Data Science Challenge

By: Miguel Ballesteros

This document contains the answer to the challenge sent by Correlation-One as an assessment of my Data Scientist skills.

Contents

Introduction

The present task is a very quick iteration through the data science pipeline to obtain insights over a particular scenario. In this case the total time taken to complete the task was around 12 hours within 4 busy days. The exercise response aims to show a good knowledge from data collection to model creation, with full awareness of the business scenario at all stages.

The document is structured as recommended in the task sheet, with sections for each separate question.

Executive Summary

To keep and improve users' engagement in the daily digest email, it is important to use topics with high click rates as the main driver when defining content, and collect additional data to create in the mid-term a better user profiling system possibly feeding a recommender system making this task more automatic and accurate.

Detailed Assessment

In this section there is a full description of the followed process at high level and the business considerations taken during the analysis, so it is contextualized to the current situation.

Data Exploration

The two datasets have records in the order of few millions, and the size is good to be handled in memory. As it will be described before, this is one of the most important criteria to use R and store all data in memory. Two quick commands show the size of the largest tables in both cases:

```
sqlite> select count(*) from email content;
10738006
```

```
PS > Get-Content .\access.log | Measure-Object -Line
  Lines Words Characters Property
  ----- ----- ---------- --------
1770781
```

Within the script and using the R library RSQLite, all tables are loaded to data frames. The articles table is de-normalized for topic and type values, with the aim to simplify later calculations. Both, id and name are preserved for the moment in case it is more convenient to use one of these later. In a second iteration the unused columns can be removed.

```
> nrow(articles)
[1] 17029
> summary(articles)
   article_id      author_id        topic_id                      topic_name       type_id
 Min.   :    1   9328   :  107   Min.   :  1.00   Entrepreneurship       :  742   Min.   : 1.00
 1st Qu.: 4258   17169  :  107   1st Qu.: 27.00   Growth Hacking         :  687   1st Qu.:18.00
 Median : 8515   9153   :  106   Median : 52.00   Planning & Forecasting:  686   Median :27.00
 Mean   : 8515   14386  :  106   Mean   : 50.85   Web Marketing          :  665   Mean   :26.68
 3rd Qu.:12772   5722   :  104   3rd Qu.: 77.00   Personnel Management   :  637   3rd Qu.:38.00
 Max.   :17029   7089   :  104   Max.   :106.00   Business Development   :  575   Max.   :49.00
                 (Other):16395                    (Other)                :13037
     type_name      submission_time
 News     : 1270   Min.   :2015-01-01 00:01:14
 Blog Post:  942   1st Qu.:2015-01-22 16:08:06
 Opinion  :  839   Median :2015-02-13 19:38:34
 Summary  :  830   Mean   :2015-02-14 21:20:08
 List     :  821   3rd Qu.:2015-03-09 20:10:24
 Webinar  :  736   Max.   :2015-03-31 22:39:51
 (Other)  :11591
```

In the case of articles (17,029 records) all values seem to be within the expected ranges. Topics and types are within the max/min values checked in the database, article ID has a consistent number and the submission times are all within the 3-month window described in the scenario. No important operations are required for cleaning beyond type conversions. There are no missing values.

```
> nrow(content)
[1] 10738006
> summary(content)
   content id          email id          user id        article id
 Min.   :       1   Min.   :      1   Min.   :    1   Min.   :    1
 1st Qu.: 2684502   1st Qu.: 262910   1st Qu.: 5004   1st Qu.: 4201
 Median : 5369004   Median : 528120   Median :10002   Median : 8411
 Mean   : 5369004   Mean   : 528000   Mean   :10001   Mean   : 8409
 3rd Qu.: 8053505   3rd Qu.: 792658   3rd Qu.:14999   3rd Qu.:12617
 Max.   :10738006   Max.   :1058436   Max.   :20000   Max.   :16982
   send time
 Min.   :2015-01-02 08:06:20
 1st Qu.:2015-01-23 09:04:52
 Median :2015-02-16 08:35:49
 Mean   :2015-02-15 07:43:31
 3rd Qu.:2015-03-10 08:18:52
 Max.   :2015-03-31 17:05:49
```

