Comparison of Statistical Learning and Predictive Models on Breast Cancer Data and King County Housing Data

Yunjiao Cai[1], Zhuolun Fu, Yuzhe Zhao, Yilin Hu, Shanshan Ding

Department of Applied Economics and Statistics, University of Delaware, Newark, DE 19716, U.S.A

ABSTRACT In this study, we evaluate the predictive performance of popular statistical learning methods, such as discriminant analysis, random forests, support vector machines, and neural networks via real data analysis. Two datasets, Breast Cancer Diagnosis in Wisconsin and House Sales in King County, are analyzed respectively to obtain the best models for prediction. Linear and Quadratic Discriminant Analysis are used in WDBC data set. Linear Regression and Elastic Net are used in KC house data set. Random Forest, Gradient Boosting Method, Support Vector Machines, and Neural Network are used in both datasets. Individual models and stacking of models are trained based on accuracy or R-squared from repeated cross-validation of training sets. The final models are evaluated by using test sets.

KEYWORDS. Machine learning, prediction, classification, regression, stacking.

1 INTRODUCTION

Statistical learning methods, such as discriminant analysis, random forests, support vector machines, and neural networks, are popular tools for classification and regression problems. The use of these methods is gradually increasing, especially in medical diagnosis and housing market. However, the classification or regression effects may be diverse by using various methods for the same data. In this study, we evaluate and compare the predictive performance of these methods based on two datasets in application. One is Breast Cancer Diagnosis in Wisconsin(WDBC), the other is House Sales in King County (KC house). We further employ model combining (stacking) techniques to achieve more robust prediction results. In the following, we give brief background regarding the two datasets.

[1] Corresponding author. E-mail: yjcai@udel.edu

1.1 WDBC dataset

Breast cancer is one of the major health problems for women around the world and it is alone to be expected to account for 30% new cancer diagnoses in women in 2017.[1] While breast cancer is easier to treat successfully when found early, detecting the cancer and predicting whether the cancer type is benign or malignant from various results of breast cancer screen is still a big challenge. The past several years has witnessed the rapid development of computational platform that allows us to use machine learning techniques to learn and diagnose tumor based on past diagnosis of patients. Numerous studies have been done on breast cancer using Wisconsin Diagnostic Breast Cancer (WDBC) dataset from the machine learning repository at the University of California at Irvine. Different techniques are developed to achieve accurate diagnosis results.

S. V. G. Reddy et al. [2] used Support Vector Machine (SVM) with polynomial and Radial Basis functions as kernels and achieved 97.13% accuracy. A. M. Elsayad et al.[3] compared results from SVM with decision tree algorithms which gives accuracy of 96.7%. S. Bagui et. al[4] also compared the performances among linear, quadratic, logistic, k nearest neighbor, and k rank nearest neighbor rules and achieved 94.2% accuracy rate. Another work[5] used multilayer neural network and obtained 92% accuracy. R. Alyami et al.[6] proposed to use SVM and artificial neural network with feature selection which gives accuracy of 97.1% and 96.7%, respectively. Another study[7] explored SVM with recursive feature elimination and principal component analysis.

While these previous works achieved decent accuracy values through various approaches, we propose to use alternative methods such as random forest, gradient boosting method, discriminant analysis, and stacking techniques to explore WDBC dataset and obtain a more robust model for prediction.

1.2 KC house dataset

In one's lifetime, the most expensive and largest purchase that the person makes is usually a home. It is very important for people to know the reasonable value of their asset. Prediction on house price will help both homeowners and homebuyers to make decisions of whether to sell or buy a house at a certain price. However, it is often difficult to determine the price of a house, as there are many factors involved, such as the age of the house, environment, location etc. In this work, we will apply several regression and predictive

methods to study house sale price in King County, Washington, USA and explore the best model for prediction.

Several researchers and teams have explored the KC house dataset. For example, feature ranking with Random Forest, RFE, and linear models was studied, and linear models were evaluated in some works. Multiple regression, lasso regression and k-Nearest Neighbors Regression were also investigated.

While previous works have shown compelling results, the R-squared values (often <0.9) might be further improved. In this study, we employ several regressions and machine learning techniques, as well as a model stacking (combining) approach to assess their prediction performance and to obtain a model that is best for prediction within our framework.

The rest of the article is organized as following. Section 2 presents brief introduction to the methods we use. Section 3 gives data description and exploratory analysis of the two datasets. Experimental results and the proposed analysis approaches are discussed in detail in Section 4 while Section 5 concludes.

2 Methods

2.1 k-Fold cross validation

k-Fold cross validation[8] is one popular way for model selection and tuning parameter determination. In this method, the dataset is divided into k nearly equaled subsets and each subset is left out but the remaining k-1 subsets are involved in training model. Each trained model will be evaluated using the subset that was left out. The overall performance is the averaged evaluation statistics from k trained models.

2.2 Linear Regression

Linear regression[9] assumes that there is approximately a linear relationship between dependent variables and independent variables. The coefficient for each independent variable can be estimated by a least square approach.[10] It is considered to be the one of the simplest while most efficient regression methods.

2.3 Elastic Net Regression

The elastic net is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods. It can be particular useful when the sample size is

much smaller than the number of predictors. Moreover, a grouping effect is encouraged and the model tends to include or exclude strongly correlated predictors together. It is believed to outperform the lasso and is a valuable tool for model fitting. For more details of this method, we refer to [11].

2.4 Random Forest

Usually, we may first consider using bagged trees method. [12] In bagging, each tree is independently constructed by bootstrap sample of the entire dataset and the majority votes from trees will be taken for prediction. However, random forest provides an improvement over bagged trees by way of a random small tweak that de-correlates the trees. Random forest is an ensemble learning method that operates by constructing a multitude of decision trees which only use a subset of all the predictors and averaging their outputs to obtain a single low-variance statistical learning model.[10] In addition to constructing each tree using a different bootstrap sample of the data, in a random forest, each node is split using the best among a subset of predictors randomly chosen at that node.[13] Random forest can lead to a substantial reduction in variance by forcing each split to consider only a subset of the predictors.[10] Thus it has advantage of robustness and resistance to overfitting. Overall, random forest is usually efficient in giving a well-perfumed model in a less time and less computationally consuming manner.

2.5 Gradient Boosting Method

The idea of boosting is to obtain an efficient classifier by combining several called inefficient classifiers.[14] Gradient Boosting Method (GBM) with trees can improve the predictions resulting from a decision tree. For boosting, each tree is built on a bootstrap data set, independent of the other trees, which are grown sequentially: each tree is grown using information from previously grown trees. Each tree is fit on a modified version of the original dataset.[10] From our experience, well-tuned boosting trees usually perform better than a random forest. There are several important parameters in this method: n.trees is number of trees. Interaction.depth is the number of split nodes. The shrinkage parameter λ, a small positive number, controls the rate at which boosting learns.

2.6 Support Vector Machines

Support vector machine(SVM) is a generalization of a simple and intuitive classifier called the maximal margin classifier, which represents the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.[15] When using SVM,[16] we can choose linear kernel as well as nonlinear kernels, like polynomial kernel and radial kernel. We can easily conduct different kernels by changing different values of parameters. Cost "C" is a general penalizing parameter for C-classification. Degree "d" is a parameter for determining a linear kernel or a polynomial kernel. A kernel with d = 1 is a standard linear kernel. A kernel with d > 1 in the support vector classifier leads to a much more flexible decision boundary. Gamma is the free parameter of the radial kernel. Technically speaking, large gamma usually leads to high bias and low variance models, and vice-versa.

2.7 Neural Network

Neural network[17] was inspired by human brain's neural network. It is "a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs."[18] The model of neural network consists of input, hidden, output layers. The idea of neural networks is that each node in the hidden layer is a function of the nodes in the previous layer, and the output node is a function of the nodes in the Hidden layer.

2.8 Discriminant Analysis

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA)[19] are commonly used supervised learning method for classification problems. They are competitors of logistic regression. It is well known that when the classes of the dependent variable are well-separated, the parameter estimates from the logistic regression model are surprisingly unstable. LDA and QDA do not suffer from this problem. The difference between LDA and QDA is: LDA requires an assumption of equal variance-covariance matrices (between the input variables) of the classes, while QDA is an extension of LDA that allows for heterogeneity of classes' covariance matrices.

2.9 Stacking

Stacking is a model ensemble technique and is the process of running two or more different analytical models and then synthesizing the results to generate a new model to improve the accuracy of predictive analytics. It involves combining multiple predictions derived by different techniques to create a stronger overall prediction. Often the stacking model (also called 2nd-level model) will outperform each of the individual models due to its smoothing nature and ability to highlight each base model where it performs best and discredit each base model where it performs poorly. For this reason, stacking is mostly effective when the base models are significantly different.

3 Data Description and Exploratory Analysis

3.1 WDBC dataset

In the dataset of Wisconsin Diagnostic Breast Cancer (WDBC) from the University of California – Irvine repository, there are 30 features computed from a digitized image of a fine-needle aspirate of a breast mass, which are all numerical variables along with 569 observations. The response variable is diagnosis which can be either one of two possible classes: malignant or benign.

We firstly plot the boxplots and density plots to examine the relationship between each predictor and the response variable. They are shown in Figure 3-1, Figure 3-2 and Figure 3-3. From the distribution plot, we find the observations in benign are more than those in malignant. From the correlation of predictors found in Figure 3-4, we can see that some predictors have high correlation with others, while many of others are uncorrelated.

