Project 2:
Predicting Ames Housing Prices

Team 4: Haziq, Guo Jun, Luka, Nixon, Ram, Saloni

Content

- Problem Statement
- Background
- Model Workflow pt. 1
  - Data cleaning
  - Feature engineering
  - Feature selection
- Model Workflow pt. 2
  - Model Prep
  - Model fit and evaluation
  - Model Improvement
- Conclusions & Recommendations

Problem Statement

- Tendency to focus on price per square foot (or meter) rather than the house as a whole

Our aim:

- Identify features that affect housing prices
- Build a model that effectively predicts sale prices of house in Ames
- Audience: Real estate agents

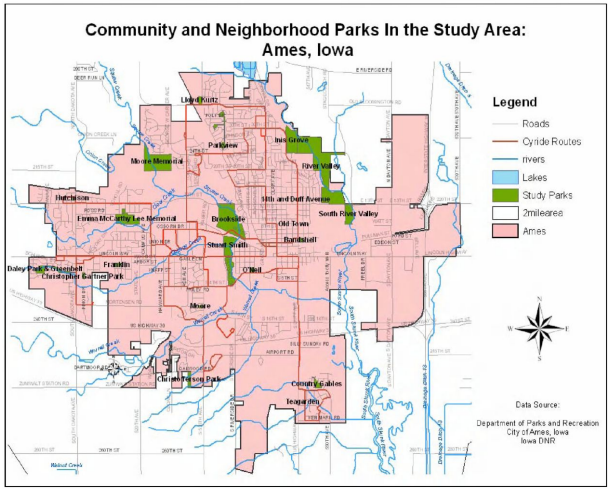
Model based on

2051
residential properties

81
house features

2006-2010

Some features



Price

How much are the houses?

Source: Ames City Assessor

$180,921

Average sale price

$12,789

Cheapest sale

$611,657

Most expensive sale

Neighborhood

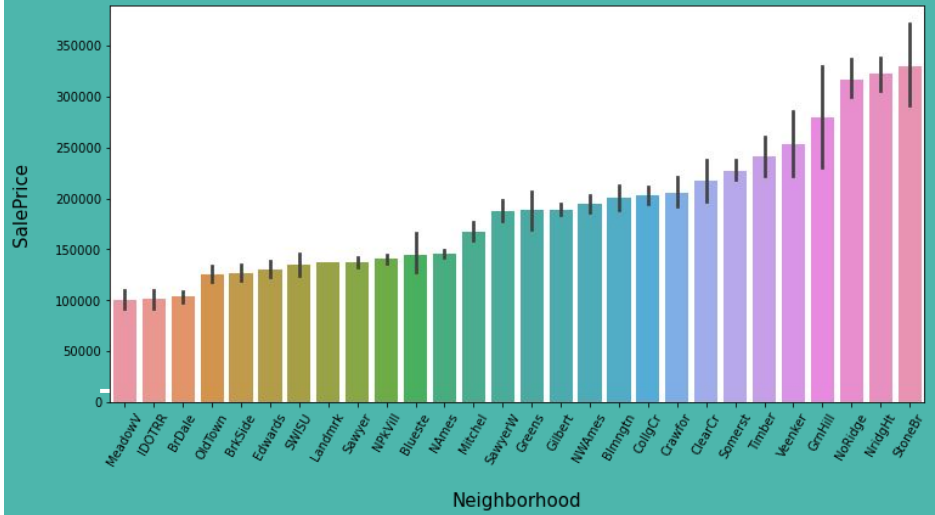
Which are the most expensive neighborhoods?

Source:
Remax Property
Rent.com

1. Stone Brooke
2. Northridge Height
3. Northridge

...

28. Meadow Village



Area

How big are the houses?

Source: Ames City Assessor

1499

Average area

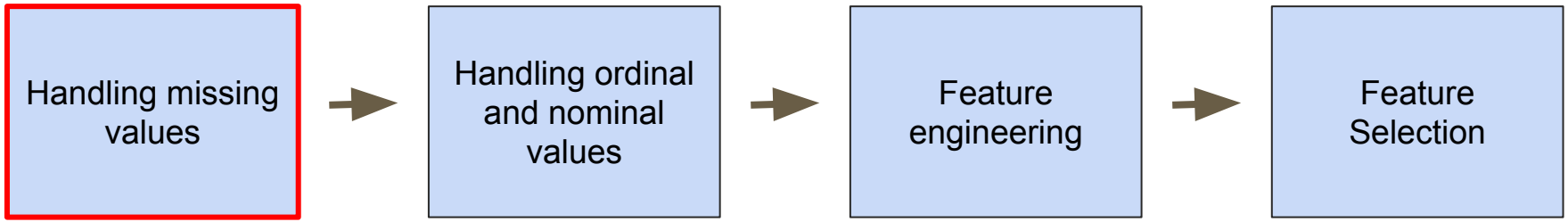
334

Smallest area

5642

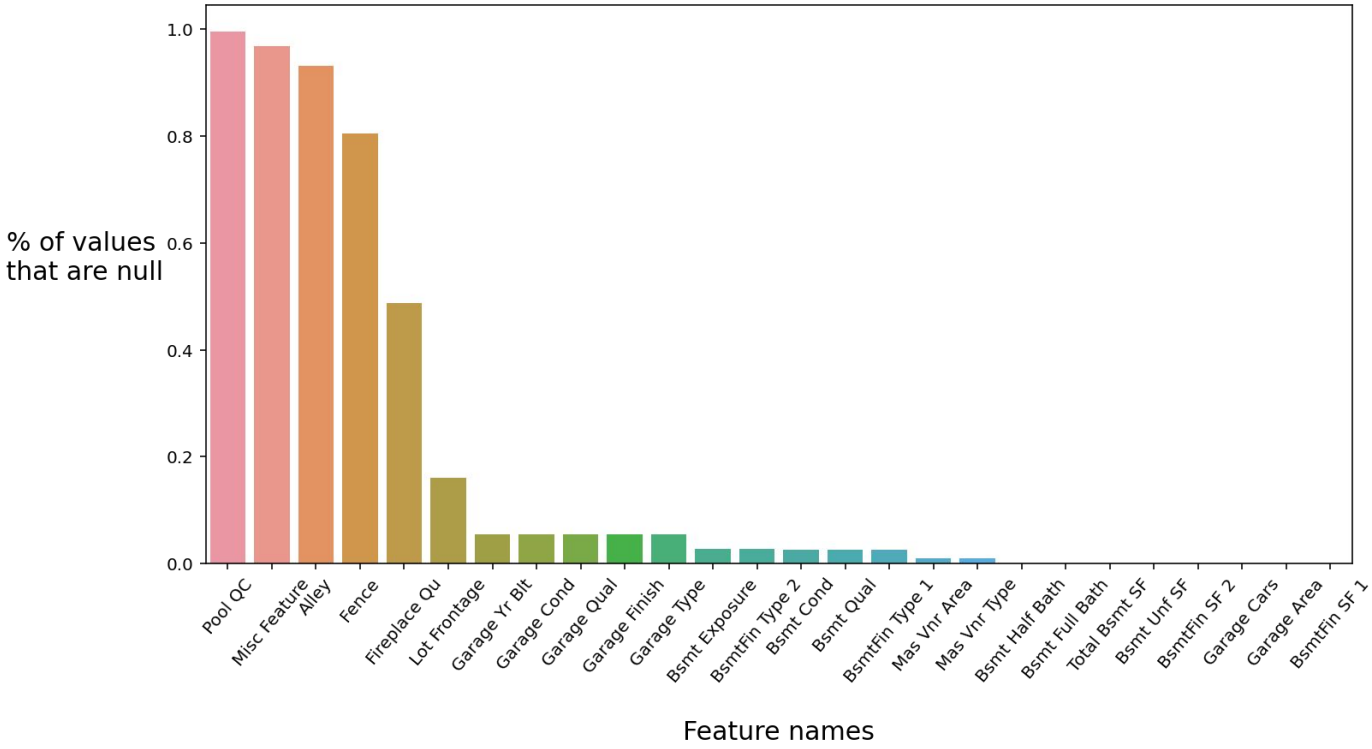
Largest area

Model Workflow pt1.



