



Housing Price Prediction Project Use Case Report



Submitted by:

Saurav Kumar

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my SME Shwetank Mishra as well as Flip Robo Technologies who gave me the opportunity to do this project on Housing Price Prediction, which also helped me in doing lots of research wherein I came to know about so many new things.

Also, I have utilized a few external resources that helped me to complete the project. I ensured that I learn from the samples and modify things according to my project requirement. All the external resources that were used in creating this project are listed below:

- 1- **Google**
- 2- **Github**
- 3- **Youtube**

INTRODUCTION

- Business Problem Framing Houses are one of the necessary needs of each and every person around the globe and therefore the housing and real estate market 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. We 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

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:

1. Which variables are important to predict the price of a variable?
2. How do these variables describe the price of the house?

● Review of Literature

Based on the sample data provided to us from our client database where we have understood that the company is looking at prospective properties to buy houses to enter the market. The data set explains it is a regression problem as we need 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. Also, we have other independent features that would help to decide which all variables are important to predict the price of the variable and how do these variables describe the price of the house.

● Motivation for the Problem Undertaken

Our main objective of doing this project is to build a model to predict the house prices with the help of other supporting features. We are going to predict by using Machine Learning algorithms. The sample data is provided to us from our client database. In order to improve the selection of customers, the client wants some predictions that could help them in further investment and improvement in selection of customers. House Price Index is commonly used to estimate the changes in housing price. Since housing price is strongly correlated to other factors such as location, area, population, it requires other information apart from HPI to predict individual housing price. There has been a considerably large number of papers adopting traditional machine learning approaches to predict housing prices accurately, but they rarely concern themselves with the performance of individual models and neglect the less popular yet complex models. As a result, to explore various impacts of features on prediction methods, this paper will apply both traditional and advanced machine learning approaches to investigate the difference among several advanced models. This paper will also comprehensively validate multiple techniques in model implementation on regression and provide an optimistic result for housing price prediction.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem We are building a model in Machine Learning to predict the actual value of the prospective properties and decide whether to invest in them or not. So, this model will help us to determine which variables are important to predict the price of variables & also how these variables describe the price of the house. This will help to determine the price of houses with the available independent variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features'). The most common form of regression analysis is linear regression, in which one finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion. For specific mathematical reasons this allows the researcher to estimate the conditional expectation of the dependent variable when the independent variables take on a given set of values. Regression analysis is also a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.

- Data Sources and their formats Data set provided by Flip Robo was in the format of CSV (Comma Separated Values). The dimension of data is 1168 rows and 81 columns. There are 2 data sets that are given. One is training data and one is testing data. 1) Train file will be used for training the model, i.e., the model will learn from this file. It

contains all the independent variables and the target variable. Size of training set: 1168 records. 2) Test file contains all the independent variables, but not the target variable. We will apply the model to predict the target variable for the test data. Size of test set: 292 records.

• Data Preprocessing Done

Data pre-processing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for building and training Machine Learning models. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. Data preprocessing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; therefore, it is extremely important that we pre-process our data before feeding it into our model. Therefore, it is the first and crucial step while creating a machine learning model. I have used some following pre-processing steps:

- a. Loading the training dataset as a dataframe
- b. Used pandas to set display I ensuring we do not see any truncated information
- c. Checked the number of rows and columns present in our training dataset
- d. Checked for missing data and the number of rows with null values
- e. Verified the percentage of missing data in each column and decided to discard the ones that have more than 50% of null values
- f. Dropped all the unwanted columns and duplicate data present in our dataframe

- g. Separated categorical column names and numeric column names in separate list variables for ease in visualization
- h. Checked the unique values information in each column to get a gist for categorical data
- i. Performed imputation to fill missing data using mean on numeric data and mode for categorical data columns
- j. Used Pandas Profiling during the visualization phase along with pie plot, count plot, scatter plot and the others
- k. With the help of Label encoding technique converted all object data type columns to numeric datatype
- l. Thoroughly checked for outliers and skewness information
- m. With the help of heatmap, correlation bar graph was able to understand the Feature vs Label relativity and insights on multicollinearity amongst the feature columns
- n. Separated feature and label data to ensure feature scaling is performed avoiding any kind of biasness
- o. Checked for the best random state to be used on our Regression Machine Learning model pertaining to the feature importance details
- p. Finally created a regression model function along with evaluation metrics to pass through various model formats

Data Inputs- Logic- Output Relationships

When we loaded the training dataset, we had to go through various data pre processing steps to understand what was given to us and what we were expected to predict for the project. When it comes to the logical part, the domain expertise of understanding how real

estate works and how we are supposed to cater to the customers came in handy to train the model with the modified input data. In the Data Science community there is a saying “Garbage In Garbage Out” therefore we had to be very cautious and spent almost 80% of our project building time in understanding each and every aspect of the data and how they were related to each other as well as our target label.

