Optimizing Used Car Prices Using Machine Learning

Andrew Perez

Western Governors University

Table of Contents

A Project Overview

A.1 Research Question/Organizational Need

Is it possible to accurately predict the price of a used vehicle with machine learning by utilizing a random forest model? If this is possible, it could offer dealerships optimized pricing which could increase sales and potentially raise profit margins.

A.2 Context & Background

Appropriate, accurate pricing is key to making a sale in any industry. It is vital in the automotive industry. What a dealership pays and lists for a used vehicle is a major component of its business model. Per Edmunds.com, "the new-vehicle department of a car dealership accounts for about 58% of a dealership's total sales but less than 26% of a dealership's total gross profit…The used-vehicle department represents only about 31% of a dealerships total sales, but profit is close to that of the new-car department: nearly 25%." (Edmunds, 2019) The used vehicle segment nearly matches the profitability of the new vehicle segment despite being listed at a significantly lower sales price. For dealerships, new vehicles sell, but used vehicles offer far more return on investment (ROI).

Optimizing how they price the used vehicle segment could have the potential to increase that ROI even further. This optimization would affect both incoming and outgoing vehicles. Determining a fair price for a vehicle would allow the dealership to evaluate trade-in vehicles and offer customers a price that allows for profitability on the lot while also being fair. If vehicles entering the lot come with built-in profitability, selling those vehicles becomes easier as the dealership has more margin to negotiate with.

A.3 Published Works Related to Research

"Where Does the Car Dealer Make Money?"

Summary:

The Edmunds article discusses the profitability dynamics within car dealerships. It emphasizes that three departments, new vehicles, used vehicles, and service and parts make up 26%, 25%, and 49% of a dealership's gross profit respectively. The article touches on customer expectations, financial aspects like dealer holdbacks, evolving commission structures, and the profitability of used vehicles. It also highlights the crucial role of the Finance and Insurance (F&I) department and the enduring significance of service bays during economic downturns, where service advisers play a key role in maintaining dealership revenue through commissions on parts and services. (Edmunds, 2019)

Relation:

This article ties directly into the proposed project. It shows the importance that used vehicle sales have in a dealership's revenue stream. Despite making up only 31% of total sales, the used-vehicle department nearly matches the profit produced by new-vehicle sales. That statistic alone shows the need for ensuring optimal, accurate used-vehicle pricing.

"Consumers three times more likely to buy used cars over new"

Summary:

The article highlights findings from a 2019 survey by AA Cars, indicating that consumers are three times more likely to purchase used cars than new ones. Of the respondents, 29% bought 'nearly new' cars, 26% chose brand new cars, and 25% opted for used cars over five years old. In total, 74% reported their most recent vehicle purchase was a used car. The survey revealed that young people, often with less disposable income, predominantly buy older vehicles, with 59% of 18–24 year-olds stating their most recent purchase was a used car over five years old. (Davies, 2019)

Relation:

This article shows the importance of used vehicles among consumers. 74% of purchasers chose a used vehicle over a brand new one showing that consumers find more value in used vehicles. When this perception is combined with the profitability of used vehicle sales, it reiterates the importance of optimizing the pricing model for such sales.

"Used-Car Shoppers Are Getting a Break As Prices Fall – Unless They Need a Loan"

Summary:

The article states that used car prices have seen a 5.5% decrease on average over the past year (2023), according to a report by Edmunds. The average transaction price dropped from $30,603 to $28,935 in the 12 months ending in the third quarter. Although still 44% higher than the Q3 2018 average of $20,085, this decline indicates a move towards a more normalized market after the frenzied conditions of the pandemic era. The disruptions in new vehicle production during the pandemic led to shortages, causing used car prices to surge in 2021. The current shift is seen as a positive development for both dealers and consumers, allowing for more stable pricing and transactions. (Hyatt, 2023)

Relation:

This article shows a return to the norm for the used vehicle market. Over the past few years, used vehicles have shifted away from being a high-value purchase for consumers as new vehicles offered more benefits with a similar price tag, assuming the new vehicles were available. With a shift back to normalized used vehicle pricing, the value should return to that market and consumers should follow. The importance of optimal pricing might have diminished due to recent environmental conditions inflating prices well beyond the norm. However, if the market is returning to a normal demand level, optimal pricing will be pushed back to the forefront as competition for sales increases.

