UNIVERSITY OF TARTU
Institute of Computer Science

Data Mining
(MTAT.03.183)

Homework Assignment XI

1: (1 points) Business understanding

NB! Don't forget to mention your project title and team members in the beginning of the report: I am working alone on project: *Your mobile can recognize your physical activity!*

Developing a business understanding within CRISP-DM consists of four tasks: identifying your business goals, assessing your situation, defining your data-mining goals and producing your project plan. For this exercise, please, develop a business understanding of your project. According to CRISP-DM you should report the following:

- Identifying your business goals
  - Background
  - Business goals
  - Business success criteria
- Assessing your situation
  - Inventory of resources
  - Requirements, assumptions, and constraints
  - Risks and contingencies
  - Terminology
  - Costs and benefits
- Defining your data-mining goals
  - Data-mining goals
  - Data-mining success criteria

Please, follow this given structure and cover all these aspects in your report. Consult this PDF-file with the chapter on Embracing the Data-Mining Process for more information on each of the deliverables. Keep the report concise and feel free to state that some aspect is not relevant in your project. If your project is not meant to benefit a business, then please specify who will benefit from the project and perform business understanding from their perspective. For instance, this could be either one or multiple individuals, organisations, or societies.

Solution:

As the question is recommended, in this task, I answer the items listed above respectively,

Identifying your business goals:

- *Background*: It is undefinable that all of us are care about our healthy and fitness. During last years, technological companies such as Sumsung and Apple and etc are presented some applications for smart pones who mange to assistant us to reach the mentioned goals easier. These softwares uses some gathered data from our body and perform some data analysis in this the gathered data. This project aims to give a practical view of this kind of analysis.

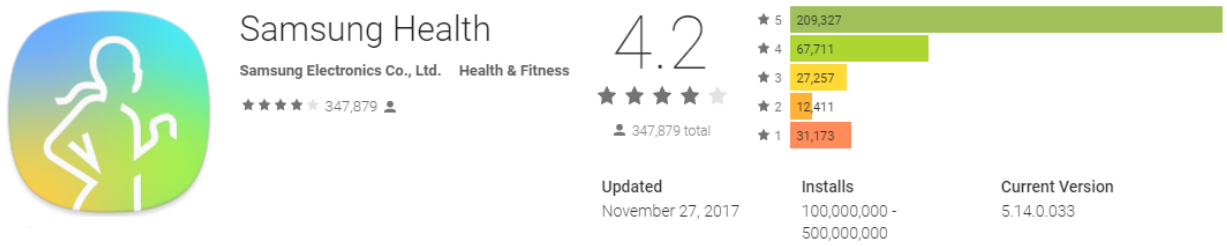

Figure 1: Samsung S-Helath page on GooglePlay (12/10/2017).

- *Business goals:* This is an academic project, but clearly its result could be used to increase life quality. Actually similar softwares are installed on most of the recently produced smartphones. For example, See Fig. 1.
- *Business success criteria*: The mentioned applications can be installed on the smart phones and number of installation and their score show the popularity and success of the applications. For instance see Fig. 1 for *Sumsung S-Health* application (10M install and Score 3.8/5).

Assessing your situation:

- *Inventory of resources:* I use the dataset *Run or Walk* from Kaggle website which is available on `https://www.kaggle.com/vmalyi/run-or-walk`. I use R for my analysis on data.
- *Requirements, assumptions, and constraints:* This is an academic project as an assignment for a course, but in the real case issues such as legal and security obligations, and etc should be considered.
- *Risks and contingencies:* Again I should emphasize that in the real cases, one need to make clear causes and reasons which might led to finish the project with delay. For this case, I think the lack of time might be a possible reason to late delivery.
- *Terminology:* For business terms, we have some terms such as *physical activities* which refer to any body movement that works your muscles and requires more energy than resting; e.g. Walking, Running. For data-mining terms I will mostly use the common analysis which we have learned in the lecture.
- *Costs and benefits:* These softwares are presented by companies free of charge and by now they are installed by more 100M people which has direct affect in companies salary. Because they are installed on recently produced Samsung smart phones which shows that they had large number of selling in last few years. One can predict an approximate benefit with some calculations.

Defining your data-mining goals:

- *Data-mining goals:* We will get to a model which can recognize our physical activity using data gathered with our smart phone.
- *Data-mining success criteria:* In this project as it can be gusset from its name, we mainly will use classification and clustering techniques to predict for new data.

