Blue Book for Bulldozers

Predicting the sales price of used bulldozers

A Springboard Data Science Capstone submission by Greg McKenzie

Why bulldozers?

What stimulates the buying & selling of bulldozers?

Infrastructure development

#BuildBackBetter

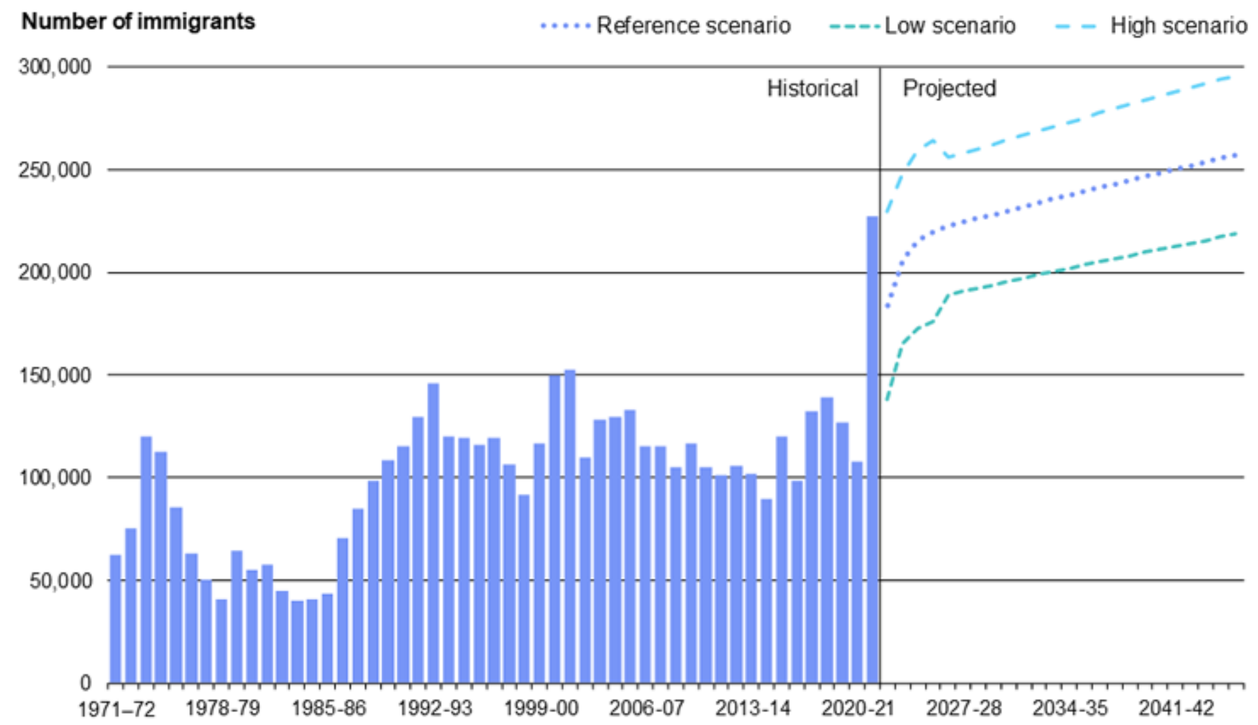
- $2.2 trillion investment

Why bulldozers?

What stimulates the buying & selling of bulldozers?

