HOUSING PRICES- ADVANCED REGRESSION TECHNIQUES

CAPSTONE 1 PROJECT

Aroonima Sinha

PROBLEM

BASIS FOR ESTIMATING SALE PRICE OF A HOUSE

The price estimation can be based on few factors or external sources such as real estate agencies. The problem for the buyer is knowing the exact amount for the purchase price of the house.

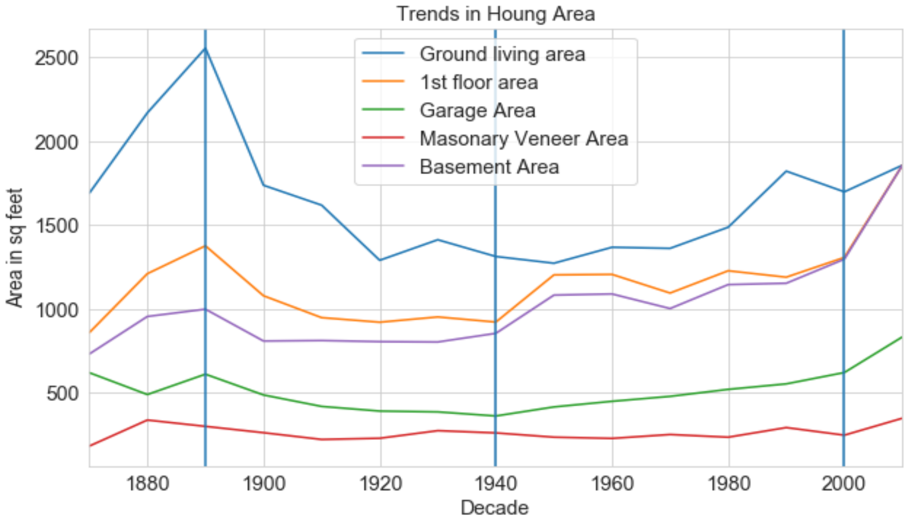
For a real estate company, which can also pose as a buyer or broker, the problem is to negotiate for the best deal.

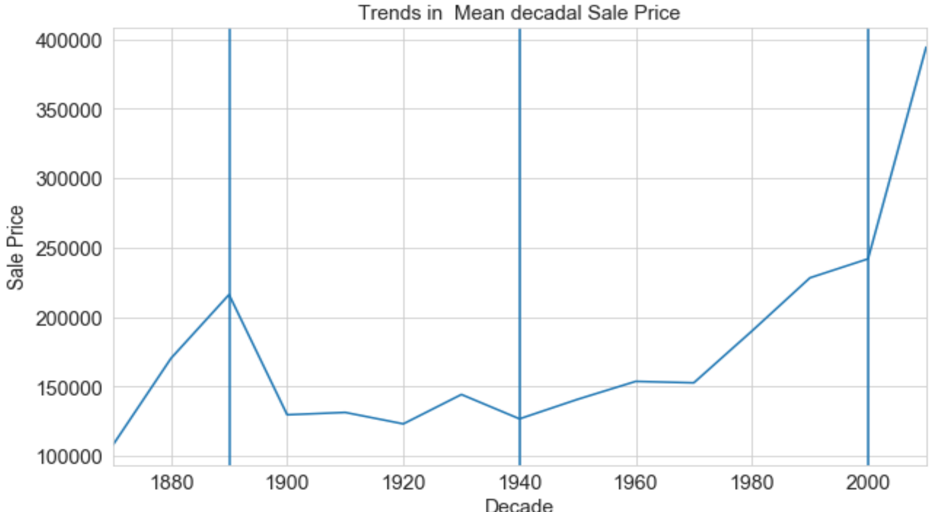
This dataset has several factors.

It becomes crucial to know the levers that drive the price and develop a model to predict them with best accuracy.

