Predicting Vehicle Condition and Market Price Using Machine Learning

A Data-Driven Approach to Vehicle Evaluation and Price Forecasting

Estarta AI & DS training

Nashat Alfarajat

Table of Contents

Introduction

In the used car market, buyers and sellers often face challenges in accurately assessing a vehicle's condition and determining its fair price. Misjudgments can lead to financial losses, mistrust, or missed opportunities. This project aims to address these challenges by leveraging machine learning to provide reliable predictions for both vehicle condition and market price.

For this purpose, a dataset of used cars was collected, containing various attributes such as mileage, age, engine specifications, and other relevant features. The dataset was carefully explored and analyzed to gain insights and understand patterns that influence both car condition and price.

Finally, predictive models were built to classify the condition of cars (excellent, good, fair, poor) and estimate their market price. By combining classification and regression tasks, this project provides a comprehensive tool to assist buyers, sellers, and dealers in making data-driven decisions in the used car market.

About the Data

The dataset used in this project, named vehicle_dataset, contains **380,180 records and 26 attributes** related to used cars. Some of the key features include price, year, manufacturer, model, cylinders, fuel type, odometer, transmission, drive type, vehicle condition, and other descriptive details.

This dataset is particularly valuable because it is real-world data, collected from actual vehicle listings. It reflects the true variability and diversity found in the used car market, including different car types, conditions, and pricing patterns. Using such authentic data ensures that the predictive models built in this project are grounded in realistic scenarios, making their insights and predictions more practical and applicable for buyers, sellers, and dealers.

By leveraging this dataset, the project can not only classify car conditions but also predict market prices with a level of accuracy and relevance that synthetic or simplified datasets would not provide.

Data Cleaning and Preparation

Missing Values

The dataset had missing values in several columns. Some columns with too many missing values were removed, while rows missing important information like manufacturer, model, or condition were dropped.
For the remaining missing values, common replacements were used: most frequent values for categorical features, typical values for numerical ones, and a "Missing" label where appropriate. After preprocessing, the dataset contained 356,036 records and 15 features, ready for analysis and modeling.

Duplicate Rows

The dataset was checked for duplicate records, but none were found. This ensured that all entries were unique and suitable for analysis.

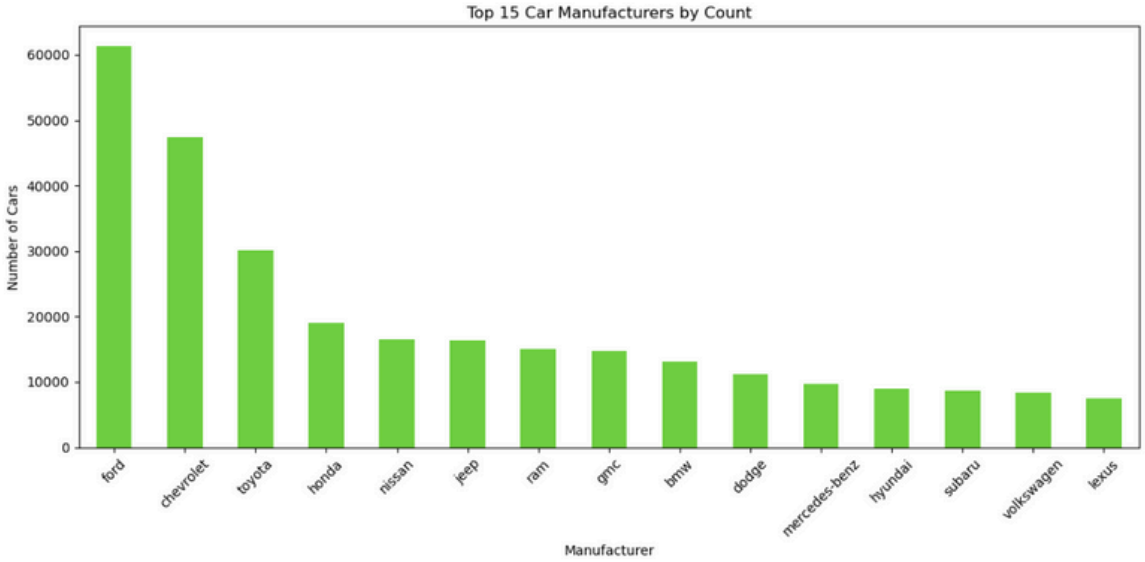
Outliers

Outliers were not removed aggressively, as extreme values can sometimes represent real, important data. Only very extreme cases were filtered out to ensure the dataset remained realistic.

