Predicting Visits of Digital Newspaper Articles

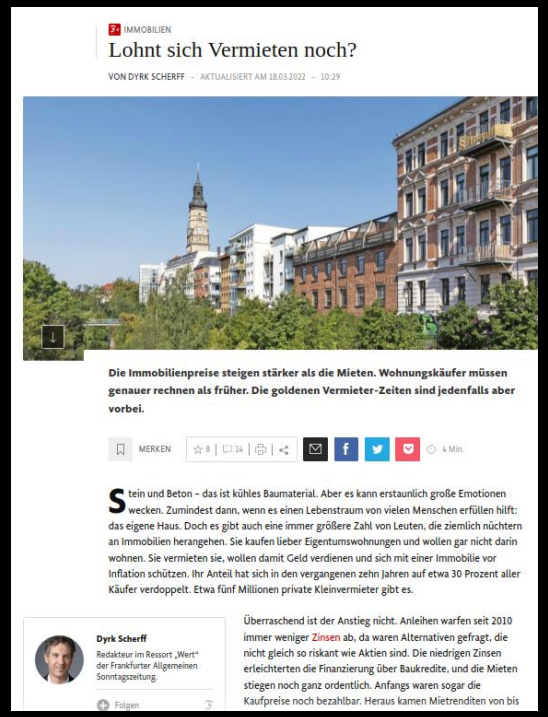
Analysis for a German publisher (Frankfurter Allgemeine Zeitung)

METIS - Regression

Fabian Paul, 19/04/2022

Introduction

Business Problem



IMMOBILIEN

Lohnt sich Vermieten noch?

VON DYRK SCHERFF - AKTUALISIERT AM 18.03.2022 - 10:29

Die Immobilienpreise steigen stärker als die Mieten. Wohnungskäufer müssen genauer rechnen als früher. Die goldenen Vermieter-Zeiten sind jedenfalls aber vorbei.

MERKEN 8 | 14 | 4 Min.

Stein und Beton – das ist kühles Baumaterial. Aber es kann erstaunlich große Emotionen wecken. Zumindest dann, wenn es einen Lebenstraum von vielen Menschen erfüllen hilft: das eigene Haus. Doch es gibt auch eine immer größere Zahl von Leuten, die ziemlich nüchtern an Immobilien herangehen. Sie kaufen lieber Eigentumswohnungen und wollen gar nicht darin wohnen. Sie vermieten sie, wollen damit Geld verdienen und sich mit einer Immobilie vor Inflation schützen. Ihr Anteil hat sich in den vergangenen zehn Jahren auf etwa 30 Prozent aller Käufer verdoppelt. Etwa fünf Millionen private Kleinvermieter gibt es.

Dyrk Scherff
Redakteur im Ressort „Wert" der Frankfurter Allgemeinen Sonntagszeitung.
Folgen

Überraschend ist der Anstieg nicht. Anleihen warfen seit 2010 immer weniger Zinsen ab, da waren Alternativen gefragt, die nicht gleich so riskant wie Aktien sind. Die niedrigen Zinsen erleichterten die Finanzierung über Baukredite, und die Mieten stiegen noch ganz ordentlich. Anfangs waren sogar die Kaufpreise noch bezahlbar. Heraus kamen Mietrenditen von bis

- Newspaper firms: Audience reach important for ad and subscription revenue
- Release >100 articles per day
- Insights on article characteristics crucial for:
  - … journalists to tailor texts to audience needs
  - … editors to decide on pay vs. free articles and release times

Objective

- Which article features are <u>significantly correlated</u> to article visits?
- Which publishing time features are <u>significantly correlated</u> to article visits?
- Is it possible to <u>predict article audience</u> before an article is even published?

Methodology

Published articles

Articles with URLs

Department, author, body-length, publishing time, social shares, ID

Overall daily traffic

Visits per day of faz.net

Article visits (target)

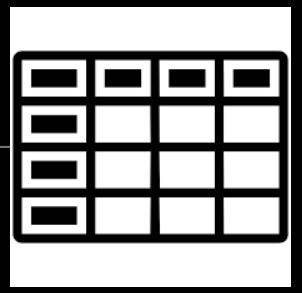
ID, visits

Source: faz.net

Feature Engineering

- *Normalize article visits for overall website daily visits*
- *Features for texts on specific topics*
- *Bins for reflecting time series and article sources*
- *Converting categorical features to dummy variables*
- *Log-transform count-data*

