Prediction of sale price for housing

Ruoyuan Qian

1 Introduction

1.1 Objective

In this report, the focus is among eight methods (multiple linear regression, ridge regression, lasso regression, elastic regression, principal component regression (PCR), k-nearest neighbors algorithm (k-NN), generalized additive model (GAM), multivariate adaptive regression spline (MARS)), which one is the best to predict sale price of house for the particular dataset and which predictors are most influential for the response SalePrice.

1.2 Data cleaning

Missing value is checked for each predictor and predictors with the number of missing value greater than 500 are excluded from the dataset. At the meantime, predictors with many zeros or near-zero observations are removed as well. Then NA's are dropped from the remaining data.

2 Exploratory data analysis (EDA)

The distribution of response SalePrice ($) is checked (Fig. 1), as we can see, it is continuous variable with a right skewed shape. Since all methods in report do not need normal distribution assumption, so the original value of response can be used in model fitting.

Scatter plots are checked for numeric variables (Fig. 2), since I treated integers as continuous variables, so gaps are introduced in some of the scatter plots, such as "GarageCars", "MoSold", "YrSold", "BsmtFullBath". There is a non-linear trend in "GarageYrBlt", "BsmtUnfSF", "YearBuilt".

Bar plots are shown for categorical variables (Fig. 3). Some categories are not equally distributed within each predictors, like " LotShape","RoofStyle","LotConfig".

Correlations between numeric predictors are visualized by heat plot (Fig. 4).

3 Models

After data cleaning, there are 44 predictors in total including 28 numeric predictors and 16 categorical predictors. After that, data is transformed to model matrix and categorical predictors are transformed as dummy variables. Finally, zero and near zero variables are checked again and excluded in the model matrix. Therefore, the final data includes 52 predictors with 28 numeric predictors and 24 dummy variables. The final data is scaled and standardized in model fitting.

Repeated cross validation (CV) is implemented in the report.

3.1 Linear Methods

3.1.1 Multiple linear regression (MLR)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$$

$\beta' s$ are estimated by least squared estimation:

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{y_i})$$

where $\widehat{Y} = \widehat{\beta_0} + \widehat{\beta_1}\widehat{X_1} + \ldots + \widehat{\beta_p}\widehat{X_p}$. Although MLR requires a Gaussian error for any inference, the report is focus on the ability of prediction, so we don't have to do transformation for response variable.

3.1.2 Ridge regression

The ridge coefficient estimation is the minimium of the loss function:

$$min(RSS + \lambda \sum_{j=1}^{p} \beta_j^2)$$

All coefficients will shrink when λ increases, but none of them will shrink to zero. So ridge regression will remain all predictors in the final model. In R, $\alpha = 0$ is fixed, and a range of λ is implemented to find the best tuning parameter with the criteria of smallest MSE through CV.

3.1.3 LASSO regression

The LASSO coefficient estimation is the minimum of the loss function:

$$min(RSS + \lambda \sum_{j=1}^{p} |\beta_j|)$$

All coefficients will shrink to zero when λ is large enough. LASSO regression will remain a subset of predictors in the final model. In R, $\alpha = 1$ is fixed, selection of best tuning parameter is similar to ridge.

3.1.4 Elastic regression

The elastic coefficient estimation is the minimum of the loss function:

$$min(RSS + \lambda_1 \sum_{i=1}^{n} \beta_j^2 + \lambda_2 \sum_{i=1}^{n} |\beta_j|)$$

All coefficients will not shrink to exact zero when λ increases. In R, α can be changed in [0,1], so combinations of a range of λ and α is implemented to find the best combination of tuning parameters with the criteria of smallest MSE through CV.

3.1.5 Principal component regression (PCR)

It includes two steps, dimension reduction and regression.

$$Z_m = \sum_{j=1}^{p} \phi_{mj} X_j, y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im}$$

The number of principal components (CP) is chosen by CV with smallest MSE.

3.2 Non-linear methods

3.2.1 K-nearest neighbors algorithm (k-NN)

The k nearest points are used to fit the line.

$$\hat{f}(x_0) = Ave(y|x \in N(x_0)) = \sum_{i=1}^{n} w(x_0, x_i) y_i$$

where $w(x, x_i) = I(x_i \in N_k(x))/K$. Tuning parameter, the number of nearset points k is chosen through CV.

3.2.2 Generalized additive model (GAM)

It allows flexible non-linearities in several variables based on their own scatter plot or degree of freedom (DF), if the points are not linear shaped or the DF is greater than 1, then a non-linear term should be considered.

$$g[E(y|X)] = \beta_0 + f_1(X_1) + ... + f_p(X_p)$$

3.2.3 Multivariate Adaptive Regression Spline (MARS)

It is a piecewise linear model while the cut points are selected by algorithm, and then the hinge functions can be written as $(h(x - c), h(c - x))$.

3.3 Results

For LASSO, plot of MSE across a sequence of λ is made (Fig. 5), and the best λ is 733.6191. For ridge model, plot of MSE across a sequence of λ (Fig. 6) shows that the best tuning parameter is $\lambda = 10575.88$. As for elastic model, 750 combinations of α and λ are checked (Fig. 7), the best pair is $\alpha = 0, \lambda = 10673.4$, since $\alpha = 0$, the result is similar as ridge. For PCR, 52 principle components (PC) are tested, and the best number of principle component (PC) is 45 through smallest MSE (Fig. 8).

For k-NN model, after testing a sequence of k from 5 to 43, the best tuning parameter is equal to 11. For GAM model, "train" function is implemented and 17 out of 52 variables are tested having a non-linear relationships with response (Fig. 9). For MARS model, 10 cut points are used to fit the model (Fig. 10).

4 Conclusion

All models are compared through MSE (Tab. 1, Fig 11, Fig. 12). With differnt metrics, the best model is different. GAM model obtained the smallest median MSE while MARS model obtained the smallest average MSE. Here we use the average MSE as the final criteria. The best model is MARS.

