Human Activity Recognition Analyst Report

**Introduction**

* Data descriptions

The R Samsung data sets consists 30 experiments subjects of total 7352 observations and of 561 different feature motion which are the raw accelerometer and gyroscope signals. Basically signals record movement in all directions of turning or accelerating during activities such as walk, walk up, walk down, running, lying and standing. Our goal is to find features as few as possible while preserving prediction accuracy and that enables us to record and analyze fewer data to save memory and processing power.

* What we do

We followed rigorous strategies and make sure every steps is reasonable. We first preprocessed data, exploratory analysis then did feature selections from different approaches. Both methods include model building, validating and performance assessments.

* Future discussion

More activities type will be took into consideration (running, jumping etc.) and it can be added into mobile application and people can be clearer about their daily activities and keep healthy.

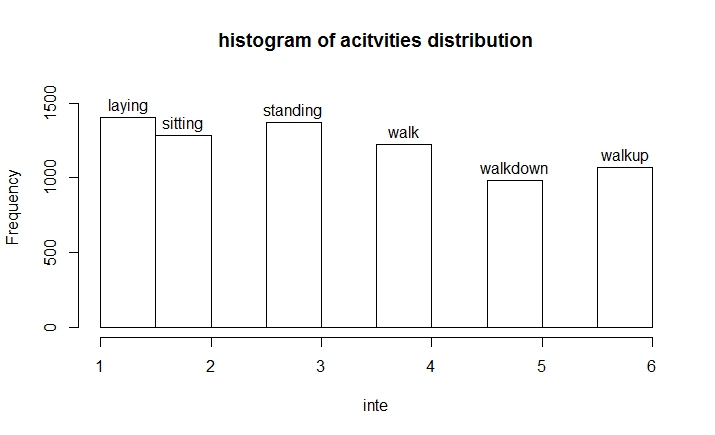
**Method**

* Data

Original data sets includes 30 volunteers within an age of 19-48 who performed six activities wearing a smartphone. Using its accelerometer and gyroscope to captured 3-axial linear acceleration and angular velocity. We used a Coursera Data Analysis Class data set of 21 subjects with 7,352 records as train data and test data from zip file of 2202 records as test set.

* Exploratory Analysis

1. 42 pairs of features have identical names and their values are very similar. t -test is performed to see if there is a big difference between two features. For example, 311 and 317 features has t -score of 6.32 and small p value, so that these duplicated features have different values. Maybe some features were recorded twice during one experiment, thus duplicated feature can’t be simply removed. All feature are renamed uniquely to be distinguishable instead.
2. Variables correlations are investigated. High correlated ones with 0.9 correlation coefficient will be dropped in method 2. Too many highly correlated variables may cause the model overfitting.
3. Distribution of all 6 activities.



