**Predictive Models of Human Activity Recognition**

Jungmin Park

DSC 383W

**Introduction:**

The work completed for this mini project addressed the problem of recognizing human’s activities based on raw accelerometer and gyroscope signals of human’s body. The goal of this mini project is to develop two predictive models using different number of features of the human activity recognition data.

The main focus will be on how accurate predictive models are in predicting and how the number of features used in the prediction is related with the resulting accuracy. The following section will outline, in detail, how the human activity recognition data is collected and prepared. The remainder of the report will discuss the dataset and its properties, description of the machine learning models used for prediction analysis and classification of human activity. The report concludes with the description of results regarding the accuracy and the confusion matrix of each predictive models used for the project.

**Data Collection and Preparation:**

This project utilized a human activity recognition dataset, which is found from the UCI Machine Learning Repository. This data was collected by conducting experiments that were carried out with a group of 20 volunteers within an age bracket of 19-48 years while wearing a smartphone on the waist. It contains 10299 records in total collected by 30 volunteers, each row corresponding to a 2.5 s segment from a volunteer.

In the human activity recognition dataset, there are total 563 columns. The first 561 columns indicate the feature variables including accelerometer and gyroscope sensor signals and time and frequency domain variables. Next, column 562 indicates the subject index of who performed the activity for each window sample, ranging from 1 to 30. Lastly, column 563 shows the six types of activity including walking, walking\_upstairs, walking\_downstairs, sitting, standing, and laying.

The dataset obtained has been randomly split into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.

**Exploratory Analysis:**

Exploratory analysis was performed by examining different kinds of plots of the human activity recognition data set. This analysis is mostly concentrating on identifying missing values and exploring the correlation between the sensor signals and activity type.

Before going through the actual analysis of data, one of the important steps is handling missing data. Missing values results in a smaller sample size and larger standard errors. So, this leads to lowering the chance to find a significant result from the data set and correctly accepting the relationship between the variables. Therefore, it is significant to inspect the missing data before doing further analysis. When checking through the human activity recognition dataset using the python function, it had no missing values included.

Figure1 below shows the number of examples of each activity type. ‘Laying’ activity had the most examples, whereas ‘Walking downstairs’ activity had the least examples included in the human activity recognition data set.

A screenshot of a cell phone

Description automatically generated

Figure 1. Bar plot of the number of examples by activity type

Figure 2 below shows the number of examples of each subject who participated in the experiments ranging from 1 to 30. Subject 25 and 21 were the top 2 subjects that had the most example data. Subject 21 was selected to further be analyzed since it has the most example data, which will produce relatively more accurate insights of the data.

A screenshot of a cell phone

Description automatically generated

Figure 2. Bar plot of the number of example data by subject

To determine the relationship between the sensor signals and activity type, the body acceleration signals in the X, Y, and Z directions were analyzed by grouping them into activity type.

Figure 3 below is the boxplot that has activity type as x-axis and the mean value of body acceleration signals in X direction as a y-axis. ‘Walking downstairs’ activity had the longest length of the box, showing how widely stretched out the values of the body acceleration signals are. The length of the box represents the interquartile range which is the range between the minimum and maximum value. It had the biggest difference in the minimum and maximum value of body signals, so it is shown to have the most body movement changes among all the other activities. In contrast, ‘Sitting’ activity had the shortest length of the box, indicating that there’s barely a difference within each value in the body acceleration signals. This shows that ‘Sitting’ activity has less body movements.

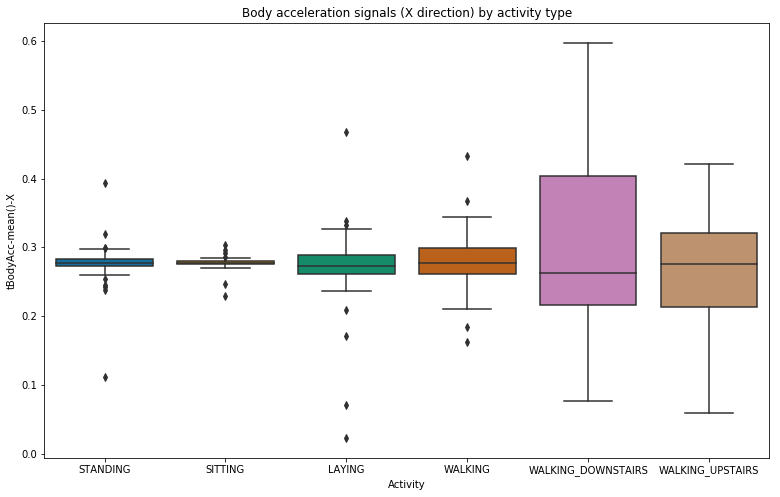


Figure 3. Box plot of body accelerations signals in X direction by activity type

Figure 4 below is the boxplot that has activity type as x-axis and the mean value of body acceleration signals in Y direction as a y-axis. Based on the plot, body acceleration signals for ‘Walking’, ‘Walking downstairs’, and ‘Walking upstairs’ activities are relatively dispersed compared to ‘Standing’, ‘Sitting’, and ‘Laying’ activities. It shows that these three activities that have longer length of box have more body movements as time pasts.