There are 10 cut points in MARS model, they are in "OverallQual", "GrLivArea", "YearBuilt", "X1stFlrSF", "BsmtFinSF1", "X2ndFlrSF", "OverallCond", "LotArea", "TotalBsmtSF", "BedroomAbvGr". The coefficients of hinge functions are shown in Tab. 2. According to Fig. 10, except "X1stFlrSF", "GrLivArea" and "BedroomAbvGr". All predictors have increasing trends when response rises. "X1stFlrSF" and "BedroomAbvGr" has a dereaseing trend all the time when the sale price increases.

Moreover, the top 10 most important variables for the MARS model are checked in Tab.3, for a decreasing order of contribution, they are: "OverallQual", "GrLivArea", "YearBuilt", "X1stFlrSF", "BsmtFinSF1", "X2ndFlrSF", "OverallCond", "LotArea", "TotalBsmtSF", "BedroomAbvGr".

Heat plot for the top 10 the most important variables is made (Fig. 13), "X2ndFlrSF" and "GrLivArea", "X1stFlrSF" and "TotalBsmtSF", "YearBuilt" and "OverallQual" are highly correlated, respectively.

Appendix - Figures and Tables

Table 1 MSE of all methods through cross validation

column	mean	sd	median	min	max	range
Lm	37652.81	13419.544	32465.29	24558.58	73047.02	48488.43
LASSO	37055.96	13296.278	32133.75	22769.02	71611.12	48842.10
Ridge	36867.69	12802.181	32101.84	22945.90	69818.05	46872.15
Elastic	36867.67	12796.652	32109.14	22938.26	69801.60	46863.33
PCR	37211.08	12937.826	31889.95	24301.60	71083.87	46782.26
Knn	38358.18	8166.486	36842.44	24604.40	55632.19	31027.79
GAM	40496.46	30754.925	28221.06	20659.37	184932.39	164273.02
MARS	30718.70	7001.811	28796.77	21003.72	55490.50	34486.78

Table 2 Hinge functions and their coefficients in MARS model

	Coefficient
(Intercept)	3.134607e+05
h(OverallQual-7)	3.649653e+04
h(7-OverallQual)	-9.993767e+03
h(GrLivArea-2872)	-6.611808e+01
h(2872-GrLivArea)	-4.677017e+01
h(X2ndFlrSF-1349)	3.326312e+02
h(YearBuilt-2007)	2.188740e+04
h(2007-YearBuilt)	-6.308092e+02
h(BsmtFinSF1-763)	5.203470e+01
h(763-BsmtFinSF1)	-7.395022e+00
h(LotArea-21780)	3.987623e-01
h(21780-LotArea)	-2.072673e+00
h(7-OverallCond)	-1.067073e+04
h(X1stFlrSF-2113)	-3.473711e+02
h(TotalBsmtSF-1626)	7.482980e+01
h(1626-TotalBsmtSF)	-2.013837e+01
h(GrLivArea-1855)	5.345370e+01
h(BedroomAbvGr-4)	-3.879188e+04
h(4-BedroomAbvGr)	3.558218e+03

Table 3 Top 10 most important variables in MARS model

	Overall
OverallQual	100.000000
GrLivArea	62.171068
YearBuilt	44.719808
BsmtFinSF1	32.817271
X1stFlrSF	32.817271
X2ndFlrSF	32.046534
OverallCond	24.157274
LotArea	20.744488

	Overall
TotalBsmtSF	18.171330
BedroomAbvGr	9.914247

Figure 1 Distridution of response (SalePrice)

