EXECUTIVE SUMMARY

Company Profile

The company is a financial media company that receives articles from a group of users and sends out personalized email digest articles to their wide network of users. On average, each user receives 179 links in a month.

The company aims to maximize customer engagement and this is measured through the number of clicks they receive for each link and user id.

Description of market

Each week, the client sends to the 20,000 registered users, personalized e-mails containing links to certain articles based on unknown high level metrics. Some of the users also contribute articles which are collected by the employees. 1,202 users (less than 5% of the users) are contributors of articles.

Of these 20,000 registered users, all the users read at least one or more article that is sent to them in the e-mail.

Summary of objectives

- Analyze the health of the business through the data provided
- Identify if and how the likelihood of a viewer's interest in an article has changed over time
- Provide sound recommendations to maximize the clicks and user engagement

DETAILED ASSESSMENT

Data Pre-processing

The given datasets are -

1. Data schema of users, articles, email_content, topics and type (SQLite)
2. Clicks on article links (Apache)

A SQL join in Pandas was done on the SQLite database to obtain a data frame that had a record of all articles that were sent to all the users along with other features such as author information, type and topic information. However, a feature of interest, indicating a click on each article link was missing and therefore, it was necessary to indicate whether the user had read the article that was sent to him.

The clicks on article links was transformed into a data frame using Pandas. Further, string manipulation was implemented to retrieve *article_id* , *user_id* and *send_time*. A *'target'* feature was added to the data frame, which has Boolean values 0 & 1 –

- 0: Article not read or missing click on article link
- 1: Article was read by the user.

On performing a left outer join on the given datasets, a final training data frame was obtained containing the dependent feature, *'target'.* This was saved as **finaltrainingdata2.csv** and is included in the deliverable package.

Data Modeling

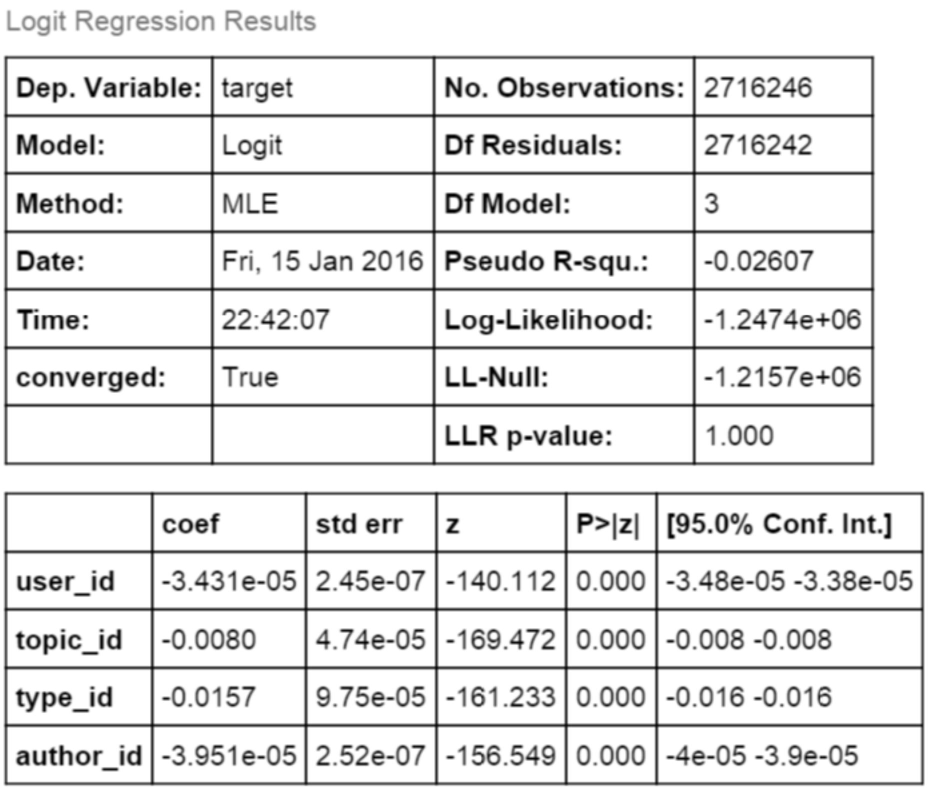
Assumption 1: Significant changes and trends happen on a monthly basis.

The objective was to see if there was a change in likelihood of prediction over a monthly basis. In the pre-processing phase, a new *send_month_time* feature was created which extracted the month from the date. This has values 1, 2 and 3 indicating months 1,2 and 3 respectively. To test the aforementioned objective, the following was the approach.

Approach in Steps

1. Segment the data based on month
2. Remove month as a feature
3. On data set for month 1, build the best prediction model
4. Use this model to predict clicks on data set of month 2
5. If the prediction error is high, a significant chance has happened which has caused the model to fail
6. Else, there is no significant change

Logistic Regression was identified as the best model with an accuracy of 83% in prediction. When the same logistic model was used in data sets for months 2 and 3, the final accuracy was as high as 83.6% which indicates that the model was a very good predictor for data sets in months 2 and 3. This is indicative of the fact that no significant changes happened over these months that would significantly impact the likelihood of an article being read by a user.



