Machine Learning Project Proposal_Group 9

Xia Xicheng, Shen Li, Li Feijiao, Wang Yifan, Jin Weiguo

Jun 11th 2019

1 Project Description

We decided to choose *House Prices: Advanced Regression Techniques* from the Kaggle competition as our project topic. Our goal is to use simple machine learning models such as *Lasso Regression* to obtain a decent result after applying feature engineering technique.

2 Data Structure

Kaggle provides both training dataset (1460×80) and test dataset (1459×79). Training datasets includes 1460 entries with 80 variables, in which 79 are explanatory variables and 1 is explained house price.
Test dataset is similar to training dataset except for the absence of house price.

3 Methodology

a Read those kernels;

b Data cleaning. Cope with missing values and abnormal values(median filter, winsorize);

c Single factor exploration (allows us to narrow our focus on plausibly important factors based on our intuition);

d Feature engineering. Dummies, linear/non-linear transformations;

e Dimension reduction. PCA;

f Scale the data. Normalization/Standardization, max_min_scalar.

g Select models and parameters. Lasso/Ridge/SVR, cross-examination/regularization/penalization/learning-rate/iteration-times

4 Performance Evaluation Criteria

Kaggle ranks all the submissions according to their relative performance. The competition score is measured by the out-of-sample *Root Mean Squared Logarithmic Error*. Currently, there are 4652 submissions from all over the world. As beginners in the machine learning field, we would like to set our target reasonably at the ranking of 3000, with a score (error) of less than 0.15. An optimistic target would be 0.13 (ranked around 2000).