**Title: Predicting a subjects activity based on Samsung phone data**

**Introduction:**

Your task is to build a function that **predicts what activity a subject is performing based on the quantitative measurements** from the Samsung phone.

* Does the analysis have an introduction, methods, analysis, and conclusions?
* Are figures labeled and referred to by number in the text?
* Is the analysis written in clear and understandable English?
* Are the names of variables reported in plain language, rather than in coded names?
* Does the analysis report the number of samples?
* Does the analysis report any missing data or other unusual features?
* Does the analysis include a discussion of potential confounders?
* Are the statistical models appropriately applied?
* Are estimates reported with appropriate units and measures of uncertainty?
* Are estimators/predictions appropriately interpreted?
* Does the analysis make concrete conclusions?
* Does the analysis specify potential problems with the conclusions?

Methods:

Data Used

The data use for this analysis were the Human Activity Recognition Using Smartphones Dataset.[1]

The database was built from recording the data produced by a smartphone (Samsung Galaxy S II) worn on the waist of 30 subjects while they performing normal daily activities, broadly grouped into 6 categories: standing, sitting, laying, walking, walking vertically up or down, such as a flight of stairs. The data were gathered from the smartphones’ embedded accelerometer and gyroscope (to captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz and were pre-processed by applying noise filters, etc.), then post-processed based on video-recorded to label the data manually.

For each activity recorded for each subject in the experiment, a broad range of data from the accelerometer and gyroscope were supplied, including 561-feature vector with time and frequency domain variables. The features selected for this database come from the accelerometer and gyroscope 3-axial raw signals processed by a time domain signals, the acceleration signal was then separated into body and gravity acceleration signals, then the body linear acceleration and angular velocity were derived in time to obtain Jerk signals and the magnitude of these three-dimensional signals were calculated using the Euclidean norm and finally, a set of factors to indicate frequency domain signals were included. These signals were used to estimate variables of the feature vector for each pattern:

'-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.

How a Gyroscope and Accelerometer are Used Together in a Smartphone[3]

When used together, a gyroscope and accelerometer provide a six-axis interpretation of movement through space. An accelerometer is used to measure sudden increases in speed within a certain range of motion. In smartphones, an accelerometer can interpret the orientation of the phone to change the display from portrait to landscape mode or interpret sudden motions such as shaking, used for some application interaction like playing certain types of games on the smartphone. Calculation of movement is possible through the measurement of the change in electronic signal as registered by the accelerometer. Depending on the angle of the device in relation to the ground, a microscopic amount of mass will move within the accelerometer circuit to cause this change in signal. These signals were captured and with pre- and post-processing, used to create the quantitative dataset used to build a function that would predicts which of the 6 activity the subject was performing (standing, sitting, laying down, walking, walking vertically up or down).

