Why Kaggle Might Be Important To People Analytics

Lyndon Sundmark, MBA

2017-05-08

Background

For any of you who have been following my blog articles on LinkedIn for the last couple of years, one of the thoughts /themes I have been sharing in probably all of them is that:

'People Analytics is what happens when Data Science meets Human Resources.'

The intersection of these two fields is People Analytics. When we:

- Ask critical HR business questions
- Search /review what information sources and data we have that may be relevant to answer those questions
- Pay attention to the quality of that data
- Conduct statistical analyses that are appropriate to the questions asked
- Interpret and document our findings
- Make recommendations based on our findings for either actions or decisions themselves that need to be made- because of a better grasp /understanding of the issues, or because the statistical analyses itself was predictive in its purpose

We are engaging in Data Science with the context / domain of human resources management and decision making. The recognition of the 'data science' part of 'that' is critical to understanding the potential contribution Kaggle can make to People Analytics.

What Is Kaggle?

So, what is Kaggle and why is that important to People Analytics?

One good description comes from Wikipedia:

https://en.wikipedia.org/wiki/Kaggle

This sums it up as follows:

'In 2010, Kaggle was founded as a platform for predictive modelling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models.'

And

'In April 2015, Kaggle released the first version of their Scripts product onto their platform. Scripts allows users to write, run, and publicly share their code on Kaggle.'

In effect Kaggle is a site on the internet at kaggle.com where users can share datasets of information. Once posted, the contributing user of the dataset or anyone else who uses Kaggle can post their code with respect to how they approached the analytics problem they were try to solve, and show their results. The datasets are uploaded as csv formatted files to Kaggle. The code uploaded and used can be code in:

- R
- Python
- Julia
- SQLite

R and Python are often the most used languages for Data Science.

The benefits and strengths of this functionality on Kaggle should be obvious:

- You have a platform where data can be shared reasonably easily
- You have a platform where many more eyes- beyond your own or your organization, and the expertise there - can be brought to bear on your data

The previous Wikipedia description indicates that Kaggle was founded as a platform for predictive modelling and analytics competitions. Indeed, the vast majority of datasets, code, and activity on Kaggle seems to revolve around those 'founding' roots. Many of these have requirements for being formally part of the analytics 'competition' and some have paid prizes.

However, in the last several months, Kaggle introduced the concept of 'organization(s)' in their platform. With this functionality, you don't necessarily have to submit data as part of an organized competition, for money, or with many restrictions- you can submit datasets for which your intent is to have others, informally and voluntarily, share their ideas of analysis freely through code examples of their own. And with this functionality you can group datasets you have submitted to be visible in one place. You can also add 'people' to your 'defined' organization who have the rights to submit other datasets to be grouped under that 'organization'. In effect, at its simplest, an organization is just a name you create to group 'your' datasets -so that they are all in one place.

While you use this functionality to 'group' them, the datasets themselves are public to all Kaggle users for download, viewing, and usage. When statisticians, data miners, data scientists look at your data, and look at the problem you are trying to solve, they can submit a 'kernel' tied to your dataset which is R code or Python code that attempts to solve the problem or answer the question. That 'kernel' too is publicly visible.

Why is this potentially important to People Analytics?

Because People Analytics is at the intersection of Data Science and HR, it requires knowledge of both domains. HR people know the context of HR and in depth, but often know very little about data science. Data Scientists often know data science and statistics in depth, but know very little, if anything, about the context of HR and HR information. You are 'fortunate' if you have both of those skills in your organization, and more fortunate still if they both exist in the same individual(s)

That really is a 'Catch 22' for People Analytics to take root and gain a proper legitimate foothold in HR in organizations. You may have lots of HR professionals who see the potential of People Analytics- but don't know how to approach their data analytically, and need the help of data science professionals. And you may have a lot of data science professionals who would like to help apply their skills to HR data to help answer / address HR questions /issues -but don't have access to HR data and to the minds of HR professionals and their knowledge of the HR context.