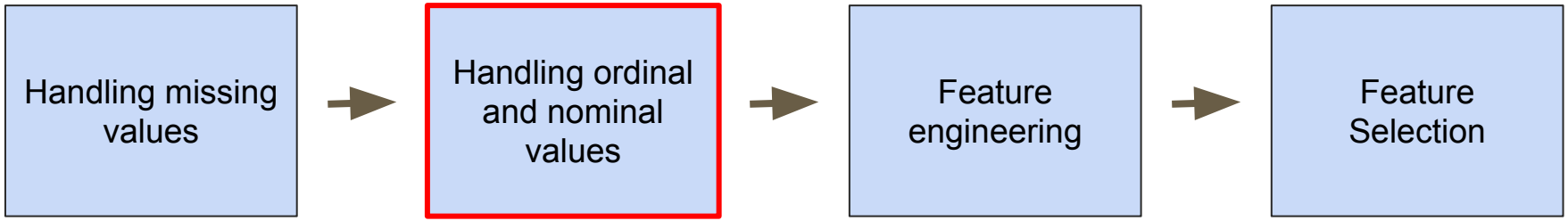
- Identifying and isolating null values
- Predicting missing values i.e. Lot frontage
- Dropping rows with missing values

Data cleaning: Handling missing values



1. Re-labelled some of the null values for categorical variables
2. Imputation for “Lot Frontage” using the median
3. Drop remaining null values

Model Workflow pt1.



- Ordinal values converted to integers
- Nominal values dummified

Data Cleaning: Ordinal Features

Exter Qual (raw data)	→	Exter Qual (ordinal data)
Ex (Excellent)		5
Gd (Good)		4
TA (Average/Typical)		3
Fa (Fair)		2
Po (Poor)		1

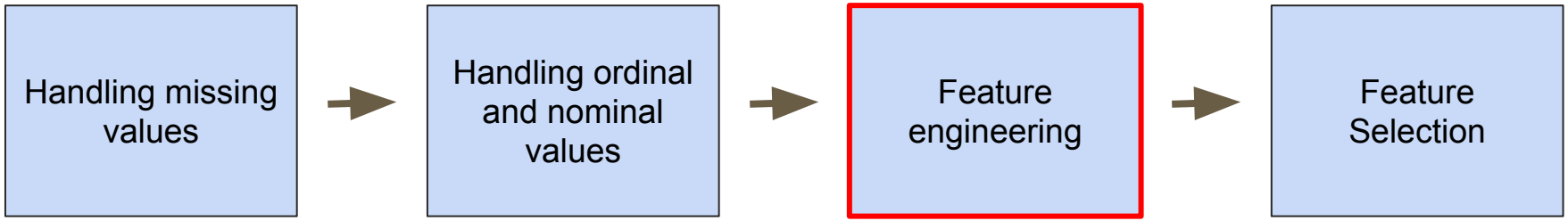
Data Cleaning: Nominal Features

Id	Roof Style
1	Flat
2	Gable
3	Gable
4	Hip
5	Gable
6	Flat

Dummification →

Id	Roof Style_Flat	Roof Style_Gable	Roof Style_Hip
1	1	0	0
2	0	1	0
3	0	1	0
4	0	0	1
5	0	1	0
6	1	0	0

Model Workflow pt1.



- Combining features
- Engineering new features i.e. age, ordinalising neighborhoods

Feature Engineering: Combining Features

1. **Combining features:**

Raw data **Combined feature**

1. Garage area + First floor area + Second floor area + Basement floor area + Masonry area = Total house area

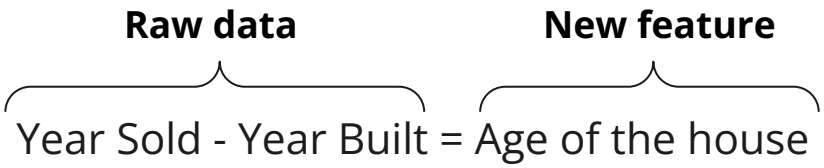
2. Total basement bathrooms + Total above grade bathrooms = Total bathrooms

	Garage Area	1st Flr SF	2nd Flr SF	Total Bsmt SF	Mas Vnr Area	combined_living_area
0	475.0	725	754	725.0	289.0	2968.0
1	559.0	913	1209	913.0	132.0	3726.0
2	246.0	1057	0	1057.0	0.0	2360.0
3	400.0	744	700	384.0	0.0	2228.0
4	484.0	831	614	676.0	0.0	2605.0

	Bsmt Full Bath	Bsmt Half Bath	Full Bath	Half Bath	total_bath
0	0.0	0.0	2	1	3.0
1	1.0	0.0	2	1	4.0
2	1.0	0.0	1	0	2.0
3	0.0	0.0	2	1	3.0
4	0.0	0.0	2	0	2.0

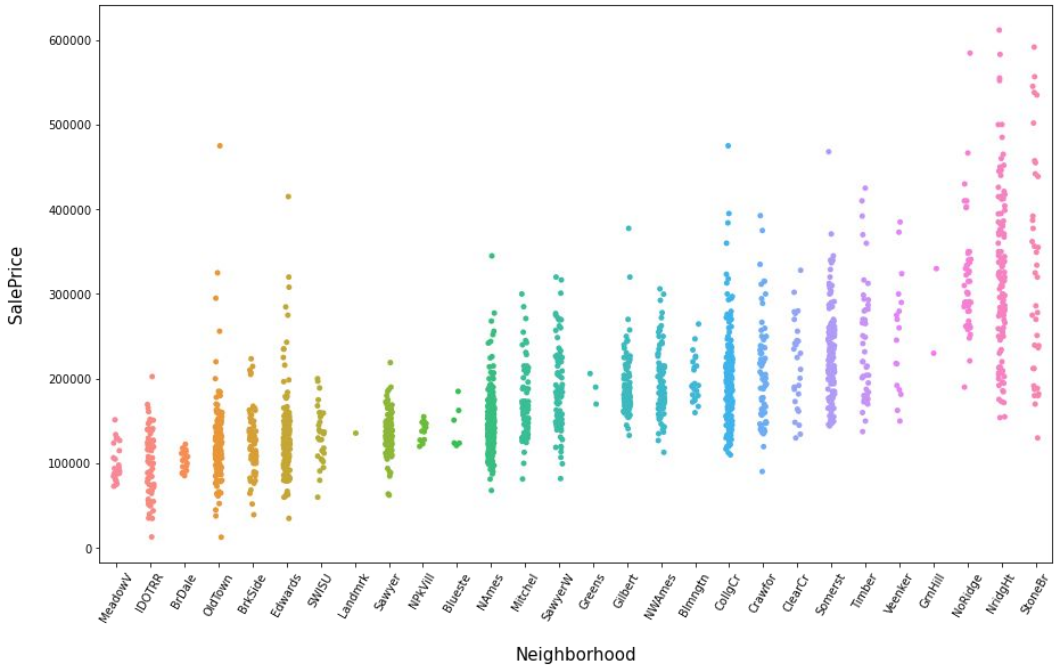
Feature Engineering: Creating New Features

