HOUSING PRICE PREDICTION PROJECT

Submitted by:

MANISH KUMAR

ACKNOWLEDGMENT

I would like to thank **FLIP ROBO TECHNOLOGY** for providing me the opportunity to build **Machine Learning Model** on **Housing Price Data**. It enhanced my data analytical capability, understanding of Machine Learning algorithms at various aspects. & I had a close study or analysis of real estate institution which enhanced my understanding of how a real estate institution works. I also would like to thank Mr. Sajid Choudhary sir for the guidance in this project.

To work on case study of **Housing Price Prediction Data**. I have used my basic knowledge of data analytics & Machine learning algorithms which I learned from DATA TRAINED institute of data science. Also I have took the help of some famous websites like www.keggle.com , www.geeksforgeeks.com & www.stackoverflow.com to learn & implement the some useful concepts of data analysis & Machine Learning Models in my research. I have also had the help of some famous YouTube Channels of Data Science like:-

1. Edureka YouTube Channel.
2. Krish Naik YouTube Channel.
3. CodeWithHarry YouTube Channel.

INTRODUCTION

- **Business Problem Framing**

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. We 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?

Business Goal:

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

  Housing Price Prediction project is to predict the price of houses in Australian Housing & Real Estate market. First of all we will understand the core concept of the use case & its motive.

  In the Housing Price Prediction use case we have understood that an US based Housing Company named **Surprise Housing** decided to enter the Australian Housing & Real Estate market. And to enter in the market the company is looking for the prospective properties to buy houses to enter in the market. The company smartly uses the data analytics to purchase the houses at a price below their actual value 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 and that data set is provided us in CSV format to build a Machine Learning Model that can predict the actual price of houses in the Australian market that would be a help making the decision whether to invest in them or not.

- ## Review of Literature

  As we already discussed in Conceptual Background of the domain Problem that Housing Price Prediction is a Machine Learning and Data Analytics Project for an US based Housing Company which wants to enter in Australian Housing & Real Estate Market.

  In this project we've done the research on various factors of Real Estate domain, we have analysed the various important variables which are useful for predicting the price of houses and also which factors affect most the price of a house.

  The data for Housing Price Prediction comes from the **Surprise Housing** Company, they have collected the data from the Australian Real Estate Market and provided us to do the related work upon. The data consist of 81 Columns (different variables for the housing or properties) & 1460 rows of data.

The research and model building on this project has done in 8 important steps:-

1. Data Collection.
2. Data Description and Understanding of the data.
3. Data Engineering.
4. Exploratory Data Analysis.
5. Data Pre-processing.
6. Machine Learning Model Building & Metric Evaluation.
7. Cross Validation of the machine Learning Models.
8. Hyper Parameter Tuning of Finalized Model.
9. Model Saving and Conclusion.

- **Data Collection.**
  The first step of every data science project starts with the data collection. In this Housing Price Prediction Project the data is collected by the **Surprise Housing** Company from the Australian Real Estate Market and provided to us in csv format.

- **Data Description and Understanding of the data.**
  At the starting of the project first of all we imported the dataset in Jupyter Notebook & than started to have a close look in the data set. At first we found that the data set came in 2 parts.
  1. Train Dataset
  2. Test Dataset

  We were instructed to do all the research and model building related work on train data & after building the model test it on test dataset as test dataset has not the price column because we needed to predict the price of test data while testing of the model.
  After that we found that the data set has total 19 columns with missing values, 43 columns had Object data type, 34 column had integer data type & 4 columns were with float data type.

- **Data Engineering.**

  After having a close look onto the data set and understanding the data there were some important data engineering process needed to perform in order to proceed further for research analysis and ML model building.

  First of all we combined the Train & Test data and merged it in one data set because we needed to perform the same Data engineering, EDA & pre-processing work on both data set so we decided to merge them and perform all together after performing all the necessary pre-processing part we would be separating those train & test data again.

  After merging the data set first we decided to fill the missing values in all 17 columns individually while doing a detailed data analysis of each column and comparing it with different columns and we filled the null values in all columns while comparing them with other important and related columns and we used multiple techniques to perform this task. 2 of the column from missing values columns had more than 95% missing data so we decided to remove those columns from the dataset along with ID column which we felt of no use in the further work.

- **Exploratory Data Analysis.**

  In EDA & Visualization of continuous columns of the data(In Distribution Plot) we can observe almost all kind of data distribution, some of the column's data distribution are heavily right skewed like **LotArea, BsmtFinSF1, BsmtFinSF2, TotalBsmtSF, 1stFlrSF, LowQualSF, GrLivArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea, MiscVal, LotFrontAge, MasVnrArea**. We also observed the left skewed data distribution in the column: - YearBlt & GarageYrBlt.

  One more point to be added here that we made a list of continuous columns based of data type of the columns which is int64 & float and while doing so we have added some of categorical column which had data type integer or

float & we can observe these column's data distribution Kernel Density Estimation is like wave because the data is distributed in very limited categories.

