HOUSING PRICE PREDICTION

Introduction

In this article, I will walk you through the process of creating a Machine Learning model in Python utilising popular machine learning libraries NumPy, Pandas, and scikitlearn to estimate house prices in the United States.

Low inventory, fierce competition, and large price increases have harmed purchasers since 2020, but quickly rising mortgage rates are making it even more difficult to find an affordable house. Housing price trends are not only a source of anxiety for buyers and sellers, but they also provide insight into the present economic condition. As a result, it is critical to estimate property prices without bias in order to assist both buyers and sellers in making judgments. This study makes use of an open-source dataset that includes 81 explanatory characteristics and 1,168 house transactions in AMES, USA. The goal is to forecast the most effective house pricing for real estate consumers based on their budgets and goals. Future prices will be projected by examining recent market trends and price ranges, as well as forthcoming changes.

Machine Learning is essential to help machines understand like people and to strengthen AI. Data science is becoming an increasingly significant tool for assisting businesses in increasing overall revenue, boosting marketing techniques, and focusing on shifting trends in house sales and purchases.

The link contains two datasets: Train Dataset and Test Dataset.

The training set has 10683 records.

Test set size: 2671 records

I developed the model using the Training set and predicted the House Prices in the Test dataset.

Predicting house prices is critical for driving real estate efficiency. As previously stated, house prices were derived by computing the acquiring and selling prices in a neighbourhood. As a result, the House Price Prediction Model is critical in closing the information gap and increasing Real Estate efficiency.

I need to model property prices using the provided independent variables. The management will then utilise this model to understand how the prices fluctuate with the variables. As a result, they can alter the firm's strategy and focus on regions that will provide significant returns. Furthermore, the model would help management comprehend the pricing dynamics of a new market. The relationship between house prices and the economy is a major motivator for forecasting house prices.

Libraries Used

There are three sets of libraries in use.

Basic Data Analysis and Visualization Libraries Data Cleaning and Feature Engineering Libraries (Data pre-processing) and Libraries for Creating ML Models.

Analytical Problem Framing

I verified the first 5 elements in both sets after loading the train and test datasets.

The train dataset has 1,168 rows and 81 columns, including our Target "Sale Price," and the Test dataset has 292 rows and 80 columns.

I discovered that it is by looking at the continuous Target column "Sale Price." a Regression issue I examined the information and unique values in each column. I noticed a massive number of missing values in some columns and more than 80% of data in certain columns as a 0 As a result, I removed those columns to avoid significant bias and variation. I made use of an Imputation method for replacing Nan values in train and test datasets. We retrieved some significant columns from the available columns. Following that, I have evaluated the dataset with plots and other visualisation techniques such as bar plot, dist plot, and so on. I also used the Target to plot features. Then I removed outliers, skewness, and identified correlation between variables before normalising the dataset with the Standard scaler. I built five models and utilised hyper parameter optimization to get the best result. Each model's actual and anticipated values have been shown using Reg plot.

The Cross Validation score was also computed, and the best model was found and saved for future predictions. I projected house prices for the Test dataset using the saved model and saved the results in a csv file. Anyone can find property values using my model if the features are similar to those in my model.

My internship company, FlipRobo Technologies, provided the data. The train and test datasets were in csv format. Another text file has been added with the essential information about the variables and their values. This greatly aided my analysis and comprehension of the dataset. My train dataset included 1168 rows and 81 columns including the target, while my test dataset had 292 rows and 80 columns removing the target. I have data of the object, float, and integer kinds.

I loaded the dataset and imported the relevant libraries.

Statistical analysis such as Shape, information, nunique, and value counts.

Removed columns with more than 80% NaN values and "0" values.

Imputation method for replacing other NaN values.

Removed ID and Utilities columns, which had all unique values and only one value, respectively.

Age was extracted from the dataset based on the year mentioned. All of the processes have been completed for both the Train and Test datasets.

Boxplot was used to discover the relationship between Categorical columns and Target. Scatterplot, Swarm plot, and Strip plot were used to discover the relationship between Numerical columns and Target. Several columns are related to the Target in a linear fashion. While certain columns lack a distinct pattern.

MODEL DEVELOPMENT AND EVALUATION

I double-checked the dataset's information for null values and datatypes.