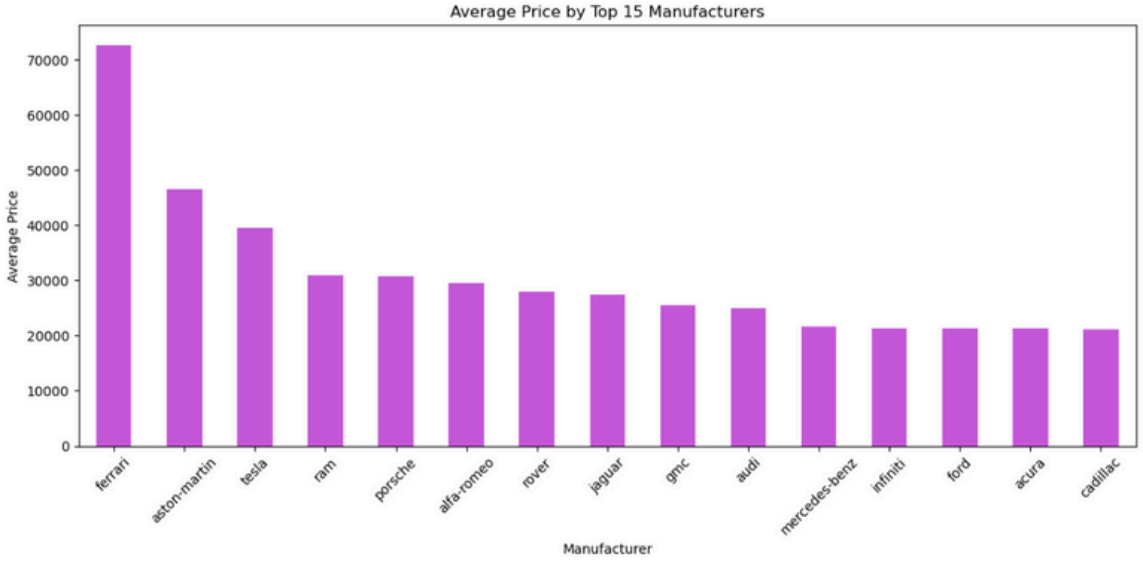
Specifically, cars with prices below $300 or above $110,000, odometer readings above 500,000 miles, or model years outside 1960–2025 were removed. This helped maintain data quality without losing meaningful information.

