Wearable Sensor Marketing Analysis

PREDICT 498, Section 57, WI2016

Northwestern University

Team Lead: R. Scott Murray - Communication/Research/Document Prep/Sr. Analyst/R programmer

Team Member: Colin Loveness - Sr. Analyst/Document Prep/Research/R programmer

Team Member: Terry Miller - Sr. Analyst/Research/Document Prep/SAS & R programmer

Team Member: Drew Staley - Sr. Analyst/Document Prep/Research/R programmer

Supervisor: Nethra Sambamoorthi, Ph.D.

# 

**Table of Contents**

[**Team Roles & Responsibilities** 1](#_Toc445649506)

[**Abstract** 1](#_Toc445649507)

[**Introduction** 2](#_Toc445649508)

[**Profile of the Organization & Opportunity** 2](#_Toc445649509)

[**Review of Literature** 4](#_Toc445649510)

[**Research Questions** 6](#_Toc445649511)

[**Data & Measurements** 7](#_Toc445649512)

[**Structural Hypotheses** 9](#_Toc445649513)

[**Methodology** 9](#_Toc445649514)

[**Variable Reduction Efforts and Feature Engineering** 10](#_Toc445649515)

[**Linear Regression Evaluation** 12](#_Toc445649516)

[**Linear Regression Fit Metrics** 15](#_Toc445649517)

[**Linear Regression Implementation Plan** 20](#_Toc445649518)

[**Logistic Regression Evaluation** 21](#_Toc445649519)

[**Logistic Modeling** 24](#_Toc445649520)

[**Model Evaluation** 29](#_Toc445649521)

[**Logistic Regression Implementation Plan** 34](#_Toc445649522)

[**Random Forest Evaluation** 35](#_Toc445649523)

[**Dashboard** 40](#_Toc445649524)

[**Conclusion** 41](#_Toc445649525)

[**Implementation** 42](#_Toc445649526)

[**Appendix A** 43](#_Toc445649527)

[**Solution Development & Implementation Report** 43](#_Toc445649528)

[**Data Quality Auditing** 43](#_Toc445649529)

[**Variable Summary Table** 46](#_Toc445649530)

[**Measurements** 46](#_Toc445649531)

[**Computational Methods** 47](#_Toc445649532)

[**Expected Methodology and Visuals Explaining Concepts** 47](#_Toc445649533)

[**Tools, Techniques, & Transformations** 48](#_Toc445649534)

[**Recommended Tools and Techniques of Data Retrieval** 49](#_Toc445649535)

[**Recommended Tools and Techniques of Data Integration** 49](#_Toc445649536)

[**Recommended Tools and Techniques of Data Transformations** 50](#_Toc445649537)

[**Recommended Tools and Techniques of Test/Validation Set-Up Procedures** 51](#_Toc445649538)

[**Project Plan** 52](#_Toc445649539)

[**Project Deliverables** 52](#_Toc445649540)

[**Project Requirements** 53](#_Toc445649541)

[**Work Breakdown Structure** 53](#_Toc445649542)

[**Project Constraints/Priority Matrix** 56](#_Toc445649543)

[**Project Assumptions** 56](#_Toc445649544)

[**Project Risks** 56](#_Toc445649545)

[**Stakeholder Register** 57](#_Toc445649546)

[**Communication Matrix** 57](#_Toc445649547)

[**Milestones Schedule** 58](#_Toc445649548)

[**Summary Budget** 58](#_Toc445649549)

[**AON** 59](#_Toc445649550)

[**Gantt Chart** 61](#_Toc445649551)

[**Network Information Table** 64](#_Toc445649552)

[**Schedule Baselines** 65](#_Toc445649553)

[**Cost** 68](#_Toc445649554)

[**Resource Loading & Leveling** 69](#_Toc445649555)

[**Project Approval Guidelines** 73](#_Toc445649556)

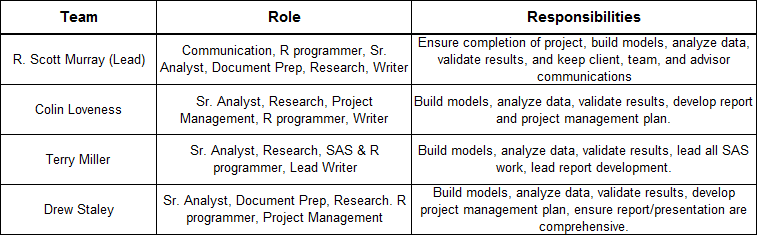
[**Authorizations** 73](#_Toc445649557)

[**Data Fields & Data Descriptions** 73](#_Toc445649558)

[**Appendix B** 77](#_Toc445649559)

[**R Code** 77](#_Toc445649560)

# **Team Roles & Responsibilities**



# **Abstract**

The purpose of this research is to identify application possibilities of 3-dimensional data in the prediction of heart rate. An expansion on prior research using the same data, it will review linear modeling, clustering, logistic regression, and random forest techniques in an effort glean insights from acceleration, gyroscopic, and magnetometer data associated with different activities and the subjects’ heart rates. The anticipated outcome of the analysis will be used to develop best practices for Trivisio to market the device to their at-risk patients, incorporating meaningful readings and appropriate device location the patient’s body. The finding could be useful in preventative care for high or medium-risk cardiac patients but will still require further age group testing.

*Keywords: Physical Activity Monitoring, Analytics, R, Heart Rate Multivariate linear regression, Logistic regression, Cluster, Random forest, Prediction, Acceleration, Gyroscopic, Magnetometer, Inertial Measurement Units, American Medical Association.*

# 

# **Introduction**

Sterling Analytics has been contracted by Trivisio to perform an evaluation of a new-to-market device in the wearable technology field designed to monitor heart rate data (bpm) related to 3-dimensional movement. Recently, nine individuals agreed to be test subjects in an analysis project to capture previously unused characteristics associated with the body’s movement in 3-dimensional space. The device is an IMU (Inertial Measurement Units) sensor which can be worn on three body locations including a wrist, ankle, or chest monitor.

Utilizing a traditional project management approach, a project plan will be lead the team through planning, design, execution, monitoring, controlling, completion, and closing of the project. The project will provide insights into daily activities of the test subjects for a better understanding of how different levels of activity intensity impact both the movement and spatial data as well as the heart rate of the individuals. Armed with this knowledge, Sterling Analytics will analyze how acceleration, gyroscopic, magnetometer, and temperature data have the ability to impact the heart rate of otherwise healthy individuals. Due to the current pressures of the wearable technology industry and the interest of individuals in monitoring their own data with the rise of other devices, along with the potential for meaningful influence on the medical industry potentially in patients who have recently suffered some type of cardiac event, performing an insightful analysis of the data in a brief period of time is of the utmost importance.

# **Profile of the Organization & Opportunity**

Trivisio exists as an engineering and development company, with headquarters in Trier, Germany. The firm, historically, demonstrates expertise in the design and manufacture of electro-optical devices, focusing on head-mounted optical controls used in a myriad of industries. In fact, Trivisio maintains primary European market share for these products, with its fully integrated production center in Germany.

The company’s core product mix, shown [here,](http://www.trivisio.com/products/hmd-nte/) consists of head-mounted video and stereoscopic optical gear used for virtual gaming, surgical guidance and military/tracking applications, among many others. Many firms contract Trivisio for development and production of its devices to sell, in conjunction with proprietary electronics, as comprehensive, private-labeled systems. Such packages could commonly be found in popular brands such as Nintendo, Sony Playstation or Medtronic medical.

But the firm has long sought to diversify its product offering, as the margins for electro-optical devices face consistent pressure from low-cost global suppliers. Existing as a small, privately-held, boutique design and production firm mandates the company possess strategic agility with respect to product migration strategies.

Recently, the company received a request from the American Medical Association to design a sensor, or sensors, which effectively explain heart rate (in beats per minute-BPM) in relation to 3-dimensional movement common in a variety of sports and functional activities. Traditional heart-rate monitors offer limited, one-dimensional assessments of a subject’s heart function. But the AMA seeks to better understand the relationship between the body’s movement and heart rate. With the emergence of telehealth and wearable sensors integral to medical diagnoses, the AMA is poised to issue guidance for the use of these types of sensors for patients exhibiting cardiac risk.

The synergies of Trivisio’s quest to diversify and the request of the AMA led to the development of a line of sensor products, called IMUs. These units, worn on distal extremities and the chest, measure several key indicators specific to the body’s movement. Ambient temperature, 3D acceleration, gyroscopic data, and magnetometer readings are measured and recorded by the IMU sensors. This provides unique body movement and temperature data, potentially useful for predicting heart rate.

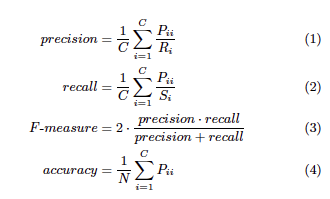
Trivisio has contracted our firm, Sterling Analytics, to launch an analytics project seeking to quantify its sensor data and the relationship to heart rate. As its core competency is product engineering and manufacturing, the company lacks the expertise to perform the statistical analyses necessary to fully execute the project. With each sensor producing a wealth of usable data, our task will be to develop statistical models using the data to predict the heart rate of subjects participating in a number of athletic and functional activities. The results of our project will determine whether, or not, Trivisio positions its sensor line in the medical-device market, useful for predicting cardiac function in relation to body movement data, along with assessing the viability of potential targeted marketing opportunities.

# **Review of Literature**

This dataset has been previously analyzed in a research paper titled “Creating and Benchmarking a New Dataset for Physical Activity Monitoring.”[[1]](#footnote-1) Attila Reiss and Didier Stricker use five different classifiers to benchmark the data set and demonstrate the difficulty of the classification tasks rooted in the high number of activities and personalization. The goal of physical activity monitoring in this paper is to estimate the intensity and to recognize activities like sitting, walking, running or cycling. The authors break their paper into the following sections: introduction, data collection, benchmark basic conditions, data processing, results and discussion and conclusion and future work.

The data collection section details the creation of the IMU dataset. It describes hardware setup, participants, data collection protocol, activity selection justification and lessons learned. The researchers found that attaching the sensors to the subjects and recording data were straightforward tasks, and all subjects reported that the sensors were comfortable and did not restrict normal movements at all. They found that recording activities outdoors took careful planning around favorable weather forecasts. Data integrity was compromised for two reasons, the first being data loss during wireless transfer, and the second being the fragile system setup, which required 2 USB-dongles, a USB-hub and a USB extension cable stored in a “companion bag,” which provided wireless connectivity for the IMUs and heart rate monitors. The first compromise doesn’t have a large impact on data quality as the frequency of observations were so high (up to 100 recordings per second from IMU sensors). The second compromise was more impactful, especially for high exertion activities like running and rope jumping. The researchers recommend consolidating the hardware to one manufacturer to simplify the setup (Trivisio’s setup requires an external USB-dongle to transmit data wirelessly) or consider local storage on the device during activity monitoring to be downloaded after the subject’s activities conclude.

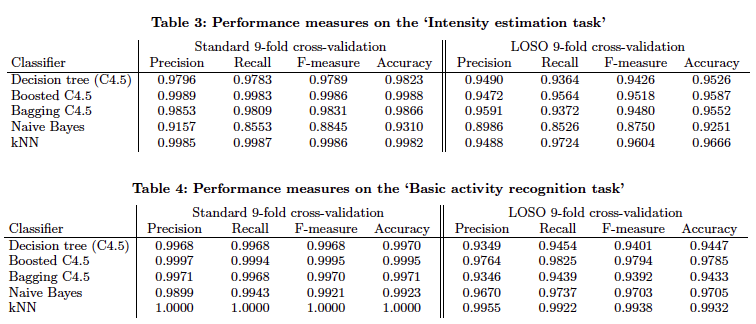
This paper benchmarks the data by defining three different classification problems: intensity estimation task (light, moderate and vigorous effort), basic activity recognition task (lying, sitting/standing, walking, running and cycling) and background activity recognition task (lying, sitting/standing, walking, running, cycling, ironing, vacuum cleaning, ascending and descending stairs, Nordic walking and jump roping). The performance measures for this project include precision, recall, F-measure and accuracy, as defined by the following four equations:



Where P*ij* is the confusion matrix of *i* and *j*, where *i* refers to the rows (annotated activities), and *j* refers to the columns (recognized activities) of the matrix[[2]](#footnote-2).

The data processing phase of this research includes data preprocessing, segmentation, feature extraction and classification. Preprocessing includes aggregating the data, using linear interpolation for missing data from wireless connectivity issues, and removing the first and last 10 seconds of each labeled activity. Segmentation resulted in a window size of 512 samples, which translates into a sliding window of length 512 seconds, shifted by one second for consecutive windows. Feature extraction of the segmented data includes signal features in time and frequency of the 3D IMU data, mean, median, standard deviation, peak acceleration and energy. Correlation between pairs of axes was useful for differentiating between activities in multiple dimensions (e.g. ascending stairs vs. simply walking on flat ground), and power ratio of the frequency bands was useful for differentiating between locomotion activities (walking or running) and cycling. The researchers extracted a total of 133 features from IMU data and 4 features from heart rate data from each data window of 512 samples. There were also no feature reductions considered, so all 137 features were included in the classification analysis. Finally, these features were applied to five different classification algorithms in the Weka toolkit: Decision Tree, Boosted Decision Tree, Bagging Decision Tree, Naive Bayes and k Nearest Neighbors.

The results and discussion of this analysis revealed that the kNN and boosted decision tree classifiers achieved the best performance. Naive Bayes performed better when there were clear class boundaries (basic activity recognition task), but the decision tree classifiers had superior performance when distinguishing between multiple activities, as in the intensity estimation and background activity recognition tasks. The best classifiers achieved 96% accuracy in intensity estimation, and any misclassifications were to neighbors and not activities in a different intensity band. The researchers found that the classifiers had difficulty distinguishing between sitting and standing (posture differences) and ironing and standing (gesticulation of subject arms during standing made distinguishing from ironing difficult). Overall, all four tasks achieved over 90% accuracy in classification, a comparable benchmark to similar classification problems of this nature. A summary of the model performance measures for each classifier in the study is represented below[[3]](#footnote-3):



# **Research Questions**

Our project will employ all the best practices of predictive modeling to fully investigate the relationship between Trivisio’s collected IMU data and the subject’s heart rate. Of particular importance will be the question of whether, or not, heart rate can be predicted or explained by acceleration, gyroscopic, temperature, or magnetometer data.

As stated previously, Trivisio intends to market the sensors as medical devices specific to at-risk cardiac patients. As FDA 510(k) compliance requires significant financial and structural commitment, the company seeks justification for the potential strategic initiative.

We will provide Trivisio with statistical justification for either proceeding with the initiative, or declining to pursue it. At minimum, multiple linear regression models will be developed and the error rate of the predetermined validation sample(s) will be evaluated. Depending on the accuracy of the model(s) developed, and subsequent performance in test-set validation, we will submit our recommendations based on these findings.

# **Data & Measurements**

Trivisio has provided IMU data from nine subjects who were monitored while performing eighteen different physical activities. The data are broken down into nine .dat files, each pertaining to a subject in the study, and include time (duration of activity monitored), heart rate, and activity ID. Each subject wears three Inertial Measurement Units which record three axis acceleration, gyroscopic and magnetometer data of the dominant wrist, ankle on the dominant side of the subject's body and chest. These data can be used for a variety of analytics applications, including activity identification and estimation of activity intensity.

The three IMUs record temperature, acceleration in three directions measured in meters per second (ms-2), gyroscopic data along three axes measured in radians per second (rad/s) and magnetometer data along three axes measured in micro Teslas (μT). A table defining each of the 33 variables in the dataset is attached to Appendix 1 of this report in Figure 1.

The nine subjects in the study have the following demographic information:

**Subject Attributes Table**

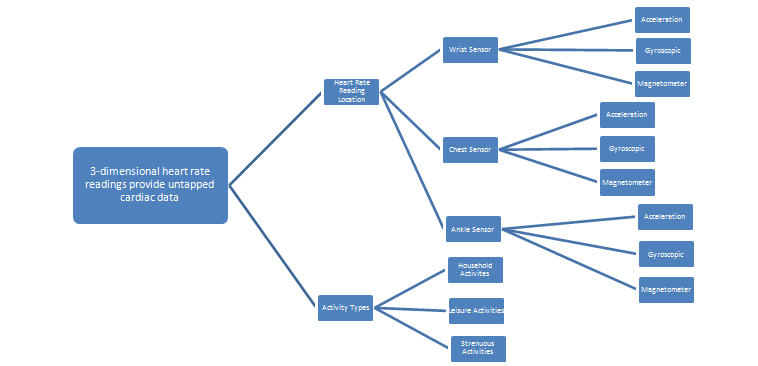
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Subject ID** | **Sex** | **Age (years)** | **Height (cm)** | **Weight (kg)** | **Resting HR (bpm)** | **Max HR (bpm)** | **Dominant hand** |
| 101 | Male | 27 | 182 | 83 | 75 | 193 | right |
| 102 | Female | 25 | 169 | 78 | 74 | 195 | right |
| 103 | Male | 31 | 187 | 92 | 68 | 189 | right |
| 104 | Male | 24 | 194 | 95 | 58 | 196 | right |
| 105 | Male | 26 | 180 | 73 | 70 | 194 | right |
| 106 | Male | 26 | 183 | 69 | 60 | 194 | right |
| 107 | Male | 23 | 173 | 86 | 60 | 197 | right |
| 108 | Male | 32 | 179 | 87 | 66 | 188 | left |
| 109 | Male | 31 | 168 | 65 | 54 | 189 | right |

The sample size and distribution of each subject’s demographics is limited, so results based on subject traits (e.g. age, sex, handedness) can’t necessarily be extrapolated to the wider public. Instead, we will use the aggregated IMU measurements to predict HR and activity, and then determine which IMU (wrist, ankle or chest) contributes the most “predictive power” for each activity to guide our client’s product development plans and marketing campaigns.

The eighteen activities each subject performs vary in aerobic effort and type. Some examples include household chores such as vacuuming and ironing, performing simple tasks around the house such as climbing stairs, sitting, lying, standing and walking, and several athletic activities, such as cycling, running, jump roping and playing soccer. The diversity of activities allows for extrapolation to a wide variety of potential markets for our client to consider, and overlap with demographics our client is segmenting for targeted advertising.

We have already acquired the data from our client, which can also be downloaded from the University of California Irvine’s Machine Learning Repository.[[4]](#footnote-4)

# **Structural Hypotheses**



# **Methodology**

Our approach to understanding the sensor data is two-fold; we are interested in predicting heart rate (bpm) from the sensor reading data, as well as accurately classifying user activity based on sensor readings. Each of these challenges requires a unique approach. Aside from the activity classification values, the entirety of the dataset consists of real numbers, which will have an impact on the methodologies implemented for modeling. The bulk of the work for this project will be conducted in the R programming environment.

Prior to the development of data models a thorough exploratory data analysis [EDA] will be conducted to assess the quality and basic summary statistics of the data. EDA is an essential process that allows the researchers and modelers to become more familiar with the nature, structure, and composition of the data. This will take an open approach and will likely be iterative in nature based upon initial findings. The EDA will likely consist of variable distribution assessment, missing value removal/imputation, correlation and collinearity assessment, and/or naïve model development for exploratory purposes, among other processes.

Additionally, for further evaluation and/or modeling, the data set will be randomly partitioned into training and test portions, comprised of a 70/30 ratio split. Models will be trained on the training (in-sample) data and evaluated on the test (out-of-sample) data to ensure model robustness and avoid potential overfitting of the models.

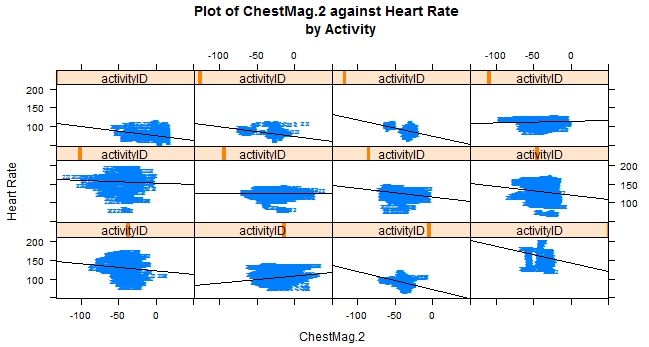
## **Variable Reduction Efforts and Feature Engineering**

Due to the two-fold approach to this analysis, variable assessment efforts will be directed in two ways to reflect the nature of regression modeling and classification efforts. With regard to the regression analysis, the distribution of the numeric predictor variables will be reviewed for potential skewing issues, such as outlier presence, among other details. Naive models were developed using forward, backward, and stepwise selection methods, using a randomized 70% training sample from the original data set, to assess variable importance and statistical significance for regression.

Forward variable selection identified HandGyro.1, ChestAccel16.2, AnkleGyro.1, AnkleGyro.2, and AnkleGyro.3 as statistically insignificant in the model (with P-values exceeding 0.05). Backward and stepwise selection found only AnkleGyro.1 to be insignificant.

With a review of these selection methods, it is apparent that most of the data within this dataset will perhaps serve our modeling efforts well. Further exploration of these matters, with respect to linear regression, will involve variable interaction effects and other effects.

An alternate method of review of high-quality predictor variables was conducted using the Lattice package in R. The Lattice package allows the user to split analysis by some defined criteria and look for variance across the levels of that variable. In our case, we modeled a simple linear regression, using Heart Rate as the response, with one predictor variable in each case, with respect to Activity. The following graphic is an example output:



These outputs are incredibly useful in allowing the analyst to quickly observe any differences across the regression that are influenced by the activity. In short, it provides a quick sense of whether the variability in activity is explained well by the predictor variable. Upon an initial review of these plots for all predictor variables, predictors of interest include HandTemp, HandAccel16.1, HandAccel16.2, HandAccel16.3, ChestAccel16.1, ChestAccel16.2, ChestAccel16.3, ChestMag.2, AnkleAccel16.1, AnkleAccel16.2, AnkleAccel16.3, and AnkleMag.1.

With respect to the classification efforts, multiple decision trees were built using the same randomized training sample used to conduct the linear regression variable assessment. A naive decision tree was constructed from the same 70% training set. The Time variable was the most important in classification splits in the naive tree, along with HandMag.1, ChestGyro.2, and HandAccel16.3. Constructing a random forest model to assess the variable importance factor (VIF) of the predictors yielded the following results (top 10 classification predictors, ranked in terms of decreasing the impurity (Gini Index):

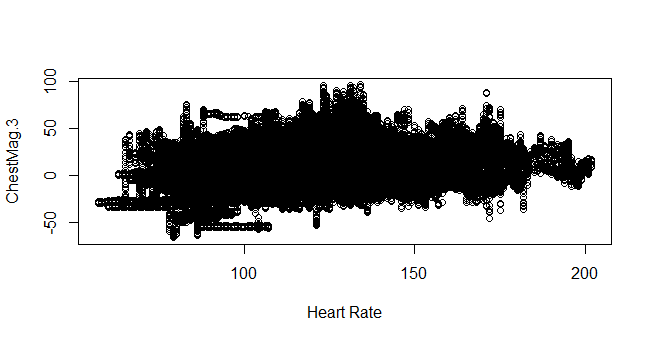
Time, HR, ChestAccel16.3, ChestTemp, HandTemp, ChestMag.2, AnkleTemp, HandAccel16.1, ChestMag.3, and HandAccel16.3.

Evidenced in this is that the temperature readings from each sensor are included and that temperature response in each region is likely correlated to an activity or set of activities. Surprisingly, the readings from the chest sensor dominate the top 10 over the hand and ankle readings. These measurements will be further explored with subsequent classification model development.

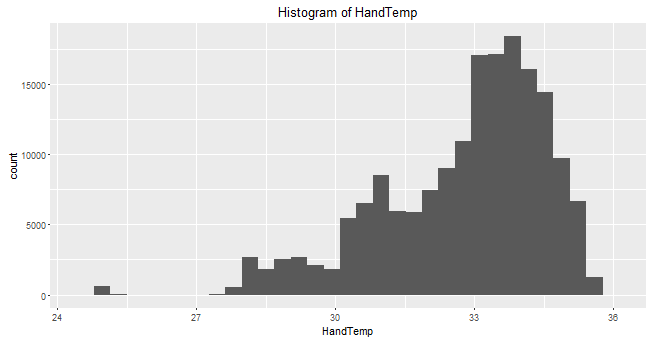
# **Linear Regression Evaluation**

To begin evaluating a multivariate linear regression model, with Heart Rate as the response variable, one must ensure a few key assumptions are satisfied. First, the relationship between response and predictor variables needs to linear. This relationship can be evaluated by examining a scatterplot, or pairs panel, between variables as shown below.

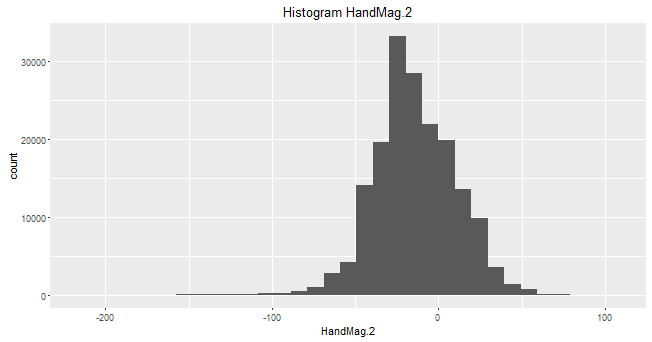
As shown, the relationship between Heart Rate and the Chest Temp sensor appears to be linear, with the suggestion of some non-linearity in the data. In fact, all of the sensors (Ankle, Chest, and Hand) exhibit the same general relationship with HR.



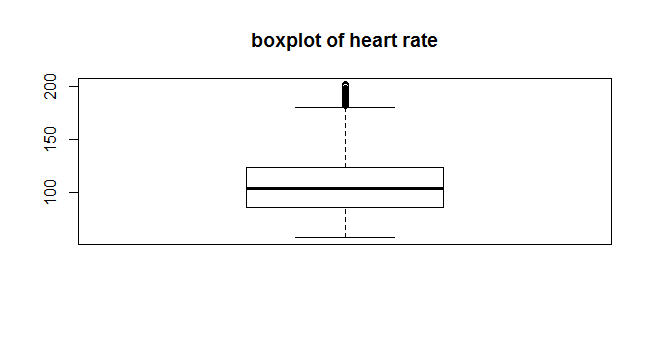
Second, one needs to evaluate the distribution of the variables to ensure each is from a multivariate normal distribution. This can best be shown in the histogram below. As can be shown, a number of the variables exhibit left, or right skew. All three temperature sensors (Hand, Ankle and Chest), Time and Heart Rate exhibit skewed tails and slight non-linearity.