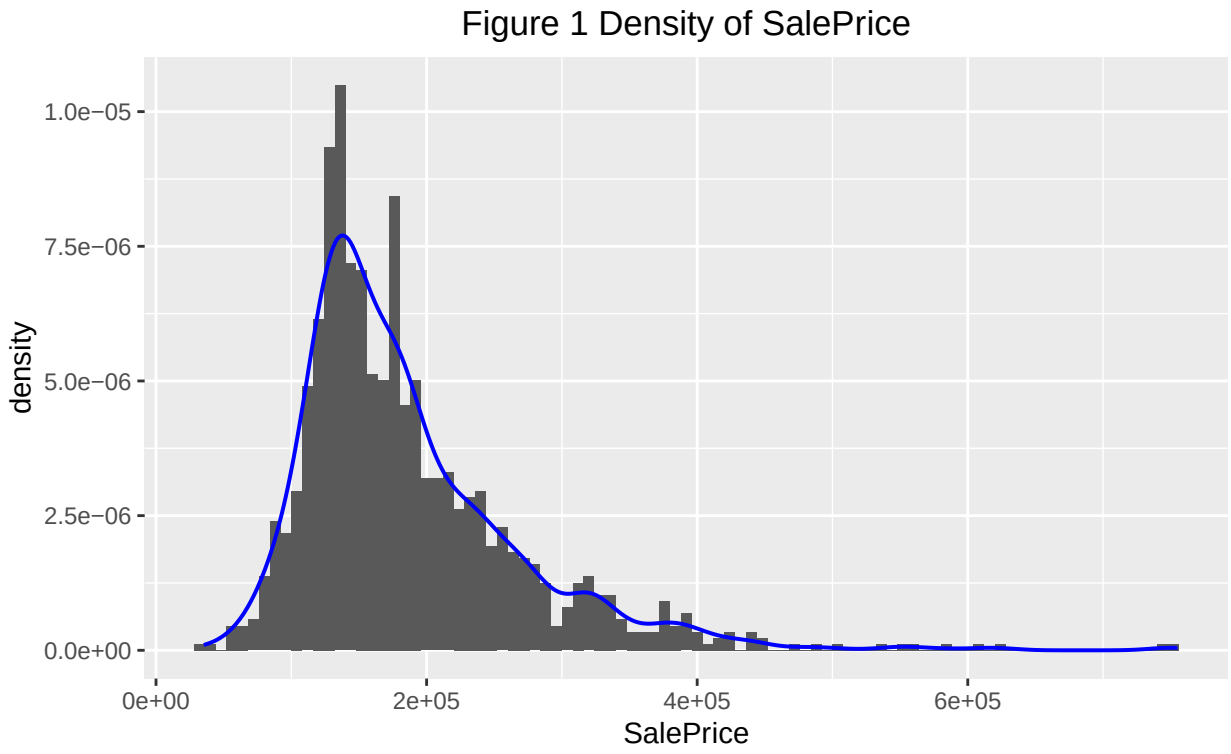

Figure 2 Scatter plots of continuous predictors against SalePrice

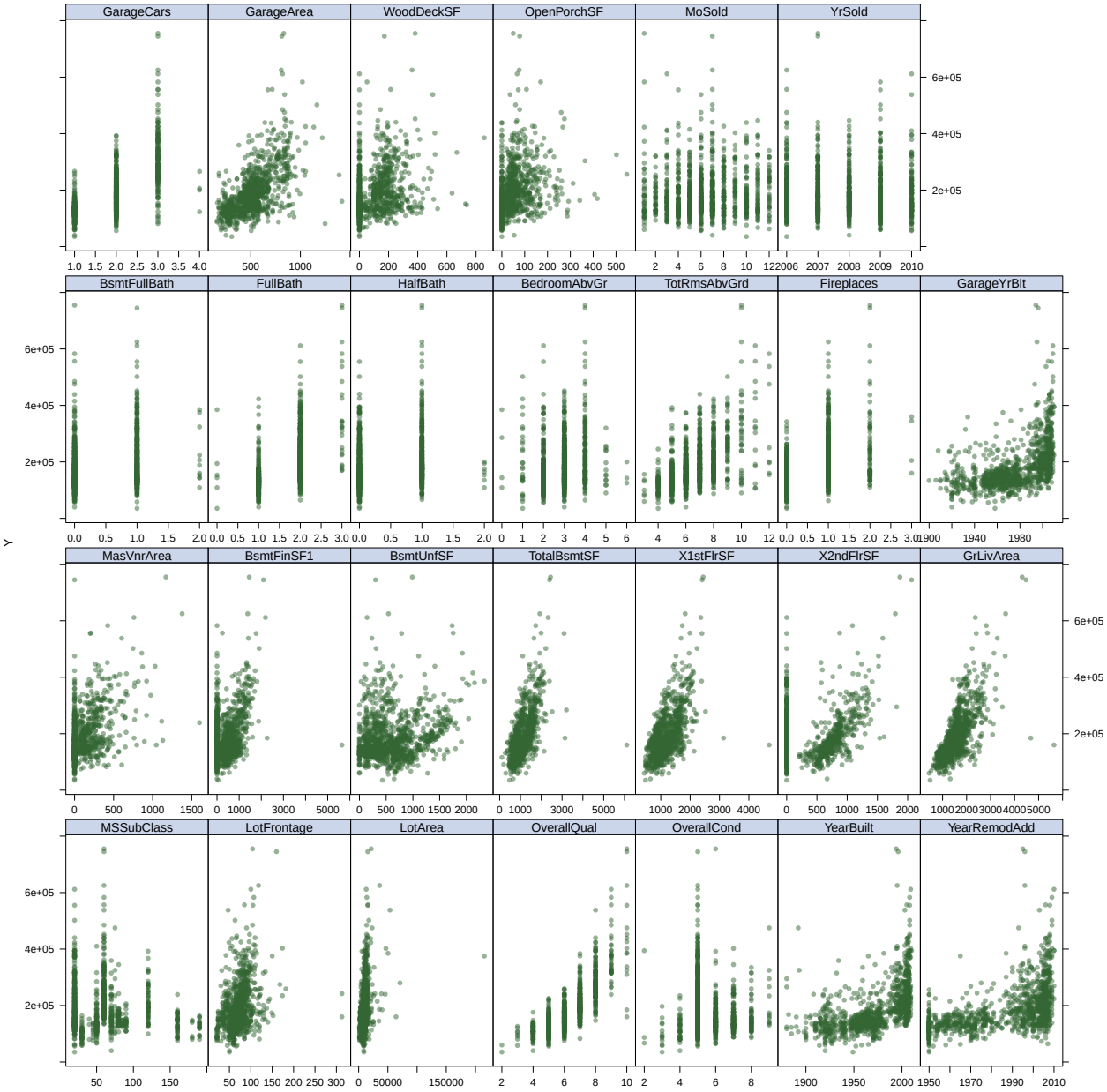

Figure 3 Bar plots of categorical predictors

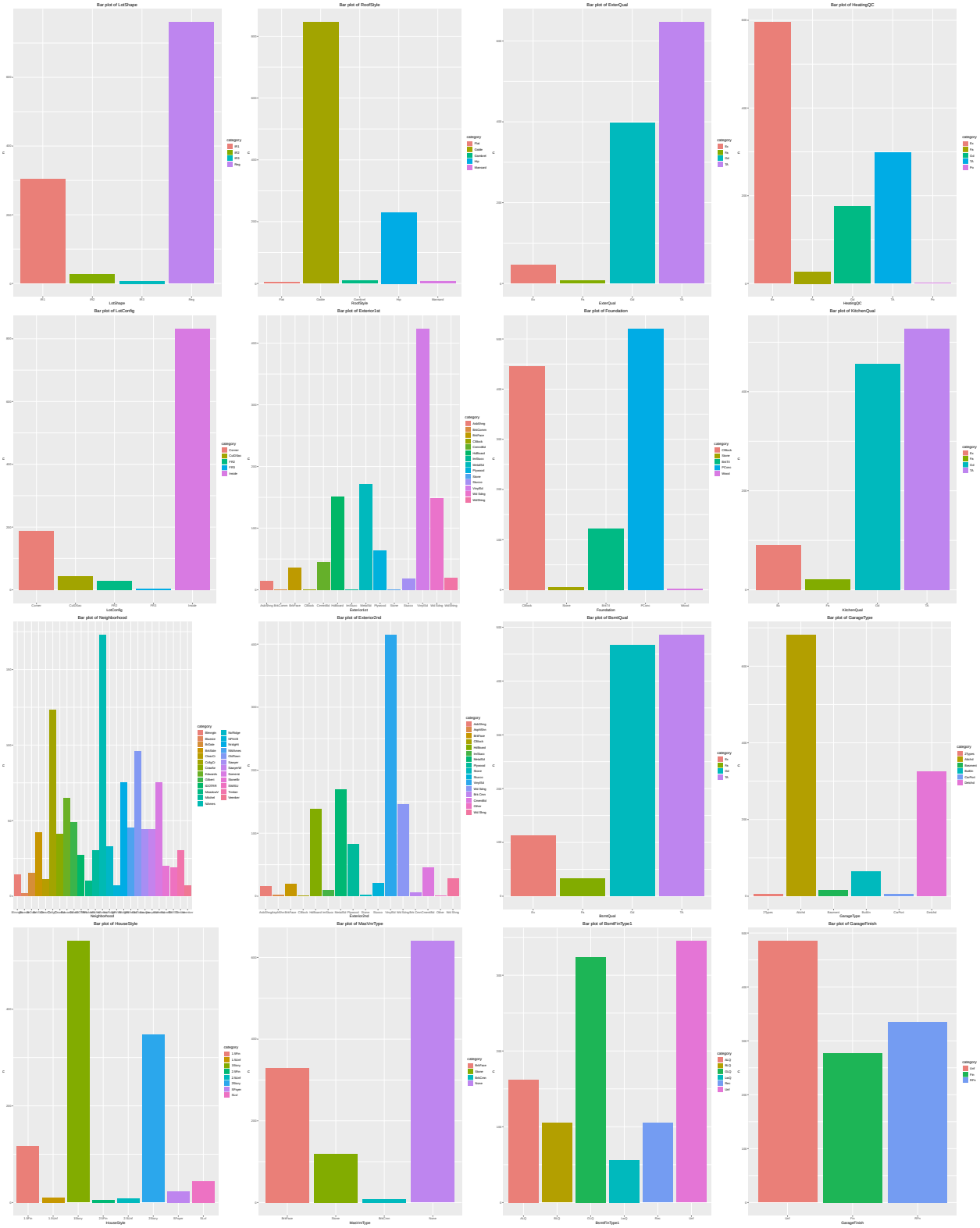

