of the data, highlighting poor predictive performance.
- **RSS**: Significantly higher (39.031) compared to other models, demonstrating the model's struggle to minimize errors.
- **Complexity Metrics**: Mallows, AIC, and BIC values are all higher, reflecting increased model complexity and poor fit.
- **Conclusion**: The Decision Tree is the worst-performing model for this target variable.

Linear Regression

- **R^2**: Near zero (0.008 on average), indicating the model captures almost no variance in the data. The high standard deviation (±0.815) highlights instability across folds.
- **RSS**: Lower (22.071) than the Decision Tree, meaning it fits the data better overall.
- **Complexity Metrics**: Mallows, AIC, and BIC are slightly lower than those of the Random Forest, indicating a better simplicity and performance trade-off.
- **Conclusion**: Linear Regression is relatively better than the Decision Tree but remains unstable and weak in explaining variance.

Random Forest

- **R^2**: Slightly better than Linear Regression (0.041 on average), indicating it captures more variance. However, its instability is evident from the high standard deviation (±0.436).
- **RSS**: Higher (23.959) than Linear Regression but still far better than the Decision Tree.
- **Complexity Metrics**: Slightly higher, AIC, and BIC than Linear Regression, suggesting increased complexity without substantial performance improvement.
- **Conclusion**: The Random Forest is slightly better than Linear Regression at capturing variance but is more complex and not significantly better overall.

General Observations

- Adjusted is slightly lower than for all models, indicating the possibility of overfitting by including unnecessary features.
- **Overall Best Model**: Linear Regression emerges as the best option for __year_resale_value despite its instability, as it achieves lower RSS, AIC, and BIC while being less complex.

3.3.3. Results for Sales_in_thousands:

Model	TSS (±std)	RSS (±std)	ESS (±std)	R^2 (±std)	Adjusted R^2 (±std)	Mallows' Cp(±std)	AIC (±std)	BIC (±std)
Linear Regression	29.845 (±19.953)	21.245 (±10.939)	16.945 (±5.723)	0.192 (±0.323)	0.163 (±0.335)	0.789 (±0.413)	77.346 (±14.998)	81.589 (±14.958)
Decision Tree	29.845 (±19.953)	43.757 (±20.164)	26.814 (±22.427)	-0.771 (±0.814)	-0.833 (±0.843)	1.635 (±0.772)	100.249 (±12.337)	104.492 (±12.303)
Random Forest	29.845 (±19.953)	24.675 (±15.215)	9.618 (±4.034)	0.134 (±0.237)	0.103 (±0.246)	0.919 (±0.574)	80.883 (±16.227)	85.126 (±16.194)

3.3.4. Result Analysis for Sales_in_thousands:

Decision Tree

- **R^2**: Negative (-0.771) on average, again performing worse than the mean prediction.
- **RSS**: Significantly higher (43.757) compared to other models, confirming poor fit.
- **Complexity Metrics**: Mallows, AIC, and BIC are the highest among the models, indicating high complexity with little to no performance gain.
- **Conclusion**: The Decision Tree performs the worst, similar to its performance for the other target variable.

Linear Regression

- **R^2**: Positive (0.192 on average), indicating the model explains some variance in the data. The standard deviation (±0.323) suggests moderate stability.
- **RSS**: Lower (21.245) than other models, indicating better fit to the data.
- **Complexity Metrics**: Achieves the lowest, AIC, and BIC, reflecting better performance and simplicity compared to the other models.
- **Conclusion**: Linear Regression is the best-performing model for this target variable.

Random Forest

- **R^2**: Slightly worse than Linear Regression (0.134 on average), with comparable standard deviation (±0.237).
- **RSS**: Higher (24.675) than Linear Regression, indicating less accurate predictions.
- **Complexity Metrics**: Higher, AIC, and BIC than Linear Regression, reflecting its increased complexity without a corresponding improvement in performance.
- **Conclusion**: Random Forest is more complex and performs slightly worse than Linear Regression.

General Observations

- **Overall Best Model**: Linear Regression again proves to be the best choice due to its simplicity and better overall performance metrics.

4. hyperparameter tuning

We optimized the performance of our regression models—Linear Regression, Decision Tree, and Random Forest—by tuning their hyperparameters to predict two target variables: __year_resale_value and Sales_in_thousands. The optimization process involved defining a hyperparameter space, applying **GridSearchCV** with 5-fold cross-validation, and evaluating each model's performance.

4.1. Hyperparameter Grids

We explored the following hyperparameter grids for each model:

Model	Hyperparameter	Values Explored
Linear Regression	No hyperparameters	(No hyperparameters to tune)
Decision Tree	max_depth	[None, 5, 10, 20]
	min_samples_split	[2, 5, 10]
Random Forest	n_estimators	[50, 100, 200]
	max_depth	[None, 10, 20]
	min_samples_split	[2, 5]

4.2. Results and Analysis

The best hyperparameter combinations obtained for each model are as follows:

Linear Regression	__year_resale_value	(No hyperparameters)
	Sales_in_thousands	(No hyperparameters)
Decision Tree	__year_resale_value	{'max_depth': 5, 'min_samples_split': 10}
	Sales_in_thousands	{'max_depth': 5, 'min_samples_split': 2}
Random Forest	__year_resale_value	{'max_depth': 20, 'min_samples_split': 2, 'n_estimators': 200}
	Sales_in_thousands	{'max_depth': 20, 'min_samples_split': 5, 'n_estimators': 100}

4.2.1. Performance Comparison

Results for Target: __year_resale_value:

