Modeling Used Car Prices in Turkey: An Explainable Regression Study

Moving Beyond Intuition with Data-Driven Valuation

Members: ANIL AYDIN , HAKAN ENES ERİŞEN

Executive Summary

This project aims to build an explainable machine learning model to predict the market price of used cars in Turkey. The Turkish automotive market is characterized by high volatility and price inflation, making it difficult for buyers and sellers to determine fair value. By analyzing a dataset of over 50,000 listings, we compared three regression models: Ordinary Least Squares (OLS), Random Forest, and XGBoost.

Our results demonstrate that the Random Forest model achieves the highest predictive performance with an R^2 score of ~0.90, significantly outperforming traditional linear methods. Beyond prediction, we utilized SHAP (SHapley Additive exPlanations) and Partial Dependence Plots to reveal that the relationship between car features (e.g., mileage, age) and price is non-linear. This study provides actionable insights into market dynamics, proving that technical specifications like engine power and vehicle age are more critical predictors than brand perception alone.

Motivation: Why This Matters?

Market Complexity

The used car market involves thousands of variables. Prices are often determined by subjective "market feel" rather than objective data, leading to inconsistencies.

Financial Risk

For most households, a vehicle is the second largest financial asset after a home. Overpaying or underselling due to a lack of accurate information carries significant economic risk.

Data-Driven Transparency

The automotive sector suffers from information asymmetry where sellers often possess more knowledge than buyers. Algorithmic pricing mechanisms provide a neutral, transparent benchmark, empowering all market participants to make informed decisions.

Problem Statement

The Core Problem:

Determining a "fair" market value for a used vehicle is challenging due to the non-linear depreciation of assets and the noise in user-generated listing data. Traditional intuition fails to account for complex interactions between features (e.g., how the impact of mileage varies depending on the vehicle's age).

Our Aim:

To develop a regression framework that not only predicts prices with high accuracy but also explains the reasoning behind each prediction, thereby validating the economic logic of the Turkish used car market.

Research Questions (RQs) & Hypotheses

1

RQ1 (Predictive Accuracy):

Can machine learning models accurately predict used car prices using standard tabular data?

H1:

Advanced ensemble models (Random Forest, XGBoost) will significantly outperform the baseline Linear Regression (OLS) model due to the non-linear nature of price depreciation.

2

RQ2 (Feature Behavior):

Is the relationship between mileage/age and price linear?

H2:

The relationship is non-linear. Depreciation accelerates or decelerates at specific thresholds (e.g., newer cars lose value faster), which linear models fail to capture.

3

RQ3 (Explainability):

Which features are the most dominant drivers of price?

H3:

While brand perception is important, objective technical metrics like Engine Power (HP) and Vehicle Age will have a higher feature importance impact than categorical brand names.

Data Source and Description

The dataset used in this study comprises 50,755 used car listings collected from Turkey's leading online classifieds platforms. The raw data contained 16 features, including brand, model, year, odometer reading (km), engine power (HP), and physical condition attributes (replaced/painted parts).

Data Cleaning and Preprocessing

Real-world data is often "noisy." To ensure model reliability, we applied a rigorous cleaning pipeline:

