Introduction

Facing with prevail violence nowadays, people are very concerned about their safeties. The motivation for our project is to explore the relationship between community safety and other covariates, trying to find out key factors which could help predict community safety. We hope our findings would offer some help for people in selecting their houses.

The data set used in the project is the Boston Housing Data from https://www.kaggle.com/jamieleech/boston-housing-dataset. There are 506 rows and 14 columns in the data set. This data set contains 1 response variable CRIM (per capita crime rate by town), 1 binary predictor variable CHAS (Charles River dummy variable, with 1 indicating tract bounds river) and 12 continuous predictor variables: RAD (index of accessibility to radial highways), ZN (proportion of residential land zoned for lots over 25,000 sq.ft.), PTRATIO (pupil-teacher ratio by town), TAX (full-value property-tax rate per $10,000), INDUS (proportion of non-retail business acres per town), NOX (nitric oxides concentration, parts per 10 million), MEDV (Median value of owner-occupied homes in $1000's), B (1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town), AGE (proportion of owner-occupied units built prior to 1940), DIS (weighted distances to five Boston employment centers), LSTAT (% lower status of the population), and RM (average number of rooms per dwelling).

With this data set, we are trying to answer some important questions. Could these 13 independent variables help us predict the crime rate? Which are the essential covariates in predicting community safety?

Exploratory Analysis/Visualization

After downloading the Boston Housing Data from the internet, we firstly standardized the variable names. Secondly, we checked the data types for each variable and converted the CHAS variable from numeric type to factor type. Finally, we also looked for missing values, and Figure 1 showed that there was no missing value in this data set.

With the data prepared for further analysis, we split the whole data set into two parts by randomly selecting 75% observations to be the training data, and the remaining 25% observations to be the test data.

In addition, we drew feature plot to explore the relationship between our response variable CRIM and all 13 predictor variables. Figure 2 showed that there may be some potential linear association between CRIM and MEDV, DIS, and AGE.

Models

In order to solve the problem, we used the caret R package to build 7 models: linear regression, ridge regression, least absolute shrinkage and selection operator model, elastic net model, principle components regression, generalized additive model, and multivariate adaptive regression splines model. We included all 13 predictor variables: CHAS, RAD, ZN, PTRATIO,

TAX, INDUS, NOX, MEDV, B, AGE, DIS, LSTAT, and RM to fully make use of all the information contained in this data set.

One important assumption of linear regression, ridge regression, lasso, elastic net, and pcr models is that the relationship between the response and predictors variables is linear. However, instead of making this linearity assumption, the gam and mars models are more flexible, which allow either linear or nonlinear relationship between the response variable and covariates. In addition, the linear regression also assumes non-correlation between predictor variables.

During model training, we applied the repeated 10-fold cross-validation to the training data and picked the optimal tuning parameter values which minimized the cross-validation error. With the selected optimal tuning parameters from cross-validation, we re-fitted our models using the whole training data to get the final models.

After fitting the 7 models, we used the resamples function to compare the performance of these models. According to the cross-validation, all models had similar performance, which showed close RMSE results. Although Figure 3 showed that the lm, pcr, ridge, lasso, and elastic net models had slightly smaller Median RMSE values, however, Table 1 also showed that the gam and mars models had slightly smaller Mean RMSE values. The lm model and the pcr model had the highest Mean RMSE values, which were greater than 5.99. Moreover, the Mean RMSE values of the ridge, lasso, elastic net, and gam models were all exceeding 5.85. In addition, the mars model showed the lowest Mean RMSE value, which was 5.58.

By further comparing the R-squared values, Figure 4 and Table 1 showed that all models also had similar results. The lm, pcr, lasso, ridge, and elastic net models had the smallest Mean R-squared values, which were all smaller than 0.58. Furthermore, the Mean R-squared value of the gam model was 0.63. Finally the mars model showed the highest Mean R-squared value, which was 0.64. Therefore, the mars model was the optimal one among the 7 competing models, because it had the lowest Mean RMSE value and the highest Mean R-squared value.

