SCHOOL OF COMPUTING

UNIVERSITY OF TEESSIDE

MIDDLESBROUGH

TS1 3BA

Machine Learning (CIS4035-N)

Machine Learning Application and Report

Preethi Mandadi - W9457747

House prices prediction

Abstract

The framework predicts the prices of houses in the United States based on square feet of space in this machine learning application. The Kaggle dataset provides a data collection of US house prices. The method calculates real estate values in the United States. It advises businesses on how to invest in real estate in the United States. Machine learning algorithms can be used to forecast costs.

Introduction

In this machine learning application, the system is predicting the prices of the houses in the USA as per square feet of area. A data set from us houses prices is get from the Kaggle dataset. The system estimates the prices of real estate of the USA. It provides the predictions to the business to invest on the property in the USA. The machine learning applications can be used to persist the prices of the any other places in changing the dataset.

Dataset URL:
https://www.kaggle.com/shree1992/housedata

Problem Statement

Real estate markets, such as those in major cities, provide an intriguing opportunity for data analysts to examine and forecast where property values are headed. Property price forecasting is getting more significant and helpful. Property prices are an excellent predictor of a country's overall market health as well as its economic health. We are wrangling a big amount of property sales records kept in an unknown format and with unknown data quality concerns based on the data given. Main objective of the machine learning model is to select the best suitable machine learning model to predict the prices of the dataset. This prediction helps for investments on lands and property.

Selections of Machine Learning algorithm

Simple Linear regression:

The link between two variables or factors is shown or predicted using linear regression models. The dependent variable is the factor that is being predicted (the factor that the equation solves for). The independent variables are the factors that are utilized to predict the value of the dependent variable.
In basic linear regression analysis, the two variables are referred to as x and y. The regression model is the equation that defines how y is connected to x.
The simple linear regression model is represented by:
$y = \beta 0 + \beta 1x + \varepsilon$

Lasso Regression:

Lasso regression is a sort of shrinkage-based linear regression. Data values are shrunk towards a central point, such as the mean, in shrinkage. Simple, sparse models are encouraged by the lasso approach (i.e., models with fewer parameters). This form of regression is ideal for models with a lot of multicollinearities or when you wish to automate elements of the model selection process, such as variable selection and parameter removal. Least Absolute Shrinkage and Selection Operator is an acronym that stands for LASSO.
L1 regularization is used in Lasso regression, and it adds a penalty equal to the absolute value of the magnitude of the coefficients. This form of regularization can lead to sparse models with few coefficients; certain coefficients may become zero, and

the model may be removed. Larger penalties provide coefficient values that are closer to zero, which is great for making simpler models. L2 regularization, on the other hand, does not result in the deletion of coefficients or sparse models (e.g., Ridge regression). As a result, the Lasso is significantly easier to understand than the Ridge.

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Multivariate Regression:

Multivariate Regression (MVR) is a supervised machine learning approach that analyses numerous data variables. With one variable y and numerous independent variables, a multivariate regression is an extension of multiple regression. We try to anticipate the outcome based on the number of independent factors.

Multivariate regression aims to create a formula that can explain how variables respond to changes in others at the same time.

One of the most basic Machine Learning Algorithms is Multivariate Regression. It belongs to the category of Supervised Learning Algorithms, in which we are given a training dataset. The most essential benefit of multivariate regression is that it aids in the understanding of connections between variables in a dataset. This will make it easier to comprehend the relationship between dependent and independent variables. A frequently used machine learning algorithm is multivariate linear regression.

Data exploration and preparation:

Because most algorithms require data to be presented in a certain way, datasets usually require some preparation before they can provide valuable information. Some datasets contain values that are either missing, invalid, or otherwise difficult to handle by an algorithm. The algorithm is unable to use data that is missing. If the data is incorrect, the algorithm will generate less accurate or even false results. Some datasets are reasonably clean but require shaping, while others simply lack meaningful business context (for example, poorly defined ID values), necessitating feature enrichment. Clean and well-curated data is produced through good data preparation, which leads to accurate results.