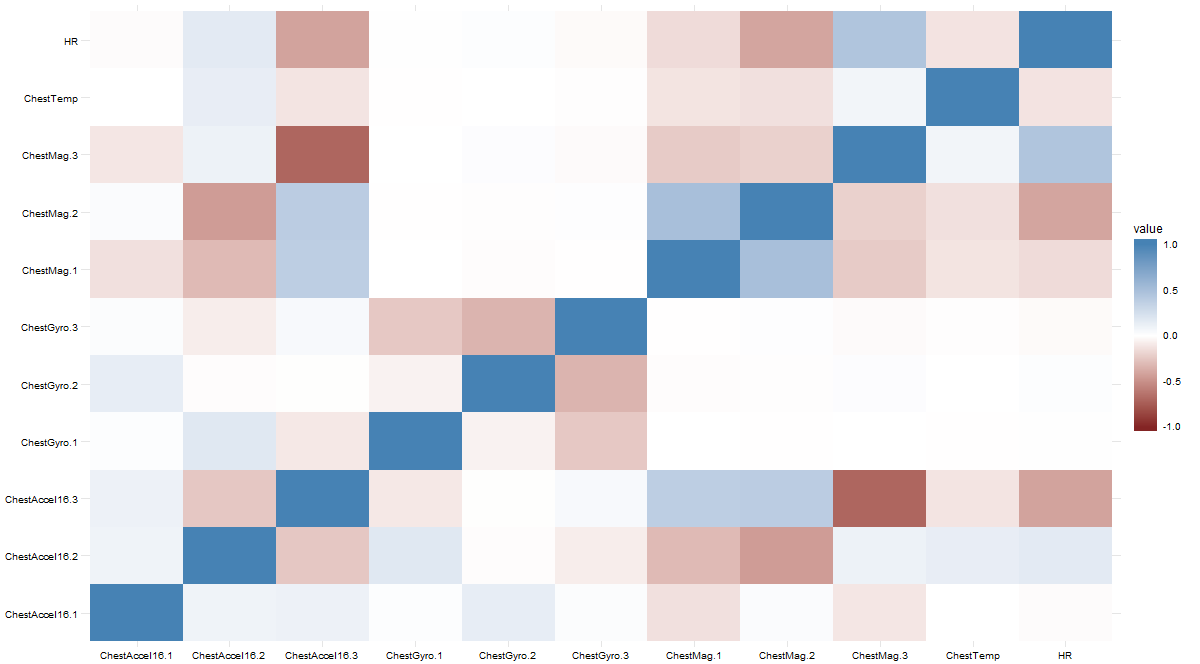
The remaining variables, not including the dummy-coded activity IDs, do exhibit normal distribution, as shown below. All of the magnetometer, accelerometer and gyroscopic readings appear to be normally distributed. This means the entire dataset does not need to be transformed, but, rather, only those variables exhibiting non-linearity need to be accounted for.



One also needs to evaluate each for variable for potential outliers, or high leverage points. Basic summary statistics of each variable are valuable EDA tools to spot potentially impactful data points. In addition, boxplots can be useful for evaluating outliers, shown below. In our dataset, the variable Heart Rate, among others does appear to be impacted by outliers and high leverage points. This will be addressed in the discussion of linear model fit metrics.



Pearson Correlation Coefficients have been generated and will be evaluated with Heart Rate as response variable, and all other variables. An example of this output can be seen in the below Heat Map of the Chest Sensors and HR. A moderately strong, positive linear relationship can be seen between HR and the Chest Magnetometer sensor. Likewise, a moderately strong, negative linear relationship can be seen between HR and the Chest Accelerometer 16.3 sensor. Generally, the temperature sensors and magnetometer sensors exhibit the highest correlation to heart rate.

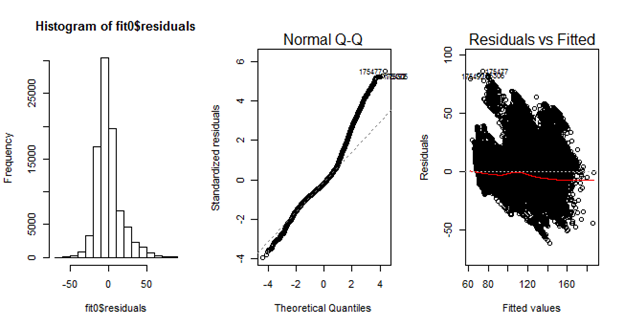


With the key assumptions of linear regression addressed, several OLS and PLS models were fitted in an attempt to determine which best predicts Heart Rate. As will be shown in the ensuing discussion, the Partial Least Squares Model, due to interpretability and parsimony, would be the model selected to best predict Heart Rate. It will be shown, however, that none is entirely effective at accounting for most of the variance of heart rate.

## **Linear Regression Fit Metrics**

First, a model was fit with all variables, except activity ID, as predictors. This resulted in an adjusted R-Square value of .6824, indicating that the model accounted for roughly 68% of the model’s variance. In addition, the performance of the model on the test data resulted in a Mean Square Error of 360.94.

To assess the fit of the model to the data, several plots were generated. First, as can be seen below, the residuals do appear to be normally distributed; a key criteria assessing model fit.



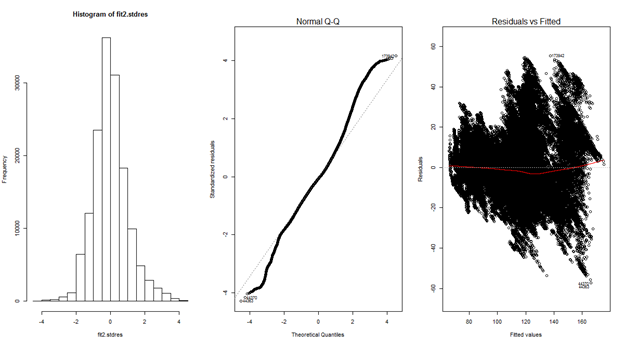
But, evaluating the QQ plot of residuals and plot of residuals versus fitted values highlights one of the problems uncovered in the exploratory data analysis. There are several highly influential observations, and outliers, that impact the overall fit of the model. This can be seen, additionally, in the Cook’s D and residual plots below. But, without having been privy to the experiment design and implementation, the justification to remove particular observations cannot be made. As a result, all points have been left in the dataset.



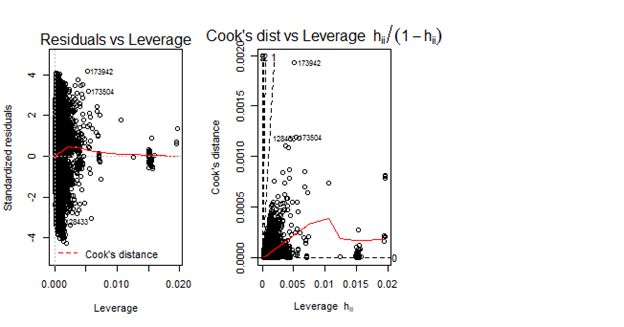
But, in order to potentially mitigate some of the non-linearity of certain variables, a second OLS regression model was fit. The variables Time, ChestTemp, HandTemp, and AnkleTemp were included as 7th order polynomials to capture the non-linear relationship with heart rate. Additionally, interaction terms were added between chest/hand/ankle temperature sensors and the magnetometers. These were selected as both were highly impactful in the original model and held high PCC values.

This approach, in conjunction with a stepwise variable selection (p<.01) led to a model with an adjusted R-Square value of .7526, certainly an improvement. In addition, the cross-validation test set performance improved the MSE to 180.40. Again, the inclusion of polynomial and interaction terms improved the model’s performance. But, interpretability remains suspect with nearly 30 variables remaining in the model. Too, the model’s fit metrics appear similar to those in the first model discussed.

As seen below, residuals appear normally distributed. And although the tails of fitted values on the QQplot have been tightened up, one can still observe the effects of outliers and high leverage points.



Furthermore, analysis of the Cook’s D and residual plots, below, indicate findings similar to those in the first model. There are several observations which are outliers, and impart high leverage. For the same reasons previously described, however, they cannot be removed.



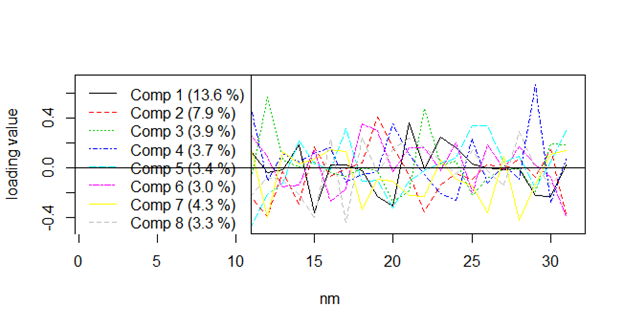
Typically, sensor data such as these with high dimensionality and collinearity concerns warrant being modeled outside of OLS regression. To mitigate these concerns, a PLS model was fit in an attempt to define the model’s shared variance. This resulted in a reasonably predictive, parsimonious model.

Using the PSLR package in R, all variables, except activity ID, were included for evaluation in the model. Using Cross-Validation it was determined that the appropriate number of components to represent the most variance within the model was 8, as shown in the below validation plot.



These 8 components represent nearly 67% of the model’s variance. This can be seen in the below score plot. Component 1 accounts for roughly 13% of the variance, followed by Component 2, with approximately 8%. These two components account for about 21% of the model’s variance. It makes intuitive sense that a regression model does not capture most of the predictive variability in heart rate as myriad of factors not captured in this dataset impact heart rate (physical condition, weather, etc.).

The loadings, or variables comprising each component, can be seen in Appendix B. Evaluating the model loadings, one can tell that the temperature sensors and magnetometers are highly impactful in each individual component.



As is the case with all model evaluations, a primary concern with this one focuses on performance of the model on the holdout sample. In this instance, the MSE of this PLS model equaled roughly 243.8. This is considerably better than the first OLS model with all predictors, but slightly worse than the OLS model with 7th order polynomials and interaction terms.

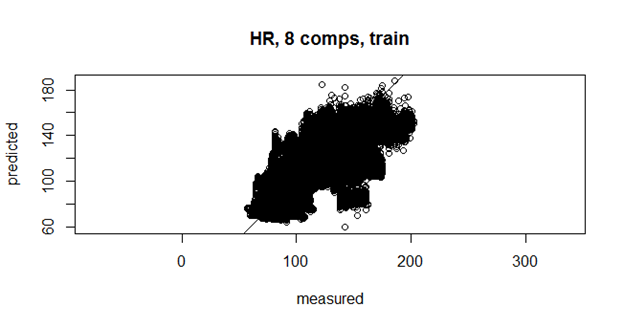
One also must evaluate the fit metrics of the model. This PLS model appears to have normally distributed residuals, as shown in Figure LR.1 below. This, in conjunction with out of sample performance, makes this model viable as the best of the linear regression models.

Although performing slightly worse than the second OLS model, this Partial Least Squares model achieves a higher degree of interpretability and is more parsimonious. With the performance on the validation set reasonably similar, a case can justifiably be made to select this model. It should be noted, however, that none does a particularly good job of predicting heart rate for reasons previously discussed. Many potential factors not accounted for in the data likely contribute to one’s heart rate.

## **Linear Regression Implementation Plan**

Based on the results of the linear regression model(s) fitted, we recommend Trivisio investigate development of a Chest sensor which incorporates temperature monitoring and magnetometer sensors. This would provide the best representation of an at risk cardiac patient’s heart rate, in relation to cost, factoring in all the potential individual sensor possibilities.

**Figure LR.1**



# **Logistic Regression Evaluation**

In support of our Cluster Analysis, we will employ a Logistic Regression Model used as a “typing” tool to classify new observations into a cluster activity by new sensor data obtained, without knowing the activity. Utilizing R and beginning the EDA for the logistic regression model first started with a review of the HR summary statistics for all participants individually and as a whole group across all activities. Understanding the prediction variable needs to be a binary measure of HR, we selected the median HR on the complete (all 9 subjects together) data set to separate the data.

**HR Median = 104**

**HR Mean = 107.3**

According to the American Heart Association[[5]](#footnote-5), the average resting heart rate of adults range from 60 to 100 beats per minute and heart rates during moderately intense activity is generally 50-69% of the maximum heart rate. Maximum heart rate data are known for each subject. With a 20% range of heart rate potentials for moderate intensity activities and the limited listing of activities our subjects could perform, we wanted to capture more of the moderate activity in the elevated category of heart rate. What you’ll see is the average HR of our 9 subjects at 54% of their individual maximum heart rates is 104.1 – very much in line with our median HR for the 9 subjects.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Subject | Max | 50% | 54% | 69% |
| 1 | 193 | 96.5 | 104.22 | 133.17 |
| 2 | 195 | 97.5 | 105.3 | 134.55 |
| 3 | 189 | 94.5 | 102.06 | 130.41 |
| 4 | 196 | 98 | 105.84 | 135.24 |
| 5 | 194 | 97 | 104.76 | 133.86 |
| 6 | 194 | 97 | 104.76 | 133.86 |
| 7 | 197 | 98.5 | 106.38 | 135.93 |
| 8 | 188 | 94 | 101.52 | 129.72 |
| 9 | 189 | 94.5 | 102.06 | 130.41 |
|  | Average | 96.39 | 104.10 | 133.02 |

Considering the mean HR is greater than the median HR in the data set, the median value was added to the lower half of the separation where the higher HR’s will encompass 104.1 and greater and the lower HR’s will be less than 104.1. A binary variable of HR.Median was created.

Further considering the varied intensities of many activities, subsets were created for both the analysis and a view into the adequacy of the original HR separation. Activities were viewed as Daily, Chores, and Intensity events.

**Daily** – Lying, Sitting, Standing, Watching TV, Comp Work, Car Driving.

**Chores** – Descending Stairs, Walking, Vacuum Cleaning, Ironing, Folding Laundry, House Cleaning.

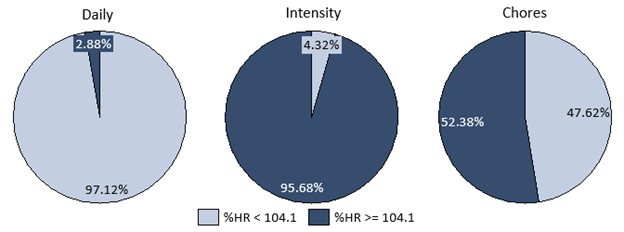
**Intensity** – Running, Cycling, Nordic Walking, Ascending Stairs, Playing Soccer, Rope Jumping.

Through the lens of these subsets, we have the sought separation of our newly created HR.Median variable. The mean values of the HR.Median for each category shows there is a clear separation of the HR activity of the participants when performing these different activities.

**Daily – HR.Median mean value = .0288**

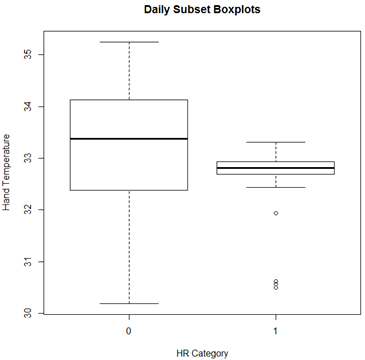
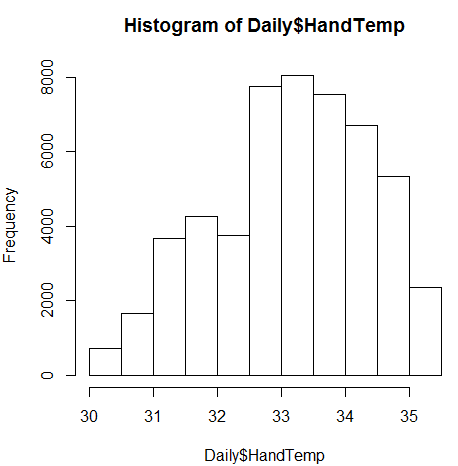
**Chores – HR.Median mean value = .4762**

**Intensity – HR.Median mean value = .9568.**



With the separation of the Daily and Intensity subsets we can examine the acceleration, gyroscopic, and magnetometer readings with an understanding of how the movements in 3-dimensional space impact the HR of the subject. The same EDA process for variable selection in the model will follow for all 4 created segments – the total set, daily, chores, and intensity.

Histograms were reviewed for all variables in the subset to view the data and boxplots for each variable against the binary variable of HR.Median such as the one below.



Following those reviews of the subset data, the data was parsed into training and test sets for all four categories and the EDA took on a trial and error exploratory method where models were built with all variables (excluding subject, HR, and activityID - for subsets) and the models were reviewed for variable importance, AIC, BIC, confusion matrix, and McFadden performance in an attempt to formulate the best factors for variables inclusion to maximize the model’s potential. While McFadden is considered a pseudo r-squared measurement with a value between 0 and 1, it operates a little different based on what is considered a good fit because it is a rho-squared value. A fit between .2 and .4 is considered very good and equivalent to that of an r-squared fit of .7 to .9 for a linear function.[[6]](#footnote-6)

## **Logistic Modeling**

Beginning with the Daily, Chores, and Intensity subsets, the modeling began with a baseline model where all variables were included with the exceptions of HR, Subject, and activityID for confounding of the model reasons. The HR is classified as the dichotomous variable, the subject is not our focus in this analysis, and the activities are already acknowledged in the subsetting of the data.

After completing the baseline model, the AIC and BIC were documented and a calculation of variable importance was reviewed. Further, the McFadden estimate or pseudo r-squared was calculated for the model. Utilizing the variable importance, variables were removed from the baseline model to create different iterations and the steps of documenting AIC, BIC, McFadden, and the new variable importance were completed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Daily** | | | | | |
|  | **Base Model** | **Model 1** | **Model 2** | **Model 3** | **Model 4** |
| AIC | 4922.4 | 5211.2 | 5276.4 | 5349.6 | 5395 |
| BIC | 5194.3 | 5372.7 | 5403.9 | 5613 | 5513.9 |
| McFadden | 0.479 | 0.445 | 0.438 | 0.433 | 0.425 |
|  |  |  |  |  |  |
| **Chores** | | | | | |
|  | **Base Model** | **Model 1** | **Model 2** | **Model 3** | **Model 4** |
| AIC | 40411 | 50031 | 40440 | 40421 | 40472 |
| BIC | 40691.4 | 50303.4 | 40633.3 | 40631.2 | 40638.5 |
| McFadden | 0.39 | 0.245 | 0.39 | 0.39 | 0.389 |
|  |  |  |  |  |  |
| **Intensity** | | | | | |
|  | **Base Model** | **Model 1** | **Model 2** | **Model 3** | **Model 4** |
| AIC | 10937 | 11380 | 10948 | 10929 | 10920 |
| BIC | 11211.5 | 11422.8 | 11084.7 | 11082.9 | 11090.9 |
| McFadden | 0.206 | 0.17 | 0.203 | 0.204 | 0.205 |

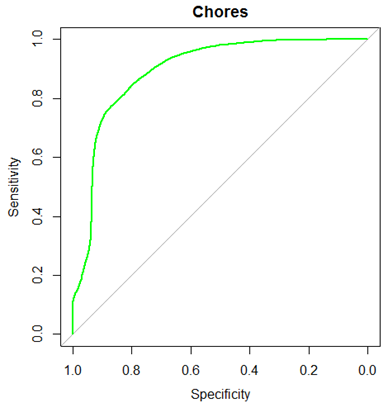
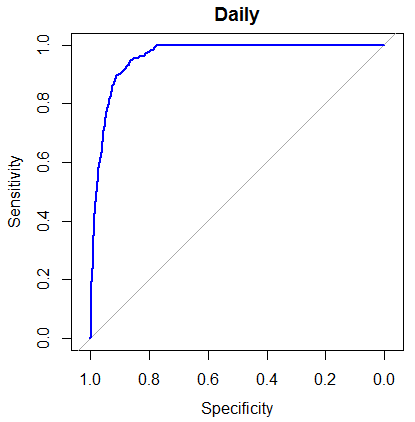
In the Daily case, the baseline model exceeded all other models in McFadden score, AIC, and BIC which also led to the highest AUC.

In the Chores case, the baseline model had the highest AIC, but Model 3 had a lower BIC and equivalent McFadden score, so with the inclusion of a penalty for factors in the model it appears Model 3 is the route we will proceed.

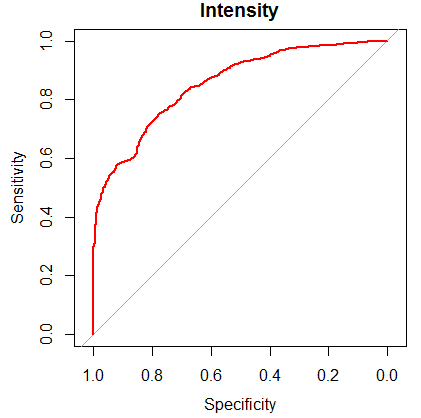
In the Intensity case, the baseline model had the highest McFadden score but lower higher AIC and BIC, resulting in the AUC for Model 4 exceeding the baseline model.

The fitted model was then used on the testing set to analyze performance generating the graphical representations below of the ROC-Chart and Lift Curve to measure the efficacy of the logistic regression model. The area under the curve indicates the degree (greater than .50) to which the model performs better than random chance.

What we see below are the ROC-charts for the Daily, Chores, and Intensity subsets including the area under the curve based on the models selected - Base Model, Model 3, and Model 4, respectively.

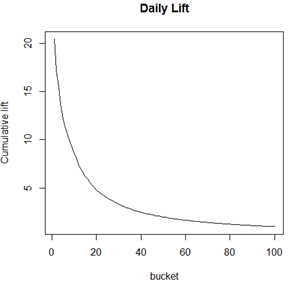


**Daily AUC = .9616 Chores AUC = .8927**



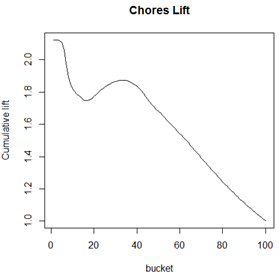
**Intensity AUC = .8539**

The Lift Curves for the 3 subsets models give us an indication of the effectiveness of the models and we see mixed results among the subsets. The lift on the Daily subset is outstanding with a KS statistic of 76.1% and predicting a cumulative percentage of 96.1% of the test data when the random percentage would be 20%.



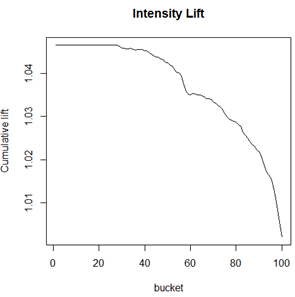
The performance of the model raises some concerns over potential accuracy issues – specifically when analyzed in unison with the Intensity lift chart we will see later.

The Chores data is effective with a KS statistic of 33.9% and a cumulative percentage of 83.9% of the test data when the random percentage would be 50%.



The analysis and graphical representations of this model do appear to have validity in usefulness and meaningfulness to Trivisio.

The Intensity data shows a performance just slightly better than random chance with a KS statistic of 2.3% and predicting a cumulative percentage of 82.3% of test data against a would-be random percentage of 80%.



## **Model Evaluation**

The next step from here is gathering insights from these models to solve Trivisio’s business proposition. Looking at the three models individually, the Daily subset list of variable importance is lengthy from using the base model, but there are clear variables of importance to the model including (in order):

* Time
* ChestTemp
* HandTemp
* AnkleTemp
* AnkleMag.1
* AnkleAccel16.1
* AnkleAccel16.3

From this information the impact is that temperature will be critical regardless of where the device is worn by the patient. We also understand that time is tracked equally among the 3 different device locations.

With that knowledge, it is likely those who perform a vast majority of their day-to-day life from activity categories in the Daily subset would be best served to have the ankle sensor for the heart rate level. This is a clear separation on the best option and better than we’ll see for other subsets, however, this may not be the best model for immediate implementation.

The Chores subset list of variable importance to the model include (in order):

* Time
* ChestTemp
* HandAccel16.2
* HandTemp
* AnkleTemp
* ChestGyro.2
* ChestAccel16.1
* HandAccel16.3
* ChestAccel16.3
* ChestMag.3

Again Time and temperature data are critical factors. To further the analysis, patients who often perform chores of moderate strenuousness but not regular intense activity and not a sedentary lifestyle of those in the Daily category, could be best served by a chest sensor in the future. This group does seem to have an alternative as the Hand sensor performs well in the acceleration categories. Patients characterized as those who perform chores of moderate intensity may have an option for which sensor is more comfortable or desirable. The Chores model has a good performance and most mimics the entire data set given heart rate levels are mixed almost evenly among the binary categories.

The Intensity subset list of variable importance to the model include (in order):

* Time
* ChestTemp
* ChestMag.1
* AnkleTemp
* HandTemp
* HandAccel16.1
* AnkleMag.1
* AnkleMag.3
* HandMag.1

At the top of the variable importance list we see Time and temperature data again. Beyond those factors, it appears the sensors on the extremities are favored for patients who intend to continue with intense physical activity. The extremity sensors have shown a propensity to capture data more attuned to this level of activity.

One of the risks Trivisio will face with utilizing this activity specific device recommendation platform is the truthfulness of the client in their intentions for the future. Few cardiac at-risk patients will want to voluntarily admit to a physician, or anyone, they intend to remain sedentary in the future. That makes assigning the appropriate sensor difficult because we see different usefulness among the three sensors for different activity levels of the wearer, however the appropriate sensor for the customer could help in predicting the likelihood of exceeding a heart rate level that may be dangerous given their prior cardiac condition.

One of the possible remedies to this risk is monitoring the frequency of measurements for categories such as acceleration data, but that may still be hit-or-miss in successfulness for Trivisio to identify the proper activity level category for the wearer because acceleration is not always intense activity and prolonged intense activity may only have changes in acceleration at the beginning and of the activity.

A second risk is the effectiveness of the Daily and Intensity subset models. To mitigate the risk of the erratic performances of those models in terms of lift and effectiveness, we may look at focusing on sensors for the whole of the subset and not based on the activity, or presumed activity level, of the patient.

