Predictive Modeling for Vehicle Selling Price Based on Market Trends

Sanjana Peruku

Harrisburg Univerisity of Science and Technology

Predictive Modeling for Vehicle Selling Price Based on Market Trends

Introduction, Motivation, and Problem Description

The automobile market is significantly affected by fluctuating market prices depending on the vehicle's condition, mileage, make, year, and more. Accurate estimation of the prices is crucial for dealers, buyers, and sellers to make effective decisions and get the most out of their transactions. Traditional methods are often subject to subjective opinions or past data, leading to inconsistencies and inaccuracies. The goal of the project is to develop a successful machine learning model capable of predicting the sale price of the vehicles in the Vehicle Sales and Market Trends Dataset based on historical sales data. Machine learning algorithms offer a data-based approach with the potential to recognize complex patterns and correlations in vehicle prices, which equates to improved accuracy and consistency in price prediction.

Related Work

Vehicle price prediction has been a widely researched area, particularly with the increased availability of large transactional databases. Industry standards like Kelley Blue Book (KBB) and Manheim Market Report (MMR) offer vehicle prices based on statistical models and real-time data for estimating prices. Academic literature has tried to utilize linear regression, decision trees, and ensemble models like XGBoost for predicting vehicle prices. All of these works indicate mileage, condition, make, and market trends as leading predictors. Further, feature engineering such as feature extraction related to date or transformation of categorical variables has been found to enhance model performance.

Dataset Description

Our dataset comprises 1,66,168 records, each representing a distinct vehicle transaction within the automotive market. The dataset offers a comprehensive view of various attributes related to vehicle details, sales transactions, and market dynamics. Key attributes include the vehicle's manufacturing year, make, model, and trim, providing insight into the specific characteristics of each vehicle. Additionally, details such as body type, transmission type, and vehicle identification number (VIN) offer further granularity in understanding vehicle specifications. The dataset also includes information on the vehicle's condition, odometer reading, exterior and interior colors, and the state where it is registered or located. Moreover, it provides data on the seller or entity selling the vehicle, market reference prices, actual selling prices, and sale dates. Overall, this dataset offers a comprehensive overview of the automotive market, facilitating analysis of pricing trends, market dynamics, and consumer behavior within the industry. The link for the dataset is https://www.kaggle.com/datasets/syedanwarafridi/vehicle-sales-data/data

Data Preprocessing and Insights

Proper data preprocessing is important in ensuring the reliability and quality of machine learning models. Preprocessing was necessary for the dataset utilized in this research to remove missing values, outliers, and categorical variables. The following procedures were performed:

- **Missing Values:** Numerical columns such as *Odometer* and *MMR* had missing values, which were filled using the median. This procedure was used to limit the effect of possible outliers on the data.

- **Outliers:** The presence of significantly high values was observed in the Odometer readings. The Interquartile Range (IQR) approach was utilized to detect and eliminate such outliers, maintain data quality, and minimize the impact of anomalous values on the model.

- **Categorical Encoding:**

  Categorical attributes like *Make*, *Model*, *Trim*, *Transmission*, and *Color* were transformed into numeric values using *Label Encoding* to make them compatible with machine learning algorithms.

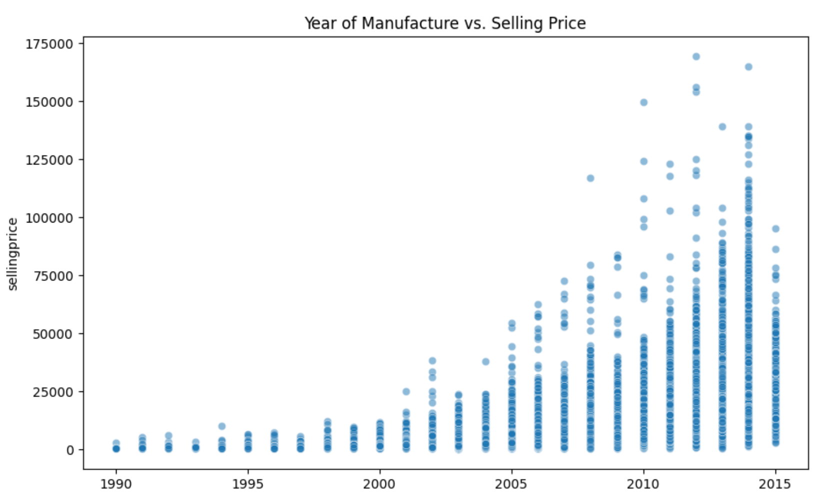
- **Condition Scaling:** The *Condition* field, which is the vehicle's condition rating, was normalized by *StandardScaler* in order to obtain zero mean and unit variance. Furthermore, *MinMaxScaler* was utilized to rescale the values into the interval between 0 and 1 for normalization purposes.

