FLIP ROBO

Housing Price Project

Submitted by:

SNEHA SANTRA

ABSTRACT

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.

INTRODUCTION

- Business Problem Framing

  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.

- Conceptual Background of the Domain Problem

  The project is sub-divided following section. These are:

  1. Loading necessary libraries
  2. Loading Dataset from a CSV file
  3. Summarization of Data to understand Dataset (Descriptive Statistics)
  4. Visualization of Data to understand Dataset (Plots, Graphs etc.)

  5.Processing the data for modeling

  6.skewness and outliers detection for better accuracy

  7.Build the model and select the right model and save it

- Review of Literature

  There are 81 columns including Mssubclass,Msxoining,lotarea,condition,lotshape,sale condition and main feature is sale price.

  There are 38 numerical features,17 discreate features,16 continious features and they are highly corelated with target variables.

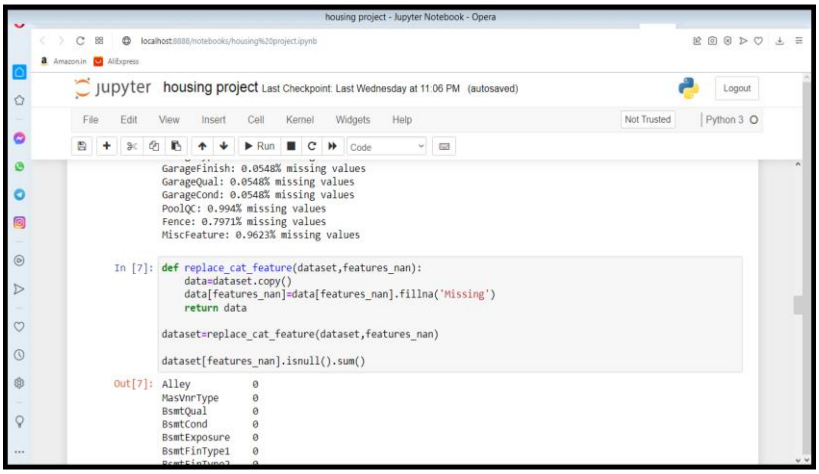
- Motivation for the Problem Undertaken

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?

Analytical Problem Framing

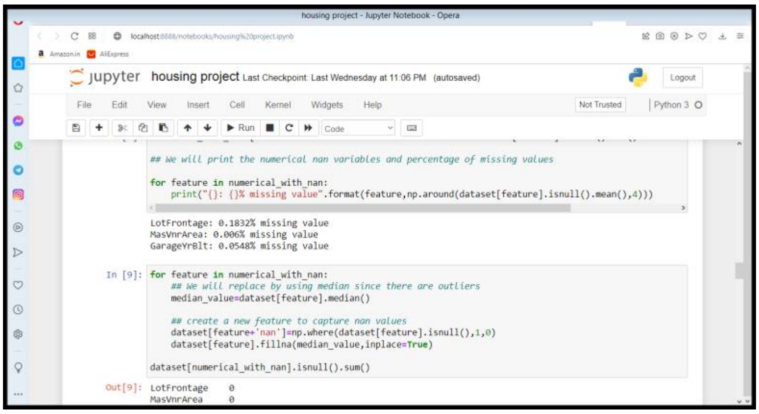
- Mathematical/ Analytical Modeling of the Problem
  Data Processing-There are some missing values in both test and train dataset so will fill it with nan values


housing project - Jupyter Notebook - Opera
localhost:8888/notebooks/housing%20project.ipynb
Amazon.in
AliExpress
jupyter
housing project Last Checkpoint: Last Wednesday at 11:06 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code

```
GarageFinish: 0.0548% missing values
GarageQual: 0.0548% missing values
GarageCond: 0.0548% missing values
PoolQC: 0.994% missing values
Fence: 0.7971% missing values
MiscFeature: 0.9623% missing values
```

```
In [7]: def replace_cat_feature(dataset,features_nan):
            data=dataset.copy()
            data[features_nan]=data[features_nan].fillna('Missing')
            return data

        dataset=replace_cat_feature(dataset,features_nan)

        dataset[features_nan].isnull().sum()
```

```
Out[7]: Alley           0
        MasVnrType      0
        BsmtQual        0
        BsmtCond        0
        BsmtExposure    0
        BsmtFinType1    0
```


housing project - Jupyter Notebook - Opera
localhost:8888/notebooks/housing%20project.ipynb
Amazon.in
AliExpress
jupyter
housing project Last Checkpoint: Last Wednesday at 11:06 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code

```
## We will print the numerical nan variables and percentage of missing values

for feature in numerical_with_nan:
    print("{}: {}% missing value".format(feature,np.around(dataset[feature].isnull().mean(),4)))
```

```
LotFrontage: 0.1832% missing value
MasVnrArea: 0.006% missing value
GarageYrBlt: 0.0548% missing value
```

