Midterm project, group 6

Ruoyuan Qian

1 Introduction

1.1 Background

In the dataset, the response variable is the SalePrice ($) of residential homes in Ames, Iowa. And there are 1460 observations with 79 explanatory variables describing a variety of aspects of these housings. Among explanatory variables, there are 37 numeric variables and 43 categorical variables.

1.2 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.

1.3 Exploratory data analysis (EDA)

1.3.1 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. Finally, NA's are dropped from the remaining data. When all works are done, there are 55 predictors in total including 28 numeric predictors and 27 categorical predictors.

1.3.2 Visualization

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), bar plots are shown for categorical variables (Fig. 3). Correlations between predictors are visualized by heat plot (Fig. 4).

2 Methods

The data for all methods is the same to keep the ability of comparison. Data is scaled and standardized in model fitting. All categorical variables are transformed to factor variavles.

2.1 Linear Methods

2.1.1 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.

2.1.2 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.

2.1.3 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.

2.1.4 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 cross validation (CV).

2.1.5 Principal component regression (PCR)

It includes two steps, dimension reduction and regression.

$$Z_m = \sum_{j=1}^{p} \phi_{mj} X_j$$

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im}$$

2.2 Non-linear methods

2.2.1 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 $w(x, x_i) = I(x_i \in N_k(x))/K$

2.2.2 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.

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

2.2.3 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 Results

3.1 Model comparison

For LASSO, plot of MSE across a sequence of λ is made (Fig. 5), and the best λ is $e^{7.035}$. For ridge model, plot of MSE across a sequence of λ (Fig. 6) shows that the best tuning parameter is $\lambda = e^{10.10}$. As for elastic model, 750 combinations of α and λ are checked (Fig. 7), the best pair is $\alpha = 0, \lambda = e^{10.10}$, which is the same as ridge, so in the model comparison, only ridge regression will be presented. For PCR, 54 principle components (PC) are tested, and the best number of principle component (PC) is 26 through 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 20 out of 54 predictors are tested for non-linear relation to response (Fig. 9). For MARS model, 10 cut points are used to fit the model (Fig. 10).

All models are compared through MSE (Tab. 1, Fig 11, Fig. 12). The best model is the one with the smallest MSE, which is MARS. K-nn method obtain the largest MSE among the all.

3.2 Model interpretation

There are 10 cut points in MARS model, they are in “OverallQual”, “GrLivArea”, “X2ndFlrSF”, “YearBuilt”, “BsmtFinSF1”, “LotArea”, “OverallCond”, “X1stFlrSF”, “TotalBsmtSF”, “SaleCondition”. The coefficients of hinge functions are shown in Tab. 2. According to Fig. 10, except “X1stFlrSF” and “SaleCondition”, all predictors have increasing trends when response rises. “X1stFlrSF” 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”, “BsmtFinSF1”, “X1stFlrSF”, “X2ndFlrSF”, “OverallCond”, “LotArea”, “TotalBsmtSF”, “SaleCondition”.

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.

4 Discussion

As for data interpretation, since the data from the Kaggle does not include the specific meaning for each variable, so I cannot interpret the model in a more detailed way. Besides, the data contains a lot of integer variables, I treated them as continuous variables, so gaps are introduced in some of the scatter plots.

Figures and Tables

Table 1 MSE of all methods through cross validation

column	mean	sd	median	min	max	range
Lm	37505.04	13940.946	33542.84	22794.36	72849.29	50054.93
LASSO	36543.47	12986.560	33204.86	22150.00	69494.01	47344.00
Ridge	36278.35	12238.432	32908.06	21886.61	66422.74	44536.14
PCR	36483.11	12114.243	32675.95	23044.04	67262.97	44218.94
Knn	38064.71	8516.354	37404.23	25491.27	58379.22	32887.94
GAM	35926.26	23224.696	27382.22	19331.35	136719.62	117388.26
MARS	33038.51	9329.652	31375.96	21072.25	58224.52	37152.27

Table 2 Hinge functions and their coefficients in MARS model

	Coefficient
(Intercept)	3.141179e+05
h(OverallQual-7)	4.077041e+04
h(7-OverallQual)	-9.109677e+03
h(2945-GrLivArea)	-5.752370e+01
h(X2ndFlrSF-1360)	3.718084e+02
h(YearBuilt-2007)	1.563304e+04
h(2007-YearBuilt)	-5.721088e+02
h(BsmtFinSF1-763)	4.820646e+01
h(763-BsmtFinSF1)	-1.379454e+01
h(LotArea-20431)	4.618026e-01
h(20431-LotArea)	-1.953780e+00
h(7-OverallCond)	-1.093203e+04
h(X1stFlrSF-2121)	-3.447610e+02
h(TotalBsmtSF-1626)	7.145797e+01
h(1626-TotalBsmtSF)	-1.678924e+01
h(SaleCondition-4)	1.813065e+04
h(4-SaleCondition)	3.657226e+03

Table 3 Top 10 most important variables in MARS model

	Overall
OverallQual	100.00000
GrLivArea	61.92524
YearBuilt	44.28836
BsmtFinSF1	32.18829
X1stFlrSF	32.18829
X2ndFlrSF	30.90316
OverallCond	23.03836
LotArea	19.41104
TotalBsmtSF	16.61787
SaleCondition	11.65620

Figure 1 Distridution of response (SalePrice)