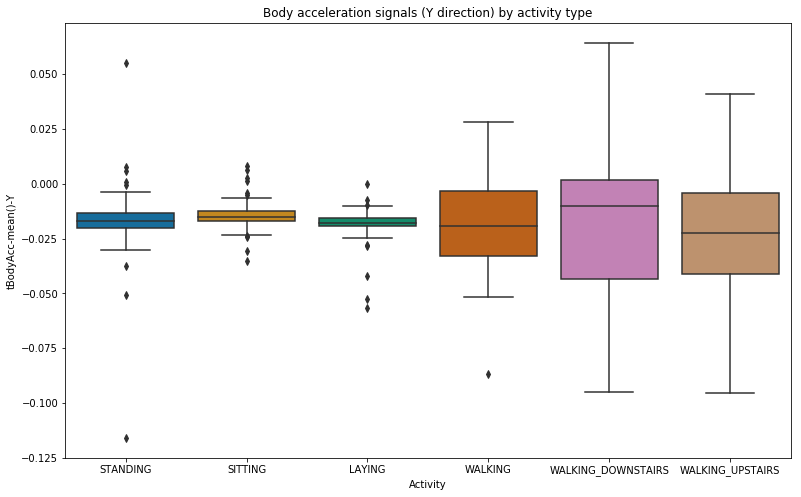


Figure 4. Box plot of body accelerations signals in Y direction by activity type

Figure 5 below is the boxplot that has activity type as x-axis and the mean value of body acceleration signals in Z direction as a y-axis. Generally, most of the boxplots of activities has a small interquartile range which indicates that the body signals in Z direction are more likely to be stable in values across the timeline.

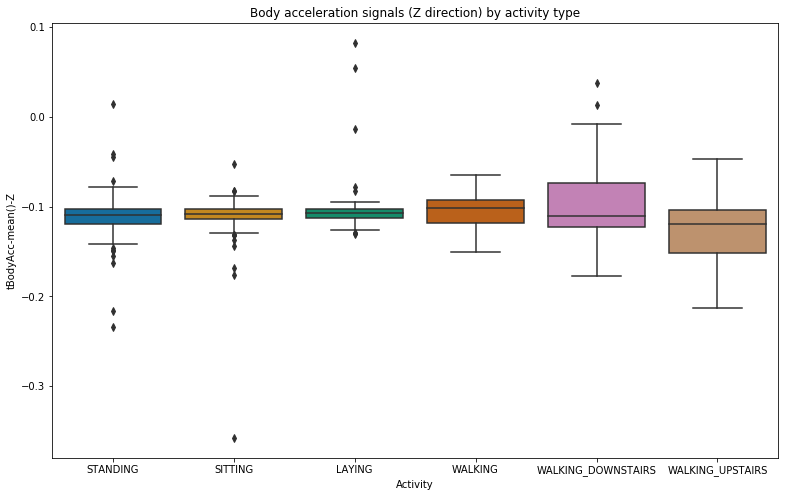


Figure 5. Box plot of body accelerations signals in Z direction by activity type

**Methods:**

As part of this project’s implementation, two predictive models were generated to classify the activity label, which are random forest classifier and multinomial logistic regression.

*Random Forest Classifier*

Random Forest Classifier is a classification algorithm that operates by constructing many decision trees while training the data and returns the result of classification or mean prediction of the individual trees. To explain in detail, random forest consists of a significant number of individual decision trees (models) that work as an ensemble. Each of the individual tree in the random forest outputs a prediction of a class and the class with the highest probability becomes the model’s prediction. The main point of the random forest classifier is that each model has low correlation with each other. Relatively uncorrelated models are able to produce ensemble predictions that are more accurate compared to any other individual predictions.

*Multinomial Logistic Regression*

Multinomial Logistic Regression is another statistical classification model that applies logistic regression to multiclass problems, which has more than two possible classes or outcomes. Logistic Regression is used to describe the given data and to explain the relationship between the dependent target variable and independent feature variables. It predicts the probabilities of the different possible outcomes of the target variable based on the given independent feature variables. This model is suitable for the human activity recognition data in that logistic regression is used when the target variable, which is activity label in this data, is categorical.

