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.

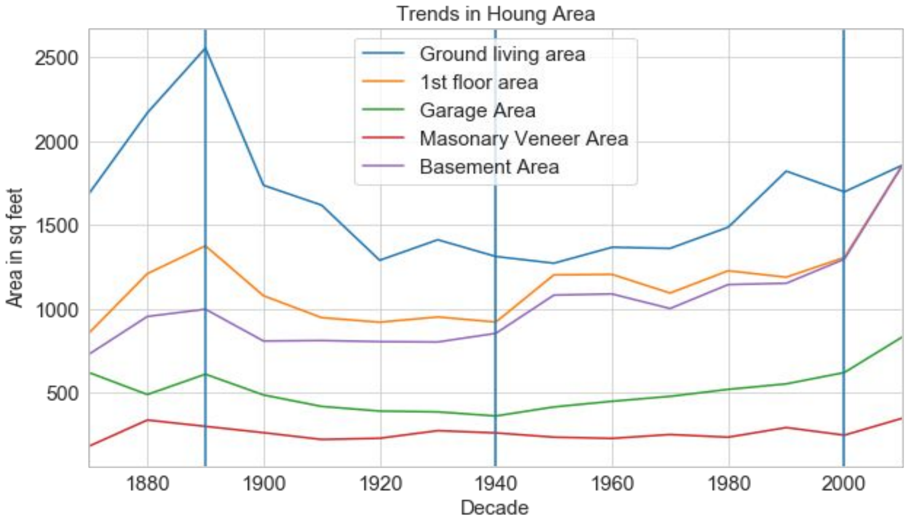
For a real estate company, which can also pose as a buyer or broker, the problem is to negotiate for the best deal.

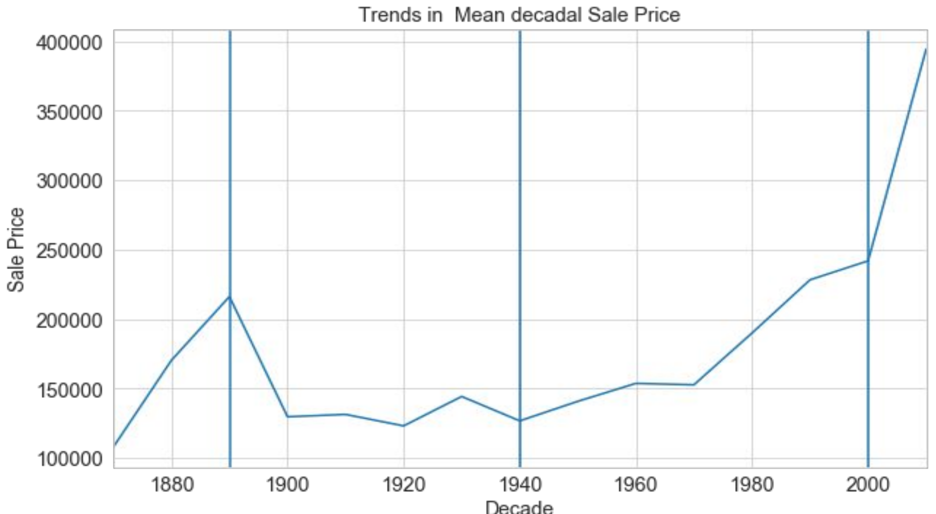
This dataset has several factors.

