Received: 13 March 2026 | Revised: 17 August 2026 | Accepted: 21 September 2026 | Published online: 3 November 2026

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2026, Vol. 00(00) 1–15
DOI: 10.47852/bonviewJCCE62029631

BON VIEW PUBLISHING

# Electric Vehicle Price Prediction Using Real-World Online Listings: An Explainable Machine Learning Approach

**Nisrean Jaber Thalji[1,*], Aryaf Al-Adwan[2], Tariq Bishtawi[3] and Hassan Kassem[4]**

[1] *Department of Intelligent Systems, Al-Balqa Applied University, Jordan*

[2] *Department of Autonomous Systems, Al-Balqa Applied University, Jordan*

[3] *Department of Computer Science, Amman Arab University, Jordan*

[4] *Department of Engineering and Computing Technology, Arab University College of Technology, Jordan*

**Abstract:** The recent gains in the adoption of electric vehicles (EVs) have triggered the emergence of secondary markets concerning EVs, where the determination of prices accurately and transparently is an extreme challenge. In the current study, we propose an explainable machine learning model to predict the advertised price (in Jordanian dinars) of used EVs, using actual online advertisements obtained in the Jordanian market. An EV-oriented and custom-made dataset of 2,165 listings was used that specifically includes battery capacity and driving range along with the typical vehicle features. The coefficient of determination ($R^2$), mean absolute error, and root mean squared error were used to assess the performance of three tree-based regressors: CatBoost Regressor, Extreme Gradient Boosting Regressor, and Random Forest Regressor trained with a leakage-safe pipeline and evaluated using grouped, repeated cross-validation. SHapley Additive exPlanations were applied to the fitted models to quantify the global contribution of each feature to price predictions and to assess the stability of these contributions across validation folds, providing a transparent account of the price-determination mechanism. The empirical data indicate that battery capacity and driving range are the most dominant EV price determinants that continue to overshadow conventional predictors of price, including brand, mileage, and year. The soundness of such findings is supported by cross-model explainability analyses. As a result, the proposed framework enhances pricing transparency and enables informed decision-making in young EV markets.

**Keywords:** explainable machine learning, SHAP, battery capacity, Jordanian automotive market, electric vehicle price prediction

## 1. Introduction

The world is moving to greener transportation, thus ushering in a new wave of electric vehicle (EV) adoption at an unprecedented pace due to the growing environmental pressure, blistering technological advances, and increasingly friendly policy platforms. With the EV market reaching the maturity phase at the international level, a secondary market of used EVs has started to form, crystallizing a new set of challenges and opportunities associated with pricing transparency, market efficiency, and consumer decision-making processes. Unlike the traditional approach to internal combustion engine (ICE) vehicles, where the depreciation curve and valuation measurements are rather straightforward, EVs are characterized by new price determinants, most notably battery capacity and driving range, which will radically alter the model of vehicle valuation [1, 2].

The accurate prediction of the price of used EVs has a significant practical implication. For consumers, good price estimations enable a more informed buying decision and are used to counter information asymmetries, which prevail in online markets. In the case of dealership networks and sellers, a pricing strategy based on data can be used to maximize revenue and inventory turnover. For policymakers and market analysts, the information on the factors influencing the valuation of EVs can shed light on general trends of technology adoption and how the market matures. Notwithstanding these requirements, the existing scholarly data on the issue of vehicle price prediction has focused significantly on traditional car models, which have not paid much attention to the specifics of the EV secondary market [3].

Over the last few years, there has been an explosion of research activity that uses machine learning (ML) methods to estimate the price of vehicles. The literature has investigated a wide range of regression approaches, such as linear regression (LR), decision trees (DT), random forests (RFs), k-nearest neighbors, gradient-boosting (GB) variants, and artificial neural networks (ANNs)

***Corresponding author:** Nisrean Jaber Thalji, Department of Intelligent Systems, Al-Balqa Applied University, Jordan. Email: n.thalji@bau.edu.jo

© The Author(s) 2026. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

with diverse geographic settings and data reflecting different compositions [4]. However, this literature has serious deficiencies. To begin with, numerous studies consider electric and conventional automobiles as a heterogeneous group, where the sets of features and modeling techniques are the same, regardless of fuel type; this presumption presupposes a price-level neutrality, which is becoming more questionable due to the dominance of battery-related characteristics in the definition of the EV value. Second, despite the inclusion of EVs in some studies, both battery capacity and driving range are seldom used as explicit predictor variables, either due to the lack of this information within the underlying data or researchers having lower priorities with EV-specific predictor variables when developing features. Third, the interpretability of the models is widely overlooked in the literature, and the focus lies more on predictive strength than on transparency regarding the individual feature's contribution to price formation—a fact that is especially consequential in consumer-stakes situations [5].

This paper aims to address these voids by providing an explainable ML model for predicting the price of EVs, which is based on real-world listings of online items in the Jordanian automotive market. It is worth clarifying at the outset that the novelty of this work does not lie in the tree-based regression algorithms or SHapley Additive exPlanations (SHAP)-based explainability techniques themselves, both of which are well established in the literature. Rather, the contribution lies in the construction of an EV-specific analytical subset containing explicit battery-related attributes, the treatment of missingness as an informative signal, and the implementation of a leakage-safe preprocessing and evaluation framework within the context of the Jordanian EV market. The study's main contributions are summarized as follows:

1) We construct an EV-specific subset of 2,165 real-world Jordanian vehicle listings that retains battery capacity and driving range as explicit predictors, coupled with a leakage-safe, cross-validation-aware preprocessing pipeline.
2) We explicitly encode ordinal battery- and condition-related attributes and model missingness in the paint- and body-condition variables as informative indicator features rather than discarding or silently imputing them.
3) We benchmark three tree-based regressors (CatBoost Regressor (CBR), Extreme Gradient Boosting Regressor (XGBR), Random Forest Regression (RFR)) under an identical preprocessing pipeline and a grouped, repeated cross-validation protocol to ensure fair and leakage-safe performance comparisons.
4) We apply SHAP to quantify the global importance of each feature and to assess the stability of these importance rankings across cross-validation folds and across the three learning paradigms.

Empirically, the results indicate that battery-related attributes exert a stronger influence on EV prices than many conventional vehicle characteristics. This finding highlights the importance of incorporating EV-specific technical features into pricing models and demonstrates the value of explainable ML analysis in understanding price-formation mechanisms.

The rest of this paper will be structured in the following way. Section 2 is a literature review of related literature in vehicle price prediction, specifically focusing on the studies that discuss EVs and those that use explainability techniques. Section 3 specifies the source of data, creation of the EV subset, full description of attributes, analysis of missing values and distribution, and the methodological framework, including feature engineering, missing value handling through explicit missingness indicators, preprocessing pipeline development, the three tree-based regression models, and the SHAP-based explainability procedure. Section 4 provides the findings of the experiment, which include SHAP-based global feature-importance analysis and cross-model importance comparison. Section 5 discusses the implications of our results, the limitations of the study, and suggestions for future research. Section 6 concludes the paper.

## 2. Related Works

Precise forecasting of vehicle prices has been a serious field of study since time immemorial because of its importance to consumers, dealerships, and market analysts. The older models of vehicle valuation were based on econometric principles and statistical models, and they usually included the brand, model, age, mileage, and condition [4]. Nevertheless, the development of large datasets and new methods of computation has also led to the appearance of ML as a potent paradigm for improving the accuracy and strength of vehicle price prediction [6].

### 2.1. Vehicle price prediction using machine learning

There has been significant expansion in the use of ML for vehicle price prediction, with various regression algorithms employed in previous studies. Common approaches include LR, DTs, RFs, GB methods such as XGBoost, and neural-network-based models [7, 8]. These models have been applied across different market settings and have demonstrated their ability to capture complex and nonlinear relationships among vehicle characteristics and prices. Comparative studies have also shown that ensemble-based approaches, particularly tree-based boosting methods, can provide strong predictive performance for vehicle price prediction tasks [7, 8].

Although the use of these ML-based vehicle price prediction models has grown, a major part of the available literature has viewed EVs and ICE vehicles as a unit group. This usually results in the usage of the same sets of features and modeling methods, implicitly supposing that the determinants of value can be used anywhere, irrespective of powertrain type [9]. This supposition though ignores the distinct nature and fast-changing technological environment of EVs, where features like battery capability and range are comparatively heavily valued [10, 11]. As a result, most studies have either overlooked EV-specific characteristics or have not emphasized them in their modeling exercises, and thus, there may be inaccuracy and limited knowledge on the EV secondary market dynamics.

Regression models, which are based on ML, can usually produce very precise predictions of observations that were not included in the training inputs [12]. Over the last few years, many of the techniques have been developed with the goal of predicting the prices of used cars. As an example, the performance of three different ML algorithms is assessed in the work reported in Reference [13]: DT, RF, and LR. The authors conclude that the RFR model has the most accurate predictions.

