General Assembly DSI Project 2
Ames, Iowa Housing Prices

Joshua Wilding
8/27/2018

GENERAL ASSEMBLY

What am I trying to accomplish?

- Predict housing prices in Ames, Iowa
- Ames is a town of about 66,000 in central Iowa
- Home of Iowa State University

How will I predict home prices?

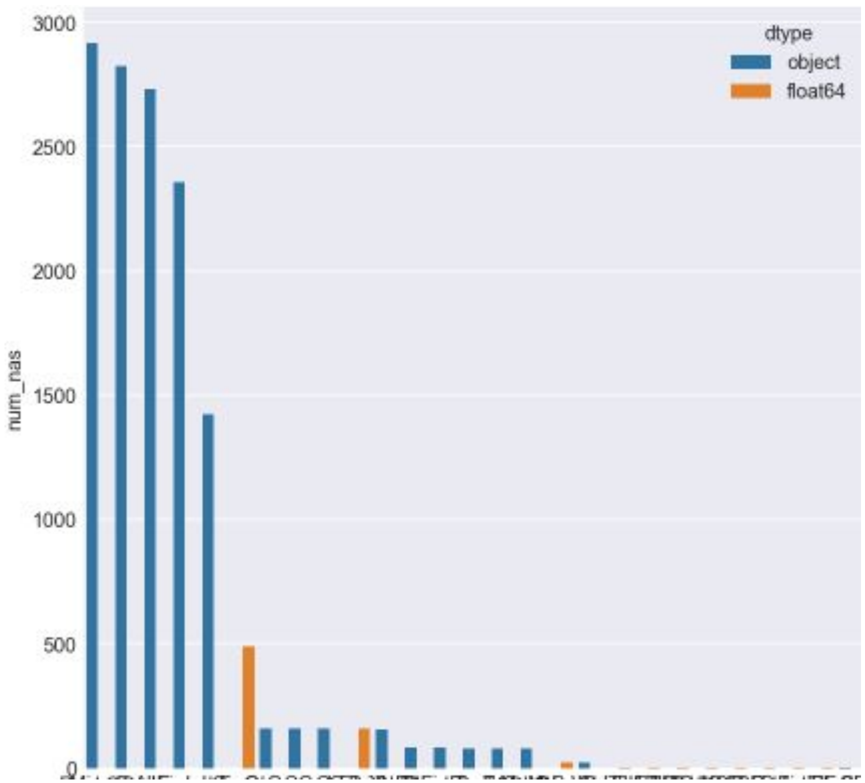
- Train a linear regression model using sale prices of 2,051 homes
  Use that to predict prices of 879 homes in test data.
- Dataset contains 80 columns of numerical and categorical data about homes sold

	Id	PID	MS SubClass	MS Zoning	Lot Frontage	Lot Area	Street	Alley
0	109	533352170	60	RL	NaN	13517	Pave	NaN
1	544	531379050	60	RL	43.0	11492	Pave	NaN
2	153	535304180	20	RL	68.0	7922	Pave	NaN
3	318	916386060	60	RL	73.0	9802	Pave	NaN
4	255	906425045	50	RL	82.0	14235	Pave	NaN
5	138	535126040	20	RL	137.0	16492	Pave	NaN
6	2827	908186070	180	RM	35.0	3675	Pave	NaN
7	145	535154050	20	RL	NaN	12160	Pave	NaN
8	1942	535353130	20	RL	NaN	15783	Pave	NaN
9	1956	535426130	60	RL	70.0	11606	Pave	NaN
10	1044	527451290	160	RM	21.0	1680	Pave	NaN
11	2752	906380150	20	RL	64.0	7488	Pave	NaN
12	807	906226060	70	RL	120.0	26400	Pave	NaN

Cleaning data

- Combined training and test data into one DataFrame for cleaning
- Many data fields contained NaN values, including both linear and non-linear columns
- Replaced all NaN values with 0
- Made sense for numerical fields
- Does not affect conversion of non-numerical fields to dummy columns