For the content data, all values are within the expected ranges. As with Articles, no important cleaning operations are required. There are no missing values.

```
> nrow(users)
[1] 20000
> summary(users)
    user id          email
 Min.   :    1   Length:20000
 1st Qu.: 5001   Class :character
 Median :10000   Mode  :character
 Mean   :10000
 3rd Qu.:15000
 Max.   :20000
```

The users table is very simple, just allows to double-check the id's range and consistency with the content reference. There are no missing values.

In the case of the access table, the unprocessed columns show that the timestamp has few values out of the upper limit. The request HTTP verb is GET for all entries, which is an expected value for an e-mail click. HTTP version is the same as well. Therefore, all remaining exploration operations have to be performed against the request portion along with the query string. After applying these operations the table reports the following summary.

```
> nrow(access)
[1] 1770780
> summary(access)
   timestamp                     request          httpstatus      bytessent       article id
 Min.   :2015-01-02 08:08:43   Length:1770780     200:1762014   Min.   :  200   Min.   :     1
 1st Qu.:2015-01-23 10:45:54   Class :character   400:   8766   1st Qu.: 1996   1st Qu.:  4206
 Median :2015-02-16 09:30:18   Mode  :character                 Median : 2987   Median :  8431
 Mean   :2015-02-15 10:10:43                                   Mean   : 3569   Mean   : 10869
 3rd Qu.:2015-03-10 09:43:20                                   3rd Qu.: 4482   3rd Qu.: 12695
 Max.   :2015-04-01 15:25:51                                   Max.   :33999   Max.   :848882
    user_id
 Min.   :    1
 1st Qu.: 5009
 Median :10001
 Mean   :10002
 3rd Qu.:15002
 Max.   :20000
```

The cleaning operations are then focused on removing the entries above the timestamp upper limit. Regarding the HTTP status code reporting an error (400 – Bad Request), all entries are preserved in case an associated pattern can be identified. Those entries also evidence the interest from the user in the topic/type/article which statistically can add more value to this particular analysis.

The code that separates the article_id and user_id values from the request string has important comments and considerations regarding the performance. For an automatized process in a production environment, such code requires some performance tunning. However, in an initial iteration it is important to complete the whole process as soon as possible and later improve the identified steps for a production deployment.

After separating the article_id and user_id values from the requests' query string and checking the summary, it made evident that all user_ids are within the expected ranges. However, the article_id maximum value is higher than the maximum article_id in the articles table, suggesting that there are values that are not eligible for this analysis. Taking a quick look at these entries, there is a perfect match to those having the HTTP 400 response giving this time a good reason to be removed.

In the same line and to keep consistency, all access entries having timestamps higher than 2015-03-31 23:59:59 are removed as part of the cleaning stage. This set accounts only for 809 entries out of 1,762,014 which does not have a significant statistical impact. The final summary is then.

```
> nrow(access)
[1] 1761208
> summary(access)
   timestamp                     request          httpstatus      bytessent       article_id
 Min.   :2015-01-02 08:08:43   Length:1761208     200:1761208   Min.   :  294   Min.   :    1
 1st Qu.:2015-01-23 10:39:45   Class :character   400:      0   1st Qu.: 2011   1st Qu.: 4187
 Median :2015-02-16 09:21:05   Mode  :character                 Median : 3000   Median : 8372
 Mean   :2015-02-15 09:37:10                                   Mean   : 3585   Mean   : 8400
 3rd Qu.:2015-03-10 09:29:52                                   3rd Qu.: 4492   3rd Qu.:12620
 Max.   :2015-03-31 21:59:54                                   Max.   :33999   Max.   :16980
    user_id
 Min.   :    1
 1st Qu.: 5009
 Median :10001
 Mean   :10002
 3rd Qu.:15001
 Max.   :20000
```

