AUGUST 2019

TEXAS
The University of Texas at Austin

PREDICTING WALMART SALES

STA S380: Predictive Modelling

Palakh Gupta, Saurabh Bodas, Rocco Lange, Robbie Geoghegan, Javeria Rangoonwala

Walmart is the world’s largest company in terms of revenue

- **OVER US$500 BILLION YEARLY REVENUE – 65% FROM THE USA**
- **LARGEST PRIVATE EMPLOYER IN THE WORLD WITH 2.2 MILLION EMPLOYEES**
- **OVER 50% PRIVATELY OWNER BY THE WALTON FAMILY**

Predicting Sales

Prediction

What are the weekly sales for a Walmart store?

What model should we use to predict sales?

6k data points aggregated from 422k

1. Weekly Sales (dependent variable)
2. Public Holiday
3. Temperature
4. Fuel price
5. Local unemployment
6. Shop size
7. CPI
8. Month

Source: Kaggle for dates 2010-02-05 to 2012-11-01

Why do we care about this?

- Sales Planning
- Demand Forecasting
- Supply Chain Management
- Financial Planning
- Internal Controls
- Marketing
- Bonuses

Overview

Data Cleaning

- Kaggle Data set
- Dataset based on stores
- Missing values
- Uncover initial patterns, characteristics, and points of interest using visual exploration

Testing approach

Model Type	Choosing Variables	Measuring Fit	Comparison
KNN	All variables and combinations based on linear regression	Testing different k's	Out-of-sample RMSE
Linear Regression	Multiple Linear Regression	Standard Error (in-sample RMSE)	
	Subset Selection	Root Sum Squares, Adjusted R2 and BIC and in-sample RMSE	
	Shrinkage (Ridge, Lasso)	Lambda and in-sample RMSE	
Trees	Regression Trees	Deviance and RMSE	
	Random Forest		
	Boosting		
	Bagging		

- **K Nearest Neighbors**
- Linear Regression
- Trees
- Next Steps

K Nearest Neighbors – All variables

k = ?

- Finding the best k:

```
#calculate best k value

out_MSE = NULL

for (i in 2:1000){

near = kknn(Weekly_Sales~.,train,out_of_sample,k=i,kernel = "rectangular")
aux = mean((out_of_sample[,1]-near$fitted)^2)

out_MSE = c(out_MSE,aux)
}

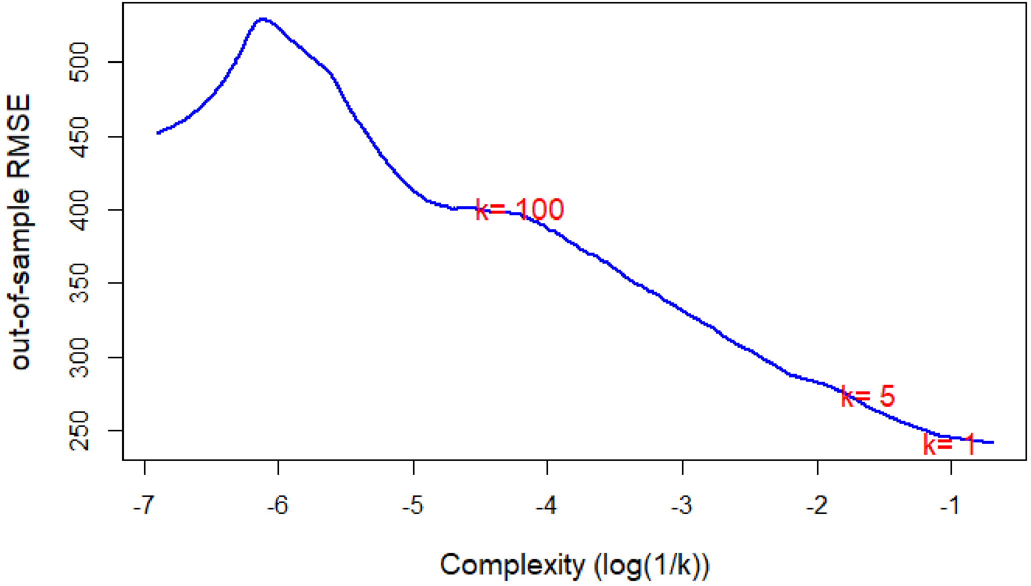

best = which.min(out_MSE)
```

K Nearest Neighbors – All variables

Out of Sample RMSE:

- k = 1, RMSE: $241,796
- k = 5, RMSE: $275,010
- k = 100, RMSE: $401,187

Best model has
RMSE >20% of avg. sales



K Nearest Neighbors – One Variable (Size)

- Finding the best k:

k = ?

```
#calculate best k value
library(kknn)
out_MSE_1 = NULL

for (i in 2:1000){

near_1 = kknn(Weekly_Sales~Size,train,out_of_sample,k=i,kernel = "rectangular")
aux_1 = mean((out_of_sample[,1]-near_1$fitted)^2)

out_MSE_1 = c(out_MSE_1,aux_1)
}

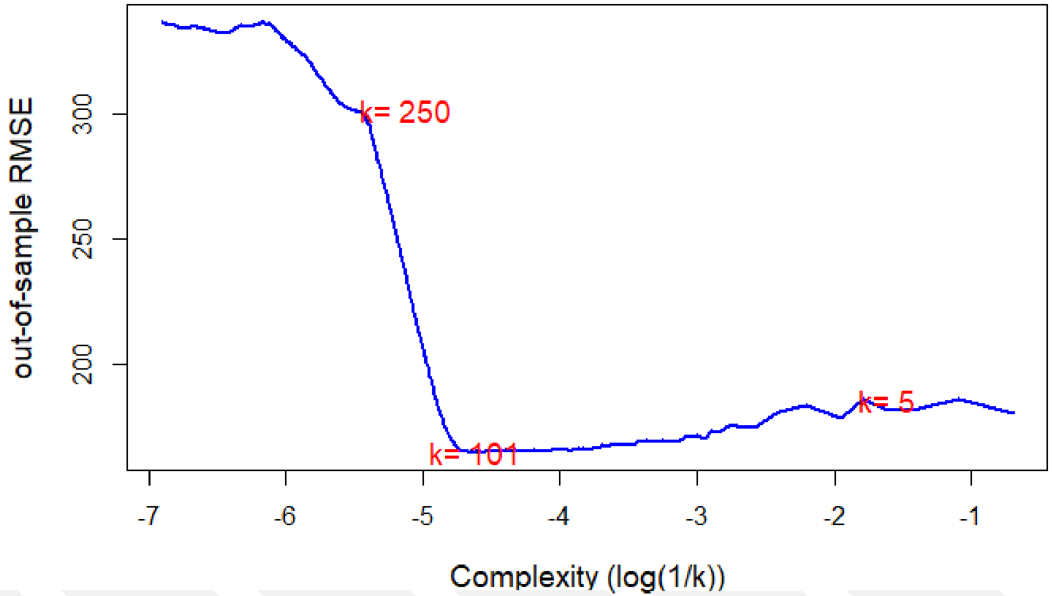

best_1 = which.min(out_MSE_1)
```

