2018-12-31

Some background:

I am a PhD who started doing “data science” on electronic police case files in the 1990-s, data preparation with an own Java parser, analytics with SPSS.

I have since co-founded a data company Smart Data Hub Ltd (www.smartdatahub.io).
Short description of what we do, we:

- Harvest open data providers and sources from the internet
- Ingest them to our AWS stack (S3)
- Run them through automated type changes (e.g. integers in string fields are recognized and typed as integers and input into data structures in which they now are integers by type)
- ..then data quality procedures written by me
    - analyze said data, outputting boolean values and diagnostic output for e.g. “table_partitionname_string_column_values_are_not_just_integers_wrongly_typed” and
    - generate “data stories”, which you can see on the linked web-page above.
- We have over 5000 datasets in our store, searchable, downloadable with one click. We pump more in.

CAPSTONE PROJECT IDEAS:

1) **Advanced dq-root cause profiler**:

NOTES from teacher: NumPy here would be good for wrangling and finger warmup

a) Data: DQ-test data from our stack’s ingestion, the booleans, the textual outputs.
b) Analytics:
    i) (descriptives) Which are the most common failing tests?
    ii) (correlation and clustering) Which tests are present together most often?
    iii) (correlation/regression) Which test failures are associated with which source(s) of which data /file type(s)?
    iv) (regression) Could we get a predictive model, which could, by source give a statistical likelihood of a test failure of a certain type, thus preparing the data engineer to check possible data problems before even putting the data to ingestion

2) Finnish national features analysis by postalcode.

a) Data: There is a rich series of data on Finnish population by postal code area (over a hundred variables, such as population by gender, age-group, income group etc), from several years now, enabling also analytics over time-series (albeit of just some years). I would like to milk that dataset dry. Not only that, our datasets have spatial dimensions, joining other datasets to this dataset would enable analytics on Finnish data never publicly done before.

b) Analytics:
    i) Correlations
    ii) Clusters and segmentation, discriminant analysis

iii) Time series analytics
iv) ...above all - predictions of where every variable is going
v) ..I could reflecti my findings with the population census statisticians, would they agree, disagree, why? This would be a nice learning opportunity.

3) **Voting behaviour of politicians**:
   a) Data: Finland has a multitude of political parties in government, nine to be exact, unlike the US (which has *de facto* 2).
   b) Their member's voting data is public data in Finland
   c) **Research questions**: how much political parties that pretend to differentiate, actually differentiate, when their member's voting behaviour is analyzed - how accurately can we predict their voting behaviour..and replace them with a bot?
   d) Analytics:
      i) Voting behaviour profile per representative. Who skip votes, either by absence or by voting "empty".
      ii) **Clustering** of representatives of different parties by their voting behaviour - can they be recognized as representatives of their parties or is there large distribution - "swing votes". If yes, who are like that?
      iii) Are there influential/characteristically party-tied voters that we could use to profile the party and predict how the party will, in majority, vote in future issues
      iv) A machine learning model, which predicts voting behaviour when taking into account votes from the typical members of the party
      v) All of the analytics from above would be made public. It would create some headlines and strengthen democracy, at the minimum revealing people who have been elected but actually do not vote at all/give empty votes...and thus might be re-considered by the electorate.

Manne Laukkanen (Mr, PhD)
Helsinki
Finland
Europe