To offer an alternative that may serve the population of Trivisio indirect patients as a whole, we did an analysis of HR on the entire set including all activities as well. This analysis eliminated the variables of HR, Subject, and activityID; the same as the subsets. In this case, activityID is removed because we hope to predict the heart rate level with the 3-dimensional data to understand the predictive power, and to some degree the physiological impact, of movement in this space - real life movement. The same steps were taken to analyze the data as whole including boxplots and histograms

The models created for the whole data set breakout with AIC, BIC, and McFadden below:

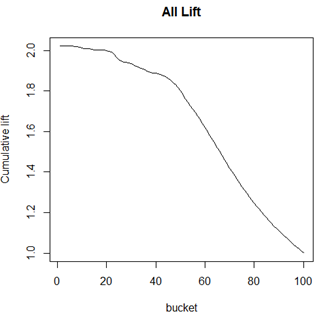
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **All** | | | | | |
|  | **Base Model** | **Model 1** | **Model 2** | **Model 3** | **Model 4** |
| AIC | 67466 | 97965 | 67826 | 67760 | 67765 |
| BIC | 67776.6 | 98266.3 | 67981.9 | 67964 | 67979.2 |
| McFadden | 0.604 | 0.425 | 0.602 | 0.602 | 0.602 |

We see little separation from one model to another, so while the base model has the best statistics, we will use Model 3 in favor of removing variables for a smaller model and less risk of over-fitting. The ROC-chart for Model 3 then represents the following:



**All AUC = .957**

The complete data set lift is effective similar to that of the Chores data with a KS statistic of 40% and a cumulative percentage of 90% of the test data when the random percentage would be 50%.



Addressing the data as a whole offers an alternative to Trivisio if there are meaningful insights within the model including the importance of variables and we see there absolutely is a clear separation.

The Complete data set list of variable importance to the model include (in order):

* Time
* ChestTemp
* ChestMag.3
* ChestMag.1
* AnkleTemp
* ChestAccel16.1
* ChestAccel16.2
* HandAccel16.2

At the top of the variable importance list remains Time and while ChestTemp is second we see two chest magnetometer readings before AnkleTemp, and in this case, HandTemp isn’t relevant in comparison. If there is uncertainty in the level of activity the patient is likely to have while wearing the sensor, the best bet is to prescribe a chest sensor to the patient. This would make intuitive sense as the subset category of Chores is most likely to perform like the data as a whole and Chores were deemed to be best characterized by the chest monitor – it offers the most universal results. Additionally, it mitigates the risk of the erratic performance resulting from the widely segmented HR data in the Daily and Intensity subsets.

## **Logistic Regression Implementation Plan**

Given the data we have from this analysis, we have a few suggestions for Trivisio on the next steps for their use of the sensors.

If Trivisio is ready to go to market with a sensor, we feel suggest they utilize the chest sensor without intentions to specialize in assigning sensors to customers based on activity level. The chest sensor provides Trivisio with strong results in predicting heart rate level with the measurement readings in the device and most uniformly fits different lifestyles.

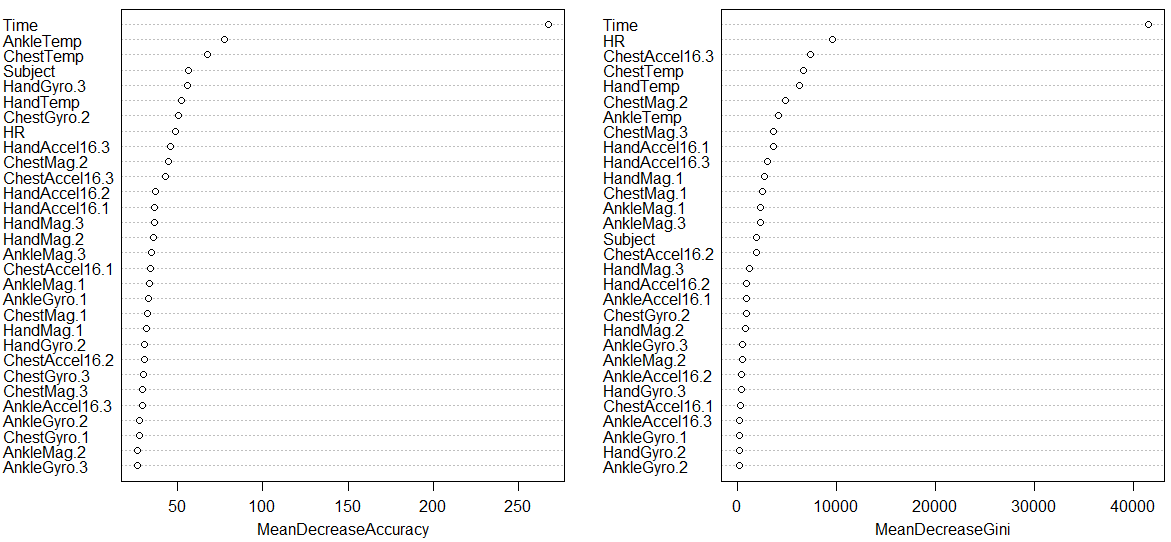
In the research performed, the test subject profiles had little variation in categories of age, height, and weight. It almost entirely discounted the population that is moderately to significantly overweight as well as the population outside of young adulthood. Considering the potential of use for this sensor in an elderly population, more research must be completed on a wide variety of test subjects. In addition to expansion of test subject physical characteristics, we also recommend more research be done on different lifestyles to gather more data specific to activity levels for examination.

Finally, the research showed little to no benefit of the gyroscopic data in the sensor. There is, of course, a cost to include that measurement sensor in the device and for the benefit, it is likely not worth the cost. It would be advised that go-forward sensors omit the gyroscopic reading altogether, unless used to further test as previously discussed. The benefits of this are potentially two-fold in both cost reduction and the possibility of a smaller, less noticeable, more comfortable, or even lighter device for the wearer.

# **Random Forest Evaluation**

The second aspect of our examination of this data was to develop a sense if the sensor readings could be utilized to develop accurate classification predictions for user activity. Multiple Random Forest models were trained and evaluated during the initial modeling process to evaluate the effectiveness of such models for the data, as well as to assess the importance of variable inclusion in the models and evaluate the key sensor inputs that contribute to the classification accuracy. These steps were completed using R.

An initial random forest model was developed using the entire randomized training (trainR) subset with the inclusion of the Subject indicator variable. The following variable importance plot (VIP) was generated:



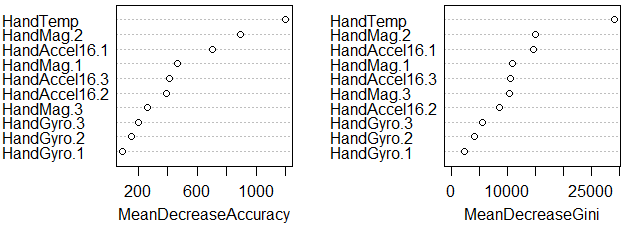
The plot indicates that the variable Time is highly important in the classification accuracy. This model accounted for 99.95% accuracy for the in-sample model, and out-of-sample accuracy was 99.94%. These results are very robust and encouraging.

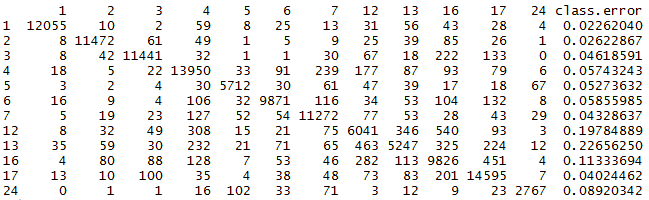
With the results of the initial random forest model, two additional models were trained to assess for any potential overfitting phenomenon and to ensure the ability to generate high accuracy predictions moving forward. First, a Support Vector Machine (SVM) model was trained and evaluated. Because SVM’s are both time- and computationally-intensive, a 20% sub-sample of the training data was taken to tune the SVM to identify the appropriate cost and gamma parameters. With a cost of 10 and a gamma value of 0.1, 94.24% prediction accuracy in the out-of-sample group was achieved.

A second evaluation model was developed using the same parameters of the initial random forest mode, but with the training and test data set roles reversed. By training the model on the 30% test sample and evaluating it on the remaining 70% of the data, we saw results in line with the baseline mode, with accuracy coming down marginally to 99.84% accuracy both in- and out-of-sample.

With overfitting and sample size issue concerns assuaged following the baseline modeling efforts, sensor-specific models were developed using the hand, chest, and ankle inputs solely to assess which of these variables were highly important and if accurate predictions could be made from the use of a single sensor. In addition, an inspection of the resultant confusion matrices provided insight into which activities were being most-often incorrectly classified.

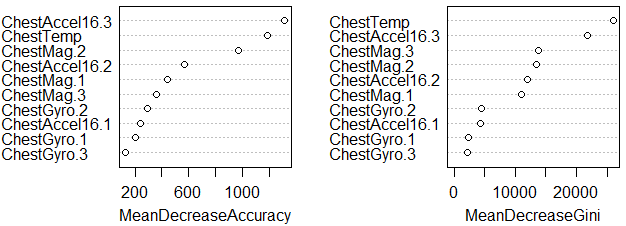
The Hand sensor VIF and confusion matrix is below:

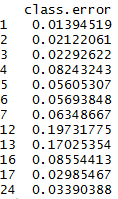
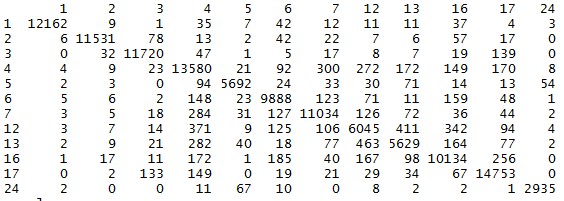




From these plots we see that HandTemp is the most indicative variable, while the most frequently misclassified activity was 13 (descending stairs), followed by 12 (ascending stairs), with class errors of 22.7% and 19.8%, respectively. This model performed well, with 92.99% accuracy for the in-sample model, and 93.02% accuracy for out-of-sample.

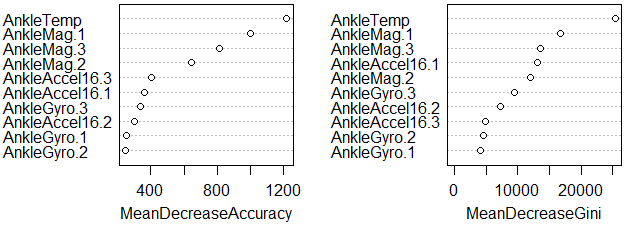
The Chest sensor VIF and confusion matrix is below:

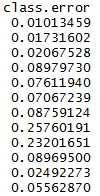
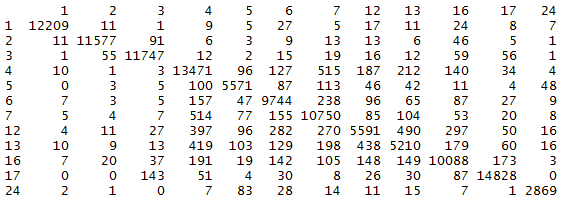




From these plots we see that ChestAccel16.3 is the most indicative variable, while the most frequently misclassified activity was 12, followed by 13, with class errors of 19.7% and 17.0%, respectively. This model performed well, with 93.69% accuracy for the in-sample model, and 93.72% accuracy for out-of-sample.

The Ankle sensor VIF and confusion matrix is below:





From these plots we see that AnkleTemp is the most indicative variable, while the most frequently misclassified activity was 12, followed by 13, with class errors of 25.8% and 23.2%, respectively. This model performed well, with 92.52% accuracy for the in-sample model, and 92.60% accuracy for out-of-sample.

With the encouraging results of the sensor-specific models, pared-down models were developed utilizing the variable importance plots to inform our subsequent variable inclusions. For each sensor type (hand, chest, ankle), we developed additional models retaining the top 5 variables (50%) in terms of mean decrease accuracy and gini. We noted similar high accuracy predictions across the board, even while eliminating half of the variable inputs, which was encouraging. Below is a table depicting the variables included in the sensor-specific sub-models, and their relative accuracy scores.

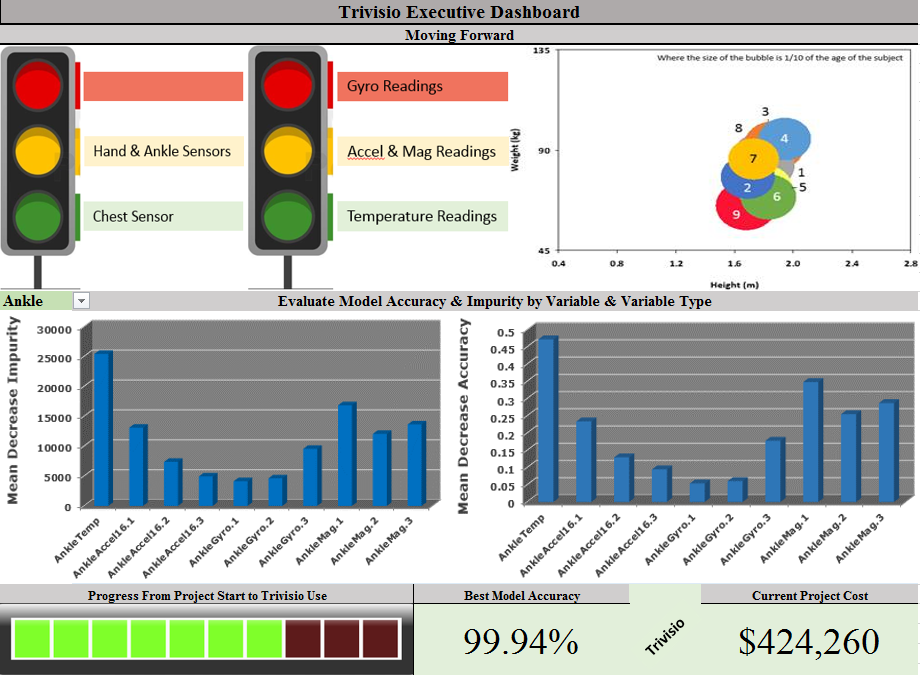
|  |  |  |  |
| --- | --- | --- | --- |
| **Sub-Model** | **Variables Included** | **In-Sample Accuracy** | **Out-of-Sample Accuracy** |
| Hand | HandTemp, HandAccel16.1, HandAccel16.3, HandMag.1, HandMag.2 | 89.89% | 89.69% |
| Chest | ChestTemp, ChestAccel16.2, ChestAccel16.3, ChestMag.2, ChestMag.3 | 91.30% | 91.40% |
| Ankle | AnkleTemp, AnkleMag.1, AnkleAccel16.1, AnkleMag.2, AnkleMag.3 | 89.00% | 89.16% |

A comprehensive summary table of the classification modeling results is included below:

|  |  |
| --- | --- |
| **Model** | **Out-of-Sample Accuracy** |
| Initial RF | 99.94% |
| SVM | 94.24% |
| Reverse RF | 99.84% |
| Hand | 93.02% |
| Chest | 93.72% |
| Ankle | 92.60% |
| Sub-Hand | 89.69% |
| Sub-Chest | 91.40% |
| Sub-Ankle | 89.16% |

While the initial, all-variable models outperformed our subsequent sensor-specific models, we saw high prediction accuracy for our subject activities across the board. With respect to the sensor-specific models, we typically experienced a 2-3% reduction in accuracy with the removal of 50% of the inputs, which is a favorable tradeoff. The chest sensor, across the board, exhibited the best results of the sensor-specific undertakings. A final observation that for all models, the gyroscopic sensor inputs typically offered little value in terms of classification prediction power.

# **Dashboard**



# **Conclusion**

Trivisio commissioned us, Sterling Analytics, to provide guidance on the potential design and development of wearable sensor technology useful in tracking performance in at-risk subjects with compromised cardiac function.  Temperature, Gyroscope, Magnetometer and Accelerometer sensors were developed for the Chest, Hand and Ankle, storing measurements from 9 subjects engaged in activities with varying intensity.  The measured sensor data comprised a grouping of variables used to potentially predict heart rate via Linear Regression.  Additionally, the sensor data was used to classify activity based on input variables via Logistic Regression, Support Vector Machine and Random Forest models.

First, predicting heart rate with the sensor data provided mixed results.  Among several linear models, a Partial Least Squares Regression model was fit, yielding 8 components accounting for 66% of the variance in heart rate.  Performance was not best of the fitted models out of sample, but very close.  Due to the interpretability of the model, and analysis of the component loadings, we conclude that a chest-derived sensor incorporating temperature, magnetometer and, design-permitting, accelerometer would be most effective in tracking cardiac performance.

Classifying activity based on sensor data proved to be much more accurate, in comparison to the linear models.  The logistic regression, SVM and forest models all predicted activity with a high degree of accuracy.  Supporting the linear regression models, these classifiers also suggest a chest-derived sensor incorporating temperature and magnetometer readings most accurately classified activity regardless of its intensity.  All modeling activity suggest the gyroscopic sensors were of little-to-no value in either predicting heart rate, or classification of activity based on intensity.

In summation, Trivisio has sought to expand its product mix due to the pressures of eroding margins in its core business, head-mounted optical devices.  Acting on inquiries from the FDA to develop a comprehensive sensor that accurately predicts HR based on 3D movement, Triviso summoned Sterling to provide feasibility consultancy grounded in the best practices of predictive modeling.  The models developed, though focused on different aspects of the dataset, all reach the same conclusion.  Trivisio should develop a chest-derived sensor which incorporates temperature, magnetometer and, design-permitting, accelerometer trackers to best satisfy the requirements as described here.  Once a design has been achieved, and prototype(s) developed, more research should be implemented to further evaluate the inferences presented here.  It is our belief, that a highly accurate multi-functional sensor can be developed to track cardiac performance in at-risk patients.

# **Implementation**

The analysis of the IMU data should provide valuable insights into new measures of heart rate calculations and predictions. With this data, it may become possible to understand the effects of body temperature, altitude, high-intensity work, and daily activities on the heart rate and with future testing the long-term effects of prolonged exposure to any of those items or a lack thereof. From the current study, the output will determine if one of the three devices is more accurate than another device (wholly or in certain functions) and whether ambient temperature, acceleration data, gyroscopic data, and magnetometer data are useful in heart rate prediction.

Intervention in the industry related to the heart rate data and availability of new prediction measurements related to the nature of different activities on performance of the heart will be the primary focus. Utilizing personalized wearable devices will bring about the ability to drive a marketing campaign as well, in the future, should the data be recognizable to the wearer in a meaningful way.

Measurement of the implementation plan will be to both run a larger scale data collection with additional subjects of more variety in age and gender as well as to combine with survey data related to self-assessment of often performed activities in patients with heart disease.

# **Appendix A**

# **Solution Development & Implementation Report**

## **Data Quality Auditing**

The data quality auditing was completed in two steps:

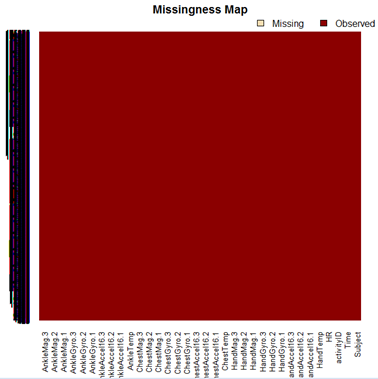
1. A complete review of the individual subject data each of the 9 participants.
2. A complete review of the complete data set merged together containing all 9 participants.

The data was originally pulled from the University of California Irvine website where we obtained a data set containing 3,850,505 instances, 52 attributes, missing values, and a data dictionary. The characteristics of the data set are multivariate and time-series. The data is consistent with what the client described including observations and variables. We’ve confirmed the data are consistent with our project needs as well. Initially reviewed in Excel as .dat files, the data was contained as space-separated files which were converted to comma-separated files in Excel. Category headers were added in the Excel file, and the data was loaded into R.

An EDA is required to identify the variable classification, and basic summary statistics (included in the Variable Summary Table below). In the initial EDA, a quick review of the first six records and last six records for each subject and for the total data set was completed using the head() and tail() function in R. This analysis of the records also presents a starting point to verify the data consistency. In these records, not much consistency verification is necessary, but we can see the timestamps and temperature data coincide with real-world expectations. Later a review of the variable distributions will be addressed based on classification variable type and our data does not require building a weighting system for sample data.

Following that first review, the removal of all transient activity (activityID = 0) from the data is completed due to the notes with the data indicating this to be transient activity picked up by the monitors and lacking meaning, the heart rate data was reviewed where we have our only missing data. The sampling frequency for the IMU devices was approximately 100Hz but the heart rate monitor was sampling at a frequency of 9Hz which led to approximately every 10th record to indicate a heart rate reading. For ease of analysis, we have removed those records with “NaN” for heart rate, effectively reducing the frequency of IMU data down to 10 measurements per second. In merging the nine data sets a new variable indicating the subject was generated to tie the records to one of the nine subjects. We are now down to 175,500 records and 34 variables which gives us just under 5% of the total instances for credible data.

Verifying the quality of the data began with a check of each individual subject data set for incomplete cases and none were found. A verification of the complete data set with all nine subjects also found no incomplete cases (see Missingness Map below).



In all ten files the variables names of those variables being utilized for research were the same and included the 34 following:

Subject, Time, activityID, HR, HandTemp, HandAccel16.1, HandAccel16.2,

HandAccel16.3, HandGyro.1, HandGyro.2, HandGyro.3, HandMag.1, HandMag.2,

HandMag.3, ChestTemp, ChestAccel16.1, ChestAccel16.2, ChestAccel16.3,

ChestGyro.1, ChestGyro.2, ChestGyro.3, ChestMag.1, ChestMag.2, ChestMag.3,

AnkleTemp, AnkleAccel16.1, AnkleAccel16.2, AnkleAccel16.3, AnkleGyro.1,

AnkleGyro.2, AnkleGyro.3, AnkleMag.1, AnkleMag.2, AnkleMag.3.

Prior to this review, variables found to have no beneficial meaning were removed and those variables include:

HandAccel6.1, HandAccel6.2, HandAccel6.3,

HandOrient.1, HandOrient.2, HandOrient.3, HandOrient.4, ChestAccel6.1,

ChestAccel6.2, ChestAccel6.3, ChestOrient.1, ChestOrient.2, ChestOrient.3,

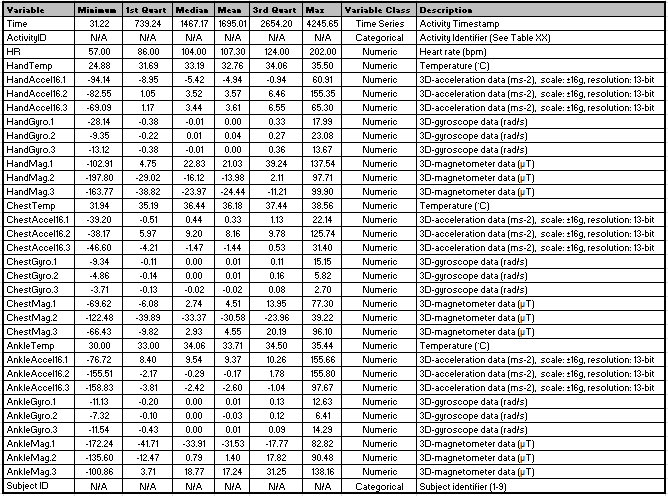
ChestOrient.4, AnkleAccel6.1, AnkleAccel6.2, AnkleAccel6.3, AnkleOrient.1,

AnkleOrient.2, AnkleOrient.3, AnkleOrient.4.

Many of these variables include orientation or are duplications or an ill-calibrated accelerometer.

Included below, in the Variable Summary Table, are the highlights of the minimum, maximum, quartiles, median, mean, variable name, description, variable class, and further confirmation of no missing values in the chart of variables.

## **Variable Summary Table**



## **Measurements**

Our strategic measurements are heart rate and activity ID. Heart rate is a continuous variable and Activity ID is a character/categorical variable that codes for different activities. The measurements we will use to predict heart rate and activity come from the IMU data, primarily temperature, acceleration, gyroscopic and magnetometer data. The last three data types are broken into separate fields for each of the three axes, so both magnitude and direction will be captured in the predictive models. This makes for a very clean dataset, although statistical transformations may be in order on the measured variables to heighten accuracy.

One transformation that will be needed is to dummy code each of the activity variables for specific applications (e.g. regression analysis within a CART model). Of course, many of these transformations will become more apparent after a thorough Exploratory Data Analysis (EDA) on the measured variables has been performed, as described in the next section. For a complete list of the measured variables that will be used in the analysis, refer to the Appendix – Data Fields & Data Descriptions. In addition to dummy variable coding, our EDA may reveal that combining the acceleration data in three directions into one acceleration field. Direction isn’t relevant for our statistical modeling and combining the IMU data may prove to be a better predictor of heart rate than the data being left in separate fields.

## **Computational Methods**

Heart rate values are continuous and are most likely suited for modeling by linear regression, while predicting activity type will likely be best estimated by some form of regression or classification model, including logistic regression, support vector machine [SVM], classification and regression tree [CART], random forest, and/or other models. Linear regression results will be interpreted for the various goodness of fit parameters, including residual/predicted value residual plots, fit plots for Y, Q-Q residual plots, Cook’s D values, R-square values, and/or other analysis of variance parameters. Out of sample goodness of fit metrics will include Mean Absolute Error (MAE) and Mean Square Error (MSE) measures.

Classification will most heavily be evaluated on accuracy/precision values, confusion matrices, and/or other means. Cluster analysis will involve a hierarchical dendrogram, signifying the distance between successively larger clusters, and a k-means visual, likely a clustergram.[[7]](#footnote-7)

## **Expected Methodology and Visuals Explaining Concepts**

As detailed further below, the expected methodology for the project will be to conduct various linear regressions to predict heart rate. A high number of explanatory variables in the data set have significant correlations with heart rate, but we will perform various variable transformations before settling on a final model. Principal components and factor analysis will be used to reduce dimensionality and account for the multicollinearity that may be present among predictors. Other types of linear regression models that may prove to have superior goodness of fit include quantile models, lasso regularized models, L1 and L2 regression models, which shrink the regressor coefficients in instances of high collinearity or dimensionality among predictors, and hierarchical regression models.