K Nearest Neighbors – One Variable (Size)

Out of Sample RMSE:

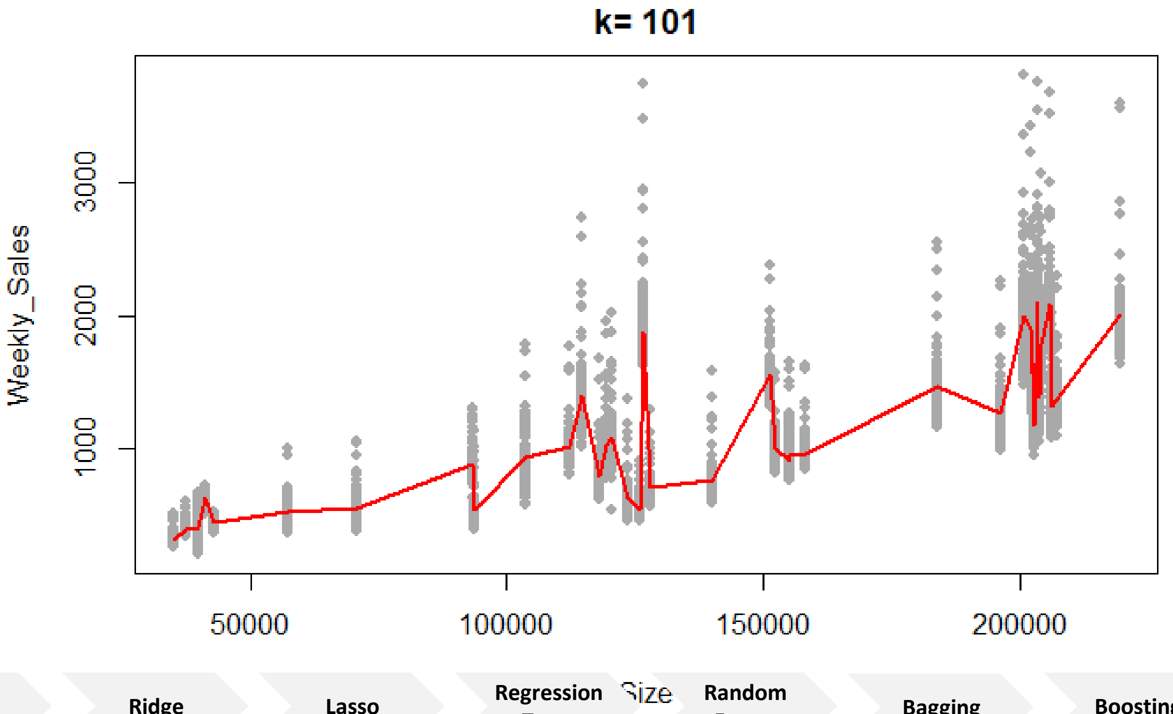
- k = 101, RMSE: $164,556
- k = 5, RMSE: $185,821
- k = 250, RMSE: $301,653

Best model has
RMSE ~15% of avg. sales



K Nearest Neighbors – One Variable (Size)

- The model:



K Nearest Neighbors – 13 Variables

k = ?

- Finding the best k

```
#calculate best k value
out_MSE_13 = NULL

for (i in 2:1000){

near_13 = kknn(Weekly_Sales~Temperature+Unemployment+Size+CPI+Jan+Feb+Mar+Apr+May+Jun+Oct+Nov+Dec,train,
out_of_sample,k=i,kernel = "rectangular")
aux = mean((out_of_sample[,1]-near_13$fitted)^2)

out_MSE_13 = c(out_MSE_13,aux)
}


best_13 = which.min(out_MSE_13)
```

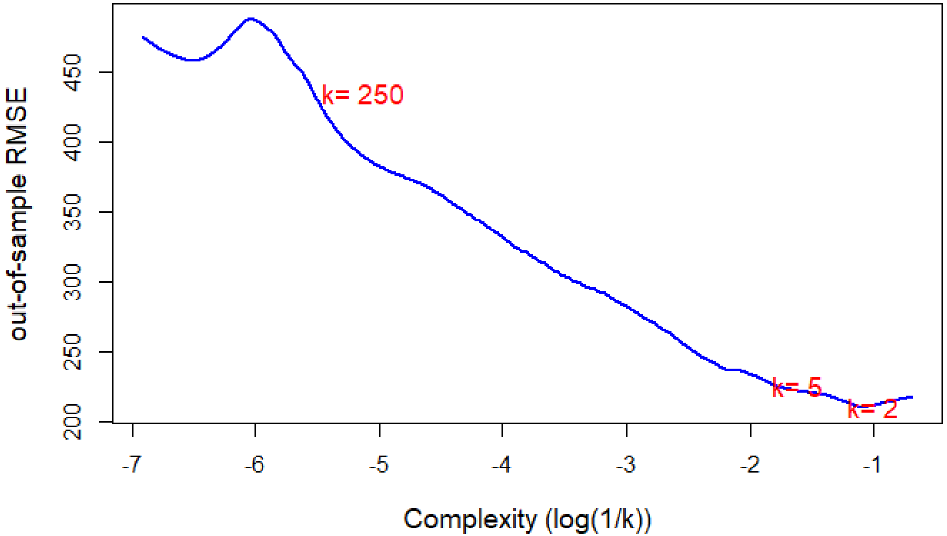
K Nearest Neighbors – 13 Variables

Out of Sample RMSE:

- k = 2, RMSE: $219,226
- k = 5, RMSE: $226,109
- k = 250, RMSE: $434,860

Best model has

RMSE >20% of avg. sales



K Nearest Neighbors – Conclusions

- Best KNN model used only Size as predictor of Weekly Sales
- The RMSE was approx. 15% of the average Weekly Sales across the data set
- Maybe we can do better with a different model selection

- K Nearest Neighbors
- **Linear Regression**
  - **Multiple Linear Regression**
  - Subset Selection
  - Shrinkage (Ridge and Lasso)
- Trees
- Next Steps

Multiple linear regression

Multiple Linear Regression Code Snippet

```
```{r}
set.seed(9)
train = sample(1:nrow(walmart),nrow(walmart)*0.8)
out_of_sample = walmart[-train,]
train = walmart[train,]
lm.fit =lm(Weekly_Sales~.-Jul,data=train)
summary(lm.fit)
```
```

- In-sample RMSE = $309.5k
- In-sample RMSE is c. 30% of average sales

Multiple Linear Regression Model Summary

```
Residuals:
    Min      1Q  Median      3Q     Max
-731.54 -233.15  -17.01  165.70 2136.31

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.219e+02  5.562e+01   3.991 6.68e-05 ***
IsHolidayTRUE  3.337e+01  1.869e+01   1.786 0.074236 .
Temperature    6.255e+00  4.338e-01  14.420  < 2e-16 ***
Fuel_Price    -1.322e+01  1.002e+01  -1.319 0.187136
Unemployment  -3.278e+01  2.590e+00 -12.657  < 2e-16 ***
Size           7.306e-03  6.920e-05 105.573  < 2e-16 ***
CPI           -1.912e+00  1.276e-01 -14.989  < 2e-16 ***
Jan            1.356e+02  2.927e+01   4.634 3.68e-06 ***
Feb            2.529e+02  2.709e+01   9.335  < 2e-16 ***
Mar            1.711e+02  2.384e+01   7.178 8.08e-13 ***
Apr            1.403e+02  2.195e+01   6.390 1.81e-10 ***
May            8.001e+01  2.157e+01   3.709 0.000210 ***
Jun            6.184e+01  1.980e+01   3.124 0.001794 **
Aug            1.643e+01  1.968e+01   0.835 0.403973
Sep           -1.461e+01  2.076e+01  -0.704 0.481548
Oct            7.424e+01  2.151e+01   3.452 0.000562 ***
Nov            2.820e+02  2.686e+01  10.501  < 2e-16 ***
Dec            4.779e+02  2.768e+01  17.266  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 309.5 on 5130 degrees of freedom
Multiple R-squared:  0.6982,	Adjusted R-squared:  0.6972
F-statistic: 698.1 on 17 and 5130 DF,  p-value: < 2.2e-16
```

Positive Relationship
Negative Relationship
No Relationship

Interpreting the predictors

Predictor Descriptions

Predictor	Coefficient	t value	Interpretation
Intercept	221	4.0	N/A (store with 0 sq feet?).
IsHolidayTrue	33	1.8	Holiday weeks drive an additional $33k sales per store, although may be noise.
Temperature	6	14.4	The hotter the temperature, the higher the sales (seasonality?).
Fuel Price	-13	-1.3	Not significant
Unemployment	-33	-12.7	The higher a store's local unemployment, the lower the sales.
Size	0.007	105.6	An additional square foot correlates with an additional $7 in sales
CPI	-1.9	-15.0	Unsure how to treat – could have adjusted sales. Can interpret as higher CPI causes lower sales – real vs. nominal dollars.
Jan, Feb, Mar, Apr, May, Jun, Oct, Nov, Dec	Positive	3 – 9	Sales in most months are higher than July. E.g. December store sales are $478k higher than July (probably due to Christmas).
Aug, Sep	0	< 1.0	Sales in Aug and Sep are similar to July.

Out of sample test with all variables – RMSE = $319,000

Multiple Linear Regression Code Snippet – Test Data

```
```{r}
lm.outofsample = lm(Weekly_Sales~., data = out_of_sample)
summary(lm.outofsample)
```
```

- Out-of-sample RMSE = $319k

Multiple Linear Regression Model Summary – Test Data

```
Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.198e+02  1.120e+02   5.535 3.79e-08 ***
IsHolidayTRUE 7.527e+01  3.833e+01   1.964   0.0498 *
Temperature   7.322e+00  9.371e-01   7.813 1.17e-14 ***
Fuel_Price   -9.956e+00  2.145e+01  -0.464   0.6426
Unemployment -2.724e+01  5.293e+00  -5.146 3.08e-07 ***
Size          7.424e-03  1.446e-04  51.346  < 2e-16 ***
CPI          -1.918e+00  2.617e-01  -7.329 4.10e-13 ***
Jan          -3.257e+02  5.035e+01  -6.469 1.40e-10 ***
Feb          -2.119e+02  4.676e+01  -4.532 6.39e-06 ***
Mar          -3.773e+02  4.658e+01  -8.101 1.26e-15 ***
Apr          -4.375e+02  4.712e+01  -9.286  < 2e-16 ***
May          -4.270e+02  5.111e+01  -8.356  < 2e-16 ***
Jun          -5.225e+02  5.637e+01  -9.268  < 2e-16 ***
Jul          -5.933e+02  5.709e+01 -10.392  < 2e-16 ***
Aug          -5.659e+02  5.837e+01  -9.695  < 2e-16 ***
Sep          -5.500e+02  5.332e+01 -10.314  < 2e-16 ***
Oct          -4.739e+02  4.877e+01  -9.718  < 2e-16 ***
Nov          -2.227e+02  4.923e+01  -4.524 6.63e-06 ***
Dec                  NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 319 on 1269 degrees of freedom
Multiple R-squared:  0.6934,    Adjusted R-squared:  0.6893
F-statistic: 168.8 on 17 and 1269 DF,  p-value: < 2.2e-16
```

