Tutorial on SHAP and LIME: Opening the Black Box of Used-Car Price Predictions

Git hub Link: https://github.com/SHANLYKS47/Machine_Learning_Tutorial

Table of Contents

1. Introduction

Modern machine-learning systems are increasingly deployed in high-value business applications such as pricing, credit scoring, fraud detection, and medical diagnostics. Although these models often deliver superior predictive performance, they tend to behave as opaque "black boxes." Their internal logic—built from nonlinear transformations and complex feature interactions—is rarely intuitive, even to the practitioners who design them.

This opacity creates a real tension: we depend on model outputs to set prices, allocate credit, and support clinical decisions, yet it is difficult to explain or justify individual predictions. Stakeholders, regulators, and end-users therefore continue to ask a fundamental question:

"Why did the model make this prediction?"

In this tutorial, we address that question in two ways:

1. General perspective: We introduce two of the most widely adopted local explainability methods—SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations)—and explain their theoretical foundations, practical algorithms, and typical use cases.

2. Concrete case study: We apply both methods to a RandomForestRegressor trained to predict used-car prices. The model uses structured features such as mileage, model year, brand, color, fuel type, and transmission configuration. We investigate what the model has learned globally and how it arrives at specific price estimates.

Using SHAP and LIME together provides a balanced and defensible interpretability strategy: SHAP offers mathematically principled, additive attributions; LIME offers fast, intuitive surrogate explanations that are highly accessible to non-technical stakeholders.

2. Why Explainability Matters in Machine Learning

Explainability is not an optional add-on; it is a core requirement for responsible machine-learning practice. Across the model lifecycle, it supports several critical objectives.

Model Validation: Explanations help determine whether the model is learning meaningful structure or merely exploiting artefacts in the data. In the used-car context, we want to confirm that the model relies on sensible drivers such as mileage, age, and brand prestige—not on accidental correlations like rare color combinations or data leakage.

Trust and Adoption: Dealers, analysts, and customers are more likely to trust and adopt a pricing system if they can understand its reasoning. Being able to show, for a specific car,

how mileage, year, and options combine to produce the predicted price is far more persuasive than reporting a single number.

Ethics and Fairness: Interpretability enables practitioners to detect unintended biases—such as over-reliance on features that proxy socioeconomic status or demographic information. In pricing and lending, this is particularly important to avoid discriminatory outcomes.

Regulatory Compliance: Many jurisdictions require organizations to justify automated decisions, especially in credit, insurance, and pricing. Explainability tools help generate audit-ready documentation describing how features influence individual outcomes.

Human–Machine Collaboration: Local explanations allow domain experts to challenge, refine, or override algorithmic decisions. For instance, a pricing analyst can review SHAP and LIME explanations for a specific vehicle and decide whether to accept, adjust, or investigate the suggested price.

SHAP and LIME both aim to quantify how much each feature contributes to a particular prediction, thereby directly addressing these business and regulatory needs.

3. SHAP: Shapley Additive Explanations

3.1 Theoretical Foundations

SHAP adapts Shapley values from cooperative game theory to the problem of model interpretability. In this framework:

- Each feature is treated as a "player" in a game.
- The model prediction is the "payout" that results from the coalition of all features.
- The Shapley value for a feature quantifies its fair share of the payout, averaging its marginal contribution over all possible feature coalitions.

SHAP inherits several desirable properties: Efficiency, Symmetry, Monotonicity, and Additivity. These give SHAP a strong theoretical footing and make its attributions particularly suitable for audit and compliance use.

3.2 Practical Algorithms

SHAP comes with a rich, intuitive visual ecosystem that helps translate quantitative attributions into human-readable stories: summary plots, force or waterfall plots, and dependence plots.

3.3 Visualization Suite

SHAP comes with a rich, intuitive visual ecosystem that helps translate quantitative attributions into human-readable stories: summary plots, force or waterfall plots, and dependence plots.