2: (2 points) Data understanding

Data understanding within CRISP-DM consists of performing four tasks: gathering data, describing data, exploring data and verifying data quality. For this exercise please develop a data understanding of your project. Report the results of the tasks according to the following structure:

- Gathering data
  - Outline data requirements
  - Verify data availability
  - Define selection criteria
- Describing data
- Exploring data
- Verifying data quality

Consult the given book chapter to understand what is expected under all these deliverables. Take inspiration from when describing and exploring the data. As a result of this exercise you should have gathered and understood the data. You should have decided which parts of the data you are potentially going to use and understood the meaning of all fields within these parts. Note that data cleaning is part of the data preparation step in CRISP-DM but you might choose to do some of it already during this task.

Solution:

Similar to previous question, I answer the items brought from CRISP-DM about the date set which will be used in my project.

- **Gathering data**
  - *Outline data requirements:* There are different sensors on smart-phones which can be used for different purposes. Sensors such as accelerometer, gyroscope, magnetometer, and etc. In the beginning I thought the easiest way is just use the corresponding documentation and I got that data of two accelerometer and gyroscope sensors are used to similar analysis. Luckily I found a dataset which is measured data of these two sensors from a person's smartphone.
  - *Verify data availability:* The required dataset (*Run or Walk*) are freely available on `https://www.kaggle.com/vmalyi/run-or-walk`. I have downloaded it, I also have read about measuring these data from my own smartphone and it seems it is not so difficult, so in the case of a problem with data, I can gather from my own smartphone or with my friend's smart watch.
  - *Define selection criteria:* Accelerometer gives changes in phones velocity in three different axes including (X, Y, Z) which are important data to analyze the move of a part of a body See Fig. 2. We see both the axes could have both positive and negative values. Similarly, a gyroscope sensor gives the rate of rotation around an axis which can give useful data to recognize type of the activity.

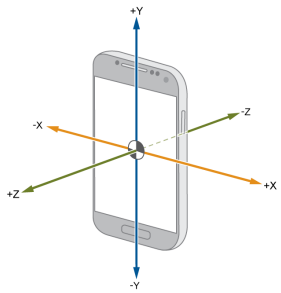

Figure 2: Measuring data by accelerometer sensor, photo from `https://www.mathworks.com/`.

	date	time	username	wrist	activity	acceleration_x	acceleration_y	acceleration_z	gyro_x	gyro_y	gyro_z
2	########	13:51:15:847724020	viktor	0	0	0.265	-0.7814	-0.0076	-0.059	0.0325	-2.9296
3	########	13:51:16:246945023	viktor	0	0	0.6722	-1.1233	-0.2344	-0.1757	0.0208	0.1269
4	########	13:51:16:446233987	viktor	0	0	0.4399	-1.4817	0.0722	-0.9105	0.1063	-2.4367
5	########	13:51:16:646117985	viktor	0	0	0.3031	-0.8125	0.0888	0.1199	-0.4099	-2.9336
6	########	13:51:16:846738994	viktor	0	0	0.4814	-0.9312	0.0359	0.0527	0.4379	2.4922
7	########	13:51:17:46806991	viktor	0	0	0.4044	-0.8056	-0.0956	0.6925	-0.2179	2.575
8	########	13:51:17:246767997	viktor	0	0	0.632	-1.129	-0.2982	0.0548	-0.1896	0.4473
9	########	13:51:17:446569025	viktor	0	0	0.667	-1.3503	-0.088	-0.8094	-0.7938	-1.4348
10	########	13:51:17:646152973	viktor	0	0	0.2704	-0.8633	0.1293	-0.4173	-0.1904	-2.6759
11	########	13:51:17:846502006	viktor	0	0	0.469	-1.074	0.0219	0.0388	1.1491	1.6982
12	########	13:51:18:46802997	viktor	0	0	0.2985	-0.7172	-0.0693	0.2326	0.4321	2.1009
13	########	13:51:18:246815025	viktor	0	0	0.6364	-1.0452	-0.24	0.1163	-0.1033	1.0822
14	########	13:51:18:446740984	viktor	0	0	0.5683	-1.2486	-0.131	-0.4556	-0.5281	-1.2407
15	########	13:51:18:646052002	viktor	0	0	0.2911	-0.7748	0.0163	-0.2345	-0.0148	-2.5884
16	########	13:51:18:846947014	viktor	0	0	0.4477	-1.1574	-0.0172	-0.1081	0.4016	0.67
17	########	13:51:19:46791970	viktor	0	0	0.2424	-0.7421	-0.0549	0.5714	-0.0506	2.1356
18	########	13:51:19:247058987	viktor	0	0	0.6028	-1.0966	-0.3046	0.1674	-0.5065	1.0156
19	########	13:51:19:446731984	viktor	0	0	0.4852	-1.3397	-0.0763	-0.8579	0.0096	-1.4015
20	########	13:51:19:645977020	viktor	0	0	0.3017	-0.8366	0.0718	-0.2701	-0.4678	-2.701

Figure 3: A snapshot of the dataset.

