Trends in Computer Usage for Profiling

Emanuelle Alemar, Jonathan Che, Joshua Core, Mishal McNeill
Carnegie Mellon University, Pittsburgh, Pennsylvania

Introduction

As our culture progresses further into the digital age, we will need better forensic tools to identify computer-based crimes. Our project focuses on applying statistical methods to determine the user of a computer. We studied people's computer usage patterns to find both general trends and deviations from the norm. We hope that our work can contribute a greater understanding to computer usage that may be useful from a forensic standpoint.

Research Questions:

- Can we determine general trends in productive computer use throughout the day?
- Can we build a model that predicts who was on the computer in a given hour of observations?

Materials & Methods

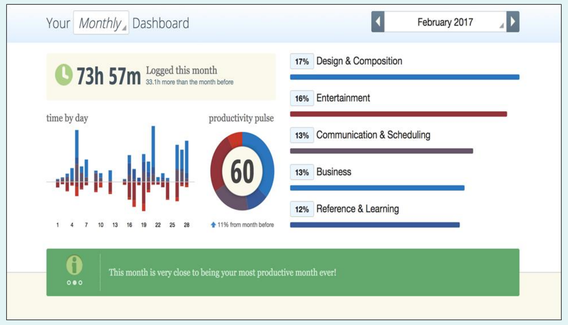


Figure 1: Screenshot of RescueTime dashboard. RescueTime is a productivity-tracking application that records computer usage.

Data Collection:

- Computer data was collected from 12 CMU summer statistics participants via RescueTime.
- Data was collected from June 20th – July 4th, 2017 (15 days).

Productivity:

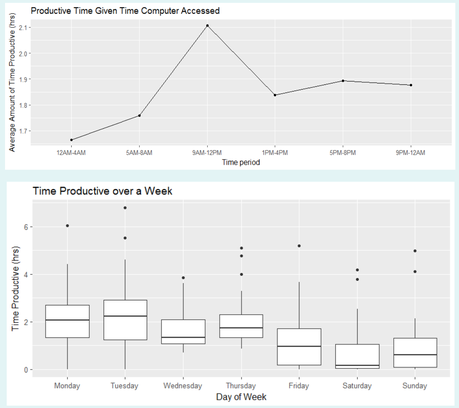
- For each day, we studied information about times the computer was used and the amounts of productive and unproductive time.
- We fit a linear regression model to predict the amount of productive time.
- We modeled percent productivity using linear regression with a logistic transformation.

Profiling:

- For each hour, we recorded information about productive time, unproductive time, and the type and number of different activities.
- We fit 12 logistic regression models to predict each of the users.
  - Each model asked the question: "Given an hour of data, is this Participant X?"
- We investigated model performance and error tradeoffs in different ways.

Productivity Relationships

We defined productive time as time spent on applications and websites typically associated with productive tasks (e.g. software development, email, etc.).



Figures 2a & 2b: Visualizations of productive time by hour and day

If a user is on the computer at all from 9AM to 12PM, they tend to have more productive time on the computer for the entire day. Also, in general, weekdays are more productive than weekends.

Productivity Modeling

Model 1: Hours of Productivity ~ Participant ID + Total Time + Accessed 9AM-12PM + # of Days into Week

Variable	Estimate	P-Value
(Intercept)	-0.848	0.095
Total Time	0.294	0.001
Accessed 9AM-12AM	0.427	0.215
# of Days into Week	-0.031	0.629

There is a significant positive relationship between total time on the computer and productive time.

Model 2: logistic(% Prod.) ~ Participant ID + Total Time + Accessed 9AM-12PM + # of Days into Week + (Total Time * Accessed 9AM-12PM)

Variable	Estimate	P-Value
(Intercept)	-3.10	1.04E-06
Total Time	0.319	1.78E-03
Accessed 9AM-12PM	2.47	6.96E-07
# of Days into Week	-0.204	1.77E-04
Total Time* Accessed 9AM-12PM	-0.378	7.67E-03

Users who are on the computer from 9AM to 12PM tend to have a higher % prod. However, as these users spend more time on the computer, their rate of productivity eventually decreases. Also, as the week goes on, productivity rates tend to decrease.