2. Creating new features:



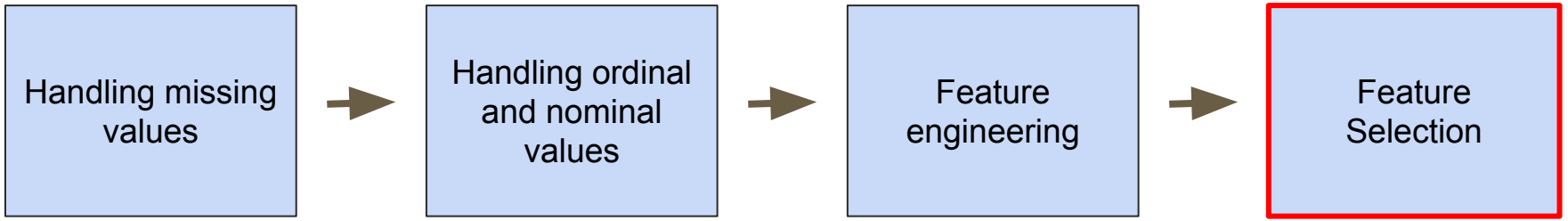
Feature Engineering: Creating New Features

3. Ordinalising Neighbourhoods using Quality/Condition features



Neighborhood	neighborhood_ranking (ordinal)
MeadowV	1
OldTown	1
SWISU	1
Edwards	1
IDOTRR	1
BrkSide	1
BrDale	1
Sawyer	2
GrnHill	2
Landmrk	2
ClearCr	2
Mitchel	2
NAmes	2
NPkVill	2
Blueste	3
SawyerW	3
NWAmes	3
CollgCr	3
Gilbert	3
Crawfor	3
Greens	3
Somerst	4
StoneBr	4
Blmngtn	4
NoRidge	4
Timber	4
NridgHt	4
Veenker	4

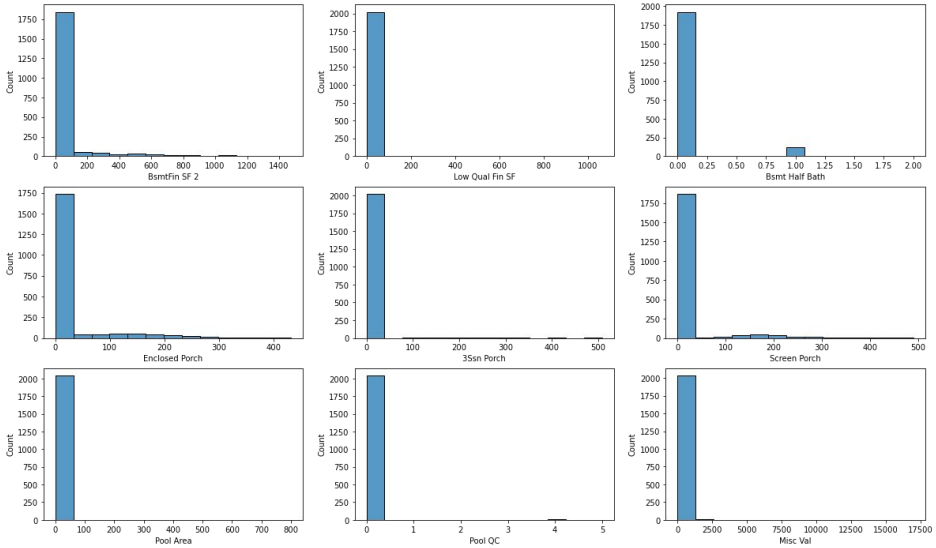
Model Workflow pt1.

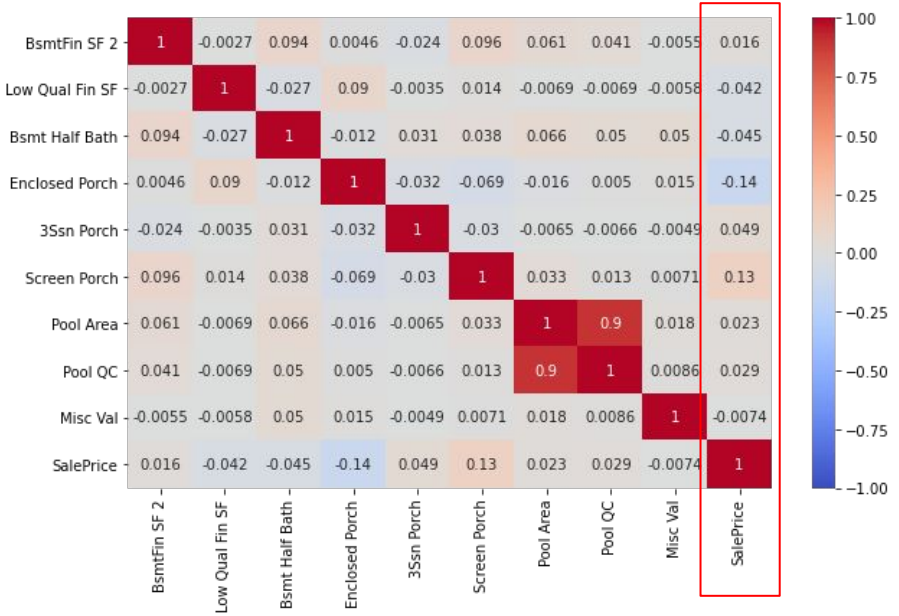


- Removing features with dominating common values and weak correlation to target feature
- Removing intercorrelated variables

Features Selection

1a. Removing features with 1 dominant value and weak correlation to SalePrice:

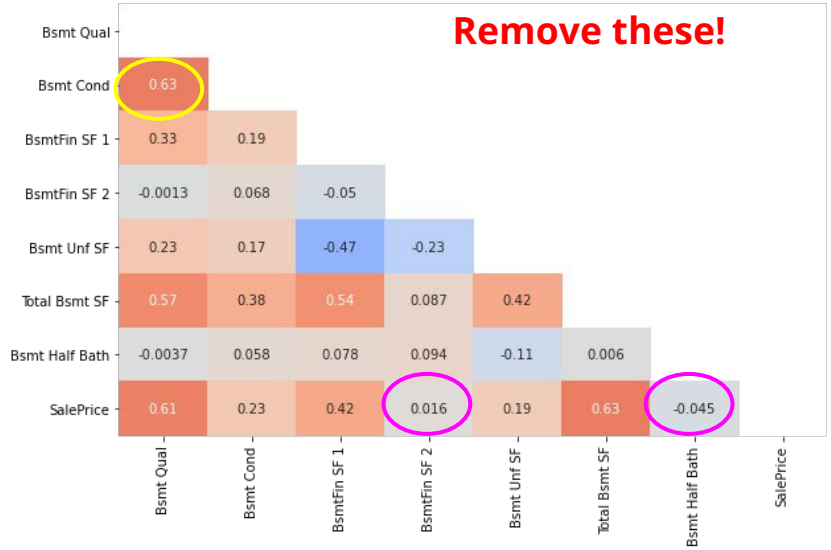




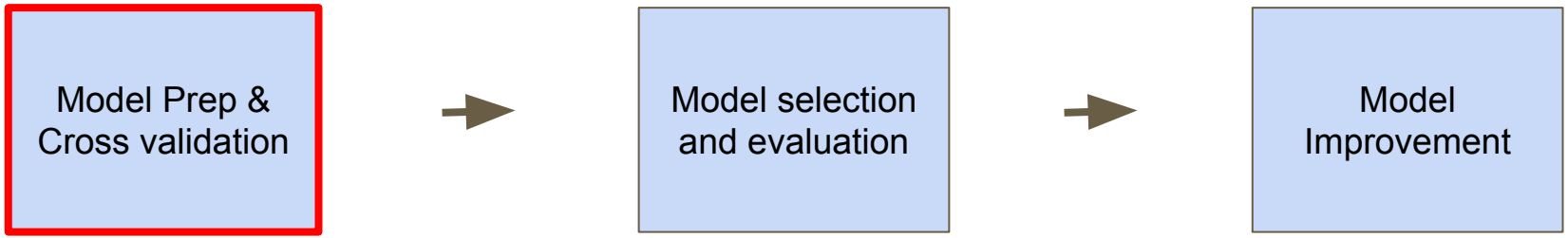

Remove these!
Overall Qual
Overall Cond
Exter Qual
Exter Cond
Bsmt Qual
Bsmt Cond
Heating QC
Kitchen Qual
Garage Qual
Garage Cond
Pool QC
Fireplace Qu
SalePrice
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Feature Selection

2. Removing intercorrelated features ('Basement' related features)



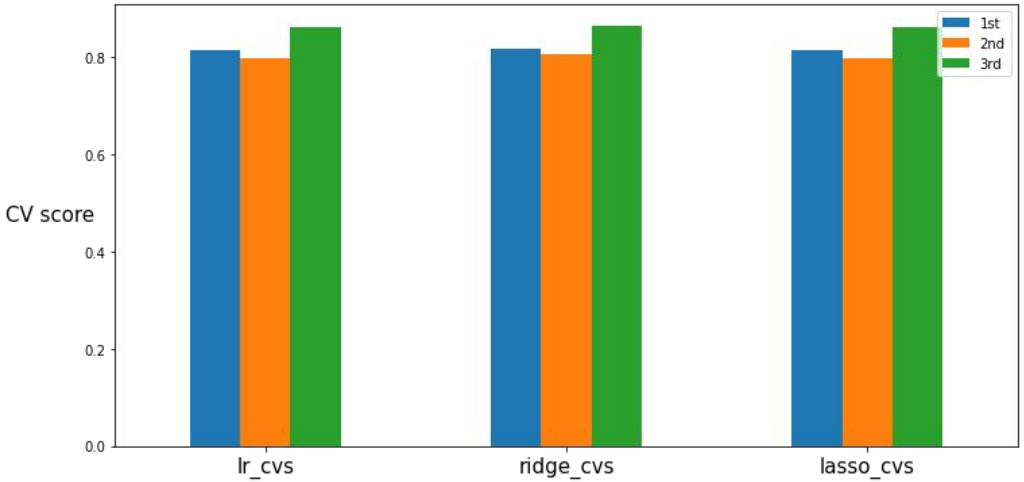
Model Workflow pt2.



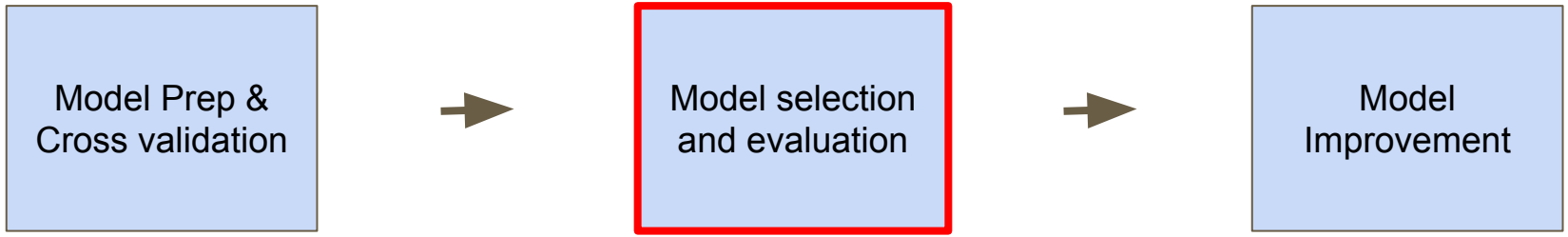
- Model utilizes 80/20 as train/test data for a robust model
- Regression models considered: Linear, Lasso, Ridge

Model Prep

Model	Cross validation score
Linear Regression	[0.815, 0.798, 0.862] **mean = 0.825**
Ridge	[0.818, 0.806, 0.865] **mean = 0.830**
Lasso	[0.816, 0.798, 0.862] **mean = 0.826**



Model Workflow pt2.

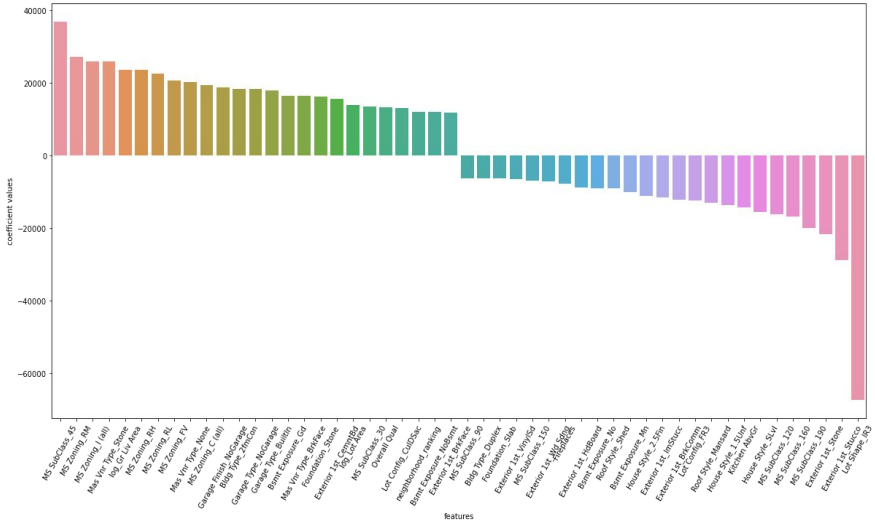


- Regularisation using Ridge and Lasso to avoid overfitting
- Ridge Regression identified as best model
- Comparison against Null Model (i.e. mean SalePrice)