KNN | Multiple Linear Reg | Subset Selection | Ridge | Lasso | Regression Trees | Random Forest | Bagging | Boosting

- K Nearest Neighbors
- **Linear Regression**
  - Multiple Linear Regression
  - **Subset Selection**
  - Shrinkage (Ridge and Lasso)
- Trees
- Next Steps

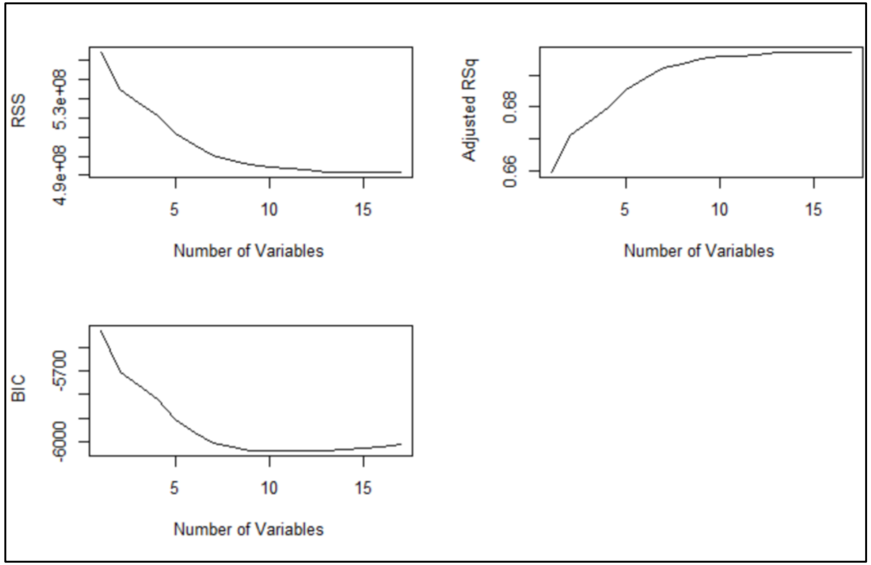
Subset selection: all variables

Subsets function model output

```
1 subsets of each size up to 17
Selection Algorithm: exhaustive
          IsHolidayTRUE Temperature Fuel_Price Unemployment Size CPI Jan Feb Mar Apr May Jun Aug Sep Oct Nov Dec
1  ( 1 )  " "           " "         " "        " "          "*"  " " " " " " " " " " " " " " " " " " " " " " " "
2  ( 1 )  " "           " "         " "        " "          "*"  " " " " " " " " " " " " " " " " " " " " " " "*"
3  ( 1 )  " "           " "         " "        " "          "*"  "*" " " " " " " " " " " " " " " " " " " " " "*"
4  ( 1 )  " "           "*"         " "        " "          "*"  "*" " " " " " " " " " " " " " " " " " " " " "*"
5  ( 1 )  " "           "*"         " "        "*"          "*"  "*" " " " " " " " " " " " " " " " " " " " " "*"
6  ( 1 )  " "           "*"         " "        "*"          "*"  "*" " " " " " " " " " " " " " " " " " " "*" "*"
7  ( 1 )  " "           "*"         " "        "*"          "*"  "*" " " "*" " " " " " " " " " " " " " " "*" "*"
8  ( 1 )  " "           "*"         " "        "*"          "*"  "*" " " "*" "*" " " " " " " " " " " " " "*" "*"
9  ( 1 )  " "           "*"         " "        "*"          "*"  "*" " " "*" "*" "*" " " " " " " " " " " "*" "*"
10  ( 1 ) " "           "*"         " "        "*"          "*"  "*" " " "*" "*" "*" " " " " " " "*" " " "*" "*"
11  ( 1 ) " "           "*"         " "        "*"          "*"  "*" "*" "*" "*" "*" " " " " " " "*" " " "*" "*"
12  ( 1 ) " "           "*"         " "        "*"          "*"  "*" "*" "*" "*" "*" "*" " " " " " " "*" "*" "*"
13  ( 1 ) " "           "*"         " "        "*"          "*"  "*" "*" "*" "*" "*" "*" "*" " " " " "*" "*" "*"
14  ( 1 ) "*"           "*"         " "        "*"          "*"  "*" "*" "*" "*" "*" "*" "*" " " " " "*" "*" "*"
15  ( 1 ) "*"           "*"         "*"        "*"          "*"  "*" "*" "*" "*" "*" "*" "*" " " " " "*" "*" "*"
16  ( 1 ) "*"           "*"         "*"        "*"          "*"  "*" "*" "*" "*" "*" "*" "*" "*" " " "*" "*" "*"
17  ( 1 ) "*"           "*"         "*"        "*"          "*"  "*" "*" "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"
```

Plotting Root Sum Squares, Adjusted R^2 and BIC based on subset selection

Error Measures Against Number of Variables



- RSS minimum = 17 variables
- Adj R^2 maximum = 16 variables
- BIC minimum = 13 variables

- Decided to use model with 13 variables

- 13 variables were:
  - Temperature
  - Unemployment
  - Size
  - CPI
  - 9 months: Jan, Feb, Mar, Apr, May, Jun, Oct, Nov, Dec