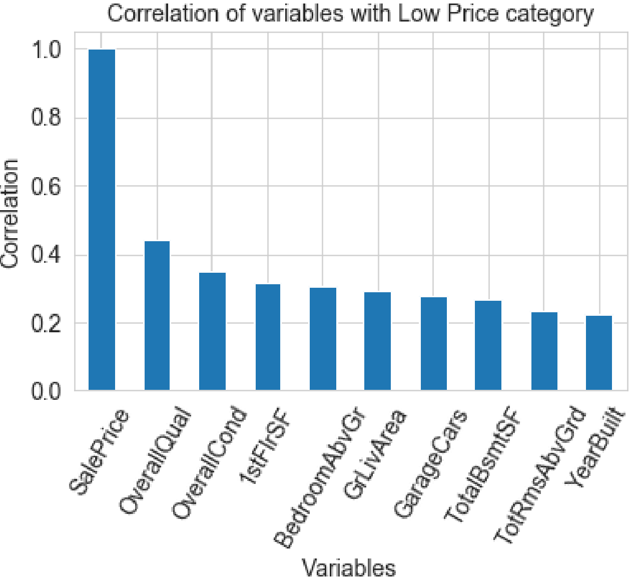
Trends in Housing Area

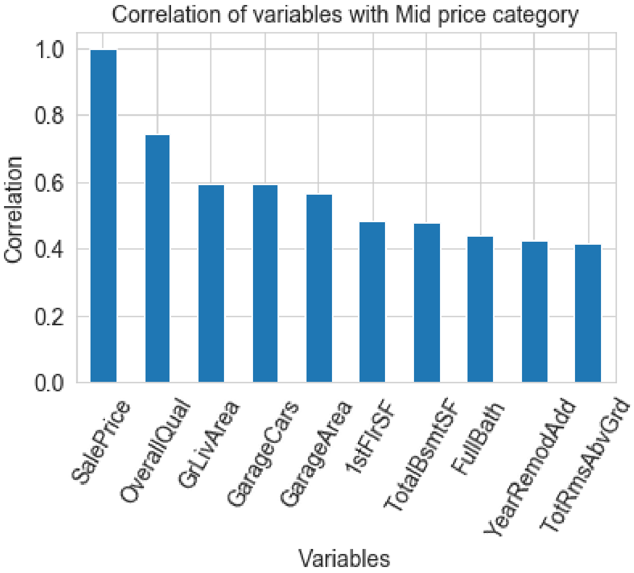
Observations: We can see 3 phases: an upward phase(till 1890), a downward and stable phase(till 1940) and an upward again (from 1940)
In 1890 the avg area of houses were big. Avg Ground living area in 1890 were biggest, which we don’t see today. We see a downward to a more stable phase till 1940. Increase in avg areas take place from 1940 with steep increase from 2000 onwards. Masonary Veneer Area shows development from 1920's.

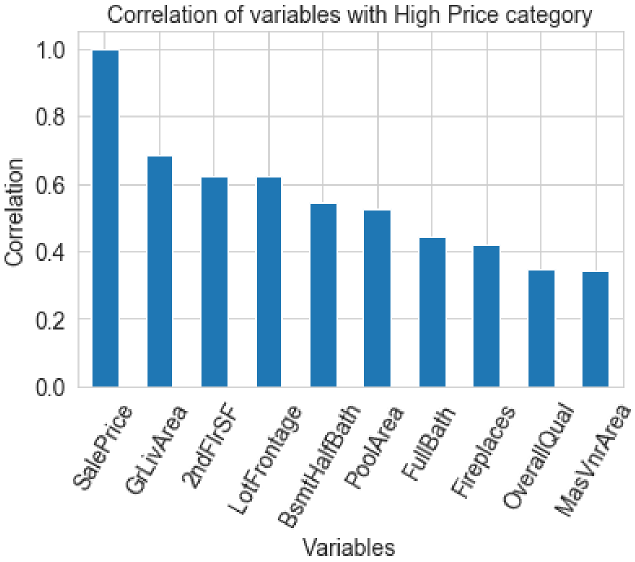




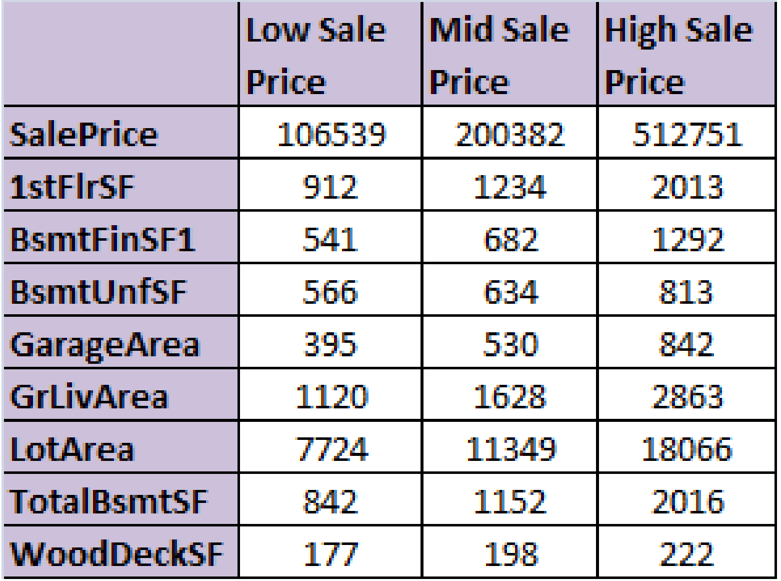
Correlation of variables with different Sale price ranges