Additionally, another big 'Catch 22' is that HR information, like many other types of information, is subject to confidentiality and privacy laws that are 'country specific'. The 'emerging' question is how can data be 'shared' while maintaining the appropriate confidentiality and compliance with privacy laws?

I think it's important that these critical questions be answered correctly. How quickly, successfully, and effectively People Analytics emerges as a discipline within HR and organizations, I think, is dependent on this. If they can be answered successfully and in compliance with privacy and confidentiality laws- Kaggle, especially with its 'organization(s)' feature can be an excellent platform to have additional analytics skills be brought to bear on your data.

A Personal Example

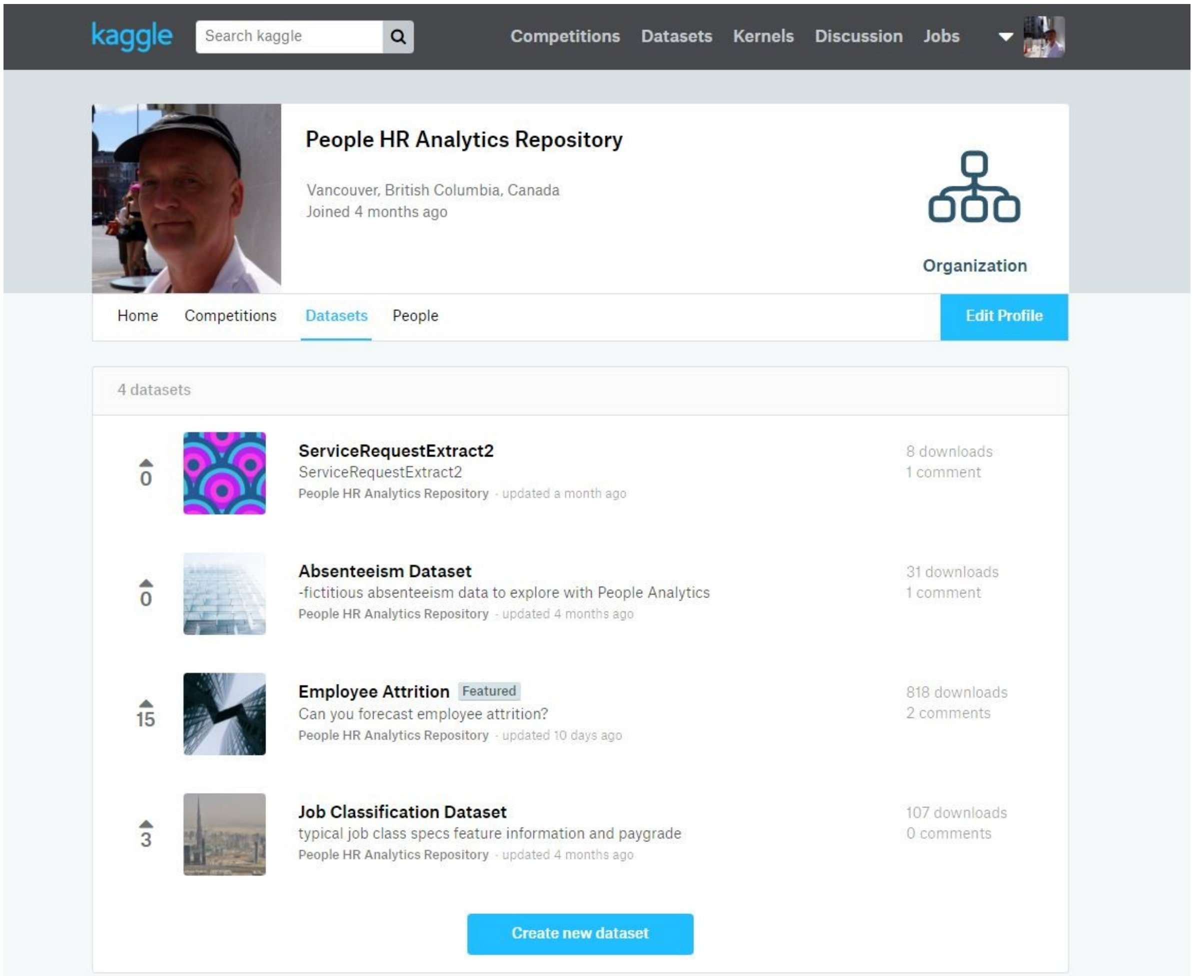
I had become aware of Kaggle about a year ago, and wanted to experience it firsthand. But I didn't want to hold any formal competition per se. I just wanted to try it out as a platform. Within the last year, the 'organization' feature was announced and I inquired with Kaggle about it. And they seem to indicate that 'yes'- creating my own organization would allow me to submit my own datasets without there needing to be an 'intent' of any formal competition.