Linear Regression		
Best Parameters		
Metrics	TSS	151.0000
	RSS	0.0000
	ESS	151.0000
	R2	1.0000
	Adjusted_R2	1.0000
	Mallows_Cp	0.0000
	AIC	-9562.6798
	BIC	-9553.6082

Decision Tree		
Best Parameters	max_depth	5
	min_samples_split	5
Metrics	TSS	151.0000
	RSS	42.5867
	ESS	108.4133
	R2	0.7180
	Adjusted_R2	0.7161
	Mallows_Cp	0.2875
	AIC	243.9617
	BIC	253.0334

Random Forest		
Best Parameters	max_depth	None
	min_samples_split	2
	n_estimators	100
Metrics	TSS	151.0000
	RSS	15.3478
	ESS	88.8396
	R2	0.8984
	Adjusted_R2	0.8977
	Mallows_Cp	0.1036
	AIC	88.8349
	BIC	97.9065

Results for Target: Sales_in_thousands:

Linear Regression		
Best Parameters		
Metrics	TSS	151.0000
	RSS	0.0000
	ESS	151.0000
	R2	1.0000
	Adjusted_R2	1.0000
	Mallows_Cp	0.0000
	AIC	-9604.3100
	BIC	-9595.2383

Decision Tree		
Best Parameters	max_depth	10
	min_samples_split	2
Metrics	TSS	151.0000
	RSS	11.7318
	ESS	139.2682
	R2	0.9223
	Adjusted_R2	0.9218
	Mallows_Cp	0.0792
	AIC	47.9978
	BIC	57.0694

Random Forest		
Best Parameters	max_depth	10
	min_samples_split	2
	n_estimators	200
Metrics	TSS	151.0000
	RSS	20.3218
	ESS	77.8652
	R2	0.8654
	Adjusted_R2	0.8645
	Mallows_Cp	0.1372
	AIC	131.5250
	BIC	140.5966

- To highlight the impact of hyperparameter tuning, we created plots comparing the performance of our models:

For __year_resale_value:

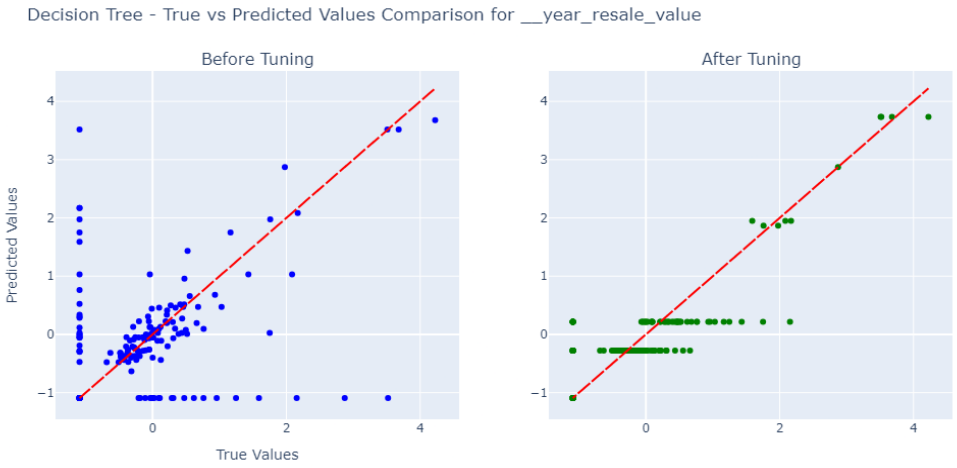

Figure: comparison between Decision Tree before and after hyperparameter tuning

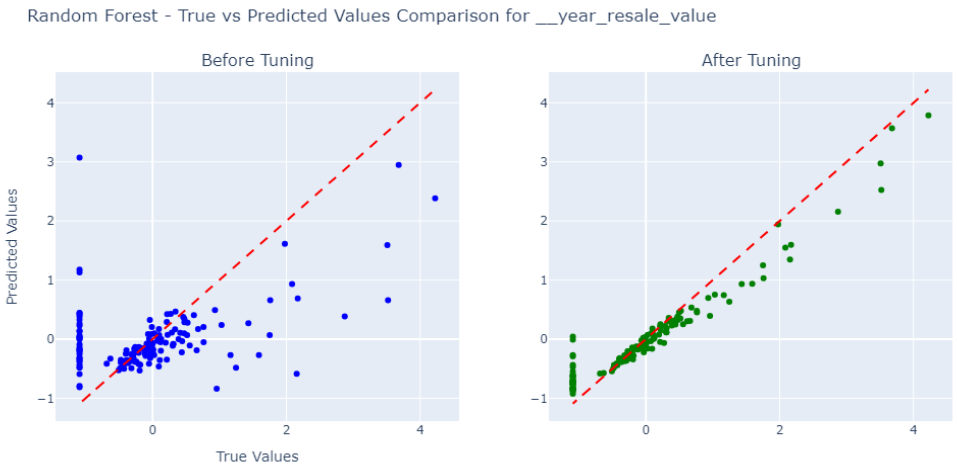

Figure: comparison between Random Forest before and after hyperparameter tuning

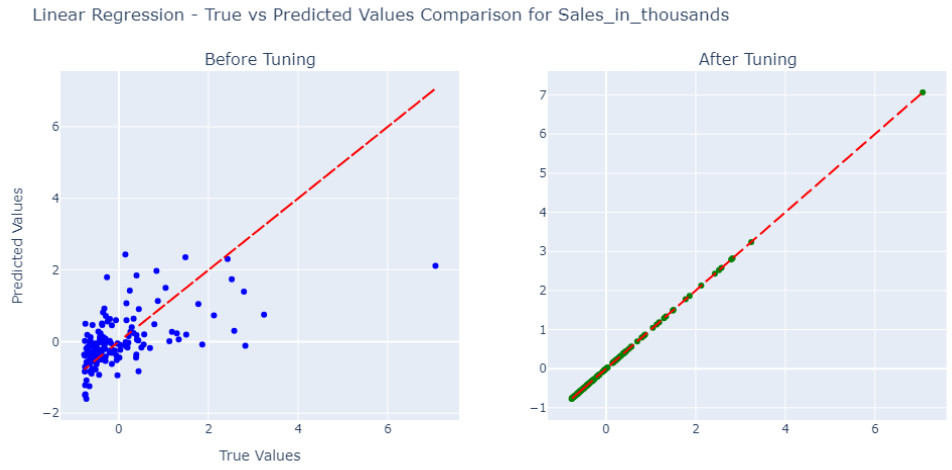

Figure: comparison between Linear Regression before and after hyperparameter tuning

For Sales_in_thousands:

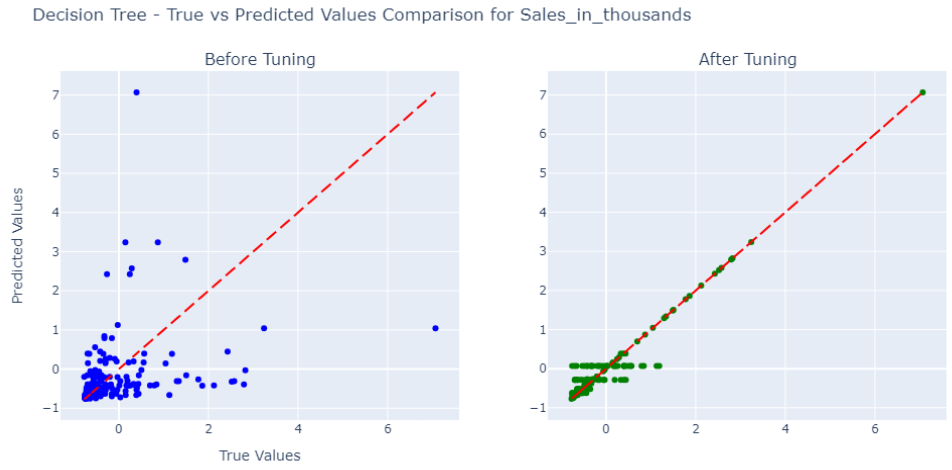

Figure: comparison between Decision Tree before and after hyperparameter tuning

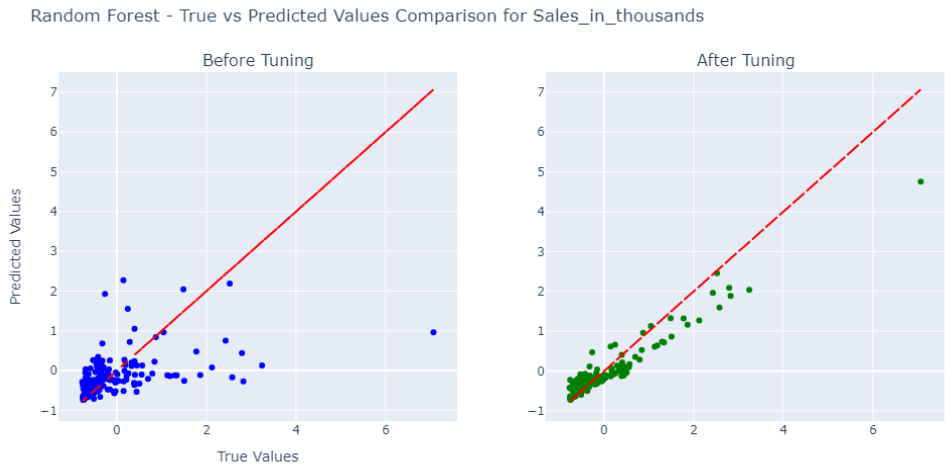

Figure: comparison between Random Forest before and after hyperparameter tuning

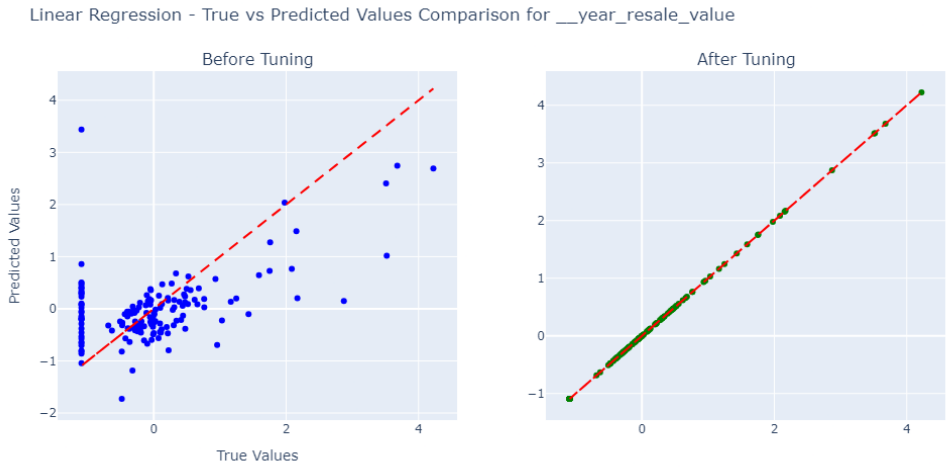

Figure: comparison between Linear Regression before and after hyperparameter tuning

4.3.2. Analysis

- **Linear Regression:** Achieved perfect R^2 and zero RSS after tuning, which raises concerns about potential overfitting. Further analysis is needed to assess generalization capabilities on unseen data.
- **Decision Tree:** Showed notable improvement in R^2 and reduction in RSS after tuning, especially for Sales_in_thousands. This indicates that hyperparameter optimization helped in reducing overfitting and enhancing predictive accuracy.
- **Random Forest:** Demonstrated significant gains in R^2 and RSS after tuning for both targets, indicating a more robust and accurate model. The increased complexity from hyperparameter optimization seems to have improved its predictive power.