Correlation of variables with Low Price category


Correlation of variables with Mid price category

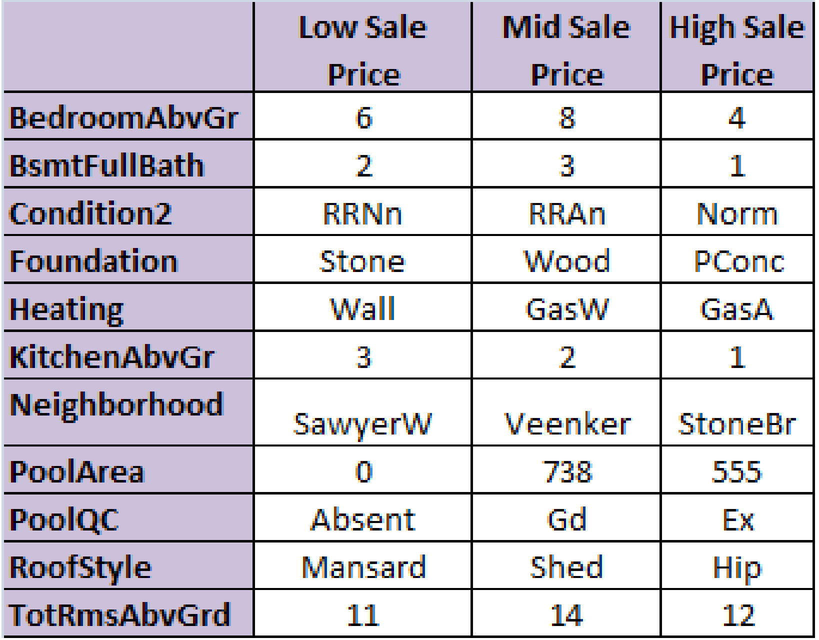

Correlation of variables with High Price category

What price can be expected with these numeric features?



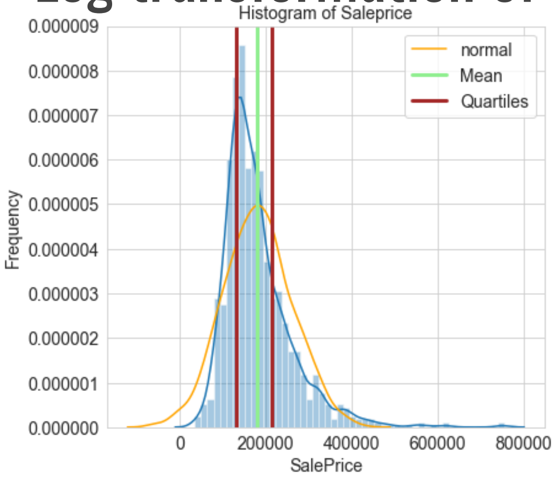
	Low Sale Price	Mid Sale Price	High Sale Price
SalePrice	106539	200382	512751
1stFlrSF	912	1234	2013
BsmtFinSF1	541	682	1292
BsmtUnfSF	566	634	813
GarageArea	395	530	842
GrLivArea	1120	1628	2863
LotArea	7724	11349	18066
TotalBsmtSF	842	1152	2016
WoodDeckSF	177	198	222

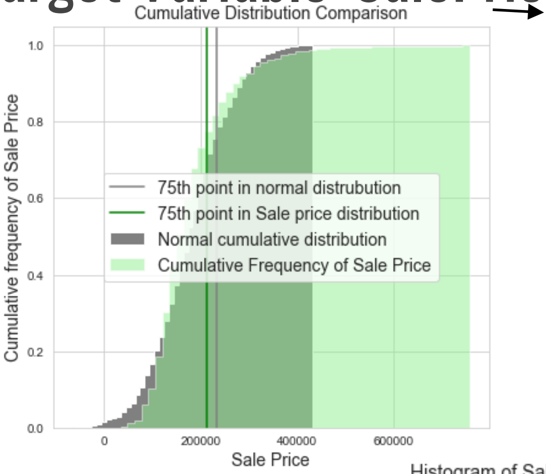
What price can be expected with these categorical features?