3.4 Strengths and Limitations

Strengths: strong theoretical guarantees; deterministic, stable explanations; unified treatment of global and local interpretability; high efficiency for tree-based models via TreeExplainer.

Limitations: higher computational cost for very large, arbitrary models; interpretation of complex interactions may require additional plots or domain expertise.

4. LIME: Local Interpretable Model-Agnostic Explanations

4.1 Local Surrogate Framework

LIME starts from a different intuition: although a model may be globally complex and nonlinear, its behavior can often be approximated locally by a much simpler model. LIME exploits this idea by fitting a local surrogate model around the instance we want to explain.

4.2 How LIME Works

LIME starts from a different intuition: although a model may be globally complex and nonlinear, its behavior can often be approximated locally by a much simpler model. LIME exploits this idea by fitting a local surrogate model around the instance we want to explain.

4.3 Strengths and Limitations

Strengths: fast, lightweight, model-agnostic, and highly intuitive. Limitations: explanations depend on sampling and can vary between runs; the local surrogate may not fully capture complex nonlinear interactions; and LIME does not enforce axiomatic properties such as efficiency or symmetry.

5. SHAP vs LIME: Comparative Analysis

5.1 Comparative Overview

SHAP is grounded in cooperative game theory, offers high stability, higher computational cost, additive feature attributions that sum to the prediction, and strong global interpretability. LIME is based on local surrogate modelling, provides medium stability due to sampling, lower computational cost, local linear surrogate coefficients that do not necessarily sum to the prediction, and is best suited for rapid exploration and stakeholder communication.

5.2 Practical Guidance for Use

Use SHAP when you need rigorous, repeatable explanations suitable for audit trails, insight into both global and local behavior, and strong performance on tree-based models. Use LIME when you need fast, exploratory explanations during model development and intuitive visualizations for demos and stakeholder communication.

5.3 Complementary Roles

In mature machine-learning workflows, SHAP and LIME are complementary rather than competing tools. SHAP delivers analytical rigour and stable attributions; LIME provides flexibility, speed, and interpretive accessibility. Using both methods on the same model offers a more complete, cross-validated understanding of model behavior.