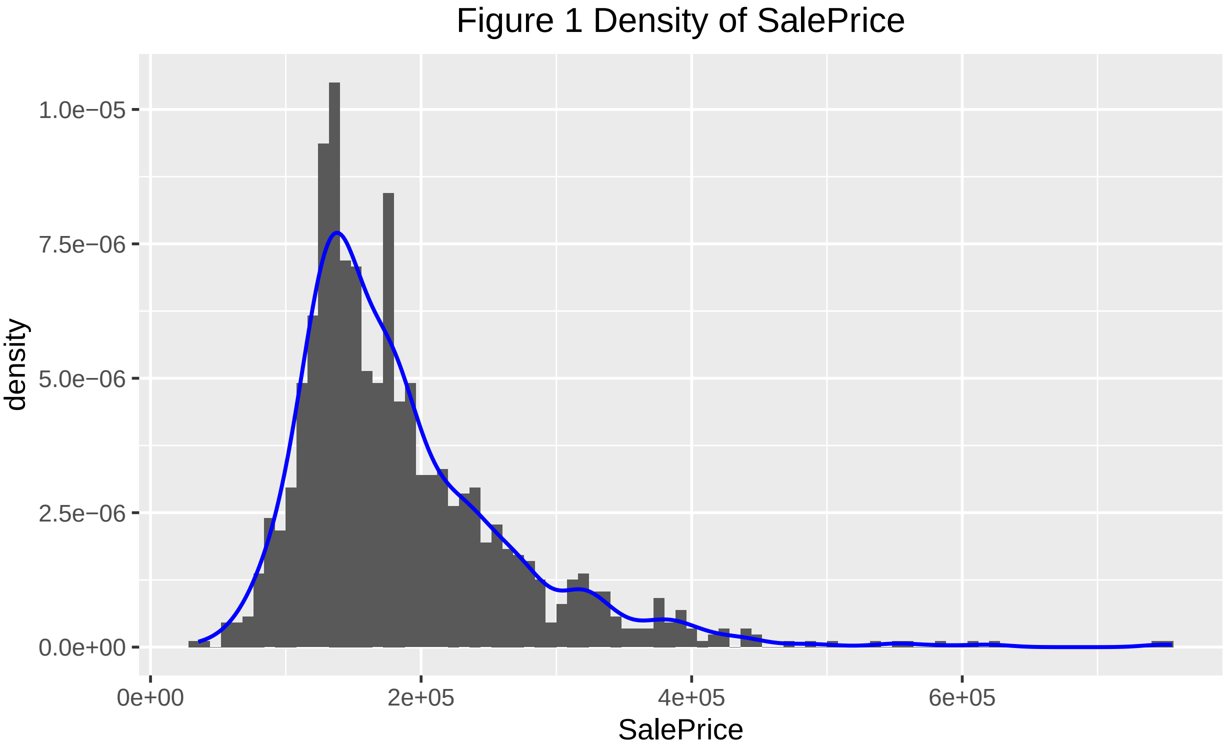


Figure 2 Scatter plots of continuous predictors against SalePrice

