581QR

**Predicting Car Prices: A Comparative Study of Gradient Boosting Machine and Decision Tree Models**

Name: Kalyan B

The dataset is taken from Kaggle.

Link to the dataset

<https://www.kaggle.com/datasets/harikrishnareddyb/used-car-price-predictions>

Introduction

The automotive market is dynamic, and predicting car prices accurately is essential for both buyers and sellers. This study aims to predict the price of used cars based on various features using two machine learning models: Gradient Boosting Machine (GBM) and Decision Tree. The goal is to identify the better performing model and understand the key predictors influencing car prices.

Data Description

The dataset used for this study contains 10,000 records with the following columns

Price: The target variable representing the price of the car.

Year: The year the car was purchased.

Mileage: The number of kilometers driven by the car.

City: The city where the car was sold.

State: The state where the car was sold.

Vin: A unique identification number for the car.

Make: The manufacturer of the car.

Model: The specific model of the car.

Packages Used

1. dplyr: For data manipulation and transformation.
2. ggplot2: For data visualization.
3. caret: For data splitting and cross-validation.
4. gbm: For Gradient Boosting Machine modeling.
5. rpart: For Decision Tree modeling.
6. DescTools: For Winsorizing the data to handle outliers.
7. forcats: For lumping rare categories in categorical variables.
8. stringr: For string manipulation and handling
9. tidyverse: A collection of R packages including dplyr, ggplot2, and others for data science tasks.

Data Cleaning and Preparation

* Data Reduction:

The original dataset was larger, and we reduced the number of rows to 10,000 for efficient modeling and analysis. This subset was chosen to ensure it represents the full range of variability in the original data.

* Removing Columns:

Removed unnecessary columnssuch as Vin because vin number is not needed for the prediction

* Handling Categorical Variables:

Reducing Categories: The 'City' column originally had many unique values, which could lead to overfitting and computational inefficiency. We used the fct\_lump function to keep only the top 50 most frequent levels and lump the rest into a single 'Other' category.

* Handling Missing Values:

Checked and confirmed there were no missing values in the dataset.

* Outlier Handling:

Used Winsorization to handle outliers in the 'Mileage' and 'Price' columns.

* Categorical Encoding:

Converted categorical variables into factors for model compatibility.

Data Visualization

1.By using a scatter plot, it has been observed that the relationship between Mileage and Price:

There is a negative correlation between Mileage and Price, indicating that as the mileage increases, the price tends to decrease. This makes intuitive sense as cars with higher mileage are generally older or have been used more extensively, which typically reduces their market value.

2.Distribution of Mileage

Histograms were used to visualize the distribution of the continuous variable Mileage:

The distribution of mileage is right-skewed, with most cars having lower mileage. There are fewer cars with very high mileage, indicating a higher number of newer or less used cars in the dataset.

3. Distribution of Price

Histograms were used to visualize the distribution of the continuous variable Price:

The distribution of car prices is also right-skewed, with a larger number of cars priced at the lower end of the spectrum. There are fewer high-priced cars, which could include luxury or premium models.

4. Distribution of Cars by State

A bar plot was used to visualize the distribution of cars across different states:

The distribution of cars by state shows variability, with certain states having a higher number of listed cars. This could be due to population density, market activity, or regional preferences. States with large urban centers are likely to have a higher number of listings.

Modeling Approach

Employed two different machine learning techniques to predict car prices:

Parameters: n.trees = 100, interaction.depth = 3

1. Gradient Boosting Machine (GBM):

* Mean Absolute Error (MAE): 2054.81
* Mean Squared Error (MSE): 8,857,565.91
* R-squared (R²): 0.9214

1. Decision Tree:

* Mean Absolute Error (MAE): 3566.25
* Mean Squared Error (MSE): 22,644,828.58
* R-squared (R²): 0.7991

Results and Interpretation

The GBM model outperformed the Decision Tree model on all performance metrics.

The R-squared value for the GBM model was significantly higher (0.9214) compared to the Decision Tree model (0.7991), indicating that the GBM model explains more variance in the car prices. The Percentage of Correct Predictions: 62.11 %

The Mean Absolute Error (MAE) and Mean Squared Error (MSE) for the GBM model were lower, suggesting better predictive accuracy and lower sensitivity to outliers.

Key Predictors

According to the GBM model, the most important predictors of car prices were:

1. Model
2. Mileage
3. Year

Conclusion

The Gradient Boosting Machine (GBM) model is more effective in predicting car prices compared to the Decision Tree model. This study highlights the importance of advanced ensemble techniques like GBM in handling complex datasets with multiple variables. Our findings suggest that the car model, mileage, and year of purchase are significant predictors of car prices. Future work could explore additional features and different machine learning algorithms to further improve predictive performance.