Predict House Price

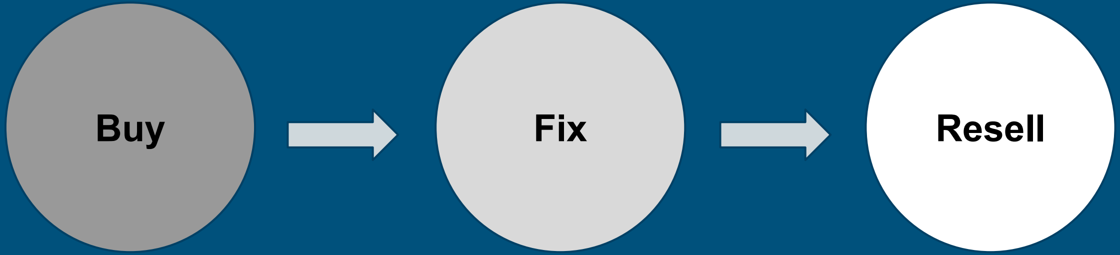
By Using Linear Regression

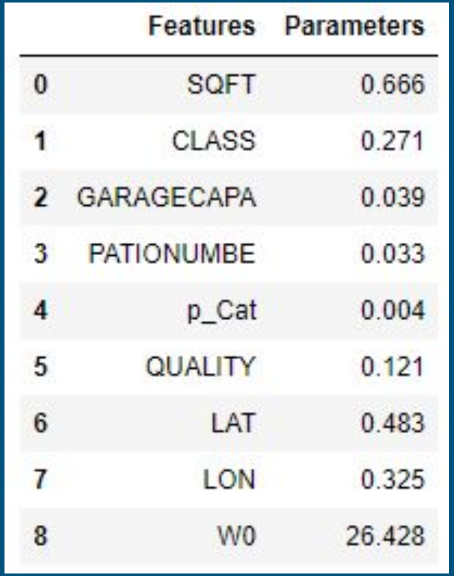
Xingliang Shu

OUTLINE

1. Business problem
2. Dataset
   a. Description
   b. Cleaning
   c. Feature Selection
3. Regression Model
   a. Cut Features
   b. Interpretation
   c. Result
4. Next

Business Problem





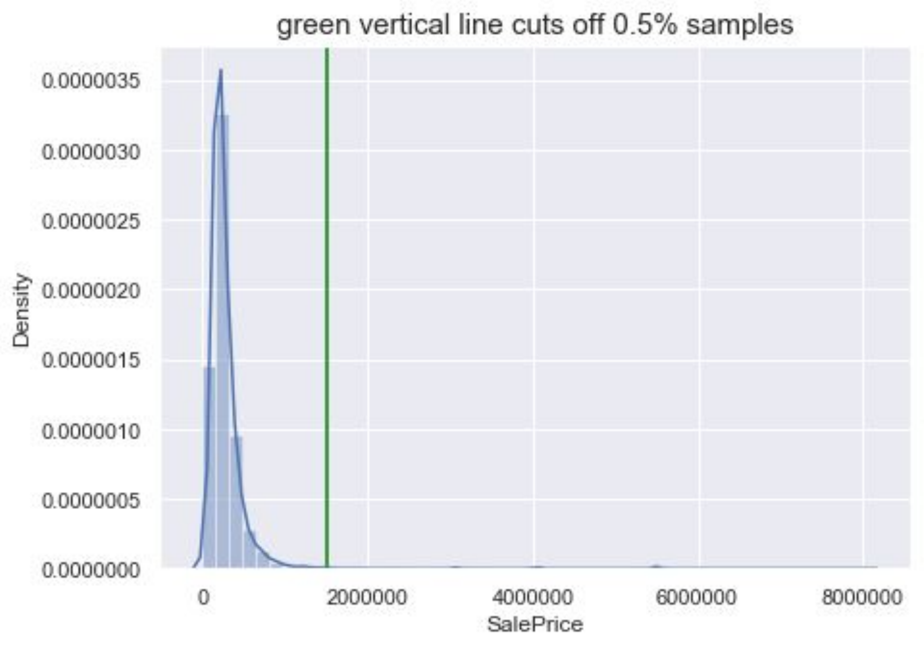
	Features	Parameters
0	SQFT	0.666
1	CLASS	0.271
2	GARAGECAPA	0.039
3	PATIONUMBE	0.033
4	p_Cat	0.004
5	QUALITY	0.121
6	LAT	0.483
7	LON	0.325
8	W0	26.428

Dataset (Description)