A.4 Data Analytics Solution

Following the CRISP-DM methodology, this project aims to develop a regression model utilizing the random forest algorithm. The model will undergo training and testing using historical data on used vehicle prices. Success will be defined by the model's ability to predict prices with a MAPE value of 5% or less. MAPE, representing the percentage difference between the model's predictions and actual values,

serves as a crucial metric for accuracy and precision. Achieving a very low MAPE value is imperative for deeming this model a valuable and precise solution for the dealership's pricing optimization needs.

A.5 Organization/Decision-Making Benefits

Accurate, Strategic Pricing

By using machine learning, the solution can recognize and utilize the complex relationships within the various vehicle features to provide more accurate pricing. With more accurate pricing, a dealership will be able to list its inventory strategically and gain an edge over its competition.

Increased Efficiency

An additional benefit of leveraging machine learning is that it allows a dealership to move from a manual, subjective decision to an automated, objective one. Using a highly adaptable model allows for prompt responses to market changes.

Increased Profits

Optimized pricing will create a positive feedback loop involving customer attraction and profits. By avoiding overpricing, a dealership is likely to attract more customers. With more customers, and the ability to avoid underpricing, a dealership will realize increased sales and profits.

Improved Decision Making

The solution will be able to handle both categorical and numerical factors that impact pricing. By adding interpretability to factors that are not easily quantified, it will provide decision-makers with an additional understanding of the importance of each factor regarding pricing.

B Data Analytics Project Plan

B.1 Goal, Objectives, Deliverables

Goal: Build a random forest model that can accurately predict a used vehicle's price while maintaining a mean absolute percentage error (MAPE) of 5% or less.

Objective 1: Collect, clean, and encode a used vehicle dataset to train and test the model.

Deliverable 1: A clean, complete, and encoded pandas data frame.

O2: Develop an optimized random forest model through hyperparameter tuning and feature importance analysis.

D2: A random forest model that has been exhaustively tuned and optimized.

O3: Communicate results in a concise and interpretable report.

D3: A written and visual report that summarizes project results.

B.2 Defined Scope of Project

In Scope:

Data Collection & Cleaning: Collection and cleaning of data, addressing missing values and outliers.

Model Development: Building a predictive model using the random forest algorithm, with feature selection and engineering for optimization.

Model Testing & Validation: Division of data for testing, testing against unseen data, and statistical analysis for model validation.

Statistical Analysis: Comprehensive statistical evaluation of model predictions using key metrics.

Hyperparameter Tuning: Iterative optimization of model hyperparameters for enhanced predictive accuracy.

Documentation: Documentation of analysis process, cleaning steps, and modeling techniques.

Out of Scope:

Cross Model Comparison: Comparative analysis involving multiple machine learning models (linear regression, support vector machines, gradient boosting, etc.)

Model Updating: Implementation of ongoing data collection, model training, and tuning.

External Data Integration: Incorporating additional external data sources beyond the initially identified dataset.

B.3 Project Planning Methodology

This project will utilize the CRISP-DM methodology and follow these phases/steps:

1. **Business Understanding (Analysis Phase)**
    - **Project Activity:** Work with dealership pricing managers to establish an understanding of current pricing strategy and assess optimization needs.
    - **Output:** A comprehensive understanding of the business problem, a potential solution, and criteria to determine the solution's viability.
2. **Data Understanding (Analysis Phase)**
    - **Project Activity:** Collect and explore specified dataset to understand its structure, quality, and relevance to the project.
    - **Output:** An understanding of the dataset, documented quality issues, and transformation needs.
3. **Data Preparation (Design Phase)**
    - **Project Activity:** Clean and preprocess the data to address issues listed in the previous phase. Handle missing values, feature engineering, and finalize dataset for modeling.
    - **Output:** A clean, prepared dataset ready to be used in the modeling phase.
4. **Modeling (Design & Development Phase)**

- **Project Activity:** Build and train a machine learning model using the random forest algorithm and the final dataset. Fine-tune model parameters to optimize results.
- **Output:** A trained random forest model.

5. **Evaluation (Testing Phase)**
    - **Project Activity:** Assess model performance against success criteria defined in the first phase. Iterate model to improve performance.
    - **Output:** Evaluation results, including model performance metrics.
6. **Deployment (Implementation Phase)**
    - **Project Activity:** If the project is successful, plan and execute the deployment of the pricing model into the dealership's operational environment. Implement monitoring to track model performance in real-world scenarios.
    - **Output:** Deployed pricing model and monitoring system/processes.