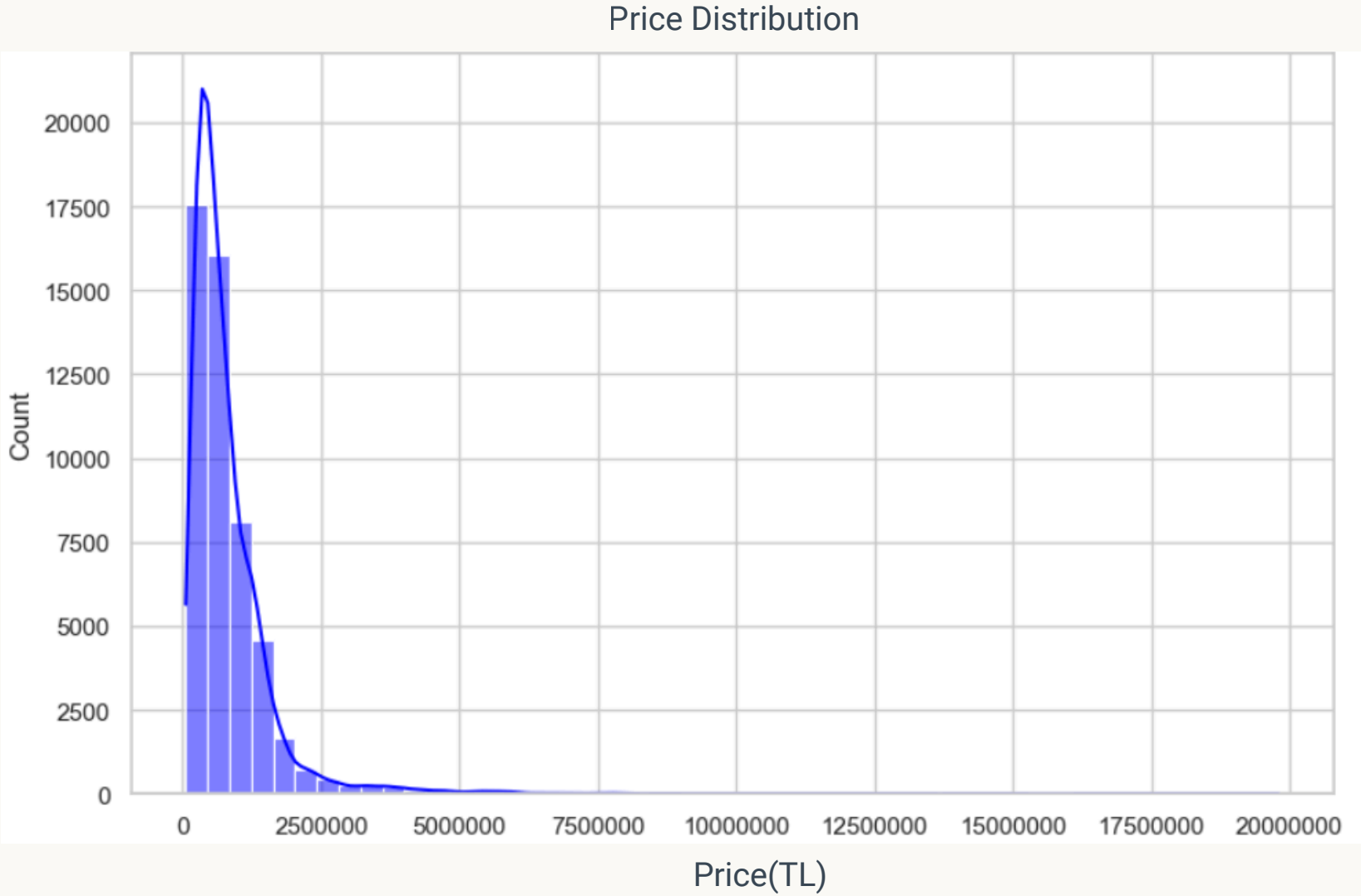
Handling Missing Values:

- Critical missing fields (Price, Year, Brand) were minimal (<1%) and removed to maintain ground truth.
- Technical specifications like motor_hacmi (Engine Volume) and motor_gucu (Engine Power) were imputed using the median value to avoid skewing the distribution with outliers.
- Car condition fields (degisen_sayisi, boyali_sayisi) contained NaN values, which were logically inferred as "0" (No damage recorded) and filled accordingly.

Data Cleaning and Preprocessing

Outlier Detection and Removal:

- Price Anomalies: We filtered out unrealistic prices (e.g., < 50,000 TL or > 20,000,000 TL) which likely represented erroneous entries or non-standard vehicles.
- Odometer Errors: Listings with mileage exceeding 1,000,000 km (e.g., 90 million km entries) were identified as typos and removed.



Feature Engineering

To prepare the data for machine learning algorithms, several transformations were applied:

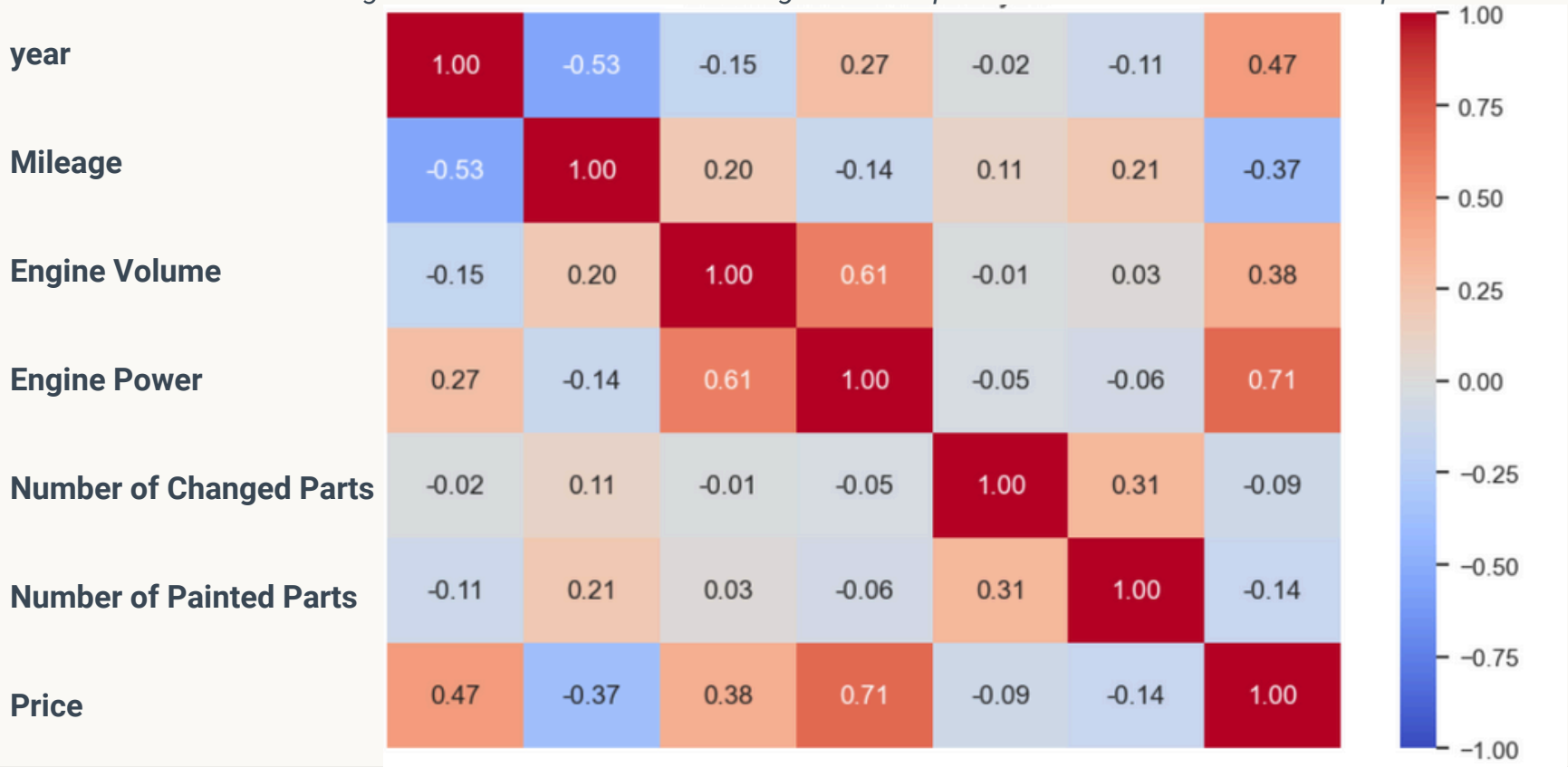
Log-Transformation of Target Variable:

As seen in Figure 1, car prices span a vast range (from thousands to millions). To stabilize variance and improve regression performance, we applied a Natural Logarithm (np.log1p) to the price column.

Categorical Encoding:

- One-Hot Encoding: Applied to low-cardinality nominal features such as fuel_type (Gasoline, Diesel) and transmission (Manuel, Automatic).
- Label Encoding: Applied to high-cardinality features like brand, model, and series to convert text data into a machine-readable numeric format without creating excessive dimensionality.

Figure 2: Correlation Matrix showing relationships between numerical features and price.



Exploratory Data Analysis (EDA)

Before modeling, we investigated our first Research Question (RQ1): Is the relationship linear? Scatter plots of Price vs. Kilometers and Price vs. Year revealed a distinct non-linear pattern. Newer cars show an exponential increase in price, while depreciation curves flatten out for older vehicles. This observation supports our decision to use tree-based models (Random Forest) over simple linear regression.