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

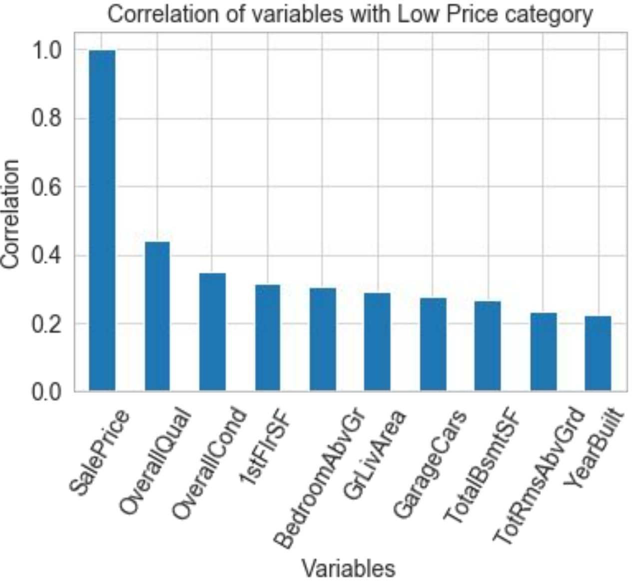
Trends in Housing Area

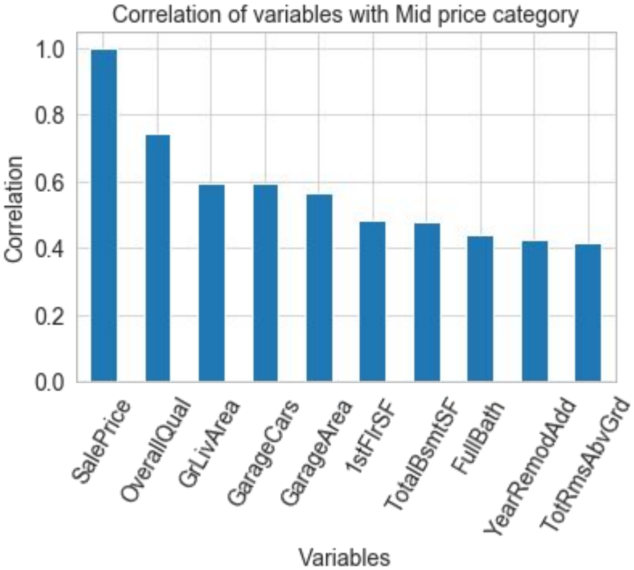
Observations: We can see 3 phases: an upward phase(till 1890), a downward and stable phase(till 1940) and an upward again (from 1940)
In 1890 the avg area of houses were big. Avg Ground living area in 1890 were biggest, which we don’t see today. We see a downward to a more stable phase till 1940. Increase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

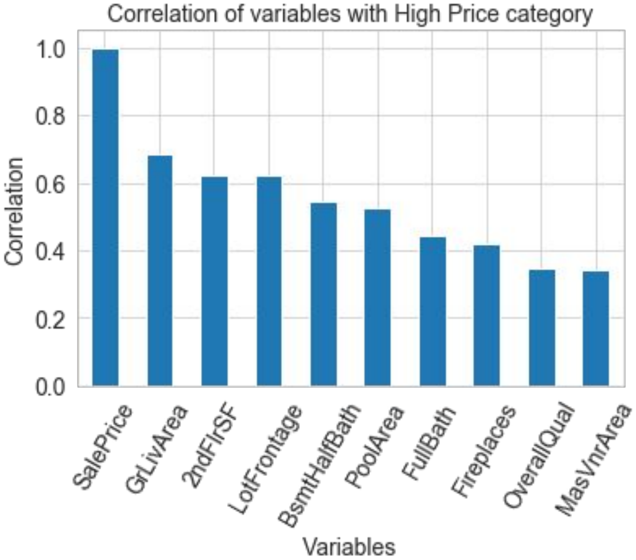




Correlation of variables with different Sale price ranges







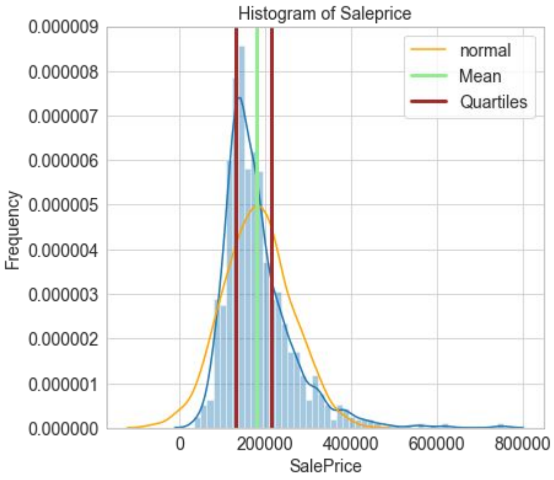
What price can be expected with these numeric features?

