FLIP ROBO

PROJECT REPORT

HOUSING: PRICE PREDICTION

SME name:

SWATANK MISHRA SIR

Prepared by:

ABHISHEK KUMAR

BATCH - 29

ACKNOWLEDGMENT

It is my sensual gratification to present this report on the HOUSING: PRICE PREDICTION project. Working on this project was an incredible experience that has given me very informative knowledge.

I would like to express my sincere thanks to my **SME SWATANK MISHRA SIR** for a regular follow-up and valuable suggestions provided throughout.

And I am also thankful to Flip Robo Technologies Bangalore for their guidance and constant supervision and for providing necessary information regarding the project and their support in completing the project.

INTRODUCTION

Business Problem Framing

Houses are one of the necessary needs of every person around the globe. Therefore, the housing and real estate market is one of the markets that is one of the major contributors to the world's economy.

It is a very large market and various companies are 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, improve their marketing strategies and focus on changing trends in house sales and purchases.

Predictive modelling, Market mix modelling, and recommendation systems are machine learning techniques used to achieve the business goals for housing companies. Our problem is related to one such housing company

We are required to build a model using machine learning algorithm. 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 a high return. Further, the model will be a good way for the management to understand the pricing of a new market.

Conceptual Background of the Domain Problem

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 company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning 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 variables?
- How do these variables describe the price of the house?

Motivation for the Problem Undertaken

The growing un-affordability of housing has become one of the major challenges for metropolitan cities around the world. To gain a better understanding of the commercialized housing market we are currently facing, we want to figure out what are the top influential factors of the housing price. Apart from the more obvious driving forces such as inflation and the scarcity of land, several variables are worth looking into. Therefore, we choose to study the house prices predicting problem on Kaggle, which enables us to dig into the variables in-depth and to provide a model that could more accurately estimate home prices. In this way, people could make better decisions when it comes to home investment.

Our object is to discuss the major factors that affect housing prices and make precise predictions for them. We use 79 explanatory variables including almost every aspect of residential homes in Ames, Iowa. Methods of both statistical regression models and machine learning regression models are applied and further compared according to their performance to better estimate the final price of each house. The model provides price prediction based on similar comparables of people's dream houses, which allows both buyers and sellers to better negotiate home prices according to the market trends.

Analytical Problem Framing

Mathematical/ Analytical Modelling of the Problem

As we have provided with two sets of data, one is for train and another test dataset. Here we need to build a machine learning model using train data set and then by using that model we will make predictions for the test data set.

Both the data sets are in CSV format, the **training dataset has 1168 rows and 81 columns** and the **test data set has 292 rows**. For similar processing of both datasets, I will combine both the data sets and do EDA, skewness treatment, data pre-processing, and other necessary steps.

And as we have to predict house prices in this problem so I will be using different regression models.

```
#lets add source column to train and test dataset
df_train["source"] = "train"
df_test["source"] = "test"

#lets combine both the datasets
df = pd.concat([df_train,df_test],ignore_index=True)
df
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolQC	Fence	MiscFeature	MiscVal	MoSold
0	127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	...	NaN	NaN	NaN	0	2
1	889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	...	NaN	NaN	NaN	0	10
2	793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	...	NaN	NaN	NaN	0	6
3	110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	...	NaN	MnPrv	NaN	0	1
4	422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	...	NaN	NaN	NaN	0	6
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1455	83	20	RL	78.0	10206	Pave	NaN	Reg	Lvl	AllPub	...	NaN	NaN	NaN	0	10
1456	1048	20	RL	57.0	9245	Pave	NaN	IR2	Lvl	AllPub	...	NaN	NaN	NaN	0	2
1457	17	20	RL	NaN	11241	Pave	NaN	IR1	Lvl	AllPub	...	NaN	NaN	Shed	700	3
1458	523	50	RM	50.0	5000	Pave	NaN	Reg	Lvl	AllPub	...	NaN	NaN	NaN	0	10
1459	1379	160	RM	21.0	1953	Pave	NaN	Reg	Lvl	AllPub	...	NaN	NaN	NaN	0	6

1460 rows × 82 columns

Data Sources and their formats

