Predicting HDB Resale Prices

Leo Siu-Yin
Metis

TABLE OF CONTENTS

03 Motivation

04 About The Project
Project Goals

06 Methodology
Project Stages
Tools Used
Data

09 Results
Model Comparison
Assumptions Test

11 Conclusion

12 Future Work

孟母三迁

–Story of how Mencius' mother moved three times to find the right place for Mencius' upbringing.

ABOUT THE PROJECT

- HDB Resale Flats have always been a hot commodity.
- What are some important factors that affect HDB Resale Prices?
- Families can decide better on their purchase of HDB Resale Flat.

Find out factors that affect prices of HDB Resale Flats.

Goal 1

Find most suitable model that fits existing data.

Goal 2

Predict HDB Resale Flat prices using the selected model.

Goal 3

PROJECT STAGES

Stage 1
Webscraping

- Obtain data from data.gov.sg.
- 2000 most recent resale transactions.

Stage 2
Data Cleaning
Exploratory Data Analysis

- Empty columns removed.
- Flat types, Storey Range, Remaining Lease changed from object to numerical representation.
- Removed highly correlated variables.
- Log resale price to remove skew.

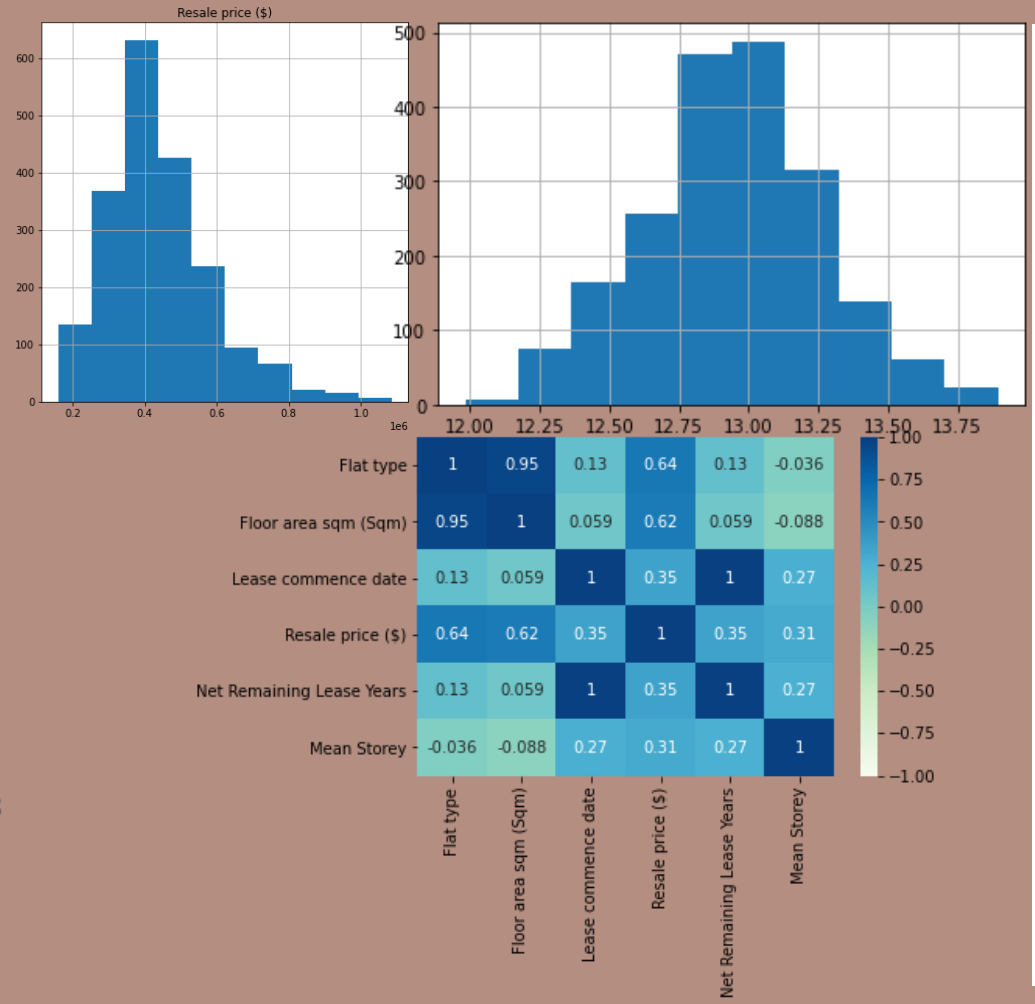
Stage 3
Model Building
Test Results

- Converted categorical variables (town) to dummy variable representation.
