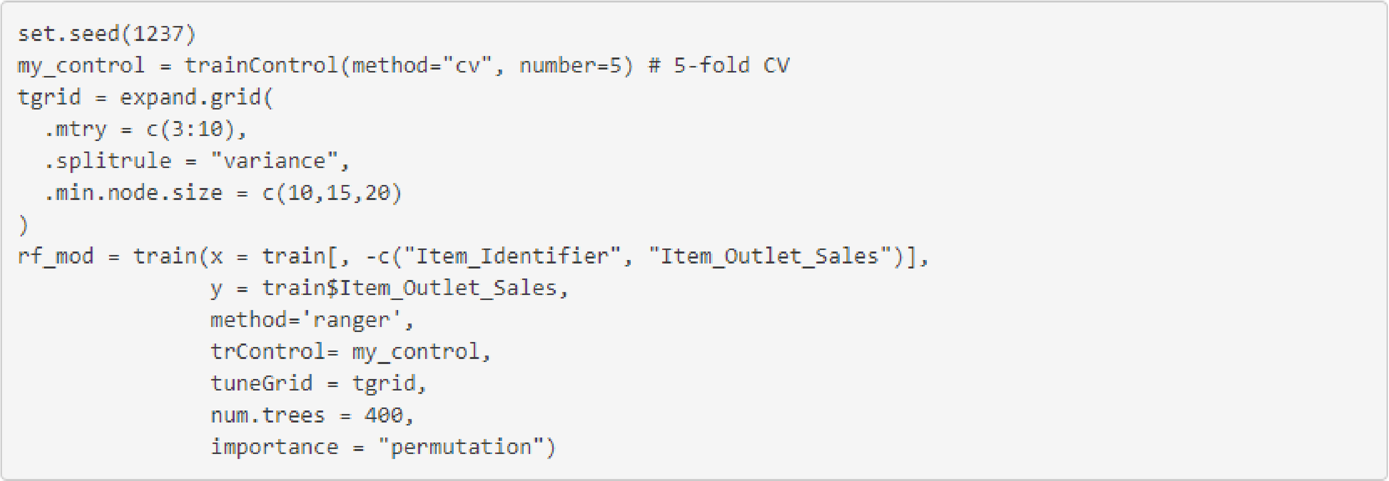
20 November 2018
20:02

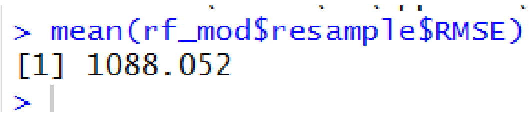
Random Forest

RandomForest is a tree based bootstrapping algorithm wherein a certain number of weak learners (decision trees) are combined to make a powerful prediction model. For every individual learner, a random sample of rows and a few randomly chosen variables are used to build a decision tree model. Final prediction can be a function of all the predictions made by the individual learners. In case of a regression problem, the final prediction can be mean of all the predictions.

We will now build a RandomForest model with 400 trees. The other tuning parameters used here are mtry — no. of predictor variables randomly sampled at each split, and min.node.size — minimum size of terminal nodes (setting this number large causes smaller trees and reduces overfitting).



```
set.seed(1237)
my_control = trainControl(method="cv", number=5) # 5-fold CV
tgrid = expand.grid(
  .mtry = c(3:10),
  .splitrule = "variance",
  .min.node.size = c(10,15,20)
)
rf_mod = train(x = train[, -c("Item_Identifier", "Item_Outlet_Sales")],
               y = train$Item_Outlet_Sales,
               method='ranger',
               trControl= my_control,
               tuneGrid = tgrid,
               num.trees = 400,
               importance = "permutation")
```