```
df.info()
```

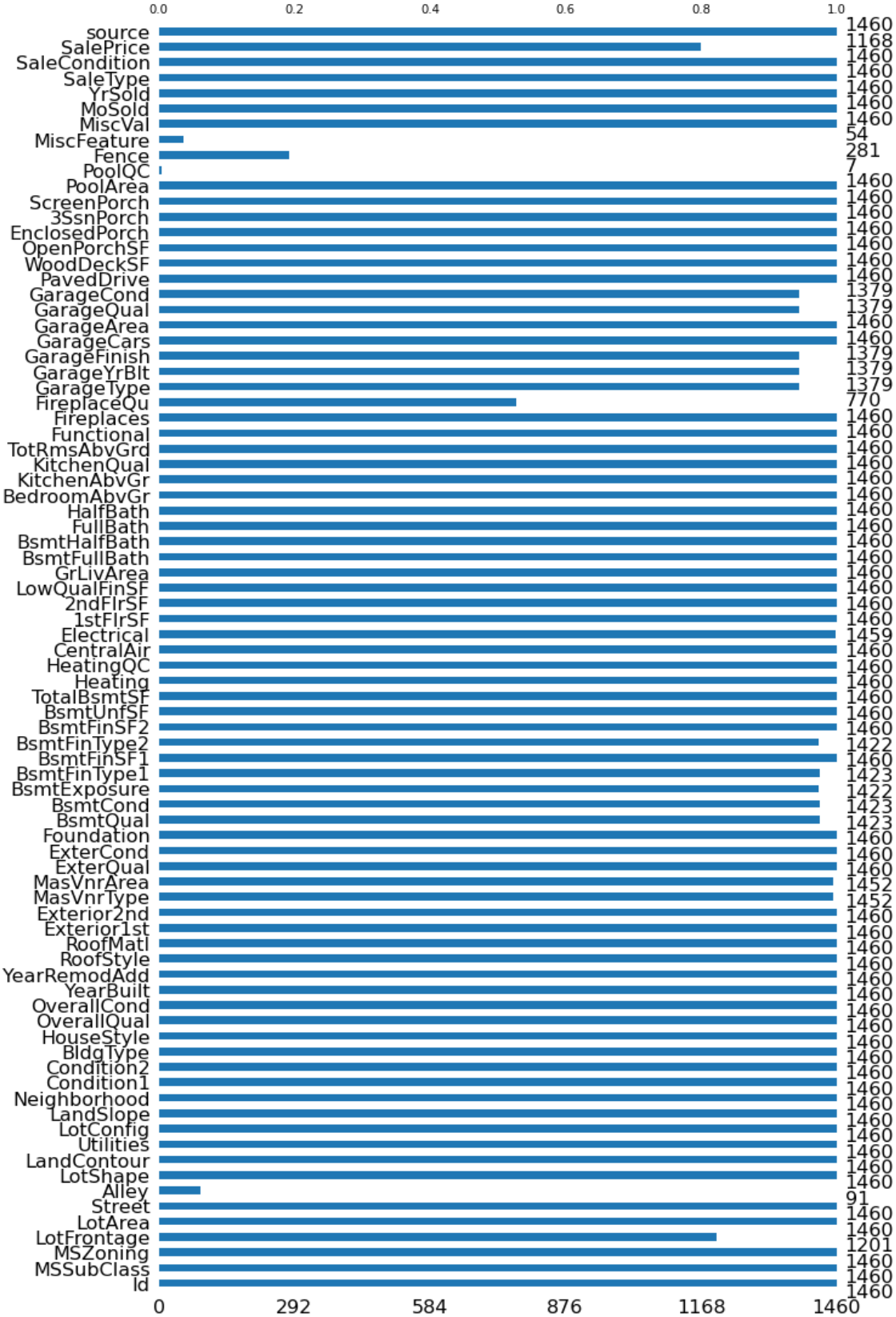
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 82 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Id             1460 non-null   int64
 1   MSSubClass     1460 non-null   int64
 2   MSZoning       1460 non-null   object
 3   LotFrontage    1201 non-null   float64
 4   LotArea        1460 non-null   int64
 5   Street         1460 non-null   object
 6   Alley          91 non-null     object
 7   LotShape       1460 non-null   object
 8   LandContour    1460 non-null   object
 9   Utilities      1460 non-null   object
 10  LotConfig      1460 non-null   object
 11  LandSlope      1460 non-null   object
 12  Neighborhood   1460 non-null   object
 13  Condition1     1460 non-null   object
 14  Condition2     1460 non-null   object
 15  BldgType       1460 non-null   object
 16  HouseStyle     1460 non-null   object
 17  OverallQual    1460 non-null   int64
 18  OverallCond    1460 non-null   int64
 19  YearBuilt      1460 non-null   int64
 20  YearRemodAdd   1460 non-null   int64
 21  RoofStyle      1460 non-null   object
 22  RoofMatl       1460 non-null   object
 23  Exterior1st    1460 non-null   object
 24  Exterior2nd    1460 non-null   object
 25  MasVnrType     1452 non-null   object
 26  MasVnrArea     1452 non-null   float64
 27  ExterQual      1460 non-null   object
 28  ExterCond      1460 non-null   object
 29  Foundation     1460 non-null   object
 30  BsmtQual       1423 non-null   object
 31  BsmtCond       1423 non-null   object
 32  BsmtExposure   1422 non-null   object
 33  BsmtFinType1   1423 non-null   object
 34  BsmtFinSF1     1460 non-null   int64
 35  BsmtFinType2   1422 non-null   object
 36  BsmtFinSF2     1460 non-null   int64
 37  BsmtUnfSF      1460 non-null   int64
 38  TotalBsmtSF    1460 non-null   int64
```

Here we have provided train and test data sets separately which are in CSV format.

The train data set is having 1168 rows and 81 columns and the test data set is with 292 rows and 80 columns. I have combined both of these data sets for EDA and equal data pre-processing for both. We will build a machine learning model using train data to predict the 'SalePrice' of houses and by using this model we will predict the house price for the test data set.

Looking at the info of our data I came to know that our data is having 38 numerical columns and 43 categorical columns.

Now let's check the missing values in the dataset:



Missing Values

Here we can conclude,

- Some columns have more than 80 % of null values so we can drop them:
  - Alley: 93.77 %
  - Misc Feature: 96.3 %
  - Pool QC: 99.52 %.
  - Fence: 80.75 %

- Also, some columns which have eye-catching null values:
  - Lot Frontage: 17.74 %
  - Fire place Qu : 47.26

Looking at the above figure we can say that our data is having a large number of missing values. And there are many ways to fill these values. I will drop some columns which are having more than 80% of missing values.