4.6. Conclusion

Hyperparameter tuning significantly improved the predictive accuracy of Decision Tree and Random Forest models for car sales forecasting. While Linear Regression achieved perfect training metrics, further validation is needed to assess its generalization and address potential overfitting

5. Interpretability and Explainability

5.1. Feature Importance Analysis

We analyzed the feature importance for our regression models—Decision Tree, Random Forest, and Linear Regression—using relevant methods to identify the most influential features for predicting the target variables: __year_resale_value and Sales_in_thousands.

5.1.1. Decision Tree and Random Forest Models

For tree-based models, we used the feature_importances_ attribute to evaluate the contribution of each feature.

Target: __year_resale_value

- **Decision Tree Notable Features**:
  - Power_perf_factor
  - Model_Cougar

- **Random Forest Notable Features**:
  - Price_in_thousands
  - Width
  - Latest_Launch_6/13/2011

Target: Sales_in_thousands

- **Decision Tree Notable Features**:
  - Model_F-Series
  - Model_Grand Cherokee
  - Latest_Launch_1/4/2012

- **Random Forest Notable Features**:
  - Manufacturer_Ford
  - Engine_size
  - Model_Caravan
  - Latest_Launch_3/6/2012
  - Latest_Launch_3/19/2012

5.1.2. Linear Regression Model

For the Linear Regression model, we analyzed the magnitude of the coefficients to determine feature impact. Features with larger coefficients were identified as having stronger influence, either positively or negatively.

Target: __year_resale_value

- **Features with High Positive Impact**:
  - Manufacturer_Porsche
  - Latest_Launch_5/10/2012
  - Latest_Launch_4/4/2011
  - Model_Carrera Cabrio

- **Features with High Negative Impact**:
  - Model_CL500
  - Model_CLK Coupe
  - Latest_Launch_6/12/2011
  - Latest_Launch_11/3/2012
  - Manufacturer_Chrysler
  - Manufacturer_Chevrolet

Target: Sales_in_thousands

- **Features with High Positive Impact**:
  - Model_F-Series
  - Model_Caravan
  - Latest_Launch_10/21/2011
  - Wheelbase

- **Features with High Negative Impact**:
  - Model_Crown Victoria
  - Manufacturer_Plymouth
  - Latest_Launch_10/9/2011
  - Latest_Launch_7/1/2012

This analysis provides valuable insights into which features most significantly influence the target variables.

5.2. Partial Dependence Plots (PDP)

To gain a deeper understanding of the relationship between individual features and the model's predictions, we utilized Partial Dependence Plots (PDP). PDPs illustrate the marginal effect of a feature on the predicted outcome, holding all other features constant.
These plots provided valuable insights into how the model's predictions change as a function of individual feature values.

For __year_resale_value target:

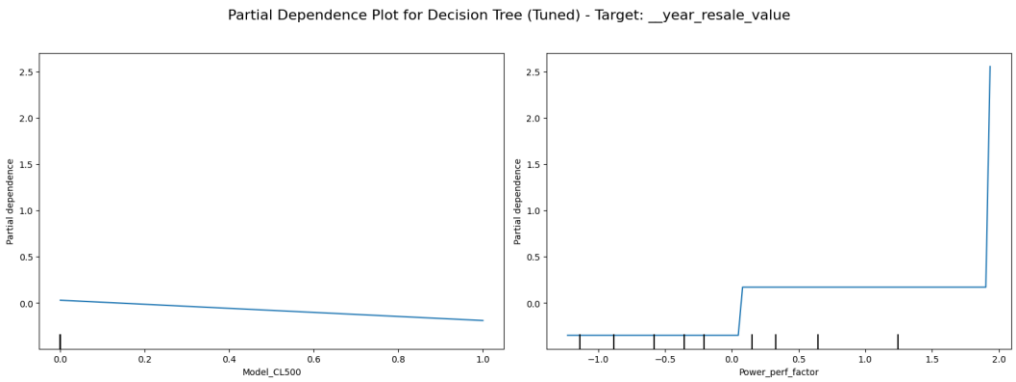

Figure: Partial Dependence of Power_perf_factor and Model_CL500 of Decision Tree model

Analysis:
for Model_CL500:

- This plot shows a negative linear relationship between Model_CL500 and the predicted resale value.
- This suggests that higher values of Model_CL500 are associated with lower resale values, according to the model.

for Power_perf_factor:

- This plot reveals a non-linear relationship between Power_perf_factor and predicted resale value.