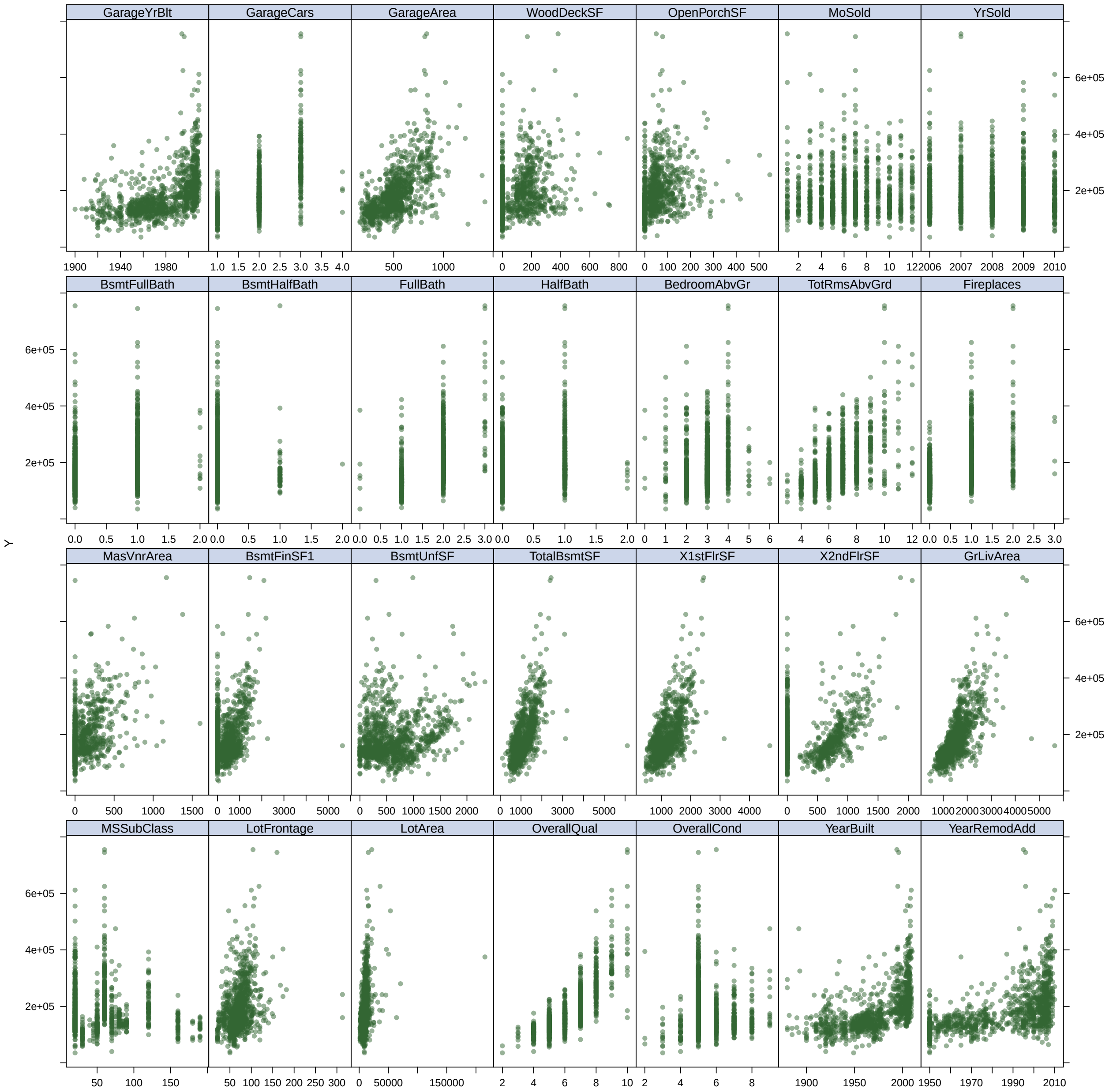
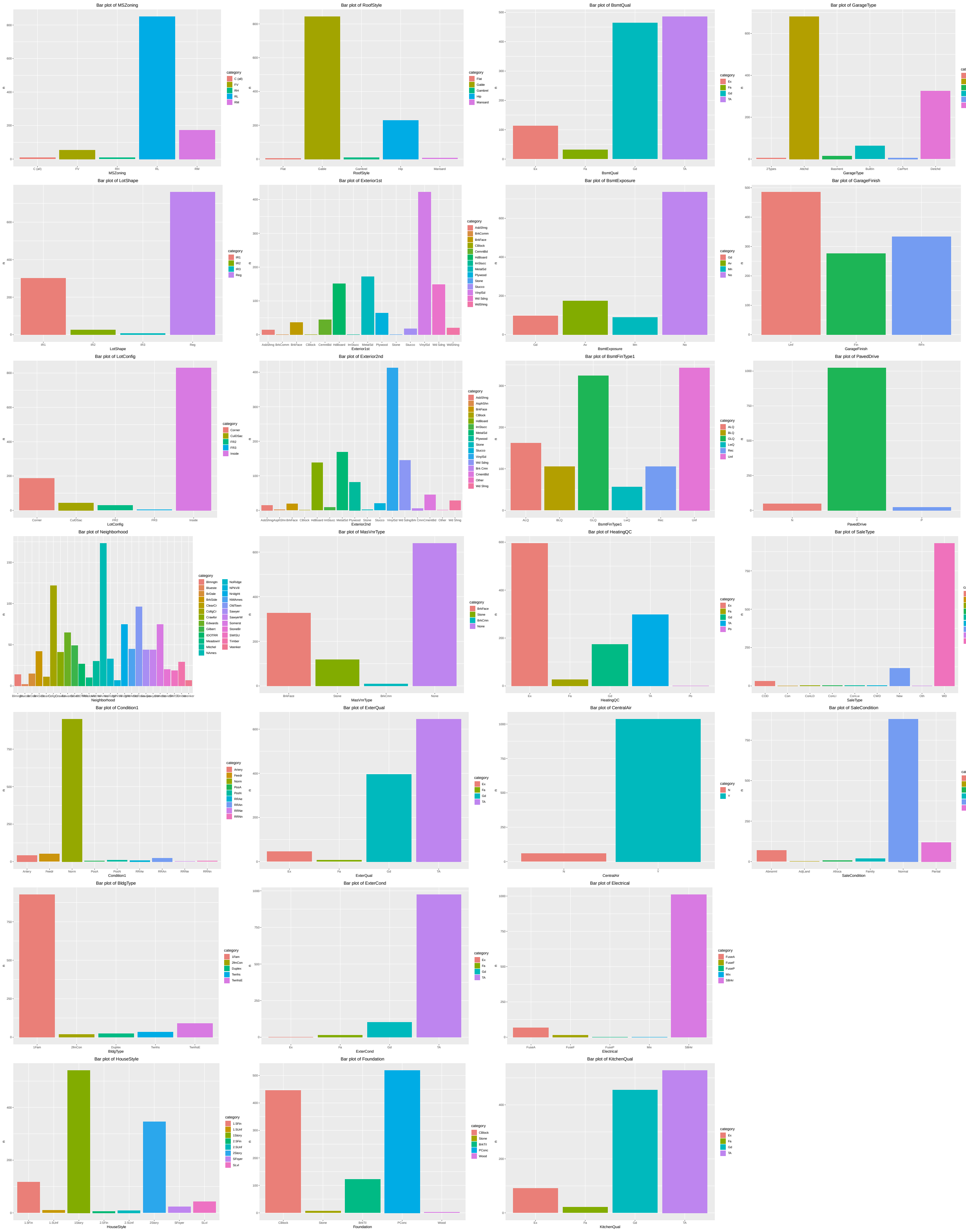


