Data Mining Final Project Presentation --- House Price Prediction

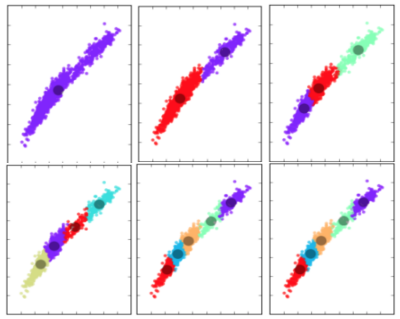
Huanyu Yang, Yirui Wang, Zewei Li, Zihao Xiao

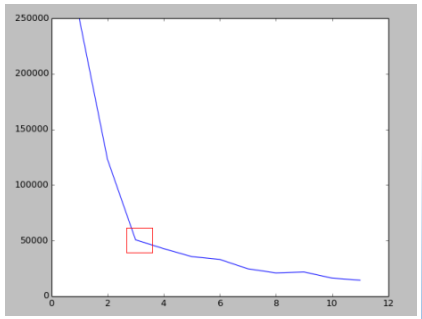
Introduction

In the real world, the house price is really a hot topic around the world. There are lots of features which can have impact on the sale price of a house. Hence, our team consider it is a good field to do data mining and analysis in order to guide real estate agents during decision making and pricing. Also, fundamentally, we can help house buyers to find the best house according to our analysis.
Even some big companies like Airbnb can use the prediction of the house prices to inform their hosts to better pricing their places.

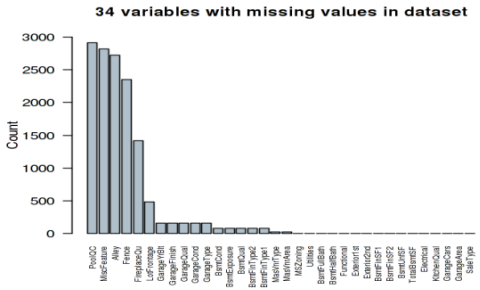
After discussion, our team choose the "House Prices: Advanced Regression Techniques" as our dataset, which is a good representative dataset from Kaggle platform.

Data Preprocess





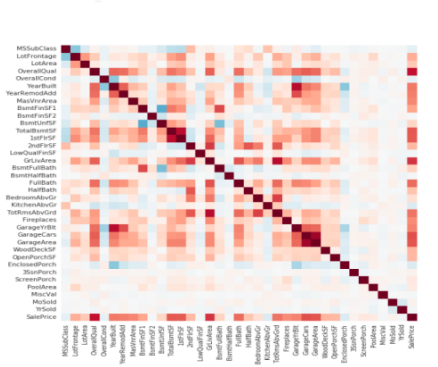
Missing Data



The missing value would influence the analysis result by acting as an improper feature that mislead the modeling process greatly and thus lead to a incorrect result. Therefore, we need to take care the missing value properly and try to fill out the missing value with proper and reasonable value.

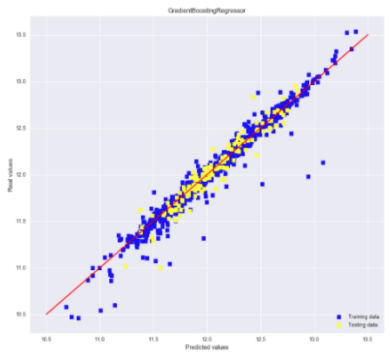
- Single imputation method
- Regression imputation
- Model based imputation

Correlation

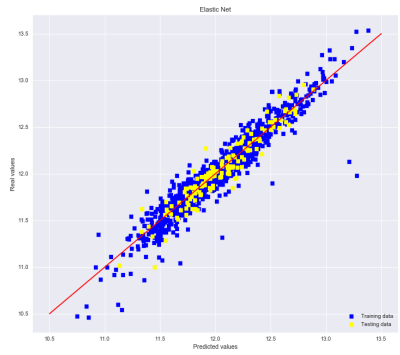


OverallQual:	0.79098	HalfBath:	0.28410
GrLivArea:	0.70862	LotArea:	0.26384
GarageCars:	0.64040	BsmtFullBath:	0.22712
GarageArea:	0.62343	BsmtUnfSF:	0.21447
TotalBsmtSF:	0.61358	BedroomAbvGr:	0.16821
1stFlrSF:	0.60585	KitchenAbvGr:	-0.13590
FullBath:	0.56066	EnclosedPorch:	-0.12857
TotRmsAbvGrd:	0.53372	ScreenPorch:	0.11144
YearBuilt:	0.52289	PoolArea:	0.09240
YearRemodAdd:	0.50710	MSSubClass:	-0.08428
GarageYrBlt:	0.48636	OverallCond:	-0.07785
MasVnrArea:	0.47749	MoSold:	0.04643
Fireplaces:	0.46692	3SsnPorch:	0.04458
BsmtFinSF1:	0.38641	YrSold:	-0.028922
LotFrontage:	0.35179	LowQualFinSF:	-0.02560
WoodDeckSF:	0.32441	MiscVal:	-0.02118
2ndFlrSF:	0.31933	BsmtHalfBath:	-0.01684
OpenPorchSF:	0.31585	BsmtFinSF2:	-0.01137