B.4 Project Timeline

Milestone	**Duration (days)**	**Projected Start Date**	**Anticipated End Date**
Data Collection/Understanding	2	12/11/2023	12/12/2023
Data Cleaning/Preparation	2	12/13/2023	12/14/2023
Model Creation/Training	1	12/15/2023	12/15/2023
Model Testing/Evaluation	3	12/16/2023	12/18/2023
Report Compilation	4	12/19/2023	12/22/2023
Stakeholder Communication	1	12/26/2023	12/26/2023
Project Closure/Completion	1	12/27/2023	12/27/2023

B.5 Project Resources/Costs

Labor:

- **Data Analyst/Scientist:** $0/hour, 90-100 hours

Data:

- **Dataset:** Free (Kaggle)

Software:

- **Jupyter Notebook:** Open Source
- **Python (pandas, NumPy, SciPy, Matplotlib, sci-kit-learn, seaborn):** Open Source

Hardware:

- **Desktop Computer:** $2,100.00
- **Network Connection:** $40.00

B.6 Success Criteria

This project's success criteria are three-pronged. First, our first objective will be deemed successful if at the completion we possess a 100% clean and completed dataset. For that to be the case, all missing values must be removed or inputted, and data types/values will be valid for modeling. Next, for our second objective, we must produce a random forest model that can predict a used vehicle price with a MAPE value of 5% or less. Lastly, our third objective's success will be determined by the completeness of the written report as well as the video presentation. For the written report to be complete, all steps and processes must be documented, and evaluation metrics must be presented and explained.

C Solution Design

C.1 Hypothesis

Null Hypothesis (H_0): A random forest model cannot accurately predict a used vehicle's price while maintaining a mean absolute percentage error (MAPE) of 5% or less.

Alternative Hypothesis (H_1): A random forest model can accurately predict a used vehicle's price while maintaining a mean absolute percentage error (MAPE) of 5% or less.

C.2 Analytical Methods

This project will employ the predictive analytics method to forecast a used vehicle's optimal sales price. There are various statistical metrics we can use alongside this method to assess our random forest model, but the key metric we will focus on is the Mean Absolute Percentage Error (MAPE).

Predictive analytics is a justified choice because we are using historical data and various features/predictors to determine an optimal price. We are attempting to find a future answer by forecasting previous results. As for MAPE, it is ideal for this project because it provides us with an interpretable metric that can be compared across various models and datasets if needed as it is scale-independent.

C.3 Tools/Environments

Python: An open-source, versatile programming language with a vast set of libraries. The following libraries have the potential to be used in this project.

pandas: Data cleaning, preprocessing, and analysis

NumPy: Scientific computing

SciPy: Used in combination with NumPy. In this case, for outlier identification through z-scoring.

sci-kit-learn: Machine learning algorithms, metrics, and model selection.

Matplotlib: Basic visualizations

seaborn: Statistic-specific visualizations

Jupyter Notebooks (IDE): Interactive, web-based IDE that allows for the combination of code, analysis, and documentation in one environment.

C.4 Evaluation Methods/Metrics

Model/Method:

- **Type:** Supervised Regression
- **Algorithm:** Random Forest
- **Performance Metrics:** Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Absolute Scaled Error (MASE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2)
- **Success Criteria:** MASE < .05
- **Justification:** Well-suited to predict continuous numerical values (prices) based on inputted features/predictors even if those predictors have complex, non-linear relationships.

Metrics:

- **MAE:** Average of the absolute difference between predicted and actual values. This metric does not consider the direction of errors, only their magnitude.
    - **Output:** Unit of target
- **MAPE:** Percentage difference between predicted and actual values. This allows us to see the accuracy of our model regardless of dataset and/or scale.
    - **Output:** Percentage between 0 and 100%
- **MASE:** Compares the performance of a forecasting model to that of a naive method. This metric considers both the direction and magnitude of errors. It pairs well with the MAPE to provide an interpretable measure of accuracy.
    - **Output:** Scale of 0.0 to 1.0+, with values less than 1.0 indicating that the model outperforms a naive method
- **MSE:** Average of the squared differences between predicted and actual values. This metric penalizes larger errors.
    - **Output:** Squared unit of target
- **RMSE:** The square root of MSE, provides an output that is the same as the unit of the target for more interpretable measure.
    - **Output:** Unit of target
- **R^2:** The proportion of the variance in the target that is predictable from the included features.
    - **Output:** Scale of 0.0 to 1.0, where a higher value indicates a better fit of the model to the data (1.0 = 100% of the target's variance can be predicted by the included features)

