Car Pricing Algorithms: An Investigation

Roman Tedeschi

Problem Statements

Can used car data be scraped from retail websites?

Which basic statistical model is best at predicting used car prices based on car features?

What is the best way of implementing these models specific to used cars?

Which Site?

Autotrader

CARMAX®

CARVANA

cars.com™

Challenge: URL Formatting

Each page url ends in a ‘page=2’, ‘page=3’, etc.

However, ‘page=’ is a dummy portion of the url which does not work with url-based parsers.

Workaround:
carvana.com/cars/nissan-sentra

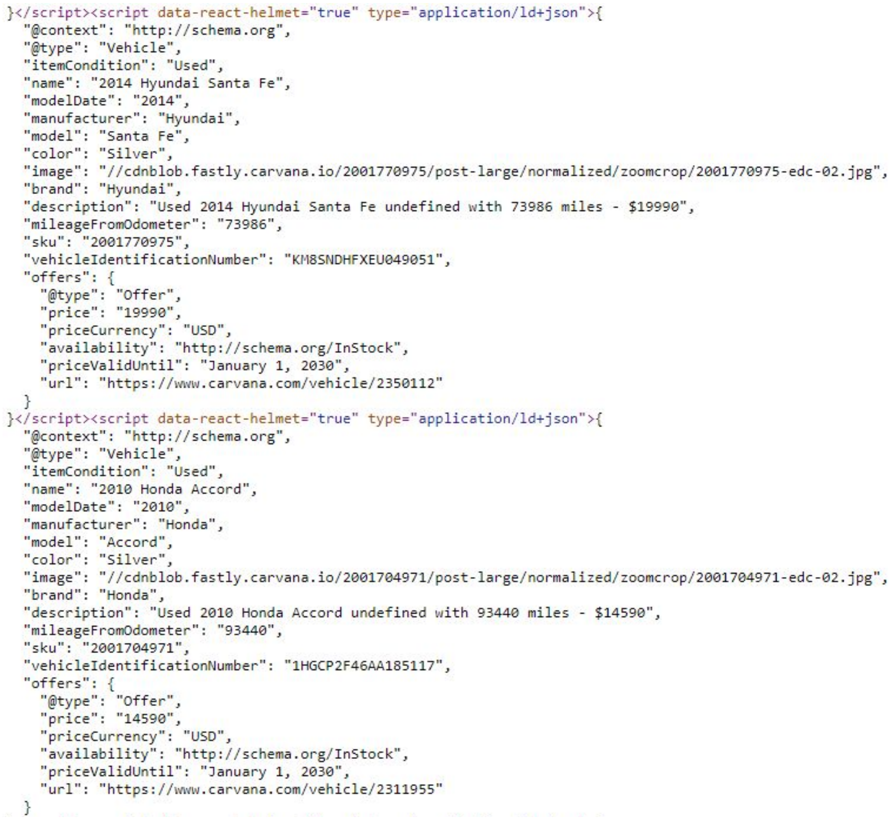
Challenge: Request Limitations

Running more than four requests per minute results in rate limitations.