	Low Sale Price	Mid Sale Price	High Sale Price
BedroomAbvGr	6	8	4
BsmtFullBath	2	3	1
Condition2	RRNn	RRAn	Norm
Foundation	Stone	Wood	PConc
Heating	Wall	GasW	GasA
KitchenAbvGr	3	2	1
Neighborhood	SawyerW	Veenker	StoneBr
PoolArea	0	738	555
PoolQC	Absent	Gd	Ex
RoofStyle	Mansard	Shed	Hip
TotRmsAbvGrd	11	14	12

Log transformation of Target Variable 'SalePrice'



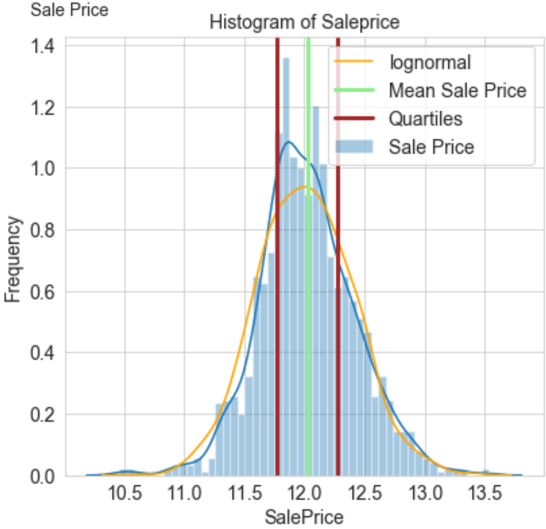


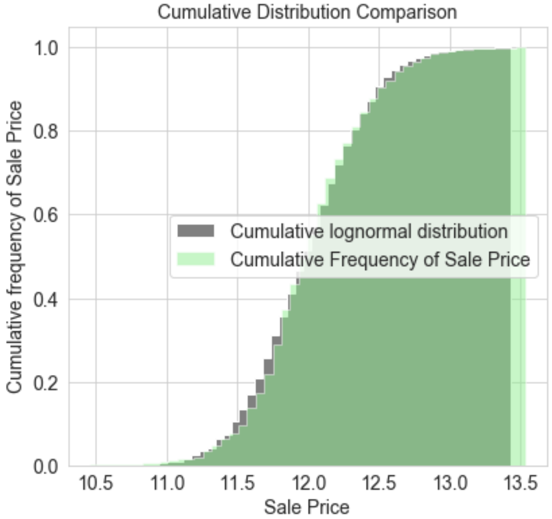
Observations:

1.The distribution is not normal. 2. Distribution of SalePrice is leptokurtic.

3. The distribution is right skewed.

4. Mean Sale Price is not a good representation and there are quite a number of outliers.





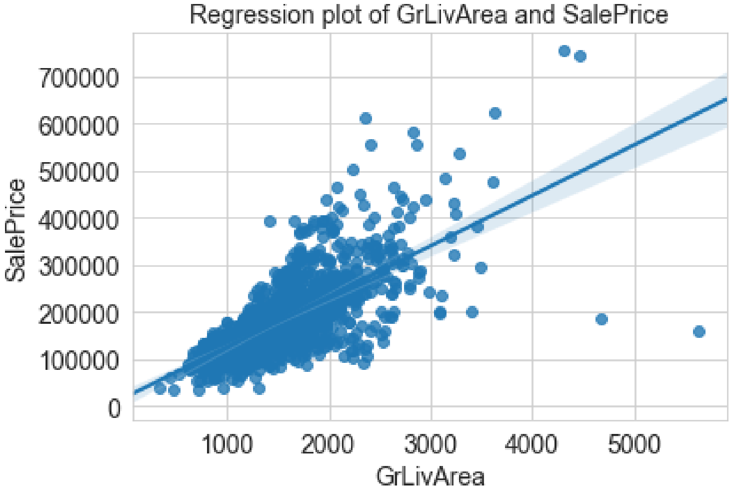
Observations:

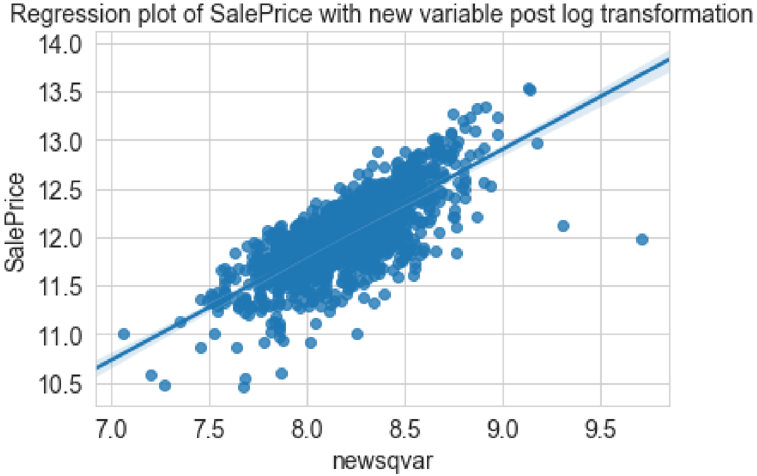
The distribution of 'SalePrice' is very close to lognormal distribution.

