Email Marketing

John

Why Email Marketing

- 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
  - When to send: day of week, hour of day
  - What to send: personalized subject, content (text, images, products to recommend)
  - Who should receive it
- Machine Learning excels at this

真题解析

- 题目: Optimize marketing emails

- 喜欢考察这类题目的公司:
  - eCommerce companies
  - Basically any companies that need to do marketing

考点

- Problem formulation
  - Open Rate/Click Rate Prediction by Time
  - Recommendation and Personalization
  - Conversion rate prediction
- Feature extraction and label/target definition
- ML end-to-end workflow
- Experiment

A Take-home Problem

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:

1. What percentage of users opened the email and what percentage clicked on the link within the email?
2. 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?
3. 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?
4. Did you find any interesting pattern on how the email campaign performed for different segments of users? Explain.

Data

```
"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:

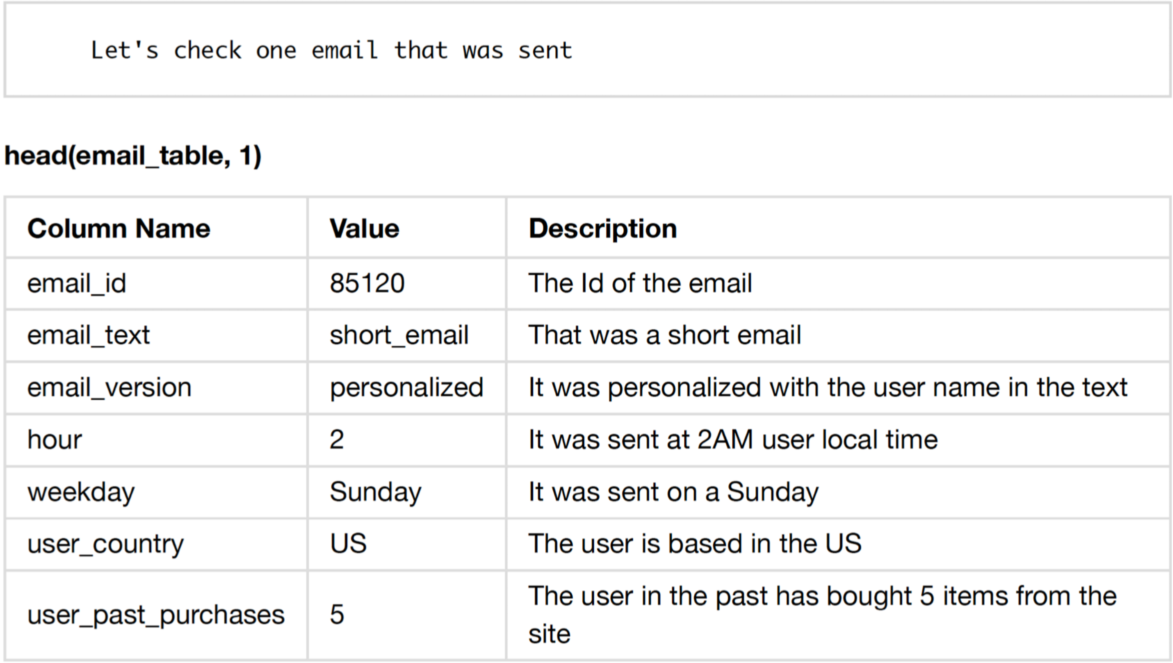
- **email_id** : the id of the emails that were opened, i.e. the user clicked on the email and, supposedly, read it.

```
    "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.

