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

**Introduction:**

**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 – the Samsung Galaxy S II – worn on the waist of 30 subjects while they performed normal daily activities, 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 and then post-processed based on video-records of the subjects used to label the activities manually.

When used together in a smartphone, 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. A gyroscope works by interpreting the shift in positioning from a set rate of rotation within the X, Y, and Z axis (left/right, up/down, forward/backward). When a smartphone is tilted toward the sky the gyroscope is able to compare this movement to its normal state and calculate the change. When a gyroscope and accelerometer are combined, it is possible to simultaneously measure acceleration and gravitational placement in the X, Y, and Z axis to create a more accurate measurement of overall movement and location through space by providing constant, cross-referenced measurements of spatial placement and acceleration.[2,3]

The 3-axial linear acceleration and 3-axial angular velocity signals were captured, and with pre- and post-processing, used to create the quantitative dataset. For each activity recorded for each subject in the experiment, a broad range of data was supplied, including a 561-feature vector with time and frequency domain variables. The acceleration signal was 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 and a set of factors to indicate frequency domain signals were included. There were no missing values in the dataset.

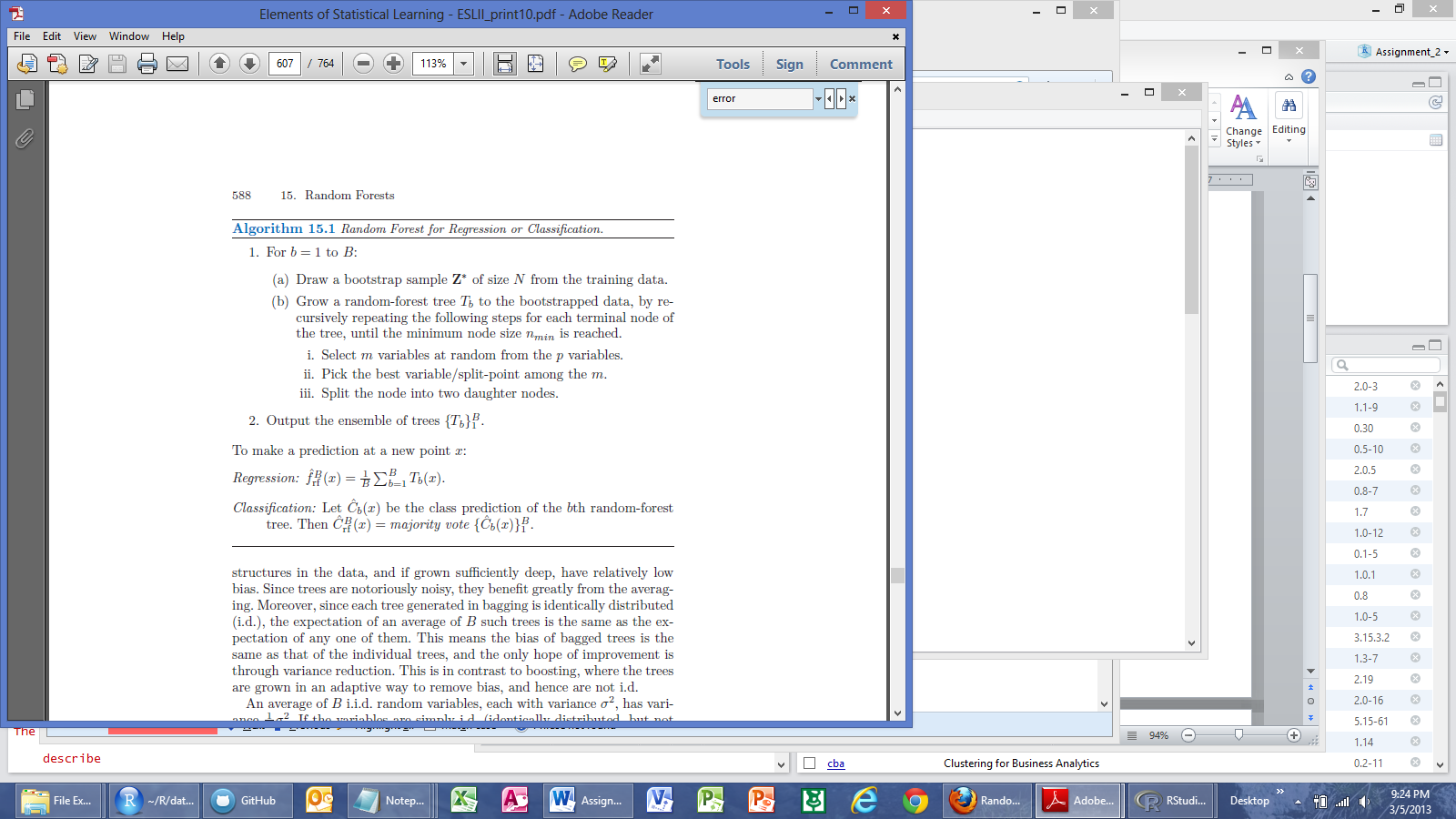
This highly technical and very robust dataset was used to build a model that predicts which of the 6 activity a subject was performing (standing, sitting, laying down, walking, walking vertically up or down).