- **Date Cleaning:**The column of *Sale Date* included unstructured data. The respective a portion of the date was pulled out, and the column was cleaned and reshaped into a standardized date format for consistency.

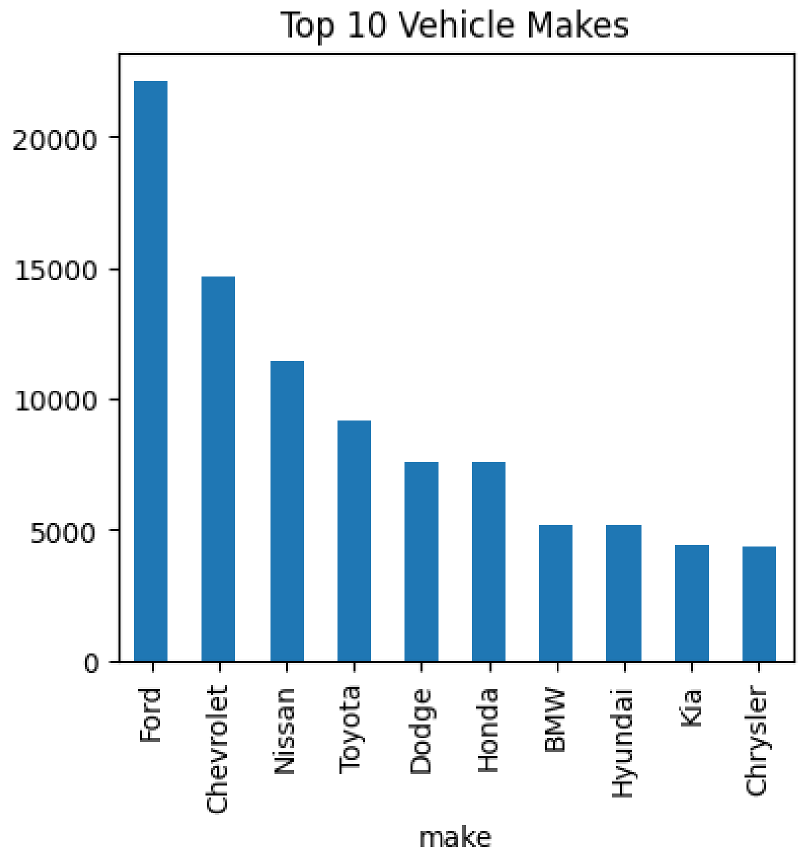
Exploratory Data Analysis (EDA)

Exploratory Data Analysis has been conducted in order to make sense of the underlying patterns and relationships in the data. Most significant results of the visual inspection are presented below

- **Year vs. Selling Price Relationship:** A non-linear relationship was discovered between *Year of Manufacture* and *Selling Price*.



- **Top Vehicle Makes and Sellers:** Bar plots highlighted the top vehicle sellers and manufacturers.



Model Selection and Training

Baseline Algorithms and Model Families

Three types of models were applied to both fit linear and non-linear relationships within the data. Each type was started with a baseline algorithm and then increasingly complex versions were adjusted to enhance performance.

- **Linear Models:**
  - **Linear Regression (Baseline):** A basic model that hypothesizes a linear correlation between features and the target variable (selling price).
- **Tree-Based Models:**
  - **Decision Tree Regressor (Baseline):** A non-linear model that divides data according to feature values to create a tree structure, thus being efficient in

detecting intricate patterns.

- **Boosting Models:**
  - **XGBoost Regressor (Baseline):** A sophisticated boosting algorithm that constructs trees in a sequential manner to rectify the mistakes of preceding trees, maximizing prediction accuracy.

