**Content**

[EXECUTIVE SUMMARY...................................1](#_Toc14658)

[1. Data Limitation and Credentials.......................2](#_Toc4151)

[2. Judgemental Bias.................................................2](#_Toc6106)

[BUSINESS UNDERSTANDING......................2](#_Toc4058)

[DATA UNDERSTANDING..............................3](#_Toc7631)

[DATA MODELING.............................................4](#_Toc27137)

[1. Ordinary Linear Squares ....................................5](#_Toc23402)

[2. Lasso/Ridge.........................................................5](#_Toc7329)

[3. Generalized Additive Models............................6](#_Toc4694)

[4. Regression Trees..................................................6](#_Toc22557)

[5. Support Vector Regression................................7](#_Toc22848)

[PERFORMANCE AND EVALUATION........8](#_Toc18657)

[TECHNICAL APPENDIX & FILES..............11](#_Toc8596)

[1. Technical appendix...........................................11](#_Toc24230)

**Contact Information**

**[Jiarui Tian jt2995@columbia.edu](mailto:Jt2995@columbia.edu)**

**Jianfeng Zhang [jz2857@columbia.edu](mailto:jz2857@columbia.edu)**

**Jing Ye [jy2878@columbia.edu](mailto:Jy2878@columbia.edu)**

**Xiaorong Yuan [xy2347@columbia.edu](mailto:Xy2347@columbia.edu)**

**Shiyu Liu [sl4258@columbia.edu](mailto:Sl4258@columbia.edu)**

**Ge Zhang [gz2262@columbia.edu](mailto:Gz2262@columbia.edu)**

**Ziwei Guo [zg2278@columbia.edu](mailto:Zg2278@columbia.edu)**

**EXECUTIVE SUMMARY**

The report is commissioned to provide an analysis and evaluation of how the machine learning techniques perform in order to predict retail sales of electronic shopping and mail-orders, a subcategory of corporate finance. In search of the simplest, most effective, and most predictive model for the sales prediction, the following six methods are applied: OLS, Lasso, Ridge, Generalized Additive Models, Regression Tree, Linear and Nonlinear Support Vector Regression.

The monthly data is about electronic shopping and mail-order houses sales, which contains both mail-delivery (telephone or TV order) and e-commerce sales from 1992 to 2017. During this time period, telephone, TV and online shopping have been growing fast. The models aim to use economic and social factors to predict the sales and to serve as a benchmark for companies involved in the e-commerce sales industry. Predictor variables are chosen mainly from macroeconomic and social angles, which include the major economic indicators and social events. They are non-seasonally-adjusted historical records of GDP, CPI, 3-Month Treasury Constant Maturity Rate, PMI, Unemployment, Crude Oil Price, Industrial Production Index, Total Nonfarm Payroll, and Holidays.

For the macroeconomic factors only, a 3-month lag is used, which means that the economic indexes three months ago is used to predict the Sales data of current month. The Holidays does not have a lag since their dates are known every year, and the holiday which may influence sales of that month remains in that month. By doing so, some macroeconomic impacts can be better represented in the data since most of them takes time to effect.

Based on the test results with same evaluation metrics across models, the models that explain the most information are Generalized Additive Models, Linear and Nonlinear Support Vector Regression, Ridge, Lasso, and through model testing the best performance attribute to Nonlinear Support Vector Regression, which have the lowest test error. The test results show that a nonlinear assumption fits slightly better to the data compared to a linear assumption, which indicates that there are few parameters that are not linearly correlated with the e-commerce sales.

Across all methods, CPI is kept in all final models, which indicates its high correlation with ecommerce sales. Other variables such as GDP, PMI, Crude Oil Price and Christmas also commonly appear in the final models. Most of these variables selected by these final models are related to the macroeconomic environments. There are also some other factors that are significant while using one method and less predictive in other methods. Limitations still exist in the model, which are attributed into the following reasons:

2. **Data Limitation and Credentials**

When adding predictive variables into the model initially, many of the predictors involve credential data, of which the impossibility of access led to a unpreventable drop in variable size.

The limitation of data access may have impacted the performance of many machine learning techniques, which are best used for variable shrinkage.

1. **Judgemental Bias**

The methods and variables to use are consulted and based on machine learning researches on sales data in the past, all data that are accessible and considered potentially contributitive are added. While we select the variables to include, a judgemental bias may happen resulting a loss of important yet uncommonly used predictors. The loss in a judgemental bias is inevitable yet influential in machine learning techniques appliance.

