Stat542 Project1

Jingyi Zhu UIN: 653703610

Xi Chen UIN: 659116404

1 Introduction

The project is aimed to predict housing price based on analyzing the housing data collected on residential properties sold in Ames, Iowa between 2006 and 2010. The dataset has 2930 rows (i.e., houses) and 83 columns. Column 1 is "PID", the Parcel identification number, the last column is the response variable, "Sale_Price", and the remaining 81 columns are explanatory variables describing (almost) every aspect of residential homes.

2 Technical details

2.1 Preparation

Before building models, we do some preparation. First, we separate test data and train data, as well as X and y in both of them. Second, we should fix the missing data from "Garage_Yr_Blt", which has 159 missing values, and these values are corresponding to have no garage. So, it only refers to "No_Garage" and "Garage_Type" tightly. Then we replace missing value by 0.
We will use the third training/test splits to select the best model of minimum RSME, then use 10 set of splits to fit that best model.

2.2 Model 1: BOOSTING TREE MODEL

We make sure X matrix is a numeric matrix by transfer categorical data into numerical data and take away “PIN” column before fit the model. The 10 sets of RMSE of boosting tree model and their accuracy are shown as below.
Result:

RMSE of boosting tree model

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
0.1190124	0.1206938	0.116919	0.1170635	0.1134271
Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
0.1333822	0.134596	0.1307949	0.1343734	0.1300949

Where sample 1，…， 10 corresponding to j=1，…， 10 in the R code represent 10 tests.
Accuracy:

Accuracy of boosting tree model

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
0.9520992	0.9655504	0.935352	0.936508	0.9074168
Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
0.9880163	0.9970074	0.9688511	0.9953585	0.9636659

the mean of 10 sets accuracy is 0.9609826.

2.3 Model 2: LASSO REGRESION MODEL

First, we remove some unnecessary variables based on experience. For example, due to very close area, the variables “Longitude” and “Latitude” can be dropped, and some variables, such as “Pool_Area”, are dropped because of specific values. Then, we transfer value from numeric to categoric data for both “Mo_Sold” and “Year_Sold”. Third, we take off outliers for numerical variables, where we assume 95% of the train column as the upper bound. Last, we fit the lasso model. The 10 sets of RMSE of linear model with lasso and their accuracy are shown as below.

Result:

RMSE of linear model

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
0.1227094	0.117493	0.121668	0.1298056	0.1124019
Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
0.1325256	0.1262792	0.120432	0.1298837	0.1237266

Accuracy:

Accuracy of linear model

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
0.9816752	0.939944	0.973344	1.038445	0.8992152
Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
0.9816711	0.9354015	0.8920889	0.9621015	0.9164933

the mean of these accuracy is 0.952038.

2.4 Model 3: GAM MODEL

To avoid curse of dimensionality when there are too many variables, we first remove some variables as below: 'Street', 'Utilities', 'Condition_2', 'Roof_Matl', 'Heating', 'Pool_QC', 'Misc_Feature', 'Low_Qual_Fin_SF', 'Pool_Area', 'Longitude','Latitude', 'Mo_Sold', 'Year_Sold', 'PID', 'Sale_Price'. Second, we take some numerical variables below to be linear terms: 'BsmtFin_SF_1', 'Bsmt_Full_Bath', 'Bsmt_Half_Bath', 'Full_Bath', 'Half_Bath', 'Bedroom_AbvGr', 'Kitchen_AbvGr', 'Fireplaces', 'Garage_Cars', which only take specific values not enough to fit a nonparametric curve. Also, for other numerical variables, we get the list of numerical variables whose nonlinear terms will be included in the gam model. Third, we generate select binary variables referring to lasso result. As we have three parts of features, we fit the gam model. The 10 sets of RMSE of GAM model and their accuracy are shown as below.

Result:

RMSE of GAM model

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
0.1200926	0.1294411	0.108904	0.1195759	0.113384

Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
0.1256247	0.1440696	0.121606	0.1311468	0.1228574

Accuracy:

Accuracy of GAM model

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
0.9607408	1.035529	0.871232	0.9566072	0.907072
Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
0.9305533	1.067182	0.997852	0.9714578	0.9100548

the mean of these accuracy is 0.9608281.

3 Findings

1. Data cleaning is very important, and we need to think about the meaning of missing value as well as the relationships with other variables;
2. We need to well know about each variable first, to determine it is a numerical variable or a categorical variable in different models;
3. From the result above, we find it very surprising that the lasso linear model has the smallest RMSE, which means it is the best model among others to predict housing price;
4. To adjust RMSE in GAM model, we find by adding select binary variables, test set RSME will decrease from about 0.14 to 0.011.

4 Conclusion

Comparing the RMSE of three models, we find boosting tree model is the best with high accuracy and the RMSE within the thresholds as required (RMSE of the first five samples should lower than 0.125, while RMSE of the last five samples should lower than 0.135).

5 Other information

- Computer system: MacBook Pro, 2.53 GHz, 4GB memory
- Running time of code: 72.510s