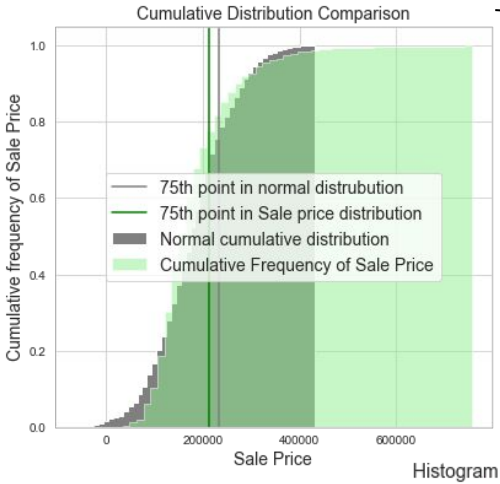
	Low Sale Price	Mid Sale Price	High Sale Price
SalePrice	106539	200382	512751
1stFlrSF	912	1234	2013
BsmtFinSF1	541	682	1292
BsmtUnfSF	566	634	813
GarageArea	395	530	842
GrLivArea	1120	1628	2863
LotArea	7724	11349	18066
TotalBsmtSF	842	1152	2016
WoodDeckSF	177	198	222

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'



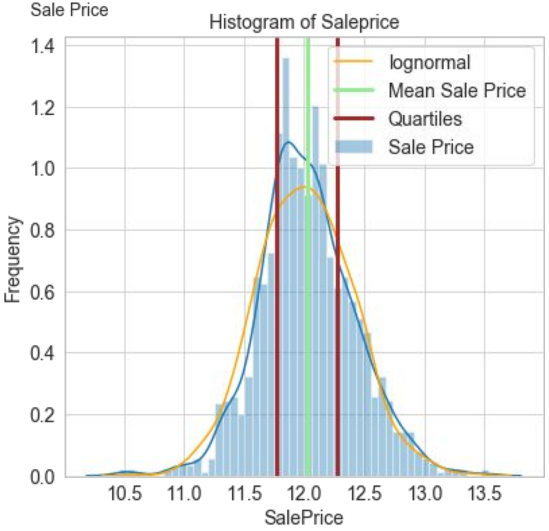


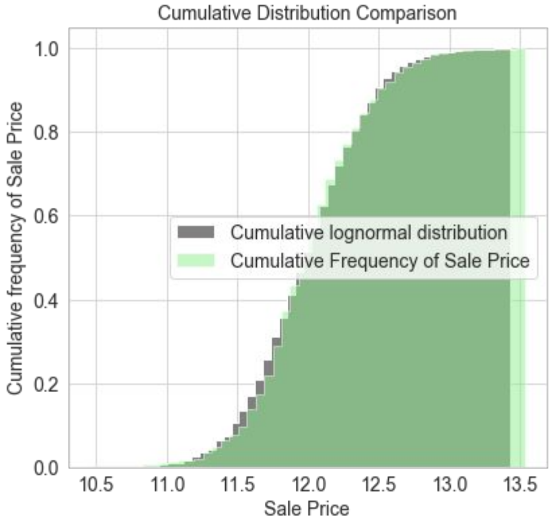
Observations:

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.





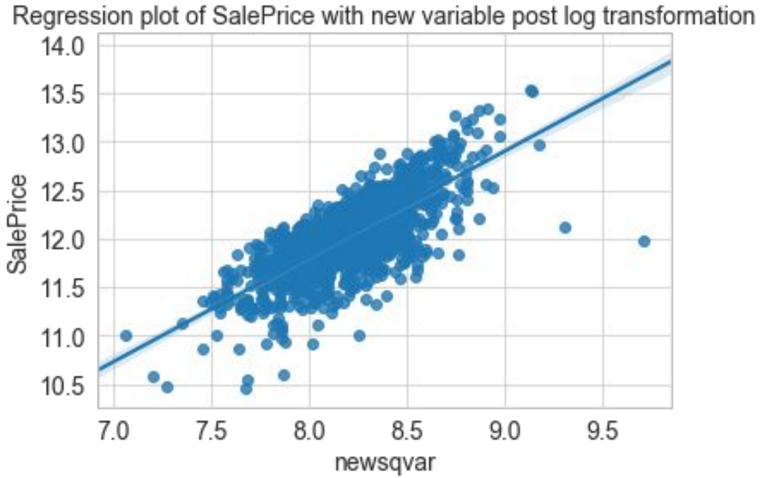
Observations:

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

Selection of Features with high coefficients.