6. Demonstration and Interpretation

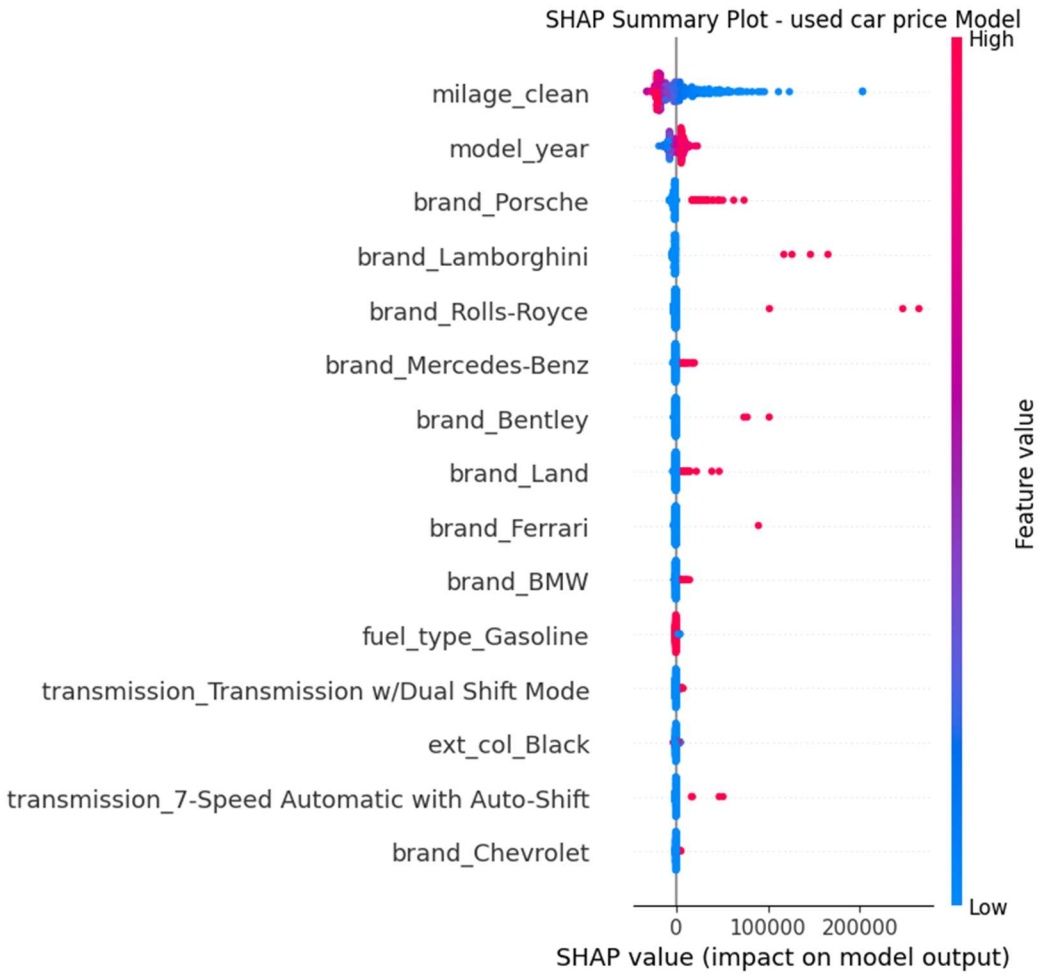
6.1 Model Setup

We train a RandomForestRegressor on a structured dataset of used cars. Features include mileage, model year, brand indicators, fuel type, transmission configuration, and exterior and interior colors. After training the model, we compute SHAP values with TreeExplainer and generate local explanations with LIME's tabular explainer configured for regression.

6.2 Global Interpretability with SHAP (Summary Plot)

Figure 1 presents the SHAP summary plot for the used-car price model. It aggregates feature effects over the entire evaluation set, where each point corresponds to a car and its horizontal position encodes the SHAP value.

Figure 1: SHAP Summary Plot – Used-Car Price Model



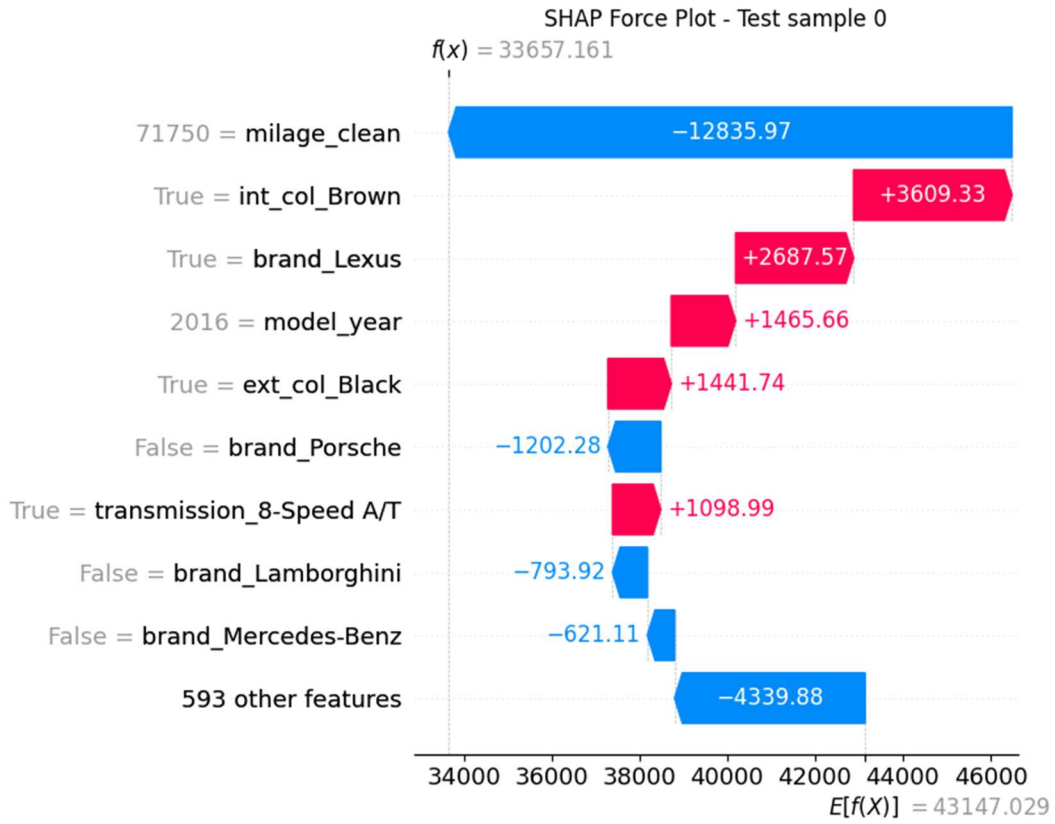
Key observations include: mileage as the dominant driver with high mileage strongly decreasing price; model year with newer vehicles increasing price; premium brands such as Porsche, Lamborghini, Rolls-Royce, Bentley, and Mercedes-Benz providing strong positive contributions; and color and configuration features exerting smaller, fine-tuning effects.

