Car Price Prediction In Jordan Market "OpenSooq Data"

Submitted By

Mones Nazih Ksasbeh

Introduction

Developing accurate and efficient car price prediction models for the Jordanian market using machine learning and data mining techniques can offer significant value to buyers, sellers, and financial institutions. The car market in Jordan is highly dynamic, with prices influenced by various factors such as make, model, year, and condition. However, accurately predicting car prices is challenging due to the lack of comprehensive market data and the variability in pricing influenced by both local and global economic factors. An analytical approach to car price prediction can provide valuable insights, helping to establish fair market values and aiding in more informed decision-making for stakeholders across the market.

Data

The dataset we'll be working with is real-world data collected using web scraping techniques from OpenSooq, the most popular platform for buying and selling cars in Jordan. This dataset reflects actual listings from the site, making it highly relevant for understanding the current car market in Jordan. By leveraging this data, we aim to develop a robust car price prediction model that can accurately reflect market trends and pricing factors.

The dataset contains 49,541 data points, each representing an individual car listing. It includes 12 attributes or features, such as the car's make, model, year, condition, and price, among others. These features will be used to analyze and predict car prices in the Jordanian market.

Table 1: Attribute Information

Attribute	Description of Attribute
Make	The brand or manufacturer of the car (e.g., Toyota, Honda).
Model	The specific model of the car (e.g., Corolla, Civic)
Year	The year the car was manufactured.
Transmission	The type of transmission (e.g., automatic, manual).
Fuel	The type of fuel the car uses (e.g., petrol, diesel).
Condition	The overall condition of the car (e.g., new, used).
Kilometers	The total distance the car has traveled, measured in kilometers.
Paint	The state of the car's paint (e.g., original paint, partly painted, repainted).
Interior Options	Features or options available in the car's interior (e.g., leather seats, air conditioning).
Exterior Options	Features or options available on the car's exterior (e.g., alloy wheels, sunroof).
Jayed	
Price	The selling price of the car.

Data Loading and Preprocessing

The first step in our car price prediction project involves loading the dataset obtained from OpenSooq, the leading online marketplace for cars in Jordan. Once the data is loaded, we proceed with comprehensive data preprocessing to ensure its quality and suitability for analysis. This preprocessing includes handling missing or inconsistent values, encoding categorical variables into numerical formats, detecting and treating outliers, and scaling or normalizing numerical features to facilitate effective analysis.

Additionally, we perform feature engineering to create new relevant features that could enhance the predictive power of our models, and feature selection to identify and retain the most impactful variables for predicting car prices.

In this data preprocessing, duplicate records are removed to ensure the dataset is free from redundancy and maintains data integrity. Additionally, car brands with fewer than 30 occurrences are considered noisy data and are excluded, leaving only brands with a count of more than 30 for further analysis.

the "Condition" column is updated based on the "Kilometers" value, where any non-zero "Kilometers" marks the car as "Used" and cars with "Kilometers "equal to 0 but with a null "Condition"are marked as "New."

The "Kilometers" column is split into `Kilometers1` and `Kilometers2`, converted to floats. A new column, "Average Kilometers", is calculated, using either a direct assignment from `Kilometers1` if it's more than 200,000 or the average of `Kilometers1` and `Kilometers2`.

Missing values in "Average Kilometers" are filled with 0 for new cars, and for used cars with an original value of 0.0 in "Average Kilometers", it is updated to 109999.5 the Median value for all cars. Finally, the `Kilometers1` and `Kilometers2` columns are dropped.

The "Price" column is converted from a string to an integer format for proper numerical analysis. Missing values in the "Paint" column are filled with "Original Paint" if the "Condition" is "New," and "Other" otherwise.

For the "Interior" and "Exterior"options, the 10 most significant interior options and 7 most significant exterior options are identified, and a new column is created, with `True` or `False` values based on whether the specified options are present. Finally, the "Interior Options" and "Exterior Options" columns are dropped from the dataset to streamline the data for further processing.

