HOUSE PRICE PREDICTION

Submitted by:

Akash chaudhary

Intenship 17

ACKNOWLEDGMENT

The project housing price prediction is a problem with different prices of houses on the behalf of the different features of the houses. The data frame was given by my internship company Flip robo. My mentor mr. sajid chaudhary help me a lot to address the problems and solutions.

INTRODUCTION

- ## Business Problem Framing

Real estate is the biggest market in the world everyone wants their own houses. It's the biggest problem to find the exact price of the property or house. Whenever a new company want to comes in the real market or some individual shifted to the new city the want to know the price of the property there. The property price could be vary on many factors like location area garage space parking area, neighbourhood, house facing etc. so because of these many factors the prices changes everytime. So the consumer or the company took the idea of the prices of houses using machine learning.

- ## Conceptual Background of the Domain Problem

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world's economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

A US-based housing company named **Surprise Housing** has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below.

The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to

predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:

- • Which variables are important to predict the price of variable?
- • How do these variables describe the price of the house?

Business Goal:
You are required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

• Review of Literature

The data has collected by the surprise housing for the purpose to predict the prices of the houses of Australia market. Because the prices are different for every houses accourding to their various features like area locality parking stories type of the houses. So the company perform the different machine learning algorithms to find out the prices on which the make there business strategies. This machine learnings algorithms includes different models like decision tree regressor, linear regression, support vector regressor, ada boost regressor knearest neighbour etc. this methods helps to check the model efficiency and help in getting the approximate price of the house. So that company and consumer get the best price.

• Motivation for the Problem Undertaken

The real estate is growing business so it make my analytical skills strong and the market can demand this type of analysis. It also enhance my personal skills. To make this type of project and being the perfectionist in such type of work could help the companies and brokers and concumer to getting the idea of the price.

Analytical Problem Framing

- **Mathematical/ Analytical Modeling of the Problem**

The mathematical problems are included this was the huge data so getting null values is not a big deal. So there is many null values we need to tackle with them because of huge data there are many rows and columns i.e.80 columns first of all we need to understand all the columns than understand which will work for us to make prediction mode for prices. After these we need to check the data is balanced or not. Checking for outliers adjust them and some other works are accured at the time of making the model. We used random forest regressor to predict the prices and putting them into the test data to predict the prices of the houses.

- Data Sources and their formats

Data source was Flip Robo technologies. The project was assigned to me by mr.sajid chaudhary.

The data was drives by Surprise Housing Company for making the machine learning model to predict the price for entering into the new Australian market for real estate. The data contains 1168 rows and 81 columns. They are both in integer and string forms.

```
#importing dataset
df= pd.read_csv('train.csv')
df
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal
0	127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
1	889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
2	793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
3	110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0
4	422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1163	289	20	RL	NaN	9819	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0
1164	554	20	RL	67.0	8777	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0
1165	196	160	RL	24.0	2280	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0
1166	31	70	C (all)	50.0	8500	Pave	Pave	Reg	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0
1167	617	60	RL	NaN	7861	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0

1168 rows × 81 columns

- **Data Preprocessing Done**

First of all we got 2 data frames test and train data we need to clean up both so first we need to clean the train data in the train data there are 1168 rows and 81 columns inclusive of nan values. So we decide to split the string and numeric columns than fill them all with different mean and mode technique. To clean the data. Than we correlate the targeted variable with some important independent variables.

- **Data Inputs- Logic- Output Relationships**

Data inputs are given in the dataset we need to filter both train and test data we can say the train data was the input and test is the output because we were need to train the model on the behalf of train data and put them in the test data. So we have done all the data cleaning steps to the test data too. After getting the best model put that model on test.

- **State the set of assumptions (if any) related to the problem under consideration**

Presumptions are includes the data was only for domestic purpose because there was not any commercial space.

- **Hardware and Software Requirements and Tools Used**

We have used jupyter notebook to make the model and imported some python libraries these are includes numpy, pandas, maths, stats, seaborn, matplotlib, sklearn etc.

Model/s Development and Evaluation

- ## Identification of possible problem-solving approaches (methods)

It was a regression problem that's why we have used linear regression. We have used random forest regressor because it gives as the best r2 score and its cross validation score was also good. And the difference between accuracy and cross validation score is very less. So we took this and also done the hyper parameter tunning to tune the data.

- ## Testing of Identified Approaches (Algorithms)

Algorithms which are used for training and testing are:

Linear regression

Lasso regression

Decision tree regressor

K Neighbors regressor

Decision tree regressor

Random forest regressor.

- ## Run and Evaluate selected models

There are the snapshots of the models that we used:

1. Lasso