With 13 variables, out of sample test RMSE = $319,300

13 variable Multiple Regression with Test Data Model Output

```
Residuals:
   Min      1Q Median      3Q     Max
-820.6 -237.9  -32.5   177.4 2286.7

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.041e+01  8.270e+01   0.368 0.713172
Temperature   7.049e+00  9.193e-01   7.667 3.47e-14 ***
Unemployment -2.678e+01  5.186e+00  -5.164 2.80e-07 ***
Size          7.429e-03  1.446e-04  51.372  < 2e-16 ***
CPI          -1.863e+00  2.538e-01  -7.341 3.78e-13 ***
Jan           2.282e+02  5.439e+01   4.195 2.91e-05 ***
Feb           3.624e+02  5.130e+01   7.065 2.63e-12 ***
Mar           1.779e+02  4.247e+01   4.189 2.99e-05 ***
Apr           1.193e+02  3.737e+01   3.192 0.001447 **
May           1.315e+02  3.562e+01   3.691 0.000232 ***
Jun           3.961e+01  3.565e+01   1.111 0.266809
Oct           8.583e+01  3.535e+01   2.428 0.015309 *
Nov           3.566e+02  4.619e+01   7.722 2.31e-14 ***
Dec           5.677e+02  5.027e+01  11.293  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 319.3 on 1273 degrees of freedom
Multiple R-squared:  0.6919,    Adjusted R-squared:  0.6888
F-statistic: 219.9 on 13 and 1273 DF,  p-value: < 2.2e-16
```

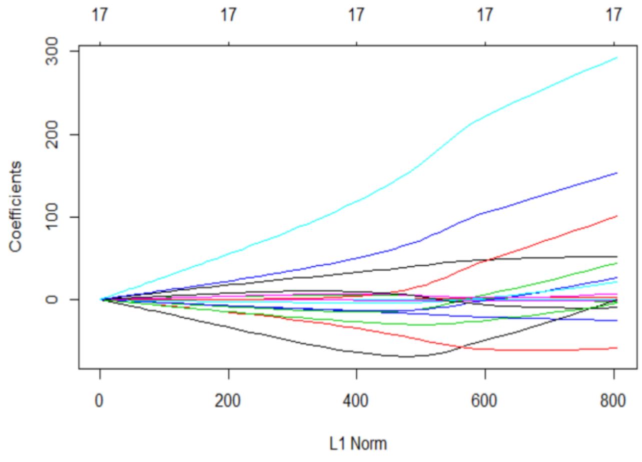
- All predictors significant except June
- Out of sample RMSE = $319,300

- Same out of sample RMSE returned for 17 variable model

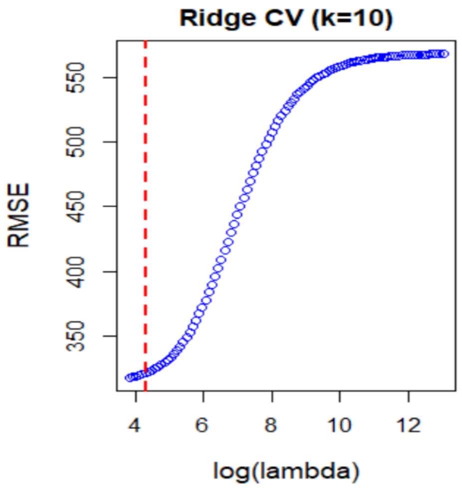
- K Nearest Neighbors
- **Linear Regression**
  - Multiple Linear Regression
  - Subset Selection
  - **Shrinkage (Ridge and Lasso)**
- Trees
- Next Steps