Data is extracted from the Kaggle datasets which is a US Prices. Getting the data and predicting the prices need a data preparation. For that data has been loaded into panda’s dataset. Axes with labels are also included in the data structure (rows and columns). Both the row and column labels align for arithmetic operations. It's possible to think of it as a duct-style container for Series objects. The basic data structure of pandas is working this way.

The act of altering raw data so that data scientists and analysts may run it through machine learning algorithms to find insights or make predictions is known as data preparation (also known as "data preprocessing").

- Data Filtering: After lording the data pandas dropna() method is uses to remove the null values
- Data Formatting: And preparing the data so that every row in the columns should give the same datatype.
- Inconsistent values: make sure that all the values are acceptable and single formatted data

Proposed Solution:

Every algorithm has a unique way of predicting the feature there is a few drawbacks for every algorithm.

Disadvantages of Multivariate Regression:

- Multivariate approaches are a little more complicated and need a lot of math.
- The output of the multivariate regression model might be difficult to comprehend at times since certain loss and error outputs are not similar.
- Smaller datasets are not well-suited to this strategy. As a result, the same cannot be said of them. For bigger datasets, the outcomes are better.

Disadvantages of Lasso Regression:

The purpose of lasso regression is to find the subset of predictors that produces the least amount of prediction error for a quantitative response variable. The lasso does this by placing a restriction on the model parameters that leads some regression coefficients to decrease toward zero. After the shrinking procedure, variables having a regression coefficient of zero are removed from the model. The response variable is most closely connected with variables having non-zero regression coefficients. Explanatory variables might be quantitative, categorical, or a combination of the two. You will use and analyze a lasso regression analysis in this course. You'll also get experience with k-fol more errored.

Several characteristics will be highly skewed. LASSO chooses at most n features for np (n number of data points, p number of features). LASSO will choose only one characteristic from a set of linked characteristics; the choice is purely random. The characteristic chosen for different boot strapped data might be rather varied. Ridge regression performs worse than prediction.

Advantages of Linear Regression:

linear Regression is a fairly simple technique that may be quickly implemented and produces good results. Furthermore, when compared to more complicated techniques, these models can be trained quickly and efficiently on systems with limited processing resources. When compared to other machine learning techniques, linear regression has a significantly reduced time complexity. Linear regression's mathematical formulae are very simple to comprehend and interpret. As a result, linear regression is a simple concept to grasp. Linear regression almost completely fits linearly separable datasets and is frequently used to determine the nature of the connection between variables. When a machine learning model fits a dataset very well, it catches the noisy data as well. This is known as overfitting. This has a detrimental effect on the model's performance and diminishes its accuracy on the test set.

Regularization is a simple approach that successfully reduces the complexity of a function, lowering the danger of overfitting.

This model using the Linear regression for the prediction of the feature price values.

Linear Regression Model predictions

Simple Linear Regression model is selected to predict the prices.

Design:

Simple linear regression is used to describe a linear connection between a response and only one explanatory variable. I'd want to forecast housing prices, and the price will be our response variable. However, we must choose a characteristic for a basic model. When I looked at the columns in the dataset, the most relevant factor seemed to be living area (sqft). When we look at the correlation matrix, we can see that pricing has the strongest correlation coefficient with living area (sqft), which confirms. As a result, I choose to use living area (sqft) as a feature, but we can choose another feature if we want to investigate the link between pricing and another characteristic.

Implementation:

I also created an empty data frame. The Root Mean Squared Error (RMSE), R-squared, Adjusted R-squared, and mean of the R-squared values generated by the k-Fold Cross Validation are included in this data frame, which are useful metrics to check performance of the model. A tighter R-squared value and a smaller RMSE indicate a better match. In the parts that follow, I'll populate this data frame with my findings.

Data preparation is completed and filtered the data by applying various panda functions.

Experiments:

Conducted several experiments to get the high-performance machine learning model.

- Tune the model by changing the test and train size of the data and changing the random state of the model.
- Tuning the model by implementing Ridge regression models with the alpha values.
- Also implemented the lasso model and tuned with alpha settings.

Modelling Results:

Taken filtered data is divided into train and test data.