In the study conducted in Reference [14], the GB approach is used to predict the value of automobiles. The authors note that the GB framework brings about a relatively low root mean squared error (RMSE) and an adequate R-squared value. In another comparative study, Khotimah et al. [15] evaluated the performance of two ML models, namely, ANN and RFR, for used car price prediction. The results showed that both models were effective, with ANN achieving a lower mean absolute error (MAE) than RFR.

### 2.2. Price determinants on electric vehicles

The development and fast growth of the EV market require special analysis of their price determinants. Contrary to ICE vehicles, the battery technology of an EV has a significant impact on its value, with some of the key aspects being battery capacity (kWh), projected driving range (km), and battery health/degradation [16, 17]. Roberson et al. [1] and Kang and Fu [2] highlighted the importance of EV-specific attributes, particularly battery capacity and driving range, in shaping consumer perceptions and resale value. EVs' depreciation curve may vary considerably compared to ICE vehicles, and it depends on the progress in battery technology, the presence of charging infrastructure sources, and changing government subsidies [18]. Thus, those models that do not explicitly consider these EV-specific characteristics may give an incomplete or false vision of the market.

In addition to resale-value studies, there is another more general line of EV research that involves battery and charging-management aspects of the same underlying technology that underlies the price determinants discussed here. There are several approaches to manage the demand EVs impose on the distribution grid, such as load-forecasting and stochastic charging-scheduling approaches [19]. Finally, resale-value models need to consider the battery health and degradation signals, which are ultimately based on accurate estimation of battery state of charge, for example, by dual extended Kalman filtering [20]. The intelligent controllers, such as Adaptive Neuro-Fuzzy Inference System (ANFIS)-based charging control [21] and other intelligent controllers for EV charging in microgrids [22], are examples of how battery capacity and range will be utilized not only at the point of sale but also at the level of grid interaction. For battery capacity to become a system-level, and not just vehicle-level, price and performance driver, energy-management frameworks that coordinate storage and Photovoltaic (PV) generation with EV charging can be used. These studies as a whole support the idea of the present work, namely, that attributes of the battery contain information beyond that already contained in traditional car attributes and therefore need to be specifically accounted for when modeling price prediction.

### 2.3. Explainable Artificial Intelligence (XAI) in EVs

Although modern ML systems can achieve impressive predictive accuracy, they have a significant limitation in the form of their relative lack of transparency, often referred to as a black-box property, when used with high stakes in consumer-focused applications like vehicle valuation [23]. The lack of transparency when it comes to internal decision pathways of a model may decrease user trust and reduce the real applicability of these advanced systems. To address this problem, a new field of study called Explainable Artificial Intelligence (XAI) has developed, which provides a collection of methods aimed at elucidating and demystifying the processes of inference of complex ML systems.

SHAP and Local Interpretable Model-Agnostic Explanations (LIME) are widely used post hoc approaches for interpreting ML predictions [24, 25]. SHAP is based on the cooperative game theory and attributes a relative credit value to each contributing input feature of a given prediction, thus distributing explainable influence fairly between features. Ribeiro et al. [5] proposed the LIME framework, which explains the predictions of any classifier or regressor to have an interpretable surrogate model that is locally fit around the case under consideration. XAI has been utilized in the automotive setup to perform a wide range of tasks, such as autonomous driving systems and vehicle diagnostics, to improve safety, reliability, or stakeholder confidence [26]. However, there is a relative lack of implementation in the context of EV price prediction. The adoption of XAI approaches will not only enable developers to reach high predictive scores but also enable them to provide clear, practical conclusions on which variables have the most considerable impact on the price of EVs, which will benefit users and players on the market in general.

This research is at the intersection of all these complementary research domains. We built a robust EV price prediction model based on state-of-the-art ML approaches, ensuring a detailed consideration of the car-specific characteristics. Most importantly, we incorporate SHAP in the modeling pipeline to ensure that we are entirely interpretable, thus providing a clear explanation of the variables that have the most significant impact on the EV valuation in the Jordanian market.

## 3. Data Source and Dataset Description

### 3.1. Data source

This study is based on a publicly available dataset hosted on Mendeley Data, which accompanies a peer-reviewed article published in the Data in Brief journal [27]. The dataset consists of real-world vehicle advertisements collected from online automotive marketplaces and reflects actual advertised prices, making it suitable for analyzing vehicle pricing behavior under realistic market conditions [28].

Each record represents a single vehicle listing and includes a comprehensive set of technical, descriptive, and usage-related attributes. These attributes cover manufacturer and model information, production year, mileage, body type, vehicle condition, transmission type, fuel type, and geographic location. In addition, the dataset contains several binary and categorical features describing insurance status, customs clearance, licensing, painting condition, and body condition. There are also EV-specific features in the dataset, which are the battery capacity and driving range [29, 30].

The initial dataset comprises 10,921 car listings characterized by 22 features, among them a target variable, which is the advertised price of the vehicle in Jordanian dinars (price_value_JOD). Of these 10,921 listings, 2,165 records were identified as EVs and retained for this study (Section 3.4); of that EV subset, 170 records lacking a valid target price were subsequently excluded prior to model training (Section 3.7.2), leaving 1,995 EV listings for modeling.

### 3.2. Dataset attributes and variable description

The dataset includes 21 explanatory variables and one target variable that will characterize vehicle identity, technical characteristics, usage, legal status, physical condition, and EV-specific properties. All the dataset attributes are presented in Table 1, where they are classified based on functional role.

### 3.3. Overview of the original dataset structure

Table 2 shows a sample of the initial dataset and shows the variety and the organization of the attributes that were recorded. This table shows that the data contains conventional, hybrid, and EV vehicles, and it will have both numerical variables, categorical variables, and textual variables along with EV-related variables. With such diversity, it is possible to undertake targeted analysis once suitable data have been selected.

**Table 1**
**Description of dataset attributes**

| Category | Attribute | Description |
|---|---|---|
| Target variable | price_value_JOD | Advertised vehicle price in Jordanian dinars |
| Vehicle identity | make | Vehicle manufacturer |
| | model | Vehicle model |
| Vehicle age and usage | year | Model production year |
| | mileage_km | Vehicle mileage in km |
| | mileage_text | Mileage expressed as categorical ranges |
| Vehicle specifications | body_type | Vehicle body type (e.g., Sedan, SUV) |
| | transmission | Transmission type |
| | engine_type | Engine configuration |
| | fuel_type | Fuel type (hlectric, hybrid, etc.) |
| Appearance and comfort | color | Exterior color |
| | interior_color | Interior color |
| | seats | Number of seats |
| Location | city | City of listing |
| Legal & status features | feat_insurance | Insurance status |
| | feat_carcustoms | Customs clearance status |
| | feat_license | Licensing status |
| Physical condition | feat_paintcondition | Paint condition |
| | feat_body_condition | Body condition |
| EV-specific features | feat_battery_capacity | Battery capacity (kWh categories) |
| | feat_battery_range | Driving range (km categories) |

**Table 2**
**Example records from the original dataset**

| price_value_JOD | make | model | year | mileage_km | mileage_text | body type | Fuel type | feat_battery capacity | feat_battery range |
|---|---|---|---|---|---|---|---|---|---|
| 11600 | Changan | Eado EV | 2022 | 74999 | 70,000–79,999 | SUV | Electric | Less than 50 kWh | 400 – 499 km |
| 22700 | Mercedes-Benz | C-Class | 2016 | 134999 | 130,000–139,999 | Sedan | Plug-in Hybrid | NaN | NaN |
| 4700 | Mazda | 2 | 2013 | 104999 | 100,000–109,999 | Hatch-back | Gasoline | NaN | NaN |
| 12700 | Hyundai | Ioniq | 2019 | 94999 | 90,000–99,999 | Sedan | Hybrid | NaN | NaN |

## 3.4. Extraction of the electric vehicle subset

As the aim of the current paper is to examine the pricing of EVs, a subsample was constructed of EVs only from the original data. The selection was done with the fuel-type attribute, where only records that were explicitly marked Electric were retained. After this filtering operation, the size of the dataset was reduced to the form of Table 3.

These findings reveal that EVs represent about 20% of the total listings, and the entire attribute framework of the information is maintained. This allows doing an EV analysis without modifying the underlying data schema.

## 3.5. Missing value analysis of the EV dataset

To inform preprocessing and modeling choices, the EV-only dataset was analyzed in terms of data completeness, type of variables, and cardinality of attributes (Table 4).

Each variable is reported in the table with the number of different values and the percentage of missing values, providing

**Table 3**
**Dataset dimensions before and after EV extraction**

| Description | Number of records | Number of attributes |
|---|---|---|
| Original dataset | 10,921 | 22 |
| EV-only dataset | 2,165 | 22 |

a quantitative basis for subsequent preprocessing and modeling decisions [31].

