2012 Michigan Undergraduate Data Mining Competition

Time Use by Employment Status and Gender

Sponsors

Yahoo! Inc.
University of Michigan
Department of Statistics
Department of Informatics

Student Team Members

Christy Duan
Yiqun Hu
Sangwon (Justin) Hyun

OUTLINE

1. Introduction
2. Data Exploration
3. Data Processing
    - Variable selection with heat map
    - Dimension reduction with principal component analysis (PCA)
4. Classification – logistic regression models
    - Model based on variable selection with heat map
    - Model based on dimension reduction based on PCA
5. Analysis of Variance (ANOVA)
6. Discussion

1. INTRODUCTION

The American Time Use Survey (ATUS) is an annual survey that records demographic information and the minutes per day that residents of the United States spend on a variety of activities. The dataset provided is a subset of ATUS data, and includes 13260 subjects from 15 to over 85 years of age.

Motivated by the recent recessions, we choose to explore employment status. Workers are defined as those actively working: Employed (active). Non-workers are defined as those who are not actively working: Employed (leave of absence e.g. disability, maternity, etc.), Unemployed (actively searching for work), Out of workforce (not actively searching for work).

Prior to 2008, there was no significant difference in unemployment rate among men and women.[11] However, beginning in August 2009, the difference grew significantly to 2.7%.[1] To account for this difference in work-life balance between men and women with different employment statuses, we focus on four classes of data: male worker, male non-worker, female worker, and female non-worker.

In this paper, we will demonstrate our approach to classifying ATUS subjects by employment status and gender. In Section 2 and 3, our basic data exploration, variable selection methods and dimension reduction methods will be discussed. In Section 4, we use obtained results to create two logistic regression models to classify the population into the four classes. In Section 5, we conduct analysis of variance (ANOVA) to provide an overview of the specific differences between the four classes. Finally, Section 6 includes our conclusions and remarks.

2. DATA EXPLORATION

After exploring the data with basic summary statistics and histograms, we encounter two main issues:

(1) High Dimensionality. The original dataset is hierarchical and has 390 columns (Tier 3) of very detailed activities, which fall into some larger categories (Tier 2), and can finally be categorized into a total of 18 groups (Tier 1). To reduce the complexity of the data, we focus on Tier 1 data. Since many Tier 3 and Tier 2 columns represent very similar activities, we do not "lose" much information when we use Tier 1 data. In Section 3, we use principal component analysis (PCA) to reduce dimensionality even further.

[1] "The Unemployment Gender Gap During the Current Recession" by Ayşegűl Şahi, Joseph Song, Bart Hobijn (http://nyfedeconomists.org/sahin/GenderGap.pdf)

(2) Predominantly Zero Entries. Histograms show data that is extremely right skewed to zero. Since the total number of minutes in a day is 1440, subjects are limited in the number of different activities they can engage in. This results in a vast majority of zeros in the columns. Also, there may be many zero entries in an activity due to non-response bias. Unfortunately, we cannot separate the non-response from the true zero entries. In the following section, we find a way to address the issue of sparse non-zero responses.

3. DATA PROCESSING

To further reduce dimensionality and address the issue of sparse non-zero responses, we take two approaches: *(1) Variable Selection – Heat Map* and *(2) Dimension Reduction – Principal Component Analysis (PCA)*.

(1) Variable Selection – Heat Map

To reduce the influence of the high proportion of zeros in the columns, we compute pairwise co-occurrence rates for variable selection. High co-occurrence rates indicate that there are activities that people engage in frequently together. Selecting activities with high co-occurrence provides denser information since they have fewer nonzero responses.

We define the "co-occurrence rate between A and B" to be the conditional probability of occurrence of one activity B, given that activity A happens. It is calculated by the equation:

$$Co-occurrence\ rate\ between\ A\ and\ B = \frac{\#\ of\ non-zero\ responses\ in\ activity\ A\ and\ B}{\#\ of\ non-zero\ responses\ in\ activity\ A}$$

Note that the co-occurrence rate between A and B is different from the co-occurrence rate between B and A, which is due to the different number of non-zero responses from A to B. (A complete table is also attached in Appendix)

Therefore, if we consider the co-occurrence rate between Personal Care (t01) and Household Activities (t02), 99.96% of the population who spent time in Personal Care (t01) also spent time in Household Activities (t02). In fact, over 99.8% of people who spent time in Personal Care (t01) spent time in other activities. This motivated us to believe that Personal Care (t01) plays an important role in peoples' lives and it should be a significant variable in our model.

