The News Lens

News Worth Knowing, Voices Worth Sharing.

Predicting readability of The News Lens' next online articles to enhance reader loyalty

Team Ideation Proposal Report - Team 4

Uniss Tseng 106077503 Elisa Wang 105078514 Patrizia Mach 106077429 Sabrina Wei 102033227

Executive Summary

In order to establish a dedicated reader base, online news website The News Lens aims to drive traffic directly to their website rather than via third-party social media, such as Facebook. Establishing this goal involves selecting a list of featured articles to display on the homepage, which are most likely to be read completely and aid in establishing a reader habit to primarily use The News Lens for its insights to current events.

Using a linear regression and predictive data mining, we establish, based on the 752 articles of data provided by The New Lens, a model that gives out predictive readability scores for any new articles. This model has been chosen for its high performance and its ease to understand and apply. The predicted scores can be sorted from highest to lowest, such that the highest scores are articles that are most likely to be read to completion and create the desired user traffic to the website.

The model using the October 2017 data shows that changes in word count, certain authors, the number of articles they write and changes in the category of the articles are related to changes in the likelihood of an article to be read fully. Therefore, it would be prudent to review and revise the model regularly as business activities and reader preferences change. For a more consistent and accurate model, it is further suggested to apply a larger database spanning for example 6 months.

Business Background

The News Lens is a company committed to bilingual online news articles with insight operating in Taiwan and Hong Kong. Due to primarily functioning online, the website layout and design is a prime concern to the editors at The News Lens.

Objective

The homepage must be engaging and draw readers to its best articles, so that the company becomes their prime insight provider to current events and thus creating a sustainably growing reader base. In order to achieve this goal, The News Lens needs to identify which articles to feature as highlight based on which articles are most likely to be read. A majority of traffic directed towards the website stems from Facebook. However, as it is currently, measuring the number of Facebook shares, page views and likes is insufficient in determining the readability of an article (readability defined as likelihood to be read to completion). By more accurately determining which articles have a better chance at being read and thus create more returning user traffic, The News Lens aims to increase its reader loyalty by forming a base of readers who come to The News Lens website regularly for insights on current news. A larger reader base then signals a powerful impact on readers to potential sponsors in order to raise revenue.

Data Mining

Amongst the user data collected by the company is a readability score. This is a ratio between the reading spending time (amount of time a webpage is viewed) and a reading scroll ratio (proportion that the reader has scrolled through). The News Lens informed us that a readability score of 0.70 is considered a benchmark score for a good and readable article. Therefore, the readability score is the continuous output variable in the model to be determined. In order to develop a model which can show the next most likely to be read articles to be featured under highlights, we employ a supervised and forward-looking task that will help determine a ranking of the articles with top readability scores which shall translate to the order in which articles should be layout on the website.

Data

The data provided included click stream data from website traffic, readability recorded data and article metadata for October 2017. Although having received over 5 million lines in data, our objective to predict the readability of the next online article constrained our dataset. The Clickstream Data contained many illegal variables pertaining to how the articles were accessed and, as such, these features had been eliminated during the cleaning process. Following a merger of the three datasets using page ID for the article, our data contained 752 lines representing articles and a 68 columns despite reduction. Due to the large amount of features compared to rows, we have applied feature selection as described in the Solutions section.

An overview of the raw clickstream and readability data is provided in the appendix.

Data Preparation

Missing values were handled by the data scientists at The News Lens directly. The focus of data preparation for the team had been in processing, combining and deriving interesting variables as potential predictors. Variables that contain duplicated information for each article and illegal variables containing information unavailable at the time of publishing have been removed to allow predictive modelling. This includes the majority of access information collected.

Derivative variables depending on existing features being collected such as the time of day and weekday have been created from the publishing time stamp. In addition, we have applied text mining techniques and obtained vectors specific to labels obtained through one-hot encoding, however these had been inconclusive due to a lack of weighting to keywords and the time required to generate a hot-words dictionary.

Finally, the readability and article data has been merged via page_id. The final data set contained a significant amount of categorical predictors for which dummy variables had been created. The variable Author_id contained 170 categorical outcomes and thus we have reduced these to 19 categories by frequency.

