Data Science Competition - Lord of the Machines

Presented by -
Aakash Kerawat
Akshay Karangale

Outline

- Problem Statement
- Hypothesis Generation
- Feature Engineering
- Modelling
- What worked what didn’t
- Suggestions

Problem Statement

- We were required to predict the click probability of links inside a mailer for email campaigns from January 2018 to March 2018 which Analytics Vidhya sends to it’s users.
- We were provided with the information about the campaigns such as the actual mail sent, date time of the mail sent, etc.

Hypothesis Generation

Hypothesis Generation

- Before looking into the data we thought of factors that might affect the probability of a user clicking a link in the email.

Hypothesis Generation

- Individuals interests
- Time duration for which user is registered on AV
- Times for which user has participated in different events.
- Is the mail still relevant
- How active is the user on AV (number of visits, participations etc.)
- Frequency / Number of mails received by user.
- Time when received.

Feature Engineering

Target Encoding

Mean of target variable w.r.t some variable

KEY POINT -
This kind of encoding prevents data leakage and contradicting feature cases.

- For each mail sent to an user, we inserted the exponential weighted mean of past “is_click” variable of that particular user.
- Along with EWM of “is_click”, we also inserted the simple mean of “is_open” variable in the same fashion.

Count of mails sent of each communication type

- We created a set of features indicating the count of number of mails sent to each user of each communication type.
- We intended to use these features as a proxy for the times the user has participated in different events

Email number

- Out of the total mails sent to a particular user, what number of mail is the current mail.

Same as encoding case -
Email number rather than count of mail

Days since first mail

- We found the number of days since the first mail was sent to the user for that particular campaign (email).
- We came up with this feature to give the information about the duration for which user is registered on AV.

Date features

- We created simple features like day of month, day of week, hour,etc from send date to capture the seasonality from the data.

Time since last email was opened

- This feature was created to capture the recency of the user.

Modelling

LightGBM

- Our final submission was obtained by taking mean of 12 LightGBM models each with different seed.
- The hyperparameters were hand tuned to optimise on the cross validation score

What didn’t work

Features that did not work

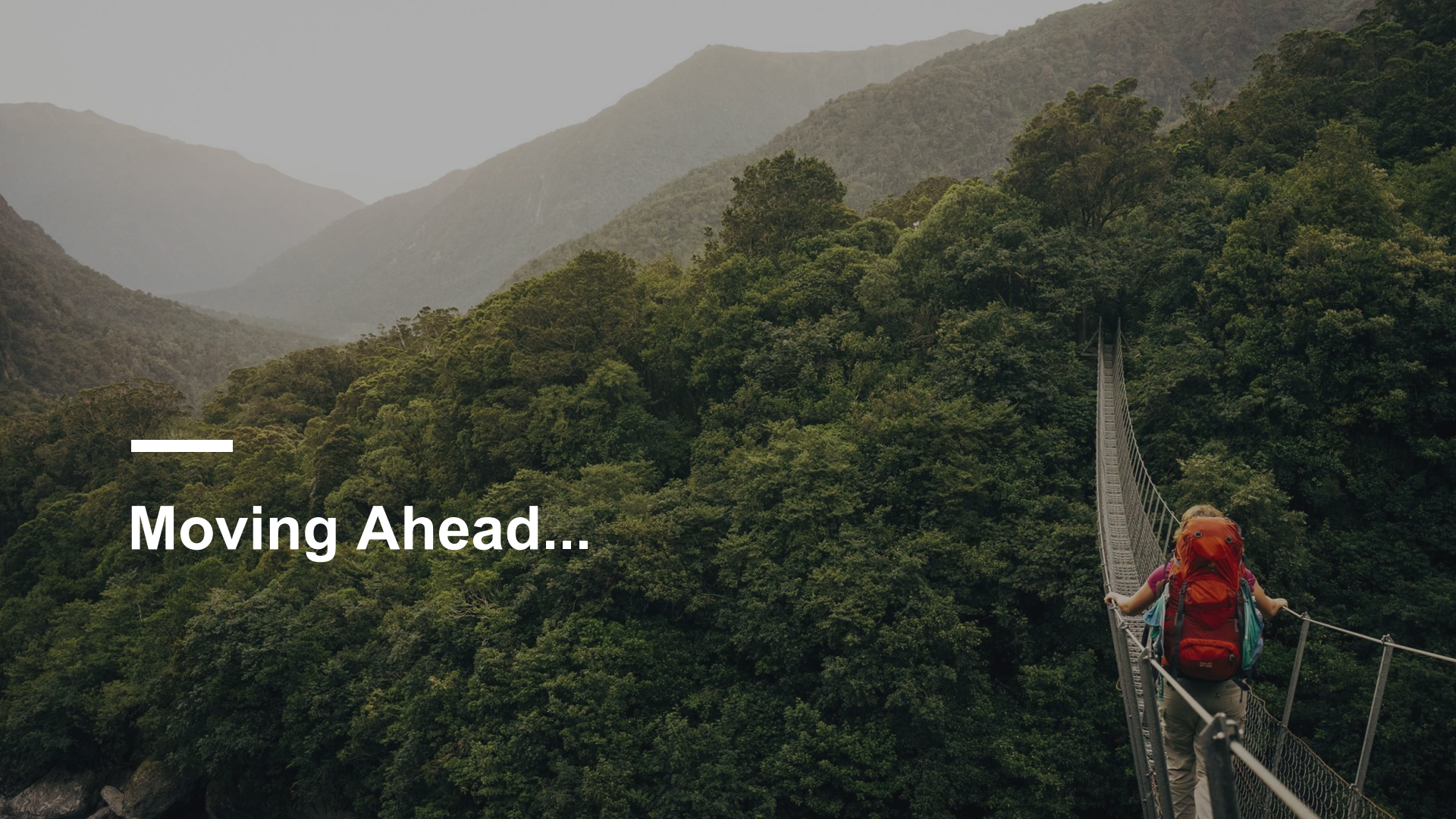
- Time since last mail was clicked
- Target encoding based on user_id and communication_type combined
- Length of subject and emails
- Countvectorizer of subject and email
- Dates parsed from emails and building date features from them

Modelling techniques that did not work

- We tried stacking with one and two layers but that did not seem to increase the score further.
- Tried different models like XGB and Catboost

Data Filtering that did not work

- In our first attempt we balanced the imbalanced target variables but that gave very poor score on LB
- We also tried making the distribution of train and test data on the basis of communication_type equal



Moving Ahead...

Further Improvement

- Data related to the user should be collected and used for the problem as the click probability would mostly be dependent on the user profile than the campaign details.
- Potential features could be:
  - Registration date on AV
  - Last seen datetime
  - Number of times user visits the site
  - Number of times the user has participated in different events

Thank you!

- Target Encoding (is_open, is_click) / exponential
- Count of comm_type mails
- Email number
- Days since first email
- Days since last email
- Dom, dow, hour etc.
- Time since last mail was opened 1 / was clicked 0
- ---
- Com_type, user combined target encoding
- subject _info
- Count_vectorizer
- Date parsed from email