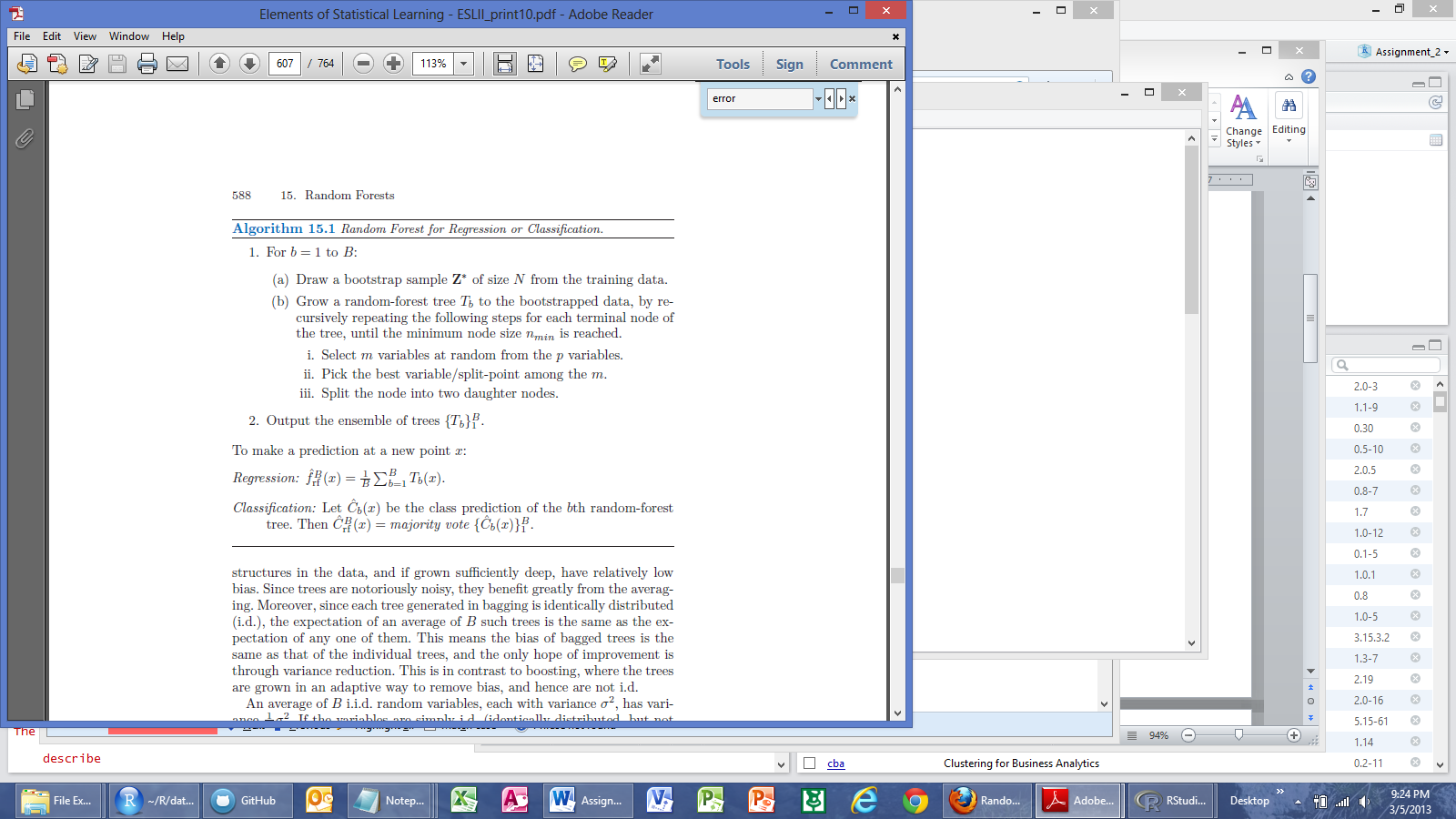
There were no missing values in the dataset.

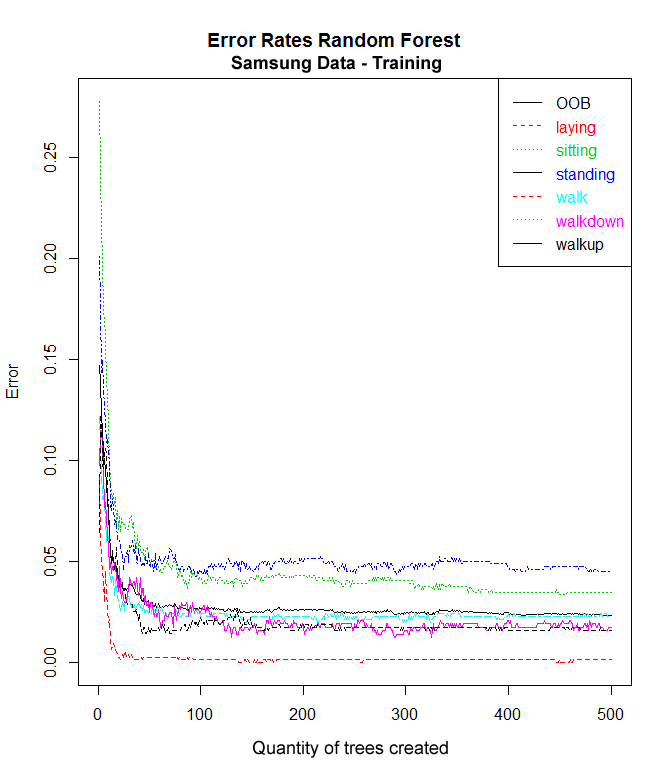
Opting for a Random Forest Model

When considering the complexity of the data and the various models types to choose from, expert opinions were reviewed regarding the usefulness of Random Forest Models in circumstances where there were a large number of variables, and with regards to variance and correlation. According to Togaware Pty, a random forest model is an ensemble of un-pruned decision trees which are often used when there are a very large number of input variables (hundreds or more). The algorithm is efficient with respect to a large number of variables since it repeatedly subsets the variables available. [4]

