Statistics 479

Project 1

Professor: Wei-Yin Loh

TA: Xiaomao Li

Name: Qiuying Li

Date: 01/12/2015

Overview:

In order to predict the INTRDVX from the Consumer Expenditure Survey, I fitted data into multiple linear regression model, randomForest model, Guide-linear split classification and Guide least squre stepwise regression. Among these four models, Guide least squre stepwise regression has the best performance.

Data description :

The Consumer Expenditure Survey (CE) program offers a continuous and comprehensive flow of data on the buying habits of American consumers. There are 25822 observations and .645 variables in the dataset. INTRDVX describes the the amount received in interest or dividend in the past 12 months. INTRDVX_ has four levels: A C D T. A means Valid blank, a blank field where a response is not anticipated ; C means"Don't know," refusal, or other nonresponse, D means Valid value, unadjusted, T means Valid value, topcoded or suppressed.

Study question:

Use the CUs with INTRDVX_ = "D" or "T" to build a model to predict the values of variable INTRDVX (interest and dividends) for which INTRDVX_ = "C"

Analysis methods:

1. Multiple linear regression ;
2. RandomForest
3. Guide-linear split classification;
4. Guide-least squre stepwise regression

1. Multiple Linear Regression

INTRDVX ~ DIRACC + DIRACC_ + AGE_REF +

AS_COMP2 + AS_COMP3 + AS_COMP4 +

BUILDING + CUTENURE + EARNCOMP +

EDUC_REF + EDUCA2 + EDUCA2_ + FAM_SIZE +
FAM_TYPE +FAMTFEDX + FAMT_EDX +

…….

+FEDRFNDX + FEDR_NDX + FEDTAXX +

FEDTAXX_ + FGOVRETX + FGOV_ETX +

FINCATAX + FINCAT_X + FINCBTAX +

FINCBT_X + FINDRETX + FIND_ETX + FINLWT21 +

FJSSDEDX + FJSS_EDX + FPRIPENX

(656 predictors in total)

Residuals:

Min	1Q	Median	3Q	Max
-23756.1	-1192.3	-137.7	898.2	23409.0

Coefficients:

Estimate Std. Error t value Pr(>ltl)

(Intercept) -1.064e+04 1.648e+05 -0.065
0.948520

BEDR_OMQC NA NA NA
NA

…….

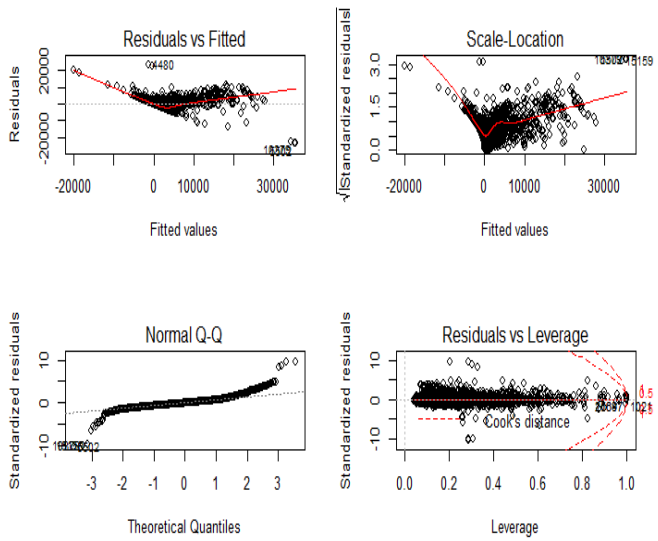
BUILDING5 3.118e+02 7.045e+02 0.443
0.658147

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' '