Development of classification models will involve training multiple model types on the training data and evaluating the out-of-sample predicted accuracy of those models against each other. As some of these models are “black box” learning methods, the inner workings w=of these models will be evaluated where practical, but the outcome of each will be used as the primary benchmark for evaluation. For random forest models, an iterative approach to variable selection will be conducted utilizing a review of the variable importance metrics, training sub-variable models, and evaluating the competitive merit of the scaled-back models against their larger subset peers. Ultimately our rationale will be to determine which variables contribute the most value in terms of predictive power to the classification of activities, and these will be conducted with respect to the entire sensor input set, along with a sensor-specific approach.

The key components of our project plan include Project Deliverables, Project Requirements, Project Constraints/Priority Matrix, Project Assumptions, Project Risks, Stakeholder Register, Communication Matrix, Milestones Schedule, Summary Budget, Work Breakdown Structure, AON, Network Information Table, Gantt Chart, Schedule Baselines, Resource Loading & Leveling, Project Approval Guidelines and Authorizations. See the project plan section in the appendix of this report for more details.

### **Tools, Techniques, & Transformations**

The primary tools used for the analysis of the PAMAP2 Physical Monitoring Data Set are R and SAS. SAS is being used primarily for the OLS multivariate regression and logistic regression modeling tasks, while R is being used to incorporate the machine learning analysis.

### **Recommended Tools and Techniques of Data Retrieval**

Data for this project has been obtained for us by the client. Our tools, both R and SAS, use import procedures such as proc import in SAS and read.csv in R to import the raw, comma delimited files into each respective programming environment. Although many tools and techniques for acquiring data are available, such as web scraping scripts in Python, these methods are not relevant or necessary for our analysis. In addition to receiving our data from the client, the dataset is readily downloadable using any web browser at UCI’s Machine Learning Repository.[[8]](#footnote-8)

### **Recommended Tools and Techniques of Data Integration**

The tools and techniques of data integration widely vary from project to project. Our analytics project focuses on a single dataset, produced in nine distinct .dat files. The peculiarities of our data have necessitated some cleaning before integration. Specifically, each .dat file contains both fields of data as well as records of data that are not relevant for our analysis. In an attempt to keep our integrated data at a manageable size, we’ve removed these fields and observations from each respective .dat file as follows:

1. Limit the number of fields in each dataset by dropping 21 fields related to orientation (invalid in this data collection) and lower resolution 3D-acceleration data, as cited in the PAM source documentation.
   1. After importing into SAS, the keep statement can be used to limit the data fields in each data set to the 33 fields of interest for our analysis.
   2. After importing into R, the subset function can be used to select certain variables: subset(subject101, select=c(“Time”,”ActivityID”,”HR”,...)
2. Limit the number of observations in the data to relevant records
   1. All of the IMU data is recorded with a frequency of 100 Hz, or 100 measurements per second. The heartrate monitor, however, is recorded with a frequency of roughly 9 Hz, or nine measurements per second. Since HR is the primary target variable, we’ve decided to exclude all observations in excess of 9 cycles per second. Another integration technique is to combine these additional records in some way (such as taking a weighted average of the 10 IMU measurements for every HR measurement), although preliminary modeling suggests the data collected at a frequency of 9 Hz is sufficient to predict activities and HR with a high degree of accuracy.
   2. Records with “N/A” values for HR are removed from each dataset. This is done in Excel by manually filtering for the non-N/A values, but can also be done in SAS using the where clause in a proc SQL statement, or by using lapply(df,na.omit) in R.

Once each individual file has been cleaned, the individual datasets are combined in R using the rbind function or with a set statement in a data step in SAS.

### **Recommended Tools and Techniques of Data Transformations**

Variable transformations are an important way of increasing a model’s predictive accuracy. Measured variables can be combined to create new variables with more meaningful inputs. A common transformation is to calculate rates or ratios from measured variables by dividing one measured variable by another. An example of a data transformation in our dataset includes some of the following:

1. Combine the three vectors of each IMU measurement (acceleration, gyroscope and magnetometer) by adding them together. We are interested in the magnitude of each measurement, and not necessarily the directional component. By adding the three directional measurements of acceleration together, for instance, we create one variable that represents the total acceleration “component” of predicting heart rate and / or activity ID. This variable reduction technique addresses the potential for multicollinearity problems in a multivariate regression that uses the three acceleration fields in isolation.
2. Ratio analysis: by transforming two variables into one by calculating the quotient may heighten predictive accuracy. One transformation under investigation is normalizing fluctuations in subject temperature by acceleration, gyroscope, or magnetometer data.
3. Performing principal components analysis or factor analysis as a means of reducing the potential for multicollinearity issues in regression. Creating new, component variables from measured variables insures that the inputs to regression are uncorrelated, or orthogonal to each other. Interestingly, PCA as an unsupervised dimensionality reduction technique has been shown to be mathematically similar to the k-means clustering technique, providing important justification for PCA based data reduction.[[9]](#footnote-9)

### **Recommended Tools and Techniques of Test/Validation Set-Up Procedures**

We’ve divided our data into two partitions, training and test sub-sets. We’ve performed this subsetting of the data in R using the following code:

sample\_size <- floor(.70\* nrow(data))

set.seed(498)

int <- sample(seq\_len(nrow(data)),size=sample\_size)

train <- data[int,]

test <- data[-int,]

Our data is separated into training (70% of observations) and test (30% of observations) data in order for our supervised learning models (regression analysis and classification algorithm, specifically the Classification and Regression Trees, or CART model). Models that are trained on the training data can then be validated by comparing their predictions to actual values using the test data.

Splitting our overall data set in this manner is critical to our analysis for multiple reasons. First, removing 30% of the observations results in a smaller data set, which will improve processing times for computationally-intensive machine learning tasks such as Random Forest or SVM. Second, training models on a subset of the available data acts as a built-in deterrent to model over-fitting occurrences. Lastly, a hold-out sample of our data allows us to evaluate our best models in practice on unseen observations, which is inherently valuable in assessing the overall predictive performance of the modeling efforts.

Simple measures of predictive potency which we will use include mean absolute error (MAE) and mean squared error (MSE). MAE sums up the difference between predicted values (resulting from test data inputs to the model) and actual values (already in the test data) on an absolute value basis, and then divides by the number of observations in the data. MSE is the same calculation but uses the square of the difference instead of the absolute value. Both measures estimate the magnitude of the difference between predictions and actual values, and can be used for selecting the most “accurate” model from a group of models. They can also be used to validate models irrespective of each other, by comparing to MAE and MSE values generated from the delta between actual values and randomly generated values. If the Model’s MAE and MSE are smaller than those generated by random values, the predictive model is providing meaningful insight.

Preliminary investigations suggest that we can achieve in excess of 99% accuracy in predicting activity ID in the test data using a random forest model. Subsequent iterations of the modeling will be focused on achieving similar levels of model accuracy during with fewer inputs to the model. If this can be achieved, the client will be able to reduce overhead by collecting fewer activity measurements, a major goal of our analytics assignment.

# **Project Plan**

## **Project Deliverables**

The following deliverables must be met upon the successful completion of the PAM Device project. Any changes to these deliverables must be approved

* How activities affect heart rate based on the location of the device
* Heart rate activity variations by activity, weight, and time of day
* Identification of IMU sensors with the largest contribution to prediction of heart rate by activity
* Model performance metrics
* Recommendations for strategy & implementation
* Dashboard

## **Project Requirements**

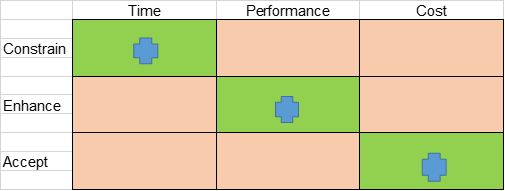
* Data accuracy and integrity.
* Clear communication of the goals and expectations.
* Implementation plan and presentation.

## **Work Breakdown Structure**

|  |  |  |
| --- | --- | --- |
| **Level** | **WBS Code** | **Element Name** |
| 1 | 1 | Physical Activity Monitoring |
| 2 | 1.1 | Initiation |
| 3 | 1.1.1 | Goals and Objectives Meeting |
| 3 | 1.1.2 | Develop Project Charter |
| 3 | 1.1.3 | Deliverable: Submit Project Charter |
| 3 | 1.1.4 | Sponsor (Trivisio) Review Project Charter |
| 3 | 1.1.5 | Project Charter Signed/Approved |
| 2 | 1.2 | Planning |
| 3 | 1.2.1 | Review Prior Work/HR background |
| 4 | 1.2.1.1 | Explore history of HR |
| 4 | 1.2.1.2 | Explore 3-D monitors |
| 3 | 1.2.2 | Create Project Scope Statement |
| 3 | 1.2.3 | Project Team Kickoff Meeting |
| 3 | 1.2.4 | Develop Project Plan |
| 3 | 1.2.5 | Project Plan Review with Trivisio |
| 4 | 1.2.5.1 | Incorporate Feedback |
| 3 | 1.2.6 | Submit Project Plan |
| 3 | 1.2.7 | Milestone: Project Plan Approval |
| 2 | 1.3 | Execution |
| 3 | 1.3.1 | Verify & Validate Data Quality |
| 3 | 1.3.2 | Exploratory Data Analysis |
| 3 | 1.3.3 | Subset Data By Activity |
| 3 | 1.3.4 | Create Testing & Training Subsets |
| 3 | 1.3.5 | Modeling Designs |
| 4 | 1.3.5.1 | Linear |
| 4 | 1.3.5.2 | Logistic |
| 4 | 1.3.5.3 | Cluster |
| 4 | 1.3.5.4 | Random Forest |
| 3 | 1.3.6 | Model Review & Validation |
| 3 | 1.3.7 | Review Implementation Strategies |
| 3 | 1.3.8 | Develop Implementation Plan |
| 3 | 1.3.9 | Create Dashboard |
| 2 | 1.4 | Control |
| 3 | 1.4.1 | Project Management |
| 3 | 1.4.2 | Project Status Meetings |
| 3 | 1.4.3 | Risk Management |
| 4 | 1.4.3.1 | Secondary model review |
| 3 | 1.4.4 | Update Project Management Plan |
| 2 | 1.5 | Closeout |
| 3 | 1.5.1 | Audit Procurement |
| 3 | 1.5.2 | Present Findings & Implementation Plan |
| 3 | 1.5.3 | Gain Formal Acceptance |
| 3 | 1.5.4 | Document Lessons Learned |
| 3 | 1.5.5 | Archive Files/Documents |

|  |  |  |  |
| --- | --- | --- | --- |
| **Level 1** | **Level 2** | **Level 3** | **Level 4** |
| 1 Physical Activity Monitoring | 1.1 Initiation | 1.1.1 Goals and Objectives Meeting  1.1.2 Develop Project Charter  1.1.3 Deliverable: Submit Project Charter  1.1.4 Sponsor (Trivisio) Review Project Charter  1.1.5 Project Charter Signed/Approved |  |
| 1.2 Planning | 1.2.1 Review Prior Work/HR background  1.2.2 Create Project Scope Statement  1.2.3 Project Team Kickoff Meeting  1.2.4 Develop Project Plan  1.2.5 Project Plan Review with Trivisio  1.2.6 Submit Project Plan  1.2.7 Milestone: Project Plan Approval | 1.2.1.1 Explore history of HR  1.2.1.2 Explore 3-D monitors  1.2.5.1 Incorporate Feedback |
| 1.3 Execution | 1.3.1 Verify & Validate Data Quality  1.3.2 Exploratory Data Analysis  1.3.3 Subset Data By Activity  1.3.4 Create Testing & Training Subsets  1.3.5 Modeling Designs  1.3.6 Model Review & Validation  1.3.7 Review Implementation Strategies  1.3.8 Develop Implementation Plan  1.3.9 Create Dashboard | 1.3.5.1 Linear  1.3.5.2 Logistic  1.3.5.3 Cluster  1.3.5.4 Random Forest |
| 1.4 Control | 1.4.1 Project Management  1.4.2 Project Status Meetings  1.4.3 Risk Management  1.4.4 Update Project Management Plan | 1.4.3.1 Secondary model review |
| 1.5 Closeout | 1.5.1 Audit Procurement  1.5.2 Present Findings & Implementation Plan  1.5.3 Gain Formal Acceptance  1.5.4 Document Lessons Learned  1.5.5 Archive Files/Documents |  |

## **Project Constraints/Priority Matrix**



As a result of the current state of the industry and product, the time constraints on the analysis intelligence is at the forefront of the project. Examining the data for meaningful insights but with a new features or upgrades requiring a technology and development team, costs are viewed as fairly accurate in the estimation and it would be expected that fluctuations would be small and can be accepted. It is of the utmost importance for the future of the product to enhance the current marketing campaign and target the appropriate audiences for each of the three device options.

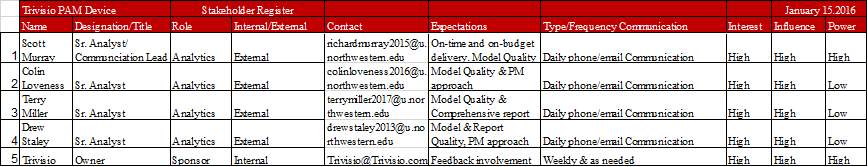
## **Project Assumptions**

* The small sample size will be representative of heart rate data for the population at least within the age range of the data participants.
* Heart rate data could vary from the analysis for those age groups significantly different than those tested.
* The participants are non-smokers and otherwise healthy individuals.
* The wrist, ankle, and chest all have equivalent ability to monitor the heart rate accurately.

## **Project Risks**

* Data quality is lacking.
* The 9 sample individuals are not representative.
* Competitors coming to market with similar devices.
* Finding meaningful results/insights

## **Stakeholder Register**



## **Communication Matrix**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Communication Type** | **Objective of Communication** | **Medium** | **Frequency** | **Audience** | **Owner** | **Deliverable** | **Format** |
| **Goals & Objectives Meeting** | **Introduce the project team and the project. Review project objectives and management approach.** | **· Conference Call** | **Once** | **· Project Sponsor**  **· Project Team** | **Project Manager** | **· Agenda**  **· Meeting Minutes** | **· Soft copy archived on project SharePoint site and project web site** |
| **Project Team Meetings** | **Review status of the project with the team.** | **· Conference Call**  **· Email** | **Semi-daily & As Needed** | **· Project Team** | **Project Manager** | **· Agenda**  **· Meeting Minutes**  **· Project schedule** | **· Soft copy archived on project SharePoint site and project web site** |
| **Project Status Reports** | **Report the status of the project including activities, progress, costs and issues.** | **· Conference Call** | **Weekly & As Needed/Requested** | **· Project Sponsor**  **· Project Team** | **Project Manager** | **· Project Status Report**  **· Project schedule** | **· Soft copy archived on project SharePoint site and project web site** |

## **Milestones Schedule**

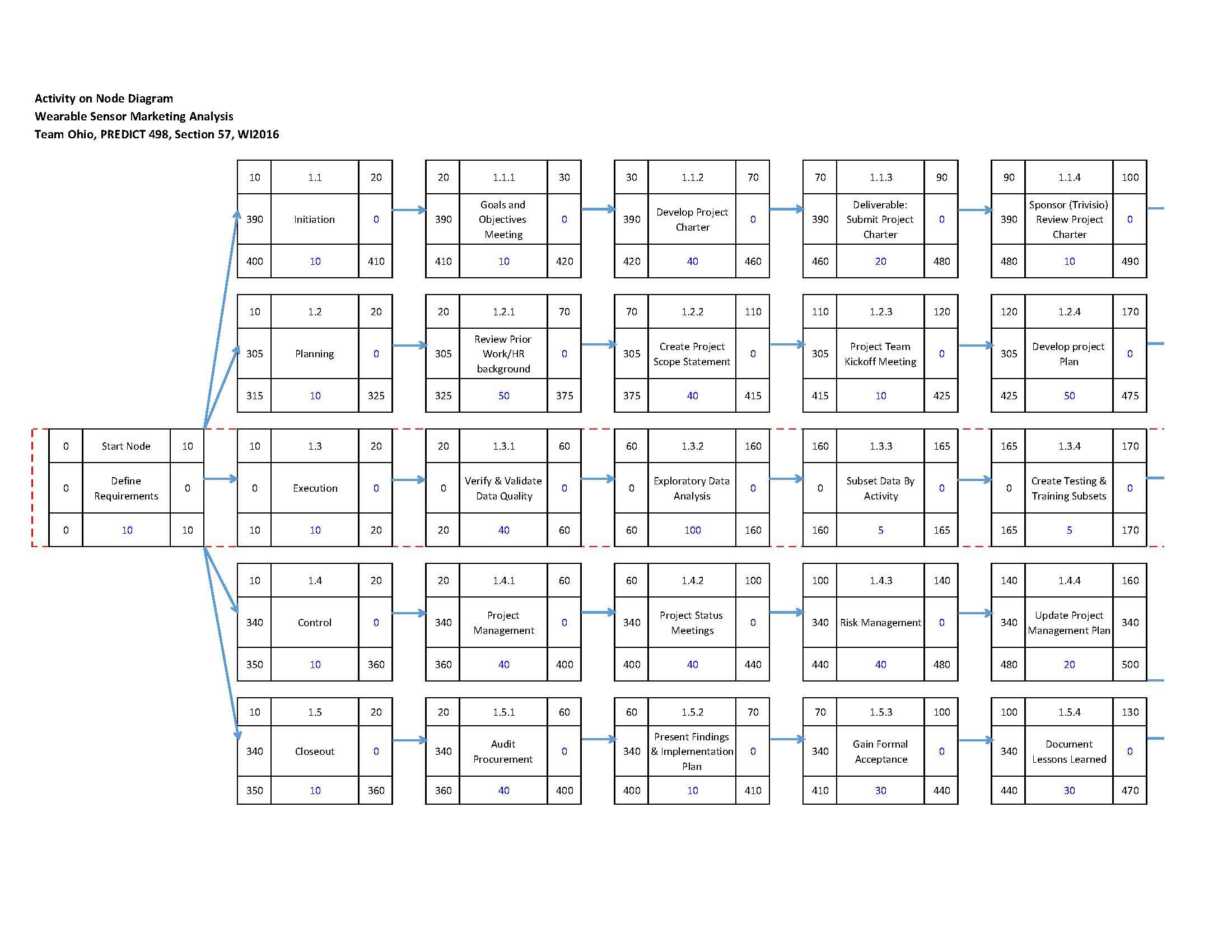
|  |  |
| --- | --- |
| **Summary Milestone Schedule – List key project milestones relative to project start.** | |
| **Project Milestone** | **Target Date (mm/dd/yyyy)** |
| · Project Start | 01/04/2016 |
| · Develop Project Plan | 01/14/2016 |
| · Data Quality Validation & EDA Completion | 02/01/2016 |
| · Model Building & Selection | 02/22/2016 |
| · Implementation Plan Complete | 02/29/2016 |
| · Present to Trivisio and Supply Dashboard | 03/06/2016 |
| · Project Complete (Address Any Problems) | 03/13/2016 |

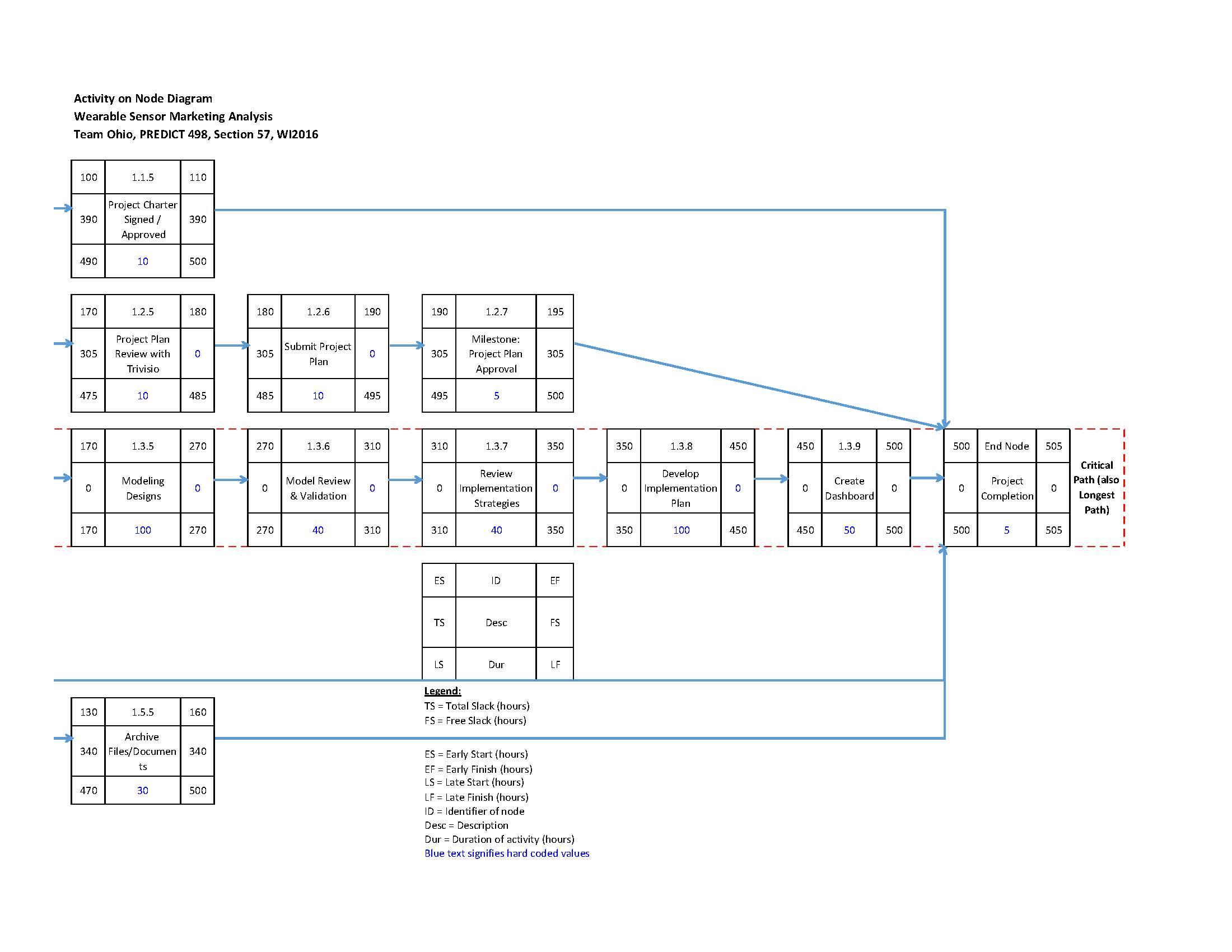
## **Summary Budget**

|  |  |
| --- | --- |
| **Summary Budget – List component project costs** | |
| **Project Component** | **Component Cost** |
| · Personnel Resources | $600,799 |
| · Dashboard Software | $18,000\* |
| **Total** | **$618,799** |

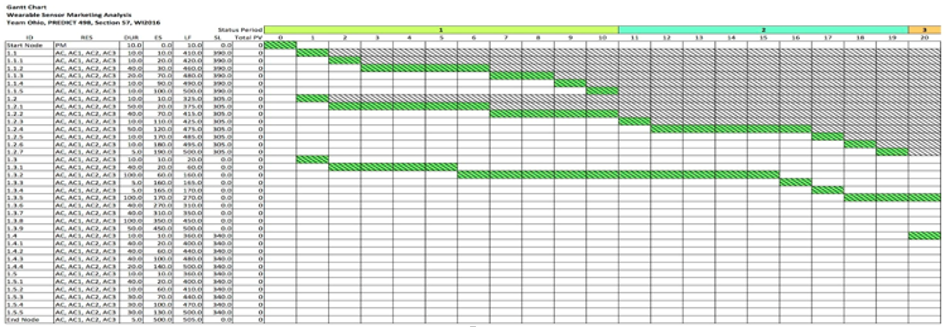
\*A typical installation with 15 to 20 users (including two or three analysts) would deploy Tableau with a perpetual license for $15,000 to $20,000 plus a 20 percent annual maintenance fee for upgrades and support. (tableau.com)