C.5 Practical Significance

To assess the practical significance of the model, a triad of the metrics listed above, MAE, MAPE, and RMSE, will be employed. In this analysis, MAE will represent the dollar amount our predictions our off by (under and overpricing.) MAPE will provide a scalable measure for that difference and will be compared to the dealership's average profit margin. If our MAPE value far exceeds the dealership's average profit margin, our solution will not be viable or successful. Lastly, RMSE is like MAE as it provides a tangible difference in pricing, but it also penalizes larger errors, which is crucial for a model that could affect the dealership's profit/loss.

C.6 Visualization Tools

With the help of Python's library Matplotlib, I will present my findings with the following visualizations:

Horizontal Bar Chart: Top Features and Their Importance on Pricing, highest to lowest

Scatter Plot: Predictions vs Actual Values

D Dataset

D.1 Source

The following dataset will be used for this project:

Dubizzle Used Car Sales Data (2022) – Ali Hassan

D.2 Source Justification

This source was chosen for the following reasons:

Completeness

Of the 20 features, only one is missing 10% or more. Every other feature has 99%+ complete and valid data. This level of completeness will allow for a comprehensive analysis and limit any biases that may arise in cleaning/preprocessing the data.

Relevance

This dataset remains relevant as it was only collected a year ago (12/05/2022). The dataset collected from the UAE reduces its relevancy to US dealerships, but the project's purpose is inherently scalable and independent of location so it should be possible to take our results and apply them to a dealership's internal dataset.

Size

This dataset has 9170 entries and 20 features. Those dimensions are sizable enough to train and test the random forest model.

Feature List

This dataset includes up to 19 features that we can use as predictors, from transmission type to color. Having a large set of vehicle specifications should improve our model's predicting performance.

D.3 Collection Methods

The dataset was directly downloaded from Kaggle and the data itself was scrapped from Dubizzle using a Python library called Beautiful Soup.

D.4 Quality/Completeness of Data

Quality

This dataset is composed of high-quality data. Values are consistent, valid, and standardized. Datatypes will likely need to be converted, but the actual data should require very little manipulation.

Completeness

As alluded to in D.2, this dataset is very complete. Two features are missing data, year and number of cylinders. Respectively, they are missing 10% and 1% of their values. Addressing the large percentage in year should be manageable through inputting the most recent year for vehicles with no usage and dropping entries that have no inferable value. Every other feature is 100% complete.

D.5 Data Governance, Privacy, Security, Ethics, Legal and Regulatory Compliances

Governance

- **Relation:** Ensure data quality, manageability, and traceability.
- **Precautions:** Perform regular data checks after manipulation, use version control when dataset changes significantly, apply consistent, appropriate naming conventions.

Privacy

- **Relation:** Ensure any personal identifiable information (PII) is not at risk of being exposed.
- **Precaution:** This dataset contains the location of each entry on a state level, but no PII data. There is potential to search for additional information through an entries title, however, that feature will not be included in the analysis or public report.

Security

- **Relation:** Ensure data is securely stored and only accessed with authorization.
- **Precaution:** Password protection for PC access.

Ethics

- **Relation:** Ensure fair and unbiased analysis.
- **Precaution:** Perform data manipulation and imputation objectively and with reasonable logic, avoid definitive conclusions without statistical significance.

Legal/Regulatory

- **Relation:** Ensure compliance with any local and federal laws or regulations.
- **Precaution:** This dataset does not include any PII or health information. The data is available for public use and completely anonymized.

E References

Davies, E. (2019, October 28). *Consumers three times more likely to buy used cars over new*. Motor Trader. https://www.motortrader.com/motor-trader-news/automotive-news/majority-buy-consumers-opt-used-new-cars-28-10-2019

Edmunds. (2019, June 13). *Where does the car dealer make money?* https://www.edmunds.com/car-buying/where-does-the-car-dealer-make-money.html

Hyatt, D. (2023, November 13). *Used-car shoppers are getting a break as prices fall-unless they need a loan*. Investopedia. https://www.investopedia.com/used-car-shoppers-are-getting-a-break-as-prices-fall-unless-they-need-a-loan-8401255