2012 Michigan Yahoo! Data Mining Competition

Alexander Flood, Albert Tan, Lo-Hua Yuan

American Time Use Survey
We spend our weekends mining data.
What do YOU do for fun?
Outline

1) Visual of how American people spend their time
   Heatmap illustrates people’s preference to spend more time socializing than working, exercising, or partaking in religious activities

2) Clustering people as more prone to work or socialize
   Gaussian mixture model (GMM) clustering on principal components segregates people into two broad personality types—what we term work-prone versus social-prone

3) Predicting time-use via classification methods
   Logistic regression, K-nearest neighbor (KNN), Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA) methods predict whether an individual is likely to spend more time working or socializing, based on demographic data

4) Limitations
   Missing values and sparsity in the data, and inferences that can be made from clustering

5) Concluding Remarks

Method

We use the 2010 American Time Use Survey dataset to show it is possible to use demographic data to predict whether a person is likely to spend more of his/her time awake working or socializing.

The 2010 American Time Use Survey dataset contains entries from 13260 participants, each with 11 demographic variables and information about time spent doing various activities. These activities are grouped into 18 major categories that are further separated into sub-tier events, giving a total of 390 specifically identified activities. However, the majority of the time-use data is very sparse.

To expedite our analysis, we condensed the sub-tier data into their respective major categories and focused on four main activity groups which contain the largest number of non-zero time entries:

1) Work & Work-Related Activities, 2) Socializing, Relaxing, and Leisure, 3) Sports, Exercise, and Recreation, and 4) Religious and Spiritual Activities. We refer to these categories as Work, Socializing, Exercise, and Religion through the remainder of this report.

In order to compare a particular individual’s time use with that of another, we performed all our analyses on the percent time awake (PTA) spent on activities.
1) Visualizations

The heatmap illustrates different proportions of time people spend on Work, Socializing, Exercise, and Religious activities. The high density of light, peach colors in the Socializing column shows that most people tend to spend more of their awake time engaged in Socializing than the other three activities. The second lightest-colored column is the Work column. Colors in Work tend to complement those in Socializing. This suggests that people who do not spend as much time socializing (in comparison to others included in the survey) spend their time working instead. Given that this heatmap suggests a possible negative correlation between Work and Socializing, we proceed to further explore the relationship between these two activities using PCA and clustering methods.

Figure 1: Heatmap of different proportions of time people spend on Work, Socializing, Exercise, and Religious activities
2) PCA and Gaussian Mixture Clustering

Dimension reduction via PCA provides another perspective on the relationship between the total number of minutes awake over the past 24 hours (time_awake) and the percent time awake (PTA) doing Work, Socializing, Exercise, and Religion. The loadings of the first principal component supports the trend observed in the heatmap that the PTA spent Working and Socializing are weighted in opposition to each other.

Even though the second principal component has a large loading for exercise, we note that the clusters as seen in Figure 2.2 are minimally influenced by the second principal component.

<table>
<thead>
<tr>
<th>Loadings:</th>
<th>Comp.1</th>
<th>Comp.2</th>
<th>Comp.3</th>
<th>Comp.4</th>
<th>Comp.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>time_awake</td>
<td>-0.476</td>
<td></td>
<td></td>
<td></td>
<td>0.873</td>
</tr>
<tr>
<td>work</td>
<td>-0.639</td>
<td>0.106</td>
<td>-0.127</td>
<td>-0.289</td>
<td>-0.694</td>
</tr>
<tr>
<td>socialize</td>
<td>0.594</td>
<td>0.116</td>
<td>-0.313</td>
<td>0.375</td>
<td>-0.628</td>
</tr>
<tr>
<td>exercise</td>
<td>-0.901</td>
<td>0.334</td>
<td>0.105</td>
<td>-0.256</td>
<td></td>
</tr>
<tr>
<td>religion</td>
<td>0.111</td>
<td>0.399</td>
<td>0.868</td>
<td>-0.126</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.1: Loadings of principal component analysis on time awake and PTA spent on work, socializing, exercise, and religion

Figure 2.2 below shows unsupervised Gaussian mixture model (GMM) clustering applied on the first two PCA principal components. We chose to cluster the data into two groups, in hopes of seeing one group correspond to people who tend to work more, and another group correspond to people who spend more time socializing. The way the two clusters in Figure 2.2 split around the center of the X-axis shows that our expectations were appropriate.

Figure 2.2: GMM clustering on the first two PCA principal components
Next, we compare the four main activity types between the two clustered groups:

![Boxplots](image)

**Figure 2.3:** Boxplots of PTA spent on Working, Socializing, Exercise, and Religion, and total time awake over the course of 24 hours, across Group 1 and Group 2

Between the two cluster groups, there appears to be a sizeable difference in the distributions of PTA spent working versus socializing. Specifically, Group 1 spends more time socializing while Group 2 spends more time working. In contrast, there do not appear to be distinct differences in the mean time spent exercising or participating in religious activities between the two clustered groups. Hence, based on these observations, we set to investigate the possibility of classifying people as either more work-prone or more social-prone based on demographic data.