- Compared results between Linear Regression, RidgeCV Regression and LassoCV Regression.
- Check test results of the final selected model. (RidgeCV regression model).

Tools Used

scikit learn
statsmodels
Seaborn
matplotlib
pandas
BeautifulSoup
Se
python

Data



OLS Regression Results

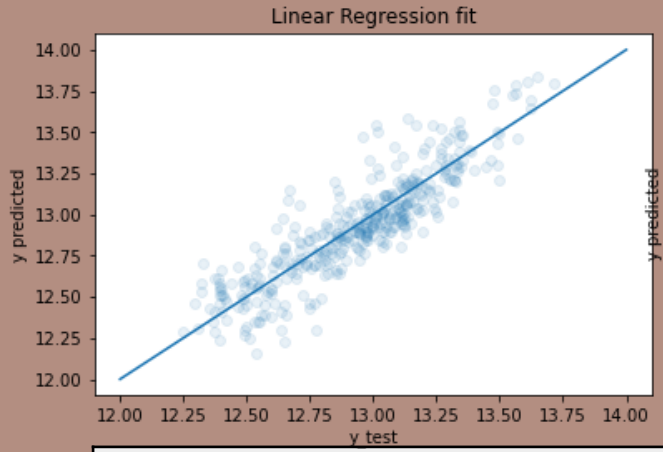
Dep. Variable:	log_Resale_Price	**R-squared:**	0.775
Model:	OLS	**Adj. R-squared:**	0.774
Method:	Least Squares	**F-statistic:**	978.6
Date:	Sun, 20 Sep 2020	**Prob (F-statistic):**	0.00
Time:	17:59:03	**Log-Likelihood:**	955.60
No. Observations:	2000	**AIC:**	-1895.
Df Residuals:	1992	**BIC:**	-1850.
Df Model:	7		
Covariance Type:	nonrobust		

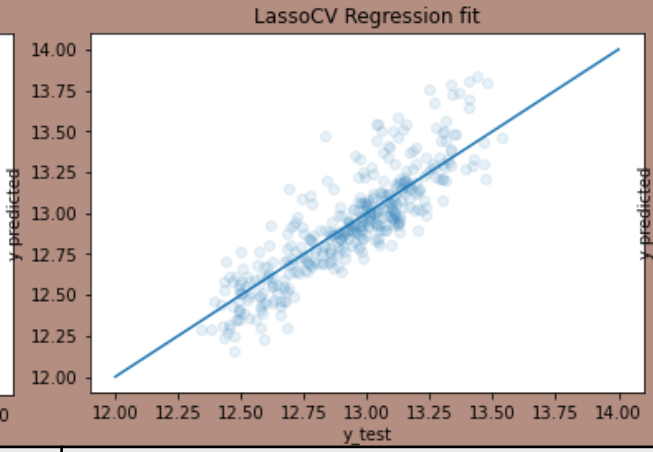
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	11.3984	0.023	502.255	0.000	11.354	11.443
Floor_Area_Sqm	0.0097	0.000	67.158	0.000	0.009	0.010
Net_Remaining_Years	0.0097	0.000	34.246	0.000	0.009	0.010
Mean_Storey	0.0107	0.001	16.908	0.000	0.009	0.012
Town_East	-0.1849	0.014	-13.249	0.000	-0.212	-0.158
Town_North	-0.3890	0.013	-30.082	0.000	-0.414	-0.364
Town_North_East	-0.2688	0.013	-20.925	0.000	-0.294	-0.244
Town_West	-0.2987	0.012	-24.158	0.000	-0.323	-0.274