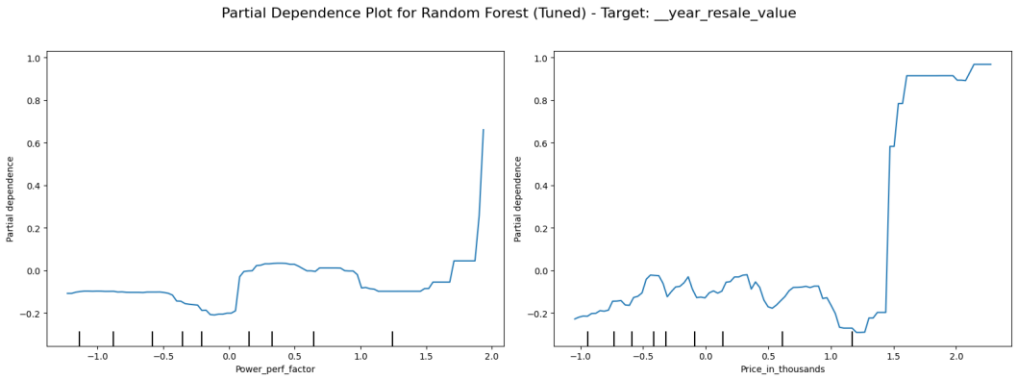

Figure: Partial Dependence of Power_perf_factor and Price_in_thousands of Random Forest model

Analysis:

For Power_perf_factor:

- This plot reveals a non-linear relationship between Power_perf_factor and the predicted resale value.
- Higher values of Power_perf_factor are strongly associated with higher resale values, especially beyond the critical threshold.

For Price_in_thousands:

- This plot also shows a non-linear relationship between Price_in_thousands and the predicted resale value.
- Higher prices are positively correlated with higher resale values, especially for premium items priced beyond a critical range.

For Sales_in_thousands target:

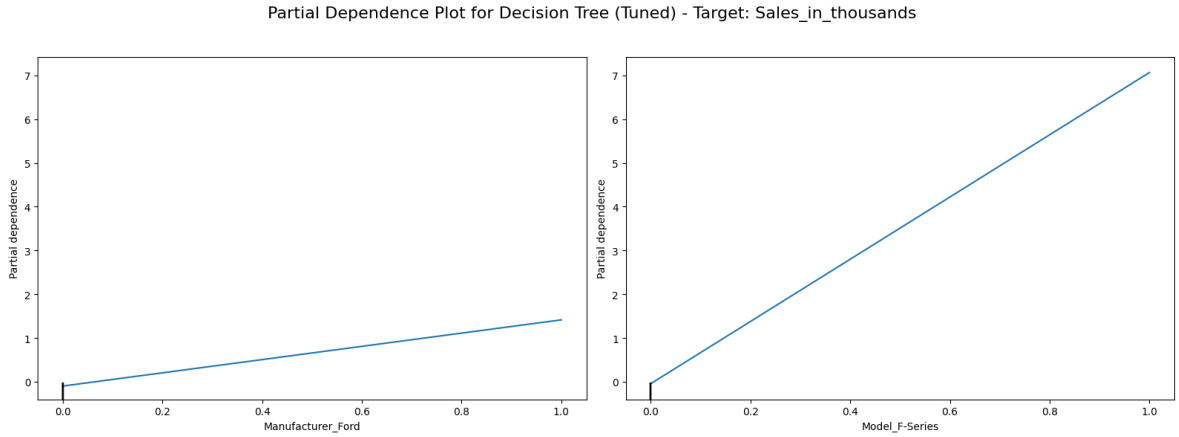

Figure: Partial Dependence of Manufacturer_Ford and Model_F-Series of Decision Tree model

Analysis:

For Manufacturer_Ford:

- Displays a positive linear relationship with the target (Sales_in_thousands).
- Higher values of Manufacturer_Ford are associated with increased sales.

For Model_F-Series:

- Exhibits a strong positive linear relationship with the target.
- The presence of Model_F-Series significantly boosts sales, indicating its strong influence on the target variable.

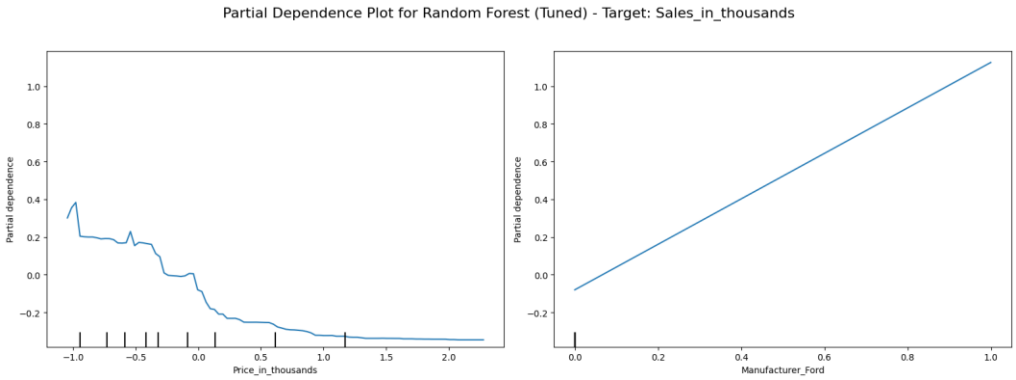

Figure: Partial Dependence of Price_in_thousands and Manufacturer_Ford of Random Forest model

Analysis:

For Price_in_thousands:

- Exhibits a negative non-linear relationship with the target (Sales_in_thousands).
- Higher prices are associated with decreased sales, suggesting that affordability plays a crucial role in driving sales.

For Manufacturer_Ford:

- Displays a strong positive linear relationship with the target (Sales_in_thousands).
- Vehicles manufactured by Ford have a significant positive impact on sales, indicating brand strength and customer preference for Ford.

5.3. Local Interpretable Model-agnostic Explanations (LIME)

We applied LIME to our models to gain insights into the rationale behind specific predictions. By visualizing the LIME explanations, we were able to identify the features that were most influential in driving the model's decision for a given instance.