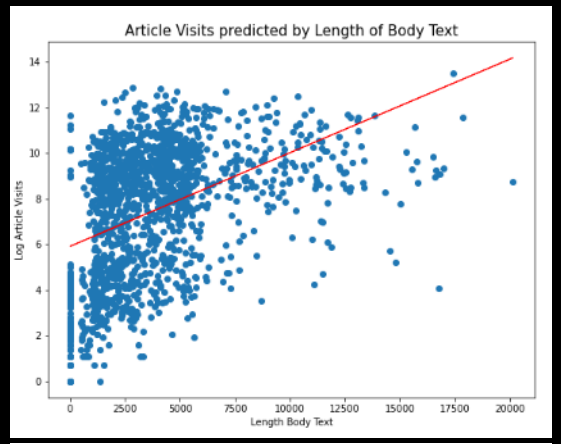
	visits_norm
visits_norm	1.000000
visits	0.984232
visits_website	0.103157
visits_norm_website	0.116332
visits_website_prevday	0.059700
visits_norm_website_prevday	0.084585
visits_website_roll7day	0.075376
visits_norm_website_roll7day	0.083561
visits_ressort	0.242568
visits_norm_ressort	0.246736
visits_ressort_prevday	0.159183
visits_norm_ressort_prevday	0.171691
visits_ressort_roll7day	0.168489
visits_norm_ressort_roll7day	0.178445
article_pubressort_18h	-0.009948
partofweek_weekend	0.045827
partofweek_monday	0.011616
partofweek_friday	-0.021926
timeofday_time_06_11	0.082548
timeofday_time_11_16	0.032796
timeofday_time_16_21	-0.114291
ressort_finanzen	0.015828
ressort_politik	0.122655
ressort_wirtschaft	-0.023170
ressort_feuilleton	0.001273
ressort_stil	0.022650
author_altenbockum	0.043195
author_looman	0.023500
author_busse	0.057118
author_knop	-0.012222
author_mohr	0.002748
author_muller	0.006470
author_siedenbiedel	0.032770
source_online	0.119331
source_weekday_print	-0.084344
source_sunday_print	0.031798
source_news_agency	0.019728
length_teaser	0.184601
length_body	0.162343
ukraine	0.185398
olympics	-0.006698
new_german_gvt	0.033953
relationship	0.008339
financial	0.014923
opinions	0.298393
sharings	0.539009
im_bright	-0.012831
im_color_red	0.042540
premium	-0.073679
author_link_personal_site	0.139586

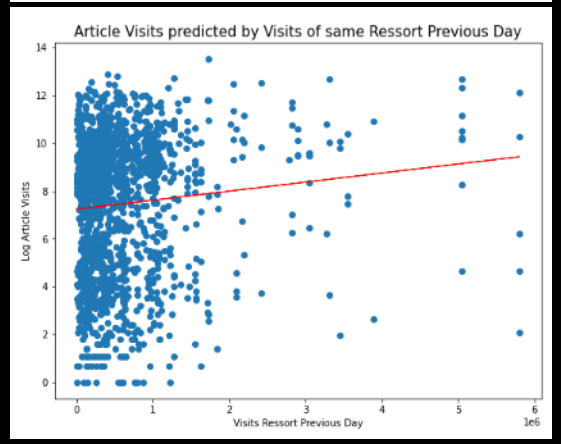


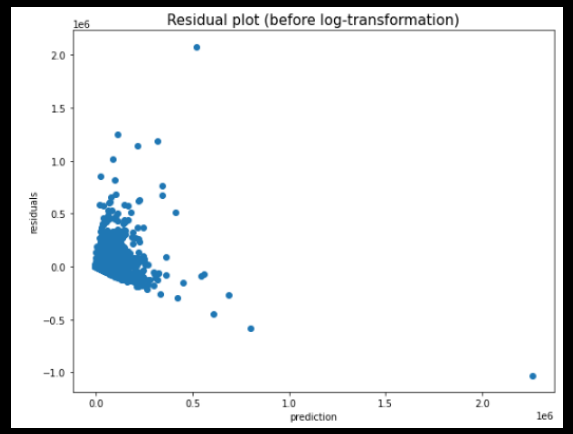
Results

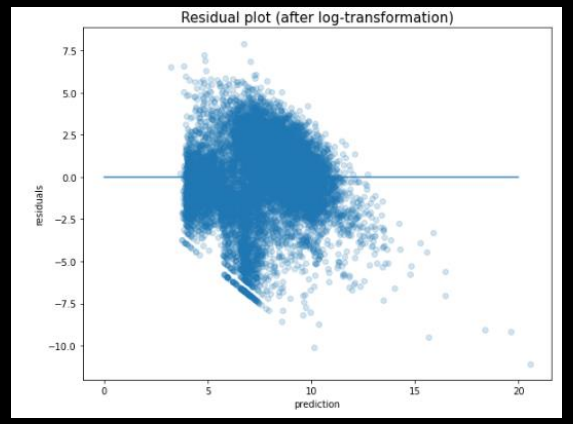
Regression Assumptions

1) *Regression is linear in parameters* ✓
2) *No perfect multicollinearity* ✓
3) *Residuals are normally distributed* ✓
4) *Errors uncorrelated across observations* ✓
5) *Equal variance of errors* ✓









	coef	P>|t|	[0.
Intercept	6.5316	0.000	6.
ukraine	0.4471	0.000	0.
ressort_finanzen	0.6353	0.000	0.
ressort_stil	0.8544	0.000	0.
source_weekday_print	-0.2582	0.000	-0.
partofweek_weekend	0.2862	0.000	0.
partofweek_monday	0.1415	0.005	0.
timeofday_time_16_21	-0.8325	0.000	-0.
visits_norm_ressort_prevday	2.135e-07	0.000	1.6e
author_altenbockum	1.6855	0.000	1.
author_knop	-1.4947	0.000	-2.
length_body	0.0004	0.000	0.
im_color_red	0.1400	0.000	0.
author_link_personal_site	1.5284	0.000	1.
premium	-1.9397	0.000	-2.