The target variable, price_value_JOD, contains 465 different values, with a missing rate of 7.85%. This indicates substantial diversity in the observed price values, while the presence of missing target values necessitates appropriate treatment before model training. The variable is treated as a continuous numerical target because vehicle price is a quantitative monetary measure.

Attributes of usage have varying features. The variable mileage_km takes 22 different values, and the missing rate is low (2.45), meaning that the information on the mileage has limited missingness and can be directly subjected to numerical treatment. Meanwhile, model year is used in 15 discrete values and is represented as a type of object, although it represents a discrete numerical value; this is the reason to use a type conversion explicitly and the possibility to derive vehicle age in the following steps.

Several EV-specific and condition-related variables have limited and definite cardinality.

Only five and six distinct categories (battery capacity and battery range) are included in them without missing values, which make them ordinal indicators of EV technical specifications that are reliable. Equally, the paint condition and body condition are four and five ordered categories with rather high missing rates (17.09% and 17.04); thus, imputation or category consolidation must be done carefully to avoid bias.

Nominal categorical variables have differing cardinality. The attributes of manufacturer (make) and model (model) demonstrate 59 and 154 different values, respectively, which is a highly

**Table 4**
**Summary of data completeness, variable types, and cardinality for the EV-only dataset**

| Attribute | Data type | Unique values | Missing (%) | Role/notes |
|---|---|---|---|---|
| price_value_JOD | Continuous numerical | 465 | 7.85 | Target variable (price) |
| mileage_km | Continuous numerical | 22 | 2.45 | Usage intensity |
| year | Discrete numerical (stored as object) | 15 | 0.00 | Converted to numeric/age derivation |
| seats | Discrete numerical (stored as object) | 6 | 0.00 | Seating capacity |
| mileage_text | Ordinal categorical | 23 | 0.00 | Ordered mileage ranges |
| feat_battery_capacity | Ordinal categorical | 5 | 0.00 | Battery size categories (kWh) |
| feat_battery_range | Ordinal categorical | 6 | 0.00 | Driving range categories (km) |
| feat_paintcondition | Ordinal categorical | 4 | 17.09 | Physical condition (paint) |
| feat_body_condition | Ordinal categorical | 5 | 17.04 | Physical condition (body) |
| make | Nominal categorical | 59 | 0.00 | High-cardinality brand |
| model | Nominal categorical | 154 | 0.55 | High-cardinality model |
| body_type | Nominal categorical | 7 | 0.00 | Vehicle form factor |
| color | Nominal categorical | 24 | 0.00 | Exterior color |
| interior_color | Nominal categorical | 22 | 0.00 | Interior color |
| city | Nominal categorical | 14 | 0.00 | Location |
| transmission | Binary categorical | 2 | 0.00 | Transmission type |
| condition | Binary categorical | 2 | 0.00 | New/used |
| feat_insurance | Binary categorical | 2 | 0.00 | Insurance status |
| feat_license | Binary categorical | 2 | 0.00 | Licensing status |
| feat_carcustoms | Binary categorical | 2 | 0.00 | Customs clearance |
| engine_type | Constant | 1 | 0.00 | No variance (excluded later) |
| fuel_type | Constant | 1 | 0.00 | Electric only (filtered) |

diversified brand and model environment. This cardinality is very high, and it indicates that naive encoding strategies are not the best and that such variables would have to be given special treatment so as to prevent overfitting and information leakage. By comparison, more modular attributes, like body_type (7 categories) and city (14 categories), have a moderate cardinality and can be better represented using conventional techniques of categorical encoding.

Lastly, several variables exhibit minor or no variability. The attributes fuel type and engine type each contain only one unique value in the EV-only dataset, reflecting that all vehicles are fully electric. Therefore, these variables do not provide any discriminative information and do not contribute to explaining price variation.

Overall, quantitative trends described in Table 4 indicate a heterogeneous dataset with high-variability continuous values, low-cardinality ordinal, and high-cardinality nominal variables. This multiplicity directly concerns the preprocessing plan used in the following methodology section, especially in connection with type conversion, choice of encoding, and choice of features.

## 3.6. Exploratory distribution analysis

*3.6.1. Price distribution*

The price distribution is highly skewed to the right, but most EVs offered are below the value of around 30,000 JOD, with only a few high-cost vehicles going above and well above 100,000 JOD. The long-tailed distribution emphasizes the heterogeneity of the EV market and encourages using powerful modeling strategies that can manage skewed target variables. The prices advertised in EVs vary, as seen in Figure 1, across the dataset.

*3.6.2. Technical and usage characteristics*

The distributions of main technical and usage-specific features of the EV subset such as battery capacity, driving range, model year, and mileage are presented in Figure 2.

The findings show that there is a prevalence of mid-range battery capacity (50–89 kWh) and medium- to long-driving ranges (300–499 km and over 500 km). The model year distribution is highly skewed around recent years (after 2019), whereas the distribution of the mileage is skewed right, and most of the vehicles are used with relatively low frequency.

*3.6.3. Manufacturer distribution*

It is extremely uneven as BYD, Volkswagen, Neta, and Tesla have a significant share of the listings. There is a second level of manufacturers (e.g., Kia and Hyundai) that are represented moderately, and other brands are less prevalent. This level implies that the effects of manufacturers are quite likely to be an essential factor in the price fluctuation of EVs and must be clearly identified in further modeling processes. Figure 3 shows the count of EV listings by manufacturer restricted to the 12 most represented.

## 3.7. Feature engineering

As a preliminary preprocessing step, an explicit type conversion procedure was applied to variables that were stored as object types despite representing numerical quantities. This step aimed to ensure consistency between the semantic meaning of each variable and its computational representation, while preserving the integrity of the original data.

The model year and seating capacity attributes were successfully converted from object to integer types. As shown in the type conversion report, both variables exhibited 0% missing values before and after conversion, and no new missing values were introduced during this process. This confirms that the conversion was lossless and that all original information was retained.

The variables mileage_km and price_value_JOD were already stored as numerical types prior to conversion. Their data types remained unchanged as float values, and their missing value proportions were preserved at 2.45% for mileage and 7.85% for the target variable, respectively. Importantly, the conversion verification confirmed that no additional missing values were introduced for these attributes.

Overall, this step ensured that all numerically meaningful variables are represented using appropriate numerical data types, without altering the original missing value structure of the dataset. Establishing correct data types at this stage provides a reliable foundation for subsequent preprocessing operations, including

