Linear Regression Comparison

Jacob Ellena

Data Scientist

jacob.ellena@gmail.com
linkedin.com/in/jellena/
github.com/jellena

So you want to automate the automaters

- What kind of effects can be seen if you cut the Data Science Department

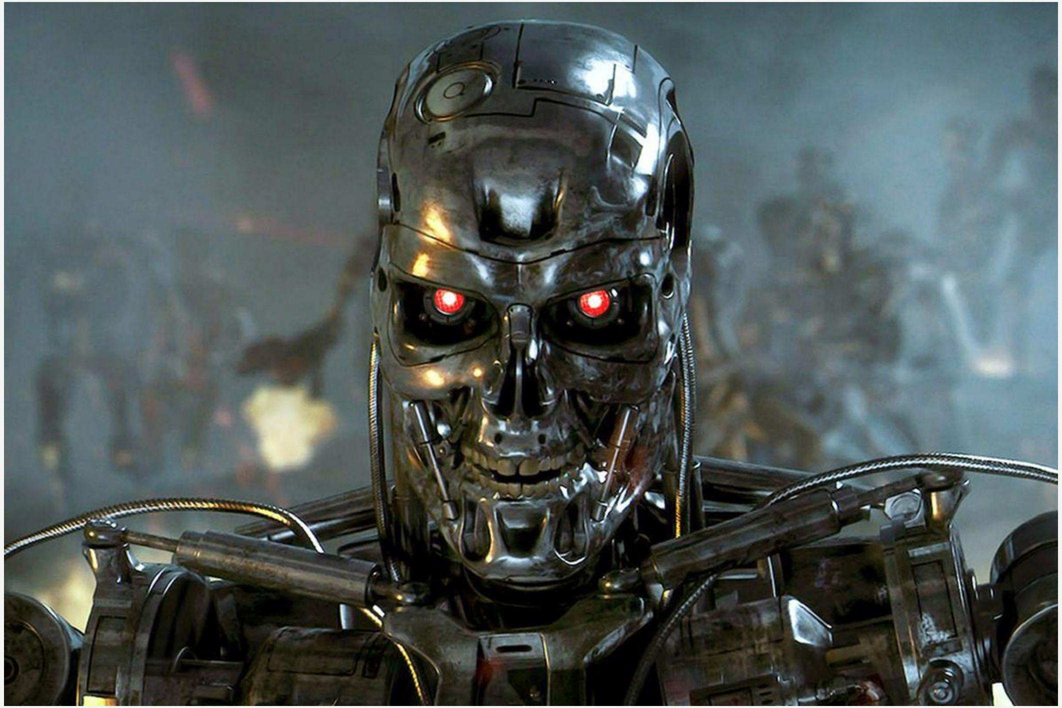
John Henry vs The Steam Drill

http://www.urchinmovement.com/2015/01/25/to-die-for-decoding-john-henry/

Garry Kasparov vs Deep Blue

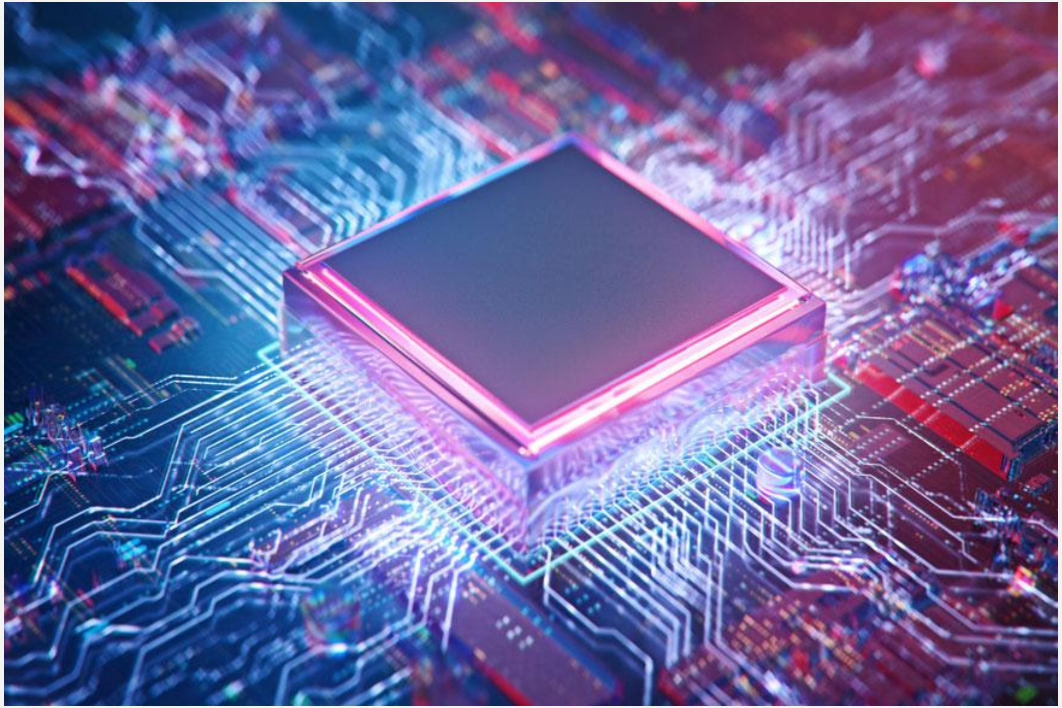
https://www.businessinsider.com/garry-kasparov-on-good-versus-great-in-chess-2017-5