A partition of 70-30 had been applied, such that 30% of the 752 lines representing articles have been randomly selected as test set and the remaining 70% used in a cross-validation of 10 folds due to the low number of articles available.

The final set of merged data appears as below:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	day	Time of Da	domain_id	type_id	main_cate	author_id	photo_num	rich_media	p_gt20_nu	h5_num	qoute_num	word_num	link_num	is_sponso	score
2	6	Evening	1	1	16	1000	0	0	13	0	1	1196	8	0	0.827632
3	5	Evening	1	2	20	64285	1	2	10	3	0	1146	1	0	0.936984
4	5	Day	1	2	3	92	6	6	13	2	0	1105	6	0	0.870739
5	1	Evening	1	2	16	64351	0	0	22	0	0	1982	5	0	1.044909
6	2	Morning	4	2	23	64347	3	3	9	0	0	1370	3	0	0.722003
7	7	Day	1	2	16	2201	3	3	13	2	0	1235	2	0	0.686346
8	2	Morning	1	2	20	64509	3	5	18	3	0	2323	6	0	0.546698
9	6	Morning	4	2	24	3396	7	8	17	0	2	2068	15	0	0.544216
10	2	Night	1	1	5	64135	2	3	42	9	4	3768	1	0	0.61373
11	5	Day	1	2	3	63739	1	1	15	4	0	1650	2	0	0.65695
12	6	Day	1	2	16	53604	1	1	33	3	0	3497	8	0	0.698723
13	5	Morning	1	2	14	64636	2	2	25	0	9	3247	35	0	0.697227
14	2	Evening	1	1	5	64135	3	3	41	9	3	4063	13	0	0.682489
15	5	Day	1	2	21	64334	6	6	21	6	3	2543	9	0	0.170908

Column O shows the continuous output variable: readability score.

Solution

In the development of the appropriate data mining model, we have applied a variety of methods on several platforms, including XLminer and RapidMiner due to the function constraints of the former. Initially we had run regression, KNN and regression tress on XLminer with a 50-30-20 data partition due to XLminer limitations. However, due to the small amount of data versus number of predictors, we see in the XLminer model scoring output that the models were either overfitting with exceedingly small average errors or did not outperform the Naïve benchmark of the average readability score.

In response, we employed RapidMiner to introduce cross validation to alleviate the problem of a small dataset and select a smaller subset of features using feature optimization where a subset of predictors had been selected using a supervised forward selection algorithm. Similar to before, we have then created models using Linear Regression, KNN, and Regression Tree in addition to Random Forest, and Ensemble to determine which of the models would perform best compared to the Naïve Rule and taking into consideration the 0.7 benchmark set by The News Lens. As can be observed in the appendix featuring RapidMiner Model Scores, the RMSE and average absolute errors of the RapidMiner models, cross validation and feature optimization has created a better fit of models.

The selected final model is Linear Regression firstly due to its parsimony and transparency in conveying relevant predictor variables. The Ensemble, although with a slightly higher performance contains a multitude and mix of the above named models and the complexity to run it compared to the insignificant improvement in error favors the following Linear Regression Model:

Model	**- 8.84E-05 * word_num- 0.088 * main_category_id_23- 0.074 * author_id_red_1- 0.323 * author_id_red_11- 0.149 * author_id_red_3+0.859**
Naïve RMSE	**0.636099782**

where main_category_id_23 refers to “Politics” and the several author IDs to specific authors as key coded by The News Lens. An example of the top 11 highest predicted readability scores is shown in appendix 6. Depending on the number of highlighted articles on the homepage required, the editor can now choose which article to feature based on the highest readability score (likelihood to be read completely).

Conclusive remarks

The solution provided depends on existing reader preferences to certain authors and categories, as such, it is recommended to revise and retrain the model regularly as large relevant current events happen and reader preferences or business activities change. We understand that the other methods have sometimes provided different optimized features, such as time of day, weekday or number of html_5 tags, which may also appear relevant when retraining the model with a larger article meta dataset of, for example, 6 months. Therefore, we suggest a larger dataset over a longer period for a more accurate and robust model.

