Shaharyar Sued
Emerson Kirby
Julius Wu
Susheel Palakurthi

Introduction

In the modern workplace, employee efficiency is arguably one of the most important factors in determining a company's success. One large factor in an employee's efficiency is the their quality of life. Personal weight management has increasingly become an issue for employees in various work forces, however, many insurance firms that provide services to large scale employers implement a method to indirectly counter this issue. Many employers provide health insurance of some sort to their employees through corporate insurance, and these insurance firms determine premiums for their clients through various data collection softwares. These softwares sustain the ability to provide necessary information about clients, and also promote client weight management subsequently affecting the quality of life and efficiency of workers.

This project aims to investigate various medical claims regarding weight gain/loss, and answer various correlation of variables questions that may arise from viewing data collected for insurance companies using real user data. Specifically, we will investigate

1. Whether there is a relationship between a client's goal weight, current measurement and gender.
2. Does working at a certain company (labor category) affect your health? In other words, are you, as a client, more likely to set a higher weight loss goal based off of where you work.
3. Does the aesthetic attractiveness of the software affect how likely a client is to use it? Can clients use this to their benefit?
4. Find 'expected' groups in user data, using medical claims concerning SBP and weight.
5. Predict a client's gender based off of their goal weight.

Data

The data being investigated belongs to a local software development company, CoreHealth Technologies. CoreHealth Technologies provides health sector insurance companies with a software that tracks and promotes client weight management and wellness. In turn, insurance companies can determine premiums for businesses they provide services to by observing usage data. Alongside determining premiums, insurers can lower their payouts, as a healthy populace will have lower health care costs. We were able to gain access to this data through a current employee.

The data was drawn from CoreHealth's relational database and therefore requires some manipulation before it could be used for analysis. Three of CoreHealth's tables were queried to gather the data; Users, Measurement, and Logins. The user table contains information provided by 40,242 unique users, who have the following useful information about them stored: Their user id, date of birth, whether they are an active account, gender, goal weight, daily steps goal, number of days of activity per week, and their locale name. Due to the nature of CoreHealth's

platform, each insurance company gets their own site, on which they can set up each client portal, which is called the locale. Therefore, each end user has a locale that they use. Users are not required to enter all their information though, so null values must be considered when performing analyses.

Each user also has their physical measurements stored in separate table, the Measurements table. It contains personal measurements such as current Height, SBP (Systolic blood pressure), DBP (Diastolic blood pressure), Weight (in KG), HDL (High-density lipoprotein), LDL (Low-density lipoprotein) and glucose levels. In addition, there is the Login table which tracks the data and time each user has logged in. This table contains 539,963 observations.

Important information regarding variables in the data:

It can be immediately noticed that all user specific measurement variables are directly correlated with weight.

Recent medical studies show that SBP has a positive direct relationship with weight, as does DBP .[1] This claim will be investigated later in this report.

Methodology & Application

Objective 1: Relationship between a client's goal weight, current measurement and gender

(Linear Regression/Variable Selection/Cross Validation)

To explore this objective, we apply a linear model to a usable subset of the data. Furthermore, to avoid overfitting the method of cross validation is applied.

A linear regression model is used to determine the relationship between a scalar dependant variable and one or more explanatory variables. The process begins with removing outliers from the data set, this is first done by creating a subset of the data under the condition that GoalWeight is between 40 and 90 (kg). This range was chosen as it seems to be fairly representative of the general population. Then a subset of that newly created set is determined under the condition that current measurement is within the same range. At this point, any outliers and high leverage points that may affect the fit of the model have been removed so that our model can display clearer results.

Secondly, a linear regression model is applied to the data, where the response variable is set as GoalWeight and the predictors are set as Measurement, Gender and Age. We've added the 'Age' variable to investigate any possible relationship outside of our proposed question. Observing a summary of the fit, we find that the 'Age' variable has a p-value beyond our threshold (0.05) so it is removed using backward selection. It should also be noted that the Multiple r-squared value is 0.83, which correlates to the model being a good fit.