Figure 3 Bar plots of categorical predictors



Bar plot of MSZoning

Bar plot of RoofStyle

Bar plot of BsmtQual

Bar plot of GarageType

Bar plot of LotShape

Bar plot of Exterior1st

Bar plot of BsmtExposure

Bar plot of GarageFinish

Bar plot of LotConfig

Bar plot of Exterior2nd

Bar plot of BsmtFinType1

Bar plot of PavedDrive

Bar plot of Neighborhood

Bar plot of MasVnrType

Bar plot of HeatingQC

Bar plot of SaleType

Bar plot of Condition1

Bar plot of ExterQual

Bar plot of CentralAir

Bar plot of SaleCondition

Bar plot of BldgType

Bar plot of ExterCond

Bar plot of Electrical

Bar plot of HouseStyle

Bar plot of Foundation

Bar plot of KitchenQual

Figure 4 Heat map for all predictors

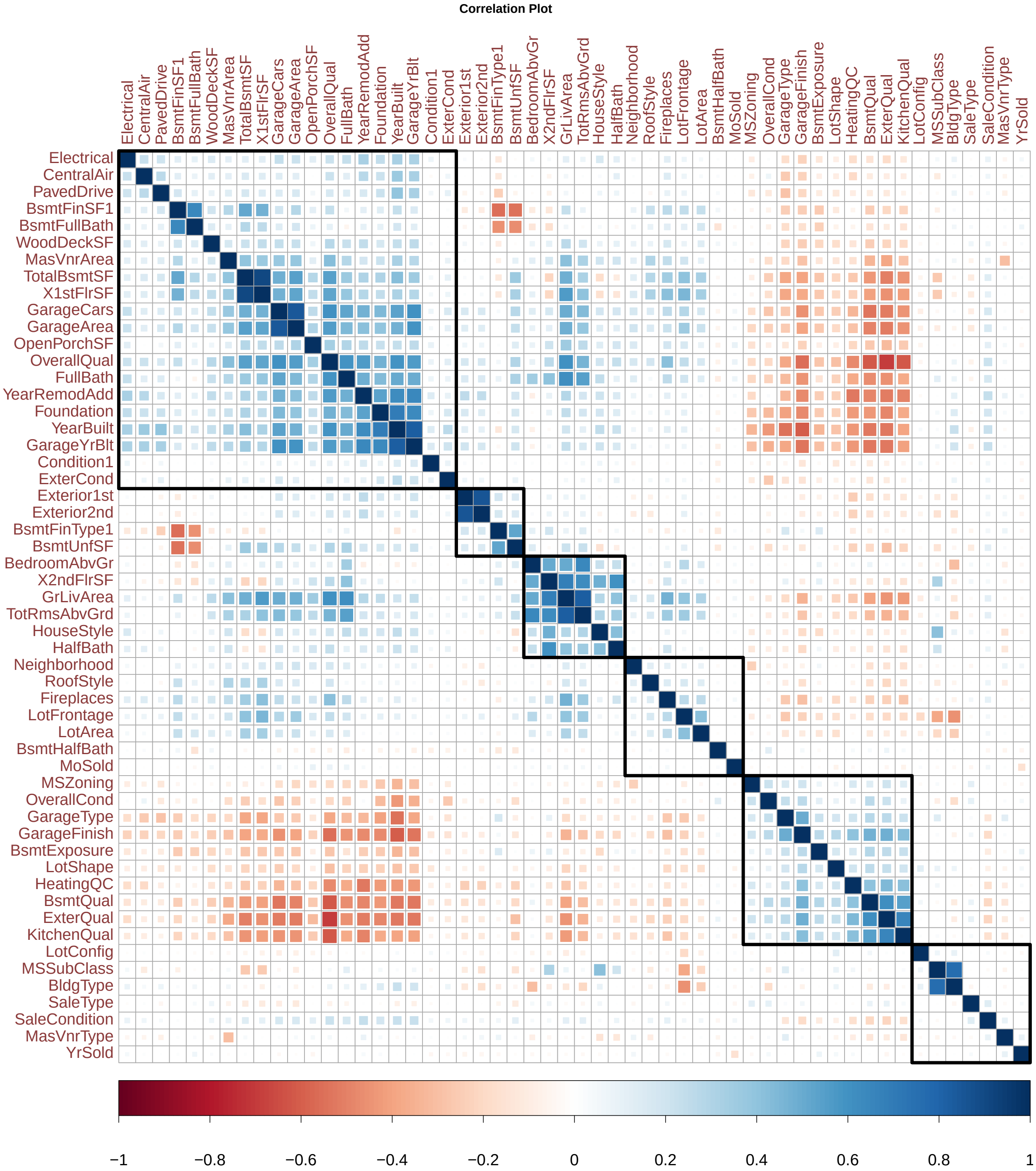


Figure 5 LASSO

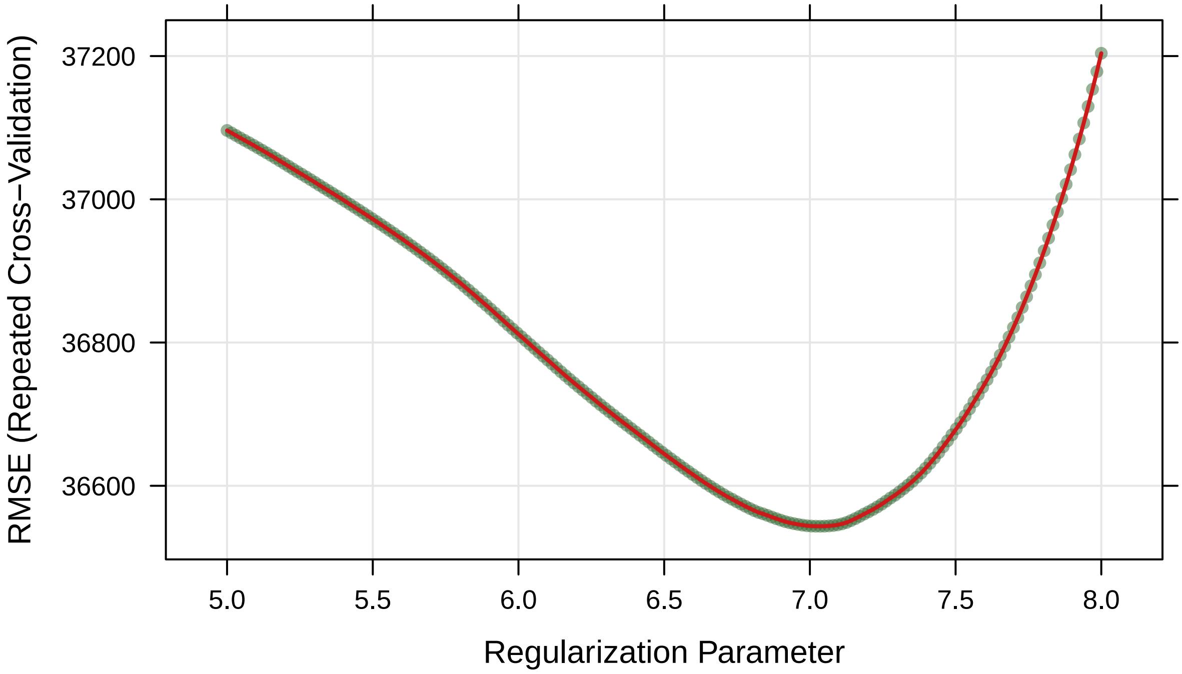


Figure 6 Ridge

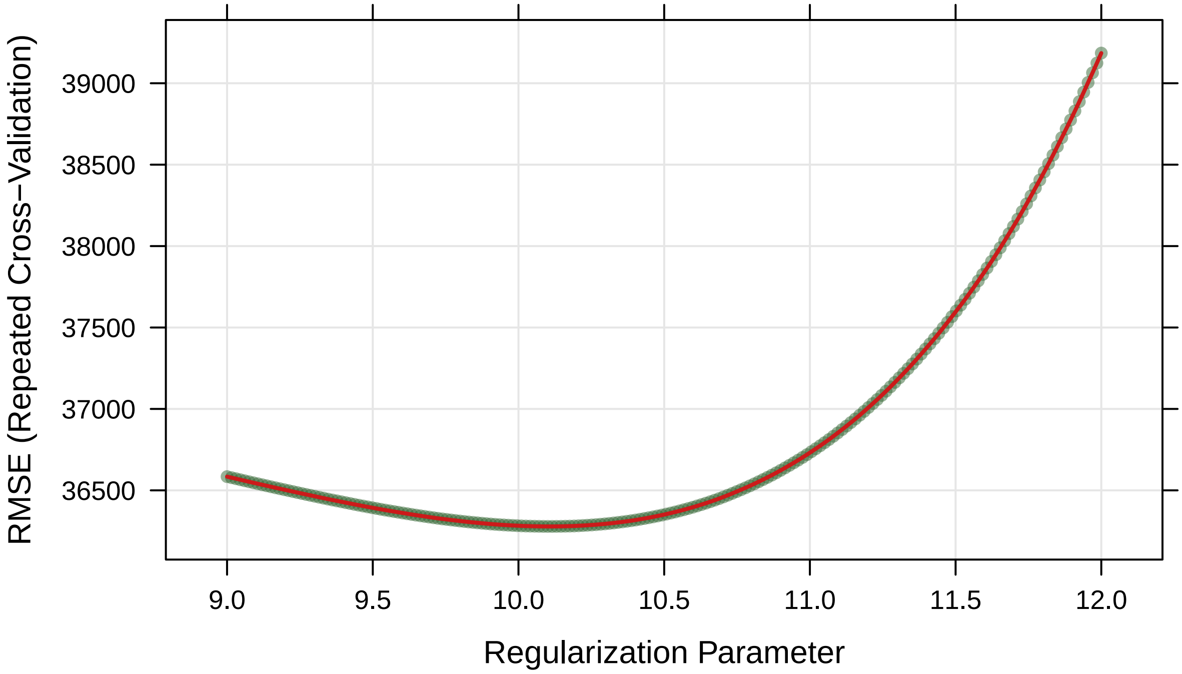


Figure 7 Elastic

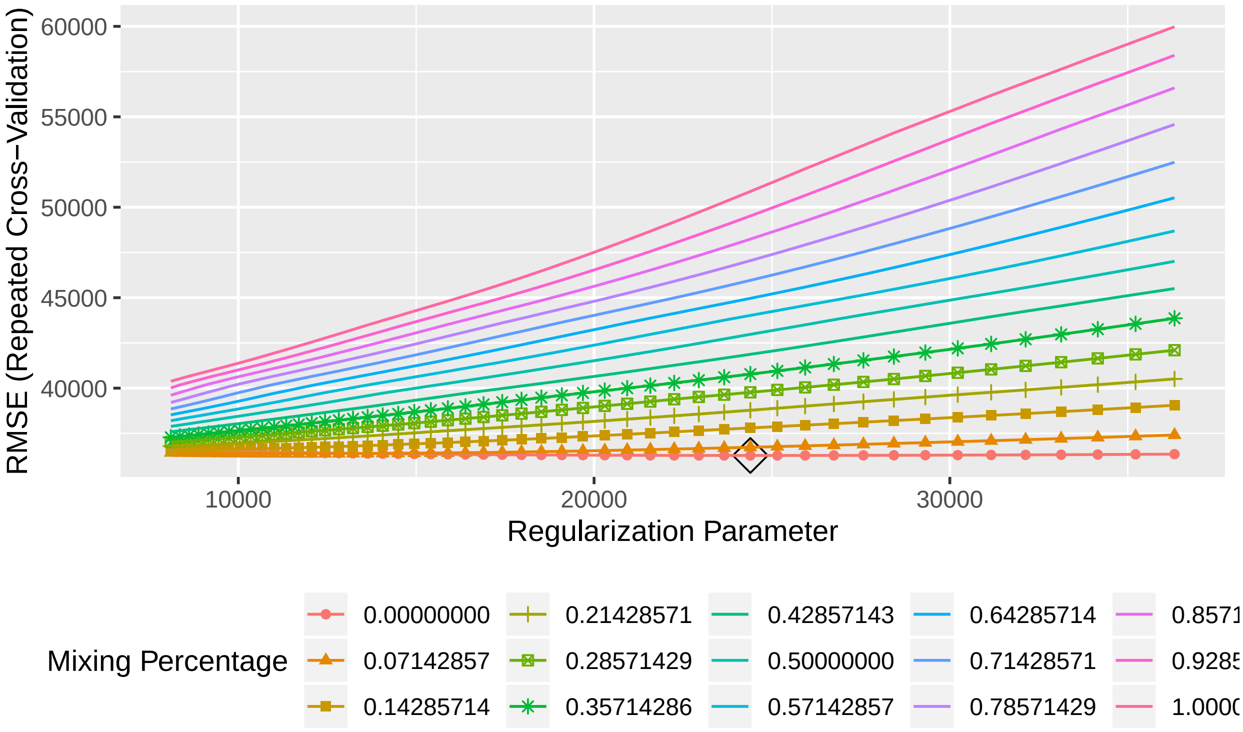


Figure 8 PCR

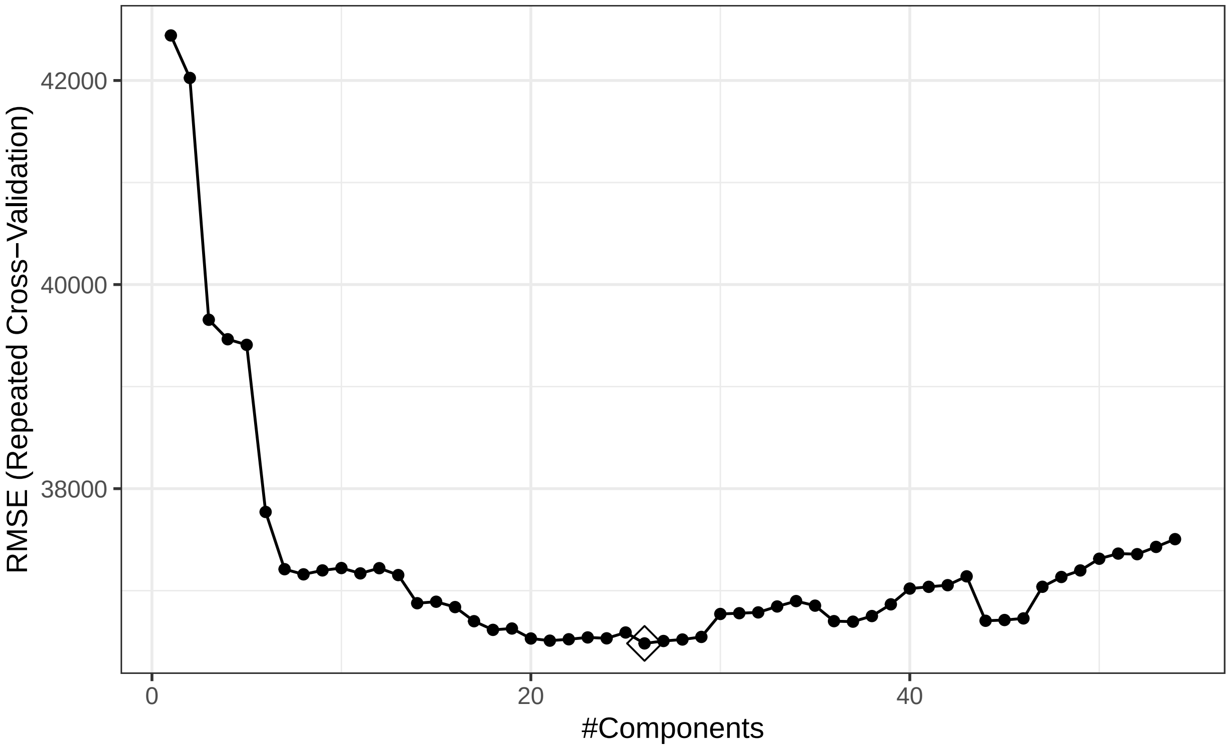


Figure 9 GAM

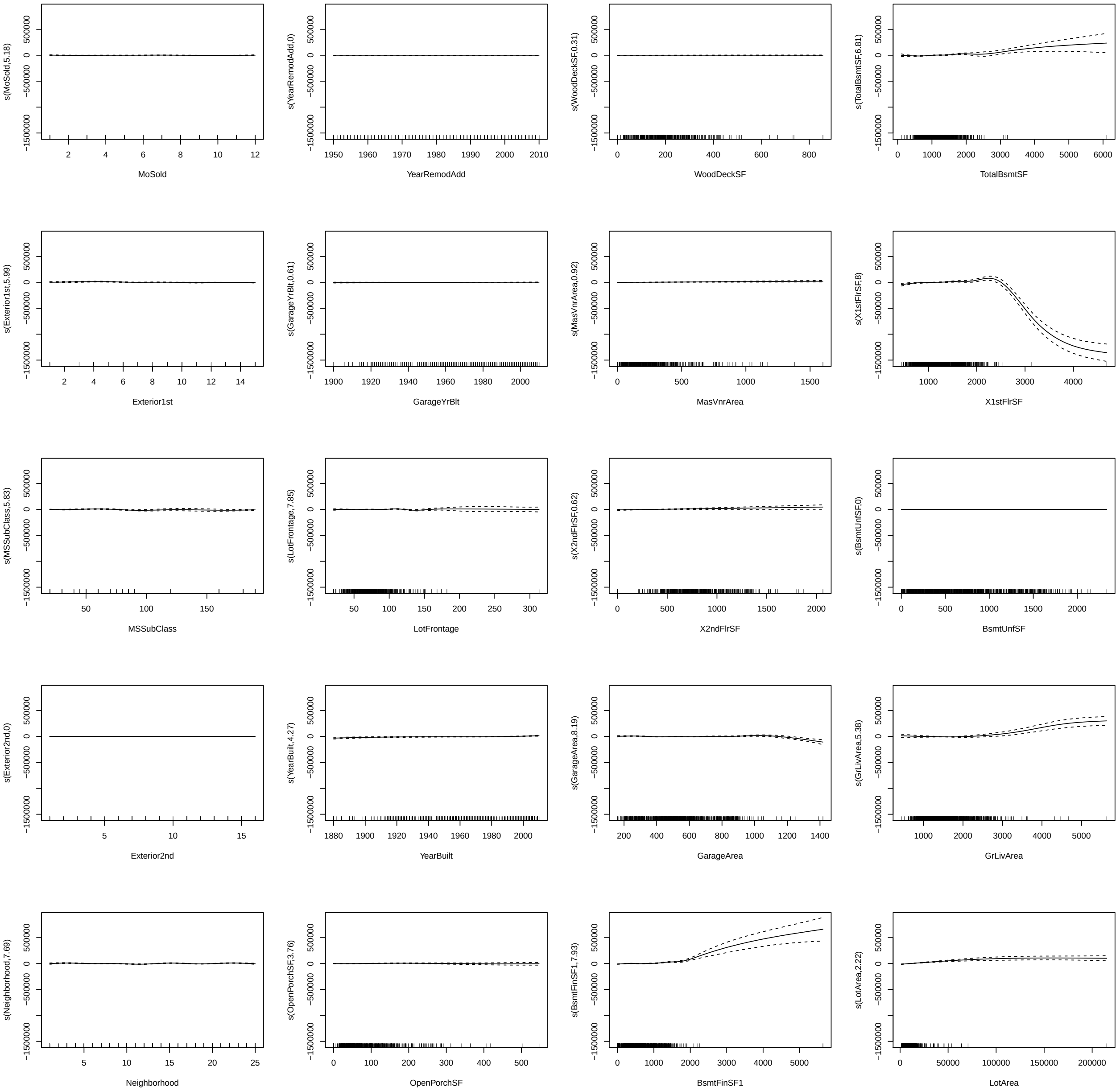

s(MoSold,5.18)
−1500000
−500000
0
500000
2
4
6
8
10
12
MoSold
s(YearRemodAdd,0)
1950
1960
1970
1980
1990
2000
2010
YearRemodAdd
s(WoodDeckSF,0.31)
0
200
400
600
800
WoodDeckSF
s(TotalBsmtSF,6.81)
0
1000
2000
3000
4000
5000
6000
TotalBsmtSF
s(Exterior1st,5.99)
2
4
6
8
10
12
14
Exterior1st
s(GarageYrBlt,0.61)
1900
1920
1940
1960
1980
2000
GarageYrBlt
s(MasVnrArea,0.92)
0
500
1000
1500
MasVnrArea
s(X1stFlrSF,8)
1000
2000
3000
4000
X1stFlrSF
s(MSSubClass,5.83)
50
100
150
MSSubClass
s(LotFrontage,7.85)
50
100
150
200
250
300
LotFrontage
s(X2ndFlrSF,0.62)
0
500
1000
1500
2000
X2ndFlrSF
s(BsmtUnfSF,0)
0
500
1000
1500
2000
BsmtUnfSF
s(Exterior2nd,0)
5
10
15
Exterior2nd
s(YearBuilt,4.27)
1880
1900
1920
1940
1960
1980
2000
YearBuilt
s(GarageArea,8.19)
200
400
600
800
1000
1200
1400
GarageArea
s(GrLivArea,5.38)
1000
2000
3000
4000
5000
GrLivArea
s(Neighborhood,7.69)
5
10
15
20
25
Neighborhood
s(OpenPorchSF,3.76)
0
100
200
300
400
500
OpenPorchSF
s(BsmtFinSF1,7.93)
0
1000
2000
3000
4000
5000
BsmtFinSF1
s(LotArea,2.22)
0
50000
100000
150000
200000
LotArea