**BUSINESS UNDERSTANDING**

With the newly emerging of e-commerce in late 20th century, companies are putting more attention on how to sell their goods and services online. From 1992 to 2017, with the rapidly popularization of internet, the monthly e-commerce and mail-orders has increased more than 2500% in sales. Unlike traditional retailing, online retailing provides customers more convenience by reducing the time and cost that they would have spent on their way to retailing stores. The development of e-commerce boosted the industry of logistics and as a counter reaction, developed logistic makes people enjoy in online retailing in an even faster way.

As more attention focused on e-commerce, the prediction of the sales of e-commerce become more important towards the firms deeply involved in such industry. Generally speaking, in retail management, forecasting serves to predict and meet the demands of consumers in retail establishments while controlling pricing and inventory. Moreover, accurate forecasting is important for retail companies that want to minimize the capital that is tied in stocks while simultaneously ensuring adequate product availability for their customers. In addition, researched evidence about the importance of forecast accuracy could aid communication both within retail companies and with other players in the supply chain.

To fit the need as mentioned above, we have this project to use some macroeconomic indices and alternative data to predict the Retail Sales of Electronic Shopping and Mail-Orders(Sales). This retail sales data comprises establishments primarily engaged in retailing all types of merchandise using non-store means, such as books and magazines, clothing and accessories, computer hardware and software, drugs, etc. In terms of predictors, compared with other indicators, macroeconomic leading indicators could provide an overview picture of domestic economic dynamics and are easily accessible. In this sense, our model is mainly constructed by using macro-economic data such as GDP, CPI, PMI, unemployment rate, etc. As outcomes, firms can efficiently and accurately predict the sales of the whole market by simply observing the operation of national economy.

Moreover, not only can we obtain sales’ prediction results, but also the relative importances of predictors in every model. The changes in predicting power of every variable across time can be served as a reference for government regulators or retailers in decision making process, such as stabilization policy and inventory management.

**DATA UNDERSTANDING**

1. **Description of Data and Data Cleaning**

Our data contains the monthly not seasonally adjusted historical records of GDP, CPI, 3-Month Treasury Constant Maturity Rate, PMI, Unemployment, Crude Oil Price (West Texas Intermediate - Cushing, Oklahoma), Industrial Production (Manufacturing) Index, Total Nonfarm Payroll, Holidays: Christmas, Thanksgiving and Super Bowl, and Retail Sales of Electronic Shopping and Mail-Orders (Sales). The longest available time period for Sales is from 01/1992 to 12/2017, as a result, the records of every other variable are also through that period and each observation is identified by time without any missing values.

The historical monthly economic indexes are downloaded from the website of the Economic Research at the St. Louis Fed. The holiday data for Christmas and Thanksgiving is organized manually, and the data for Super Bowl is collected by checking the Wikipedia for its historical dates. The historical monthly sales are downloaded from the website of the Census Bureau, part of the U.S. Department of Commerce. Original economic indexes are in the form of a single column in csv files identified by time. So, we combine each variable by column. Original sales data is in the form of a row among 61 other retail sales categories in a xlsx file with the columns containing the monthly sales in a year. And every year is separated in different sheets. We extract the row of retail sales of Electronic Shopping and Mail-Orders, transpose each row into column and combine into one column.

1. **Data Manipulation**

Due to the time lag of release of economic statistics, the current monthly data is usually only available three months later. To make our model more applicable, we match the economic statistics three months ago with current sales. And since the schedule of holiday is fixed and known, we still match current holiday data with current sales. So that we are able to build model to use currently available economic statistics to predict sales of the next third month. After data manipulation, we have records of 8 economic indexes from 01/1992 to 10/2017 corresponding to records of holiday and sales from 03/1992 to 12/2017.

1. **Independent Variables and Dependent Variables**

Independent variables are the 8 economic indexes and three holidays. Since the days of holidays are almost fixed every year, we treat the three holidays as dummy variables. If a holiday is scheduled in a month, then the corresponding dummy variable of that month takes value 1, or it takes 0. Dependent variable is the sales, the target number we try to predict.

Since the responses and predictions are quantitative , we choose Mean Square Error and R-Squared to measure the performance of different models. Mean Square Error measures the average of the squares of the deviations. R-Squared is defined as the percentage of the dependent variable variation that is explained by a model.