The tails are matching, though 'SalePrice' appears to be bimodal.

The range above the upper quartile has normalised to quite an extent.

Independent variable GrLivArea and its transformation:



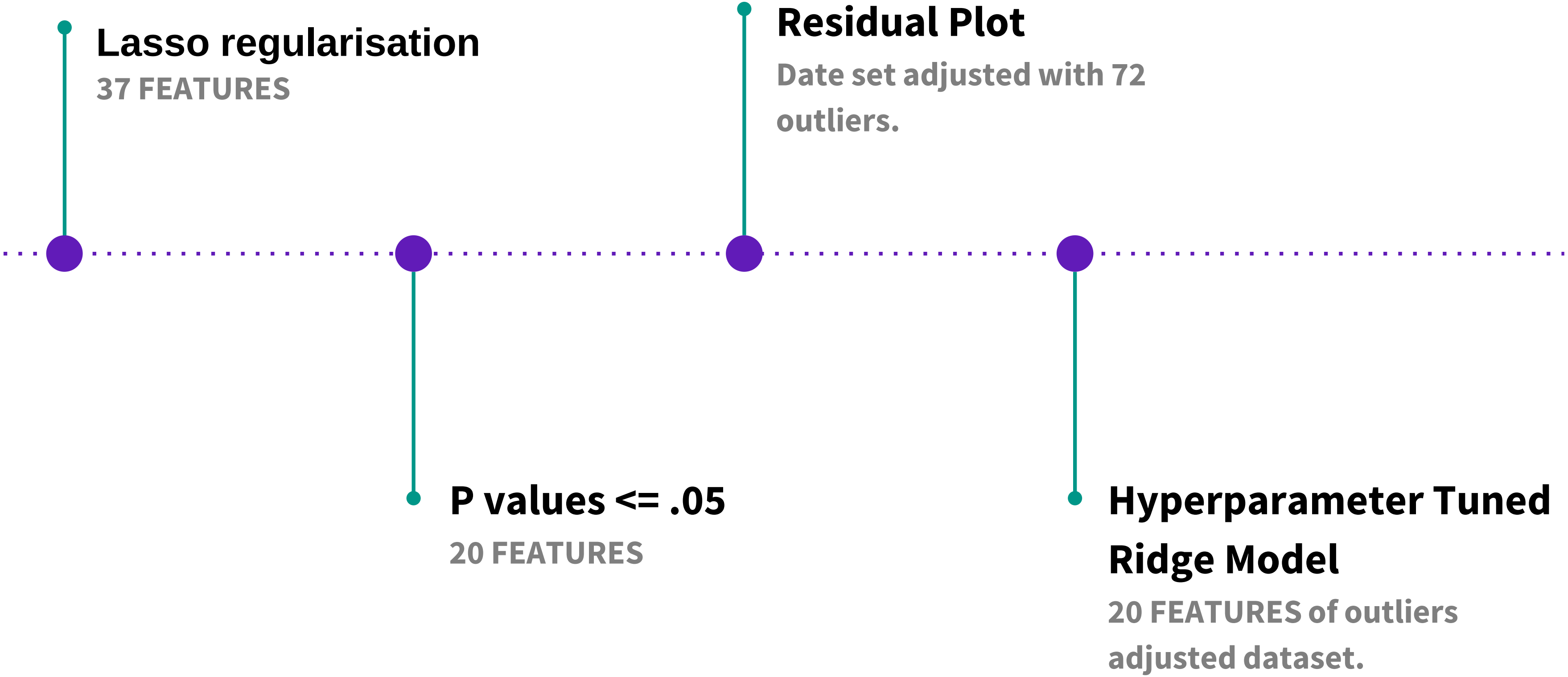


We create a new variable- 'newsqvar' which is a combination of 'TotalBsmtSF','1stFlrSF' and 'GrLivArea'. This new variable has a strong linear relation with SalePrice and a constant variance as well. The correlation also has improved from .70 to .76.

Transforming labels of categorical variables.

Basis our observations we find that there are categories, sensitive to Average Sale Price. For including these variables in the prediction models we need to assign them numerical labels. Here, have assigned the labels based on average SalePrice.