Exploring the Data: Visualizations and Key Insights

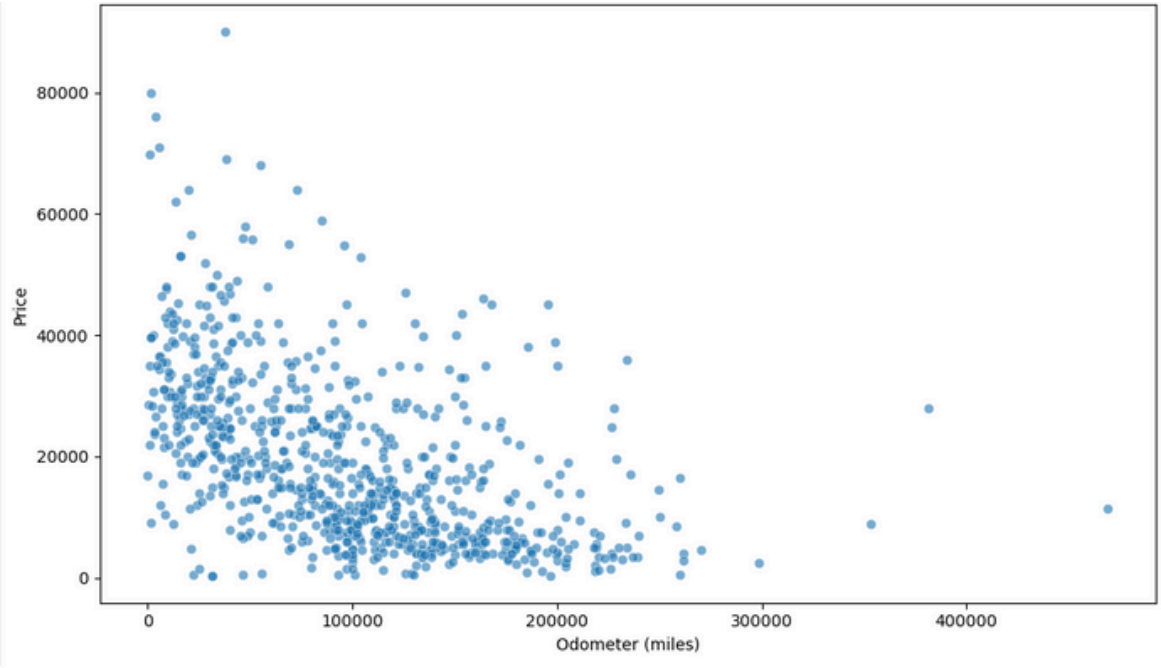
Top 15 Car Manufacturers by Count



Average Price by Top 15 Manufacturers



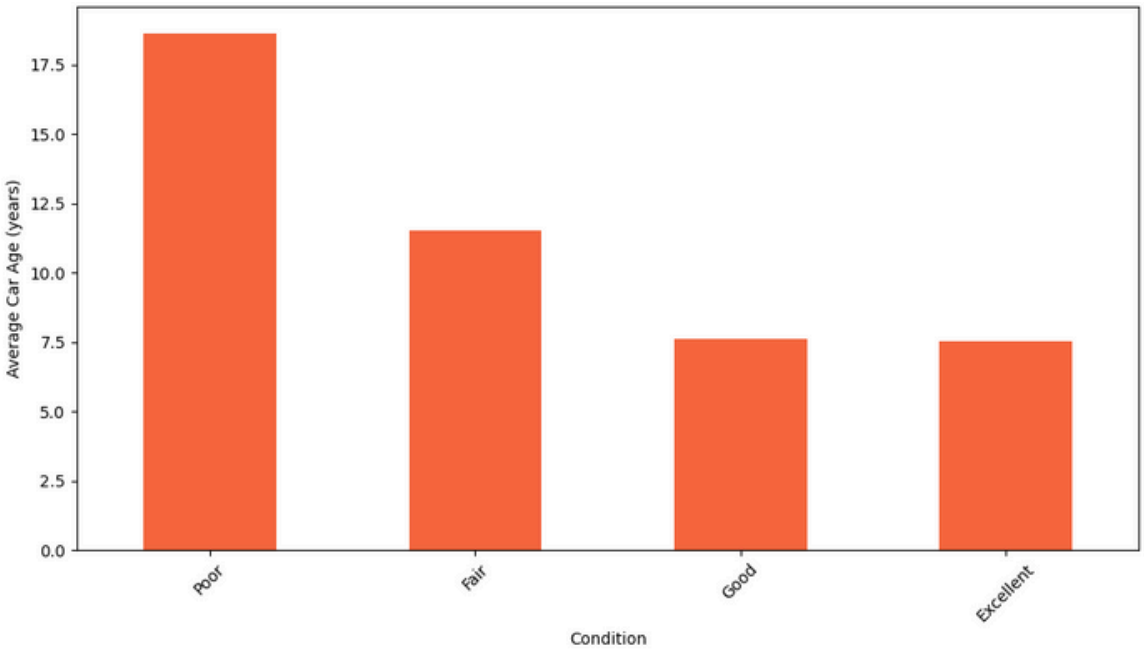
Price vs Odometer (Sample of 1000)



Feature Engineering

A new feature, Car Age, was created by subtracting the car's year of manufacture from the current year. This helps the model understand how a car's age affects its condition and price.

Average Car Age by Condition



Classification: Predicting Car Condition

The goal of this step was to classify car condition into four categories: Fair, Excellent, Good, and Poor. Categorical features and the target variable were first encoded so that the models could process them. The dataset was then split into training and testing sets.

Two models were trained for this task. The first model was a Decision Tree, chosen for its simplicity and interpretability. Feature selection was left to the model itself since tree-based methods can naturally handle irrelevant features, and the number of features was not large enough to require dimensionality reduction.

