Project:

I try to do the same experiment as Experiment 2 in the Homework pdf.

I try to detect the ‘on’ and ‘off’ genes using the random forest classification model and predict the expression level using the linear regression model.

Preprocess and Normalization:

Due to slow processing speed of my Laptop, I only pick data on Chromosome 2.

To preprocess the data, I manually save genes in gencode.v7.annotation.gtf(https://www.gencodegenes.org/human/release_7.html), which have in expression score in data of expression level (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRikenCage/wgEncodeRikenCageGm12878NucleusPapPlusSignalRep1.bigWig%20) and have active modification in all ten datasets shown at the end of the report. Technically, I used mergeByOverlaps() function in IRanges package to find subset of genes in annotation that has valid data points in both modification and expression datasets, by matching the ranges. The preprocess leaves us 1844 observations in total. For the expression level and modification, a single gene could include several active expressions and modification, I took the mean expression scores and modification scores, defined as [Sum of all scores]/[Length of active ranges].

I did the same normalization as Homework pdf. Expression level is transformed with log2 and pseudocount (0.1) is applied on all 10 predictors.

Modeling:

Classification

40% data is split to train random forest and the rest 60% is used as test data.

AUROC is used to measure the accuracy of the model.

Decrease in Gini index is calculated to measure the importance of different variables.

Regression

Due to the small volumes of observations for chr2, I run the regression on all data without excluding ‘off’ genes.

40% data is split to train random forest and the rest 60% is used as test data.

The importance of 10 predictors is measured by contribution to R2

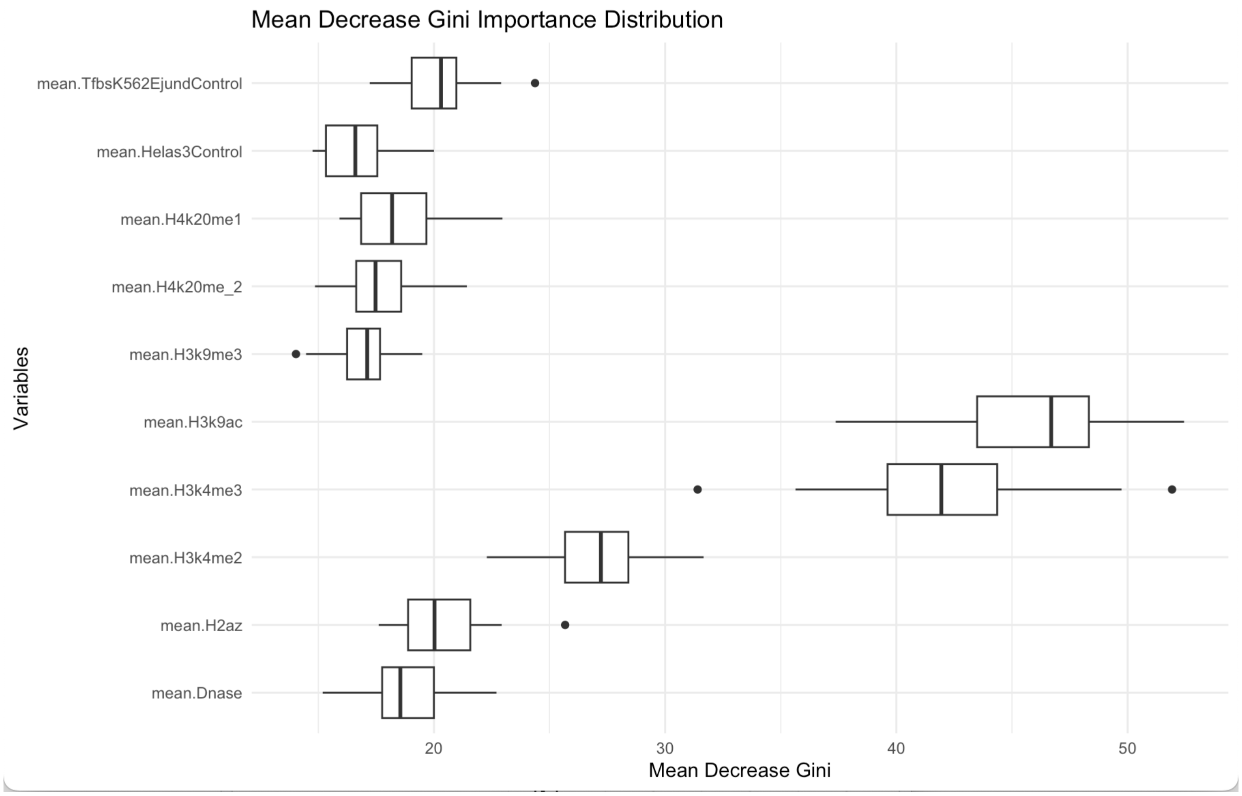
The performance of regression is measured Pearson Correlation coefficient (PCC) between the predicted value and the observed value gene expressions: r = cor(Y, $\widehat{Y}$) and the root-mean-square error in prediction (RMSE).