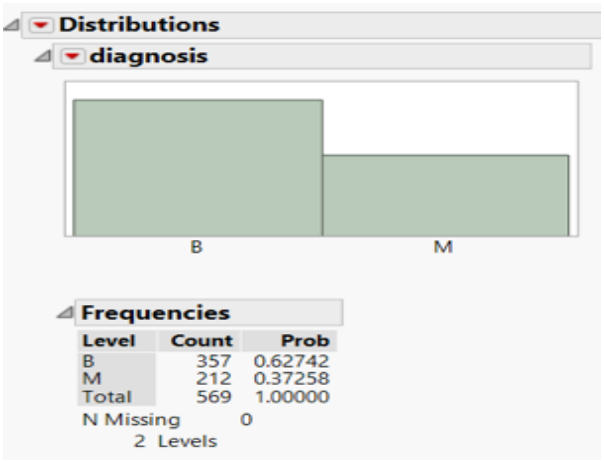

Figure 3-1 Distribution of B(benign), M(malignant) in WDBC dataset

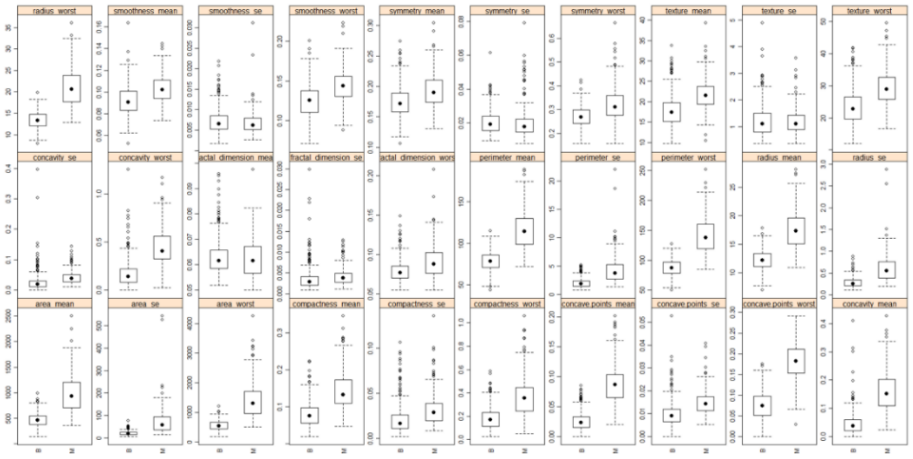

Figure 3-2 Boxplots of Predictors, group by diagnosis in WDBC dataset

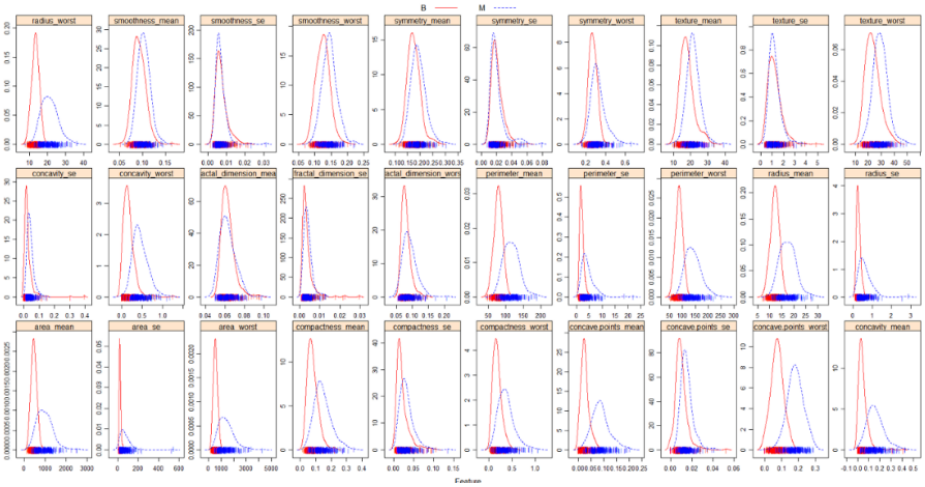

Figure 3-3 Predictors density plot, group by diagnosis in WDBC dataset

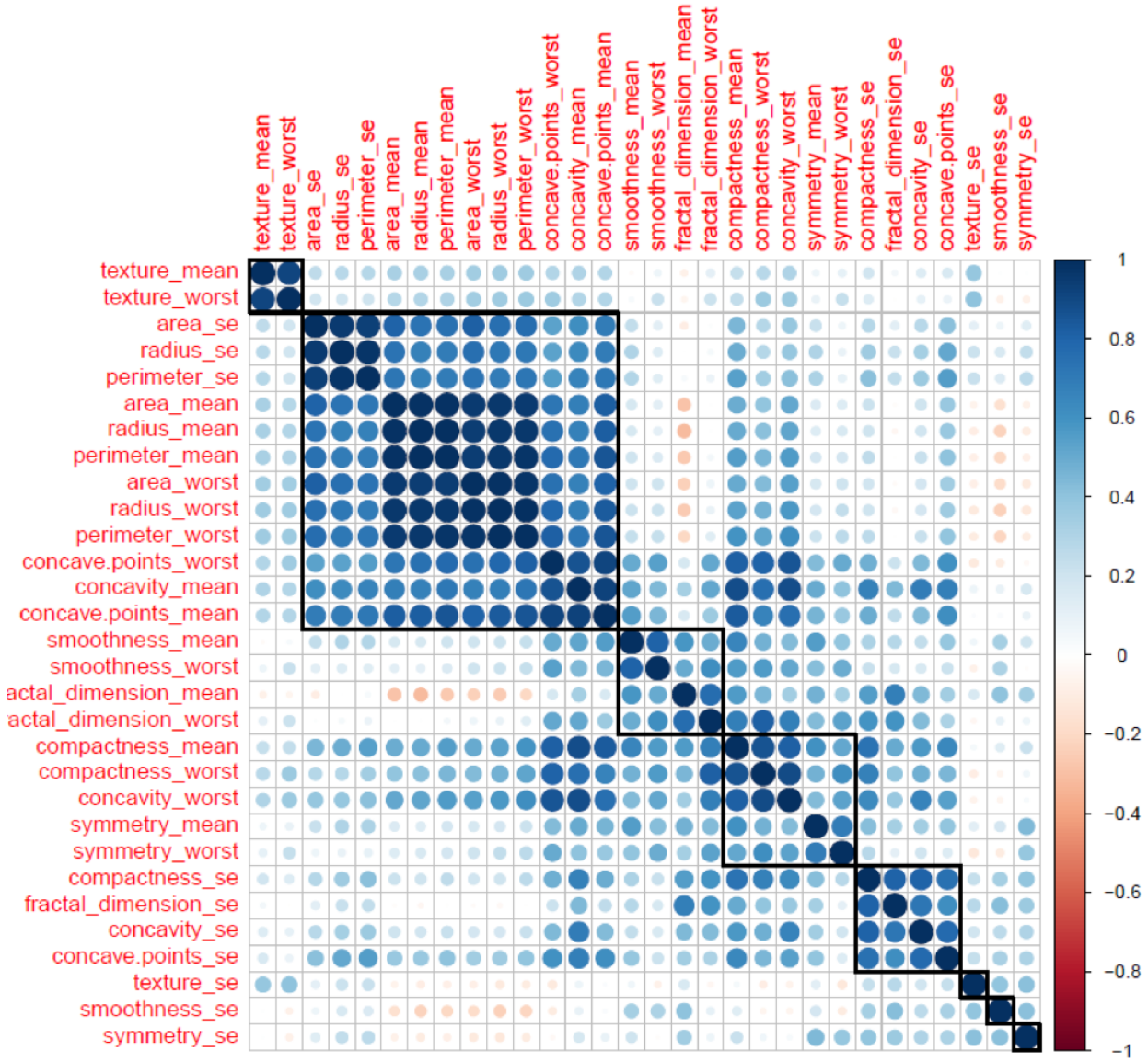

Figure 3-4 Correlation of predictors in WDBC dataset

In the WDBC dataset, variable "id" is just for identification of a patient, so we exclude it from analysis. Then the dataset is randomly partition into 70% training observations and 30% testing observations. The model selection is evaluated based on the training set and the final model is obtained with the best prediction performance for the testing set.

3.2 KC house dataset

The KC house dataset includes homes sold between May 2014 and May 2015 in King County, Washington. There are 19 house predictors plus the price and the id columns, along with 21613 observations. The response variable is price.

We first build scatterplot for each attribute versus price, see Figure 3-5 and Figure 3-6.

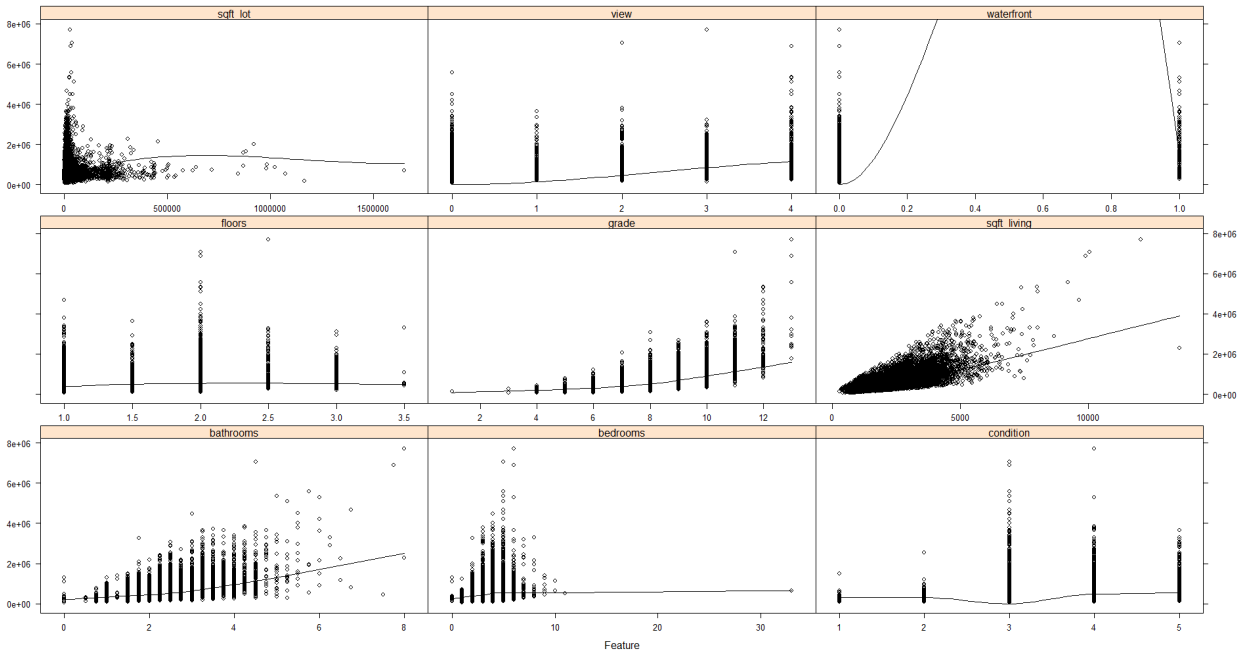

Figure 3-5 Scatterplots in kc house dataset

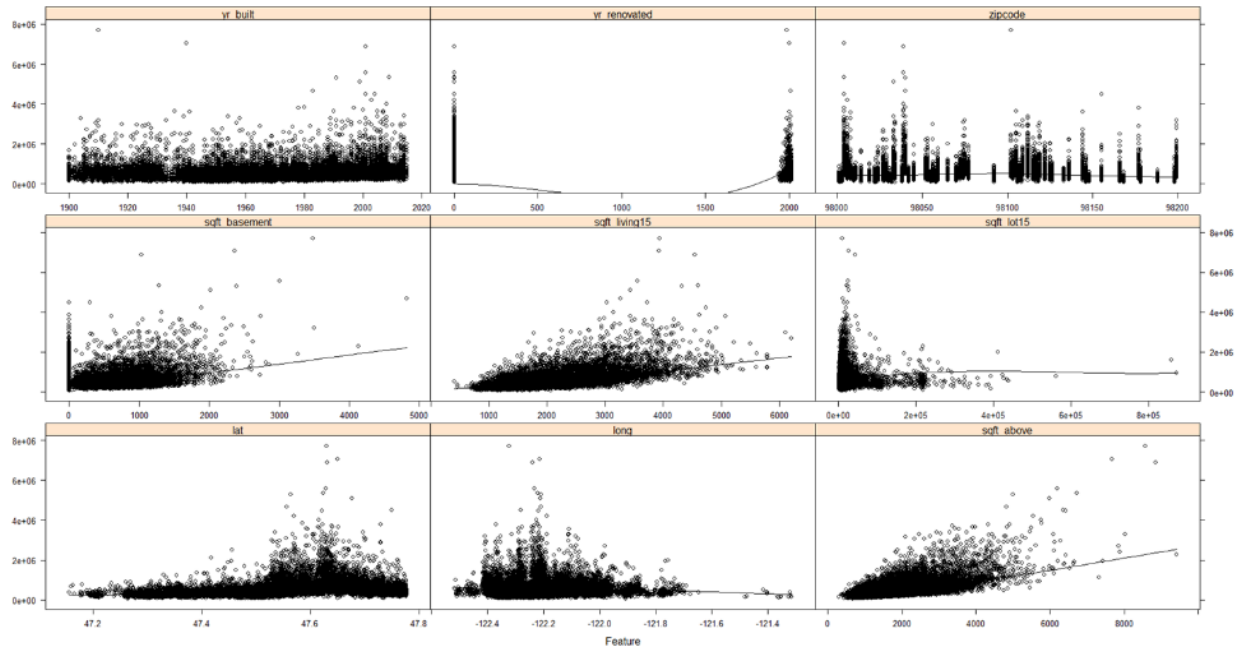

Figure 3-6 Scatterplots (continued) in kc house dataset

Some variables, such as view, floors, and grade, are categorical, but we treat them as numeric. The reason is that we believe the difference between each level is the same and we can keep the ordinal information. Note that “zipcode”, longitude and latitude are spatial variables.

The correlation matrix is shown in Figure 3-7. It indicates that some of variables are correlated.

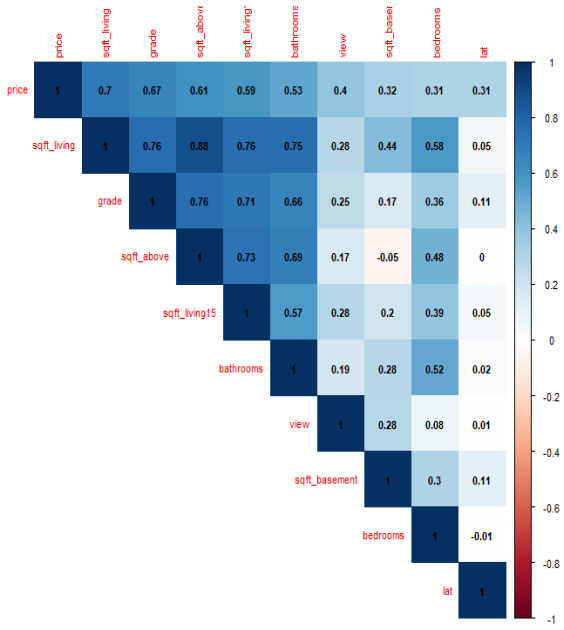

Figure 3-7 Correlation of specific variables in kc house dataset

We also notice that most houses are never renovated, as shown in Figure 3-8.

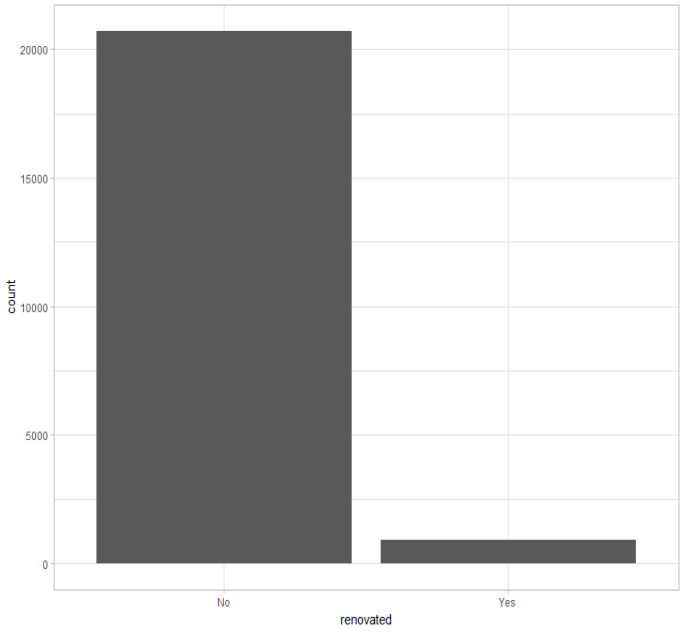

Figure 3-8 Histogram of renovated house in kc house dataset

Then we preprocess the data as followings. Variable “id” is just for identification of a specific house, so we exclude it from the study. We change variable “date” to the number of days since 1970.01.01 as numerical variable and exclude “sqft_living” in that it can be expressed by

"sqft_above+sqft_basement". "zipcode" is coded as categorical variable with 70 levels. As shown in Figure 3-9, it can be considered as 70 geometrical clusters, which may be related to "price".

