**Text

Description automatically generated with medium confidence**

**Housing Price Prediction Project**

A picture containing text, clipart

Description automatically generated

Submitted by:

Rajashri Sadafule Darveshi

**ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to the sources Medium, Analytics Vidhya, TowardsDataScience, Data Training Projects and youtube channel’s which helped me to for this project. I would like to thank FlipRobo for giving me this opportunity to develop and accomplish this project.

**INTRODUCTION**

**Problem Statement:**

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?

**Problem Understanding:**

* House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house.
* House Price prediction, is important to drive Real Estate efficiency. As earlier, House prices were determined by calculating the acquiring and selling price in a locality.
* Therefore, the House Price prediction model is very essential in filling the information gap and improve Real Estate efficiency.
* The aim is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities.
* By analysing previous market trends and price ranges, and also upcoming developments future prices will be predicted. Cost of property depending on number of attributes considered.
* Now as a data scientist our work is to analyse the dataset and apply our skills towards predicting house price.

**What is Housing Price Prediction?**

* Prediction house prices are expected to help people who plan to buy a house so they can know the price range in the future, then they can plan their finance well. In addition, house price predictions are also beneficial for property investors to know the trend of housing prices in a certain location.

**Importance of Housing Price Prediction:**

* House Price prediction, is important to drive Real Estate efficiency. As earlier, House prices were determined by calculating the acquiring and selling price in a locality. Therefore, the House Price prediction model is very essential in filling the information gap and improve Real Estate efficiency

**Data Sources:**

The training data and testind data for this project are available in csv file.

**About the data:** Details of dataset are as follows

* Number of data points in train data:1168
* Number of features in train data: 81
* Number of data points in test data: 292
* Number of features in test data: 80

We will understand all features are their details of the dataset are as follows:

1. MSSubClass: Identifies the type of dwelling involved in the sale.
2. MSZoning: Identifies the general zoning classification of the sale.
3. LotFrontage: Linear feet of street connected to property
4. LotArea: Lot size in square feet
5. Street: Type of road access to property
6. Alley: Type of alley access to property
7. LotShape: General shape of property
8. LandContour: Flatness of the property
9. Utilities: Type of utilities available
10. LotConfig: Lot configuration
11. LandSlope: Slope of property
12. Neighborhood: Physical locations within Ames city limits
13. Condition2: Proximity to various conditions (if more than one is present)
14. BldgType: Type of dwelling
15. HouseStyle: Style of dwelling
16. OverallQual: Rates the overall material and finish of the house
17. OverallCond: Rates the overall condition of the house
18. YearBuilt: Original construction date
19. YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)
20. RoofStyle: Type of roof
21. RoofMatl: Roof material
22. Exterior1st: Exterior covering on house
23. Exterior2nd: Exterior covering on house (if more than one material)
24. MasVnrType: Masonry veneer type
25. MasVnrArea: Masonry veneer area in square feet
26. ExterQual: Evaluates the quality of the material on the exterior
27. ExterCond: Evaluates the present condition of the material on the exterior
28. Foundation: Type of foundation
29. BsmtQual: Evaluates the height of the basement
30. BsmtCond: Evaluates the general condition of the basement
31. BsmtExposure: Refers to walkout or garden level walls
32. BsmtFinType1: Rating of basement finished area
33. BsmtFinSF1: Type 1 finished square feet
34. BsmtFinType2: Rating of basement finished area (if multiple types)
35. BsmtFinSF2: Type 2 finished square feet
36. BsmtUnfSF: Unfinished square feet of basement area
37. TotalBsmtSF: Total square feet of basement area
38. Heating: Type of heating
39. HeatingQC: Heating quality and condition
40. CentralAir: Central air conditioning
41. Electrical: Electrical system
42. 1stFlrSF: First Floor square feet
43. 2ndFlrSF: Second floor square feet
44. LowQualFinSF: Low quality finished square feet (all floors)
45. GrLivArea: Above grade (ground) living area square feet
46. BsmtFullBath: Basement full bathrooms
47. BsmtHalfBath: Basement half bathrooms
48. FullBath: Full bathrooms above grade
49. HalfBath: Half baths above grade
50. Bedroom: Bedrooms above grade (does NOT include basement bedrooms)
51. Kitchen: Kitchens above grade
52. KitchenQual: Kitchen quality
53. TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
54. Functional: Home functionality (Assume typical unless deductions are warranted)
55. Fireplaces: Number of fireplaces
56. FireplaceQu: Fireplace quality
57. GarageType: Garage location
58. GarageYrBlt: Year garage was built
59. GarageFinish: Interior finish of the garage
60. GarageCars: Size of garage in car capacity
61. GarageArea: Size of garage in square feet
62. GarageQual: Garage quality
63. GarageCond: Garage condition
64. PavedDrive: Paved driveway
65. WoodDeckSF: Wood deck area in square feet
66. OpenPorchSF: Open porch area in square feet
67. EnclosedPorch: Enclosed porch area in square feet
68. 3SsnPorch: Three season porch area in square feet
69. ScreenPorch: Screen porch area in square feet
70. PoolArea: Pool area in square feet
71. PoolQC: Pool quality
72. Fence: Fence quality
73. MiscFeature: Miscellaneous feature not covered in other categories
74. MiscVal: $Value of miscellaneous feature
75. MoSold: Month Sold (MM)
76. YrSold: Year Sold (YYYY)
77. SaleType: Type of sale
78. SaleCondition: Condition of sale