As I mentioned at the beginning of this blog article, if you have been following my blog articles over the last couple of years, you will have come across several People Analytics in R examples. These examples included both links to datasets on my Microsoft Onedrive space as well as showing R code and output.

I thought the best way to try out and experience Kaggle would be to try to upload those datasets and include my code as Kaggle kernels. To do that I needed to create an organization first. The name of my organization on Kaggle is

People HR Analytics Repository found at

https://www.kaggle.com/HRAnalyticRepository



People HR Analytics Repository

Vancouver, British Columbia, Canada
Joined 4 months ago

4 datasets

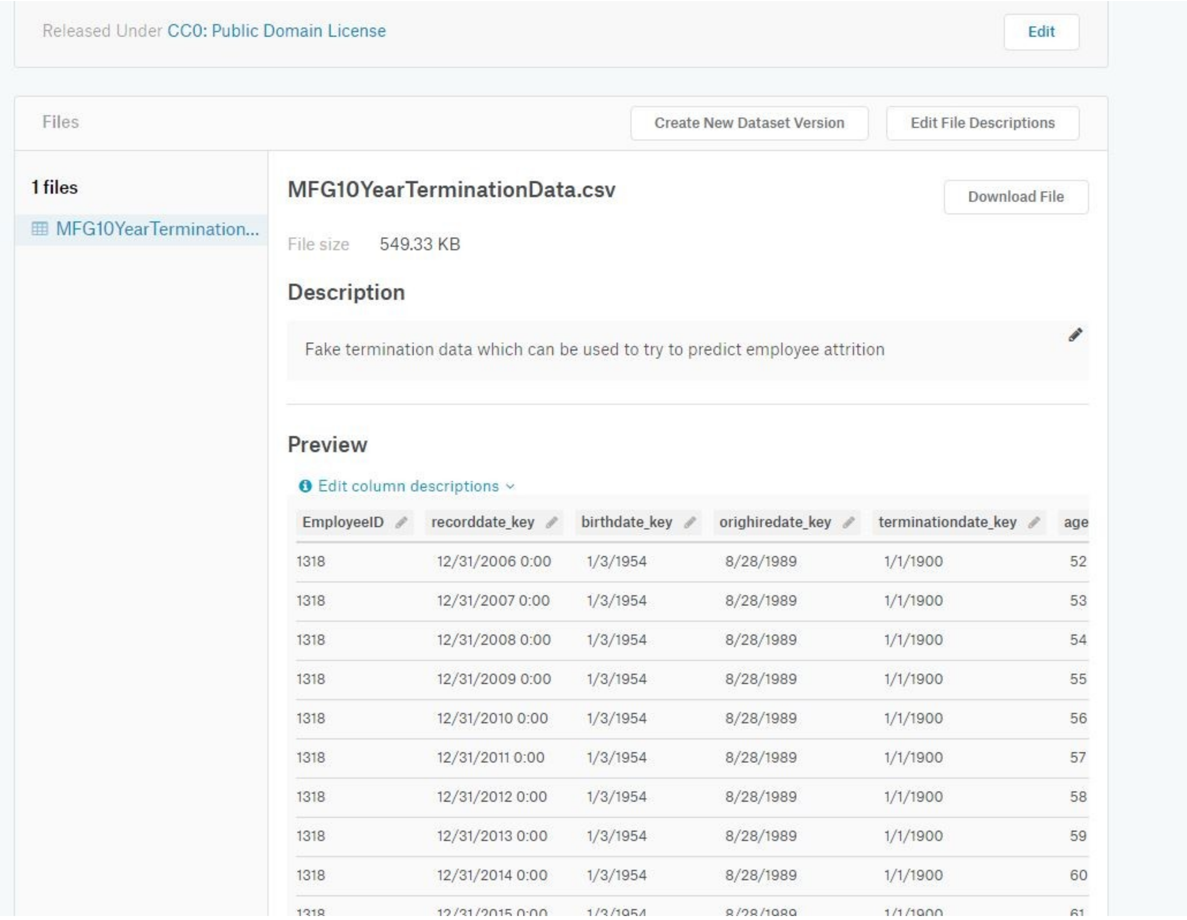
0 **ServiceRequestExtract2**
ServiceRequestExtract2
People HR Analytics Repository · updated a month ago
8 downloads
1 comment