**Figure 1**
**Distribution of advertised prices for EV**

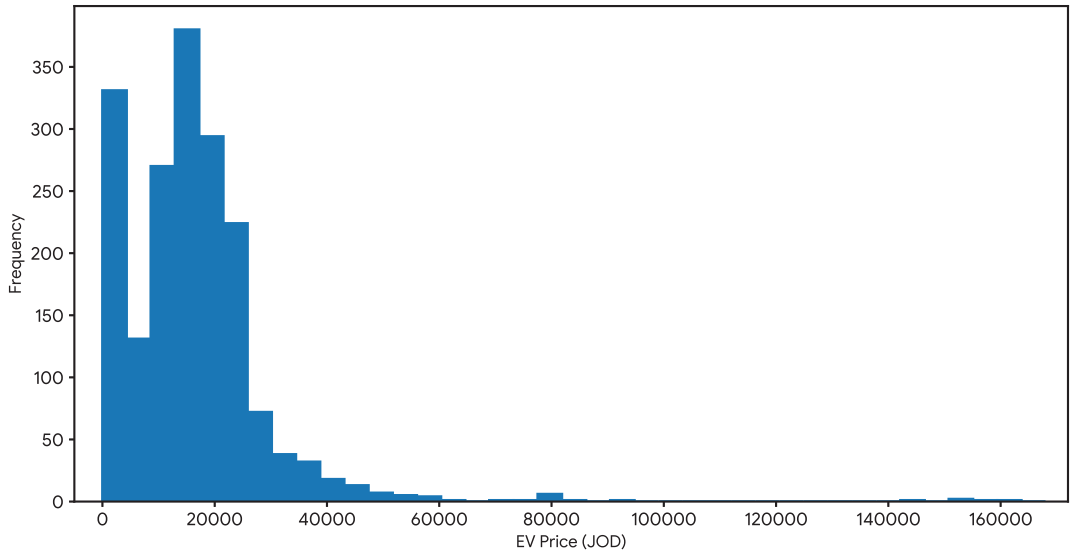


**Figure 2**
**Distributions of key technical and usage-related characteristics of the EV**

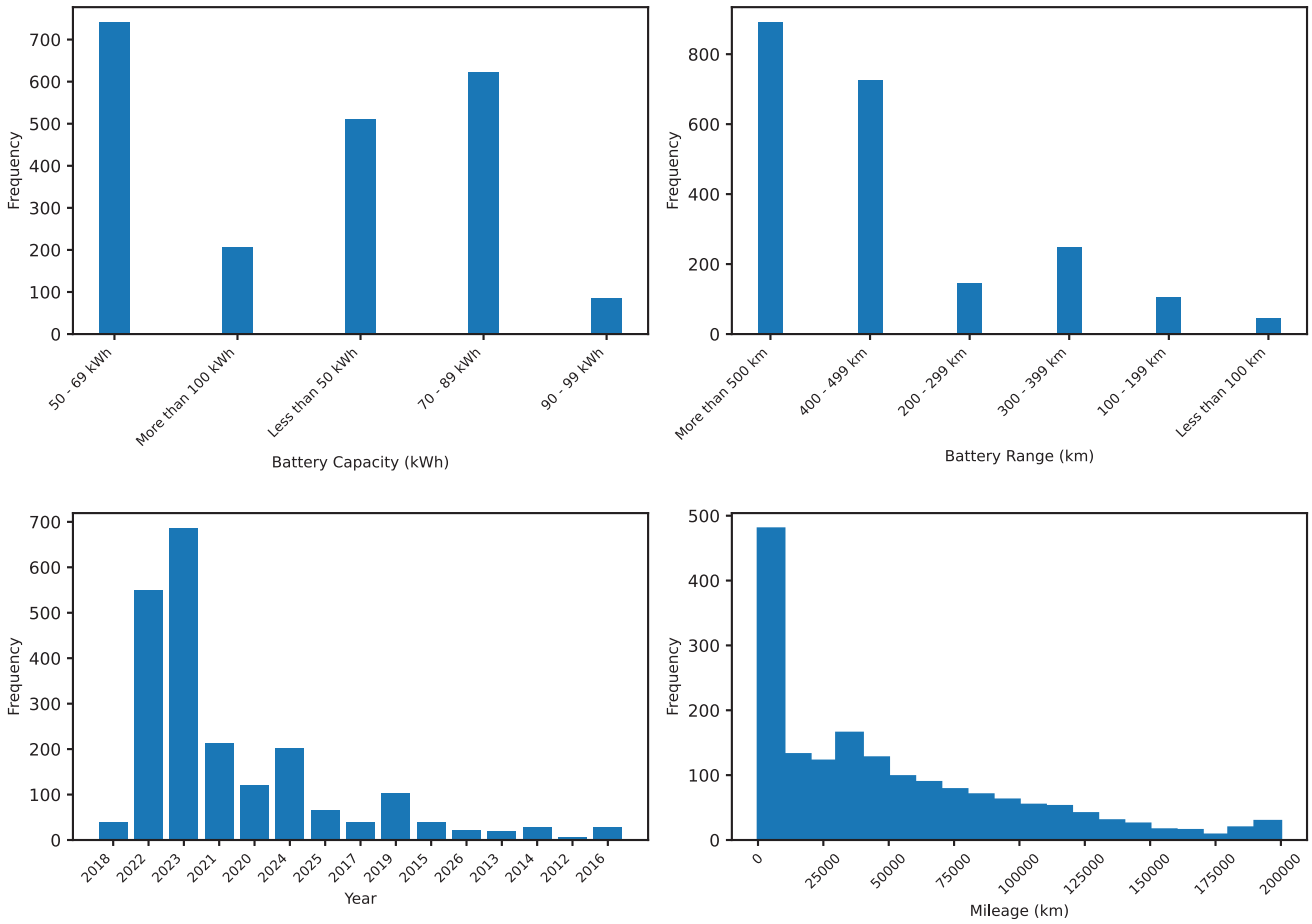
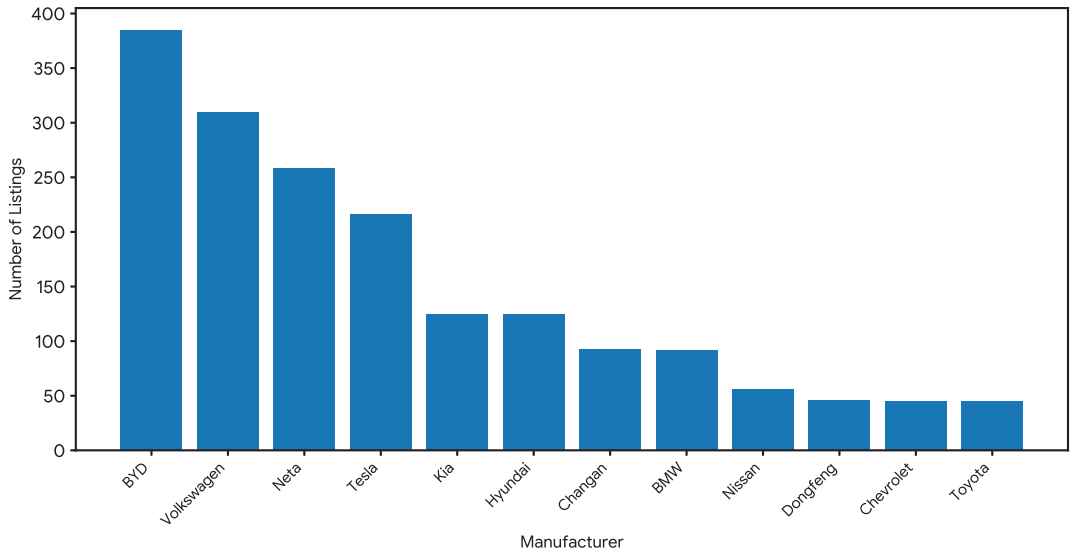



**Figure 3**
**Distribution of EV listings by manufacturer (top 12 shown)**



feature transformation, encoding, and model training. Table 5 summarizes the data type conversion results and confirms that the process did not introduce any new missing values.

*3.7.1. Coding procedure*

The only condition-related attributes that were ordinally encoded were the following: feat_paintcondition and feat_body_condition since the categories of these attributes reflect natural levels of qualification that can be interpreted as a natural sequence. These condition variables could not be represented by nominal categorical variables, so they had to be encoded using ordinal representation to maintain the inherent ranking structure and semantic sense in them. Separately from these condition attributes, the two EV-specific battery variables (feat_battery_capacity and feat_battery_range) were likewise ordinal in nature, since the source dataset already expressed them as discrete, increasing bands (Table 4 reports 5 and 6 unique bands, respectively, for battery capacity and driving range). These bands were mapped to increasing integer codes using the same explicit, dictionary-based approach described below for the condition variables, so that the natural ordering of battery size and range was preserved without assuming equal spacing between adjacent bands.

This ordinal representation was retained because battery-capacity and driving-range variables were provided in the source dataset as predefined ordered categories rather than continuous numerical measurements. The categories reflect increasing levels of EV technical capability and therefore possess an inherent ranking structure that is meaningful from both a technical and market perspective. Ordinal encoding preserves this ordering while retaining the original measurement scale of the source data. In contrast, one-hot encoding would treat adjacent categories as unrelated nominal labels and would discard the natural progression from lower-capacity and shorter-range vehicles to higher-capacity and longer-range vehicles. The fact that battery capacity and driving range subsequently emerged as leading SHAP predictors reflects their predictive importance but does not alter the underlying measurement scale on which these variables were collected.

The first step involved applying the transformation to a working copy of the EV dataset following the conversion of the data into types, which is necessary to make sure that no changes were made to the original subset of data. To achieve traceability and reproducibility, the original textual values of the two features were maintained by forming companion raw columns (feat_paintcondition_raw and feat_body_condition_raw). This design enables direct checking of the correspondence between the encoded numerical values and their initial semantic labels.

Second, the ordinal variables were categorized according to adequately interpretable levels of quality that were found in the data. To map three categories that are meaningful to an increasing quality scale, the paint condition was mapped as follows:

1) Total repaint → 1
2) Partially repainted → 2
3) Original Paint → 3

In a similar way, the body condition was coded as four ordered levels of quality:

1) Poor (severe body damages) → 1
2) Fair (body needs work) → 2
3) Good (body only has minor blemishes) → 3
4) Excellent with no defects → 4

The battery-capacity variable consisted of five naturally ordered categories ranging from less than 50 kWh to more than 100 kWh, while the driving-range variable consisted of ordered categories ranging from 200–299 km to more than 500 km. Because both variables represent increasing levels of EV technical capability rather than unordered nominal labels, they were encoded using monotonically increasing integer values that preserved their natural ordering. Lower battery-capacity and range categories were assigned smaller integer values, whereas higher-capacity and longer-range categories were assigned progressively larger values. Following this transformation, both variables were treated as ordinal predictors throughout the preprocessing pipeline and were included in the leakage-safe median-imputation procedure applied to numerical and ordinal features. This approach preserves ordering information while avoiding the dimensionality expansion that would result from one-hot encoding.

Third, explicit dictionary-based mapping was used in the implementation of the encoding, ensuring a deterministic and fully transparent transformation. This corresponds directly with the number of unique raw categories reported in Table 4: feat_paintcondition contained four raw categories and feat_body_condition contained five. In both variables, the additional category labeled "Other" was intentionally excluded from the ordinal mapping, as it does not correspond to a distinct ordinal quality level. Consequently, although Table 4 reports four and five unique raw categories, respectively, only three and four ordinal levels were retained in the final encoding.