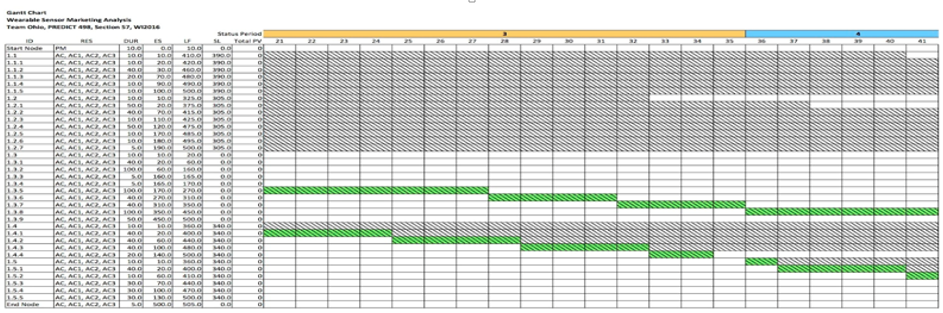
## **AON**

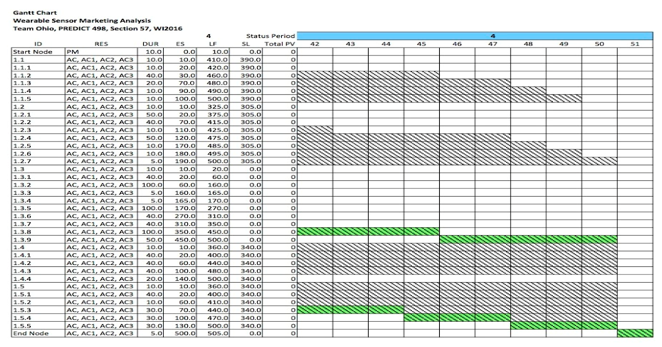




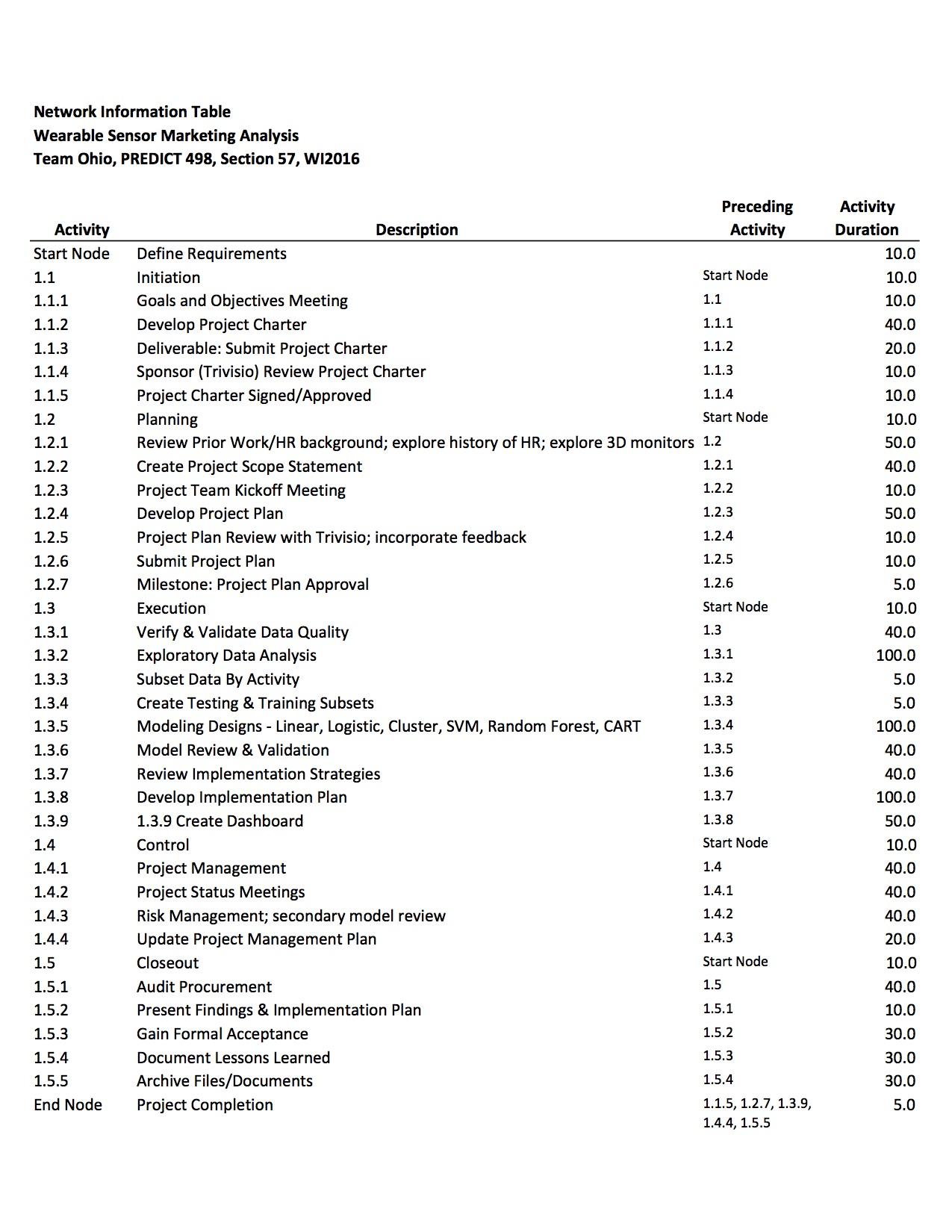
## **Gantt Chart**



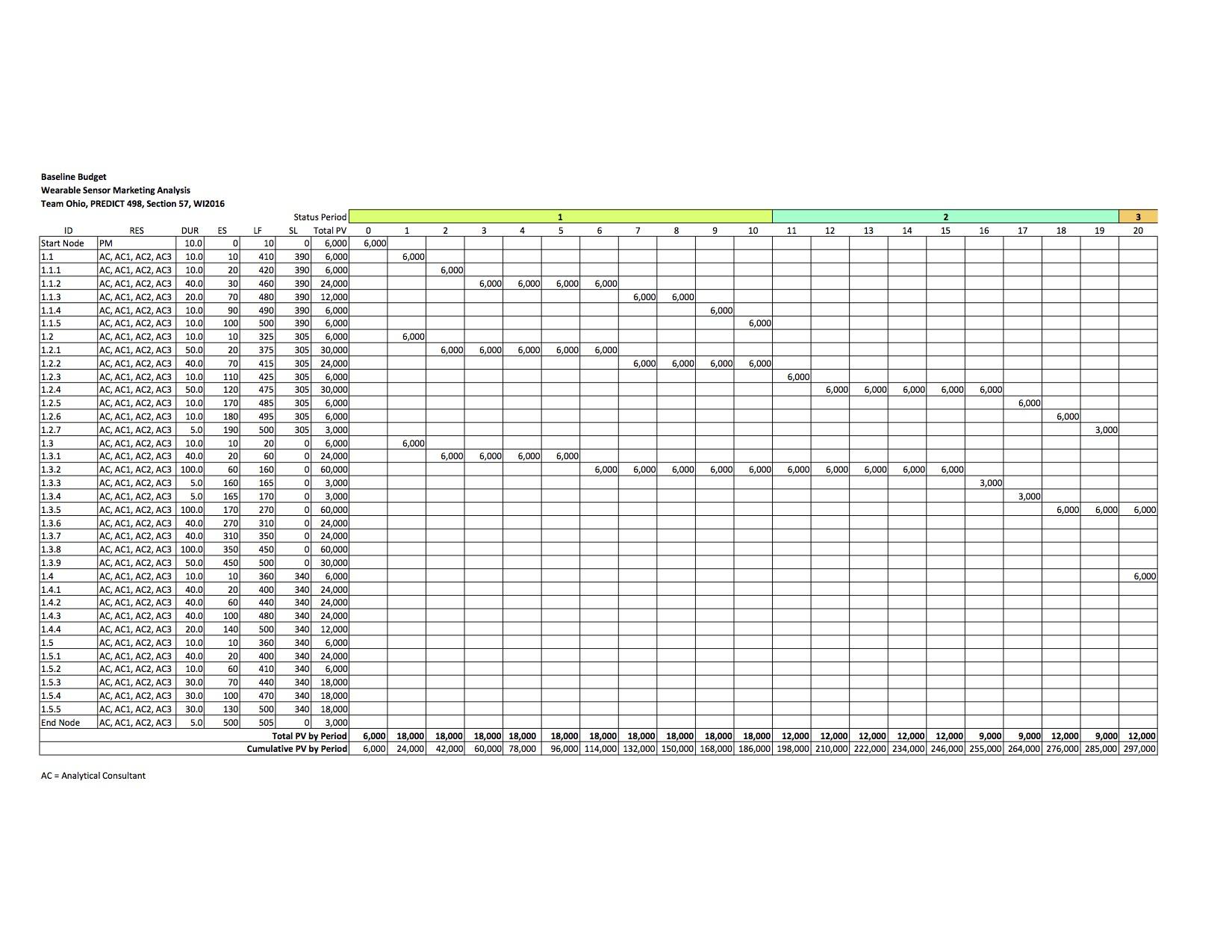


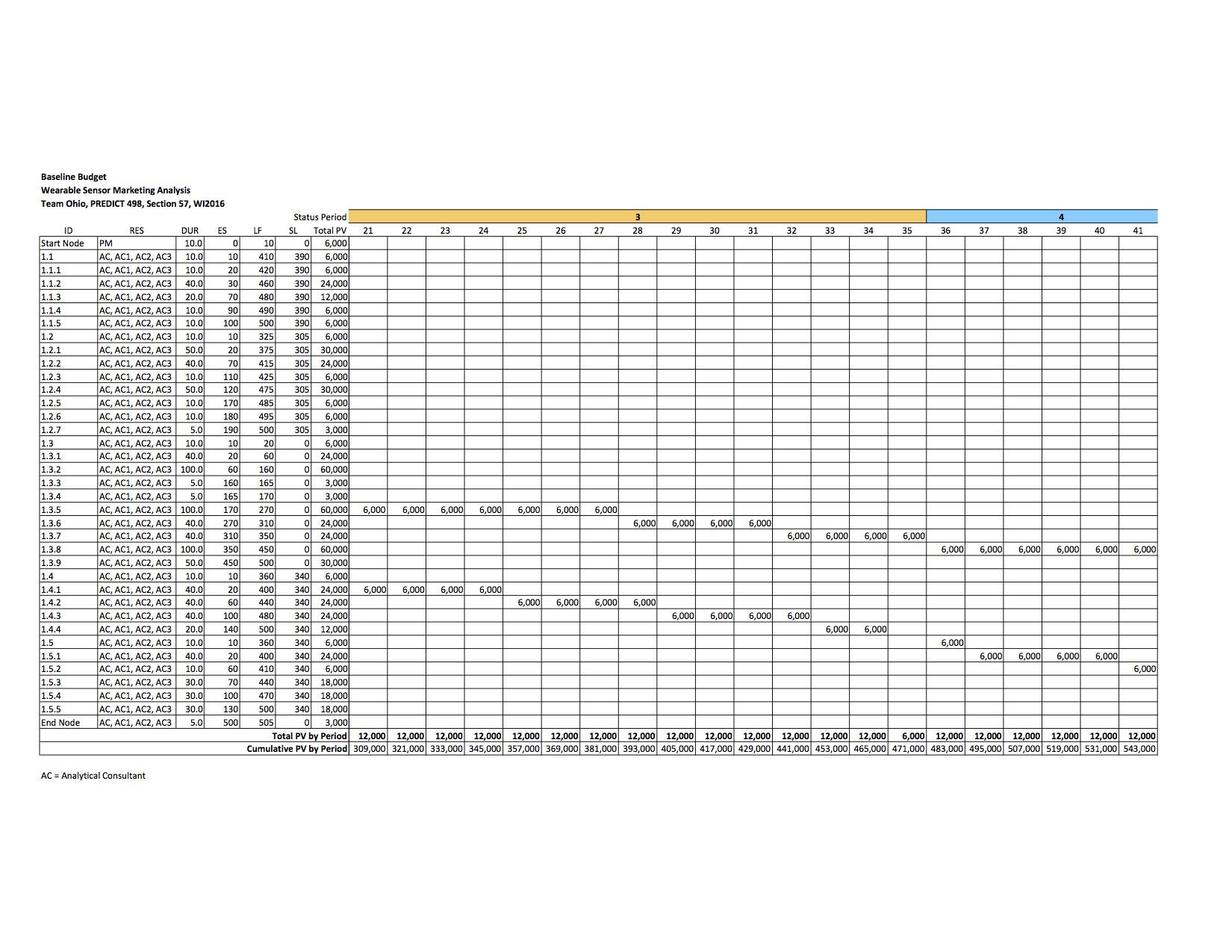


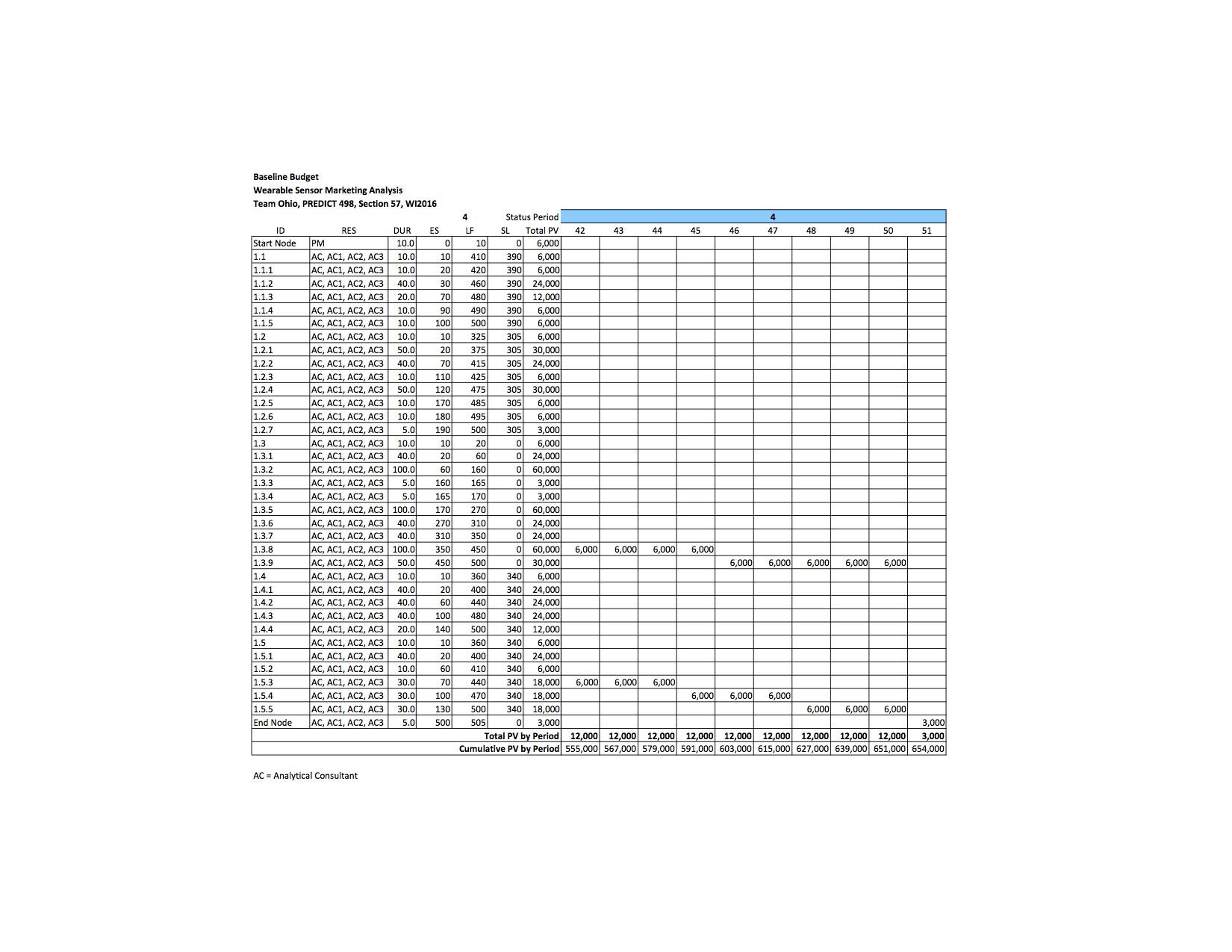
## **Network Information Table**



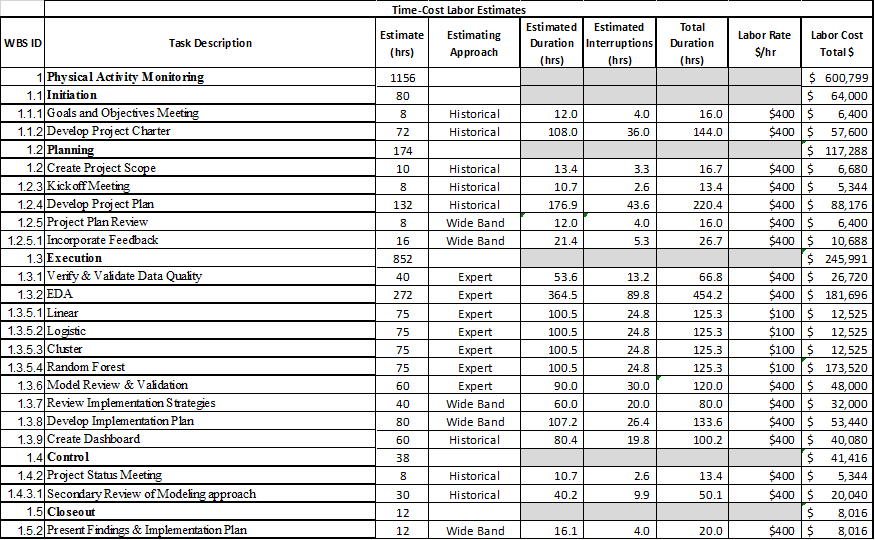
## **Schedule Baselines**



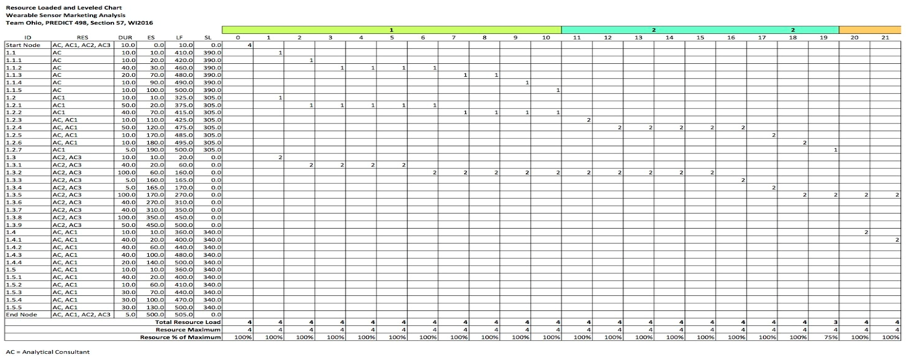


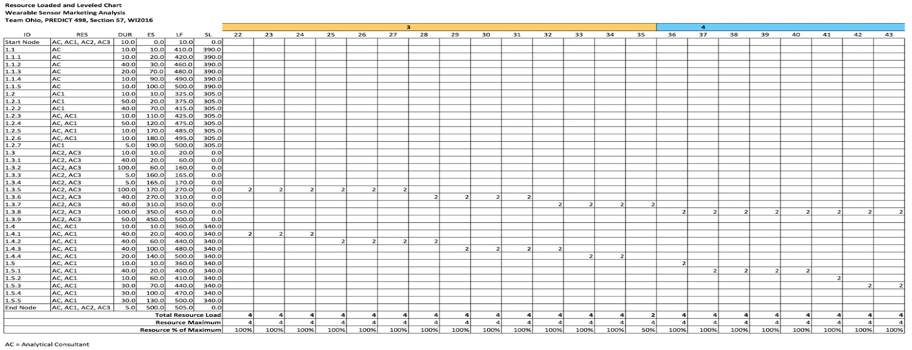


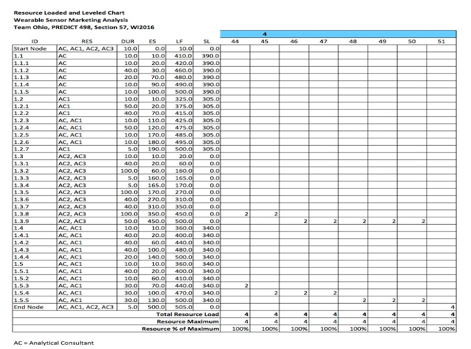
## **Cost**

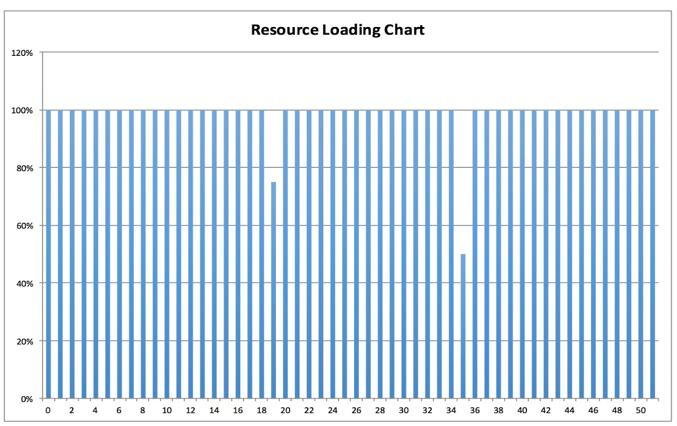


## **Resource Loading & Leveling**









## **Project Approval Guidelines**



## **Authorizations**

Acceptance

Approved by:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

John Doe

Trivisio Chief Analytics Officer (CAO)

# **Data Fields & Data Descriptions**

|  |  |  |
| --- | --- | --- |
|  | **Data Field** | **Data Description** |
| 1 | Time | The duration of the activity performed by the subject in seconds (s) |
| 2 | activityID | A unique identifier for each of the 18 activities in the study |
| 3 | HR | The heart rate of the subject in beats per minute (BPM) |
| 4 | HandTemp | Body temperature measured by the wrist IMU in degrees Celsius |
| 5 | HandAccel16.1 | Acceleration of the wrist IMU along the x-axis in meters per second squared (ms-2) |
| 6 | HandAccel16.2 | Acceleration of the wrist IMU along the y-axis in meters per second squared (ms-2) |
| 7 | HandAccel16.3 | Acceleration of the wrist IMU along the z-axis in meters per second squared (ms-2) |
| 8 | HandGyro.1 | Rotation of the wrist IMU about the x-axis in radians per second (rad/s) |
| 9 | HandGyro.2 | Rotation of the wrist IMU about the y-axis in radians per second (rad/s) |
| 10 | HandGyro.3 | Rotation of the wrist IMU about the z-axis in radians per second (rad/s) |
| 11 | HandMag.1 | Strength of the magnetic field measured by the wrist IMU along the x-axis in micro Teslas (μT) |
| 12 | HandMag.2 | Strength of the magnetic field measured by the wrist IMU along the y-axis in micro Teslas (μT) |
| 13 | HandMag.3 | Strength of the magnetic field measured by the wrist IMU along the z-axis in micro Teslas (μT) |
| 14 | ChestTemp | Body temperature measured by the chest IMU in degrees Celsius |
| 15 | ChestAccel16.1 | Acceleration of the chest IMU along the x-axis in meters per second squared (ms-2) |
| 16 | ChestAccel16.2 | Acceleration of the chest IMU along the y-axis in meters per second squared (ms-2) |
| 17 | ChestAccel16.3 | Acceleration of the chest IMU along the z-axis in meters per second squared (ms-2) |
| 18 | ChestGyro.1 | Rotation of the chest IMU about the x-axis in radians per second (rad/s) |
| 19 | ChestGyro.2 | Rotation of the chest IMU about the y-axis in radians per second (rad/s) |
| 20 | ChestGyro.3 | Rotation of the chest IMU about the z-axis in radians per second (rad/s) |
| 21 | ChestMag.1 | Strength of the magnetic field measured by the chest IMU along the x-axis in micro Teslas (μT) |
| 22 | ChestMag.2 | Strength of the magnetic field measured by the chest IMU along the y-axis in micro Teslas (μT) |
| 23 | ChestMag.3 | Strength of the magnetic field measured by the chest IMU along the z-axis in micro Teslas (μT) |
| 24 | AnkleTemp | Body temperature measured by the ankle IMU in degrees Celsius |
| 25 | AnkleAccel16.1 | Acceleration of the ankle IMU along the x-axis in meters per second squared (ms-2) |
| 26 | AnkleAccel16.2 | Acceleration of the ankle IMU along the y-axis in meters per second squared (ms-2) |
| 27 | AnkleAccel16.3 | Acceleration of the ankle IMU along the z-axis in meters per second squared (ms-2) |
| 28 | AnkleGyro.1 | Rotation of the ankle IMU about the x-axis in radians per second (rad/s) |
| 29 | AnkleGyro.2 | Rotation of the ankle IMU about the y-axis in radians per second (rad/s) |
| 30 | AnkleGyro.3 | Rotation of the ankle IMU about the z-axis in radians per second (rad/s) |
| 31 | AnkleMag.1 | Strength of the magnetic field measured by the ankle IMU along the x-axis in micro Teslas (μT) |
| 32 | AnkleMag.2 | Strength of the magnetic field measured by the ankle IMU along the y-axis in micro Teslas (μT) |
| 33 | AnkleMag.3 | Strength of the magnetic field measured by the ankle IMU along the z-axis in micro Teslas (μT) |

**Appendix B-PLS Component Loadings**

**Loadings:**

**Comp 1 Comp 2 Comp 3 Comp 4 Comp 5 Comp 6 Comp 7 Comp 8**

**Time 0.402 0.292 0.318 0.232 0.141**

**HandTemp -0.214 -0.456 0.385 -0.165 0.487 0.140**

**HandAccel16.1 -0.290 0.209 0.134 -0.219 0.152**

**HandAccel16.2 -0.144 -0.166 0.148**

**HandAccel16.3 -0.187 -0.169 -0.121 0.698 -0.170 -0.250 0.219**

**HandGyro.1 -0.177 0.229 -0.140 0.169**

**HandGyro.2 -0.393 0.207 0.111**

**HandGyro.3 -0.130 0.154 0.214 -0.480**

**HandMag.1 0.273 -0.224 0.137 -0.162 0.580 -0.310 -0.202**

**HandMag.2 -0.147 0.263 -0.204 -0.130 0.376 -0.291**

**HandMag.3 0.131 -0.235 0.479 -0.476 0.258 0.153 -0.220**

**ChestTemp -0.390 0.573 -0.106 -0.215 -0.402**

**ChestAccel16.1 0.113 -0.122 -0.155 0.133**

**ChestAccel16.2 0.189 -0.294 0.222 -0.138 -0.211**

**ChestAccel16.3 -0.359 0.168 0.119 0.147 -0.401**

**ChestGyro.1 0.160 -0.266 0.148 0.231**

**ChestGyro.2 -0.113 0.315 -0.166 0.133 -0.437**

**ChestGyro.3 -0.105 0.355 -0.336 0.247**

**ChestMag.1 -0.227 0.410 0.298**

**ChestMag.2 -0.309 0.172 -0.278 0.357 -0.319 -0.104 0.149**

**ChestMag.3 0.361 -0.186 0.107 -0.107 0.162 -0.219 0.241**

**AnkleTemp -0.355 0.478 0.159 -0.227 0.161**

**AnkleAccel16.1 0.246 -0.138 -0.207**

**AnkleAccel16.2 0.158 -0.262 0.198**

**AnkleAccel16.3 -0.222 0.229 0.343 -0.198 -0.144 0.136**

**AnkleGyro.1 -0.119 0.340 0.187 -0.361**

**AnkleGyro.2 -0.145**

**AnkleGyro.3 0.167 -0.423 0.296**

**AnkleMag.1 -0.221 -0.192 0.673 -0.159 -0.128**

**AnkleMag.2 -0.232 0.150 0.193 -0.286 0.111 -0.126**

**AnkleMag.3 -0.367 0.184 0.304 -0.380 0.138**

**Comp 1 Comp 2 Comp 3 Comp 4 Comp 5 Comp 6 Comp 7 Comp 8**

**SS loadings 1.084 1.288 1.237 1.332 1.577 1.835 1.260 1.244**

**Proportion Var 0.035 0.042 0.040 0.043 0.051 0.059 0.041 0.040**

**Cumulative Var 0.035 0.077 0.116 0.159 0.210 0.269 0.310 0.350**

# **Appendix B**

## **R Code**

# CAPSTONE - PAM

# TEAM OHIO

# PRED 498 - WINTER 2016

setwd("/Users/Scott/Dropbox/Northwestern/PRED 498 - Capstone/Project")

getwd() # Verify working directory

data <- read.csv("PAM\_Master.csv", sep=",", header=TRUE)

sum(is.na(data))# cHECK NUMBER MISSING VARS

# Load Packages

library(lattice) # Graphics

library(plyr) # Graphics

library(rpart) # Tree

library(ggplot2) # Graphics

library(gbm) # Generalized Boosted Regression Models

library(adabag) # Classification with Boosting and Bagging

library(randomForest) # Random Forest

library(e1071) # Support Vector Machine (SVM)

library(rpart.plot) # Clean rpart plots

library(pastecs) # 'Fancy' summary statistics

library(Hmisc)

library(stats)

library(dplyr)

library(rpart)

library(tree)

library(rattle)

library(aod)

library(caret)

library(sensitivity)

library(pROC)

library(lmtest)

library(pscl)

library(Amelia)

library(gains)

library(lift)

