Predictive Modeling of Used Car Prices using XGBoost

Srikar Pottabathula
Department of Computer Science
Georgia State University
Atlanta, Georgia, USA
spottabathula1@student.gsu.edu

Monique Gaye
Department of Computer Science
Georgia State University
Atlanta, Georgia, USA
mgaye3@student.gsu.edu

Project Repository: https://github.com/srikar161720/used_car_price_prediction

Abstract

This study presents a comprehensive data mining approach to predict used car prices using a large-scale dataset from Craigslist vehicle listings. The dataset comprised 426,880 records with 26 features, representing diverse vehicle characteristics across the continental United States. A systematic four-phase methodology was employed: data cleaning and preprocessing, exploratory data analysis, feature engineering with baseline modeling, and advanced model development. Rigorous data preprocessing reduced the dataset to 348,001 high-quality records while eliminating all missing values through strategic imputation and removal techniques. Multiple regression algorithms were evaluated, including Ridge Regression, Random Forest, and XGBoost, with hyperparameter tuning performed via randomized search. The final tuned XGBoost model achieved superior performance with a test Mean Absolute Error of $2,979, Root Mean Squared Error of $4,924, and R^2 score of 0.848, demonstrating the effectiveness of gradient boosting techniques for price prediction. Feature importance analysis revealed that vehicle year, odometer reading, and manufacturer were the most influential predictors. This research contributes to the automotive market domain by providing an accurate, data-driven pricing model that can support both buyers and sellers in making informed decisions.

Keywords *— Data Mining, Machine Learning, Price Prediction, Used Car Valuation, XGBoost, Random Forest, Ridge Regression, Regression Analysis, Feature Engineering, Ensemble Methods, Data Preprocessing, Supervised Learning, Hyperparameter Tuning*

I. INTRODUCTION

The used car market represents a significant segment of the automotive industry, with millions of transactions occurring annually in the United States. Accurately determining the fair market value of used vehicles is crucial for both buyers seeking to avoid overpayment and sellers aiming to price competitively. Traditional valuation methods often rely on expert appraisals or simplified blue book estimates, which may not capture the complex interplay of factors influencing actual market prices.

Machine learning techniques offer a data-driven alternative that can model non-linear relationships and interactions between vehicle characteristics. With the availability of large-scale online marketplace data, it has become feasible to develop predictive models that learn from actual market transactions rather than relying solely on heuristic rules.

This research aims to develop and evaluate machine learning models for predicting used car prices using data from Craigslist, one of the largest classified advertisement platforms in the United States. The specific objectives of this study are: (1) to clean and preprocess a large-scale real-world dataset containing vehicle listings, (2) to identify the most important features influencing used car prices through exploratory analysis, (3) to compare the performance of multiple machine learning algorithms including linear and ensemble methods, and (4) to develop an optimized predictive model that achieves high accuracy on previously unseen data.

II. MATERIALS AND METHODS

A. Data Explanation and Characterization

The dataset consists of Craigslist used car and truck listings scraped across the United States [1]. The original dataset contained 426,880 records with 26 features, representing vehicle postings from April to May 2021. The dataset size was approximately 1.5 GB, with a memory footprint of 84.7 MB when loaded into pandas DataFrames.

Features in the dataset included both numerical and categorical variables. Numerical features comprised the target variable (price), vehicle year, odometer reading, and geographic coordinates (latitude and longitude). Categorical features included region, manufacturer, model, condition, number of cylinders, fuel type, title status, transmission type, drive type, vehicle type, paint color, and state. Additionally, several metadata features were present, including listing URLs, posting dates, and vehicle identification numbers (VINs).

Initial data profiling revealed significant data quality challenges. The price variable, serving as the target for prediction, exhibited extreme outliers with values ranging from $0 to over $3.7 billion, indicating data entry errors. Missing values were prevalent across multiple features, with some categorical features missing over 70% of their values. The model feature demonstrated extremely high cardinality with 22,637 unique values, presenting challenges for categorical encoding.

B. Data Cleaning and Preprocessing

A comprehensive data cleaning pipeline was implemented to address data quality issues while maximizing information retention. The preprocessing workflow consisted of several systematic steps executed in a specific order to ensure data integrity.

