Modeling house prices

Presented by:
Jason Wu

EDA

- Removed any houses that weren't residential

```
house[house.MiscFeature =="Othr"]
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal
705	706	190	RM	70.0	5600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	Othr	3500
873	874	40	RL	60.0	12144	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	Othr	0

- Removed variables that were made up of predominantly one value

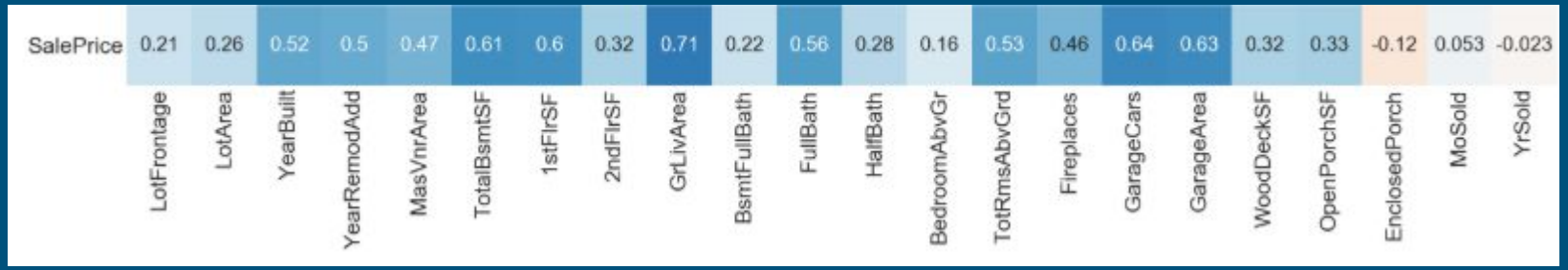
```
(count      1448
 unique        2
 top        Pave
 freq       1444
 Name: Street, dtype: object, count       1448
 unique          2
 top        AllPub
 freq         1447
 Name: Utilities, dtype: object)
```

- Original dataframe dimensions: 1460, 81
- New dataframe dimensions: 1448, 69

Can a house’s fixed variables predict price?

- Choosing the variables were highly subjective (roof style changeable?)
- Assigning the correct values to the different missing values
  - Missing area values should be “0”
  - Missing features should be “None”
- Dependant variable type is continuous
  - Regression problem

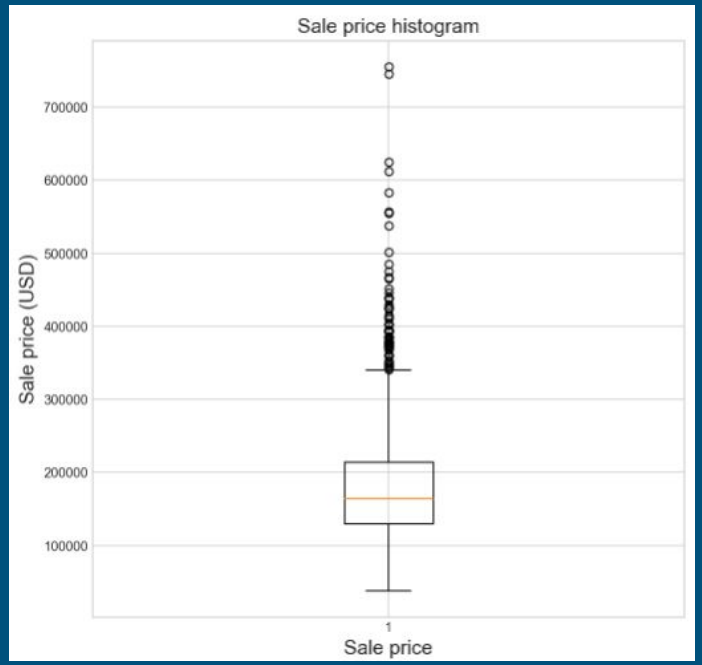
Can a house's fixed variables predict price?