Hyperparameter Tuning Using Optuna

In order to enhance the performance of the baseline models, *Optuna* was used to automate hyperparameter tuning. Optuna is an optimization platform that systematically searches for optimal hyperparameters by trial and error.

Two versions of each of the three learning approaches were created with Optuna-optimized hyperparameters, six optimized models in total:

Model	**Brief Description**
Ridge Regression	A linear model with L2 regularization to reduce overfitting by penalizing large coefficients.
Lasso Regression	A linear model with L1 regularization, which can also reduce coefficients to zero for feature selection.
RandomForest Regressor	An ensemble of multiple decision trees that averages predictions to reduce overfitting and improve stability.
LightGBM Regressor	A gradient boosting algorithm optimized for speed and efficiency, suitable for large datasets.
XGBoost Regressor	An optimized gradient boosting algorithm that handles missing data and improves computational performance.
CatBoost Regressor	A gradient boosting algorithm designed to handle categorical features efficiently without heavy preprocessing.

Ensemble Learning

After single model performance, a Voting Regressor was implemented to blend the forecasts of Ridge, RandomForest, and XGBoost. This multi-model ensemble strategy tried to capitalize on the strengths of every model while compensating for the weakness of any one algorithm.

Test and Evaluation

- **Testing Approach:** The data was divided into training (90%) and test (10%) sets to make sure that the models were tested on unseen data, replicating real-world prediction contexts. Random seeds were fixed to make sure multiple training runs are consistent.

- **Validity of the Approach:** The adopted strategy is legitimate since train-test splitting is standard practice for regression problems. Evaluation across several model families (tree-based, linear, boosting) also supported the fact that gains in performance were stable across methods.

- **Evaluation Metrics:** Model performance was evaluated using:

  - Mean Squared Error (MSE): Primary metric to penalize large prediction errors.
  - Mean Absolute Error (MAE): Provides an intuitive measure of average error in dollars.
  - R-squared (R^2): Indicates how well the model explains variance in selling price.

  Performance analysis included output comparison of predicted vs. actual prices for baseline, tuned, and ensemble models.

Model	**MSE**	**MAE**	**R^2**
Linear Regression	2,640,062.06	1,045.68	0.9698
Decision Tree Regressor	4,425,812.91	1,366.62	0.9494
XGBoost Regressor	2,370,479.43	912.13	0.9729
Ridge Regression (Tuned)	2,640,061.61	1,045.68	0.9698
Lasso Regression (Tuned)	2,640,061.68	1,045.67	0.9698
RandomForest Regressor (Tuned)	2,190,170.50	941.12	0.9749
LightGBM Regressor (Tuned)	2,325,123.06	906.31	0.9734
CatBoost Regressor (Tuned)	3,788,325.52	1,053.30	0.9567
Ensemble Model	**2,175,224.14**	**937.40**	**0.9751**

Performance analysis of all models is combined in above Table.

The baseline models (*Linear Regression*, *Decision Tree Regressor*, and *XGBoost Regressor*) provided the first insight into the data, with *Linear Regression* providing a very high R^2 of 0.9698. However, tree-based and boosting models, such as *XGBoost* and *RandomForest*, performed better than the linear models in terms of lower Mean Squared Error (MSE) and higher R^2.

Hyperparameter optimization with *Optuna* subsequently optimized the performance of the *RandomForest* and *LightGBM* models, which suggests fine-tuning tree-based models maximizes predictive performance.

The best generalization overall was the final Ensemble Model by voting from the combinations of *Ridge*, *RandomForest*, and *XGBoost*, at the lowest MSE of 2,175,224, lowest MAE at 937.40, and highest R^2 at 0.9751.

This verifies that the integration of heterogeneous models exploits their strengths and enhances predictive accuracy, which is the best solution for vehicle price prediction in this project.