eg. 10,000 cars with full data:
2,500 / 60 = 42 hours

Workaround:
Set a 15 second sleep timer

Schema.org



```
}</script><script data-react-helmet="true" type="application/ld+json">{
  "@context": "http://schema.org",
  "@type": "Vehicle",
  "itemCondition": "Used",
  "name": "2014 Hyundai Santa Fe",
  "modelDate": "2014",
  "manufacturer": "Hyundai",
  "model": "Santa Fe",
  "color": "Silver",
  "image": "//cdnblob.fastly.carvana.io/2001770975/post-large/normalized/zoomcrop/2001770975-edc-02.jpg",
  "brand": "Hyundai",
  "description": "Used 2014 Hyundai Santa Fe undefined with 73986 miles - $19990",
  "mileageFromOdometer": "73986",
  "sku": "2001770975",
  "vehicleIdentificationNumber": "KM8SNDHFXEU049051",
  "offers": {
    "@type": "Offer",
    "price": "19990",
    "priceCurrency": "USD",
    "availability": "http://schema.org/InStock",
    "priceValidUntil": "January 1, 2030",
    "url": "https://www.carvana.com/vehicle/2350112"
  }
}</script><script data-react-helmet="true" type="application/ld+json">{
  "@context": "http://schema.org",
  "@type": "Vehicle",
  "itemCondition": "Used",
  "name": "2010 Honda Accord",
  "modelDate": "2010",
  "manufacturer": "Honda",
  "model": "Accord",
  "color": "Silver",
  "image": "//cdnblob.fastly.carvana.io/2001704971/post-large/normalized/zoomcrop/2001704971-edc-02.jpg",
  "brand": "Honda",
  "description": "Used 2010 Honda Accord undefined with 93440 miles - $14590",
  "mileageFromOdometer": "93440",
  "sku": "2001704971",
  "vehicleIdentificationNumber": "1HGCP2F46AA185117",
  "offers": {
    "@type": "Offer",
    "price": "14590",
    "priceCurrency": "USD",
    "availability": "http://schema.org/InStock",
    "priceValidUntil": "January 1, 2030",
    "url": "https://www.carvana.com/vehicle/2311955"
  }
```

Web Scraping: The Data

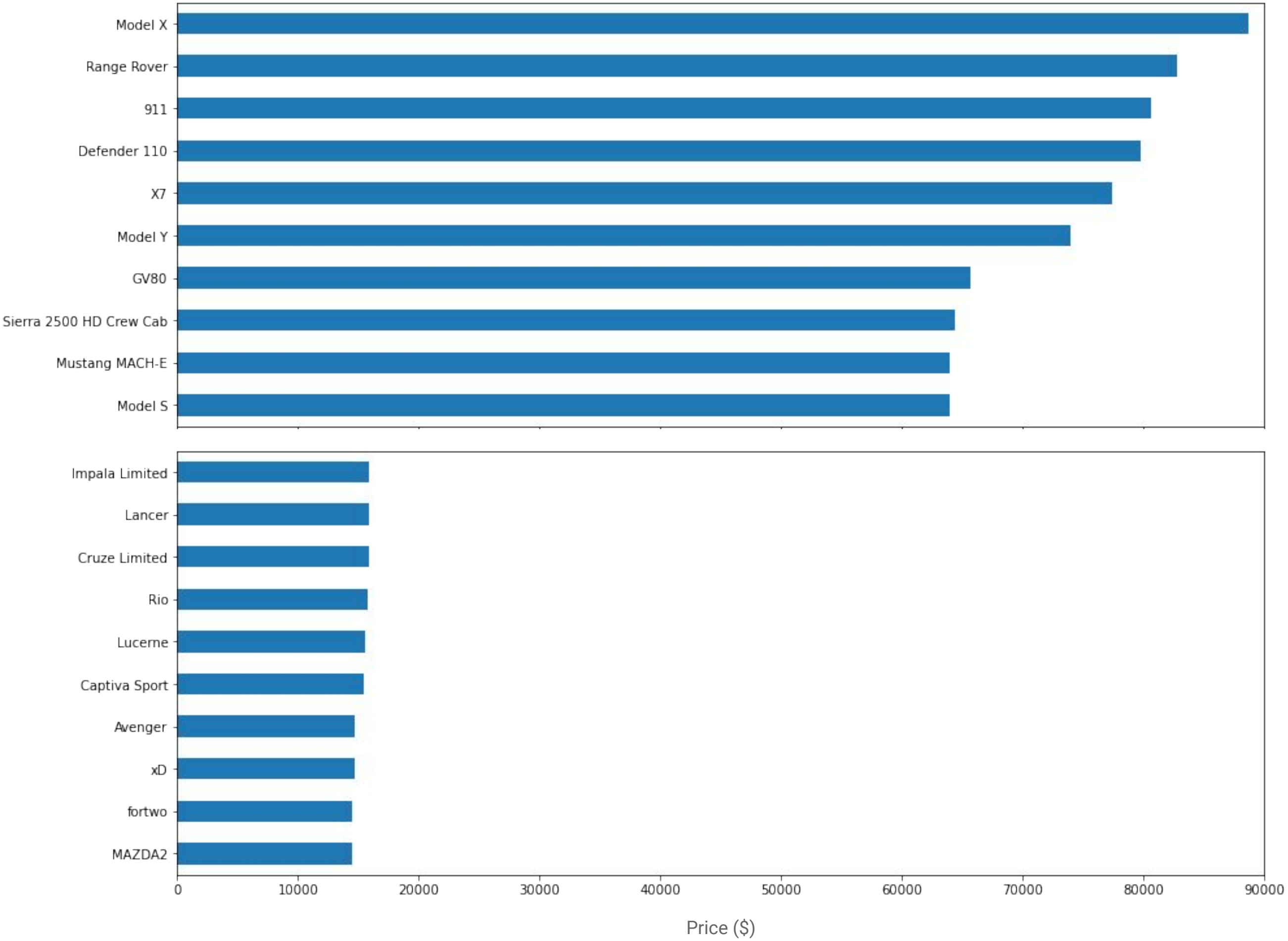
After web scraping, a DataFrame was created with 10,659 unique cars.