# EDA

class(data)

names(data)

str(data)

summary(data)

stat.desc(data)

# Recode Activity and Subjects as Factor

data$activityID <- as.factor(data$activityID)

data$Subject <- as.factor(data$Subject)

# Coded HR binary separation at .1 above the median for the 9 subjects

HR.Median = 0

HR.Median <- ifelse(data$HR >=104.1, HR.Median+1, HR.Median)

data <- cbind(data, HR.Median)

summary(data)

missmap(data)

# Subset Activities

# Lying, Sitting, Standing, Watching TV, Comp Work, Car Driving

Daily <- subset(data, activityID=="1"|activityID=="2"|activityID=="3"|

activityID=="9"|activityID=="10"|activityID=="11")

summary(Daily)

# Descending Stairs, Walking, Vacuum Cleaning, Ironing, Folding Laundry, House Cleaning

Chores <- subset(data, activityID=="4"|activityID=="13"|activityID=="16"|

activityID=="17"|activityID=="18"|activityID=="19")

summary(Chores)

# Running, Cycling, Nordic Walking, Ascending Stairs, Playing Soccer, Rope Jumping

Intensity <- subset(data, activityID=="5"|activityID=="6"|activityID=="7"|

activityID=="12"|activityID=="20"|activityID=="24")

summary(Intensity)

# HR.Median means

# Daily mean = .02883

# Chores mean = .4762

# Intensity mean = .9568

# Daily Boxplots

boxplot(Time ~ HR.Median, data=Daily, xlab="HR Category", ylab="Time", main="Daily Subset Boxplots")

#Temp

boxplot(HandTemp ~ HR.Median, data=Daily, xlab="HR Category", ylab="Hand Temperature", main="Daily Subset Boxplots")

boxplot(ChestTemp ~ HR.Median, data=Daily, xlab="HR Category", ylab="Chest Temperature", main="Daily Subset Boxplots")

boxplot(AnkleTemp ~ HR.Median, data=Daily, xlab="HR Category", ylab="Ankle Temperature", main="Daily Subset Boxplots")

#Accel

boxplot(HandAccel16.1 ~ HR.Median, data=Daily, xlab="HR Category", ylab="HandAccel16.1", main="Daily Subset Boxplots") #Min separation

boxplot(HandAccel16.2 ~ HR.Median, data=Daily, xlab="HR Category", ylab="HandAccel16.2", main="Daily Subset Boxplots") #Many outliers

boxplot(HandAccel16.3 ~ HR.Median, data=Daily, xlab="HR Category", ylab="HandAccel16.3", main="Daily Subset Boxplots") #best of 3?

boxplot(ChestAccel16.1 ~ HR.Median, data=Daily, xlab="HR Category", ylab="ChestAccel16.1", main="Daily Subset Boxplots") #Many outliers

boxplot(ChestAccel16.2 ~ HR.Median, data=Daily, xlab="HR Category", ylab="ChestAccel16.2", main="Daily Subset Boxplots") #clear separation

boxplot(ChestAccel16.3 ~ HR.Median, data=Daily, xlab="HR Category", ylab="ChestAccel16.3", main="Daily Subset Boxplots") #Good separation

boxplot(AnkleAccel16.1 ~ HR.Median, data=Daily, xlab="HR Category", ylab="AnkleAccel16.1", main="Daily Subset Boxplots") #OK separation

boxplot(AnkleAccel16.2 ~ HR.Median, data=Daily, xlab="HR Category", ylab="AnkleAccel16.2", main="Daily Subset Boxplots") #min separation

boxplot(AnkleAccel16.3 ~ HR.Median, data=Daily, xlab="HR Category", ylab="AnkleAccel16.3", main="Daily Subset Boxplots") #min separation

#Gyro

boxplot(HandGyro.1 ~ HR.Median, data=Daily, xlab="HR Category", ylab="HandGyro1", main="Daily Subset Boxplots") #many outliers

boxplot(HandGyro.2 ~ HR.Median, data=Daily, xlab="HR Category", ylab="HandGyro2", main="Daily Subset Boxplots") #many outliers

boxplot(HandGyro.3 ~ HR.Median, data=Daily, xlab="HR Category", ylab="HandGyro3", main="Daily Subset Boxplots") #many outliers

boxplot(ChestGyro.1 ~ HR.Median, data=Daily, xlab="HR Category", ylab="ChestGyro1", main="Daily Subset Boxplots") #many outliers

boxplot(ChestGyro.2 ~ HR.Median, data=Daily, xlab="HR Category", ylab="ChestGyro2", main="Daily Subset Boxplots") #many outliers

boxplot(ChestGyro.3 ~ HR.Median, data=Daily, xlab="HR Category", ylab="ChestGyro3", main="Daily Subset Boxplots") #many outliers

boxplot(AnkleGyro.1 ~ HR.Median, data=Daily, xlab="HR Category", ylab="AnkleGyro1", main="Daily Subset Boxplots") #many outliers

boxplot(AnkleGyro.2 ~ HR.Median, data=Daily, xlab="HR Category", ylab="AnkleGyro2", main="Daily Subset Boxplots") #many outliers

boxplot(AnkleGyro.3 ~ HR.Median, data=Daily, xlab="HR Category", ylab="AnkleGyro3", main="Daily Subset Boxplots") #many outliers

#Magnet

boxplot(HandMag.1 ~ HR.Median, data=Daily, xlab="HR Category", ylab="HandMag1", main="Daily Subset Boxplots") #inverse

boxplot(HandMag.2 ~ HR.Median, data=Daily, xlab="HR Category", ylab="HandMag2", main="Daily Subset Boxplots") #inverse

boxplot(HandMag.3 ~ HR.Median, data=Daily, xlab="HR Category", ylab="HandMag3", main="Daily Subset Boxplots") #inverse

boxplot(ChestMag.1 ~ HR.Median, data=Daily, xlab="HR Category", ylab="ChestMag1", main="Daily Subset Boxplots") #min separation

boxplot(ChestMag.2 ~ HR.Median, data=Daily, xlab="HR Category", ylab="ChestMag2", main="Daily Subset Boxplots") #min separation

boxplot(ChestMag.3 ~ HR.Median, data=Daily, xlab="HR Category", ylab="ChestMag3", main="Daily Subset Boxplots") #min separation

boxplot(AnkleMag.1 ~ HR.Median, data=Daily, xlab="HR Category", ylab="AnkleMag1", main="Daily Subset Boxplots")

boxplot(AnkleMag.2 ~ HR.Median, data=Daily, xlab="HR Category", ylab="AnkleMag2", main="Daily Subset Boxplots")

boxplot(AnkleMag.3 ~ HR.Median, data=Daily, xlab="HR Category", ylab="AnkleMag3", main="Daily Subset Boxplots")

# Daily Histograms

hist(Daily$Time)

hist(Daily$HR)

#Temp

hist(Daily$HandTemp)

hist(Daily$ChestTemp)

hist(Daily$AnkleTemp)

#Accel

hist(Daily$HandAccel16.1)

hist(Daily$HandAccel16.2)

hist(Daily$HandAccel16.3)

hist(Daily$ChestAccel16.1)

hist(Daily$ChestAccel16.2)

hist(Daily$ChestAccel16.3)

hist(Daily$AnkleAccel16.1)

hist(Daily$AnkleAccel16.2)

hist(Daily$AnkleAccel16.3)

#Gyro

hist(Daily$HandGyro.1)

hist(Daily$HandGyro.2)

hist(Daily$HandGyro.3)

hist(Daily$ChestGyro.1)

hist(Daily$ChestGyro.2)

hist(Daily$ChestGyro.3)

hist(Daily$AnkleGyro.1)

hist(Daily$AnkleGyro.2)

hist(Daily$AnkleGyro.3)

#Magnet

hist(Daily$HandMag.1)

hist(Daily$HandMag.2)

hist(Daily$HandMag.3)

hist(Daily$ChestMag.1)

hist(Daily$ChestMag.2)

hist(Daily$ChestMag.3)

hist(Daily$AnkleMag.1)

hist(Daily$AnkleMag.2)

hist(Daily$AnkleMag.3)

# Intensity Boxplots

boxplot(Time ~ HR.Median, data=Intensity, xlab="HR Category", ylab="Time", main="Intensity Subset Boxplots")

#Temp

boxplot(HandTemp ~ HR.Median, data=Intensity, xlab="HR Category", ylab="Hand Temperature", main="Intensity Subset Boxplots")

boxplot(ChestTemp ~ HR.Median, data=Intensity, xlab="HR Category", ylab="Chest Temperature", main="Intensity Subset Boxplots")

boxplot(AnkleTemp ~ HR.Median, data=Intensity, xlab="HR Category", ylab="Ankle Temperature", main="Intensity Subset Boxplots")

#Accel

boxplot(HandAccel16.1 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="HandAccel16.1", main="Intensity Subset Boxplots")

boxplot(HandAccel16.2 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="HandAccel16.2", main="Intensity Subset Boxplots")

boxplot(HandAccel16.3 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="HandAccel16.3", main="Intensity Subset Boxplots")

boxplot(ChestAccel16.1 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="ChestAccel16.1", main="Intensity Subset Boxplots")

boxplot(ChestAccel16.2 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="ChestAccel16.2", main="Intensity Subset Boxplots")

boxplot(ChestAccel16.3 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="ChestAccel16.3", main="Intensity Subset Boxplots")

boxplot(AnkleAccel16.1 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="AnkleAccel16.1", main="Intensity Subset Boxplots")

boxplot(AnkleAccel16.2 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="AnkleAccel16.2", main="Intensity Subset Boxplots")

boxplot(AnkleAccel16.3 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="AnkleAccel16.3", main="Intensity Subset Boxplots")

#Gyro

boxplot(HandGyro.1 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="HandGyro1", main="Intensity Subset Boxplots")

boxplot(HandGyro.2 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="HandGyro2", main="Intensity Subset Boxplots")

boxplot(HandGyro.3 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="HandGyro3", main="Intensity Subset Boxplots")

boxplot(ChestGyro.1 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="ChestGyro1", main="Intensity Subset Boxplots")

boxplot(ChestGyro.2 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="ChestGyro2", main="Intensity Subset Boxplots")

boxplot(ChestGyro.3 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="ChestGyro3", main="Intensity Subset Boxplots")

boxplot(AnkleGyro.1 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="AnkleGyro1", main="Intensity Subset Boxplots")

boxplot(AnkleGyro.2 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="AnkleGyro2", main="Intensity Subset Boxplots")

boxplot(AnkleGyro.3 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="AnkleGyro3", main="Intensity Subset Boxplots")

#Magnet

boxplot(HandMag.1 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="HandMag1", main="Intensity Subset Boxplots")

boxplot(HandMag.2 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="HandMag2", main="Intensity Subset Boxplots")

boxplot(HandMag.3 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="HandMag3", main="Intensity Subset Boxplots")

boxplot(ChestMag.1 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="ChestMag1", main="Intensity Subset Boxplots")

boxplot(ChestMag.2 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="ChestMag2", main="Intensity Subset Boxplots")

boxplot(ChestMag.3 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="ChestMag3", main="Intensity Subset Boxplots")

boxplot(AnkleMag.1 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="AnkleMag1", main="Intensity Subset Boxplots")

boxplot(AnkleMag.2 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="AnkleMag2", main="Intensity Subset Boxplots")

boxplot(AnkleMag.3 ~ HR.Median, data=Intensity, xlab="HR Category", ylab="AnkleMag3", main="Intensity Subset Boxplots")

# Intensity Histograms

hist(Intensity$Time)

hist(Intensity$HR)

#Temp

hist(Intensity$HandTemp)

hist(Intensity$ChestTemp)

hist(Intensity$AnkleTemp)

#Accel

hist(Intensity$HandAccel16.1)

hist(Intensity$HandAccel16.2)

hist(Intensity$HandAccel16.3)

hist(Intensity$ChestAccel16.1)

hist(Intensity$ChestAccel16.2)

hist(Intensity$ChestAccel16.3)

hist(Intensity$AnkleAccel16.1)

hist(Intensity$AnkleAccel16.2)

hist(Intensity$AnkleAccel16.3)

#Gyro

hist(Intensity$HandGyro.1)

hist(Intensity$HandGyro.2)

hist(Intensity$HandGyro.3)

hist(Intensity$ChestGyro.1)

hist(Intensity$ChestGyro.2)

hist(Intensity$ChestGyro.3)

hist(Intensity$AnkleGyro.1)

hist(Intensity$AnkleGyro.2)

hist(Intensity$AnkleGyro.3)

#Magnet

hist(Intensity$HandMag.1)

hist(Intensity$HandMag.2)

hist(Intensity$HandMag.3)

hist(Intensity$ChestMag.1)

hist(Intensity$ChestMag.2)

hist(Intensity$ChestMag.3)

hist(Intensity$AnkleMag.1)

hist(Intensity$AnkleMag.2)

hist(Intensity$AnkleMag.3)

# BOXPLOTS (LATTICE)

boxplot(HR ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of HR by Activity")

boxplot(HandTemp ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of Hand Temp by Activity")

boxplot(HandAccel16.1 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of HA16-1 by Activity")

boxplot(HandAccel16.2 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of HA16-2 by Activity")

boxplot(HandAccel16.3 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of HA16-3 by Activity")

boxplot(HandGyro.1 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of HandGyro1 by Activity")

boxplot(HandGyro.2 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of HandGyro2 by Activity")

boxplot(HandGyro.3 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of HandGyro3 by Activity")

boxplot(HandMag.1 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of HandMag1 by Activity")

boxplot(HandMag.2 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of HandMag2 by Activity")

boxplot(HandMag.3 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of HandMag3 by Activity")

boxplot(AnkleTemp ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of Ankle Temp by Activity")

boxplot(AnkleAccel16.1 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of CA16-1 by Activity")

boxplot(AnkleAccel16.2 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of CA16-2 by Activity")

boxplot(AnkleAccel16.3 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of CA16-3 by Activity")

boxplot(AnkleGyro.1 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of AnkleGyro1 by Activity")

boxplot(AnkleGyro.2 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of AnkleGyro2 by Activity")

boxplot(AnkleGyro.3 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of AnkleGyro3 by Activity")

boxplot(AnkleMag.1 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of AnkleMag1 by Activity")

boxplot(AnkleMag.2 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of AnkleMag2 by Activity")

boxplot(AnkleMag.3 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of AnkleMag3 by Activity")

boxplot(AnkleTemp ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of Ankle Temp by Activity")

boxplot(AnkleAccel16.1 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of AnkleA16-1 by Activity")

boxplot(AnkleAccel16.2 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of AnkleA16-2 by Activity")

boxplot(AnkleAccel16.3 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of AnkleA16-3 by Activity")

boxplot(AnkleGyro.1 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of AnkleGyro1 by Activity")

boxplot(AnkleGyro.2 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of AnkleGyro2 by Activity")

boxplot(AnkleGyro.3 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of AnkleGyro3 by Activity")

boxplot(AnkleMag.1 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of AnkleMag1 by Activity")

boxplot(AnkleMag.2 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of AnkleMag2 by Activity")

boxplot(AnkleMag.3 ~ activityID, data=data, xlab = "Activity",

main = "Boxplot of AnkleMag3 by Activity")

# HISTOGRAMS

hist(HR, xlab = "Heart Rate", main = "Heart Rate Distribution")

hist(HandTemp, xlab = "HandTemp", main = "HandTemp Distribution")

hist(HandAccel16.1, xlab = "HandAccel16.1", main = "HandAccel16.1 Distribution")

hist(HandAccel16.2, xlab = "HandAccel16.2", main = "HandAccel16.2 Distribution")

hist(HandAccel16.3, xlab = "HandAccel16.3", main = "HandAccel16.3 Distribution")

# VARIABLE MEANS BY ACTIVITY

# HR Mean by ACTIVITY

aggregate(data[, 3,2], list(activityID=data$activityID), mean)

# HandTemp Mean by ACTIVITY

aggregate(data[, 4,2], list(activityID=data$activityID), mean)

# HandAccel16.1 Mean by CT

aggregate(data[, 5,2], list(activityID=data$activityID), mean)

# HandAccel16.2 Mean by ACTIVITY

aggregate(data[, 6,2], list(activityID=data$activityID), mean)

# HandAccel16.3 Mean by ACTIVITY

aggregate(data[, 7,2], list(activityID=data$activityID), mean)

# HandGyro.1 Mean by ACTIVITY

aggregate(data[, 8,2], list(activityID=data$activityID), mean)

# HandGyro.2 Mean by ACTIVITY

aggregate(data[, 9,2], list(activityID=data$activityID), mean)

# HandGyro.3 Mean by ACTIVITY

aggregate(data[, 10,2], list(activityID=data$activityID), mean)

# HandMag.1 Mean by ACTIVITY

aggregate(data[, 11,2], list(activityID=data$activityID), mean)

# HandMag.2 Mean by ACTIVITY

aggregate(data[, 12,2], list(activityID=data$activityID), mean)

# HandMag.3 Mean by ACTIVITY

aggregate(data[, 13,2], list(activityID=data$activityID), mean)

# ChestTemp Mean by ACTIVITY

aggregate(data[, 14,2], list(activityID=data$activityID), mean)

# ChestAccel16.1 Mean by ACTIVITY

aggregate(data[, 15,2], list(activityID=data$activityID), mean)

# ChestAccel16.2 Mean by ACTIVITY

aggregate(data[, 16,2], list(activityID=data$activityID), mean)

# ChestAccel16.3 Mean by ACTIVITY

aggregate(data[, 17,2], list(activityID=data$activityID), mean)

# ChestGyro.1 Mean by ACTIVITY

aggregate(data[, 18,2], list(activityID=data$activityID), mean)

# ChestGyro.2 Mean by ACTIVITY

aggregate(data[, 19,2], list(activityID=data$activityID), mean)

# ChestGyro.3 Mean by ACTIVITY

aggregate(data[, 20,2], list(activityID=data$activityID), mean)

# ChestMag.1 Mean by ACTIVITY

aggregate(data[, 21,2], list(activityID=data$activityID), mean)

# ChestMag.2 Mean by ACTIVITY

aggregate(data[, 22,2], list(activityID=data$activityID), mean)

# ChestMag.3 Mean by ACTIVITY

aggregate(data[, 23,2], list(activityID=data$activityID), mean)

# AnkleTemp Mean by ACTIVITY

aggregate(data[, 24,2], list(activityID=data$activityID), mean)

# AnkleAccel.1 Mean by ACTIVITY

aggregate(data[, 25,2], list(activityID=data$activityID), mean)

# AnkleAccel.2 Mean by ACTIVITY

aggregate(data[, 26,2], list(activityID=data$activityID), mean)

# AnkleAccel.3 Mean by ACTIVITY

aggregate(data[, 27,2], list(activityID=data$activityID), mean)

# AnkleGyro.1 Mean by ACTIVITY

aggregate(data[, 28,2], list(activityID=data$activityID), mean)

# AnkleGyro.2 Mean by ACTIVITY

aggregate(data[, 29,2], list(activityID=data$activityID), mean)

# AnkleGyro.3 Mean by ACTIVITY

aggregate(data[, 30,2], list(activityID=data$activityID), mean)

# AnkleMag.1 Mean by ACTIVITY

aggregate(data[, 31,2], list(activityID=data$activityID), mean)

# AnkleMag.2 Mean by ACTIVITY

aggregate(data[, 32,2], list(activityID=data$activityID), mean)

# AnkleMag.3 Mean by ACTIVITY

aggregate(data[, 33,2], list(activityID=data$activityID), mean)

# Correlation Matrix

data\_corr <- data[4:33]

cor(data\_corr, data\_corr)

# Lattice Plots

# Dependent ~ HR

# 1

xyplot(HR ~ HandTemp | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "HandTemp",

ylab = "Heart Rate",

main = "Plot of Hand Temp against Heart Rate

by Activity")

xyplot(HR ~ HandAccel16.1 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "HandAccel16.1",

ylab = "Heart Rate",

main = "Plot of HandAccel16.1 against Heart Rate

by Activity")

xyplot(HR ~ HandAccel16.2 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "HandAccel16.2",

ylab = "Heart Rate",

main = "Plot of HandAccel16.2 against Heart Rate

by Activity")

xyplot(HR ~ HandAccel16.3 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "HandAccel16.3",

ylab = "Heart Rate",

main = "Plot of HandAccel16.3 against Heart Rate

by Activity")

xyplot(HR ~ HandGyro.1 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "HandGyro.1",

ylab = "Heart Rate",

main = "Plot of HandGyro.1 against Heart Rate

by Activity")

xyplot(HR ~ HandGyro.2 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "HandGyro.2",

ylab = "Heart Rate",

main = "Plot of HandGyro.2 against Heart Rate

by Activity")

xyplot(HR ~ HandGyro.3 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "HandGyro.3",

ylab = "Heart Rate",

main = "Plot of HandGyro.3 against Heart Rate

by Activity")

xyplot(HR ~ HandMag.1 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "HandMag.1",

ylab = "Heart Rate",

main = "Plot of HandMag.1 against Heart Rate

by Activity")

xyplot(HR ~ HandMag.2 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "HandMag.2",

ylab = "Heart Rate",

main = "Plot of HandMag.2 against Heart Rate

by Activity")

xyplot(HR ~ HandMag.3 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "HandMag.3",

ylab = "Heart Rate",

main = "Plot of HandMag.3 against Heart Rate

by Activity")

# CHEST

xyplot(HR ~ ChestTemp | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "ChestTemp",

ylab = "Heart Rate",

main = "Plot of Chest Temp against Heart Rate

by Activity")

xyplot(HR ~ ChestAccel16.1 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "ChestAccel16.1",

ylab = "Heart Rate",

main = "Plot of ChestAccel16.1 against Heart Rate

by Activity")

xyplot(HR ~ ChestAccel16.2 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "ChestAccel16.2",

ylab = "Heart Rate",

main = "Plot of ChestAccel16.2 against Heart Rate

by Activity")

xyplot(HR ~ ChestAccel16.3 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "ChestAccel16.3",

ylab = "Heart Rate",

main = "Plot of ChestAccel16.3 against Heart Rate

by Activity")

xyplot(HR ~ ChestGyro.1 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "ChestGyro.1",

ylab = "Heart Rate",

main = "Plot of ChestGyro.1 against Heart Rate

by Activity")

xyplot(HR ~ ChestGyro.2 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "ChestGyro.2",

ylab = "Heart Rate",

main = "Plot of ChestGyro.2 against Heart Rate

by Activity")

xyplot(HR ~ ChestGyro.3 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "ChestGyro.3",

ylab = "Heart Rate",

main = "Plot of ChestGyro.3 against Heart Rate

by Activity")

xyplot(HR ~ ChestMag.1 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "ChestMag.1",

ylab = "Heart Rate",

main = "Plot of ChestMag.1 against Heart Rate

by Activity")

xyplot(HR ~ ChestMag.2 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "ChestMag.2",

ylab = "Heart Rate",

main = "Plot of ChestMag.2 against Heart Rate

by Activity")

xyplot(HR ~ ChestMag.3 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "ChestMag.3",

ylab = "Heart Rate",

main = "Plot of ChestMag.3 against Heart Rate

by Activity")

# ANKLE

xyplot(HR ~ AnkleTemp | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "AnkleTemp",

ylab = "Heart Rate",

main = "Plot of Ankle Temp against Heart Rate

by Activity")

xyplot(HR ~ AnkleAccel16.1 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "AnkleAccel16.1",

ylab = "Heart Rate",

main = "Plot of AnkleAccel16.1 against Heart Rate

by Activity")

xyplot(HR ~ AnkleAccel16.2 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "AnkleAccel16.2",

ylab = "Heart Rate",

main = "Plot of AnkleAccel16.2 against Heart Rate

by Activity")

xyplot(HR ~ AnkleAccel16.3 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "AnkleAccel16.3",

ylab = "Heart Rate",

main = "Plot of AnkleAccel16.3 against Heart Rate

by Activity")

xyplot(HR ~ AnkleGyro.1 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "AnkleGyro.1",

ylab = "Heart Rate",

main = "Plot of AnkleGyro.1 against Heart Rate

by Activity")

xyplot(HR ~ AnkleGyro.2 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "AnkleGyro.2",

ylab = "Heart Rate",

main = "Plot of AnkleGyro.2 against Heart Rate

by Activity")

xyplot(HR ~ AnkleGyro.3 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "AnkleGyro.3",

ylab = "Heart Rate",

main = "Plot of AnkleGyro.3 against Heart Rate

by Activity")

xyplot(HR ~ AnkleMag.1 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "AnkleMag.1",

ylab = "Heart Rate",

main = "Plot of AnkleMag.1 against Heart Rate

by Activity")

xyplot(HR ~ AnkleMag.2 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "AnkleMag.2",

ylab = "Heart Rate",

main = "Plot of AnkleMag.2 against Heart Rate

by Activity")

xyplot(HR ~ AnkleMag.3 | activityID,

data=data, as.table = TRUE, pch = 50,

panel = function(x,y,...){

panel.xyplot(x,y,...)

fit <- lm(y ~ x)

panel.abline(fit)

},

xlab = "AnkleMag.3",

ylab = "Heart Rate",

main = "Plot of AnkleMag.3 against Heart Rate

by Activity")

# Randomize observations and split into train/test groups

sample\_size <- floor(.70\* nrow(data))

set.seed(498)

int <- sample(seq\_len(nrow(data)),size=sample\_size)

train <- data[int,]

test <- data[-int,]