Ridge regression

Standardized Ridge Coefficients



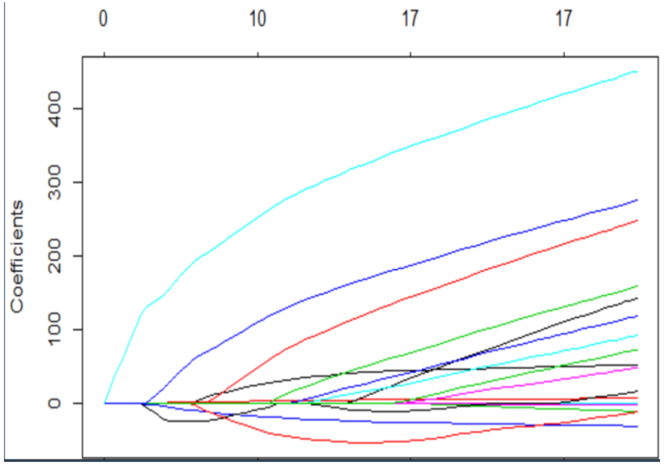
RMSE for Different Log Lambda Levels



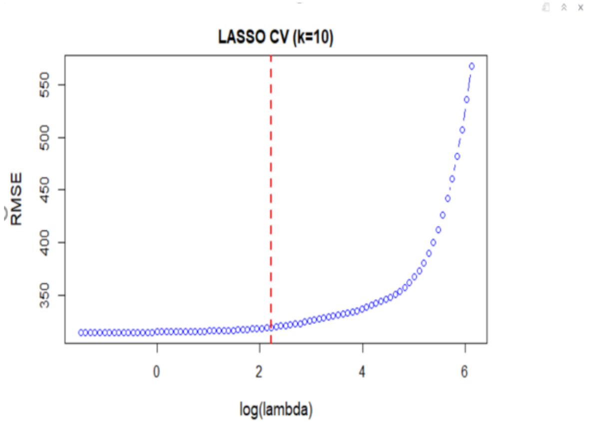
Out of sample RMSE is $319,000

Lasso regression

Standardized Lasso Coefficients



RMSE for Different Log Lambda Levels



Out of sample RMSE is $319,000

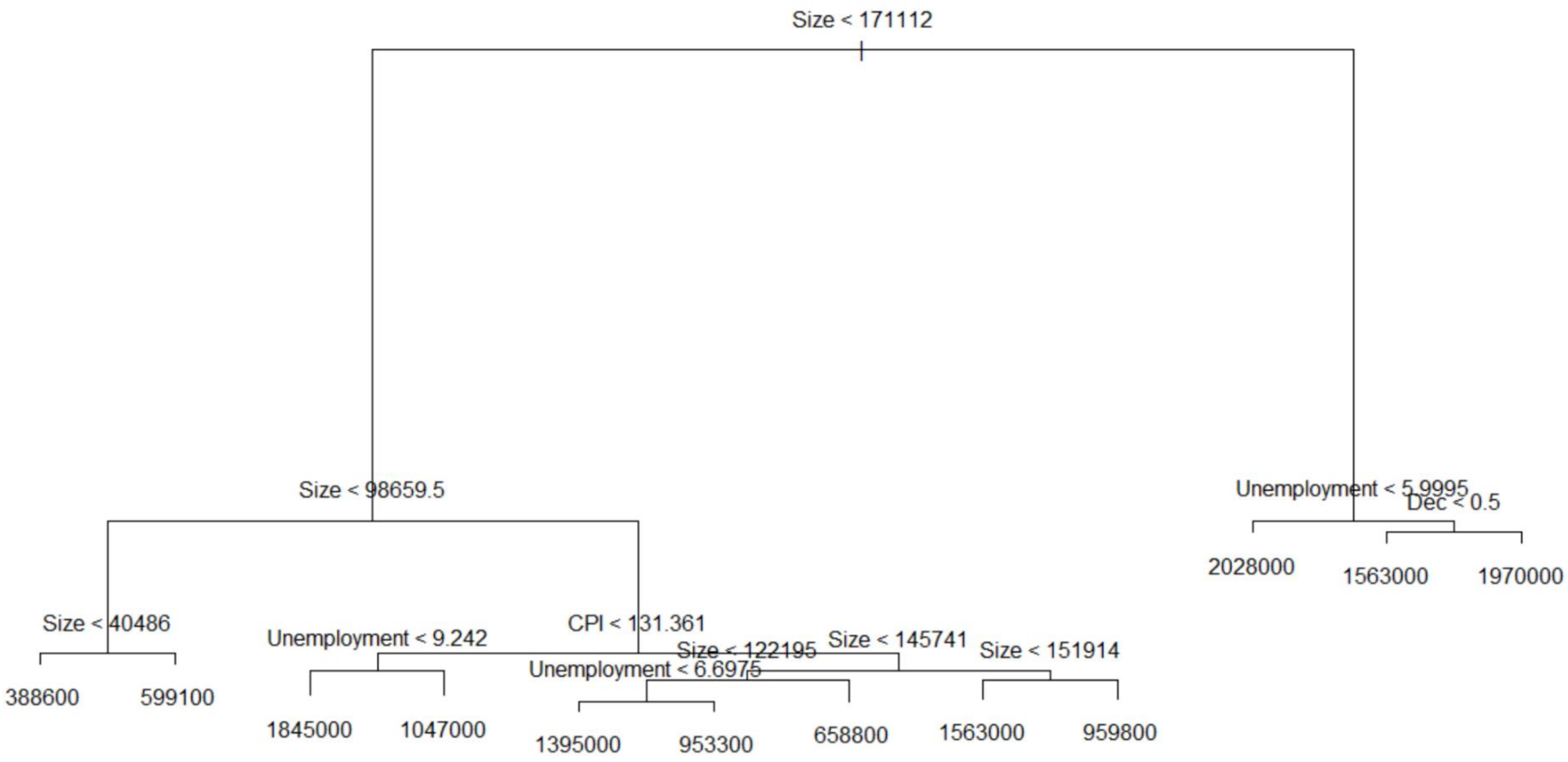
(with 11 non-zero coefficients)

- K Nearest Neighbors
- Linear Regression
- **Trees**
  - **Regression Trees**
  - Random Forest
  - Boosting
  - Bagging
- Next Steps

Regression Trees

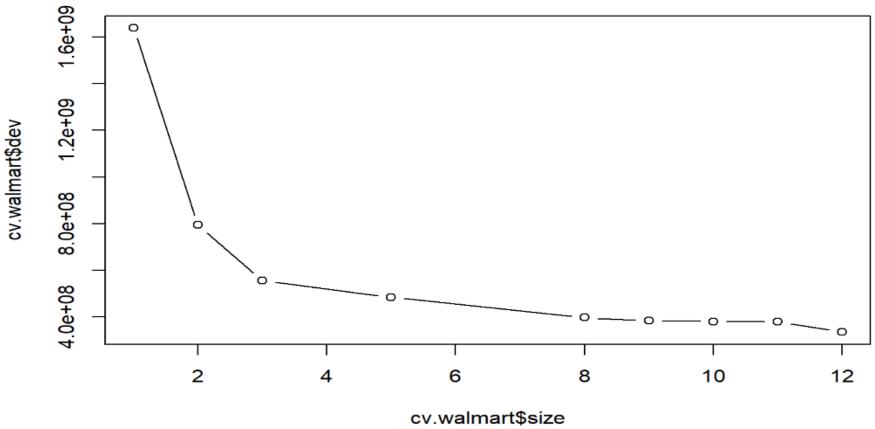
- Variables used in tree construction: Size, CPI, Unemployment, and Dec
- 12 nodes
- Residual mean deviance (RSS): 61,800

```
##
## Regression tree:
## tree(formula = Weekly_Sales ~ . - Jul, data = walmart, subset = train_1)
## Variables actually used in tree construction:
## [1] "Size"         "CPI"           "Unemployment" "Dec"
## Number of terminal nodes:  12
## Residual mean deviance:  61800 = 317400000 / 5136
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1015.00  -135.60   -37.93     0.00    99.45  1848.00
```


Size < 171112
Size < 98659.5
Unemployment < 5.9995
Dec < 0.5
2028000
1563000
1970000
Size < 40486
Unemployment < 9.242
CPI < 131.361
Size < 122195
Size < 145741
Size < 151914
Unemployment < 6.6975
388600
599100
1845000
1047000
1395000
953300
658800
1563000
959800

Cross-validation

- Cross-validation (K=10) chooses the most complex tree (with 12 nodes) – based on the minimum deviance
- The alpha value ($k) = -inf (full, unpruned tree)
- Test RMSE = 248444
- The predicted sales will be within $248444 of the true value




```
$size
[1] 12 11 10  9  8  5  3  2  1

$dev
[1]  334.0810  379.7313  379.7313  382.8082  396.5910  483.3512
[7]  554.3823  793.4233 1639.2737

$k
[1]      -Inf  17.47625  17.50351  17.86648  20.68575  29.44519
[7]  37.32663 238.98517 845.92328

$method
[1] "deviance"

attr(,"class")
[1] "prune"         "tree.sequence"
```