These consisted of 404 different car models.

The Chevrolet Camaro had the most cars at 59, while many others just met the chosen minimum of 11.

Mean Price per Model

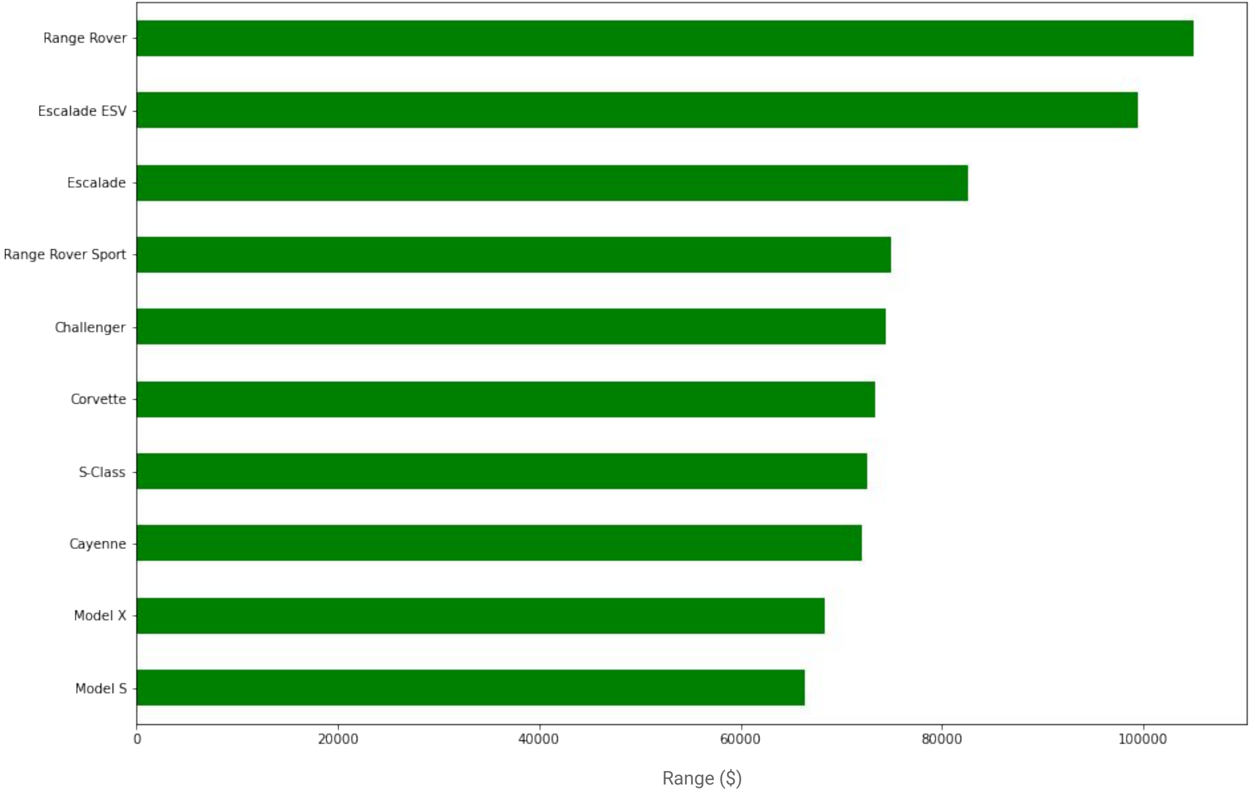
10 most and least expensive



Price Range per Model

10 largest ranges

Mean range was $21,108



Features

Name	A summary of the car year, manufacturer, and mode
Year	The model year of the car
Manufacturer	The name of the company that manufactured the car
Model	The model name of the car
Trim	The trim level of the car
Mileage	The odometer reading of the car at the time it was put up for resale (miles)
Color	The color of the car
Price	The price for which the car is being offered for resale (US dollars)

Modeling

(Isn’t r^2 supposed to be between 0 and 1?)

1 | Linear regression on all features and all cars

- Astronomical coefficients
- Duplicated data frame
- Data leakage

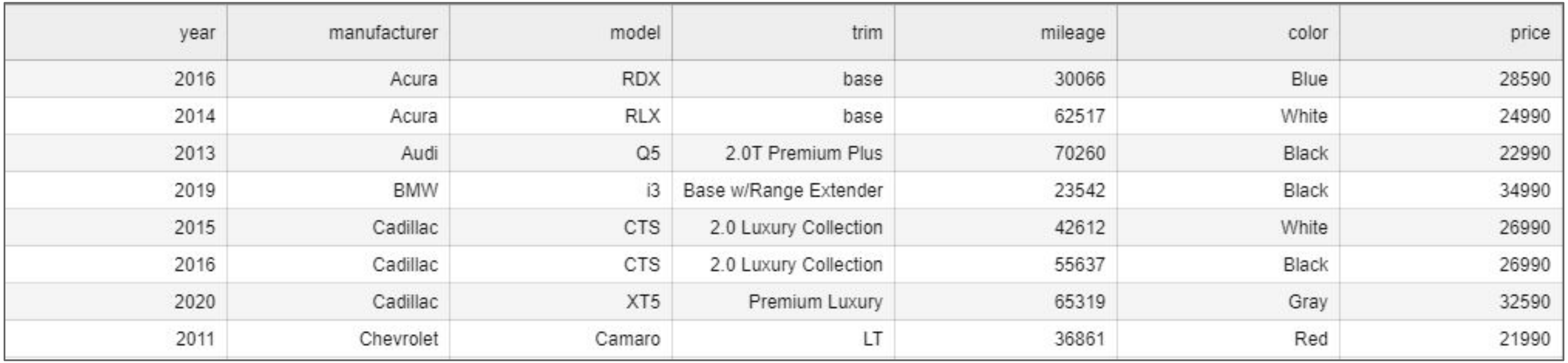
2 | Decision tree on all features and all cars

