House Prices: Advanced Regression Techniques

By Lesi Wei_a1797017

Faculty of Engineering, Computer and Mathematical Sciences, The University of Adelaide

Report submitted for COMP SCI 7209 Big Data Analysis and Project at the Faculty of Engineering, Computer and Mathematical Sciences, University of Adelaide towards the Master of Data Science

THE UNIVERSITY of ADELAIDE

SUB CRUCE LUMEN

Abstract

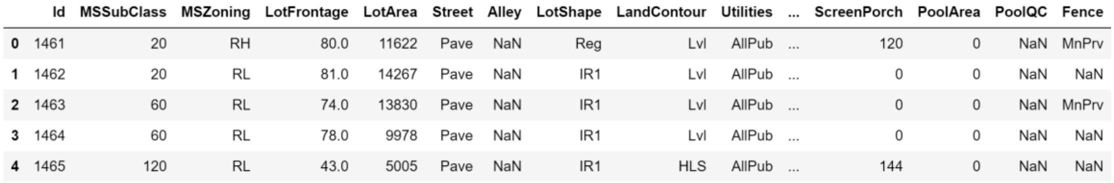
The prediction Root Mean Square Error for the final model is 0.11701. The final model is a combination of xgboost regressor, light gradient boosting regressor, ridge regression, random forest, and gradient boosting.

Introduction

For most people, buying a home is still one of the important purchase decisions that individuals make in their lives. Potential purchasers usually need years of saving to make a down payment and make regular payments for a period of thirty years or fifth years. Certain considerations like location, insurance, school and facilities surrounding, and home maintenance are additional costs related to home ownership. Also, what kind of house is worth more than others is advantageous for potential buyers, and what are some key aspects to determine the house price. Likewise, for existing homeowners seeking to sell and maximize investment, their decision-making process would greatly benefit if they could predict their home 's price before listing it.

House Price Dataset

The dataset is aims to predict sale prices of houses in Ames, Iowa by using various aspects of residential homes. The dataset is consisting of mainly categorical variables stored as integers and factors, and it is including 79 features and 1460 houses. (Kaggle,2020) Those features contain different type of data from categorical to numerical data, and it also have ordinal data in it. The process to able to predict the house price is first clean the data to a stage that is can be used for model building, then fit it with various model, then compare their error rate to see which one is the best model to predict the house price.



	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	ScreenPorch	PoolArea	PoolQC	Fence
0	1461	20	RH	80.0	11622	Pave	NaN	Reg	Lvl	AllPub	...	120	0	NaN	MnPrv
1	1462	20	RL	81.0	14267	Pave	NaN	IR1	Lvl	AllPub	...	0	0	NaN	NaN
2	1463	60	RL	74.0	13830	Pave	NaN	IR1	Lvl	AllPub	...	0	0	NaN	MnPrv
3	1464	60	RL	78.0	9978	Pave	NaN	IR1	Lvl	AllPub	...	0	0	NaN	NaN
4	1465	120	RL	43.0	5005	Pave	NaN	IR1	HLS	AllPub	...	144	0	NaN	NaN

Data Cleaning

Some of missing data in this dataset is due to some of the feature is not included in the certain house unit. Especially for the 'Alley','PoolQC','Fence','MiscFeature ' features, with more than 80% of data is missing, which indicated that more than 80% of house unit not have those features, so drop those features is appropriate for the sale price prediction. For other features that is missing value, using their mean, median, or mode to fill out the missing value which is a normal technique is being used. For example, ‘LotFrontage’ is missing 486 observations, after looking at its mean, median and model, using median is more appropriate in this case. There are some ordinal data categorical that indicates one of the rating for a house feature, such as Evaluates the quality of the material on the exterior, Evaluates the general condition of the basement, etc. So when the house do not have a such features, it will recorded as NA for the specific features.

