HOUSING PROJECT

Submitted by:

RUCHITA PARMAR

ACKNOWLEDGMENT

Particularly this Housing Project is for predicting the selling price of a home with in the specific time frame of _ to _. Here we were focusing on data science processing and to find strong guide. Without examining the data, pattern detection will be imperfect and without pattern detection, you cannot draw any conclusions about your data. In this we were traditionally considered few factors like exterior appearance, square footage, number of bedrooms and bathrooms, and number of-floors to be the best explanatory variables.

INTRODUCTION

- Business Problem Framing

  Available dataset is simple and the project aim is for predicting housing price where we need to develop a regressive model based on exploratory available variables.

- Conceptual Background of the Domain Problem

  Since the data is only observational in nature, there is no causal inference can be made against the relationship between the explanatory variables and the response variables. These data constitute a census for the time period and city observed, and thus the scope of inference is limited. Finally, this data may not be important predictors of a final selling price.

- Review of Literature

  There are several parameters in this model that predict a positive or negative influence on the sales price of a house. Not all ideal assumptions hold when using multiple-linear regression because the end goal is to retrieve the best model.

- Motivation for the Problem Undertaken

  We still need to predict the trend of housing prices. Is housing price prediction suitable for linear regression? Is it suitable for combined methods? Or some other method ? Hope we can make our own exploration and discovery.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  The price function which is described as the sum of the individual objectives (criteria), and the goals are the prices of comparable properties. The model integrates with the inductive and deductive approach overcomes many of the assumptions of the best known statistical approaches. The evaluation of the proposed model is performed by comparing the results obtained by the application, to the same case study, of a multiple linear regression model and a nonlinear regression method based on penalised spline smoothing model. The comparison shows, first of all, the best interpretation capabilities of the proposed model.

- Data Sources and their formats

  "Data collection" is the initial step before starting to analyse the patterns or useful information in data. The data which is to be analysed must be collected from different valid sources. Here we were using different data sources as mentioned below:-

  Survey method: - The survey method can be obtained in both online and offline mode like through website forms and email. Then that survey answers are stored for analysing data. Examples are online surveys or surveys through social media polls.

  Observation method:- The observation method is a method of data collection in which the researcher keenly observes the behaviour and practices of the target audience using some data collecting tool and stores the observed data in the form of text, audio, video, or any raw formats. In

this method, the data is collected directly by posting a few questions on the participants.

Experimental method :- Experimental method is the process of collecting data through performing experiments, research, and investigation. The most frequently used experiment methods are CRD, RBD, LSD, FD.

CRD- Completely Randomized design is a simple experimental design used in data analytics which is based on randomization and replication. It is mostly used for comparing the experiments.
RBD- Randomized Block Design is an experimental design in which the experiment is divided into small units called blocks. Random experiments are performed on each of the blocks and results are drawn using a technique known as analysis of variance (ANOVA). RBD was originated from the agriculture sector.
LSD – Latin Square Design is an experimental design that is similar to CRD and RBD blocks but contains rows and columns. It is an arrangement of NxN squares with an equal amount of rows and columns which contain letters that occurs only once in a row. Hence the differences can be easily found with fewer errors in the experiment. Sudoku puzzle is an example of a Latin square design.
FD- Factorial design is an experimental design where each experiment has two factors each with possible values and on performing trail other combinational factors are derived.

Internal source:- These types of data can easily be found within the organization such as market record, a sales record, transactions, customer data, accounting resources, etc. The cost and time consumption is less in obtaining internal sources.

External source:- The data which can't be found at internal organizations and can be gained through external third party resources is external source data. The cost and time consumption is more because this contains a huge amount of data. Examples of external sources are Government publications, news publications, Registrar General of India, planning commission, international labour bureau, syndicate services, and other non-governmental publications.

Other sources:- Satellites data: Satellites collect a lot of images and data in terabytes on daily basis through surveillance cameras which can be used to collect useful information.

