Comparing Lasso Regression and Graient Boosting Regressor on Stock Price estimation

Frederik Bonfanti
Filinta Yilmaz
M. Faisal Sulaimankhel
Nawid Hbrahimgel
Thomas P. Crilly

2019-3-21

Abstract

Real Estate is still one of the biggest investments for people, so its important. Predicting housing prices can seem easy at first, but it is dependent on way more things than is ibvisou and the relations are not always easy to see. THe Keggle competition House Prices: Advanced Regression Techniques provides a good starting point for applying linear regression models to the real estate market. In this paper we show extensively how to deal with the data and compare lasso regression with the gradient boosting regressor.

1 Introduction

We used two approaches of linear regression on the same dataset for housing prices in - Say a lot about housing prices, the features etc. - Why is it relevant? - What is regression? What kinda approaches are there?

Regression is a statistical machine learning approach that can be used to estimate relationships between variables; trying to predict dependent variable based on other independent variables. The most common loss function is the squared error loss funciton, but there are many different approaches and modifictions of the core linear regression idea.

- What approaches did we choose and why?

- Tell a lot about lasso regression, how it works, upsides, downsides - Tell a lot about gradient boosting regressor, how it works, upsides, downsides

- How did we do it? - What is our data? How much do we have? - Test set split - Validation Methods - How did we test the methods? - What result did we arrive at (Should that already be here?)

2 Methods

2.1 Technologies Used

Notebook server for Collaboration Pythong with the sklearn package (right?) and pandas

2.2 Dataset

Our dataset comes from the Kaggle comepetition "House prices adanced regression techniques [2] All in all it
- 50% of the data was randomly chosen and held back as a test dataset

2.3 Data Origin

The Dataset used here is the Ames Dataset [1] which is based of housing sales from between 2006 and 2010 in Ames, Iowa. It is based on a real dataset from the municipality in Ames, but has been cleaned up to make it more suitable for the task of linear regression. This cleanup consistent of a filtering to only include residential properties and removal of all previous sales of a property that has been sold multiple times. Also, some of the categorical features have been modified to be more easily understandable and provide some more information. This only concerns naming and no effective change on the data though. The dataset contains 2930 observations with 80 features, of those are roughly half categorical and half continuous. Features include many different area measurements of the house, such as basement size, living room size etc. and various counts of available facilities (such as bathrooms, kitchens, bedrooms above ground...). Categorical features include data about the location (neihgbourhood, street name etc.), the kind of heating, the style of the roof top and others. Non of the data have been changed, the only changes have been the removal of some columns or rows. Therefore this dataset can be considered a real dataset.

2.4 Data Preprocessing

We are using a real dataset, that means that we first have to deal with outliers and missing data. Both the training and test dataset were treated together in the following ways: Missing Data:
Two different approaches were used to deal with missing data in our dataset. To understand what was needed, first we analysed how much data was missing on a per-feature-basis. Some features showed a huge amount of missing data, while others had fewer rows with missing entries. [insert missing data graphics here] Features were split into two categories based on the percentage of missing data, with the split set at 50%. Features that had more than 50% of entries missing were removed from the dataset, while we used replacement by the mean or media for the missing entries in the remaining features.
Distribution:

2.5 Outliers

Outliers are datapoints that are very different from the rest of the data. Formally defined one would usually assume a distribution (e.g. the normal distribution) on a given datarow and assume everything below and above a certain percentile an outlier. We chose to discard outliers, because they do not provide much vlaue for the linear regression approaches that were used to analyze the data later. Thus we removed all rows which had features below the 5 percentile or above the 90 percentile.

3 Result and Discussion

3.1 Data Preprocessing

A whole lotta charts, discussing step by step what we did and why

3.2 Lasso Regression

A whole lotta charts, discussing step by step what we did and why
We used the root mean square error as the loss function for lasso regression.

3.3 Gradient Boosting Regressor

A whole lotta charts, discussing step by step what we did and why
We also used the root mean square error as the loss function for the gradient boosting regressor.

3.4 Comparison

3.5 Further Research

3.6 Conclusion

References

[1] Dean De Cock. Ames, iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3), 2011.

[2] www.kaggle.com. Kaggle: House prices: Advanced regression techniques, 2018. [Online; accessed 26-March-2019].