Predicting the Price of Used Vehicles

Jonathan Daughtry

Online Vehicle Shopping is Increasing in Popularity

An estimated 34.1 million used vehicles were sold in 2020, and that number is expected to increase to 41 million in 2021 (CNBC)

Car shoppers visit an average of 4.2 websites when car shopping (CoxAuto)

Less than ⅓ of younger consumers want to do in person vehicle shopping, and more respondents are interested in contactless service (McKinsey)

18% of consumers would buy a vehicle sooner if there was an online purchase option (ThinkwithGoogle)

48% of consumers say they want to handle price negotiations online (Cars.com)

The Problem:

With the seemingly endless variety of used vehicles for sale, how can buyers and sellers guarantee the best value for their purchase?

The Solution:

Create a model that can give an accurate price suggestion based on vehicle properties.

Who Might Care?

Large online retailers

CarGurus

vroom

Autotrader

CARVANA

CARMAX

Local Used Car Companies

Individuals who want to buy or sell

Data Information

Data collected for all vehicles sold on Craigslist
Number of vehicle records used: 183,514

Target Variable:
Price range: $100 - $125,000

Determining Variables:
Mileage range: 1,000-150,000
Year range: 1919 - 2020
Manufacturer
Number of cylinders
Transmission
Drive
State
Color

Data Cleaning: https://github.com/dawgtree/CapstoneTwoProject/blob/main/2nd_Capstone_Data_Wrangling.ipynb

Data Source: https://www.kaggle.com/austinreese/craigslist-carstrucks-data

Data Preparation

In order to clean the data:

Null values were turned into ‘Unknown’

All years after 2020 were deleted

All prices below $100 and above $125,000 were deleted.

Any rows that did not have Manufacturer, Odometer, Year, and Price were deleted

All columns other than what were listed were deleted.

Data Exploration: Year and Mileage

Price decreased as amount of miles driven increased, an inverse relationship

Price increased as year of vehicle increased, a direct relationship

Data Exploration: Categorical Variables

Manufacturer, number of cylinders, transmission, drive, state, color were all categorical

Hypothesis testing determined all could be important in affecting price

EDA: https://github.com/dawgtree/CapstoneTwoProject/blob/main/2nd_Capstone_EDA.ipynb

Machine Learning Modeling Selection

Models Used:

Random Forest Regression

AdaBoost Regression

Gradient Boosting Regression

Best Hyperparameters:

Random Forest - max depth of 60, 9 max features, and 200 estimators

AdaBoost - learning rate of 0.05, an exponential loss function, and 50 estimators

Gradient Boosting - learning rate of 0.25, a max depth of 10, and 200 estimators

Modeling: https://github.com/dawgtree/CapstoneTwoProject/blob/main/2nd_Capstone_Modeling.ipynb

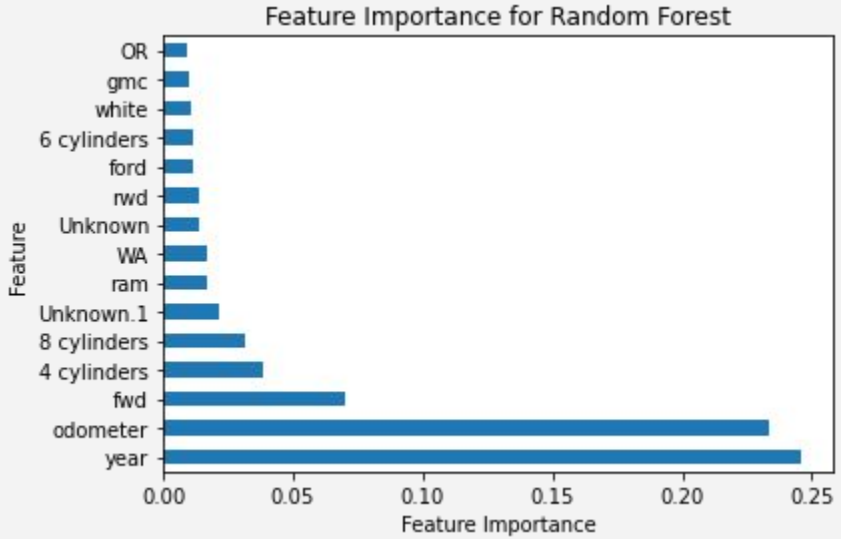
Model Comparison

	Time Taken in seconds (per 100 samples)	RMSE	MAE	Adj R^2
Random Forest	0.009	5282	2821	0.83
AdaBoost	0.003	9372	6517	0.46
Gradient Boost	0.002	5629	3355	0.81

Time was determined to be the least important factor

Random Forest scored the best, AdaBoost scored the worst

Random Forest Feature Importance



Year and odometer(mileage) were the most important features

Drive and number of cylinders were the next most significant features

*Note “Unknown.1” was unknown type of drive, “Unknown” was unknown number of cylinders

Conclusions

Random Forest was the best model, but Gradient Boosting was close enough that further adjustment of hyperparameters could make Gradient Boosting the best.

Features could be reduced to year, odometer, number of cylinders, and type of drive, but accuracy would be reduced.

Time differences were insignificant, especially considering that purchasing a vehicle is often the second largest investment a person makes in their life.

Future Research/Improvements

Be wary of multicollinearity between year and mileage. Further modeling could be done using each one separately or use models that punish that.

Further adjust and expand hyperparameters.

Take car model into account, but it would require classifying over 15,000 different types of models in the dataset.

Update modeling as new data comes in since dataset is continuously being updated.

Recommendations

This model is ready to be used as a foundation for businesses to use to develop a pricing range.

Suggest using 95% confidence interval on price ranges of specific vehicle.

Range will provide seller to adjust based on unlisted vehicle traits (condition, extra features, accident history, etc.)

Users should feel that range gives ability to have personal input.

Thank You!

Any Further Questions?

Jonathan Daughtry
Email: daughtryje@gmail.com
https://www.linkedin.com/in/jonathan-daughtry/
https://github.com/dawgtree
Full detailed report of project:
https://github.com/dawgtree/CapstoneTwoProject/blob/main/Capstone%20Two%20Final%20Report.pdf