- *Describing data:* As I mentioned above, dataset is taken from `https://www.kaggle.com/vmalyi/run-or-walk`. It has 11 columns and 88588 rows. The columns are including $date, time, username, wrist, activity, acceleration_x, acceleration_y, acceleration_z, gyro_x, gyro_y, gyro_z$. The dataset is gathered for both left and right wrist and it is taken for two activities including *running* and *walking*. An snapshot of the data is shown in Fig. 3. We see that, the dataset is included with a "time" column that gives precise data about the exact time of data gathering.
- *Exploring data:* I checked the data and in first evaluation I could not see any problem with the dataset. The activity *walking* and *left wrist* are shown with 0 and the activity *running* and *right wrist* are shown with 1. So, I expected to see values 0 or 1 for two columns *wirst* and *activity* which all values are 0 or 1; See Fig. 4. Similarly, the columns $acceleration_x, acceleration_y, acceleration_z, gyro_x, gyro_y, gyro_z$ are data measured with two target sensors and could have both positive and negative values.
- *Verifying data quality:* I went through the dataset with some quick evaluation and it seems they is no serious problem and the current data have enough quality to target analysis. As already mentioned above, based on my search on the Internet, the data of these two sensors should be enough to get the target of the project. In the case of any problem I will take the time of assistance to get further help regard the problem.

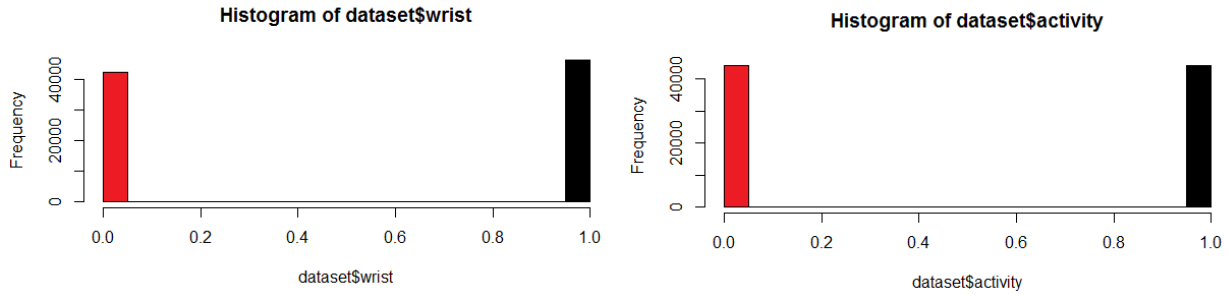

Figure 4: Histogram of the columns *wrist* and *activity*.

3: (1 points) Setting up and planning your project

Please perform the following tasks:

- Create a project repository either in GitHub or Bitbucket.
- Register you project by adding a new entry into the List of projects. Please follow the instructions given there on slide 2, this helps to keep the list tidy. In your homework report include a direct link to the slide with your project (can be copied from the address bar of your web browser). Add this link to the front page of your project repository as well.
- Make a detailed plan of your project with a list of tasks. Specify how many hours each team member is going to contribute to each task.
- Add the results from business understanding, data understanding and planning to your project repository. Report the links to where these results are listed.
- Prepare to pitch your project at the practice session. The slide of your project within the List of projects will be shown and you will be given 2 minutes to explain to others what your project is about and what your plan is. This will be followed by 2 minutes of questions and discussion about your project. Feel free to add a visualization to your slide if you think this helps during the presentation.

Solution:

From the description of the question, it seems the main goal of this task is to update our teacher assistants regard to the team member activities, collaboration and additionally the progress the project. Since, currently I am alone in this project, so I mostly will upload the result of my project and will write a README about the project.

URL of my Git: `https://github.com/atapoor/Data_Mining_2017`

Regards,

Shahla Atapoor