IST 718 - Big Data Analytics

# Predictive Modelling on Activity Recognition for Wearable Devices

**Group 13**

Anjali Shahi[1] [1]iSchool SU ashahi@syr.edu

Akash Kandarkar[2] [2]iSchool SU aakandar@syr.edu

Chaitanya Kunapareddi[3] [3]iSchool SU ckunapar@syr.edu

# Contents

# 1. Abstract

We are all aware of the benefits of exercising, but do we understand how much exercise is ample to remain fit? Also, how do we make sure that we are continuously maintaining our fitness while we are swamped with mundane routines? Detecting human movements is an important task in various areas such as healthcare, fitness, and eldercare. Motion sensors can provide users, doctors, and related persons with a better understanding about daily physical activities.

| **Attribute Name** | **Attribute Type** | **Recorded From** | **Description** | |
|---|---|---|---|---|
| X, Y, Z (Axis columns) | Double | Accelerometer/ gyroscope | Since the data is recorded from devices which are 3D objects, these 3 columns will help get device position in all 3 dimensions | |
| Users | Categorical - string | Accelerometer/ gyroscope | The data was recorded via a survey across 9 users of devices, this column contains labels of all 9 users | |
| Model | Categorical - string | Accelerometer/ gyroscope | This column gives us the information whether the sensor has recorded data from a phone or from a watch | |
| Sensor | Categorical-String | | Accelerometer/ gyroscope | This column gives us the information, data was captured using which of the sensors |
| BjerkMean | Int | | Accelerometer/ gyroscope | This column contains the meaning of positions where the person performs the jerk which determines that there is change of activity or any activity Is being performed |
| GT | Categorical - string | Accelerometer/ gyroscope | This is our dependent column which gives information about a person's activity, whether the person was standing, sitting, walking, biking, going down or upstairs and null (no activity) | |

Smart watches have motion sensors built-in to track the movement of the user. We are going to use such data recorded from smart watches to analyze user lifestyle and whether the activities performed are sufficient to help them maintain their health. We are extracting data from accelerometers and gyroscopes present in the smartwatch sensor for this purpose. To consolidate the object, given the position of the device attached, we would predict the activity being performed.

By the end of this project, we aim to analyze user lifestyle, predict their activities using machine learning and statistical components while enhancing the predictions as well.

## 2. Dataset Description

In the dataset for this project, we will be using data recorded by two sensors, the accelerometer and gyroscope. Each of the sensors are present in smart watches of 2 models Samsung gear and LG watch. The data has around **708,604 records** with **14 attributes**. Listed below are some of the columns which will be used as independent variables and the final column is the prediction column.

To view the dataset files, Click Here. Since the data has over 14 columns, we are going to use the subset and few columns will be additionally derived such as time of the day, activity category based on activity.

## 3. Data Exploration

- Upon exploring the data, we can determine that the data has no null or missing values which means there is no discrepancy, the data is captured perfectly for every movement. The null in the activity column indicates no activity (ex: watch removed or sleeping).
- Irrespective of the activity being performed the jerk time i.e., human reaction time to the activity is around half a second, which means response to change in activity is uniform.
- Looking at the x angles of the axis, as compared to the Y and Z, we can say that a lot of motion is sideways rather than forward/backward like travelling distance or vertical like jumping, which is

quite different do day to day activities. In **Figure 1** The two spikes in x axis show that going down the stairs there is quite an inclination while sitting there is no change. Similarly for z axis we can see in **Figure 2** that biking makes you move forward while other motions happen in the same place.