This method does not involve using a data-driven fitting process and does not leak information, making the preprocessing pipeline fully reproducible. All categories that failed to reflect a distinct ordinal level, particularly the term Other, were purposefully omitted in the ordinal scale and transformed into missing data (NaN) to avoid the possibility of introducing arbitrary or misleading numerical correlations.

Lastly, the transformation was checked by comparison of the number of unique categories and the proportion of missing values before and after encoding. The fact that the number of separate categories is fewer clearly shows that the non-ordinal entries were carefully excluded but not lost accidentally.

The resulting ordinal features have semantically meaningful numerical representations that are well suited to further

**Table 5**
**Summarize the results of the data type conversion**

| Column | Before_Dtype | After_Dtype | Missing_%_Before | Missing_%_After | Newly_Introduced_NaNs_% |
|---|---|---|---|---|---|
| year | object | int64 | 0.00 | 0.00 | 0.0 |
| seats | object | int64 | 0.00 | 0.00 | 0.0 |
| mileage_km | float64 | float64 | 2.45 | 2.45 | 0.0 |
| price_value_JOD | float64 | float64 | 7.85 | 7.85 | 0.0 |

missing value processing and predictive modeling and yet are methodologically transparent and interpretable.

*3.7.2. Handling missing values*

After converting the types and ordinal encoding, a structured missing value handling plan was implemented to overcome unfinished entries and retain most data and its semantic meaning. The strategy adopted has a significant distinction of explicitly separating between missing values in the target variable and missing values in explanatory features since they pose essentially different methodological consequences.

Table rows that had a value of NA in the target variable (price_value_JOD) were dropped before model training because the target cannot be sensibly filled and any such attempt would create noise. This led to the elimination of 170 records, leaving the dataset of 2,165 EV listings down to 1,995. Notably, no records were dropped because of the absence of feature values and, therefore, data exclusion was strictly confined to the case when the application of supervised learning was impossible.

In the case of numerical and ordinal explanatory variables, median imputation was used to deal with the missing values, so that the evaluation would be leakage safe. These usage-related variables (mileage_km) were utilized using this strategy, along with ordinal variables based on battery specifications and the state of the vehicle. The reason why median imputation was chosen is that it is strong against skewed distribution and outliers, which are particularly common in vehicle-related data.

The nominal variables such as manufacturer, model, body type, color, interior color, city, type of transmission, insurance status, licensing status, and customs clearance were coded as a constant placeholder (Unknown). This methodology avoids the deletion of records and has a constant feature space between training and verification folds, and the model can be specifically trained to learn the impact of the absence of categorical information.

Condition-related attributes (feat_paintcondition and feat_body_condition) were treated with special weight (the reason why they are missing was also a major issue of interest), and omissions in ordinal encoding were left deliberately caused by the semantic ambiguity of certain attributes (e.g., Other). To preserve the informational value of missingness itself, explicit missingness indicator variables were added to both attributes. These binary flags are coded as to whether the original condition information was absent and were retained as distinct input features. In this design, the learning algorithm can extract the possible pricing signals directly related to the absence of condition disclosures and not interpret missing values as random noise.

The entire imputation pipeline based on distributional statistics (e.g., median values) was performed in a single preprocessing pipeline based on a column-wise transformation strategy. Consequently, all imputation parameters were inferred only based on the training data in every cross-validation fold and thus prevented information leakage and biased evaluation of the models. Overall, the missing value handling approach adopted strikes a balance between data completeness, semantic validity, and methodological rigor. The preprocessing framework allows preserving the structure of the EV market data by selectively excluding records, using strong imputation, and explicitly modeling missingness to enable predictive modeling using the data [32].

## 3.8. Preprocessing pipeline and feature matrix construction

All engineered features were combined into a single preprocessing pipeline that was constructed by a column-wise transformation plan to produce a model-ready input representation. The pipeline explicitly distinguishes numerical/ordinal, missingness indicator, and categorical variables, which should be preprocessed in a consistent manner that does not leak features, and generate features in a way that is reproducible.

Numerical and ordinal variables, such as mileage, model year, seating capacity, battery capacity, battery driving range, and the ordinal condition attributes, were reduced through a median-imputation step that provides a strong treatment of missing values. Meanwhile, the binary missingness indicators of paint and body condition were preserved with the help of the pass-through mechanism, which guarantees that the lack of condition information would be preserved as an explicit model input [33].

Categorical nominal attributes (e.g., manufacturer, model, body type, exterior/interior color, city, transmission, legal status, and insurance-related variables) were filled in with a constant category of Unknown and then transformed with one-hot encoding with unknown-category handling active. This ensures that the feature space is consistent when training and splitting evaluation and eliminates failure in the case of unknown categories. In addition, as identified in Table 4, the engine_type and fuel_type attributes were confirmed to be constant across the entire EV-only subset (a single category each, reflecting that all retained records are electric) and were therefore explicitly excluded from the feature matrix, since a constant column carries no discriminative information for price prediction. Upon running the preprocessing pipeline, the resulting feature matrix consisted of 1,995 EV listings whose features resulted in 288 encoded features/columns, capturing the expansion created by the one-hot encoding and without any loss of the entire traceability of the feature construction process. The fitted encoder was explicitly accessed to extract feature names, making them understandable and enabling future explainability analysis. This completed matrix is the standardized contribution to all the downstream models and evaluation processes.

The categorical mileage variable (mileage_text) was excluded from the final feature set because it is a coarse discretization of the continuous mileage variable (mileage_km) and therefore does not provide additional information beyond that already contained in the original numerical measurement [34]. In fact, mileage_text is directly derived from mileage_km by grouping mileage values into predefined intervals. Retaining both variables would introduce redundant information, while replacing the continuous variable with its categorical counterpart would result in a loss of measurement precision. Consequently, the continuous mileage variable (mileage_km) was retained as the more informative representation of vehicle usage intensity.

## 3.9. Regression models

To measure predictive performance and provide a fair and reproducible comparison, three ensemble-based regression models have been applied in the preprocessing pipeline. Each model was trained and tested with the same preprocessing structure to guarantee that the features are represented in a similar manner, and the performance is estimated without any bias.

1) Random Forest Regressor
   RFR is an ensemble algorithm that is a bagging technique used in the prediction of values based on multiple DTs as a means of reducing variance and enhancing the accuracy of the prediction. It is very effective in capturing nonlinear relationships and interactions between features in tabular data.

2) Extreme Gradient Boosting Regressor
   XGBR is a GBgradient boosting model through which trees are constructed one after another to reduce the prediction error. It is characterized by great predictive power and computing speed.
3) CatBoost Regressor
   CBR is a GB model that supports tabular data, and it supports categorical features by default. It employs ordered boosting to minimize overfitting and enhance the performance of generalization.
   Unless otherwise noted, each of the three models was trained using its library-default hyperparameter configuration, without additional tuning, so that the reported cross-model comparison reflects baseline learner behavior rather than the effect of a tuning budget applied unevenly across models. Table 6 summarizes the key hyperparameter values used for each model; the full configuration is available in the released code repository.

## 3.10. Model performance evaluation metrics

Three conventional predictive performance evaluation measures were employed and calculated with the grouped repeated cross-validation structure to guarantee fair and objective comparison of the predictive performance only on the held-out validation folds to avoid information leakage.

The coefficient of determination ($R^2$) is used to estimate the extent of the target variable variance that the model can account for. When the values are high, the better the predictive performance where 1.0 shows perfect prediction and the value of values that are closer to 0 or below 0 denotes poor performance.

MAE: It is the average amount of difference between the actual and predicted price. It can be directly interpreted in the form of Jordanian Dinar (JOD) and gives a clear reflection of mean prediction error.

Root mean squared error (RMSE): RMSE is used to assess the magnitude of the errors of prediction but with more weight on large errors. It can be applied to detect material prediction deviations and the general reliability of the model.

## 3.11. Explainability and stability validation using SHAP

SHAP was applied to enhance model transparency and ensure scientific validity. It was used to quantify global feature importance and assess the stability of feature contributions across validation folds.

Specifically, for each model trained in each fold of the cross-validation, a tree-based SHAP explainer was fitted, and the mean absolute SHAP value of each feature was calculated on the held-out validation set of the current fold. The mean absolute SHAP values were subsequently averaged per fold and repeated to define one global importance value per feature per model, and fold-to-fold values were also saved to characterize the importance feature's contribution stability. This approach to aggregation guarantees that the feature-importance rankings reported will account for the model's performance over the entire train/validation cross-validation scheme and not over just one split, which might not be representative.

This approach enables comprehensive interpretation of overall model behavior while ensuring that the derived explanations remain consistent, reliable, and robust across different validation splits.