Future, *The Elements of Statistical Learning* states:

The idea in Random Forest™ (Algorithm 15.1) is to improve the variance reduction of bagging by reducing the correlation between the trees, without increasing the variance too much. This is achieved in the tree-growing process through random selection of the input variables. Specifically, when growing a tree on a bootstrapped dataset: Before each split, select m ≤ p of the input variables at random as candidates for splitting.[5]





Therefore, since Random Forest™[10] models are well suited for dataset with similar circumstances to the Human Activity Recognition Using Smartphones Dataset, it was determined that a Random Forest model would be used to build a function that predicts what activity a subject is performing based on the quantitative measurements from the Samsung phone. Since the target/outcome variable (activity) was categorical, a Classification type of Random Forest model was used. The model training and validation were performed using the R package randomForest[7] while utilizing the interactive user interface R package Rattle[8].

Preparing the Model

The Testing dataset for the analysis was defined as the observations associated with the last 4 subjects (id 27 through 30). These data were split from the full dataset and set aside. The remaining data (subject ids 1 through 26) were split into three subgroups used for Training (70%), Validation (15%) and the final 15% was used as a “preliminary” Testing dataset. When the model is evaluated, the default dataset is the Validation data. The “preliminary” Testing dataset can be selected to provide an unbiased error estimate. The same seed (42) was used each time the data was partitioned, so that the partitioned data did not change during modeling.

All of the 561 features (which was a vector of variables with values ranging from +1 to -1), were explored and measured for correlation, but ultimately all were included in the Random Forest models. The Random Forest algorithm builds multiple decision trees from different samples of the dataset, and while building each tree, random subsets of the available variables are considered for splitting the data at each node of the tree. A simple majority vote is then used for prediction in the case of classification (and average for regression). Random Forests are generally robust against over fitting. [4].

While working the interactive user interface R package Rattle[8], the default is to build 500 trees and to select the square root of the number of variables as the subset to choose from at each node. In the dataset used, the default is 23 since there were 561 features. However, variations in the number of variables were used to evaluate additional model performance – the quantities selected were 13, 23, 33, 43, 53 – the default square root of the number of variables +/- increments of 10 variables. Generally, the resulting models were not very sensitive to the changes in this parameter.

Results:

An estimate of the error rate was provided as the out-of-bag (OOB) estimate. This applies each tree to the data that was not used in building the tree to give an estimate of the error rate. [4]

Error Rates are measured and reported as subsequent trees are generated, up to the default (or modified) 500 trees. The Error Rates for the OOB and all activities (standing, sitting, laying, walking, walking vertically up or down) decreased substantially within the first 100 trees built (see Figure 1).



Conclusions:



**References**

1. Human Activity Recognition Using Smartphones Data Set, Version 1.0, Jorge L. Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto, Smartlab - Non Linear Complex Systems Laboratory, DITEN - Università degli Studi di Genova, Via Opera Pia 11A, I-16145, Genoa, Italy, www.smartlab.ws URL: <http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>
2. <http://www.wisegeek.com/how-are-a-gyroscope-and-accelerometer-used-together.htm>
3. How Are a Gyroscope and Accelerometer Used Together? Wise Geek, Clear answers for common questions URL: <http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm>
4. Rattle 2.6.25 r42 - The R Analytical Tool To Learn Easily (A GNOME Data Miner Built on R) 2013-01-22 Copyright (C) 2006-2013 Togaware Pty Ltd. Users guide and help topics.
5. The Elements of Statistical Learning (2nd ed.), Hastie, Tibshirani and Friedman (2008). Springer-Verlag. 763 pages, URL: <http://www-stat.stanford.edu/~tibs/ElemStatLearn/> Chapter 15 Random Forests
6. Data Science with R, Building Models|A Template, 13th February 2013, Graham.Williams@togaware.com, Togaware Pty Ltd.
7. <http://cran.r-project.org/web/packages/randomForest/randomForest.pdf>
8. <http://rattle.togaware.com/> and <http://cran.r-project.org/web/packages/rattle/rattle.pdf>
9. ROC curve Wikipedia page URL: <http://en.wikipedia.org/wiki/Receiver_operating_characteristic>
10. Random Forests(tm) is a trademark of Leo Breiman and Adele Cutler and is licensed exclusively to Salford Systems for the commercial release of the software. The trademarks also include RF(tm), RandomForests(tm), RandomForest(tm) and Random Forest(tm).