# EDA TREE WITH RANDOMIZED DATA

# Shuffle the dataset

set.seed(1234)

g <- runif(nrow(data))

dataR <- data[order(g),]

sample\_size <- floor(.70\* nrow(dataR))

set.seed(498)

int <- sample(seq\_len(nrow(data)),size=sample\_size)

trainR <- data[int,]

testR <- data[-int,]

# Clean Up

rm(sample\_size)

rm(dataR)

rm(g)

length(trainR)

length(trainR$activityID)

dim(trainR)

names(trainR)

# All variables considered

PAM.rpart <- rpart(activityID ~ trainR[,3:33],

data = trainR,

method = "class")

plotcp(PAM.rpart)

printcp(PAM.rpart)

PAM.rpart2 <- prune(PAM.rpart, cp = 0.013)

plot(PAM.rpart2, uniform = TRUE, margin = .2)

text(PAM.rpart2, use.n = TRUE, cex = 0.75)

prp(PAM.rpart2, extra=1, uniform=F, type=3)

#-------------------------------------------------------/

# LINEAR REGRESSION

#-------------------------------------------------------/

library(MASS)

# Working Models

fit <- lm(HR~.-c(Subject+Time+HR.Median),

data=trainR)

fit$coefficients

# Forward Selection

forwardfit<- stepAIC(fit, direction="forward")

summary(forwardfit) # show results

# Backward Selection

backfit<- stepAIC(fit, direction="backward")

summary(backfit) # show results

# Stepwise Selection

stepfit<- stepAIC(fit, direction="both")

summary(stepfit) # show results

#-------------------------------------------------------/

# RANDOM FOREST

#-------------------------------------------------------/

# Using randomized train sample

# Strip out SubjectID

library(pROC)

rf.forest =randomForest(as.factor(activityID)~.-(Subject+HR.Median),

data=trainR,

mtry=10,

importance =TRUE,

type="classification")

# IN-SAMPLE STATISTICS

# Accuracy: 99.94%

rf.train.cm <- rf.forest$confusion

sum(diag(rf.train.cm))/sum(rf.train.cm)

# VARIABLE IMPORTANCE

varImpPlot(rf.forest, sort=TRUE, n.var=min(30, nrow(rf.forest$importance)),

type=NULL, class=NULL, scale=TRUE,

main=deparse(substitute(x)))

rf.forest$importance

RF.Imp <- rf.forest$importance

write.csv(RF.Imp, file = "RF.Imp.csv",row.names=TRUE)

# OUT-OF-SAMPLE EVALUATION

# Accuracy: 99.94%

yhat.rf.forest <- predict(rf.forest,newdata=testR)

rfconfusion <- table(yhat.rf.forest,testR$activityID)

sum(diag(rfconfusion))/sum(rfconfusion)

summary(yhat.rf.forest)

# --------------------- REVERSE MODELS

reverse.forest =randomForest(as.factor(activityID)~.-(Subject+HR.Median),

data=testR,

mtry=10,

importance =TRUE,

type="classification")

# IN-SAMPLE STATISTICS

# Accuracy: 99.84%

reverse.train.cm <- reverse.forest$confusion

sum(diag(reverse.train.cm))/sum(reverse.train.cm)

# VARIABLE IMPORTANCE

varImpPlot(reverse.forest, sort=TRUE, n.var=min(30, nrow(reverse.forest$importance)),

type=NULL, class=NULL, scale=TRUE,

main=deparse(substitute(x)))

reverse.forest$importance

# OUT-OF-SAMPLE EVALUATION

# Accuracy: 99.84%

yhat.reverse.forest <- predict(reverse.forest,newdata=trainR)

reverse.confusion <- table(yhat.reverse.forest,trainR$activityID)

sum(diag(reverse.confusion))/sum(reverse.confusion)

summary(yhat.reverse.forest)

# ---------------------

# SENSOR-SPECIFIC MODELS

# ---------------------

# HAND SENSOR

Hand.forest =randomForest(as.factor(activityID)~HandTemp+

HandAccel16.1+

HandAccel16.2+

HandAccel16.3+

HandGyro.1+

HandGyro.2+

HandGyro.3+

HandMag.1+

HandMag.2+

HandMag.3,

data=trainR,

mtry=10,

importance =TRUE,

type="classification")

# IN-SAMPLE STATISTICS

# Accuracy: 92.99%

Hand.train.cm <- Hand.forest$confusion

sum(diag(Hand.train.cm))/sum(Hand.train.cm)

# VARIABLE IMPORTANCE

varImpPlot(Hand.forest, sort=TRUE, n.var=min(30, nrow(Hand.forest$importance)),

type=NULL, class=NULL, scale=TRUE,

main=deparse(substitute(x)))

Hand.RF.Imp <- Hand.forest$importance

write.csv(Hand.RF.Imp, file = "Hand.RF.Imp.csv",row.names=TRUE)

# OUT-OF-SAMPLE EVALUATION

# Accuracy: 93.02%

yhat.Hand.RF <- predict(Hand.forest,newdata=testR)

rf.Hand.confusion <- table(yhat.Hand.RF,testR$activityID)

sum(diag(rf.Hand.confusion))/sum(rf.Hand.confusion)

summary(yhat.Hand.RF)

#-------------------------------------------------------/

# Chest SENSOR

Chest.forest =randomForest(as.factor(activityID)~ChestTemp+

ChestAccel16.1+

ChestAccel16.2+

ChestAccel16.3+

ChestGyro.1+

ChestGyro.2+

ChestGyro.3+

ChestMag.1+

ChestMag.2+

ChestMag.3,

data=trainR,

mtry=10,

importance =TRUE,

type="classification")

# IN-SAMPLE STATISTICS

# Accuracy: 93.69%

Chest.train.cm <- Chest.forest$confusion

sum(diag(Chest.train.cm))/sum(Chest.train.cm)

# VARIABLE IMPORTANCE

varImpPlot(Chest.forest, sort=TRUE, n.var=min(30, nrow(Chest.forest$importance)),

type=NULL, class=NULL, scale=TRUE,

main=deparse(substitute(x)))

Chest.RF.Imp <- Chest.forest$importance

write.csv(Chest.RF.Imp, file = "Chest.RF.Imp.csv",row.names=TRUE)

# OUT-OF-SAMPLE EVALUATION

# Accuracy: 93.72%

yhat.Chest.RF <- predict(Chest.forest,newdata=testR)

rf.Chest.confusion <- table(yhat.Chest.RF,testR$activityID)

sum(diag(rf.Chest.confusion))/sum(rf.Chest.confusion)

summary(yhat.Chest.RF)

#-------------------------------------------------------/

# ANKLE SENSOR

Ankle.forest =randomForest(as.factor(activityID)~AnkleTemp+

AnkleAccel16.1+

AnkleAccel16.2+

AnkleAccel16.3+

AnkleGyro.1+

AnkleGyro.2+

AnkleGyro.3+

AnkleMag.1+

AnkleMag.2+

AnkleMag.3,

data=trainR,

mtry=10,

importance =TRUE,

type="classification")

# IN-SAMPLE STATISTICS

# Accuracy: 92.52%

Ankle.train.cm <- Ankle.forest$confusion

sum(diag(Ankle.train.cm))/sum(Ankle.train.cm)

# VARIABLE IMPORTANCE

varImpPlot(Ankle.forest, sort=TRUE, n.var=min(30, nrow(Ankle.forest$importance)),

type=NULL, class=NULL, scale=TRUE,

main=deparse(substitute(x)))

Ankle.RF.Imp <- Ankle.forest$importance

write.csv(Ankle.RF.Imp, file = "Ankle.RF.Imp.csv",row.names=TRUE)

# OUT-OF-SAMPLE EVALUATION

# Accuracy: 92.60%

yhat.Ankle.RF <- predict(Ankle.forest,newdata=testR)

rf.Ankle.confusion <- table(yhat.Ankle.RF,testR$activityID)

sum(diag(rf.Ankle.confusion))/sum(rf.Ankle.confusion)

summary(yhat.Ankle.RF)

#-------------------------------------------------------/

# VARIABLE SUBSET RANDOM FOREST MODELS

#-------------------------------------------------------/

# HAND

# Eliminated bottom 5 variables for

# MeanDecrease GINI and ACCURACY

# Vars left: HandTemp, HandMag.2, HandAccel16.1,

# HandMag.1, and HandAccel16.3

# HAND SENSOR with top 5 VARS

Hand.sub.forest =randomForest(as.factor(activityID)~HandTemp+

HandAccel16.1+

HandAccel16.3+

HandMag.1+

HandMag.2,

data=trainR,

importance =TRUE,

type="classification")

# IN-SAMPLE STATISTICS

# Accuracy: 89.89%

Hand.sub.train.cm <- Hand.sub.forest$confusion

sum(diag(Hand.sub.train.cm))/sum(Hand.sub.train.cm)

# VARIABLE IMPORTANCE

varImpPlot(Hand.sub.forest, sort=TRUE, n.var=min(30, nrow(Hand.forest$importance)),

type=NULL, class=NULL, scale=TRUE,

main=deparse(substitute(x)))

# OUT-OF-SAMPLE EVALUATION

# Accuracy: 89.69%

yhat.Hand.sub.RF <- predict(Hand.sub.forest,newdata=testR)

rf.Hand.sub.confusion <- table(yhat.Hand.sub.RF,testR$activityID)

sum(diag(rf.Hand.sub.confusion))/sum(rf.Hand.sub.confusion)

summary(yhat.Hand.sub.RF)

# ------------------------------\*

# CHEST

# Eliminated bottom 5 variables for

# MeanDecrease GINI and ACCURACY

# CHEST SENSOR with top 5 VARS

Chest.sub.forest =randomForest(as.factor(activityID)~ChestTemp+

ChestAccel16.2+

ChestAccel16.3+

ChestMag.2+

ChestMag.3,

data=trainR,

importance =TRUE,

type="classification")

# IN-SAMPLE STATISTICS

# Accuracy: 91.30%

Chest.sub.train.cm <- Chest.sub.forest$confusion

sum(diag(Chest.sub.train.cm))/sum(Chest.sub.train.cm)

# VARIABLE IMPORTANCE

varImpPlot(Chest.forest, sort=TRUE, n.var=min(30, nrow(Chest.forest$importance)),

type=NULL, class=NULL, scale=TRUE,

main=deparse(substitute(x)))

# OUT-OF-SAMPLE EVALUATION

# Accuracy: 91.40%

yhat.Chest.sub.RF <- predict(Chest.sub.forest,newdata=testR)

rf.Chest.sub.confusion <- table(yhat.Chest.sub.RF,testR$activityID)

sum(diag(rf.Chest.sub.confusion))/sum(rf.Chest.sub.confusion)

summary(yhat.Chest.sub.RF)

# -------------------------------------------------\*

# ANKLE

# Eliminated bottom 5 variables for

# MeanDecrease GINI and ACCURACY

# Vars left: HandTemp, HandMag.2, HandAccel16.1,

# HandMag.1, and HandAccel16.3

# ANKLE SENSOR with top 5 VARS

Ankle.sub.forest =randomForest(as.factor(activityID)~AnkleTemp+

AnkleMag.1+

AnkleAccel16.1+

AnkleMag.2+

AnkleMag.3,

data=trainR,

importance =TRUE,

type="classification")

# IN-SAMPLE STATISTICS

# Accuracy: 89.00%

Ankle.sub.train.cm <- Ankle.sub.forest$confusion

sum(diag(Ankle.sub.train.cm))/sum(Ankle.sub.train.cm)

# VARIABLE IMPORTANCE

varImpPlot(Ankle.forest, sort=TRUE, n.var=min(30, nrow(Ankle.forest$importance)),

type=NULL, class=NULL, scale=TRUE,

main=deparse(substitute(x)))

# OUT-OF-SAMPLE EVALUATION

# Accuracy: 89.16%

yhat.Ankle.sub.RF <- predict(Ankle.sub.forest,newdata=testR)

rf.Ankle.sub.confusion <- table(yhat.Ankle.sub.RF,testR$activityID)

sum(diag(rf.Ankle.sub.confusion))/sum(rf.Ankle.sub.confusion)

summary(yhat.Ankle.sub.RF)

### CLUSTER

# Look at stuff from Marketing to cluster by activity

# Run models based on activity clusters

#--------------------------------------------------------------\*

# KNN

#--------------------------------------------------------------\*

library(class)

cluster <- knn(trainR, testR, trainR$activityID, k = 1, l = 0, prob = FALSE, use.all = TRUE)

cluster

#--------------------------------------------------------------\*

# SVM

#--------------------------------------------------------------\*

# Sample down trainR set to 20%

sub20.trainR <- trainR[sample(nrow(trainR), .2\*(nrow(trainR))), ]

# TUNE SVM

tune.out=tune(svm,as.factor(activityID) ~ . -Subject, data=sub20.trainR,

kernel = "linear",

ranges=list(cost=c(0.001, 0.01, 0.1, 1,5,10,100),

gamma=c(0.1,0.5,1,2,3,4,5,10)))

tune.out

# FIT MODEL

svmfit = svm(as.factor(activityID) ~ . -Subject,

data=sub20.trainR,

kernel = "linear", gamma=0.1, cost=10)

svmpred = predict(svmfit,testR)

mean(svmpred == testR$activityID)

#94.24% Accuracy

# LOGIT MODELS

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# Daily Sample 70/30

sample\_sizeDaily <- floor(.70\* nrow(Daily))

set.seed(498)

intDaily <- sample(seq\_len(nrow(Daily)),size=sample\_sizeDaily)

trainDaily <- Daily[intDaily,]

testDaily <- Daily[-intDaily,]

# EDA TREE WITH RANDOMIZED DATA

#Shuffle the dataset

set.seed(1234)

gDaily <- runif(nrow(Daily))

dataRDaily <- Daily[order(gDaily),]

sample\_sizeDaily <- floor(.70\* nrow(dataRDaily))

set.seed(498)

intDaily <- sample(seq\_len(nrow(Daily)),size=sample\_sizeDaily)

trainDaily <- Daily[intDaily,]

testDaily <- Daily[-intDaily,]

# Model Base

PAMlogitDaily <- glm(HR.Median ~ .-HR-activityID-Subject, data=trainDaily, family="binomial")

summary(PAMlogitDaily)

# AIC 4922.4

BIC(PAMlogitDaily)

# BIC 5194.34

varImp(PAMlogitDaily)

pR2(PAMlogitDaily) #McFadden .479

# Model 1

PAMlogitDaily1 <- glm(HR.Median ~ .-HR-Subject-activityID-HandGyro.1-HandGyro.2-HandGyro.3-HandMag.2-ChestAccel16.1

-ChestGyro.1-ChestGyro.3-ChestMag.3-AnkleAccel16.1-AnkleGyro.1-AnkleGyro.2-AnkleGyro.3-AnkleMag.2,

data=trainRDaily, family="binomial")

summary(PAMlogitDaily1)

# AIC 5211.2

BIC(PAMlogitDaily1)

# BIC 5372.68

varImp(PAMlogitDaily1)

pR2(PAMlogitDaily1) #McFadden .445

# Model 2

PAMlogitDaily2 <- glm(HR.Median ~ .-HR-Subject-activityID-HandGyro.1-HandGyro.2-HandGyro.3-HandMag.2-ChestAccel16.1

-ChestGyro.1-ChestGyro.3-ChestMag.3-AnkleAccel16.1-AnkleGyro.1-AnkleGyro.2-AnkleGyro.3-AnkleMag.2-HandAccel16.2-

HandAccel16.3-ChestAccel16.3-ChestMag.1, data=trainRDaily, family="binomial")

summary(PAMlogitDaily2)

# AIC 5276.4

BIC(PAMlogitDaily2)

# BIC 5403.86

varImp(PAMlogitDaily2)

pR2(PAMlogitDaily2) #McFadden .438

# Model 3

PAMlogitDaily3 <- glm(HR.Median ~ .-HR-Subject-activityID-Time, data=trainDaily, family="binomial")

summary(PAMlogitDaily3)

# AIC 5349.6

BIC(PAMlogitDaily3)

# BIC 5613.04

varImp(PAMlogitDaily3)

pR2(PAMlogitDaily3) #McFadden .433

# Model 4

PAMlogitDaily4 <- glm(HR.Median ~ .-HR-Subject-activityID-Time-HandGyro.1-HandGyro.2-HandGyro.3-HandMag.2-ChestGyro.1

-ChestGyro.3-ChestMag.2-ChestMag.3-AnkleGyro.1-AnkleGyro.2-AnkleGyro.3-AnkleMag.2-HandAccel16.2-HandMag.3-

HandAccel16.3-ChestAccel16.3-ChestMag.1, data=trainRDaily, family="binomial")

summary(PAMlogitDaily4)

# AIC 5395

BIC(PAMlogitDaily4)

# BIC 5513.92

varImp(PAMlogitDaily4)

pR2(PAMlogitDaily4) #McFadden .425

anova(PAMlogitDaily, PAMlogitDaily1, PAMlogitDaily2, PAMlogitDaily3, PAMlogitDaily4, test = "Chisq")

lrtest(PAMlogitDaily1, PAMlogitDaily3)

predDaily <- predict(PAMlogitDaily, newdata=testDaily, type="response")

accuracy <- table(predDaily, testDaily[,"HR.Median"])

sum(diag(accuracy))/sum(accuracy)

# 0.07664045

table(testDaily$HR.Median, predDaily > .5)

# FALSE TRUE

# 0 14988 60

# 1 372 94

predDailyTrain <- predict(PAMlogitDaily, type="response")

table(trainDaily$HR.Median, predDailyTrain > .5)

# FALSE TRUE

# 0 35025 148

# 1 742 283

# Calculate Odds Ratios

exp(coef(PAMlogitDaily))

# ROC

a <- roc(HR.Median ~ predDaily, data = testDaily)

plot(a, col="blue", main="Daily")

# Create Lift Chart

plotLift(predDaily, testDaily[,"HR.Median"], cumulative=TRUE, n.buckets=100, main="Daily Lift")

gains(testDaily[,"HR.Median"], predDaily, groups=10, optimal=TRUE, percents=TRUE)

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# Intensity Sample 70/30

sample\_sizeIntensity <- floor(.70\* nrow(Intensity))

set.seed(498)

intIntensity <- sample(seq\_len(nrow(Intensity)),size=sample\_sizeIntensity)

trainIntensity <- Intensity[intIntensity,]

testIntensity <- Intensity[-intIntensity,]

# EDA TREE WITH RANDOMIZED DATA

#Shuffle the dataset

set.seed(1234)

gIntensity <- runif(nrow(Intensity))

dataRIntensity <- Intensity[order(gIntensity),]

sample\_sizeIntensity <- floor(.70\* nrow(dataRIntensity))

set.seed(498)

intIntensity <- sample(seq\_len(nrow(Intensity)),size=sample\_sizeIntensity)

trainRIntensity <- Intensity[intIntensity,]

testRIntensity <- Intensity[-intIntensity,]

# Model Base

PAMlogitIntensity <- glm(HR.Median ~ .-HR-Subject-activityID, data=trainIntensity, family="binomial")

summary(PAMlogitIntensity)

#AIC 10937

BIC(PAMlogitIntensity)

# 11211.48

varImp(PAMlogitIntensity)

pR2(PAMlogitIntensity) #McFadden .206

# Model 1

PAMlogitIntensity1 <- glm(HR.Median ~ ChestTemp+Time+ChestMag.1+AnkleTemp, data=trainRIntensity, family="binomial")

summary(PAMlogitIntensity1)

#AIC 11380

BIC(PAMlogitIntensity1)

# 11422.8

varImp(PAMlogitIntensity1)

pR2(PAMlogitIntensity1) #McFadden .170

# Model 2

PAMlogitIntensity2 <- glm(HR.Median ~ .-HR-Subject-activityID-HandGyro.1-HandGyro.2-HandGyro.3-HandMag.3-ChestAccel16.1

-ChestGyro.1-ChestGyro.2-ChestGyro.3-ChestMag.2-ChestMag.3-AnkleAccel16.1-AnkleAccel16.2-AnkleAccel16.3-AnkleGyro.1-AnkleGyro.2-AnkleGyro.3

, data=trainRIntensity, family="binomial")

summary(PAMlogitIntensity2)

#AIC 10948

BIC(PAMlogitIntensity2)

# 11084.7

varImp(PAMlogitIntensity2)

pR2(PAMlogitIntensity2) #McFadden .203

# Model 3

PAMlogitIntensity3 <- glm(HR.Median ~ .-HR-Subject-activityID-HandGyro.1-HandGyro.2-HandGyro.3-HandMag.3-ChestGyro.1

-ChestGyro.3-ChestMag.2-ChestMag.3-AnkleAccel16.1-AnkleAccel16.2-AnkleAccel16.3-AnkleGyro.1-AnkleGyro.2-AnkleGyro.3

, data=trainRIntensity, family="binomial")

summary(PAMlogitIntensity3)

#AIC 10929

BIC(PAMlogitIntensity3)

# 11082.88

varImp(PAMlogitIntensity3)

pR2(PAMlogitIntensity3) #McFadden .204

# Model 4

PAMlogitIntensity4 <- glm(HR.Median ~ .-HR-Subject-activityID-HandGyro.1-HandGyro.2-HandGyro.3-HandMag.3

-ChestGyro.3-ChestMag.2-ChestMag.3-AnkleAccel16.1-AnkleAccel16.2-AnkleAccel16.3-AnkleGyro.1-AnkleGyro.2

, data=trainRIntensity, family="binomial")

summary(PAMlogitIntensity4)

#AIC 10920

BIC(PAMlogitIntensity4)

# 11090.89

varImp(PAMlogitIntensity4)

pR2(PAMlogitIntensity4) #McFadden .205

predIntensity <- predict(PAMlogitIntensity4, newdata=testIntensity, type="response")

accuracy <- table(predIntensity, testIntensity[,"HR.Median"])

sum(diag(accuracy))/sum(accuracy)

# 0.0001201

table(testIntensity$HR.Median, predIntensity > .5)

# FALSE TRUE

# 0 20 721

# 1 18 15895

predIntensityTrain <- predict(PAMlogitIntensity4, type="response")

table(trainIntensity$HR.Median, predIntensityTrain > .5)

# FALSE TRUE

# 0 31 1625

# 1 19 37183

# Calculate Odds Ratios

exp(coef(PAMlogitIntensity4))

# ROC

b <- roc(HR.Median ~ predIntensity, data = testIntensity)

plot(b, col="red", main="Intensity")

# Create Lift Chart

plotLift(predIntensity, testIntensity[,"HR.Median"], cumulative=TRUE, n.buckets=100, main="Intensity Lift")

gains(testIntensity[,"HR.Median"], predIntensity, groups=10, optimal=TRUE, percents=TRUE)

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# Chores Boxplots

boxplot(Time ~ HR.Median, data=Chores, xlab="HR Category", ylab="Time", main="Chores Subset Boxplots")

#Temp

boxplot(HandTemp ~ HR.Median, data=Chores, xlab="HR Category", ylab="Hand Temperature", main="Chores Subset Boxplots")

boxplot(ChestTemp ~ HR.Median, data=Chores, xlab="HR Category", ylab="Chest Temperature", main="Chores Subset Boxplots")

boxplot(AnkleTemp ~ HR.Median, data=Chores, xlab="HR Category", ylab="Ankle Temperature", main="Chores Subset Boxplots")

#Accel

boxplot(HandAccel16.1 ~ HR.Median, data=Chores, xlab="HR Category", ylab="HandAccel16.1", main="Chores Subset Boxplots")

boxplot(HandAccel16.2 ~ HR.Median, data=Chores, xlab="HR Category", ylab="HandAccel16.2", main="Chores Subset Boxplots")

boxplot(HandAccel16.3 ~ HR.Median, data=Chores, xlab="HR Category", ylab="HandAccel16.3", main="Chores Subset Boxplots")

boxplot(ChestAccel16.1 ~ HR.Median, data=Chores, xlab="HR Category", ylab="ChestAccel16.1", main="Chores Subset Boxplots")

boxplot(ChestAccel16.2 ~ HR.Median, data=Chores, xlab="HR Category", ylab="ChestAccel16.2", main="Chores Subset Boxplots")

boxplot(ChestAccel16.3 ~ HR.Median, data=Chores, xlab="HR Category", ylab="ChestAccel16.3", main="Chores Subset Boxplots")

boxplot(AnkleAccel16.1 ~ HR.Median, data=Chores, xlab="HR Category", ylab="AnkleAccel16.1", main="Chores Subset Boxplots")

boxplot(AnkleAccel16.2 ~ HR.Median, data=Chores, xlab="HR Category", ylab="AnkleAccel16.2", main="Chores Subset Boxplots")

boxplot(AnkleAccel16.3 ~ HR.Median, data=Chores, xlab="HR Category", ylab="AnkleAccel16.3", main="Chores Subset Boxplots")

#Gyro

boxplot(HandGyro.1 ~ HR.Median, data=Chores, xlab="HR Category", ylab="HandGyro1", main="Chores Subset Boxplots")

boxplot(HandGyro.2 ~ HR.Median, data=Chores, xlab="HR Category", ylab="HandGyro2", main="Chores Subset Boxplots")

boxplot(HandGyro.3 ~ HR.Median, data=Chores, xlab="HR Category", ylab="HandGyro3", main="Chores Subset Boxplots")

boxplot(ChestGyro.1 ~ HR.Median, data=Chores, xlab="HR Category", ylab="ChestGyro1", main="Chores Subset Boxplots")

boxplot(ChestGyro.2 ~ HR.Median, data=Chores, xlab="HR Category", ylab="ChestGyro2", main="Chores Subset Boxplots")

boxplot(ChestGyro.3 ~ HR.Median, data=Chores, xlab="HR Category", ylab="ChestGyro3", main="Chores Subset Boxplots")

boxplot(AnkleGyro.1 ~ HR.Median, data=Chores, xlab="HR Category", ylab="AnkleGyro1", main="Chores Subset Boxplots")

boxplot(AnkleGyro.2 ~ HR.Median, data=Chores, xlab="HR Category", ylab="AnkleGyro2", main="Chores Subset Boxplots")

boxplot(AnkleGyro.3 ~ HR.Median, data=Chores, xlab="HR Category", ylab="AnkleGyro3", main="Chores Subset Boxplots")