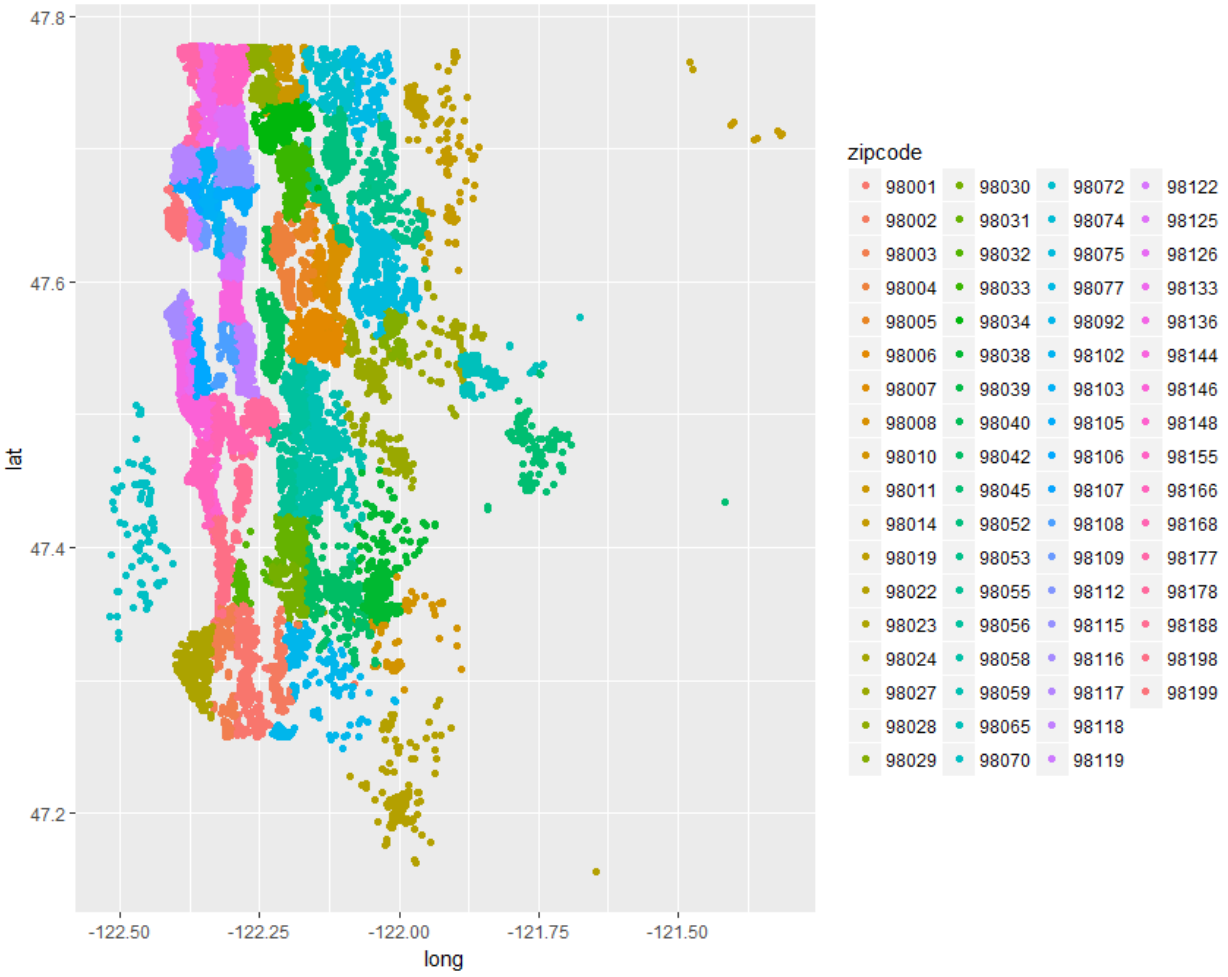

Figure 3-9 Cluster map for zipcode in KC house dataset

If a house is never renovated, we then set "yr_renoveted" equals to "yr_built". We use ln(price), a natural logarithm of "price", as the response variable so it is better normally distributed.

The dataset is randomly partitioned into 70% training observations and 30% testing observations. The model selection is evaluated based on the training set and the final model is obtained with the best prediction performance for the testing set. Since the training set still contains large numbers of observations that increases the computational cost in using some methods, we randomly choose 10% of the training observations and use them in parameter tuning in Support Vector Machines and Neural Network.

4 Statistical Modeling and Analysis

4.1 WDBC dataset

4.1.1 Random Forest

In random forest, it is not necessary to use external folds of cross-validation(CV) to get an unbiased estimate of the parameters, which can instead be estimated internally. So, we use the 10-repeated out-of-bag(OOB) CV to tune the parameters. We select the number of predictors (mtry) to be included in an individual tree and the best value is mtry=4, as shown in Figure 4-1. We also see from Figure 4-2 that 1000 trees are enough for error rate to saturate. The OOB CV and test accuracy are 0.966 and 0.936, respectively. As we can see, random forest can easily get a high accuracy for this dataset.

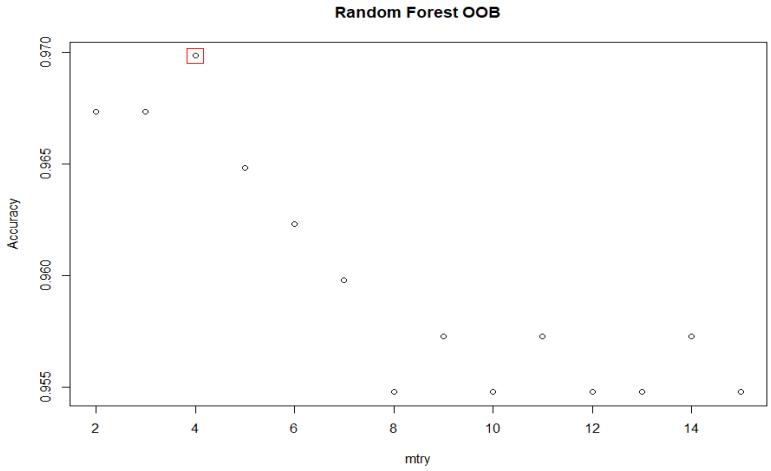

Figure 4-1 Out-of-bag accuracy vs. mtry in Random Forest in WDBC dataset

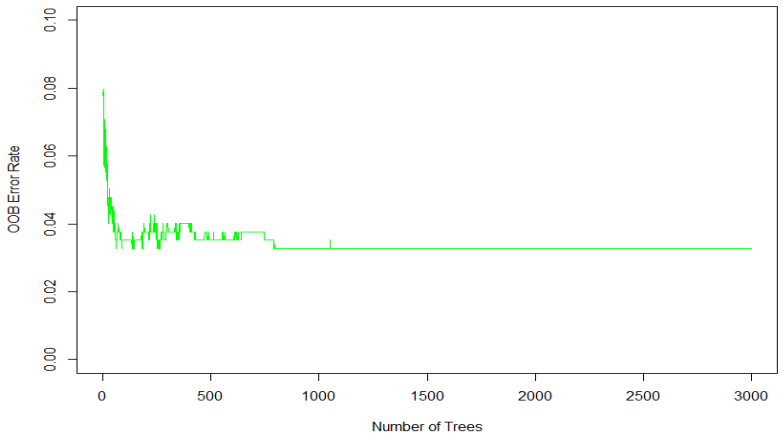

Figure 4-2 Out-of-bag Error Rate vs. Number of Trees in Random Forest in WDBC dataset

4.1.2 Gradient Boosting Method

From 10 repeated 10-folds CV, by trying different combinations, the best values for the following parameters are n.trees equals 2000, Interaction.depth is equal to 10. We choose the shrinkage parameter λ to be 0.03 and n.minobsinnode to be 10, as shown in Figure 4-3, Figure 4-4 and Figure 4-5. The training CV accuracy is 0.977 that outperforms random forest.

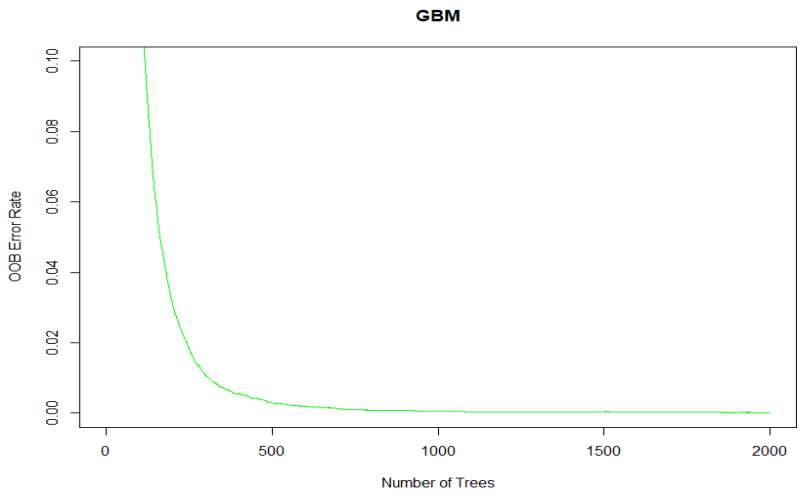

Figure 4-3 Out-of-bag Error rate vs. Number of Trees in GBM in WDBC dataset

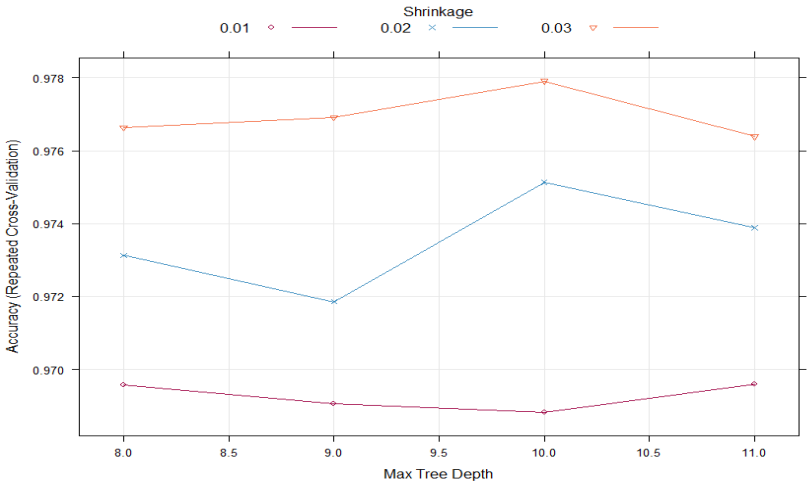

Figure 4-4 Curve plot of parameter tuning in GBM in WDBC dataset

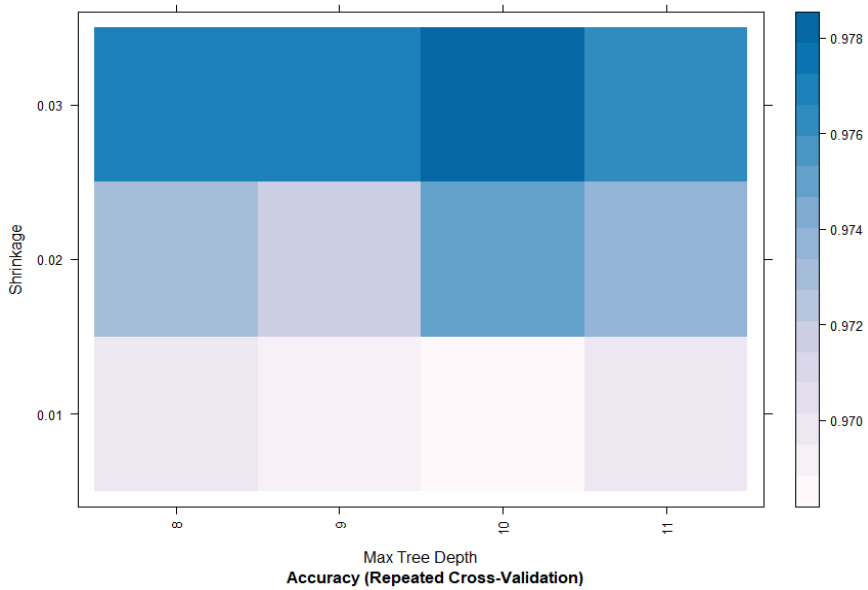

Figure 4-5 Heat map of parameter tuning in GBM in WDBC dataset

4.1.3 Support Vector Machines

Before conducting SVM models, we center and scale the dataset. From 10 repeated 10-folds CV, for SVM with linear kernel, the best values for cost is 0.5, as shown in Figure 4-6.

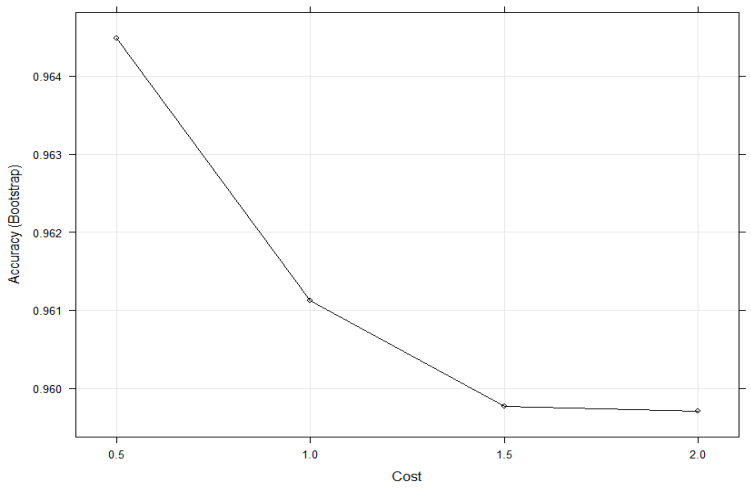

Figure 4-6 Accuracy vs. Cost in SVM with linear kernel in WDBC dataset

For SVM with polynomial kernel, the best values for the following parameters are Poly degree = 4, scale = 0.05 and gamma =0.3, as shown in Figure 4-7 and Figure 4-8.