Logit Regression Results

Dep. Variable:	target	**No. Observations:**	2716246
Model:	Logit	**Df Residuals:**	2716242
Method:	MLE	**Df Model:**	3
Date:	Fri, 15 Jan 2016	**Pseudo R-squ.:**	-0.02607
Time:	22:42:07	**Log-Likelihood:**	-1.2474e+06
converged:	True	**LL-Null:**	-1.2157e+06
		LLR p-value:	1.000

	coef	**std err**	**z**	**P>|z|**	**[95.0% Conf. Int.]**
user_id	-3.431e-05	2.45e-07	-140.112	0.000	-3.48e-05 -3.38e-05
topic_id	-0.0080	4.74e-05	-169.472	0.000	-0.008 -0.008
type_id	-0.0157	9.75e-05	-161.233	0.000	-0.016 -0.016
author_id	-3.951e-05	2.52e-07	-156.549	0.000	-4e-05 -3.9e-05

Based on the absolute value of z statistic, the top three features of importance are *type_id, topic_id* and *author_id.*

Based on the absolute value of z statistic, the top three features of importance are *type_id, topic_id* and *author_id.*

TOOLS USED

- Language: Python
- Environment: ipython notebook
- Packages: pandas, sklearn, sqlite3
- Visualization toolkit: Seaborn, Pandas

TECHNIQUES IMPLEMENTED

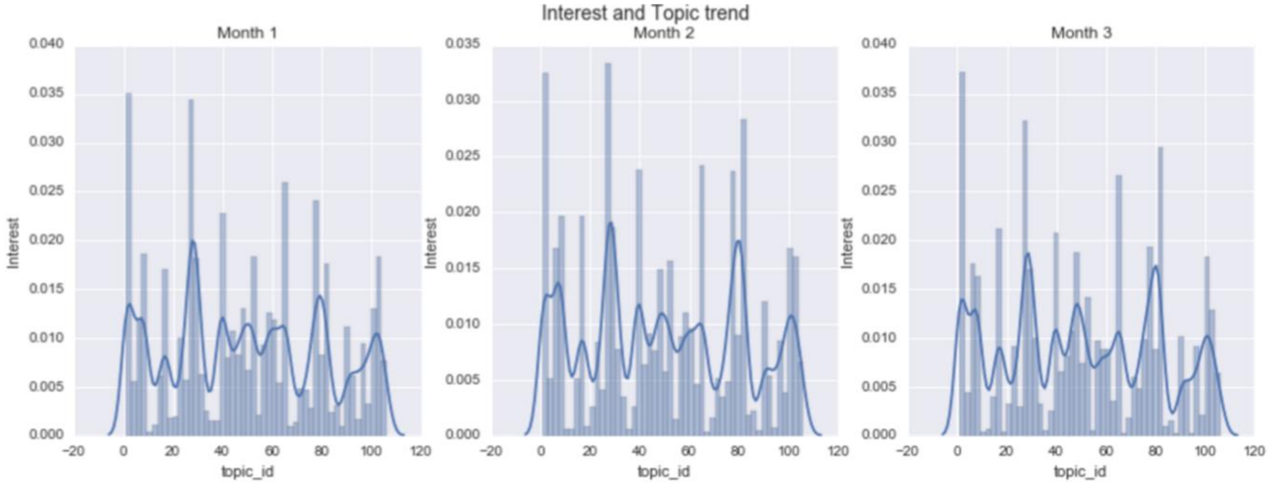
- Naïve Bayes technique
- Logistic Regression

PLOTS

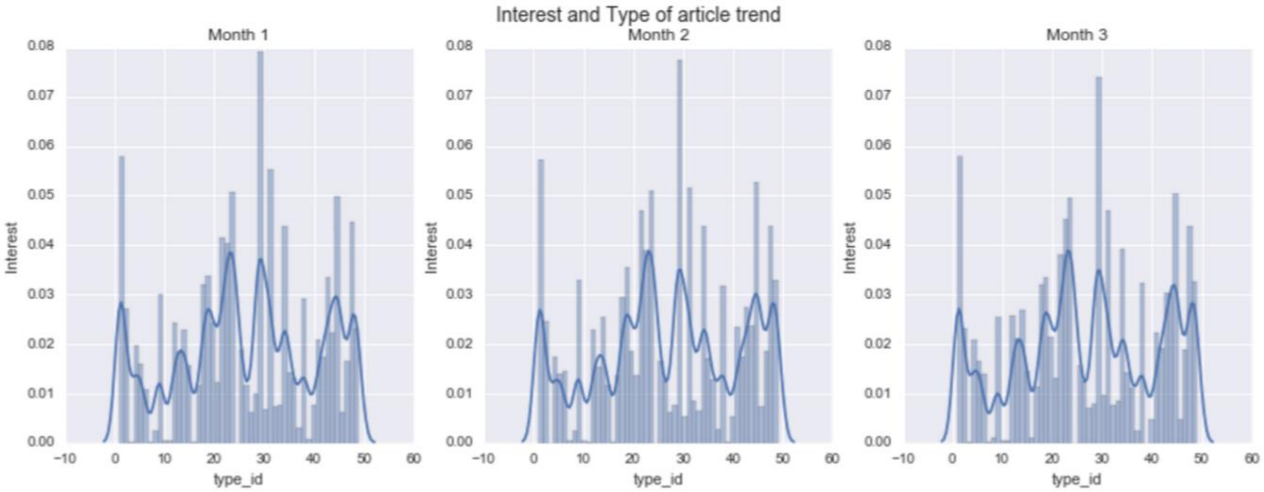
The three plots used are:

- Interest vs. Topic of article trend
- Interest vs. Type of article trend
- Interest vs. Author of article trend

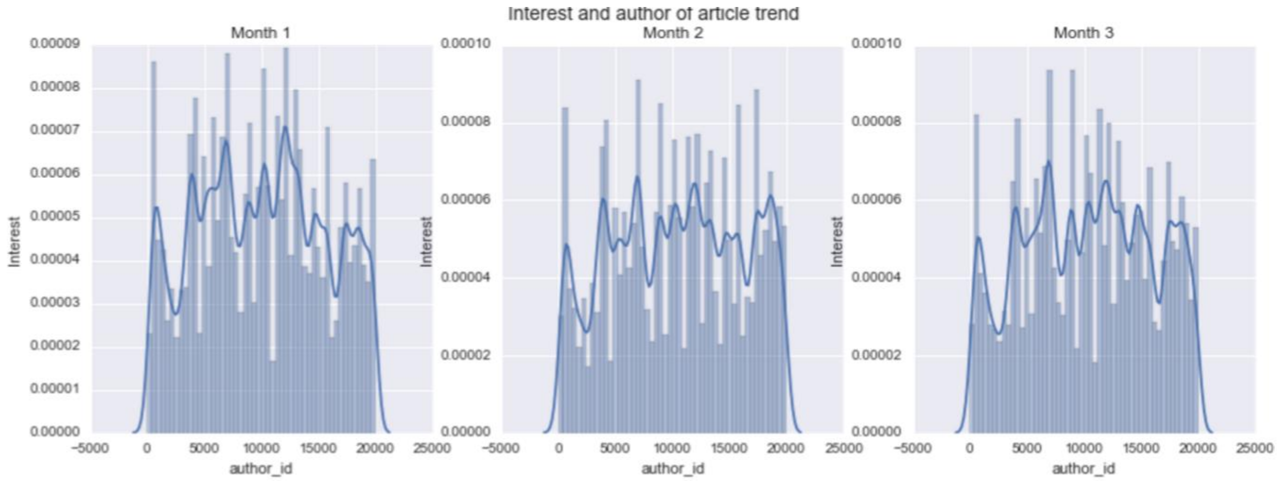
Interest is calculated as the count of the total number of topics of a type, topic or author that have been read



The pattern over the months 1, 2 and 3 are similar and haven't changed by much. This is indicative that the interest over a certain set of topics of articles still remain the same.



The pattern over the months 1, 2 and 3 are similar and haven't changed by much. This is indicative that the interest over a certain set of types of articles still remain the same.



The pattern over the months 1, 2 and 3 are similar and haven't changed by much. This is indicative that the interest over a certain set of authors of articles still remain the same.

COMMERCIAL RECOMMENDATION

- Segment the users into multiple groups who share similar interests in topics and types.
- Send only links that they are most likely to read; this is a customer retention strategy; if the client does not monitor their users' interests and send links that they may not read, the users may lose interest.
- Provide incentives to encourage more authors to write interesting articles that will be read by most users.

OTHER USEFUL DATA

- **Time spent on the link**
  One of the assumptions that was made on the data was that if the link was clicked, the article was read. However, it cannot be a true measure of engagement. We may need more information to assess the user's interest in that article. Time spent on the article is a good measure of whether the article was actually read or not.

- **User rating of article**
  User rating of the article is a direct measure of whether a particular user liked the article or not. We can use this ratings to even analyze the topics and types that are losing interest over time (if they are) and keep the client's authors informed of such trends.

- **Analyzing influence of networks**
  Some articles gain a lot of attention because of several reasons. If there is an evident network of user groups who interact with each other by sharing articles, it would be useful to identify ‘influencers’ of the group who can spread a good article and popularize it. This way we can keep our influencers satisfied with articles they receive and they can help us maximize clicks by sharing articles across their networks

- **Phrases used in the title of the article**
  Sometimes, the title of an article has a lot to do with catching attention and gaining a click. An analysis of the common phrases, jargons or text can help the client identify which words can capture attention and maximize clicks. This can be communicated to the authors and trendier words can be use