Durbin-Watson:	1.879
Jarque-Bera (JB):	1496.196
Prob(JB):	0.00
Cond. No.	1.79e+07

	features	vif
0	ukraine	1.374584
1	ressort_finanzen	1.067039
2	ressort_stil	1.025103
3	source_weekday_print	3.391542
4	partofweek_weekend	1.227778
5	partofweek_monday	1.167582
6	timeofday_time_16_21	2.169754
7	length_body	3.076015
8	visits_norm_ressort_prevday	1.826100
9	author_altenbockum	1.014736
10	author_knop	1.009439
11	im_color_red	1.489527
12	author_link_personal_site	2.185529
13	premium	2.964761

Results

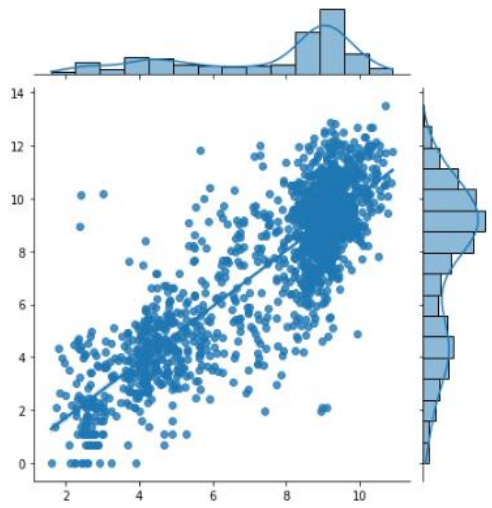
Model Selection – Random Forest Regressor

Focus prediction – without social shares and comments

	coef
Intercept	6.5316
ukraine	0.4471
ressort_finanzen	0.6353
ressort_stil	0.8544
source_weekday_print	-0.2582
partofweek_weekend	0.2862
partofweek_monday	0.1415
timeofday_time_16_21	-0.8325
visits_norm_ressort_prevday	2.135e-07
author_altenbockum	1.6855
author_knop	-1.4947
length_body	0.0004
im_color_red	0.1400
author_link_personal_site	1.5284
premium	-1.9397

```
Linear Regression Mean Average Error (MAE): 1.6928
Linear Regression Root Mean Squared Error (RMSE): 2.1989
Linear Regression R2 Score (R2): 0.4380

Lasso MAE: 2.8285
Lasso RMSE: 2.1995
Lasso R2: 0.4377

Ridge MAE: 1.6931
Ridge RMSE: 2.1990
Ridge R2: 0.4379

ElasticNet MAE: 2.8299
ElasticNet RMSE: 2.1994
ElasticNet R2: 0.4377

Polynomial MAE: 1.5258
Polynomial RMSE: 1.9587
Polynomial R2: 0.5540

Random Forest Regressor MAE: 2.9756
Random Forest Regressor RMSE: 1.5529
Random Forest Regressor R2: 0.7197

Gradient Boosting MAE: 3.0002
Gradient Boosting RMSE: 1.5584
Gradient Boosting R2: 0.7177
```

Random Forest Prediction vs. Test Values



Focus correlation – with social shares and comments

```
Random Forest Regressor MAE: 3.1121
Random Forest Regressor RMSE: 1.1650
Random Forest Regressor R2: 0.8422
```

Conclusions

Results

- Article features:
  - ✚ Link to authors personal website, text length, author Altenbockum, departments Stil or Finanzen, Ukraine war, text length, images with red color, comments, social shares
  - ➖ Paid-articles, author Knop, source weekday print
- Publishing time features:
  - ✚ Publishing day Monday or weekend
  - ➖ Publishing time 4pm-9pm
- It is possible to predict article visits before an article is published with a MAE of 19.6 visits using a random forest regression

Limitations

- No information on position individual articles were placed on website
- No information on total amount of time articles were placed on website
- Time series trends may not be properly captured
- Count data (Poisson distributions) may require specialized models beyond pure log-transformations of features

Future Work

1. → Include information on position of articles on website
2. → Include information on amount of time articles were placed on website
3. → Apply ARIMA model to a wider time window (at least 1y) to capture trends
4. → Apply specific Poisson Regression Models

Thank you for your attention!