```
Lasso()
fit score : 0.8511948744276692
r2 score 0.6684146324912148
mean absolute error 21901.392323257453
root mean squered error 46663.0109290083
```

2. Decision tree regressor

```
DecisionTreeRegressor()
fit score : 1.0
r2 score 0.6242793239685194
mean absolute error 29901.58219178082
root mean squered error 49671.54024345867
```

3. KNeighbors regressor

```
KNeighborsRegressor()
fit score : 0.8533746915569522
r2 score 0.7199214423294367
mean absolute error 23573.731506849315
root mean squered error 42885.94755376299
```

4. Linear regressor

```
LinearRegression()
fit score : 0.8511948873517692
r2 score 0.6683973494795625
mean absolute error 21901.471561144077
root mean squered error 46664.22700627232
```

5. Decision tree regressor

```
RandomForestRegressor()
fit score : 0.9776608111049723
r2 score : 0.8315798298858463
mean absolute error:  19973.497054794523
root mean squared error : 33256.162251102476
Cross_val_score for Lasso(alpha=5) is 0.7619980245913898
```

So we have chose the decision tree regressor because of its best fit r2 score and cross validation score.

- **Key Metrics for success in solving problem under consideration**

The key metrics which are used are:

Fit score

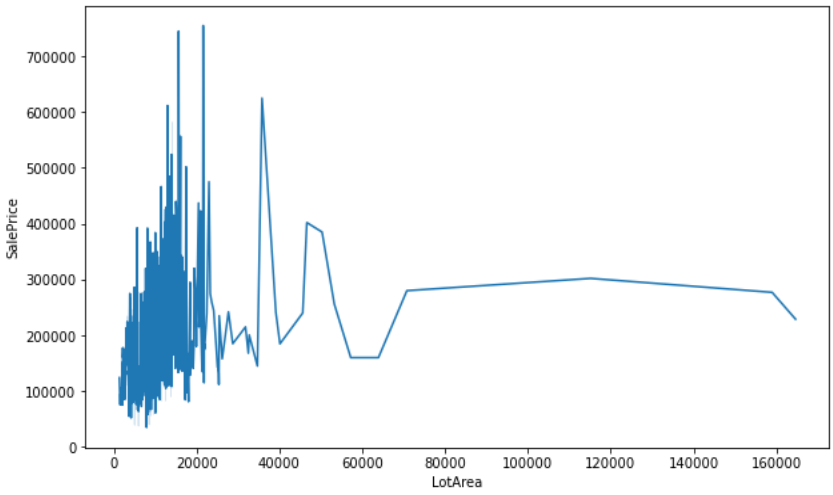
R2_score

And also check the cross val score. Because the difference between the fit score and cross validation score indicates models accuracy.