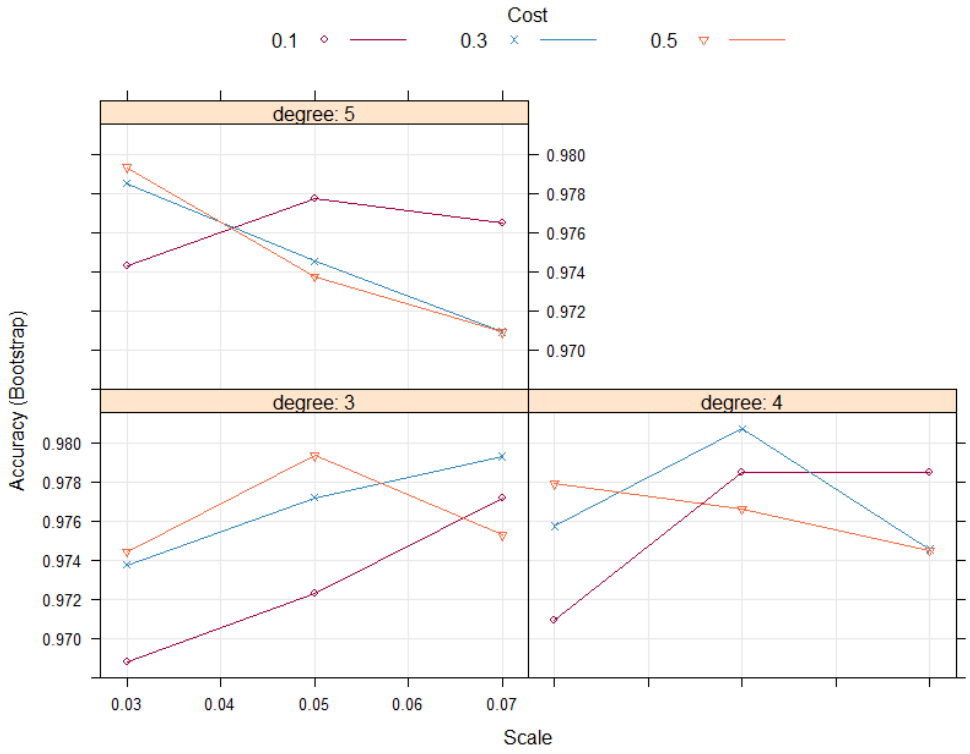

Figure 4-7 Curve plots of Accuracy vs. scale, grouped by degree in SVM with polynomial kernel in WDBC dataset

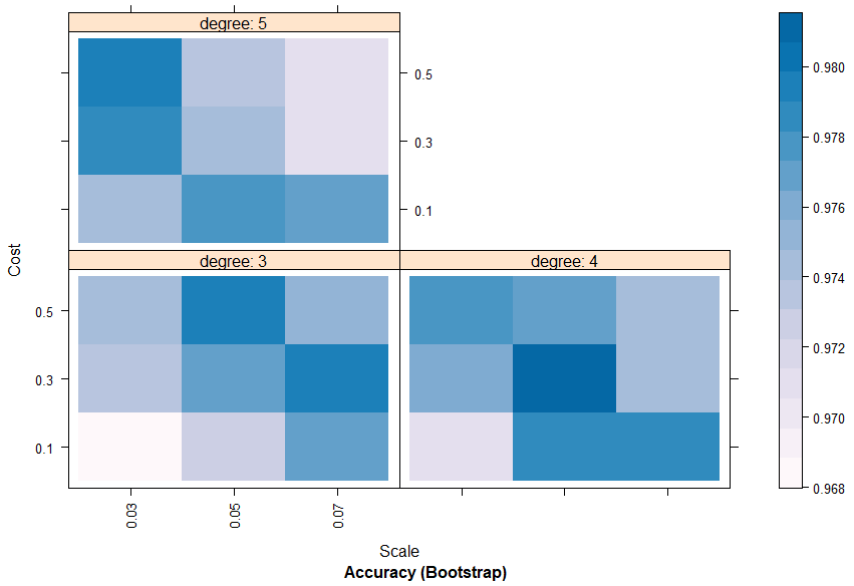

Figure 4-8 Heat map of parameter tuning in SVM with polynomial kernel in WDBC dataset

On the other hand, the best values of SVM with radial kernel for the following parameters are sigma = 0.03 and cost = 8, as shown in Figure 4-9 and Figure 4-10.

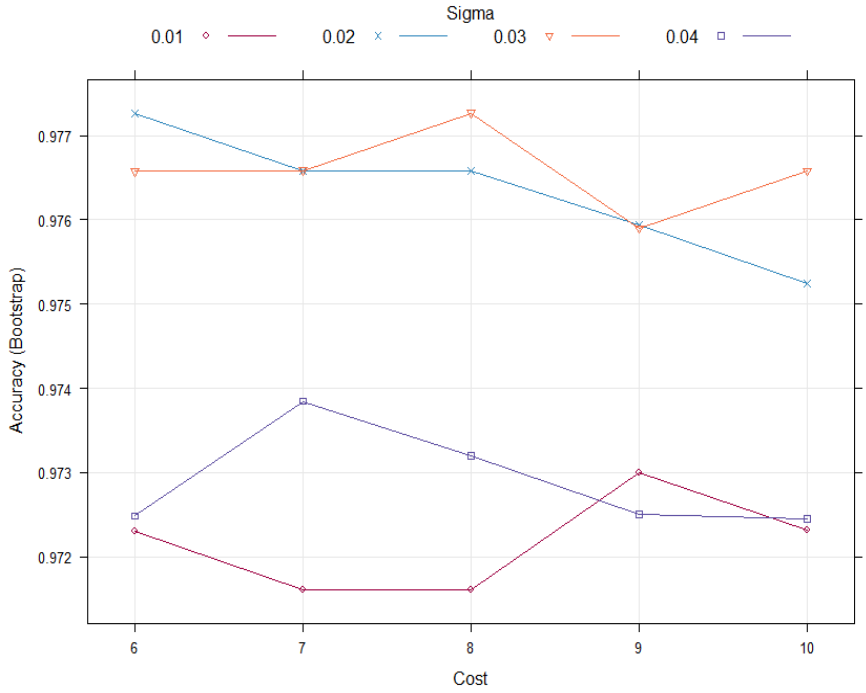

Figure 4-9 Curve plots of parameter tuning in SVM with radial kernel in WDBC dataset

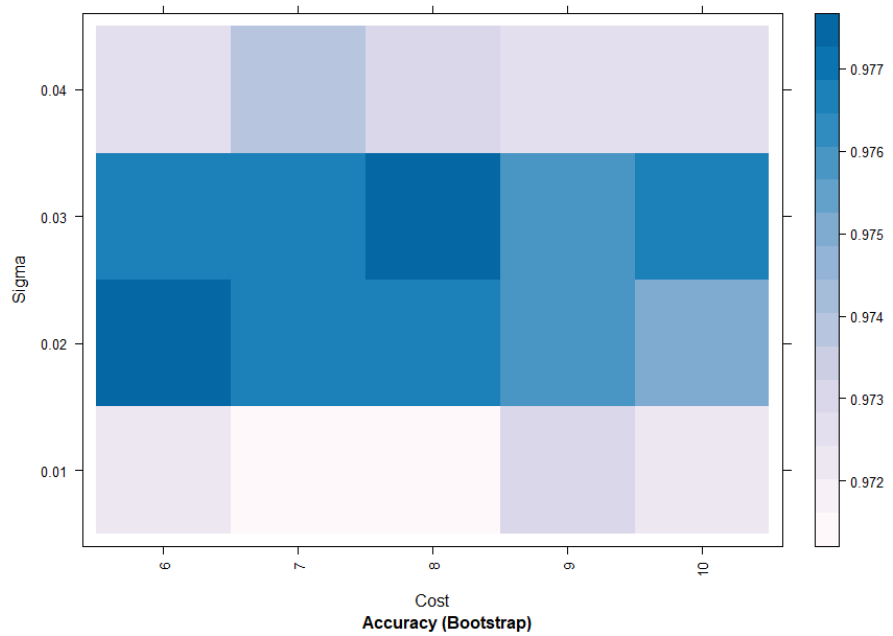

Figure 4-10 Heat map of parameter tuning in SVM with radial kernel in WDBC dataset

All the SVM models have high CV accuracy. They are robust, flexible and have other principles of operation than random forest. They can be candidates for the best model.

4.1.4 Neural Network

Before conducting neural networks, we center and scale the dataset. For Neural Network, we choose to use only one hidden layer and based on 10 repeated 10-folds CV, the best values for the following parameters are: decay, which is the parameter for weight decay, is 1.2, size, which is the number of units in hidden layers, is 6, determined based on the prediction accuracy, as shown in Figure 4-11.

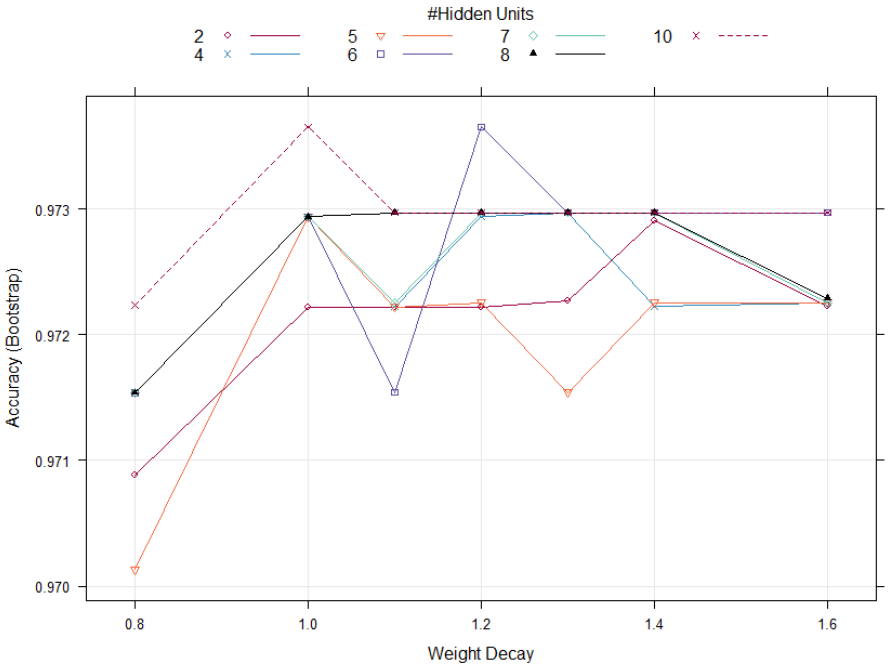

Figure 4-11 Curve plots of Accuracy in Neural Network in WDBC dataset

4.1.5 Discriminant Analysis

Based on the 10 repeated 10-folds CV, for LDA, the training CV accuracy is 0.958. For QDA, the training CV accuracy is 0.960. QDA performs slightly better than LDA.

4.1.6 Stacking

So far, the above models we use all achieve relatively high accuracy, as summarized in Figure 4-12. We consider the prediction from above selected individual models as independent variables and use linear regression to find a good linear combination of these models and to potentially better predict the response.

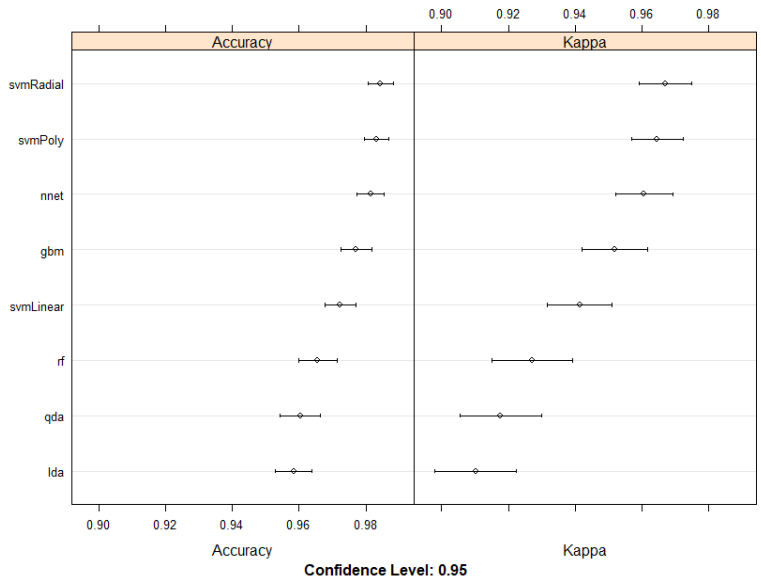

Figure 4-12 Comparison of individual models used in WDBC dataset

We first check their correlation, as shown in Figure 4-13.They are mostly not correlated.

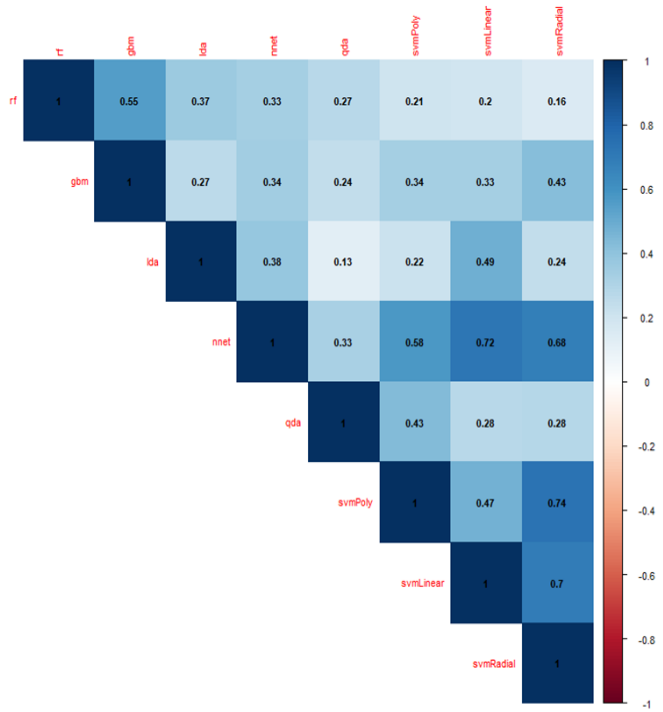

Figure 4-13 Correlation table of models used in WDBC dataset

Based on the 10 repeated 10-folds CV, the CV accuracy of the linearly ensemble model is 0.9887. The result comparing the individual models and linearly ensemble model is shown in Figure 4-14.

