Ames, IA: A Window Into Pricing Tier 3 Cities Across USA

January 1, 2022

By: Harry Dzeba

Overview

- A private equity fund is looking for real-estate opportunities in America’s smaller cities

Problem:

- The Fund needs a data-driven house-pricing model to uncover undervalued assets
- The Fund is also unfamiliar with the value-maximizing factors in such towns.

Ames, IA

Project objectives:

Create pricing model

Linear Regression (in different versions) was used to distill more than 80 columns of data related to house characteristics into a model that accurately predicts house prices

Test the model

In addition to using the hold-out data, the model was entered into a Kaggle competition to make sure it's stacks well against the competitors

Use the model

Out of a few versions of the model, one was chosen to accurately gauge how much various house features affect the selling price.

Creating the model, part 1: Data Cleaning

Task

Initially, there were too many overlapping columns that would confuse the model

Cure:

- Drop unnecessary features
- Merge features to pack more info
- Transform features

Creating the models, part 2: Regression

- 3 basic regression models were created, OLS, Ridge and Lasso

- Generally speaking, the scores on all of them were high, with the r2 of around 0.9 (right)
- There was very little dispersion among the model in terms of the r2 score

Conclusion:

- **Such high and consistent scores on testing data confirm that the transformed data fed into the models is of high quality**

```
====== OLS =======
0.9036392752570962
0.8984750357094147

===== Ridge ======
0.8956861698373847
0.9007843705777905

===== Lasso ======
0.8957970302793162
0.8999868541547633
```

Testing the model with unseen data on Kaggle competition leaderboard
(our top score highlighted in red)

- Our error of $20,090 would've been in the top 15 of all the scores submitted

5 submissions for harry dzeba — Sort by Se

All Successful Selected

Submission and Description	Private Score	Public Score
kaggle4_project_model.csv 8 hours ago by harry dzeba simple regression where only log is log(y). to be used for the presentation	26634.36179	30538.37447
kaggle4_ridge.csv 4 days ago by harry dzeba same as kaggle2_ridge, except scoring in RidgeCV set to 'neg_mean_squared_error'	20090.41442	25615.98250
kaggle3_ridge.csv 4 days ago by harry dzeba Ridge, 30 columns, alpha = 39	20272.79776	25682.85384
kaggle2_ridge.csv 4 days ago by harry dzeba Ridge regression, only 20 features, alpha = 34	20121.94149	25660.75257
kaggle1.csv	20180.51681	25519.48133

1	▲ 22	Dereje Workneh	18012.90...
2	▲ 8	Scott Armstrong	18390.87...
3	▲ 21	Matt DeVay	18605.75...
4	▲ 7	Richard Ling	18611.59...
5	▼ 2	Griffin	18728.08...
6	▼ 2	weisja4	18870.77...
7	▲ 1	Marina Baker	19096.59...
8	▲ 20	Jon Godin	19154.74...
9	▲ 33	James Dargan	19194.90...
10	▼ 4	Stephanie Caress	19436.17...
11	▼ 9	rhys	19476.99...
12	▼ 3	Jeong Huh	19764.75...
13	▼ 8	Luke McKinley	19842.11...
14	▲ 53	ramin vafadary	19892.68...
15	▲ 5	Kalema Faisal	20135.01...

Model selection: 16 candidates

- R^2 in 0.89-0.90 for all model
- Test, train and cross-validate scores are consistent
- Model number 8 has the best Kaggle score with the fewest features, while number 16 has best test scores and is the most suited for everyday business use