Feature selection was performed first, removing seven features deemed unsuitable for predictive modeling. URL-based features (url, region_url, image_url) were removed as they provided no predictive value. The description field, while potentially informative, was excluded as natural language processing was beyond the project scope. The inclusion of the county feature was as error during dataset creation and was dropped. The VIN feature exhibited extremely high cardinality with each vehicle having a unique identifier, making it unsuitable as a predictor. The size feature

was removed due to 72.34% missing values, where imputation would introduce excessive bias.

Outlier removal was performed using a multi-layered approach combining domain knowledge with statistical methods. For the price variable, records with zero or negative prices (32,895 records) were removed as invalid entries. Percentile-based filtering removed values below the 1st percentile ($150) and above the 99th percentile ($68,747.48), eliminating 7,864 records. Additionally, IQR-based filtering using a 1.5×IQR threshold removed 4,734 extreme outliers. Similar procedures were applied to the year and odometer features, with domain-based constraints ensuring year values fell between 1900 and 2025, and odometer readings were positive. Geographic coordinates were validated against continental United States boundaries (latitude: 24.5°N to 49.4°N, longitude: -125°W to -66°W), removing 3,281 records with invalid coordinates.

Missing value imputation employed a strategic multi-tier approach to preserve maximum information. For categorical features with moderate to high missingness (cylinders, condition, drive, paint_color, type, manufacturer, title_status, model, fuel, transmission), missing values were replaced with an '*unknown*' category rather than deletion. This approach retained all records while allowing models to learn patterns associated with missing information, which may itself be informative about vehicle characteristics or seller behavior. Numerical features received targeted imputation: odometer values were imputed using the median grouped by manufacturer and year, leveraging the relationship between vehicle age, brand, and typical mileage. Year values were imputed using the mode grouped by manufacturer and model, utilizing the fact that certain model lines are associated with specific production years.

Following all cleaning procedures, the final dataset retained 348,001 records (81.52% of original data) across 19 features (73.08% of original features), with zero missing values. TABLE I summarizes the data transformation from original to cleaned state. This high retention rate while achieving complete data integrity demonstrates the effectiveness of the strategic cleaning approach.

Metric	Original	Cleaned
Total Rows	426,880	348,001
Total Columns	26	19
Missing Values	1,655,336	0
Row Retention	-	81.52%

TABLE I
DATA CLEANING SUMMARY

C. Data Analysis and Mining

Exploratory Data Analysis. Comprehensive exploratory data analysis was conducted to understand feature distributions, identify patterns, and examine relationships with the target variable. Statistical summaries revealed that the cleaned price distribution exhibited a right-skewed pattern with a mean of $18,077 and standard deviation of $12,595, ranging from $150 to $56,255. The year variable showed a left-skewed distribution concentrated in recent model years (2008-2020), with the mean year being 2012 with a standard deviation of 5.45 years. Odometer readings demonstrated the expected positive correlation with vehicle age.

Analysis of categorical features revealed important patterns in the used car market. The fuel type distribution showed overwhelming dominance of gasoline vehicles (84.6%), with diesel (5.6%), hybrid (1.3%), electric (0.4%), and other fuel types representing similarly small segments. Transmission types were similarly concentrated, with automatic transmissions accounting for 77.9% of listings. Among vehicle conditions, 'good' (31.7%) and 'excellent' (24.4%) were most common, though condition information was missing for 37.2% of listings. Geographic analysis indicated that listings were concentrated in California, Florida, Texas, and New York, reflecting both population density and active vehicle markets in these states.

Figure 1 illustrates the price distribution after cleaning, showing the transformation from the heavily skewed raw data to a significantly lesser skewed distribution suitable for regression modeling. The histogram demonstrates the concentration of prices in the $5,000 to $30,000 range, representing the mainstream used car market.