P values <= .05

Selection of features with p-values less than .1

Residual Plot

Date set adjusted with high biased points identified by residual plot.

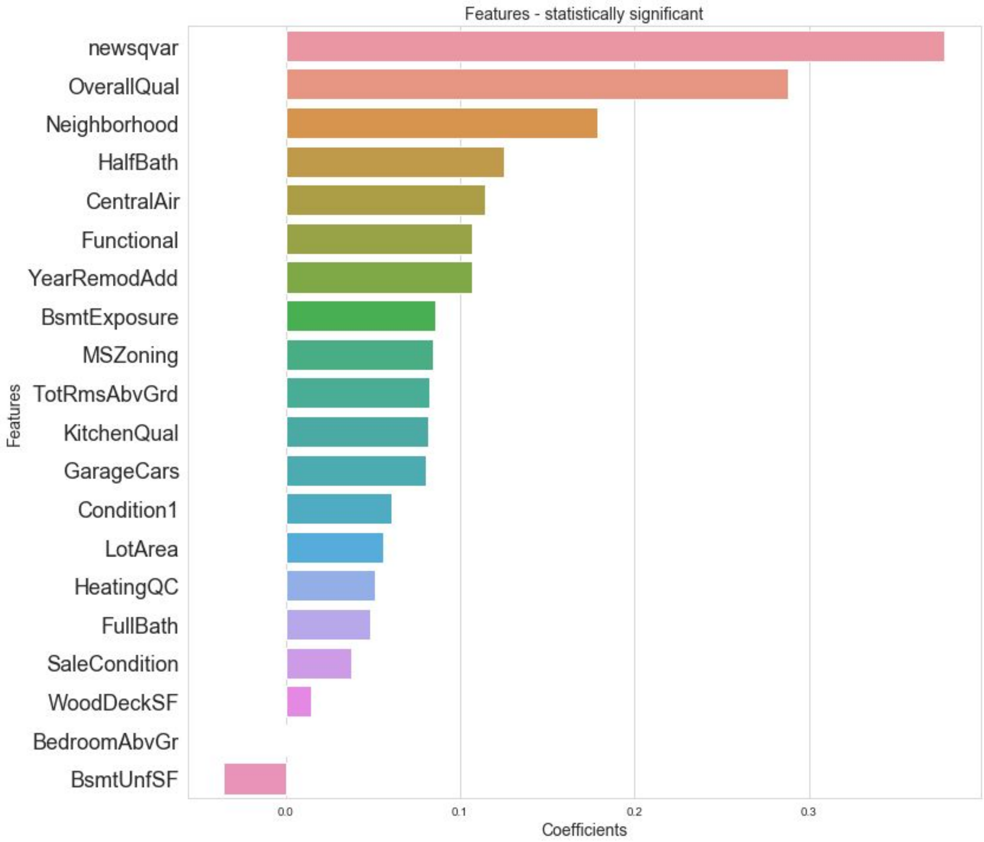
Hyperparameter Tuned Ridge Model

Best alpha value picked based on Grid SeArch CV

RIDGE MODEL- SCORE & FEATURES

SCORE

	Root Mean Squared Logarithmic error		
Algorithm	Cross Validated RMSE on Train Set	RMSE on Train Set	RMSE on test set
Linear Regression	0.094	0.092	0.13
Ridge	0.109	0.106	0.13



SUPPORT VECTOR MACHINES (SVR)

