Annotation and Feature Engineering

Introduction to Data Science Algorithms

Jordan Boyd-Graber and Michael Paul

HOUSES, SPOILERS, AND TRIVIA

Original Data

zip	state	county	2007	2013
10467	NY	bronx	335.2	294
60640	IL	cook	254.6	174.2
94109	CA	san francisco	707100	822.6
37211	TN	davidson	139800	141.7

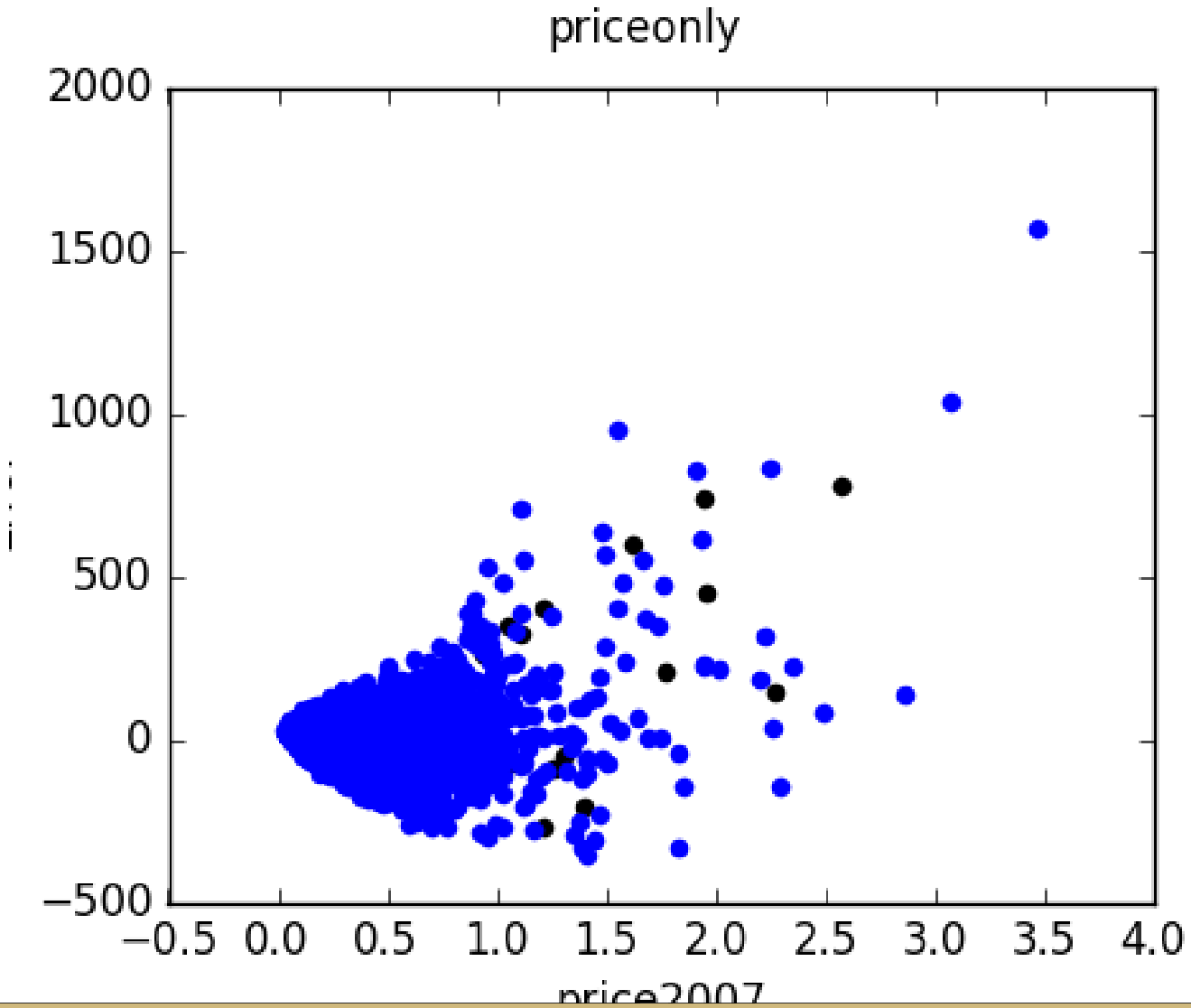
Goal: Predict post-crash prices in 2013 (in thousands)

Start simple . . .

$$p_{2013} = 911.4 \cdot p_{2007} - 20.1 \tag{1}$$

Error: 4302 (train), 4826 (dev)

Where are we making errors?



