Executive Summary

Project: Prediction of Used Car Price

Team Members

Song Gao, Navdeep Rana, Yutong Duan, and Xuzhi Tang

Overview

This project aims to use a machine learning algorithm to predict the price of used cars based on features such as mileage, model year, brand, accident history, etc. Being able to predict the price of used cars would allow used car dealers, car owners, online marketplaces to determine the best selling price, and help insurance companies to calculate premiums based on car value. We used data obtained from a Kaggle competition page to train different regression models, including Linear Regressions, Polynomial Regression, K-nearest neighbor regressions, Random Forest, and XGBoost regressions. We used cross-validation, and several KPI's (RMSE, R^2 score, MAPE, etc.) to compare different regression models.

Dataset

We used the "Used Car Price Prediction Dataset" data obtained from the Kaggle website with URL **https://www.kaggle.com/datasets/taeefnajib/used-car-price-prediction-dataset**. This data is publicly available for use under the CC BY 4.0 license.

This data has 4,009 instances where each row represents a unique vehicle listing and includes nine distinct features providing valuable insights into the world of automobiles. The features in this data set include:

- Brand & Model: Identifies the manufacturer and model of the vehicle.
- Model Year: Indicates the year of manufacture, important for assessing depreciation, technological advancements, and the vehicle's age.
- Mileage: Represents the distance traveled by the vehicle, used to estimate wear and tear.
- Fuel Type: Specifies whether the car uses gasoline, diesel, electric, or hybrid fuel.
- Engine: Includes details on engine power (horsepower), fuel injection type, cylinder count, and engine capacity.
- Transmission: Includes the transmission type (manual, automatic) and number of gears (e.g., 6-speed).
- Exterior & Interior Colors: Explore the aesthetic aspects of the vehicles, including exterior and interior color options.
- Accident History: Indicates whether the car has been involved in any accidents.
- Clean Title: Shows whether the vehicle has a clean legal title.

- Price: The target variable represents the vehicle's listed price.

Stakeholders

- Used Car Dealers
- Car Owners Looking to Sell
- Online Car Marketplace Platforms
- Insurance Companies
- Bank and Financial Institutions

Key Performance Indicators (KPIs)

The following Key Performance Indicators (KPIs) were used to evaluate the performance of the predictive models:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R^2 (Coefficient of Determination)
- Mean Absolute Percentage Error (MAPE)
- Symmetric Mean Absolute Percentage Error (SMAPE)

Assumptions

- EDA for the price suggested a highly skewed plot with outliers beyond $40000. Thus, to evaluate the best model for car price prediction, we have considered the price range to be $1000-$400000 for now.
- The electronic cars and the cars with other fuel types have different features. Due to very small electronic cars data present in the given dataset we have excluded the electronic cars.
- EDA showed car colors, accident status, clean title to have no correlation with price so we have used features with correlation beyond 0.4.

Models Used

- To evaluate the best model for car price prediction, we started with Linear Regression as a baseline model. Then, we tested other models like Polynomial Regression, Random Forest Regressor, XGBoost Regressor, and K-Nearest Neighbors (KNN).
- Each model was assessed based on R^2, MAE, RMSE, and other relevant metrics.
- We performed hyper-parameter tuning where necessary and incorporated feature engineering to improve the model performance.

The following table summarizes the key results obtained across the models:

Model	MAE	RMSE	R^2	MAPE	SMAPE
Linear Regression	$10389	$19366	0.77	27%	26%
Polynomial Regression	$7795	$14418	0.85	21.8%	21.2%
KNN Regression	$9006	$18007	0.84	24.9%	22.9%

Random Forest	$8407	$16058	0.87	21.6%	20.5%
XGBoost	$6945	$11254	0.90	22.2%	20.6%

Results

The model evaluation highlights that XGBoost is the best-performing model, with an R^2 of 0.90 and the lowest errors (MAE: 6,945, RMSE: 11,254). Random Forest follows with an R^2 of 0.87 and slightly higher error metrics. While Polynomial Regression and KNN showed moderate improvements over Linear Regression, they were less effective in handling complex patterns or required extensive tuning.

Conclusion

XGBoost is the most suitable model for predicting car prices due to its accuracy, robustness, and ability to manage data complexity.

Future Iterations

- Future iterations of this car price prediction project could explore several directions for improvement. Incorporating additional features, such as economic indicators, local market trends, or seasonal variations, could provide a deeper context for the pricing dynamics. Including these external factors might improve the model's ability to capture fluctuations in car prices beyond the available dataset.
- Enhancing the feature engineering process is another avenue for improvement. Creating interaction terms, using polynomial features judiciously, or applying feature selection techniques could better capture non-linear relationships and reduce redundancy in the data. Additionally, introducing advanced time-series methods, especially if the dataset spans multiple years, could offer insights into price trends over time.
- Experimenting with deep learning models, such as neural networks, may also be beneficial, particularly for handling complex patterns in large datasets. While tree-based models like Random Forest and XGBoost performed well, testing gradient boosting alternatives or stacking models might lead to further performance gains.
- Lastly, improving the hyperparameter tuning process through techniques like Bayesian optimization or automated machine learning (AutoML) could ensure the models are optimized effectively. Exploring ensemble strategies, such as blending predictions from multiple algorithms, could also enhance robustness and accuracy.