I have 3 float data, 35 INT data, and 43 Object data. I then went over the missing data.

There is a lot of missing information. I removed the columns that had more than 80% missing values.

To replace Nan values, I utilised an imputation technique.

To eliminate outliers, I utilised the Percentile technique.

Skewness was removed using a Power transformer.

Ordinal Encoder was used to convert category columns to numerical columns.

Pearson Correlation was employed to determine the relationship between variables.

To scale and normalise the data, I used Standard Scaler.

I created models to estimate house prices using various Machine Learning algorithms.

I have a Regression Problem because our target is Sale Price, which is continuous. I built the models using five different techniques and calculated the R2 and CV scores for each one. I finally opted on the Random Forest Regressor model because it has the smallest difference between r2 and CV Score.

Linear Regression, KNN Regressor, Random Forest Regressor, XGB Regressor, and Gradient Boosting Regressor were all used.

First, I chose the best Random state with the highest score and ran a traintest-split to fit the model. My linear regression model score is 85.69 percent, and my CV score is 79.97 percent. I tweaked with the best parameters, but the score stayed unchanged.

I found the optimal random state that produces the highest score and then fitted the model. KNN Regressor had an R2 score of 82.75 percent and a CV score of 73.43 percent. I experimented with several options, but the score did not increase.

XGB Regressor had an R2 score of 88.92 percent and a CV score of 83.30 percent. After adjusting the Hyper parameters, the score did not improve.

Gradient boosting had the greatest r2 value of 90.78 percent among all other models.

However, the CV score was quite low, and the difference is more than my best model. GBR's CV score is 83.14 percent.

I started by finding the best random state and fitting it to the model. I received an 89.89 percent score and an 83.92 percent CV Score. As a result, I chose the random forest as the best model and saved it. During the fitting of the final model, my score increased to 90.02 percent.

The model's hyper parameter adjustment did not improve my score.

The regplot of real versus predicted for the Random Forest model demonstrates that it is a good model.

I obtained an accuracy of 90.02 percent while saving the Final model. As a result, my model gets kept for future forecasts.

I used the r2 score as the model's accuracy score. To calculate the error rate in the model, I utilised mean squared error and mean absolute error. I utilised root mean squared error and chose the model with the lowest amount as the best match. I also used Cross Validation Score to cross-check with the r2 score and pick the best model with the smallest difference between the r2 score and the cv Score mean.

I used barplots to depict categorical data counts.

MS Zoning-in residential low density zoning classification has a higher count. Lot Config- inner lot is selected by more people as a lot configuration. Neighbourhood-northwest Ames has the highest count, while blue stem has the lowest for physical location within ames city borders. Condition1 & Condition2- Normal condition has a higher count for both condition1 and condition2, which are close to each other. More people favour single-family detached buildings.

House Style-one story dwelling has the highest count of all sorts. Roof style-gable has the most numbers as a sort of roof and the fewest counts as a shed. Outside 1st and Exterior 2nd- the majority of the houses in the dataset have vinyl siding as the exterior covering of the house and asphalt shingles imitation shicco, with brick common having the fewest counts. Exergual & Extercont-the outside material's quality is normal or typical. Foundation-Poured concrete and cinderblock have the highest count as foundation types, while wood has the lowest. BSMTQuality-a greater number of people prefer a basement height of 80 to 89 inches, while the least number prefers a basement height of 70 to 79 inches.

For walkout or garden level walls, BSMTExposure-no exposure has the highest count and minimum exposure has the lowest. The BSMF type-unfinished followed by good living quarters has a higher count than the basement finished type and low quality. Heating-gas A has a higher count. The majority of the homes in the sample have central air conditioning. More residences utilise an electrical-standard circuit breaker, and just a handful use a mixed type of electrical system. More people prefer the sale type-warranty deed as the type of sale. Sale Condition-the most common sale condition is having the greatest counts and the fewest counts of adjacent land purchase is a condition of sale.

I used distplot to examine the distribution of numerical data.

Skewness can be found in practically all number columns. I'll use the power transformer to correct the skewness.

To determine the relationship between category and goal, I used a boxen plot.

Residential Low density Zone property has the highest price and demand. Paved roads are more desirable and more expensive.