Figure 4 Heat map for all predictors

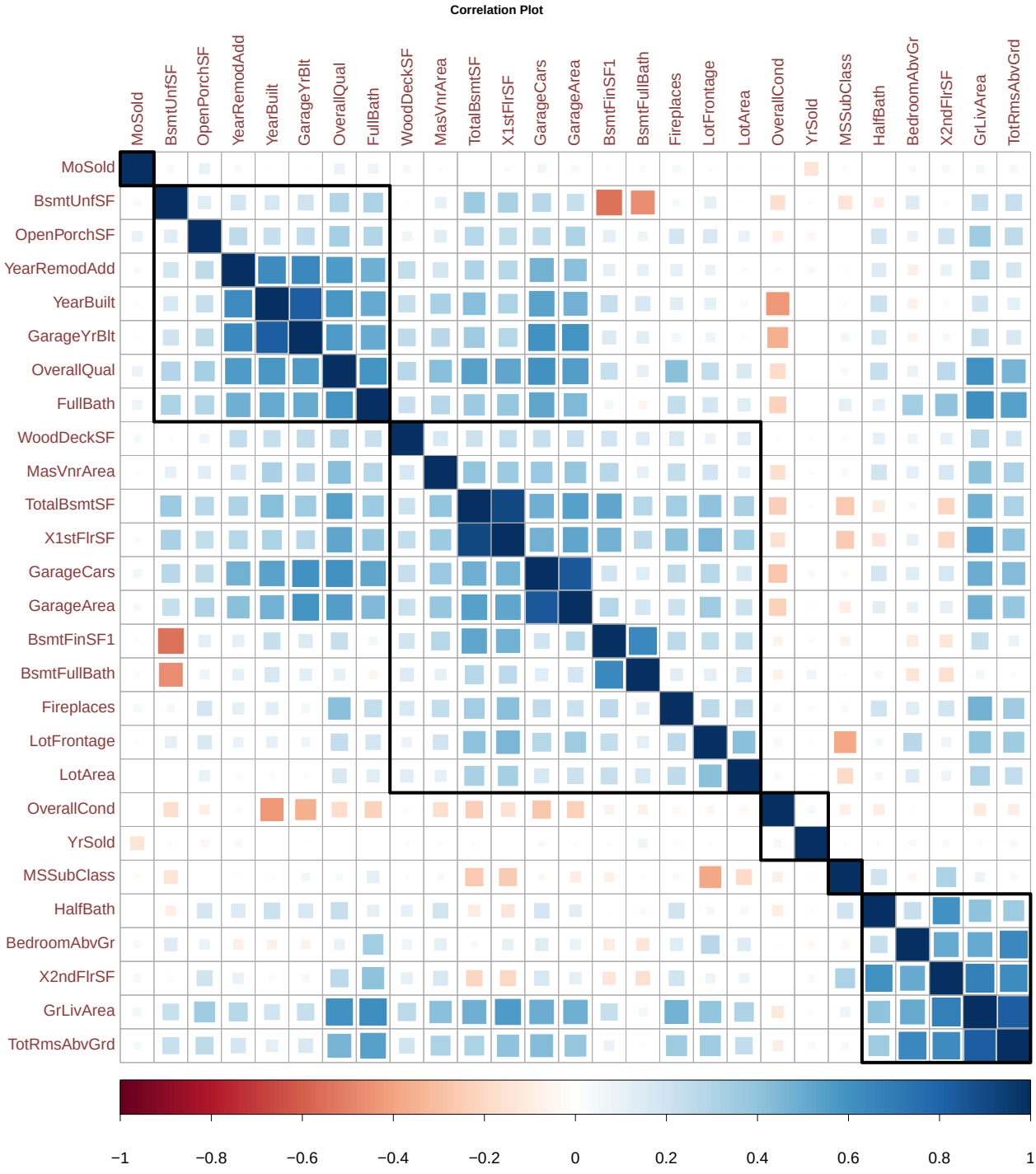

Figure 5 LASSO

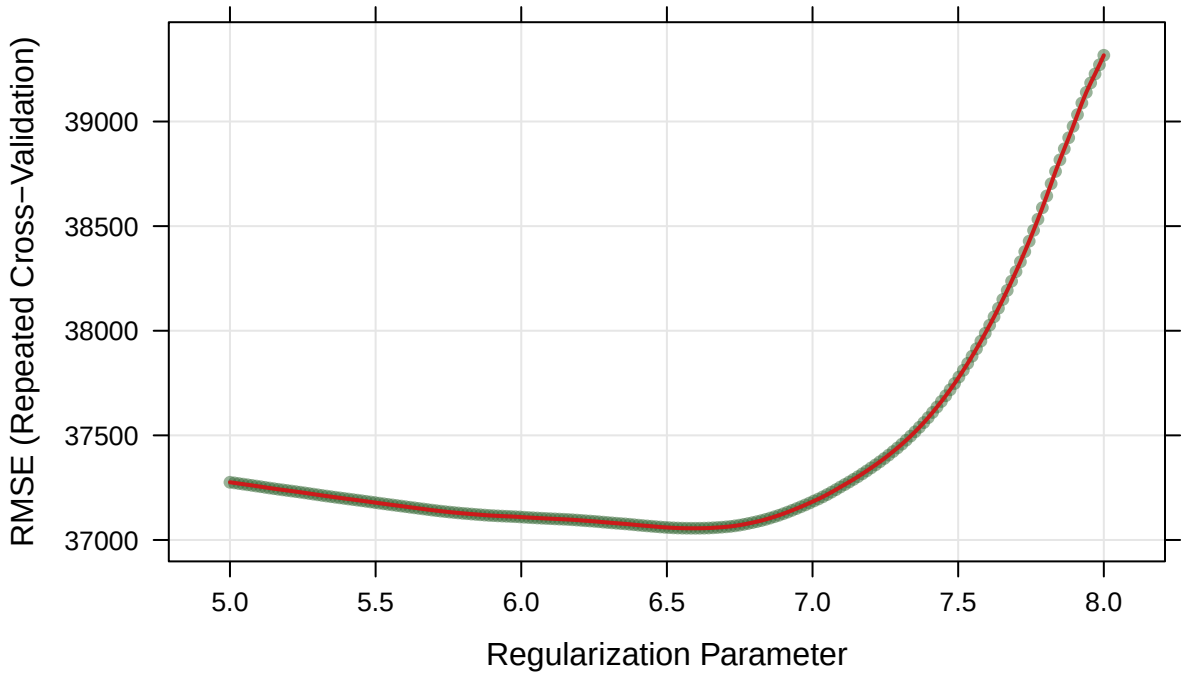

Figure 6 Ridge

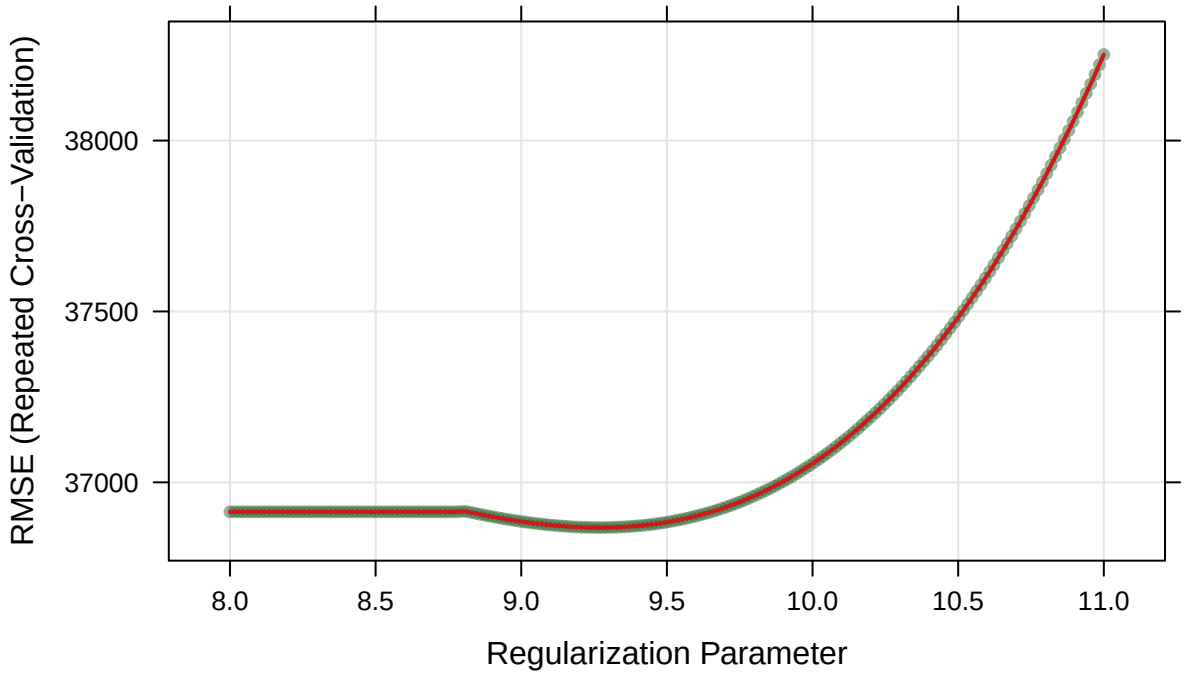