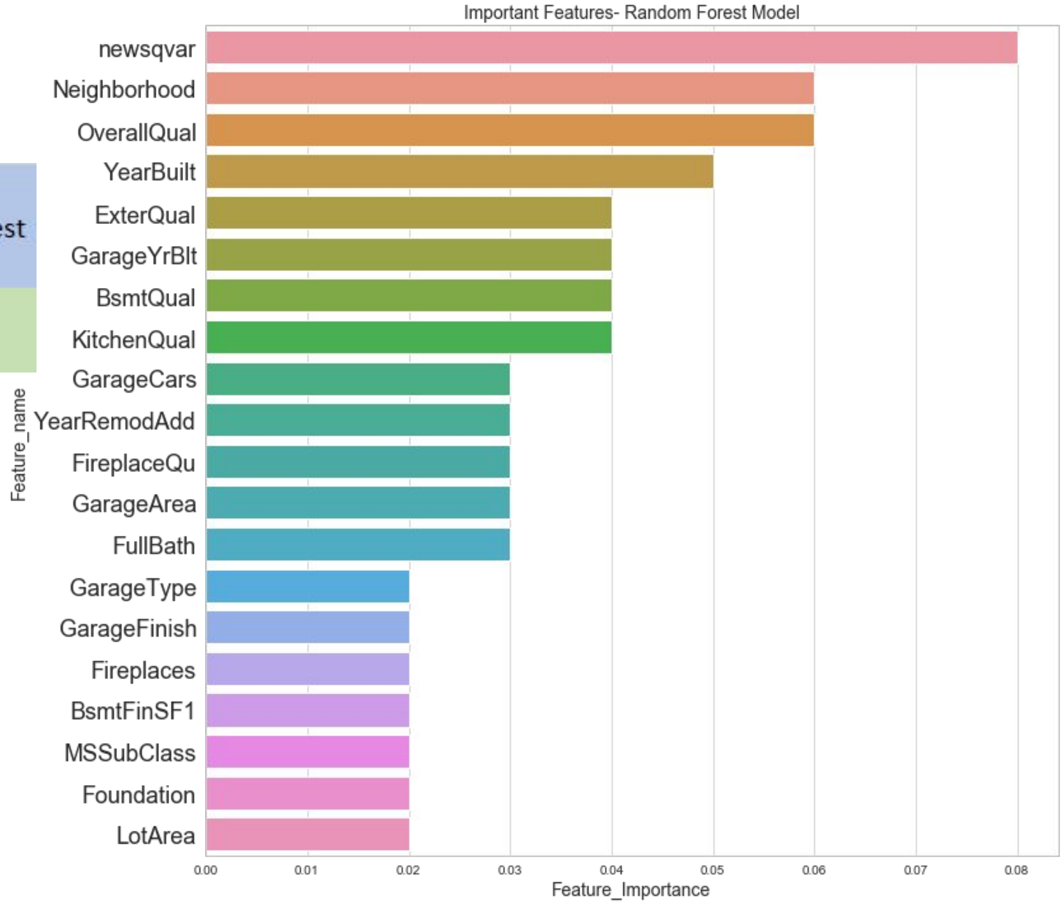
SCORE

Algorithm	Root Mean Squared Logarithmic error Cross Validated RMSE on Train Set	RMSE on Train Set	RMSE on test set
Support Vector Machines(for Regression)	0.101	0.094	0.14

RANDOM FOREST MODEL- SCORE

SCORE

	Root Mean Squared Logarithmic error		
Algorithm	Cross Validated RMSE on Train Set	RMSE on Train Set	RMSE on test set
Random Forest Regressor	0.137	0.103	0.17



GRADIENT BOOSTING REGRESSOR MODEL- SCORE

SCORE

	Root Mean Squared Logarithmic error		
Algorithm	Cross Validated RMSE on Train Set	RMSE on Train Set	RMSE on test set
Gradient Boosting Regressor	0.113	0.099	0.15

REAL WORLD APPLICATION OF MODEL

1. The client will be able to predict sale price of a house.
2. Various aspects or features that have a strong influence on price can be known.
3. The client can be in an advantageous position while negotiating.
4. The model can be useful to real estate agents and online companies as it would save additional costs and time in further examination and research.
5. Having an idea of the most influential features would enable the client to plan and effect changes in the property vis a vis the cost and expected return from investment. One can also decide what features need to be included for the house construction / renovation as per budget.

Contact

aroonima12@gmail.com