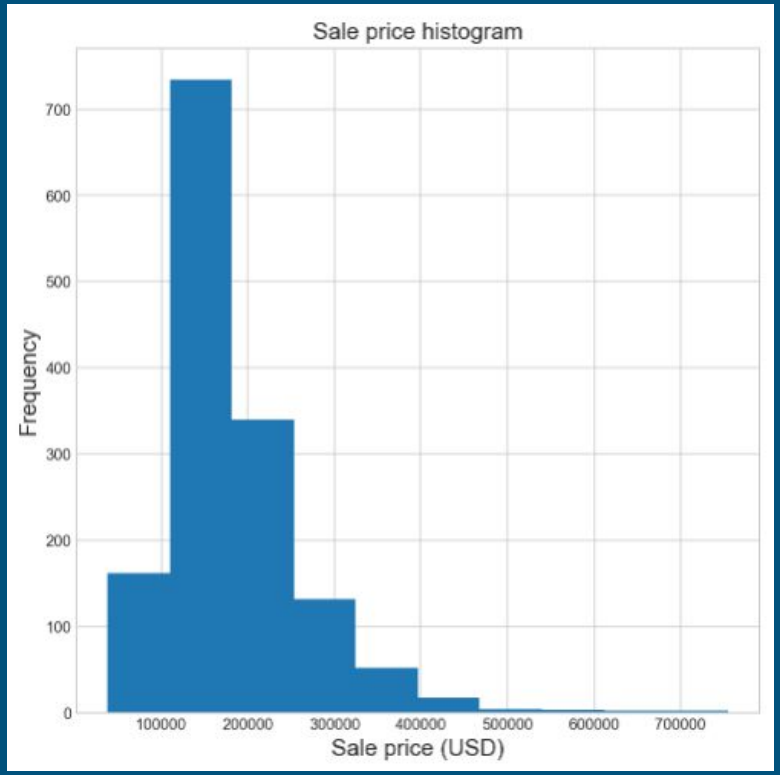


- Most continuous independent variables are positively correlated with sale price

Can a house's fixed variables predict price?

- There are a few extreme house prices





Can a house’s fixed variables predict price?

Feature engineering

- Collated all the living spaces up and made a total living area variable
- Age of a house when sold
- Age of a house when renovated
  - Age of a house when renovated has a slight negative correlation to sale price

Can a house’s fixed variables predict price?

Fitting a model

- Lasso regression performed best in terms of R2 value for linear regression models.
  - Training R2=0.83
  - Testing R2=0.86

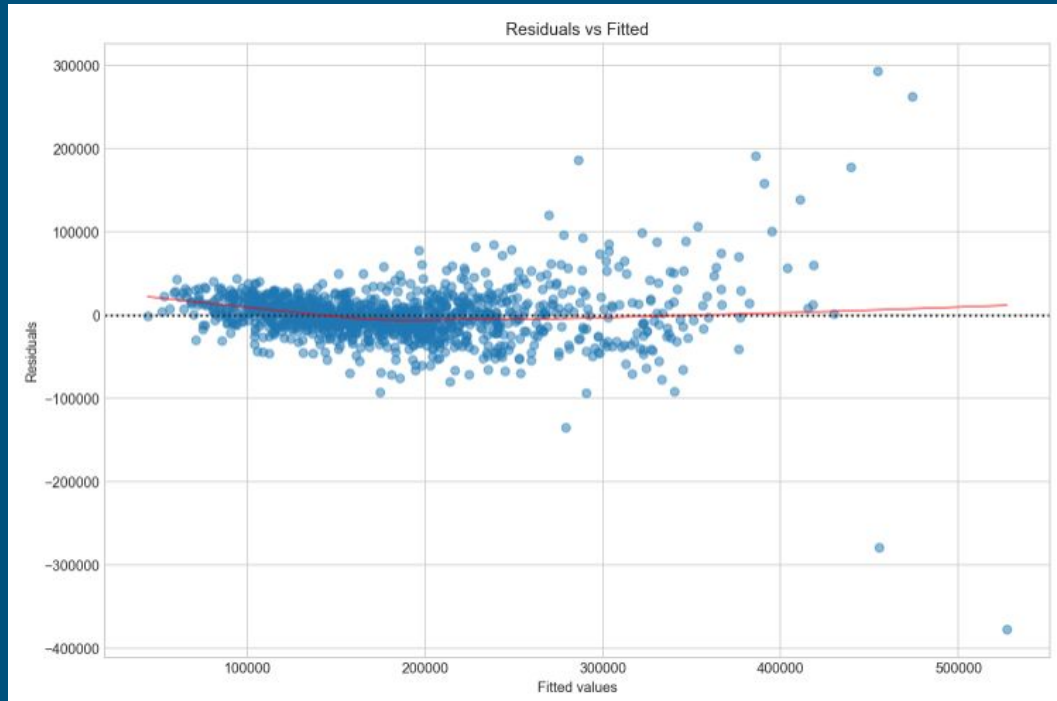
- Random forest regressor returned R2 of 0.96 for training and 0.88 for testing

Lasso model will be better suited for answering the question

Can a house's fixed variables predict price?

