Marketing Email Campaign

Goal

Optimizing marketing campaigns is one of the most common data science tasks. Among the many possible marketing tools, one of the most efficient is using emails.

Emails are great cause they are free and can be easily personalized. Email optimization involves personalizing the text and/or the subject, who should receive it, when should be sent, etc. Machine Learning excels at this.

Challenge Description

The marketing team of an e-commerce site has launched an email campaign. This site has email addresses from all the users who created an account in the past.

They have chosen a random sample of users and emailed them. The email let the user know about a new feature implemented on the site. From the marketing team perspective, a success is if the user clicks on the link inside of the email. This link takes the user to the company site.

You are in charge of figuring out how the email campaign performed and were asked the following questions:

- What percentage of users opened the email and what percentage clicked on the link within the email?

- The VP of marketing thinks that it is stupid to send emails to a random subset and in a random way. Based on all the information you have about the emails that were sent, can you build a model to optimize in future email campaigns to maximize the probability of users clicking on the link inside the email?

- By how much do you think your model would improve click through rate (defined as # of users who click on the link / total users who received the email). How would you test that?

- Did you find any interesting pattern on how the email campaign performed for different segments of users? Explain.

Data

We have 3 tables downloadable by clicking **here**.

The 3 tables are:

```
"email_table" - info about each email that was sent
```

Columns:

- **email_id** : the Id of the email that was sent. It is unique by email
- **email_text** : there are two versions of the email: one has "long text" (i.e. has 4 paragraphs) and one has "short text" (just 2 paragraphs)
- **email_version** : some emails were "personalized" (i.e. they had the name of the user receiving the email in the incipit, such as "Hi John,"), while some emails were "generic" (the incipit was just "Hi,").
- **hour** : the user local time when the email was sent.
- **weekday** : the day when the email was sent.
- **user_country** : the country where the user receiving the email was based. It comes from the user ip address when she created the account.
- **user_past_purchases** : how many items in the past were bought by the user receiving the email

---

```
"email_opened_table" -  the id of the emails that were opened at least once.
```

Columns:

- **email_id** : the id of the emails that were opened, i.e. the user clicked on the email and, supposedly, read it.

```
    "link_clicked_table" -  the id of the emails whose link inside was clicked
at least once. This user was then brought to the site.
```

Columns:

- **email_id** : if the user clicked on the link within the email, then the id of the email shows up on this table.

Example

```
    Let's check one email that was sent
```

head(email_table, 1)

Column Name	Value	Description
email_id	85120	The Id of the email
email_text	short_email	That was a short email
email_version	personalized	It was personalized with the user name in the text
hour	2	It was sent at 2AM user local time
weekday	Sunday	It was sent on a Sunday
user_country	US	The user is based in the US
user_past_purchases	5	The user in the past has bought 5 items from the site

```
    Let's check if that email was opened
```

subset(email_opened_table, email_id == 85120) >

<0 rows> (or 0-length row.names) # Nop. The user never opened it.

```
    We would obviously expect that the user never clicked on the link, since you
  need to open the email in the first place to be able to click on the link
  inside. Let's check:
```

subset(link_clicked_table, email_id == 85120)

<0 rows> (or 0-length row.names) # The user obviously never clicked on the link.