Overall, the global SHAP behavior confirms that the model is learning realistic pricing relationships: mechanical and temporal attributes dominate, while brand and configuration influence value in line with market expectations.

6.3 Local Interpretability with SHAP (Force/Waterfall Plot)

Figure 2 decomposes the prediction for a single vehicle (Test Sample 0) using a SHAP force/waterfall plot. Starting from the baseline price, the plot shows how each feature pushes the prediction up or down to arrive at the final estimate.

Figure 2: SHAP Force/Waterfall Plot – Test Sample 0



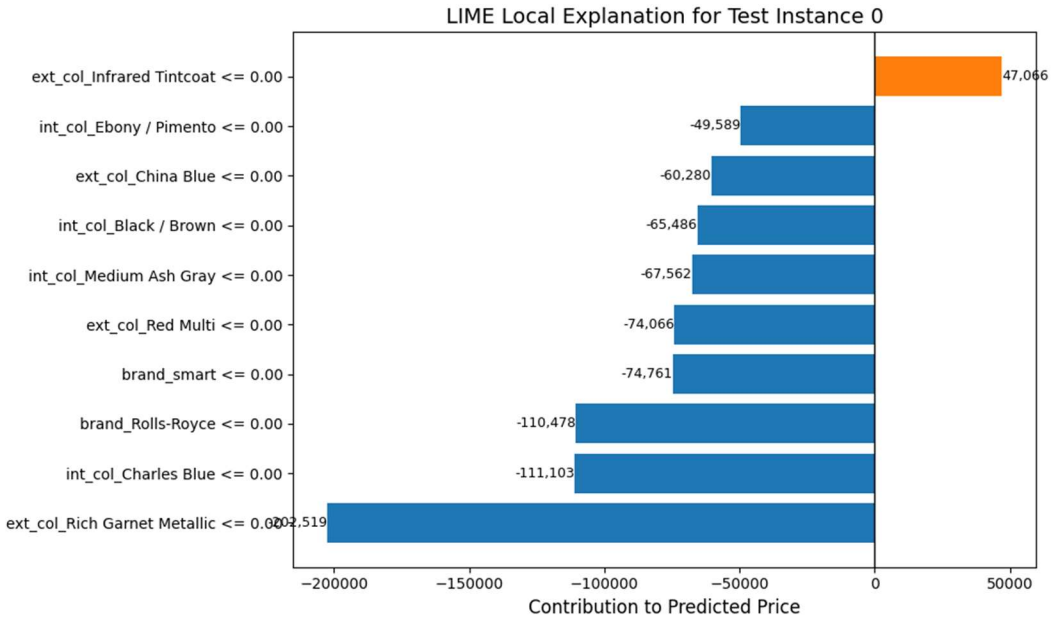
For this instance, mileage of 71,750 miles contributes a large negative SHAP value, indicating substantial depreciation. Interior color Brown and the Lexus brand provide positive contributions, as do the relatively recent model year and a desirable black exterior. The absence of ultra-luxury brands such as Porsche or Lamborghini contributes small negative adjustments, while the combined effect of remaining features accounts for the residual difference.