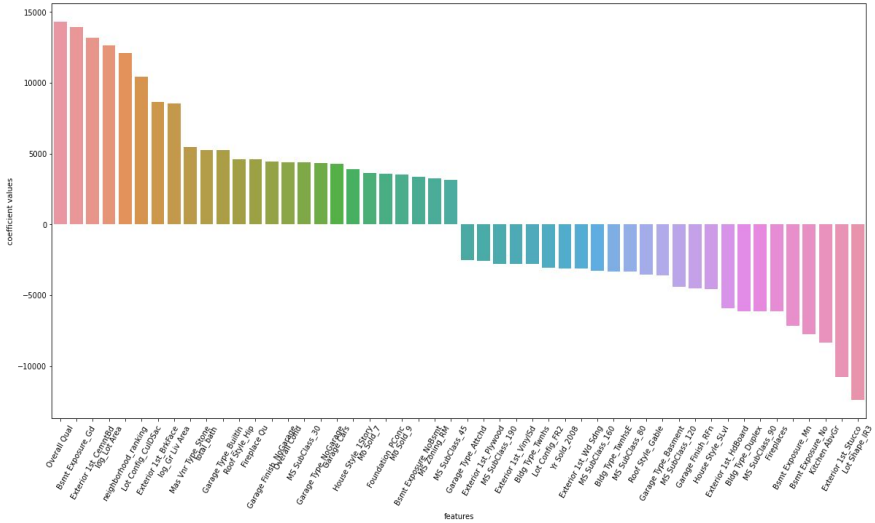
Model	R^2 score (Train set)	R^2 score (Test set)	RMSE
Linear	0.866	0.836	31,069
Ridge	0.861	0.842	30,507
Lasso	0.866	0.837	30,963