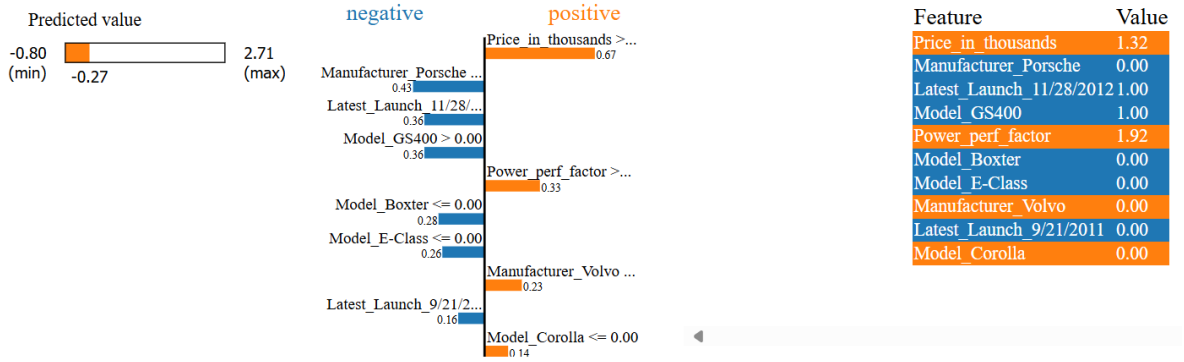

Figure: LIME plot of Random Forest model targeting __year_resale_value

Analysis:
The two features with the strongest contributions (positive and negative) are:

- **Power_perf_factor**: Drives the prediction up.
- **Manufacturer_Porsche**: Drags the prediction down.

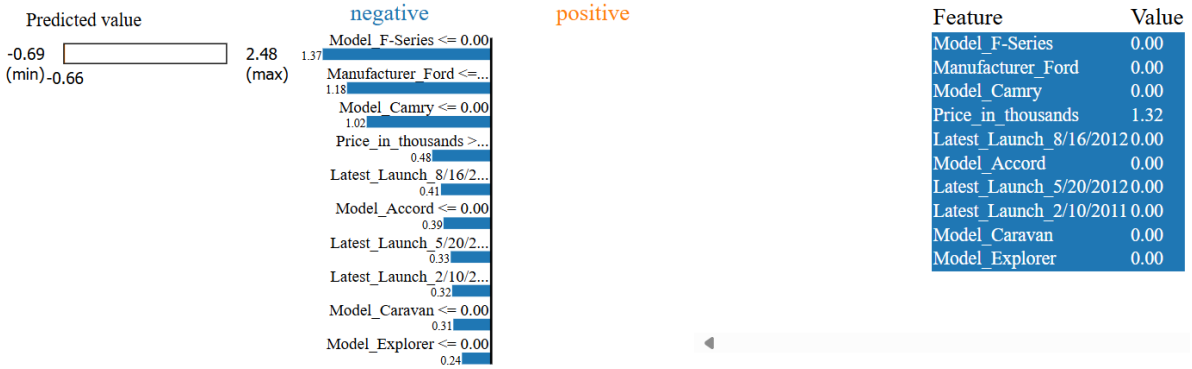

Figure: LIME plot of Random Forest model targeting Sales_in_thousands

Analysis:

- **Model_F-Series (1.37)**: The most significant negative factor.
- **Manufacturer_Ford (1.18)** and **Model_Camry (1.02)** also reduce the prediction significantly.
- No positive contributors are highlighted in this instance, indicating all factors are dragging the prediction down.

5.4. SHapley Additive exPlanations (SHAP)

We applied SHAP to our models to explain individual predictions

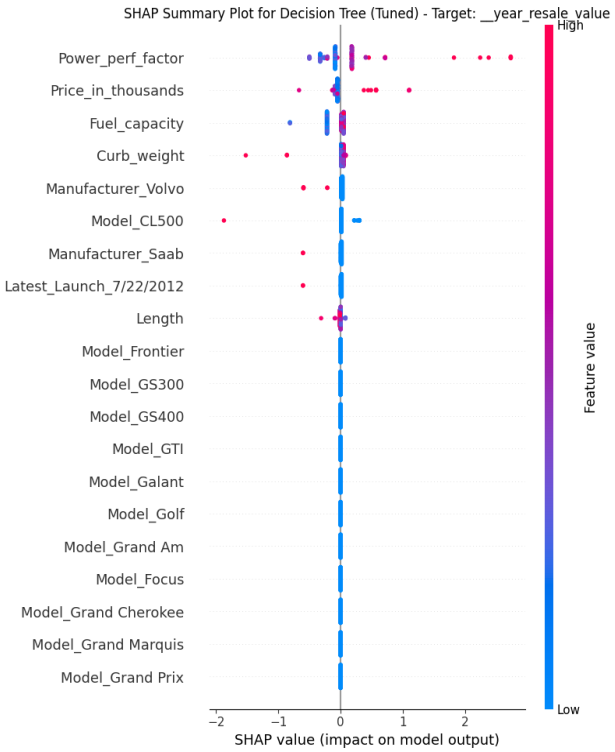

Figure: SHAP plot of Decision Tree model targeting __year_resale_value

Analysis:
The magnitude of the SHAP values determines how much each feature contributes to the model's predictions:

- **Power_perf_factor**: Dominates in importance with a wide range of SHAP values.
- **Price_in_thousands**: Has a substantial impact, reflecting its importance in determining resale value.
- **Fuel_capacity** and **Curb_weight**: Moderate contributors, showing some influence on resale value predictions.
- Features like **Model_CL500**, **Manufacturer_Saab**, and **Latest_Launch_7/22/2012** have smaller but still notable impacts.

