Flat-tery Will Get You Nowhere

The effects of dimensional reduction

What does a data scientist actually do?

1. We do $(synonym of science) with $(synonym of data)

What does a data scientist actually do?

1. We do $(synonym of science) with $(synonym of data)
2. We add value to your company (by magic presumably)

What does a data scientist actually do?

1. We do $(synonym of science) with $(synonym of data)
2. We add value to your company (by magic presumably)
3. Mumbles a bit…self deprecatory joke

What does a data scientist actually do?

1. We do $(synonym of science) with $(synonym of data)
2. We add value to your company (by magic presumably)
3. Mumbles a bit...self deprecatory joke
4. I try to draw lines in a clever way

A better answer

I’m going to take you through a classic data science problem and show you how I try to work through it.

The Ames Iowa Dataset

Data on houses sold in Ames Iowa

Contains 81 columns that cover an exhaustive list of characteristics

39 columns contain numeric data

A few kinds of data

Discrete: counts of something. The data typically consists of whole numbers.

Continuous: Measurements of something. Can be whole numbers or fractions

Ordinal: ranking of things. There’s a clear order to 1st, 2nd or 3rd place in a race, but they don’t actually measure or count things and you can’t do arithmetic operations with them

Categorical: Data that puts things in groups no clear order or arithmetic meaning

Linear Regression Model

A line that goes through the data and which has the minimum possible distance from each point.

This line can be in an arbitrary number of dimensions.

My first attempt

The first model I made simply ignored the non-numeric data

Underfit: this model did not have enough data to make good predictions

Actually came out really well: 82% of the variance in the model was accounted for in the features present. It did even better at predicting the test data: 85%

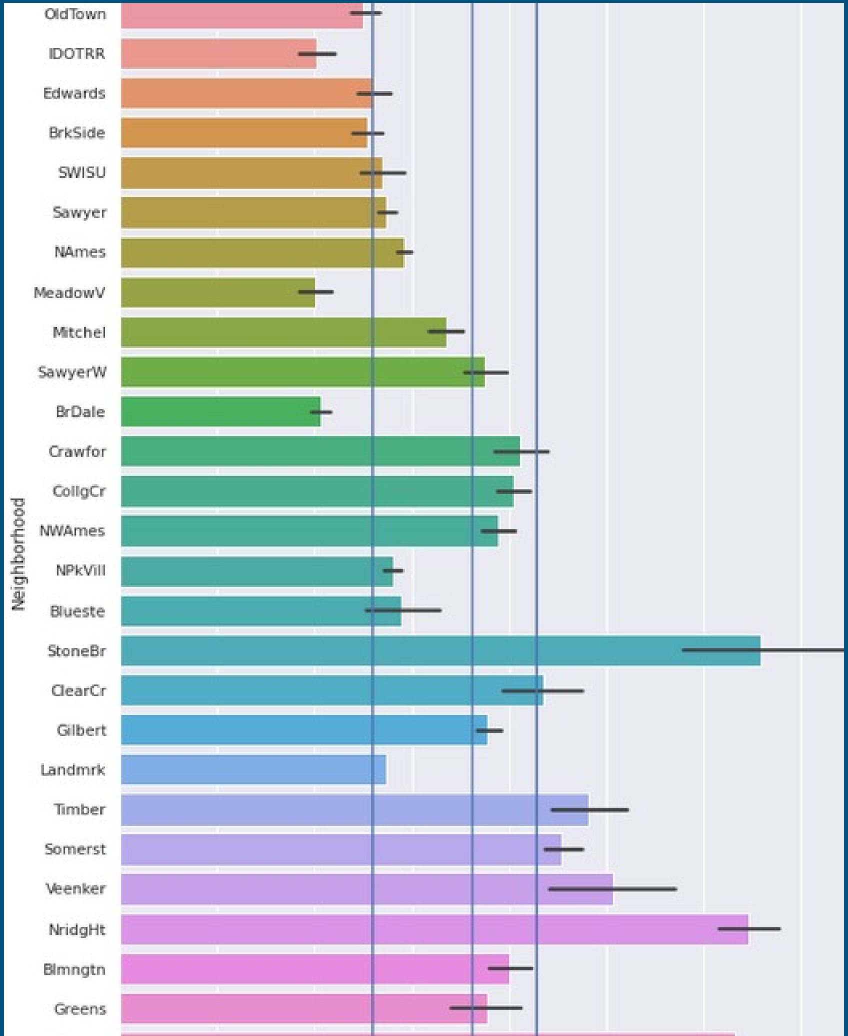
My second attempt

1. Take all categorical data and encode it with one hot encoding
2. 263 Features
3. Overfit..the model did a great job with the training data: 94% but a much worse job with new information 77%

Feature engineering

Attempt to reduce the total dimensionality of the data while retaining the meaningful data.