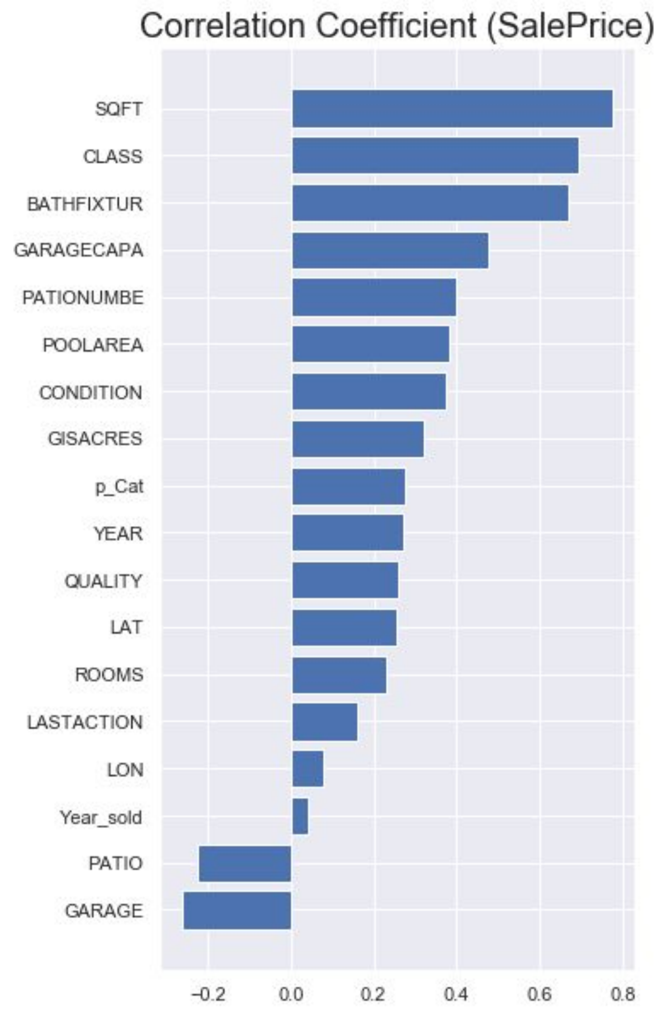
- 52,918 samples
- 49 features

Dataset (Cleaning)

- **Original samples: 52,918**
- Duplicated samples: 16,038
- Remove sale price < 1k
- Cut off extreme samples
- **Total removed: 16,416**
- **Remain: 36,502**