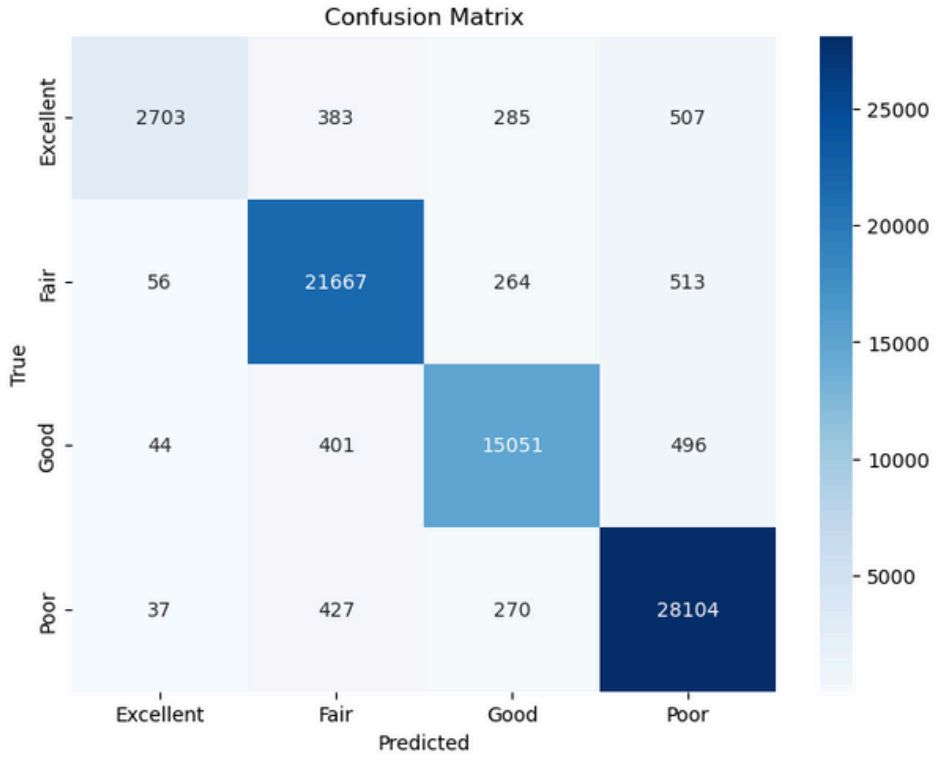
Decision Tree Results

The Decision Tree model achieved an accuracy of 89%, with strong precision and recall for most categories. While it performed well overall, there was room for improvement, especially for the Excellent category, which had slightly lower scores compared to others.

Since tree-based models performed well, using a boosted tree-based model was a natural next step to improve performance. This led to the choice of XGBoost, which often provides better accuracy and handles complex patterns more effectively.

XGBoost Results

The XGBoost model improved performance, achieving an accuracy of 94%. Most categories were predicted very well, and a confusion matrix is provided to visualize detailed results. This shows that boosting significantly enhanced the model's ability to classify car conditions compared to the simple Decision Tree.



Regression: Predicting Car Price

The next step was to predict the market price of cars using the available features. The model was trained on all relevant features except the condition and the price, which is the target variable.

Similar to classification, categorical features were encoded, and the dataset was split into training and testing sets. A tree-based model was chosen for regression due to its ability to capture non-linear relationships between features and price.

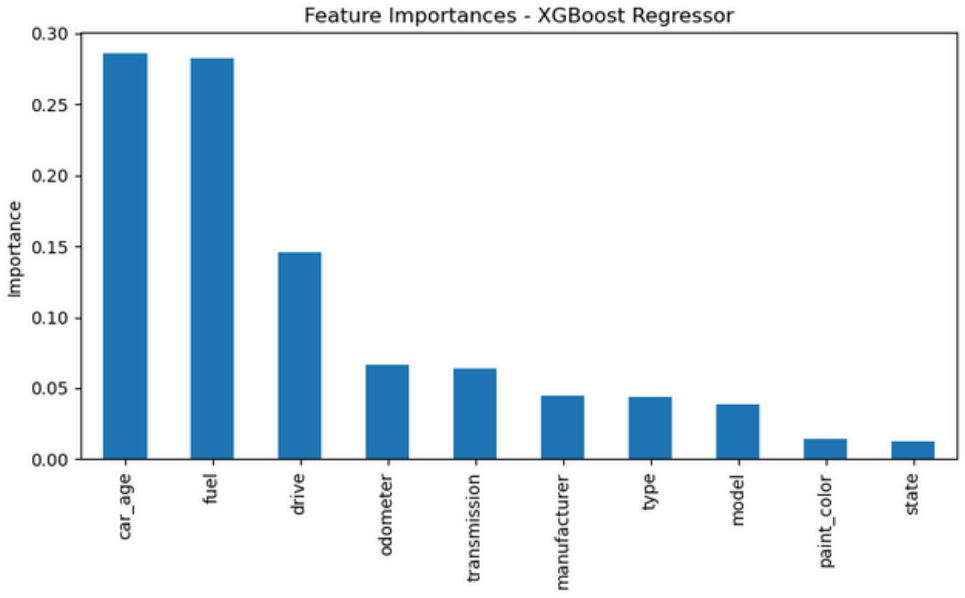
The regression model allows estimating a car's market value based on its attributes, helping buyers and sellers make informed decisions.