The SHAP waterfall plot therefore offers a transparent, quantitative narrative: start from the baseline price, apply a large mileage discount, then apply uplifts for a premium brand, relatively new model year, and attractive colors to arrive at the final predicted price.

6.4 Local Surrogate Interpretability with LIME

Figure 3 shows the LIME explanation for the same test instance. LIME fits a simple local linear surrogate model around this car and ranks features by their contribution to the surrogate prediction.

Figure 3: LIME Local Explanation – Test Instance 0



Several exterior and interior color indicators dominate the local linear approximation, along with conditions describing the absence of particular brands. The strongest positive local contributor is the condition associated with not having the exterior color Infrared Tintcoat, while the absence of other colors such as Rich Garnet Metallic contributes negatively. These coefficients reflect how the decision boundary behaves in the immediate neighborhood of this car rather than exact Shapley-style contributions.

LIME's explanation differs from SHAP because it explains the behavior of the local surrogate, not the full RandomForestRegressor. Nonetheless, it confirms that color-related features and certain brand indicators play a non-trivial role in the local decision surface, which can be valuable for debugging unexpected local behaviors.

7. Application Example: Motorsport Analytics

Beyond used-car pricing, SHAP and LIME can be applied to many other domains. In motorsport analytics, teams predict lap times under varying conditions such as tyre temperature, fuel load, aerodynamic configuration, and track temperature. Using SHAP, engineers can identify global drivers of lap time and quantify interactions. Using LIME, they can diagnose why a particular lap was predicted to be unusually slow or fast and detect outlier sensor readings. The same toolkit supports both strategic planning and real-time decision-making.

8. Best Practices and Common Pitfalls

Best Practices: Use SHAP for high-stakes or regulated decision contexts; use LIME for early-stage development and rapid debugging; always cross-check explanations with domain

expertise; and combine local methods (SHAP, LIME) with global diagnostics such as partial dependence plots or accumulated local effects.

Common Pitfalls: Over-interpreting noise, especially in LIME explanations; ignoring feature correlations; treating surrogate models as exact representations of the underlying model; and presenting explanations without communicating uncertainty or limitations.

9. Conclusion

In this report, we opened the black box of a RandomForestRegressor trained to predict used-car prices and used SHAP and LIME to understand its behavior. Globally, SHAP summary plots show that the model's primary drivers—mileage, model year, and brand prestige—are consistent with real-world pricing dynamics. Locally, SHAP force plots provide exact additive decompositions of individual price predictions, while LIME offers complementary local surrogate explanations that are fast and intuitive.

Together, SHAP and LIME form a powerful interpretability toolkit that enhances trust and adoption of predictive models, supports regulatory compliance and ethical review, and empowers domain experts to collaborate effectively with machine-learning systems. By integrating these methods into the development and deployment of models—whether for used-car pricing, motorsport analytics, or other applications—practitioners can design solutions that are not only accurate but also transparent, accountable, and aligned with human judgement.

10. References

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

SHAP Documentation: https://shap.readthedocs.io

LIME Documentation: https://github.com/marcotcr/lime