- K Nearest Neighbors
- Linear Regression
- **Trees**
  - Regression Trees
  - **Random Forest**
  - Bagging
  - Boosting
- Next Steps

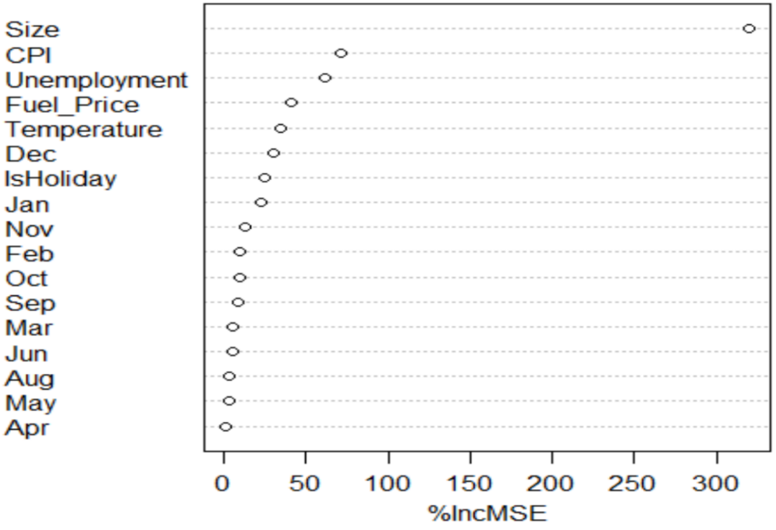
Random Forest

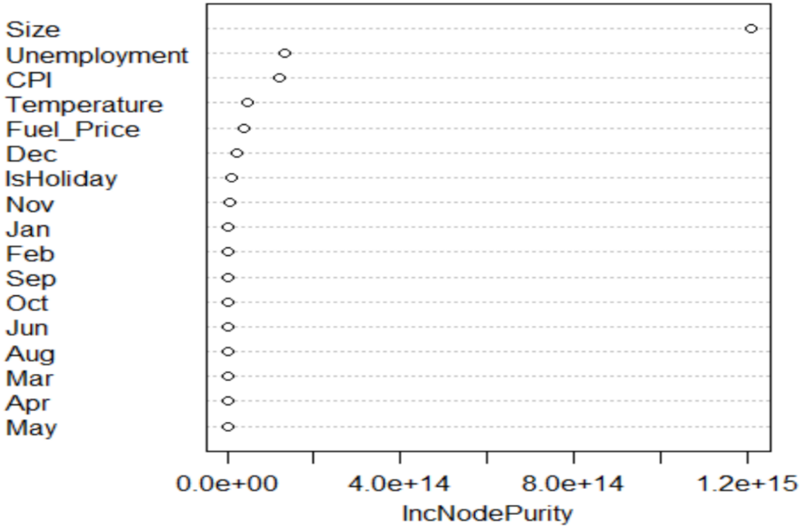
- Variables used at each spilt: 6
- No. of trees: 500

```
Call:
 randomForest(formula = Weekly_Sales ~ . - Jul, data = walmart,
mtry = 6, importance = TRUE, subset = train_1)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 6

          Mean of squared residuals: 16606668132
                    % Var explained: 94.78
```

- Reported RMSE: 98392 (much better than that of the regression tree)
- Plotting the most significant variables





- K Nearest Neighbors
- Linear Regression
- **Trees**
  - Regression Trees
  - Random Forest
  - **Bagging**
  - Boosting
- Next Steps

Bagging

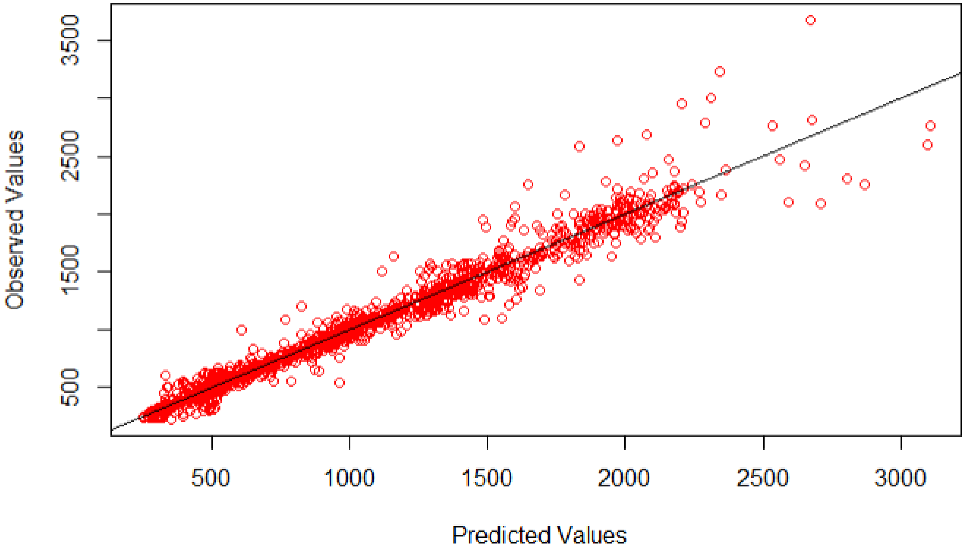
RMSE = ?

```r
library (randomForest)
set.seed(9)
sapply(data, class)
bag.data =randomForest(Weekly_Sales~.-Jul,data=data,subset=train,mtry=16,importance=TRUE,ntree = 1000)
bag.data
importance(bag.data)
plot(importance(bag.data))
varImpPlot (bag.data, sort= "TRUE")
yhat.bag = predict(bag.data,newdata=test)
plot(yhat.bag , test$Weekly_Sales,xlab = " Predicted Values", ylab ="Observed Values" , col= 'red')
abline (0,1)
(mean((yhat.bag - test$Weekly_Sales)^2))^(1/2)
```