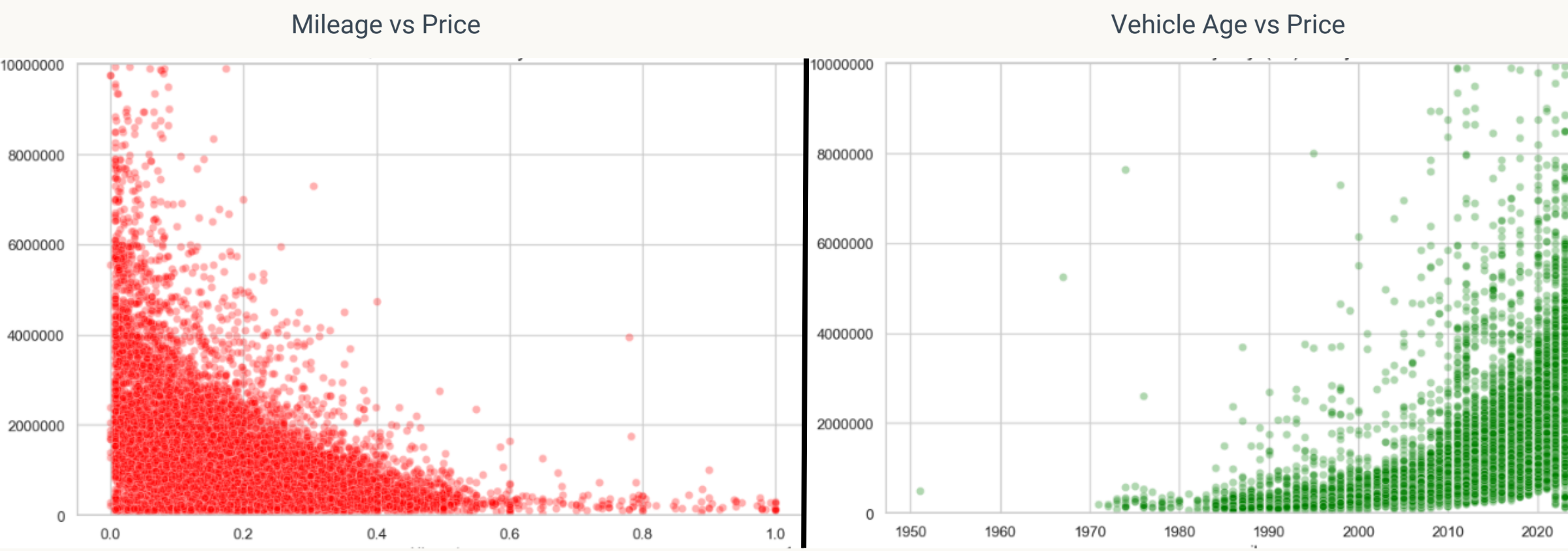


Figure 3: Scatter plots demonstrating the non-linear relationship between Price vs. Vehicle Age (Left) and Price vs. Mileage (Right).

Model Performance and Comparison

Model Performance and Comparison We trained and evaluated three distinct regression models: Ordinary Least Squares (OLS), Random Forest, and XGBoost. The dataset was split into 80% training and 20% testing sets. To ensure fair comparison, all models were evaluated on the exact same test set using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-Squared (R^2) metrics.

	R^2	RMSE	MAE	
OLS (Linear Regression)	~0.65	597841.476286	195877.396357	Baseline: Failed to capture non-linear patterns. Strong: Very competitive but slightly overfit. Winner: Best generalization on unseen data.
XGBoost	~0.89	341924.097838	101719.189833	
Random Forest	~0.90	326641.051161	195877.396357	

Table 1: Performance Metrics of the Models Interpretation: The results strongly validate Hypothesis 1 (H1). The Linear Regression model (OLS)