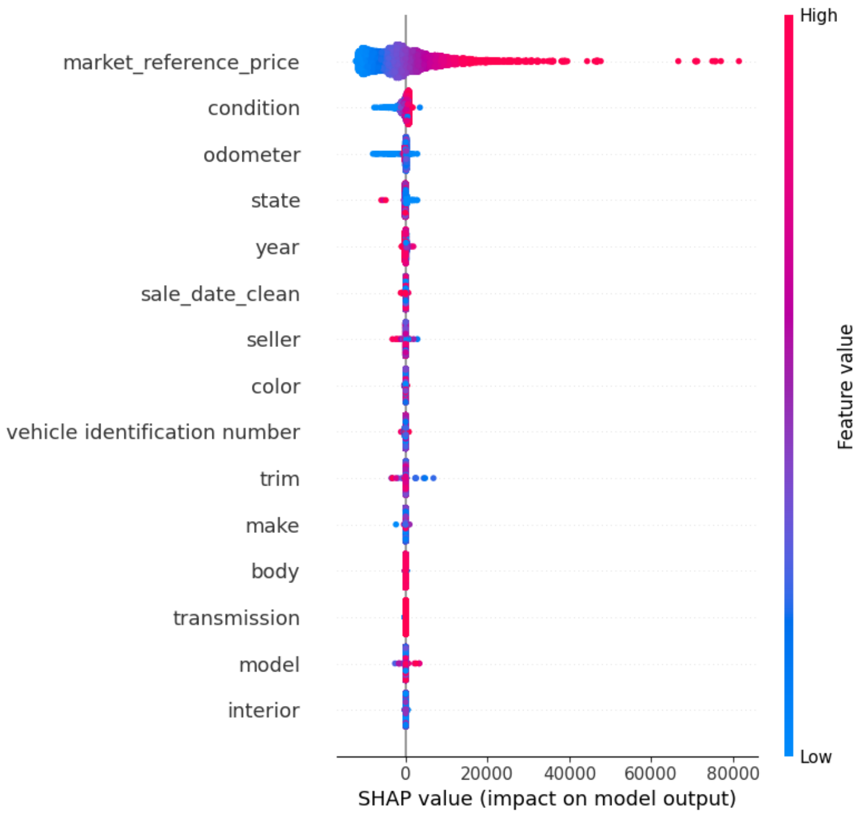
Performance Comparison

- **Linear Regression:** Baseline performance, limited by linear assumptions.

- **Decision Tree:** Increased accuracy but susceptible to overfitting.
- **XGBoost:** Single best model performance, dealing well with non-linearity.
- **Ensemble Model:** The best overall performance with the least MSE and best R^2 value.

Shap Analysis

SHAP analysis identified condition, odometer, and MMR as the most significant



predictors.

Validity and Reproducibility

Validity of the method was guaranteed through robust testing on unseen test data, enabling performance of the models to be assessed in a realistic prediction environment. Evaluation metrics, including *Mean Squared Error (MSE)*, *Mean Absolute Error (MAE)*, and R^2, demonstrated **consistency across multiple runs**, confirming the stability and reliability of the modeling process. To achieve reproducibility, **random seeds were fixed uniformly** during model training and data splitting, removing the randomness introduced by random initialization. In addition, **extensive documentation and code comments** were kept throughout the project, allowing future researchers to **reproduce the workflow and results precisely**.

Potential Future Work

While the current model has achieved robust performance, the following can be explored to further enhance predictive accuracy and robustness:

- **Addition of External Variables:** Addition of additional variables such as *regional demand*, *fuel prices*, and *economic conditions* could offer a better understanding of the vehicle price drivers involved.
- **Deep Learning Models:** Exploration of *neural networks* and other *deep learning architectures* has the potential to help capture **highly complex, non-linear relationships** not handled by standard models.
- **Advanced Feature Engineering:** Further improvement can be made by *building interaction terms* between key features, *including polynomial features*, or *extracting time-based features* from *sale dates* to better pick up on seasonality or trends.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2009.

[2] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, San Francisco, CA, USA, 2016, pp. 785-794.

[4] Manheim, "Market Pricing Trends and Vehicle Valuation," [Online]. Available: https://www.manheim.com. [Accessed: 18-Feb-2025].

[5] Kelley Blue Book, "Vehicle Valuation and Market Analysis," [Online]. Available: https://www.kbb.com. [Accessed: 18-Feb-2025].

[6] S. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf. [Accessed: 18-Feb-2025].

[7] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '19)*, Anchorage, AK, USA, 2019, pp. 2623–2631. [Online]. Available: https://doi.org/10.1145/3292500.3330701. [Accessed: 18-Feb-2025].