FEATURE SELECTION - RIDGE MODEL



Lasso regularisation

37 FEATURES

P values <= .05

20 FEATURES

Residual Plot

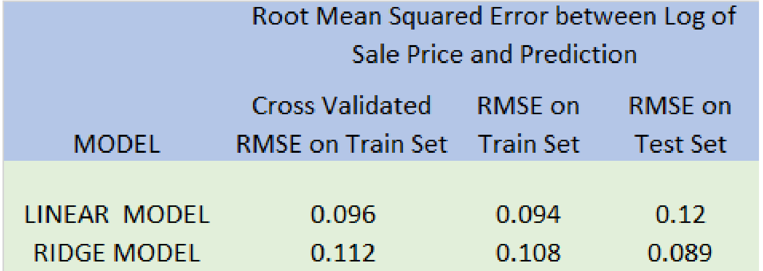
Date set adjusted with 72 outliers.

Hyperparameter Tuned Ridge Model

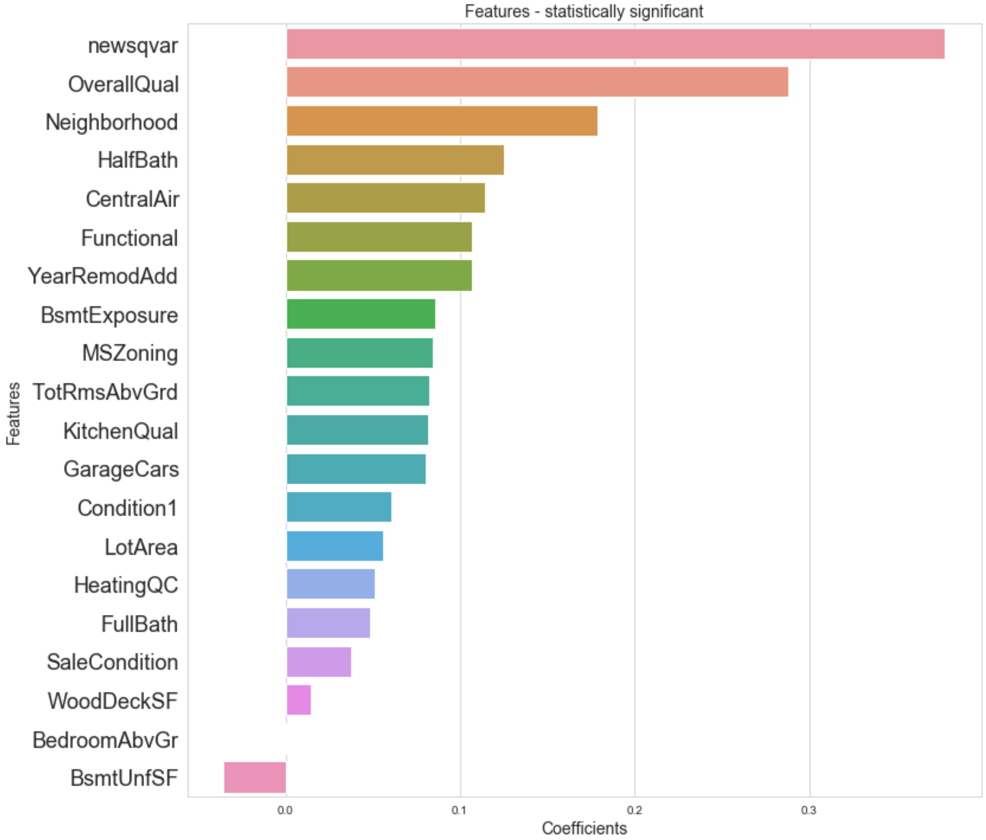
20 FEATURES of outliers adjusted dataset.

RIDGE MODEL- SCORE & FEATURES

SCORE

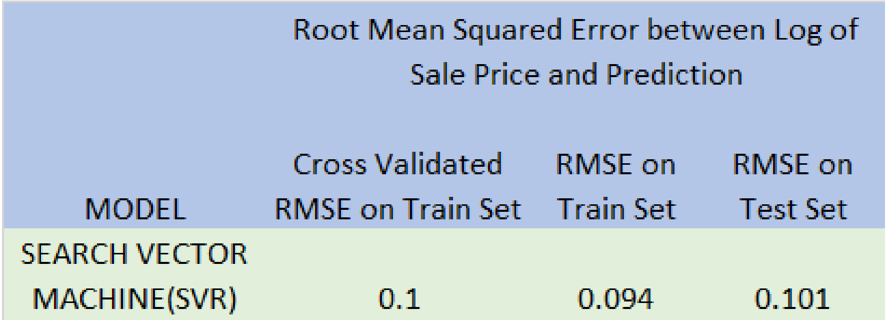


	Root Mean Squared Error between Log of Sale Price and Prediction		
MODEL	Cross Validated RMSE on Train Set	RMSE on Train Set	RMSE on Test Set
LINEAR MODEL	0.096	0.094	0.12
RIDGE MODEL	0.112	0.108	0.089