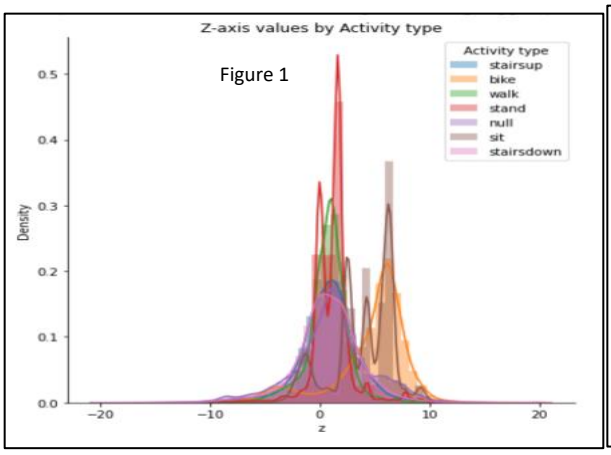



Figure 1

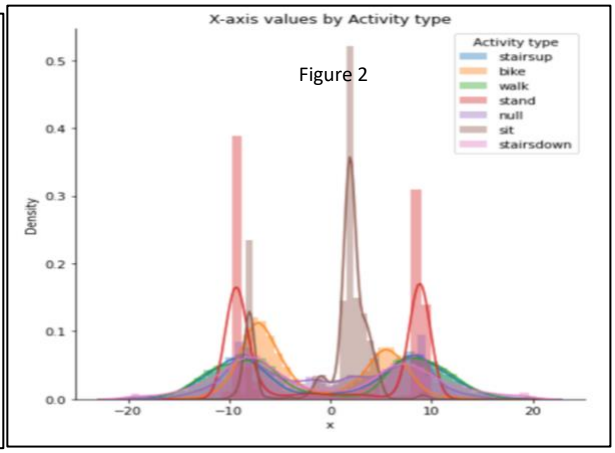



Figure 2

- In **Figure 3** We have noticed there are more moving activities than static (stand, sit and null), amongst users but only a few of them move a lot while the others are 'lazy'. This gives us a compressed group of people to target making our models application much clear.

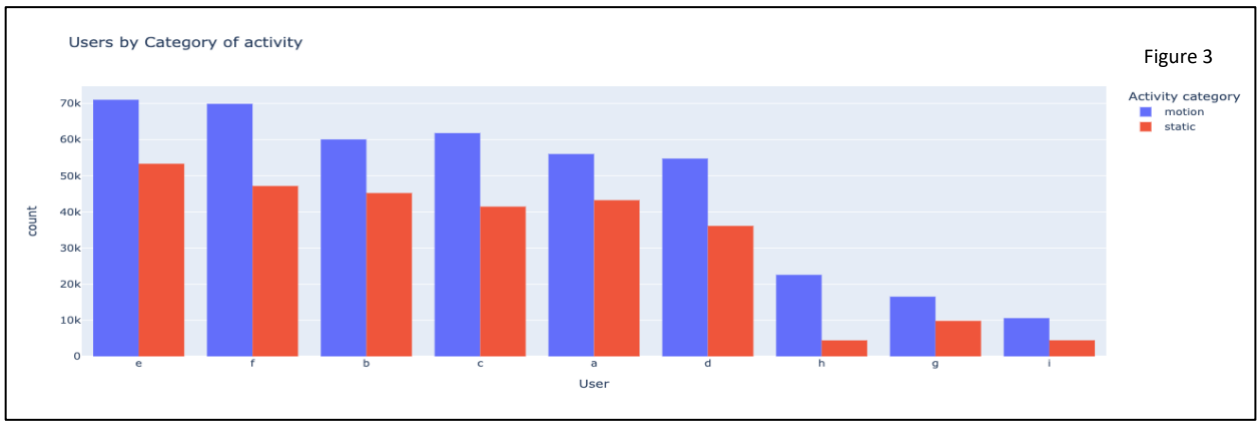



Figure 3

- In **Figure 4** for a given range of users an activity distribution plot will help us determine how much of the historical data has been recorded for each. For the user ‘H’, ‘G’, and ‘I’ we have very few records and could trigger biasing in prediction or interpretation of grouped results.

