Used Car Price Prediction

Harsh Agarwal Abhinav Verma Bharat Agarwal

November 24, 2024

Abstract

This study uses a dataset from Cars.com with 4,009 vehicle listings to develop predictive models for used car prices. Data preprocessing ensures quality and relevance, while exploratory analysis reveals the relationship between vehicle features and pricing. Feature engineering, including derived metrics and categorical encoding, enhances model inputs. Advanced machine learning techniques like Random Forest and CatBoost predict prices based on vehicle attributes. An ablation study identifies key features, highlighting the importance of model year and mileage. The results show the models' accuracy in predicting car prices, offering valuable insights for buyers and sellers. This report outlines the methodologies and discusses the broader implications for automotive sales analytics.

Keywords:
used car prices, predictive modeling, machine learning, feature engineering, data analysis

1 Introduction

The used car market plays a vital role in the automotive industry, offering diverse options for consumers. Accurate price prediction is essential for ensuring equitable transactions and effective inventory management. The advent of online car sales platforms has provided a wealth of data that can be utilized to refine these predictions.

This project leverages a dataset from Cars.com, which includes detailed information on 4,009 vehicles, to develop predictive models using advanced machine learning algorithms. The study commences with an exploratory data analysis (EDA) to identify patterns and relationships, followed by feature engineering to optimize the inputs for modeling. We employ Random Forest and CatBoost algorithms, known for their robustness in handling complex datasets.

Our approach also includes ablation studies to assess the influence of different features on the model's accuracy, allowing for iterative refinements. The results aim to enhance understanding of the factors affecting car prices, providing valuable insights for both academic research and practical applications in the used car market. The following sections detail our methods, present findings, and discuss the broader implications for pricing strategies in the automotive sector.

2 Exploratory Data Analysis (EDA)

The dataset used in this study is derived from Cars.com, encompassing a diverse range of used vehicles. It comprises 50,000 entries, each representing a unique vehicle listing. The dataset includes several attributes that describe various aspects of each vehicle, which are crucial for the accurate prediction of used car prices.

Features Overview:

- **Brand:** Manufacturer of the vehicle.
- **Model:** Specific model of the vehicle.
- **Model Year:** Year the vehicle was manufactured.
- **Mileage:** Total miles the vehicle has been driven.
- **Fuel Type:** Type of fuel the vehicle uses (e.g., Gasoline, Diesel, Hybrid).
- **Engine:** Description of the vehicle's engine, including power output and engine type.
- **Transmission:** Type of transmission system in the vehicle (e.g., Automatic, Manual, Dual Shift Mode).
- **Exterior and Interior Color:** Color of the vehicle's exterior and interior.
- **Accident History:** Information on whether the vehicle has been involved in any accidents.
- **Clean Title:** Indicates whether the vehicle has a clean title, affecting its legal status and resale value.
- **Price:** Listed price of the vehicle, which is the target variable for our predictive modeling.

Exploratory Data Analysis was conducted to gain insights into the dataset and understand the underlying distributions and relationships of

features. This section presents the results of the analysis including data quality, feature distributions, and correlations among features.

2.1 Data Quality

Our initial analysis revealed no missing values across all features, indicating high data quality and completeness. Additionally, the examination of unique values in each feature provided insights into the diversity of data, particularly in categorical features like brand, model, and color.

2.2 Feature Distributions

- **Price Distribution:** The price of vehicles showed a right-skewed distribution, suggesting that while most used cars are priced on the lower end, there are a few high-priced outliers in the market.
- **Model Year Distribution:** The distribution of the model year is left-skewed, indicating a higher number of newer vehicle listings in the dataset. This skewness reflects a market trend towards selling relatively newer used cars.
- **Mileage Distribution:** Mileage exhibited a right-skewed distribution, with most vehicles having lower mileage, which correlates with the higher numbers of newer models.

2.3 Correlation Analysis

The correlation analysis focused on numeric features, revealing several expected relationships:

- A strong negative correlation between the model year and mileage, indicating that newer cars tend to have lower mileage.
- The price is moderately negatively correlated with mileage and positively with the model year, suggesting that prices decrease with increased mileage and older model years.