Urbanization

Chart 16: Immigration to Ontario, 1971 to 2046

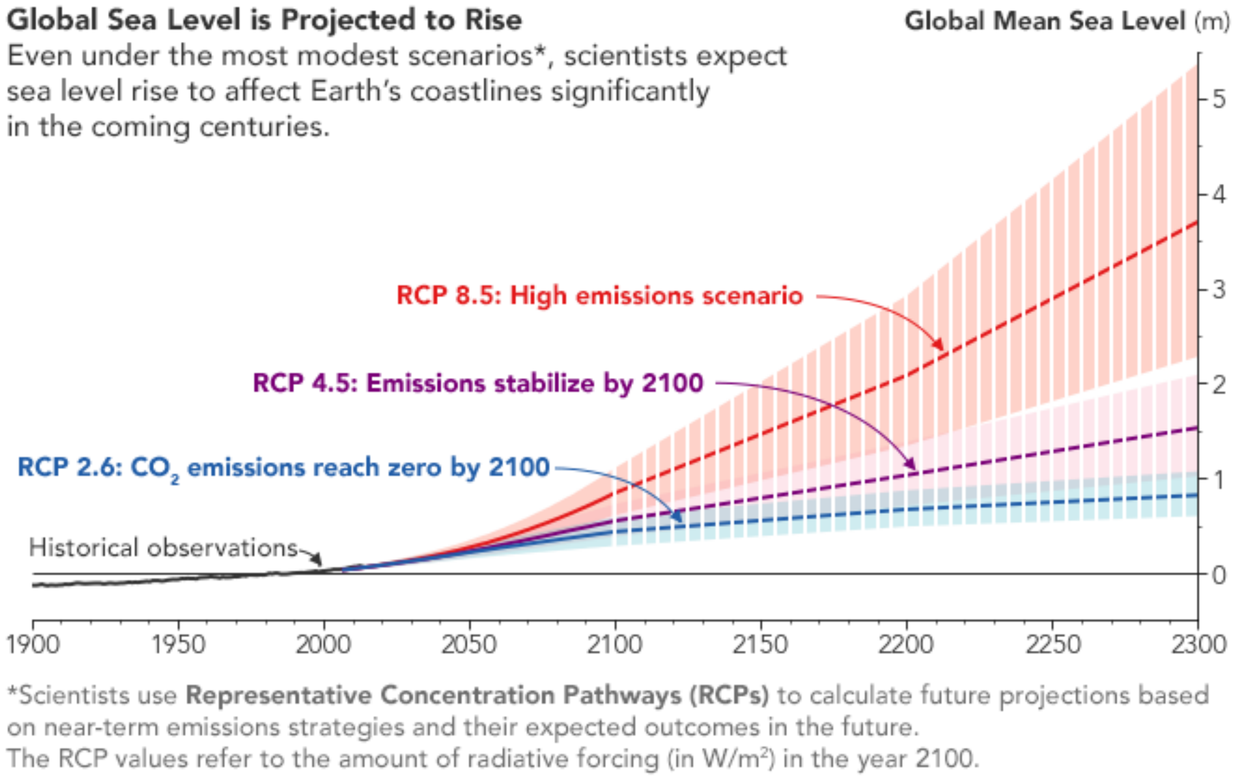


Sources: Statistics Canada for 1971–2022, and Ontario Ministry of Finance projections.

Why bulldozers?

What stimulates the buying & selling of bulldozers?

Natural Disasters



Who might care?

Where is a price prediction useful?

Insurance Companies

Who might care?

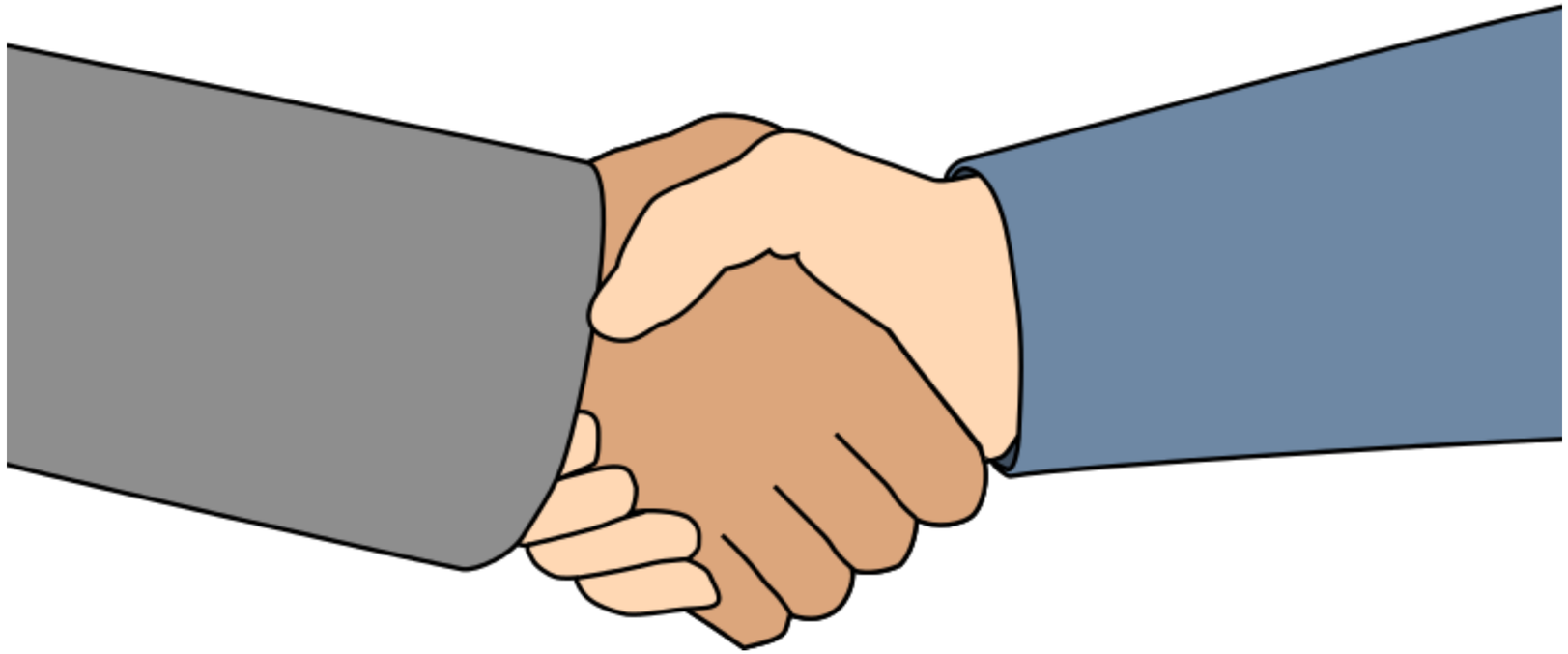
Where is a price prediction useful?