The correspond Boxplot of these right & left skewed column shows the outliers within these columns. We need to remove these outliers from the column but the problem is if we will remove the outliers from these columns with IQR method we will lead to a heavy data loss because we had observed that the number of columns with outliers are very large and some of the columns have very large amount of outliers so we are going to remove the outliers with z-score method & after that we will remove the skewness of the data.

- **Data Pre-Processing**

  In Data Pre-Processing we have removed all the outliers from the data set with the help of z-score & we also had compared the by re-visualising the distribution & outliers and found & observed that data has been much sorted from the previous distribution.

  Encoding of the Object Data Type columns are done. At first we made a list of all the object data type and after that we used Ordinal encoder in a FOR LOOP to convert all the object data types columns into float data type column. We confirmed encoding of columns by rechecking data information.

  We have 2 columns in the data set in which only 1 kind of data is present that means their value counts is 1, which is common for all the other columns and target column so we do not need these column for model building hence it is better to drop these columns from the data set.

  These columns are `Utilities` & `PoolArea`.
  In `Utilities` column there was only one category with the name of `AllPub` & after encoding it became 0.0

  In PoolArea there were values other than 0 but after removal of outliers all other values got removed and only 0 left.

Hence both these column in the data set will add no value in the prediction because they are saturated columns and their relation is common with other column. So it is better to remove these column from the data set. And we did the same.

- **Machine Learning Model Building & Metric Evaluation.**

  We have built 7 Machine Learning Regression models for Housing Price prediction & 2 regularization models of Linear Regression model. Out of which Linear Regression model has given the highest accuracy score which is 93.05%, it's 2 regulaization model has also given almost same accuracy score. Support Vector Regressor has given the worst score which is -8%.

  Accuracy Score of models is as follows:-

  - **Linear Regression** Model r2 accuracy score is = `0.9305102798580087`
  - **Lasso Regression** Model r2 accuracy score is = `0.9307027865352177`
  - **Ridge Regression** Model r2 accuracy score is = `0.9304986522738888`
  - **Polynomial Regression** Model r2 accuracy score is = `0.9108296062033461`
  - **K-Nearest Neighbors Regressor** Model r2 accuracy score is = `0.831133171293738`
  - **Decision Tree Regressor** Model r2 accuracy score is = `0.791015886831979`
  - **Random Forest Regressor** Model r2 accuracy score is = `0.9103908113045052`
  - **XG Boost Regressor** Model r2 accuracy score is = `0.8968271345378613`
  - **Support Vector Regressor** Model r2 accuracy score is = `-0.08007104818505972`

- **Cross Validation of the machine Learning Models.**

  We have successfully cross validated all the 9 models with 20 folds **(CV = 20)** and got the mean score of each model as follows:-

- **Linear Regression Model** Cross Validation Mean Score = `0.8766040250632579`.
- **Lasso Regression Model** Cross Validation Mean Score = `0.8768639308207126`.
- **Ridge Regression Model** Cross Validation Mean Score = `0.8764875865804139`.
- **Polynomial Regression Model** Cross Validation Mean Score = `0.8419583403051343`.
- **K-Nearest Neighbors Regressor Model** Cross Validation Mean Score = `0.8114889371456041`.
- **Decision Tree Regressor Model** Cross Validation Mean Score = `0.7315444021902423`.
- **Random Forest Regressor Model** Cross Validation Mean Score = `0.8721079772546659`.
- **XG BOOST Regressor Model** Cross Validation Mean Score = `0.8641200787246086`.
- **Support Vector Regressor Model** Cross Validation Mean Score = `-0.06280947105551629`.

After cross validating of all the models we have analyzed that **Lasso Regularization** of **Linear Regression** Model & also **Linear Regression** has given the best accuracy score in both the model accuracy score & model cross validation score. Over all every model has given good cross validation score except **Support Vector Regressor** Model, this model has given worst score on both model accuracy & cross validation score.

We have observed that though **Linear Regression, Lasso Regression, Ridge Regression, Random Forest & XG Boost** model have performed best and given the highest score but the **K-Nearest Neighbors Regressor** has given the minimum difference between `Model accuracy score` & `Model cross validation score`. Which proves that **KNN** model is not overfit or underfit model it has very less difference between `accuracy score` & `cross validation score`.

Hence we have finalized **K-Nearest Neighbors Regressor** model for our **`Housing Price Prediction Project`**.

- **Hyper Parameter Tuning of Finalized Model.**

  In Hyper parameter tuning of K-Nearest Neighbors algorithm we got some good productivity as we significantly increased the model's accuracy by more than 3% with the help of hyper parameter tuning.

  So, here we have completed the machine learning model building & we ended up by building **`K-NearestNeighborsRegressor`** algorithm's model for the **`Housing Price Prediction Project`**.

- **Model Saving and Conclusion.**

  After successfully copleting the HyperParameter Tuning of the model, we have successfully loaded the saved model in the local system using **Pickle** and also predicted the test data from the loaded model when we had predicted housing price value of test data. We had concluded this project by making a data frame of predicted price of test data.
  Here our project of Housing Price Prediction is completed with conclusion which we concluded by predicting the price of housings of Hosuing test data.

  And we have saved the predicted price of housing in a Data Frame named **Predicted_price**.