0 **Absenteeism Dataset**
-fictitious absenteeism data to explore with People Analytics
People HR Analytics Repository · updated 4 months ago
31 downloads
1 comment

15 **Employee Attrition** Featured
Can you forecast employee attrition?
People HR Analytics Repository · updated 10 days ago
818 downloads
2 comments

3 **Job Classification Dataset**
typical job class specs feature information and paygrade
People HR Analytics Repository · updated 4 months ago
107 downloads
0 comments

You will notice that I have 4 datasets included there. One of the things that Kaggle does is keep you apprised of activity with respect to your datasets- how many times it's been downloaded etc. When you click on the dataset listed and scroll down the page presented you can see a sample of the data:



Released Under CC0: Public Domain License

MFG10YearTerminationData.csv

File size 549.33 KB

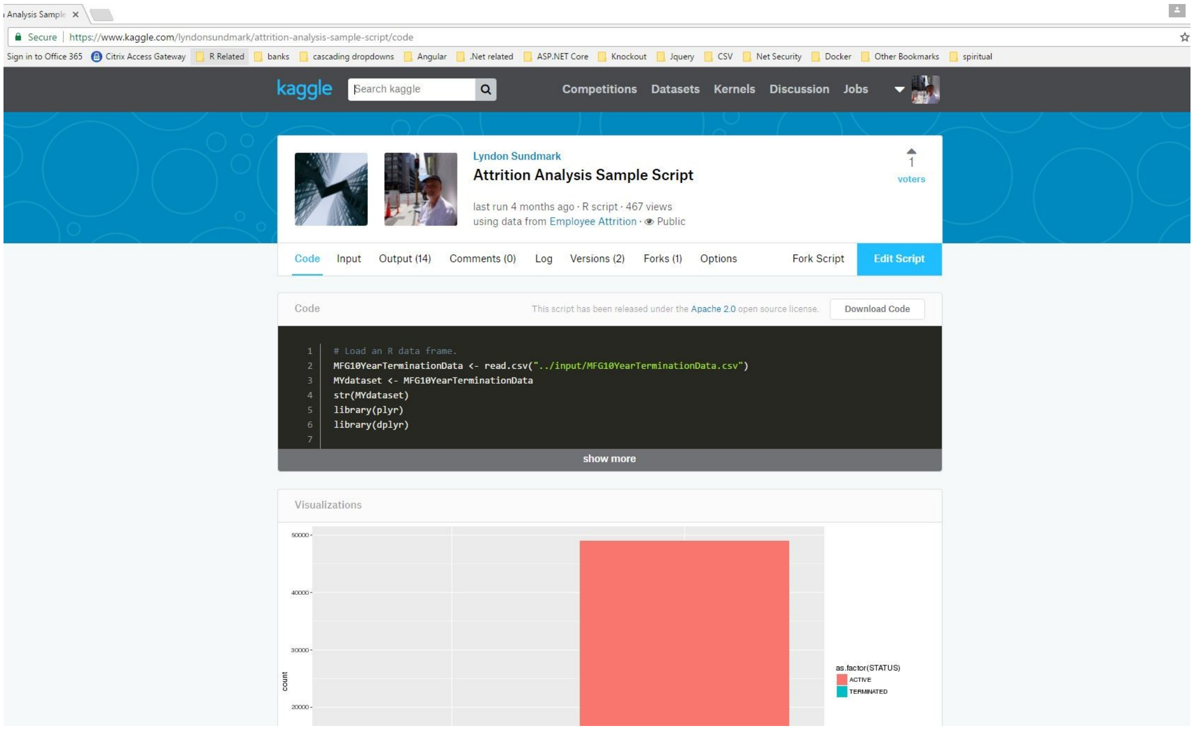
Description

Fake termination data which can be used to try to predict employee attrition

Preview

EmployeeID	recorddate_key	birthdate_key	orighiredate_key	terminationdate_key	age
1318	12/31/2006 0:00	1/3/1954	8/28/1989	1/1/1900	52
1318	12/31/2007 0:00	1/3/1954	8/28/1989	1/1/1900	53
1318	12/31/2008 0:00	1/3/1954	8/28/1989	1/1/1900	54
1318	12/31/2009 0:00	1/3/1954	8/28/1989	1/1/1900	55
1318	12/31/2010 0:00	1/3/1954	8/28/1989	1/1/1900	56
1318	12/31/2011 0:00	1/3/1954	8/28/1989	1/1/1900	57
1318	12/31/2012 0:00	1/3/1954	8/28/1989	1/1/1900	58
1318	12/31/2013 0:00	1/3/1954	8/28/1989	1/1/1900	59
1318	12/31/2014 0:00	1/3/1954	8/28/1989	1/1/1900	60
1318	12/31/2015 0:00	1/3/1954	8/28/1989	1/1/1900	61

At the top of the page you can see among other things ‘kernels’ for that dataset listed and you can click on those lists to see the kernel and its output:



Lyndon Sundmark

Attrition Analysis Sample Script

last run 4 months ago · R script · 467 views
using data from Employee Attrition · Public

This script has been released under the Apache 2.0 open source license.

```
# Load an R data frame.
MFG10YearTerminationData <- read.csv("../input/MFG10YearTerminationData.csv")
MYdataset <- MFG10YearTerminationData
str(MYdataset)
library(plyr)
library(dplyr)

```