Number of NaN values by column



Feature engineering

- Created a DataFrame with dtype and number of unique values for each field
- Made dummy columns out of non-numerical fields and numerical fields with less than 10 unique values
- Standardized all other fields.
- Generated polynomial interaction terms for all columns, including dummies.

	column	unique_count	num_nas	dtype	make_dummy	standardize
6	Street	2	0	object	True	False
42	Central Air	2	0	object	True	False
12	Land Slope	3	0	object	True	False
10	Utilities	3	0	object	True	False
66	Paved Drive	3	0	object	True	False
51	Half Bath	3	0	int64	True	False
7	Alley	3	2732	object	True	False
49	Bsmt Half Bath	4	2	float64	True	False
28	Exter Qual	4	0	object	True	False
8	Lot Shape	4	0	object	True	False
53	Kitchen AbvGr	4	0	int64	True	False
61	Garage Finish	4	159	object	True	False
9	Land Contour	4	0	object	True	False
74	Fence	5	2358	object	True	False
41	Heating QC	5	0	object	True	False
33	Bsmt Exposure	5	83	object	True	False
73	Pool QC	5	2917	object	True	False
29	Exter Cond	5	0	object	True	False

Feature elimination

- Now have 39,000 features, including interaction terms and dummies
- Most are useless
- Calculated correlation coefficient between each feature and sale price
- Kept top 300 features based on strength of correlation

	feature	abs(correlation to SalePrice)
141	Overall Qual Gr Liv Area	0.837152
130	Overall Qual^2	0.825539
144	Overall Qual Garage Area	0.813247
131	Overall Qual Year Built	0.806902
132	Overall Qual Year Remod/Add	0.804740
5	Overall Qual	0.800207
142	Overall Qual TotRms AbvGrd	0.795420
138	Overall Qual 1st Flr SF	0.792151
137	Overall Qual Total Bsmt SF	0.768630
331	Gr Liv Area Garage Area	0.754659
342	TotRms AbvGrd Garage Area	0.719328
163	Year Built Gr Liv Area	0.716450
143	Overall Qual Garage Yr Blt	0.708581
184	Year Remod/Add Gr Liv Area	0.707879
292	1st Flr SF Garage Area	0.705365
330	Gr Liv Area Garage Yr Blt	0.697483

