Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project

Dean De Cock Truman State University Journal of Statistics Education
Volume 19, Number 3(2011)

-Rishabh Mehta (801257231)

Problem Statement:

Every day, thousands of residences are sold. Every buyer has various queries, such as: What is the genuine price that this home deserves? Is the price I'm paying reasonable? A machine learning approach is developed in this work to forecast a house price based on data about the house (its size, the year it was built in, etc.).

Data:

https://www.kaggle.com/code/ammar111/house-price-prediction-an-end-to-end-ml-project/data

Motivation:

- To apply data preprocessing and preparation techniques in order to obtain clean data
- To build machine learning models able to predict house price based on house features
- To analyze and compare models' performance in order to choose the best model

Survey of Related Work:

Related Papers:

1. Alkhatib, K., Najadat, H., Hmeidi, I., & Shatnawi, M. K. A. (2013). Stock price prediction using k-nearest neighbor (kNN) algorithm. International Journal of Business, Humanities and Technology, 3(3), 32-44

2. Feng, Y., & Jones, K. (2015, July). Comparing multilevel modelling and artificial neural networks in house price prediction. In Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2015 2nd IEEE International Conference on (pp. 108-114). IEEE.

Survey of Related Work:

Related Paper 1: Stock price prediction using k-nearest neighbor (kNN) algorithm

1. In this article, a Bayesian regularized artificial neural network was utilized to forecast future stock values. Previous stock information, as well as other financial technical data, are fed into the model. The model's output is the closing price of the related stocks the next day.

2. This sort of network's weights are probabilistic in nature. This enables the network to automatically punish highly complicated models (with many hidden layers). As a result, the model's overfitting will be reduced.

3. This data set spans 734 trade days (4 January 2010 to 31 December 2012).Each data point contained the following daily statistics: low price, high price, opening price, closure price, and trading volume. This data was separated into training and test data containing 80% and 20% of the original data, respectively, to ease model training and testing. Six additional variables were established to depict financial indications in addition to the daily-statistics variables in the data.

Survey of Related Work:

Related Paper 1: Stock price prediction using k-nearest neighbor (kNN) algorithm

1. The performance of the model were evaluated using mean absolute percentage error (MAPE) performance metric. MAPE was calculated using this formula:

$$MAPE = \frac{\sum_{i=1}^{r}(\text{abs}(y_i - p_i)/y_i)}{r} \times 100$$

where pi is the predicted stock price on day i, yi is the actual stock price on day i, and r is the number of trading days. When applied on the test data, The model achieved a MAPE score of 1.0561 for MSFT part, and 1.3291 for GS part.

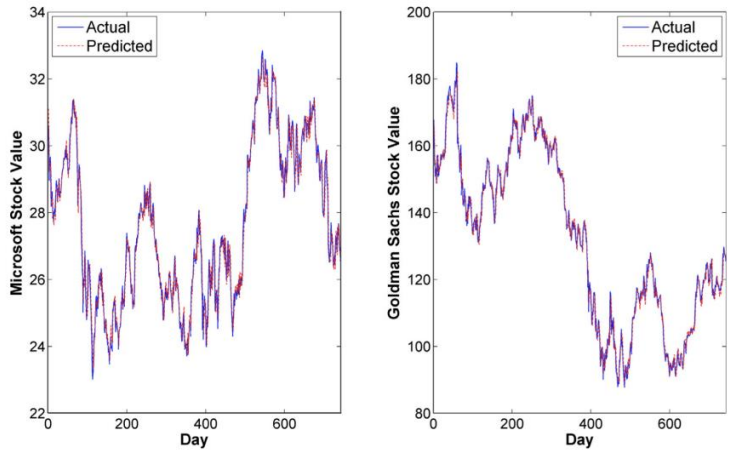


Figure 1: Predicted vs. actual price

Survey of Related Work:

Related Paper 2: House Price Prediction Using Multilevel Model and Neural Networks

1. To predict house prices, two models were built: a multilevel model (MLM) and an artificial neural network model (ANN). These two models were compared to each other and to a hedonic price model (HPM).

2. The macro-level equation, which describes the links between neighborhoods, and the micro-level equation, which specifies the interactions between homes inside a specific neighborhood, are both integrated into the multilevel model. The hedonic pricing model is a method for estimating home values based on characteristics like the number of bedrooms, size, and other factors.