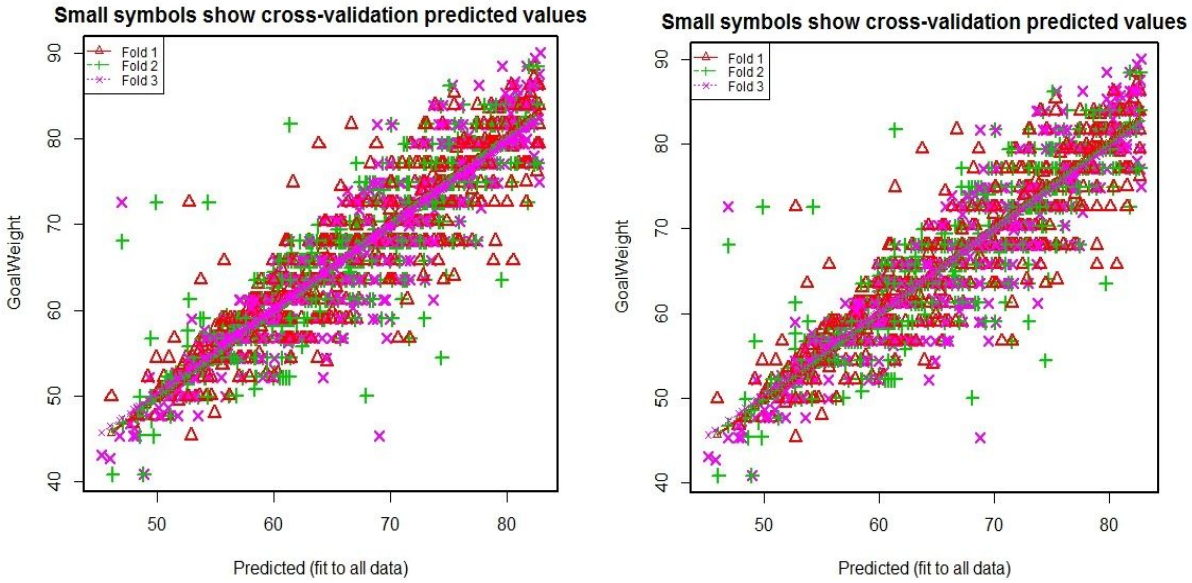
Finally, considering a plot of the model, a clear linear relationship between GoalWeight and Measurement is observed. It can be concluded that as a user's current measurement for

weight increases, their goal weight increases linearly. This would mean that most users are realistic in choosing a goal weight based on their level of activity; a goal weight that is not overly optimistic (Over 10 kg loss or gain).

Cross Validation:

Cross validation is the process used in variable selection to better estimate the error in a model. Although there are many techniques such as splitting data in two halves, K-fold cross validation, leave one out cross validation (LOOCV), etc; the basic idea behind it is to create a number of partitions in the data to create training and testing sets and comparing their model of fit estimates against each other. In some cross validation methods, folds (K) are used to separate the data into K subsets in order to perform cross validation. In the selection of K, a high K introduces high variance and less bias while a low K introduces low variance and more bias. When cross validation is being done, one K subset is held as the test set while the K-1 subset is the training set.

In order to assess the model, cross validation was done on both linear models with and without the Age variable. The following graphs are the results from the K-Fold Cross Validation (with K = 3):



Left: Graph depicting cross validation results with 3 folds with linear model:

GoalWeight ~ Measurement + Gender + Age

Right: Graph depicting cross validation results with 3 folds following the linear model:

GoalWeight ~ Measurement + Gender

From these two graphs, two things are observed:

(1) Each folds' estimates follow a similar upward pattern.
(2) Even without the Age variable, the models follow a similar trendline.

To best check the variance among each model, mean squares were calculated. Both models had a mean square of 17.9. This is surprising because when backward selection was performed, Age was shown not to be a significant variable in the model. The mean square of the latter model should be higher. To further check the importance of the Age variable, the function "varImp" in the caret package was used and the results showed that the Age variable had a value of 1.0 which represented a scalar defining a statistical significance.

Objective 2: Predicting Goal Weight w/ Locale Name (Employer)

(Regression Tree)

The overall goal of this objective is to create a method for the insurance company overseeing the data to predict the goal weight of a user by using previous data from the employer. In a sense, this answers whether or not employees of certain companies are trying to lose more weight than others.

To begin, a decision tree is created with GoalWeight as the response variable and Locale Name (employer) as the predictor. Decision trees are a method of constructing prediction models, and since our response is of continuous numeric type, we are creating a regression tree. When the data being analysed contains many features which interact in vastly unknown and complicated non-linear ways, assembling a global model can be difficult. An alternative approach to this 'non-linear' regression is to subdivide or partition space into smaller regions where interactions are more manageable. This partitioning is called recursive partitioning; decision trees use the 'tree' to represent this partitioning .[3] Furthermore, due to simplicity and accessibility of a useful plot, we chose to implement this method.

