Problem

Real estate is a very important pillar of the economy. However, despite a large amount of data available, we do not have precise measurements of house prices. Therefore, this study aims to apply machine learning to predict home selling prices based on various economic indicators.

Introduction

Modeling uses machine-learning algorithms, where the machine learns from the data and uses them to predict new data. The most frequently used model for predictive analysis is Linear Regression. As we know, the proposed model for accurately predicting future outcomes has applications in economics, business, the banking sector, the healthcare industry, e-commerce, entertainment, sports, etc. One such method used to predict house prices is based on multiple factors. A potential homebuyer considers several factors like location, living space and land space, number of bedrooms and bathrooms, and most importantly the house price.

The Data used for this project consists of 600,000 house prices from all over the US, which are listed at Zillow.com. The dataset has 1 table with a total of 16 categorical features 10 numerical features and 2 special features. This makes it a good dataset for learning how to extract features from location and DateTime features for spatiotemporal modeling. And a large number of dataset samples makes it a “tall” dataset which will make it possible to achieve useful predictive accuracy for a wide range of locations and home types. The first step is to clean this data and ensure fits for purpose. Then look at which features would provide interesting insights and answer questions that would help to predict the value of a house. Finally, dive into building a model capable of predicting house prices.

Criteria for success: Main metrics for model evaluation in regression:

R Square/Adjusted R Square

R Square measures how the model can explain much variability independent variable. It is the square of the Correlation Coefficient(R) and that is why it is called R Square.

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

R Square is calculated by the sum of the squared prediction error divided by the total sum of the square which replaces the calculated prediction with the mean. The R Square value is between 0 and 1 and a bigger value indicates a better fit between prediction and actual value.

R Square is a good measure to determine how well the model fits the dependent variables. However, it does not take into consideration of overfitting problem. If your regression model has many independent variables, because the model is too complicated, it may fit very well with the training data but performs badly for testing data. That is why Adjusted R Square is introduced because it will penalize additional independent variables added to the model and adjust the metric to prevent overfitting issues.

Mean Square Error(MSE)/Root Mean Square Error(RMSE)

While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for fit.

$$MSE = \frac{1}{N} \sum\nolimits_{i=1}^{N} (y_i - \hat{y}_i)^2$$

MSE is calculated by the sum of the square of prediction error which is real output minus predicted output and then divided by the number of data points. It gives you an absolute number on how much your predicted results deviate from the actual number. You cannot interpret many insights from one single result but it gives you a real number to compare against other model results and help you select the best regression model.

Root Mean Square Error(RMSE) is the square root of MSE. It is used more commonly than MSE because firstly sometimes MSE value can be too big to compare easily. Secondly, MSE is calculated by the square of error, and thus square root brings it back to the same level of prediction error and makes it easier for interpretation.

Mean Absolute Error(MAE)

Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of the square of error in MSE, MAE is taking the sum of the absolute value of error.

$$MAE = \frac{1}{N} \sum\nolimits_{i=1}^{N} |y_i - \hat{y}_i|$$

Compare to MSE or RMSE, MAE is a more direct representation of the sum of error terms. MSE gives larger penalization to big prediction errors by squaring them while MAE treats all errors the same.

The resulting deliverables are:

- An initial model
- A notebook (Jupyter Notebook file)
- A write-up detailing the requirements, methodology, and resulting model
- A slide deck
- A project report

REFERENCES

1. Wu, S. (2021, June 5). What are the best metrics to evaluate your regression model? Medium. Retrieved August 26, 2022, from https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b
2. Alyousfi, A. (2018, December). House price prediction using machine learning techniques. Retrieved August 26, 2022, from https://ammar-alyousfi.com/assets/documents/house-price-prediction.pdf
3. ANAND G. RAWOOL, DATTATRAY V. ROGYE, SAINATH G. RANE, DR. VINAYK A. BHARADI (2021). House Price Prediction Using Machine Learning. IRE Journals, 4(11), 29–33.
4. Farah Khanum, Nisarga .P .M, Navitha Pawar, Vijayalakshmi .D, R. Anitha. (2021). Real Estate House Price Prediction using Machine Learning. International Journal of Engineering Science and Computing (IJESC), 11(07), 28453–28454.