Analysis and data processing of the boston hosuing from 1978

F. Wieser, Noah Fournier, Nikita Pond, Arshia Tajlili Moghanjoghi

Contents

1 Introduction

The aim of the project was to predict the parameters provided, the Boston housing data, with the independent parameters Longitude and Latitude. Essentially creating a position basis and relating a prediction at every location of this set of longitude and latitude. To do this, we had to find a way of finding these values from just the longitude and latitude. The aim was carried out by using the lon and lat, as a test set and create a model that would then attempt to find a variable. Using the concept of a nested keras, creating a subsequent model which includes the first prediction as nested variable to produce a second layer of model. Ultimately in order to have a comparison basis and as a method of optimizing the choice of variables the second layer was carried out for a selection of variables, changing the order of variable finding to improve accuracy. The procedure was then optimized for the prediction of house prices as accurately as possible. Using these prediction a person pointing on to a map at a certain longitude and latitude in the Boston area would be able to find the house price in a relatively accurate way. The accuracy relativity is decided based on how low the degrees of freedom are for the training variable. The interactive map of Boston were the house prices were shown on the input of the longitude and latitude on a mouse click on a certain position.

2 main block

All of the modules were loaded and ran in pandas. As mentioned in the introduction the order of these variables was important. Therefore by creating an order correlating matrix it was possible to find which factors correlate the most in comparison to other variables. It was found that certain variables, such as "DIS" could be found to a high degree of accuracy ($\approx 5.5\%$ error) using only positional data as they are heavily correlated with the positional data. Whereas other variables, such as "CMEDV" needed far more input variables to return an accurate number. The comparison between the data and the trained module was done in several ways, Firstly there was a comparison between a heat map created on the map of Boston and the data set on the longitude and latitude. The pandas module was used to plot the heatmap with the help of **follium.map**. The second sort of comparison was done based on the trained neural network in the figure 2. The model prediction vs the true

value is shown below. The green marker being around zero shows a higher confidence in the model prediction which was created using a sequencial model on tensor flow.

3 Conclusion

The model worked quite well after the nested kernels were used. Initially the model did not return very accurate data when it was trained using sklearn library. The "CMEDV" only depending on the coordinates and trained using an svc returned an error of around 12 %. However with the nested kernels the model significantly improved and relationships between the Longitude and Latitude and other variables was discovered. The heat maps showed clearly the relationship between the variables and the coordinates. Some other training modules were used which later on could be used for a comparison. The random tree training seems like a promising module to work with in the further future.

