Correlation One Data Science Test

Lei Yang

leiyang@berkeley.edu

Questions

1. What is your one -sentence executive summary?
2. What is your detailed assessment (for a technical audience)? Please quantify, use technical jargon.
3. What tools did you use?
4. What techniques did you try?
5. What three plots did you make to best explain the data?
6. What is your commercial recommendation for business unit heads who are non-technical?
7. What other data would you like to see about the platform? What questions would this additional data help you answer?

What is your one -sentence executive summary?

- The overall effectiveness of emailing link is reducing, although more links will increase click; improvement can be made by analyzing individual’s click history and customize the email content.

What is your detailed assessment (for a technical audience)? Please quantify, use technical jargon.

- Please see the attached python notebook for details

What tools did you use?

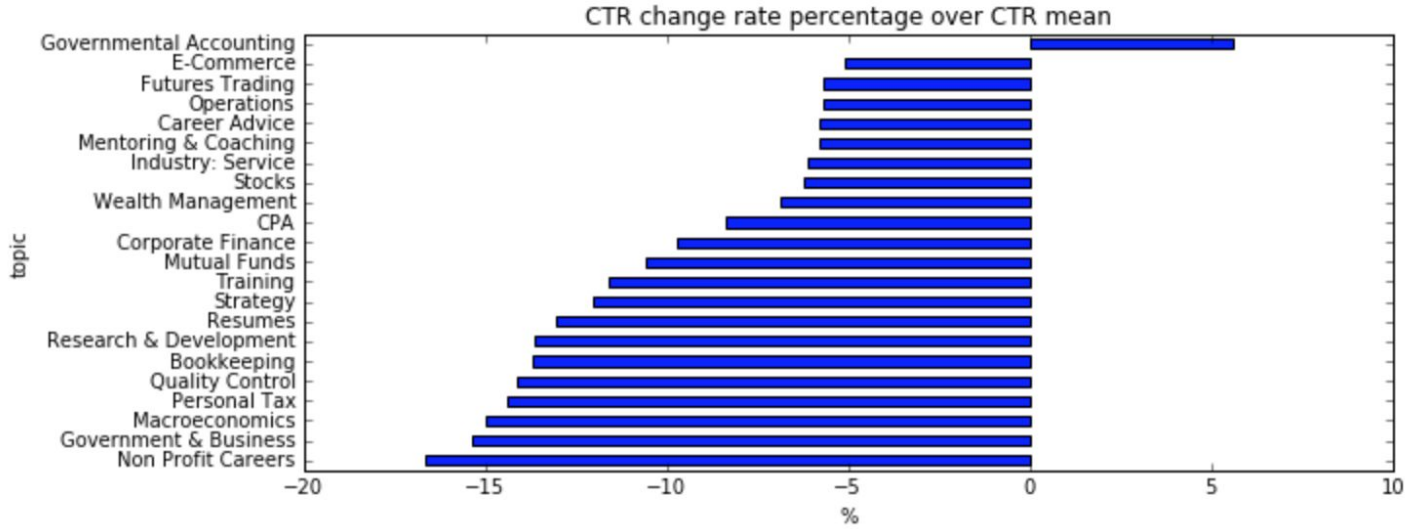
- Python for scripting
- Pandas, numpy for data analysis
- Matplotlib for charting
- Jupyter notebook for presenting
- SQLite for data storage and retrieving

What techniques did you try?

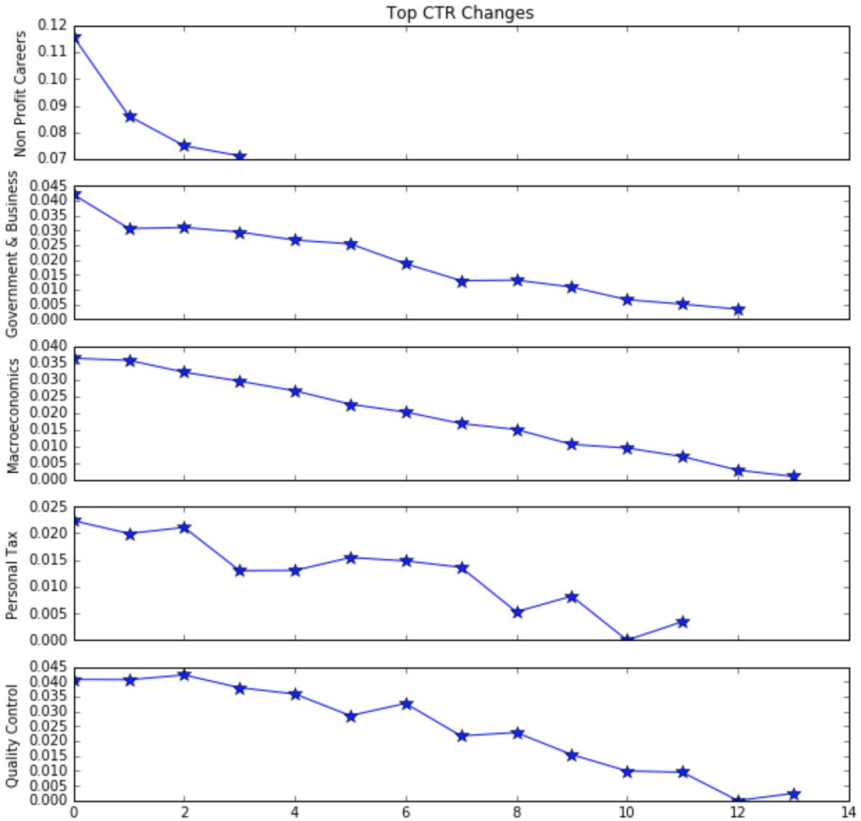
- Data pre-processing (summarize, transform, join, match etc.) is done through SQL statement
- Data manipulation (grouping, modeling, charting etc.) is done with Pandas dataframe
- Click-through rate (CTR) is defined for clicking likelihood
- Linear regression model is used to evaluate CTR change rate
- All analysis is done in Python notebook, if the data quantity is big, MapReduce framework can be used to parallelize the computation.
- **It is also possible to build a network model between authors, with outgoing arrow represent click, then a page-rank can run to find the popular authors, whose articles can provide insight on user preference.**

