

HOUSING PRICE PREDICTION

Submitted by:

SHAILESH SAXENA

**ACKNOWLEDGMENT**

* Youtube
* Kaggle
* Krish naik videos

**INTRODUCTION**

* **Business Problem Framing**

**The US based Surprise Housing 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.**

* **Conceptual Background of the Domain Problem**
* **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**.

**.**

* **Motivation for the Problem Undertaken**

**Objective behind to make this project, this domain**

**Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. In this project, house prices will be predicted given explanatory variables that cover many aspects of residential houses. The goal of this project is to create a regression model that are able to accurately estimate the price of the house given the features.**

**Analytical Problem Framing**

* **Mathematical/ Analytical Modeling of the Problem**

**we will suggest some of its advantages and important features which will clear your mind why use the RF Algorithm in machine learning.**

* **Random forest algorithm can be used for both classifications and regression task.**
* **It provides higher accuracy through cross validation.**
* **Random forest classifier will handle the missing values and maintain the accuracy of a large proportion of data.**
* **If there are more trees, it won’t allow over-fitting trees in the model.**
* **It has the power to handle a large data set with higher dimensionality.**
* **Handle data without scaling**
* **Data Sources and their formats**

**A US-based housing company named Surprise Housing has decided to enter the Australian market. The company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file .**

* **Data Preprocessing Done**

**1)Handling the missing values.**

**2)Removing a few columns having more than 30% of null values.**

**3)Replace the null values by taking mode or the most frequent value of the column.**

**4)Handle all the Categorical data**

* **State the set of assumptions (if any) related to the problem under consideration**
* **1)All the NA values are equal to mod of the columms.**
* **2)Scaling not required.**
* **Hardware and Software Requirements and Tools Used**
* **import numpy as np**
* **import pandas as pd**
* **import seaborn as sns**
* **import matplotlib.pylab as plt**
* **from sklearn.model\_selection import cross\_val\_score**
* **from sklearn.model\_selection import cross\_val\_predict**
* **from sklearn.metrics import confusion\_matrix**
* **import warnings**
* **from sklearn.model\_selection import train\_test\_split**
* **from sklearn.metrics import mean\_absolute\_error**
* **from sklearn.metrics import mean\_squared\_error**
* **from sklearn.metrics import r2\_score**
* **from sklearn import linear\_model**
* **from sklearn.ensemble import RandomForestRegressor**

**Model/s Development and Evaluation**

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

**Using random forest regreessor**

* **Testing of Identified Approaches (Algorithms)**

**# Train - Test split**

**#just drop the outcome columns**

**#specify input and output attributes**

**#X is the input and y is the output**

**X = df\_final.drop('SalePrice', axis=1)**

**y = df\_final['SalePrice']**

**x\_train,y\_train,y\_train,y\_test = train\_test\_split(x, y, random\_state = 42, test\_size=0.15)**

**print(X\_train.shape,X\_test.shape,y\_train.shape,y\_test.shape)**

**rfg = RandomForestRegressor(random\_state=85)**

**#Training the model**

**rfg.fit(X\_train,y\_train)**

**y\_predict = rfg.predict(X\_test)**

**#Lets find the rmse and r2\_score using sklearn.metrics**

**from sklearn.metrics import r2\_score**

**from sklearn.metrics import mean\_squared\_error**

**print("RMSE is :",np.sqrt(mean\_squared\_error(y\_test,y\_predict)))**

**print("r2 score is : ",r2\_score(y\_test,y\_predict))**

RMSE is : 27531.06604002789

r2 score is : 0.8821046871309735

* **Key Metrics for success in solving problem under consideration**

**Metrics used are as follow**

* **Root mean squared eroor**
* **R2 error**
* **Visualizations**

**Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.**

* **Interpretation of the Results**
* Visiualiation of data tell us necessary features of the problem.

**CONCLUSION**

**Key Findings and Conclusions of the Study**

* **1)Random Forest Regressor gives about 90% accuracy without ensemble technique.**
* **2)Random Forest Regressor work better than Linear Regression.**

* **Limitations of this work and Scope for Future Work**
* **Need more time to find the better accuracy.**
* **Need to apply bagging and boosting technique to find the better result.**
* **Scaling the data may gives the better result with lasso and Ridge.**
* **There are another ways to implement the model.**