3. The dataset has about 65,000 items in total. The dataset was split into a training set that comprises information about home sales from 2001 to 2012 and a test set that contains information about home sales in 2013 to enable model training and testing.

Survey of Related Work:

Related Paper 2: House Price Prediction Using Multilevel Model and Neural Networks

1. Three scenarios were used to evaluate the MLL, ANN, and HPM models. Locational and measurable neighborhood characteristics were not present in the data for the first scenario. Grid references for the position of the houses were included in the data for the second scenario. Measured neighborhood characteristics were included in the data for the third scenario.

2. The models were compared in goodness of fit where R 2 was the metric, predictive accuracy where mean absolute error (MAE) and mean absolute percentage error (MAPE) were the metrics, and explanatory power.

COMPARISONS OF GOODNESS-OF-FIT

	R^2 (training set)	R^2 (test set)
HPM1	0.39	0.23
MLM1	0.75	0.75
ANN1	0.39	0.23
HPM2	0.43	0.3
MLM2	0.75	0.75
ANN2	0.41	0.26
HPM3	0.68	0.65
MLM3	0.75	0.74
ANN3	0.69	0.67

COMPARISON OF PREDICTIVE ACCURACY

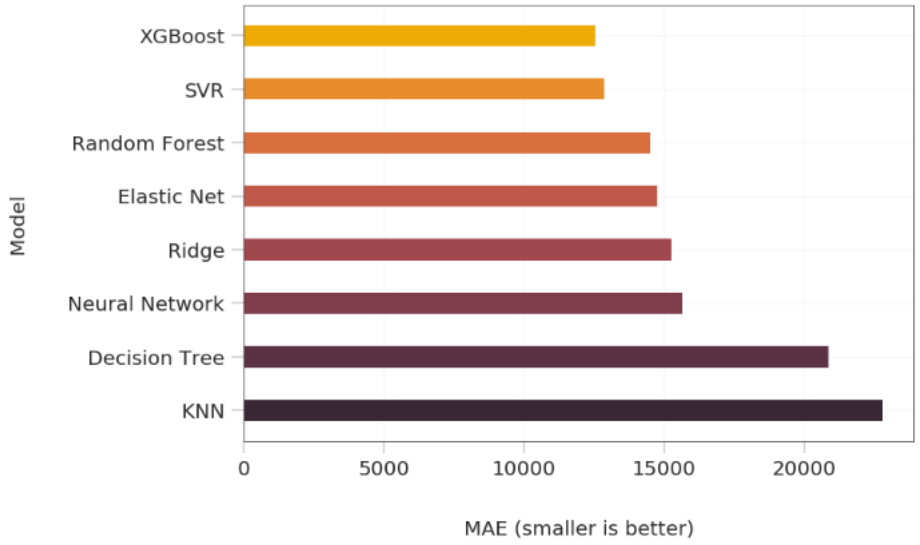
Test set	MAE (lnP)	MAPE (lnP)	MAE (raw price)	MAPE (raw price)
HPM1	0.319	5.89%	80.4	30.9%
MLM1	0.178	3.29%	48.6	17.5%
ANN1	0.318	5.85%	80.1	30.0%
HPM2	0.304	5.61%	77.0	29.4%
MLM2	0.178	3.29%	48.6	17.5%
ANN2	0.313	5.76%	79.0	29.8%
HPM3	0.210	3.89%	25.3	20.7%
MLM3	0.178	3.30%	48.8	17.6%
ANN3	0.216	4.00%	55.7	20.9%

Figure 4: Model performance comparison

Summary of the Method:

This study develops many regression models to forecast a house's price given a few of its characteristics. The models are evaluated and compared to identify the one that performs best. The steps in this paper's data science workflow are as follows: collecting the data, cleaning and preparing it, exploring it and creating models, assessing the findings, and conveying them visually.

Model	MAE
XGBoost	12556.68
Support Vector Regression (SVR)	12874.93
Random Forest	14506.46
Elastic Net	14767.91
Ridge	15270.46
Neural Network	15656.38
Decision Tree	20873.95
K-Nearest Neighbors (KNN)	22780.14



Plan:

Week	Plan
1	Studying the dataset and research papers in detail
2	Data Preparation and Exploratory Data Analysis
3	Prediction Type and Modeling
4	Model Building, Evaluation, Analysis
5	Report Preparation, Presentation.