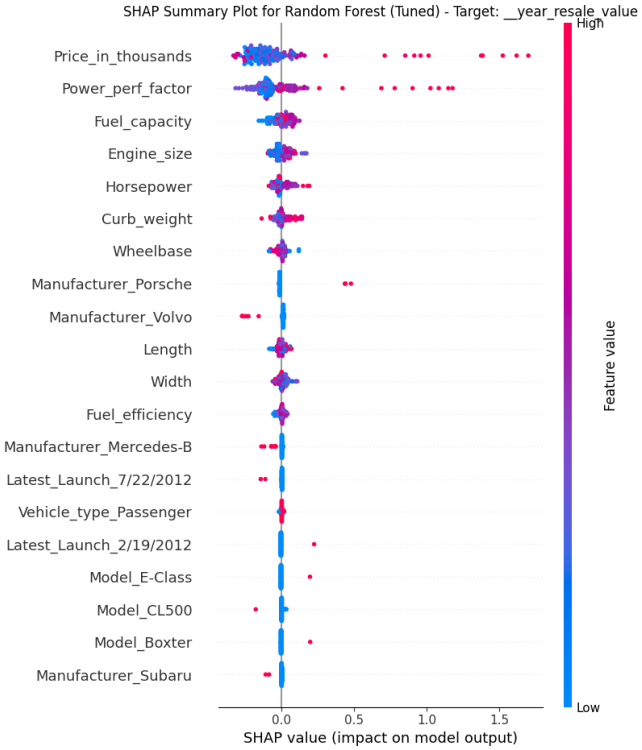

Figure: SHAP plot of Random Forest model targeting __year_resale_value

Analysis:

- **Top drivers**:
  - **Price_in_thousands**: Dominates the model with the widest SHAP value distribution.
  - **Power_perf_factor**: Strongly linked to performance, contributing significantly to predictions.

- **Secondary contributors**:
  - **Fuel_capacity**, **Engine_size**, and **Horsepower**: Reflect the importance of fuel and performance-related features in determining resale value.
  - **Curb_weight** and **Wheelbase**: Influence the resale value, albeit to a lesser extent.

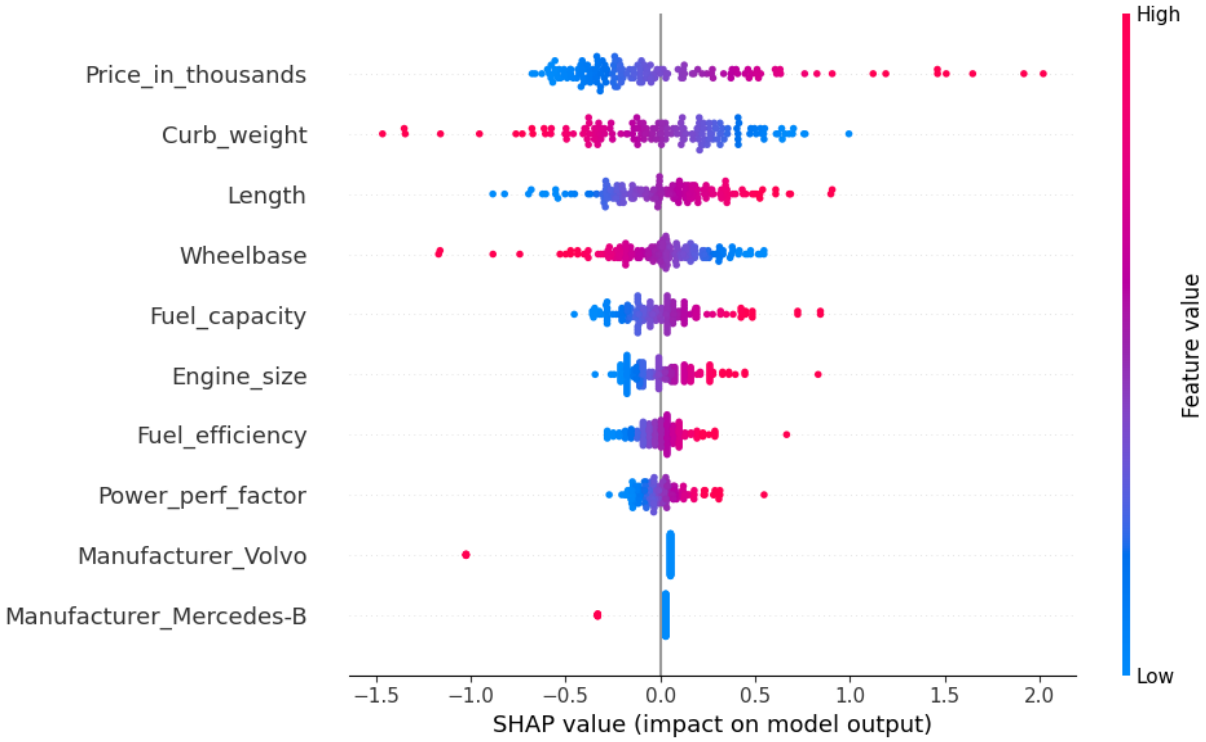

Figure: SHAP plot of Linear Regression model targeting __year_resale_value

Analysis:

- Price_in_thousands has the most significant positive and negative influence on the model predictions.
- Curb_weight and Length also play substantial roles but show moderate impact variability.
- Features like Manufacturer_Volvo and Manufacturer_Mercedes-B have a concentrated distribution around zero, indicating limited influence overall.