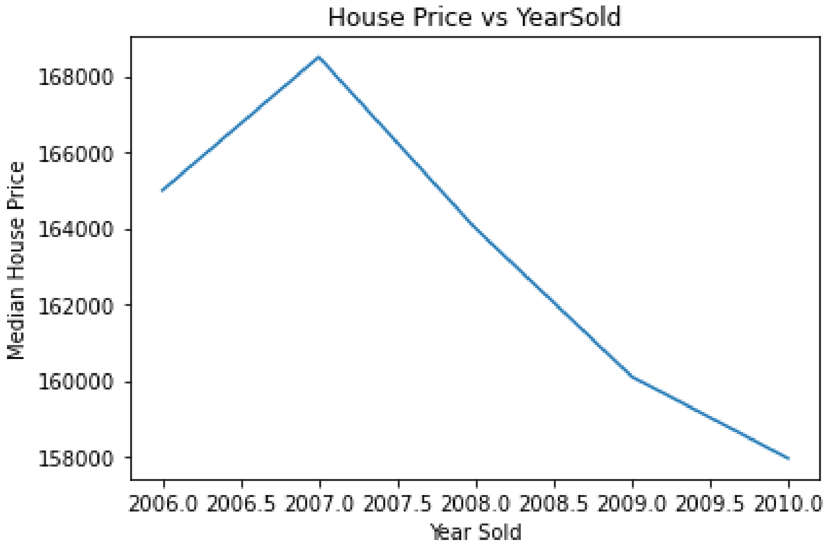
```
In [9]: for feature in numerical_with_nan:
            ## We will replace by using median since there are outliers
            median_value=dataset[feature].median()

            ## create a new feature to capture nan values
            dataset[feature+'nan']=np.where(dataset[feature].isnull(),1,0)
            dataset[feature].fillna(median_value,inplace=True)

        dataset[numerical_with_nan].isnull().sum()
```

```
Out[9]: LotFrontage    0
        MasVnrArea     0
```

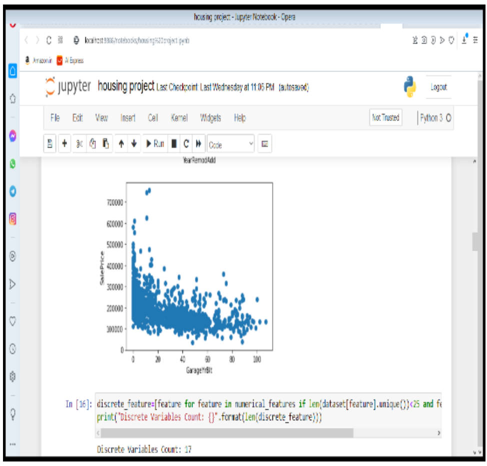
Now there is no missing values

And we drop some columns to scalling the dataset .

EDA-

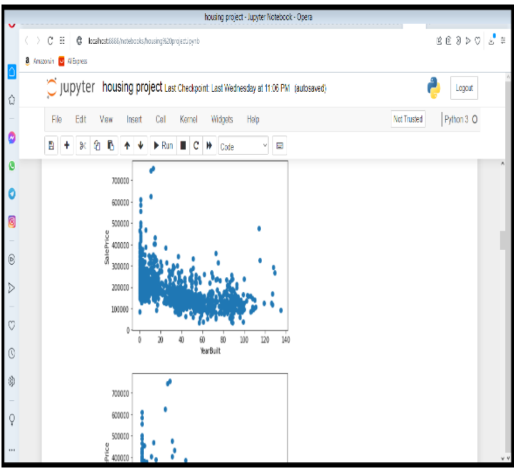


This is the relation between houseprice and yearsold.