Feature Selection

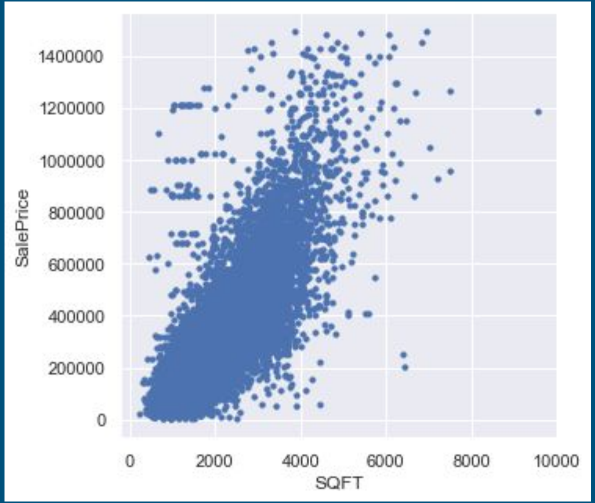


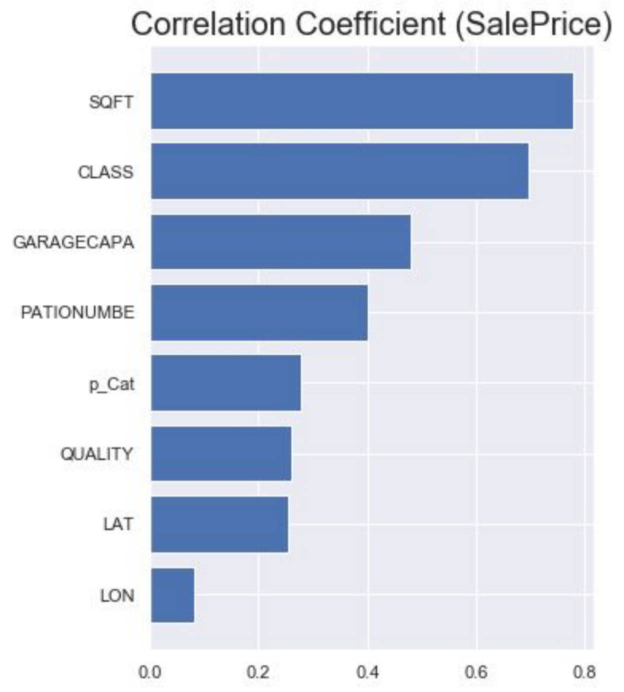
Dataset Split

- 70% train
- 30% test

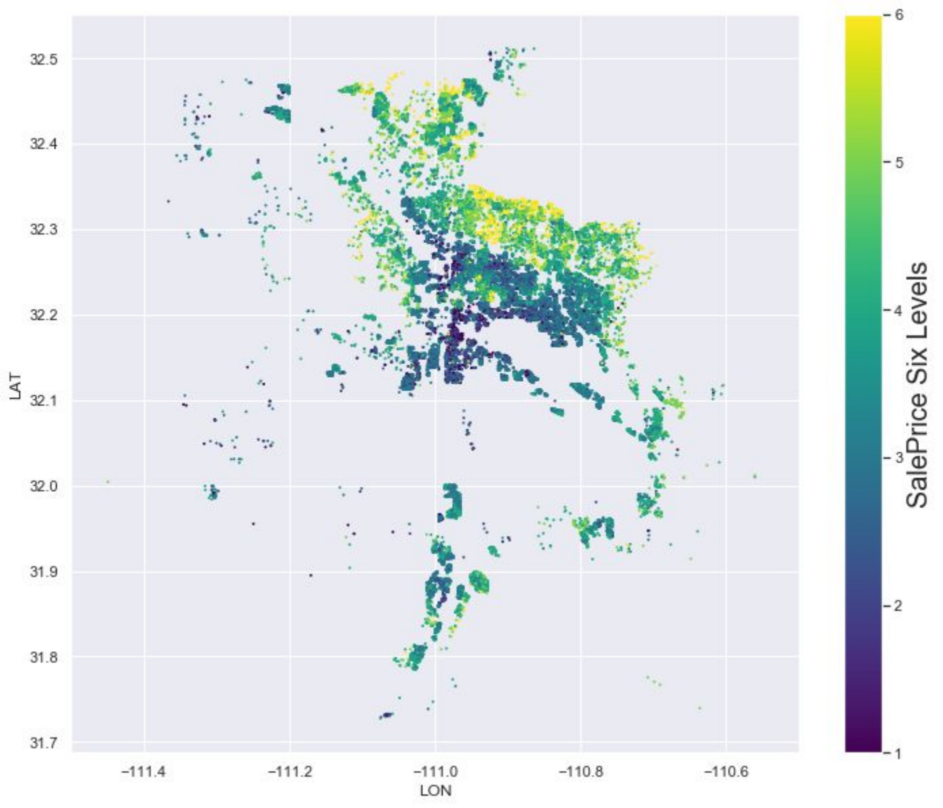
Regression Model

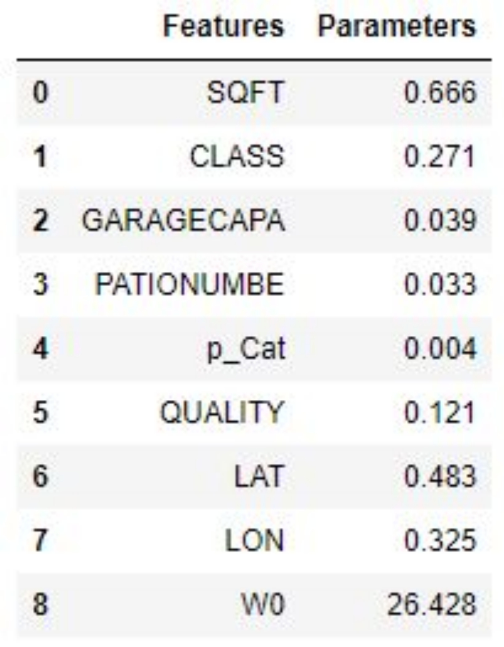
- Optimal selected feature (afte regression)
- log(SQFT), log(SalePrice)





Regression Model (Interpretation)





	Features	Parameters
0	SQFT	0.666
1	CLASS	0.271
2	GARAGECAPA	0.039
3	PATIONUMBE	0.033
4	p_Cat	0.004
5	QUALITY	0.121
6	LAT	0.483
7	LON	0.325
8	W0	26.428

Regression Model (Result)