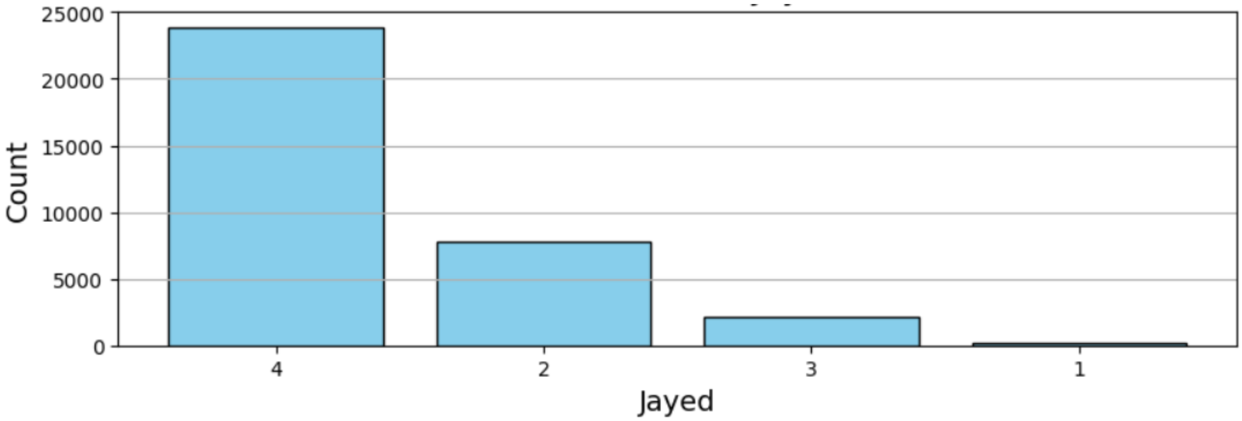
Due to the large number of missing values in the `Jayed` column, I developed a machine learning model to predict and fill these missing values. This approach ensures that the dataset remains complete without discarding valuable data. By leveraging the existing features in the dataset, the model can estimate the missing values in the `Jayed` column, allowing for more accurate analysis and preserving the integrity of the dataset. This method minimizes the potential bias and inaccuracies that could arise from simply dropping or filling the missing values with arbitrary assumptions.

Exploratory Data Analysis and Visualization

After preprocessing the data, we conduct an exploratory data analysis (EDA) to gain deeper insights into the dataset's structure and underlying patterns. This involves generating descriptive statistics and visualizations for various attributes such as car make, model, year, condition, and price.

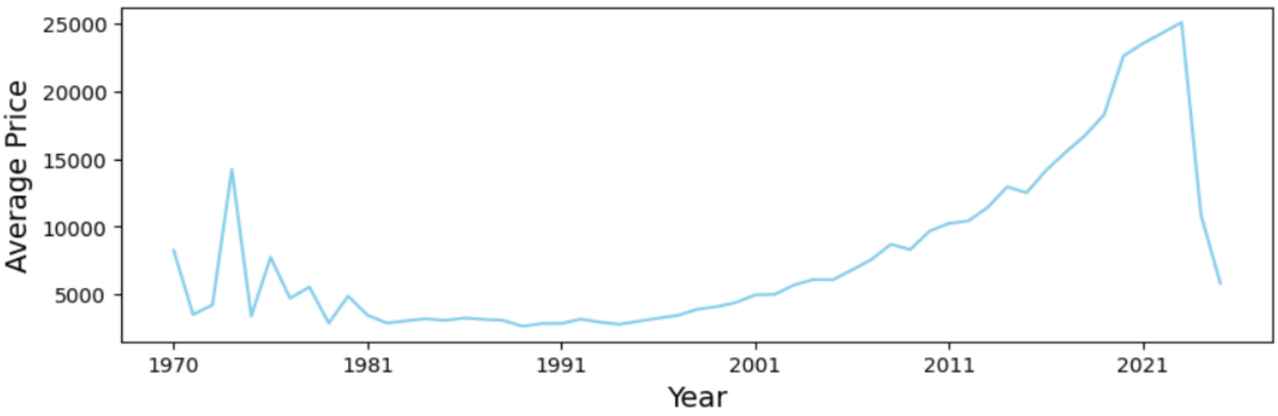
Through techniques like histograms, box plots, scatter plots, and correlation matrices, we explore the distributions and relationships between different features, uncovering trends and anomalies that inform our modeling approach. This thorough understanding of the data assists in refining our feature selection and guides the development of robust predictive models tailored to the nuances of the Jordanian car market.

Figure 1: Distribution of Jayed



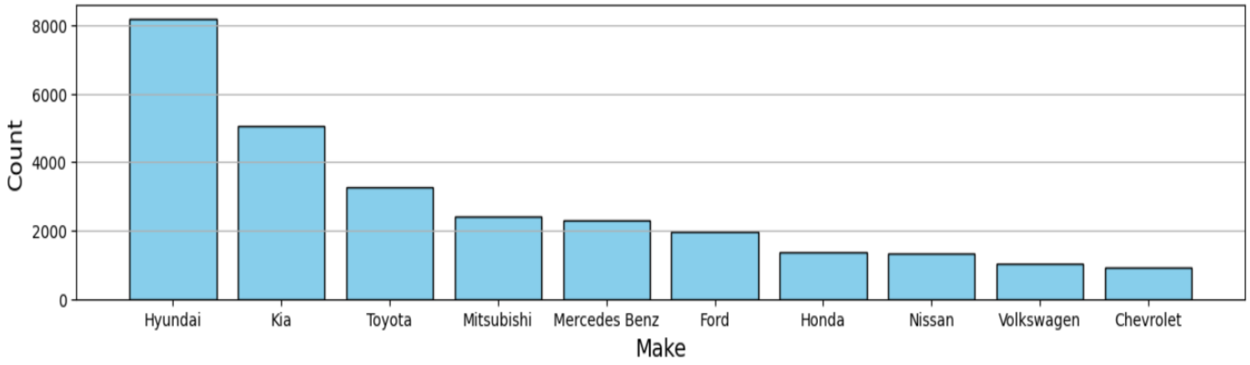
As shown in the bar plot, the most common value in the "**Jayed**" column is 4, while the least common is 1