On the other hand, if we take a look at Government Services and Civic Obligations (t10) and Household Activities (t02), only 0.66% of the population who spent time in Government Services and Civic Obligations (t10) also spent time in Household Activities (t02). Likewise, less than 1.5% of people who spent time in Government Services and Civic Obligations (t10) did spent time in other activities. That is to say, Government Services and Civic Obligations (t10) is not a very common activity in peoples' daily lives and it may be reasonable to omit it since it may not be a significant variable in our model.

Figure 1-1 gives the heat map of co-occurrence rate (column-wise) for all Tier 1 activities. For example, the lower right corner is given by the probability that t50 happens provided that t01 happens; while the upper left corner is given by the probability that t01 happens provided that t50 happens. The diagonal is calculated by the number of non-zero responses in activity A divided by the number of non-zero responses in activity A, which is 100%.

Figure 1-2 is a hierarchical clustering heat map, which allows us to visualize the different levels of co-occurrence more easily. Table 1 below gives a summary of these levels.

Selecting the most significant eight activities with high- and medium- co-occurrence rates, we obtain a dataset with a much smaller dimension. We do not lose much information as the eliminated columns represent the more trivial activities in most people's daily life - they do not co-occur with other activities - and will not have a large impact on our work-balance analysis. A classification model will be built on these selected columns in Section 4.

FIGURE 1-1. Heat map of pairwise co-occurence rates for Tier 1

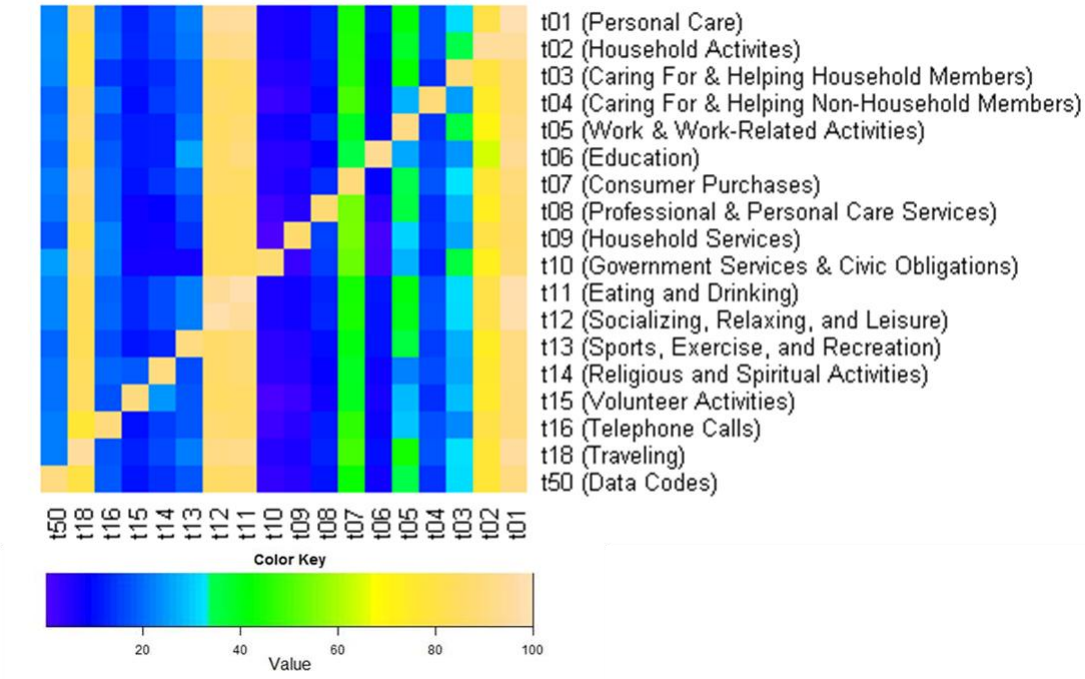


FIGURE 1-2. Hierarchical clustering heat map of pairwise co-occurence rates for Tier 1