# 4. Result Analysis and Discussion

## 4.1. Model performance evaluation

$R^2$, MAE, and RMSE were used to assess the predictive performance in a five-fold grouped cross-validation that

**Table 6**
**Hyperparameter configuration of the three regression models**

| Model | Hyperparameter | Value |
|---|---|---|
| RFR (RandomForestRegressor) | n_estimators | 100 |
| | max_depth | None (nodes expanded until leaves are pure or min_samples_split reached) |
| | min_samples_split/min_samples_leaf | 2 / 1 |
| | max_features | 1.0 (all features considered at each split) |
| | random_state | 42 |
| XGBR (XGBRegressor) | n_estimators | 100 |
| | max_depth | 6 |
| | learning_rate | 0.3 |
| | subsample/colsample_bytree | 1.0 / 1.0 |
| | objective | reg:squarederror |
| | random_state | 42 |
| CBR (CatBoostRegressor) | iterations | 1000 |
| | depth | 6 |
| | learning_rate | 0.03 (auto) |
| | l2_leaf_reg | 3 |
| | loss_function | RMSE |
| | random_state | 42 |

was repeated three times to guarantee reliable estimation that is leakage-safe. Grouped cross-validation was implemented to mitigate the risk of information leakage arising from duplicate or highly similar vehicle listings. Rather than defining groups using a single attribute, a composite vehicle-identity key was constructed from multiple stable vehicle descriptors, including manufacturer, model, production year, transmission type, body type, engine type, battery-capacity category, driving-range category, city, exterior color, interior color, seating capacity, and vehicle condition. In addition, vehicle mileage was discretized into 100-km intervals and incorporated into the grouping key, ensuring that vehicles with nearly identical mileage values were treated as belonging to the same grouping structure. Listings sharing the same composite vehicle profile were assigned to the same group throughout the validation process, thereby preventing duplicate or near-duplicate observations from appearing simultaneously in both the training and validation folds. This grouping strategy provided a more rigorous and realistic assessment of model generalization performance while supporting the leakage-safe evaluation framework adopted in this study. Better performance was observed with CBR, as exhibited in Table 7, because it has the highest $R^2$ (0.8386) and the lowest MAE (2550.36 JOD) and RMSE (4460.62 JOD). XGBR performed slightly better than RFR, but both showed lower accuracy than CBR and higher variability.

In general, GB models, especially CBR and XGBR, were the most accurate and stable models, which supports their effectiveness to their ability to model the data of used car prices.

## 4.2. Global feature-importance analysis (SHAP – CBR model)

The analysis was performed to find the strongest determinants of the price of EVs by using SHAP. The fully preprocessed EV dataset was analyzed, and global importance was determined as the average absolute SHAP value of all observations. Figure 4 demonstrates the 15 most influential characteristics.

The findings show that EV-focused technical qualities prevail in the price-formation process. Battery capacity is the most significant feature, and its mean absolute SHAP value is 1819.04, and then there is battery driving range that is 1251.43. The two variables have a significant role to play in price prediction than any other attribute, which illustrates that EV pricing in the Jordanian market is highly influenced by the properties of energy storage. This observation is consistent with the intuitive economic theory in that battery capacity and range directly show the usability of the vehicle, frequency of charging batteries, and expenses over long periods of ownership.

The third significant characteristic that arises is the model year (650.97), and it means that the vehicle age is a significant factor in the value of EVs. Newer EVs are more expensive because of better battery technology, longer range, and lower chances of degradation. Equally, seating capacity (507.33) and mileage (455.93) are significant contributors to price variability, which reflects car size and intensity of use in the two cases, respectively. There is also an observable effect of geographic location. The indicator city Amman (506.07) is ranked in the list of the best features, implying the existence of systematic price differences between the capital listing and the rest of the areas. This must indicate increased demand, purchasing power, and liquidity in the market of Amman relative to the rest of the cities.

The effects of manufacturers are evident, but the secondary ones are compared to technical EV features. Neta (407.04), Tesla (362.96), Mercedes-Benz (333.27), BYD (247.82), and BMW (230.33) are the brands that feature in the top 15 features, meaning that the perception of the brand and brand positioning affect the prices of EVs, albeit at a reduced magnitude compared to battery-related features.

Price differentiation is also caused by vehicle body type. Hatchback (292.25) and SUV (191.03) indicators represent the structural and functional dissimilarity with influence on consumer inclinations and pricing. Moreover, the existence of interior color (Black) (238.95) implies that some esthetic elements probably have a slight, albeit noticeable, impact on advert prices.

Lastly, the fact that the paint missing flag (182.70) is among the leading predictors is also confirmed by the fact that information about this missing feature has its own predictive power.

Listings that do not have information on the condition of the paint appear to adopt systematic pricing behavior, providing a rationale for the previous decision to explicitly modeling missingness as opposed to either discarding these records or filling in such imputations quietly. Overall, the SHAP analysis confirms the dominant role of battery-related attributes while highlighting the secondary contributions of age, usage, manufacturer, and location factors.

## 4.3. Global SHAP-based cross-model feature-importance comparison

A cross-paradigm SHAP-based comparison was performed between three regression models—CBR, XGBR, and RFR—to evaluate the strength of feature importance when using three learning paradigms. Mean absolute SHAP values were calculated according to each model and then normalized, which allowed the comparison of models fairly.

The normalized SHAP scores of the most significant features selected in the three models are given in Figure 5. A definite and logical trend can be identified, as battery capacity and battery driving range topped the importance lists in all models. Both features have the maximum normalized SHAP values, which means that the characteristics of batteries are the key determinants of the price of EVs in the Jordanian market.

The other secondary factors, including model year, vehicle mileage and seating capacity, also have a significant impact on

**Table 7**
**Overview of model performance**

| | $R^2$ | | MAE | | RMSE | |
|---|---|---|---|---|---|---|
| Model | Mean | Std | Mean | Std | Mean | Std |
| CBR | 0.8386 | 0.02 | 2550.36 | 121.72 | 4460.62 | 393.81 |
| RFR | 0.8190 | 0.04 | 2698.61 | 162.71 | 4884.87 | 528.42 |
| XGBR | 0.8290 | 0.03 | 2588.43 | 131.41 | 4687.34 | 425.04 |

**Figure 4**
**CBR feature importance (SHAP, top 15)**

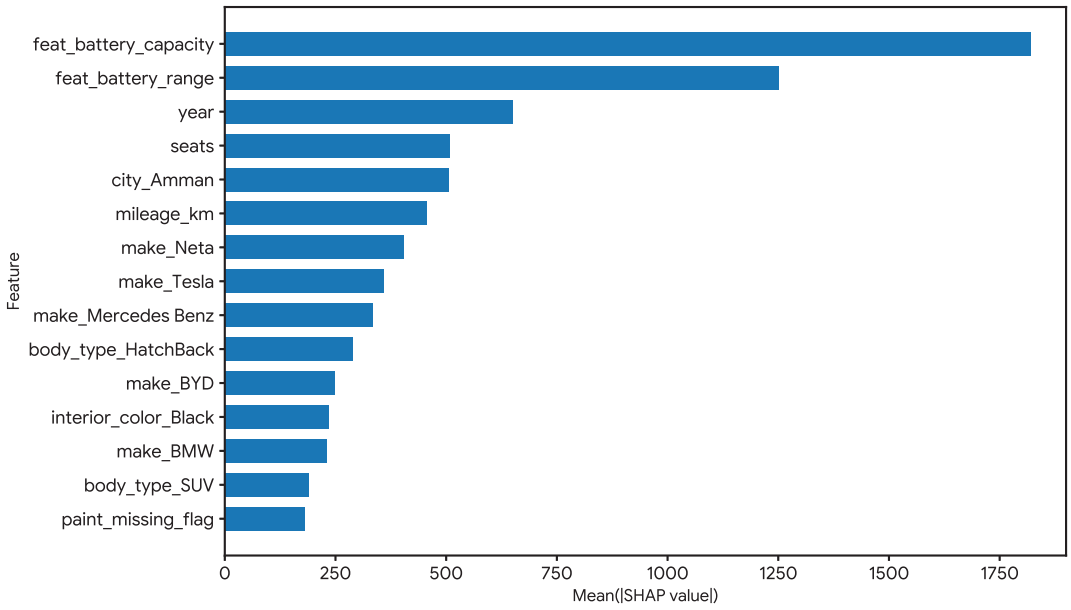
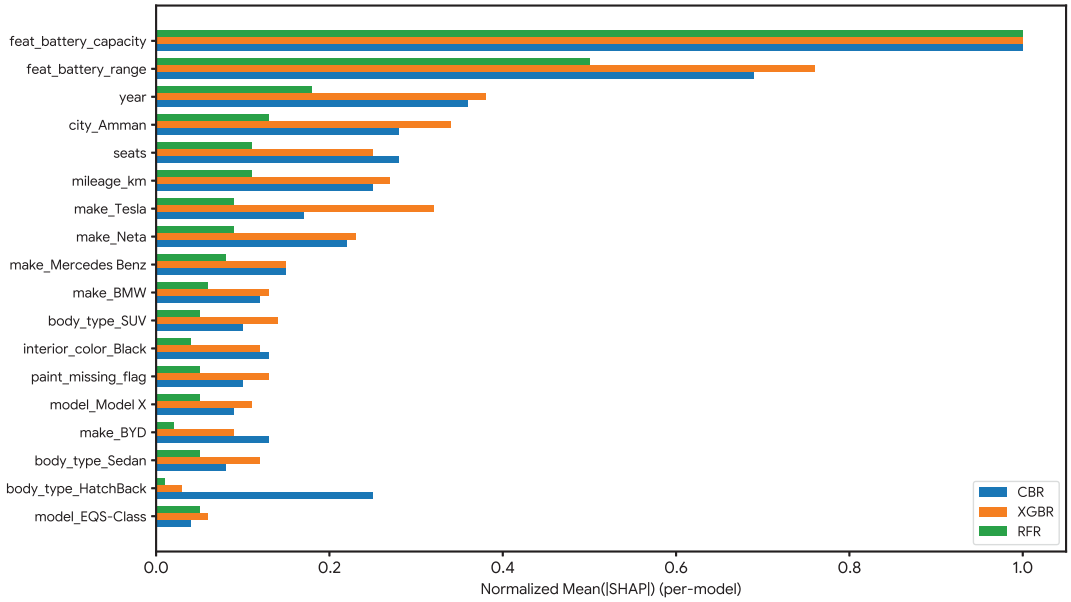



