House Value Estimate

Project 3 - By Carol

BACKGROUND

I have just joined a new “full stack” real estate company in Ames, Iowa.

My goals for this year:

1. Develop an algorithm to reliably estimate the value of residential houses based on fixed characteristics.

2. Identify characteristics of houses that can be cost-effectively renovated. And evaluate the mean dollar value of the renovations.

TABLE OF CONTENTS

ABOUT

AMES HOUSING DATA

1460 rows, 81 columns

38 numerical, 43 categorical features

Fixed and Renovate-able characteristics

DATA PREPROCESSING

Inspect **missing values**, and fill them with appropriate values. (0, mean or most frequent)

Remove non-residential entries from the dataset

convert **categorical columns** to dummy coded columns

Use StandardScaler to **standardize the predictors**

DATA PREPROCESSING

- Different tools for preprocessing

missing values

pandas .fillna()

Sklearn .simpleImputer(strategy=)

categorical columns

pandas .get_dummies()

Sklearn .OneHotEncoder()

BUSINESS ACUMEN

VS

CORRELATION MATRIX

’OverallQual’
’OverallCond’
’YearBuilt’
’YearRemodAdd’
’GrLivArea’
’FullBath’
’Bedroom’
’KitchenQual’
’TotRmsAbvGrd’

50% of my guessing are in the correlation matrix

Q: Can I 100% trust my business acumen? No.

feature	corr_with_SalePrice
OverallQual	0.789031
GrLivArea	0.708658
GarageCars	0.642135
GarageArea	0.629831
TotalBsmtSF	0.612088
1stFlrSF	0.603300
FullBath	0.556674
TotRmsAbvGrd	0.533355
YearBuilt	0.517602
YearRemodAdd	0.501186
MasVnrArea	0.471056
Fireplaces	0.463264

Fixed features

STRATEGY 1

Included full set of features to avoid missing any important features.

Renovate-able features

STRATEGY 2

Removed irrelevant features before modeling (VarianceThreshold)

FEATURE ENGINEERING (continue)

How feature engineering change the performance of the model
(for Renovate-able features)

Used full set features	Num of features: 142 Model score: 0.034	**Features 40 ↓**
VarianceThreshold (threshold=0.006)	Num of features: 102 Model score: 0.057	**Score 68% ↑**

MODELING - FIXED FEATURES

MODEL	MEAN SCORE
Ridge / RidgeCV	0.80 / 0.81
Lasso / LassoCV	0.79 / 0.70
ElasticNet / ElasticNetCV	0.75 / 0.58

STRATEGY: Try all the models, and find out the one performs the best.

MODELING - FIXED FEATURES (continue)

My choice - Lasso

```
Lasso_Cross validated scores: [0.84969164 0.79471072 0.81455472 0.84866657 0.66703215]
Lasso_Cross validated mean score: 0.7949311595456072
```

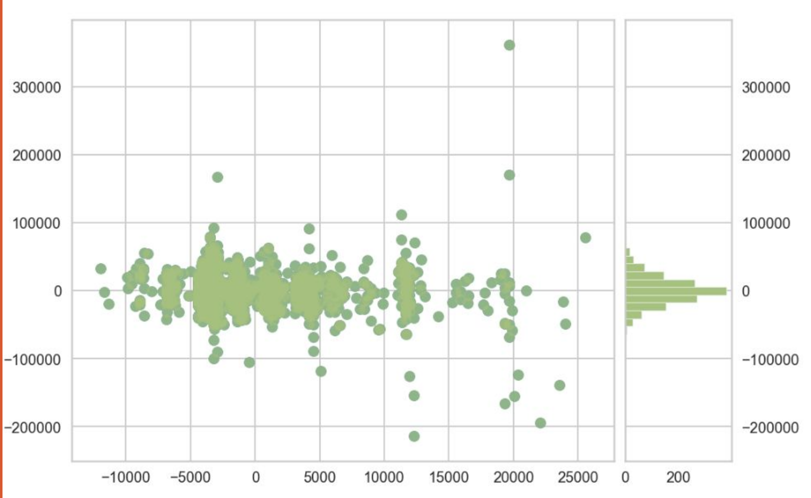
```
# Check what features have been exclued by Lasso model
coefs = list(l.coef_)
coefs_list = list(zip(features, coefs))

coefs_zero = []
for elem in coefs_list:
    if elem[1] == 0.0:
        coefs_zero.append(elem)
coefs_zero
```

```
[('MSZoning_RL', 0.0),
 ('LandContour_Lvl', -0.0),
 ('LotConfig_Inside', 0.0),
 ('Condition1_Norm', 0.0),
 ('MasVnrType_Stone', 0.0),
 ('Foundation_CBlock', -0.0)]
```