According to the results of linear regression, MEDV, RAD, DIS, and ZN were significant predictors of CRIM at 5% level. With respect to the lasso and elastic net models, RAD, MEDV, LSTAT, DIS, and B were the most important predictor variables. Regarding our final chosen mars model, MEDV and RAD played important roles in predicting the response.

Let’s take a look at the limitations of these models. The most important limitation of the linear regression, ridge regression, lasso, elastic net, and pcr models is that they are not suitable when the relationship between the response and predictor variables is actually nonlinear, because they are not flexible enough to capture the underlying truth.

Another crucial limitation of linear regression and ridge regression is that if there are some redundant or irrelevant predictor variables, these two models cannot do variable selection. Although the pcr model can do dimension reduction, however, it has an important limitation that the principle components may not be highly predictive of the response variable because the response variable does not supervise the identification of the principal components. In addition, the linear regression is also not suitable if some predictor variables are highly correlated.

One important limitation of gam and mars models is that the computation is much more complicated and it is much slower to train these models. In addition, the caret package does not allow interaction terms for gam model implementation, which may make the model not flexible enough.

Conclusions

Comparing the cross-validation performance, we finally chose mars model as the optimal one. In addition, it also had the smallest mean squared error for the test data compared with other models, which further confirmed our choice. According to the result of mars model, MEDV (Median value of owner-occupied homes in $1000's) and RAD (index of accessibility to radial highways) were most important predictor variables of the response variable CRIM (per capita crime rate by town). We could see that MEDV was negatively associated with CRIM, which was as expected. One possible explanation could be that communities with higher median home values may had better security facilities, which led to lower crime rate. In addition, RAD was positively associated with CRIM, which was also as expected. It could be interpreted as better accessibility to radial highways would allow the criminals to flee more easily. Therefore, it could be a potential incentive for crimes.