Figure 2: Average Price Vs Year



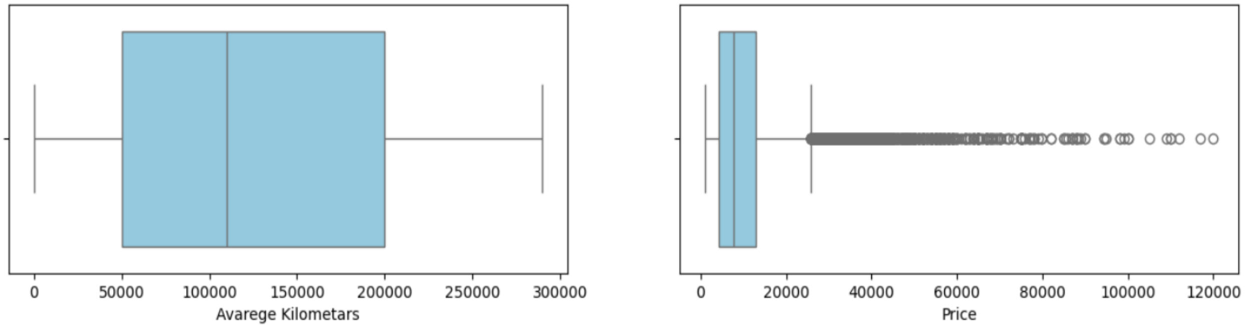
In this line chart, we can observe that car prices generally increase over the years. However, it's worth noting that there are some older cars with unusually high prices, which may be outliers. Additionally, there are a few cars from 2023 with lower prices, possibly due to the lack of data for 2023 cars.

Figure 3: Distribution of Brands



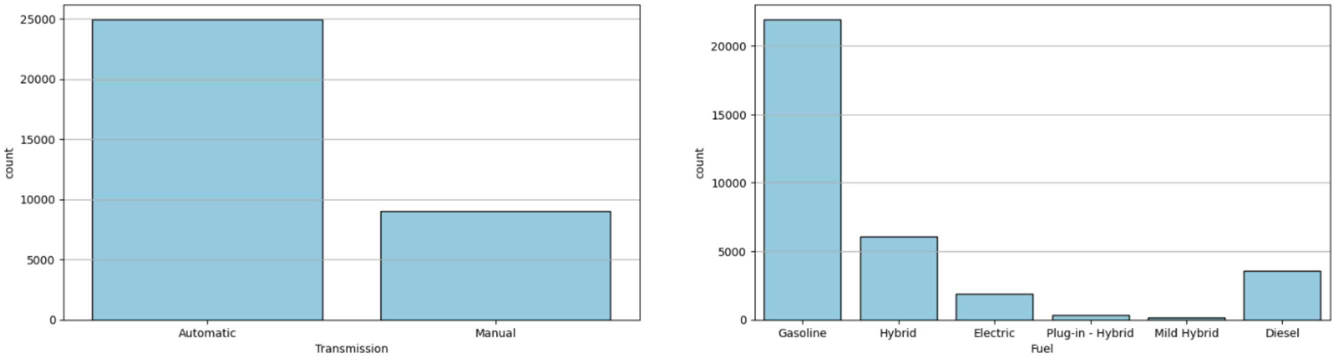
In this bar chart, we can see the top 10 most frequent car brands in the data set. It is obvious that Hyundai is the most prevalent car brand in Jordan.