- Data Pre-processing Done

Handling Missing data: To handle missing data in the datasets. If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset

By deleting the particular row: The first way is used to commonly deal with null values. In this way, we just delete the specific row or column which consists of null values. But this way is not so efficient and removing data may lead to loss of information which will not give the accurate output.

By calculating the mean: In this way, we will calculate the mean of that column or row which contains any missing value and will put it on the place of missing value. This strategy is useful for the features which have numeric data such as age, salary, year, etc. Here, we will use this approach.

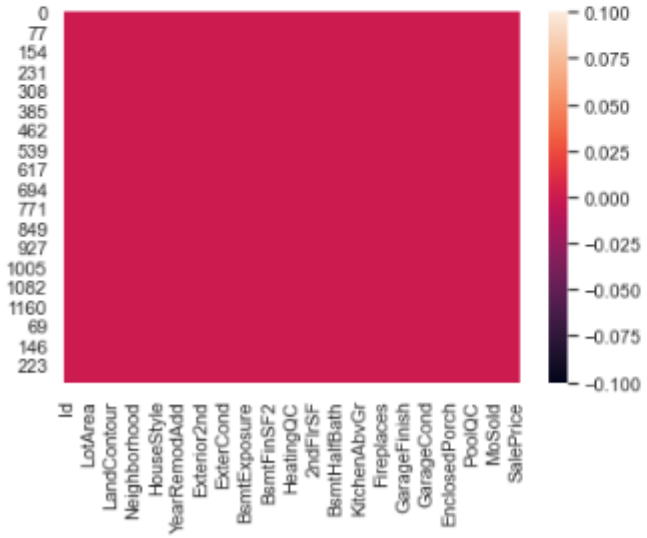
remove/replace the null values

```
1 for column in numeric_data.columns:
2     train_data[column].fillna(train_data[column].mean(), inplace=True)
```

```
1 for column in cat_data.columns:
2     train_data[column].fillna(train_data[column].mode()[0], inplace=True)
```

```
1 sns.heatmap(train_data.isnull())
```

```
<AxesSubplot:>
```



Encoding Categorical data: Categorical data is data which has some categories such as, in our dataset Since machine learning model completely works on mathematics and numbers, but if our dataset would have a categorical variable, then it may create trouble while building the model. So it is necessary to encode these categorical variables into numbers

```
from sklearn import preprocessing
from sklearn.preprocessing import LabelEncoder
le=preprocessing.LabelEncoder()
```

```
#apply the encoding on catagorical columns
for i in cat_data:
    train_data[i]=le.fit_transform(train_data[i].astype("str"))
```

- Data Inputs- Logic- Output Relationships

  The basic idea is this: formal concept lattices offer powerful, well-studied, analytical techniques for classifying, visualising and analysing binary relations, revealing implicit hierarchical structure and/or natural clustering and dependencies between the objects of the relation. Since the set of axioms in any given input/output logic is just a binary relation between formulae, it ought to be possible to apply results from FCA to the study of forms of conditionality that are not naturally assimilated to the model based on inference relations and/or conditionals.

- Hardware and Software Requirements and Tools Used

  Here I used Python and Jupyter notebooks for the completion of the Housing project. As Jupyter notebooks were easy to follow and show your working steps.

  Libraries: These are frameworks in python to handle commonly required tasks. I explored to familiarise themselves with these libraries:

  Pandas — For handling structured data which provide packages fast, flexible, and expressive data structures designed to make working with "relational" or "labelled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python.

  Scikit Learn — For machine learning

  NumPy — For linear algebra and mathematics

  Seaborn — For data visualization

  matplotlib:- It is a comprehensive library for creating static, animated, and interactive visualizations in Python.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

Problem Framing: This is the selection of the type of problem, e.g. regression or classification, and perhaps the structure and types of the inputs and outputs for the problem.

Data Understanding: Data understanding means having an intimate grasp of both the distributions of variables and the relationships between variables.

Data Cleaning: Statistical methods are used for data cleaning; for example:

Outlier detection- Methods for identifying observations that are far from the expected value in a distribution.
Imputation - Methods for repairing or filling in corrupt or missing values in observations.

