Predicting Human Activity using Smartphone Sensors

By: David Estoque

Human Activity Recognition Data

- Experiment carried out by:
    - Smartlab
        - Non-Linear Complex Systems Laboratory
    - CETpD
        - Technical Research Centre for Dependency and Autonomous Living
            - Polytechnic University of Catalonia
- 30 subjects were tracked performing six activities:
    - Walking
    - Walking Upstairs
    - Walking Downstairs
    - Sitting
    - Standing
    - Laying
- Data was collected using the accelerometer and gyroscope of a Samsung Galaxy S II Smartphone
    - Captured 3-axial linear acceleration and 3-axial angular angular velocity

Data Attributes

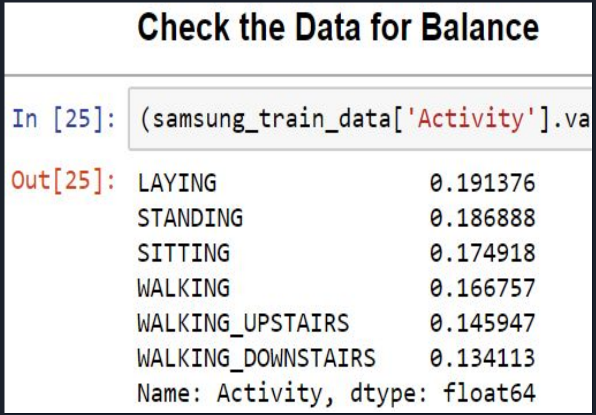
- Total of 561 features
- All features were normalized and bounded between -1 and 1
- Triaxial acceleration signal obtained from phone's accelerometer
    - Separated into:
        - Body acceleration- tBodyAcc-XYZ
        - Gravity Acceleration- tGravityAcc-XYZ
- The 't' represents time domain signals
- Body linear velocity and angular velocity derived to obtain jerk signals
    - tBodyAccJerk-XYZ and tBodyGyroJerk-XYZ
        - Magnitude of signals collected as well
            - tBodyAccMag
            - GravityAccMag
            - tBodyAccJerkMag
            - tBodyGyroMag
            - tBodyGyroJerkMag

Data Attributes

- The signals were then separated into the 3-axial signals of X, Y, and Z vectors
- These signals were then estimated to create a set of variables:
  - Variables related to Central Tendency
    - Mean, median, quartiles
      - Example, tBodyAcc-mean()-X
  - Distribution Measurements
    - Skewness, kurtosis
      - Example, fBodyBodyGyroJerkMag-kurtosis()
  - Correlation coefficients between signals
    - Example, tBodyAccJerkMag-arCoeff
- All these led to a features count of 561

Data Exploration

- The Target Variable was the “Activity” column
  - “Subject” column removed
- The data contained a Test set and Training set
  - Training Set Size
    - 7352 rows, 563 columns
  - Test Set Size
    - 40% of training set data
    - No need to do split the test data
    - 2947 rows, 563 columns



Check the Data for Balance

```
In [25]: (samsung_train_data['Activity'].va

Out[25]: LAYING                0.191376
         STANDING              0.186888
         SITTING               0.174918
         WALKING               0.166757
         WALKING_UPSTAIRS      0.145947
         WALKING_DOWNSTAIRS    0.134113
         Name: Activity, dtype: float64
```