1. **Exploratory Data Analysis**

To get an overview of the data, we plot the Pearson correlation as Figure 1. From the correlation plot, we observe high correlations between dependent variable sales and almost all independent variables expect Unemployment. But there also exists high correlations between pairs of independent variables, such as GDP and CPI, Industrial Production and Payroll. So we are supposed to take collinearity into consideration when constructing predicting models.

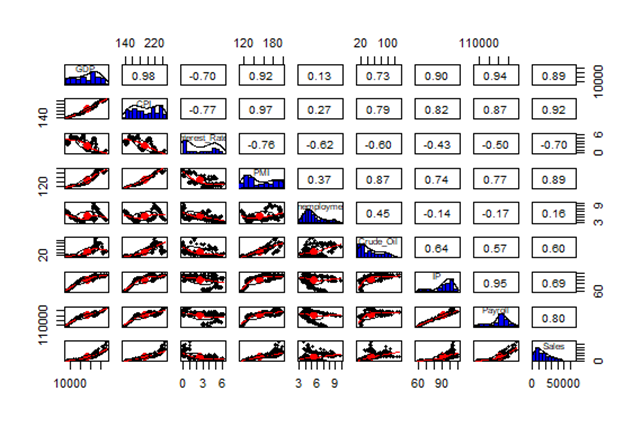


Figure 1: Correlation among all variables and distribution of individual variables

**DATA MODELING**

In order to predict the monthly sales data using a 3-month lag, the following OLS (Ordinary Linear Squares) model and five machine learning techniques are applied and compared:

1. **Ordinary Linear Squares**

OLS model fits observed data with a straight line and the coefficient of every variable implying its relative importance in predicting the response in the model. So, it is used as our benchmark model to predict sales. However, it has many assumptions, including the linear relationship between each pair of independent variable and dependent variable, same variance of error terms, non-correlation among error terms and non-existence of exact linear relationship among the independent variables. The fitted model may produce questionable predictions when any of those assumptions are violated, which is the situation we need to take into account and discuss later. The data is first randomly divided into training and test datasets with 80/20 rule. In order to make interpretation easier for estimated coefficients, we further scale each independent variable to the same unit by subtracting the mean and dividing by the standard deviation.

The regression result shows that among all eight economic indexes and three holiday dummy variables, Christmas, Unemployment, Industrial Production Index, Crude Oil Price, Thanksgiving and GDP these six variables turn out to be most significant, in which Unemployment, IP index, Crude Oil Price have a negative correlation with the sale. It seems counterintuitive that IP index has a negative relationship with sale. After checking the VIF value of each variable, we conclude that there exists high collinearity among variables, which can also be seen from the correlation plot in Figure 1. This explains why this model produce irrational coefficient estimator for IP Index in the sense of economics.

To tackle this problem, we choose to use forward stepwise function to select variables into the model, because in this method the remaining predictors will be modified such that they are uncorrelated to the selected predictors after every step. We get a sequence of 8 models and choose one with lowest training error and satisfy non-collinearity between variables. The final model includes three variables: CPI, Crude Oil Price and Christmas. CPI and Christmas have positive correlations with the sale, while Oil Price shares a negative relationship. The R-square for training and test data are 0.916 and 0.897 respectively, and RMSE for test data is 3402.23.

1. **Lasso/Ridge**
   1. *Ridge:*

Ridge regression produce more stable coefficients than OLS by adding L2 regularization (sum of squared coefficients) to the OLS function to penalize residuals. So, we do not consider collinearity among variables when preparing data for Ridge regression, since all coefficients become stable when the value of model parameter, lambda, increases. And we choose the best lambda by cross validation. As a result, the model of Ridge regression may fit the training data less well than the OLS, while it generalizes better because of smaller sensitivity to extreme variance in the data. Compared with Lasso, Ridge regression does not shrink the model, but changes the weights of every coefficients. So, our model still contains 11 independent variables with non-zero coefficients. The R-square for training and test data are 0.951 and 0.931 respectively, and RMSE for test data is 2783.732.

* 1. *Lasso:*

Since Lasso poses L1 regularization to OLS function, Lasso shrinks some of the coefficients to zero, which provides an alternative way to do variable selection and perfectly addresses the collinearity problem in OLS. Like Ridge regression, all numeric variables need to be scaled in order to prevent penalizing some coefficients more than others. By using cross validation, we choose the best tuning parameter and get the result that only GDP, CPI, PMI, Unemployment, Crude Oil Price, IP Index, Christmas, Thanksgiving, Super Bowl are selected into Lasso model and the R-square for training and test data are 0.953 and 0.944 respectively, and RMSE for test data is 2508.182.

