AIN 429 TERM PROJECT REPORT

Arda Deniz Ayyıldız 2210765018

Mahmut Arda Gümüş 2200765006

Introduction

The prediction of car prices is a critical task in the automotive industry and consumer market. This report explores the application of machine learning techniques, including Random Forest and Gradient Boosting models, to estimate car prices based on various

attributes such as mileage, car age, and engine size. The primary objective of this study was to evaluate the effectiveness of these models in achieving reliable price predictions.

Method

Dataset

The dataset used in this study contained various attributes of cars, including both numerical and categorical features. Non-numeric features were excluded during preprocessing to simplify the analysis. Missing values were handled by dropping rows containing them.

Preprocessing

1. **Data Splitting:** The dataset was split into training and testing sets using an 80-20 ratio.
2. **Normalization:** Numeric features were scaled to ensure uniformity across the dataset.

Model Selection

Three machine learning models were utilized:

1. **Random Forest Regressor:** A robust ensemble method that combines multiple decision trees.
2. **Gradient Boosting Regressor:** A boosting algorithm that builds models iteratively to minimize prediction errors.
3. **Linear Regression:** A statistical approach that models the relationship between a dependent variable and one or more independent variables.

Hyperparameter Tuning

RandomizedSearchCV was employed to optimize the hyperparameters of the Random Forest model. Parameters such as the number of estimators, maximum depth, and minimum samples split were tuned using a five-fold cross-validation approach.

Evaluation Metrics

The models were evaluated using the following metrics:

- **Mean Squared Error (MSE):** Measures the average squared difference between actual and predicted values.

- **R-squared (R^2):** Assesses the proportion of variance explained by the model.

Results

Random Forest Model

- **Initial Results:** The Random Forest model showed moderate accuracy with MSE and R^2 values indicating potential for improvement.
- **Optimized Results:** After hyperparameter tuning, the optimized model achieved a training score of 0.341 and a testing score of -0.027, reflecting better alignment with the dataset characteristics.

Gradient Boosting Model

- **Performance Metrics:** The Gradient Boosting model was evaluated at various learning rates. The optimal performance was observed with a learning rate of 0.01, yielding an MSE of 0.281 and an R^2 value of -0.2782.

Linear Regression Model

- **Performance Metrics:** The linear regression model yielded an MSE of 0.395 and an R^2 value of -0.004, indicating poor predictive accuracy for the dataset. This result highlights the limitation of linear regression in modeling complex relationships present in the data.

Visualizations

1. **Prediction vs Actual Prices:** A scatter plot demonstrated the alignment between actual and predicted prices for the Random Forest model.
2. **Residual Plot:** Highlighted the residuals (errors) for further analysis of model performance.

Discussion

The Random Forest, Gradient Boosting, and Linear Regression models showcased varying levels of predictive accuracy. While the Random Forest model benefited from hyperparameter tuning, Gradient Boosting displayed sensitivity to the learning rate. The

linear regression model struggled with the dataset's complexity, as reflected in its poor performance metrics.

Despite the moderate performance metrics of ensemble models, the insights gained from hyperparameter tuning and error analysis can guide further enhancements. For instance, advanced feature engineering and the inclusion of categorical variables through encoding could potentially improve model efficacy.

Future Work

1. **Feature Engineering:** Incorporating domain knowledge to create derived features could enhance predictive performance.
2. **Categorical Feature Encoding:** Investigating encoding techniques such as one-hot or ordinal encoding for categorical variables.
3. **Model Ensemble:** Combining multiple models through stacking or bagging to leverage the strengths of each.
4. **Dataset Augmentation:** Expanding the dataset to include more diverse and balanced samples.

References

1. IEEE Xplore. (2022). Predictive Modeling of Car Prices Using Machine Learning Techniques. https://ieeexplore.ieee.org/document/9800719