Missing values from some categorical columns which are many null values have been replaced by 'None'. Missing values from numerical columns which are having a large number of null values are replaced with '0'.

And for columns are having fewer missing values I have replaced categorical data with mode and numerical data with the mean of that particular column.

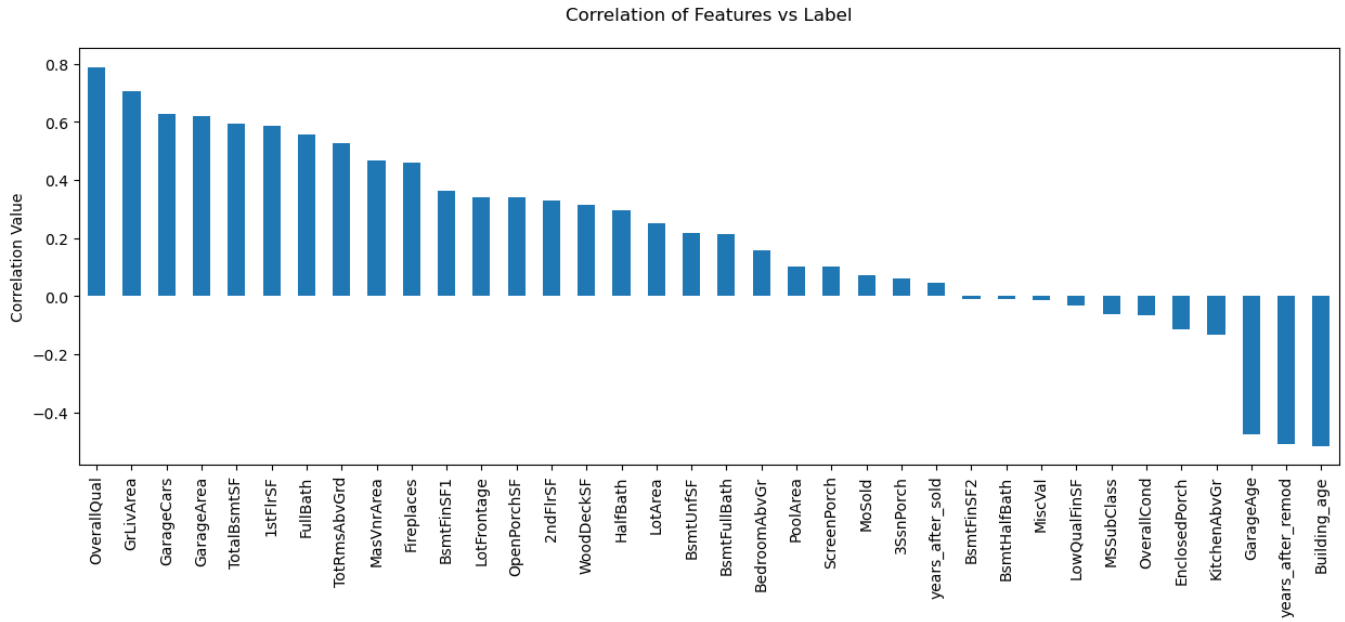
Data Pre-processing Done

Data pre-processing is a very important process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. I have used the following pre-processing steps

- Filling or Treating Missing values
- Handling Multi colinearity
- Outliers treatment
- Encoding
- scaling
- Skew-ness treatment

Data Inputs- Logic- Output Relationships

To analyze the relationship between our features and the target variable I have done an EDA to know the contribution of various features to the prediction. And got to know that which are the important features and which are not much. For EDA I have used different plots like distribution plots, scatter plots, box plots, strip plots, heat-map, etc.