Comparison on coefficient

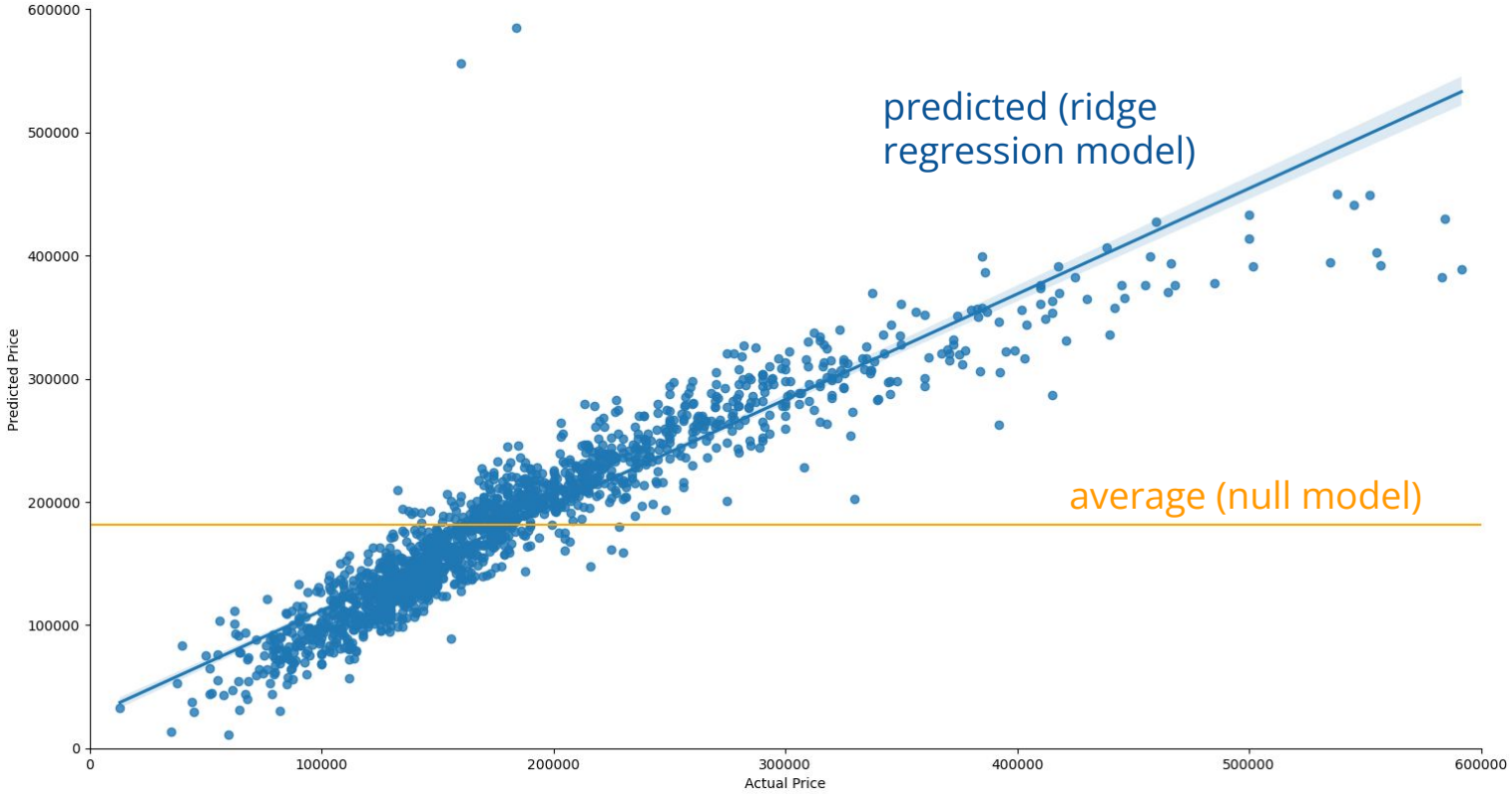
Linear



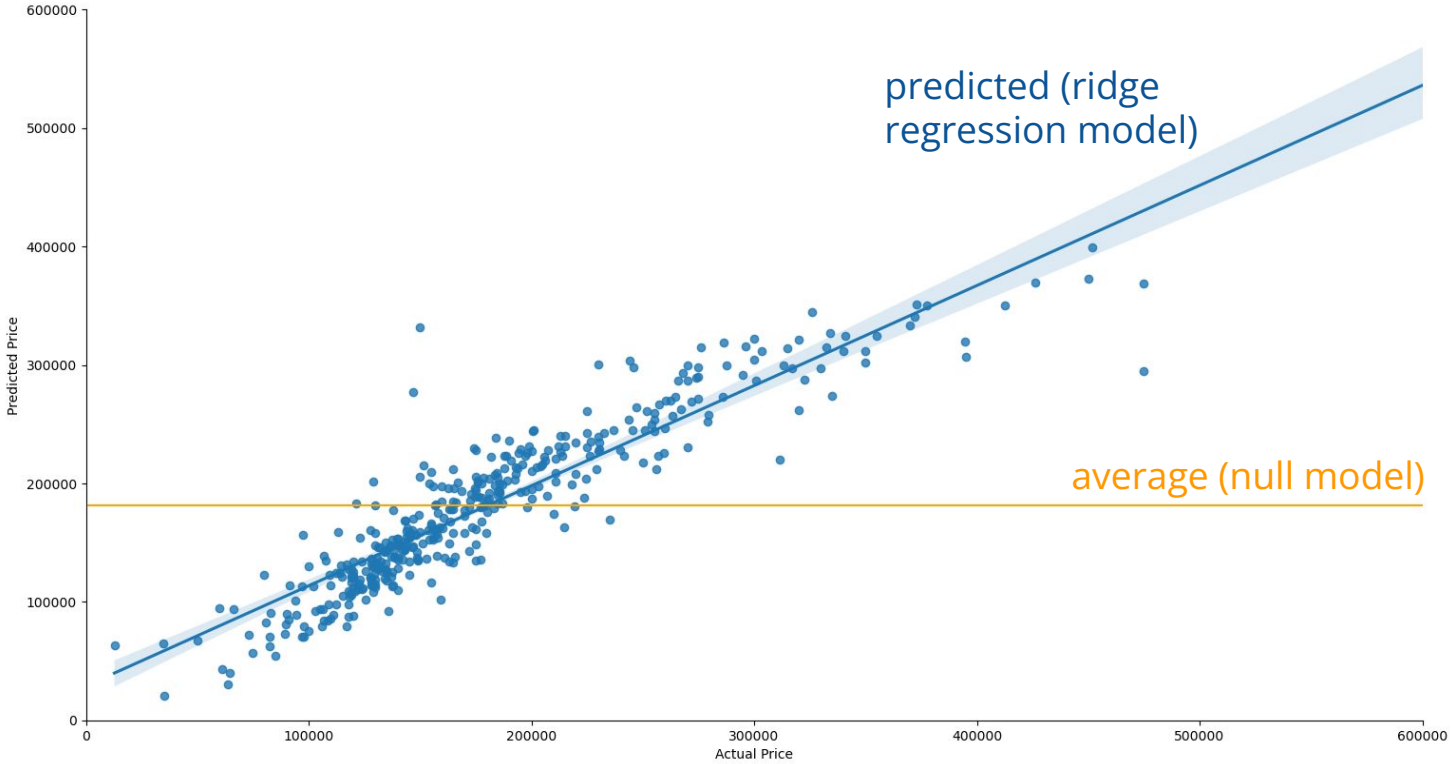
Ridge (after regularisation)



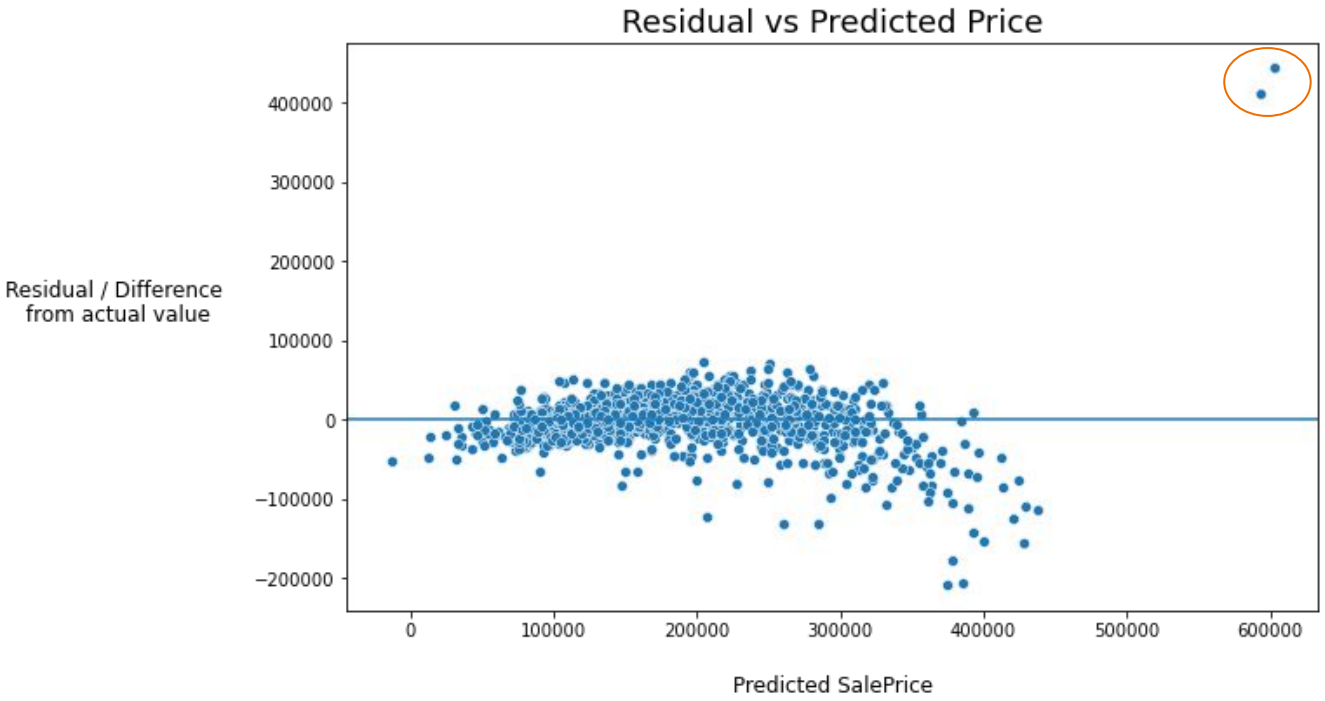
Model Comparison (Train Set)



Model Comparison (Test Set)



Model Comparison (Residual vs Training Set)