1. **Generalized Additive Models**

By plotting the predictors against the response, we find that most of the variables have non-linear relations with the monthly sales. In this case, the non-linear Generalized Additive Models would be a suitable choice to build the predictive model.To fit the model reasonably, we first look at plots that depict the relations between each predictor and the response and find that a smoothly quadratic curve would be a better fit for the relations between predictors GDP, CPI, PMI and Payroll. To examine this idea quantitatively, we use the cross-validations to find the best degree of freedom in each sub-model as Table 1:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Predictors | GDP | CPI | Interest Rate | PMI | Unemployment | Crude Oil | IP Index | Payroll |
| DF | 5.73 | 57.12 | 34.04 | 16.11 | 6.81 | 13.45 | 14.56 | 34.59 |

Table 1: Best degree of freedom in each sub-model in

Table 1 shows the cross-validation results for each smoothing spline. It should be noticed that the degree of freedom in the smoothing spline of CPI is 57.12, which is obviously the highest among variables. This high degree of freedom will lead to a very flexible sub-model of CPI. Though having increased the risk of overfitting, this situation is not that surprising and it could be explained by the dramatically fluctuation of Consumer Price from 1992 to 2017. Then we fit the Generalized Additive Model using these parameters. 80% of the original data is used as the training data and the other 20% data are put on hold. The model gives a AIC of 4542.193. Then the trained model is used to predict the sales based on the on-hold data. Following a standard test procedure, we obtained a test MSE of 2781133 or a RMSE of 1667.673.

1. **Regression Trees**

The regression tree assumes data independency and no other specific assumptions are necessary. Therefore, no pre-processing methods is applied. The data is splitted into training and testing datasets with 80/20 rule by random sampling. The regression model are first build to use all 11 variables to predict Sales within the training dataset. The pre-set parameters specify maxdepth of 8, which prevents the tree from further growing after reaching the height of 8; a minbucket and minsplit of 5, which assign the smallest observations in terminal node to be 5 and in parent node, correspondingly; finally a tuning parameter complexity parameter of 0.01, which is the minimum improvement of r-squared under each split. The first fitted model contains 3 parameters that are eventually used to grow the three: GDP, PMI and CPI.

The 5-fold cross-validation is used to prune the complexity parameter (cp) under the 3-variable model. When processing the 5-fold cross-validation, the sample size are 199, 197, 198, 200 and 198. Then set the split parameter to 0.8. Under the criteria of Root Mean Square Error (RMSE), the complexity parameter (cp) is a useful parameter to improve efficiency and accuracy of models, which also specifies how the cost of a tree is penalized by the number of terminal nodes. Hence, the final value used for the model is complexity parameter (cp) = 0.0814, with Root Mean Square Error (RMSE) = 5241.323, R-squared = 0.8442, and Mean Absolute Error (MAE) = 3852.083. Mean Absolute Error equals to 3852.083, which interpret the average absolute difference between outcome and predictors. The R-squared equals to 0.8442, which implies that 84.42% of the variability has been explained by the optimal model, which turns out to be a good fit across the training data.

After the processes of pruning parameters and cross-validation, fit the test data to the final model with selected predictors and optimal complexity parameter generated by regression tree and cross-validation. And R-squared of comparing the predicted response to original independent variable is 0.7508, which means 75.08% of variation can be explained by this model. 75.08% is still reasonable and acceptable, although the value of R square decreases when comparing it to the result of training set, 0.8442. And the RMSE value is 5670.108.

1. **Support Vector Regression**

The support vector machine is well-know as a specific classification algorithms while it can also be used as a regression method, maintaining the primary features which characterize the algorithm. The major idea for this algorithm is the same as others, minimizing the error and individualizing the hyperlane which maximizes the margin. There is a motivation to find and optimize the bounds given for regression as the same way as with the classification approach. To be more precise, it relies on defining the loss function that ignores the errors which are located within the distance of true values. Under this part, both the linear and radial support vector regressions are built and compared on the values of best performance, MSE, and R-squared values.

