Problem

You're leading the product team at a content reading platform such as Beedly. Over time you've collected some data about content and user interactions with the content. You now want to build a model that predicts virality of a given article.

The data collected is described below and can be downloaded from Kaggle:

Content:
The file **shared_articles** contains information about the articles shared in the platform.

Each article has its sharing date (timestamp), the original url, title, content in plain text, the article language (Portuguese - pt or English - en) and information about the user who shared the article (*author*).

There are two possible event types at a given timestamp:

- CONTENT SHARED: The article was shared in the platform and is available for users.
- CONTENT REMOVED: The article was removed from the platform.

Interactions:

This file **user_interactions** contains logs of user interactions on shared articles. It can be joined to **shared_articles.csv** by `contentId` column.

The eventType values are:

- VIEW: The user has opened the article.
- LIKE: The user has liked the article.
- COMMENT CREATED: The user created a comment in the article.
- FOLLOW: The user chose to be notified on any new comment in the article.
- BOOKMARK: The user has bookmarked the article for easy return in the future.

Some internal analysis has revealed that the metric representing virality is described as follows:

VIRALITY = 1* VIEW + 4*LIKE + 10*COMMENT + 25*FOLLOW + 100*BOOKMARK

You're to build a model that can predict virality of a new article being posted on the platform so that the news feed product can your model to showcase new articles.

Deliverables

The scope of the project is broad, please feel free to aggressively adjust it given your time constraints. Please don't spend a lot of time chasing the last bit of accuracy. What we would like to see:

- All the code you wrote to solve the problem, including the model and feature generation.
- A short document answering the following questions:
    - What features did you consider?
    - What model did you use and why?
    - What was your evaluation metric for this?
    - What features would you like to add to the model in the future if you had more time?
    - What other things would you want to try before deploying this model in production.

That said, please feel free to set your own time constraints and scope down as needed and (please do not ruin your weekend for this!). We sincerely appreciate the fact that you're willing to spend any time whatsoever on this. Thank you and have fun!

Hints

To play with the data, please feel free to use some of the previous notebooks on Kaggle.