Residual standard error: 2734 on 2142 degrees of freedom

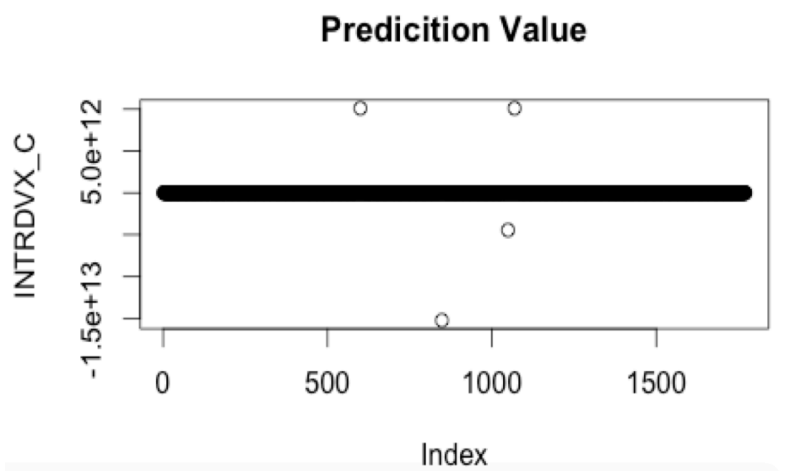
Multiple R-squared: 0.7707, Adjusted R-squared: 0.6964

F-statistic: 10.38 on 694 and 2142 DF, p-value: < 2.2e-16



Output Description

1. Large R^2 and small p-value might due to the hundreds of predictors.
2. QQ plot shows a heavy tails on the two side
3. Independence and normality test are failed




```
> summary(INTRDVX_C)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-1.524e+13 -2.380e+02  1.534e+03  2.901e+08  3.467e+03  1.010e+13
```

Linear model conclusion

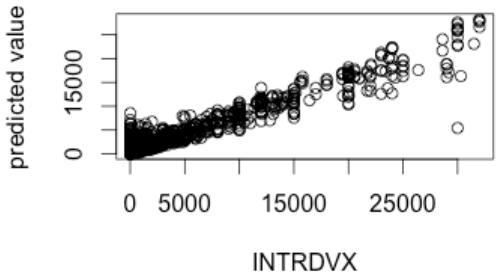
Assumptions for the model were not satisfied, and nagatives values are not expected.

Thus, multiple linear model is not correct.

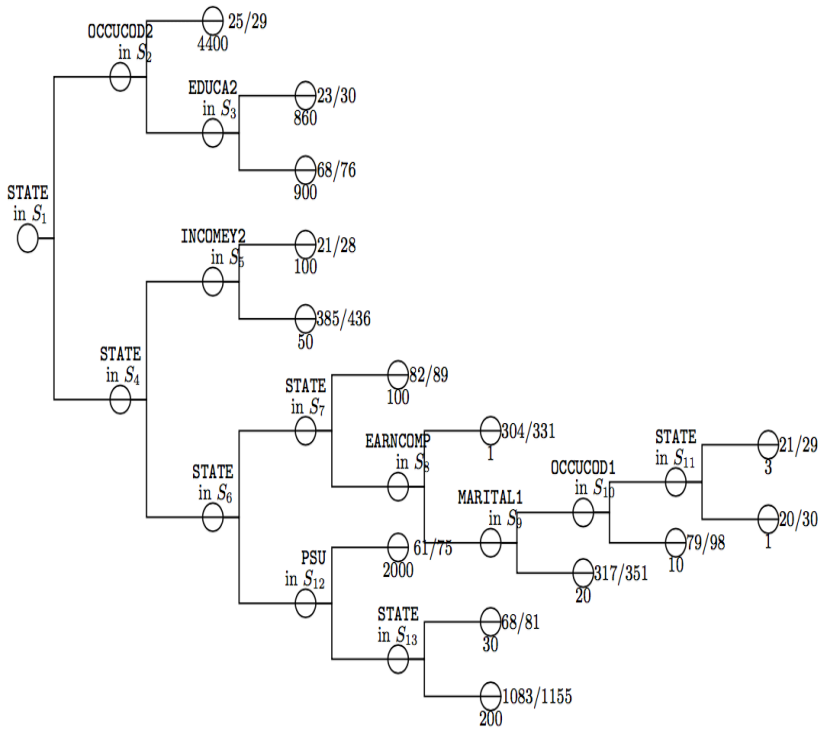
2. RandomForest model:

- Considering that linear regression model is not accurate, use random Foreset based on the linear model.
- However, randomForest model can only handle dataset contains 53 categories. So I decided to select 38 important variables based on the ranking to be the predictors.
- Below are the 38 selected variables

```
AS_COMP1   AS_COMP2   AS_COMP4   FINDRETX  FPRI_ENXT   FRRETIRX   FSALARYX      FSSIX
       9         10         12         69         75         76         77         81
 INC_HRS2   INC_RANK  INCOMEY12  MISC_AXXT   LUMPSUMX  LUMP_UMXD  LUMP_UMXT  OCCUCOD12
      86         87         99        111        112        114        115        128
OCCUCOD214  OCCUCOD23   OTHRINCX  OTHR_NCXD   ALCBEVPQ   FEEADMPQ   POV_CY_D   FSMPFRMX
     140        143        150        152        215        304        368        562
 NETRENTX  NETR_NTXD   OTHREGBX   OTHREGX  OTHREGX_D   RETSURVX  RETS_RVXD   RETSRVBX
     599        600        621        624        625        636        637        639
```

RandomForest:Prediction Plot	Min. 1st Qu. Median Mean 3rd Qu. Max. 7.663 244.5 632. 2046. 1802. 28370
	**Conclustion of RandomForest** All the prediction values are positive, which indicates that the prediction is more accurate, because INTERDVX has to be positive. Thus the randomForest model is more reliable than the multiple linear model.

3. Guide-Classification : Linear split

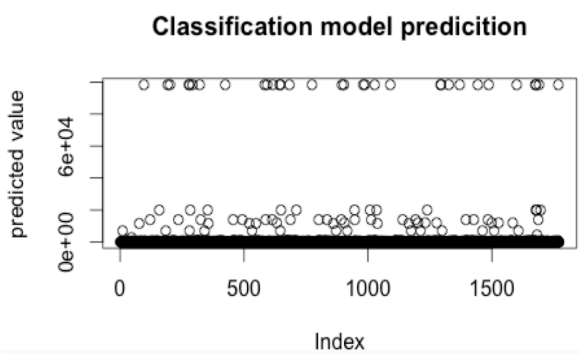


GUIDE 0.50-SE classification tree for predicting INTRDVX using linear split priority, estimated priors and unit misclassification costs. At each split, an observation goes to the upper branch if and only if the condition is satisfied. For splits on categorical variables, values not present in the training sample go to the right. Set $S_1 = \{10, 2, 23, 33\}$. Set $S_2 = \{1, 10, 12, 2, 4, 7\}$. Set $S_3 = \{13, 14, 15\}$. Set $S_4 = \{12, 16, 17, 34, 4, 45, 47, 8\}$. Set $S_5 = \{2, 3, 5\}$. Set $S_6 = \{13, 15, 25, 26, 27, 31, 32, 36, 39, 42, 49, 53, 55\}$. Set $S_7 = \{13, 53\}$. Set $S_8 = \{4, 5, 7, 8\}$. Set $S_9 = \{3, 5\}$. Set $S_{10} = \{12, 2, 5, 6, 7\}$. Set $S_{11} = \{15, 25, 36, 42, 49, 55\}$. Set $S_{12} = \{1313, 1420\}$. Set $S_{13} = \{11, 20, 24\}$. Predicted classes (based on estimated misclassification cost) printed below terminal nodes; #misclassified/sample size beside each node.

```
Number of cases used for tree construction =  2838
Number misclassified =  2557
Resubstitution est. of mean misclassification cost =    0.9009866102889392
```

The misclassified number is fairly larger, more importntly, the resubtitution estemation of mean misclassification cost is colse to 1, which suggest model is not accurate

Prediction of Guide classification

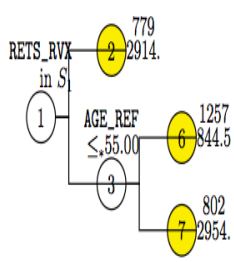


Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	20	50	2340	200	98340

Conclusion : Guide classification

There is a huge variation in the predicted values. Moreoever, there is a huge misclassification cost, So the Guide-Classification is not a good candidate model.

4.Guide-Regression:Linear Squre Stepwise



GUIDE 0.50-SE piecewise linear least-squares regression tree with stepwise variable selection for predicting INTRDVX. At each split, an observation goes to the upper branch if and only if the condition is satisfied. The symbol '≤,' stands for '≤ or missing'. Set S_1 = {C, D, T}. Number in italics beside terminal node is sample mean of INTRDVX. Number above terminal node is sample size.