Auction Houses

Who might care?

Where is a price prediction useful?

Rental Agencies



Who might care?

Where is a price prediction useful?

Bulldozer Manufacturers

Data

53 features, 412,698 observations kaggle

- **saleprice (target)**: what the machine sold for at auction
- **SalesID**: the uniue identifier of the sale
- **MachineID**: the unique identifier of a machine. A machine can be sold multiple times
- **saledate**: the date of the sale
- **state**: US state in which the bulldozer was sold

  and 48 more…

Data

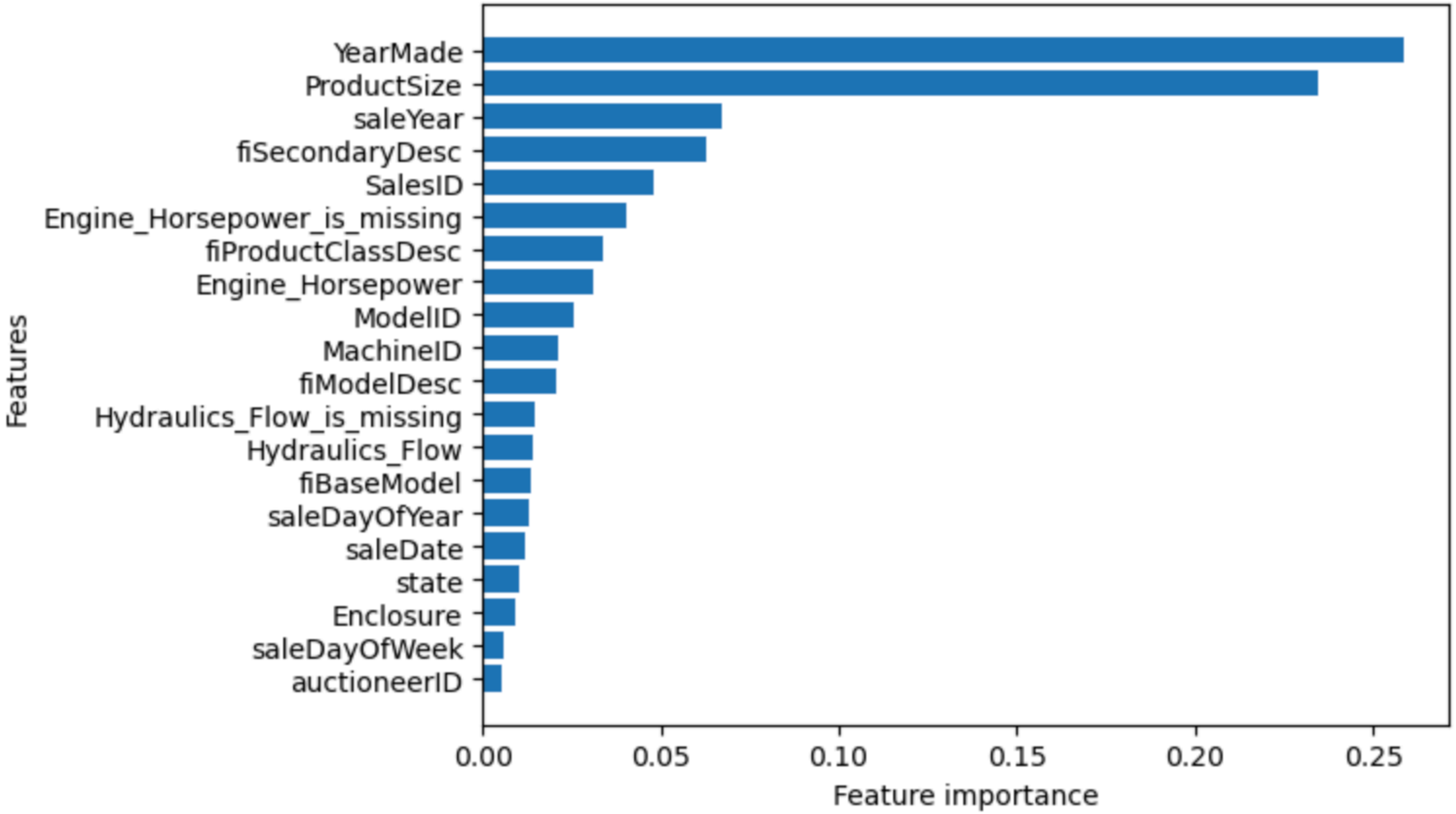
53 features, 412,698 observations kaggle