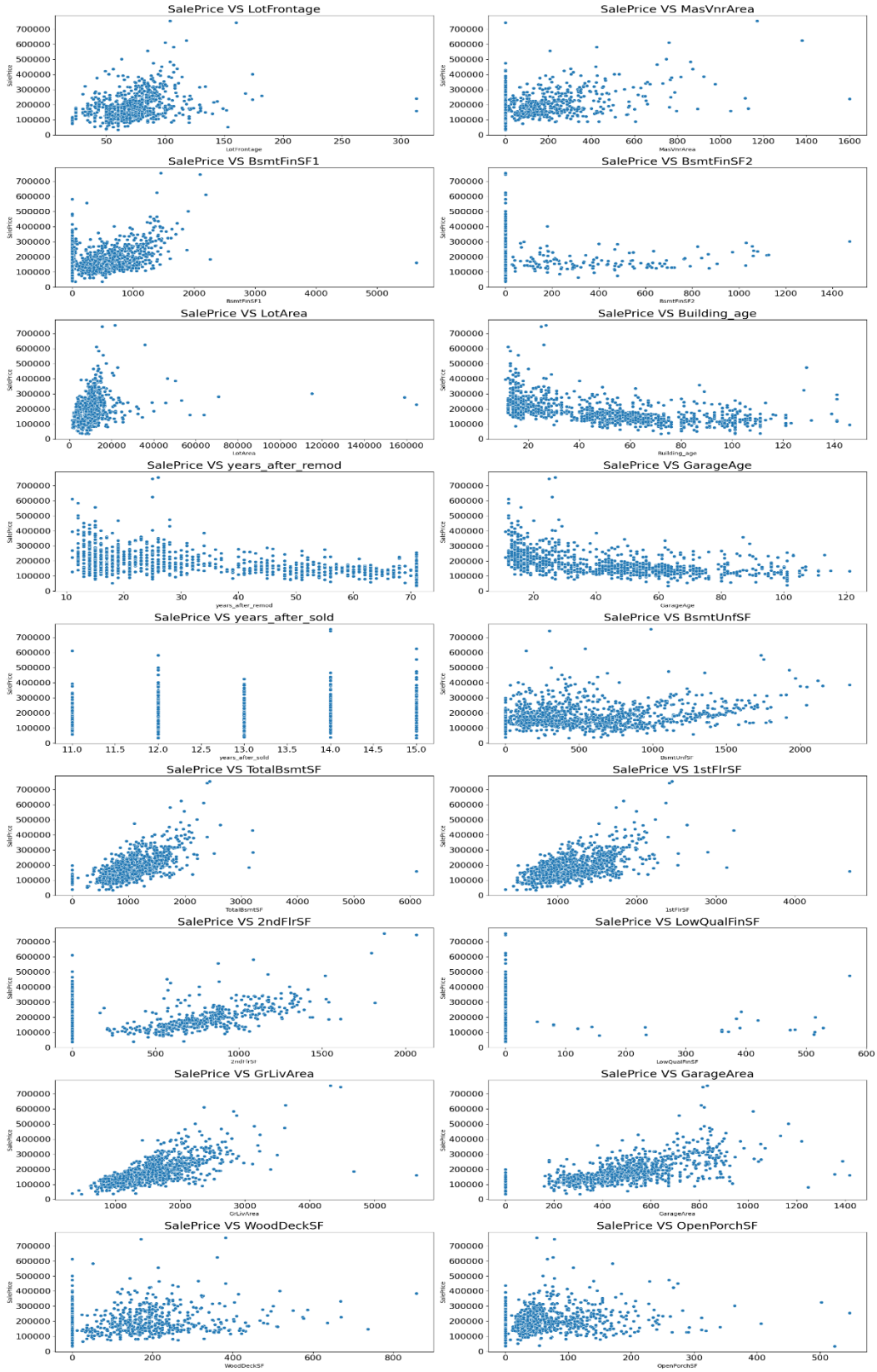


EDA:

The above plot shows us the relation or contribution of various numerical features with our target variable.

- looking at the above plot we can say that the features like ‘Overall quality, ‘grlivArea’, ‘GarageCars’, GarageArea’ are having maximum positive relation with the target variable.
- And features like ‘Building age’, ‘years_after_remod’, ‘GarageAge’ are in negative relation with the target variable.

Now we plot Scatter plots of some numerical column features vs SalePrice:

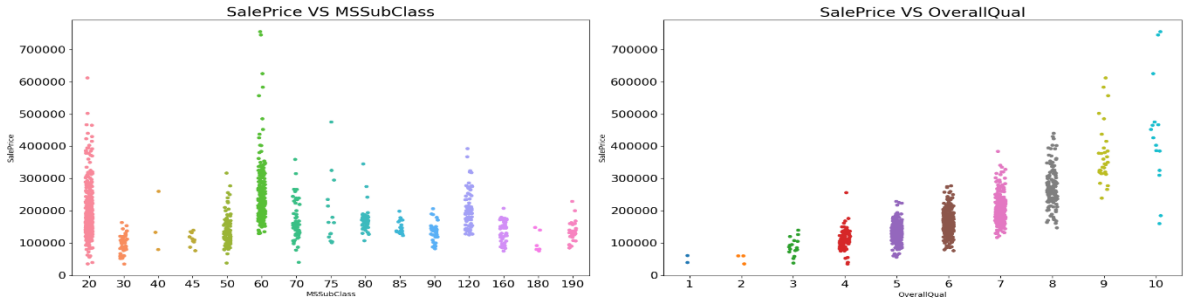


❖ **Observations:**

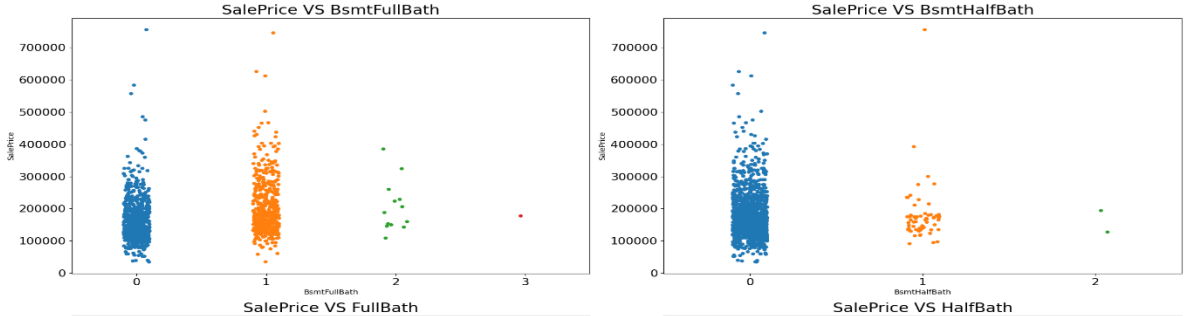
The above scatter plots are showing the relation between some numerical features vs sales price.