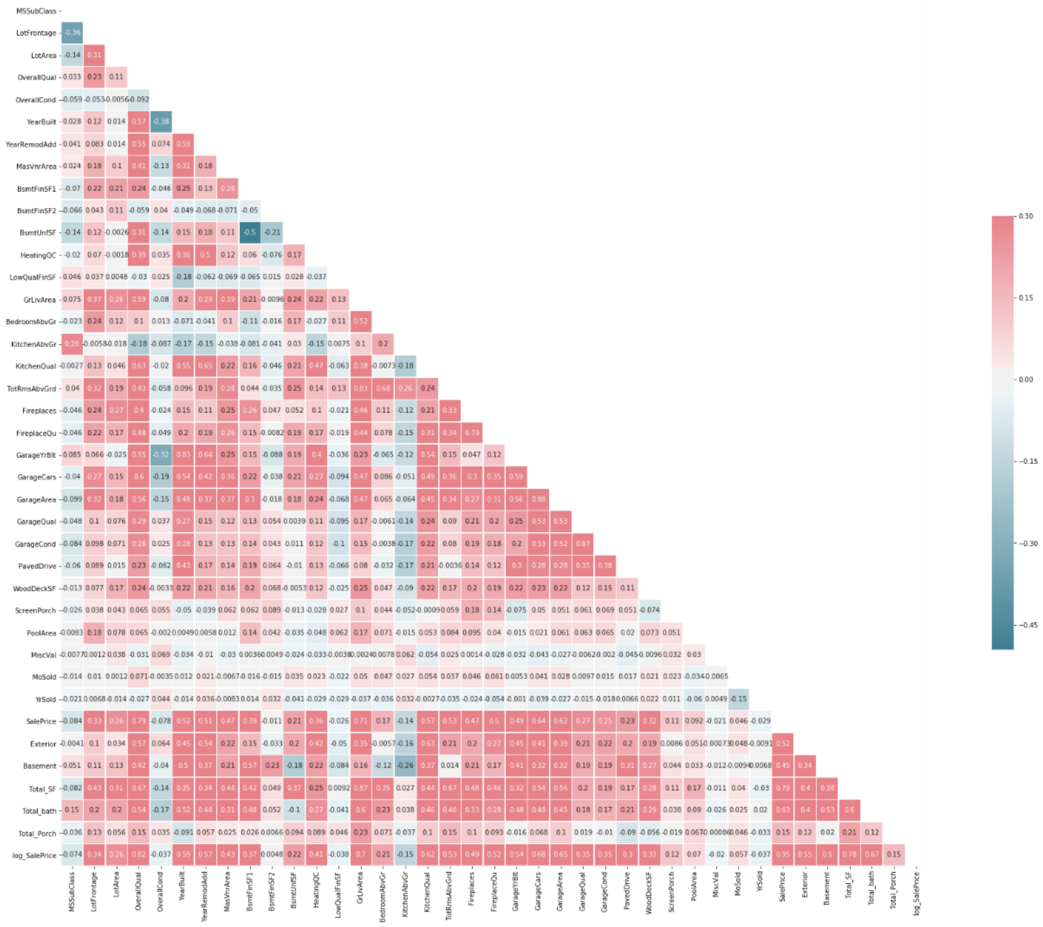
The basic approach to clean ordinal data is given ‘NA’ a text meaning and transform those rating in text into number which can be fit in to model in the later use. So a dictionary is be created, to include all the mapping attributes according to the ordinal rating variables. Then use the map function in pandas to map the descriptive rating into a numerical rating scale. And there are some features describe small feature of the house unit which it can be combine together as one feature to get bigger impact in the model building stage. For the rating of exterior, can be combine together as one rating to reflect the overall score of house unit's exterior, and the same thing can be done for the basement rating score. And the features also include first floor square feet, second floor square feet, basement square feet, which can be add up together as the total square feet for the house unit. Same thing can be done for the number of bathroom, and total porch size in the house unit.

Since there still a lot of categorical variables that is describe one of the feature or function of the house unit. Those variables can be hard to use when building the regression model, so transfer those categorical variables into dummy variables will help to build the regression model.

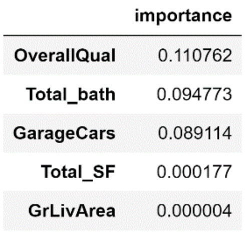
After look at the sale price distribution, the data is showing a right skew, it is indicated that a log transformation will be preform to normalized the response variable which is the sale price in this case.

Modeling and Evaluation

The model is built by different methods. A correlation matrix is being implemented to determine some of the features that is highly correlated to the log transformed sale price. From the correlation matrix, can concluded that OverallQual, GrLivArea,Total_SF,Total_bath are highly correlate to the



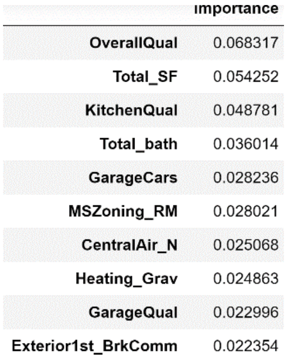
log-sale price, so a multiple linear regression model is implemented based on the variable above. After split the data into training set and testing set. The rooted mean square error of the multiple linear regression is 0.19155 and r square score is 0.75. Also look at what is contribute the most for the model by using their coefficients. Then several models are being built.



	importance
OverallQual	0.110762
Total_bath	0.094773
GarageCars	0.089114
Total_SF	0.000177
GrLivArea	0.000004

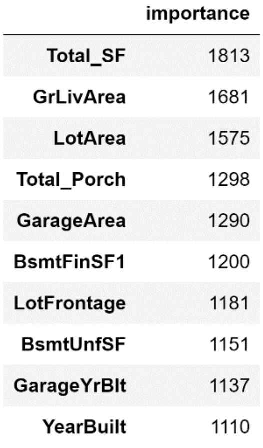
Extreme Gradient Boosting

The first one is an Extreme Gradient Boosting model which an alternative tree-based modeling method to predictive the sale price. XGBoost and Scikit- learn's built-in GridSearchCV are used. With top 10 features of the importance shown below and compare with the linear regression model is shown that OverallQual ,Total_SF and Total_bath are still important contributor to the EXG boost model. And it has a RMSE of 0.11768.