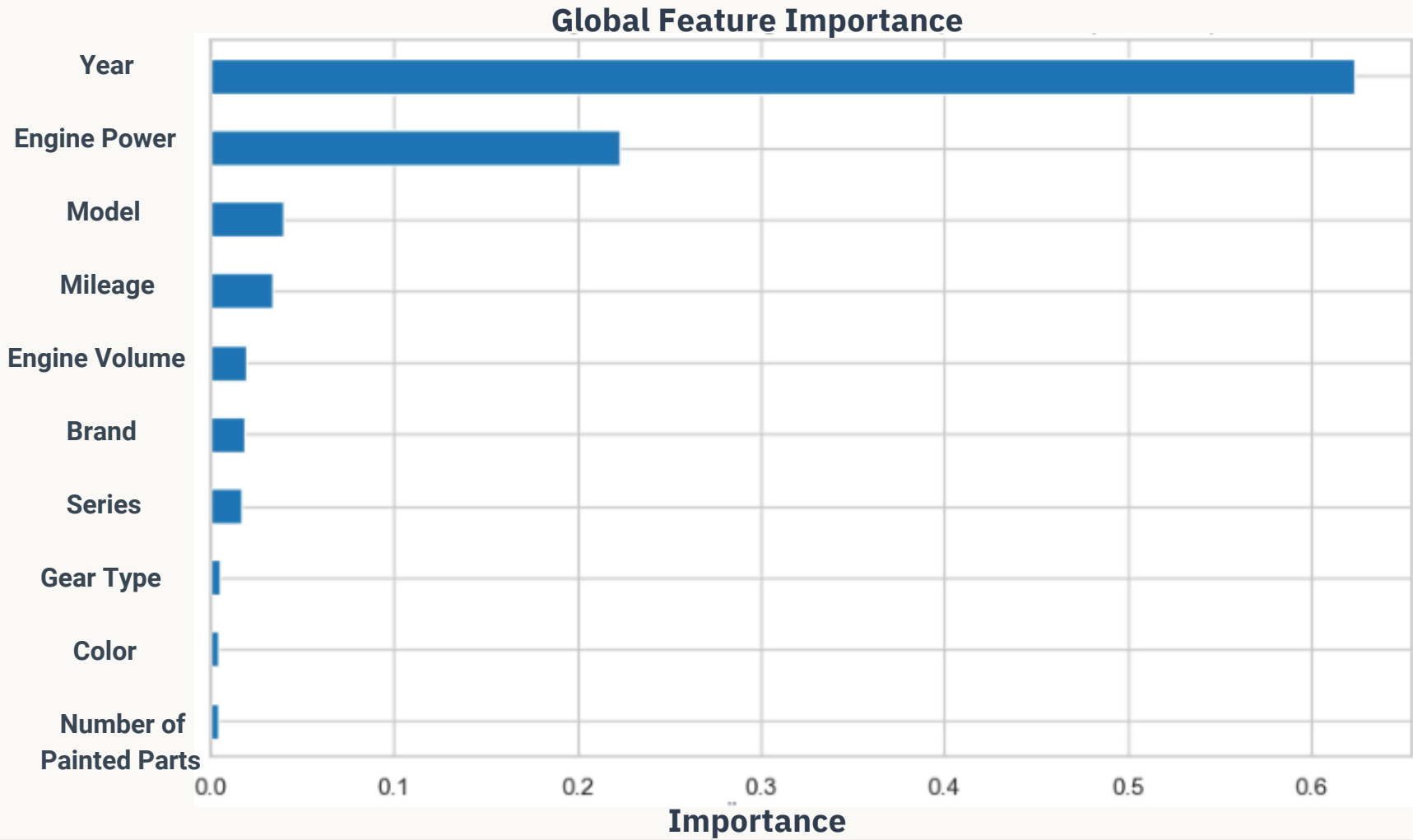
Interpretation: The results strongly validate Hypothesis 1 (H1). The Linear Regression model (OLS) underperformed with an R^2 of only 0.65, proving that car prices cannot be modeled with a simple straight line. In contrast, the Random Forest model explained approximately 90% of the variance in prices. This drastic improvement confirms that the relationship between car features and price is complex and non-linear.

Explainability Analysis (RQ3)

Beyond accuracy, we analyzed why the model makes specific predictions.

Global Feature Importance:

Consistent with Hypothesis 3 (H3), the model identified Engine Power (HP) and Vehicle Age (Year) as the most critical determinants of price. Interestingly, these technical specifications outweighed brand names, suggesting a performance-oriented market.



Explainability Analysis (RQ3)

