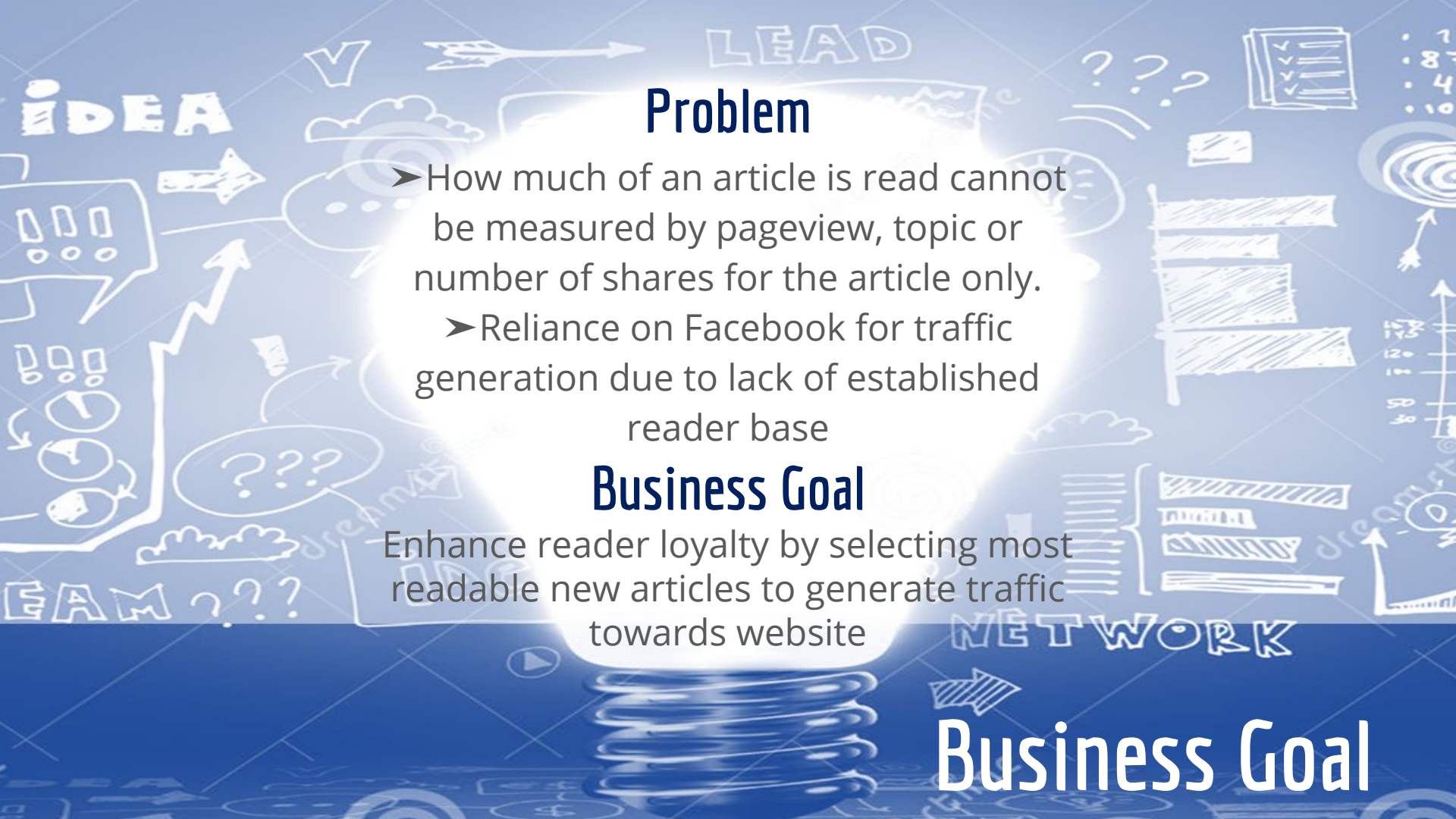
Predicting readability of the next online articles to enhance reader loyalty

Team 4: Uniss Tseng (106077503), Elisa Wang (105078514), Sabrina Wei(102033227), Patrizia Mach (106077429)



Problem

➤How much of an article is read cannot be measured by pageview, topic or number of shares for the article only.