2.4 Figures

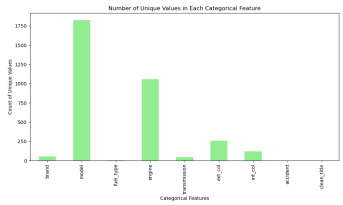

Figure 1: Number of Unique Values in Each Feature

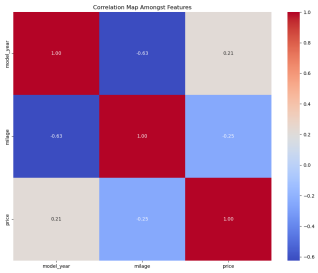

Figure 2: Correlation Map Amongst Features

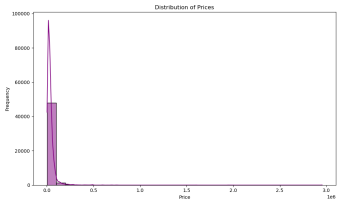

Figure 3: Distribution of Prices

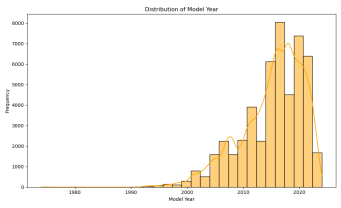

Figure 4: Distribution of Model Year

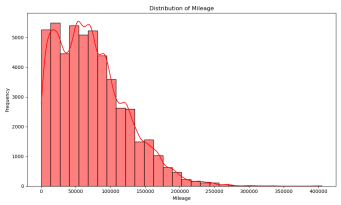

Figure 5: Distribution of Mileage

3 Feature Engineering

Feature engineering was essential for this particular project as the categorical features consisted a lot of Unique Values, so it wasn't possible to have one-hot encoding on them, Additionally, there's empirical knowledge of how certain features help in feature engineering. In this project, several new features were derived from the dataset, each serving a specific purpose:

3.1 Damage Indicator

A binary variable, `damage`, indicates accident history: 0 for "None reported," 1 otherwise, as accident history significantly affects car value.

3.2 Vehicle Age

The `age` of the car, calculated as the difference between the current year (2024) and the

`model_year` of the car. Older cars generally have lower prices, making this feature valuable for predicting car prices.

3.3 Exterior and Interior Colors

The popularity and neutrality of exterior and interior colors (`ext_common`, `ext_neutral`, `int_common`, `int_neutral`) were encoded to reflect preferred color schemes, capturing their influence on consumer preferences and car pricing.

3.4 Transmission Type

The `transmission` feature was categorized into 'Automatic,' 'Manual,' and 'Semi-Automatic,' with binary indicators (`Automatic`, `Manual`, `Semi`) added to aid model training, reflecting its impact on driving experience and market value.

3.5 Engine Popularity and Specifications

Popular engine configurations were identified, and a binary indicator (`popular_engine`) was created to highlight high-demand engines. `horsepower` and `engine_capacity` were extracted from engine descriptions using regular expressions to quantify engine power and size.

3.6 Brand Categorization

Vehicle brands were categorized as 'Luxury,' 'Affordable,' or 'Sports' (`Luxury`, `Affordable`) based on market segments, reflecting brand prestige and positioning, key factors influencing pricing.

3.7 Fuel Type

A binary variable (`gasoline`) indicates whether a car uses gasoline, reflecting its impact on running costs and environmental considerations, which are crucial for valuation. **Reduction of Dimensionality:** Finally, non-essential columns were dropped to focus on the most impactful features and reduce overfitting.

These engineered features incorporate domain knowledge and empirical observations into the dataset, thereby allowing the model to learn more complex patterns and improve its performance.

4 Methodology

This section outlines the methodologies used to develop predictive models for car price estimation. Various machine learning algorithms were employed, each tailored through hyperparameter tuning to optimize performance.

4.1 Model Training

Three different types of regression models were trained using the prepared features:

- **CatBoostRegressor**: A gradient boosting framework that handles categorical variables natively. It is known for its speed and efficiency at handling large datasets.
- **XGBoostRegressor**: An implementation of gradient boosted decision trees designed for speed and performance.
- **RandomForestRegressor**: An ensemble method that uses multiple decision trees to ensure lower risk of overfitting than a single decision tree and improve prediction accuracy.

For each model, the dataset was split into training and testing sets with 80% of the data used for training and 20% reserved for testing. This split ensures that the models are evaluated on unseen data, providing a measure of their generalization ability.

4.2 Hyperparameter Tuning

Hyperparameter tuning was performed using Grid Search to optimize each model's configuration. This process involves selecting a combination of parameters that results in the best model performance as measured by the root mean squared error (RMSE).