To validate the stability of our model, I iterated all the process 30 times and take mean of all above index including PCC, RMSE, AUROC and Decrease in Gini Index.

Results:

For the classification model, the mean AUROC is 83%, very close to the experiment 2.

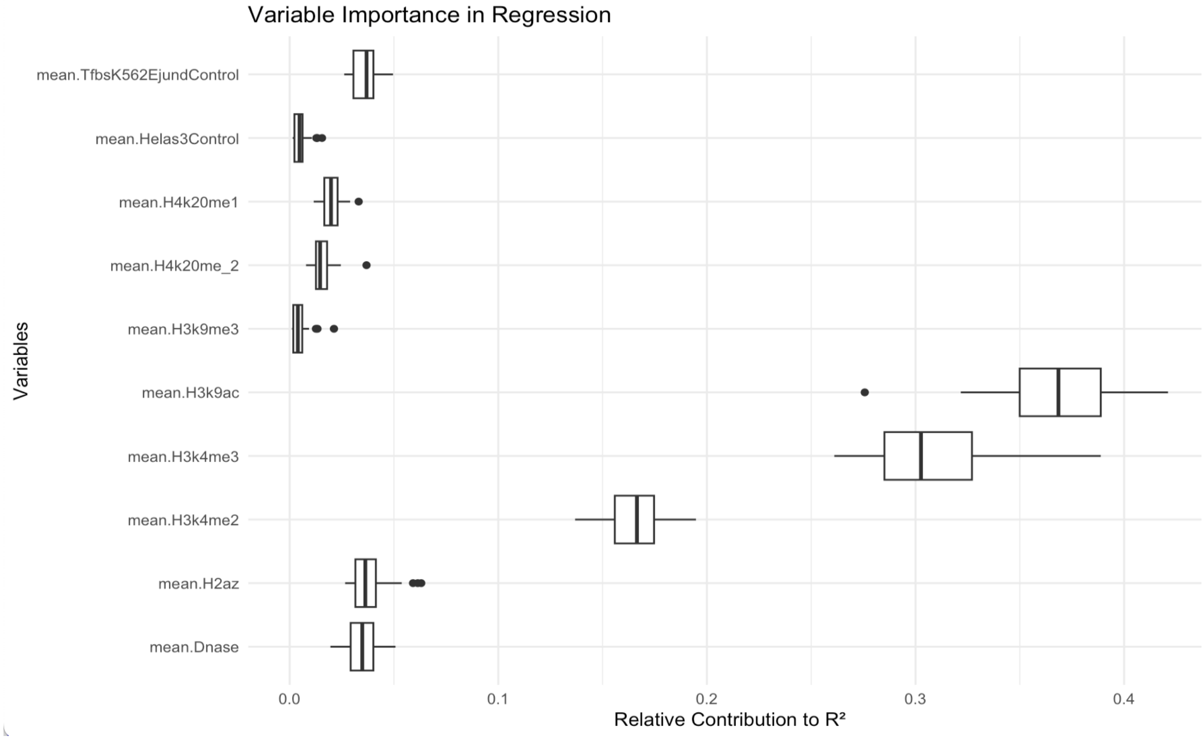
From the Decrease in Gini, we can see H3k9ac, H3k4me3, and H3K4me2 are relatively influencing in detecting on/off compared to all other modifications. H2az and Dnase which are important predictors in Dong et al(2012), are relatively less important here. For 5 predictors of low importance in Dong et al(2012), TfbsK562EjundControl surprisingly shows to be similarly important to H2az and Dnase, while the rest 4 predictors all have low importance.