Opting for a Random Forest Model:

Expert opinions were reviewed regarding the best type of model to use when there are a large number of variables and the data are complex. It was clear that Random Forest Models are quite usefulness in such circumstances. 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) – there were 561 in this dataset. 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 below) 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™[6] models are well suited for dataset similar 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 (subject ids 27 through 30 – 1,485 observations were used to test the model). These data were split from the full dataset and set aside. The remaining data (subject ids 1 through 26 – 4,106 observations were used to build the model) 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 was evaluated, the default evaluation dataset was the Validation data. The “preliminary” Testing dataset could then be selected to provide an unbiased error estimate. The same seed (42) was used each time the data was partitioned.

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. Reducing the number of features included in the Random Forest model reduces both the correlation between any two trees in the forest (increasing the correlation increases the forest error rate) and the strength of each individual tree in the forest (a tree with a low error rate is a strong classifier, increasing the strength of the individual trees decreases the forest error rate). Increasing the number of features included increases both the correlation and the strength[6].

While working the interactive user interface R package Rattle[8], 500 is the default number off trees to build. The number of features to choose from at each node is automatically calculated as the square root of the number of all variables available to the model. In the dataset used, there were 561 features so the number of features to choose from at each node was set at 23. 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.

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 with classification models. Random Forests models do not overfit.[6]

An estimate of the error rate was provided as the out-of-bag (OOB) estimate. The OOB 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 each subsequent tree is generated, up to the default quantity of trees. The Error Rates for the OOB and all activities decreased substantially within the first 100 trees built (see Figure 1). The OOB estimate of error rate was 2.34% on the Training data.

**Results:**

Once the Random Forests model was created, the final step in the modeling process was to Score the Test data using the model. The following chart quantifies the results:



Several of the activities were predicted with a high degree of accuracy. The activities that were challenging for the model to predict were standing and sitting. The Hit and Miss percentages illustrate how laying, walking, walking vertically up or down, were all accurately classified, while sitting and standing were more often misclassified.

Random Forest models do not have the same concerns for potential confounders as other types of models. However, reducing the number of features used to build the model nodes reduces both the correlation and the strength of the model, while increasing the number increases both the correlation and the strength. Further testing should be conducted to evaluate if the model could be improved by changes in this parameter.

**Conclusions:**

A Random Forest model, an ensemble of un-pruned decision trees, was used to predict which of the 6 activity a subject in the Human Activity Recognition Using Smartphones Dataset was performing (standing, sitting, laying down, walking, walking vertically up or down). The Random Forest model was an excellent choice because the algorithm benefits from a large number of variables and yet the model was very fast, returning results in just over 2 minutes.

However, not all activities were predicted with equal accuracy. Adjustments to the model should be explored. A useful option may be to rerun the model using only those variables that were most important in the original run, or perhaps the error could be balanced by setting different weights. Another option is looking at interactions between variables.

**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. Random Forests URL: <http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm>

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).

1. R package randomForest: <http://cran.r-project.org/web/packages/randomForest/randomForest.pdf>
2. R package rattle: <http://rattle.togaware.com/> and <http://cran.r-project.org/web/packages/rattle/rattle.pdf>
3. ROC curve Wikipedia page URL: <http://en.wikipedia.org/wiki/Receiver_operating_characteristic>
4. OOB error estimate Random Forests, Leo Breiman and Adele Cutler, URL: <http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#ooberr>