#Magnet

boxplot(HandMag.1 ~ HR.Median, data=Chores, xlab="HR Category", ylab="HandMag1", main="Chores Subset Boxplots")

boxplot(HandMag.2 ~ HR.Median, data=Chores, xlab="HR Category", ylab="HandMag2", main="Chores Subset Boxplots")

boxplot(HandMag.3 ~ HR.Median, data=Chores, xlab="HR Category", ylab="HandMag3", main="Chores Subset Boxplots")

boxplot(ChestMag.1 ~ HR.Median, data=Chores, xlab="HR Category", ylab="ChestMag1", main="Chores Subset Boxplots")

boxplot(ChestMag.2 ~ HR.Median, data=Chores, xlab="HR Category", ylab="ChestMag2", main="Chores Subset Boxplots")

boxplot(ChestMag.3 ~ HR.Median, data=Chores, xlab="HR Category", ylab="ChestMag3", main="Chores Subset Boxplots")

boxplot(AnkleMag.1 ~ HR.Median, data=Chores, xlab="HR Category", ylab="AnkleMag1", main="Chores Subset Boxplots")

boxplot(AnkleMag.2 ~ HR.Median, data=Chores, xlab="HR Category", ylab="AnkleMag2", main="Chores Subset Boxplots")

boxplot(AnkleMag.3 ~ HR.Median, data=Chores, xlab="HR Category", ylab="AnkleMag3", main="Chores Subset Boxplots")

# Chores Histograms

hist(Chores$Time)

hist(Chores$HR)

#Temp

hist(Chores$HandTemp)

hist(Chores$ChestTemp)

hist(Chores$AnkleTemp)

#Accel

hist(Chores$HandAccel16.1)

hist(Chores$HandAccel16.2)

hist(Chores$HandAccel16.3)

hist(Chores$ChestAccel16.1)

hist(Chores$ChestAccel16.2)

hist(Chores$ChestAccel16.3)

hist(Chores$AnkleAccel16.1)

hist(Chores$AnkleAccel16.2)

hist(Chores$AnkleAccel16.3)

#Gyro

hist(Chores$HandGyro.1)

hist(Chores$HandGyro.2)

hist(Chores$HandGyro.3)

hist(Chores$ChestGyro.1)

hist(Chores$ChestGyro.2)

hist(Chores$ChestGyro.3)

hist(Chores$AnkleGyro.1)

hist(Chores$AnkleGyro.2)

hist(Chores$AnkleGyro.3)

#Magnet

hist(Chores$HandMag.1)

hist(Chores$HandMag.2)

hist(Chores$HandMag.3)

hist(Chores$ChestMag.1)

hist(Chores$ChestMag.2)

hist(Chores$ChestMag.3)

hist(Chores$AnkleMag.1)

hist(Chores$AnkleMag.2)

hist(Chores$AnkleMag.3)

# Chores Sample 70/30

sample\_sizeChores <- floor(.70\* nrow(Chores))

set.seed(498)

intChores <- sample(seq\_len(nrow(Chores)),size=sample\_sizeChores)

trainChores <- Chores[intChores,]

testChores <- Chores[-intChores,]

# EDA TREE WITH RANDOMIZED DATA

#Shuffle the dataset

set.seed(1234)

gChores <- runif(nrow(Chores))

dataRChores <- Chores[order(gChores),]

sample\_sizeChores <- floor(.70\* nrow(dataRChores))

set.seed(498)

intChores <- sample(seq\_len(nrow(Chores)),size=sample\_sizeChores)

trainRChores <- Chores[intChores,]

testRChores <- Chores[-intChores,]

# Model Base

PAMlogitChores <- glm(HR.Median ~ .-HR-activityID-Subject, data=trainRChores, family="binomial")

summary(PAMlogitChores)

#AIC 40411

BIC(PAMlogitChores)

# 40691.42

varImp(PAMlogitChores)

pR2(PAMlogitChores) #McFadden .3902

# Model 1

PAMlogitChores1 <- glm(HR.Median ~ .-HR-activityID-Subject-Time, data=trainRChores, family="binomial")

summary(PAMlogitChores1)

#AIC 50031

BIC(PAMlogitChores1)

# 50303.42

varImp(PAMlogitChores1)

pR2(PAMlogitChores1) #McFadden .245

# Model 2

PAMlogitChores2 <- glm(HR.Median ~ .-HR-Subject-activityID-HandGyro.1-HandGyro.2-HandGyro.3-HandMag.3-ChestGyro.1

-ChestGyro.3-AnkleAccel16.2-AnkleAccel16.3-AnkleGyro.1-AnkleMag.1

, data=trainRChores, family="binomial")

summary(PAMlogitChores2)

#AIC 40440

BIC(PAMlogitChores2)

# 40633.29

varImp(PAMlogitChores2)

pR2(PAMlogitChores2) #McFadden .3895

# Model 3

PAMlogitChores3 <- glm(HR.Median ~ .-HR-Subject-activityID-HandGyro.1-HandGyro.3-HandMag.3-ChestGyro.1

-ChestGyro.3-AnkleAccel16.2-AnkleAccel16.3-AnkleMag.1

, data=trainRChores, family="binomial")

summary(PAMlogitChores3)

#AIC 40421

BIC(PAMlogitChores3)

# 40631.2

varImp(PAMlogitChores3)

pR2(PAMlogitChores3) #McFadden .3898

# Model 4

PAMlogitChores4 <- glm(HR.Median ~ .-HR-Subject-activityID-HandGyro.1-HandGyro.2-HandGyro.3-HandMag.2-HandMag.3-ChestMag.1-ChestGyro.1

-ChestGyro.3-AnkleGyro.1-AnkleGyro.3-AnkleAccel16.2-AnkleAccel16.3-AnkleMag.1

, data=trainRChores, family="binomial")

summary(PAMlogitChores4)

#AIC 40472

BIC(PAMlogitChores4)

# 40638.51

varImp(PAMlogitChores4)

pR2(PAMlogitChores4) #McFadden .389

predChores <- predict(PAMlogitChores3, newdata=testChores, type="response")

accuracy <- table(predChores, testChores[,"HR.Median"])

sum(diag(accuracy))/sum(accuracy)

# 0.0000488203

table(testChores$HR.Median, predChores > .5)

# FALSE TRUE

# 0 9060 1765

# 1 1910 7748

predChoresTrain <- predict(PAMlogitChores3, type="response")

table(trainChores$HR.Median, predChoresTrain > .5)

# FALSE TRUE

# 0 21072 3868

# 1 4642 18211

# Calculate Odds Ratios

exp(coef(PAMlogitChores3))

# ROC

c <- roc(HR.Median ~ predChores, data = testChores)

plot(c, col="green", main="Chores")

# Create Lift Chart

plotLift(predChores, testChores[,"HR.Median"], cumulative=TRUE, n.buckets=100, main="Chores Lift")

gains(testChores[,"HR.Median"], predChores, groups=10, optimal=TRUE, percents=TRUE)

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# Complete Boxplots

boxplot(Time ~ HR.Median, data=data, xlab="HR Category", ylab="Time", main="All Subjects Subset Boxplots")

#Temp

boxplot(HandTemp ~ HR.Median, data=data, xlab="HR Category", ylab="Hand Temperature", main="All Subjects Subset Boxplots")

boxplot(ChestTemp ~ HR.Median, data=data, xlab="HR Category", ylab="Chest Temperature", main="All Subjects Subset Boxplots")

boxplot(AnkleTemp ~ HR.Median, data=data, xlab="HR Category", ylab="Ankle Temperature", main="All Subjects Subset Boxplots")

#Accel

boxplot(HandAccel16.1 ~ HR.Median, data=data, xlab="HR Category", ylab="HandAccel16.1", main="All Subjects Subset Boxplots")

boxplot(HandAccel16.2 ~ HR.Median, data=data, xlab="HR Category", ylab="HandAccel16.2", main="All Subjects Subset Boxplots")

boxplot(HandAccel16.3 ~ HR.Median, data=data, xlab="HR Category", ylab="HandAccel16.3", main="All Subjects Subset Boxplots")

boxplot(ChestAccel16.1 ~ HR.Median, data=data, xlab="HR Category", ylab="ChestAccel16.1", main="All Subjects Subset Boxplots")

boxplot(ChestAccel16.2 ~ HR.Median, data=data, xlab="HR Category", ylab="ChestAccel16.2", main="All Subjects Subset Boxplots")

boxplot(ChestAccel16.3 ~ HR.Median, data=data, xlab="HR Category", ylab="ChestAccel16.3", main="All Subjects Subset Boxplots")

boxplot(AnkleAccel16.1 ~ HR.Median, data=data, xlab="HR Category", ylab="AnkleAccel16.1", main="All Subjects Subset Boxplots")

boxplot(AnkleAccel16.2 ~ HR.Median, data=data, xlab="HR Category", ylab="AnkleAccel16.2", main="All Subjects Subset Boxplots")

boxplot(AnkleAccel16.3 ~ HR.Median, data=data, xlab="HR Category", ylab="AnkleAccel16.3", main="All Subjects Subset Boxplots")

#Gyro

boxplot(HandGyro.1 ~ HR.Median, data=data, xlab="HR Category", ylab="HandGyro1", main="All Subjects Subset Boxplots")

boxplot(HandGyro.2 ~ HR.Median, data=data, xlab="HR Category", ylab="HandGyro2", main="All Subjects Subset Boxplots")

boxplot(HandGyro.3 ~ HR.Median, data=data, xlab="HR Category", ylab="HandGyro3", main="All Subjects Subset Boxplots")

boxplot(ChestGyro.1 ~ HR.Median, data=data, xlab="HR Category", ylab="ChestGyro1", main="All Subjects Subset Boxplots")

boxplot(ChestGyro.2 ~ HR.Median, data=data, xlab="HR Category", ylab="ChestGyro2", main="All Subjects Subset Boxplots")

boxplot(ChestGyro.3 ~ HR.Median, data=data, xlab="HR Category", ylab="ChestGyro3", main="All Subjects Subset Boxplots")

boxplot(AnkleGyro.1 ~ HR.Median, data=data, xlab="HR Category", ylab="AnkleGyro1", main="All Subjects Subset Boxplots")

boxplot(AnkleGyro.2 ~ HR.Median, data=data, xlab="HR Category", ylab="AnkleGyro2", main="All Subjects Subset Boxplots")

boxplot(AnkleGyro.3 ~ HR.Median, data=data, xlab="HR Category", ylab="AnkleGyro3", main="All Subjects Subset Boxplots")

#Magnet

boxplot(HandMag.1 ~ HR.Median, data=data, xlab="HR Category", ylab="HandMag1", main="All Subjects Subset Boxplots")

boxplot(HandMag.2 ~ HR.Median, data=data, xlab="HR Category", ylab="HandMag2", main="All Subjects Subset Boxplots")

boxplot(HandMag.3 ~ HR.Median, data=data, xlab="HR Category", ylab="HandMag3", main="All Subjects Subset Boxplots")

boxplot(ChestMag.1 ~ HR.Median, data=data, xlab="HR Category", ylab="ChestMag1", main="All Subjects Subset Boxplots")

boxplot(ChestMag.2 ~ HR.Median, data=data, xlab="HR Category", ylab="ChestMag2", main="All Subjects Subset Boxplots")

boxplot(ChestMag.3 ~ HR.Median, data=data, xlab="HR Category", ylab="ChestMag3", main="All Subjects Subset Boxplots")

boxplot(AnkleMag.1 ~ HR.Median, data=data, xlab="HR Category", ylab="AnkleMag1", main="All Subjects Subset Boxplots")

boxplot(AnkleMag.2 ~ HR.Median, data=data, xlab="HR Category", ylab="AnkleMag2", main="All Subjects Subset Boxplots")

boxplot(AnkleMag.3 ~ HR.Median, data=data, xlab="HR Category", ylab="AnkleMag3", main="All Subjects Subset Boxplots")

#Complete Histograms

hist(data$Time)

hist(data$HR)

#Temp

hist(data$HandTemp)

hist(data$ChestTemp)

hist(data$AnkleTemp)

#Accel

hist(data$HandAccel16.1)

hist(data$HandAccel16.2)

hist(data$HandAccel16.3)

hist(data$ChestAccel16.1)

hist(data$ChestAccel16.2)

hist(data$ChestAccel16.3)

hist(data$AnkleAccel16.1)

hist(data$AnkleAccel16.2)

hist(data$AnkleAccel16.3)

#Gyro

hist(data$HandGyro.1)

hist(data$HandGyro.2)

hist(data$HandGyro.3)

hist(data$ChestGyro.1)

hist(data$ChestGyro.2)

hist(data$ChestGyro.3)

hist(data$AnkleGyro.1)

hist(data$AnkleGyro.2)

hist(data$AnkleGyro.3)

#Magnet

hist(data$HandMag.1)

hist(data$HandMag.2)

hist(data$HandMag.3)

hist(data$ChestMag.1)

hist(data$ChestMag.2)

hist(data$ChestMag.3)

hist(data$AnkleMag.1)

hist(data$AnkleMag.2)

hist(data$AnkleMag.3)

# Complete Sample 70/30

sample\_size <- floor(.70\* nrow(data))

set.seed(498)

int <- sample(seq\_len(nrow(data)),size=sample\_size)

train <- data[int,]

test <- data[-int,]

# EDA TREE WITH RANDOMIZED DATA

#Shuffle the dataset

set.seed(1234)

g <- runif(nrow(data))

dataR <- data[order(g),]

sample\_size <- floor(.70\* nrow(dataR))

set.seed(498)

int <- sample(seq\_len(nrow(data)),size=sample\_size)

trainR <- data[int,]

testR <- data[-int,]

PAMlogit <- glm(HR.Median ~ .-HR-Subject-activityID, data=trainR, family="binomial")

summary(PAMlogit)

# AIC 67466

BIC(PAMlogit)

# 67776.6

varImp(PAMlogit)

pR2(PAMlogit) #McFadden .604

# Model 1

PAMlogit1 <- glm(HR.Median ~ .-HR-Subject-Time-activityID, data=trainR, family="binomial")

summary(PAMlogit1)

# AIC 97965

BIC(PAMlogit1)

# 98266.3

varImp(PAMlogit1)

pR2(PAMlogit1) #McFadden .425

# Model 2

PAMlogit2 <- glm(HR.Median ~ .-HR-Subject-activityID-HandAccel16.1-HandAccel16.3-HandGyro.1-HandGyro.2-HandGyro.3-HandMag.2-ChestAccel16.3

-ChestGyro.1-ChestGyro.2-ChestGyro.3-AnkleAccel16.2-AnkleAccel16.3-AnkleGyro.1-AnkleGyro.2-AnkleGyro.3-AnkleMag.2

, data=trainR, family="binomial")

summary(PAMlogit2)

#AIC 67826

BIC(PAMlogit2)

# 67981.9

varImp(PAMlogit2)

pR2(PAMlogit2) #McFadden .602

# Model 3

PAMlogit3 <- glm(HR.Median ~ .-HR-Subject-activityID-HandAccel16.1-HandGyro.1-ChestAccel16.3

-ChestGyro.1-ChestGyro.2-ChestGyro.3-AnkleAccel16.2-AnkleGyro.1-AnkleGyro.2-AnkleGyro.3-AnkleMag.2

, data=trainR, family="binomial")

summary(PAMlogit3)

#AIC 67760

BIC(PAMlogit3)

# 67964

varImp(PAMlogit3)

pR2(PAMlogit3) #McFadden .602

# Model 4

PAMlogit4 <- glm(HR.Median ~ .-HR-Subject-activityID-HandAccel16.1-HandGyro.1-HandMag.2-ChestAccel16.3

-ChestGyro.1-ChestGyro.2-AnkleGyro.1-AnkleGyro.2-AnkleGyro.3-AnkleMag.2

, data=trainR, family="binomial")

summary(PAMlogit4)

#AIC 67765

BIC(PAMlogit4)

# 67979.2

varImp(PAMlogit4)

pR2(PAMlogit4) #McFadden .602

pred <- predict(PAMlogit3, newdata=test, type="response")

accuracy <- table(pred, test[,"HR.Median"])

sum(diag(accuracy))/sum(accuracy)

# 0.000018993

table(test$HR.Median, pred > .5)

# FALSE TRUE

# 0 24183 2416

# 1 3121 22931

predTrain <- predict(PAMlogit3, type="response")

table(train$HR.Median, predTrain > .5)

# FALSE TRUE

# 0 55934 5850

# 1 7229 53836

# Calculate Odds Ratios

exp(coef(PAMlogit3))

# ROC

d <- roc(HR.Median ~ pred, data = test)

plot(d, col="black", main="All")

# Create Lift Chart

plotLift(pred, test[,"HR.Median"], cumulative=TRUE, n.buckets=100, main="All Lift")

gains(test[,"HR.Median"], pred, groups=10, optimal=TRUE, percents=TRUE)

# LINEAR MODELING

# EDA #

hand<- grep('^Hand', names(data), value = TRUE)

chest<- grep('^Chest', names(data), value = TRUE)

ankle<- grep('^Ankle', names(data), value = TRUE)

# Hist #

par(mfrow=c(3,4))

qplot(data$HandTemp, geom="histogram")

qplot(data$HandAccel16.1, geom="histogram")

qplot(data$HandAccel16.2, geom="histogram")

qplot(data$HandAccel16.3, geom="histogram")

qplot(data$HandGyro.1, geom="histogram")

qplot(data$HandGyro.2, geom="histogram")

qplot(data$HandGyro.3, geom="histogram")

qplot(data$HandMag.1, geom="histogram")

qplot(data$HandMag.2, geom="histogram")

qplot(data$HandMag.3, geom="histogram")

par(mfrow = c(1,1))

#heatmap hand#

require(reshape)

require(scales)

require(ggplot2)

hand.cor <-cor(data[,c(3, 4:13)])

handmelt <- melt(hand.cor, varnames = c("x", "y"))

handmelt <- handmelt[order(handmelt$value),]

ggplot(handmelt, aes(x=x, y=y)) + geom\_tile(aes(fill = value))+scale\_fill\_gradient2(low=muted("red"), mid="white", high = "steelblue", guide = guide\_colorbar(ticks=FALSE, barheight = 10,main="heatmap hand variables and heart rate"), limits=c(-1,1))+theme\_minimal()+labs(x=NULL,y=NULL)

#heatmap chest#

chest.cor <-cor(data[,c(3, 14:23)])

chestmelt <- melt(chest.cor, varnames = c("x", "y"))

chestmelt <- chestmelt[order(chestmelt$value),]

ggplot(chestmelt, aes(x=x, y=y)) + geom\_tile(aes(fill = value))+scale\_fill\_gradient2(low=muted("red"), mid="white", high = "steelblue", guide = guide\_colorbar(ticks=FALSE, barheight = 10), limits=c(-1,1))+theme\_minimal()+labs(x=NULL,y=NULL)

#heatmap ankle#

ankle.cor <-cor(data[,c(3, 24:33)])

anklemelt <- melt(ankle.cor, varnames = c("x", "y"))

anklemelt <- anklemelt[order(anklemelt$value),]

ggplot(anklemelt, aes(x=x, y=y)) + geom\_tile(aes(fill = value))+scale\_fill\_gradient2(low=muted("red"), mid="white", high = "steelblue", guide = guide\_colorbar(ticks=FALSE, barheight = 10), limits=c(-1,1))+theme\_minimal()+labs(x=NULL,y=NULL)

#Pairs hand, chest, ankle#

pairs(data[,c(3,14:23)])

pairs(data[,c(3,4:13)])

pairs(data[,c(3,24:33)])

## Model 1 All Variables ##

set.seed(19)

train <- sample(1:nrow(data),0.70\*nrow(data))

fit0 <- lm(HR~.-activityID-HR.Median, data = data, subset = train)

fit1 <- lm(HR~.-HR.Median-Time-HandAccel16.1-HandTemp-ChestAccel16.3-activityID-HandGyro.1-ChestGyro.1-AnkleAccel16.1-AnkleGyro.1-AnkleGyro.2-AnkleGyro.3+poly(ChestTemp,7)+poly(AnkleTemp,7)+poly(HandTemp,7)+poly(Time,7), data = data, subset = train)

fit2 <- step(fit1)

summary(fit2)

mean((total$HR -predict(fit2 , data))[-train ]^2)

summary(fit0)

mean((total$HR -predict(fit0 , data))[-train ]^2)

## fit metrics fit 2##

fit2.stdres = rstandard(fit2)

qqnorm(fit2.stdres,

ylab="Standardized Residuals",

xlab="Normal Scores",

main="HR Prediction")

qqline(fit2.stdres)

par(mfrow=c(2,3))

hist(fit2.stdres)

plot(fit2, which=2)

plot(fit2, which=1)

hist(fit2, which = 4)

plot(fit2, which=5)

plot(fit2, which=6)

par(mfrow=c(1,1))

## Model 2 Best Subset ##

library (leaps)

set.seed (17)

train=sample(c(TRUE,FALSE),nrow(data),rep=TRUE)

test=(!train)

regfit.best=regsubsets(HR~.-HR.Median-activityID,data= data[train,],nvmax =19)

test.mat=model.matrix(HR~.-HR.Median-activityID,data=data[test,])

val.errors =rep(NA ,19)

for(i in 1:19){

coefi=coef(regfit.best,id=i)

pred=test.mat[,names(coefi)]%\*% coefi

val.errors [i]= mean((total$HR[test]-pred)^2)

}

which.min(val.errors)

coef(regfit.best,19)

## Model 3 Ridge ##

x=model.matrix(HR~.-HR.Median-activityID,data)[,-1]

y=data$HR

library(glmnet)

grid =10^seq(10,-2,length=100)

ridge.mod=glmnet(x,y,alpha=0,lambda=grid)

set.seed (1)

train=sample(1:nrow(x),nrow(x)/2)

test=(-train )

y.test=y[test]

# CVal Ridge #

set.seed (12)

cv.out=cv.glmnet(x[train,],y[train],alpha=0)

plot(cv.out)

bestlam=cv.out$lambda.min

bestlam

ridge.pred=predict(ridge.mod,s=bestlam ,newx=x[test ,])

mean((ridge.pred-y.test)^2)

out=glmnet(x,y,alpha =0)

predict (out,type="coefficients",s=bestlam )[1:20 ,]

## Lasso ##

lasso.mod =glmnet(x[train ,],y[train],alpha =1,lambda =grid)

plot(lasso.mod)

set.seed (11)

cv.out=cv.glmnet(x[train ,],y[train],alpha =1)

plot(cv.out)

bestlam=cv.out$lambda.min

lasso.pred=predict(lasso.mod ,s=bestlam ,newx=x[test ,])

mean((lasso.pred -y.test)^2)

out=glmnet(x,y,alpha =1, lambda =grid)

lasso.coef=predict(out ,type ="coefficients",s=bestlam )[1:30 ,]

lasso.coef[lasso.coef!=0]

## PCR ##

library (pls)

set.seed (2)

pcr.fit=pcr(HR~.-HR.Median-activityID, data=data ,subset=train,scale =TRUE,validation ="CV")

validationplot(pcr.fit,val.type="MSEP")

pcr.pred=predict (pcr.fit ,x[test ,], ncomp =25)

mean((pcr.pred-y.test)^2)

## PLS ##

library (pls)

set.seed (1)

x=model.matrix(HR~.-HR.Median-activityID,data)[,-1]

y=total$HR

train=sample(1:nrow(x),nrow(x)\*.70)

test=(-train )

y.test=y[test]

grid =10^seq(10,-2,length=100)

pls.fit=plsr(HR~.-HR.Median-activityID, data=data ,subset=train ,scale=TRUE,validation ="CV")

summary(pls.fit)

pls.pred=predict(pls.fit ,x[test ,], ncomp =8)

mean((pls.pred -y.test)^2)

pls.fit=plsr(HR~.-HR.Median-activityID, data=data,scale=TRUE ,ncomp =8)

summary(pls.fit)

## Fit and Plots PLS ##

plot(RMSEP(pls.fit), legendpos = "topright")

plot(pls.fit, ncomp = 8, asp = 1, line = TRUE)

plot(pls.fit, plottype = "scores", comps = 1:8)

plot(pls.fit, plottype = "correlation", comps =8)

explvar(pls.fit)

plot(pls.fit, "loadings", comps = 1:8, legendpos = "topleft", xlab = "nm") > abline(h = 0)

pls.fit$loadings

1. Reiss, A. and Stricker, D. (2012). *Creating and Benchmarking a New Dataset for Physical Activity Monitoring*. Kaiserslautern, Germany: German Research Center for Artificial Intelligence (DFKI). [↑](#footnote-ref-1)
2. Reiss, A. and Stricker, D. (2012). *Creating and Benchmarking a New Dataset for Physical Activity Monitoring*. Kaiserslautern, Germany: German Research Center for Artificial Intelligence (DFKI), p. 4 [↑](#footnote-ref-2)
3. Reiss, A. and Stricker, D. (2012). *Creating and Benchmarking a New Dataset for Physical Activity Monitoring*. Kaiserslautern, Germany: German Research Center for Artificial Intelligence (DFKI), p. 7. [↑](#footnote-ref-3)
4. UCI Machine Learning Repository, PAMAP2 Physical Activity Monitoring Data Set, accessed on January 9, 2016 at <https://archive.ics.uci.edu/ml/datasets/PAMAP2+Physical+Activity+Monitoring> [↑](#footnote-ref-4)
5. <http://www.heart.org/HEARTORG/HealthyLiving/PhysicalActivity/FitnessBasics/Target-Heart-Rates_UCM_434341_Article.jsp#.VtjQpWVIgy8> [↑](#footnote-ref-5)
6. Louviere, Jordan J., David A. Hensher, Joffre D. Swait, and Wiktor Adamowicz. *Stated Choice Methods: Analysis and Applications*. Cambridge: Cambridge UP, 2000. Print. [↑](#footnote-ref-6)
7. Schonlau, M., *Visualizing non-hierarchical and hierarchical cluster analysis with clustergrams* [Electronic Format]. Retreived from http://www.schonlau.net/publication/04compstat\_clustergram.pdf [↑](#footnote-ref-7)
8. UCI Machine Learning Repository, PAMAP2 Physical Activity Monitoring Data Set, accessed on January 9, 2016 at <https://archive.ics.uci.edu/ml/datasets/PAMAP2+Physical+Activity+Monitoring> [↑](#footnote-ref-8)
9. Ding, C. & He, X. (2004). K-means clustering via Principal Component Analysis. ICML '04 Proceedings of the twenty-first international conference on Machine learning, p. 29. Retrieved from http://www.vision.caltech.edu/wikis/EE148/images/c/c2/KmeansPCA1.pdf [↑](#footnote-ref-9)