- Method 1, desirability score
- I grouped the data into four groups based on the mean sale price
- Turned 25 columns into 4



Method 2 :Aggregation

The data set included two categorical features pertaining to roof: Roof Style and Roof Mtrl

6 features each for a total of 36 values

I hypothesized that not all possible combinations were present

By grouping them together I was able to cu that t about 15

Results:
I was able to reduce the total dimensionality of the data by over half
This did not improve results over the simple dropping non numeric data

Why:

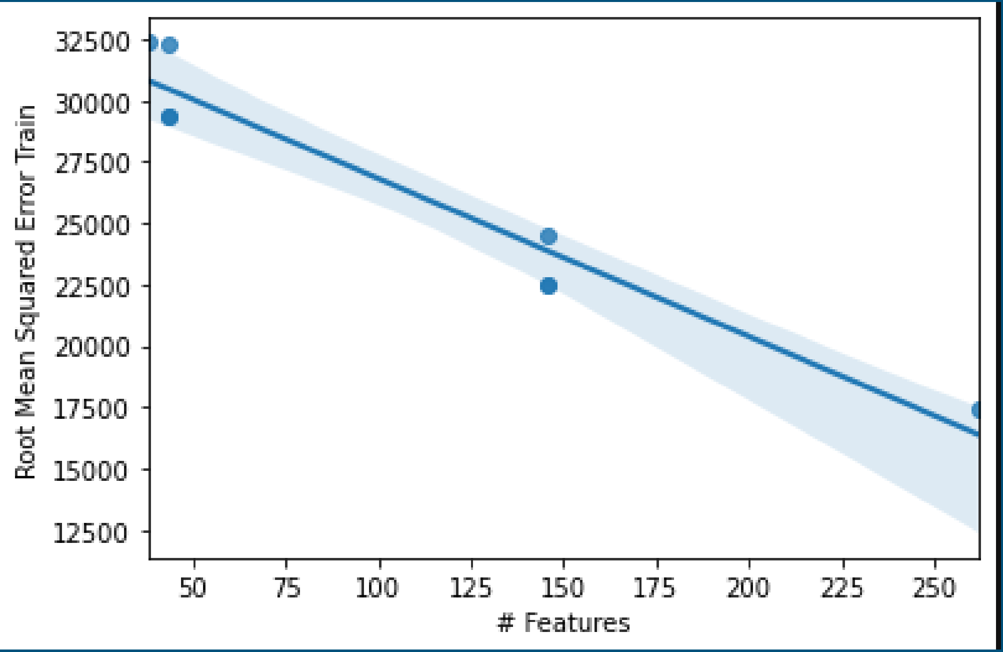
-There was a lot of numeric data

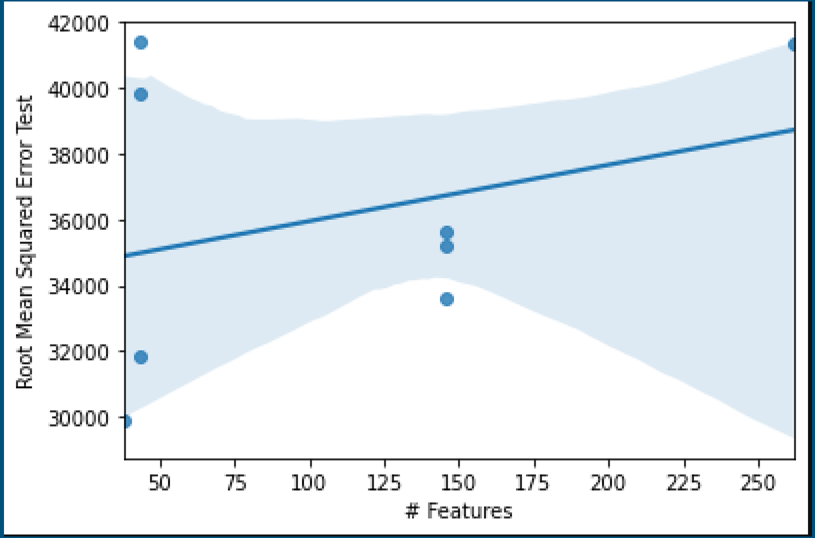
-Some of the data was likely redundant

Several of linear models compared:

	Model	# Features	R2 Training Data	R2 Test Data	Mean Squared Error Train	Mean Squared Error Test	Root Mean Squared Error Train	Root Mean Squared Error Test
0	Underfit	38	0.8274203264852614	0.8574404716022525	1047828067.7968353	894017622.2843026	32370.17250180844	29900.12746267652
1	Overfit	262	0.9480446618012022	0.7756839765035066	302902197.7328259	1709187344.2759871	17404.085662074463	41342.31904811324
2	reduced onehot	146	0.9480446618012022	0.7849465219114472	504518913.2209439	1267222410.866267	22461.498463391617	35598.06751589567
3	reduced numeric	43	0.9212572911121156	0.7894898250668314	859289303.1545429	1714364272.709657	29313.63681214842	41404.882232771255
4	reduced 1hot ridgeCV (Best)	146	0.85320553187025	0.7653648236174785	506663066.0746632	1240450578.9059782	22509.177374454695	35220.03093277998
5	reduced 1hot lassoCV (Best)	146	0.9209226427975585	0.8084733158525769	602339984.5117174	1128588612.4128618	24542.61568194632	33594.472944412475
6	reduced numeric ridgeCV(best)	43	0.853115423925284	0.7759227396051694	860149118.5720009	1586310994.7344778	29328.298937579057	39828.51986622749
7	reduced numeric lassoCV(best)	43	0.8056179127104846	0.8583320324631964	1043420588.3012797	1015110760.7379383	32302.021427478492	31860.80288909773

Dimensionality and fitness





Why am I sharing this?

- What I do is hard. It’s an iterative process
- I was seeing significant improvements, which I will use to further refine my model
- Specifically, I’m going to see what features are most correlated and filter to hopefully get to approximately 20 features encoding different data
- I’m also going to try to boost the signal with sensible combination of data
- You’ve already heard about how successful our methods can be. I wanted to show you a little more about how the sausage is made
- I hope you have a better understanding of the work that I do and that this is a WORK. It’s a painstaking process, even for experts.
- So you should contract with the best