In an analysis of the correlations between selected variables and the readability score output variable, we note the strongest relevant correlation to be the negative correlation between word count and readability, indicating that long articles are less likely to be completely read.

Other correlations of interest have been included in the appendix.

Appendix

1. List of Variables as per final merged Dataset

1. **rich_media** includes photo, facebook post, twitter, etc.
2. **p_gt20_num** is the number of html tag “p” includes more than 20 words
3. **h5_num** is the number of html h5 tags
4. **weekday** represents the day of the week in which an articles was published
5. **time of day** has been derived from the time stamp of publishing to indicate whether an articles was published during morning, day, evening or night.
6. **Word count** is the number of words in an article
7. **is_sponsor** , 1 if the article has been commissioned by a sponsor.
8. **reading scroll ratio** is a ratio of how much readers scroll down the article page
9. **reading spending time** represents how long the readers spend reading the article
10. **readability score** combines the performance of reading scroll ratio and reading spending time

2. Raw data examples prior to merger with article metadata

Clickstream Data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	page_id	title	pub_datetime	domain	domain_i	type	type_id	main_ca	main_category	author	author_id	photo_num	rich_media_num	p_gt20_num	h5_num	qoute_num	word_num	link_num	is_sponsor
2	52472	總裁現出「我有舉過	2017-01-03 07:51	臺灣	1	評論	2	生活	14	精選書摘	34176	8	8	14	8	0	1301	3	0
3	56995	宋朝平盧怎麼點？防	2017-01-01 09:37	臺灣	1	評論	2	生活	14	精選書摘	34176	3	3	14	1	0	1576	3	0
4	56997	宋朝的書節很像西方	2017-01-07 10:20	臺灣	1	評論	2	生活	14	精選書摘	34176	2	2	20	1	3	1663	3	0
5	58042	「不可以XXX！」父母	2017-01-06 08:32	臺灣	1	評論	2	生活	14	精選書摘	34176	1	1	39	16	0	2727	4	0
6	58047	九種典型的「認知扭	2017-01-06 17:01	臺灣	1	評論	2	生活	14	精選書摘	34176	2	2	44	12	0	2768	4	0
7	58288	從「Lagu」這個越南	2017-01-03 11:25	臺灣	1	評論	2	生活	14	精選書摘	34176	1	2	31	0	0	2214	2	0
8	57403	創造大衛勉伊	2017-01-10 07:56	臺灣	1	評論	2	藝文	20	精選書摘	34176	2	7	24	1	4	2920	24	0
9	57443	大衛勉伊「柏林三部	2017-01-10 07:56	臺灣	1	評論	2	藝文	20	精選書摘	34176	1	1	36	1	0	4979	43	0
10	57446	被王爾德揭開深藏的	2017-01-02 18:24	臺灣	1	評論	2	藝文	20	精選書摘	34176	1	1	45	2	0	4519	4	0
11	57657	紀傳《遺愁傳》：我	2017-01-03 07:56	臺灣	1	評論	2	藝文	20	精選書摘	34176	1	1	42	4	0	5400	5	0
12	58053	纸上明治村：臺北城	2017-01-01 17:06	臺灣	1	評論	2	藝文	20	精選書摘	34176	6	6	22	5	0	3505	4	0
13	58054	纸上明治村：百貨公	2017-01-01 09:37	臺灣	1	評論	2	藝文	20	精選書摘	34176	6	6	22	5	0	3561	4	0
14	58343	惡僧是這樣誕生的：-	2017-01-07 18:20	臺灣	1	評論	2	藝文	20	精選書摘	34176	1	1	38	2	0	3913	3	0
15	58520	她讓擺置在總司美索	2017-01-05 12:05	臺灣	1	評論	2	藝文	20	精選書摘	34176	2	2	22	1	0	1988	11	0
16	58822	表總遭境男孩希望成	2017-01-09 12:42	臺灣	1	評論	2	藝文	20	精選書摘	34176	2	2	24	1	1	1960	6	0
17	57839	所有集韓暴力都不可	2017-01-06 07:56	臺灣	1	評論	2	社會	16	精選書摘	34176	1	1	42	3	5	5444	6	0
18	57844	金錢並不總是能夠激	2017-01-06 17:06	臺灣	1	評論	2	社會	16	精選書摘	34176	1	1	45	3	2	6497	5	0