Add in square of 2007 price (capture outliers)

$$p_{2013} = 698.5 \cdot p_{2007} + 167.3 \cdot p_{2007}^2 + 18.7 \quad (2)$$

Error: 3424 (train), 3295 (dev)

Add in states?

$$p_{2013} = 781.8 \cdot p_{2007} + 137.5 \cdot p_{2007}^2 + 38.6 \quad (3)$$

State	β
AK	38.7
CO	8.7
FL	-99.0
IL	-44.7
MD	-61.3
TX	3.6
VA	-35.2

Error: 2579 (train), 2477 (dev)

Add Poverty Column

zip	state	county	poverty	2007	2013
10467	NY	bronx	27.1	335.2	294
60640	IL	cook	14.6	254.6	174.2
94109	CA	san francisco	10.6	707100	822.6
37211	TN	davidson	15.2	139800	141.7

Error was 2579 (train), 2478 (dev)

Now 2579 (train), 2486 (dev)

Add Poverty Column

zip	state	county	poverty	2007	2013
10467	NY	bronx	27.1	335.2	294
60640	IL	cook	14.6	254.6	174.2
94109	CA	san francisco	10.6	707100	822.6
37211	TN	davidson	15.2	139800	141.7

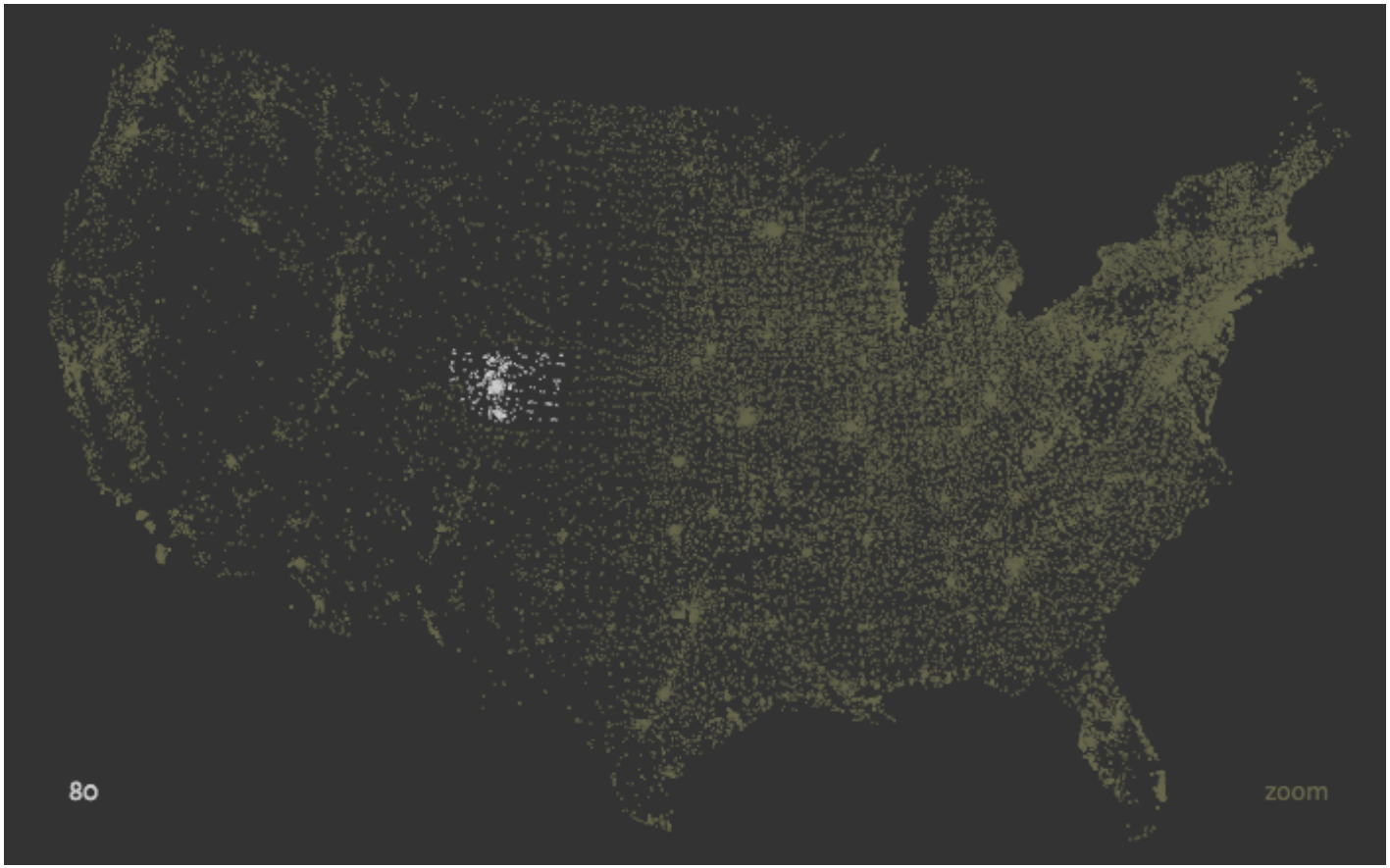
Error was 2579 (train), 2478 (dev)

Now 2579 (train), 2486 (dev)

States are too big, counties too small

$\beta = 51.9$

States are too big, counties too small



$\beta = 27.2$

States are too big, counties too small

(cannot fit without imputation)

States are too big, counties too small

Error: 2168 (train), 2062 (test)