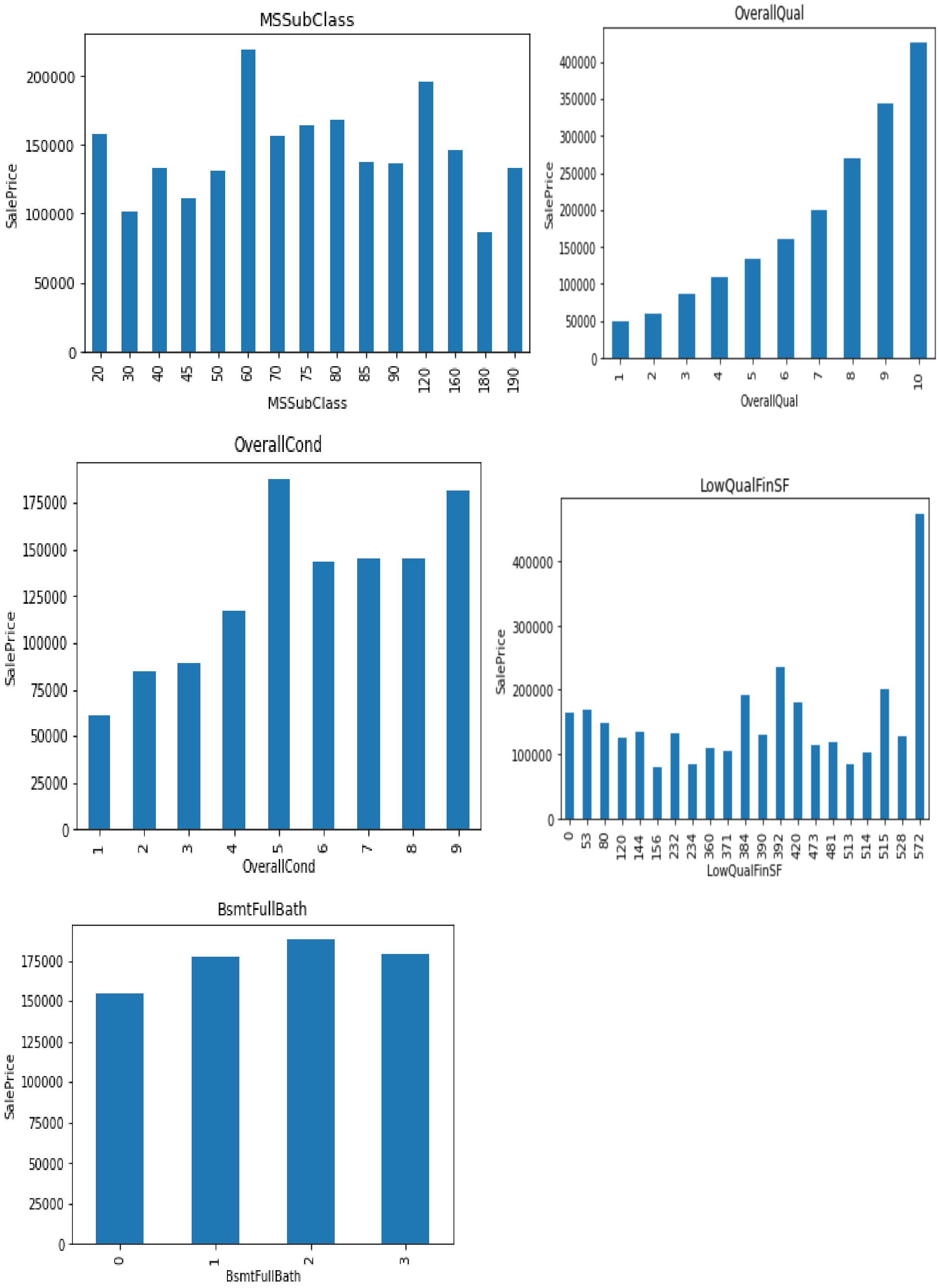


```
In [16]: discrete_feature=[feature for feature in numerical_features if len(dataset[feature].unique())<25 and fe
         print("Discrete Variables Count: {}".format(len(discrete_feature)))
```