Figure 10 MARS

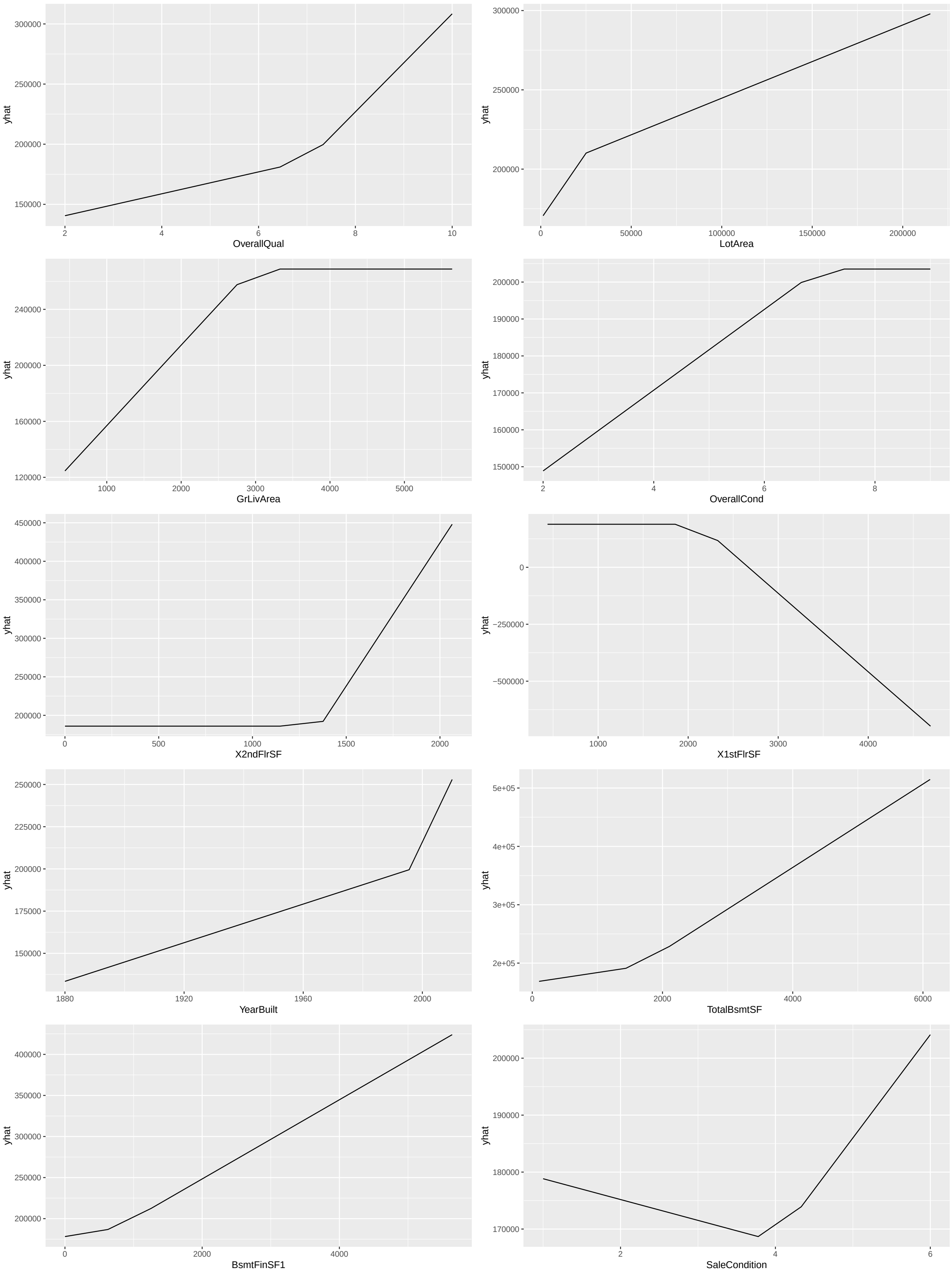


Figure 11 Box plot of all methods through CV

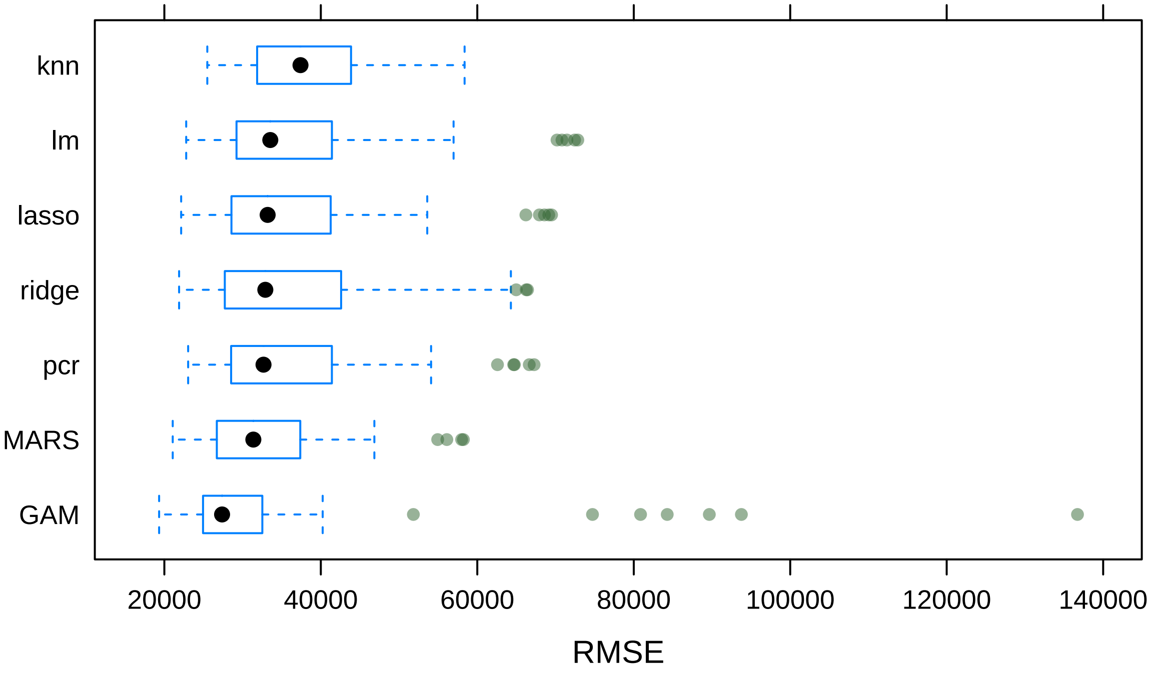


Figure 12 Box plot of all methods through CV

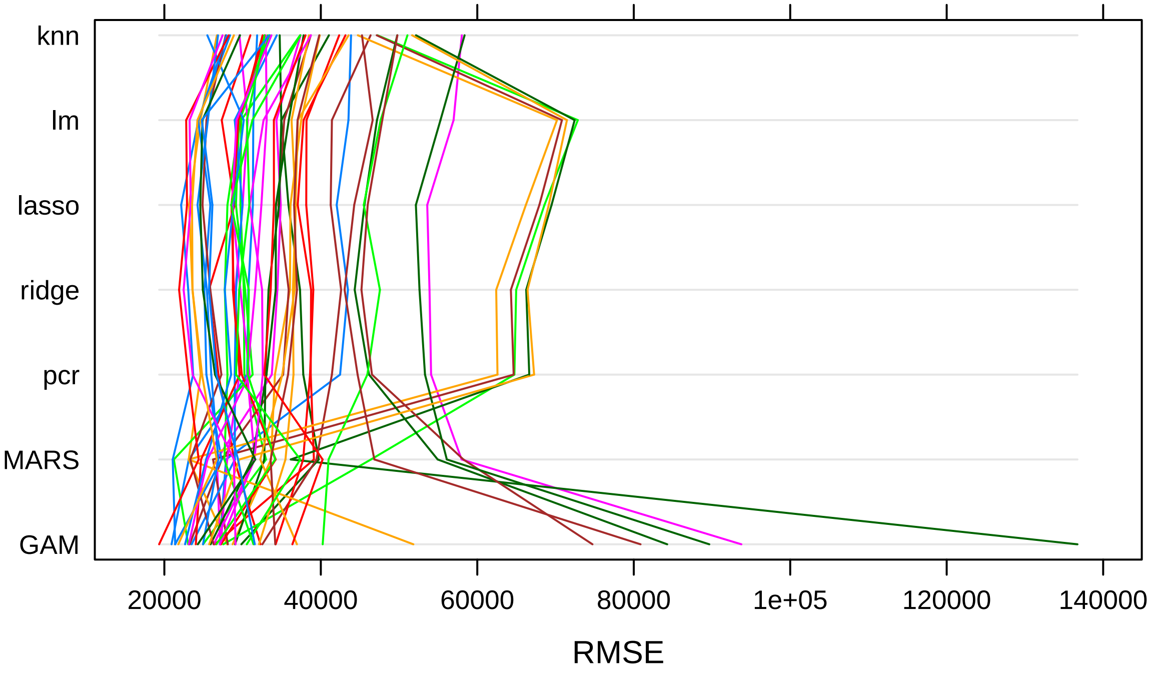