**Figure 5**
**SHAP global importance comparison (top union features, normalized)**




all the models by means of the impact of both technological depreciation and the intensity of usage.

Although there is a small variation in the relative position of their features across models, this indicates high cross-model agreement as their features appear in the rank of the top-ranked features. Variables with moderately high importance and some distinguishable differences across the models are manufacturer- and location-related variables (e.g., brand indicators (e.g., Tesla, Neta, Mercedes-Benz)) and the location of listing. This is an indication that brand positioning and regional market influences pricing although less steadily than the technical specifications.

Lastly, attributes of appearance and condition, including body type, interior color, and markers of missing condition, show smaller yet significant amounts. It is interesting to note that missingness flags were also included in the list of influential features, which implies that the lack of condition information in itself has price-related signals. Overall, the high degree of agreement across models supports the robustness of the reported feature-importance patterns and strengthens confidence in the resulting explanations.

## 5. Discussion

The empirical results of the research provide a detailed explanation of the price-formation processes in the young secondary market of EVs in Jordan. Using an ML framework that can be explained, we move beyond predictive accuracy to discover the logic behind the valuation of EVs. That part of the work

interprets these findings within the framework of existing literature, explains their wider implications to the stakeholders, and outlines research directions that can be taken in future studies.

## 5.1. Interpretation of results and hypotheses

The major finding of this study is the great prevalence of the battery-related features, namely, battery capacity and driving range, as the most significant predictors of EV prices. This result soundly supports our initial hypothesis that the valuation of EVs differs fundamentally from that of internal combustion engine (ICE) vehicles. Although the traditional automotive literature, for example, Chen et al. [3] and Noor and Jan [4], focuses on brand identity, mileage, and vehicle age as the key factors that define the resale value, our SHAP analysis demonstrates that the shift in the EV industry is a paradigm shift.

Battery-related attributes consistently emerged as the strongest predictors of EV price, exceeding the influence of conventional vehicle characteristics such as age, mileage, and brand. This finding suggests that technical capability and energy-storage performance play a central role in value formation within emerging EV markets.

## 5.2. Jordan implications on the EV market

The findings have far-reaching consequences for various stakeholders of the Jordanian automotive ecosystem. For the consumers, the information asymmetry, which is usually common in online listing platforms, can be checked by the transparency provided by our explainable models. Understanding battery capacity as the most important value-creating factor, buyers will be able to make more conscious decisions, focusing their inspections on battery health as opposed to either the esthetic condition or brand name.

In the case of dealerships and online marketplaces, the addition of EV-specific functions to pricing algorithms is increasingly important rather than merely optional. As we show, models that fail to consider battery specifications are likely to produce inaccurate valuations and hence create market inefficiencies. Further, the fact that the predictive sign of missingness is identified (e.g., the paint_missing_flag) suggests that the listing transparency is itself a value attribute; the seller who does not provide important condition information may be automatically punished by the market pricing mechanism.

## 5.3. Constraints and future research implications

Although the salient findings stand out, there are limitations associated with this study that open fruitful paths for inquiry in the future. To start with, although battery capacity and range were reported, the dataset did not contain actual measurements of State of Health or battery degradation percentages, which are usually not publicized in online listings. Future studies must strive to incorporate real-time diagnostics information from vehicle onboard systems to hone the price forecasts.

Second, the present study deals with a snapshot of the Jordanian market. Longitudinal studies will be necessary to track the changes in the depreciation curves of EVs over time as battery technology reaches new plateaus. Furthermore, the inclusion of external economic variables, including the change in electricity prices, government subsidies, and prices of gasoline, would be a more comprehensive approach to the determinants of EV resale value.

Another promising direction for future research is the incorporation of free-text vehicle descriptions that frequently accompany online listings. Such descriptions may contain valuable information regarding battery condition, maintenance history, charging characteristics, accident history, optional equipment, and other factors that are not captured by the structured attributes available in the current dataset. Natural language processing techniques could be used to extract these signals and transform them into structured predictive features. Furthermore, semi-supervised learning approaches may provide an effective mechanism for integrating information derived from textual descriptions and external battery-related datasets when detailed battery-health measurements are unavailable.

One important limitation of this study is its focus on the Jordanian market, which may have unique economic, cultural, and regulatory characteristics. Consequently, the findings may not be directly generalizable to other markets. Future studies should examine the proposed relationships in different countries and contexts to assess the robustness and external validity of the findings and to determine whether similar patterns emerge across diverse market environments. Furthermore, the target variable (advertised price) exhibited a right-skewed distribution that is commonly observed in used-vehicle markets, with a relatively small number of high-priced listings. Consequently, RMSE is more sensitive than MAE to prediction errors associated with these higher-priced vehicles, as larger errors receive greater weight in its calculation. No logarithmic transformation of the target variable was applied in this study in order to preserve the direct interpretability of predictions in Jordanian dinars. Therefore, the reported RMSE values should be interpreted in light of the underlying right-skewed distribution, particularly when comparing model performance.

Another promising direction for future research is the explicit quantification of prediction uncertainty. While the present study focuses on point predictions, methods such as Quantile Regression Forests and Conformal Prediction could be incorporated to generate prediction intervals alongside predicted prices. Such interval-based estimates would provide buyers, sellers, dealerships, and market analysts with a clearer understanding of the uncertainty associated with EV valuations and could further enhance the practical usefulness and transparency of the proposed framework.

Lastly, though the performance of tree-based models was high, future research might investigate utilizing deep-learning architecture, including Tabular Transformers, with other more advanced XAI methods like Integrated Gradients could be used to increase the level of explanation granularity. The geographic scope, which should be extended to cover a comparative study against many different Middle Eastern markets, would produce useful data regarding the variations in EV adoption and valuation on a regional level.

# 6. Conclusion

In this study, we have managed to build and strictly test an interpretable ML model, which is supposed to be used in predicting the prices of EVs, with the use of real-world online advertisements for the Jordanian market. We find strong emphasis on the fact that EV-specific technical features, in particular, battery capacity and driving range are the key price determinants, significantly outperforming conventional price determinants, which are brand, mileage, and model year. This is a paradigmatic change in valuation models, highlighting the technological

peculiarities of EVs and the need to systematically consider these variables in predictive models to accurately reflect the market. The fact that these metrics of battery-related performance have been consistently demonstrated across CBR, XGBR, and the RFR models bears testimony to the quality of our findings and provides a clear understanding of the factors specific to EV value in the context of secondary markets that are only developing.

Furthermore, the introduction of SHAP has provided not only high predictive fidelity but has also shed light on the inner mechanics of the model and thus alleviated the ubiquitous black-box problem so common with high-tech ML algorithms. The resulting open-minded, interpretive framework can provide actionable insights to a spectrum of interested parties, can give consumers finer insights into the matter, can help dealerships to make pricing strategies more precise, and can supply policymakers with valuable intelligence on what drives EV adoption and what infrastructural consideration is necessary. By proving the efficiency of XAI in the high-stakes consumer setting, this article builds confidence in AI-based solutions and lays the foundation for more transparent and efficient secondary EV markets on a global scale.

## Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

## Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

## Data Availability Statement

The data that support the findings of this study are openly available in Mendeley Data at https://doi.org/10.17632/ddcz486x5t.2.