Data Selection: The process of reducing the scope of data to those elements that are most useful for making predictions is called data selection.

Data Preparation: Data can often not be used directly for modelling. Some transformation is often required in order to change the shape or structure of the data to make it more suitable for the chosen framing of the problem or learning algorithms. Data preparation is performed using statistical methods. Some common examples include:

Scaling - Methods such as standardization and normalization.
Encoding - Methods such as integer encoding and one hot encoding.
Transforms - Methods such as power transforms like the Box-Cox method.

Model Evaluation: The planning of this process of training and evaluating a predictive model is called experimental design. This is a whole subfield of statistical methods.
Resampling Methods - Methods for systematically splitting a dataset into subsets for the purposes of training and evaluating a predictive model.
Model Selection: One among many machine learning algorithms may be appropriate for a given predictive modelling problem. The process of selecting one method as the solution is called model selection.

Model Presentation: Once a final model has been trained, it can be presented to stakeholders prior to being used or deployed to make actual predictions on real data. A part of presenting a final model involves presenting the estimated skill of the model.

Model Predictions: Finally, it will come time to start using a final model to make predictions for new data where we do not know the real outcome.

- Testing of Identified Approaches (Algorithms)

In a very general way, recommender systems are algorithms aimed at suggesting relevant items to users depending on industries. In this, we were using different datasets which are extracted from the provided datasets by using the concept relevant to the dataset. The dataset consists of some features. This dataset is used for forecasting.

Linear Regression:- A relationship is established between independent and dependent variables by fitting them to a line. This line is known as the regression line and represented by a linear equation Y= a *X + b.

KNN (K- Nearest Neighbors) Algorithm: - This algorithm can be applied to both classification and regression problems. Apparently, within the Data Science industry, it's more widely used to solve classification problems. It's a simple algorithm that stores all available cases and classifies any new cases by taking a majority vote of its k neighbors. KNN can be easily understood by comparing it to real life. Things to consider before selecting K Nearest Neighbours Algorithm. KNN is computationally expensive. Variables should be normalized, or else higher range variables can bias the algorithm. Data still needs to be pre-processed.

- Run and Evaluate selected models

Housing project is a regression base problem so we perform the regression model like

R^2

Mean Square Error MSE

Mean absolute Error MAE

1. **RandomForestRegressor:-** A collective of decision trees is called a Random Forest. To classify a new object based on its attributes, each tree is classified, and the tree “votes” for that class. The forest chooses the classification having the most votes.

RandomForestRegressor

```
reg_rf = RandomForestRegressor()
reg_rf.fit(x_train, y_train)
y_pred = reg_rf.predict(x_test)
```

```
print('MAE:', metrics.mean_absolute_error(y_test, y_pred))
print('MSE:', metrics.mean_squared_error(y_test, y_pred))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
print("R squar score",metrics.r2_score(y_test, y_pred))
```

```
MAE: 0.14557817243136642
MSE: 0.04523702600763003
RMSE: 0.21268997627445924
R squar score 0.6639796299148218
```

2. **Linear Regression:-** A relationship is established between independent and dependent variables by fitting them to a line. This line is known as the regression line and represented by a linear equation Y= a *X + b.

LinearRegression

```
lr = LinearRegression()
lr.fit(x_train, y_train)
y_pred = lr.predict(x_test)
print('MAE:', metrics.mean_absolute_error(y_test, y_pred))
print('MSE:', metrics.mean_squared_error(y_test, y_pred))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
# The coefficients
print("Coefficients: \n", lr.coef_)
print("Coefficient of determination: %.2f" % r2_score(y_test, y_pred))

```

```
MAE: 0.14646328395237426
MSE: 0.04577902028784824
RMSE: 0.21396032409736213
```

3. **KNN (K- Nearest Neighbors) Algorithm: -** This algorithm can be applied to both classification and regression problems. Apparently, within the Data Science industry, it's more widely used to solve classification problems. It’s a simple algorithm that stores all available cases and classifies any new cases by taking a majority vote of its k neighbors. KNN can be easily understood by comparing it to real life. Things to consider before selecting K Nearest Neighbours Algorithm. KNN is computationally expensive. Variables should be normalized, or else higher range variables can bias the algorithm. Data still needs to be pre-processed.