Readability Data:

	A	B	C	D	E	F	G	H	I	J	K
1	Reading scrolling ratio	Reading spending time ratio	Readability score	Reliabilityof readabi	Page views of Part 1	median of reading time	median of finished readin	Five parts of rea	Page views of Par	Page views of Par	Reading prog
2	[3.5以上為佳] 讀者閱讀	[0.6以上為佳] 讀者閱讀文章	[0.7以上為佳] 綜	0.000	包含instant article，讀者閱讀文章秒數的中	讀者閱讀完成花費秒數的中位數					
3	4.653	0.928	1.045	0.127	1482	182.884	197.002	98_95_93_92_87	0	0	1241
4	4.480	0.912	0.980	0.074	473	42.002	46.072	97_96_95_93_67	0	0	423
5	4.447	0.908	0.967	0.065	502	54.679	60.252	96_94_92_90_72	0	0	460
6	4.443	0.803	0.950	0.160	1065	30.241	37.667	97_94_90_86_77	0	0	950
7	4.354	1.009	0.949	0.065	2417	22.226	22.034	95_92_89_86_74	0	0	594
8	4.396	0.856	0.941	0.092	150	32.440	37.897	95_94_90_88_72	0	0	134
9	4.392	0.859	0.940	0.088	294	40.502	47.129	97_95_92_87_68	0	0	263
10	4.372	0.904	0.940	0.038	753	106.090	117.304	98_94_91_82_73	0	0	646
11	4.360	0.916	0.937	0.023	4084	106.731	116.581	97_93_90_82_73	0	0	3603
12	4.275	0.954	0.912	0.045	79	35.146	36.845	88_88_88_86_77	0	0	69
13	4.298	0.892	0.911	0.021	4711	70.567	79.124	95_93_91_87_63	0	83	1942
14	4.312	0.846	0.909	0.068	990	68.224	80.596	93_91_88_83_77	0	0	836
15	4.275	0.932	0.909	0.025	151	28.621	30.696	95_93_90_86_63	0	0	138
16	4.313	0.841	0.909	0.073	1667	52.778	62.744	96_94_90_84_67	0	0	1412
17	4.299	0.847	0.904	0.062	337	79.314	93.595	93_90_87_85_76	0	0	291
18	4.255	0.917	0.899	0.020	134	50.200	54.716	91_91_86_83_76	0	0	98
19	4.279	0.835	0.895	0.066	6888	70.776	84.805	95_90_86_83_74	0	2080	2762
20	4.258	0.854	0.891	0.040	240	29.217	34.229	92_90_88_84_73	0	0	213
21	4.264	0.806	0.886	0.087	241	109.926	136.385	95_91_85_80_75	0	0	208

3. XLminer Model Scoring Output

Naïve-Training

RMS Error	Average Error
0.201724	4.3361E-16

Linear Regression
Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
9.0618	0.15524	2.3E-16

KNN
Training Data Scoring - Summary Report (for k = 9)

Total sum of squared errors	RMS Error	Average Error
0.00235	0.0025	-4E-18

Regression Trees
Training Data scoring - Summary Report (Using Full-Grown Tree)

Total sum of squared errors	RMS Error	Average Error
7.13174	0.13772	3.2E-18

Naïve-Validation

RMS Error	Average Error
0.200702	-3.639E-05

Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
5.22972	0.15212	-0.0007

Validation Data Scoring - Summary Report (for k = 9)

Total sum of squared errors	RMS Error	Average Error
6.22008	0.1659	0.01779

Validation Data scoring - Summary Report (Using Full-Grown Tree)

Total sum of squared errors	RMS Error	Average Error
8.4436	0.19329	0.01836

Naïve-Test Data

RMS Error	Average Error
0.238865	2.2182E-05

Test Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
6.14001	0.20232	-0.0235