- **saleprice (target)**: what the machine sold for at auction
- **SalesID**: the unique identifier of the sale
- **MachineID**: the unique identifier of a machine. A machine can be sold multiple times
- **saledate**: the date of the sale
- **state**: US state in which the bulldozer was sold

  and 48 more…

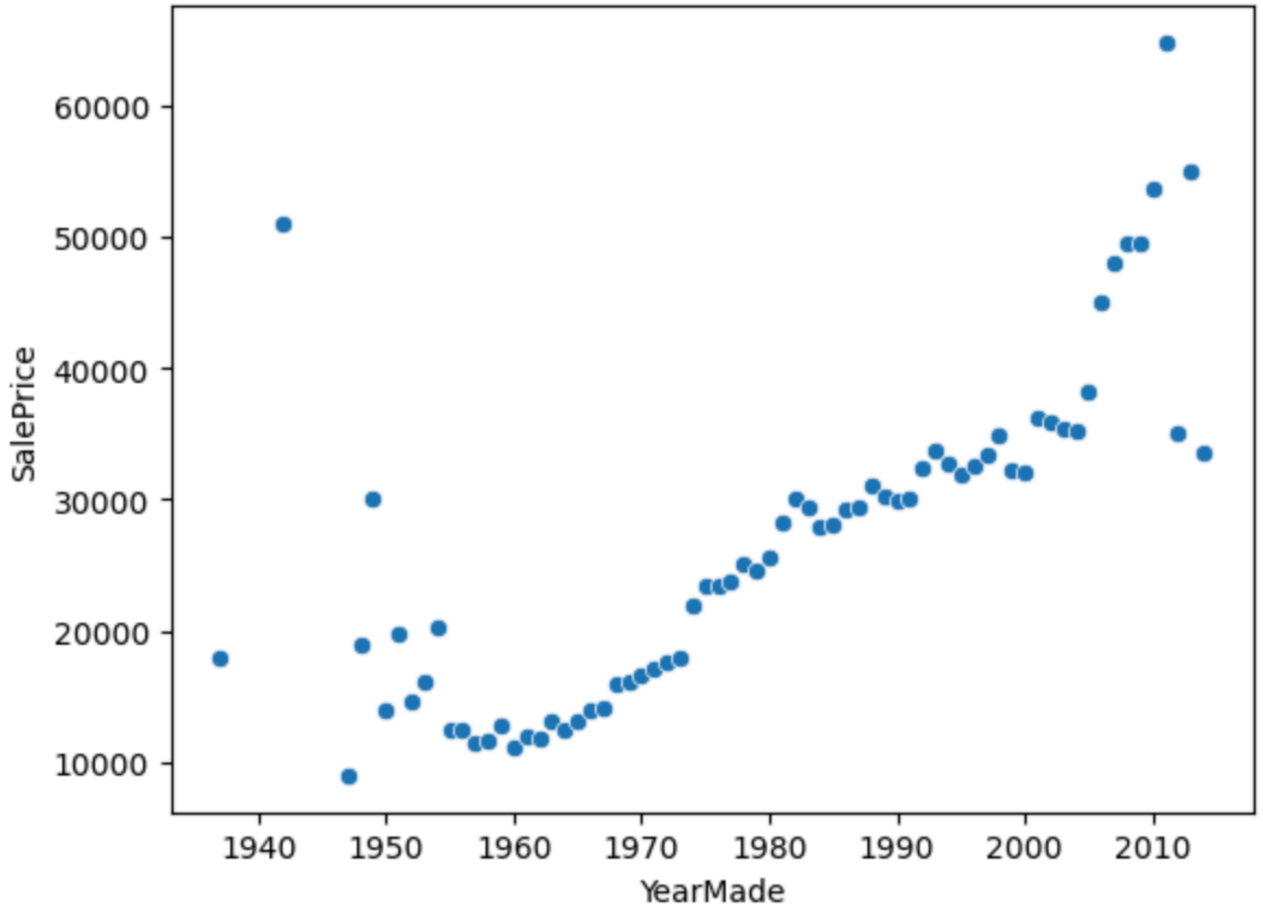
Data

Ranked feature importances



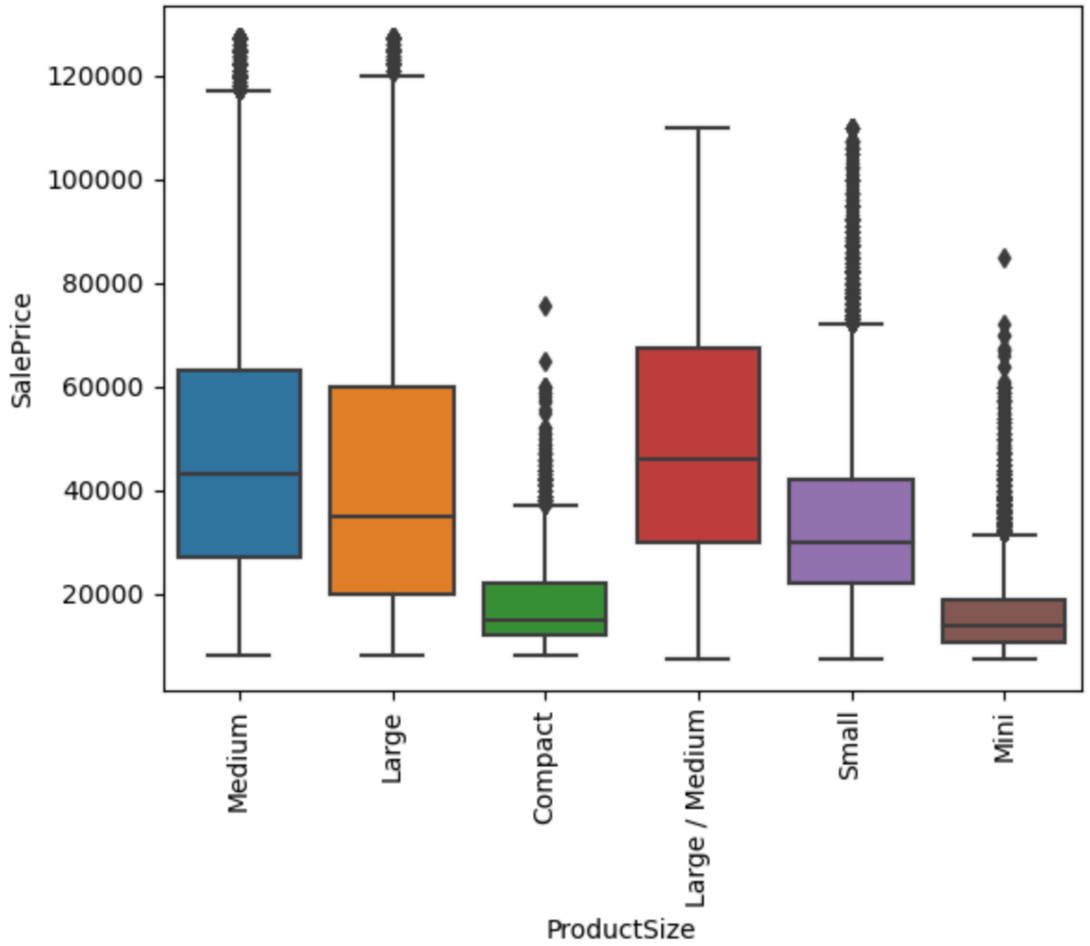
Data

Average sale price by year made



Data

Sale price by product size

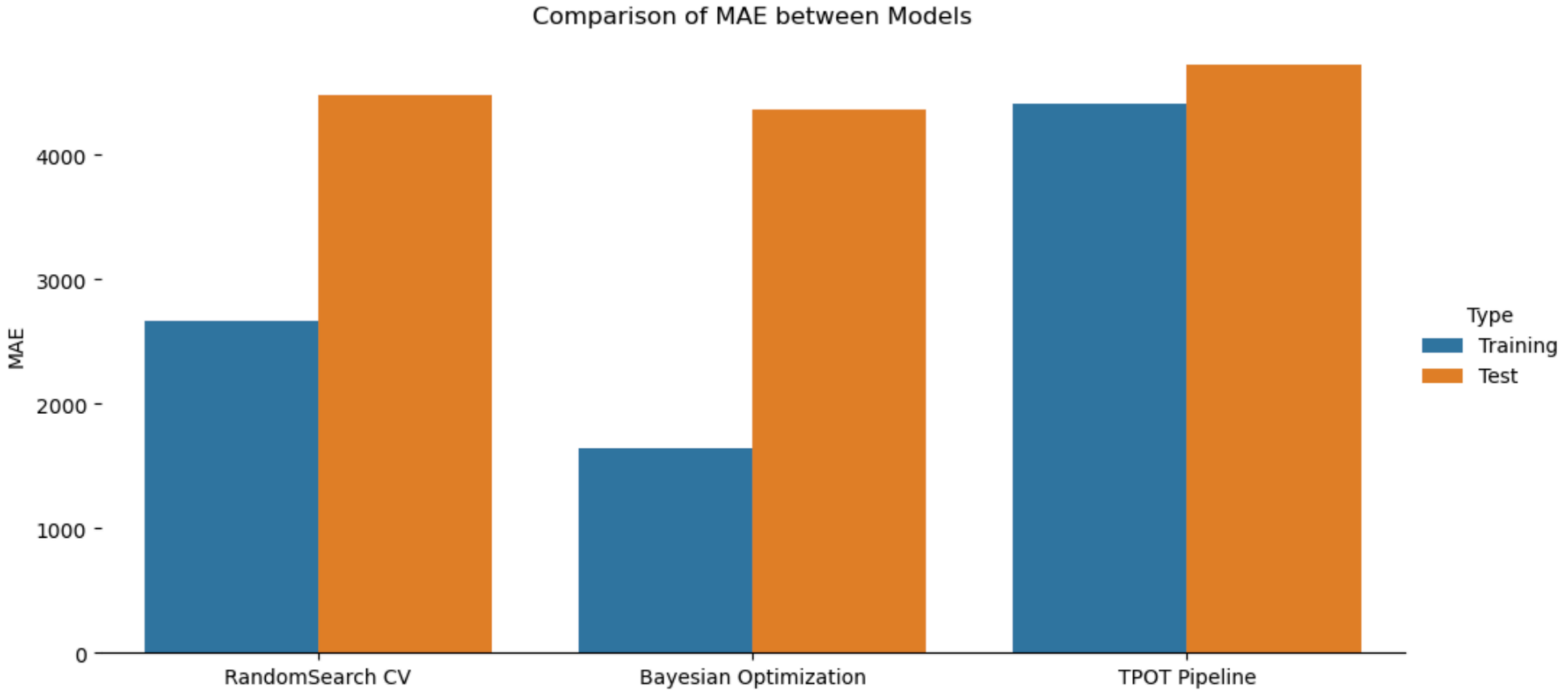


Modelling

Regression Algorithms used

- **RandomForestRegressor** with hyper parameters tuned via RandomSearchCV
- **RandomForestRegressor** with hyper parameters tuned via Bayesian Optimization
- **XGBRegressor** with hyper parameters tuned via TPOT

Modelling


Comparison of MAE between Models
MAE
4000
3000
2000
1000
0
RandomSearch CV
Bayesian Optimization
TPOT Pipeline
Type
Training
Test

Ideal Model

Scores & Precision

How well does our model predict against new, unseen data?

- **Mean Absolute Error**: $4,364.20
- **R2**: 0.91
- **Mean Squared Error**: 48,115,859.18

Ideas to Improve the Model in the Future

- As my training set only includes entries up to 2012, it would be nice to refresh this training data to reflect more current prices.
- Because I was tuning this through my personal laptop computer, prolonged hyperparamter tuning was difficult. To improve this model, I would rent virtual machine space to get a more exhaustive hyperparater tuning.
- Enriching this data with population growth metrics relative to the area in which the Bulldozer was sold my expose new patterns in sales prices