- If the linear feet of street-connected to the property is more, then the sales price is higher.
- The Masonry veneer area increases the price of the house also gets higher.
- If the house is with more area of type 1 finished the price is higher.
- There are less number of houses with type 2 finish as the area for type two is given zero for most houses.
- As the lot size gets higher the price also will increases.
- The buildings which have been built long back are having fewer sales price compared to new buildings. Similarly, if Building modification has been done long back then the price is less.
- If garages have been built recently the building is having a higher sales price
- We can see almost all the buildings have been sold in recent years.
- We can see that in most of the cases the Unfinished basement area is below 1000 square feet it will also tell us that the Sales prices will increase as the total basement area increases.
- We can see that most houses are having more area on the first floor compared to the second floor And some of the houses are not having rooms on the second floor.
- Sales prices increase with the floor area.
- We can see most of the houses are with very less area finished with low quality, and also it is observed that more area finished with low quality causes a reduction in sale price.
- And as above grade (ground) living area and garage area increases the sale price also increases.

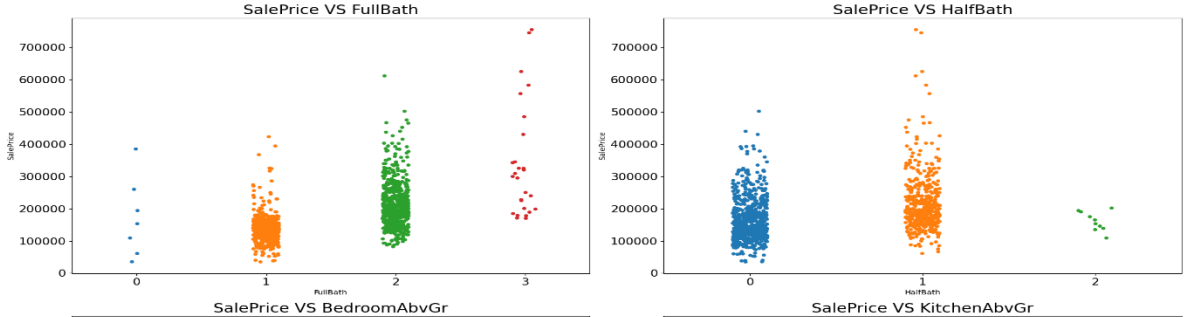
After this I have decided to plot a Strip plot of some numerical features vs SalePrice:


SalePrice VS MSSubClass
SalePrice
MSSubClass

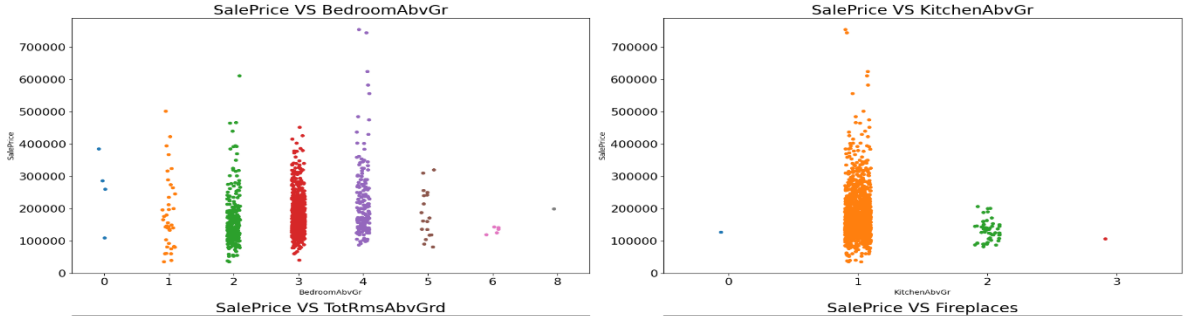
SalePrice VS OverallQual
SalePrice
OverallQual


SalePrice VS BsmtFullBath
SalePrice
BsmtFullBath

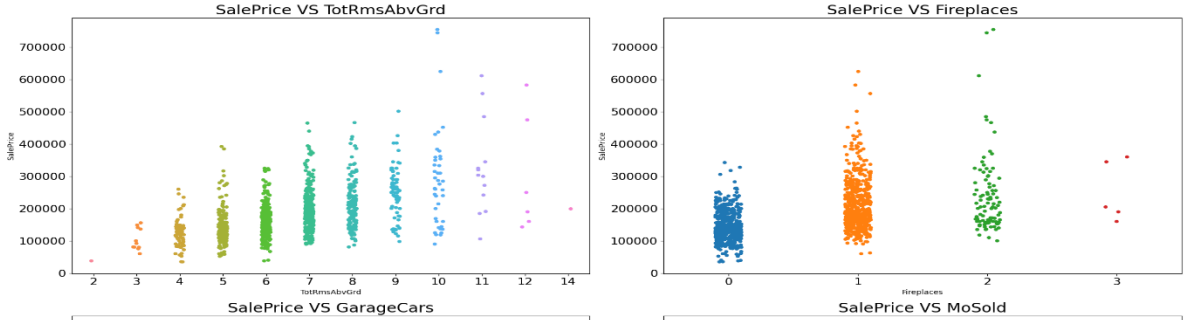
SalePrice VS BsmtHalfBath
SalePrice
BsmtHalfBath


