IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Cars Price Prediction Using Machine Learning

R.Azhagusundaram[1],
"Department of Computer Science and
"Engineering, Bharath Institute of Higher
"Education and Research, "Chennai, India ".

P. Harsha vardhan[2],
Department of Computer Science and
Engineering,Bharat Institute of Higher
Education and research,Chennai,India. .

R. Sai Krishna[3], Department
of Computer Science and
Engineering,Bharat Institute of Higher
Education and research,Chennai,India
.

R. Lokesh[4],
Department of Computer Science and
Engineering,Bharat Institute of Higher
Education and research,
Chennai,India.

P. Simhadri[5]
Department of Computer Science
and"Engineering,Bharat Institute of
Higher"Education and
research,Chennai,India".

Abstract : In the contemporary automotive sector, the precise evaluation of used cars' resale value stands as a pivotal task for diverse stakeholders, encompassing car dealerships, insurance entities, and individual car owners. Conventional methods of car valuation often hinge on manual data analysis and subjective judgment, fostering inconsistencies and inaccuracies. To surmount these challenges, the presented project introduces CarValueML: an AI-driven car price estimation system bolstered by Gradio. CarValueML leverages machine learning algorithms to scrutinize an extensive dataset of car attributes, including make, model, year, mileage, condition, and features. The system employs varied data preprocessing techniques to meticulously cleanse and prepare the data for both training and inference. Subsequently, sophisticated machine learning models, such as random forests and gradient boosting trees, undergo training on the refined data, yielding predictions of car prices with remarkable precision. Gradio, an accessible web framework, serves as the bedrock for CarValueML's user interface. Facilitating the creation of interactive and user-friendly interfaces for machine learning models, Gradio extends CarValueML's accessibility to a broad spectrum of users, encompassing those without technical expertise. The Gradio interface enables users to input car specifications and receive instantaneous price estimations, concurrently providing insights into the factors influencing car prices, thereby augmenting user comprehension of the valuation process. CarValueML outshines traditional car valuation methods on multiple fronts. Firstly, it attains superior accuracy compared to manual estimation techniques, mitigating the risks associated with overvaluation or undervaluation. Secondly, its consistent and objective approach eradicates the inherent subjectivity present in manual valuations. Thirdly, its capability to efficiently handle large volumes of car data renders it suitable for extensive applications within the automotive industry. Fourthly, Gradio's intuitive interface broadens its accessibility, extending its utility to a more extensive user base, including those lacking technical acumen. Lastly, it furnishes insights into the factors influencing car prices, thereby enriching user comprehension of the valuation process. In conclusion, CarValueML presents an innovative and efficacious methodology for car price estimation, seamlessly integrating AI and Gradio to provide accurate, consistent, and user-friendly valuations. The system harbors substantial potential for diverse applications within the automotive industry.

Keywords : Artificial Intelligence, Gradio, Machine Learning, Mean Squared Error, Root Mean Error .

I. INTRODUCTION

A.Introduction of the study

The prediction of used car prices holds paramount importance within the automotive industry, exerting considerable influence on facets like car sales, insurance pricing, and individual car ownership. Traditional methodologies for forecasting used car prices have hitherto relied on manual data analysis and subjective judgment, resulting in inconsistencies and inaccuracies. However, the integration of machine learning (ML) has ushered in a revolutionary paradigm shift, offering a more precise and standardized approach.

ML algorithms, capable of scrutinizing extensive datasets, excel at identifying intricate patterns and relationships that elude manual detection. This enables them to furnish more astute predictions regarding used car prices, encompassing a broader spectrum of factors than traditional approaches. Furthermore, ML models exhibit the adaptability to continuously assimilate new data, ensuring the enduring accuracy of their predictions over time.

CarValueML, a cutting-edge AI-powered car price estimation system harnessed by Gradio, exemplifies this evolution. Leveraging advanced ML algorithms, CarValueML employs a user-friendly interface to deliver accurate and consistent valuations of used cars. The system's comprehensive dataset, encompassing attributes like make, model, year, mileage, condition, and features, empowers it to meticulously assess the determinants of car prices. Gradio's intuitive interface facilitates real-time price estimations, making the process accessible to a diverse audience, including non-technical users. CarValueML's prowess in efficiently handling substantial volumes of car data positions it aptly for large-scale applications, and its transparent insights into price-influencing factors augment user comprehension of the valuation process.