Appendices

Figure 1. Missing Values

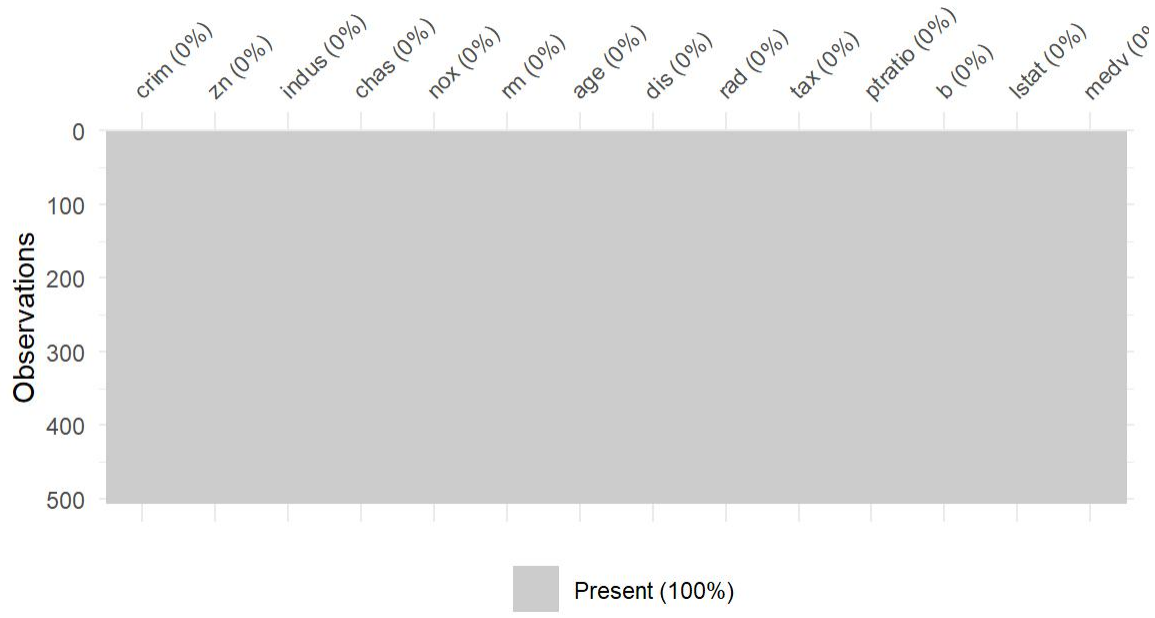

Figure 2. Feature Plot

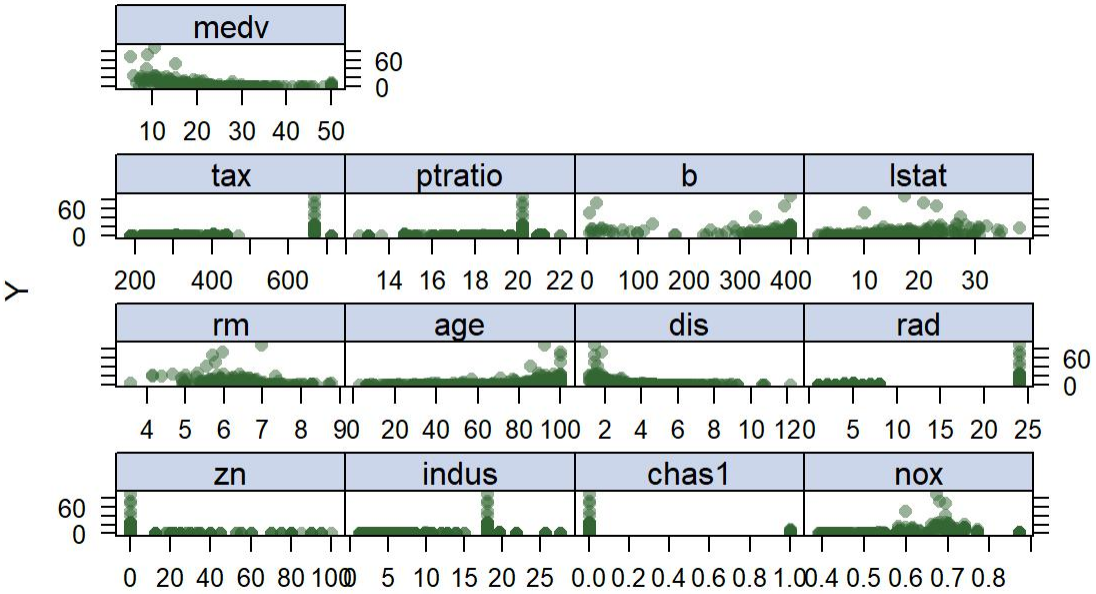

Figure 3. Median RMSE Values

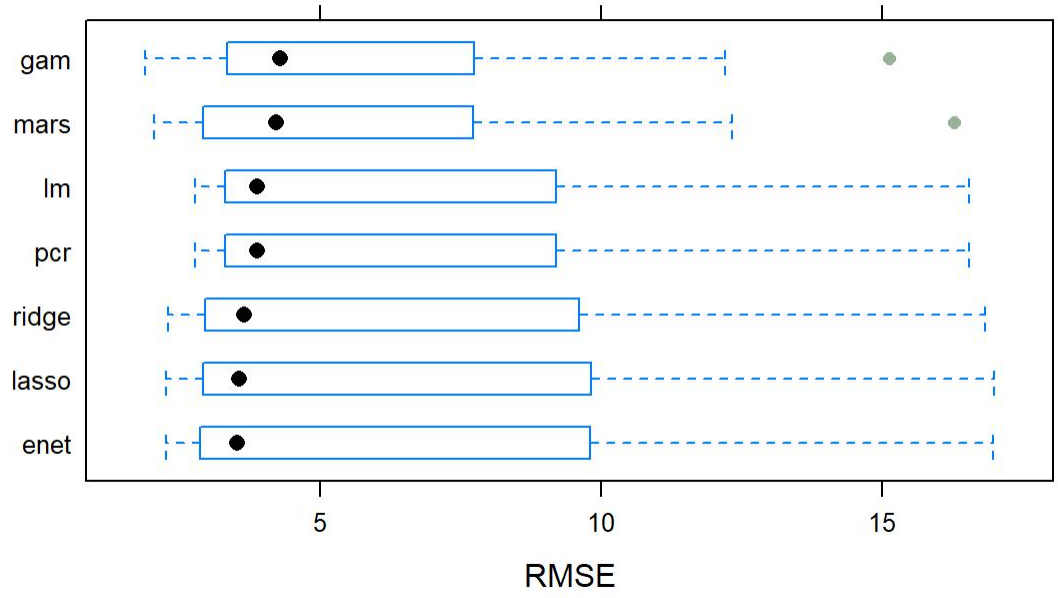

Figure 4. Median R-squared Values

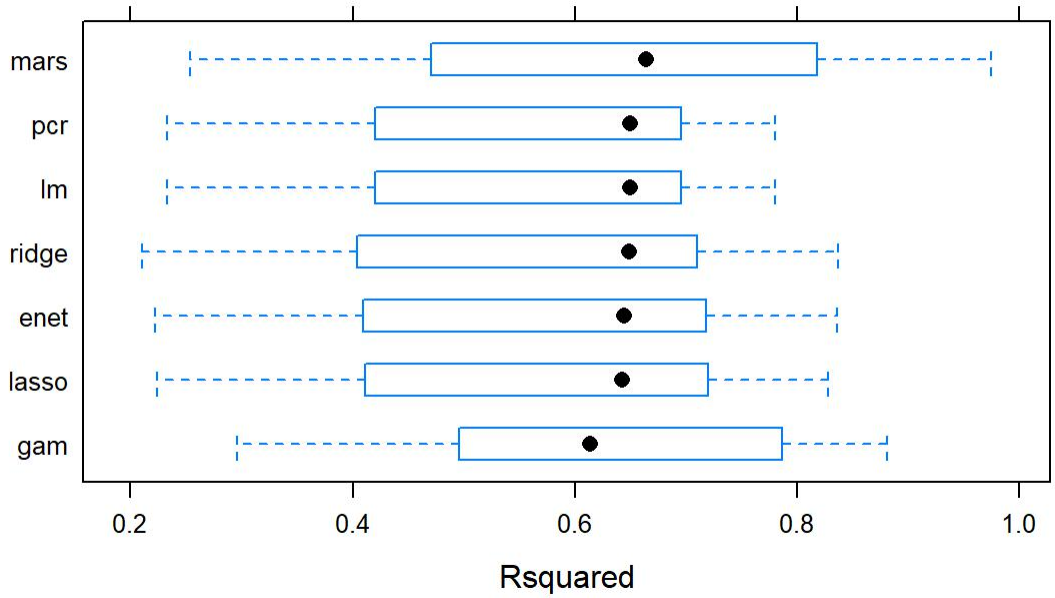

Table 1. Model Comparison by Cross Validation

RMSE

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
lm	2.781304	3.321123	3.888241	5.999115	8.745633	16.54133	0
ridge	2.306043	2.991830	3.654703	5.893753	9.047621	16.82551	0
lasso	2.266606	2.924485	3.566532	5.879788	9.243247	16.98289	0
enet	2.262869	2.878359	3.535544	5.871994	9.221960	16.96686	0
pcr	2.781304	3.321123	3.888241	5.999115	8.745633	16.54133	0
gam	1.897762	3.390709	4.290127	5.855060	7.588959	15.12059	0
mars	2.048382	2.941739	4.231846	5.582078	7.666572	16.28212	0

Rsquared

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
lm	0.2330454	0.4220000	0.6491904	0.5669691	0.6911455	0.7800115	0
ridge	0.2107030	0.4100340	0.6486207	0.5696873	0.7081521	0.8370814	0
lasso	0.2237487	0.4177951	0.6419694	0.5694642	0.7163039	0.8281398	0
enet	0.2222461	0.4151329	0.6440967	0.5706420	0.7140506	0.8354291	0
pcr	0.2330454	0.4220000	0.6491904	0.5669691	0.6911455	0.7800115	0
gam	0.2957810	0.4982447	0.6132422	0.6259010	0.7794797	0.8810599	0
mars	0.2541778	0.4722400	0.6640774	0.6442416	0.8096733	0.9742654	0