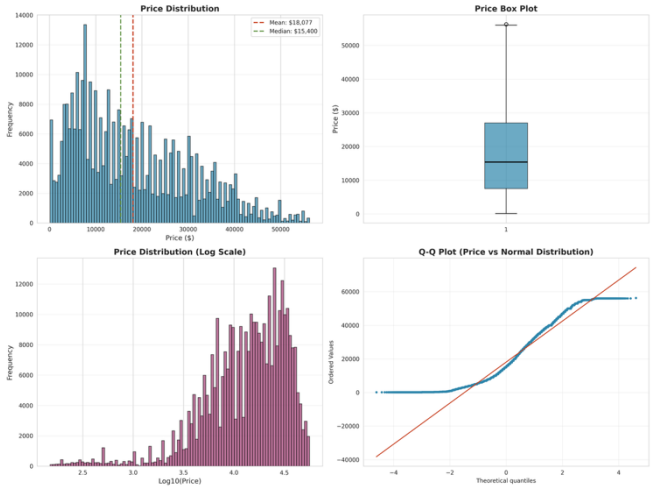

Figure 1: Price Distribution after Cleaning

Feature Engineering and Selection. Feature engineering focused on preparing categorical variables for machine learning algorithms while managing computational complexity. High-cardinality categorical features posed significant challenges: the model feature contained 21,970 unique values, and the region feature had 403 unique categories. Direct one-hot encoding of these features would create an impractically large feature space.

A cardinality reduction strategy was employed, retaining only the top 50 most frequent categories for both model and region features, with all other values mapped to an 'other' category. This approach preserved information about the most common vehicles and regions while dramatically reducing feature dimensionality. Following cardinality reduction, one-hot encoding with *drop_first=True* was applied to all categorical features, resulting in 199 binary features.

The feature set for modeling included year, manufacturer, model (reduced cardinality), condition, cylinders, fuel type, odometer, title status, transmission, drive type, vehicle type, paint color, and region (reduced cardinality). Geographic coordinates (latitude, longitude) were excluded from the final model to focus on vehicle characteristics rather than location-based pricing.

Baseline Models. Two baseline models were developed to establish performance benchmarks: Ridge Regression [2] and Random Forest [3]. Ridge Regression served as a linear baseline, employing L2 regularization with alpha=1.0 to prevent overfitting. This model assumed linear relationships between features and price, providing a simple interpretable baseline.

Random Forest represented a non-linear ensemble baseline. The model was configured with 100 estimators, maximum depth of 20, and minimum samples per leaf of 5, balancing model complexity with computational feasibility.

Due to computational constraints in the Google Colab environment with L4 GPU, the Random Forest model was trained on a stratified random sample of 50,000 records from the training set.

Candidate Models. XGBoost, an advanced gradient boosting framework, was selected as the primary candidate model due to its proven effectiveness in regression tasks and ability to capture complex non-linear relationships [4]. Two XGBoost configurations were evaluated: a base model with manually selected hyperparameters and an optimized model with hyperparameters tuned via RandomizedSearchCV.

The base XGBoost model was configured with 300 estimators, learning rate of 0.05, maximum depth of 8, subsample ratio of 0.8, and column subsample ratio of 0.8. The histogram-based tree method was employed for efficient handling of the large feature space. This configuration provided a strong starting point based on general best practices for gradient boosting.

Hyperparameter optimization was performed using RandomizedSearchCV with 3-fold cross-validation on the 50,000-record training sample. The search space included n_estimators (200, 300, 400), max_depth (6, 8, 10), learning_rate (0.03, 0.05, 0.1), subsample (0.7, 0.8, 1.0), colsample_bytree (0.7, 0.8, 1.0), and min_child_weight (1, 3, 5). Fifteen random combinations were evaluated with negative mean absolute error as the optimization metric.

D. Evaluation and Interpretation

Model performance was evaluated using three complementary metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2). MAE quantifies the average magnitude of prediction errors in dollar terms, providing an interpretable measure of typical prediction accuracy. RMSE penalizes larger errors more heavily through squaring, making it sensitive to outliers and large deviations. R^2 measures the proportion of variance in car prices explained by the model, with values closer to 1 indicating superior predictive power.

All models were trained on 278,400 records (80% of cleaned data) and evaluated on a held-out test set of 69,601 records (20% of cleaned data). The train-test split employed random stratification with a fixed random seed (42) to ensure reproducibility. Model comparison focused primarily on test set performance to assess generalization capability.

Feature importance analysis was conducted using the XGBoost model's built-in gain-based feature importance scores. These scores measure the average improvement in prediction accuracy when a feature is used for splitting nodes across all trees in the ensemble. High importance scores indicate features that consistently contribute to reducing prediction error.