Table

Description automatically generated

Housing Price Prediction dataset having 1168 rows and 81 features.

Where **SalePrice** is the resultant feature

Features names are as follow.

Text, letter

Description automatically generated

**Exploratory Data Analysis:**

Dataset contains Categorical and Numericle type data.

Above details features details we get the datatypes of features. This gives the information about the dataset which includes indexing type, column type, contains null values and memory usage.

|  |  |  |
| --- | --- | --- |
|  |  |  |

The dataset contains the details of the employees who are working in an organization. The dataset contains both dependent and independent variables and also contains both categorical and numerical data. In this dataset "**SalesPrice**"" is our target variable which has continous data. So this is a "**Regression type**" problem in which we need to predict the house price for given independanat features.

We an see number of unique contain present in each feature.

|  |  |
| --- | --- |
|  |  |

#### **Detect the missing values:**

The dataset has missing values we can see with isnull().sum() function and with heatmap graph.

Graphical user interface

Description automatically generated with medium confidence

In dataset ID is just for serial number not giving any information so we will drop ID column. Then 'Alley’, 'MiscFeature’, 'PoolQC' these columns are having more than 80% NA data so we will drop these columns

**Statistical Analysis of dataset:**

**We will use describe() method for calculating some statistical data like percentile, mean and std of the numerical values of the Series or DataFrame. As our dataset having both numeric and object series and also the DataFrame column sets of mixed data types.** Describe methode uses columns contain continuous type of data

## Graphical user interface Description automatically generated

We can observe the following things.

* While checking the info of the datasets, found some columns with more than 80% null values
* Features LotFrontage, MasvnrArea, GarageyrBlt will fill NA with mean values
* Now remaining null catogoricle features values will replace with NA
* While checking for null values I found null values in most of the columns and I have used imputation method to replace those null values (mode for categorical column and mean for numerical columns).

Text

Description automatically generated

Correlation heatmap is graphical representation of correlation matrix representing correlation between different variables. As in this dataset more than 75 features present so with heatmap its difficult to correlate the features.

Table

Description automatically generated with low confidence

**Visualizing the outliers using the lineplot: Bivariate Analysis** we can do analysis with salesprice label

Diagram

Description automatically generated

Diagram

Description automatically generated

**Visualizing the outliers using the boxplot:**

Text, calendar

Description automatically generated

* Box plot for each pair of categorical features that shows the relation with the median sale price for all the sub categories in each categorical feature. And also for continuous numerical variables I have used reg plot to show the relationship between continuous numerical variable and target variable.
* Found that there is a linear relationship between continuous numerical variable and SalesPrice.
* we can observe these features are having outliers 'LotFrontage', 'LotArea','MasVnrArea’, 'BsmtFinSF1', 'BsmtFinType2', 'BsmtFinSF2', 'BsmtUnfSF', 'TotalBsmtSF', '1stFlrSF', 'LowQualFinSF', 'GrLivArea’, 'GarageArea’, 'WoodDeckSF', 'OpenPorchSF', 'EnclosedPorch', '3SsnPorch', 'ScreenPorch', 'PoolArea', 'Fence’, 'MiscVal’, 'SaleType’, 'SaleCondition' we will try to remove outliers with zscore

**Removing outliers by Zscore and IQR Methode**

Chart, scatter chart

Description automatically generated

Dataloss of 4.2% with zscore.

