Price predictions: Home Sales for Ames Iowa Housing

January 14, 2020
Prepared by: Noah Zuckerman
Prepared for: Kaggle Competition

Build a regression model to better predict house sale prices in the Ames region

Detailed housing dataset with 80 features describing many different elements of housing in Ames from roof material, to pool area to the configuration type of the lot

80 columns of features

	Id	PID	MS SubClass	MS Zoning	Lot Frontage	Lot Area	Street	Alley	Lot Shape	Land Contour	Utilities	Lot Config	Land Slope	Neighborhood	Condition 1	Condition 2	Bldg Type
0	2658	902301120	190	RM	69.0	9142	Pave	Grvl	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	2fmCon
1	2718	905108090	90	RL	NaN	9662	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	Duplex
2	2414	528218130	60	RL	58.0	17104	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam
3	1989	902207150	30	RM	60.0	8520	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Norm	Norm	1Fam
4	625	535105100	20	RL	NaN	9500	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam

The model predictions will be evaluated based on the Mean-Square-Error metric (e.g. accuracy)

Model building methodology

Data cleaning

- Data completeness
- Remove columns/features
- Replace NA values in test/train
- Remove NA values in train

EDA and Feature Engineering

- Label encode all string type features in train & test
- Investigate feature correlation
- Investigate cross-correlation between features
- Create interaction terms
- Identify outliers

Model Preparation & Preprocessing

- Train/test split
- Cross value score linear vs. lasso algorithm
- Scale data
- Feature analysis using model cross validation (linear, lasso, knn's) across multiple feature subsets

Model Evaluation

- Select two best performing feature subsets and scale down 'saleprice' as target vector
- Evaluate performance across linear, lasso, ridge and knn algorithms (both cross value score and fit & score)
- Evaluation metrics include R^2, MSE and test-train variance

Production Model Selection

- First selected Lasso model due to low bias (90% on train data) and low relative variance (3.5% variance)
- Re-evaluated selected model in order to improve model ability to generalize across other housing datasets
- Selected Ridge model and tuned hyperparameters using gridsearch

Conclusions and Recommendations

- Created two submissions of price predictions with respective MSEs of 35,970 and 37,449
- Identified desirable housing features with large impact on price
- Identified most desirable neighborhoods for prospective buyers

Modeling and Model Evaluation

Model Eval 1

- 6 highest correlation features
- LinearRegression
- Train score = 0.753

Model Eval 2

- 2 highest correlation features
- 2 interaction features
- LinearRegression
- Train score = 0.765

Model Eval 3

- 14 most weighted features from lasso
- LinearRegression
- Train score = 0.78

Model Eval 4

- 4 datasets
- LinearRegression
- Best train score = 0.78

Model Eval 5

- 2 datasets
- LR, Lasso, Ridge, KNN
- **Lasso tr = 0.87 D1**
- **Lasso te = 0.91 D1**
- **Ridge tr = 0.884 D1**
- **Ridge te = 0.885 D1**

Production Model 1

- Y = log(saleprice)
- Full dataset (D1)
- Lasso
- Hyperparameter tuning
- Train score = 0.848
- Submission MSE = 35,971

**best KAGGLE score*

Production Model 2

- Y = log(saleprice)
- Full dataset (D1)
- Ridge
- Hyperparameter tuning
- Train score = 0.849
- Submission MSE =
- 37,450

Primary Findings

- Selected the ridge regression model as production model due to its ability to achieve ~85% R^2 while minimizing variability to < 1%
- This model also automates the requirement of feature engineering which increases the efficiency of model building giving a dataset of ~80 features
- Through exploratory analysis of model coefficients it is clear that the features with the largest impact on home price are unsurprisingly:

Total home size

- Lot area
- Total basement sqft
- 1st floor sqft
- 2nd floor sqft
- Ground living area
- Garage # of cars

Condition of home

- Year built
- Overall quality of home
- Overall condition of home
- Year re-modelled
- Home functionality rating

Total home size

- Basement full bath
- Number of fireplaces
- Miscellaneous feature values
- Roof Material
- Number of Full Baths
- Total rooms above ground
- Screen Porch

**Caveat to this is that the model shows these features are correlated to higher prices but that does not necessarily indicate their relationships is causal. It could be possible that large highly priced homes with higher sqft just happen to already have these features on average. The model could be picking up and predicting this trend.*

Recommendations for homebuyers

1 Consider these elements

```
{'lot_area': 0.021209554049469427,
 'overall_qual': 0.10538490910608841,
 'overall_cond': 0.0445330351951265,
 'year_built': 0.050634491121384496,
 'year_remod/add': 0.027639280007791137,
 'roof_matl': -0.021945790403600066,
 'total_bsmt_sf': 0.021555123343044953,
 '1st_flr_sf': 0.04180340954290668,
 '2nd_flr_sf': 0.027345655502209285,
 'gr_liv_area': 0.05625692945160828,
 'bsmt_full_bath': 0.027538767217553346,
 'full_bath': 0.02102395327691179,
 'totrms_abvgrd': 0.018552082519389765,
 'functional': -0.019769242093420156,
 'fireplaces': 0.023675099059379003,
 'garage_cars': 0.028639748829315324,
 'screen_porch': 0.019632765993459134,
 'misc_val': -0.02606914882902194}
```

2 Consider these neighborhoods

Neighborhood	
StoneBr	329675.736842
NridgHt	322831.352459
NoRidge	316294.125000
GrnHill	280000.000000
Veenker	253570.588235
Timber	241051.354167
Somerst	227183.900000
ClearCr	217490.074074
Crawfor	205901.211268
CollgCr	202497.216667
Blmngtn	200417.681818

Next steps...

Given additional time:

- Additional feature engineering using PIPE method and polynomials
- Further refine algorithm and fine tune hyperparameters using gridsearch
- Investigate other algorithms to use