A subset of the original data is created under the condition that GoalWeight is between 40 and 90 (kg). This range was chosen as it seems to be fairly representative of the general population. A new filter is applied to the data so that it contains only observations from the following locales; CVI Sweden, UK, Fairport, Singapore, France, Scottsville, Espana, Deutschland and West Henrietta. Once the regression tree is fitted to the data, it is pruned to automatically select the complexity parameter associated with the smallest cross-validated error; this is to avoid overfitting. Finally, a plot is created of the regression tree, and is easily readable, showing a distinct prediction ability as expected.

In this analysis, we've only considered different international departments of a single company. The intention is to show that this approach can be used when creating a model for multiple companies of varying backgrounds. Knowing the background and general work type of a potential customer, insurance companies can use this approach on on-hand data to determine

the goal weight predictions thus pre-determine premiums/plans. The result in our case shows that the Singapore department of CVI seeks to lose the most weight out of all departments, as it has the lowest average of goal weight.

Objective 3: Effect of software theme on user activity

(Logistic Regression & Random Forest)

Ideally, Core Health's end users are engaged and active on the site, frequently recording their activity and weight loss progress. However, fostering user engagement is easier said than done. Many factors affect user engagement, one of which is the look and feel of the site. The Core Health platform was built with a very customisable user interface, as they serve many different customers in 8 different countries. Each company has different styling they want implemented on the system, so in order to satisfy the customers while still keeping complexity to a minimum, Core Health has devised different "skins", or themes, that can be applied to alter the look and functionality. CoreHealth is interested in predicting the probability a user is engaged based on the theme of their site.

Logistic regression is used to determine the probability (the odds) of a binary response, using more than one predictor variable. We are interested in predicting the odds a user has been active on the site using just the skin of the site, although the model could be modified to use other predictor variables as well. In CoreHealths case, the binary response variable is "isActive" and is equal to 1 if the user has logged in within the past two weeks and is equal to 0 otherwise. The dataset contains the name of the 8 different themes possible for users, which are used as predictor variables; 5b_Green, Andromeda, Contempo, CoreLite, Freshness, leapfrog, lwl2016, Rotator. Each theme has a different look, **a sample of which have been displayed in the appendix.** The model is fit with the glm function in R using half the dataset for training & half for testing. All themes are significant except Contempo. Surprisingly, this model has a misclassification rate of only 12.9%, meaning we can predict whether a user is active with an accuracy of approximately 87%.

To further ratify the relationship between the software theme and user activity, the method of random forests is applied. Random forests is a classification/regression method that constructs a multitude of decision trees. The main feature of Random Forests that is being sought for here is its ability to give estimates of what variables are important in the classification.

After the application of the random forest model on the dataset containing the IsActive variable and software theme, an importance summary shows that Theme has a IncNodePurity of 1734; which implies that Theme is a significant predictor in regards to the isActive variable. IncNodePurity is the total decrease in node impurities, measured by the Gini Index from splitting on the variable, averaged over all trees. So in essence, the larger the increase in purity the better the predictor is for the model. The random forest model was applied to isActive as the binary-case response, Theme, Gender and Days Per Week Activity as predictors. The importance

table as a result of the fit determines that Theme is the most significant of all predictors, further validating our findings for the logistic regression application.

Objective 4: SBP and Weight claims

(Hierarchical Clustering & K-means clustering)

Myriad medical papers have confirmed the claim that systolic blood pressure has a direct relationship with weight. This claim makes sense, the higher an individual's weight the higher their blood pressure will be as a result. Subsequently, it can be justified that individuals with high blood pressure may be classified as overweight, while the opposite is true for individuals with normal - optimal blood pressure.

The intention with the data on hand is to determine whether specific groups exist within provided SBP values, and whether or not there is a relationship between individual weight and said groups. Our data contains information about user weight and SBP/DBP (systolic & diastolic blood pressure). There are, however, some users that do not disclose all information i.e some users will input measurements for current weight but not for SBP/DBP and vice versa. The first task is to filter the data for only SBP measurements and remove outliers. To remove outliers, a subset of the data is created and a condition is set to include only SBP measurements between 110 and 150 as this is within the range of low-normal, normal - high .[2] Since we are concerned with the existence of groups in this data set, hierarchical clustering can be performed as it will group together data points with the least distance between each other.

As a result, hierarchical clustering shows there exists 2, almost 3 groups within the data. It can be claimed that the groups represent individuals with high, normal and low SBP (which was expected). This observation can be used further...

Now to investigate the direct relationship between SBP and Weight, K-Means clustering is performed on the SBP data with the intention of determining three groups (which is what we determined using H-Clustering). The K-means clustering algorithm will partition the SBP data into three groups; it will first determine three centroids and then assign all observations to the nearest centroid. Once three groups have been created, the mean value for each is calculated and the algorithm is reset. This process continues until no changes occur. A summary of the K-Means Centers shows that the three groups created have mean SBP values of 136, 117 and 97 which all correspond to high, normal and low blood pressure respectively.