SEARCH VECTOR MACHINES (SVR)

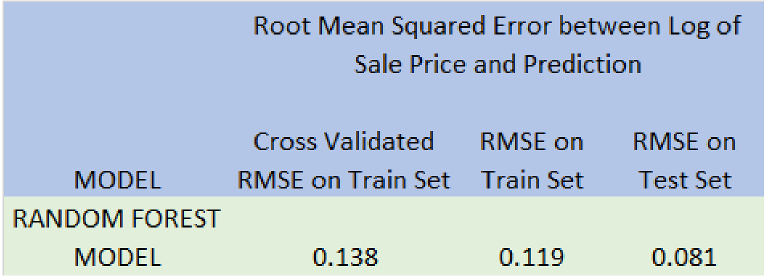
SCORE



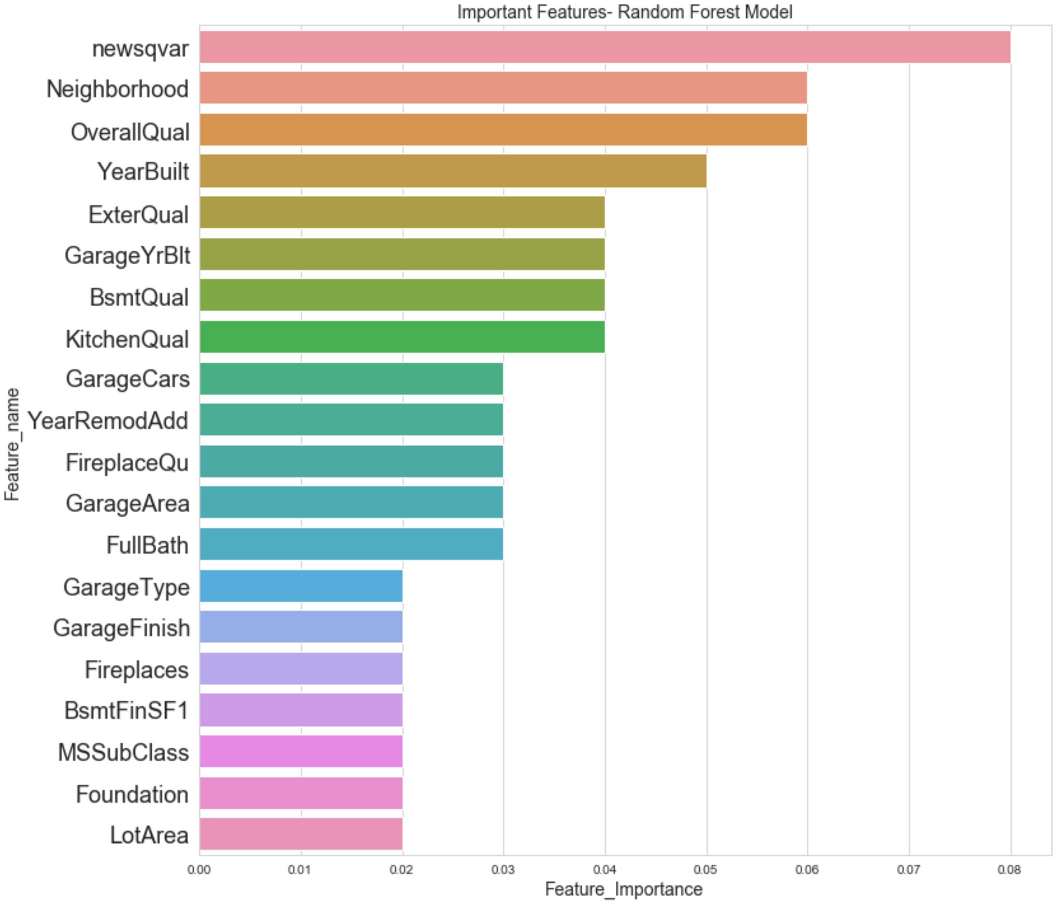
	Root Mean Squared Error between Log of Sale Price and Prediction		
MODEL	Cross Validated RMSE on Train Set	RMSE on Train Set	RMSE on Test Set
SEARCH VECTOR MACHINE(SVR)	0.1	0.094	0.101