Figure 4: Distribution price and average kilometers



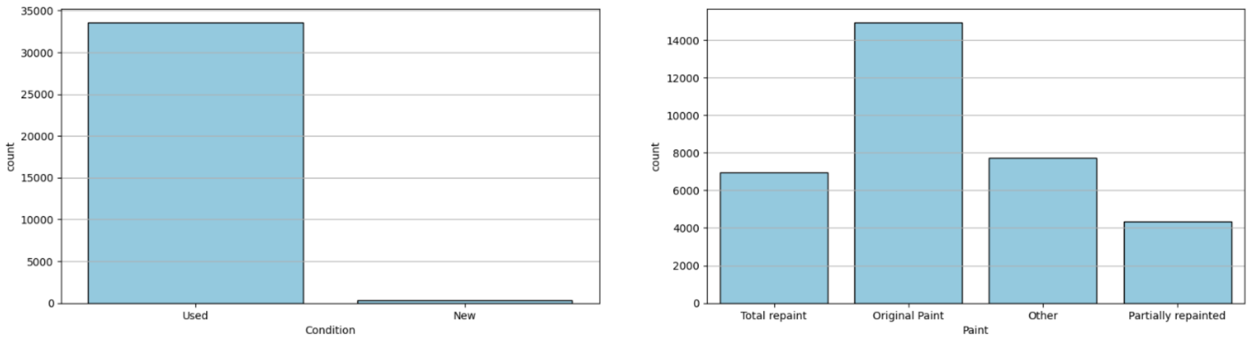
In the two box plots showing the distribution of average kilometers and price features, we can observe that there are many outliers in the price feature. We have the option to either address these outliers or leave them as they are. On the other hand, in the average kilometers plot, there are no outliers, which is a positive result.

Figure 5: Distribution categorical Features



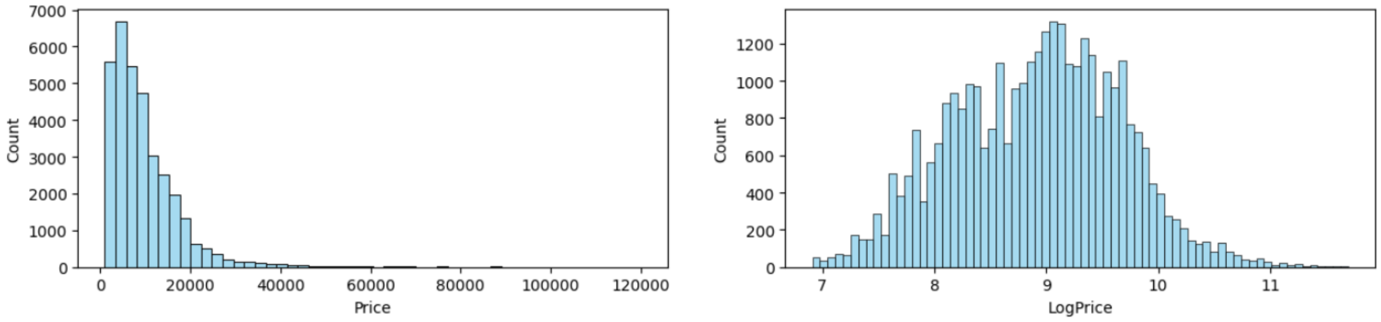
In the two bar charts displaying the distribution of two features - Fuel and Transmission - we observe that most cars in Jordan are automatic. In the fuel section, we notice that gasoline is the most frequent type. Perhaps in the future, electric and hybrid cars may become more prevalent.

Figure 5: Distribution categorical Features



In this bar chart, we can see the distribution of car conditions and paint types, most of the cars on OpenSooq are used, with only a few being new. When it comes to paint, the majority of the cars have original paint.

Figure 5: Price Log Transformation



When looking at charts that describe the price distribution before and after logarithmic transformation, we can see that the log transformation handles the skewed data distribution, obvious in the price.

Model Development

In the model development phase, we implemented and tested three different regression models to predict car prices in the Jordanian market. The models selected for this experiment were the **Decision Tree Regressor**, **Random Forest Regressor**, and **XGBoost Regressor**. The Decision Tree Regressor provides a simple yet powerful model that captures non-linear relationships within the data. The Random Forest Regressor, an ensemble method, builds upon multiple decision trees to improve accuracy and reduce overfitting. Finally, the XGBoost Regressor, known for its high performance and efficiency, leverages gradient boosting to optimize model predictions. By comparing the performance of these models, we aimed to identify the most accurate and reliable model for car price prediction.

The dataset is split into features and target variables, with features excluding the "**Price**" column. The data is divided into training and testing sets to evaluate model performance. Different regression models, including XGBoost, DecisionTree, and RandomForest, are then trained and tuned using hyperparameter optimization techniques such as "**RandomizedSearchCV**" and "**GridSearchCV**". Each model's performance is assessed through metrics like Mean Squared Error (MSE) and R-squared. The results are compared to identify the best-performing model based on its accuracy and predictive capabilities.

Table 2: Results Summary

Model	***Decision Tree***	***RandomForest***	***XGBoost***
R^2	**0.77**	**0.82**	**0.83**
MSE	**17246051.59**	**13507987.47**	**13111162.16**