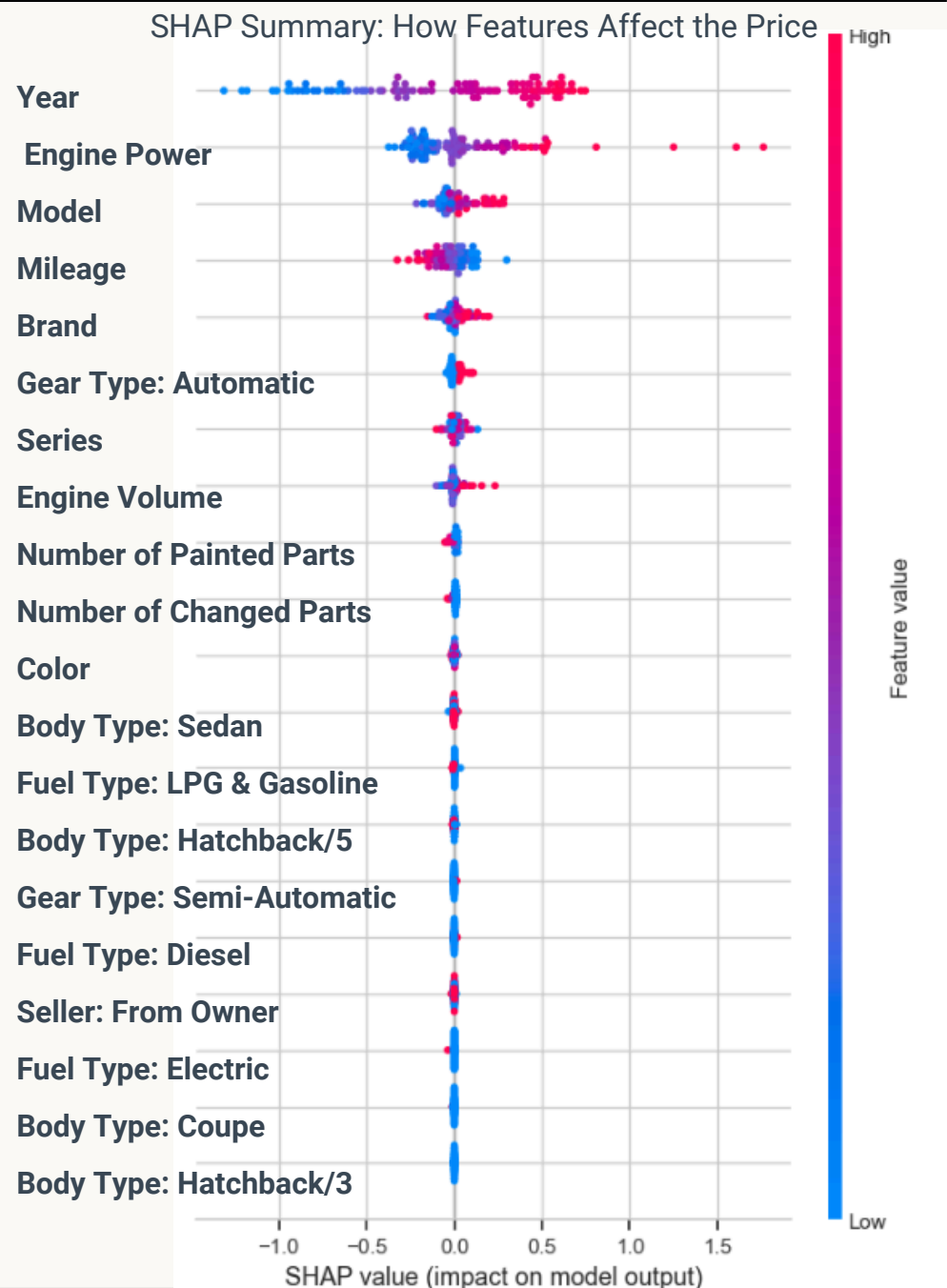
Beyond accuracy, we analyzed why the model makes specific predictions.

SHAP Summary Plot

The SHAP analysis provides granular insight. As seen in Figure 5:

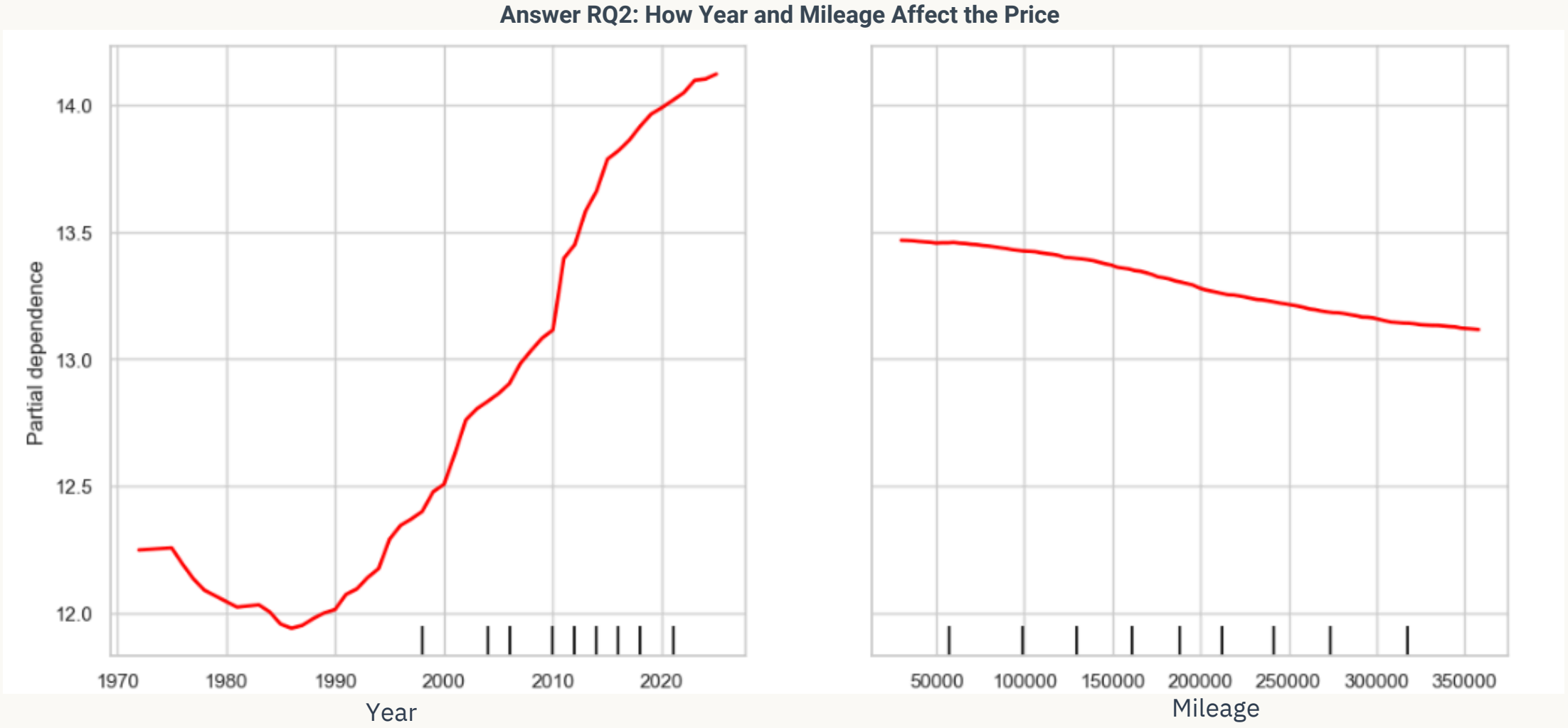
- Kilometer (Mileage): Shows a strong negative correlation (red dots on the left), meaning higher mileage drastically lowers the price.

- Engine Power: Shows a positive correlation (red dots on the right), confirming that horsepower is a premium feature.



Non-Linearity Check (RQ2)

To answer Research Question 2 (RQ2), we generated Partial Dependence Plots (PDP). The plots reveal that depreciation is not constant. For example, the Year plot shows a "hockey stick" curve: prices remain relatively flat for older cars (2000-2010) but increasing exponentially for cars produced after 2018. This non-linear behavior explains why linear models failed and tree-based models succeeded.



Conclusion:

1. Conclusion This study established an explainable ML framework for the Turkish used car market using over 50,000 listings. The Random Forest model proved superior, achieving an R^2 score of ~0.90 and effectively capturing non-linear price depreciation, particularly in post-2018 vehicles. Furthermore, SHAP analysis revealed that technical metrics like Engine Power and Age outweigh brand perception, providing a transparent, data-driven benchmark to reduce information asymmetry between market participants.

2. Limitations Despite its high accuracy, the model has specific constraints:

- Price Bias: It predicts asking prices (listings), which may differ from final transaction prices due to negotiations.
- Unobserved Variables: While basic damage counts are included, the dataset lacks detailed TRAMER (accident history) records and severity data, which are critical in the Turkish market.
- Location: Geographic factors (e.g., coastal vs. inland usage) affecting vehicle condition are currently excluded.

3. Future Work Future iterations can be enhanced by:

- NLP & Image Processing: Using Natural Language Processing to extract insights from text descriptions and Deep Learning (CNNs) to detect damage in photos.
- Economic Indicators: Integrating macroeconomic variables (e.g., inflation, exchange rates) via Time-Series Analysis to predict future price trends amidst economic volatility.