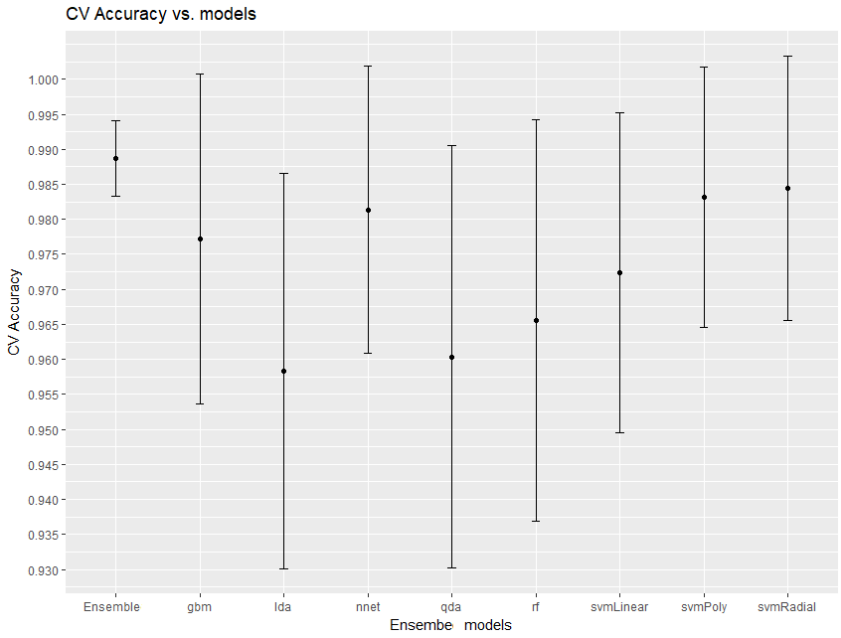

Figure 4-14 CV accuracy comparison between individual models and the linearly ensemble model in WDBC dataset

As a result, the CV accuracy from the linearly ensemble model is higher, and its uncertainty is smaller than any other individual models. It indicates that stacking may give a better and more robust prediction. Then we consider using stacking by random forest, SVM with kernels, and neural network.

For stacking by fandom forest, the optimal parameters are as following: mtry=3, ntree=200. The CV accuracy is 0.9924, as shown in Figure 4-15 and Figure 4-16.

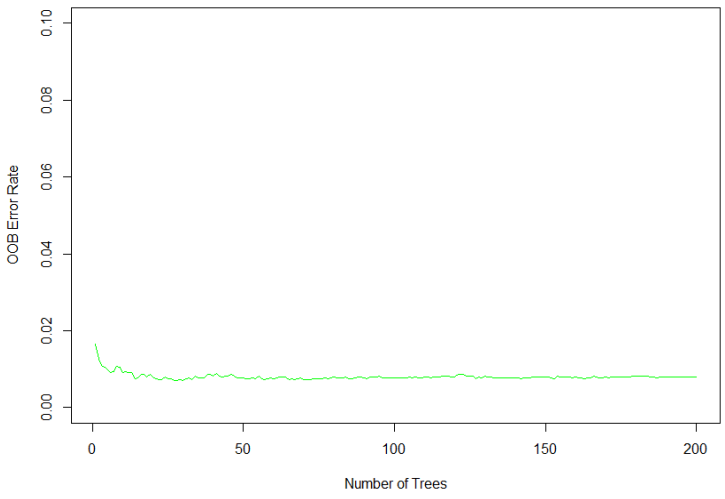

Figure 4-15 Out-of-bag error rate vs. number of trees in stacking model by random forest in WDBC dataset

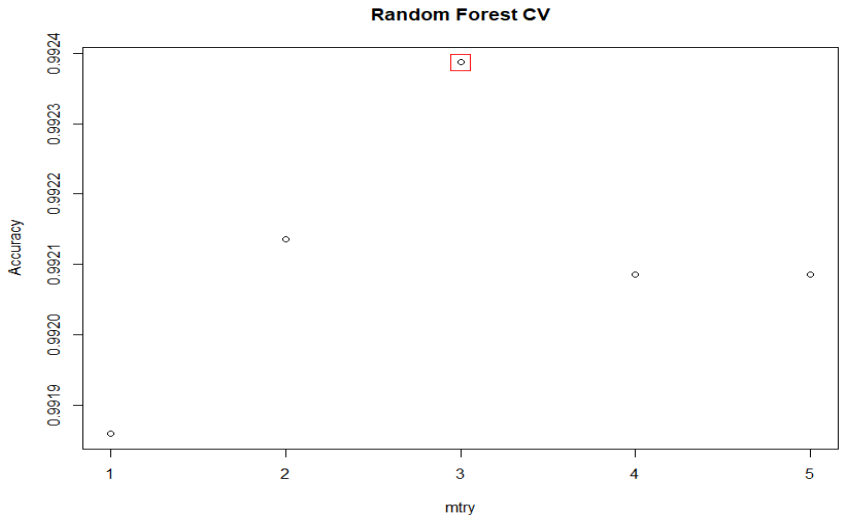

Figure 4-16 CV accuracy vs. mtry using random forest in stacking model in WDBC dataset

For SVM with linear kernels, cost=3. CV accuracy=0.9881, as shown in Figure 4-17.

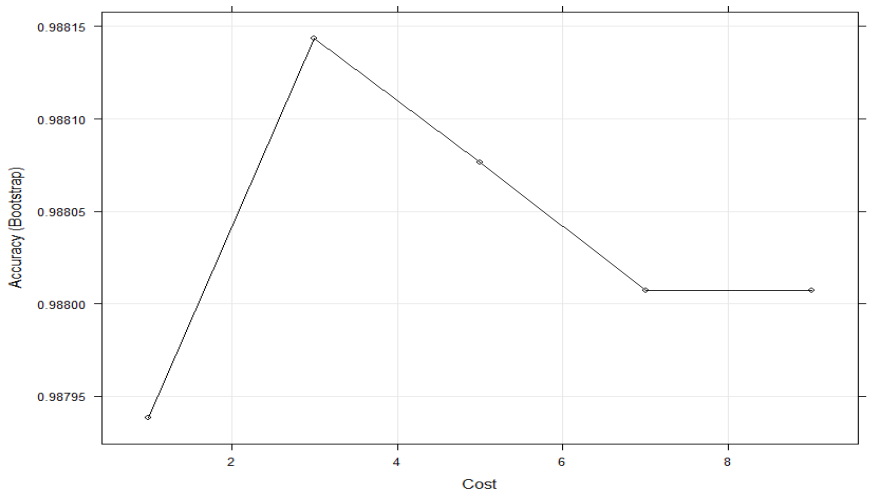

Figure 4-17 CV accuracy vs. cost using SVM with linear kernel in stacking model in WDBC dataset

For SVM with polynomial kernels, d=4, gamma=0.5, cost=30. CV accuracy=0.9917, as shown in Figure 4-18 and Figure 4-19.

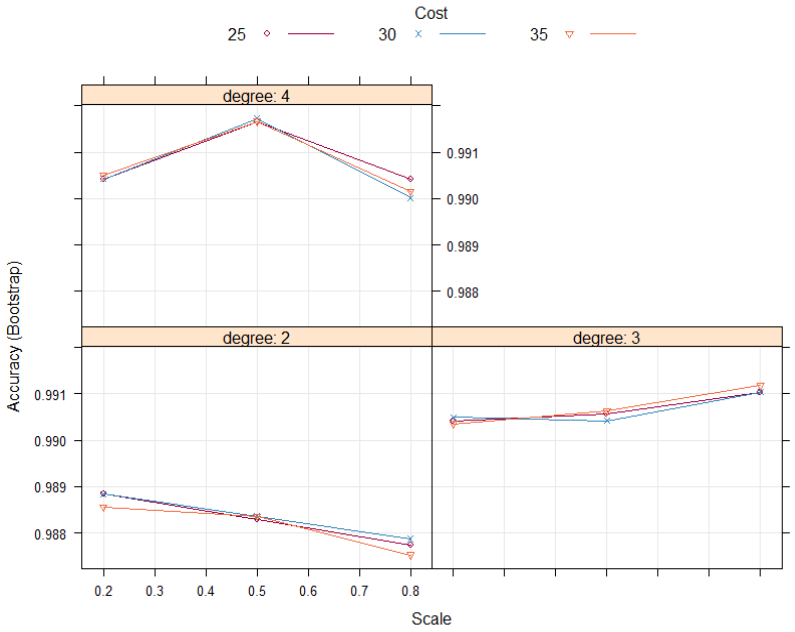

Figure 4-18 Curve plots of CV accuracy using SVM with polynomial kernel in stacking model in WDBC dataset

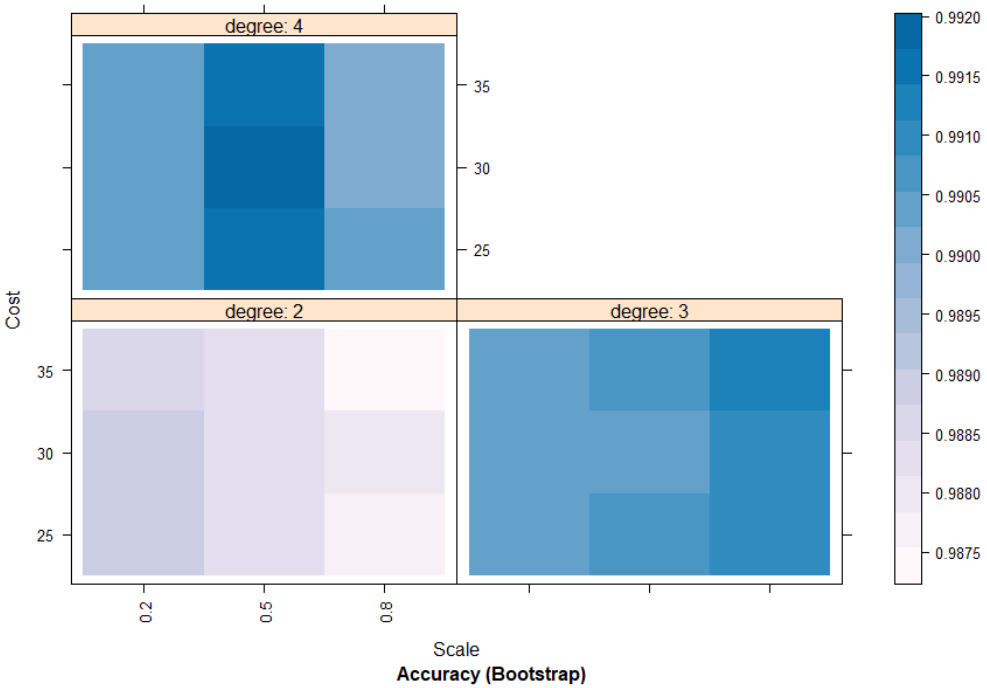

Figure 4-19 Heat map of CV accuracy using SVM with polynomial kernel in stacking model in WDBC dataset

For SVM with radial kernel, gamma=3, cost=3. CV accuracy=0.9936, as shown in Figure 4-20 and Figure 4-21.

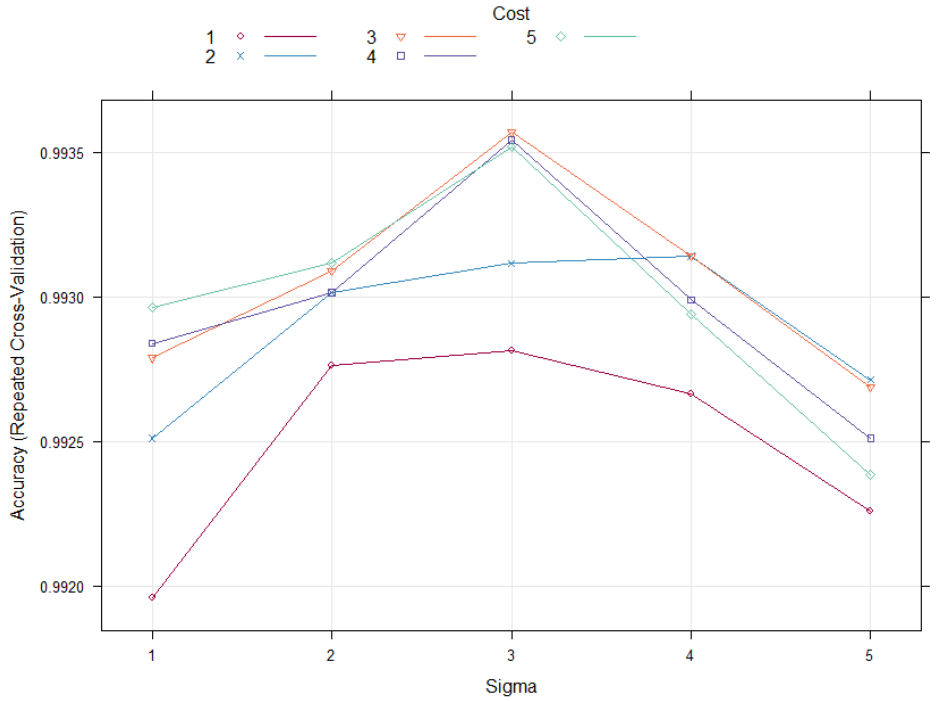

Figure 4-20 Curve plots of CV accuracy using SVM with radial kernel in stacking model in WDBC dataset

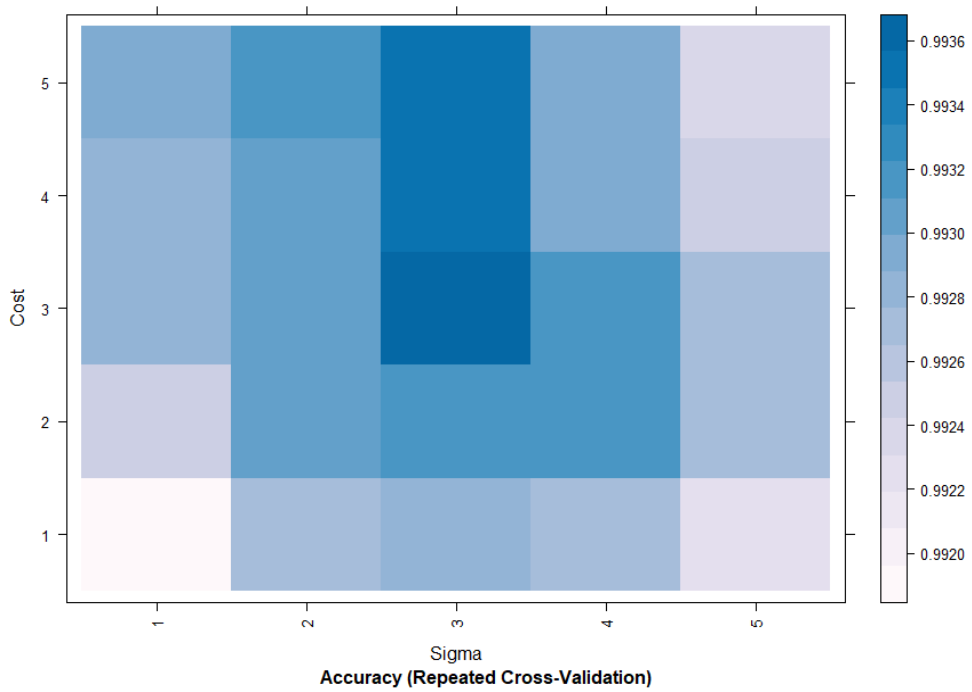

Figure 4-21 Curve plots of CV accuracy using SVM with radial kernel in stacking model in WDBC dataset

For Neural Network, decay=0.01, the only one hidden layer's size=9. CV accuracy=0.9930, as shown in Figure 4-22 and Figure 4-23.