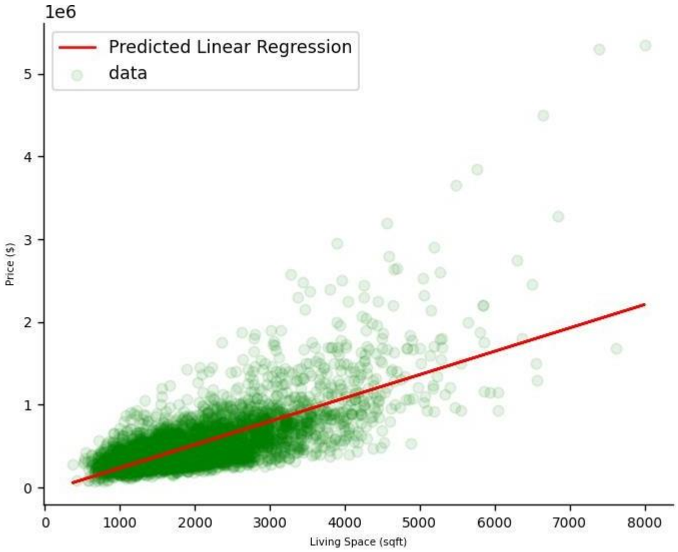
By taking scikit learn train_test_split split the data on 80% as train data and remaining 20% as a test data.

Divided the data into X and Y train and test data and fitted in to a model before that from Scikit learn Linear regression model is imported.

Fitted the model in to the data and stated the predictions.

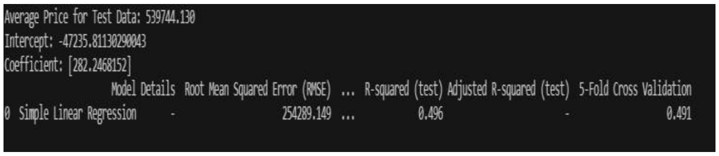
Trained the model with the data and tested the results.

From mat plot lib data predictions is seen.



Evaluation

As the model predicts the output, the model is evaluated using R square mean and adjusted mean. The Root Mean Squared Error (RMSE), R-squared, Adjusted R-squared, and mean of the R-squared values generated by the k-Fold Cross Validation are included in this data frame, which are useful metrics to check performance of the model



```
Average Price for Test Data: 539744.130
Intercept: -47235.81130290043
Coefficient: [282.2468152]
                 Model Details  Root Mean Squared Error (RMSE)  ...  R-squared (test) Adjusted R-squared (test)  5-Fold Cross Validation
0  Simple Linear Regression     -                     254289.149  ...             0.496                         -                    0.491
```

Conclusion

The linear regression (all characteristics, no preprocessing) is the best in the evaluation table. However, I have reservations regarding its validity. The ridge regression is second preferred model, however other models may be beneficial depending on the scenario. I used to scikit-learn in this kernel because it has all the built-in functions I need.

The regression line is significant because it improves the accuracy of dependent variable estimate and permits the estimate of a response variable for people whose carrier variable values are missing from the data. The author also concluded that there are two approaches for forecasting a variable: within the range of values of the sample's independent variable (interpolation) or beyond this range (extrapolation) (extrapolation). The author suggested the first technique since it is safe, however there are issues about how to establish the linearity of the relationship between the two variables.

Future Work

As per the experience with the work, the project needs to implement more and get high percentage predictions score with the advance regression models like in Un supervised learning algorithms such as random forest and deep learning models.

Implementing the various machine learning algorithms and deep learning models to predict the feature of the real estate.

Learning and building models for the more complex and higher frequent data sets.

Reference

Mathumitha Mahendran (2020), Fine-Tuning your Linear Regression Model, Fine-Tuning your Linear Regression Model | Jigsaw Academy.

Saishruthi Swaminathan (2018), Linear Regression — Detailed View, https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86

Jason Brownlee (2016), Linear Regression for Machine Learning, https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86

SRIRAM PARTHASARATHY, Top 5 Predictive Analytics Models and Algorithms, Top 5 Predictive Analytics Models and Algorithms | Logi Analytics Blog

Burhan Y. Kiyakoglu Predicting House Prices | Kaggle

Nagesh Singh Chauhan, From Data Pre-processing to Optimizing a Regression Model Performance, From Data Pre-processing to Optimizing a Regression Model Performance - KDnuggetss