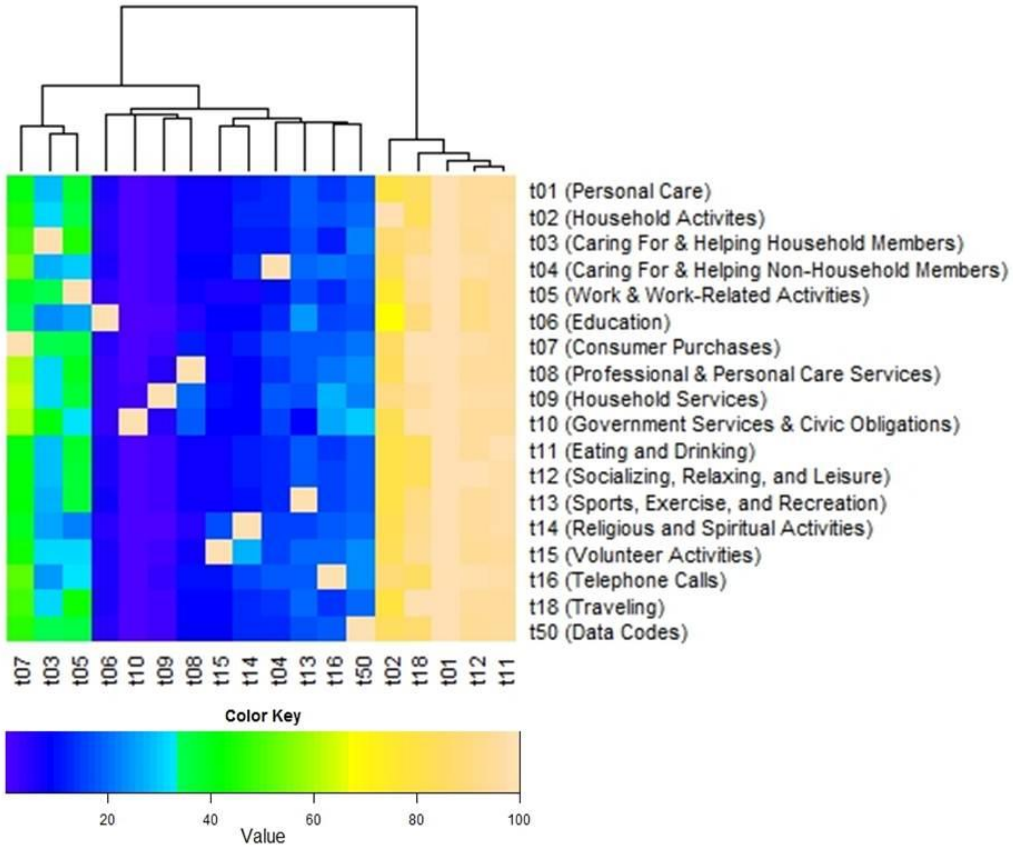


TABLE 1. Tier 1 activities by co-occurrence rate

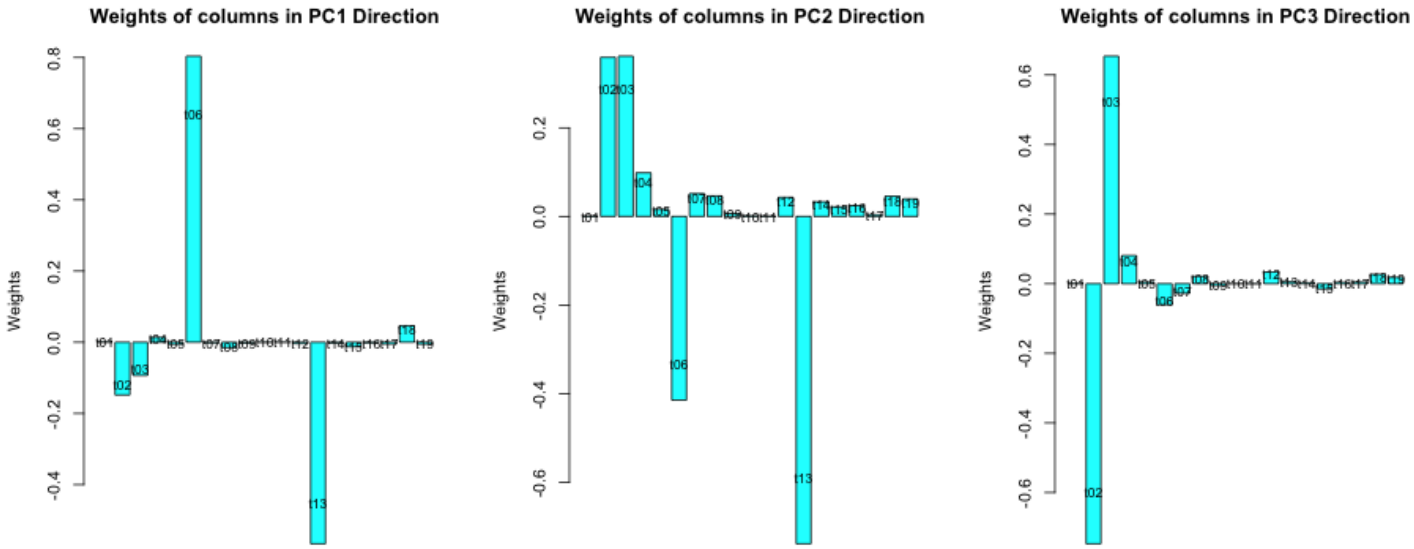
Co-occurrence Rate	High (Yellow)	Medium (Green)	Low (Blue/Purple)
Columns	* Personal Care (t01) * Household Activities (t02) * Eating and Drinking (t11) * Socializing, Relaxing, Leisure (t12) * Traveling (t18)	* Caring for & Helping Household members (t03) * Work & Work-Related Activities (t05) * Consumer Purchases (t07)	t04, t06, t08, t09, t10, t13, t14, t15, t16, t50

(2) Dimension Reduction – Principal Component Analysis (PCA)

Another approach to create a "more manageable" data set is dimension reduction using PCA. This method takes linear combinations of the original variables to create new variables, or principal components (PCs). The transformed dataset has fewer columns than the original, while retaining a large amount of the original variance. These PCs will be used for classification in Section 4. Before we proceed to classification, we make some observations of the PCA results.

Let us focus on the PCA results on Tier 1. Figure 2 shows the weights given to each column when transforming the original data into the first three PCs:

FIGURE 2. Column weights in 3 Principal Components (PCA on Tier 1)



PC1, which accounts for 41.4% of the total variation, weighs Education (t06) against Sports, Exercise and Recreation (t13). PC2, which accounts for 20.07% of the variation, also gives the most weight to Education (t06), and Sports, Exercise and Recreation (t13). However, there is more weight given to Household Activities (t02) and Caring for and Helping Household Members (t03). PC3 accounts for 12.13% of the variation, and gives the most weight to Household Activities (t02) and Caring for and Helping Household Members (t03). Together, these three principal components account for 73.6% of the total variation.

The fact that the weights in are mostly given to the four columns Household Activities (t02), Caring for and Helping Household Members (t03), Education (t06) and Sports, Exercise and Recreation (t13) is noticeable. This information supports our heat map variable selection, where Household Activities (t02) and Caring for and Helping Household Members (t03) were considered significant variables.

For the purposes of classification, we conduct PCA on all tiers, and selected the number of PCs that captured ~70-80% of the total variance in the data [Figure 3, Table 2]. For Tier 1, we chose 3, 4, and 5 PCs. For Tier 2, we chose 5, 7, and 10 PCs. For Tier 3, we chose 9, 12, 16 PCs. In the next section, classification models will be built on these PCs.