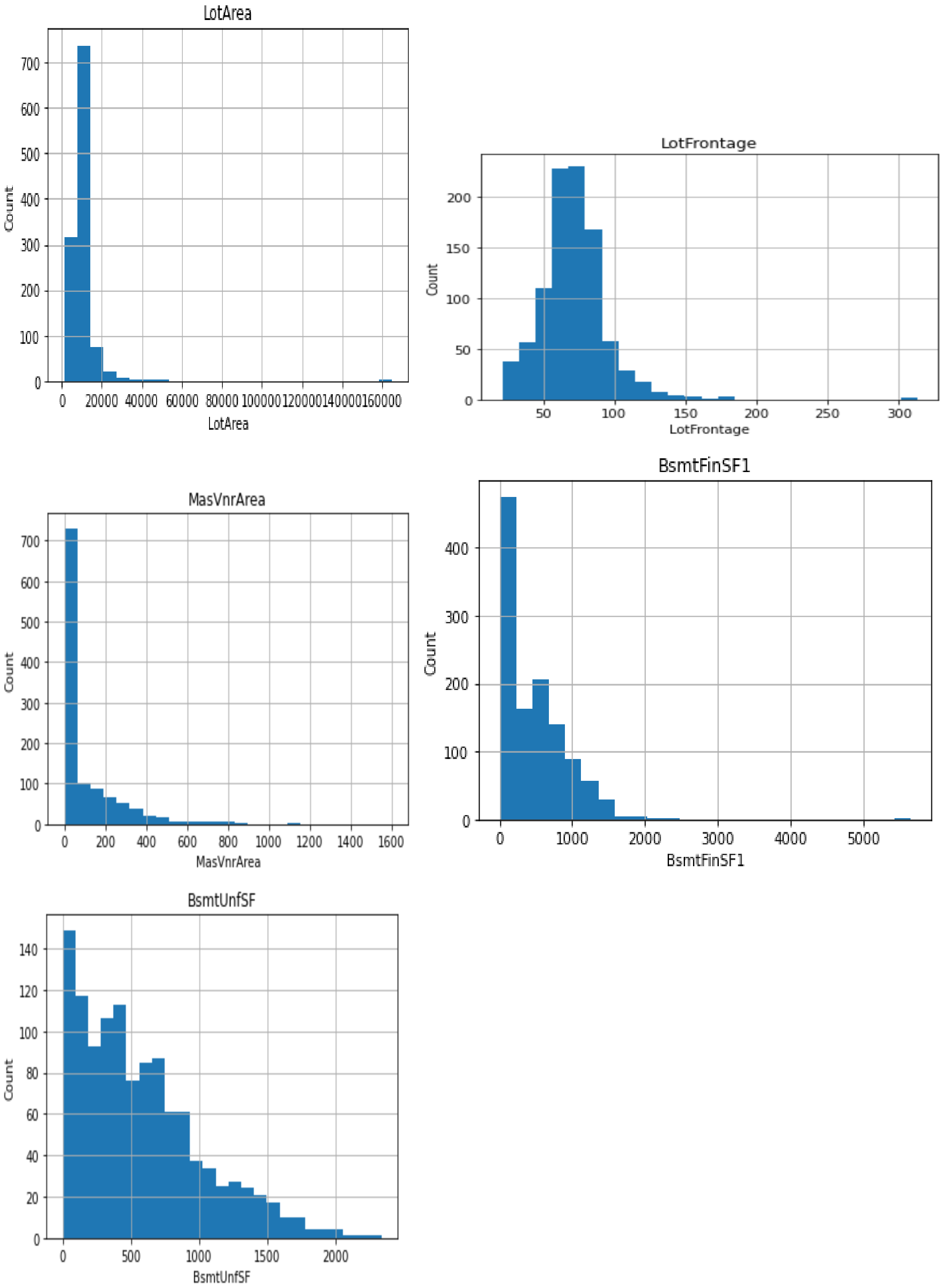
```
Discrete Variables Count: 17
```



This is the relation between saleprice vs yearbuild and garageYrBlt



This is the relation between salePrice and discrete features.