B.Problem Statement

The precision of used car price prediction stands as a critical imperative in the automotive sector, wielding influence over car sales, insurance pricing, and individual ownership. Conventional methods for predicting used car prices, reliant on manual analysis and subjective judgment, foster inconsistencies and inaccuracies. This predisposes used cars to potential overvaluation or undervaluation, incurring financial losses for dealerships, insurers, and car owners. Moreover, traditional methods often falter in capturing the intricate relationships between car attributes and their impact on market prices.

The imperative for a more accurate and uniform prediction methodology has spurred the exploration of machine learning (ML) techniques. ML algorithms possess the capability to analyze extensive historical car sales data, discerning patterns and relationships that elude manual identification. Through harnessing these patterns, ML models can deliver more informed predictions, incorporating a wider array of factors than conventional methods. Crafting and implementing effective ML models for used car price prediction, however, presents myriad challenges.

Firstly, gathering comprehensive and high-quality historical car sales data is pivotal, necessitating information on attributes like make, model, year, mileage, condition, market location, and sale prices. This data must undergo meticulous cleaning and preprocessing to ensure its compatibility with ML algorithms. Secondly, feature engineering, the transformation of raw data into meaningful features, requires deliberate consideration to effectively represent the complex relationships between car attributes and market prices. Thirdly, the selection of an appropriate ML algorithm and optimization of its parameters are critical for accurate predictions, considering the distinct strengths and weaknesses of various algorithms and their dependence on data characteristics. Finally, rigorous evaluation and validation of the trained ML model are indispensable to ensure its applicability and reliability in real-world scenarios, encompassing testing its performance on unseen data and assessing its predictive accuracy under diverse market conditions.

C.Research Objective

The primary goals of our project encompass:

Analyzing the efficacy of diverse machine learning algorithms for used car price prediction:

i. Identifying and evaluating a range of ML algorithms suitable for used car price prediction, including linear regression, random forests, gradient boosting trees, and neural networks.

ii. Comparing the performance of different algorithms using metrics such as mean squared error (MSE) and root mean error (RMSE).

iii. Investigating the impact of hyperparameter tuning on the performance of each algorithm.

Developing a robust feature engineering pipeline for used car price prediction:

i. Identifying the relevant car attributes influencing used car prices, including make, model, year, mileage, condition, and features.

ii. Implementing data cleaning and preprocessing techniques to handle missing values, outliers, and data inconsistencies.

iii. Exploring diverse feature engineering techniques, such as feature scaling and selection, to enhance the representation of car attributes and improve model performance.

Investigating the impact of data quality and quantity on machine learning model performance:

Evaluating the sensitivity of different machine learning algorithms to variations in data quality and quantity.

II. LITERATURE SURVEY

Review on Previous Techniques:

In the exploration of predictive models for used car prices, various studies have employed diverse machine learning techniques, each contributing valuable insights. Monburinon et al. (2018) conducted a study on a German e-commerce site, employing gradient boosted regression tree and multiple linear regression. The pinnacle of accuracy, marked by a mean absolute error (MAE) of 0.28, was achieved with the gradient boosted regression tree, underscoring the significance of meticulous parameter adjustments and categorical data encoding.

Gegic et al. (2019) focused on predicting used car prices in Bosnia, employing Support Vector Machine, Random Forest, and Artificial Neural Network. Their innovative approach involved pre-calcification of prices using Random Forest, resulting in commendable accuracies reaching up to 87.38%.

Noor and Jan (2017) set an impressive benchmark with a 98% accuracy utilizing Multiple Linear Regression on data from a Pakistani used cars website. Their success was attributed to effective variable selection, emphasizing the importance of attribute relevance in simplifying model complexity.

K.Samruddhi and Kumar (2020) proposed a supervised machine learning model using K-Nearest Neighbors, achieving an 85% accuracy on Kaggle's dataset. The study delved into cross-validation and parameter adjustments to enhance predictive performance.

