Capstone Proposal

Lára Hrönn Pétursdóttir

December 12th 2016

All the people of earth have this in common: One day is one earth's rotation. None has more or less time in a day - just more or less sunlight.

The project's domain background —
We have probably all heard: "I´m an "A" type or "B" type". But are we really?
There are many forms used to divide people into personalities or types. Some as simple as categorising people in to A, B and even C types and other more complicated. Known within the field of psychology is for example the Psychometric Evaluation of the Revised Temperament and Character Inventory (TCI). TCI is a set of tests designed to identify the intensity of and relationships between the seven basic personality dimensions of Temperament and Character, which interact to create the unique personality of an individual. Basic personality factors are important predictors of risky health behaviours where they often result in a lifestyle more open to certain health issues

In this field (and others) it could be helpful to find simplified labels that represent how people spend their days. That would help include the influence of lifestyle in wide array of research and evaluations.

http://www.simplypsychology.org/personality-a.html
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2810834/
http://psychobiology.wustl.edu/what-is-the-tci/
http://www.sciencedirect.com/science/article/pii/S0191886901000800

A problem statement — Can we be divided into groups / types / personalities based on how we spend our days? Could these groups be used as labels for further research to see how your lifestyle affects different aspects of our life such as health, family, happiness or even consumer behaviour.

Other interesting questions to answer include:

How do daily activities differ by:
- labor force status
- income
- household composition
- geographical region
- disability status

The datasets and inputs —
The American Time Use Survey (ATUS) is the Nation's first federally administered, continuous survey on time use in the United States. The goal of the survey is to measure how people divide their time among life's activities.

In ATUS, individuals are randomly selected from a subset of households that have completed their eighth and final month of interviews for the Current Population Survey (CPS). ATUS respondents are interviewed only one time about how they spent their time on the previous day, where they were, and whom they were with. The survey is sponsored by the Bureau of Labor Statistics and is conducted by the U.S. Census Bureau.

The major purpose of ATUS is to develop nationally representative estimates of how people spend their time. Many ATUS users are interested in the amount of time Americans spend doing unpaid, nonmarket work, which could include unpaid childcare, eldercare, housework, and volunteering.

The survey also provides information on the amount of time people spend in many other activities, such as religious activities, socializing, exercising, and relaxing. In addition to collecting data about what people did on the day before the interview, ATUS collects information about where and with whom each activity occurred, and whether the activities were done for one's job or business. Demographic information—including sex, race, age, educational attainment, occupation, income, marital status, and the presence of children in the household—also is available for each respondent. Although some of these variables are updated during the ATUS interview, most of this information comes from earlier CPS interviews, as the ATUS sample is drawn from a subset of households that have completed month 8 of the CPS.

https://www.kaggle.com/bls/american-time-use-survey
;

A solution statement — Create simplified labels for people depending on how they spend their days. That can be used for different applications (mental and physical healthcare, education, marketing and so on).

A benchmark model —
Depending on the amount of clusters if Friedman's & Rosenman's research on Personality types will be used or if TCI will be used to compare the results with the seven dimensions of personality traits.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3477961/
http://hendrix.imm.dtu.dk/services/jerne/brede/WOEXT_71.html

A set of evaluation metrics —
R^2 - To evaluate the how features fit the data
Explained Variance - to measure how much variance in the data is explained be the components.
Silhouette score - to determine how many clusters are optimal
Sample points - predict to which cluster a person belongs to and discuss if that is reasonable

An outline of the project design —
I will base my project on the Udacity project Customer Segments

- Data Exploration
  - What information does it contain
  - Statical description of the dataset
  - Distribution and relevance and correlation
  - Deal with NaN
  - Remove features if necessary
  - Merging with other dataset if necessary

- Data Preprocessing
  - Feature and scaling
  - Identify outliers and decide if they should be removed or not

- Feature Transformation
  - Apply PCA
  - Dimensionality Reduction and use cumulative explained variance ration to decide how many dimensions are necessary.
  - Visualize a Biplot
  - Clustering, use silhouette coefficient to decide on amount of clusters (or how many "types of people" there are depending on how they spend their 24 hours)
  - Cluster Visualization
  - Data Recovery - find the cluster centre points that correspond to the "types of people".