Data Science Technical Test

Elisenda Vila

June 16th, 2015

Retention calculation

- 2nd day, 7th day and 10th day retention
  - Using user_id

datetime	Retention_Day2	Retention_Day7	Retention_Day10
2015-04-14	24.00 %	4.00 %	4.00 %
2015-04-15	30.77 %	0.00 %	7.69 %
2015-04-16	18.18 %	4.55 %	4.55 %
2015-04-17	26.92 %	3.85 %	3.85 %

Table : Retention according to user_id

Retention calculation

- Using client_mobile_device_aid

datetime	Retention_Day2	Retention_Day7	Retention_Day10
2015-04-14	37.50 %	6.25 %	6.25 %
2015-04-15	30.43 %	0.00 %	17.39 %
2015-04-16	20.00 %	5.00 %	5.00 %
2015-04-17	28.00 %	4.00 %	4.00 %

Table : Retention according to client_mobile_device_aid

Retention calculation

- As you might see, metrics differ when using different identifiers (user_id or client_mobile_device_aid)
- Why?
  - client_mobile_device_aid is not a unique identifier. Multiple users share the same client_mobile_device_aid

user_id	client_mobile_device_aid
3068771952811311104	6f050ff9-06d1-4d8a-b330-abc45e058366
3068901772492800000	6f050ff9-06d1-4d8a-b330-abc45e058366

Table : Duplicity example: 2015-04-14

Understanding retention
Compleating tutorial

- How much of the variation in the 2nd day retention can be explained by whether or not the user completed the tutorial on the 1st day?

Understanding retention
Compleating tutorial

- How much of the variation in the 2nd day retention can be explained by whether or not the user completed the tutorial on the 1st day?
- We should consider a generalized linear model since the response variable takes values TRUE / FALSE

Understanding retention
Compleating tutorial

- How much of the variation in the 2nd day retention can be explained by whether or not the user completed the tutorial on the 1st day?
- We should consider a generalized linear model since the response variable takes values TRUE / FALSE
- The best model is obtained by recoding **funnel_1d** variable into a logical variable that indicates if tutorial phase has been achieved (score higher than 2116)

Understanding retention
Compleating tutorial

- How much of the variation in the 2nd day retention can be explained by whether or not the user completed the tutorial on the 1st day?
- We should consider a generalized linear model since the response variable takes values TRUE / FALSE
- The best model is obtained by recoding **funnel_1d** variable into a logical variable that indicates if tutorial phase has been achieved (score higher than 2116)
- This model explains 10.24 % of the total variance

Understanding retention
Compleating tutorial

- If we include variables **funnel_5min** and **funnel_1hour** and it's interaction, the variance explained by the model achieves 11