	importance
OverallQual	0.068317
Total_SF	0.054252
KitchenQual	0.048781
Total_bath	0.036014
GarageCars	0.028236
MSZoning_RM	0.028021
CentralAir_N	0.025068
Heating_Grav	0.024863
GarageQual	0.022996
Exterior1st_BrkComm	0.022354

Light Gradient Boosting



	importance
Total_SF	1813
GrLivArea	1681
LotArea	1575
Total_Porch	1298
GarageArea	1290
BsmtFinSF1	1200
LotFrontage	1181
BsmtUnfSF	1151
GarageYrBlt	1137
YearBuilt	1110

Light GBM is a fast, high-performance gradient boosting framework based on decision tree algorithm, it splits the tree leaf wise with the best fit whereas other boosting algorithms split the tree depth wise or level wise rather than leaf-wise. So when growing on the same leaf in Light GBM, the leaf-wise algorithm can reduce more loss than the level-wise algorithm and hence results in much better accuracy. Also, it is surprisingly very fast, compare to other boosting algorithms Look at the top 10 important features above, it is different compare to two model above, with the OverallQual drop out of top 10 features, and GrlivArea and Total_SF are still consider the top features in the model. And it has a RMSE of 0.127737, which is the lowest model so far.

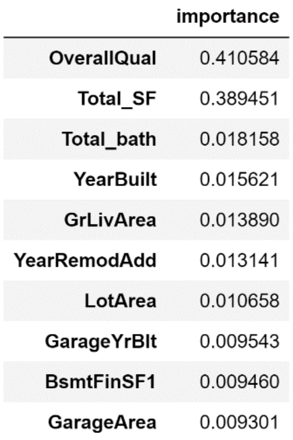
Ridge Regression

Regularized regression is explored using the Ridge method. The L2 penalty is applied and hyperparameter tuning of lambda is conducted. A various range of alphas is being selected to test the model. After fit the model and calculated the RMSE is 0.160803.

Random Forest

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. It is a methods that combines multiple predictions from different algorithms to predict more accurate prediction. The structure of a

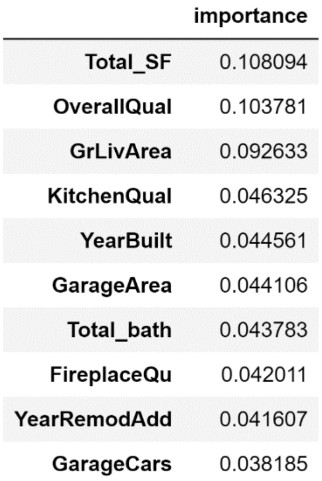
Random Forest is running the parallel tree with no interaction with each other, and as a result output the mean of all the tree predictions. And the top 10 features is shown on the right side, compare with other models, OverallQual, Total_SF, Total_bath GrLivArea are still the top contributor for the model. After model fit, the RMSE is 0.1464770.



	importance
OverallQual	0.410584
Total_SF	0.389451
Total_bath	0.018158
YearBuilt	0.015621
GrLivArea	0.013890
YearRemodAdd	0.013141
LotArea	0.010658
GarageYrBlt	0.009543
BsmtFinSF1	0.009460
GarageArea	0.009301

Gradient Boosting

Gradient Boosting is a technique for changing over weak learners into strong learners. In boosting, each new tree is a fit on a changed form of the original data set, and Subsequent trees help classify observations that are not all around arranged by the previous trees. Expectations of the final model is thusly the weighted aggregate of the prediction made by the previous tree models. The top 10 features of the Gradient Boosting model is shown on the left side, which is still shown that Total_SF,OverallQual, GrLivArea being the top influence for the gradient boosting model. And the RMSE for the model is 0.124656. It is the most accurate model so far of all the model.

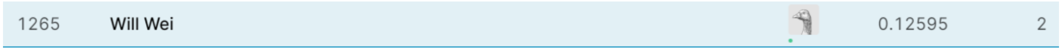


	importance
Total_SF	0.108094
OverallQual	0.103781
GrLivArea	0.092633
KitchenQual	0.046325
YearBuilt	0.044561
GarageArea	0.044106
Total_bath	0.043783
FireplaceQu	0.042011
YearRemodAdd	0.041607
GarageCars	0.038185

Stack Model

This is a procedure to consolidate the best of the best of multiple algorithms which can give more steady forecasts with less change than what we get with a solitary regressor. The StackingCVRegressor is one such algorithm that permits to altogether utilize numerous regressors to predict. In general, the stack model will preform better than the single model, it will combine all the model built above to predict the sale price. After fitting the model, the RMSE is 0.11701, which is the best model overall. So the stack Model is consider the final model to this house price prediction.

After submitting the final prediction model into Kaggle, it shown that the score is 0.12595, which is the root mean square error, and the ranking is top 25%.



1265	Will Wei	0.12595	2

Reference

Bakshi, C. (2020, June 09). Random Forest Regression. Retrieved August 26, 2020, from https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

House Prices: Advanced Regression Techniques. (n.d.). Retrieved August 26, 2020, from https://www.kaggle.com/c/house-prices-advanced-regression-techniques

KhandelwalPranjal, P. (2020, March 27). Light GBM vs XGBOOST: Which algorithm takes the crown. Retrieved August 26, 2020, from https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/