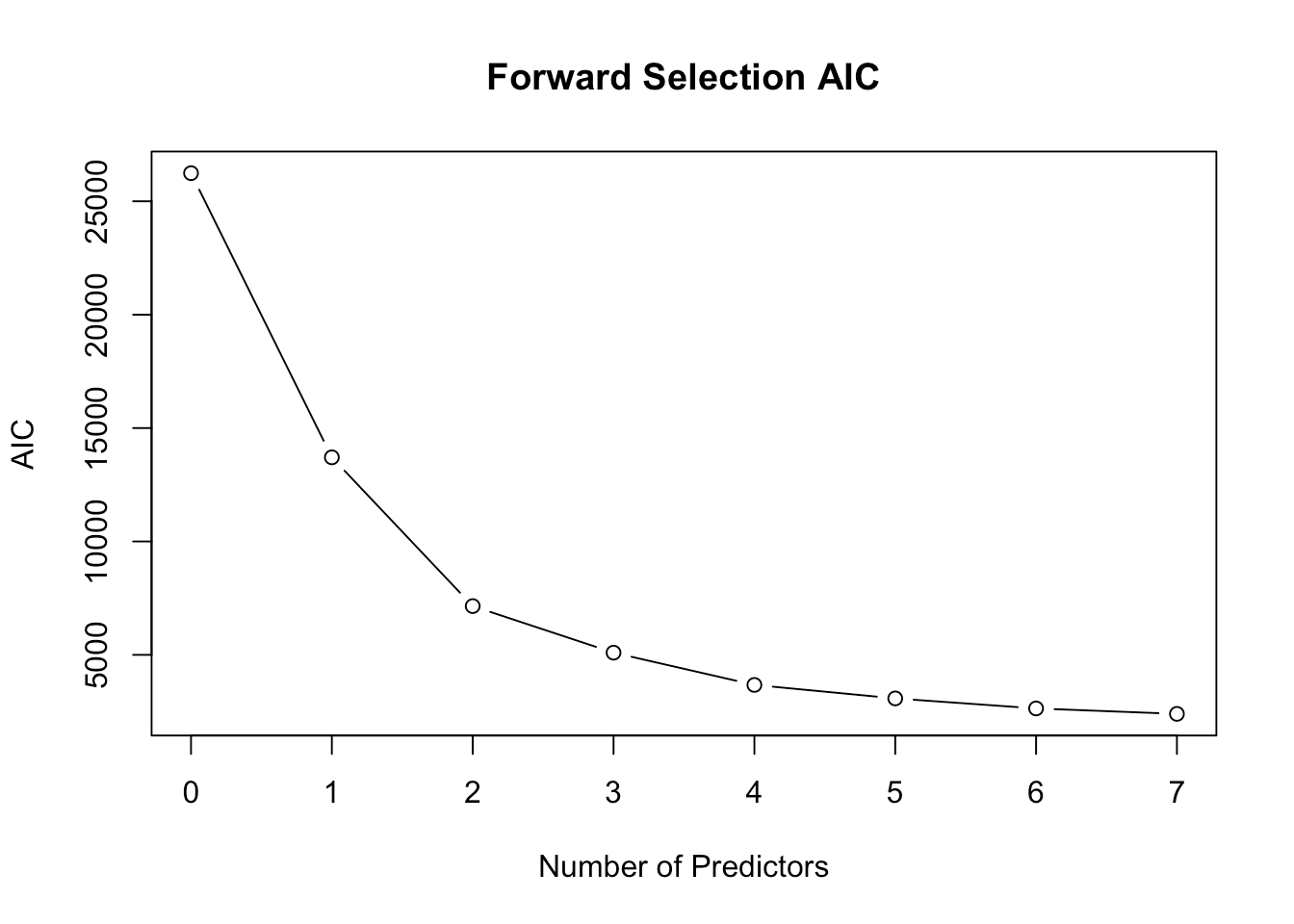
* Statistical Modeling

1. **Random Forest with Stepwise Multinomial Selections**. Wrapper method (Random Forest modeling) used to pick out candidate features then do recursive feature elimination through stepwise. By looking at feature importance (Univariate Feature Selection) stepwise the applied to screen through 50 top features. Finally, train the model by Random Forest with best 3 or 4 features to see performances.
2. **Lasso approach**. Highly correlated variables are screened out (Filter Method), then used Lasso Multinomial Classification (Embedded Method) in glmnet with penalization technique. cv.glmnet with 10 fold cross-validation is applied then different lambda values (lambda.1se, lambda.min) will return a simpler model and a small prediction error model with final selected features, coefficient.

**Analysis**

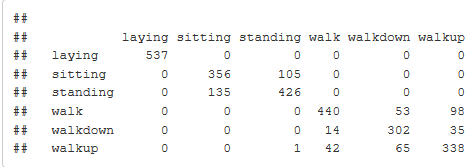
* Random Forest with Stepwise Regression

Model was trained first with 10 tress setup then Gini importance criterion ranked permutation importance. Stepwise regression (AIC criterion) filtered through top 50 features and results showed below:



Graph Explanation: AIC significantly drop after 4 predictors and 4 total features might be good where over 80% accuracy is preserved.

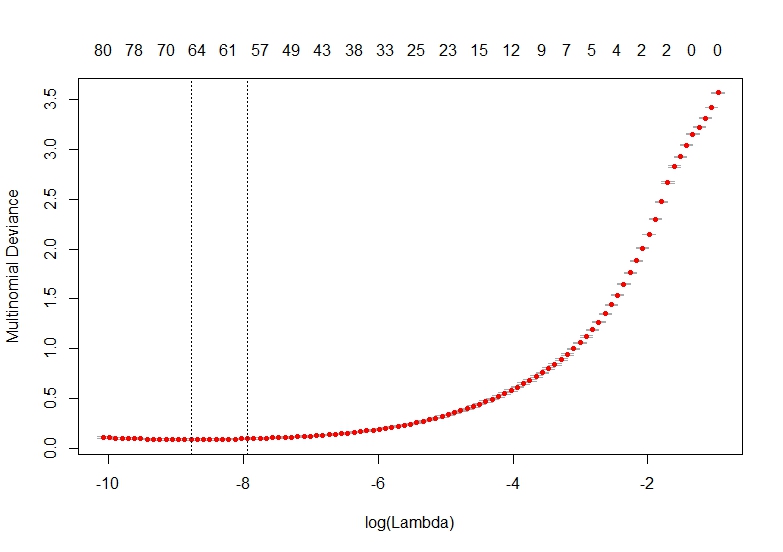
We have the following confusion matrix as assessments.



