House Prices: Advanced Regression Techniques

Introduction to Machine Learning Project

Zihan Zhao RIN: 661983787 Section 01

Lally School of Management, Rensselaer Polytechnic Institute, Troy, NY

Abstract

The project is a regression problem that using the provided data of comprehensive aspects of a house to predict the house price. In this project, three benchmark kernel notebooks are chosen to discuss about their feature approaches and model approaches. Then the basic steps of exploratory data analysis are conducted in detail, including general information on the meaning of data, the missing value, the skewness of the features and target, and the correlations among features. In the modeling section, random forest is used to select features according to their importance, and the most two important features are 'Overall Quality' and 'GrLivArea', which are also two features having highest correlation with target variable. This project also compares the performance of various models trained relatively by all features and selected features. The results imply that the lasso regression model trained by all features has the best performance. And the kernel ridge regression model performs worst as this house price predicting project tends to be a linear problem. Moreover, using the selected features to train the model, though increasing the speed of model training and reducing the computation complexity, fails to contribute to the performance of models. Therefore, better methods for feature selection are worth to discuss to improve the performance of model.

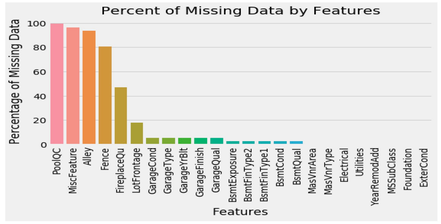
Benchmark Kernel Notebooks

Kernel Name	Feature Approach	Model Approach	Train/Test Performance
A Detailed Regression Guide with House-pricing	1.Filling missing value of numerical features with mode. 2.Encoding categorical features and creating dummy features. 3.Fixing the skewness of all data using boxcox transformation. 4.Sorting the correlation between each feature and target feature. 5.Dropping useless features and creating new features by combination.	Linear Regression model Regularization models: Lasso, Ridge, Elastic Net GBR LightGBM XgBoost Stacked model of Lasso, Ridge, Elastic Net, XgBoost, LightGBM Blended model of all previous model	RMSE score of train dataset: 0.06332 RMSE score of test dataset: 0.10583
Regularized Linear Models	1.Filling missing value of numeric features with mean of the columns. 2.Fixing the skewed numeric data by taking log transformation. 3.Creating dummy variables for categorical features. 4.Ranking the coefficient of each variable in the Lasso models.	Lasso Ridge XgBoost	RMSE score of test dataset: 0.12096
#1 House Prices Solution [top 1%]	1.Using PCA and Kmeans to deduct the data dimensions and divide the features into 5 clusters. 2.Filling missing value of numerical features with 0 or median grouped by other features. 3.Filling missing values with "None" and creating dummy features for categorical features. 4.Fixing the skewness of all data using boxcox transformation. Eliminating outliers. 5.Dropping useless features and creating new features by combination.	Linear Regression model Regularization models: Lasso, Ridge, Elastic Net GBR LightGBM XgBoost Stacked model of Lasso, Ridge, Elastic Net, XgBoost, LightGBM Blended model of all previous model	RMSE score of train dataset: 0.054402 RMSE score of test dataset: 0.10649

Exploratory Data Analysis

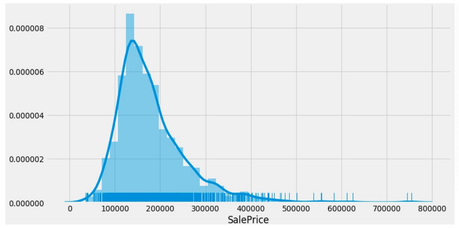
Training Dataset Description:
1460 entities, and 79 features with 35 numeric data and 44 categorical data. Basically, there are three segments of features: building, space, location. Building features describe the physical characteristic of the house including overall qualities, house style, overall condition. Space features presents the space properties of the house such as first floor square feet, total basement square feet. And location features describe the surrounding environment of the house including neighborhood, street, alley.

Missing Values:

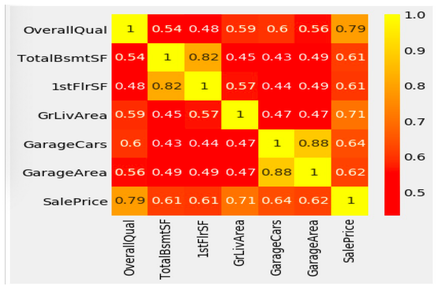


PoolQC has 1432 missing values with a percentage of 99.52, as a result, this feature is supposed to be dropped. In addition, MiscFeature, Alley, Fence also has high percentages of missing value.

Skewness of Target Feature:
In order to create a more fitted linear regression model, it is of importance that response variable SalePrice is multivariate normally distributed. The following visualization present the distribution of target feature SalePrice, whch is not normally distributed, with skewness of 1.88 and kurtosis of 6.54.