RANDOM FOREST MODEL- SCORE

SCORE

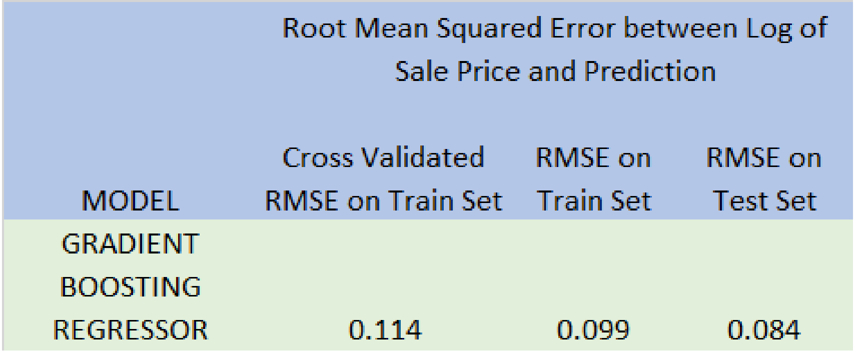


	Root Mean Squared Error between Log of Sale Price and Prediction		
MODEL	Cross Validated RMSE on Train Set	RMSE on Train Set	RMSE on Test Set
RANDOM FOREST MODEL	0.138	0.119	0.081



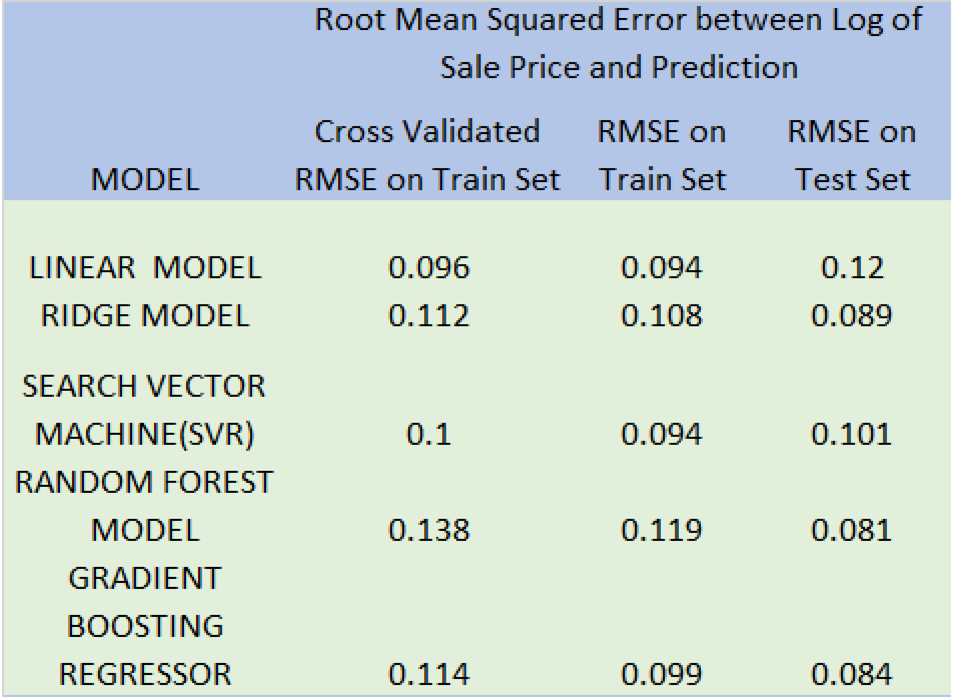
GRADIENT BOOSTING REGRESSOR MODEL- SCORE

SCORE



	Root Mean Squared Error between Log of Sale Price and Prediction		
MODEL	Cross Validated RMSE on Train Set	RMSE on Train Set	RMSE on Test Set
GRADIENT BOOSTING REGRESSOR	0.114	0.099	0.084

SUMMING UP



	Root Mean Squared Error between Log of Sale Price and Prediction		
MODEL	Cross Validated RMSE on Train Set	RMSE on Train Set	RMSE on Test Set
LINEAR MODEL	0.096	0.094	0.12
RIDGE MODEL	0.112	0.108	0.089
SEARCH VECTOR MACHINE(SVR)	0.1	0.094	0.101
RANDOM FOREST MODEL	0.138	0.119	0.081
GRADIENT BOOSTING REGRESSOR	0.114	0.099	0.084

Contact

aroonima12@gmail.com