Research Questions

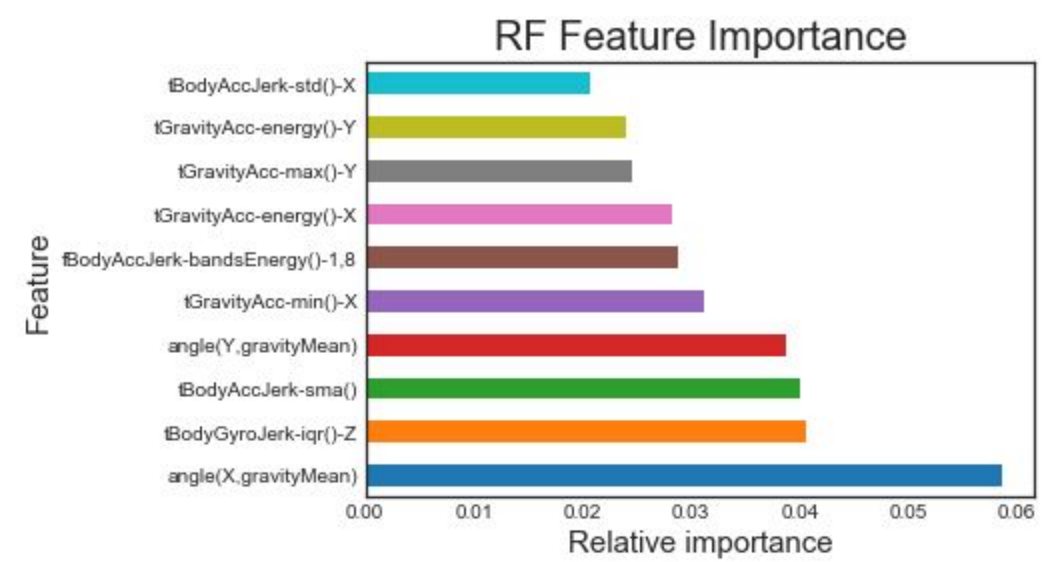
- Can models accurately predict human activity?
- What models are the most accurate?
- What are the top 10 features for predicting Human Activity?

Random Forest Classifier

Model	Random Forest					
Variable Size	561	300	200	100	50	20
Variable Size (%)	100	53.5	35.7	17.8	8.9	3.6
Test (%)	90.7	89.3	91.3	90.4	87.4	83.0
Train (%)	90.8	90.8	91.0	90.3	89.3	87.3
Runtime (s)	4.73	3.4	2.88	1.87	1.28	0.86

- RFC with defaults
  - Yielded 90% accuracy
  - Run time 4.8 seconds
- Using feature importance I examined the top 20, 50, 100, 200, 300 features
- Weakest performing was top 20
  - Accuracy 83%
- Top performing was an RFC with 200 Features
  - 91.3 percent accuracy
  - Run time 2.8 seconds

10 Most Important RFC Features



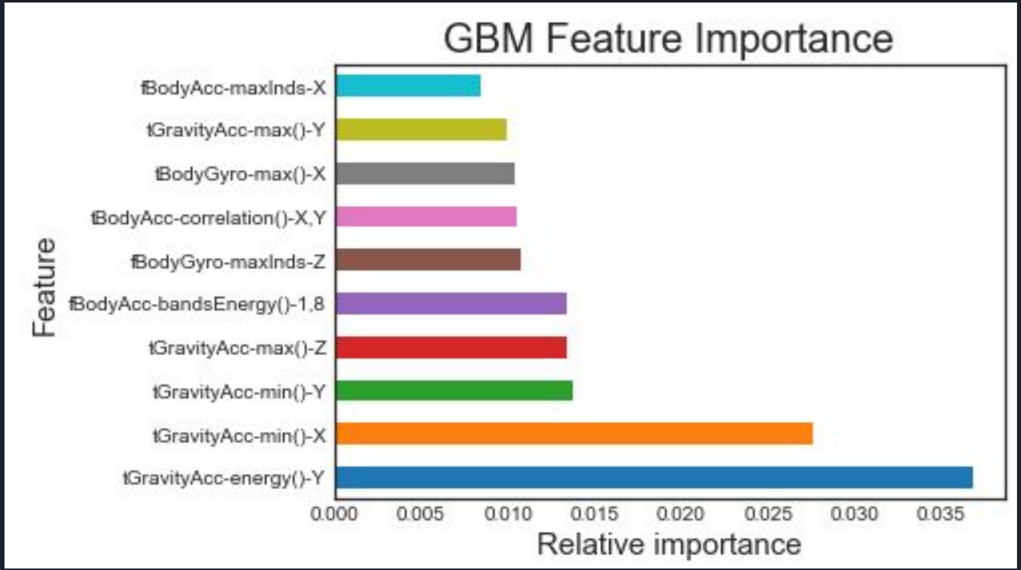
- Used feature importance built-in function from the Random Forest SKLearn Package

Gradient Boosting

Model	GBM	
Variable Size	561	200
Variable Size (%)	100	35.7
Test (%)	94.5	94.5
Train (%)	92.9	92.5
Runtime (s)	1720	734

- GB with all features has high accuracy
- Like SVC I examined GBM with a variable size of 200
    - Similar accuracy and 58% faster!