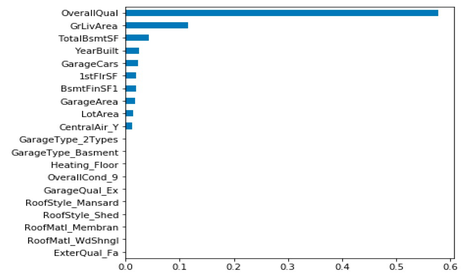
Correlation among Features:



The graph presents the pairs with correlation of 0.6 and above. We can find strong correlations do exist between several pairs of features such as OverallQual and GrLivArea. Therefore, we may reduce the dimension of data by dropping one of featues in the highly correlated pairs. The target feature SalePrice and feature OverallQual have the highest correlation of 0.79, which implies the strong positive linear relationship.

Modeling Results

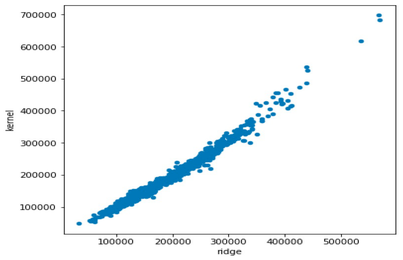
Random Forest:
Considering that the features are not independent from each other and highly correlated pairs may be dropped, random forest model is employed to indicate how important each feature is in the decision trees. According to the result, the importance of 40 variables are assigned to 0 and random forest selects only 18 features, which are 'LotFrontage', 'LotArea', 'OverallQual', 'YearBuilt', 'YearRemodAdd', 'BsmtFinSF1', 'BsmtUnfSF', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF', 'GrLivArea', 'FullBath', 'TotRmsAbvGrd', 'GarageYrBlt', 'GarageCars', 'GarageArea', 'MSZoning_RM', 'CentralAir_Y'. The following graph presents the top 10 and last 10 important features selected by random forest. As we can see that the features 'Overall Quality' and 'GrLivArea' are the two most important features, which is correspond to the finding of the correlation heatmap.



Lasso Regression:
The cross-validation rmse error mean of lasso model on train dataset is 0.111 with the standard deviation of 0.0144. The rmse mean of lasso model trained by selected 18 features is 0.1367 with the standard deviation of 0.0128. The increase of rmse error may because lasso itself filters the unimportant features and reduces the complexity of training already.

Ridge Regression:
The ridge model has the cross-validation rmse error mean on train dataset of 0.1124 with the standard deviation of 0.0141, The results of model trained by selected features are 0.1367 with standard deviation of 0.0117. Though ridge does not select features, it applied penalty to the sum of squares of all the regression coefficient, which help assure the generalization of model.

Kernel Ridge Regression:
The kernel ridge has the cross-validation rmse error mean on train dataset of 0.1223 with the standard deviation of 0.0105, and the rmse mean of model trained by selected features is 0.1356 with the standard deviation of 0.0117. Kernel ridge performs worst among the four models in this project.
In addition, the linear relationship of the predictions of two models are visualized. According to the graph, the predictions of two models are accordant for most samples and outliers are few.



Gradient Boosting Regression:
Gradient Boosting has the cross-validation rmse error mean on train dataset of 0.1141 with the standard deviation of 0.0116, and the rmse mean and standard deviation of Gradient Boosting model trained by selected features are relatively 0.1396 and 0.0127.

Feature Engineering

Before training different models, feature engineering plays a role in determining the optimal performance of models.

- All the skewed numeric data are fixed by taking log transformation or boxcox transformation as normally distributed variables contribute to make linear regression model more fitted.
- Moreover, Outliers, which has large living areas but a very low prices, and useless feature 'Utilities', which is same for all samples, are dropped.
- Numeric features 'MSSubClass', 'OverallCond', 'YrSold' and 'MoSold' are transferred to categorical features according to their realistic meanings.
- Finally, missing values of categorical features are replaced with 'None' with the meaning 'do not have'. Missing values of numeric features related to area are assigned to 0 to mean 'do not have'. Missing values of features 'LotFrontage' are replaced with median group by feature 'Neighborhood'. And missing values of rest features are replaced with mode. As a result, the skewness and missing values of features are well fixed to guarantee the later training of models.

Conclusion

- Through reviewing the benchmark kernel notebooks, specifically, comparing the second notebook, which has the worst score, with the other two kernels, we can summarize that a better prediction needs detailed feature engineering including dealing with the missing values, outliers and useless or correlated features. In addition, trying different models and different combination of stacking and blending is also of significance to build better models
- After comparing the performance of different models trained by all features and selected features, lasso regression has the best performance with rmse mean of 0.1111 and kernel ridge has the largest rmse mean of 0.1223 as this is a linear regression problem.
- Moreover, the features selected by random forest seems to have no help in improving the performance of models especially for lasso regression model. This may because lasso filters again the important features which are already selected out by random forest and affects the performance of prediction. However, using the selected features do increases the speed of model training and reduces the computation complexity. Therefore, better methods for feature selection are worth to discuss to improve the performance of model and consider the risk of overfitting at the same time.

Appendix

https://github.com/rpi-intro-ml-app-fall-2019/final-project-zihan97.git