US Used Vehicle Price Prediction Using Machine Learning

Project by Lokesh Balasubramaniam

LXB240006

Course: CS 6375.002

Introduction

- Objective: Predict the price of used cars in 20 U.S. cities using machine learning.
- Dataset: Sourced from Kaggle.
- Focus: Data preprocessing, feature engineering, and model evaluation.

Dataset Overview

- Total columns: 66 (after preprocessing, reduced to key features).
- Key Features: Body type, fuel type, mileage, horsepower, age of vehicle, etc.
- Data Cleaning: Addressed null values, removed irrelevant columns.

Preprocessing

- 1. Removal of unnecessary columns (e.g., VIN, dealer zip code).
- 2. Handling missing data using mean aggregation by groups.
- 3. Creation of derived features like vehicle age.
- 4. Correlation matrix analysis for feature selection.

Feature Engineering

- Categorical Features: Encoded using OneHotEncoder.
- Numerical Features: Scaled using StandardScaler.
- Features selected for modeling: Body type, fuel type, horsepower, mileage, year, etc.

Machine Learning Models

- Models Used:
- 1. Linear Regression
- 2. Ridge Regression
- 3. Random Forest
- 4. Gradient Boosting
- Cross-validation for parameter tuning.
- Handling Outliers: Log transformation of target variable.

Evaluation Metrics

- Metrics used:
- R^2 (coefficient of determination)
- - RMSE (Root Mean Squared Error)
- Example:
- Random Forest: R^2 = 0.92, RMSE = 9,512.11
- Gradient Boosting: R^2 = 0.88, RMSE = 10,235.60