## Author Contribution Statement

**Nisrean Jaber Thalji:** Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Aryaf Al-Adwan:** Methodology, Validation, Investigation, Writing – review & editing. **Tariq Bishtawi:** Software, Validation, Writing – review & editing. **Hassan Kassem:** Software, Validation, Writing – review & editing.

## References

[1] Roberson, L., Pantha, S., & Helveston, J. P. (2024). Battery-powered bargains? Assessing electric vehicle resale value in the United States. *Environmental Research Letters*, *19*(5), 054053. https://doi.org/10.1088/1748-9326/ad3fce

[2] Kang, W., & Fu, X. (2026). The pricing of used electric vehicles via machine learning and hedonic price model: A comparative study. *Computers & Industrial Engineering*, *216*, 111964. https://doi.org/10.1016/j.cie.2026.111964

[3] Chen, S., Wang, Z., Zhang, N., Huang, A., & Zou, C. (2026). Comparative study of machine learning models for price prediction. *Advances in Economics, Management and Political Sciences*, *262*, 120–132. https://doi.org/10.54254/2754-1169/2026.BL32880

[4] Noor, K., & Jan, S. (2017). Vehicle price prediction system using machine learning techniques. *International Journal of Computer Applications*, *167*(9), 27–31. https://doi.org/10.5120/ijca2017914373

[5] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

[6] Bergmann, S., & Feuerriegel, S. (2025). Machine learning for predicting used car resale prices using granular vehicle equipment information. *Expert Systems with Applications*, *263*, 125640. https://doi.org/10.1016/j.eswa.2024.125640

[7] Nandan, M., & Ghosh, D. (2023). Pre-owned car price prediction by employing machine learning techniques. *Journal of Decision Analytics and Intelligent Computing*, *3*(1), 167–184. https://doi.org/10.31181/jdaic10008102023n

[8] Cui, B., Ye, Z., Zhao, H., Renqing, Z., Meng, L., & Yang, Y. (2022). Used car price prediction based on the iterative framework of XGBoost+LightGBM. *Electronics*, *11*(18), 2932. https://doi.org/10.3390/electronics11182932

[9] Kumar, A. (2024). A comprehensive review of an electric vehicle based on the existing technologies and challenges. *Energy Storage*, *6*(5), e70000. https://doi.org/10.1002/est2.70000

[10] Sheldon, T. L., & Dua, R. (2024). Are automakers overcharging consumers for electric vehicle batteries? *Energy Strategy Reviews*, *55*, 101538. https://doi.org/10.1016/j.esr.2024.101538

[11] Cole, C., Droste, M., Knittel, C., Li, S., & Stock, J. H. (2023). Policies for electrifying the light-duty vehicle fleet in the United States. *AEA Papers and Proceedings*, *113*, 316–322. https://doi.org/10.1257/pandp.20231063

[12] Swati, Vijarania., Agarwal, M., A, & Le, D.-N. (2024). Machine learning approach for predicting the price of used cars. In S. Dalal, N. Dahiya, V. Jaglan, D. Koundal, & D.-N. Le (Eds.), *Reshaping intelligent business and industry* (pp. 107–115). Scrivener Publishing. https://doi.org/10.1002/9781119905202.ch7

[13] Li, C. (2024). Machine learning-based models for accurate car prices prediction. *Highlights in Business, Economics and Management*, *40*, 416–421. https://doi.org/10.54097/9zcpv779

[14] Hankar, M., Birjali, M., & Beni-Hssane, A. (2022). Used car price prediction using machine learning: A case study. In *2022 11th International Symposium on Signal, Image, Video and Communications* (pp. 1–4). https://doi.org/10.1109/ISIVC54825.2022.9800719

[15] Khotimah, E. K., Swasono, D. I., & Fajarianto, G. W. (2025). Forecasting used car prices using machine learning. *IT Journal Research and Development*, *9*(2), 123–139. https://doi.org/10.25299/itjrd.2025.18031

[16] Etxandi-Santolaya, M., Mora-Pous, A., Canals Casals, L., Corchero, C., & Eichman, J. (2024). Quantifying the impact of battery degradation in electric vehicle driving through key performance indicators. *Batteries*, *10*(3), 103. https://doi.org/10.3390/batteries10030103

[17] Andreassen, G. L., & Lind, J. T. (2024). Climate, technology and value: Insights from the first decade with mass-consumption of electric vehicles. *Environmental and Resource Economics*, *87*(7), 1783–1844. https://doi.org/10.1007/s10640-024-00872-z

[18] Olabi, A. G., & Abdelkareem, M. A. (2022). Renewable energy and climate change. *Renewable and Sustainable*

*Energy Reviews*, *158*, 112111. https://doi.org/10.1016/j.rser.2022.112111

[19] Irani, A., Alefy, Jalili., B, & Babazadeh, V. (2025). Optimization strategies for plugged-in hybrid EVs: Addressing challenges, load forecasting, and smart charging plans. *Advances in Engineering and Intelligence Systems*, *4*(1), 1–20. https://doi.org/10.22034/aeis.2025.498063.1270

[20] Christian, N., & Ling, L. (2023). A dual extended Kalman filter for the state of charge estimation of lithium-ion batteries. *Advances in Engineering and Intelligence Systems*, *2*(4), 1–10. https://doi.org/10.22034/aeis.2023.412040.1123

[21] Mahdavi, Z., Samavat, T., Javanmardi, A. S. J., Dashtaki, M. A., Zand, M., Nasab, M. A., . . . , & Khan, B. (2024). Providing a control system for charging electric vehicles using ANFIS. *International Transactions on Electrical Energy Systems, 2024*(1), 9921062. https://doi.org/10.1155/2024/9921062

[22] Rastgoo, S., Mahdavi, Z., Azimi Nasab, M., Zand, M., & Padmanaban, S. (2022). Using an intelligent control method for electric vehicle charging in microgrids. *World Electric Vehicle Journal*, *13*(12), 222. https://doi.org/10.3390/wevj13120222

[23] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[24] Li, M., Sun, H., Huang, Y., & Chen, H. (2024). Shapley value: From cooperative game to explainable artificial intelligence. *Autonomous Intelligent Systems*, *4*(1), 2. https://doi.org/10.1007/s43684-023-00060-8

[25] Kuznietsov, A., Gyevnar, B., Wang, C., Peters, S., & Albrecht, S. V. (2024). Explainable AI for safe and trustworthy autonomous driving: A systematic review. *IEEE Transactions on Intelligent Transportation Systems*, *25*(12), 19342–19364. https://doi.org/10.1109/TITS.2024.3474469

[26] Tahir, H. A., Alayed, W., Hassan, W. U., & Haider, A. (2024). A novel hybrid XAI solution for autonomous vehicles: Real-time interpretability through LIME–SHAP integration. *Sensors*, *24*(21), 6776. https://doi.org/10.3390/s24216776

[27] Thalji, N. J., & Tahat, K. (2026). JUCars-2024: A dataset of used car listings in Jordan. *Data in Brief*, *66*, 112682. https://doi.org/10.1016/j.dib.2026.112682

[28] Phu, B. Q., Phuc, P. H., & Son, P. T. (2026). Car price analysis using data collected from an online sales platform. *International Journal of Information Technology and Computer Science*, *18*(1), 124–139. https://doi.org/10.5815/ijitcs.2026.01.07

[29] Amin, A., Amin, M. S., Park, H., & Lee, D. (2025). Electric vehicle range prediction models: A systematic review of machine learning, mathematical, and simulation approaches. *World Electric Vehicle Journal*, *16*(11), 607. https://doi.org/10.3390/wevj16110607

[30] Jodlbauer, H., Tripathi, S., Bachmann, N., & Brunner, M. (2024). Unlocking hidden market segments: A data-driven approach exemplified by the electric vehicle market. *Expert Systems with Applications*, *254*, 124331. https://doi.org/10.1016/j.eswa.2024.124331

[31] Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*(1), 140. https://doi.org/10.1186/s40537-021-00516-9

[32] van den Broeck, J., Cunningham, S. A., Eeckels, R., & Herbst, K. (2005). Data cleaning: Detecting, diagnosing, and editing data abnormalities. *PLoS Medicine*, *2*(10), e267. https://doi.org/10.1371/journal.pmed.0020267

[33] Verdonck, T., Baesens, B., Óskarsdóttir, M., & vanden Broucke, S. (2024). Special issue on feature engineering editorial. *Machine Learning*, *113*(7), 3917–3928. https://doi.org/10.1007/s10994-021-06042-2

[34] Luengo, J., García-Gil, D., Ramírez-Gallego, S., García, S., & Herrera, F. (2020). *Big data preprocessing: Enabling smart data*. Springer. https://doi.org/10.1007/978-3-030-39105-8

**How to Cite:** Thalji, N. J., Al-Adwan, A., Bishtawi, T., & Kassem, H. (2026). Electric Vehicle Price Prediction Using Real-World Online Listings: An Explainable Machine Learning Approach. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE62029631