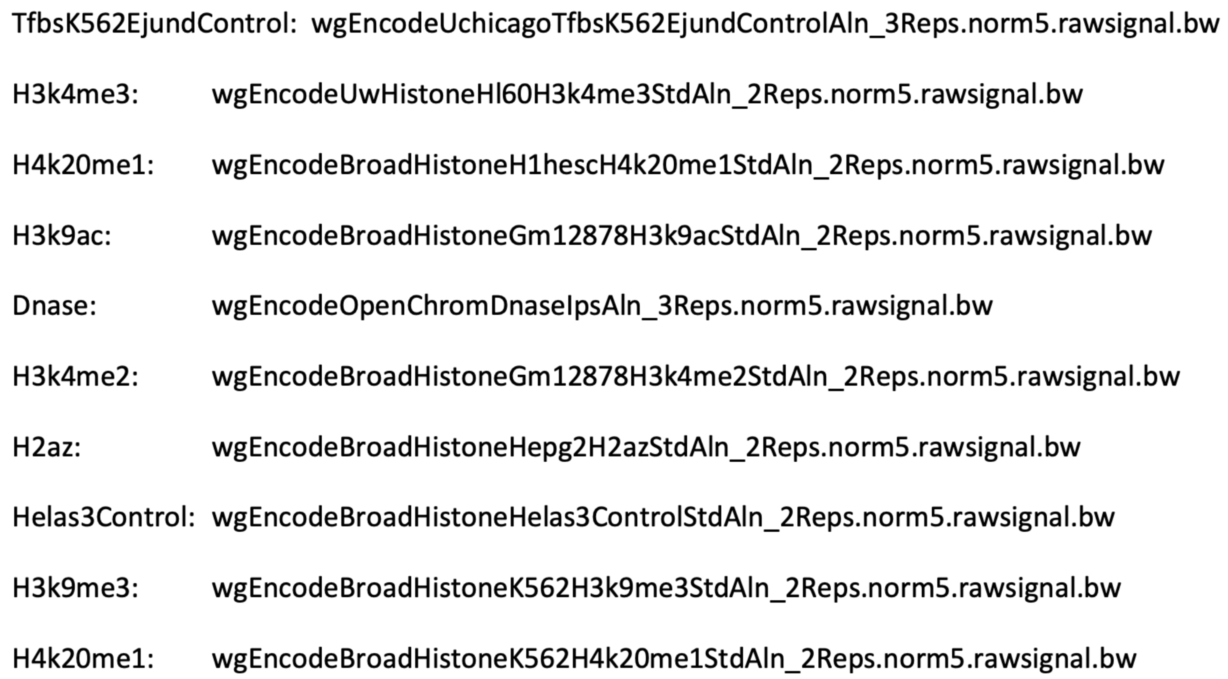
For regression, the average of Pearson Correlation is 0.555 close to result of experiment 2, but the average RMSE is 2.435 which is slightly higher than that of experiment 2.

The importance of variables situation is extremely similar to that in classification model with H3k9ac, H3k4me3 and H3k4me2 ranking 1st, 2nd and 3rd respectively.

Results from both models are highly consistent with result of experiment 2 in Homework pdf, even if only data of chr2 is used.


Variable Importance in Regression
Variables
mean.TfbsK562EjundControl
mean.Helas3Control
mean.H4k20me1
mean.H4k20me_2
mean.H3k9me3
mean.H3k9ac
mean.H3k4me3
mean.H3k4me2
mean.H2az
mean.Dnase
0.0
0.1
0.2
0.3
0.4
Relative Contribution to R²

Data of 10 predictors (https://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/signal/jan2011/bigwig/):



TfbsK562EjundControl:	wgEncodeUchicagoTfbsK562EjundControlAln_3Reps.norm5.rawsignal.bw
H3k4me3:	wgEncodeUwHistoneHl60H3k4me3StdAln_2Reps.norm5.rawsignal.bw
H4k20me1:	wgEncodeBroadHistoneH1hescH4k20me1StdAln_2Reps.norm5.rawsignal.bw
H3k9ac:	wgEncodeBroadHistoneGm12878H3k9acStdAln_2Reps.norm5.rawsignal.bw
Dnase:	wgEncodeOpenChromDnaseIpsAln_3Reps.norm5.rawsignal.bw
H3k4me2:	wgEncodeBroadHistoneGm12878H3k4me2StdAln_2Reps.norm5.rawsignal.bw
H2az:	wgEncodeBroadHistoneHepg2H2azStdAln_2Reps.norm5.rawsignal.bw
Helas3Control:	wgEncodeBroadHistoneHelas3ControlStdAln_2Reps.norm5.rawsignal.bw
H3k9me3:	wgEncodeBroadHistoneK562H3k9me3StdAln_2Reps.norm5.rawsignal.bw
H4k20me1:	wgEncodeBroadHistoneK562H4k20me1StdAln_2Reps.norm5.rawsignal.bw