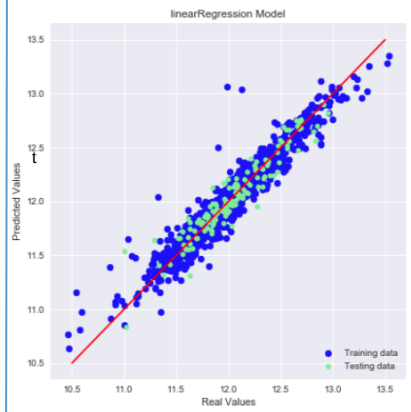
Method & Model



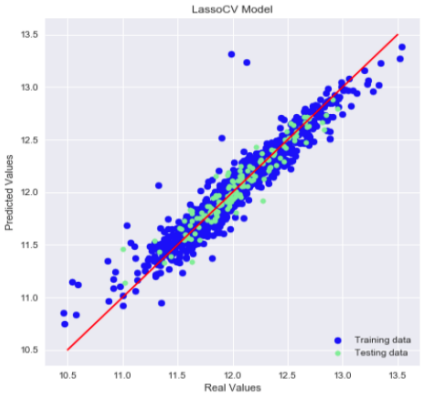
Gradient Boosting Regression



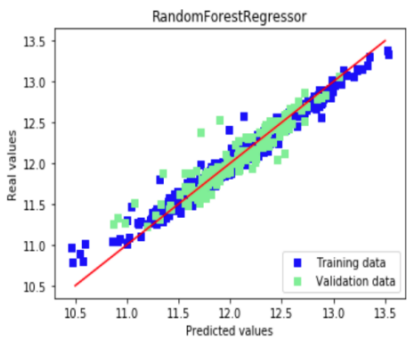
Elastic Net



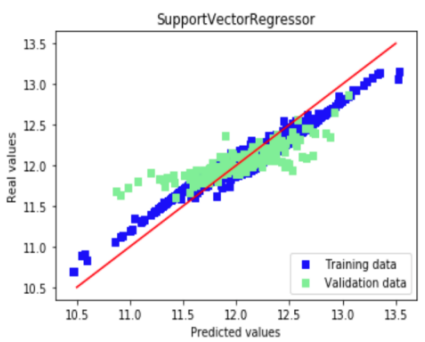
Linear Regression



Lasso CV

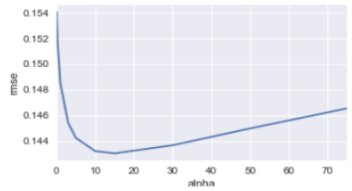


Random Forest Regression

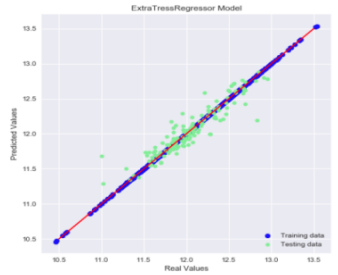


Support Vector Regression

Method & Model



We choose the alpha value which has least value of rmse. In this dataset, the alpha value is "10".



ExtraTreesRegressor Model

The cross validation score is 0.858874576346.

This model uses ensemble method which can combine the different models and reduce the bias.

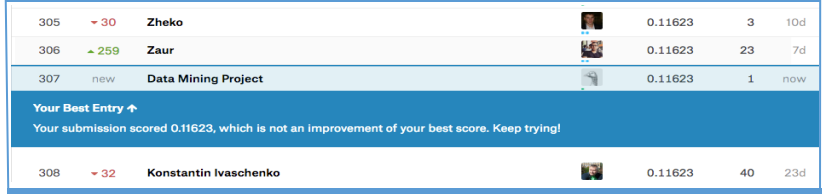
The figure shows the predicted values are close to real values.

Model Average

Model Averaging

10

Achievement



Summary

Through this project, we realized that a real-world data set which involves complex features is much more complicated than some general data sets like MNIST. Prior to performing regression training, there are lots of feature engineering jobs to do which definitely have large impact on our prediction model. Also the model selection is not pretty deterministic. We see from the experiments that lasso regression could easily get good performance but other approaches like random forest need to well tune hyper-parameters to get decent results.