SalePrice VS FullBath
SalePrice
FullBath

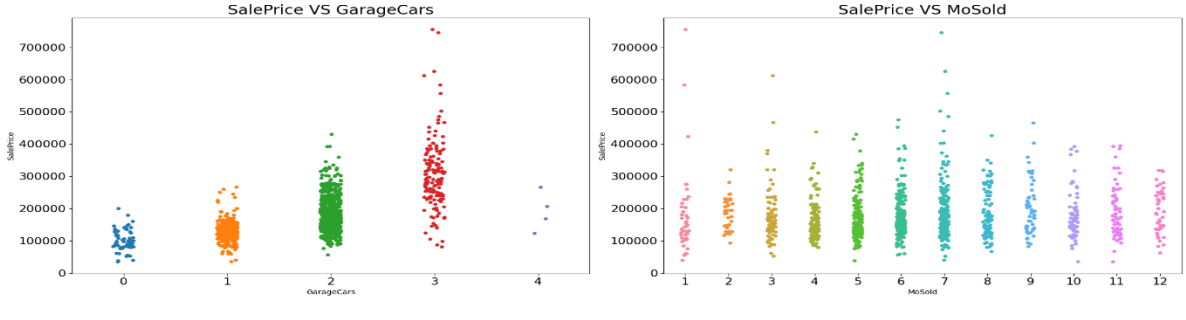
SalePrice VS HalfBath
SalePrice
HalfBath


SalePrice VS BedroomAbvGr
SalePrice
BedroomAbvGr

SalePrice VS KitchenAbvGr
SalePrice
KitchenAbvGr


SalePrice VS TotRmsAbvGrd
SalePrice
TotRmsAbvGrd

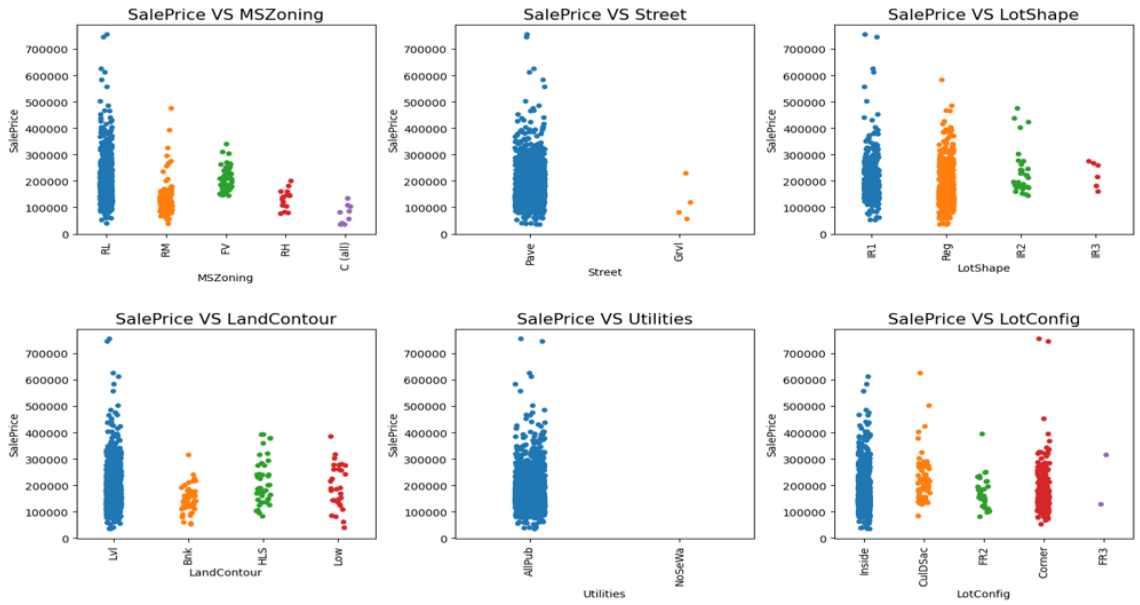
SalePrice VS Fireplaces
SalePrice
Fireplaces


SalePrice VS GarageCars
SalePrice
GarageCars

SalePrice VS MoSold
SalePrice
MoSold

Observations:

- Looking at the above strip plot we can say that more MSSubClass are 20 and 60 and have higher sale prices also.
- We can see there is a good linear relation between OverallQual and SalePrice, that is as quality increases the price of the house also gets higher.
- Most of the housing data shows basement full bathrooms as 0 and 1, and it seems like the number of basement bathrooms is not affecting our sales price.
- Most of the houses are not having basement half bathrooms.
- There is some relation is observed between full bathrooms above grade and salePrice. A large count of houses is with 1 - 2 full bathrooms above grade.
- Many houses are with zero and 1 half-bathrooms, and very few with 2 half_bathrooms.
- A Large number of houses are having 2 to 4 bedrooms and have higher prices. And very few houses are with more than 5 bedrooms which are having a lower price.
- Most of the houses are having single kitchens and some houses with two kitchens. The sale price is higher in the case of houses with a single kitchen.
- We can observe some linear relation between Total rooms above grade and Sale Prices as the number of rooms increases the price also goes up.
- Some houses are not having fireplaces and some are with 1 to 2 fireplaces, very few houses are haveing 3 fireplaces.
- Sales price of the house increases with the Size of garage in car capacity. But as the size of the garage increases beyond 3 the price comes down.
- We can see that the MoSold column is having the data from every column and it may not have any significant impact on our target variable.