KNeighborsRegressor

```
knn = KNeighborsRegressor(n_neighbors=10)
knn.fit(x_train, y_train)
y_pred = knn.predict(x_test)
print('MAE:', metrics.mean_absolute_error(y_test, y_pred))
print('MSE:', metrics.mean_squared_error(y_test, y_pred))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

```
MAE: 0.1849246845793168
MSE: 0.06576809904637564
RMSE: 0.25645291779657264
```

DecisionTreeRegressor

```
dt =DecisionTreeRegressor()
dt.fit(x_train, y_train)
y_pred = dt.predict(x_test)


print('MAE:', metrics.mean_absolute_error(y_test, y_pred))
print('MSE:', metrics.mean_squared_error(y_test, y_pred))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

```
MAE: 0.20587244781635364
MSE: 0.08304118182266543
RMSE: 0.2881686690510705
```

4. **Lasso and Ridge:-** Ridge and lasso regression allow you to regularize ("shrink") coefficients. This means that the estimated coefficients are pushed towards 0, to make them work better on new data-sets ("optimized for prediction"). This allows you to use complex models and avoid over-fitting at the same time

Lasso and Ridge

]:
```
clf = linear_model.Lasso(alpha=0.1)
clf.fit(x_train, y_train)

y_pred = clf.predict(x_test)

print("MAE",mean_absolute_error(y_test, y_pred))
print("MSE",mean_squared_error(y_test, y_pred))
print("Rsqur score",r2_score(y_test, y_pred))
```

```
MAE 0.14919408647742297
MSE 0.049552141506423286
Rsqur score 0.6319269767050323
```

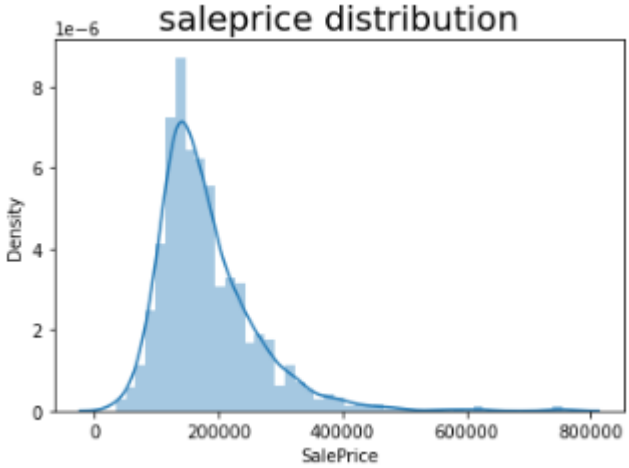
]:
```
rdg = linear_model.Ridge(alpha=0.1)
rdg.fit(x_train, y_train)

y_pred = rdg.predict(x_test)

print("MAE",mean_absolute_error(y_test, y_pred))
print("MSE",mean_squared_error(y_test, y_pred))
print("Rsqur score",r2_score(y_test, y_pred))
```

```
MAE 0.1464283003225948
MSE 0.045745592331827525
Rsqur score 0.6602020021715462
```