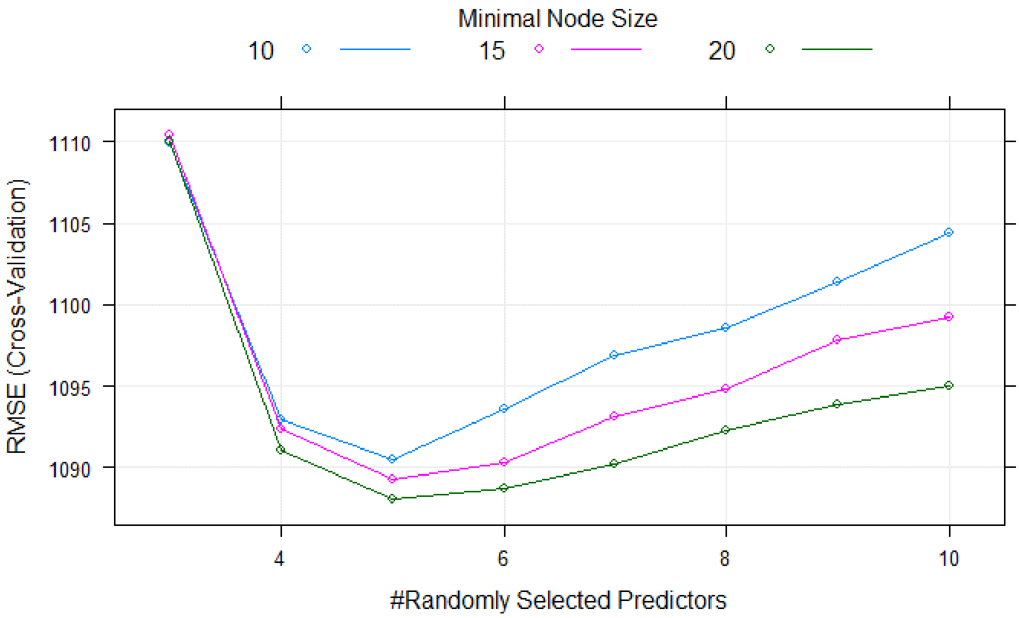
```
> mean(rf_mod$resample$RMSE)
[1] 1088.052
> 
```

Our score on the leaderboard has improved considerably by using RandomForest when we compare it with 1127.269 in Linear regression.

Now let’s visualize the RMSE scores for different tuning parameters.

Best Model Parameters

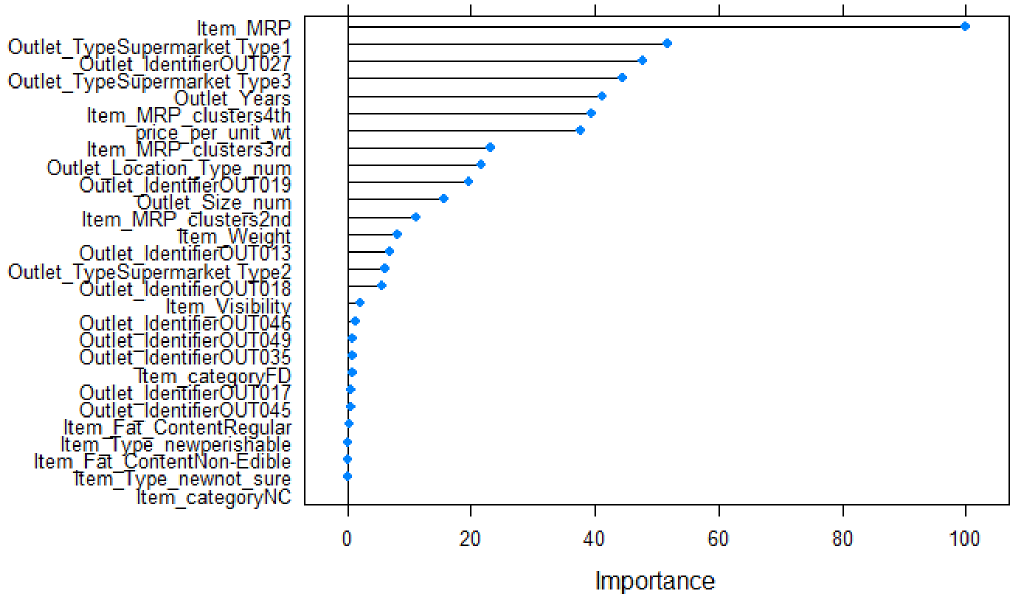
```
plot(rf_mod)
```



As per the plot shown above, the best score is achieved at mtry = 5 and min.node.size = 20.

Variable Importance

Let's plot feature importance based on the RandomForest model



As expected Item_MRP is the most important variable in predicting the target variable. New features created by us, like price_per_unit_wt, Outlet_Years, Item_MRP_Clusters, are also among the top most important variables. This is why feature engineering plays such a crucial role in predictive modeling.