Gradient Boosting Feature Importance



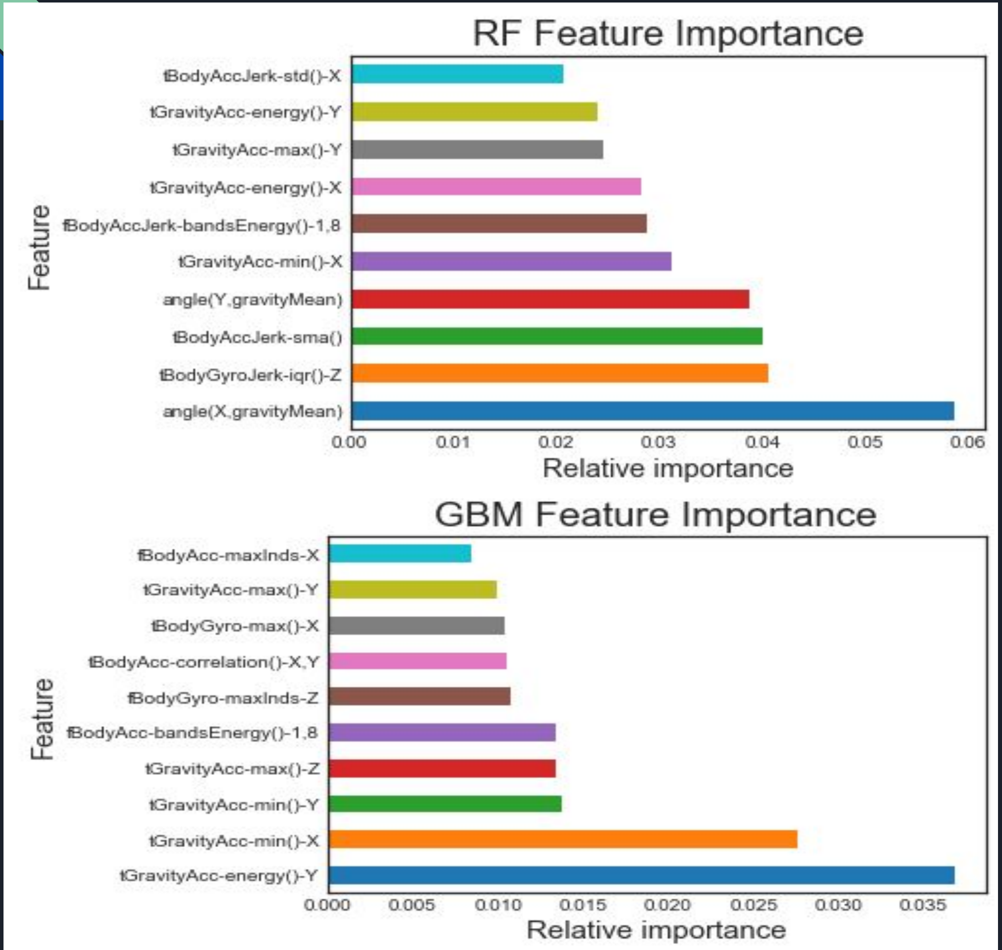
- Feature importance obtained using the feature_importances_ built-in function in sklearn

Support Vector Classifier

Model	SVM		
Variable Size	561	200	100
Variable Size (%)	100	35.7	17.8
Test (%)	96.4	96.2	93.5
Train (%)	94.1	92.6	91.8
Runtime (s)	13.6	4.4	2.3

- Using all the features:
  - 96 percent accuracy
  - 13.6 s
- 200 features was the top performing RF, therefore, I looked at the SVC of the top 200 features
  - 96 percent accuracy
    - Slightly lower than all features
  - 4.6 s