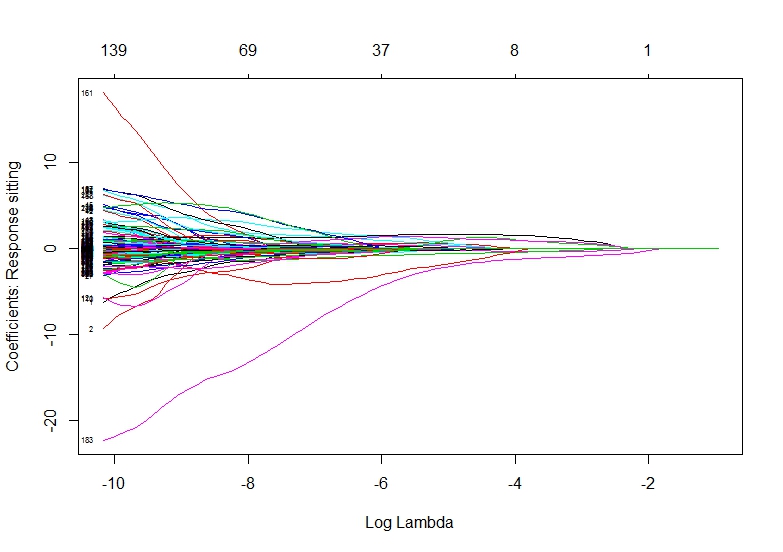
Conclusion: Final predict accuracy is 0.8140482 which is pretty good.

* Lasso and Elastic-Net Regularized Generalized Linear (glmnet) Models

After checking Pearson Correlation, 0.95 or higher correlation variables are remove which help to avoid overfitting. Penalized Maximum Likelihood in Lasso is used to compute regularization parameter lambda with input of 10 folds cross-validation and max 10000 iterations.



Graph Explanation: As lambda increase, the deviance increase significantly. Certain lambda value (lambda.1se, lambda.min) enables model to be simple with low predict error



Graph Explanation: We roughly see how coefficients are penalized during Lasso process and only 5 variables left in the model.

**Results**

Filtering out high correlation variables was not performed for Random Forest method because better trees were built when including all columns, except for low-unique valued. However, it might better for SVM, KNN and Lasso to exclude highly-correlated features. Dimension reduction technique (SVD, PCA for instances) might also pay dividends for future model development.

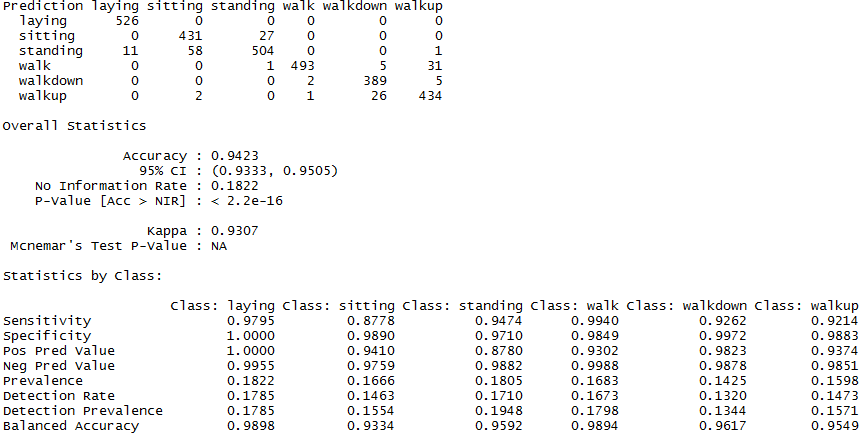
Correlation Analysis: 284 high features have correlation of 0.95 or higher. Lasso process handled with the rest 277 features.

Lambda.1se return 5 features with coefficients blow:



Accuracy on test data (2202 Obs) of 0.9494401.

Confusion Matrix and Statistics:



**Conclusion and Discussion**

UCI machine learning web described that 70% subjects are assigned to train set and 30% to test set. Researchers may assumed that different subjects may perform same activity when recording different values. We trained model on certain subjects (21 of 30 subjects) may help the model to be robust and more general when predicting unknown subjects (9 of 30 subjects).

My analysis has shown that it is feasible to develop activity prediction functions from accelerometer and gyroscope data, based on known and accepted techniques of Random Forest and Lasso. – Which have

1. Acceptable performance in the range: 80% to 95% correct classification rate across a range of sub-samples.

2. Levels of cross-validation error which seem to indicate the models will be robust to future data sample from similar studies.

3. Practical implementable characteristics, both models are simple with few features.

4. Random Forest with Stepwise selection is a greedy approach which aim pursing high accuracy and Lasso aims to build model which focusing on avoiding bias and keep it simple and robust.

We are encouraged to build on these results and it will be of interest to practitioners and researchers whose work is either directly or indirectly concerned with studying the Activities of Daily Living.

Further studies should be generalized to similar classes of problems in other domains – where categorical information must be obtained quickly and reliably under possibly noisy conditions from large numbers of inputs. We need to learn more about physical knowledge and it help to us to come up with better strategies when making decision of dropping highly correlated features.

**Reference**

glmnet.

https://cran.r-project.org/web/packages/glmnet/index.html

Preprocessing about high correlation variables.

<http://topepo.github.io/caret/pre-processing.html#identifying-correlated-predictors>

cv.glmnet

<https://www.rdocumentation.org/packages/glmnet/versions/2.0-5/topics/cv.glmnet>

An introduction to feature selections

<http://machinelearningmastery.com/an-introduction-to-feature-selection/>