Figure 7 Elastic

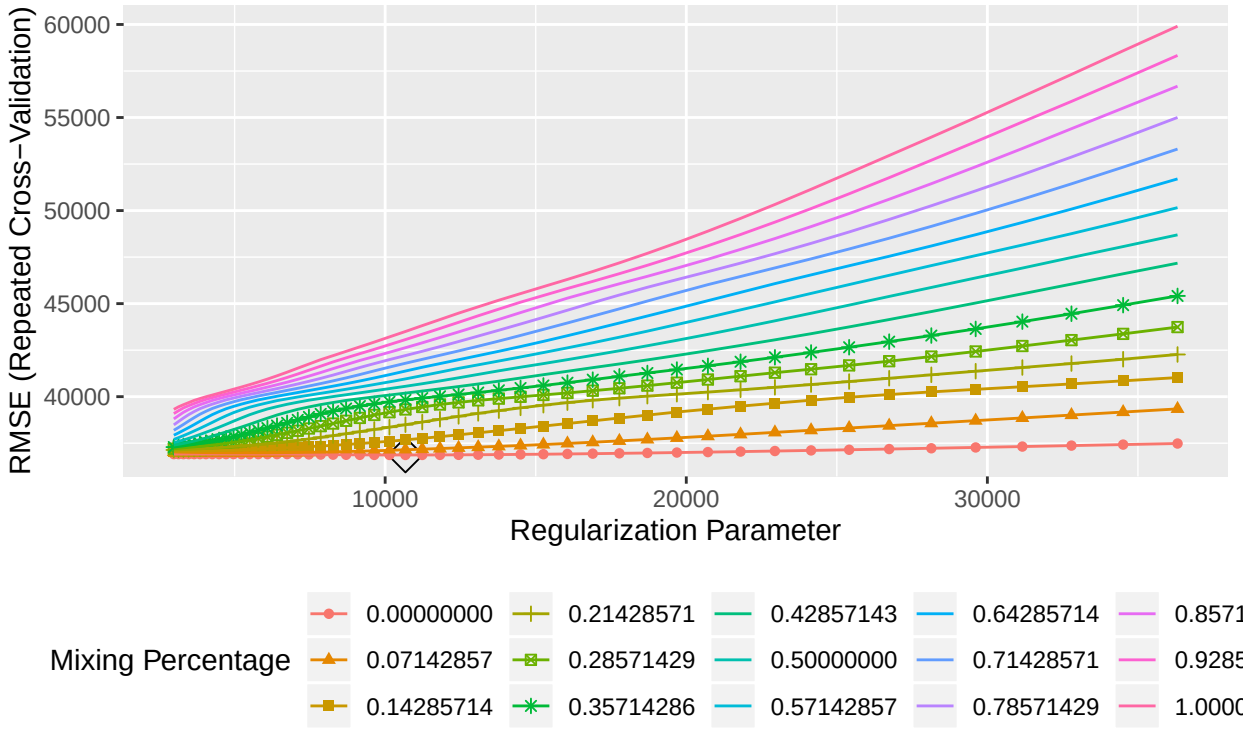

Figure 8 PCR

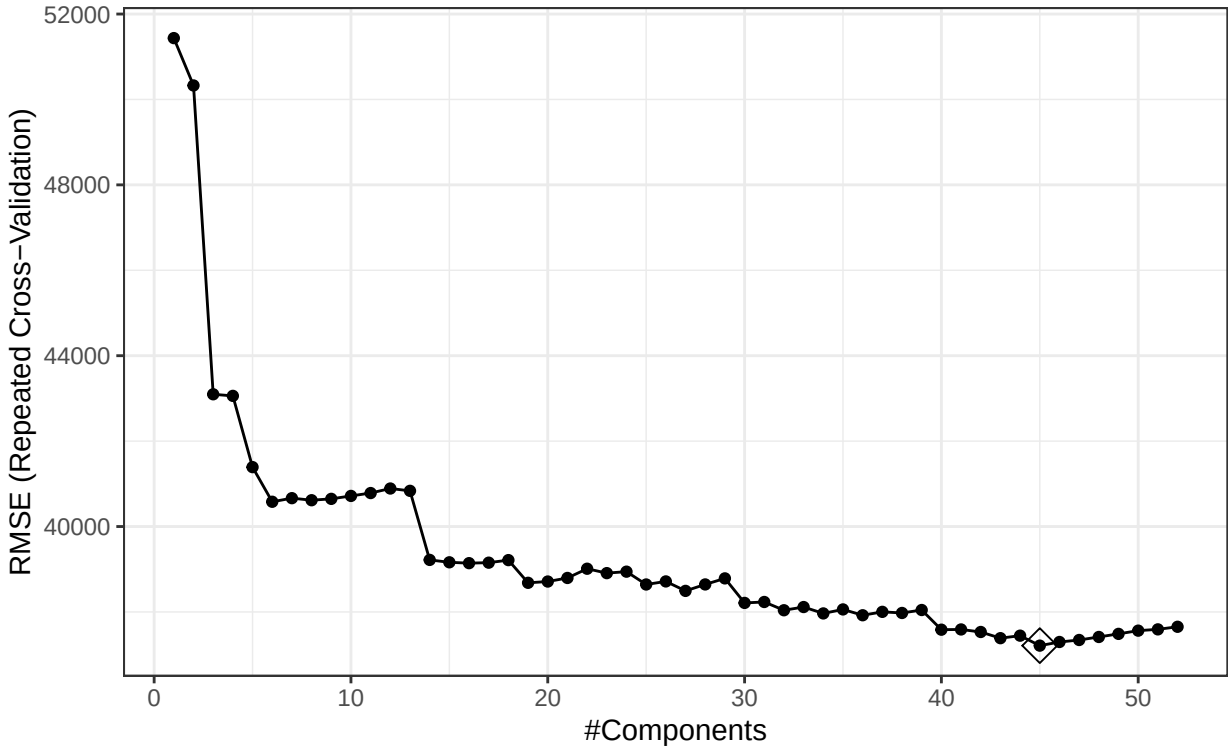

Figure 9 GAM

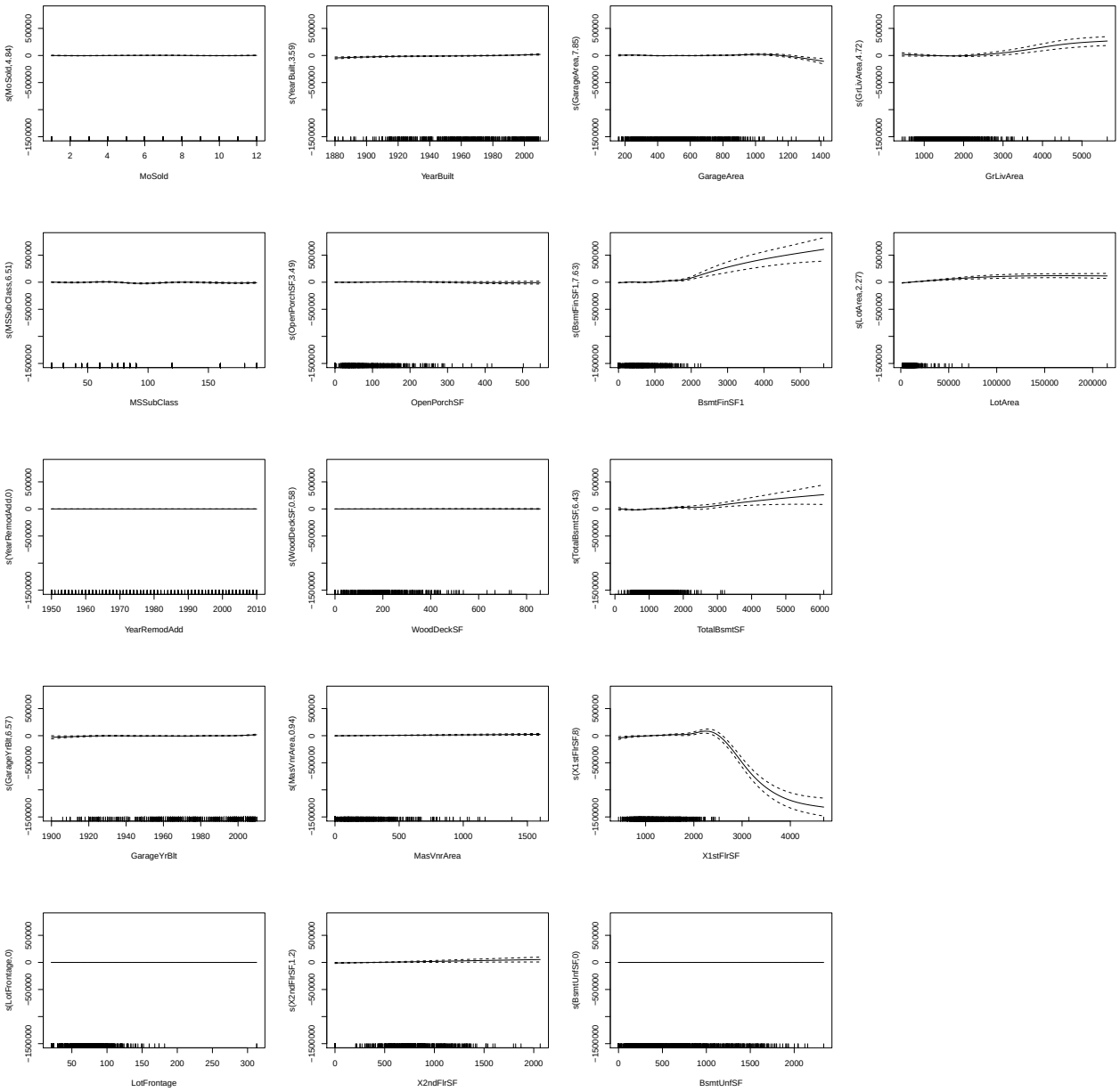
s(MoSold,4.84)
MoSold
s(YearBuilt,3.59)
YearBuilt
s(GarageArea,7.85)
GarageArea
s(GrLivArea,4.72)
GrLivArea
s(MSSubClass,6.51)
MSSubClass
s(OpenPorchSF,3.49)
OpenPorchSF
s(BsmtFinSF1,7.63)
BsmtFinSF1
s(LotArea,2.27)
LotArea
s(YearRemodAdd,0)
YearRemodAdd
s(WoodDeckSF,0.58)
WoodDeckSF
s(TotalBsmtSF,6.43)
TotalBsmtSF
s(GarageYrBlt,6.57)
GarageYrBlt
s(MasVnrArea,0.94)
MasVnrArea
s(X1stFlrSF,8)
X1stFlrSF
s(LotFrontage,0)
LotFrontage
s(X2ndFlrSF,1.2)
X2ndFlrSF
s(BsmtUnfSF,0)
BsmtUnfSF