XGBoost for Estimating the car price

The XGBoost regression model shows strong predictive performance:

- The **MAE (3,345.72)** indicates that, on average, the model's predicted price is about $3,345 away from the actual price.
- The **RMSE (5,607.85)** is slightly higher, reflecting that larger errors are more heavily penalized, but still **reasonable given the range of car prices**.
- The **R^2 value (0.845)** means the model explains about 84.5% of the variation in car prices, showing that it captures most of the important patterns in the data.

Overall, these results suggest that the model can reliably estimate used car prices with good accuracy, but could be improved for our model.

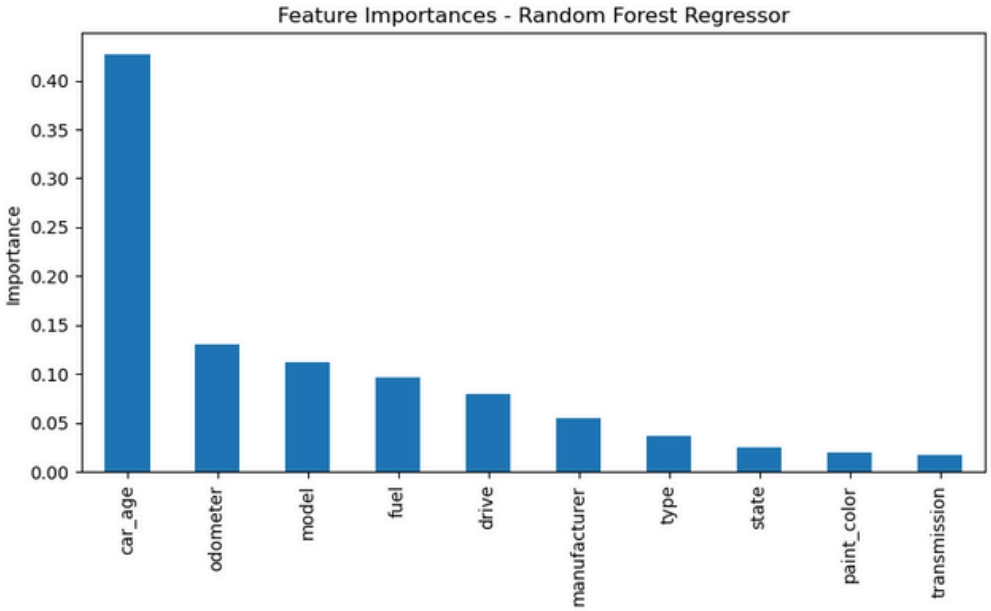


Improving Price Prediction: Random Forest

To improve the regression results, a Random Forest model was trained. This model achieved better performance than XGBoost:

- **MAE: 2,004.46**
- **RMSE: 4,524.44**
- **R^2: 0.8990**

These results indicate that the Random Forest model predicts car prices more accurately, with smaller errors and a higher proportion of variance explained. Feature importance was also analyzed, showing which attributes most influence car prices, such as car age, odometer reading, and manufacturer. A visualization of feature importance is provided to highlight the key drivers of vehicle value.



Linear Regression Performance

Linear regression models were also tested for predicting car prices, but they did not perform as well as tree-based models. This is because many relationships between the target variable (price) and the features are non-linear. For example, price does not increase uniformly with mileage or age, and interactions between features like manufacturer and fuel further complicate the patterns. As a result, linear models struggled to capture these complexities, making ensemble tree-based methods like Random Forest and XGBoost more effective for this dataset.

Conclusion

This project aimed to predict used car conditions and market prices using a real-world dataset. For classification, the models successfully categorized car condition into four classes: Excellent, Good, Fair, and Poor. A Decision Tree provided a solid baseline, achieving 89% accuracy, while XGBoost improved performance to 95%, demonstrating the effectiveness of boosted tree-based methods for classification tasks.

For price prediction, regression models were trained to estimate market value based on car attributes. XGBoost achieved good results, but the Random Forest model performed even better, with an R^2 of 0.8990, MAE of 2,004.46, and RMSE of 4,524.44. Feature importance analysis highlighted key drivers of price, such as car age, odometer reading, manufacturer, and condition. Linear regression models were tested but performed poorly due to the non-linear relationships present in the data.

Overall, the project demonstrates that tree-based machine learning models are well-suited for both classification and regression tasks in the used car market, providing accurate predictions and valuable insights for buyers, sellers, and dealers.