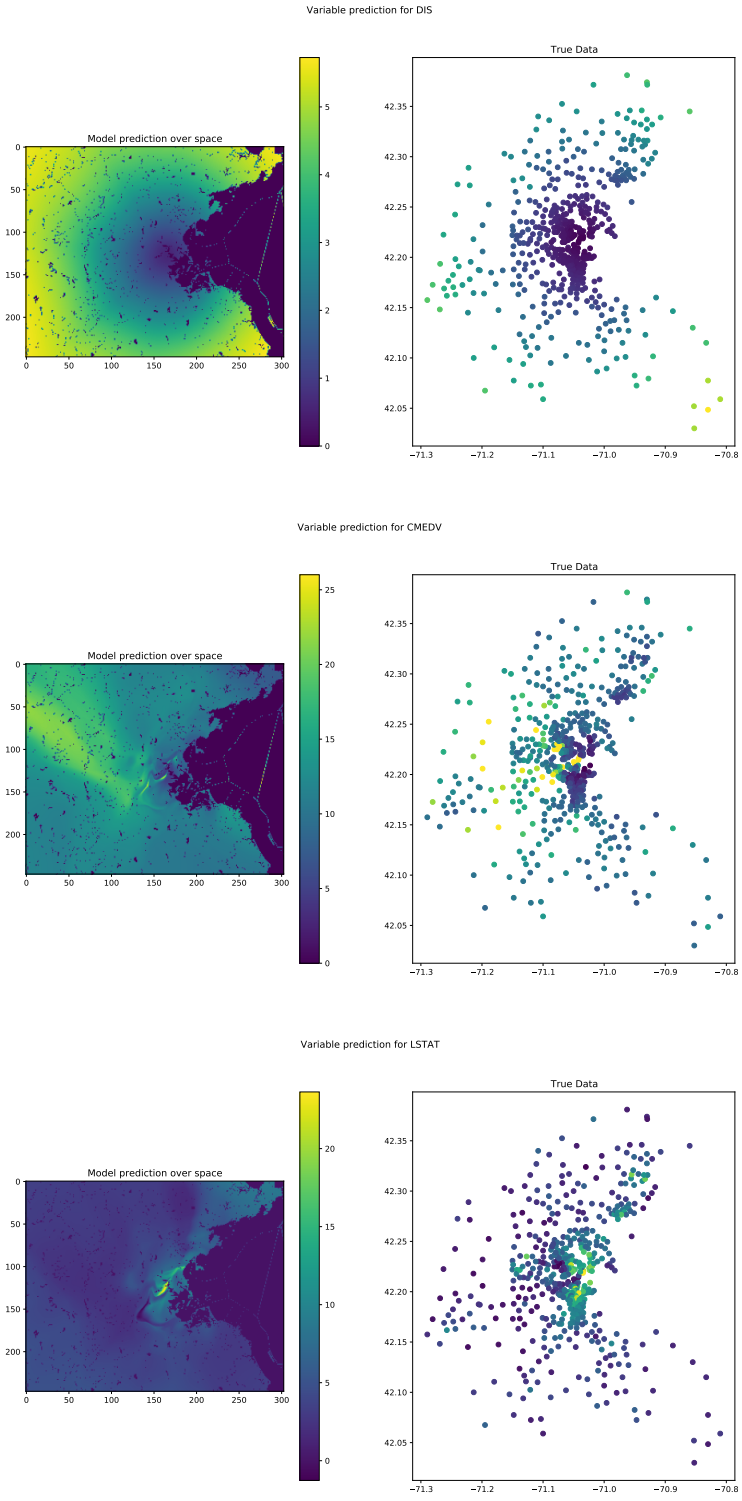

Figure 1. The heat map on the left with the predicted results for the three parameters with the actual data on the right hand side presented in the coordinate grid

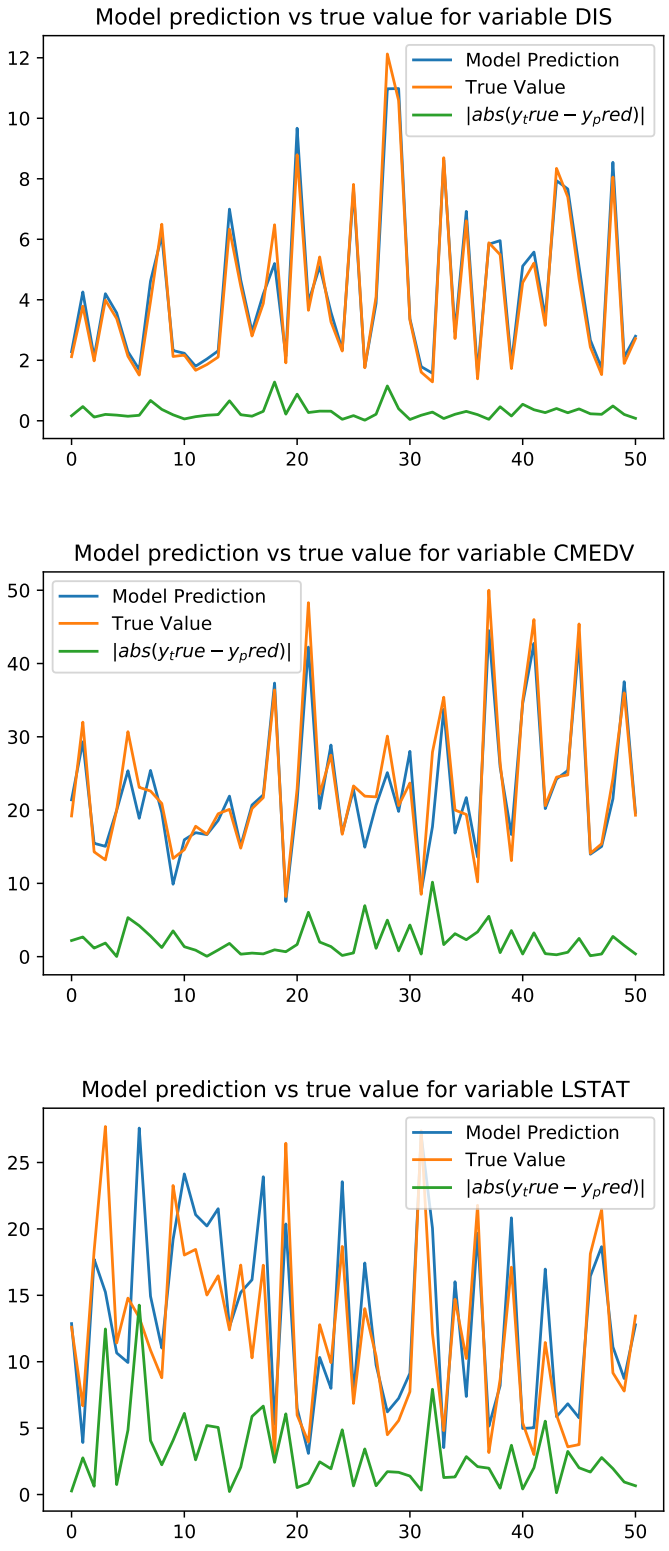

Figure 2. The comparison of the training set data trained with the prediction to the actual data plot