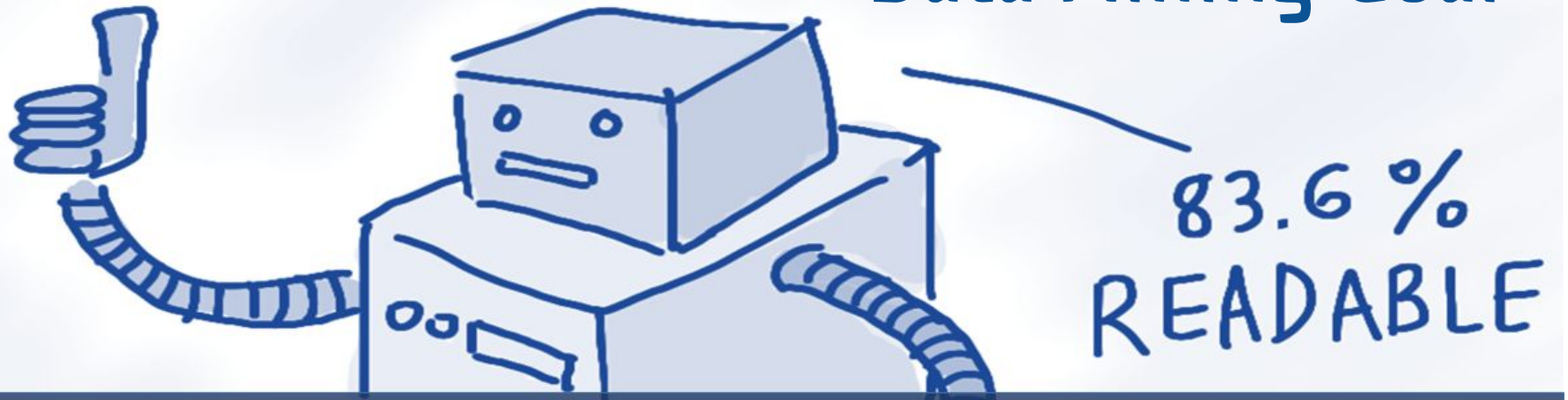
➤Reliance on Facebook for traffic generation due to lack of established reader base

Business Goal

Enhance reader loyalty by selecting most readable new articles to generate traffic towards website

Business Goal

Data Mining Goal



Predict readability

- Reading scroll ratio
- Reading spending time

→ Readability Score

A supervised task.
A forward-looking and predictive task.

Readability Score of each article is compared and ranked to select top most likely to be read the most.

DATA

	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_nur	rich_media	p_gt20_nu	h5_num	qoute_nur	word_num	link_num	is_sponsor	score
2															827632
3															936984
4															870739
5															044909
6															722003
7															686346
8															546698
9															544216
10															0.61373
11															0.65695
12															698723
13															697227
14		Evening													682489
15	5	Day	1	2	21	64334	6	6	21	6	3	2543	9	0	0.170908

What is a row	Article
Input variables	word_num, main_category_23, author_id_red_1, author_id_red_11, author_id_red_3,
Output variable	Readability score (continuous)
Partitioning	70% - Cross validation of 10 folds 30% - Test set

PROCESS

- Reduce Categories
- Create Dummies
- Data partition
- Feature Optimization
- Cross Validation
- Run linear regression on data partitions

XlMiner

■Naive Rules as benchmark

■Linear Regression

■KNN

■Regression Trees

Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
9.0618	0.155243	2.34963E-16

Validation Data Scoring - Summary Repo

Total sum of squared errors	RMS Error	Average Error
5.229717	0.152119	-0.000686471

Test Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
6.140012	0.20232	-0.023489557

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

Total sum of squared errors	RMS Error	Average Error
0.00235	0.0025	-3.91236E-18

	Average error	RMSE
Training	4.33607E-16	0.201724
Validation	-3.639E-05	0.200702
Test	2.21816E-05	0.238865

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

Total sum of squared errors	RMS Error	Average Error
7.43411	0.22262	-0.024242695

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

Total sum of squared errors	RMS Error	Average Error
	0.137722	3.17417E-18

scoring - Summary Report (Using Full-Grown Tree)

RMS Error	Average
0.19329	

Overfitting!