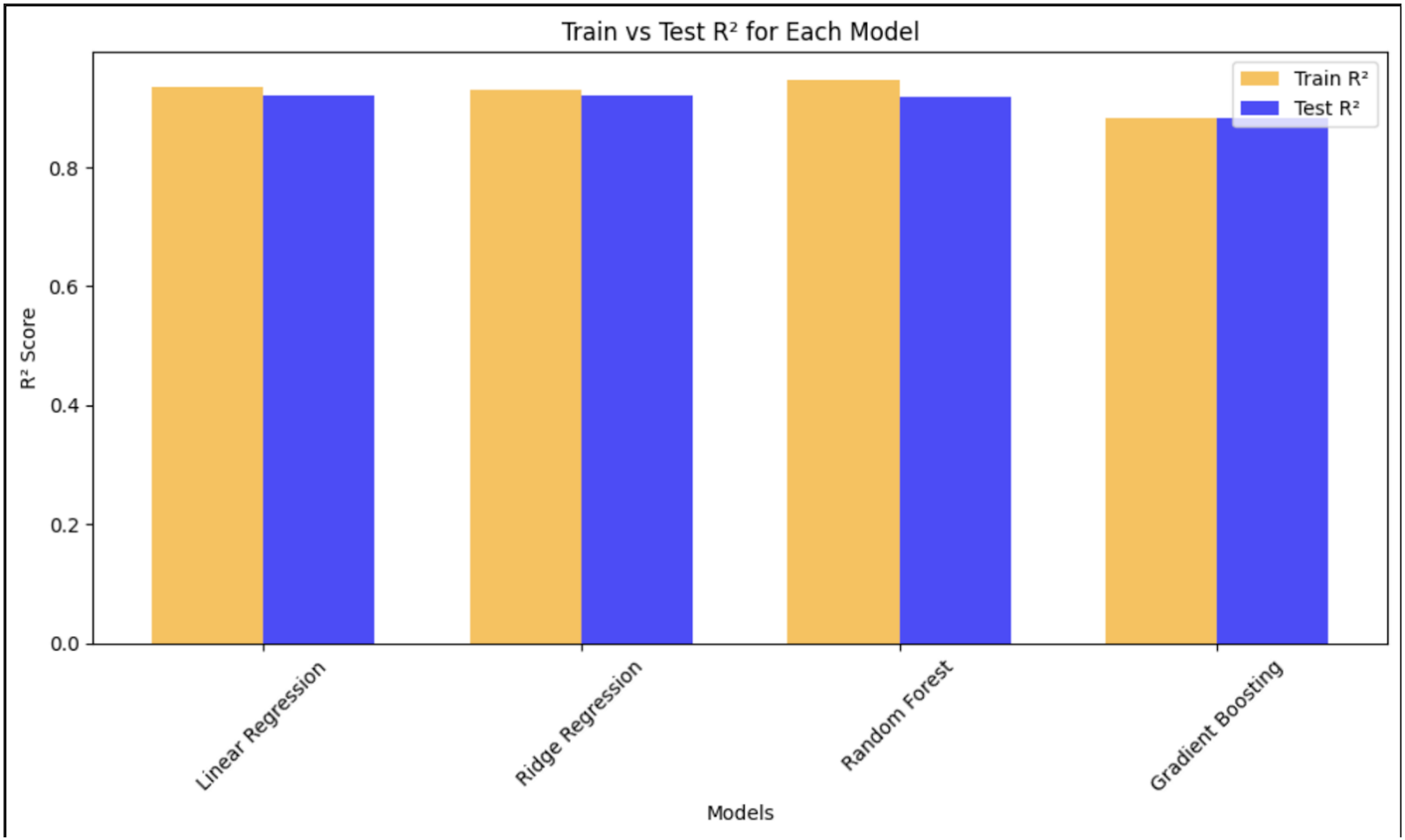
Results

- Best Performing Model: Random Forest.
- Key Observations:
- Ensemble models outperform regression models.
- Price predictions improved significantly after handling outliers.

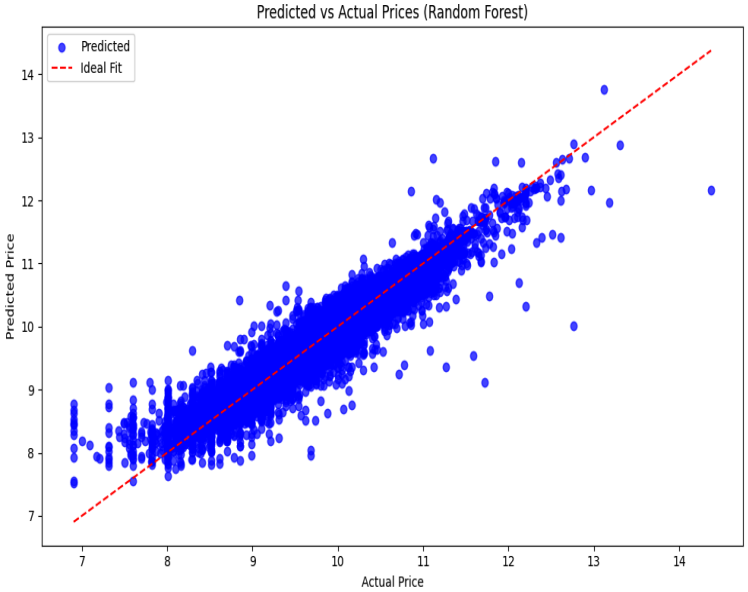
Visual Insights

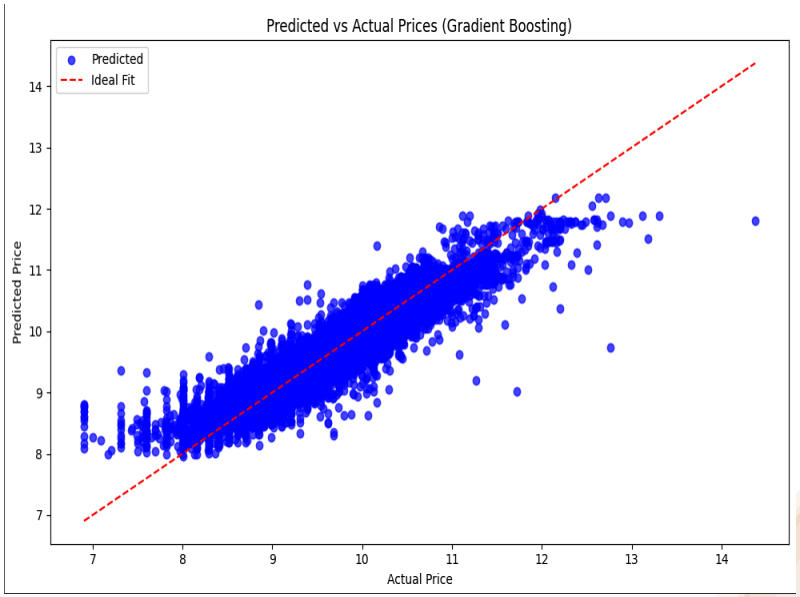
- Predicted vs Actual Price (Graphs for Random Forest and Gradient Boosting).
- Feature Importance: Highlight most significant predictors like mileage, age, horsepower.

Comparison of ML Models



Predicted Price vs Actual price





Conclusion

- Random Forest provides the best accuracy for used car price prediction.
- Practical Implications:
- Aid consumers and businesses in estimating car resale value.
- Enhance pricing strategies in the automobile market.

Future Work

- Explore more features (e.g., regional economic factors).
- Test additional models (e.g., XGBoost, LightGBM).
- Deploy model for real-time price predictions.

References

- - Kaggle Dataset: https://www.kaggle.com/datasets/ananaymital/us-used-cars-dataset
- - scikit-learn Documentation: https://scikit-learn.org
- - Spark API Documentation: https://spark.apache.org/docs/latest/api/python/index.html