Project Proposal

Zizhou Sang and Yuqian Cao

Mar 7, 2020

1 Background and motivation

Data science has been a popular topic these days. As beginners who seek to march into the industry, we may be capable of speaking pros and cons of several algorithms, but lack experience with the entire workflow. Even if an oracle had revealed us which algorithm best suits our problem, we do not know what data preprocessing steps should be taken, and which set of variables and parameters should be used to yield the best solution.

Therefore, in forming project ideas, after balancing the trade-off between interestingness of question and depth of exploration, we finally chose to focus on the latter, and decided our research should be built on a well-structured and reasonably complex dataset. Eventually, we chose a classic Kaggle competition: the House Prices: Advanced Regression Techniques.

2 Dataset description

Our house price training set contains 1460 observations and 81 attributes, including the sale price, basic information (size, shape, type, utilities availability, year built), quality (rates of the overall material, internal quality, external quality, basement quality) and some special features that the house might have (near to railroad).

The goal is to use the above training set to train our models, then run on the testing set, which consists of 1459 observations with identical structure except the `SalePrice`. Kaggle will return the standardized MSE score and overall ranking after submitting the predictions.

3 Research questions

We mainly have three research questions as below:

1. Model accuracy. How precise we can be about the final prediction?

   This is the most common question and is what the data competition intended to. In fact, our main goal is try to build a precise model using whatever justified approaches.

2. Model interpretability. What are the factors that contribute most to the house price?

   Besides accuracy we also want to build interpretable models that can provide practical guides to people in real world. For example, a house seller might want to know which attributes matter most so he or she can improve the selling.

3. Model generality. How much of our findings can be migrated to other datasets?

   Besides this problem we also wonder how well does our refined model perform on other house datasets, compared to other models. Will our experience from this competition enable us to shortcut to top in other datasets as well? We hope to find out!

4 Plan for data analysis

We made the plan below according to our research questions:

1. Model accuracy. We basically want to test most of available regression algorithms on our dataset. So far we have run algorithms covered in the lecture plus the SVM. In the future we will test on other popular methods like XGBoost and stacked regression, as well as refined parameter tuning, feature extraction. The whole procedure will be more of an exploration on methods focusing on one dataset.

2. Model interpretability. We will run different algorithms that can tell the importance of variables and compare. Besides the attributes we also concern the reason of different answer output from different algorithms.

3. Model generality. We will migrate our findings to other housing dataset such as Airbnb and see how well our experience help to solve similar problems. If time permit we will try other non-housing data as well and see whether our progress has general value.

5 Current results

We expect future response of research question 1 to be of similar form as below:

Table 1: Scoreboard

Method	MSE($\times 10^{10}$)	Score	Ranking	Percentage
SVM-lin ($c = 25$)	0.16258	0.14630	2547	55%
Ridge ($10\lambda^*$)	0.16258	0.16178	3108	67%
Ridge (λ^*)	0.17994	0.17065	3242	70%
PLS ($n = 6$)	0.18909	0.17266	–	–
LASSO (λ^*)	0.18336	0.18095	–	–
OLS	0.19128	0.18753	3544	76%
PCR ($n = 6$)	0.18336	0.20511	–	–

[1] We performed a random split of original training data, using 70% of it as training set and the other 30% as testing set. All methods are run on our training set and MSE is calculated from predictions of the testing set.

[2] The Score is the returned result from Kaggle after submission.

[3] λ^* represents the best λ selected from 10-fold cross validation.