Data Transformations

The applied data transformations rely mostly on basic mathematical operations and focused on aggregating values by the fields topic, content type and week as time variable. Most transformations are designed to provide insights regarding the level of engagement of users. With the available data, such level of engagement can be measured mostly by the number of clicks and their proportion to the total sent links

The most transformed table is the articles one which includes the aggregated values of number of sent links in e-mails, clicks associated to these links, and a relationship between these two values as click rate. For quick time analysis the week number was added to verify if there are important changes. The week is preferred to month or days to have a good set to compare values (14 items). With month it is not easy to verify sequential or consistent changes in only 3 periods, and daily is not a reliable metric for this scenario considering that people not necessarily process e-mails in the same way during two consecutive days, but do week by week.

In the case of the table content the topic and type values were added to initially explore different conditions or patterns. For users, the field e-mail was split to obtain the domain and check potential common behaviors from similar ones, but almost all values are unique.

Finally, there are two fact tables for topics and content types. These fact tables basically contain basic statistics for these dimensions as these seem to be of particular interest by the business unit.

Results

From the time perspective there are no significant findings that can be taken into account to define further actions. In general terms the analyzed behaviors do not change significantly with the time, so this dimension for the moment can be ignored. However, it is important to consider that the covered period is only 3 months and it may be missing seasonal interests or behaviors, so a more complete analysis should cover at least one year to have more reliable data. A plot shown in a later section supports this conclusion.

From the content type perspective, no specific patterns were found. This suggests that this is not a relevant factor affecting directly the level of engagement or the click rate.

Conversely, the topic is an important factor and has a direct effect on the rate of clicks a link has. This is the most engaging factor with topics having high click rates. The top 10 are displayed for illustration proposes (table topic_facts).

	topic_name	topic_clicks	topic_links	click_rate
1	Public Finance	84650	205136	0.41265307

2	Consumer Behavior	67332	205622	0.32745523
3	Unemployment	57857	182016	0.31786766
4	Industry: Retailing	56165	180725	0.31077604
5	Outsourcing	21508	70407	0.30548099
6	Budgeting	56709	189127	0.29984614
7	Industry: Pharmaceutical	16920	57250	0.29554585
8	Franchises	16607	56924	0.29173986
9	Advertising	83607	317325	0.26347436
10	Organizational Learning	14978	61056	0.24531578

From the users perspective the rate of clicks is not bad, averaging on 0.17 and showing a normal distribution suggesting that any improvement most likely will impact the vast majority of users.

Tools

This task was totally performed in R. The main reasons for this choice were.

- The datasets are not bigger than an average PC resources and those can fit easily in memory for quick and convenient use
- Contains sufficient resources through libraries to perform all operations. However, some string operations are known to not have the best performance, but the extra times can be tolerated within this context. At the same time the designed pipeline avoid reprocessing steps, increasing the productivity when focused on each stage.
- The tool and libraries are easy to install and deploy, so it can easily be executed in a different environment.
- The language is broadly known in the Data Science field.

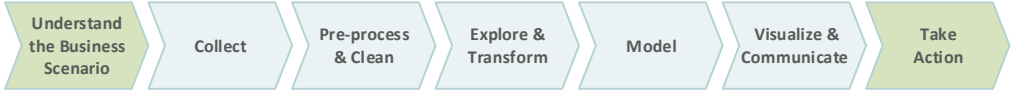
Occasionally some verifications were done using PowerShell commands, SQLite command interface, and simple tools like notepad.

In addition to R, Tableau is also used to explore the data and quickly visualize data to take decisions when writing the code or planning the plots included in the automatic code.

To write the report, Word is the most convenient and universal tool in most companies.

