

HOUSE PRICE PREDICTION

Submitted by:

SHIVAM NAMDEV GADEKAR

**ACKNOWLEDGMENT**

This project is based on predicting house price on number of factors. For this project, I took help of my father who is friends with real estate agent. His information helped to consider more factor which should be in list and I told these factors also plays important role in House price prediction. Also I like to thank Mr. Krish Naik for helping in this EDA.

Coming to project, out of all data, there are some features which does not play an important role in defining price for house. To get clarity for those features, I looked into real estate website, House Websites that highlights important aspects that people are looking for and will pay more for more efficient and effective service.

**INTRODUCTION**

* Business Problem Framing

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.

* Conceptual Background of the Domain Problem

Houses are one of the necessary needs 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.

* Review of Literature

Project is divided into 4 stages. In first stage basic data reading, processing duplicates, unique and missing values.

In second stage, performing extensive EDA to get maximum output from data describing important features that plays important role in price prediction.

In third stage, separating out target variable, performing VIF methods to remove multi-collinearity from data. Encoding categorical features into numerical processable format.

In final stage, applying all models with varying random state for finding best possible solution. After finalizing model, performing parameter tuning for finding put best parameters. And finally deploying trained model for production.

* Motivation for the Problem Undertaken

Describe your objective behind to make this project, this domain and what is the motivation behind.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

For this model, I have chosen below models for predicting Salary. As for finding best model, I have run through all models with random state varying from 0-101.

* 1. Linear Regression
  2. Lasso
  3. Ridge
  4. ElasticNet
  5. SVR
  6. DecisionTreeRegressor
  7. RandomForestRegressor
  8. KNeighborsRegressor
  9. SGDRegressor
  10. GradientBoostingRegressor
  11. AdaBoostRegressor
  12. ExtraTreesRegressor

After running these models, Found that RandomForestRegressor with random state 6, is working good.

* Data Sources and their formats

Data was in csv format and sourced from Surprise Housing.

* Data Preprocessing Done

1. Checking Duplicate Values
2. Checking missing Values
   1. Dropping Missing Values with more than 15% of missing values
   2. Replacing missing values with median and Mode Imputation.
3. Handling Outliers –
   1. Checking outliers, with plots and replacing outliers.

* Data Inputs- Logic- Output Relationships

There are some features, which doesn’t play an important role in defining price of house. For modelling, I have dropped some of the features based on Variance Inflation Factor method.

* State the set of assumptions (if any) related to the problem under consideration
  + Numerical missing values are replaced by median.
  + Categorical missing values are replaced with mode.
* Hardware and Software Requirements and Tools Used
  + Intel i7-11800,
  + Jupyter Note Book
  + Python 3

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

* Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing.

* Run and Evaluate selected models

Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

* Key Metrics for success in solving problem under consideration

What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

* Visualizations

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

If different platforms were used, mention that as well.

* Interpretation of the Results

Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Describe the key findings, inferences, observations from the whole problem.

* 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.