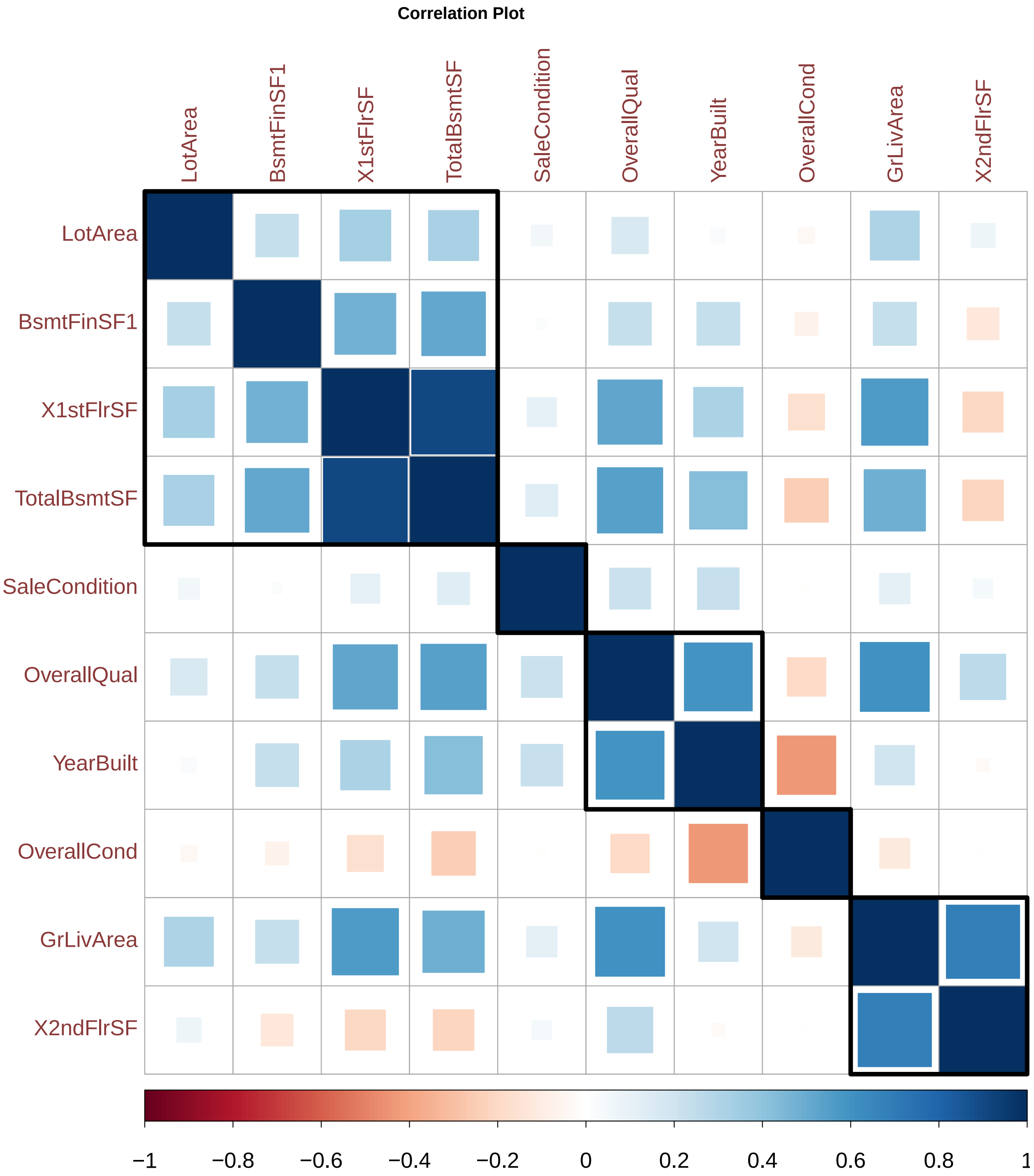


Figure 13 Heat map for top 10 most important variables


Correlation Plot
LotArea
BsmtFinSF1
X1stFlrSF
TotalBsmtSF
SaleCondition
OverallQual
YearBuilt
OverallCond
GrLivArea
X2ndFlrSF