- **Visualizations**

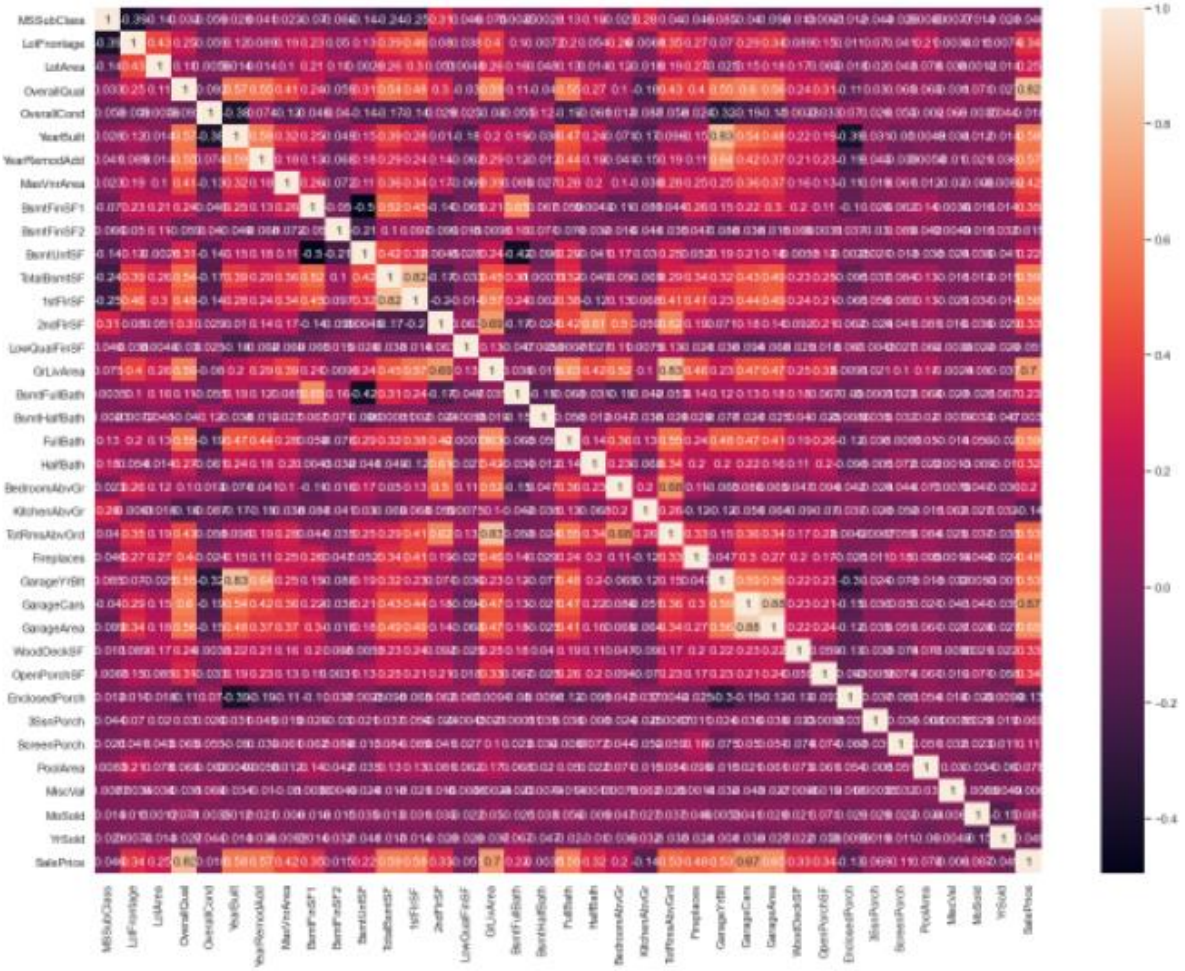
  There is another way called Visualization, to understand the data. With the help of data visualization, we can see how the data looks like and what kind of correlation is held by the attributes of data. It is the fastest way to see if the features correspond to the output.

1. -Price Distribution Graph -



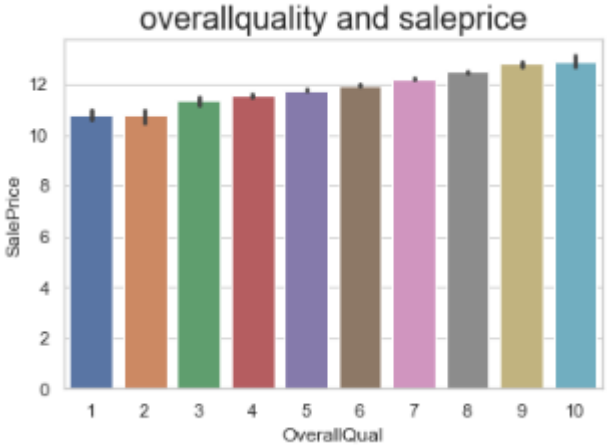
In our target variable skewness present (right skew)