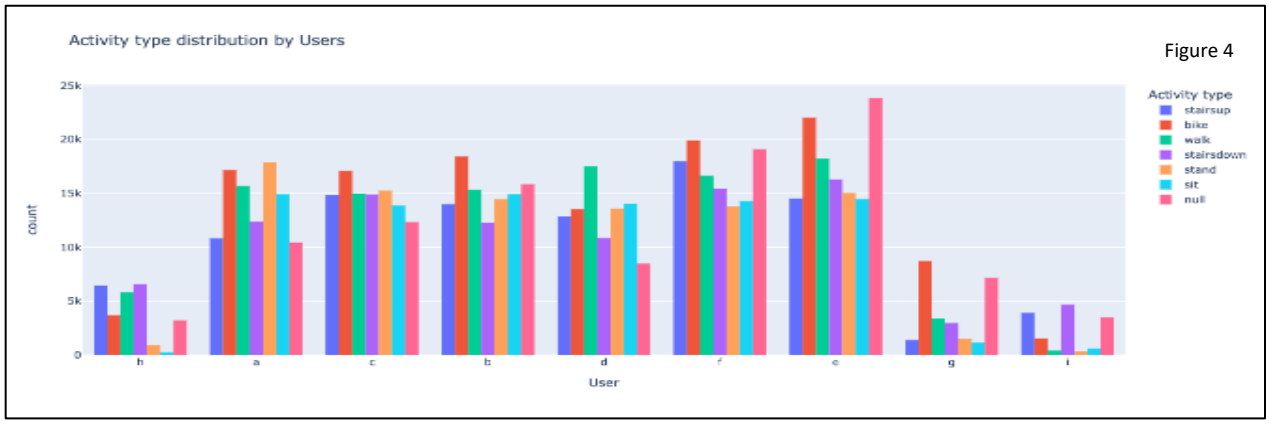



Figure 4

- In **Figure 5** we intend to show the balance between classes. We have 6 types of activities captured and these would be the target column for the model. The activities are Biking, walking, Stairs up, Stairs down, stand, sit and Null (no activity). We found that all activities were equally performed by the users in a day

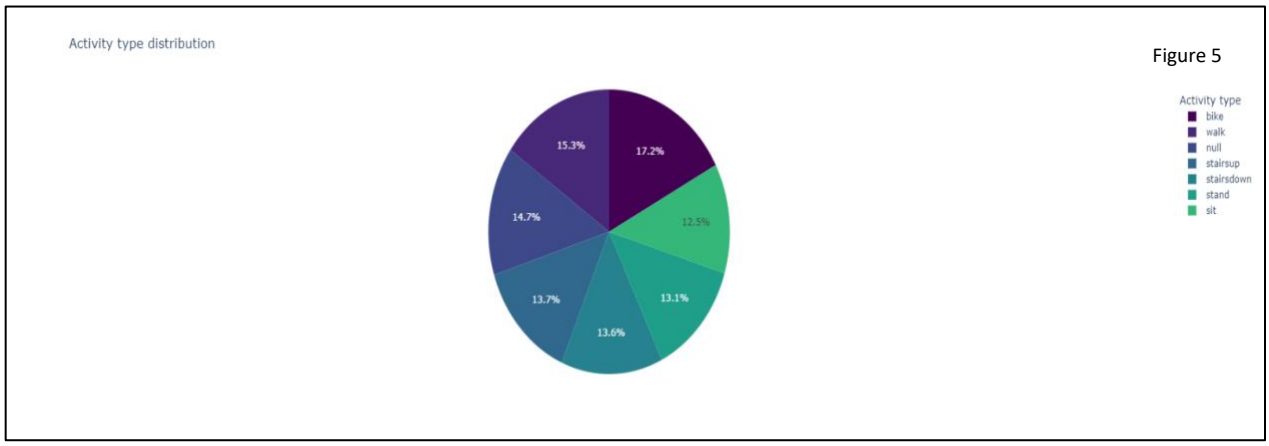



Figure 5

## 4. Data Cleaning

- Since there weren't any null values, we did not opt to drop any of the row data.
- There were some columns such as *'Arrival Time'* & *'Creation Time'* these were converted into timestamp format, so that a derived column can be created *'Time of day'*. This column consists of information whether an activity was performed in the morning, afternoon or evening based on the time column.
- Activities *'gt'* were labelled into ordinal numbers before hand which made it less complex for the pipeline to run.

## 5. Methodology

- *Data transformations:* Given the limited number of features in the dataset we decided to create a few additional categorical columns for deeper analysis and understanding of the data. The column *'gt_category'* was created based on User activity column (GT) categorizing the activities as *'Static'* or *'Motion'* activities.