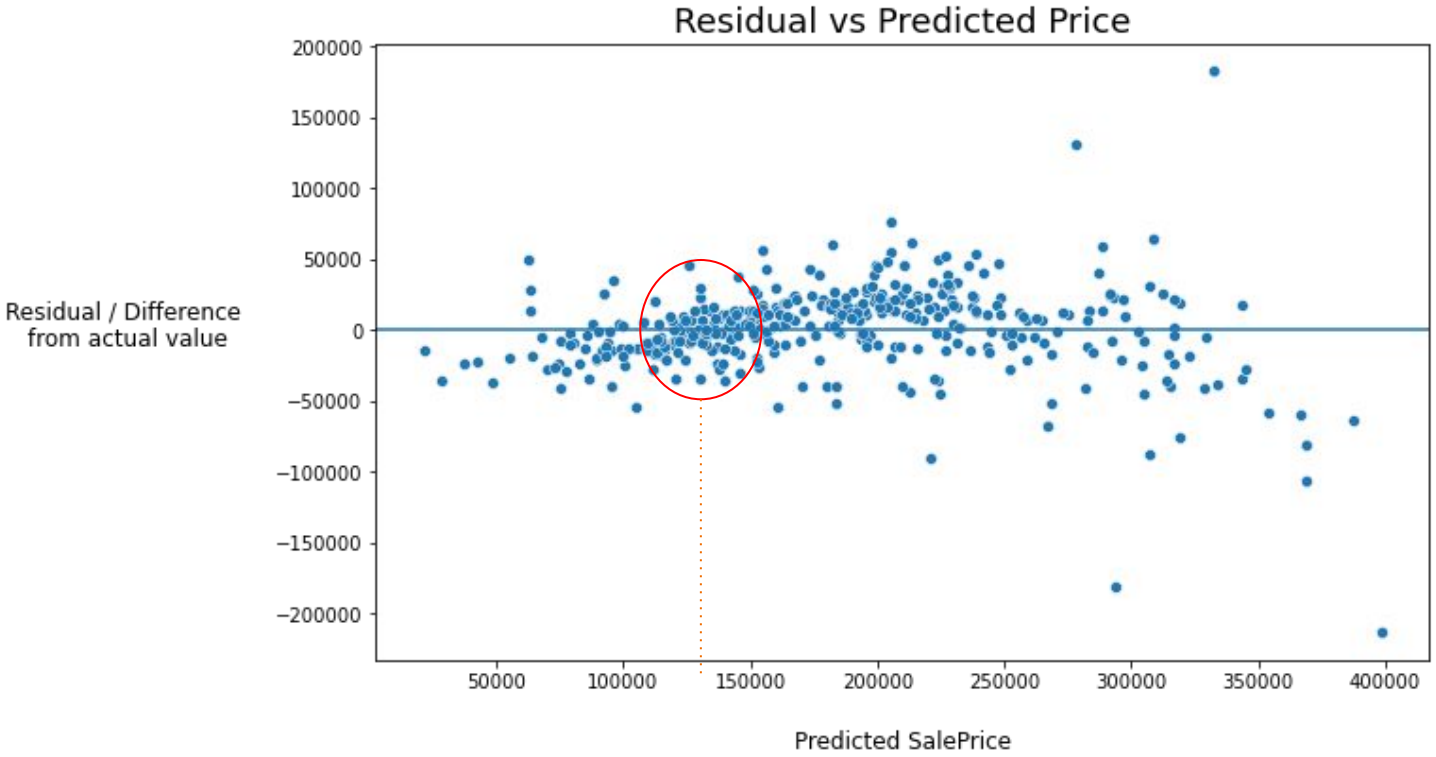


- Residual value

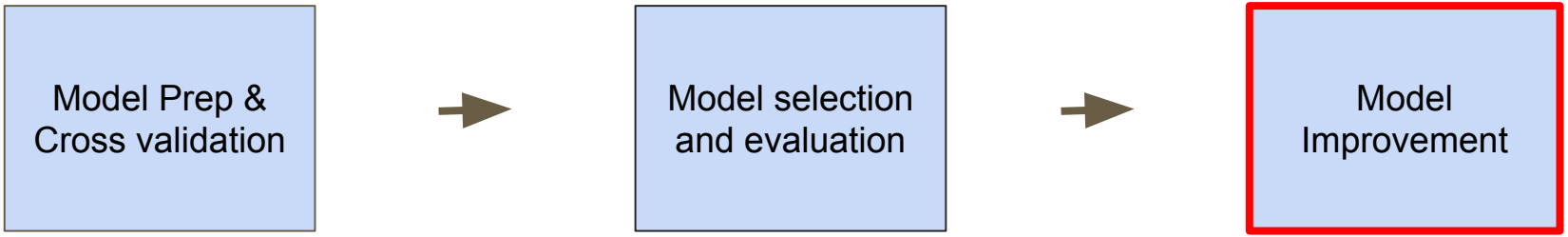
- Robust model

- Outliers kept

Model Comparison (Residual vs Test Set)



Model Workflow pt2.



- Hyperparameter tuning (alpha = 100)
- Logarithmic transformation

Model Improvement: Hyperparameter (Alpha) Tuning

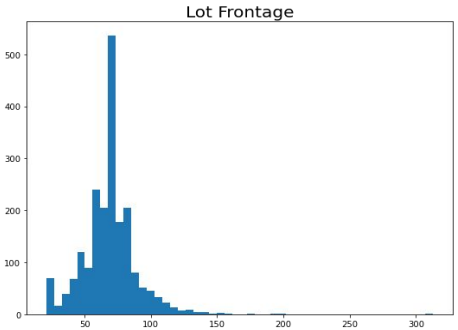
Model: RidgeCV	R^2 score (Train set)	R^2 score (Test set)	RMSE
Alpha = 5	0.865	0.834	30,773
Alpha = 40	0.860	0.843	30,455
Alpha = 100	0.855	0.844	30,370
Alpha = 120	0.854	0.844	30,382

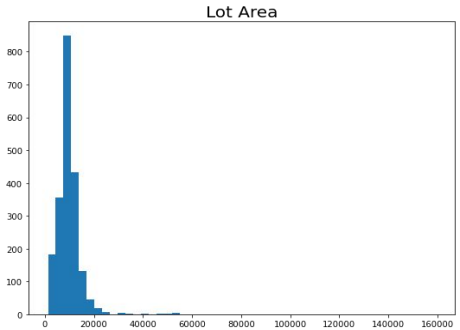
- Alpha = 100 ⇒ best compromise for maximum accuracy and minimum error
- 84.4% of variations can be explained by our model.

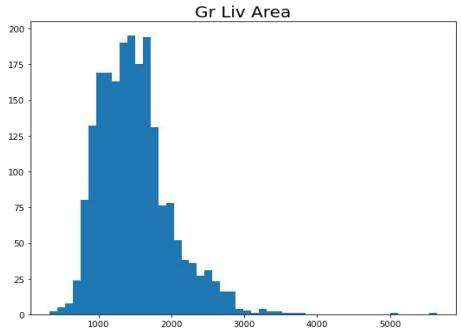
$$
$$

Model Improvement: Logarithmic Transformation

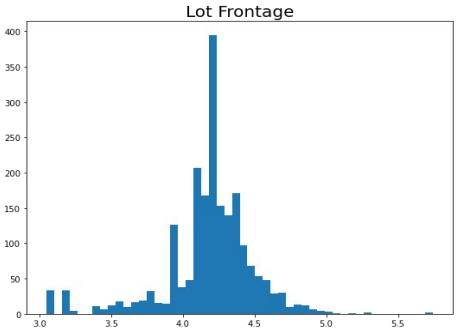
Before

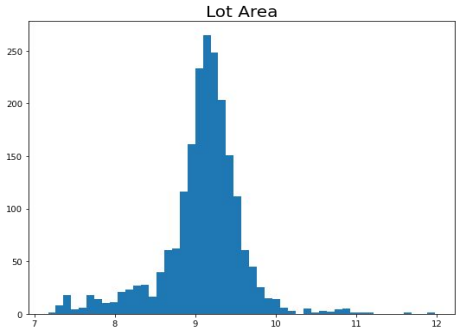


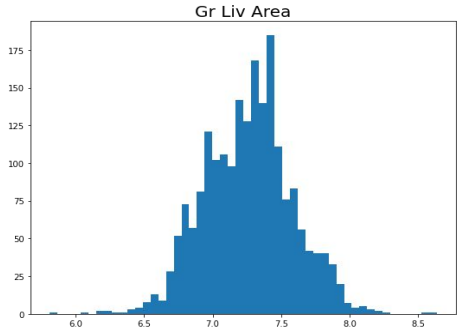




After

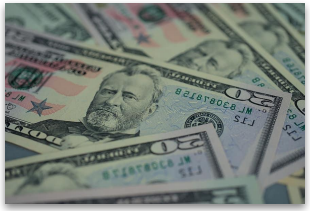






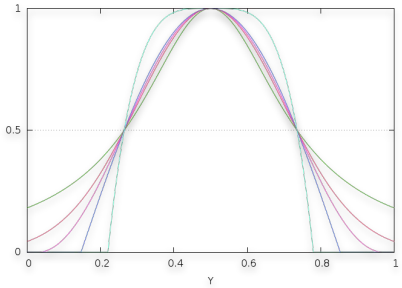
Future Model Optimisation:

- Data on family income / per-capita income



- Data on average household age

- Feature selection technique such as variance thresholding



Considerations for Agents

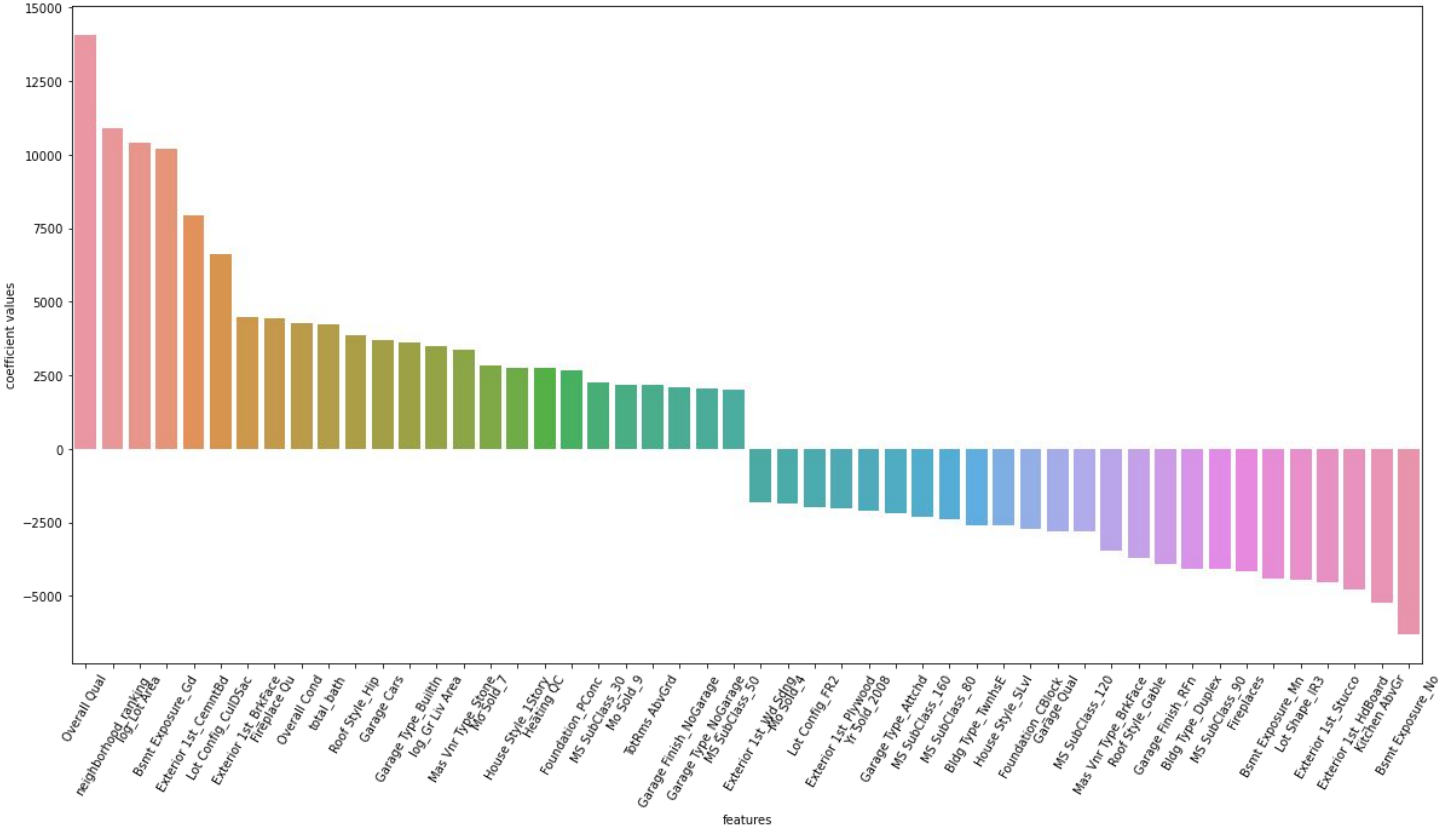
Which features to focus on for higher sale prices and profits

Price ?

Location ?

House ?

Top 50 House Features



Bottom 10 Factors

Rank	Feature	Coefficient value
1	Basement Exposure - No	-6288
2	Kitchen Above great	-5239
3	Exterior - Hardboard	-4800
4	Exterior - Stucco	-4536
5	Lot shape - irregular	-4469
6	Basement Exposure- minimum exposure	-4404
7	Fireplaces	-4185
8	MS subclass 90 - duplex	-4070
9	Building type - duplex	-4070
10	Garage Finish - FN	-3936

Avoid these units

Worst	
Exterior	Hardboard, Stucco
Lot Shape	Irregular
Building Type	Duplex
Basement Exposure	No, Minimum
Garage Finish	Rough Finished

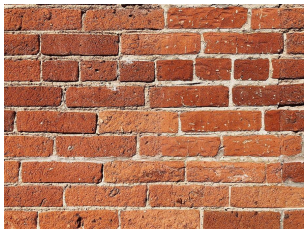
Top 10 Factors

Rank	Feature	Coefficient value
1	Overall Quality	14050
2	Neighborhood - ranking	10892
3	Lot Area	10395
4	Basement Exposure - Good	10214
5	Exterior - Cement Board	7919
6	Lot Configuration - Cul de Sac	6623
7	Exterior - Brickface	4478
8	Fireplace Quality	4417
9	Overall Condition	4273
10	Total bath	4210

Choose these units with features

Specific		Generic
Neighborhood	StoneBrook, Northridge	Overall Quality
Basement Exposure	Good	Lot Size
Exterior	Cement, Brickface	Overall Condition
Garage	Built-in	Fireplace Quality
Lot Configuration	Cul de sac	
Bath	Total Bath	

Recommendations



StoneBrook

Cul De Sac (Dead End)

Conclusion

Our Approach

- Using model to predict sale prices of properties 85% accuracy
- Provide real estate agents top performing features to improve prices of houses
- Focus on improving worst features in a house to bring up prices

Future Improvements (Why our model?)

- Advanced model for price prediction
- Shows how features can affect sale price
- Can be updated with more features, eg. age of buyer/seller, income of buyer/seller
- Able to cater specifically to every income groups
- Explore other relationships between features

Kishan Analytics

Kishan Analytics

Thank you