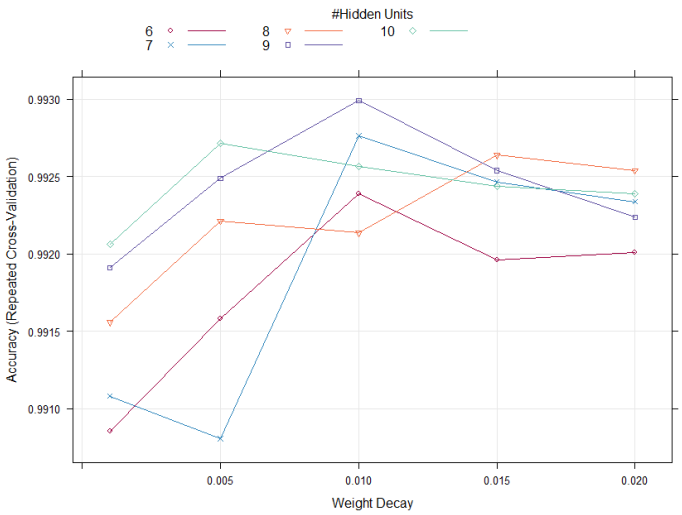

Figure 4-22 Curve plots of CV accuracy using Neural Network in stacking model in WDBC dataset

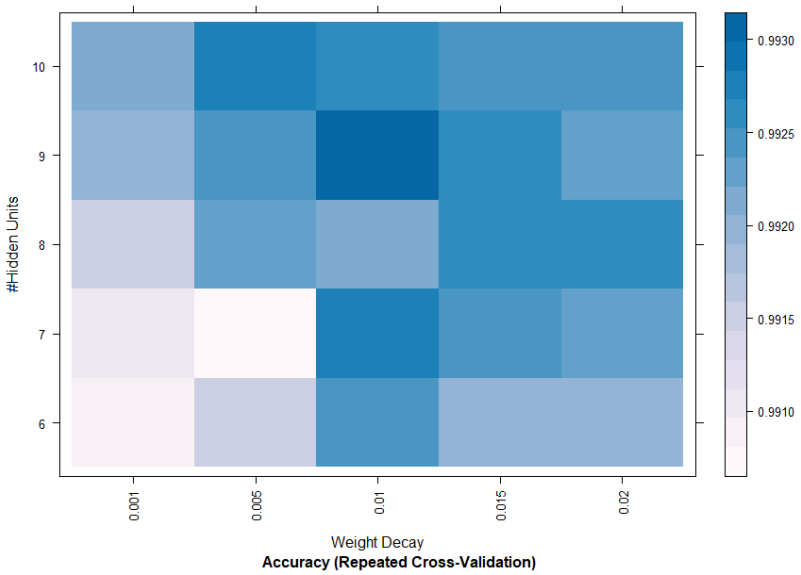

Figure 4-23 Heat map of CV accuracy using Neural Network in stacking model in WDBC dataset

Overall, the best stacking model evaluated by CV is stacking by SVM with radial kernel, whose CV accuracy is 0.9936.

4.2 KC house dataset

4.2.1 Linear Regression

We implement linear regression using "ln(price)" as a response variable and all other variable as predictors. From the 5-folds cross-validation (CV) with 5 repeats using different seeds, the CV R-squared is 0.878, which is relatively high considering the complexity of the dataset and simplicity of the model. We check the diagnostic plots as shown in Figure 4-24. There are no obvious outliers with high leverage and the residual plot shows little nonlinear pattern.

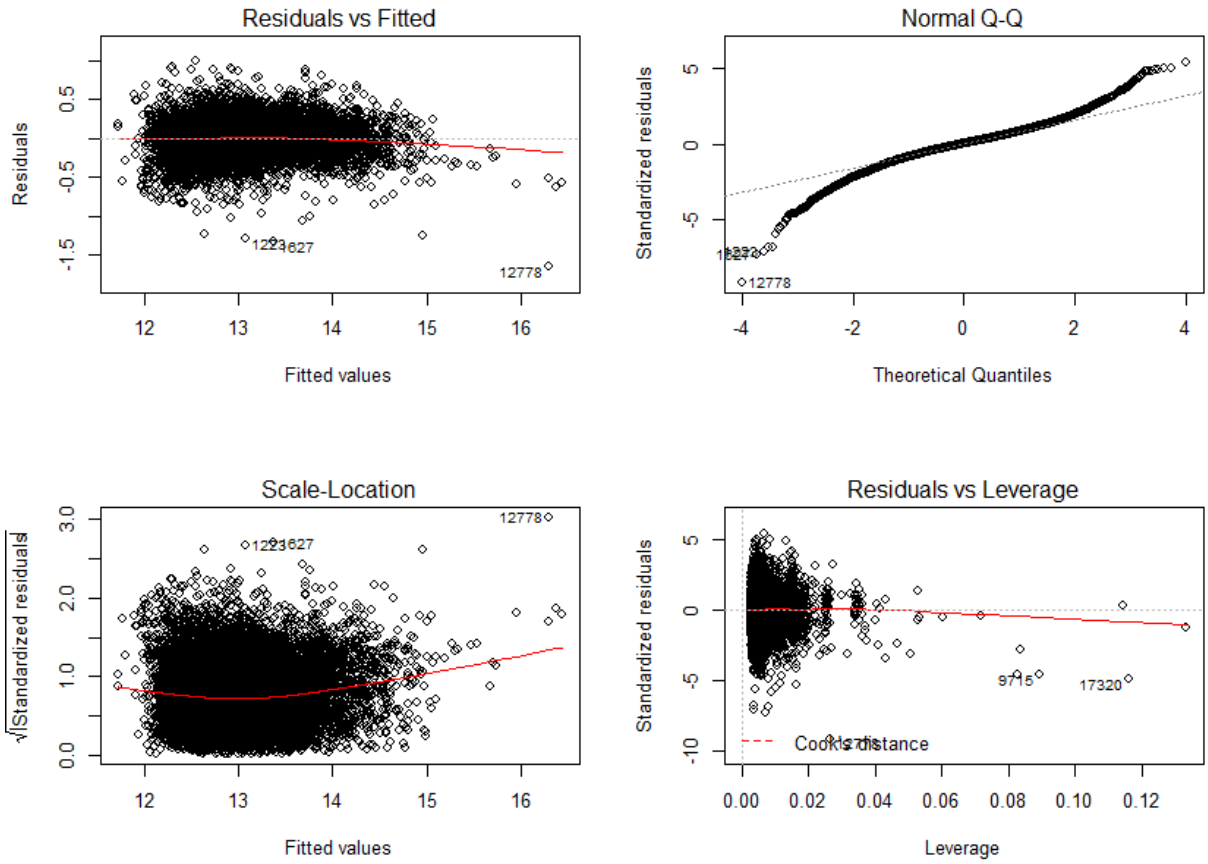

Figure 4-24 Diagnostic plots for linear model of ln(price) vs. predictors in KC house dataset

The linear model is computationally favorable and good for interpretation. Thus the linear model might be one of the candidates for the best model.

4.2.2 Elastic Net Regression

We use 5 repeated 5-folds CV to tune the regulation parameter "lambda" and mixing percentage parameter "alpha". The best values are alpha =0.9 and lambda=8.86e-5, as shown in Figure 4-25 and Figure 4-26. The CV R-squared is 0.877, which is close to that in the linear model.

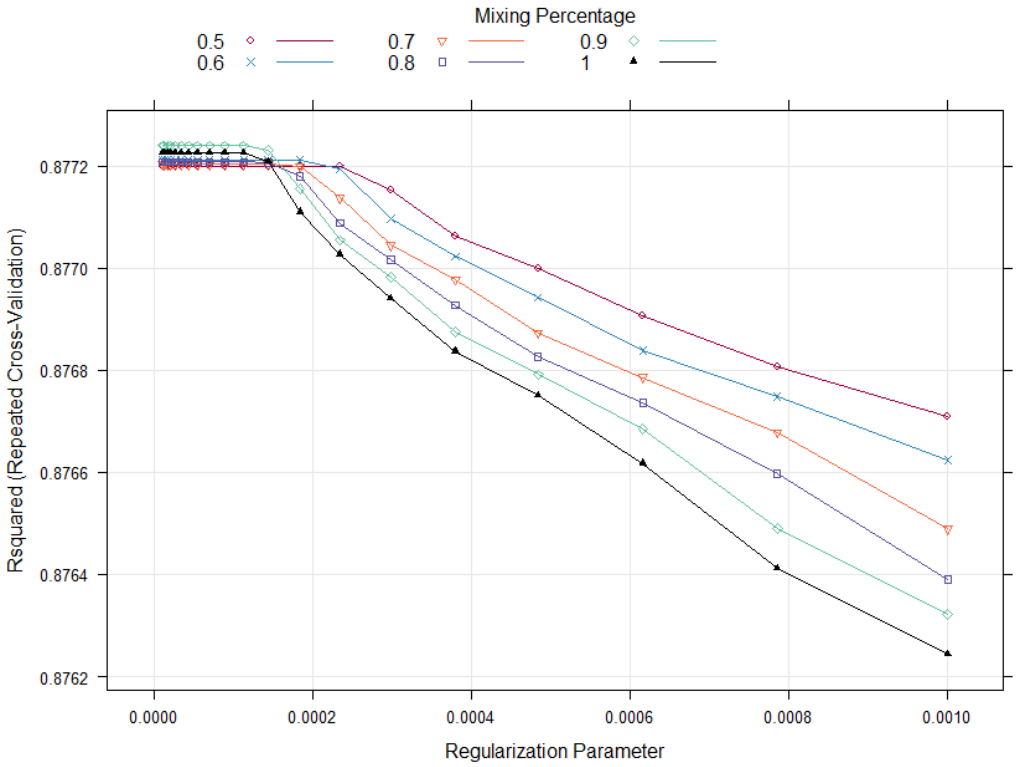

Figure 4-25 Curve plot of parameter tuning using Elastic Net in KC house dataset

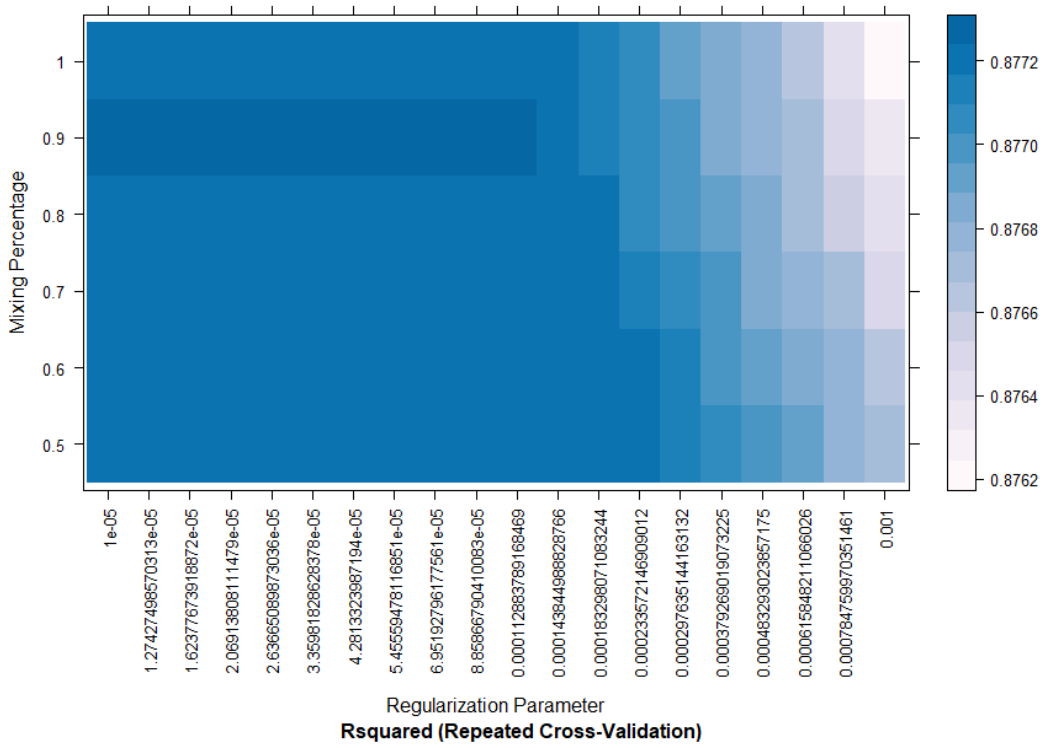

Figure 4-26 Heat map of parameter tuning using Elastic Net in KC house dataset

4.2.3 Random Forest

We use 5 repeated out-of-bag(OOB) CV to tune how many predictors (mtry) to be included in an individual tree and the best value is mtry=30 as shown in Figure 4-27.