4.2.1 CatBoost Parameters

- **Iterations**: Number of boosting stages to be run.
- **Depth**: Depth of each tree.
- **Learning Rate**: Step size shrinkage used to prevent overfitting.
- **Random Strength**, **Bagging Temperature**, and **L2 Leaf Reg**: Regularization parameters to manage model complexity and robustness.

4.2.2 XGBoost Parameters

- **N Estimators**: Number of gradient boosted trees.
- **Max Depth**: Maximum depth of a tree.
- **Learning Rate**: Boosting learning rate.
- **Subsample** and **Colsample Bytree**: The fraction of samples and features to be used for each tree, a strategy to prevent overfitting.
- **Min Child Weight**, **Reg Alpha**, and **Reg Lambda**: Parameters that help in adding more regularization to the model.

4.2.3 Random Forest Parameters

- **N Estimators**: Number of trees in the forest.
- **Max Features**: The number of features to consider when looking for the best split.
- **Max Depth**, **Min Samples Split**, and **Min Samples Leaf**: These parameters control the size and complexity of the trees.

5 Ablation Studies

Ablation studies provide insights into the significance of individual features by observing the effect on model performance when each feature is systematically removed.

5.1 Methodology

For each feature, we removed it from the dataset and retrained the model to assess the impact on the Root Mean Squared Error (RMSE).

5.2 Results and Insights

The results of the ablation studies are illustrated in Figure 6, which shows the increase in RMSE when each feature is removed. This is complemented by the feature importance rankings directly derived from the CatBoost model, shown in Figure 7.

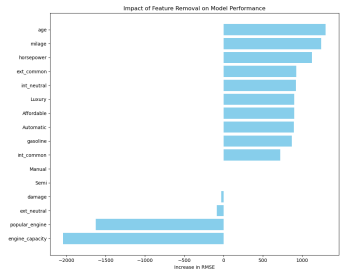

Figure 6: Impact of Feature Removal on Model Performance

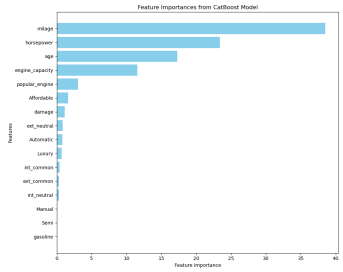

Figure 7: Feature Importances from CatBoost Model

5.2.1 Key Observations

- **Age and Mileage**: The removal of 'age' and 'mileage' resulted in the most significant increase in RMSE, underscoring their critical role in predicting vehicle prices. These features likely capture the depreciation and usage aspects of the vehicle pricing models.
- **Engine Features**: horsepower and enginecapacity also show a marked impact on RMSE when removed, highlighting the importance of engine characteristics in the valuation of used cars.
- **Less Impactful Features**: Features such as 'automatic', 'manual', and 'gasoline' have minimal impact when removed, suggesting that while these features contribute to the model, they are not primary drivers of price prediction in this context.

5.3 Implications

The ablation study not only validates the importance of top features as identified by the feature importances but also helps in refining the model by focusing on the most influential factors.

6 Results

The evaluation of the three models on the validation dataset yielded the following Root Mean Squared Errors (RMSE), which serve as a measure of the accuracy of the predictions:

- **CatBoost RMSE:** 64,230.44
- **XGBoost RMSE:** 63,620.39
- **Random Forest RMSE:** 62,523.12

These results indicate that the Random Forest model achieved the lowest RMSE, suggesting it was the most accurate in predicting the prices of used cars from the validation dataset. Conversely, the CatBoost model, despite its powerful handling of categorical features, did not perform as well as the other models.

7 Conclusion

The outcomes of the ablation studies and RMSE comparisons suggest that the Random Forest model not only offers the best performance in terms of prediction accuracy but also benefits from robustness against overfitting, given its ensemble nature. The XGBoost model also shows competitive performance, making it a valuable alternative for scenarios where execution speed is a priority.

7.1 Model Deployment

For deploying the model to predict prices on a test set, an ensemble method could be considered. By averaging predictions from all three models, the strengths of each could be leveraged, leading to more stable and accurate predictions across various data distributions. This approach helps mitigate the weaknesses of individual models, as ensemble methods often outperform single models in such scenarios.

However, if computational resources or time constraints are critical, selecting the Random Forest model may be the most practical choice due to its superior performance and simplicity.

- The GitHub Repository for the Project is attached Here.

References

[1] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, 2009.

[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[4] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, Springer, 2013. Chapter on Feature Engineering.

[5] "Feature Engineering Techniques for Machine Learning," Towards Data Science.