The price and range of a somewhat irregular lot shape are higher than those of a normal lot shape, which is desired by more clients.

Prices are greater in the Northridge and Northridge Heights communities.

Normal proximity to diverse conditions is more expensive than any other.

1 Family detached dwellings are in higher demand and cost more than other types.

Two-story homes are more expensive.

Hip roofs are more expensive.

Poured concrete foundations are more expensive than slab foundations.

The Sale Price for Standard Circuit Breakers & Romex (Sbrkr) of Electrical system (Electrical) is Maximum.

The Sale Price is high for a completely finished (Fin) interior of the garage (Garage Finish).

The highest Sale Price is for a newly constructed and sold home (New) and a contract with a 15% down payment on ordinary terms (Con) of the kind of sale (Sale Type).

The Sale Price is maximal if the home was not completed when it was last assessed (related with New Homes) (Partial) Condition of sale (Sales Condition).

I visualised numerical columns with target using swarm plot, strip plot, and scatter plot.

The fee rises as the garage space grows up to three automobiles.

In general, as the size of the garage increases, so does the price.

There is no pattern for Wood Deck SF or OpenPorhSF.

The price has decreased throughout the years.

Price lowers as garage age increases.

As the 2ndFlrSF climbs, so does the price.

Price rises as ground level area increases.

The price and demand for 0 and 1 full baths in the basement are considerable.

If the full bath is 3, the price is higher than 0,1 or 2 for 4 bedroom residences above ground level.

The price for one kitchen above ground level is greater.

As the number of rooms above grade increases, so does the price.

If there are two fireplaces, the price is more.

2-STORY 1946 AND NEWER and 1-STORY 1946 AND NEWER ALL STYLES are selling for a higher price.

Lot Frontage & Lot Area- There is no set pattern for the number of linear feet of street connecting to the property and the lot size in square feet.

Overall Qual- As the overall quality of the house's materials and finish improves, so does the sale price.

Overall Cond- As the overall condition improves, so does the sale price; average-rated houses are in higher demand, and the sale price rises.

Price has an increasing trend with MasVnrArea, BsmtFinsF1, BsmtUnfSF, TotalBsmtSF, and 1stFlrSF.

CONCLUSION

Higher mortgage rates mean that many buyers can no longer afford homes in certain price ranges. The issue is that even basic single-family homes now cost the same as opulent pads did a few years ago, so purchasers are forced to choose between waiting for additional inventory to become available or moving to a more inexpensive place.

I employed machine learning techniques to anticipate housing prices in this project report. I described the step-by-step approach for analysing the dataset and determining the correlation between the features. As a result, we estimated and compared the performance of each model using several performance metrics. Then we saved the test dataset's data frame of projected prices.

I've noticed that certain characteristics, such as Overall Cond, Exergual, and so on, add the most to the house's price. In addition, factors such as the number of years since construction have a negative impact on the price. The value decreased with time.

The model is declared to have passed the test based on the results of the tests that have been performed. The advancement of computing technology has enabled the examination of social data that could not previously be recorded, processed, and evaluated. The power of visualisation has aided us in comprehending data through graphical representation. One of the most critical tasks is to eliminate missing numbers and zero values and replace them with the appropriate mean, median, or mode. This study is an exploratory attempt to estimate house prices using five machine learning algorithms and then compare their results.

I hope that our study has taken a tiny step forward by making methodological and empirical contributions to property appraisal and proposing an alternate approach to home price valuation. Future study could include combining additional property transaction data from a bigger geographical area with more attributes, as well as analysing various property kinds other than housing construction.

There were numerous outliers in the sample, and data loss was significant when utilising the Z-score approach.

We were forced to adopt the Percentile method, which is less effective than the Z-score method in removing outliers.

There was a lot of skewness in the dataset, which would damage the model again because we need to transform it.

This study used only a few easy regression methods to a few complex ones, rather than all advanced algorithms.

To avoid data leakage, I did not integrate the Train and Test datasets.

There were a few columns that had more 0 values and a couple that had more missing values.

Those columns must be removed.

There was multicollinearity, and the columns with the strongest correlation with the Target had to be eliminated to avoid it.

Despite these limitations and disadvantages, my model performs well, with an accuracy of 90.02 percent with the Random Forest model and a CV Score of 83.92 percent.