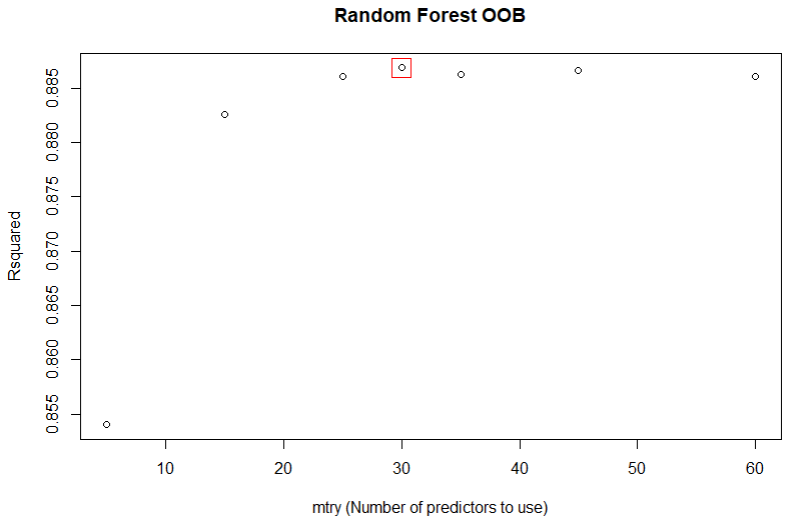

Figure 4-27 Dot plot of parameter tuning using Random Forest in KC house dataset

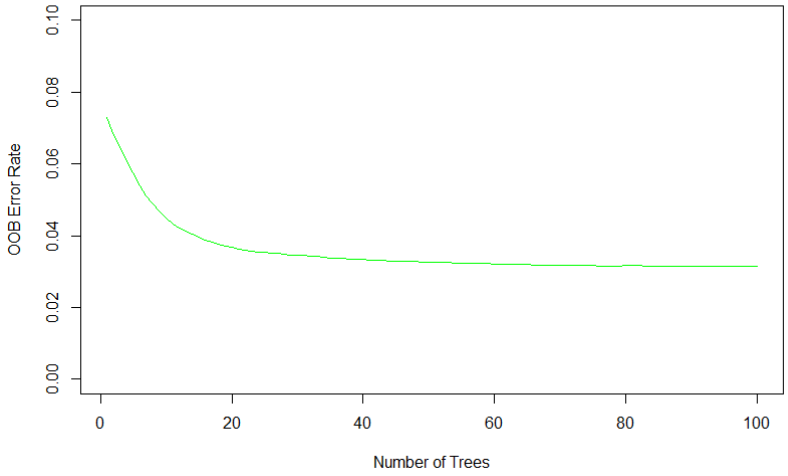

Figure 4-28 Out-of-bag Error Rate vs. Number of Trees using Random Forest in kc dataset

We also see from Figure 4-28 that 100 trees are enough for error rate to saturate. The 5 repeated 5-folds CV R-squared is 0.888, The prediction performance is improved compared to the previous models.

4.2.4 Gradient Boosting Method

From 5 repeated 5-folds CV, as shown in Figure 4-29, n.trees = 1500 is enough and the best values of parameters are interaction.depth = 12, shrinkage = 0.03 and n.minobsinnode = 10, as

shown in Figure 4-30 and Figure 4-31. The CV R-squared is 0.907. Thus, the gradient boosting method is also a good candidate for the best model.

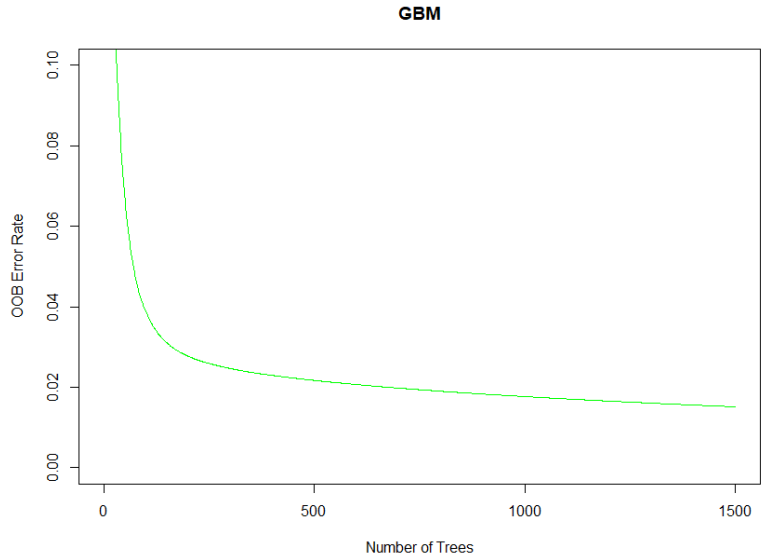

Figure 4-29 Out-of-bag Error Rate vs. Number of Trees using Boosting Trees in kc dataset

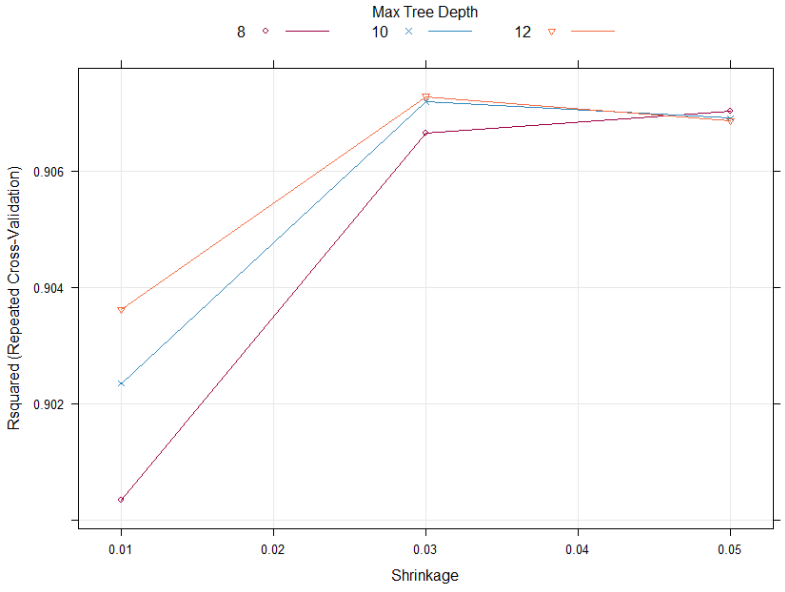

Figure 4-30 Curve plot of parameter tuning using GBM in kc house dataset

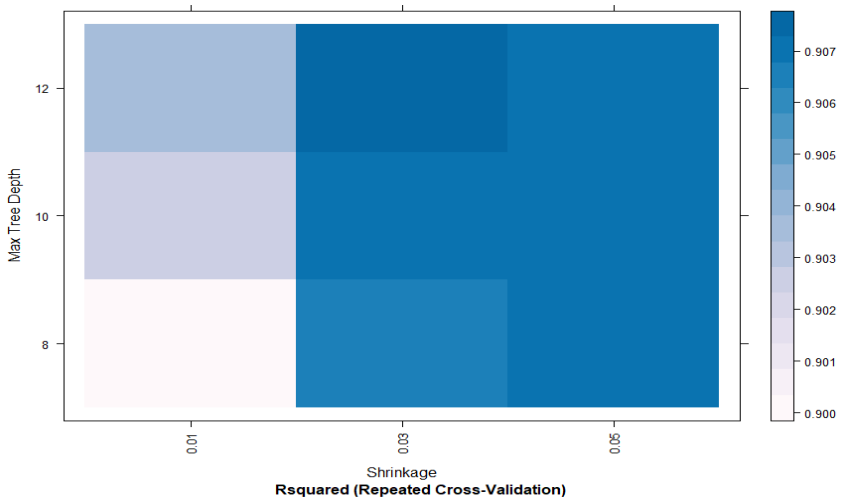

Figure 4-31 Heat map of parameter tuning using GBM in kc house dataset

4.2.5 Support Vector Machines

From 5 repeated 5-folds CV, for SVM with polynomial kernel, the best values for the following parameters are poly degree = 3, scale = 0.001 and C =10, as shown in Figure 4-32 and Figure 4-33. The CV R-squared is 0.892.

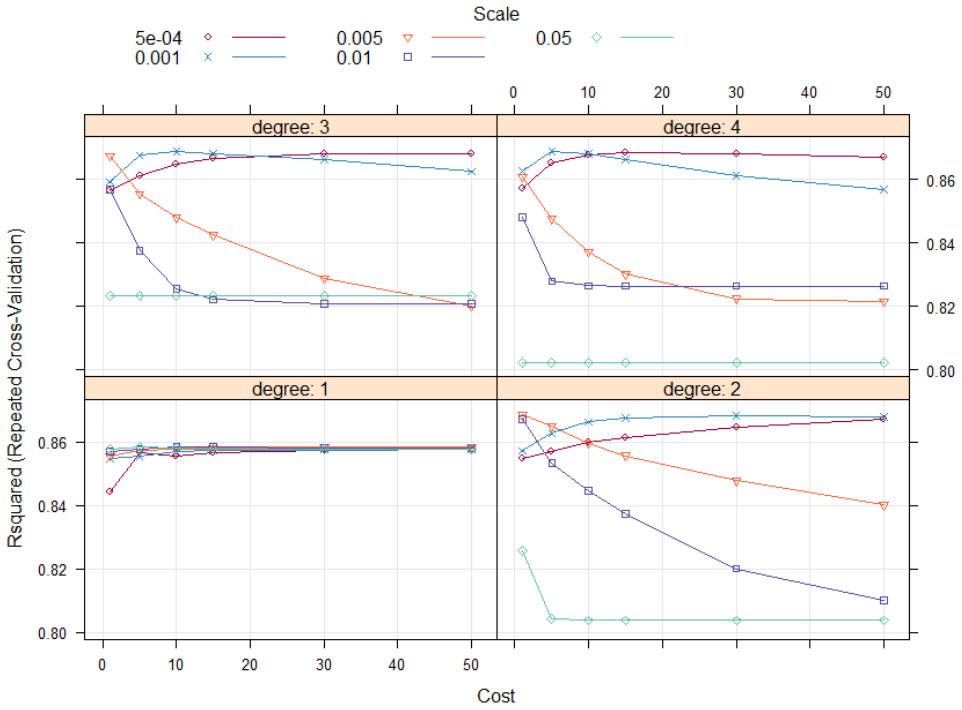

Figure 4-32 Curve plot of parameter tuning using SVM with Poly in kc house dataset

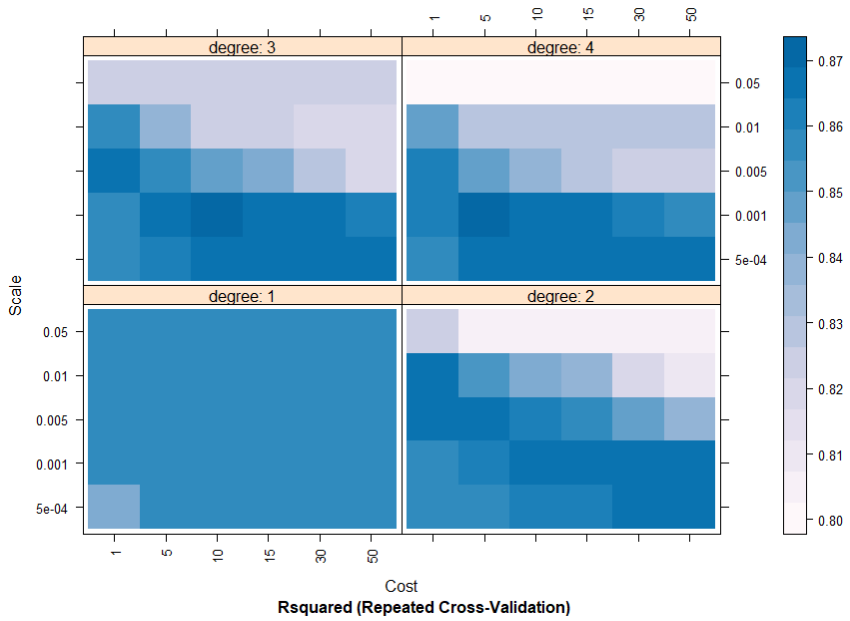

Figure 4-33 Heat map of parameter tuning using SVM with Poly in kc house dataset

On the other hand, the best values of SVM with radial kernel for the following parameters are sigma(gamma) = 5e-04 and C = 30, as shown in Figure 4-34. The CV R-squared is 0.893.

