Data-Driven Predictive Modeling for Used Vehicle Pricing

Anil Kumar Mureboina
Vishwa Sujit Reddy Challa
Jahnavi Konduru
Manasa Golipally
Team 2

ABSTRACT

This paper proposes developing a robust, data-driven predictive model for estimating used vehicle pricing by leveraging a comprehensive dataset of over 400,000 used car listings aggregated from Craigslist. The overarching goal is to provide consumers with reliable pricing insights that can help them identify potential bargains, reality-check listings against expected prices, and anchor their expectations to make informed decisions. After collecting this voluminous, unstructured dataset, the proposed workflow involves systematically loading the data into Databricks DBFS for scalable storage and processing. As the raw data has inconsistencies and missing values, it then undergoes meticulous cleaning and wrangling with PySpark to handle nulls and anomalies. Before building models, substantial exploratory data analysis will be conducted using diverse statistical techniques and visualizations to derive deeper insights from the data. Armed with the insights from exploration, the most appropriate regression models will be selected, instantiated, trained, and thoroughly evaluated using the machine learning capabilities of Spark MLlib on the Databricks platform. The modeling phase focuses on optimizing key evaluation metrics like R-squared, mean squared error, mean absolute error, and root mean squared error to improve predicted accuracy. Given the scale of data, distributed model training leverages the elastic computer resources of the Databricks cloud infrastructure to enable rapid iteration and refinement of models. To further boost model performance, the practice of ensemble modeling will be employed whereby multiple models are strategically combined to improve overall results. The final phase involves rigorously assessing the trained models on test data, fine-tuning their hyperparameters, and selecting the best performing ones for production deployment. At the tail end of the machine learning pipeline, the superior models operationalize the extracted knowledge from the diverse dataset into an extremely useful instrument that can generate reliable estimates of used car prices to empower more informed buying decisions. In summary, this project delivers an end-to-end, scalable machine learning system - leveraging Databricks, PySpark and MLlib - to convert raw, unstructured data into actionable pricing insights for the used automobile market. Predictive capabilities enhance transparency for consumers and facilitate fact-based decisions.

Keywords: used car pricing, predictive modeling, machine learning, data analytics, PySpark, Spark MLlib, Databricks, exploratory data analysis, ensemble models.

References

Pudaruth, S. (2014, 01). Predicting the Price of Used Cars using Machine Learning Techniques. *International Journal of Information & Computation Technology, 4*, 753-764. Retrieved from https://www.researchgate.net/publication/319306871_Predicting_the_Price_of_Used_Cars_using_Machine_Learning_Techniques

Reese, A. (2021, May 6). Used cars dataset. Kaggle. https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data