- r^2 train/test: 1.0 / 0.81
- RMSE: $2,998

3 | Random forest on all features and all cars

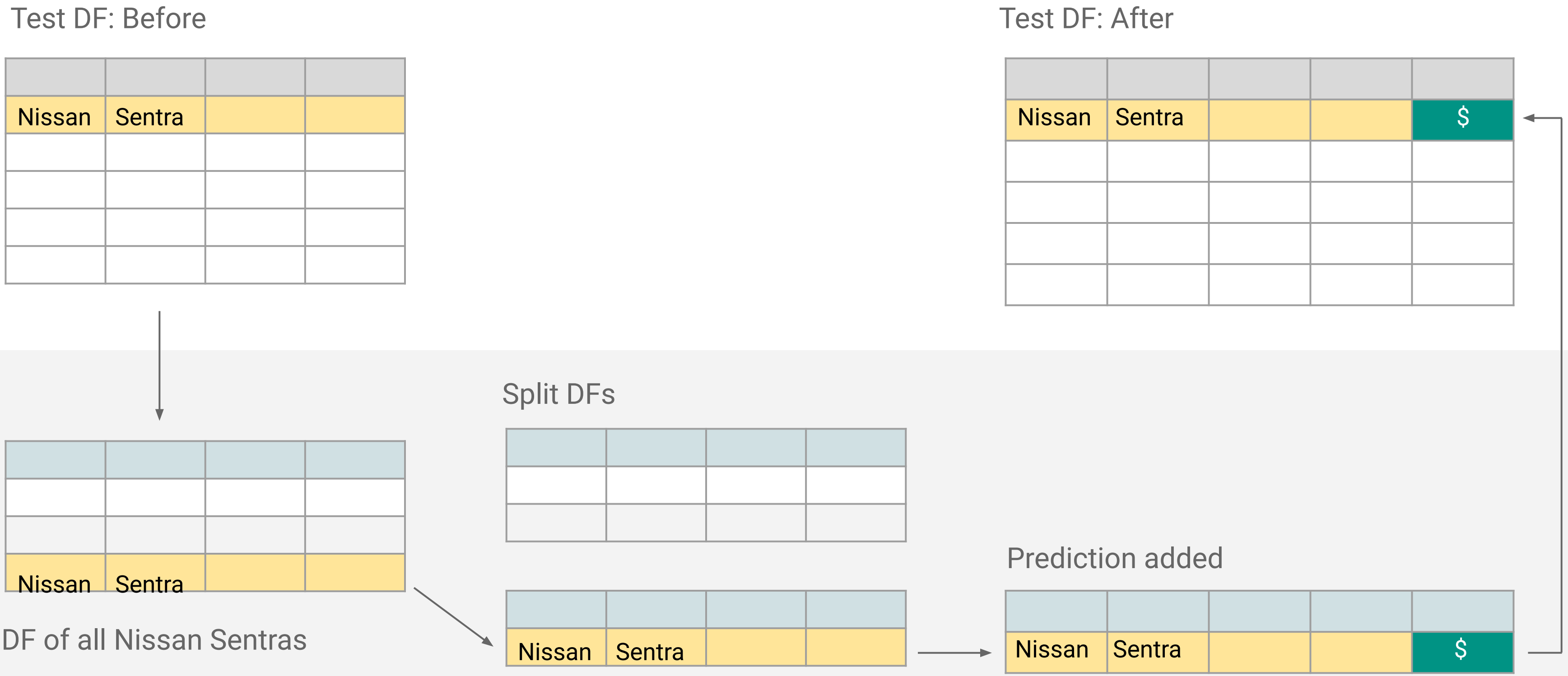
- r^2 train/test: 0.98 / 0.86
- RMSE: $3,086

Is One Model Optimal?



year	manufacturer	model	trim	mileage	color	price
2016	Acura	RDX	base	30066	Blue	28590
2014	Acura	RLX	base	62517	White	24990
2013	Audi	Q5	2.0T Premium Plus	70260	Black	22990
2019	BMW	i3	Base w/Range Extender	23542	Black	34990
2015	Cadillac	CTS	2.0 Luxury Collection	42612	White	26990
2016	Cadillac	CTS	2.0 Luxury Collection	55637	Black	26990
2020	Cadillac	XT5	Premium Luxury	65319	Gray	32590
2011	Chevrolet	Camaro	LT	36861	Red	21990

Function Steps



Test DF: Before

Nissan Sentra

DF of all Nissan Sentras

Nissan Sentra

Split DFs

Nissan Sentra

Prediction added

Nissan Sentra $

Test DF: After

Nissan Sentra $

The ‘Per Model’ Model

Two out of the 51 test cars had to be dropped since they were not models where Carvana had over ten cars for sale

1 | Linear Regression

RMSE: $2,293

2 | Decision Tree

RMSE: $3,342

3 | Random Forest

RMSE: $2,063

Conclusion

For highly specific inventory like cars, one algorithm was not the best choice.

Rather, the data should be grouped by like features that share correlative value.

Best Overall Model: Random Forest

Tends to work well on data where some features are a subset of another feature.

Able to handle unseen categorical values better.

Limitations

This exercise was simply an investigation into a snapshot of available car prices.

The models could be improved via:

- *access to more features*
- *larger data set*
- *profit maximization*
- *supply and demand*
- *inventory costs*
- *seasonal adjustments*

Thank you! Questions?