A picture containing plate, drawing

Description automatically generated

The online ad industry is large. Despite many consumers stating they rarely click on ads, the industry’s success is powered by the long tail of the internet. In fact, the industry is expected to post 8% annual growth reaching more than $160B in revenue per year through 2023.

An important metric when selling ads, or bidding on ads is the “click through rate (CTR).” According to Wikipedia, CTR is the ratio of users who click on a specific link to the number of total users who view a page, email, or advertisement. It is commonly used to measure the success of an online advertising campaign for a particular website as well as the effectiveness of email campaigns. As a result, click prediction systems are essential and widely used for sponsored search and real-time bidding.

In this case, you are a business analyst working for [BookSends.com](https://booksends.com/). This service has more than 300,000 email users who periodically get offers for free and heavily discounted e-book offers. Your company makes money selling ad placements in the emails, getting paid for both CTR and conversion, when someone clicks then completes a discounted book offer.

Your boss “Carrie Trish Reynolds” has asked you to model the click through rates using historical data for 50,000 emails. She wants to know the CTR in the data, any insights learned through EDA, what features are important as well as how good your model is. This will aid her in improving future emails and informing the CPM (rate to charge advertisers). Carrie is extremely data fluent and will expect your presentation to be both engaging and one that demonstrates consistent logical workflow, understanding of the methods applied and how the model performs on unseen data.

## Data Notes

* This data is 100% synthetic though realistic. The business application is 100% real and the data itself has been vetted to be similar to actual data.
* The size of the data in this space is many 1000s of times bigger with CTR models often trained on tens of millions of records and with many more features such as IP address, more granularity on browser, and device type among other individual personal identifiers.
* CTR can be very low creating an “unbalanced” data set (lots of 0s, not many 1s). To improve results in your model the data science task of “over sampling” has already been performed on this data set, though it Is covered in the book.

## Example *Abridged* Data

Data has been provided in the class repository. Example code to read it in.

adData <- read.csv('Case2\_final\_adTech\_data.csv')

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Email\_ID | Subject\_Marketability\_Score | Customer\_Location | Total\_Past\_Communication | Word\_count | Total\_Links |
| 1 | 3.83 | Illinois | 12 | 287 | 1 |
| 8 | 4.87 | Wisconsin | 12 | 256 | 3 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Total\_Images | Signup\_Site | Genre | Sent\_Hour | BRowser | Y\_click |
| 3 | C | H | 00:00:00 | Chrome | 0 |
| 3 | A | B | 00:00:00 | Chrome | 1 |

## Data Dictionary

1. Email\_ID – unique Identifier for the email
2. Subject\_Marketability\_score – Using another algorithm, this is the expected “hotness” of the subject of the email; higher numbers indicate an higher expected CTR
3. Customer\_Location – Using IP Address, the state of the recipient was identified
4. Total\_Past\_Communucations – total number of past emails recieved
5. Word\_Count – Email text varies by length of book title, description and reviews to attract clicks and conversions. The company’s belief is that too little text doesn’t help sell to readers but too much is overwhelming
6. Total\_Links – total number of offers in the link
7. Total\_Images – total number of images used in the email
8. Signup\_Site – the company operates multiple email lists, for broader appeal including genre. The categories correspond to different digital properties owned by the company
9. Genre – the main genre of the email offering coded to factors such that “Mystery” is A etc.
10. Sent\_Hour – This is a timestamp
11. Browser – In leiu of personally identifiable information like specific device information this is the browser used to open the email
12. Y\_click – This is the dependent variable, 1 means the email was opened and clicked. Ignoring or opening but not clicking is 0.

## Criteria for Success

The case & presentation will be evaluated on the following criteria.

* **Organization** – Was the presentation well organized?
* **Delivery** – Was the content delivered clearly and persuasively with the audience in mind?
* **Documentation** – Was the data mined to support the conclusion?
* **Data Mining Process** – Did the you approach the problem similar (as applicable) to steps outlined in page 19 of the book or as covered in the lecture?