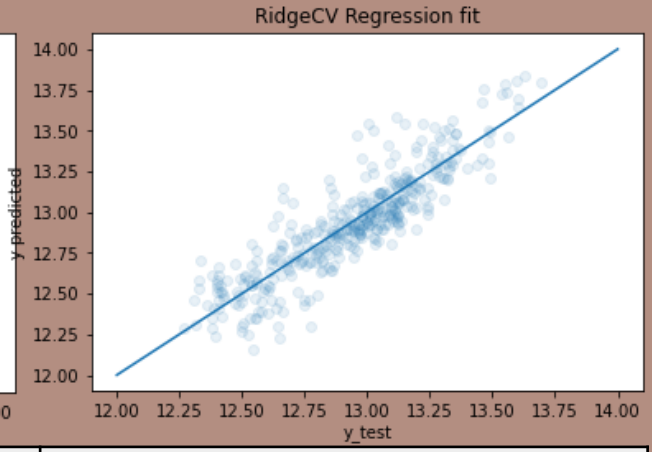
Omnibus:	208.465	**Durbin-Watson:**	0.762
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	351.428
Skew:	0.721	**Prob(JB):**	4.88e-77
Kurtosis:	4.463	**Cond. No.**	887.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Model Comparison



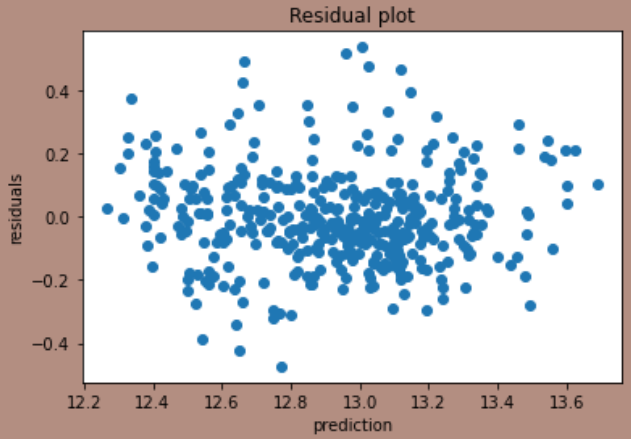


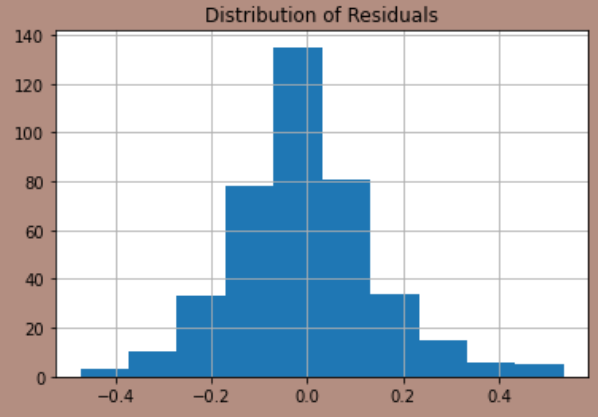


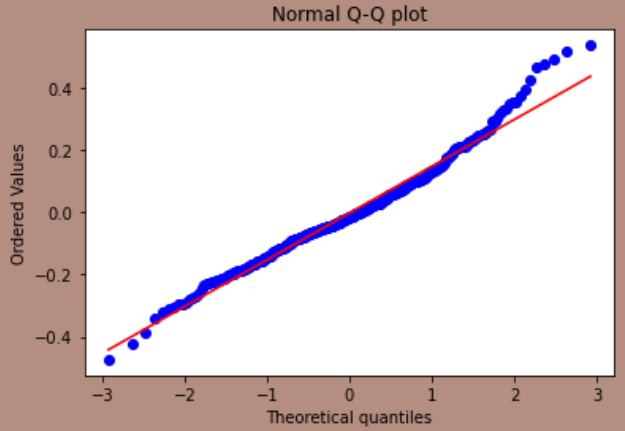
Linear Regression	LassoCV Regression	RidgeCV Regression
MAE 0.115170	MAE 0.121675	MAE 0.114181
MSE 0.022781	MSE 0.026426	MSE 0.022711

- Points of observed data are most symmetric around the line of predicted data from the RidgeCV Regression Model.