Now here are some examples of strip plots of all the categorical features vs SalePrice:

Note: Here I have represented some part of my EDA, for more details of EDA please go through GitHub

Before building the regression model I removed outliers using Zscore and skewness. And after encoding the objective data type columns our data got ready for model development.

Model/s Development and Evaluation

For this project, we need to predict the prices of houses, which means our target column is continuous so this is a regression problem. I have used various regression algorithms and tested them for prediction. By doing various evaluations I have selected '**XGBoost Regressor model**' as the best suitable algorithm for our final model as it is giving good r2-score and the least difference in r2-score and CV-score among all the algorithms used, other algorithms are also giving me better accuracy but some are over-fitting and some are with an under-fitting problem which may be because of less amount of data
For getting good performance as well as accuracy and to check my model from over-fitting and under-fitting I have used KFold cross-validation.

I have used the following algorithms and evaluated them

- LinearRegression
- DecisionTree Regressor
- RandomForest Regressor
- XGBRegressor
- ExtraTreeRegressor

- LightGBM

Key Metrics for success in solving the problems under consideration

I have used the following metrics for evaluation:

- I have used mean absolute error which gives the magnitude of difference between the prediction of observation and the true value of that observation.
- I have used root mean square deviation as one of the most commonly used measures for evaluating the quality of predictions.
- I have used the r2 score which tells us how accurate our model is.

Hyperparameter Tuning

I have done hyperparameter tuning for 'XGBoost Regressor model' for the parameters like 'learning_rate' , ‘max_depth’, 'min_child_weight', 'subsample', 'colsample_bytree', 'n_estimators' and 'objective' using GridSearchCV.

```
1 gsearch.best_params_
```

```
{'colsample_bytree': 0.3,
 'learning_rate': 0.1,
 'max_depth': 3,
 'min_child_weight': 0.5,
 'n_estimators': 500,
 'objective': 'reg:squarederror',
 'subsample': 0.7}
```

- These are the best parameter values.

After running the code for the above-mentioned parameters, I got the values that are indicated in the above figure as the best parametric values for our final model.

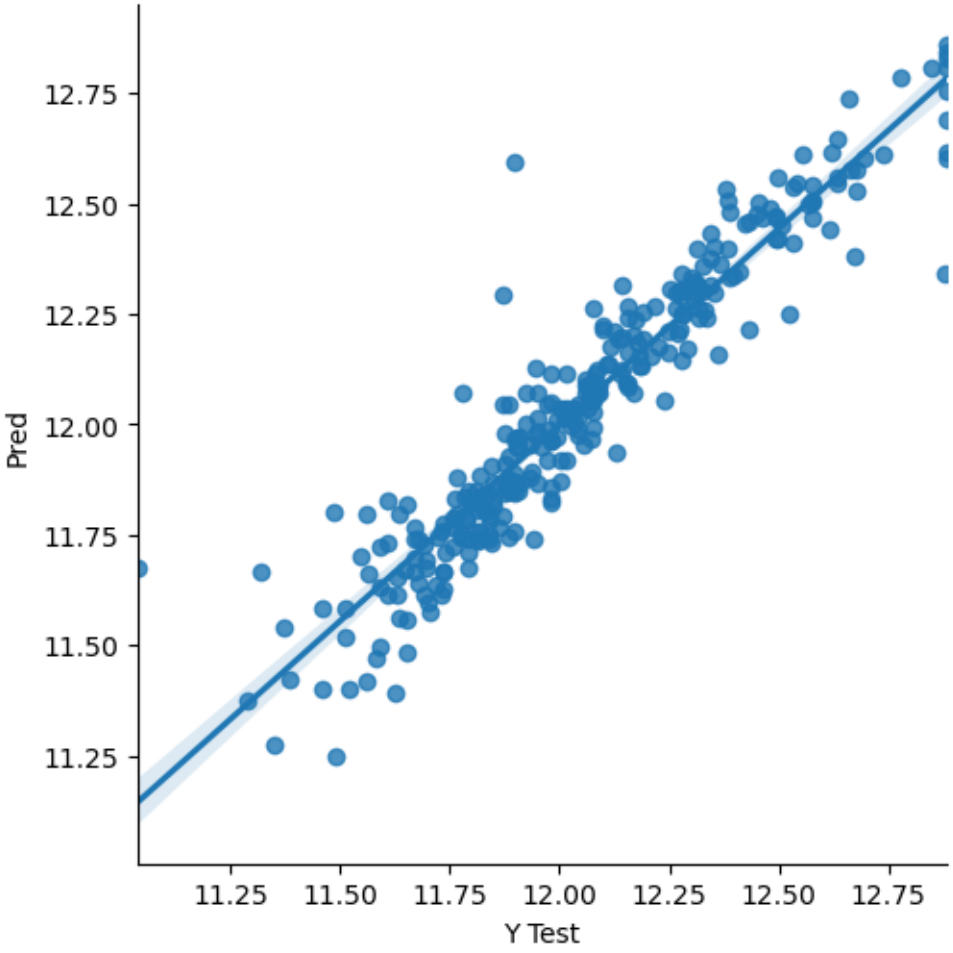
Final Model

```python
#lets train and test our final model with best parameters

model = XGBRegressor(
        objective = 'reg:squarederror',
        colsample_bytree = 0.3,
        learning_rate = 0.1,
        max_depth = 3,
        min_child_weight = 0.5,
        n_estimators = 500,
        subsample = 0.7)

model.fit(x_train,y_train)
pred = model.predict(x_test)

r2score = r2_score(y_test,pred)*100

#evaluation
mse = mean_squared_error(y_test,pred)
rmse = np.sqrt(mse)
mae = mean_absolute_error(y_test,pred)
print("MAE :", mae)
print("RMSE :", rmse)
print('------------------------------')

# r2 score

print(f" \nr2 Score:", r2score,"%")
```

```
MAE : 0.07631909015083167
RMSE : 0.11446485569565193
------------------------------

r2 Score: 88.99648754522569 %
```

