PROJECT REPORT DATA SCIENCE FALL 2023

Predicting Car Prices Based on Different Features

Abdul Rafey, Syed Zain Aamir, and Asad Jabbar

Department of Computer Science, National University of Computer and Emerging Sciences, Lahore

Submission Date: November 27, 2024

Abstract

Car price prediction is a vital aspect of the automotive market, aiding in transparency and informed decision-making for buyers and sellers. This study utilized a dataset sourced from Kaggle to develop a machine learning model for accurate car price predictions. The research analyzed multiple car features, such as engine specifications, body design, and fuel type, to determine their influence on pricing. Using exploratory data analysis and model training techniques, the project identified key trends and relationships within the dataset. The results demonstrate the potential for data-driven methodologies in enhancing price estimation accuracy.

1 Introduction

The automotive industry has witnessed significant growth and technological advancements. Pricing a car accurately remains a challenge, influenced by numerous factors like technical specifications and market trends. This study focuses on building a predictive model using machine learning techniques to estimate car prices based on various attributes. By doing so, it contributes to informed decision-making for manufacturers and consumers. The dataset includes comprehensive details about cars, serving as the foundation for analysis and model development.

2 Methodology

The project followed these steps:

1. **Data Collection**: The dataset was sourced from Kaggle.
2. **Preprocessing**: Data cleaning included handling missing values, encoding categorical variables, and normalizing numerical features.
3. **Exploratory Data Analysis (EDA)**: Key trends and correlations were visualized to understand the relationships between variables like engine size, fuel type, and car price.
4. **Model Selection and Training**: Various regression models (Linear Regression, Random Forest, and Gradient Boosting) were implemented to predict car prices.
5. **Evaluation**: Models were evaluated using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to identify the best-performing model.

3 Experiments

The dataset included 11915 rows and 16 columns, with significant features such as:

- **CarName**: The car's name and brand.
- **Enginesize**: Engine displacement in cubic centimeters.
- **Horsepower**: Engine power output.
- **Fueltype**: Gas or diesel fuel used by the car.
- **Price**: Target variable representing the car's market value.

Figures were created to visualize distributions and correlations. For instance, a positive correlation was observed between engine size and price. Multiple machine learning models were trained and tested to determine the best predictor.

4 Results & Discussion

The analysis of the dataset and the performance of various models led to the selection of Gradient Boosting as the most suitable algorithm for predicting car prices. Here is a detailed discussion of the results and why Gradient Boosting outperformed other models.

Model Comparison

Three machine learning models were tested: XGBoost, Random Forest, and Gradient Boosting. Their performance was evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and *R^2* score, with the following results:

Model	MAE	RMSE	*R^2*
XGBoost	10994	23024	0.832
Random Forest	11758	22109	0.845
Gradient Boosting	10676	22054	0.846

Key Insights

1. **Gradient Boosting's Strengths**

   Gradient Boosting achieved the lowest MAE (10,676) and RMSE (22,054), indicating its predictions had the smallest average error and it was the most accurate at minimizing large prediction errors. It also achieved the highest R^2 value (0.846), meaning it explained 84.6% of the variance in car prices, slightly outperforming Random Forest and XGBoost.

2. **Performance of Other Models**

   a. XGBoost: XGBoost, with an R^2 of 0.832, performed reasonably well but had the highest RMSE and MAE values. Its performance may have been affected by challenges in optimizing its hyperparameters on this dataset.
   b. Random Forest: Random Forest showed a strong performance with an R^2 of 0.845 and a lower RMSE than XGBoost. However, it was slightly less precise than Gradient Boosting, especially for complex, non-linear relationships in the data.

3. **Why Gradient Boosting Outperformed**

   Gradient Boosting's iterative and adaptive nature was key to its success. Specific factors include:

   a. **Error Minimization**: The model iteratively refined its predictions by minimizing residual errors from prior iterations.
   b. **Feature Interactions**: Gradient Boosting handled interactions like the combined effects of engine size and fuel type more effectively than the other models.
   c. **Regularization and Tunability**: Hyperparameter tuning allowed for precise control over learning rate, maximum depth, and number of estimators, leading to optimal performance.

Interpretations and Implications

The superior performance of Gradient Boosting indicates that car prices are influenced by a complex set of features. For instance:

- **Horsepower**: Higher horsepower were strongly correlated with increased prices, reflecting their importance as luxury indicators.
- **Body Type and Fuel Type**: Sedans and cars with diesel engines commanded higher prices, aligning with market demand trends.

Limitations and Improvements

While Gradient Boosting demonstrated impressive results, certain limitations were observed:

- **Dataset Size**: The dataset contained only 11915 entries, which may have restricted the model's ability to generalize to real-world scenarios.
- **Feature Diversity**: Including additional variables such as mileage, market demand, or geographic factors could improve the model's predictive power.
- **Overfitting Risk**: Although hyperparameter tuning minimized overfitting, Gradient Boosting's inherent complexity makes it sensitive to overtraining on small datasets.

Future research could focus on incorporating larger and more diverse datasets, exploring deep learning models, and using advanced feature engineering techniques to improve performance further.

Conclusion and Future Work

This project successfully developed a machine learning model to predict car prices using key features. The results highlight the significance of data-driven approaches in the automotive industry. Future research could enhance the model's robustness by including additional features like mileage and market demand trends.

References

Car Price Prediction Dataset, Kaggle