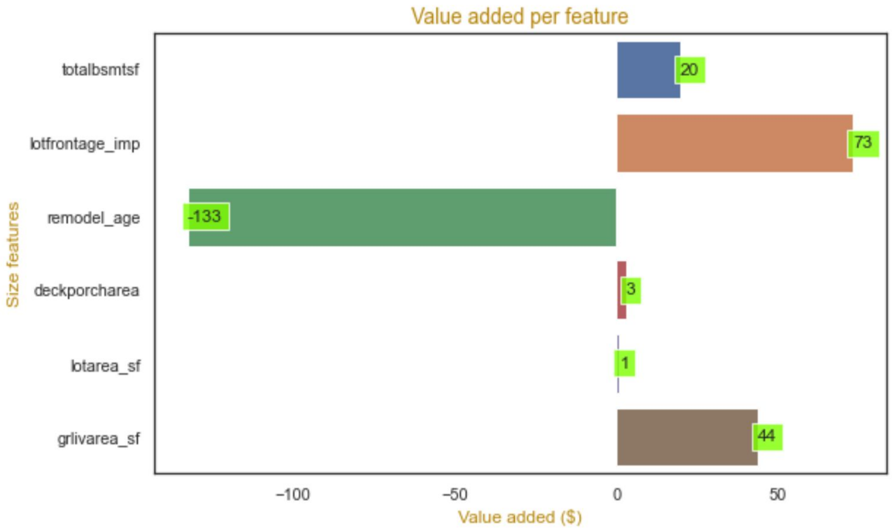
No	MODEL	FTRS	TEST R2	TEST RMSE	COMMENTS
1	LinReg	28	0.894	26173	nghbhood, deck and bsmt insignificant. shouldn't be
2	Same as 1 + dummy nghood	29	0.895	26601	nhood dummies are significant now
3	Same as 2 + undo deck/bsmt transformation	29	0.898	25154	big improvement in rmse. both bsmt and deck now significant
4	Same as 3 + cross-val k=8	29	0.899	23357*	- rmse improvement not to be trusted as cross_val_predict employs multiple models at the same time -cross_val does prove the model is solid
5	Same as 3 + Ridge	29	0.900	25119	- alpha=51, not very high - Ridge reg. didn't do much
6	Same as 3 + Lasso	29	0.897	25266	- alpha=0.021, insignificant -Lasso reg. didn't do much either
7	LinReg with optimal feature selector to cut down 9 features	20	0.900		- not strictly 'optimal' since not all among 10m combinations were tested - amazing result considering 9 features were cut
8	Same as 7 + Ridge	20	0.901	24926	- alpha=32. great result for so few features
9	Same as 7 + Lasso	20	0.900	24931	- alpha=0. no extra benefits from Lasso
10	LinReg with optimal feature selector to cut down 2 features	27	0.901		- no special improvement with 27 features optimized for highest r2 - unlike cutting 9 features - two, three or four features can be cut with all the possible combinations tested
11	LinReg with optimal feature selector to cut down 3 features	26	0.901		- again no special improvement
12	LinReg with optimal feature selector to cut down 4 features	25	0.902	24959	- the highest r2 - a lot of processing power required to try every 20-feature combo
13	Same as 12 + cross validation (k=8)	25	0.892		- no special improvement
14	Same as 11 + 4 dummies for years	30	0.902	25220	- year dummies add 4 columns but don't improve scores
15	Same as 14 + Ridge	30	0.902	25195	- Ridge regularization doesn't make year dummies more predictive
16	LogReg with only the dependent variable log transformed	28	0.903	24408	- the model chosen for the project - test scores better, kaggle worse than our top few model -reason for that is unknown, need more data. cross-val r2 is 0.892 - in line with other cross-val scores - overall scores about in line with other models, yet more interpretable

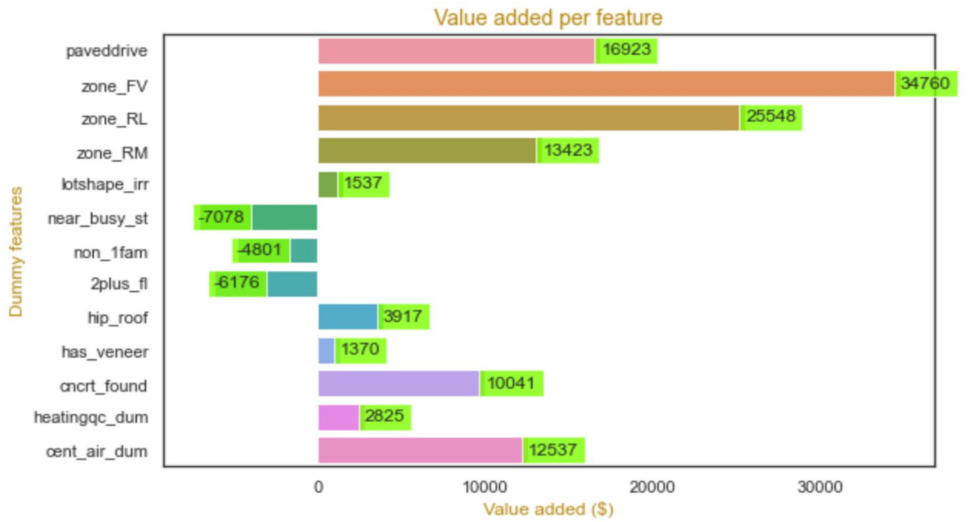
Using the model beyond price prediction

- scoring 1 point higher on garage and overall quality scales increases property value by more than $10,000.

Return on investment: which upgrade to choose?

- Before selling the house, the model can provide guidance as to which upgrade would be most profitable, after taking into account current cost estimates
- Central air and concrete foundation can increase average house's value by more than $10,000





In conclusion . . .

Recommendation 1	• The model shows that some features of the house such as garage and central air upgrades get unlock lots of value
Recommendation 2	• While relatively high degree of accuracy is necessary, the focus of this project should be interpretability and usability of the model
Recommendation 3	• Upgrades to central air, foundation, as well as garage and overall quality can yield high return
Recommendation 4	• The model can integrate future data, or be deployed to other towns, then be used to guide future buying/upgrading decisions

Further work to be done

As we add more data, we'll encounter new challenges:

1. As prices fluctuate and more years worth of data are added, we must find a way to adjust for macro house price data without burdening the model with extra dummy variables for each year
2. Test how the model performs in other (similar) cites. If the coefficients obtained are noticeably different, decide if that is due to prevailing local conditions, or just due to statistical noise.