- **Visualizations and Interpretation of the Results**

The visualizations are includes:

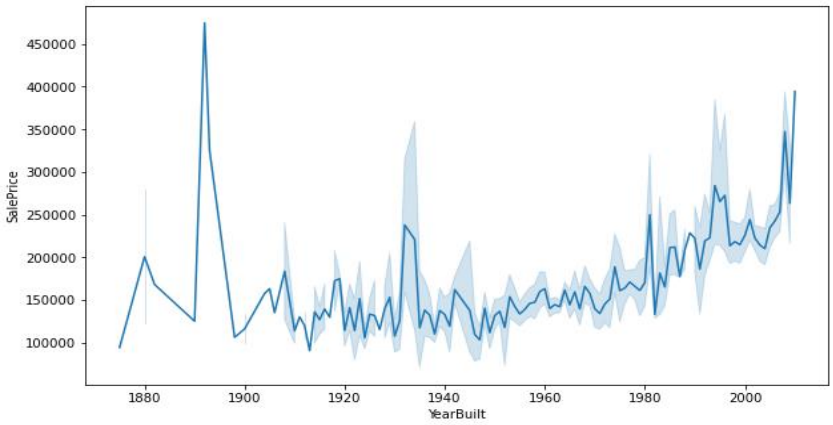
1. Lot area and sales price



property area under 20000 sq.ft. are more in the town and maximum plotting price lies under 6000 too 40000.

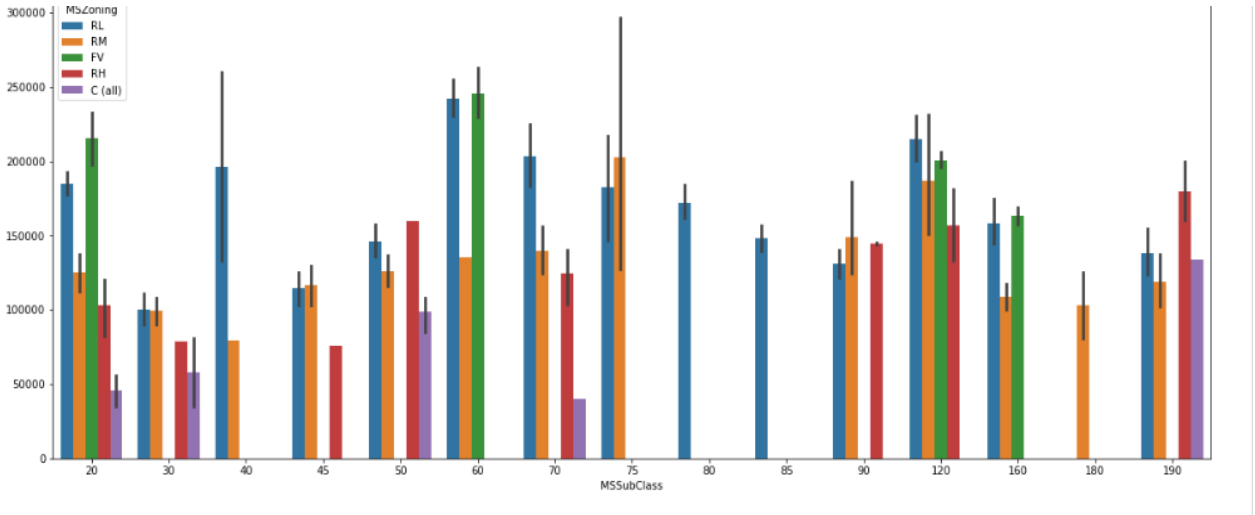
so there is the demand of 0-20000 square feet plot area.

2. Year built and sales price.



newly built house prices are higher than old house which are made before 1980's. only some very old unique houses prices are high maybe they builted with unique style.

3. Subclass, ms zone and sales price.

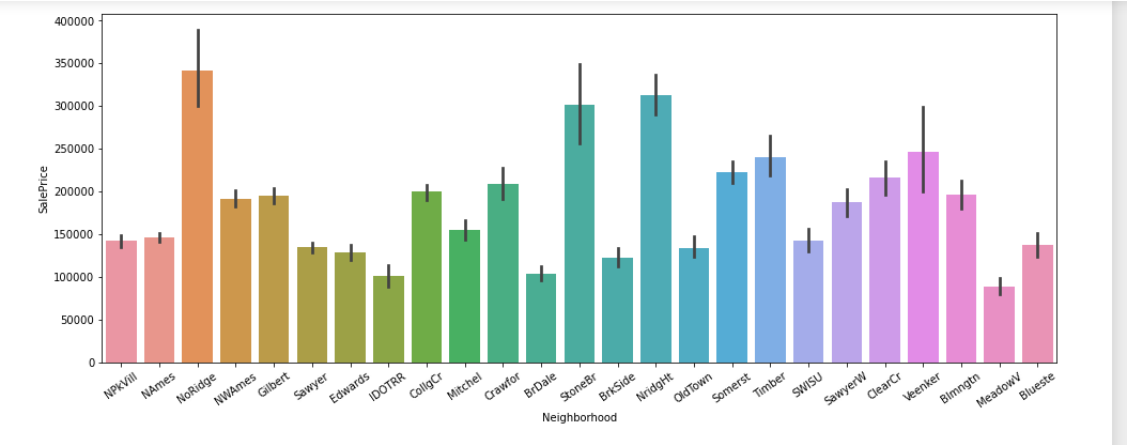


there are diffrent types of zones each zone having diffrent price ranges. in the barplot we can see RL zone areas having huge demand and prices are also as demand high.

dwellings in houses are have average demand and sales

commercial spaces doesn't have high demands.

4. Neighborhood and salesprice

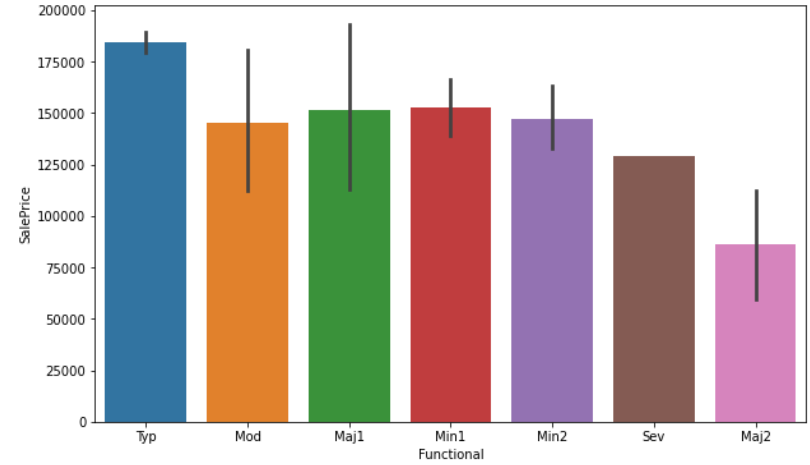


noRidge,stoneBR and NridgeHT is kind of porsh area of something thire is price of houses are very high

tomber,somerst,veenker are having average pricebetween 20000-25000 so these area are more beneficial.

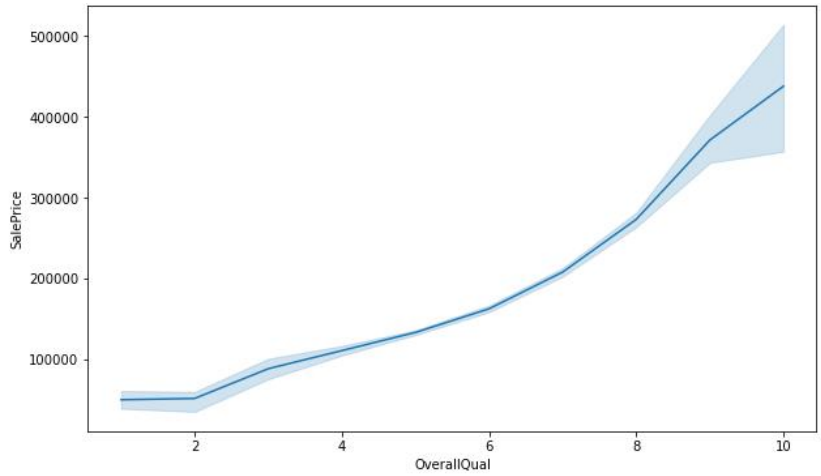
meadwV,Idotrr anf Brdale available in low price range

5. Functionality and sales price



