STAT 542 Project 1 Report

Bin Feng

1 INTRODUCTION

In this project, we are asked to analyze the housing data collected on residential properties sold in Ames, Iowa between 2006 and 2010. The goal is to predict the final price of a home (in log scale) with those explanatory variables. Three prediction models were built to achieve the goal: 1) linear regression model based on LASSO; 2) boosting tree model; 3) generalized additive model (GAM).

Four R code files were developed as follow:

1) mymain.R (serve as the main file to source and run the three prediction models and write predictions to txt submission files);
2) DataPrepLASSO.R (sub file to prepare data for LASSO regression);
3) DataPrepBoost.R (sub file to prepare data for boosting tree regression);
4) DataPrepGAM.R (sub file to prepare data for GAM regression).

Prediction evaluation is based on Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted price and the logarithm of the observed sales price. Prediction performance is based on the minimal RMSE from the three prediction models. Full credit with be given to submissions with minimal RMSE less than:

1) 0.125 for the first 5 training/test splits and;
2) 0.135 for the remaining 5 training/test splits.

The reason for different RMSE threshold between the first 5 training/test splits and the remaining 5 training/test splits is that one extreme house case (No. 1554) is added to the last 5 splits, which can have a significant negative influence on the overall prediction performance.

2 DATA PREPARATION

2.1 LASSO

The Ames House data contains both numerical and categorical information. Some properties (e.g. Overall_Qual) contain essential information for the prediction while others (e.g. PID) have limited or even negative effect on the performance. To arrange the data for LASSO regression, following preparations were conducted:

1) NA values were detected and replaced with 0
2) Some variables with negative effects were dropped
3) Winsorization was conducted at the threshold of 95% to treat extreme values
4) Dummy coding was performed for categorical variables

2.2 Boosting Tree

The Ames House data contains both numerical and categorical information. Some properties (e.g. Overall_Qual) contain essential information for the prediction while others (e.g. PID) have

limited or even negative effect on the performance. To arrange the data for boosting tree, following preparations were conducted:

1) NA values were detected and replaced with 0
2) Extreme values were dropped from training dataset
3) Some categorical variables behave as rankings were coded into numerical variables
4) Some variables with negative effects were dropped
5) Dummy coding was performed for the rest categorical variables

2.3 Generalized Addictive Model (GAM)

The Ames House data contains both numerical and categorical information. Some properties (e.g. Overall_Qual) contain essential information for the prediction while others (e.g. PID) have limited or even negative effect on the performance. To arrange the data for GAM, following preparations were conducted:

1) NA values were detected and replaced with 0
2) Extreme values were dropped from training dataset
3) Some variables with negative effects were dropped
4) Some numerical variables were selected as linear variables, meaning only linear term of these variables will be kept in the prediction. This is because they only take a handful of unique values, which are not enough to fit a nonparametric curve.
5) The rest numerical variables were selected and will fit a nonparametric curve in the prediction.
6) Some categorical variable levels were selected based on LASSO regression with the "1se" lambda value.

3 MODEL RUNNING TIME

All three prediction models were run a MacBook Pro. Detailed computer configuration are as follow:

macOS Mojave (version 10.14.3), 2.6 GHz Intel Core i7, 16 GB 2400 MHz DDR4, Radeon Pro 560X 4096 MB, Intel UHD Graphics 630 1536 MB.

Running time for all three prediction models are summarized in Table 1 as below:

TABLE 1 Running time for three prediction models

Dataset No.	LASSO/s	Boosting Tree/s	GAM/s
1	**1.590382**	42.22115	47.76227
2	**1.492658**	49.85570	44.09633
3	**1.193442**	52.45167	26.27821
4	**1.454671**	53.14427	25.60333
5	**1.244313**	51.90762	27.08329
6	**1.131448**	52.35834	46.17251
7	**1.379873**	51.01782	47.10455
8	**1.218985**	50.07540	42.86314
9	**1.545740**	51.26410	25.59622
10	**1.165624**	50.97464	24.86547

Note that running time for LASSO is much smaller than the ones for Boosting Tree and GAM. This is because LASSO is a global regression model, meaning parameters are calculated for the whole dataset rather than small intervals between nodes. Boosting Tree requires such long time is because 10,000 trees were built for each dataset and predictions were evaluated based on all these trees. GAM needs such long time is because is perform spline regression (which needs to determine parameters locally within each nodes) on every variable selected.

4 MODEL PREDICTION ACCURACY

Model prediction accuracy (RMSE) is calculated based on the evaluation metric provided in the introduction section. Prediction accuracy for all three models are summarized in Table 2 as below:

TABLE 2 Prediction accuracy (RMSE) for three prediction models

Dataset No.	LASSO	Boosting Tree	GAM
1	0.1276809	**0.1192998**	0.1655591
2	0.1205030	**0.1198989**	0.1347752
3	0.1420143	**0.1195387**	0.1265634
4	0.1386641	**0.1160097**	0.1666590
5	0.1201118	**0.1106235**	0.1373261
6	0.1331143	**0.1293919**	0.1685218
7	**0.1289814**	0.1296621	0.1453750
8	**0.1212651**	0.1267216	0.1377239
9	0.1349361	**0.1287774**	0.1695683
10	0.1249204	**0.1200348**	0.1432870

For the 10 datasets, we see that the minimum RMSEs from three prediction models are all below the required thresholds. Also note that in most cases, Boosting Tree has the best prediction performance among the three, especially in the first 5 datasets.

We know that the remaining 5 datasets has an extreme case that will significantly influence the prediction performance. We notice that LASSO may have a better performance because of winsorization and global regression.

GAM always has the worst performance in my prediction while in Professor's example, it has the best prediction accuracy. Such discrepancy may be due to the differences in the choice of categorical variables. My model uses variables selected by LASSO with "1se" lambda which will results in a more concise and stable model. If selecting variables based on LASSO with "min" lambda, a better performance may be achieved.