Figure 10 MARS

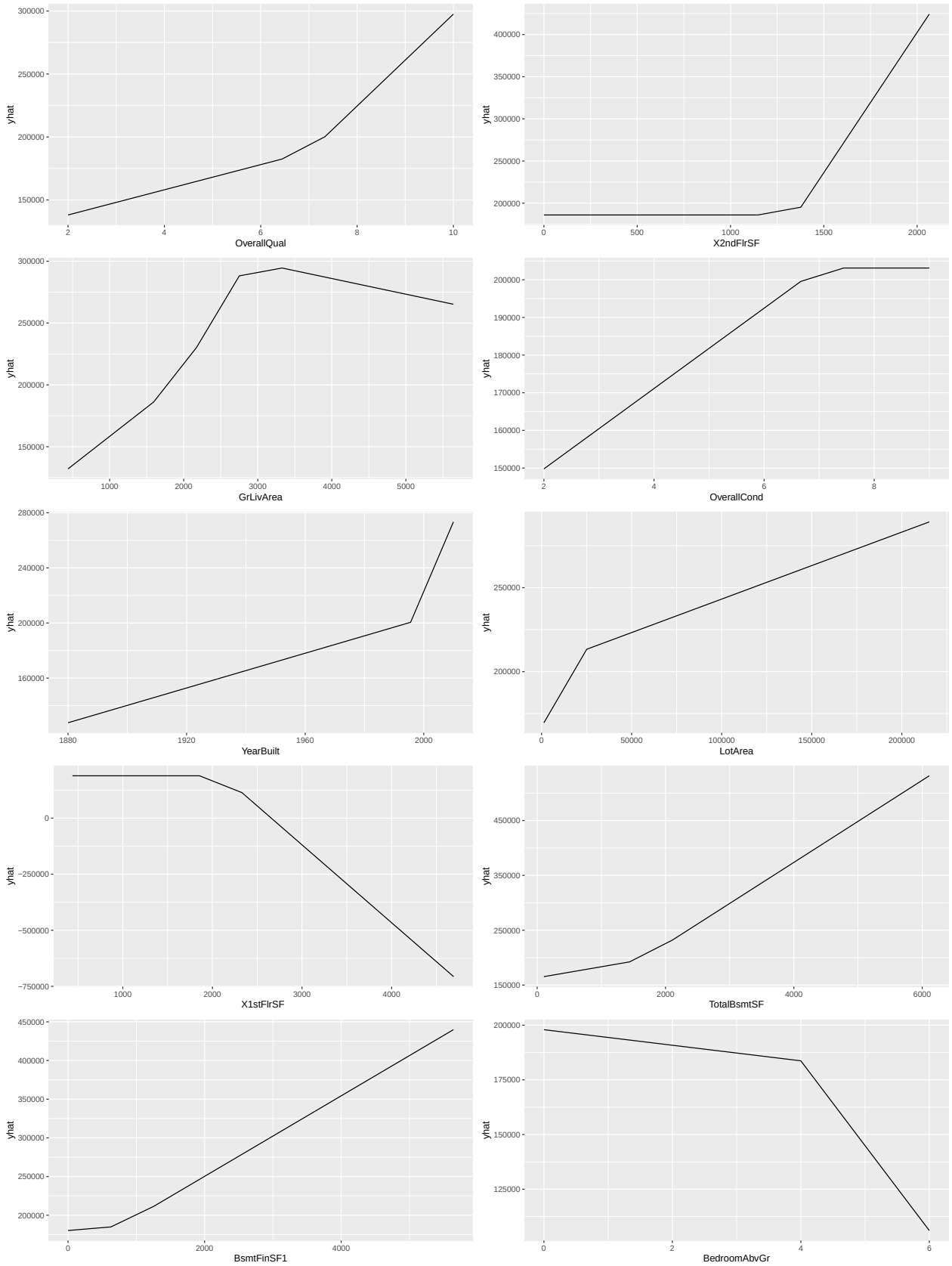

Figure 11 Box plot of all methods through CV

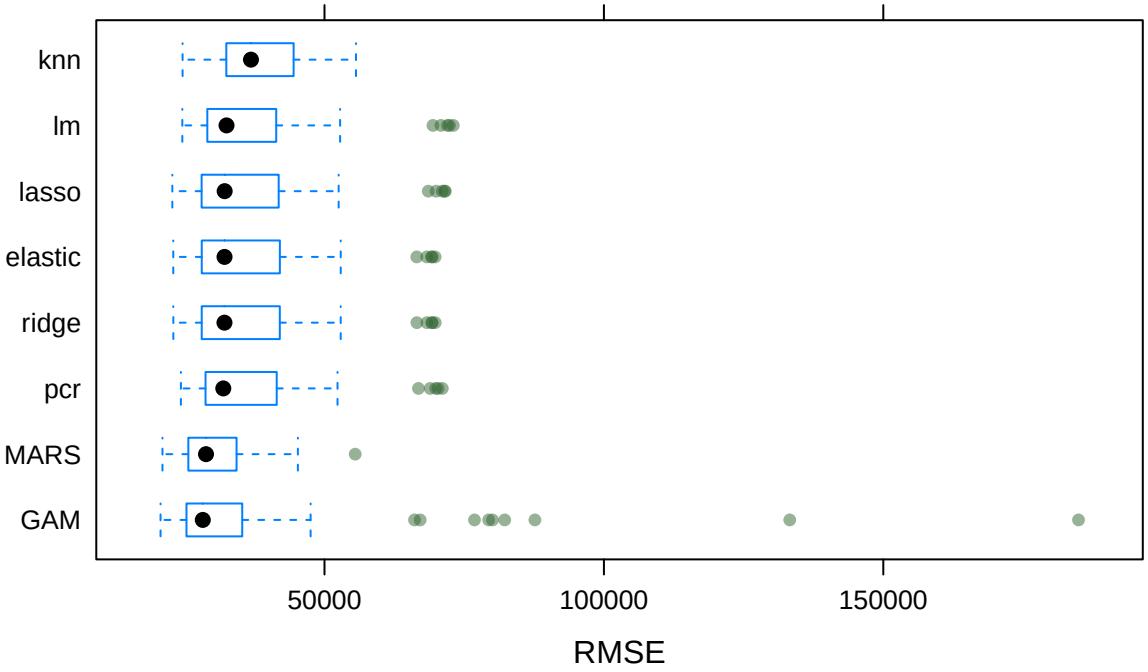

Figure 12 Box plot of all methods through CV

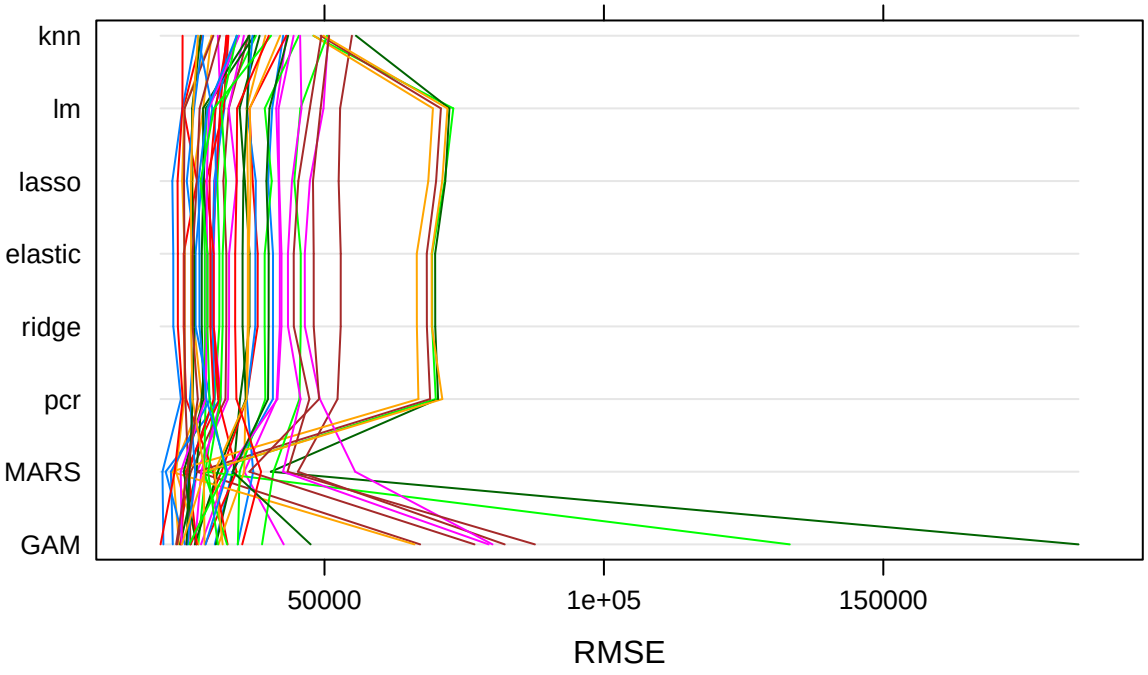

Figure 13 Heat map for top 10 most important variables