houses with typ function having high price as compare to other functions. and mod,maj1,min1 and min2 having average price between 125000 to 150000 so this unctions are most considerable.

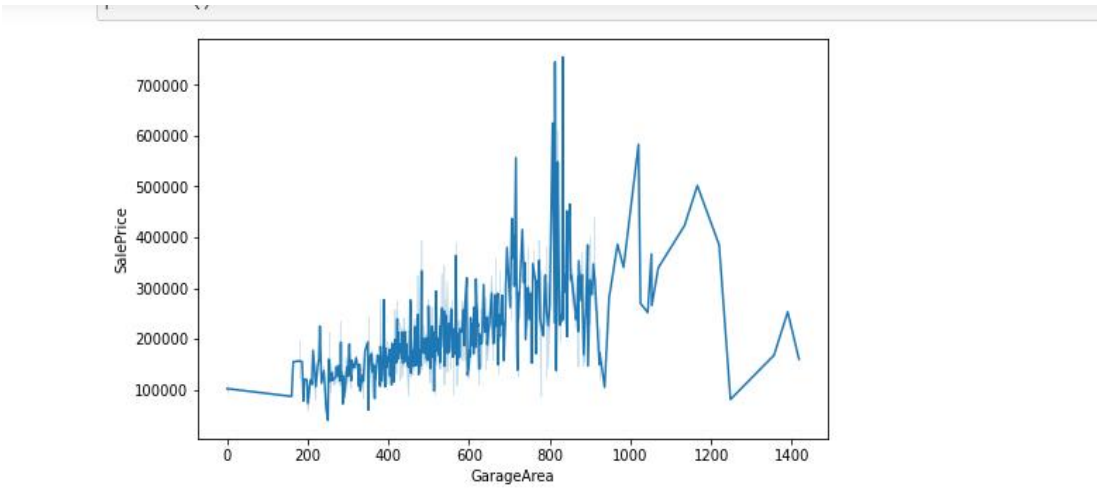
6. Overall quality and sales price



overall quality increases with sales price with property.

overallquality between 4 to 8 are having good market.

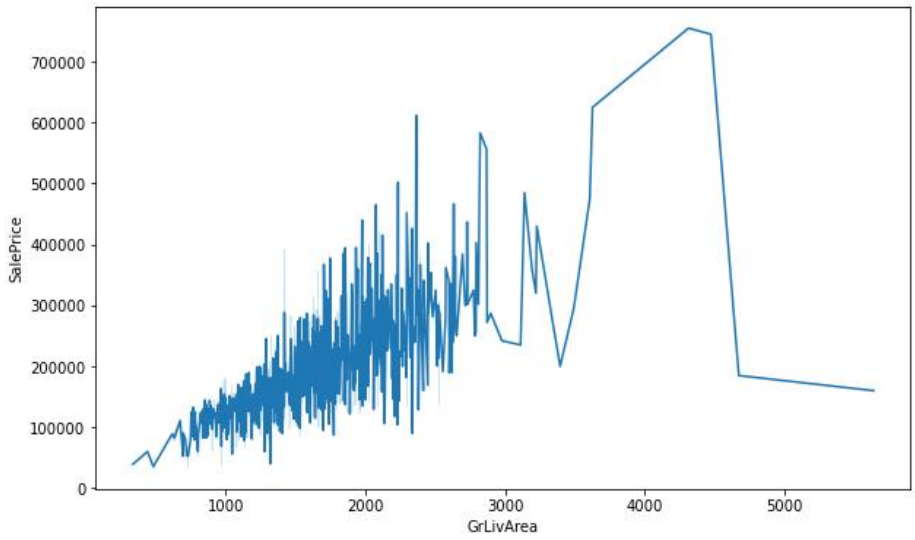
7. Garage area and sales price



the slope is upwards. it shows the incresing trend.

the chart showing us the greater the ground area will increase the price.

8. Gr living area and sales price



there is much fluctuation in the data but thee trend is upwards shifting

the bars shows the greater the ground living area will higher the price.

CONCLUSION

- **Key Findings and Conclusions of the Study**

The real estate is the biggest market in the world and pricing or the property is the toughest thing so we have used the linear regression because it's a regression problem. And used the different machine leaning algorithms to predict the prices of the property on the bases of the previous data and the factors which effect the price of the property like area, neighbourhood, garage space, facing, pool area, parking capacity etc.

The conclusion is many factor as I mentioned earlier could effect the prices of the property. The newly built houses have more demand and there prices are also little high. Parking capacity also impacts the price if the garage area is large the price will increase.

- ## Learning Outcomes of the Study in respect of Data Science

Because the data is huge the number of rows and columns are also high so these make my understanding of data increased. I have tried many things to clean up the data than find with are used for prediction and make the model powerful I have correlate many useful independent variables with the targeted variable so that I can understand the data more. After visualisation I have correlate the data with target in numbers to understand which independent variables gives how much impact on the targeted variable. So after all of these I used standard scaler to scale the data after scalling the data train the model than used the different machine learning models. And get the best score in random forest regressor than doing hyper parameter tuning etc. these steps improve my machine learning and model building skill. And these model surely will help the consumers, companies, property agents to estimate the price of a particular property.

- ## Limitations of this work and Scope for Future Work

Because we get the good r2 score and we have choose the best one model. We can assume the estimated price are almost correct. But we all know property prices are very dynamic it could change with many other factors. So we can took the estimated prices but also need to see the market trend.