Other studies explored techniques such as Artificial Neural Network (Gongqi, Yansong, & Qiang, 2011), Support Vector Machines (Listiani, 2009), and Multivariate Regression (Kuiper, 2008), each addressing specific challenges and showcasing varying levels of accuracy.

Nabarun Pal (2018) utilized Random Forest to predict used car prices from Kaggle's dataset, achieving a notable training accuracy of 95.82% and testing accuracy of 83.63%. Feature selection played a pivotal role in augmenting prediction accuracy.

Jian Da Wu (2017) introduced a comprehensive system for used car price prediction, comparing conventional artificial neural networks (ANN) with adaptive neuro-fuzzy inference systems (ANFIS). ANFIS, incorporating fuzzy logic and adaptive neural network capabilities, outperformed in accurately predicting prices.

Review on Pre-Processing:

Supervised machine learning's effectiveness hinges on meticulous preprocessing techniques, crucial for optimal model performance. Feature scaling, highlighted by James et al. (2013), ensures uniform contribution of features to model training, enhancing the convergence speed of algorithms like Support Vector Machines and k-Nearest Neighbors.

Class imbalance, a common challenge in binary classification, was addressed by Chawla et al. (2002) and Sun et al. (2009) through oversampling, undersampling, and techniques like SMOTE, promoting better generalization and model robustness.

Dealing with missing data, as explored by Little and Rubin (2019) and Schafer (1997), involves imputation methods such as mean imputation, regression imputation, and advanced techniques like k-NN imputation, maintaining data integrity and completeness.

Effective encoding of categorical variables, crucial for machine learning, is detailed by Chen and Guestrin (2016) and Greenwell (2017). Techniques like one-hot encoding, label encoding, and target encoding enhance the interpretability of categorical features, facilitating their integration into supervised models.

For high-dimensional datasets, Jolliffe (2002) and van der Maaten and Hinton (2008) propose dimensionality reduction techniques such as Principal Component Analysis, t-Distributed Stochastic Neighbor Embedding, and Linear Discriminant Analysis, improving computational efficiency and model interpretability.

III. RELATED WORKS

Existing System:

In the realm of car price estimation, historical practices predominantly relied on conventional methodologies before the emergence of advanced machine learning algorithms. Conventional approaches encompassed statistical and rule-based methods, with popular choices including linear regression and decision trees. These techniques aimed to establish a linear correlation between diverse features and the target variable—namely, the car price. Often, manual rules were integrated to accommodate specific scenarios.

However, these traditional methodologies faced challenges when confronted with the intricacies of the automotive market. Linear regression assumes a linear relationship between input features and output, potentially overlooking the nonlinear dependencies inherent in car price determination. Decision trees, while widely used, may struggle to capture complex interactions and be susceptible to overfitting, particularly with diverse datasets.

A significant drawback of traditional methods lies in their inadequate handling of categorical variables and their inability to adapt to evolving market dynamics. The simplistic nature of these approaches impedes their ability to discern nuanced patterns and evolving trends, resulting in suboptimal performance in car price estimation.

Disadvantages:

The utilization of the CatBoost algorithm in the AI-driven car price prediction system, CARVALUEML, harnesses one of the most efficient gradient boosting libraries. This library excels at managing categorical features with minimal preprocessing. However, despite its advantages in terms of speed, accuracy, and handling categorical data seamlessly, specific challenges are associated with its application in car price prediction.

Model Complexity and Interpretability: CatBoost, akin to other machine learning models, can become intricate as it incorporates decision trees to enhance prediction accuracy. This complexity may render the model a "black box," making it challenging to interpret the reasoning behind specific predictions, thereby hindering user understanding.

Overfitting Risk: Despite CatBoost's mechanisms to mitigate overfitting, the risk persists, especially when the model is trained on non-representative datasets or excessively tuned for training data performance. Overfitting can lead to inaccurate price predictions on unseen data.

Data Quality and Representation: The performance of CatBoost is contingent on the quality and representativeness of the data it is trained on.

