House Prices: Advanced Regression Techniques

Abstract:

The document contain a one page report mentioning the analysis of the competition taken from Kaggle, It contains information from loading/Reading our data set to the predictions that we have submitted on Kaggle Competition as our project to Big Data Lab held at CINECA

Group Members Luca Pedretti, Sana Intakhab, Giulia Capestro, Luca Bonafede

2/7/2018

HOUSE PRICES: ADVANCED REGRESSION TECHNIQUES

STEPS TAKEN FOR DATA ANALYSIS FOR KAGGLE COMPETITION

Following steps are taken using "R" for our Project

1. Search the topic on Kaggle Competition & Selected House Prices (Advanced Regression Techniques).
2. Download the data set from Kaggle Competition; Load the data in R in order to start our analysis.
3. Data Preparation has been carried out in order to visualize variables present in data set and also to take a look on missing values and we replaced all numerical variables' missing values as 0. Now our data is ready to for analysis.
4. The first step taken is Feature Engineering that is the creation of feature functions relevant to a specific ML algorithm and domain. Feature functions can be thought of as composites of variables that can help quantify the relationships between inputs, variables, or values specific to a given domain. We also add some new variables.
5. Ordered Categorical values are converted Numerical type values using Qualitative Scale Method. (e.g. Very Poor, Poor, Fair, Good, Very Good)
6. Binarization of the not ordered categorical values are done to make ease in running algorithms.
7. We have taken some different weighted parts in order to predict the sale prices of the houses.
8. We have used 4 techniques to predict the Sale Prices of the House that are as follows:
    - Multi-linear Regression
    - Random Forest
    - Gradient Boost and Extreme Gradient Boosting Xgb
    - Support Vector Machine SVM (Linear, Polynomial & Radial)
9. After running the algorithms we have the predictions, we submitted our competition. Score: 0.12884 (Top 20%).. 981
10. Results of Predictions from high to low are Random forest, XG Boost, Multi Linear Regression.