A plot of the K-Means clustering results and user weight shows that there is no significant association with weight. Which is interesting as it opposes the claims using real human data. A possible reasoning behind this may be inflated/biased data, since users enter their own measurements themselves.

Shaharyar Sued
Emerson Kirby
Julius Wu
Susheel Palakurthi

Objective 5: Predict the gender of a user given their Goal Weight

(Linear Discriminant Analysis)

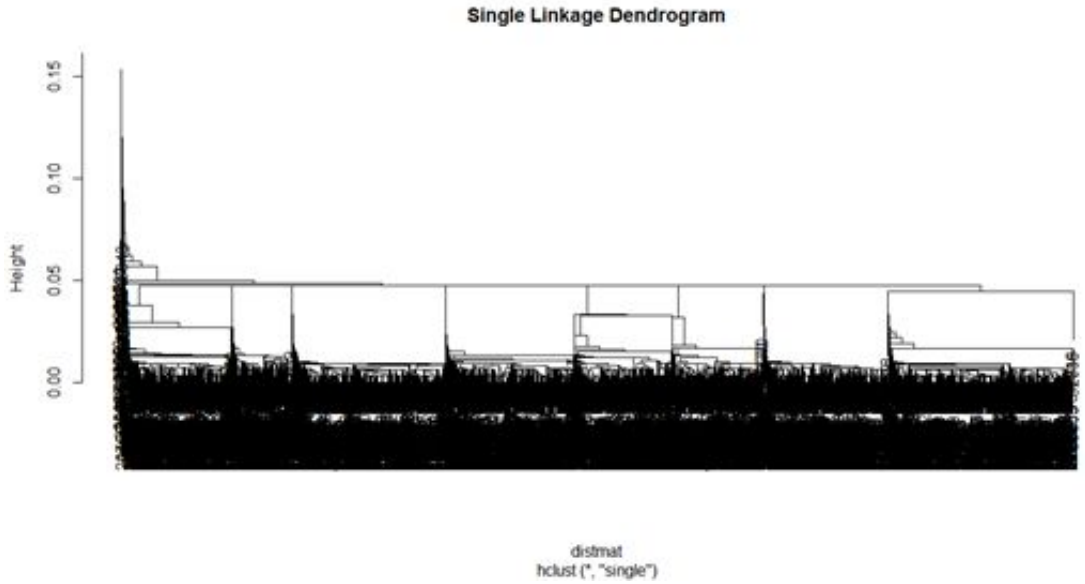
Due to the nature of how data is collected through the software, users are found to not provide information for all variables. It is highly likely that insurance companies will require to know the gender of their clients. This objective seeks to demonstrate a method in which a user's gender can be determined based off of goal weight entries.

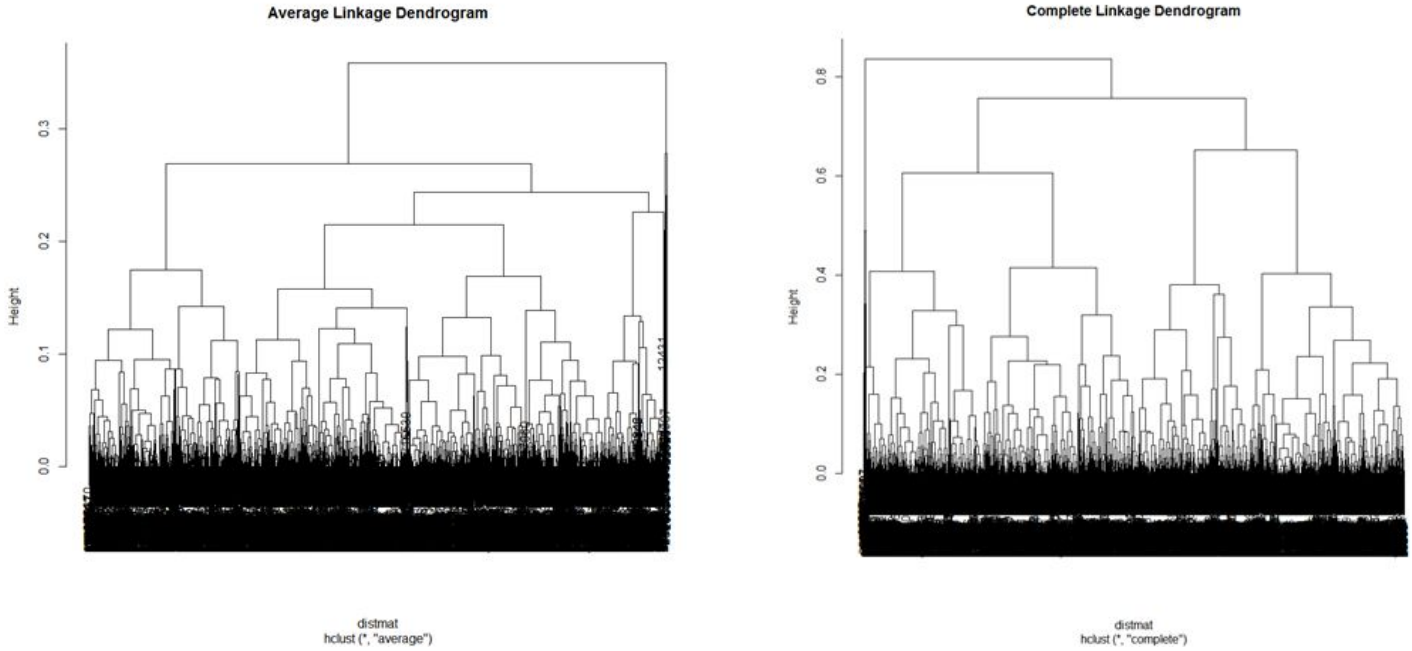
Immediately, it is justifiable to state this problem can be approached with a classification model. Linear Discriminant Analysis is a method used in statistics to find a linear combination of features that characterizes or separates two or more classes of objects or events .[4] The result of LDA, the linear combination, can be used a linear classifier. The goal of classification is to identify which class or group an object belongs to based on its characteristics, and the result of LDA does this by making a classification decision based on the value of a linear combination of the characteristics .[4]