RMSE Error = 117,156

```
> importance(bag.data)
                %IncMSE  IncNodePurity
IsHoliday     53.938146     14147507.5
Temperature   28.637920     22474641.7
Fuel_Price    45.998429     21760824.3
Unemployment  73.380665     94875044.8
Size         485.903843   1329515962.7
CPI           69.416612     96309754.6
Jan           75.085749      3318843.5
Feb           18.682794      1225099.5
Mar           13.229784       393984.5
Apr           -5.419644       586885.9
May           12.431284       226449.7
Jun           32.132903       584715.9
Aug           35.370550       558352.9
Sep           44.009534      1131941.2
Oct           23.133230       445067.7
Nov           48.740459      6531659.9
Dec           64.154475     31982539.4
```

Holiday and Temperature prove to be huge drivers for our model

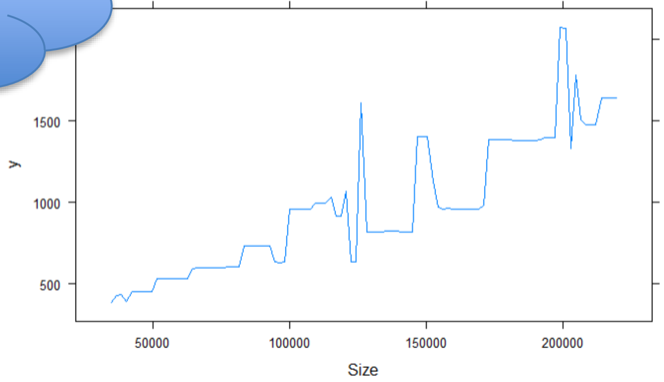


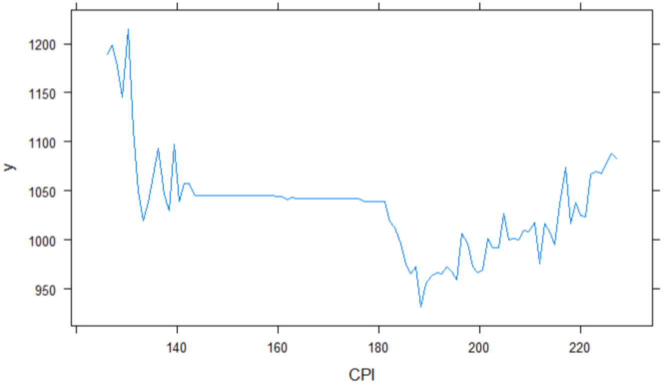
- K Nearest Neighbors
- Linear Regression
- **Trees**
  - Regression Trees
  - Random Forest
  - Bagging
  - **Boosting**
- Next Steps

Boosting

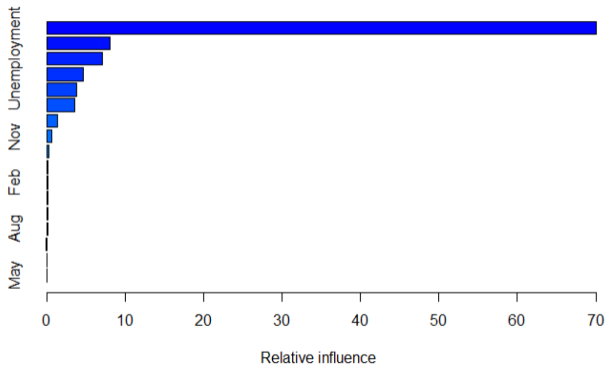
RMSE = ?

```
#separating training and test data
attach(data)
data$IsHoliday <- as.factor(data$IsHoliday)
train=sample(1:nrow(data),size=nrow(data)*0.80)
data[train,]

boost.data=gbm(Weekly_Sales~.-Jul,data=data[train,],distribution=
"gaussian",n.trees=10000,interaction.depth=4,shrinkage = 0.2)
summary(boost.data)
boost.data

par(mfrow=c(1,2))
plot(boost.data,i="CPI")
plot(boost.data,i="Size")
test = data[-train,]
yhat.boost=predict(boost.data ,newdata =data[-train ,], n.trees=5000)
yhat.boost
test
data$Weekly_Sales
(mean((yhat.boost-test$Weekly_Sales)^2))^(1/2)
```





RMSE Error = 108,000



Highest Influential variables

Size

CPI

	var <fctr>	rel.inf <dbl>
Size	Size	70.07906474
CPI	CPI	8.01439070
Unemployment	Unemployment	7.06716300
Temperature	Temperature	4.64872772
Fuel_Price	Fuel_Price	3.81546296
Dec	Dec	3.58239309
IsHoliday	IsHoliday	1.30559730
Nov	Nov	0.61599930
Jan	Jan	0.24598318
Sep	Sep	0.12928339

- K Nearest Neighbors
- Linear Regression
- Trees
- **Next Steps**

Comparison of models

Model Type	Choosing Variables	Out-of-sample RMSE
KNN	Testing multiple combinations	$165k
Linear Regression	Multiple Linear Regression	$319k
	Subset Selection	$319k
	Shrinkage: Lasso	$319k
	Shrinkage: Ridge	$320k
Trees	Regression Trees	$248k
	Random Forest	**$98k**
	Boosting	$108k
	Bagging	$117k

Random Forest Model

What are the biggest predictors for sales:

- Store Size is by far the biggest predictor for store sales
- CPI and Unemployment are the second and third most important predictors
- Dec is the month with the highest predictor importance for sales

Next Steps

- **Missing data:** collecting the missing “mark down” data to determine if reduced prices are associated with increased sales
- **More detailed data:** product information rather than store information could provide more useful information for making business decisions like managing inventory, making marketing decisions and understanding product trends
- **Reducing the error:** incorporating other data points such as local advertising budgets, local competitor information or additional local demographic could reduce the error