Test Data Scoring - Summary Report (for k = 9)

Total sum of squared errors	RMS Error	Average Error
7.43411	0.22262	-0.0242

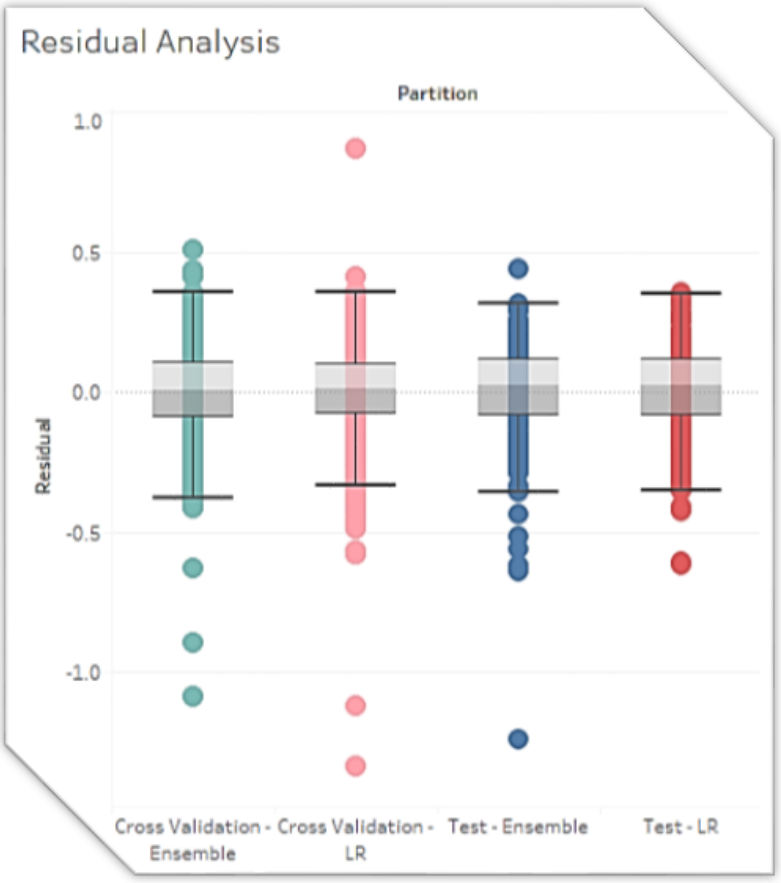
Test Data scoring - Summary Report (Using Full-Grown Tree)