My choice - LassoCV

```
LassoCV_Cross validated scores: [0.06829743 0.02228852 0.08197088 0.07966626 0.03511698]
LassoCV_Cross validated mean score: 0.057468014082305084
```

Use ResidualsPlot (from YellowBrick) to check if the linear regression model appropriate for the data



DECISION MAKING SUPPORT

FEATURE IMPORTANCE

When deciding whether or not buy a house, make sure the important features are fully evaluated and taken into consideration because they greatly contribute to the house price

TOP 20 FIXED FEATURES

```
[('GrLivArea', 29986.09196485551),
 ('GarageCars', 11614.585600937358),
 ('BsmtQual_Ex', 11314.178190257364),
 ('Neighborhood_NridgHt', 10135.1688997426),
 ('2ndFlrSF', 8864.175183587739),
 ('Neighborhood_NoRidge', 8585.569713848436),
 ('Neighborhood_StoneBr', 6940.0279197283935),
 ('HouseStyle_1Story', 6625.856411009754),
 ('BsmtFullBath', 6316.079036167664),
 ('LotArea', 6090.373274204094),
 ('YearRemodAdd', 6088.571828581061),
 ('Condition2_PosN', -5943.695124451557),
 ('KitchenAbvGr', -5712.594972942175),
 ('GarageType_Attchd', 5520.632547530351),
 ('Neighborhood_Edwards', -5293.274748266839),
 ('YearBuilt', 5109.390062067734),
 ('BedroomAbvGr', -4710.065329694738),
 ('Fireplaces', 4573.545070287801),
 ('HouseStyle_2Story', -4520.536369540194),
 ('GarageYrBlt', -4446.4860057672795)]
```

TOP 10 RENOVATE-ABLE FEATURES

y top features

Weight?	Feature
+3684.693	KitchenQual_Ex
+2668.091	OverallCond
+2136.466	BsmtExposure_Gd
+989.416	Exterior1st_BrkFace
+425.248	OverallQual
+227.502	ScreenPorch
+4.092	Exterior2nd_ImStucc
-12.460	BsmtFinType1_Unf
-131.656	BsmtExposure_No
-334.840	<BIAS>

DECISION MAKING SUPPORT (CONTINUE)

ESTIMATE HOUSE PRICE

Use the first models I to estimate the price for a target house, and then use the estimated price as the budget guide to decide whether or not purchase the target house

Estimated price = prediction from the 1st model

RENOVATION BUDGETING

If the company aim to renovate a house, we can use the second model to evaluate the effect on house price change after the renovation.

Estimated price change = Predicted residuals - residuals before renovation

The budget of renovation should less than (Estimated price change - acceptable profit of the project)

CREDITS

MASUM RAB

COOL IDEAS

- Statsmodels for modeling
- SimpleImputer
- OneHotEncoder
- VarianceThreshold
- Pipeline
- FeatureImportances(yellowbrick.model_selection)
- ELI5

THANKS!

Do you have any questions?

Please contact: carolcheng.au@gmail.com