Why Lasso?

- Lasso reduced variables from 136 to 68
  - more interpretability
- Random forest model showed signs of overfitting
  - test R2 > train R2
- The relationship between dependant and independant appears to be linear



Can a house’s fixed variables predict price?

Improving the model

- Linear regression is susceptible to being heavy influenced by outliers
  - Removing them may help improve the model
- New Lasso R2 0.88 for both train and test
  - Suggests the new model is more generalized

What fixed variables affect house price the most?

	variable	coef	abs_coef
15	TotalLivArea	34358.729399	34358.729399
56	Neighborhood_NridgHt	9430.472451	9430.472451
2	MasVnrArea	8554.974379	8554.974379
17	Age	-7994.106395	7994.106395
62	Neighborhood_StoneBr	7130.434648	7130.434648
128	SaleType_New	7114.891839	7114.891839
16	Ageofsale	-6757.153044	6757.153044
1	LotArea	6126.549966	6126.549966
55	Neighborhood_NoRidge	5727.011277	5727.011277
6	BedroomAbvGr	-5672.576542	5672.576542

- Engineered features play a big role in price prediction
- 1 square feet increase in Total living area increases house price by US$34,358
- Every year since built date, a house depreciates by US$7,994
- Northridge Heights most expensive neighborhood

What should we renovate?

- Most remaining variables used
- Effectiveness of renovation measured against the error of our fixed variables model (SSE)
- Most of these variables have rankings associated with them and were changed accordingly

What should we renovate?

- Ridge regression used because I wanted to see the effect of all the rennovatable features

```
Ridge train score, 0.160639486956606 13
Ridge test score, 0.09102209625943125
```

- Low R2 values for both, the lower test score suggests the model may be slightly overfit with the training data

What should we renovate?

- Each unit increase of the Overall Condition of a house will **increase** the value by US$6,085
- Each unit increase in Basement Condition **decreases** the value by US$3,002
- Increasing the living conditions of a finished Basement does increase the price but doesn't make up for the drop
    - No point renovating basement

	variable	coef	abs_coef
1	OverallCond	6085.216377	6085.216377
0	OverallQual	4621.634371	4621.634371
6	BsmtExposure	3058.163601	3058.163601
5	BsmtCond	-3002.343004	3002.343004
7	BsmtFinType1	2950.796699	2950.796699
2	ExterQual	1568.669979	1568.669979
25	Exterior1st_CemntBd	-1087.712686	1087.712686
13	FireplaceQu	-1054.135505	1054.135505
32	Exterior1st_VinylSd	-936.894407	936.894407
3	ExterCond	-841.891865	841.891865

What should we renovate?

I believe the focus should be on increasing kitchen quality. This will return US$792USD for every unit increase and can be done with cheap refurbishment.

Although Exterior quality yields a good return US$1568 for every improvement, I believe the cost to improve it will outweigh the return.

Fire places don't appear to add any value to a property and actually decrease a property's value by quite a bit.

2	ExterQual	1568.669979	1568.669979
25	Exterior1st_CemntBd	-1087.712686	1087.712686
13	FireplaceQu	-1054.135505	1054.135505
32	Exterior1st_VinylSd	-936.894407	936.894407
3	ExterCond	-841.891865	841.891865
40	Electrical_SBrkr	-804.455530	804.455530
9	Fireplaces	-799.807483	799.807483
18	RoofStyle_Hip	796.954163	796.954163
12	KitchenQual	792.694238	792.694238