Exploratory Hierarchical Clustering

As part of an initial exploratory analysis, hierarchical clustering was performed in the hopes of teasing out some interesting . The clustering was performed on the users GoalWeight, DailyStepsGoal and DaysPerWeekActivity data with single, average, and complete linkage. Initially, the dataset had 7,289 users who had non-null values for GoalWeight, DailyStepsGoal, and DaysPerWeekActivity. After removing outliers, (users who had a GoalWeight > 90 kg, users who had a GoalWeight <= 40kg, and users who had a DailyStepsGoal > 100000) the dataset contained 6,001 observations. The dissimilarity matrix for this analysis was computed using Gower's metric as we have mix of continuous and discrete variables. Using Gower's metric in R means the data is automatically scaled. Viewing the dendrograms, one may notice that single linkage is not very informative but there appears to be some interesting looking groups in the Complete and Average linkage dendrograms.

Shaharyar Sued
Emerson Kirby
Julius Wu
Susheel Palakurthi





For example, the Complete linkage dendrogram appears to have nice groups at height = 5, which warranted further investigation. However, when the plot of is viewed, there exist no distinguishable groups and this analysis ultimately ends up not providing any useful information.

Shaharyar Sued
Emerson Kirby
Julius Wu
Susheel Palakurthi

Citations:

Unused: Akahoshi, M., et al. *American Journal of Epidemiology: Correlation between Systolic Blood Pressure and Physical Development in Adolescence*. 144 Vol. Oxford University Press, 07/01/1996. Web. 30 Mar. 2017.

1. Song, Young-Hwan. "The Correlation of Blood Pressure with Height and Weight in Korean Adolescents Aged 10-19 Years; The Korean National Health and Nutrition Examination Surveys (2009-2011)." *Korean Journal of Pediatrics* 57.1 (2014): 35–40. *PMC*. Web. 30 Mar. 2017.

2. Ovbiagele B, Diener H, Yusuf S, Martin RH, Cotton D, Vinisko R, Donnan GA, Bath PM, PROFESS Investigators FT. Level of Systolic Blood Pressure Within the Normal Range and Risk of Recurrent Stroke. *JAMA.* 2011;306(19):2137-2144. doi:10.1001/jama.2011.1650

3. Shalizi, Cosma, comp. Regression Trees. Pittsburgh: Carnegie Mellon U, 2006. Statistics 36-350: Data Mining, Fall 2006. Web. 04 Apr. 2017.

4. Martos, Gabriel. *Discriminant Analysis in R*. N.p.: n.p., n.d. Research Techniques. Web. 03 Apr. 2017. <https://rstudio-pubs-static.s3.amazonaws.com/35817_2552e05f1d4e4db8ba87b334101a43da.html>

STAT 311 Project: CoreHealth Technologies

Shaharyar Sued
Emerson Kirby
Julius Wu
Susheel Palakurthi