RidgeCV Model







- Residuals are normally distributed around 0 (mean) - errors have a constant variance.
- Randomness in residuals - model chosen makes little systematic errors.
- Q-Q plot formed a relatively straight line - real data is not too skewed but some extremities exist.
- Little or no concern on issues of multicollinearity based on selected independent variables.

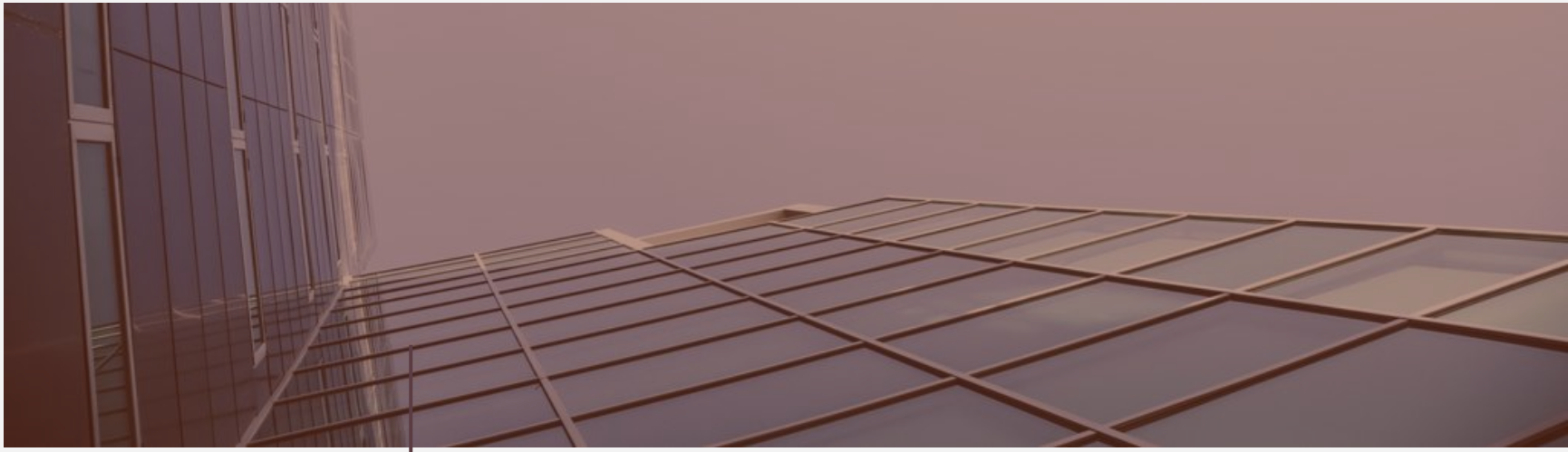
CONCLUSION

39,000

Difference between predicted and actual prices

1. Floor Area
2. Remaining Lease
3. Storey of flat

1. Central
2. East
3. North-East
4. West
5. North



Future Work

Availability of market, good schools, proximity of transport lines could be incorporated into the study of HDB Flats Resale prices.

THANK YOU

Follow Project Updates on

https://www.linkedin.com/in/syleo/
https://github.com/syleo22/SiuYin_Projects
syleo22@gmail.com

Appendix-
Town Definition

North
Sembawang
Woodlands
Yishun

North-East
Ang Mo Kio
Hougang
Punggol
Sengkang
Serangoon

West
Bukit Batok
Bukit Panjang
Choa Chu Kang
Clementi
Jurong East
Jurong West
Tengah

Central
Bishan
Bukit Merah
Bukit Timah
Central
Geylang
Kallang/ Whampoa
Marine Parade
Queenstown
Toa Payoh

East
Bedok
Pasir Ris
Tampines