Great! We have got improved our accuracy from 85.90% to 88.99% after Hyperparameter tuning

Let’s see the final Actual Vs Predicted sample



By looking at the above graph we can say our model is performing better now. So I will use this model for predicting the house prices using test data set.

CONCLUSION

- Column "OverallQual" is having highest (79%) correlation with target variable followed by the columns: 'GrLivArea'(71%), 'GarageCars'(63%) & 'GarageArea' (62%). So, we can say these features are directly proportional to the Target variable.
- column 'GarageArea' is highly correlated (88%) with column 'GarageCars'.
- column 'GarageAge' is highly correlated (83%) with column 'Building_age'.
- column 'TotRmsAbvGrd' is higly correlated (83%) with column 'GrLivArea'.
- The difference between accuracy score and cross-validation score of XGBoost Regressor model is very less compared to other models.
- So, we can conclude that the **'XGBoost Regressor model'** is our best-fitting model.
- After tunning the model, we got almost 90% of accuracy.

Key Findings and Conclusions of the Study

1. Most of the houses are belong to the Residential Low-Density zone. and many houses from this zone are having higher prices than other zones.
2. We can observe almost all houses are having paved streets and very few are having gravel streets.
3. Large numbers of houses are having a General shape of property slightly irregular or regular. very few are having an irregular shape.
4. Looking at the SalePrice vs LandContour plot we can say that most of the houses are nearly Flat/Leveled.
5. Almost all houses are with all kinds of utilities.
6. More number of lots are inside or at corners.
7. We can see most of the houses are having gentle slopes, and houses with severe slopes are having slightly lower prices.
8. Houses which are located in Northridge are having more prices compared to other locations.
9. Looking at the plot for SalePrice vs Condition1 we can see that most houses bear normal conditions.
10. Looking at the plot for SalePrice vs Condition2 we can see that most houses have normal conditions and very few with other conditions.
11. Most houses are Single-family Detached and are having higher sale prices than other categories.
12. Looking at the plot of SalePrice vs HouseStyle we can see that the houses which are having a style of dwelling 1-story and 2-story are had higher prices than other types.
13. Many houses are having roof styles with gable and hip. and a very less number of houses are having shed.
14. Many houses are having roof material as standard (Composite) Shingle and houses with roof material as Standard (Composite) Shingle and Wood Shingles are having higher prices.
15. Many houses are having Vinyl Siding as 1st and 2nd covering on the house and are also having higher prices; houses with hardboard and cement shielding are also having higher prices.
16. We can observe that the houses are with four Masonry veneer types that are, Brick Common, Brick Face, Cinder Block, Stone. Houses with Brick Common are having a lower price.

17. The prices of houses are higher when materials used for the exterior are good or excellent.
18. It is observed that the present conditions of exterior material are mostly average/typical and good and prices for the same are higher.
19. Many houses are having cinder blocks and Poured Concrete foundations and very fewer houses are having wood foundations, houses with Poured Concrete foundations are having higher prices.
20. Basement quality is mostly average or good and the houses with excellent basement quality are having more prices.
21. Most of the houses are having average/Typical basement conditions and very few houses are with poor basement conditions.
22. It seems like basement exposure is not strongly related to the sale price.
23. Most houses are having Heating type as Gas forced warm air furnace and the Sale price of houses are higher whenever the quality of heating is excellent.
24. Most houses are having central air conditioning and are having more prices than houses that are without air conditioning.
25. Most of the houses with Standard Circuit Breakers & Romex electrical systems and are having higher sale prices as well. Very less number of houses are with Mixed type of electrical systems.
26. Most houses are with good and average kitchen quality, houses are having higher prices when kitchen quality is excellent.
27. Most houses are with typical functionality and a very less number of houses are having severely damaged functionality.
28. In very rare cases fireplace is a prefabricated fireplace in the basement and ben franklin Stove and these houses are having lower prices.
29. In most of the cases garage is attached to the house only. And when the garage is attached prices are higher and it seems like garage finish does not affect much to sale prices. Most garages are of typical/average quality and conditions.
30. Most of the houses are having paved driveways
31. Many houses are having Sale type of Warranty Deed - Conventional and just constructed and sold and are having higher prices.

Limitations of this work and Scope for Future Work

During this project I have faced the problem of the low amount of data; we can improve the model accuracy with a large amount of data. And many columns are with same entries more than 80% of rows this will lead to a reduction in our model performance.

One more issue is there are a large number of missing values present in this data set, so we have to use correctly fill those missing values.

We can still improve our model accuracy with some feature engineering and by doing some extensive hyper parameter tuning.