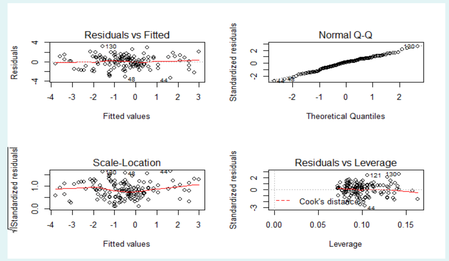


Figure 3: Regression assumption plots for Model 3a

Profiling Users

Variable	Number of Models that used Variable
Hour	3
Longest Activity Time	11
Number of Activities	9
Productive Time	6
Total Time	10
Unproductive Time	11
Used Google Chrome	11

Variable	Number of Models that used Variable
Used Windows Explorer	9
Visited Facebook	10
Visited Gmail	11
Visited Google	6
Visited Outlook	10
Visited Reddit	10
Visited Youtube	10

Tables 4a & 4b: Variables used by logistic regression models

The 12 logistic regression models were fit stepwise, so they could use different variables. In general, though, most of the models found similar variables to be useful.

Evaluating Profiling Models

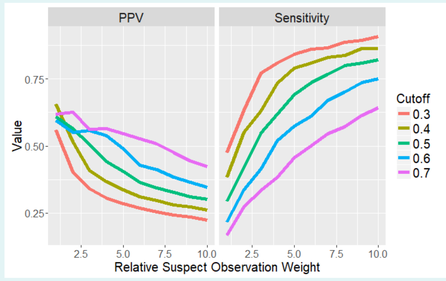


Figure 5: Average performance of models by weight and cutoff

Positive Predictive Value (PPV): How often observations classified as user X are actually from user X.

Sensitivity: The proportion of observations from User X that are correctly identified as such.

There is a tradeoff between PPV and sensitivity. Adjusting the weights assigned to observations of user X and the classification cutoff for logistic regression can control this tradeoff. In a forensic context, we prefer a higher PPV at the expense of sensitivity to avoid false accusations.

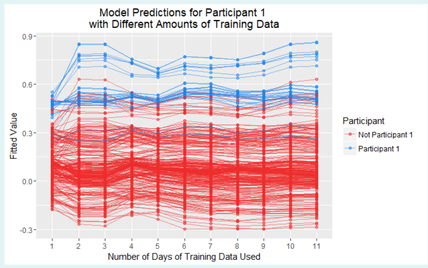


Figure 6: Model performance with different training set sizes

After about two days of training data, the fitted values predicted by our model do not change or improve with more training data. This is generally true for all 12 models. This indicates that user habits may not significantly vary on a day-to-day basis.

Discussion

In our project, we looked into common computer usage trends as well as methods to differentiate between users. In general, people seem to be more productive in the morning and early in the week. We also found that more computer activity overall is associated with a higher percent of productive computer usage. Our findings are quite intuitive, and match many of our expectations.

Despite the fact that people's habits appear to be similar, our models still do a reliable job of distinguishing between users. This indicates that there are individual computer usage habits that differ between users. The variables chosen by stepwise regression seem to indicate that while users visit similar sites, the way in which they do so is specific to them.

Finally, we looked into the feasibility of using profiling models like ours in real-world applications. We found that we only need approximately two days of data for our models to recognize trends, which would be good for forensic problems where data may be limited. We also examined the tradeoff between PPV and sensitivity, finding that observation weights and regression cutoffs can be intuitively changed to optimize for any desired combination of PPV and sensitivity.

Limitations/Next Steps

Limitations:

- Small, nonrandom sample
- Limited information collected by RescueTime

Next Steps:

- Larger random population samples
- Collecting more detailed computer usage data
- Implementing different classification algorithms

Conclusion

- Observing productivity trends produced useful factors to distinguish users.
- Studying percent productivity was more effective than just studying the amount of time productive.
  - Users are typically more productive in the morning and early in the week.
- Fitting a model to each user is effective for profiling.
  - PPV and sensitivity are good measures for evaluating the performance on these models.
  - Users differ in their website-visiting habits even if they visit the same websites.
  - User habits can be identified after two days of training data.

If there are any further questions, feel free to contact Jonathan Che at jche18@amherst.edu

Acknowledgements

We would like to thank our TA, Ciaran Evans, for all of his assistance, as well as the CMU Statistics Department, CSAFE, and NIST.