Total sum of squared errors	RMS Error	Average Error
7.07404	0.21716	-0.0287

4. RapidMiner Model Scoring Output

Ensemble Test	Linear Regression Test	Random Forest Test	Single Tree Test	Naïve Test	kNN Test
root_mean_squared_error: 0.182 +/- 0.000	root_mean_squared_error: 0.156 +/- 0.000	root_mean_squared_error: 0.173 +/- 0.000	root_mean_squared_error: 0.177 +/- 0.000	root_mean_squared_error: 0.201 +/- 0.00	root_mean_squared_error: 0.181 +/- 0.000
absolute_error: 0.129 +/- 0.129	absolute_error: 0.121 +/- 0.099	absolute_error: 0.127 +/- 0.117	absolute_error: 0.144 +/- 0.103	absolute_error: 0.159 +/- 0.123	absolute_error: 0.148 +/- 0.105
squared_error: 0.033 +/- 0.112	squared_error: 0.024 +/- 0.045	squared_error: 0.030 +/- 0.075	squared_error: 0.031 +/- 0.041	squared_error: 0.040 +/- 0.064	squared_error: 0.033 +/- 0.042
Validation	**Validation**	**Validation**	**Validation**	**Validation**	**Validation**
root_mean_squared_error: 0.163 +/- 0.030 (mikro: 0.165 +/- 0.000)	root_mean_squared_error: 0.168 +/- 0.040 (mikro: 0.173 +/- 0.000)	root_mean_squared_error: 0.174 +/- 0.032 (mikro: 0.177 +/- 0.000)	root_mean_squared_error: 0.175 +/- 0.037 (mikro: 0.179 +/- 0.000)	root_mean_squared_error: 0.211 +/- 0.034 (mikro: 0.214 +/- 0.000)	root_mean_squared_error: 0.178 +/- 0.033 (mikro: 0.181 +/- 0.000)
absolute_error: 0.124 +/- 0.013 (mikro: 0.124 +/- 0.109)	absolute_error: 0.120 +/- 0.018 (mikro: 0.120 +/- 0.124)	absolute_error: 0.129 +/- 0.018 (mikro: 0.129 +/- 0.121)	absolute_error: 0.128 +/- 0.016 (mikro: 0.129 +/- 0.125)	absolute_error: 0.162 +/- 0.024 (mikro: 0.162 +/- 0.140)	absolute_error: 0.130 +/- 0.015 (mikro: 0.130 +/- 0.125)
squared_error: 0.027 +/- 0.011 (mikro: 0.027 +/- 0.071)	squared_error: 0.030 +/- 0.015 (mikro: 0.030 +/- 0.106	squared_error: 0.031 +/- 0.012 (mikro: 0.031 +/- 0.101)	squared_error: 0.032 +/- 0.013 (mikro: 0.032 +/- 0.105)	squared_error: 0.046 +/- 0.016 (mikro: 0.046 +/- 0.104)	squared_error: 0.033 +/- 0.012 (mikro: 0.033 +/- 0.100)
Training	**Training**	**Training**	**Training**	**Training**	**Training**
root_mean_squared_error: 0.148 +/- 0.007 (mikro: 0.149 +/- 0.000)	root_mean_squared_error: 0.164 +/- 0.004 (mikro: 0.165 +/- 0.000)	root_mean_squared_error: 0.167 +/- 0.003 (mikro: 0.167 +/- 0.000)	root_mean_squared_error: 0.172 +/- 0.004 (mikro: 0.172 +/- 0.000)	root_mean_squared_error: 0.213 +/- 0.004 (mikro: 0.213 +/- 0.000)	root_mean_squared_error: 0.162 +/- 0.004 (mikro: 0.162 +/- 0.000)
absolute_error: 0.110 +/- 0.004 (mikro: 0.110 +/- 0.100)	absolute_error: 0.118 +/- 0.002 (mikro: 0.118 +/- 0.115)	absolute_error: 0.121 +/- 0.002 (mikro: 0.121 +/- 0.114)	absolute_error: 0.122 +/- 0.001 (mikro: 0.122 +/- 0.121)	absolute_error: 0.161 +/- 0.003 (mikro: 0.161 +/- 0.140)	absolute_error: 0.117 +/- 0.002 (mikro: 0.117 +/- 0.112)
squared_error: 0.022 +/- 0.002 (mikro: 0.022 +/- 0.063)	squared_error: 0.027 +/- 0.001 (mikro: 0.027 +/- 0.080)	squared_error: 0.028 +/- 0.001 (mikro: 0.028 +/- 0.093	squared_error: 0.030 +/- 0.001 (mikro: 0.030 +/- 0.100)	squared_error: 0.045 +/- 0.002 (mikro: 0.045 +/- 0.103	squared_error: 0.026 +/- 0.001 (mikro: 0.026 +/- 0.079)

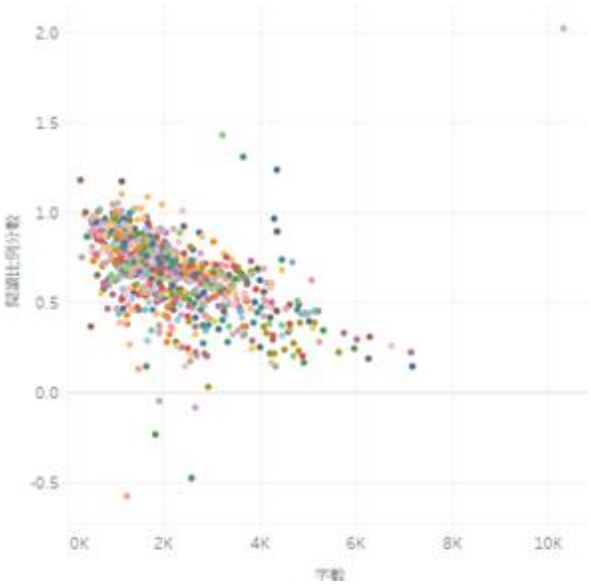
5. Linear Regression and Ensemble Residual Analysis