### 3) Predicting a person’s time-use

#### 3.1. Generating the indicator variable for work-prone and social-prone

We first rescale and recenter the PTA spent on working and socializing to have mean zero and unit variance. The purpose of rescaling and recentering is to enable us to compare an individual’s time awake spent socializing and working relative to the rest of the population. For instance, if an individual’s PTA spent socializing is greater than his PTA spent working, but his socializing time is less than the population average, then he would be classified as work-prone rather than social-prone\(^1\). It should be noted that individuals who are unemployed or out of the workforce are removed before generating the indicator variable to improve the robustness of our analysis by reducing the bias towards classifying more people as social-prone. Individuals who are unemployed or out of the workforce would appear to be social-prone because they work 0 hours a week.

\(^1\) The average PTA spent socializing is higher than working
3.2. Classification analysis

Classification analysis was employed to predict, based on demographic data, whether an individual is social-prone or work-prone. The demographic attributes used are age, gender, urban (i.e. whether the individual lives in a city or suburb/rural area), working status, working time, household size, and number of household children.

The classification methods we used are logistic regression, KNN, LDA, and QDA. For the logistic regression, we used the AIC criterion to select the best variables to predict whether someone is work-prone or social-prone. The AIC criterion selected the model with the following predictors: female, employed but on leave of absence (e.g. maternity leave, disability), work hours, and household size. Table 3.1 summarizes the results from the logistic regression analysis based on the model selected by the AIC.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient Estimates</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.652521</td>
<td>&gt; 0.001</td>
</tr>
<tr>
<td>Female</td>
<td>-0.101992</td>
<td>0.1437</td>
</tr>
<tr>
<td>On leave of absence</td>
<td>-2.124360</td>
<td>&gt; 0.001</td>
</tr>
<tr>
<td>Work Hours</td>
<td>0.018051</td>
<td>&gt; 0.001</td>
</tr>
<tr>
<td>Household Size</td>
<td>0.046705</td>
<td>0.0371</td>
</tr>
</tbody>
</table>

Table 3.1: Logistic regression analysis results: The estimated coefficients for the variables and their corresponding p-values. The base class is work-prone.

From Table 3.1, we see that females tend to be more social-prone than males. In addition, those employed but on leave of absence have a higher tendency to be social-prone compared to those who are actively employed. Both work hours and household size are estimated to have a positive effect on the probability of an individual being work-prone. The observation on household size can be attributed to the fact that household size is a proxy to number of dependent children in the household; the correlation between household size and household children is roughly 0.85. Hence, the employed members of the household are more work-prone because they need to cover the expenses of their children.

<table>
<thead>
<tr>
<th>Model</th>
<th>Logistic</th>
<th>KNN</th>
<th>LDA</th>
<th>QDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction error</td>
<td>0.4210526</td>
<td>0.4681221</td>
<td>0.4278438</td>
<td>0.4499151</td>
</tr>
</tbody>
</table>

Table 3.2: The prediction error of various classification models

Next, we compared the prediction error rate between logistic regression, KNN, LDA, and QDA. (Using leave-one-out cross-validation, we chose 25 neighbors for KNN.) Table 3.2 shows the prediction error rates for the four models. It appears that linear classifiers (logistic and LDA models) appear to work better than a non-linear classifier (QDA) and a non-parametric classifier (KNN). The logistic classification model yields the lowest test error while KNN yields the highest classification error. The prediction errors of the linear classifier models are moderately less than 50%. This implies that Yahoo! can employ linear classification methods on demographic data to better predict whether someone is social-prone or work-prone.
4) Limitations

4.1. Data

Missing values
- There appears to be missing values on demographic data and time-use variables. We removed approximately forty percent of the observations that contain these missing values to facilitate our analysis. However, the remaining observations may not be representative of the whole population.

Sparsity of the data
- The data is sparse, at best. To reduce the sparsity of the data, we combined various time-use variables into their major categories. Hence, in the process we may have lost important information on the individual time activity variables.

Data collection method
- Respondents to the survey were required to document exactly 1440 minutes in their responses (twenty-four hours) across almost 400 activities. This may have influenced individuals to change their responses to satisfy this constraint.

4.2. Data Analysis

Clustering error
- Since the GMM clustering was performed on the principal components, some individuals in the “work-prone” cluster may have not actually worked very much. For instance, time awake also had a large negative loading in the first principal component, so an individual who does not spend a lot of time working, but also doesn’t sleep much, may end up in the work-prone cluster.

5) Conclusion

We find that demographic variables such as age, gender, work status, household size, etc., do in fact help to predict whether someone is social-prone or work-prone. This suggests that Yahoo can leverage readily accessible demographic information on their customers to more effectively target advertising and offers. In addition, we also find that linear classification models appear to work best for prediction.