Continuing on the below post, I am going to use a support vector machine (SVM) to predict combined miles per gallon for all 2019 motor vehicles.

Part 1: [Using Decision Trees and Random Forest to Predict MPG for 2019 Vehicles](https://blog.alpha-analysis.com/2019/06/predicting-mpg-for-2019-vehicles-using-r.html)

Part 2: [Using Gradient Boosted Machine to Predict MPG for 2019 Vehicles](https://blog.alpha-analysis.com/2019/06/using-gradient-boosted-machine-to.html)

The raw data is located on the [EPA government site](https://www.fueleconomy.gov/feg/download.shtml)

The variables/features I am using for the models are: Engine displacement (size), number of cylinders, transmission type, number of gears, air inspired method, regenerative braking type, battery capacity Ah, drivetrain, fuel type, cylinder deactivate, and variable valve.

There are 1253 vehicles in the dataset (does not include pure electric vehicles) summarized below.

fuel\_economy\_combined eng\_disp num\_cyl transmission  
 Min. :11.00 Min. :1.000 Min. : 3.000 A :301   
 1st Qu.:19.00 1st Qu.:2.000 1st Qu.: 4.000 AM : 46   
 Median :23.00 Median :3.000 Median : 6.000 AMS: 87   
 Mean :23.32 Mean :3.063 Mean : 5.533 CVT: 50   
 3rd Qu.:26.00 3rd Qu.:3.600 3rd Qu.: 6.000 M :148   
 Max. :58.00 Max. :8.000 Max. :16.000 SA :555   
 SCV: 66   
 num\_gears air\_aspired\_method  
 Min. : 1.000 Naturally Aspirated :523   
 1st Qu.: 6.000 Other : 5   
 Median : 7.000 Supercharged : 55   
 Mean : 7.111 Turbocharged :663   
 3rd Qu.: 8.000 Turbocharged+Supercharged: 7   
 Max. :10.000   
   
 regen\_brake batt\_capacity\_ah   
 No :1194 Min. : 0.0000   
 Electrical Regen Brake: 57 1st Qu.: 0.0000   
 Hydraulic Regen Brake : 2 Median : 0.0000   
 Mean : 0.3618   
 3rd Qu.: 0.0000   
 Max. :20.0000   
   
 drive cyl\_deactivate  
 2-Wheel Drive, Front :345 Y: 172  
 2-Wheel Drive, Rear :345 N:1081  
 4-Wheel Drive :174   
 All Wheel Drive :349   
 Part-time 4-Wheel Drive: 40   
   
   
 fuel\_type   
 Diesel, ultra low sulfur (15 ppm, maximum): 28   
 Gasoline (Mid Grade Unleaded Recommended) : 16   
 Gasoline (Premium Unleaded Recommended) :298   
 Gasoline (Premium Unleaded Required) :320   
 Gasoline (Regular Unleaded Recommended) :591   
   
   
 variable\_valve  
 N: 38   
 Y:1215

Starting with an untuned base model:

set.seed(123)  
m\_svm\_untuned <- svm(formula = fuel\_economy\_combined ~ .,  
 data = test)  
  
pred\_svm\_untuned <- predict(m\_svm\_untuned, test)  
  
yhat <- pred\_svm\_untuned  
y <- test$fuel\_economy\_combined  
svm\_stats\_untuned <- postResample(yhat, y)

svm\_stats\_untuned  
 RMSE Rsquared MAE   
2.3296249 0.8324886 1.4964907

Similar to the results for the untuned boosted model.  I am going to run a grid search and tune the support vector machine.

hyper\_grid <- expand.grid(  
 cost = 2^seq(-5,5,1),  
 gamma= 2^seq(-5,5,1)   
)  
e <- NULL  
  
for(j in 1:nrow(hyper\_grid)){  
 set.seed(123)  
 m\_svm\_untuned <- svm(  
 formula = fuel\_economy\_combined ~ .,  
 data = train,  
 gamma = hyper\_grid$gamma[j],  
 cost = hyper\_grid$cost[j]  
 )   
   
 pred\_svm\_untuned <-predict(  
 m\_svm\_untuned,  
 newdata = test  
 )  
   
 yhat <- pred\_svm\_untuned  
 y <- test$fuel\_economy\_combined  
 e[j] <- postResample(yhat, y)[1]  
 cat(j, "\n")  
}  
  
which.min(e) #minimum MSE

The best tuned support vector machine has a cost of 32 and a gamma of .25.