Terminators vs Everyone



https://www.theverge.com/2017/9/27/16374734/terminator-sequel-release-date-2019-james-cameron-tim-miller-movie

Data Scientist vs Pure Processing Power



https://www.premiumbeat.com/blog/processing-power-resolve-vs-premiere-pro/

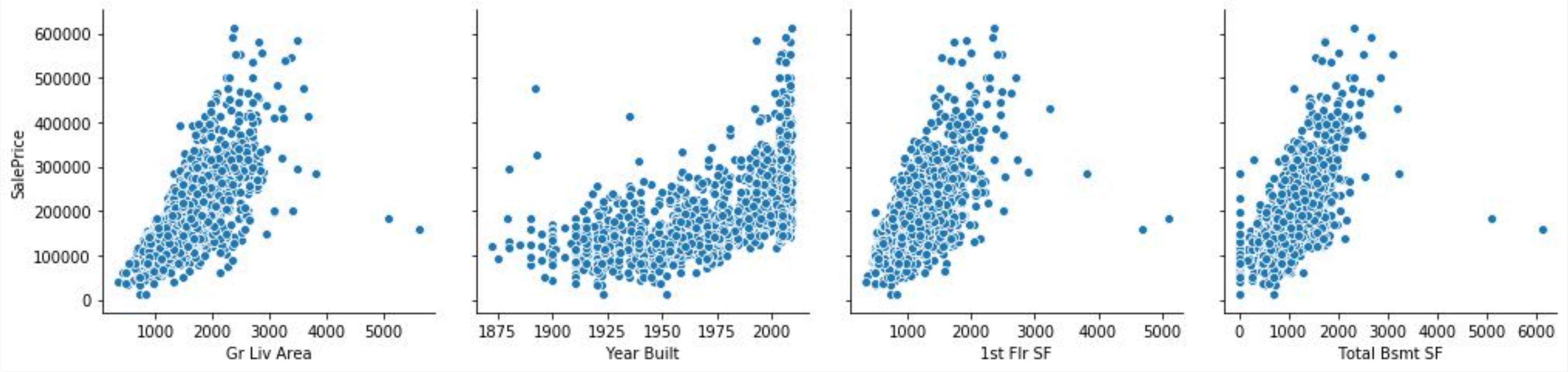
The Test

To compare the efficacy between brute force processing or human intuition three scenarios were created

Each is a Linear Regression will submit predictions to an unseen data set and be scored

The Human Regression

Data was analyzed and paired with outside reach features were manually selected



The Random Regression

A random number of random features was selected to run a regression

This process was repeated 1,000,000 times

The Combined Regression

The previously selected list is combined with a random number of selections

This process was repeated 1,000,000 times

The Results

Model Type	R^2 Score	Kaggle Score	Number of Features
Human	0.8449	37596.53	6
Random	0.7118	41270.68	17
Combination	0.8640	35887.03	16

Conclusion

- The highest performing model was the fusion of man and machine
- The Combination score was over 10% more accurate
- To remove human component of analysis at this point would be foolhardy

Further Research

- Compare levels of R^2 of different numbers of n
- Improved data cleaning with data transformations
- More feature engineering
- Develop random model to include all related dummies if one is selected
- Redevelop models with ElasticNet to account or large numerical data and numerous dummy columns

Questions?

Thank You