Techniques

- The Data Science pipeline is the main driver when planning all steps in the analysis

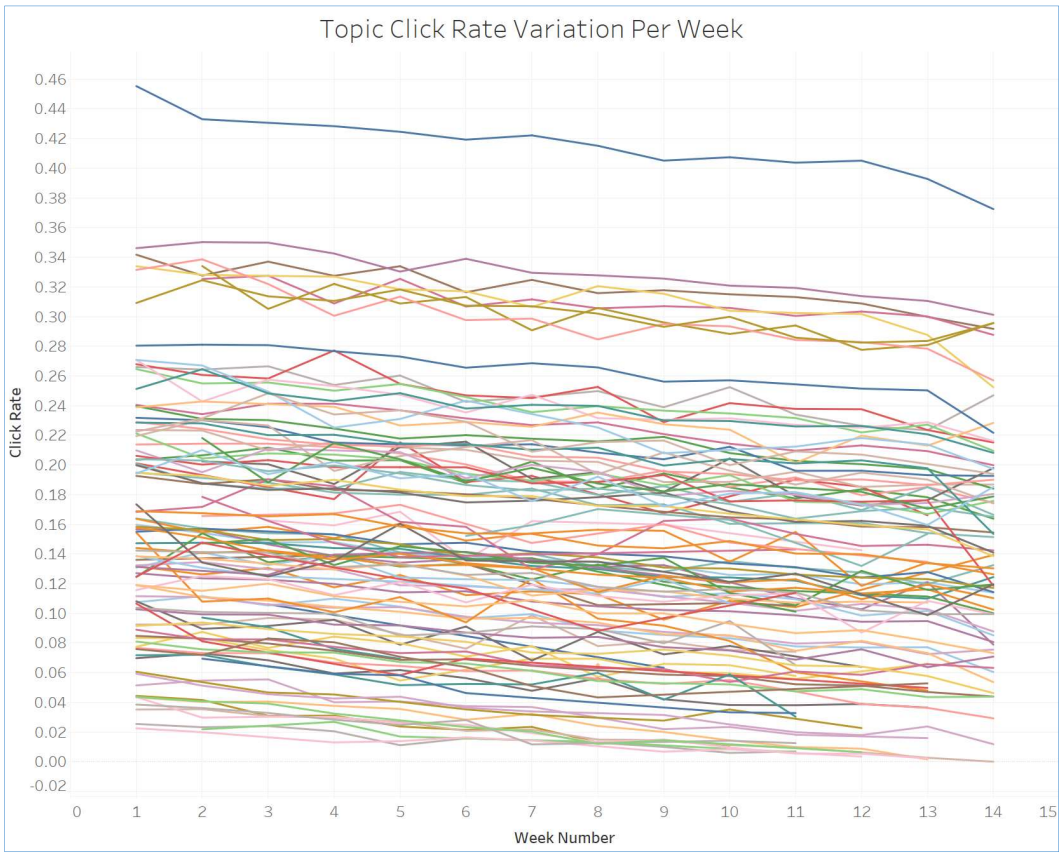


- Visual Analytics: While getting progress and needing to understand data quickly, a tool like Excel or Tableau is helpful to improve the view on the current situation, or to plan the steps to follow.
- Data Mining Techniques: Clustering and other unsupervised approaches help understanding or surfacing new patterns. Clustering technique was planned inintially to be applied in this task, but it was not possible due to time restrictions.