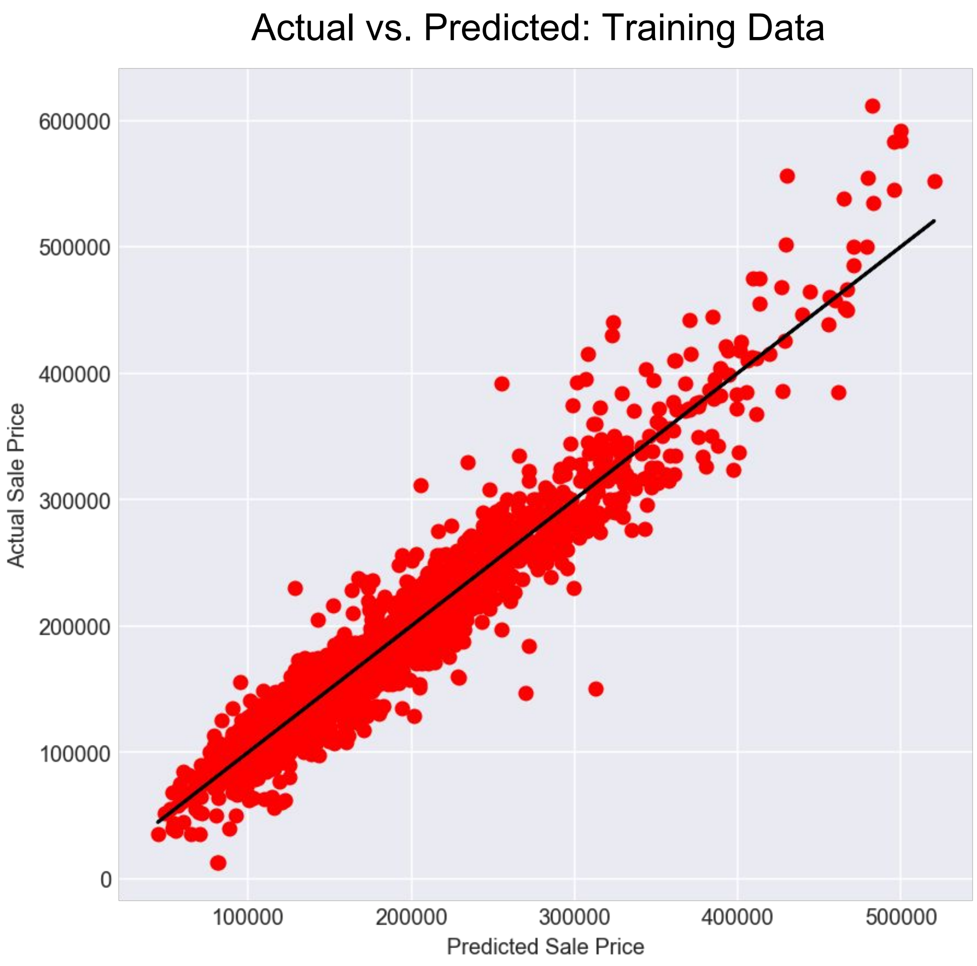
Running the model

- Chose a Lasso regulated linear regression model with 3-fold cross validation
- Ended up with 114 non-zero coefficients and an intercept of $191,545
- R^2 of 0.923 on training data.
- 5-fold cross validation of model reveals minor overfitting problem
- Average R^2 of 0.880

	feature	linear_coef	magnitude
0	Overall Qual Gr Liv Area	3.702670e+04	3.702670e+04
8	Overall Qual Total Bsmt SF	3.445434e+04	3.445434e+04
54	Exter Qual_TA	-2.191371e+04	2.191371e+04
105	Total Bsmt SF 1st Flr SF	-2.181223e+04	2.181223e+04
104	Mas Vnr Area Gr Liv Area	-2.011421e+04	2.011421e+04
129	Roof Style_Hip Garage Cars_3.0	1.781758e+04	1.781758e+04
259	Exter Qual_Gd Garage Cars_3.0	-1.718155e+04	1.718155e+04
7	Overall Qual 1st Flr SF	1.656464e+04	1.656464e+04
130	Half Bath_1 Garage Cars_3.0	1.650148e+04	1.650148e+04
244	Bsmt Qual_Ex Bsmt Exposure_Gd	1.629979e+04	1.629979e+04
34	1st Flr SF Gr Liv Area	-1.544256e+04	1.544256e+04
98	Mas Vnr Area TotRms AbvGrd	1.535333e+04	1.535333e+04
74	Mas Vnr Area Garage Area	1.497182e+04	1.497182e+04
51	Garage Cars_3.0 Garage Qual_TA	-1.458548e+04	1.458548e+04
114	Overall Qual BsmtFin SF 1	1.392208e+04	1.392208e+04

Results

- Actual vs predicted price is linear, but slightly trumpet shaped
- Could be better at predicted more expensive houses
- RMSE of 21,923 for training data
- RMSE of 24,289 for test data
- Model is a decent predictor of price



Things to Consider/Improve Upon

- Dealing with NaN values could have been more specific to each data field
- Dummy column selection criteria was arbitrary. 10 might not be right number of unique values
- Interaction terms between dummy values was bloated and clunky
- Selecting top 300 features based on correlation coefficients was arbitrary
- Not sure if that was the right amount
- Could have gotten to know the data better to explore relationships and understand real-world impact of each field

Thanks for listening.

Data Source: https://www.kaggle.com/c/dsi-us-5-project-2-regression-challenge/data

GENERAL ASSEMBLY