Target sales in thousands

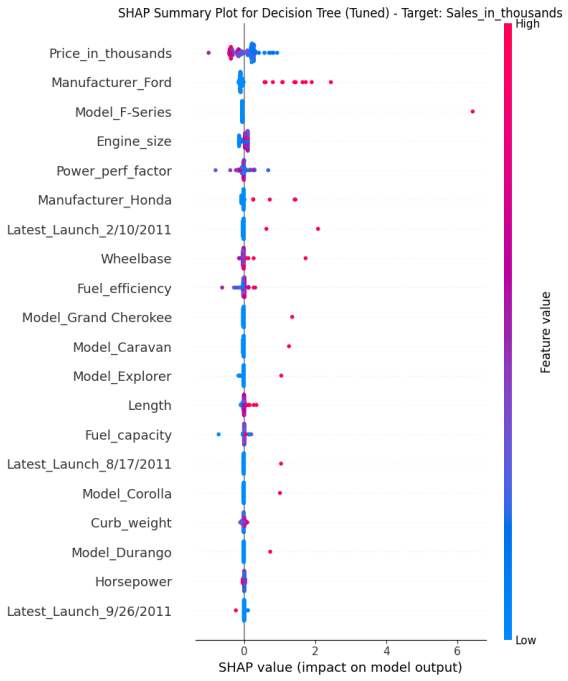

Figure: SHAP plot of Decision Tree model targeting Sales_in_thousands

Analysis:

- **Price_in_thousands** consistently has the largest impact (both positive and negative).
- **Manufacturer_Ford** and **Model_F-Series** also exhibit wide SHAP value distributions, indicating their significant but variable influence.
- Features like **Wheelbase**, **Fuel_efficiency**, and **Latest_Launch_2/10/2011** have lower impacts, as their SHAP values are tightly clustered near 0.

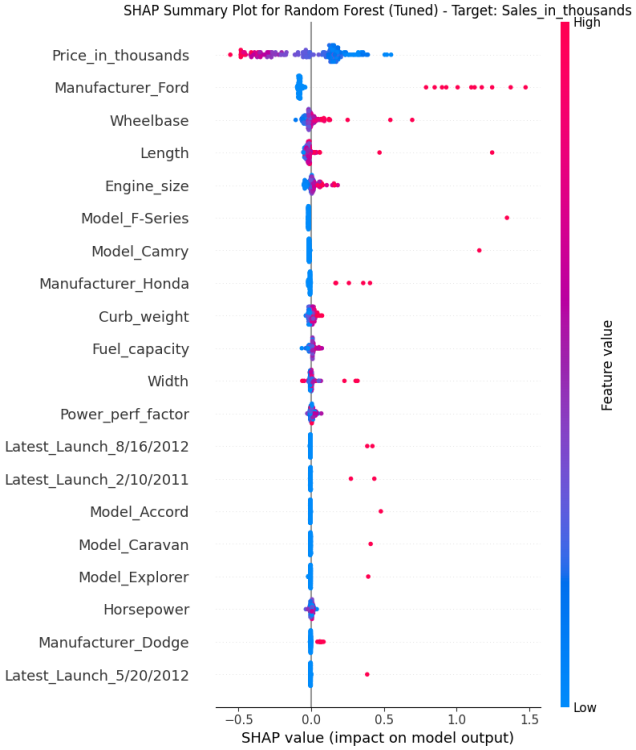

Figure: SHAP plot of Random Forest model targeting Sales_in_thousands

Analysis:

- **Price_in_thousands** dominates in terms of importance and strongly drives model predictions.
- The presence of specific manufacturers (**Ford**, **Honda**) plays a role, but their impact is less variable.
- The model emphasizes pricing and key physical characteristics (e.g., **Wheelbase**, **Length**).

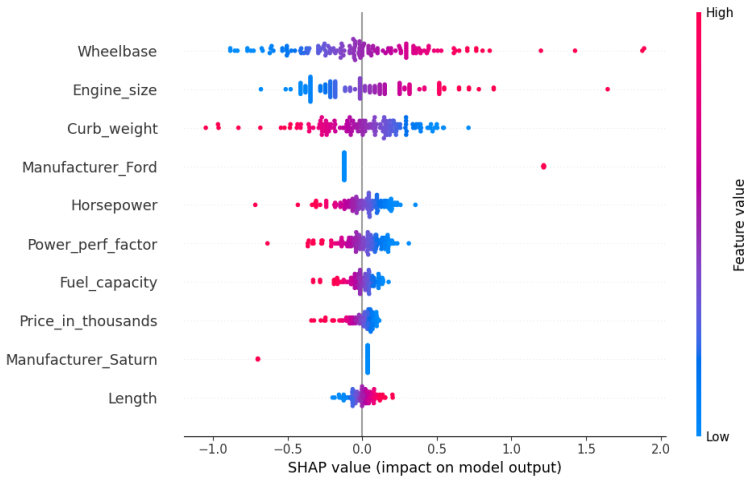

Figure: SHAP plot of Linear Regression model targeting Sales_in_thousands

Analysis:

- Wheelbase is the most important feature, followed by **Engine_size**, **Curb_weight**, and **Horsepower.**
- **Fuel_capacity** and **Price_in_thousands** are also moderately influential.
- Manufacturer-related features (**Manufacturer_Ford**, **Manufacturer_Saturn**) have minimal impact compared to the top numerical features.

6. Final Conclusion

- For both __year_resale_value and Sales_in_thousands, **Linear Regression** is the best model. It balances simplicity, lower complexity metrics (, AIC, and BIC), and better data fit (lower RSS).
- The Decision Tree consistently performs the worst across both targets.