FIGURE 3. Cumulative proportion of variation of the first *n* PCs

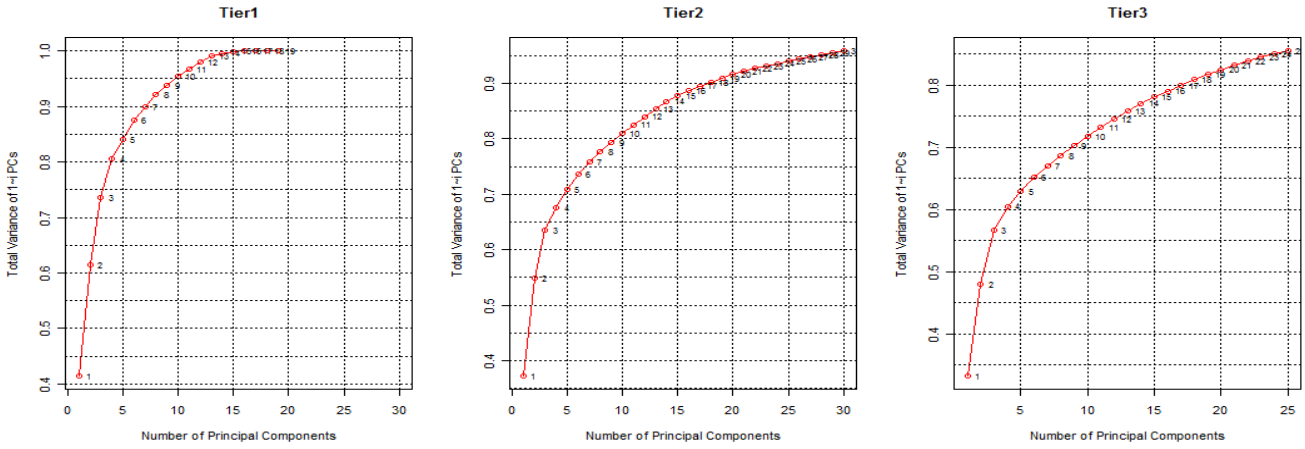


TABLE 2. Cumulative proportion of variation that *n* PCs account for in each tier

	Tier 1			Tier 2			Tier 3		
# PCs	3	4	5	5	7	10	9	12	16
% Variation	73.6%	80.5%	84.1%	70.9%	75.8%	80.9%	70.3%	74.6%	79.1%

4. CLASSIFICATION

For classification, we split the dataset into a training set and test set, which we used consistently for all classification. The training set includes 80% of data randomly selected from each class. The test set includes 20% of data randomly selected from each class. We use logistic regression for a binary classification on all six pairwise relationships among the four classes - male worker (mw), male non-worker (mnw), female worker (fw), and female non-worker (fnw).

(1) Classification on Heat Map Selected Variables

In this approach, we use the eight Tier 1 variables carefully selected from the heatmap in order to conduct logistic regression. Table 3 shows the classification error rates on the training and test sets:

TABLE 3. Classification error rates from logistic regression on heat map selected variables

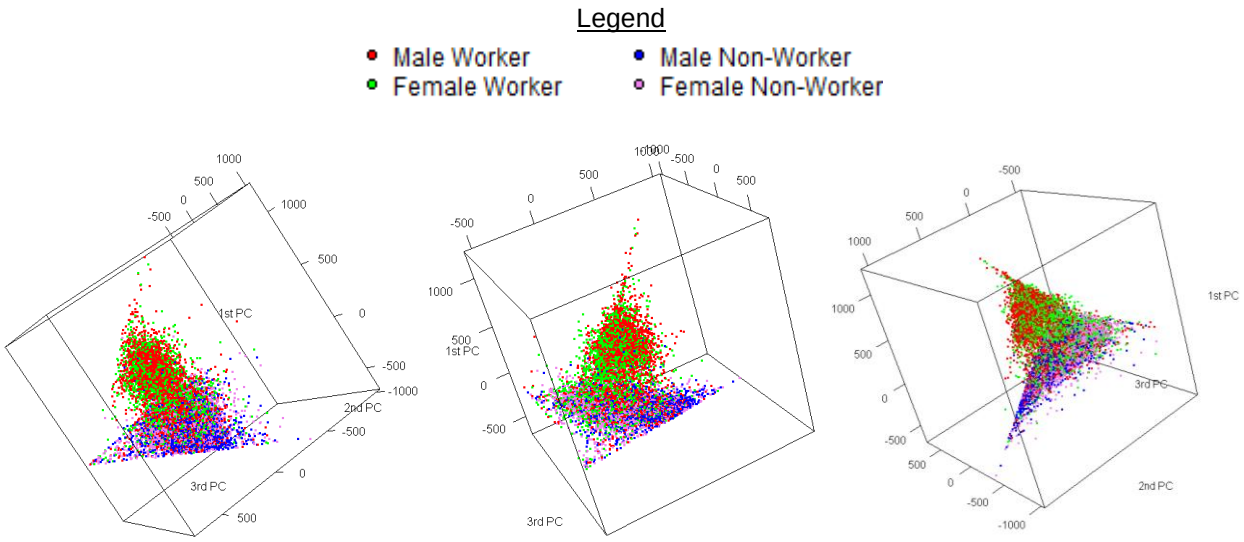
	mw-fw	mw-mnw	mw-fnw	fw-mnw	fw-fnw	mnw-fnw
Training Error	40.7%	23.5%	21.8%	22.6%	25.4%	35.2%
Test Error	40.6%	23.9%	23.5%	23.9%	27.9%	34.2%