III. RESULTS

TABLE II presents the comprehensive performance comparison across all evaluated models (Ridge Regression, Random Forest, XGBoost with fixed parameters, and XGB with hyperparameter tuning). The results demonstrate a clear progression in predictive accuracy from simple linear models to sophisticated gradient boosting ensembles.

Model	Train MAE	Train RMSE	Train R^2	Test MAE	Test RMSE	Test R^2
Ridge	4957.17	6965.72	0.694	4990.25	7031.31	0.690
RF	2791.53	4492.37	0.872	3542.62	5653.14	0.800
XGB (base)	3122.78	4797.74	0.855	3512.94	5490.96	0.811
XGB (tuned)	1723.99	2684.13	0.954	2979.29	4924.44	0.848

TABLE II
MODEL PERFORMANCE COMPARISON

Ridge Regression established the baseline with a test MAE of $4,990 and R^2 of 0.690, explaining approximately 69% of price variance through linear relationships. The substantial gap between Ridge and ensemble methods confirms that used car pricing involves significant non-linear patterns and feature interactions that linear models cannot capture.

Random Forest demonstrated substantial improvement over Ridge, achieving a test MAE of $3,543 (28.9% reduction) and R^2 of 0.800. This performance gain validates the importance of modeling non-linear relationships and feature interactions in price prediction. The gap between training (MAE: $2,792, R^2: 0.872) and test performance indicates some overfitting, though the model generalizes reasonably well.

XGBoost base configuration achieved competitive performance with test MAE of $3,513 and R^2 of 0.811, slightly outperforming Random Forest. The base XGBoost model demonstrated better balance between training and test performance compared to Random Forest, suggesting superior regularization through its gradient boosting framework.

The tuned XGBoost model achieved the best overall performance with test MAE of $2,979, RMSE of $4,924, and R^2 of 0.848. Hyperparameter optimization reduced test MAE by 15.2% compared to the base XGBoost configuration and 40.3% compared to Ridge Regression. The optimal hyperparameters were: 400 estimators, learning rate of 0.1, maximum depth of 10, subsample ratio of 0.8, column subsample ratio of 1.0, and minimum child weight of 1. These parameters balance model complexity with generalization, achieving an R^2 of 0.848 while maintaining reasonable training-test performance gap.

Feature importance analysis from the tuned XGBoost model revealed that vehicles with 4-cylinder engines, gas fuel type, and fwd (front-wheel drive) drivetrain were the most influential predictors. Figure 2 displays the top 15 features ranked by importance.

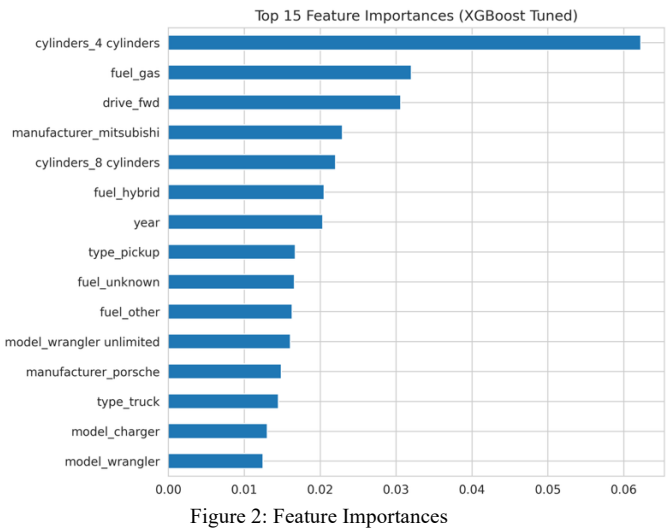

Figure 2: Feature Importances

Figure 3 presents a scatter plot of actual versus predicted prices for the test set using the tuned XGBoost model. The tight clustering of points around the diagonal reference line demonstrates strong prediction accuracy across the price range. The model performs consistently well for vehicles in the $10,000-$40,000 range, which represents the majority of the used car market. Some variance is observed at the extremes, particularly for very low-priced vehicles (under $5,000) and high-end vehicles (over $45,000), where data is sparser and pricing factors may be more idiosyncratic.