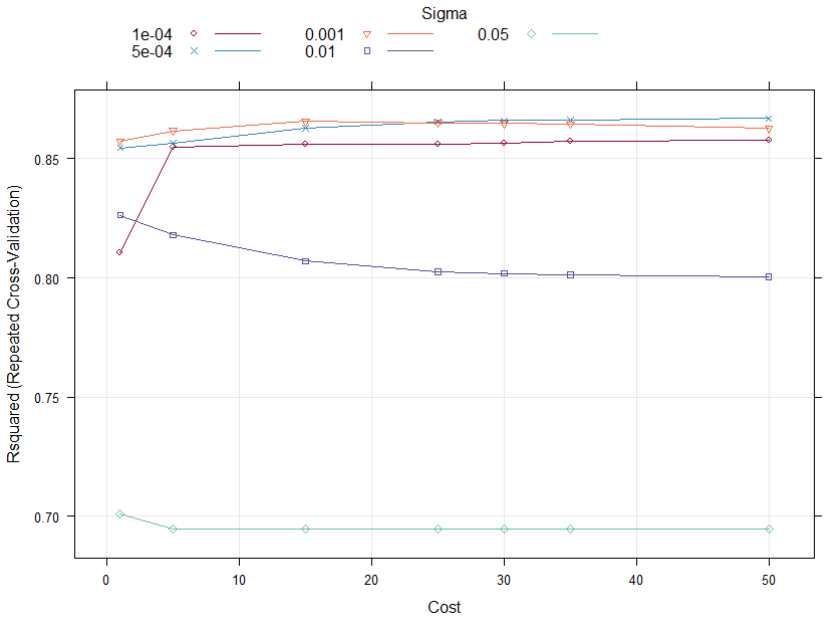

Figure 4-34 Curve plot of parameter tuning using SVM with Radial in kc house dataset

Both SVM models have high CV R-squared, so they also become good candidates for the best model.

4.2.6 Neural Network

We only use one hidden layer in this problem. We then use 5 repeated 5-folds CV to tune the number of perceptions "size" in the hidden layer as well as regularization parameter "decay", the best values are size=9, decay=0.001, as shown in Figure 4-35. The CV R-squared is 0.884, which is also substantially a good candidate for best model.

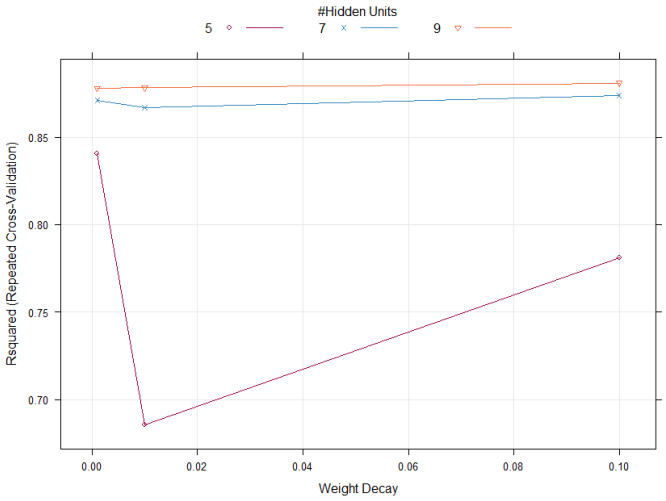

Figure 4-35 Curve plot of parameter tuning using Neural Network in kc dataset

4.2.7 Stacking

We selected linear model, gradient boosting trees, SVM with polynomial and radial kernels and neural network to make a linearly ensemble model. We first check their correlation, as shown in Figure 4-36. Most of them are highly correlated, so we preprocess the prediction of these models using principal component analysis and then stack them. From 5 repeated 5-folds CV, the R-squared is 0.909, which is in favor of all the individual models.

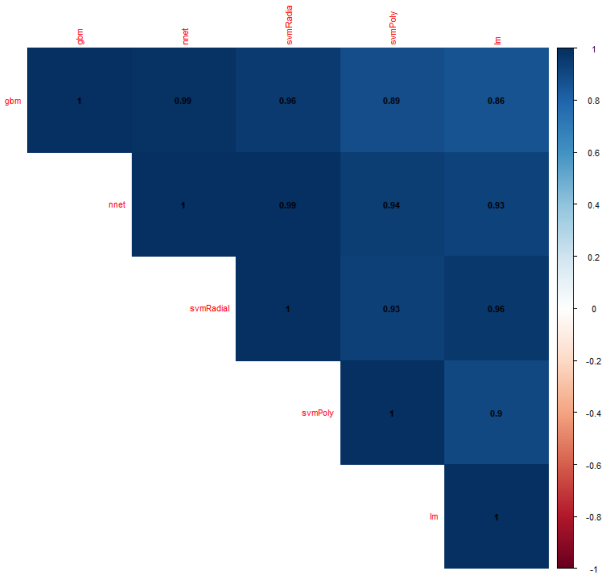

Figure 4-36 Correlation Table for ensemble models in KC house dataset

5 Conclusions

5.1 WDBC dataset

We have demonstrated using six different approaches to obtain a best predictive model for the WDBC data. The comparison is shown in Table 1. Based on the accuracy from training CV, the best model is the stacking model using SVM with radial kernel, which gives CV accuracy of 0.994. Thus, we choose it as our final model. The test accuracy of the model is 0.965. It is a little worse than the test accuracy of SVM with linear or polynomial models alone, but we believe our final model is more robust and the small difference of test error is due to randomness using different seeds in training the model.

Table 1 Comparison of Accuracy for models in WDBC dataset

Model / Accuracy	*Random Forest*	*Gradient Boosting*	*SVM Linear*	*SVM Polynomial*	*SVM Radial*	*Neural Network*	*LDA*	*QDA*
cross-validation	0.966	0.977	0.972	0.983	0.984	0.981	0.958	0.960
Test set	0.936	0.947	**0.982**	**0.982**	0.971	0.977	0.953	0.936
Model	*Linearly Ensemble*	*Stacked*	*Stacked*	*Stacked*	*Stacked*	*Stacked*		

Accuracy	*Model*	*Random Forest*	*SVM Linear*	*SVM Polynomial*	*SVM Radial*	*Neural Network*		
cross-validation	0.989	0.992	0.988	0.992	**0.994**	0.993		
Test set	0.971	0.965	0.971	0.965	0.965	0.959		

5.2 KC house dataset

We have demonstrated using different methods to obtain the best predictive model for the KC house data, the comparison is shown in Figure 5-1.

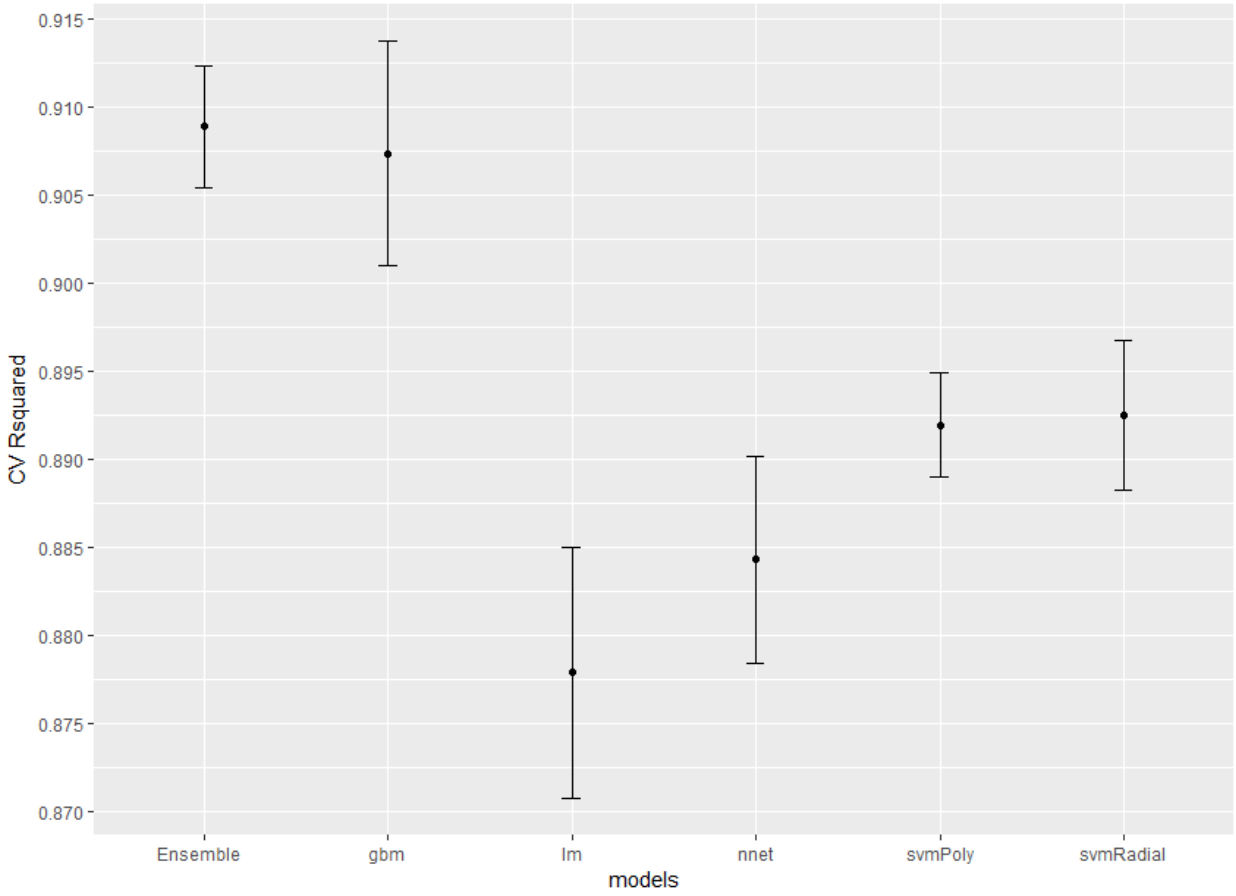

Figure 5-1 Comparison of CV R-squared in KC house dataset

Based on CV R-squared values, the best predictive model is the linearly ensemble model using linear regression, random forest, gradient boosting, SVM with polynomial and radial kernels, and neural network.

We also calculated test mean-squared error (test-MSE) using test data set for all the models, and define the "test R-squared" as "test-MSE * Number of observations in test set / Total sum of squares in test set". The results are also shown in Table 2. The gradient boosting is slight better than others in terms of test R-squared.

Table 2 Comparison of mean-squared-error for models in KC house dataset

R-squared	*Linear Regression*	*Random Forest*	*Gradient Boosting*	*SVM Polynomial*	*SVM Radial*	*Neural Network*	*Ensemble Model*
cross-validation	0.8747	0.888	0.907	0.892	0.893	0.884	**0.909**
Test set	0.867	0.866	**0.914**	0.896	0.900	0.893	0.913

In conclusion, we select the linearly ensemble model as the best model for prediction for the KC house dataset. Though the test R-squared is lower than gradient boosting, it has highest CV R-squared and is based on other five models including gradient boosting, thus we believe it is more robust for future test.

6 Reference

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer Statistics, 2017," (in English), *Ca-a Cancer Journal for Clinicians,* Article vol. 67, no. 1, pp. 7-30, Jan-Feb 2017.

[2] K. T. R. S. V. G. Reddy, V. V. Kumari, and K. V. Varma, "An SVM Based Approach to Breast Cancer Classification using RBF and PolynomialKernel Functions with Varying Arguments," *International Journal of Computer Science and Information Technologies,* vol. 5, no. 4, pp. 5901–5904, 2014.

[3] H. A. E. A. M. Elsayad, "Diagnosis of Breast Cancer using Decision Tree Models and SVM " *International Journal of Computer Applications,* vol. 83, no. 5, pp. 19-29, 2013.

[4] S. Bagui, S. Bagui, and R. Hemasinha, "The Statistical Classification of Breast Cancer Data," *International Journal of Statistics and Applications,* vol. 6, no. 1, pp. 15-22, 2016.

[5] D. K. U. Rani, "Parallel Approach for Diagnosis of Breast Cancer using Neural Network Technique," *International Journal of Computer Applications,* vol. 10, no. 3, pp. 1-5, 2010.

[6] R. Alyami *et al*., "Investigating the effect of Correlation based Feature Selection on breast cancer diagnosis using Artificial Neural Network and Support Vector Machines," (in English), *2017 International Conference on Informatics, Health & Technology (Iciht),* Proceedings Paper p. 7, 2017.

[7] Z. Y. Yin, Z. Y. Fei, C. M. Yang, A. Chen, and Ieee, "A novel SVM-RFE based biomedical data processing approach: basic and beyond," (in English), *Proceedings of the Iecon 2016 - 42nd Annual Conference of the Ieee Industrial Electronics Society,* Proceedings Paper pp. 7143-7148, 2016.

[8] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, 1995, vol. 14, no. 2, pp. 1137-1145: Stanford, CA.

[9] D. A. Freedman, *Statistical models: theory and practice*. cambridge university press, 2009.

[10] G. James , Witten , D., Hastie , T., Tibshirani , R, *An Introduction to Statistical Learning with Applications*. Springer 2013.

[11] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B-Statistical Methodology,* vol. 67, pp. 301-320, 2005.

[12] L. Breiman, "Bagging Predictors," *Machine Learning,* journal article vol. 24, no. 2, pp. 123-140, August 01 1996.

[13] A. Liaw and M. Wiener, "Classification and regression by randomFores," *R NewsR News,* vol. 2/3, pp. 18-22, 2002.

[14] R. H. Wang, "AdaBoost for Feature Selection, Classification and Its Relation with SVM*, A Review," *International Conference on Solid State Devices and Materials Science,* vol. 25, pp. 800-807, 2012.

[15] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* journal article vol. 20, no. 3, pp. 273-297, September 01 1995.

[16] A. Ben-Hur and J. Weston, "A User's Guide to Support Vector Machines," *Data Mining Techniques for the Life Sciences,* vol. 609, pp. 223-239, 2010.

[17] G. A. Carpenter, "NEURAL NETWORK MODELS FOR PATTERN-RECOGNITION AND ASSOCIATIVE MEMORY," (in English), *Neural Networks,* Review vol. 2, no. 4, pp. 243-257, 1989.

[18] M. Caudill, "Neural networks primer, part I" *AI Expert,* 1989.

[19] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop.*, 1999, pp. 41-48: IEEE.