(Note: mw:male worker, fw:female worker, mnw: male nonworker, fnw: female nonworker)

(2) Classification on Principal Components (PCs)

In this approach, we use the selected PCs from Tier 1, Tier 2, and Tier 3 data to conduct logistic regression on the PCs for a binary classification on all six possible pair-wise relationships among the four classes.

FIGURE 4. Different perspectives of covariance PCA Tier 1 data projected onto the 1st, 2nd, 3rd PCs



In Figure 4, the 3-dimensional graphical is shown from different perspectives. There is a clear difference between the locations of the employed persons' data points and unemployed population. There is also some sense of division between genders *within* the employed or unemployed data points, although less apparent.

We record the training and test error to evaluate the performance of our classification in Table 4:

TABLE 4. Classification error rates averaged over logistic regression results from nine different PCs for Tier 1, Tier 2, Tier 3

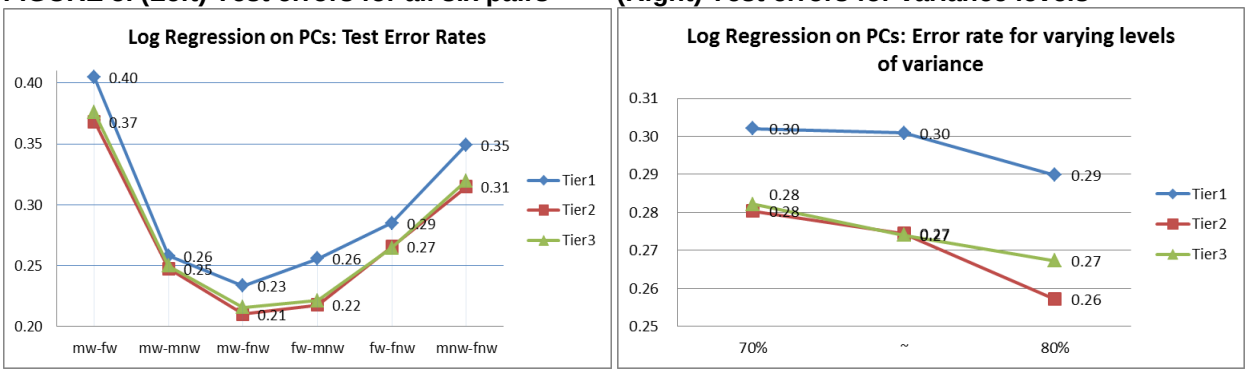
	mw-fw	mw-mnw	mw-fnw	fw-mnw	fw-fnw	mnw-fnw
Training Error	37.4%	24.1%	20.8%	20.5%	24.7%	33.3%
Test Error	38.3%	25.2%	21.9%	23.2%	27.2%	32.8%

(Note: mw:male worker, fw:female worker, mnw: male nonworker, fnw: female nonworker)

Classification Results

From Table 3 and 4 we can see that the classification errors are within an acceptable level. Classification using PCs tend to have slightly lower error rate compared to heat map selected variables. Overall, there is a consistent pattern between the two approaches – classification for workers versus non-workers perform the best regardless of gender, with test errors ~23%. Classification between workers of different genders, and classification between non-workers of different genders performs slightly less well, with test errors ~35%. Also, the error rates drop uniformly as you use more PCs(account for more variation).