While my initial intention was to just experience Kaggle with datasets that I was familiar with, and try running my R code as kernels- It wasn't long before I started receiving the above-mentioned notifications that others were interested and downloading my datasets to try out their own hand at analysis of my data.

Eventually – on one of them- another Kaggle user used a different data science algorithm on my data and came with higher accuracy of predicted results than I had. The 5 algorithms I had used had varying levels of accuracy of prediction, but his beat all of them.

And the bonus was that I hadn't explicitly asked anyone to take a look at my data to see if they could 'better' my prediction. They did this voluntarily for the 'challenge' on their part. And I learned a new algorithm which could be brought to bear on my data. Bonus!

Such is the spirit and intent of Kaggle as was described in the opening Wikipedia description.

Caveat

One thing that can't be stressed enough is that in submitting any dataset to Kaggle, that you have

- the authority to do so
- that you are complying with all privacy and confidentiality laws in your jurisdiction.

One thing that I asked Kaggle is – ‘is it possible to delete a dataset once you have submitted it?’. The answer a few months ago, was ‘No’. At present, you can’t delete it once uploaded. You can update it however. So, know the answers to the above bullet points, before you upload anything.

In my case, I didn’t have to worry about the above bullet points because all my datasets are fictitious, and I created them.

Closing Comments

What my own personal experience with Kaggle illustrated for me was the power of having ‘additional minds’ and ‘eyes’ on my data. While I know/knew several the appropriate algorithms to use on my data to answer the questions I had, you can never know ‘everything’. The extra eyes on this was rewarding.

To the extent that many organizations don’t necessarily have both the data science and HR skills in their organization, Kaggle as a platform can be very useful to People Analytics depending on your situation, and providing that you can be fully compliant with the privacy /confidentiality laws in your jurisdiction.