What three plots did you make to best explain the data?

- Overall click rate of all topics (except one) are decreasing, with below having significant decreasing rate

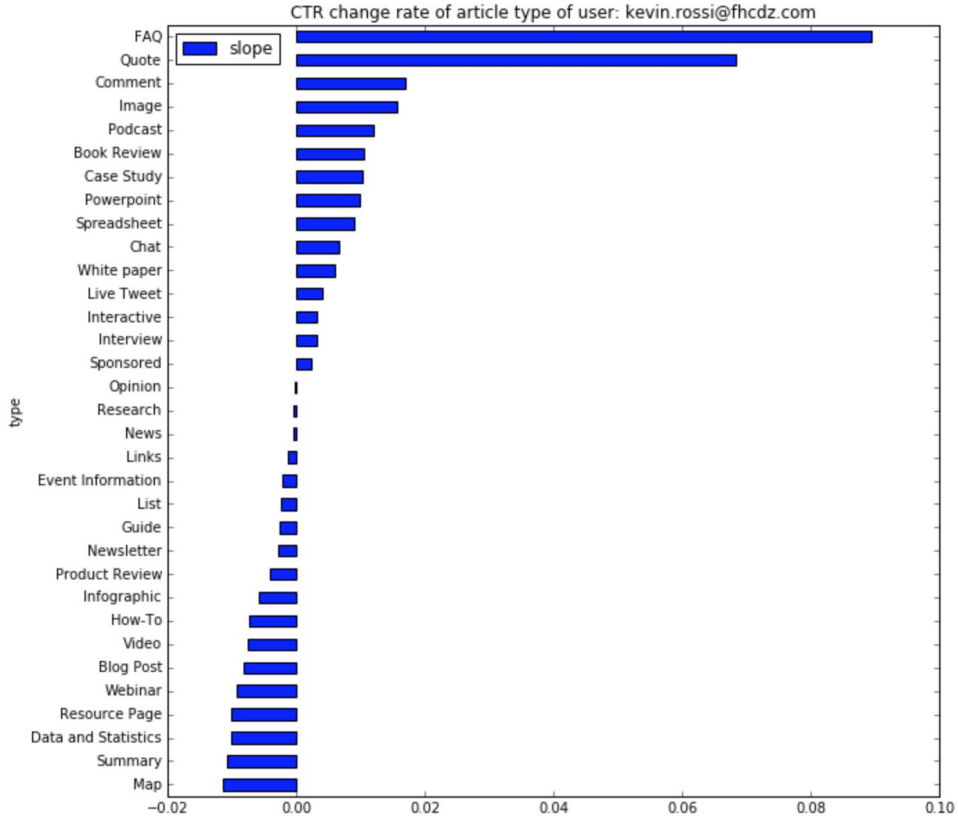


What three plots did you make to best explain the data?



- Top 5 topics with biggest decreasing click rates

What three plots did you make to best explain the data?



- Individually, significant CTR increasing and decreasing can be found for both article topic and type

What is your commercial recommendation for business unit heads who are non-technical?

- Overall effectiveness of engaging customer by email is decreasing
- User preference on article topic and type can be extracted from history data
- Email content should be customized for each user based on their likes and dislikes.
- Emailing frequency should be evaluated, too much emails would cause user fatigue and can become counter-productive
- More data should be collected for more granular analysis

What other data would you like to see about the platform? What questions would this additional data help you answer?

- Time spent on actual reading after open the article
  - to evaluate how user digest the article
  - to detect immediate close
- Link order in the email
  - to build learn-to-rank model based on relative preference extracted based on user clicking
- More article attributes, such as length, title
  - to enable more granular analysis
  - to evaluate the effect of title wording on user perception
  - to generate accurate summary that helps user to determine
- Link(s) within the article that was clicked by user
  - to help gain more insight on user preference