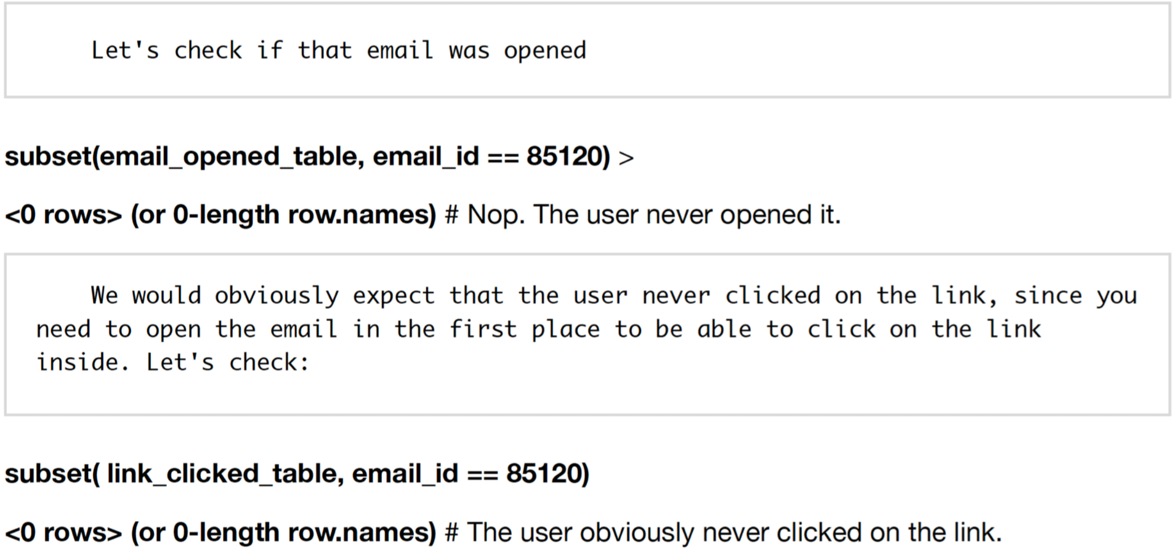
Data Examples



```
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.

What models can you built given data?

- Open Rate (Given Received) by Sending Time
- Click Through Rate (Given Received) Model by Sending Time
- Click Through Rate Model (Given Open) by Sending Time

Question 2

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?

- Given Question 3, let’s build Click Through Rate Model (Given Received) by Sending Time

Training Data for Click Through Rate Model

(Given Received)

- Model population
  - All emails
- Labels
  - 1 if clicked, otherwise 0
- Features and Insights
  - email_text : two values: "long text" and "short text”.
  - email_version : two values: "personalized" (i.e. they had the name of the user receiving the email in the incipit, such as "Hi John,"), "generic" (the incipit was just "Hi,").
  - hour : the user local time when the email was sent. Time should be a determining factor.
  - weekday : the day when the email was sent. Weekday should be a determining factor.
  - user_country : the country where the user receiving the email was based. People in different countries may have different email open/click habit; but first of all, they are at different Time Zones.
  - user_past_purchases : how many items in the past were bought by the user receiving the email. More purchase, higher the click rate.

Click Trough Rate Model Algorithms

- Supervised Methods
  - Logistic regression (Regularized)
  - Neural Networks
  - Random Forest
  - GBDT (XGBoost)

Feature transformation

- For Logistic Regression
  - email_text : 0/1 indicator
  - email_version : 0/1 indicator
  - hour : one hot.
  - weekday : one hot
  - user_country : one hot
  - user_past_purchases : binning and discretize
- For Tree Based Methods
  - email_text : 0/1 indicator
  - email_version : 0/1 indicator
  - user_country : one hot

Click (Through) Rate and Sampling

- Click Rates are low
    - Typically <1% ~ few percent
- Down Sampling Negatives (similar to fraud detection)
    - Effect of down sampling
        - AUC won’t be affected
        - Precision is increased at the same recall
        - Base fraud rate is higher after sampling, so model output probability is higher.
    - Scores need to be scaled to before down sampling

$$p_{scaled} = \frac{p}{p + (1 - p) \times neg\_ds\_factor}$$

neg_ds_factor is negative downsampling factor (e.g. 5 times), in downsampled data one negative represents neg_ds_factor examples in original data.

Click Trough Rate Optimization

- Time Optimization
  - Send email to each user that their best send time (highest click through rate), time being day of week, hour of day
- Audience Selection
  - Select audience segments with higher click through rate
  - Segment features can be
    - country
    - user_past_purchases
    - Our click through rate model

Question 3

By how much do you think your model would improve click through rate. How would you test that?

- How to estimate?
  - We can use our click through rate model to estimate the average click through rate for the audience we selected, and at the specific sending time
- How do we test?
  - We can run real time test, compare estimated click through rate from model with actual click through rate
  - What test should we run?

Question 4

- Did you find any interesting pattern on how the email campaign performed for different segments of users?
  - This is an analytics question, please dig into data.

Other Modeling Related Questions

- Cross-validation
- Hyper-parameter tuning
- Feature Selection
- Regularization
- Explain any ML methods you mentioned, why do you want to use them?
- Comparison between ML methods, pros and cons under different scenarios
- Open/click rate is likely low, do you want to apply sampling?
  - What are the effects on final conversion rate after sampling, how to correct?
  - How sampling affect AUC, precision and recall?

Other Concerns

- How to model open rate and click rate, and how to optimize them? How to optimize balancing open and click?
- If this is an eCommerce company, how to come up with a list of products to recommend to customer?
    - This is a recommendation/personalization problem
        - Logic to come up with item list
            - Recent add-to-cart items
            - Similar items to the add-to-cart items by collaborative filtering
        - A conversion model will help: rank item by conversion rate
    - How to predict conversion rate of the recommended items
        - Similar to ads conversion rate model: one-sided or pair-wised model
- Opt-out
    - Send too many emails to customers, customers may opt-out
        - How to predict opt-out rate?
        - How to optimize balancing open, click, and opt-out?