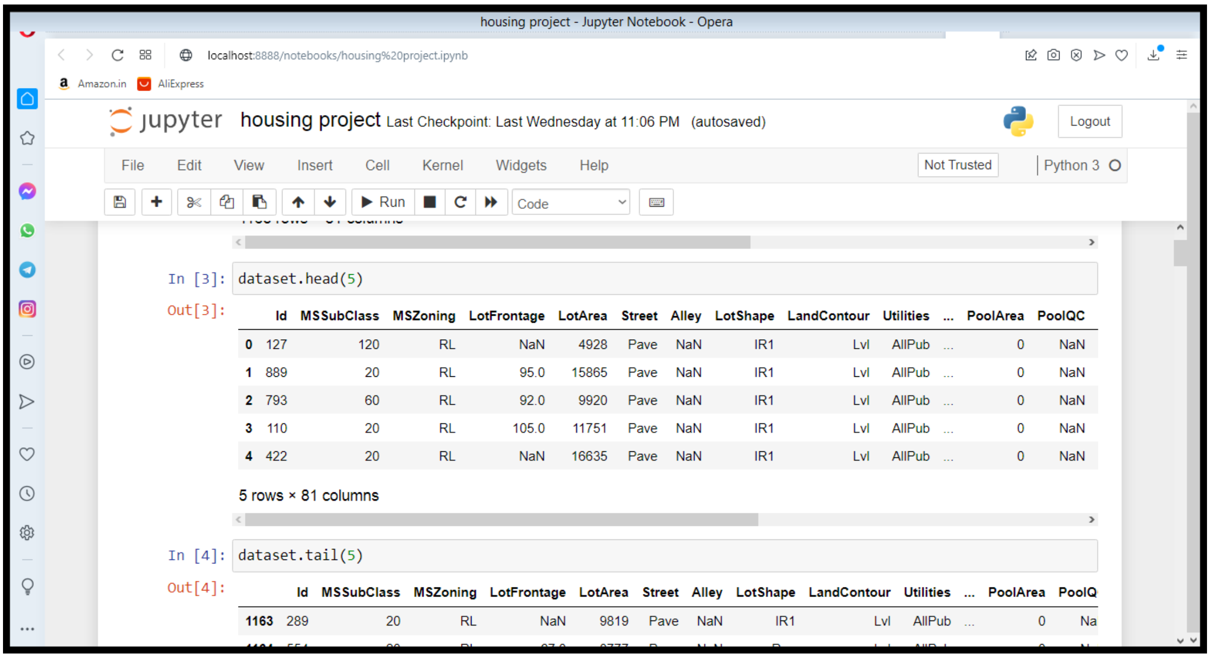
This is the plot/graph between count and continious features.

- ## Data Sources and their formats

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.

This data consist of text and train data which is scrapped and There are 81 columns and 1460 entries. For the classification problem under

consideration we have used some object and some float and some int columns for input variable and our output column is **sale price and sale condition** that is also object type.


housing project - Jupyter Notebook - Opera
localhost:8888/notebooks/housing%20project.ipynb
Amazon.in
AliExpress
jupyter
housing project Last Checkpoint: Last Wednesday at 11:06 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code

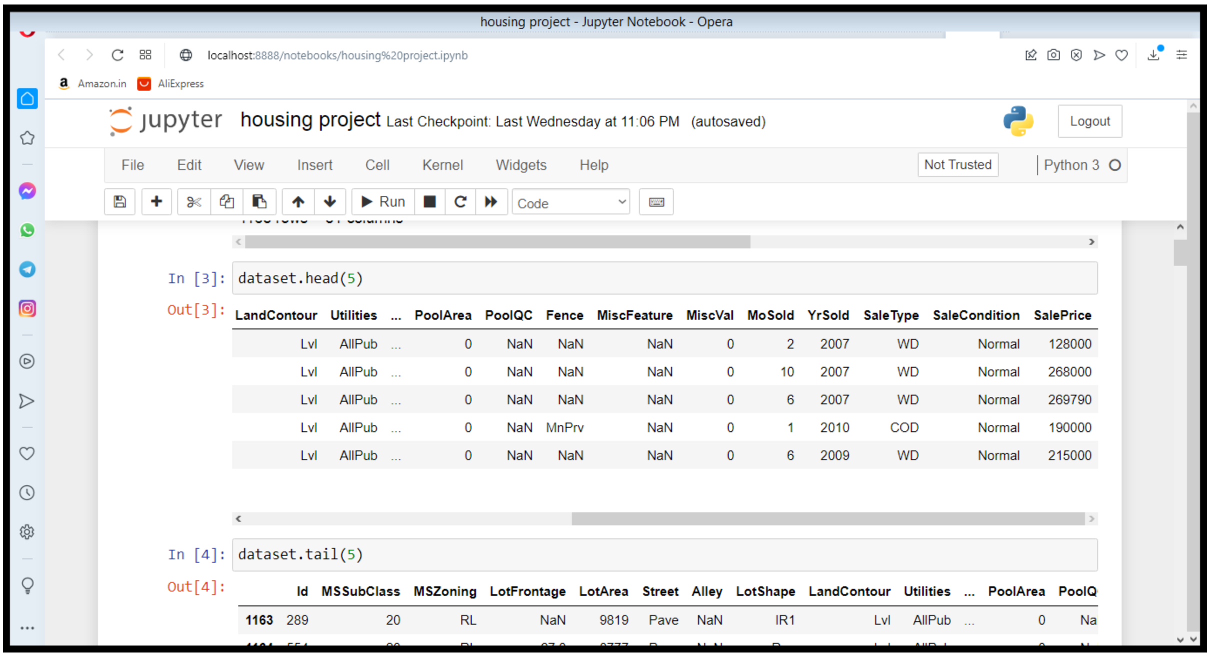
```
In [3]: dataset.head(5)
Out[3]:
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC
0	127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN
1	889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN
2	793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN
3	110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN
4	422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN

5 rows × 81 columns