  Similarly, column '*time_of_day'* was created by categorizing the 'Arrival time' column values as '*Morning'*, '*Afternoon'* and '*Evening'*.
- *Sampling*: Since we performed 3 use cases in this project, we had to sample our data w.r.t the case. The data towards model of watch used be the user was imbalanced and we had to sample it down so that the number of Samsung and Lg users are same (as shown in **figure 6** below).

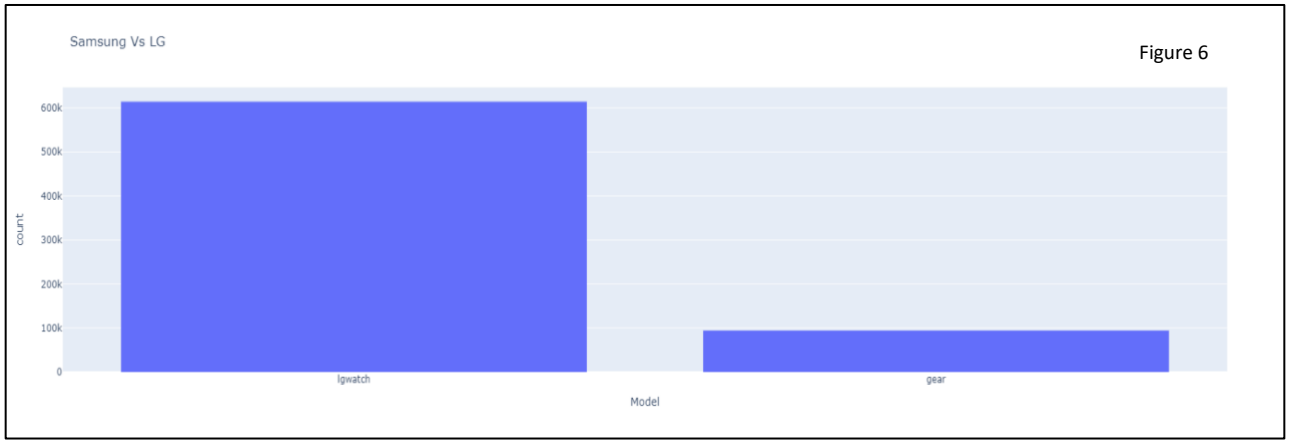



Figure 6

- *String Indexing*: this function was highly used for the machine to understand the columns. The string indexing helped us convert categorical columns such as time of day (Morning, Afternoon

and Evening) , sensor in watch (Accelerometer and Gyroscope), model of watch(Samsung Gear or LG) and whether the event was motion or static into ordinal numerical values.

- *Standardization*: this method helped in scaling down the jerk values into uniform values. The jerk column had varied range of values which was dependent on the amount of jerk applied to perform an activity or to stop an activity. Standardizing had bought the variation down hence keeping the values in uniform range.
- *Feature Selection*: Another important technique used. This helped us select the right columns for the use case. For example, including the time-of-day feature really did not help us when we tried to compare devices.
- *Modelling and scoring*: the final stage of any machine learning project. Teaching the machine, the huge knowledge base of tracked device data using the random forest and logistic regression algorithms. To compare the best fit algorithm for a given case we intend to analyze the model performance based on precision and accuracy metrics. To get the best metrics, we have organized a set of parameters for grid search that help us fine tune the models.