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

Total sum of squared errors	RMS Error	Average Error
7.074036	0.217164	-0.028748897

Methods

Performance Evaluation

RapidMiner

■ Linear Regression ■Naïve Test ■ Regression Trees
■ Ensembles ■Ensemble Opt. Feature ■ Random Forest Test

Ensemble Test
PerformanceVector:
root_mean_squared_error: 0.182 +/- 0.000
absolute_error: 0.129 +/- 0.129
squared_error: 0.033 +/- 0.112
prediction_average: 0.638 +/- 0.224

Ensemble Opt. Feature
PerformanceVector:
root_mean_squared_error: 0.192 +/- 0.000
absolute_error: 0.133 +/- 0.138
squared_error: 0.037 +/- 0.132
prediction_average: 0.638 +/- 0.22

Linear Regression Test
PerformanceVector:
root_mean_squared_error: 0.156 +/- 0.000
absolute_error: 0.121 +/- 0.099
squared_error: 0.024 +/- 0.045

Random Forest Test	Single Tree Test	Naïve Test
PerformanceVector:	PerformanceVector:	PerformanceVector:
root_mean_squared_error: 0.173 +/- 0.000	root_mean_squared_error: 0.177 +/- 0.000	root_mean_squared_error: 0.201 +/- 0.00
absolute_error: 0.127 +/- 0.117	absolute_error: 0.144 +/- 0.103	
squared_error: 0.030 +/- 0.075	squared_error: 0.031 +/- 0.041	

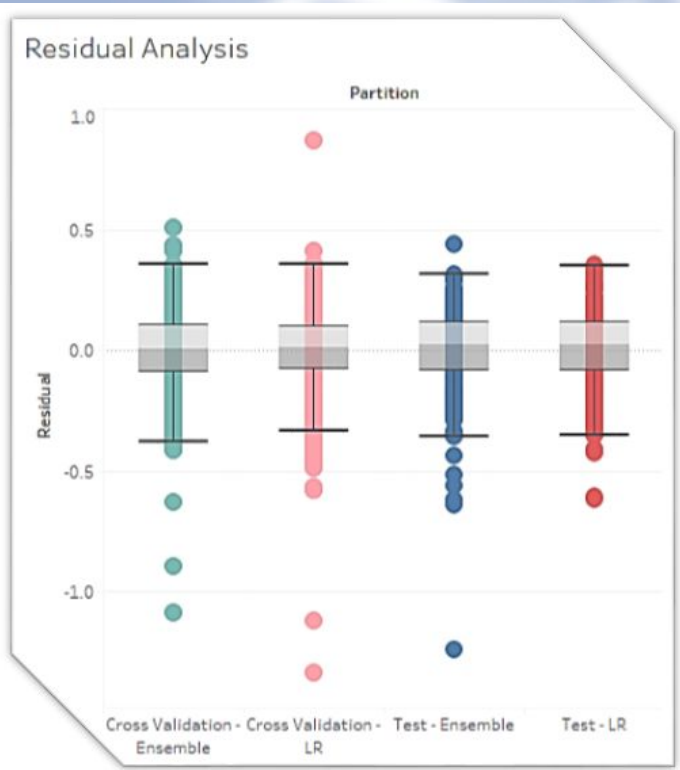
Methods

Performance Evaluation

PROCESS

- Reduce Categories
- Create Dummies
- Data partition
- **Feature Optimization**
- **Cross Validation**
- **Run linear regression on data partitions**

RapidMiner



■Naïve Test ■ Regression Trees
m Forest Test

Linear Regression	Single Tree	Random Forest	Ensemble
0.164	0.172	0.167	0.148
0.168	0.175	0.174	0.163
0.156	0.177	0.173	0.182

Methods

Performance Evaluation

RapidMiner

■Linear Regression

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_re_3 +0.859
Naïve RMSE	0.636099781797528