```
In [4]: dataset.tail(5)
Out[4]:
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQ
1163	289	20	RL	NaN	9819	Pave	NaN	IR1	Lvl	AllPub	...	0	Na


housing project - Jupyter Notebook - Opera
localhost:8888/notebooks/housing%20project.ipynb
Amazon.in
AliExpress
jupyter
housing project Last Checkpoint: Last Wednesday at 11:06 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code

```
In [3]: dataset.head(5)
Out[3]:
```

LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2007	WD	Normal	128000
Lvl	AllPub	...	0	NaN	NaN	NaN	0	10	2007	WD	Normal	268000
Lvl	AllPub	...	0	NaN	NaN	NaN	0	6	2007	WD	Normal	269790
Lvl	AllPub	...	0	NaN	MnPrv	NaN	0	1	2010	COD	Normal	190000
Lvl	AllPub	...	0	NaN	NaN	NaN	0	6	2009	WD	Normal	215000

```
In [4]: dataset.tail(5)
Out[4]:
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQ
1163	289	20	RL	NaN	9819	Pave	NaN	IR1	Lvl	AllPub	...	0	Na

- Data Preprocessing Done

What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

- ## Data Inputs- Logic- Output Relationships

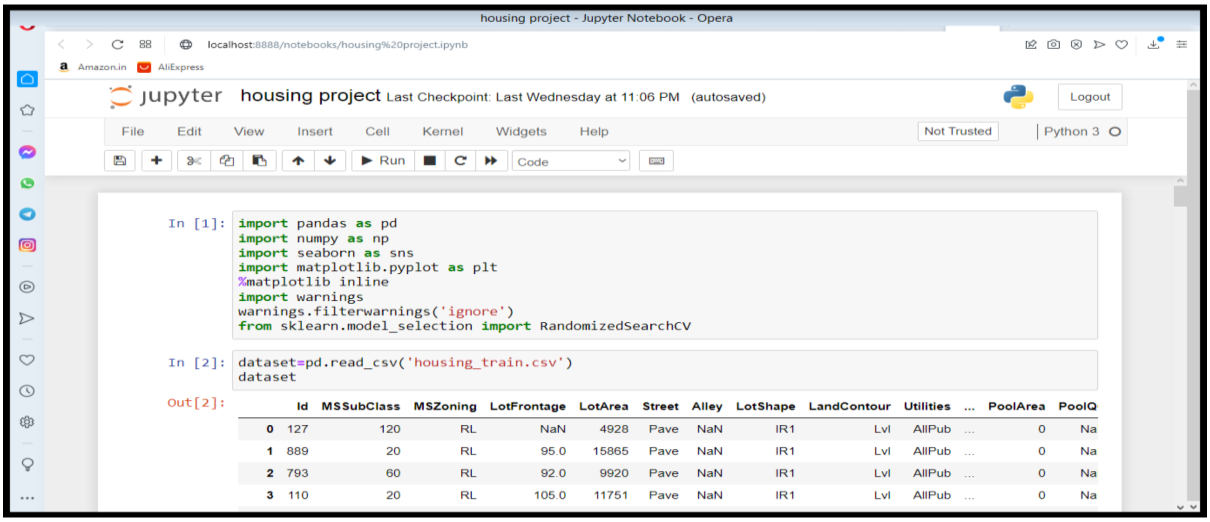
Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

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

Here, you can describe any presumptions taken by you.

- ## Hardware and Software Requirements and Tools Used

Here we use lots of liaberies like pandas,numpy,matplot,seaborn, and we use python language for the coding purpose and import some other metrics liaberies also for model building like sklearn metrics ,classification report,accuracy report etc.


housing project - Jupyter Notebook - Opera
localhost:8888/notebooks/housing%20project.ipynb
Amazon.in
AliExpress
jupyter
housing project Last Checkpoint: Last Wednesday at 11:06 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code