Proposed System:

The CarValueML system introduces a paradigm shift in car price estimation by leveraging advanced machine learning techniques to overcome the limitations of traditional methodologies. Drawing on insights from traditional methods, the project aims to enhance accuracy, adaptability, and robustness in predicting car prices.

CarValueML incorporates state-of-the-art machine learning algorithms, with CatBoost emerging as the algorithm of choice due to its seamless handling of categorical features. This leads to improved model performance and the ability to capture complex relationships within the data.

The system employs advanced feature engineering and preprocessing techniques to optimize model performance. One-hot encoding ensures effective representation of categorical variables, while standard scaling normalizes numerical features, contributing to a stable and efficient training process.

Unlike traditional methodologies with fixed rules, the machine learning models in CarValueML exhibit a higher degree of flexibility and adaptability. They can learn from data, capturing intricate patterns and adjusting to evolving market dynamics—a crucial feature in the ever-changing automotive market.

The integration of advanced machine learning algorithms, coupled with robust preprocessing techniques, aims to enhance the accuracy of car price predictions. The models are designed not only to fit the training data effectively but also to generalize well to unseen data, ensuring reliable estimations for a diverse range of cars and market conditions.

CarValueML is complemented by a user-friendly interface built using Gradio, simplifying the deployment of machine learning models and ensuring accessibility and usability for a wider audience.

Advantages:

Incorporating the CatBoost algorithm into the proposed system, CARVALUEML, for car price prediction offers numerous advantages due to CatBoost's unique features and capabilities.

Superior Handling of Categorical Features

Direct Processing: CatBoost can handle categorical variables directly, eliminating the need for preprocessing into numerical values.

Reduced Preprocessing Time: By avoiding extensive preprocessing of categorical data, CatBoost shortens data preparation, enabling quicker iterations and model development.

Robustness to Overfitting

Ordered Boosting: CatBoost employs ordered boosting and other techniques to combat overfitting, enhancing reliability and ensuring the model generalizes well to unseen data.

High Performance and Efficiency

Speed: CatBoost is optimized for speed without compromising accuracy, making it suitable for scenarios where fast model training and prediction are crucial.

Scalability: It efficiently handles large datasets, crucial for the extensive data involved in car price prediction.

Accuracy

Advanced Algorithms: CatBoost's advanced algorithms, particularly in handling categorical and numerical data, often result in higher accuracy levels compared to other machine learning models.

Ease of Use

Accessibility: CatBoost is relatively straightforward to implement and integrate into existing systems, making it accessible to developers and data scientists of varying expertise levels.

IV. METHODOLOGY

Modules in CarValueML:

1. Dependencies

2. Data Ingestion Module

3. Data Transformation Module

4. Model Trainer

5. Training Pipeline

6. Prediction Pipeline

7. Gradio GUI

8. Github For Version Control

9. Notebook

Pre-processing steps:

Effective pre-processing is paramount for the success of machine learning models, with the CarValueML project exemplifying a meticulous approach to enhance input data quality. A key focus is the dissection of registration numbers, emphasizing the extraction and isolation of state information. This not only converts raw data into a more meaningful representation but also enables the model to discern regional nuances in car pricing, fostering a more nuanced and accurate estimation.

An additional crucial feature engineering step involves breaking down car names into two components: Brand Trustness (Brand Name) and Model. This strategic division facilitates a deeper dataset understanding by isolating essential brand and model information. More than an organizational tactic, this decomposition represents a deliberate effort to unveil underlying data patterns, empowering the model to capture nuanced relationships between brand, model, and pricing. This granularity aims to enrich the dataset, providing the model with heightened discriminative power.

Beyond feature engineering, the preprocessing pipeline delves into categorical features, where one-hot encoding emerges as a pivotal step. This transformation converts categorical variables into a binary matrix, assigning a unique binary representation to each category. Essential for effective interpretation of categorical data, one-hot encoding ensures states, brand names, and car models are presented in a machine-learning-friendly format, enhancing overall predictive accuracy.