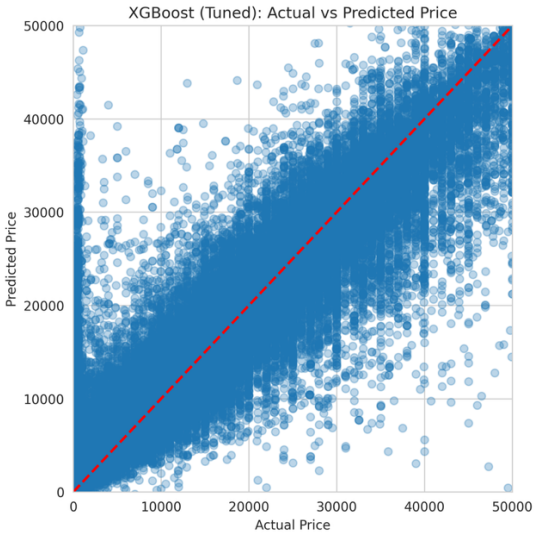

Figure 3: Price Distribution after Cleaning

IV. DISCUSSION AND CONCLUSION

This research successfully developed a high-performance machine learning model for predicting used car prices using large-scale marketplace data. The final tuned XGBoost model achieved a test MAE of $2,979 and R^2 of 0.848, demonstrating that gradient boosting techniques can effectively capture the complex relationships governing used vehicle pricing.

The systematic four-phase methodology proved effective in handling real-world data challenges. The data cleaning pipeline successfully managed extreme outliers and missing values while retaining 81.52% of original records, demonstrating that strategic imputation can preserve information while maintaining data quality. The tiered missing value strategy, which treated missingness as potentially informative rather than purely problematic, represents a pragmatic approach to incomplete real-world datasets.

The progression from Ridge Regression ($R^2 = 0.690$) through Random Forest ($R^2 = 0.800$) to XGBoost ($R^2 = 0.848$) validates the importance of non-linear modeling and ensemble methods for price prediction. The 40.3% reduction in MAE from Ridge to tuned XGBoost quantifies the value of sophisticated machine learning techniques over traditional linear approaches. Feature importance analysis confirmed that vehicle age, mileage, and brand are the primary pricing factors, aligning with domain knowledge about depreciation patterns.

Several limitations should be noted. First, the temporal snapshot nature of the data (April-May 2021) limits generalizability to other time periods, particularly given market volatility in used car prices during and after the COVID-19 pandemic. Second, training Random Forest and XGBoost on a 50,000-record sample rather than the full training set was necessary due to computational constraints, potentially limiting model performance. Third, the exclusion of text descriptions and the reduction of high-cardinality features (model, region) discarded potentially valuable information for prediction accuracy.

Future work could address these limitations through several approaches. Incorporating temporal dynamics by training on multi-period data would enable modeling of seasonal effects and market trends. Leveraging distributed computing frameworks such as Apache Spark would enable training on the full dataset without sampling, likely improving model performance. Natural language processing of vehicle descriptions could extract additional features related to vehicle condition, modifications, and seller credibility. Alternative advanced techniques such as neural networks or stacking ensembles may yield further performance improvements.

In conclusion, this research demonstrates that machine learning, particularly gradient boosting methods, can achieve accurate price predictions for used vehicles. The developed XGBoost model provides a foundation for practical applications including automated vehicle valuation tools, fair pricing recommendations for buyers and sellers, and market analysis for dealerships and financial institutions. The methodology and findings contribute to the growing body of work applying data mining techniques to automotive market prediction.

ACKNOWLEDGMENT

The dataset used in this study was created by Austin Reese and obtained from Kaggle, consisting of used car and truck listings scraped across the United States on Craigslist [1].

All phases of this study have been performed and executed in the Google Colab IDE using an L4 GPU at runtime. The L4 GPU used for this project is a freely available computational resource on Google Colab, usable through their Student Subscription plan.

REFERENCES

[1] A. Reese, "Used Cars Dataset," *www.kaggle.com*. https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data

[2] "sklearn.linear_model.Ridge — scikit-learn 0.23.2 documentation," *scikit-learn.org*. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html

[3] scikit-learn, "3.2.4.3.2. sklearn.ensemble.RandomForestRegressor — scikit-learn 0.20.3 documentation," *Scikit-learn.org*, 2018. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[4] "XGBoost Parameters — xgboost 1.2.0-SNAPSHOT documentation," *xgboost.readthedocs.io*. https://xgboost.readthedocs.io/en/latest/parameter.html