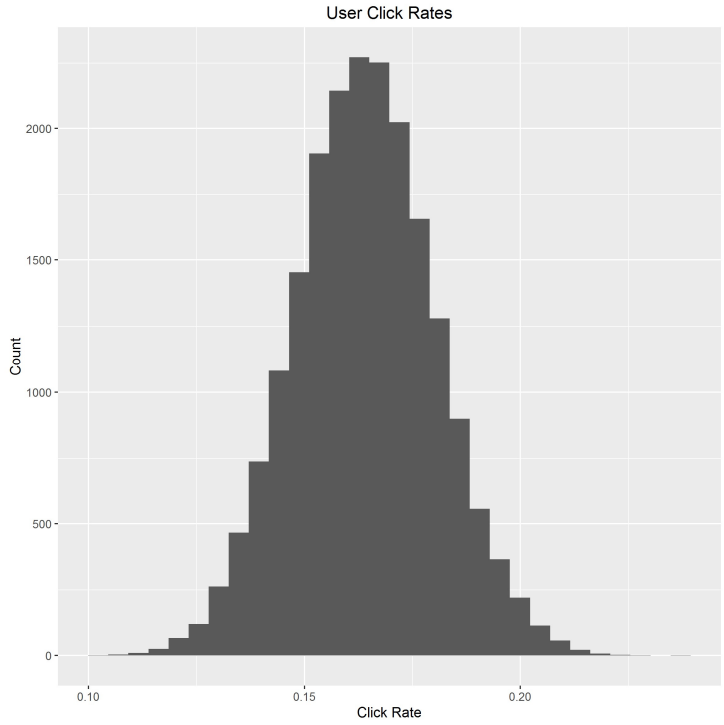
- Structured code and logging: The code is split in different files with specific proposes. Additionally, it contains logging mechanisms to ease the troubleshooting tasks.
- Source Control: GitHub is a convenient tool to keep track of changes and an alternative backup for code and associated files.

Plots

The first plot shows the evolution of click rate per week. The general tendency in this case is that click rates are decreasing for most topics. No particular reasons can be associated at this point. The global effect is however constant and some positive trends compensate the low tendency.