Simultaneously, numerical feature scaling is integral to the preprocessing steps. Scaling guarantees numerical features share a similar scale, preventing certain features from dominating due to inherent differences in magnitude. This is particularly crucial for algorithms reliant on distance metrics, ensuring each feature contributes proportionally to the model's learning process. The incorporation of feature scaling underscores a commitment to data integrity and algorithmic robustness, bolstering the model's stability and performance.

Project phases:

Requirement Analysis:

The initial phase revolves around the meticulous collection of system requirements. This critical step encompasses the generation of comprehensive documents and thorough requirement reviews. The primary goal is to establish a clear understanding of the project's prerequisites.

System Design:

With the gathered requirements as a foundation, the system specifications undergo a transformation into a tangible software representation. This phase places a significant emphasis on refining algorithms, defining data structures, and structuring the software architecture to align with the specified requirements.

Coding:

As the design takes shape, the coding phase commences, where programmers meticulously craft the code to provide a comprehensive blueprint of the product. Here, system specifications are translated into machine-readable code, laying the groundwork for the subsequent implementation phase.

Implementation:

The implementation phase involves the actual coding and programming of the software. The output at this stage typically includes libraries, executables, and comprehensive software documentation. This phase marks a crucial step in bringing the envisioned product to life.

Testing:

In this integral phase, all individual programs or models are integrated and rigorously tested to ensure the overall system aligns with the specified software requirements. Testing efforts are dedicated to both verification and validation, ensuring the software's reliability and functionality.

Maintenance And Deployment:

The maintenance phase, the longest in the development process, focuses on updating the software to address evolving customer needs, adapt to changes in the external environment, rectify errors and oversights not identified during testing, and enhance overall software efficiency. Deployment activities are also carried out during this phase, facilitating the transition from development to practical application.

The analysis of preprocessing techniques implemented in the CarValueML project underscores a strategic and impactful approach to augmenting the dataset's quality and interpretability, thereby bolstering the machine learning model's performance.

Fig.1 Architecture diagram flow

Registration Number Segmentation:

The decision to segment registration numbers to isolate state information demonstrates a judicious choice. This not only simplifies the representation of geographical factors but also empowers the model to discern regional variations in car prices. This approach aligns seamlessly with the project's goal of delivering nuanced estimations that mirror diverse market conditions.

Car Name Deconstruction:

Deconstructing car names into Brand Trustworthiness (Brand Name) and Car Model represents a well-thought-out feature engineering endeavor. This granularity allows the model to differentiate between various brands and models, capturing pricing variations linked to specific car attributes. It enriches the dataset, furnishing the model with more discerning features.

One-Hot Encoding for Categorical Features:

The use of one-hot encoding for categorical features like states, brand names, and car models reflects a prudent decision. This method ensures the model can interpret and utilize categorical data effectively, eliminating the need for cumbersome preprocessing steps. It streamlines the representation of categorical variables, thereby enhancing the model's overall efficiency.

Scaling of Numerical Features:

Scaling numerical features emerges as a pivotal preprocessing step, ensuring that all features contribute proportionally to the model's learning process. By normalizing numerical features to a similar scale, the risk of certain features dominating others due to differences in magnitude is mitigated. This step is indispensable for algorithms sensitive to variations in feature magnitudes.

V. RESULTS AND DISCUSSIONS

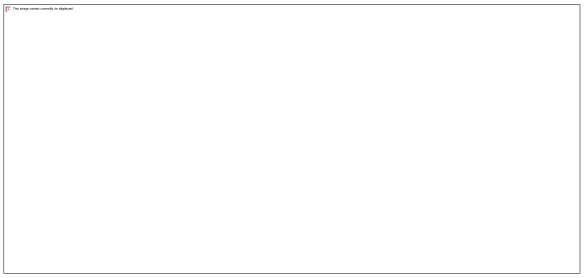

Fig.2 Univariate Analysis

Fig.3 Bi variate Analysis

The CarValueML project takes a cutting-edge approach to estimating used car prices through the implementation of the powerful CatBoost algorithm in machine learning. With a focus on enhancing user experience, the algorithm's proficiency in handling categorical features is highlighted. Employing preprocessing techniques such as one-hot encoding and standard scaling ensures the model's adaptability to diverse datasets, making it robust for various features.