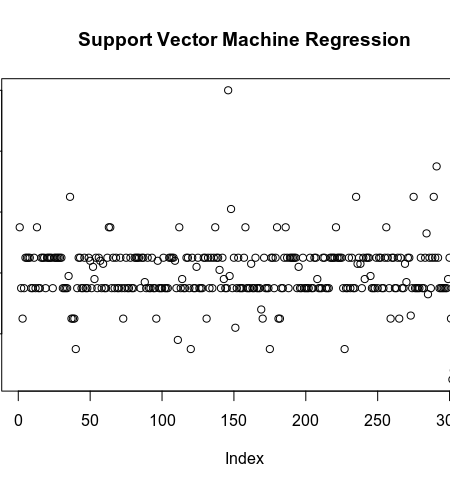
I am going to run this combination:

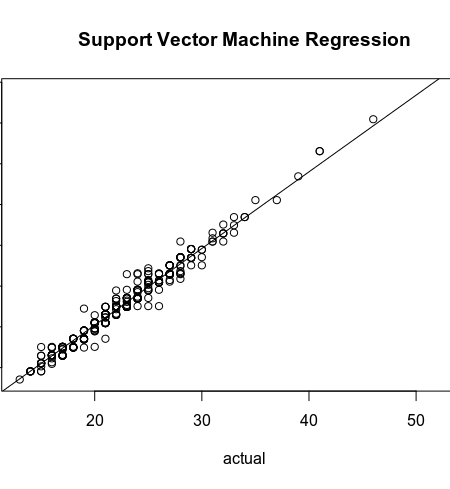
set.seed(123)  
m\_svm\_tuned <- svm(  
 formula = fuel\_economy\_combined ~ .,  
 data = test,  
 gamma = .25,  
 cost = 32,  
 scale=TRUE  
 )   
  
pred\_svm\_tuned <- predict(m\_svm\_tuned,test)  
  
yhat<-pred\_svm\_tuned   
y<-test$fuel\_economy\_combined  
svm\_stats<-postResample(yhat,y)

svm\_stats  
 RMSE Rsquared MAE   
0.9331948 0.9712492 0.7133039

The tuned support vector machine outperforms the gradient boosted model substantially with a MSE of .87 vs a MSE of 3.25 for the gradient boosted model and a MSE of 3.67 for the random forest.

summary(m\_svm\_tuned)  
  
Call:  
svm(formula = fuel\_economy\_combined ~ ., data = test, gamma = 0.25, cost = 32, scale = TRUE)  
  
  
Parameters:  
 SVM-Type: eps-regression   
 SVM-Kernel: radial   
 cost: 32   
 gamma: 0.25   
 epsilon: 0.1   
  
  
Number of Support Vectors: 232

[](https://i0.wp.com/s3-us-west-1.amazonaws.com/alpha-analysis.com/Pictures/SVM/svm_residuals.png?ssl=1)

[](https://i1.wp.com/s3-us-west-1.amazonaws.com/alpha-analysis.com/Pictures/SVM/svm_pred_vs_actual.png?ssl=1)

sum(abs(res)<=1) / 314  
[1] 0.8503185

The model is able to predict 85% of vehicles within 1 MPG of EPA estimate. Considering I am not rounding this is a great result.

The model also does a much better job with outliers as none of the models predicted the Hyundai Ioniq well.

tmp[which(abs(res) > svm\_stats[1] \* 3), ] #what cars are 3 se residuals  
 Division Carline fuel\_economy\_combined pred\_svm\_tuned  
641 HYUNDAI MOTOR COMPANY Ioniq 55 49.01012  
568 TOYOTA CAMRY XSE 26 22.53976  
692 Volkswagen Arteon 4Motion 23 26.45806  
984 Volkswagen Atlas 19 22.23552