Graphical user interface, text

Description automatically generated

Dataloss of 86% with IQR which is very high.

We removed outliers with dataloss of 4.2% with zscore.which is less than 5% using zscore.

**Distplot before Removing outliers by Zscore:**

**Diagram

Description automatically generated**

Diagram

Description automatically generated

Still some feature are having skewness or outlier present**.**

|  |  |
| --- | --- |
| **Table  Description automatically generated** | **Text  Description automatically generated** |

**Observation**: After removing applying Zscore method data loss is 4.2%. Which is less than 5%. Still some feature are having skewness or outlier present. Now we will do **power transformation technique to treat the skewness in the data yeo-johnson**

**yeo-johnson method:** Removed the skewness using yeo-johnson method.

**Distplot power transformation technique ‘yeo-Johnson’**

Diagram

Description automatically generated

Diagram

Description automatically generated

* The looks normal compare to the old data but still in some features skewness is present. After applying power transformation technique to treat the skewness in the data yeo-Johnson, skewness is decresed but some of columns still having lots of outlier so we will drop them.
* These are some columns MiscVal, PoolArea, ScreenPorch, 3SsnPorch, EnclosedPorch, BsmtFinSF2, Exterior2nd.
* After droping MiscVal, PoolArea, ScreenPorch, 3SsnPorch, EnclosedPorch, BsmtFinSF2, Exterior2nd these columns now our data is cleaned.

**Pearson’s correlation coefficient :**

Pearson’s correlation coefficient to check the correlation between dependent and independent features

|  |  |
| --- | --- |
|  |  |

**Observation**: from corelation we can observe these are features are positively corelated with Sales price ,OverallQual, GrLivArea, GarageCars, GarageArea, TotalBsmtSF, 1stFlrSF, FullBath, TotRmsAbvGrd, YearBuilt, YearRemodAdd, MasVnrArea, Fireplaces, GarageYrBlt,Foundation, BsmtFinSF1, penPorchSF, 2ndFlrSF, LotFrontage, WoodDeckSF, HalfBath, LotArea, GarageCond, CentralAir, Electrical, PavedDrive, SaleCondition,BsmtUnfSF, BsmtFullBath, HouseStyle

And negatively corelated with Sales price these features are LotShape, BsmtExposure, HeatingQC, GarageType, GarageFinish, KitchenQual, BsmtQual, ExterQual And Utilities having all NAN so we will drop Utilities

# Standard scaler

## Scaled the data using standard scalarizaion method to overcome with the issue of data biasness.

## Text Description automatically generated

# Observation of Exploratory Data Analysis:

# Statistical analysis like checking shape, nunique, value counts, info describe etc

# While checking the info of the datasets I found some columns with more than 80% null values, so these columns will create skewness in datasets so I decided to drop those columns.

# Then while looking into the value counts I found some columns with more than 85% zero values this also creates skewness in the model and there are chances of getting model bias so I have dropped those columns with more than 85% zero values.

# While checking for null values I found null values in most of the columns and I have used imputation method to replace those null values (mode for categorical column and mean for numerical columns).

# In Id and Utilities column the unique counts were 1168 and 1 respectively, which means all the entries in Id column are unique and ID is the identity number given for perticular asset and all the entries in Utilities column were same so these two column will not help us in model building. So I decided to drop those columns.

# And all these steps were performed to both train and test datasets separately and simultaneously.

# Model Preparation

## For model preparation we will Separate the features and label variables into x and y

# Graphical user interface Description automatically generated with medium confidence

**Encoding the categorical columns using label encoder**

Graphical user interface, text

Description automatically generated

# Training and Testing Data

Separate data into a training set and a test set . This is a very standard approach in Machine Learning. The random\_state parameter is simply a seed for the algorithm to use (if we didn't specify one, it would create different training and test sets every time we run it) **Find for which state we are getting best accuracy with LinearRegression. Below we can see at 40 random state we are getting best accuracy score 89.7%.**

**Now with 40 random state we done train test split for taining and testing data.**

**Text

Description automatically generated**

* Using Linear Regression we find R2 score and its corresponding random state
* Done the train\_test\_split data with 75 training and 25 testing data

****

**Building Machine Learning Models:**

* Since SalePrice was my target and it was a continuous column so this particular problem

was regression problem.