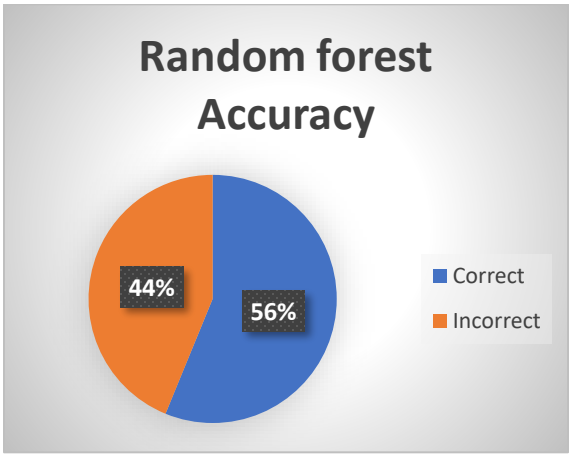
# 6. Model-Predictions

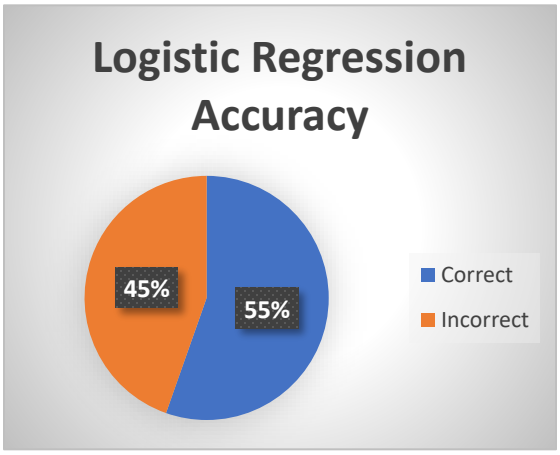
## 6.1 Case 1- Activity Prediction

- The motive of this use case is to create a generic model that could recognize and classify the type of user activity based on positions detected by accelerometer and gyroscope sensor which give us the tilt (x), direction (y) , and plane (z) of the watch to further optimize device performance and promote physical fitness.
- To achieve this goal, we have started off with creating two generic models the random forest and logistic regression model. The explanatory columns along with their transformations are listed in the table below.

| Explanatory Variable | Transformation | Before | After |
|---|---|---|---|
| X, Y, Z | None | - | - |
| Device | String Indexer | Samsung, LG | 0,1 |
| Sensor | String Indexer | Acc, Gyr | 0,1 |
| Jerk Mean | Standardization | 0.355368205 | 0.760172154 |

- For this use case we found both models to perform the same, but the random forest had better outcomes. The error rate is slightly higher in logistic regression, also the random forest upon error analysis showed much better class prediction for imbalanced classes.

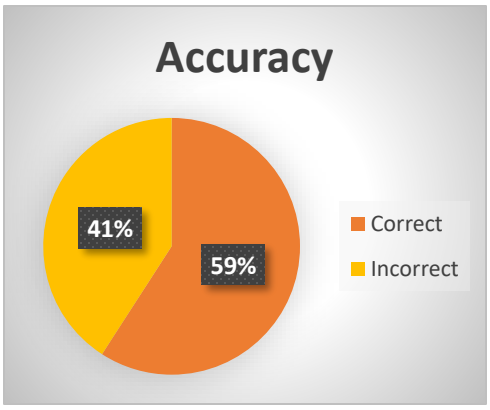


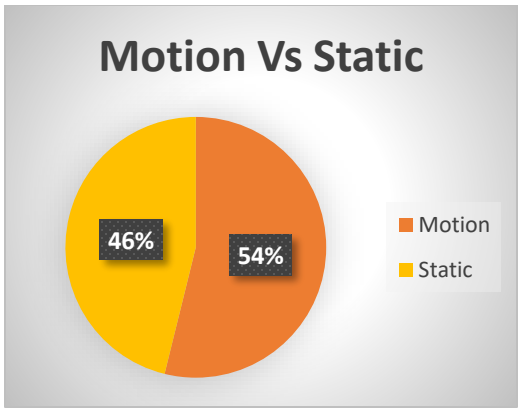




- These results helped us in creating a baseline model as the random forest model that would be used for comparing with the model performances for the other 2 use cases and would give a worst-case accuracy of 56%.