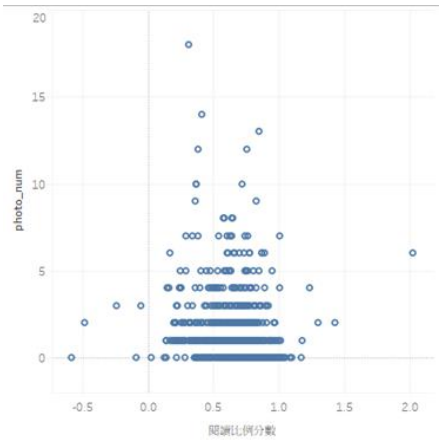
6. Predicted Readability Score Sorted from Highest to Lowest

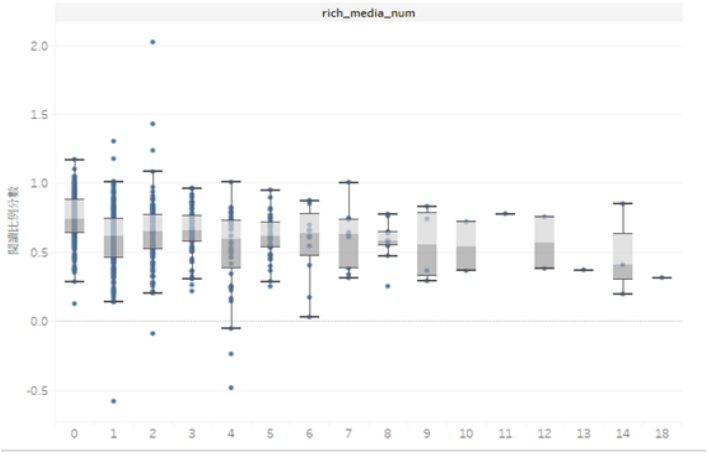
score	predictio... ↓
1.176	0.835
0.999	0.827
0.959	0.816
0.803	0.815
0.909	0.812
0.980	0.811
0.766	0.806
0.934	0.804
1.002	0.803
0.867	0.802
1.013	0.801
0.912	0.799

7. Correlations of Interest

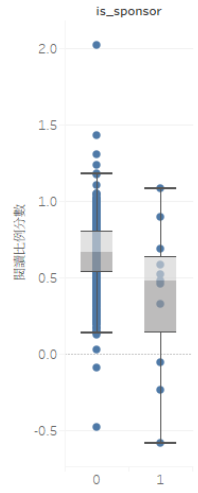


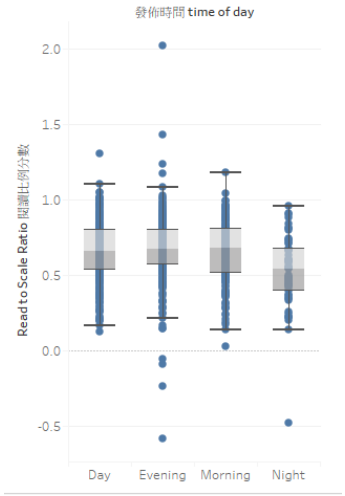
↑Word Count vs. Readability correlation: Indicates shorter articles better chance at reading to completion.





↑Photo or Rich Media content seems not linked to Readability: Company mentioned they spend a lot of cost to produce the picture. However, observing the behavior here, we can review cost and benefit of informatics again to see if it could gain benefit or not





← Night Articles vs Readability: Indicate Less reading outside office hours. Company could set a post strategy based on the behavior

↑ Sponsored Articles vs Readability: Indicate sponsored articles has lower readability. Company need to think how to improve the readability, so that they can have a better power to get more sponsors.