* Using Linear Regression we find R2 score and its corresponding random state
* Done the train\_test\_split data with 75 training and 25 testing data
* Now we will check accuracy with following Repressor algorithm and finalize one model
* DecisionTreeRegressor()
* Linear Regression
* RandomForestRegressor()
* KNeighborsRegressor()
* AdaBoostRegressor()
* Lasso()
* Ridge()
* ExtraTreesRegressor()
* XGBRegressor()
* GradientBoostingRegressor()

**Fitting the data to various model and checking the accuracy:**

Graphical user interface, text, application

Description automatically generated

**Here with following comparision table of Training score, Test Score, Mean Square Error and Cross Validation Score**

Table

Description automatically generated

**Observation**:

* Following are the result over all 9 algorithms with respect to Training Score,Test Score,Mean Square Error and the Cross Validation Score
* DecisionTree having 100% accuracy which is showing over fitting and maximum difference in test score and CV score. XGBRegressor is also somewhat showing overfitting.
* Linear Regression, Ridge and Lasso are underfitting model as training score is less than testing score.
* We can see GradientBoosting, Extra Tree and RandomForest regressor having comparative less Mean Square Error.
* GradientBoosting, Extra Tree and RandomForest regressor Also test score and Cross Validation Score difference is also less.
* So we will go for Hyper parameter tuning for GradientBoosting, Extra Tree and RandomForest regressor model and will choose best model out of them

# Hyper Parameter Tunning RandomForest Regressor:

Graphical user interface, text, application, email

Description automatically generated

* **RandomForest Regressor**, after tuning the model with best parameters we can see the decreased accuracy from 89.30% to 88.87% and Cross Validation Score almost same Also Mean Square Error values has increased which means error has increased so we will not go for this model

**Hyper Parameter Tunning ExtraTreesRegressor:**

**Graphical user interface, text, application

Description automatically generated**

* **Extra Tree Regressor**, after tuning the model with best parameters we can see the decresed accuracy from 90.05% to 89.28% and Cross Validation Score almost same Also Mean Square Error values has increased which means error has increased so we will not go for this model.

**Hyper Parameter Tunning GradientBoosting Regressor**:

Graphical user interface, text, application, email

Description automatically generated

# Finally we selected GradientBoosting Regressor, after tunning the model with best parameters we can see the incresed accuracy from 90.55% to 91.39% and Cross Validation Score from 86.10% to 87.21% Also Mean Square Error values has reduced which means error has reduced.

# Saving the model and predictions using saved model:

# Save best model using .pkl as follows.

# Now after saving the best model, loading my saved model and predicting the test values.

# Predicted the SalePrice for test dataset(25% of train dataset) using saved model of train dataset, and the predictions look good.

# Also Predicted the SalePrice for test dataset using saved model of train dataset.

Graphical user interface, text, application

Description automatically generated

**Best Model Saving:**

* Predicted the SalePrice for test dataset(25% of train dataset) using saved model of train dataset, and the predictions look good.

Chart, scatter chart

Description automatically generated

# Conclusion::Best Model

* In this project report, we have used machine learning algorithms to predict the house prices.
* We have mentioned the step by step procedure to analyze the dataset and finding the correlation between the featuers. Thus we can select the features which are not correlated to each other and are independent in nature.
* Those feature sets were then given as an input to nine algorithms
* Hence we calculated the performance of each model using different performance metrics and compared them based on these metrics. Then we have also saved the dataframe of predicted prices of test dataset.
* To conclude, the application of machine learning in property research is still at an early stage. We hope this study has moved a small step ahead in providing some methodological and empirical contributions to property appraisal, and presenting an alternative approach to the valuation of housing prices.
* Future direction of research may consider incorporating additional property transaction data from a larger geographical location with more features, or analysing other property types beyond housing development.

We can observe both original and predicted attrition values are same. Conclusion is **GradientBoosting** as best model.

Chart, scatter chart

Description automatically generated

**Thank You**