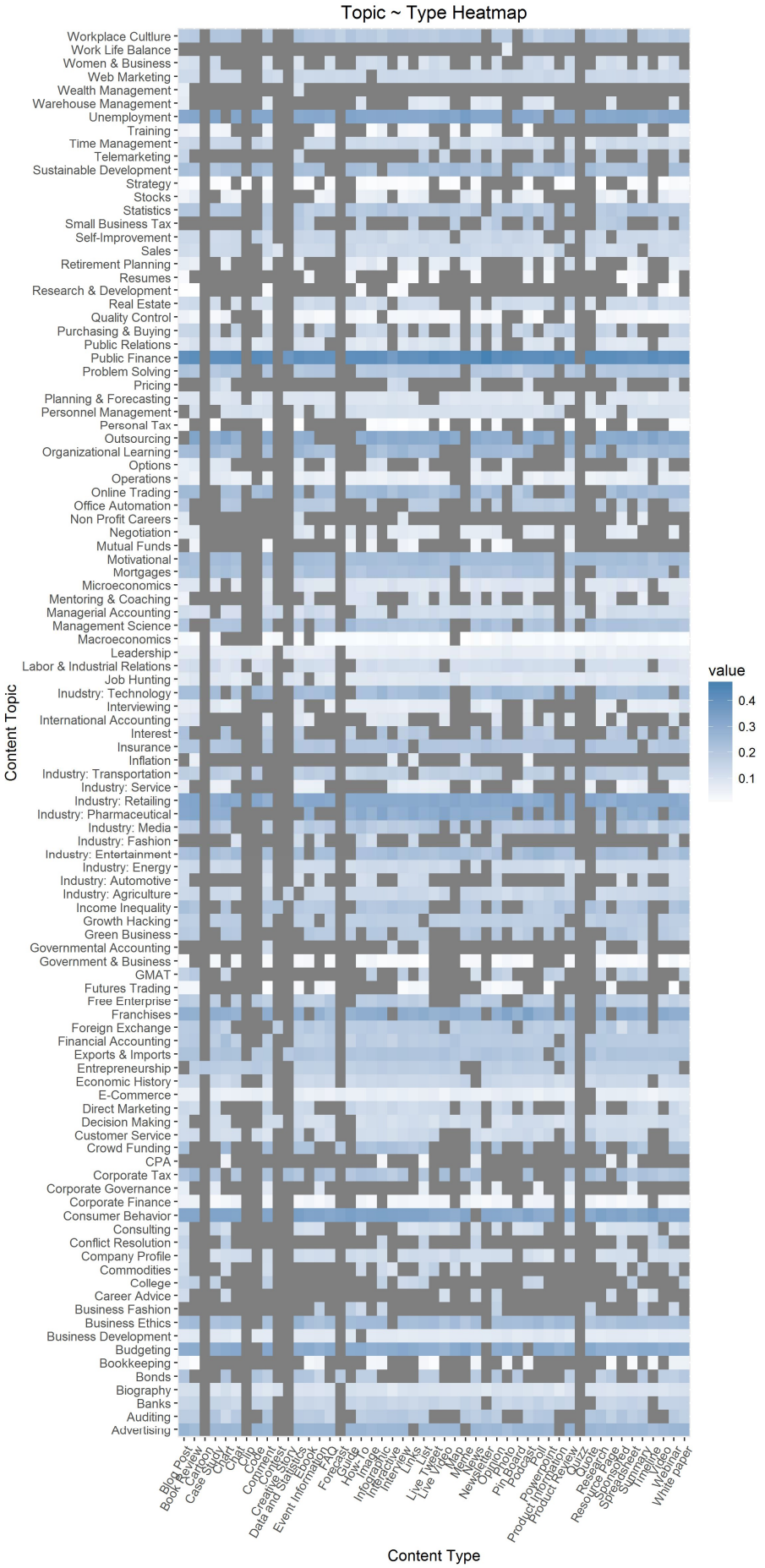


The second plot shows the users' click distribution evidencing a very normally distributed values.



The third plot is a heat map showing a comparison between the topics and content types. This plot explores the relationship between these factors and checks for patterns in case a particular combination have positive impact on user engagement levels.

Just with the naked eye it is possible to notice the evident pattern in rows, confirming that common behaviors are associated only to topics and not to content types.



In this plot it is clear then that the main driver for user engagement is the topic.

Commercial Recommendations

As found, the main driver to keep users interest in the sent links is the topic. A list with the proportion of click rates is provided as the topic_facts table. The recommended actions are.

- Increase the proportion of articles about the topics in the top 20 of the topics list, in all e-mails. Such emphasis is expected to boost the click rate value
- Basically all users have some degree of engagement, and any encouraging action potentially will impact all users in the same proportion.
- Check for the existence of the data listed in the "Additional Data" section below, with the aim to create more accurate and complete dataset to later target all e-mails more accurately.

Future Work

The results shown in this document are the dedication of few hours, so there are many areas of improvement from this point.

Additional Data

- User demographics: Users cannot be profiled with the available data. This may provide more understanding about user preferences and users segmentations. A simple set of demographics should be enough to create more complete feature vectors and apply clustering techniques with the aim to better target the email receivers, or even rely the e-mail generation tool to a recommender system.
- Article Authors: Articles are not profiled per the author or publisher. This data is key to understand if particular content sources are more reliable or perceived as more credible or preferred by the end-users, so the future e-mails can increase the proportion of articles from these sources.
- Device Type: Having the device type will help to have a good picture on how easy the content is to read for different publishers. This value is key to complement user habits and target more appropriate content according to the user's preferred device. This can also feed a recommender system. The device value can easily be taken from the field user_agent in the Web server.
- Location: This value helps getting an idea of how geographically are distributed the users, potential topics of interest, important events at country or regional level, etc. Additionally, the location data is useful to calculate the most common time during the day the user consumes the content. Adjusting the time zone is key to have an accurate picture.

Other Questions

- What topics are similar in terms of click-effectiveness?
- How relevant is the time between the sent-content action and the end-user-click action? Is this evidencing end-user engagement conditions?
- Are the email contents well-balanced in terms of topics?

Conclusion

This short task allows to show a small set of skills from a Data Scientist, but in general terms those are applied in a methodological sequence that increases the chance to find and provide business insights.

The most relevant finding in this case was to confirm that the topic is a relevant factor to define digest email contents.

The other findings basically support the topic importance and suggest the effect a positive or negative action will have in the scenario. With some clear areas of improvement, the analysis is limited due to time restrictions, but it can grow significantly, especially with the existence of additional data that helps boost other analysis dimensions.