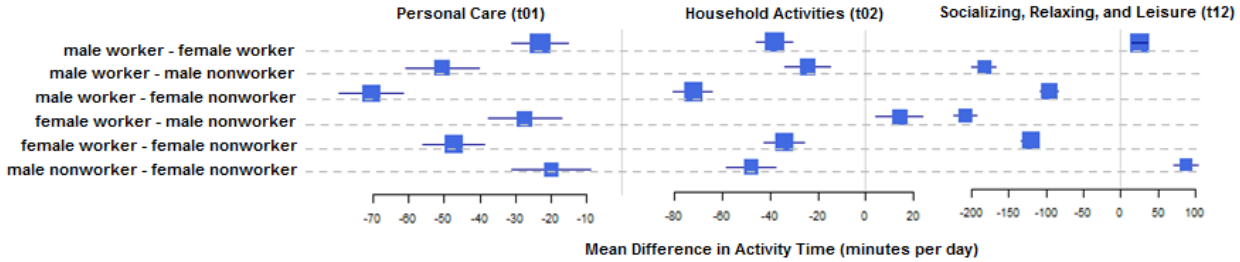
FIGURE 5. (Left) Test errors for all six pairs (Right) Test errors for variance levels



5. ANALYSIS OF VARIANCE (ANOVA)

We use analysis of variance (ANOVA) to test for significant differences in mean activity time spent by male workers, female workers, male non-workers, and female non-workers at $\alpha = 0.05$. Due to unequal variances, Welch's t-test and the Games-Howell post-hoc test were used. There were significant differences for all of the activity times except for Household Services (t09) and Government Services & Civic Obligations (t10). Additionally, ANOVA confirms that the eight variables selected from the heat map in Section 3 are significant.

FIGURE 6. Forest plot of 95% confidence intervals and mean differences in Tier 1 activity times



For instance, the results show that male workers spend an average of 50.593 minutes/day more than female workers for Work & Work-Related Activities (t05). From the forest plot in Figure 4, we see that even though male workers work longer than female workers, female workers continue with 38.379 more minutes/day of Household Activities (t02) labor than male workers, on average. Finally, we see that male workers spend an average 24.722 more minutes/day on Socializing, Relaxing, and Leisure (t12) than female workers. Female workers choose to spend their non-working time differently. Instead, they spend an average of 23.151 more minutes/day on Personal Care (t01) than male workers.

The tests show us general trends in activity times for these classes in Table 5. If relationships are not significantly different, they are grouped as "workers", "non-workers", "female", "male", or "all other classes." (Refer to Table 6 in the appendix for specific mean differences) :

TABLE 5. General trends in activity times for each class

Activity	Time Use Trend
Personal Care (t01)	female non-workers > male non-workers >female workers > male workers
Household Activities (t02)	female non-workers > female workers > male non-workers > male workers
Caring For & Helping Household Members (t03)	female > male workers > male non-workers
Caring For & Helping Non-Household Members (t04)	non-workers > workers
Work & Work-Related Activities (t05)	male workers > female workers > male non-workers > female non-workers
Education (t06)	male non-workers > female non-workers > workers
Consumer Purchases (t07)	female>male
Professional & Personal Care Services (t08)	female non-workers > all other classes
Eating & Drinking (t11)	all other classes > female workers
Socializing, Relaxing, and Leisure (t12)	male non-workers > female non-workers > male workers > female workers
Sports, Exercise, and Recreation (t13)	male non-workers > male workers > female
Religious and Spiritual Activities (t14)	female non-workers > all other classes
Volunteer Activities (t15)	female non-workers > all other classes
Telephone Calls (t16)	female non-workers > female workers & male non-workers > male workers
Traveling (t18)	workers > non-workers

6. DISCUSSION

Based on the results that we obtained from the classification models and ANOVA analysis, we saw clear evidence of significantly different work-life balance among the four classes of people. We also obtained good results for our logistic regression classification, for variable selection by heat maps and dimension reduction by PCA. However, there are limitations:

The variable selection method based on co-occurrence rate is subjective. Even though these variables were chosen appropriately for the logistic regression model, more rigorous variable selection criteria must be developed.

Also, there is a trade-off between variable selection by heat maps and dimension reduction by PCA. While selecting highly co-occurring columns is easier to do, it gives a slightly higher classification error. On the other hand, classifying using PCs, demand a slightly more complicated analysis at the first stage, but gives a better classification results. Also, interpretation of the columns becomes more difficult when PCA is used – we can observe the PCs and column weights (which are interesting enough) and possibly use them to support some conjectures, but alone are not enough to draw conclusions.

This paper is a good starting point for future analysis that compares the behaviors of male workers, female workers, male non-workers, and female non-workers during a recession. We recommend that, in the future, time series analysis be conducted using more data from ATUS to observe trends.