FINESSE Modeling Challenge

Engineering & Modeling Social Data

Goal & Context

Over the last years, the birthplace of fashion shifted from catwalks and magazines to social media. Consequently, social data constitutes the heart of FINESSE.
While somewhat cleaned up and pre-processed, this sample data set is supposed to expose you to the messy and variable nature of social media and challenge your thinking around feature engineering, data engineering and modeling.
The data set is a subset of Instagram posts during April 2020.

The goal with this data set is to establish the top five trending fashion posts across Instagram in this April 2020 subset. There is a lot one can do with the data set and it is easy to get lost in either one of feature engineering, data engineering or modeling. While I'd like to see you show off the way you code, I'm also interested in the way you think. To that end, please also outline & rationalize any features or modeling that you would implement were this not a timed challenge. Please clearly state your reasoning behind why certain modeling architectures would work best here, how you would modify them (or not) and how you would go about implementing them. This task should be treated as a hybrid of both a coding challenge as well as a research / implementation proposal.

Deliverables

How much you code up vs write up is totally up to you to show the best of your skills. The time requirement should be that of a regular, albeit challenging, coding take-home. The deliverables should include:

- A solid sample of coding work in either creative feature engineering, data engineering and / or modeling
- A written plan & reasoning (in any format that works best for you) for architecting a successful model including further feature engineering, data engineering and modeling (were there no time-constraints).

Data Structures

Attached to this challenge, you should find three different .CSV files.

- **post_metrics_and_comments**
  This is your main file. It includes metrics for Instagram posts including the post url, username, caption, image urls, max likes, max comments, max views (if video), followers of the posting profile, followings of the posting profile, all comments (delimited by the character "|"), and the date of posting.

- **hashtag_top_appearances**
  This table includes the amount of days a post was trending across leading streetwear community hashtags on Instagram. Trending, in this case, means appearing in the “top posts” section of certain pre-filtered, trending streetwear hashtags. This table includes the post_url and the associated days_in_hashtag_top_section metric.

- **raw_post_metrics**
  This table includes main post metrics collected on certain posts across different times in the four week period to monitor the posts’ growth (intervals between consecutive data collection attempts vary). The table includes the post_url, the num_likes, num_comments and num_views (if video) of the post on date_time_collected (the date & time the data was collected on the specific post).

Notes

This exercise is intended mostly to give an idea of how you approach engineering a system from scratch and think about its scalability & success. I’m more interested in seeing your process and structure as there is no one right answer. The documents should give us a great foundation for our follow-up call at which point you can walk me through your reasoning and deliverables step-by-step.

Most importantly, have fun and don’t hesitate to ask any questions!