*Feature Selection Process*

Feature selection is one of the key concepts in machine learning that significantly impacts the performance of the predictive model. Varying the number of features used in the prediction resulted in different accuracy. A python function that takes the number of features as a parameter and returns the accuracy of the predictive model was made to show the relationship between the number of features selected and the corresponding accuracy. The numbers of features that

**Results:**

The primary metric collected during testing for each predictive model was accuracy. Each predictive model was tested at a 30 percent test fraction using all 561 features of the human activity recognition dataset. The accuracy of random forest classifier resulted in 90 percent and Figure 6 below is the confusion matrix table of random forest classifier with respective labels. Activity types were labeled to numbers from 1 to 6 (1-Walking, 2-Walking Upstairs, 3-Walking Downstairs, 4-Sitting, 5-Standing, 6-Laying). Random forest classifier perfectly predicted ‘Laying’ activity. On the other hand, it shows a poor performance when predicting ‘Walking’ activity, resulting total 97 errors.

A close up of a keyboard

Description automatically generated

Figure 6. Confusion matrix of Random Forest Classifier

On the other hand, multinomial logistic regression showed an accuracy of 95 percent. Performance of this model was better than random forest classifier, exceeding the model by around 5 percent in accuracy. Figure 7 is the confusion matrix table of multinomial logistic regression that is shown below. Same as above, activity types were labeled to numbers from 1 to 6. Multinomial logistic regression perfectly predicted ‘Laying’ activity, whereas it poorly predicted ‘Standing’ activity. However, multinomial logistic regression model mostly showed great performance in prediction compared to the random forest classifier model.

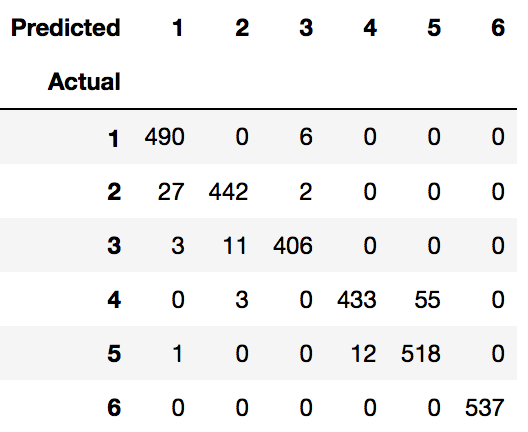


Figure7. Confusion matrix of Multinomial Logistic Regression

Number of features used in the prediction was varied to find the relationship between the number of features and the corresponding accuracy. The number of features that have shown significant changes in resulting accuracy was recorded.

Figure 8 below is the accuracy plot of the multinomial logistic regression predictive model. As shown from the graph, it shows a radical increase until 41 features used for the prediction. The left green line indicates 41 features being used and it is the point where it obtains over 80% accuracy. Starting from this green line, the plot shows a gradual increase in the accuracy. The right red line indicates 129 features, which is the number of features required to obtain 90% accuracy.

A screenshot of a cell phone

Description automatically generated

Figure 8. Accuracy plot of Multinomial Logistic Regression

**Conclusion:**

This human activity recognition analysis project implemented two predictive models using different classifier algorithms that effectively predict the overall activity type of the given sensor signals. The optimal classification model that has shown the best performance was multinomial logistic regression, resulting in higher prediction rate compared to random forest classifier model. It was generally easy to classify the ‘Laying’ activity, both of the classification model showing perfect prediction. On the other hand, it was relatively hard to classify ‘Walking’ and ‘Standing’ activities. This exploration of activity type analysis classification has brought a lot of insights of what kinds of models are good at predicting activity type.

Also, it is shown that the number of features used for prediction has a positive correlation with accuracy. The accuracy increased as more and more feature variables were used for predicting activity type. This analysis of finding relationship between the number of features used and accuracy shows that the number of features is an important factor that greatly affects the performance of the predictive model.

**References:**

1. “Logistic Regression.” Wikipedia. Wikimedia Foundation. URL: <https://en.wikipedia.org/wiki/Logistic_regression>.

2. “A Beginner's Guide to Selecting Machine Learning Predictive Models in Python.” Medium. URL: <https://towardsdatascience.com/the-beginners-guide-to-selecting-machine-learning-predictive-models-in-python-f2eb594e4ddc>.

3. “Understanding Random Forest.” Medium. URL: <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>.

4. “Random Forest.” Wikipedia. URL: <https://en.wikipedia.org/wiki/Random_forest>.

5. “Multinomial Logistic Regression.” Wikipedia. URL: <https://en.wikipedia.org/wiki/Multinomial_logistic_regression>.

6. “Logistic Regression - Detailed Overview.” Medium. URL: <https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc>.

7. “What Is Logistic Regression?” Statistics Solutions. URL: <https://www.statisticssolutions.com/what-is-logistic-regression/>.

8. Handling missing data. URL: <http://www.emgo.nl/kc/handling-missing-data/>.