As for support vector machine, the lower performance indicates the lower-error model. At the same time, we also perform a 5-fold cross validation to help select the best choices of parameters: cost and Gamma for the SVM with an radial kernel. If the computed cost value is small, the margin will be wide and the model will perform better. If we increase the value of cost, the number of training error is possible to be reduced. However, this comes a risk of overfitting the training data at the price of a more irregular decision boundary. The value of Gamma governs the tradeoff between error which contains the bias, and variance in the model. A lower value for Gamma implied that the support vector has larger influence of predictors and model is less likely to overfit the train dataset.

The linear support vector regression performs better because of the lower value of best performance which equals to 7466976. Another parameter to compare the overall fit of train dataset is the MSE value, equalling to 6966581. As what has been discussed before, the RMSE is an important value to evaluate the performance of the model for test dataset. The value of RMSE for test dataset is about 2303.02 which is twice as much as the value of non-linear model. The R-squared value of train dataset IS approximately 0.9557 which implies that around 95.57% of variability within the train dataset is explained by the model. The R-squared value for test dataset is about 0.9115, indicating that about 91.15% of the variability is explained by the model and there is a good fit within the test data. In conclusion, this method performs well with an acceptable value of R-squared.

As for nonlinear support vector regression, it makes possible to perform linear separation by mapping the data into higher dimensional kernel induced feature space. Even though the model of this method has larger best performance, other values of the method are more competitive than the linear model, with MSE for train dataset of 686635.4. Compared to the previous method, this value is significantly smaller. The R-squared value for training dataset is as high as 0.9956. Besides, there exist lower RMSE for test dataset of 1067.666 and higher R-squared value of 0.9901 for test dataset which implies that about 99.29% of the variability is explained by the non-linear model.

Both methods perform well with R-squared values close to 1. In other words, the models applied with Support Vector Regression can well explain the variability of the dataset. Generally, the second method of non-linear model is preferred. Because it has significantly lower values of MSE for train dataset and RMSE for test dataset. To be more precise, its MSE value is about 10 percent and RMSE is about half of the values of previous linear method.

**PERFORMANCE AND EVALUATION**

The time series of predictions of different models and corresponding responses from hold-out dataset are plotted as Figure 2.

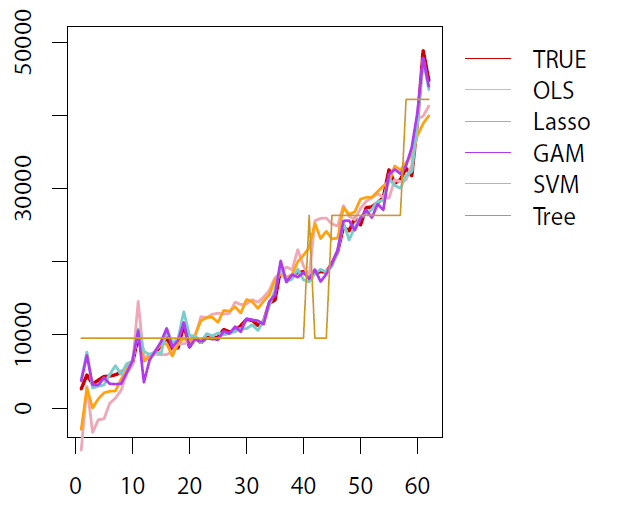
****

Figure 2: Time series of predictions of different models comparing with observed responses

The graph above represents the performance of typical prediction models being used in this project. The red thick line represents the test data which is being sampled from the overall observations. As we see, GAM and SVM(non-linear) give the closest predictions to the true data. This can be explained by the non-linearity of the variables: as most variables do have a relatively flexible relation with reponses, the non-linear models naturally outperform others. The linear models, such as OLS and Lasso Regression, have a worse performance at both ends of the time period. Both models show negative sales prediction at the very beginning of the period. One can roughly see a linear trend in the predion curves of OLS and Lasso. This trend shows the linearity character of linear models as well as the reason why these and not meet the requirements for this data. Lastly, the Regression Tree model also does not give a better. Only three terminal nodes are in the model and they clearly cannot fit the flexibility of the data.