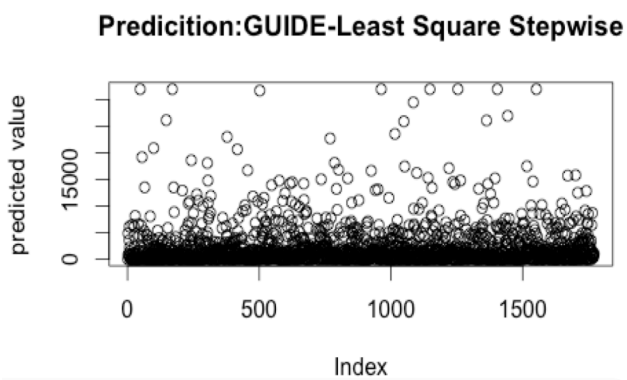
Plot description: In order to visualize the fitted regression function and the data at the same time, fitting a piecewise simple linear model in the guide, where the best single regressor is selected to fit a straight line in each node, as shown in the picture.

```
Proportion of variance (R-squared) explained by tree model =  .6189
```

```
Number of terminal nodes of final tree: 3
Total number of nodes of final tree:    5

Regression tree:

 Node 1: RETS_RVX = "C", "D", "T"
   Node 2: INTRDVX-mean =  2.91405E+03
 Node 1: RETS_RVX /= "C", "D", "T"
   Node 3: AGE_REF <=   55.00000 or NA
     Node 6: INTRDVX-mean =  8.44452E+02
   Node 3: AGE_REF >   55.00000 and not NA
     Node 7: INTRDVX-mean =  2.95427E+03
```



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.0	1.0	955.5	2335.0	2705.0	32000.0

Conclusion : Guide Regression

1. R^2 = 0.6189, which is relatively large.
2. Predictied values are all positive with a small variance

   Thus, Guide regression model is a good candidate model.

5. Models Comparisons

Models	Advantages	Disadvantages
Linear regression	● Simple and easy to use	● Hard to deal with the missing values ● Inaccuarate when containing too many variables
RandomForest	● More accurate than linear model ● Faster than linear model	● RandomForest can only handle less than 53 categories
Guide (Linear split Classofocatopn & Least Square Regression)	● Guide is able to deal with many unseen cases ● High accuracy for the prediction ● Guide is powerful to the missing values ● Running Guide is faster than linear models. ● Powerful to huge dataset	● Guide-classification for this dataset has relatively large variation. ● Guide-classification has huge misclassification cost for this dataset.

Conclusion:

1. Based on the discussion above, I decided to drop linear regression model for the inaccuracy.
2. Guide-least square stepwise regression model has the best performance.

   According to the summary information of the predicted for 3 different models below, Guide-least square stepwise has the smallest standard deviation, so the this model has the smallest variation to the mean.In addition, over 60% variation in the response can be explained by the model.

Models	Summary of the predicted INTRDVX	sd
RandomForest	Min. 1st Qu. Median Mean 3rd Qu. Max. 7.663 244.5 632. 2046. 1802. 28370	5690．23
Classification	Min. 1st Qu. Median Mean 3rd Qu. Max. 1 20 50 2340 2003.0 98340	13259.28
Guide-Regression	Min. 1st Qu. Median Mean 3rd Qu. Max. 1.0 1.0 955.5 2335.0 2705.0 32000.0	4024.159

Below are the 95% confidence intervals for the INTRDVX.

Models	95% Confidence Interval
RandomForest	（-22982.82，13246.2）
Guide-Classification	（-24178.42，28858.7）
Guide-Regression	（-5552.352，10222.35）

3. INTRDVX describes the the amount received in interest or dividend in the past 12 months. INTRDVX_ = “D” or “T” menas people answered the questions, and there are 2922 observations of it. INTRDVX_ = “C” means peope who did not answer the questions. In order to predict the the INTRDVX of people who did not answer the question, we fitted the valid data into the guide least squre regression model, and the we have 95% confidence that the true INTRDVX , which is the amount of the interest or dividend received of the people who did not answer the questions are between （-5552.352，10222.350）.