```
In [1]: import pandas as pd
        import numpy as np
        import seaborn as sns
        import matplotlib.pyplot as plt
        %matplotlib inline
        import warnings
        warnings.filterwarnings('ignore')
        from sklearn.model_selection import RandomizedSearchCV
```

```
In [2]: dataset=pd.read_csv('housing_train.csv')
        dataset
```

```
Out[2]:
      Id  MSSubClass  MSZoning  LotFrontage  LotArea  Street  Alley  LotShape  LandContour  Utilities  ...  PoolArea  PoolQ
   0  127        120        RL          NaN     4928    Pave    NaN       IR1          Lvl     AllPub  ...         0     Na
   1  889         20        RL         95.0    15865    Pave    NaN       IR1          Lvl     AllPub  ...         0     Na
   2  793         60        RL         92.0     9920    Pave    NaN       IR1          Lvl     AllPub  ...         0     Na
   3  110         20        RL        105.0    11751    Pave    NaN       IR1          Lvl     AllPub  ...         0     Na
```

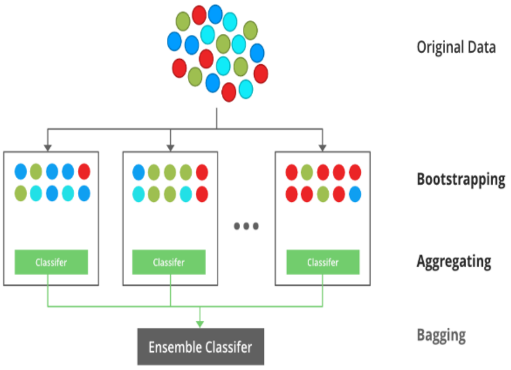
Model/s Development and Evaluation

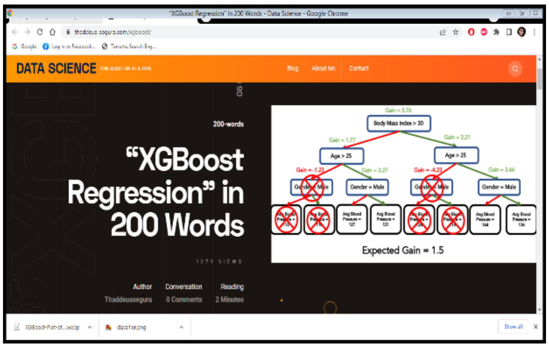
- Identification of possible problem-solving approaches (methods)

  We use xgboost classifier and regressor for the model building as it is lassoregression problem and finally we save the classifier model for gd accuracy.

- Testing of Identified Approaches (Algorithms)

  We use xgboost classifier and regressor




"XGBoost Regression" in 200 Words - Data Science - Google Chrome

thaddeus segura.com/xgboost/

DATA SCIENCE ONE QUESTION AT A TIME

200-words

"XGBoost Regression" in 200 Words

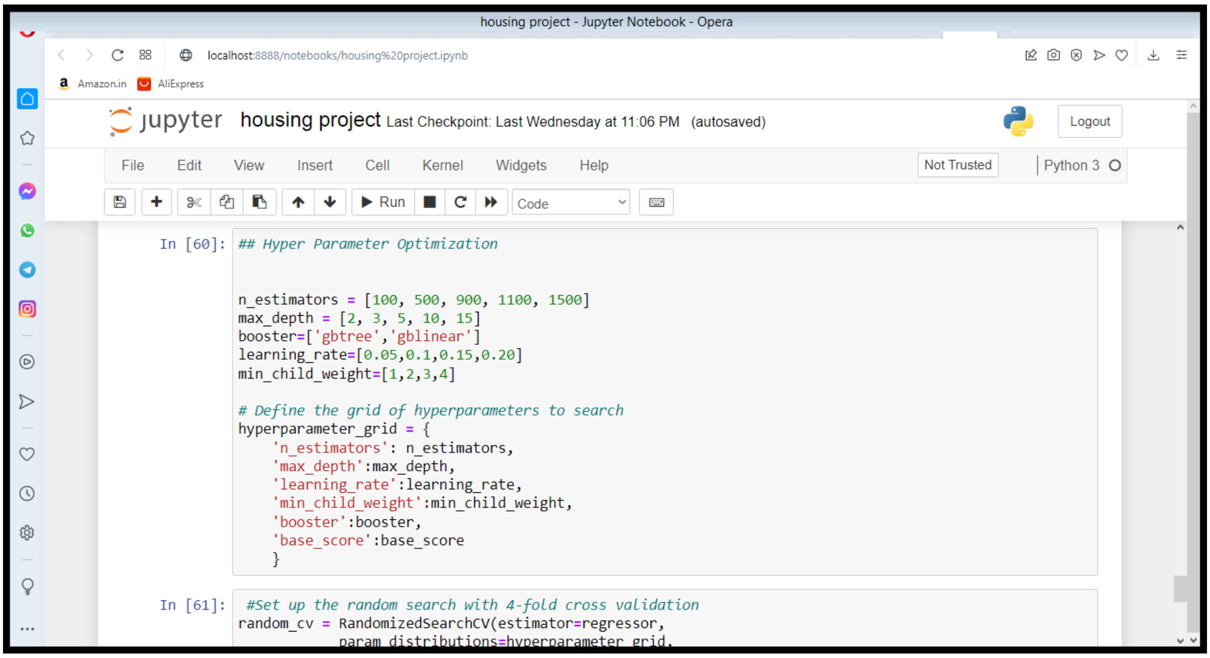
1271 VIEWS

Author Thaddeussegura

Conversation 0 Comments

Reading 2 Minutes

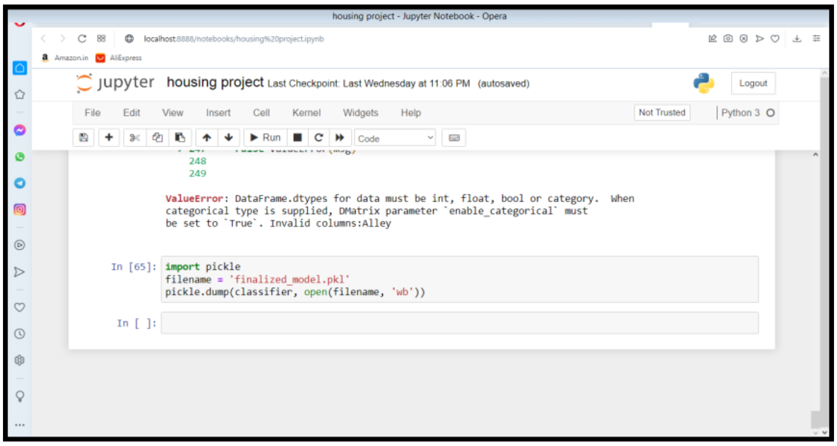
- Run and Evaluate selected models


housing project - Jupyter Notebook - Opera
localhost:8888/notebooks/housing%20project.ipynb
Amazon.in
AliExpress
jupyter
housing project Last Checkpoint: Last Wednesday at 11:06 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code