2. Correlation graph - We can easily understand the positive and negative correlation



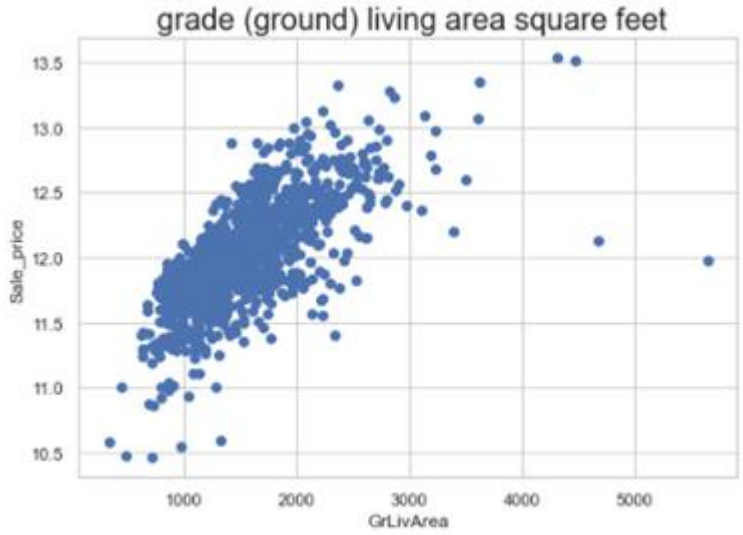
```
SalePrice       1.000000
OverallQual     0.818551
GrLivArea       0.697364
GarageCars      0.671597
GarageArea      0.647206
FullBath        0.593204
TotalBsmtSF     0.592753
1stFlrSF        0.577722
YearBuilt       0.575768
YearRemodAdd    0.570256
TotRmsAbvGrd    0.531997
GarageYrBlt     0.528149
Fireplaces      0.481744
MasVnrArea      0.419153
BsmtFinSF1      0.348917
Name: SalePrice, dtype: float64

----------------------
MSSubClass      -0.045982
YrSold          -0.048638
LowQualFinSF    -0.050755
EnclosedPorch   -0.133181
KitchenAbvGr    -0.141049
Name: SalePrice, dtype: float64
```

3. Overalluality and saleprice



In this graph overallquality is increase than saleprice also increase

4. GrLivArea and saleprice:



We see that increases in living area correspond to increases in price.

- Interpretation of the Results

  A major role played by data in building an accurate, efficient model. Better the data will be better the model. We need to perform data processing followed by data analysis and then visualizing the data. In pre-processing we dealt with Null values, categorical variables and standardize the data.

  Data Visualization is a method to represent the data in various graphical format which makes data to easily understandable, even huge amount of data can easily be understood just by looking the graph and plot. Here we used MatPlotLib library for data visualization some basis plotting techniques are Scatter Plot, Line Plot, Bar Chart.

CONCLUSION

- Key Findings and Conclusions of the Study

  With the help of just a Random Forest Classifier, we were able to predict the house prices in a fairly good. Also, we discussed the Data Analysis and Data Visualization. We performed rescaling, normalizing, binarizing, and standardizing the data. We discovered how we can make raw data suitable for our machine learning algorithm to learn from using data pre-processing techniques. Not all ideal assumptions hold when using multiple-linear regression because the end goal is to retrieve the best model. Variance inflation factors, normality, and diagnostics for influential points are to the wayside in the real world when seeking the best predictive model.

With this, we able to use new analytical techniques in property research. This study is an exploratory attempt to use different algorithms in estimating housing prices, and then compare their results. Our study has shown that advanced machine learning algorithms like SVM, RF and GBM, are promising tools for property researchers to use in housing price predictions. Many conventional estimation methods produce reasonably good estimates of the coefficients that unveil the relationship between output variable and predictor variables. These methods are intended to explain the real-world phenomena and to make predictions, respectively.

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

  List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

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

  What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.