Using Machine Learning to Predict Car Value

WGU Data Science Capstone Project
By Keniyah Chestnut

Introduction

My name is Keniyah Chestnut. I have a background in Information Technology and Data Science, with experience in technical support and data analytics. I'm currently completing my Master's in Data Science at Western Governors University, where I've been focused on applying machine learning techniques to real-world problems like used car price prediction.

The Problem

Can we predict the value of a used car based on its features?

Using a dataset of 4,000 vehicles, this project explores whether we can build a model to accurately estimate car prices.
The dataset includes key attributes such as:

- Make and model
- Year
- Fuel type
- Engine type
- Accident history
- Color
- Title status

Hypothesis: These variables, when analyzed using a Random Forest regression model, can be used to reliably predict used car prices.

Null Hypothesis:
There is no statistically significant relationship between vehicle features and used car prices.

The Data Analysis Process

brand	0
model	0
model_year	0
mileage	0
fuel_type	170
engine	0
transmission	0
ext_color	0
int_color	0
accident	113
clean_title	596
price	0

dtype: int64

Summary Table of Missing Values in the Dataset

Step 1: Clean the Data

Since I am using a dataset provided by the university, the first step was to clean it before analysis.

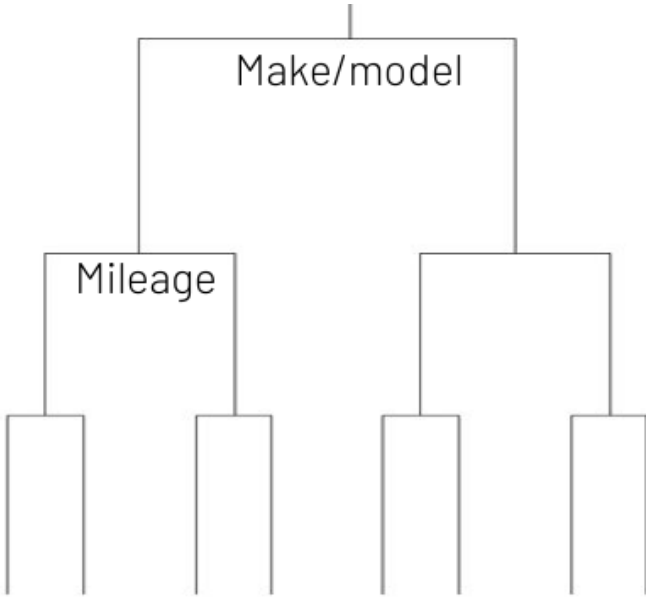
Handling missing data:

- I labeled missing values as "unknown."
- I assumed the missing data might be intentional. For example, some sellers may have chosen not to report past accidents or rebuilt titles.
- However, without knowing how the data was collected, it is difficult to tell whether the missing entries are meaningful or just incomplete.

Other cleaning steps included:

- Removing non-numeric characters from mileage values
- Dropping extreme outliers, such as cars priced over $100,000

Step 2: Analyze the Data



Random Forest Model Visualization

Using a Random Forest Regressor

- Regression is a method used to identify relationships between variables and make predictions.
- A Random Forest is a group of decision trees. Each tree looks at different parts of the data and makes its own prediction.

For example, let's say we are trying to predict the value of a used Toyota Corolla:

- One tree might estimate the value at $10,000 based on the make and model.
- Another tree might notice the car has between 100,000 and 150,000 miles and adjust the value to around $8,000.
- A third tree might consider that the car has no accidents reported and raise the estimate to $8,500.

Each tree contributes to the final prediction by averaging its results with the others, helping improve accuracy and reduce bias.

Limitations of Random Forest Regression

Advantages of Random Forest Regressors:

- Powerful tool that can handle many variables at once
- Automatically identifies which features are most important
- Produces clear, reliable predictions based on patterns in the data

Disadvantages of this Tool:

- Predictions can be inaccurate if the dataset is too small
- Too many unique or rare entries can make it hard for the model to learn effectively

Findings

Model Used: Random Forest Regressor
Train/Test Split: 60/40
Number of Estimators: 100

Evaluation Metrics:

- Mean Absolute Error (MAE): $8,025
- Mean Squared Error (MSE): 132,764,065
- R^2 Score: 0.699

Interpretation:

- The model predicts car prices with an average error of around $8,025.
- Approximately 70% of the variance in car prices is explained by the selected features.
- Key features used include brand, mileage, model year, fuel type, and title status.

Advantages:

- Handles many variables and interactions well.
- Performs well with mixed data types.

Limitations:

- Model may struggle with rare or high-priced vehicles over $100,000.
- Less interpretability compared to simpler models.

Proposed Action: Improve Model Accuracy

The current model predicts used car values within approximately $8,000, which limits its usefulness for precise pricing. To improve performance, the following actions are recommended:

- Limit the analysis to a narrower vehicle price range
- Increase the size and diversity of the dataset
  - Some makes and models appeared only once, reducing prediction accuracy
  - More frequent entries improve model learning and generalization
- Exclude rare or high-value vehicles (over $100,000) to avoid skewing results

Expected Benefits of the Study

This model provides a general sense of whether a car is high or low in value. However, for cars priced under $10,000, prediction accuracy is not yet precise enough for individual pricing.

By implementing the proposed improvements, we can build a more accurate model that:

- Automates the process of pricing used vehicles
- Reduces the time and effort required for manual research
- Helps ensure cars are priced competitively and accurately in the market

Sources

All data and tools used were provided by WGU.