With the objective of predicting housing sale prices accurately we had to make sure that a model was built that understood the customer priorities trending in the market imposing those norms when a relevant price tag was generated. I tried my best to retain as much data possible that was collected but I feel discarding columns that had lots of missing data was good. I did not want to impute data and then cause biases in the machine learning model from values that did not come from real people.

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

The assumption part for me was relying strictly on the data provided to me and taking into consideration that the separate training and testing datasets were obtained from real people surveyed for their preferences and how reasonable a price for a house with various features inclined to them was.

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

I have used both statistical and analytical approaches to solve the problem which mainly includes the pre-processing of the data and EDA to check the correlation of independent and dependent features. Also, before building the model, I made sure that the input

data was cleaned and scaled before it was fed into the machine learning models. For this project we need to predict the sale price of houses, meaning our target column is continuous so this is a regression problem. I have used various regression algorithms and tested for the prediction. By doing various evaluations I have selected Random Forest Regressor as the best suitable algorithm for our final model as it is giving a good r^2 -score and least difference in r^2 -score and CV-score among all the algorithms used. Other regression algorithms are also giving me good accuracy but some are over-fitting and some are under-fitting the results which may be because of less amount of data. In order to get good performance as well as accuracy and to check my model from overfitting and under-fitting I have made use of the K-Fold cross validation and then hyper parameter tuned the final model. Once I was able to get my desired final model I ensured to save that model before I loaded the testing data and started performing the data pre-processing as the training dataset and obtaining the predicted sale price values out of the Regression Machine Learning Model.

● Testing of Identified Approach

- A. Linear Regression Model
- B. Support Vector Regression Model
- C. Decision Tree Regression Model
- D. Random Forest Regression Model
- E. K Nearest Neighbours Regression Model
- F. Gradient Boosting Regression Model
- G. AdaBoost Regression Model

Here we select Gradient Boosting Regression Model for model building.

```
In [154... x_train, x_test, y_train, y_test = train_test_split(princi_comp, y, test_size=0.20, random_state=96)

#training our model
gb=GradientBoostingRegressor()
gb.fit(x_train, y_train)

#Predicting y_test
y_gb = gb.predict(x_test)
gb_r2 = r2_score(y_test, y_gb)

#R2 Score
print("R2 score : ", gb_r2)

#Cross Validation Score
cross_score_gb = cross_val_score(gb,princi_comp,y,cv=5)
print(cross_score_gb)
print("Cross Validation Score : ",cross_score_gb.mean()*100)

#Error Estimation Methods
print("Mean Absolute Error :",mean_absolute_error(y_test,y_gb))
print("Mean Square Error : ",mean_squared_error(y_test,y_gb))
print("Root Mean Square Error :",np.sqrt(mean_squared_error(y_test,y_gb)))

R2 score : 0.9221551147020572
[0.90327303 0.8927663 0.81959897 0.86619444 0.82126927]
Cross Validation Score : 86.06204017085692
Mean Absolute Error : 12392.209552649687
Mean Square Error : 311436757.8400126
Root Mean Square Error : 17647.570876469446
```

AFTER HYPERPARAMETER TUNING

```
In [274... gbr =GradientBoostingRegressor(learning_rate= 0.1, max_depth= 3, max_features= 'sqrt', min_samples_split= 4)

gbr.fit(x_train,y_train)

pred = gbr.predict(x_test)

r2_score(y_test,pred)
```

Out[274... 0.9019982426838569

Predicted With test data ..

```
In [279]: #predicting values for test dataset

predicted_values = gbr.predict(princi_comp_test)

data = pd.DataFrame(predicted_values,columns=['Predicted Sales Prices'])

data
```

```
Out[279]:
```

	Predicted Sales Prices
0	129332.094866
1	248162.591677
2	198204.218687
3	231336.942742
4	149888.491065
...	...
798	126673.396774
799	105189.898801
800	147983.939059
801	48944.199044
802	197586.029768

803 rows × 1 columns

Conclusion

Post model building and choosing the appropriate model I went ahead and loaded the testing dataset. After applying all the data pre processing steps as the training dataset I was then able to get the predicted sale price results. Since the values were in array format, I converted them into a dataframe and merged it with the original testing dataframe that consisted only our feature columns. Once the testing dataset with feature columns and predicted label was formed, I exported the values in a comma separated values file to be accessed as needed

- Learning Outcomes of the Study in respect of Data Science

The above study helps one to understand the business of real estate. How the price is changing across the properties. With the Study we can tell how multiple real estate amenities like swimming pool, garage, pavement and lawn size of Lot Area, and type of Building raise decides the cost. With the help of the above analysis, one can sketch the needs of a property buyer and according to need we can project the price of the property.

- Limitations of this work and Scope for Future Work During this project I have faced a problem of low amount of data. Many columns are with the same entries in more than 80% of rows which lead to 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 fill those missing values in the correct manner. We can still improve our model accuracy with some feature engineering and by doing some extensive hyperparameter tuning on it.