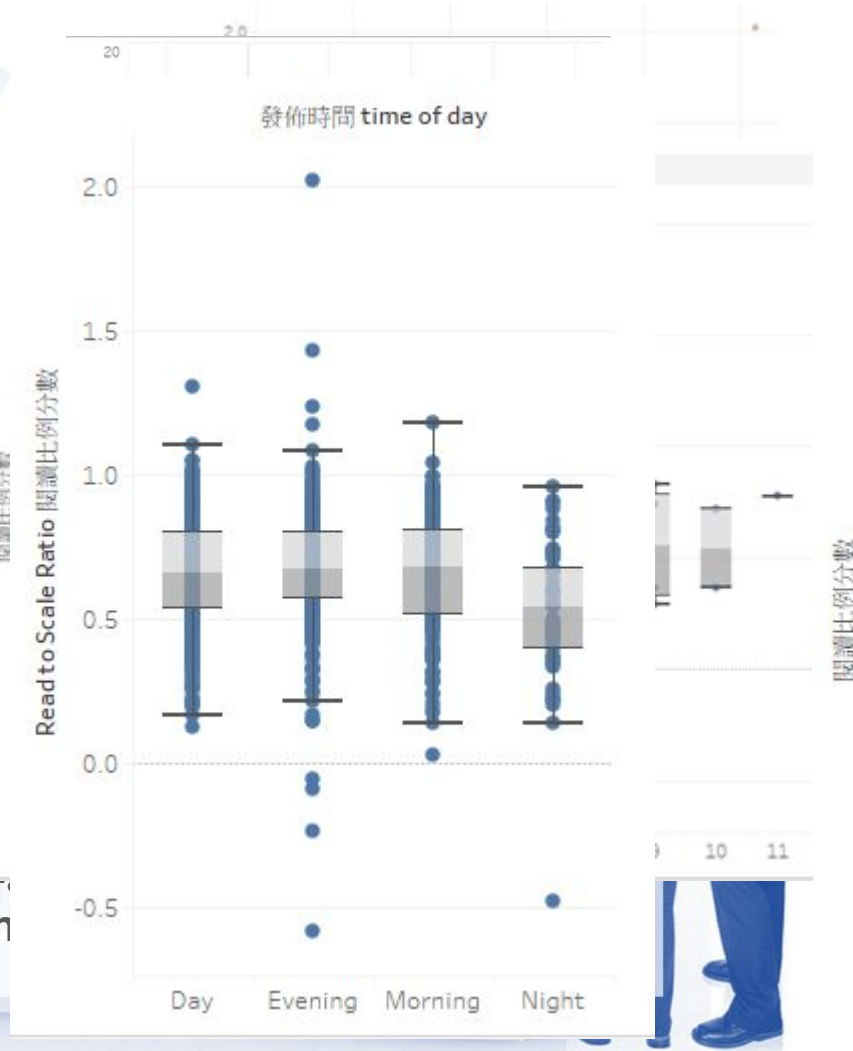
	Score	Predicted
1	1.176068	0.834521
2	0.99851	0.826917
3	0.959334	0.816483
4	0.803143	0.814538
5	0.909099	0.811974
6	0.97985	0.811443
7	0.765636	0.805696
8	0.934385	0.803574
9	1.002382	0.802689
10	0.866768	0.801717
11	1.012772	0.800744
12	0.912426	0.799329

Methods

Performance Evaluation

Recommendation

- **Solution should be reviewed as re**
  current events:
  Use, update and retrain model with
- **Word Count vs. Readability correl**
  Indicates shorter articles better char
- **Photo or Rich Media content not l**
  Review cost vs benefit of informatics
- **Sponsored Articles vs Readability**
  Indicate lower readability
- **Night Articles vs Readability**
  Indicate Less reading outside office
- **Review batch of new articles for h**
  strategically place on website for h
  sponsored content



Implementation & Production Considerations

Implementation

- **Data Scientist:**
  Use, update and retrain model.
- **Marketing & Editors:**
  Use interpreted results to improve articles.
  Evaluate article effectiveness.
- **Business development:**
  Use prediction for long-term decision making

Production

- **Solution does not need to be real-time**
- **Model should be re-investigated regularly**
  Changes in trends
  Changes to business model