R Programming Report

Group 24: Shiqi Huang, Zhijie Huang, Yizhi Jiang, Jin Sun, Jincong Wang

1. Methods:

In this project, we use 9 models (1) Linear Regression, (2) Linear Regression with transformed data, (3) Lasso, (4) Bagging, (5) Random Forest, (6) Boosted Tree, (7) Support Vector Machine, (8) MARS and (9) Regression Tree to predict sales (Y). We used 96 variables in the training data set to train the models, and then used the test data to evaluate the models. For each model, we summarized a 24*8 table of RMSE of each category in each store. After we get 8 tables, which include 8 sets of RMSE(except SVM's RMSE table), we compare and derive the lowest RMSE for each category in each store.

(1)Linear regression with log transformation and forward subset selection:

We apply log transformation to variable sales and price. In this way, we can add polynomials of independent variables into the to make the model more accurate. Besides, we use forward subset selection to select the best number of X variables to get the minimum training RMSE. After applying the new linear model to the test data, we found that it has better performance than the original linear model in almost all categories.

(2)Lasso：

In this model, we use cross validation to get the best lambda, predict the model, and then get RMSE for each category in each store. We try to change grids to make our model better. We try to use 3 grids, which are [10^-2, 10^10], [10^-6, 10^10], [10^-8, 10^10]. But after trying these three grids, we find that the performance of 3 grids are similar.

(3)Bagging & Random Forests:

Unlike bagging, random forest improves variance by reducing correlation between trees. Random forest has another tuning parameter, called mtry, the number of features that can be searched at each split point. As a rule of thumb, we use one-third of the total predictors as our mtry for our project. Then, we tuned the number of trees, from 500 to 10,000, for all the categories in each store in order to get the lowest training RMSE.

(4)Boosted Tree:

In order to reduce the RMSE of the output, our first step is tuning shrinkage. The default value is not an idea index in this case, thus we changed the shrinkage to 0.002 and RMSE reduced significantly. Then we tune the number of tree from 5000 to 10000, but the results are not improved. Therefore we keep the number of tree around 5000 to have a simpler model.

(5)Support Vector Machine:

Although support vector machine is widely used in the classification problem, we would like to see how this method performed in this dataset. In order to improve the performance of the support vector regression, our group used a list of epsilon to test the best tuning parameter. In

addition, we also changed the value of cost, which means the cost of a violation to the margin to train the models.

2. Results:

RMSE	Store1	Store2	Store3	Store4	Store5	Store6	Store7	Store8
Carbonated Beverages	Boosted Tree	Boosted Tree	Linear with Transformation	Linear with Transformation	Random Forest	Linear	Boosted Tree	Linear with Transformation
Cigarettes	Bagging	Boosted Tree	Bagging	Boosted Tree	Bagging	Boosted Tree	Boosted Tree	Boosted Tree
Coffee	Boosted Tree	Boosted Tree	Random Forest	MARS	MARS	MARS	Linear with Transformation	MARS
Cold Cereal	Lasso	MARS	Linear	Boosted Tree	MARS	Boosted Tree	Boosted Tree	Boosted Tree
Deodorant	Lasso	MARS	Lasso	Random Forest	Boosted Tree	Lasso	Lasso	MARS
Diapers	Lasso	Boosted Tree	Boosted Tree	Boosted Tree	Boosted Tree	Bagging	Lasso	Bagging
Face Issue	Boosted Tree	Lasso	Boosted Tree	Boosted Tree	Lasso	Lasso	Lasso	Boosted Tree
Frozen Dinner Entre	MARS	MARS	Boosted Tree	Boosted Tree	Boosted Tree	MARS	Boosted Tree	Boosted Tree
Frozen Piazza	Lasso	MARS	MARS	Boosted Tree	MARS	Boosted Tree	Lasso	Linear with Transformation
Hot Dog	MARS	Boosted Tree	Boosted Tree	Boosted Tree	Boosted Tree	Boosted Tree	Bagging	Boosted Tree
Laundry Detergent	Boosted Tree	Boosted Tree	Lasso	Random Forest	Boosted Tree	Boosted Tree	Lasso	Lasso
Margarine & Butter	Lasso	Boosted Tree	Random Forest	Boosted Tree	Random Forest	Boosted Tree	Random Forest	Boosted Tree
Mayonnaise	MARS	Linear with Transformation	Bagging	Linear with Transformation	Bagging	Linear with Transformation	Random Forest	Linear with Transformation
Mustard & Ketchup	MARS	Boosted Tree	Random Forest	Linear with Transformation	Random Forest	Boosted Tree	Boosted Tree	Linear with Transformation
Paper Towel	Lasso	Boosted Tree	Lasso	Boosted Tree	Boosted Tree	Lasso	Lasso	Boosted Tree
Peanut Butter	Lasso	Lasso	Bagging	Linear with Transformation	Boosted Tree	MARS	Random Forest	Lasso
Shampoo	MARS	Lasso	Boosted Tree	Bagging	Boosted Tree	Random Forest	Boosted Tree	MARS
Soup	Boosted Tree	Boosted Tree	Boosted Tree	Boosted Tree	Boosted Tree	Boosted Tree	Boosted Tree	Lasso
Spaghetti Sauce	Boosted Tree	Boosted Tree	Bagging	Boosted Tree	Lasso	Boosted Tree	Bagging	Boosted Tree
Sugar Substitute	Lasso	Lasso	Bagging	Bagging	Boosted Tree	Lasso	Linear with Transformation	Boosted Tree
Toilet Tissues	Lasso	Lasso	MARS	Boosted Tree	MARS	Boosted Tree	Linear	MARS
Tooth Paste	Lasso	Bagging	Bagging	MARS	Lasso	Lasso	Bagging	Lasso
Yogurt	Lasso	Boosted Tree	Linear	Linear	Boosted Tree	Boosted Tree	Boosted Tree	Boosted Tree
Beer	Boosted Tree	Boosted Tree	Boosted Tree	Boosted Tree	Boosted Tree	Boosted Tree	Lasso	MARS

We compare the performance of each model. For each category in each store, we apply the model with the smallest RMSE. For all categories in all of the 8 stores, there are 82 categories using Boosted Tree, 28 categories using LASSO, 25 categories using MARS, 17 categories using Bagging, 13 categories using Linear regression with transformation, 12 categories using Random Forest and 5 categories using Linear regression.

Boosted Tree	82
LASSO	38
MARS	25
Bagging	17
Linear regression with transformation	13
Random Forest	12
Linear regression	5

3. Conclusion:

(1) We find that Boosted Tree is the best model among all 9 methods. Tree methods perform well in general, especially Boosted Tree. Lasso also performs well.
(2) SVM is not a good method to build regression model. It is more suitable for classification.
(3) Building a single tree might lead to high variance, thus leading to large MSE. Therefore, regression tree, without transforming it into better methods such as bagging and random forest, is definitely a bad method to build a model.
(4) Linear methods (including Linear regression and Linear regression with transformed Y and Price) are not performing well compared with other non-linear models. It indicates that, to some extent, the relationship between sales and 96 predictors are non-linear.
(5) As two methods that choose specific predictors without using all predictors, Boosted Tree and LASSO are 2 frequently used methods. It indicates that not all variables have effect on sales Y, which means some variables do not influence sales.