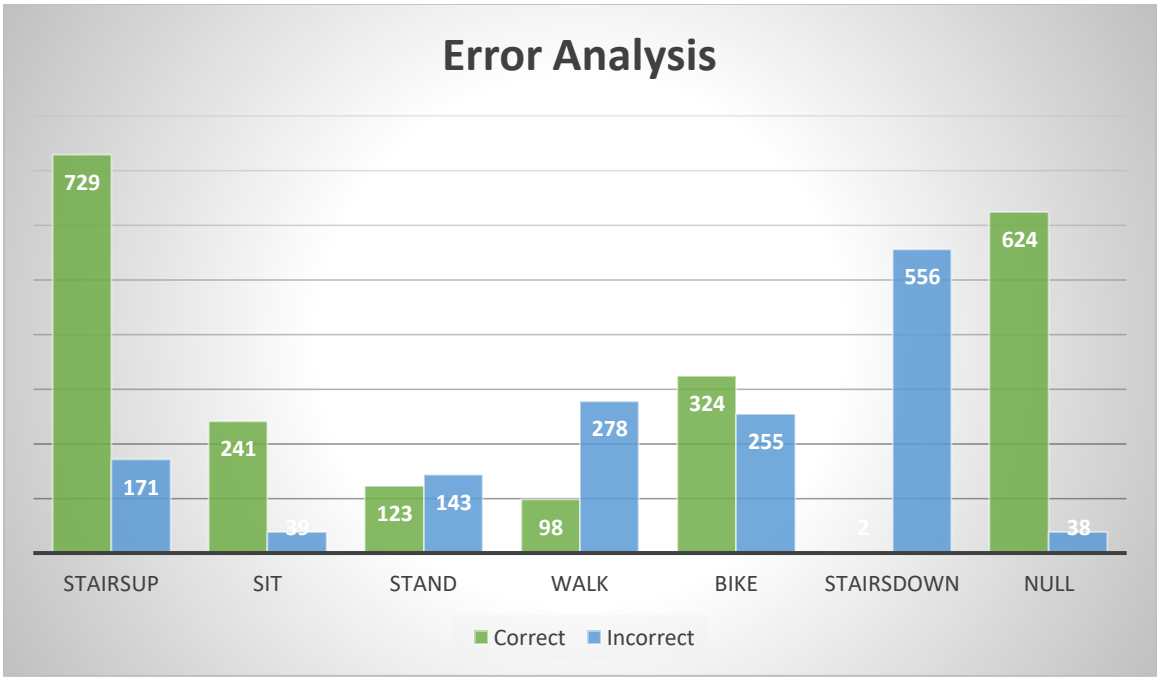
## 6.2 Case2 - User based activity tracking

- The motive of this use case to create a model that recognize an activity performed by a user at a given time of the day. This would help us track the user activity and suggest them to do something to stay fit, example, suggesting the user to go for a walk when recognized that the user is sitting idle for long time.
- To achieve this goal, we have subset the data for User 'F' and time of day as 'Afternoon'.
- Similar transformations were performed as above. But the results generated were quite different. Our model was 60% sure of the outcomes, hence gaining good accuracy. The user performed 5% more motion-based activities rather than static.



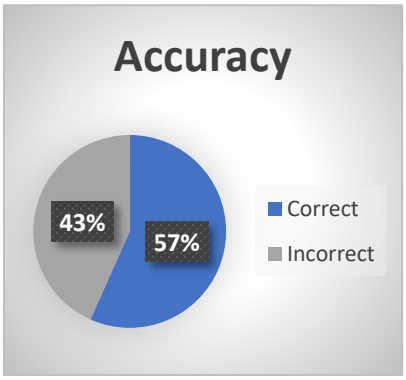

- The error analysis made us understand what predictions where rightly achieved. The model was good at tracking the user activity such as stairs up, sitting, biking, no activity etc. but was bad at predicting activities such as standing, walking, stairs down.