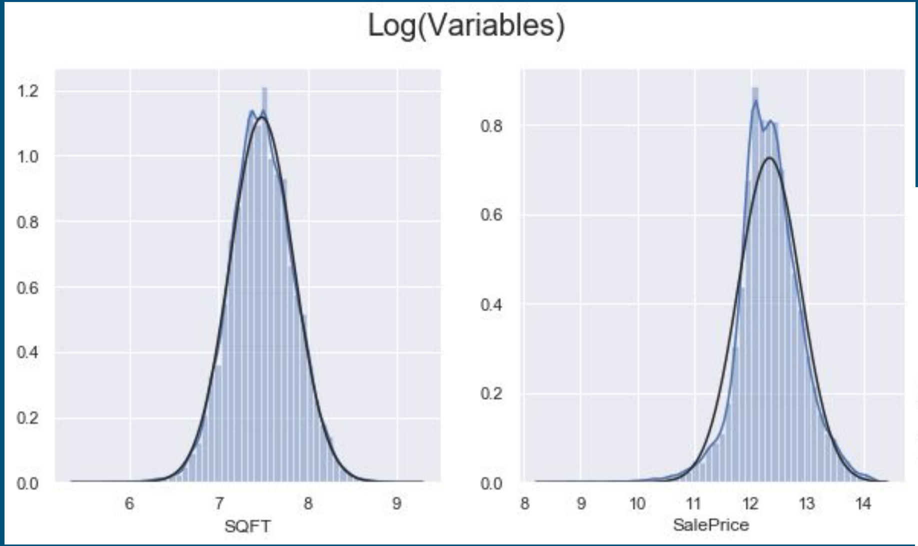
- Sale price = exp[parameters.dot(features)]
- **R square: 0.725**
- **Accuracy: 80%**

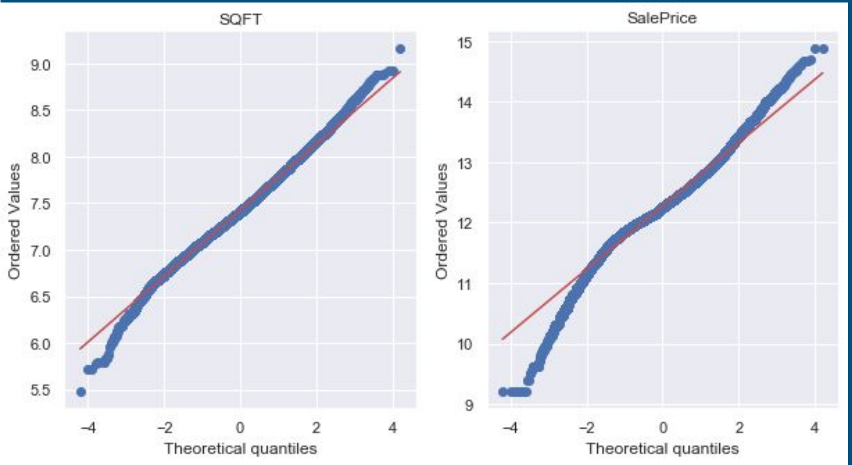
Summary

- Capture 80% house price
- Next
  1. Better feature selection
  2. Better models

