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

The aim of our paper is to compare two commong variations of the linear regression machine learning algorithm. Linear Regression is one of the most basic machine learning algorithms, and it works by gradually learning linear relationships between the input features and the output. Even though this is a basic algorithm, it has time and again been proven to be one of the most powerful and builds a baseline for many variations and other approaches. 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.
To evaluate these machine learning technologies, it was chosen to attempt the keggle competition "House Prices: Advanced Regression Techniques" [3] In this paper two common machine learning algorithms are considered.
The first one is lasso regression, which works like linear regression but adds one term to the loss function. This term is equivalent to the sum of the absolute values of the weights. Adding the weight of the features to the loss function is a common way to make the algorithm "prefer" solutions which smaller weights; in other words it is more likely to converge to a simpler solution. Lasso regression, as opposed to ¡TODO: OTHER SIMILAR REGRESSION METHOD HERE¿

applies the absolute value of the weights instead of the square, this makes it more likely to drop the weights of uncorrelated features completely to zero, and thus removes those terms from the regression equation. This is very similar to L1 normalization.
The second machine learning algorithm we used is the gradient boosting regressor. Gradient boosting works by assembling a number of weak learners step by step to create one strong learner.

The dataset we are using is the Amex housing prices dataset, consisting of real propery sales prices with many different features for each property. Housing prices have already previously been used as one of the most common ways to test regression methods, as they tend to have a great availability of data, many different features - both categorical and numerical - and are based on real data.

The approach used in this report starts by analyzing the dataset, removing outliers and processing missing data
- 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

For collaboration on this project we decided to use a centrally hosted IPython Jupyter Notebook server, this enabled us to use the power and availability of a provisioned server while all being able to collaborate in real time. The programming language we used was python 3 with the common data processing and machine learning packages such as sklearn, numpy and pandas.

2.2 Dataset

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.

2.3 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 high proportion of missing values, while others had fewer observations in which the value was missing.

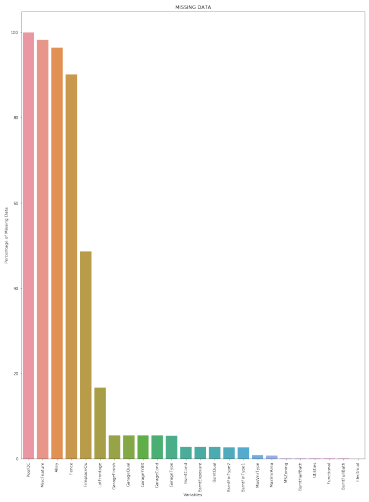
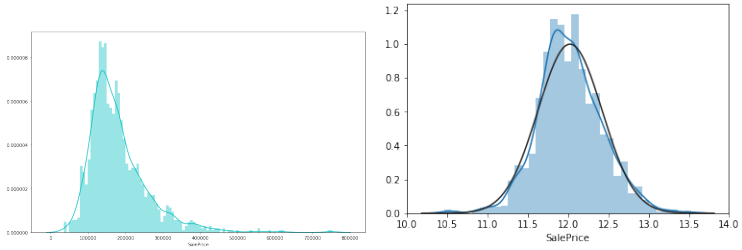


Figure 1: Missing data per feature

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:
The target value is the sales price. For linear regression with the sum of squared errors loss function, we are essentially assume the target to be normally distributed. By looking at the distribution of the price data**??** we can see that it is not normally distributed. To make up for this the logarithm of the price was taken, which left us with a distribution which comes very close to the normal distribution**??**.

2.4 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.



(a) Distribution of price (b) After normalization

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

Lasso regression is a technique used in machine learning to analyze regression methods, which allows both variable selection whilst performing regularisation which increases the accuracy and reduces overfitting [2]. Lasso regression is a combination of regression with the l1 normalization, equivalent to minimising the sum of the weights. Due to the advancements in computing power lasso regression has become more relevant and variations have been made. One such is the LARS algorithm which yields an efficient way to solve lasso regression whilst also connecting lasso regression to forward stagewise regression. Another variation is nearly isotonic regression, where isotonic regression given a data set will find to minimise the sum of their differences squared. The problem with this is that it assumes a monotone series. To fix this, nearly isotonic regression assumes a nearly monotone sequence which is half L1 penalty on differences. This allows us to compare nearly monotone assumptions where the best can be used to solve isotonic regression.
To compare

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] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

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