The results of model applied to test dataset are summarized as table 2 and table 3 as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Test RMSE | Training R2 | Test R2 | Test R2 Compared with OLS |
| OLS | 3402.23 | 0.916 | 0.897 | - |
| Ridge | 2783.732 | 0.951 | 0.931 | Higher |
| Lasso | 2508.182 | 0.953 | 0.944 | Higher |
| GAM | 1667.673 | 0.991 | 0.979 | Higher |
| Regression Tree | 5241.323 | 0.844 | 0.751 | Lower |
| SVR-Linear | 2303.023 | 0.9557 | 0.911 | Higher |
| SVR-Non-Linear | 1067.666 | 0.9956 | 0.993 | Higher |

Table 2: Comparisons about the performance coefficients among different models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable Importance | OLS | Ridge | Lasso | GAM | Regression  Tree | SVR |
| GDP | - | 4 | 8 | 8 | 1 | 1 |
| CPI | 1 | 3 | 1 | 9 | 2 | 3 |
| Interest Rate | - | 10 | - | 11 | - | 11 |
| PMI | - | 2 | 4 | 7 | - | 6 |
| Unemployment | - | 8 | 6 | 6 | - | 5 |
| Crude Oil Price | 3 | 6 | 5 | 4 | - | 9 |
| Industrial Production Index | - | 5 | 3 | 3 | - | 8 |
| Total Nonfarm Payroll | - | 11 | - | 10 | - | 2 |
| Christmas | 2 | 1 | 2 | 1 | - | 4 |
| Thanksgiving | - | 7 | 7 | 2 | - | 7 |
| Super Bowl |  | 9 | 9 | 5 |  | 10 |

Table 3: Variable importances measured in different models

From table 2, we could say that most algorithms except the Regression Tree performs better than the OLS one on each coefficient which we use to determine the model performances. Especially for the Non-linear Support Vector Regression model, its R-squared value is pretty high and almost equal to 1, which indicates an absolute great performance for explaining the variability of the datasets.

Observing table 3, we could conclude that different models place different weights on variables. Comparing to our benchmark model, the OLS, most models put high importance on variable CPI except GAM. Most models also give high rank to the appearance of Christmas while only Lasso and GAM models assign high ranks to Crude Oil Price as the OLS model. And all models deploy a low rank on Interest Rate. Comparing with the best-performing SVR model, most of the variable importance can find similar rank as other models, but the importance of Total Nonfarm Payroll is especially addressed with a rank of 2 and the importance of Crude Oil Price is less addressed by a rank of 9.

The changes in relative importance provides some clues for previous ignored signals from economic indexes to the ecommerce retail sales. Besides well-known sales season related with Christmas, regulators and retailers are also advised by the Support Vector Regression model to pay close attention to Total Nonfarm Payroll and Unemployment index to get a preview of ecommerce sales of the following third month. To get a more general idea and the trend of ecommerce sales, the Tree Regression model is a more straightforward method and provides predictions in three steps. It takes the first split at GDP, second one at CPI and third one at PMI.

**TECHNICAL APPENDIX & FILES**

1. **Technical appendix**

Step 1:

Scale training and test data of numerical predictors: GDP, CPI, Interest Rate, PMI, Unemployment, Crude Oil Price, IP Index, Payroll. Turn 3 independent variable holiday into dummy variables by assigning 1 to that variable in a month if the holiday happens in that month, otherwise its value is assigned 0. Sales data is left unscaled.

Step 2:

Randomly assign 80 percent of 309 observations as training set and the remaining as 61 observations as testing set and save predictor variables and the independent variable separately for future use

Step 3:

Fit the OLS, Ridge and Lasso model using all independent variables in training data set and use residual plot to check the assumption of OLS. Since high collinearity exist among predictors, we employ stepwise model selection, and choose the largest step where the VIF of each predictor in the model is less than 10. Ridge and Lasso model do not have the concern of collinearity.

Step 4:

For Ridge and Lasso Model, choose the parameter (lambda) by employing 10 folds cross-validation. Plug in the best lambda into two models and make prediction using tuned models.

For regression tree, control the parameters by setting the maximum depth of any node of the final tree to 8, the minimum number of observations that must exist in a node in order for a split to be attended to 5, the minimum number of observations in any terminal leaf node to 5, and the complexity parameter to 0.01. Then improve the accuracy of prediction by pruning the derived tree with the complexity parameter “anova”.

For GAM, use smoothing spline model to fit the data and use cross-validation to find the best degrees of freedom in each submodel. Then fit the overall GAM with the training data.

As for Support Vector Regression, use tune method and 5-fold cross-validation to the best values of cost and gamma for the minimum best performance. Besides, fit the test dataset and obtain the measurement coefficients.

Step 5:

Compute the Root Mean Square Error and R-Squared of predicted values.