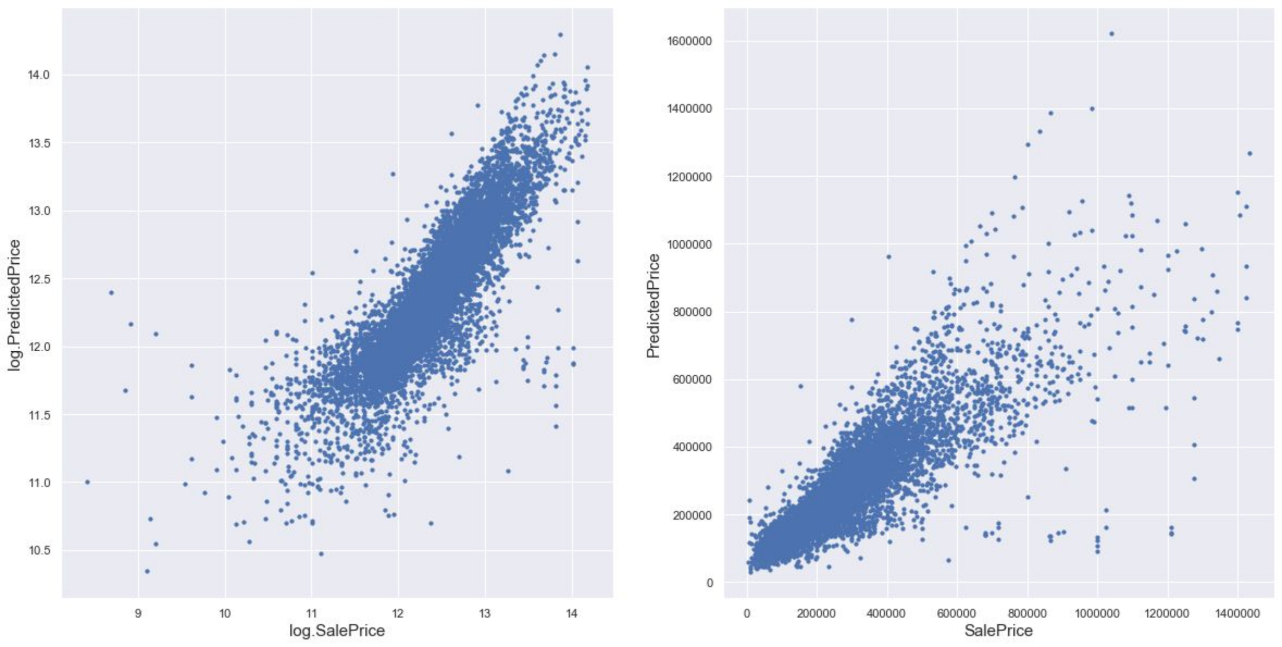
Q & A

Appendix (Transform Function for Normality)

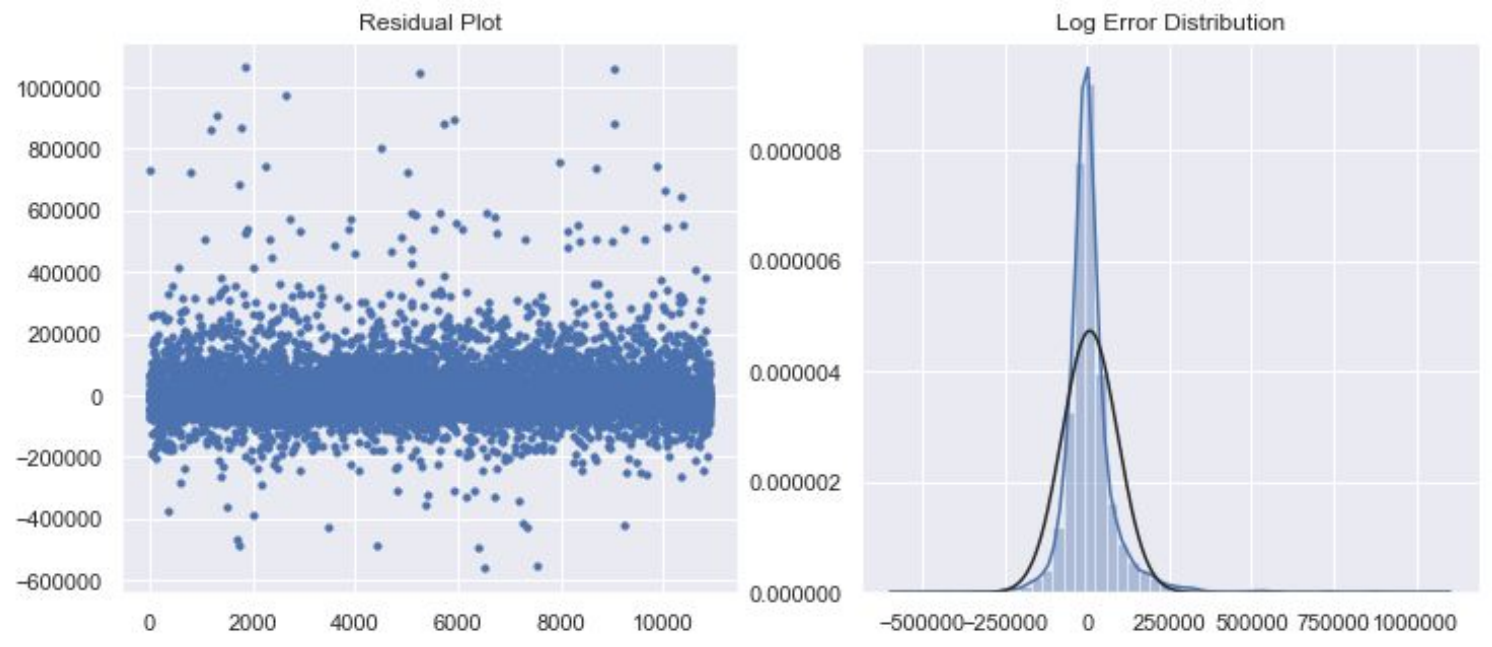




Appendix (Target vs Prediction)



Appendix (Residual Plot)



Residual Plot

Log Error Distribution