Integrating Gradio for the user interface introduces an interactive dimension to CarValueML, aligning with the goal of broad accessibility. Gradio's deployment streamlining facilitates user engagement through a seamless web interface. Code modularity and clarity are evident in well-organized components like the Predict Pipeline and CustomException class, emphasizing maintainability and scalability. GitHub's use for version control underscores transparency and collaborative project management.

The experimentation and analysis phase played a crucial role in refining the model's performance. The team's exploration of diverse algorithms and preprocessing techniques, as evidenced in the literature survey, demonstrates a thoughtful and informed approach, providing valuable insights for ongoing enhancements.

VI. CONCLUSION AND FUTURE WORKS

In conclusion, the CarValueML project marks a significant advancement in used car price estimation, integrating advanced machine learning techniques, user-centric design, and robust testing methodologies. The project's primary aim is to offer users an accurate and intuitive platform for evaluating their vehicle's market value. The strategic adoption of the CatBoost algorithm showcases its prowess in handling categorical features, ensuring superior predictive performance. Comprehensive preprocessing techniques, including one-hot encoding and standard scaling, enhance the model's adaptability across diverse datasets.

The integration of Gradio for the user interface improves accessibility and engagement, aligning with the project's goal of widespread application. The literature survey provides valuable insights, informing decision-making and contributing to model refinement. Rigorous testing methodologies attest to the dedication to delivering a reliable and error-free application.

Prioritizing code modularity and clarity promotes collaboration, maintainability, and scalability, with GitHub enhancing transparency in version control. As CarValueML evolves, continuous refinement, user feedback, and exploration of new features will contribute to its growth and effectiveness. CarValueML not only offers a sophisticated solution for used car price estimation but also exemplifies collaborative and iterative development, positioning it as a valuable tool in automotive valuation.

Future Works

Feature Engineering and Data Enrichment:

Exploring additional features and data sources for a more nuanced understanding of car prices.

Dynamic Model Updating:

Implementing a mechanism for regular model updates to adapt to evolving market conditions and user preferences.

User Feedback Integration:

Establishing a feedback loop for continuous improvement based on real-world user experiences.

Integration of External Datasets:

Accessing external datasets to provide additional context and refine the accuracy of price estimations.

Interpretability and Explainability:

Enhancing the interpretability of model predictions to build user trust and understanding.

Localized Market Models:

Tailoring the model to specific geographic regions for more accurate estimations.

Advanced User Interface Features:

Enriching the user interface with advanced features for a comprehensive understanding of factors influencing car values.

Collaboration with Automotive Industry:

Establishing partnerships for valuable insights, industry expertise, and potential data enrichment.

REFERENCES

[1] Python Documentation - https://docs.python.org/3/

[2] PyCharm - https://www.jetbrains.com/pycharm/

[3]Numpy Documentation - https://numpy.org/doc/stable/

[4]Pandas Documentation - https://pandas.pydata.org/docs/

[5] Hunter, J. D. (2007). MatPlotLib: a 2D Graphics environment. Computing in Science and Engineering, 9(3), 90–95. https://doi.org/10.1109/mcse.2007.55

[6] Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. arXiv (Cornell University). https://arxiv.org/pdf/1810.11363.pdf

[7] Torlay, L., Perrone-Bertolotti, M., Thomas, E., & Baciu, M. (2017b). Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. Brain Informatics, 4(3), 159–169. https://doi.org/10.1007/s40708-017-0065-7

[8] Abid, A., Abdalla, A., Ali, A., Khan, D. A., Alfozan, A., & Zou, J. (2019). Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild. arXiv (Cornell University). https://arxiv.org/pdf/1906.02569.pdf

[9] Pattabiraman, V., & Ganesh, M. (2019). Used Cars Price Prediction using Supervised Learning Techniques. International Journal of Engineering and Advanced Technology, 9(1s3), 216–223. https://doi.org/10.35940/ijeat.a1042.1291s319

[10] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). SciKit-Learn: Machine Learning in Python. HAL (Le Centre Pour La Communication Scientifique Directe). https://hal.inria.fr/hal-00650905.