```
In [60]: ## Hyper Parameter Optimization


         n_estimators = [100, 500, 900, 1100, 1500]
         max_depth = [2, 3, 5, 10, 15]
         booster=['gbtree','gblinear']
         learning_rate=[0.05,0.1,0.15,0.20]
         min_child_weight=[1,2,3,4]

         # Define the grid of hyperparameters to search
         hyperparameter_grid = {
             'n_estimators': n_estimators,
             'max_depth':max_depth,
             'learning_rate':learning_rate,
             'min_child_weight':min_child_weight,
             'booster':booster,
             'base_score':base_score
             }

In [61]:  #Set up the random search with 4-fold cross validation
         random_cv = RandomizedSearchCV(estimator=regressor,
                     param_distributions=hyperparameter_grid,
```

This is hyper parameter optimization and 4-fold cross validation for set up the random search and we choose the best parameter

- Key Metrics for success in solving problem under consideration
  We choose xgboost classifier as the accuracy of the hyper parameter accuracy of classifier is best .


housing project - Jupyter Notebook - Opera
localhost:8888/notebooks/housing%20project.ipynb
Amazon.in
AliExpress
jupyter
housing project Last Checkpoint: Last Wednesday at 11:06 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code

```
    248
    249

ValueError: DataFrame.dtypes for data must be int, float, bool or category.  When
categorical type is supplied, DMatrix parameter `enable_categorical` must
be set to `True`. Invalid columns:Alley

In [65]: import pickle
         filename = 'finalized_model.pkl'
         pickle.dump(classifier, open(filename, 'wb'))

In [ ]:
```

CONCLUSION

- Key Findings and Conclusions of the Study

After the Final Submission of test data, my accuracy score was 90%

Feature engineering helped me increase my accuracy.

Amazingly xgboost classifier worked better than all other Ensemble models.

▪ Learning Outcomes of the Study in respect of Data Science

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.