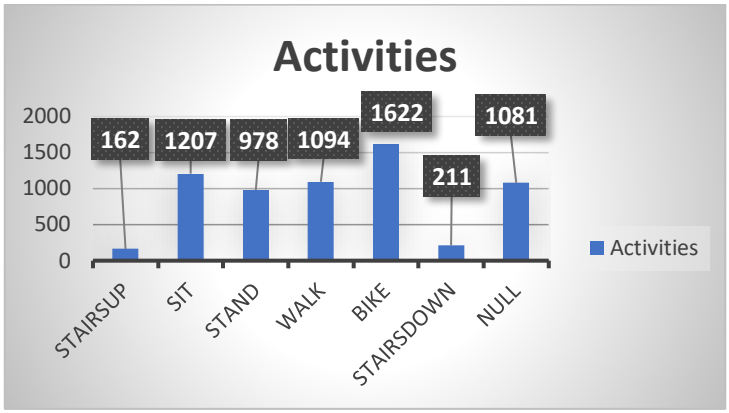



## 6.3 Case 3 - Model Comparison

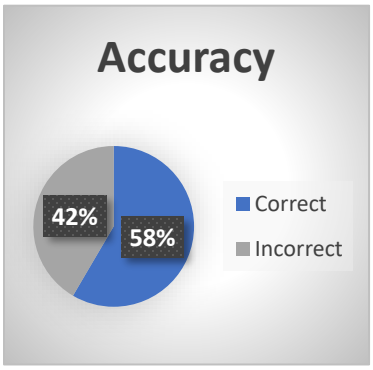
- The motive of this use case to create predictive modelling for Samsung and LG devices to propose better business results and increase sales of the companies.
- To achieve this goal, we have subset the data based on the device type. Now we have two datasets for modelling and finally would evaluate them based on their accuracy.
- We have considered the same transformations as above for each dataset.

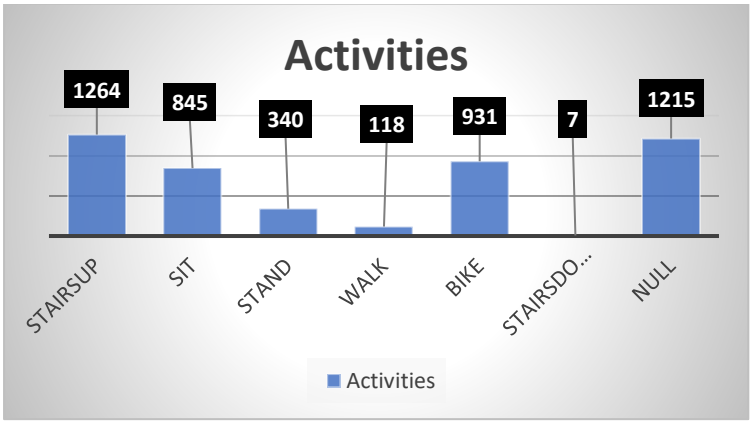
- Given below are the results for LG watch.




- The above charts show us that LG watch is good at predicting all activities of the user, except when the user goes upstairs or downstairs. This model is poor at predicting movements along the Y axis plane, but for the other activities our random forest model is 57% sure that it is being performed.
- Given below are the results for Samsung watch.




- The above charts show us that Samsung watch is good at predicting only a few activities of the user but considers most of the activities as not being performed. This random forest model has 58% accuracy which is approximately same as LG watch
- But due to better predictability of all activities LG watch is considered the better.

## 7. Conclusion

- Random forest model and logistic regression model are both good at prediction for multiclass classification.

| Model | Accuracy |
|---|---|
| Random Forest | 56% |
| Logistic Regression | 55% |

- User activity tracking was achieved using random forest model. This helped us track and inform user about their day-to-day activities and promote fitness. It was healthy to see user had good movements every day.

| Model | Accuracy |
|---|---|
| Activity Prediction | 59% |
| Motion Vs Static | 54% |

- Finally, the last model of device comparison. We derive to the fact that LG watches are good products. Samsung must work on their activity by capturing much accurate data or from more sensors.

| Model | Accuracy |
|---|---|
| LG | 57% |
| Samsung | 58% |

- With the help of the above modelling techniques our goal was to generate an analysis that serves the following purposes:

1. Using big data analytics to monitor physical activities of users and promote better fitness.
2. User activity classification based on available features as well as analyzing specific user activity at different time periods in a day.
3. Fine tuning classification model accuracy to help business leaders to develop more accurate wearable devices.

# 8. Appendix

## 8.1 Reference


- Pyspark classification:https://spark.apache.org/docs/latest/ml-classification-regression.html#classification
- Smartwatch: https://www.cashify.in/explained-sensors-in-smartwatch
- Sensors in Smartwatch: https://illumin.usc.edu/fitness-trackers-how-they-work-and-their-highly-anticipated-future/


## 8.2 Project Files

- PPT link - https://www.canva.com/design/DAFT0aUb_Z4/4-tQbBukdswJqTzH46EgtQ/view?utm_content=DAFT0aUb_Z4&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton
- Code link - https://colab.research.google.com/drive/1CQ-WljaUvY4PNa6L3ylzoQES5Zryj1E7?usp=sharing