Using Machine Learning to Predict Car Value

WGU Capstone project
by Eric Williams

A bit about me...

- Born and raised in Salt Lake City, Utah
- Mathematics degree from University of Utah
- Former high school and middle school teacher
- Student at WGU studying Data Analytics

My girlfriend and I at Niagara Falls in 2024

The Problem

Given a dataset of 4,000 cars and prices, can we create a model that predicts car value?
The dataset includes

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

Hypothesis: By analyzing the above variables in a Random Forest regression, a model can be created to accurately predict the value of used cars.

The Data Analysis Process

Step 1: Clean the data

- Because I am using a dataset provided by the university, the first step is to clean the data
- What to do with missing data?
  - Label it as "unknown"
  - My assumption is that the data is missing for a reason
    - People may not want to report their past accident status or rebuilt titles
  - Without knowing the context of the data collection process, it's hard to know whether or not the data is incomplete for a specific reason
- Other data cleaning steps included:
  - Stripping mileage values of non-numeric data
  - Removing outliers

brand	0
model	0
model_year	0
milage	0
fuel_type	170
engine	0
transmission	0
ext_col	0
int_col	0
accident	113
clean_title	596
price	0
dtype: int64	

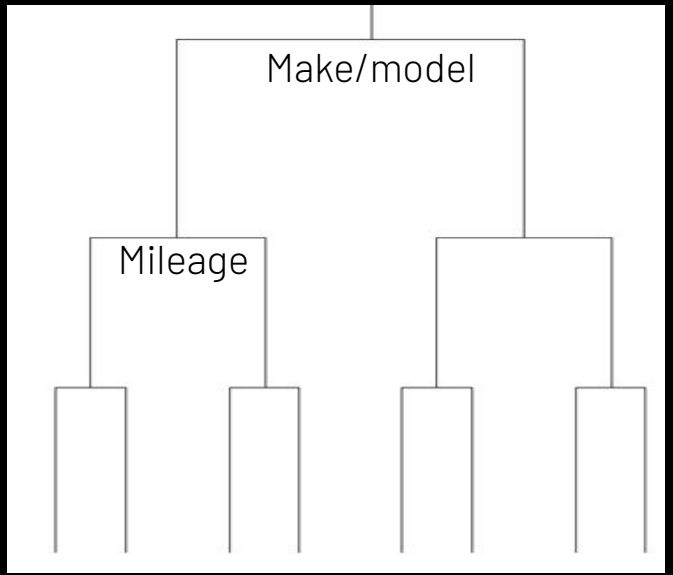
Table summary of missing data in the dataset

Step 2: Analyze the data

Using a Random Forest Regressor

- A regression is a tool to help us find relationships between variables and make predictions
- A random forest is a bunch of small trees that take in data and make decisions

For example, we might start with a used Toyota Corolla:
The first tree might say: “this car’s value on average is $10,000.”
The second tree might say “This car has between 100,000 and 150,000 miles on it, so the value is closer to $8,000.”
The third tree might say: “There are no accidents on record for the car, so its value is closer to $8,500.”



A visualization of Random Forests

Limitations of Random Forest Regression

Advantage of Random Forest regressors:

- High powered, multi-dimensional tool
- Analyzes many variables and weighs importance
- Makes concrete predictions

Disadvantage of this tool:

- If there isn't enough data to train the model, predictions will be off
- If we have too many unique entries, they will be hard to predict

Findings

When analyzing cars with value between $2,000 and $100,000:

- The model was able to predict car value within $8,182
- The model was able to explain 69% of variance in price

Limitation: This analysis only included car values under $100,000

Proposed Action: More Data Needed

A model that can only predict value within about $8,000 is not useful. However, we can improve this model by:

- Limiting the analysis to a smaller range of vehicles
- Including more data
  - Some car entries were unique for the training, or had never been seen before by the model when making a prediction
  - The more times a model sees a make and model, the more accurate the predictions will be
- Exclude rare or expensive vehicles

Expected Benefits of the Study

In short, this model is great for predicting generally if a car is valuable or not

However, if many cars are priced under $10,000, the values are not predicting accurately enough to give the value of cars

With the improvements mentioned in the last slide, we can create a model that can automate car prices. If we can build a more accurate model, this will:

- Save time, effort, and research when it comes to determining car values
- Will ensure the cars are priced accurately

Sources

No sources were used besides official WGU materials.