PREDICTING HOUSE PRICES WITH MACHINE LEARNING

Abstract:

In this study, the machine learning algorithms k-Nearest-Neighbours regression (k-NN) and Random Forest (RF) regression were used to predict house prices from a set of features in the Ames housing data set. The algorithms were selected from an assessment of previous research and the intent was to compare their relative performance at this task

Contents:

INTRODUCTION:

Accurately estimating the value of real estate is an important problem for many stakeholders including house owners, house buyers, agents, creditors, and investors. It is also a difficult one. Though it is common knowledge that factors such as the size, number of rooms and location affect the price, there are many other things at play.

BACKGROUND:

The field of Data Science is rather young, having taken form over the last halfcentury as a discipline distinct from statistics. It is also rapidly growing with many interesting advancements in recent years, most notably within Machine Learning (ML). This has resulted in an increase in media attention as well as funding of AI related businesses and research projects.

MACHINE LEARNING ALGORITHMS.

In this study two machine learning algorithms were compared against each other in order to investigate which one is more successful in predicting housing prices. As mentioned in the previous section, Baldominos et al.

K-NEAREST NEIGHBOURS REGRESSION:

k-Nearest neighbours (k-NN) is a non-parametric algorithm that can be used for both classification and regression problems. The algorithm relies on the assumption that any item in the data set should be have a similar value for the prediction target if they share similar values for other features.

PREDICTING HOUSE PRICES WITH MACHINE LEARNING

RANDOM FOREST REGRESSION:

Random forest is an algorithm which can be used both for classification and regression. Random forest models are constructed by using a collection of decision trees based on the training data. Instead of taking the target value from a single tree, the Random forest algorithm makes a prediction on the average prediction of a collection of trees.

METHODS:

In order to answer the question about what machine learning method is better to use for the house price problem the algorithms k-NN and Random Forest, as motivated in section 2.1, have been compared in terms of their prediction accuracy. Instead of implementing the algorithms from scratch for this study, algorithms from the scikit-learn library have been used.

RESULTS:

By following the method stated in the previous section, the following results have been obtained. To begin with, the two methods have been tested with different values for the selected hyperparameters as described in the previous section

DISCUSSION:

In this study the two machine learning regression algorithms k-Nearest neighbour and Random forest have been compared when trained and tested with the Ames housing data set. This has been done in order to study how accurately they, as machine learning methods, predict the prices for the house pricing problem.

CONCLUSIONS:

The research question for this study is to study how well house prices can be predicted by using k-Nearest neighbour and Random forest regression. In this study we have found that the Random forest regression algorithm performs better at predicting house prices than the k-Nearest neighbour algorithm.