Feature Importance Comparisons



- 6 Shared Important Features:
  - tGravityAcc-energy
  - tGravityAcc-max
  - fBodyAccJerk-bandEnergy
  - tGravityAcc-min
  - Angle(, gravityMean) (2)
- Top GBM Feature importance
  - tGravityAcc-Energy()-Y
- Top RF Feature Importance
  - angle(X, gravityMean)

Conclusions

- Results
- Findings
- Limitations
- Practical Applications

Results and Findings

Model	Random Forest						SVM			GBM	
Variable Size	561	300	200	100	50	20	561	200	100	561	200
Variable Size (%)	100	53.5	35.7	17.8	8.9	3.6	100	35.7	17.8	100	35.7
Test (%)	90.7	89.3	91.3	90.4	87.4	83.0	96.4	96.2	93.5	94.5	94.5
Train (%)	90.8	90.8	91.0	90.3	89.3	87.3	94.1	92.6	91.8	92.9	92.5
Runtime (s)	4.73	3.4	2.88	1.87	1.28	0.86	19.1	6	3.7	1720	734

- Gradient Boosting
    - Provides a high amount of accuracy
    - Runtime is extremely long about 28 minutes compared to the others
    - Allows for examination of feature importance
- Random Forest
    - Very Accurate
    - Allows for examination of feature importance
    - Runtime is relatively fast
- Support Vector Machine
    - Highest accuracy at 96%
    - Does not allow for examination of feature importance
- Which is best??
-

Worst Performing Model by Activity (RFC 20)

col_0 / Activity	LAYING	SITTING	STANDING	WALKING	WALKING_DOWNSTAIRS	WALKING_UPSTAIRS
LAYING	537	0	0	0	0	0
SITTING	0	380	111	0	0	0
STANDING	0	111	421	0	0	0
WALKING	0	0	0	414	41	41
WALKING_DOWNSTAIRS	0	0	0	38	335	47
WALKING_UPSTAIRS	0	0	0	98	13	360

Top Performing Model by Activity (SVM)

col_0 / Activity	LAYING	SITTING	STANDING	WALKING	WALKING_DOWNSTAIRS	WALKING_UPSTAIRS
LAYING	537	0	0	0	0	0
SITTING	0	435	54	0	0	2
STANDING	0	16	516	0	0	0
WALKING	0	0	0	492	3	1
WALKING_DOWNSTAIRS	0	0	0	4	410	6
WALKING_UPSTAIRS	0	0	0	18	2	451

Accuracy by Activity

	Accuracy by Activity (%)					
Model	RF		SVM		GBM	
Variable Size	20	200	200	561	200	
Activity						Activity cum. Avg
Laying	**100%**	**100%**	**100%**	**100%**	**100%**	*100.0%*
Sitting	**77%**	**91%**	**95%**	**96%**	**93%**	*90.5%*
Standing	**79%**	**87%**	**90%**	**91%**	**90%**	*87.4%*
Walking	**75%**	**88%**	**97%**	**96%**	**94%**	*90.2%*
Walking Downstairs	**86%**	**95%**	**99.5%**	**99%**	**98%**	*95.7%*
Walking Upstairs	**80%**	**87%**	**97%**	**98%**	**92%**	*90.9%*

Limitations

- With a dataset with so many features there may be many relationships that weaken the model
- Difficult to truly find the most important features
- A deeper dive into the features may improve the true accuracy of the models
- RBM and RF use different algorithms to choose important features
  - Thus, producing varying results
- Consider using XGBoost

Practical Applications

- Fitness tracking applications
    - Can use the model to improve the accuracy of fitness activity detection
    - Better estimates for energy expenditure
- Mobile phone companies
    - Can use the model to evaluate the accuracy of their hardware in detecting physical movement
        - Improve phone hardware efficiency
    - Increase the accuracy human behavior monitoring

Resources

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013. Bruges, Belgium 24-26 April 2013.https://www.kaggle.com/uciml/human-activity-recognition-with-smartphones/home