Project Write-up – Predicting Visits of Digital Newspaper Articles

Abstract

The goal of this project was to use regression models to predict visits of digital newspaper articles in to support editors in choosing which articles to publish. Additionally, the model may provide evidence for editors which article characteristics may be important for attracting high numbers of readers. My analysis was based on all articles of German publisher Frankfurter Allgemeine Zeitung published online 01/01/22 - 01/04/22 and their corresponding visit figures. In feature engineering, I leveraged meta information of articles itself and their images to create a set of categorical and numerical features. In terms of modeling, I used linear regression, lasso, ridge, polynomial regression, random forest regressor and gradient boosting regressor.

Design

The project topic constitutes a central question of current (data-driven) journalism. Feature data was scraped from a publically available newsticker website listing all historically published articles and from each articles' corresponding URL. The paywall was circumvented using a digital subscription. Target data was provided from within the company. Target data was further enriched collecting overall daily traffic figures published at industry association website. Extracting insights on article properties (such as publication time, length, author, topics or images) contributing to high reader numbers via machine learning models may support journalists and editors to reach larger audiences. It may help to evaluate article popularity ex-ante their publication online and thus assist in the choice of articles to be provided for-free or paid.

Data

The dataset contains 15,477 articles with 85 features for each, 47 of which are categorical. A few feature highlights include "topic on ukraine-war", "article length", "source weekday print", "publishing-time of day", "day of week", "previous day visits", "author" or "image color". Nearly 40% of individual features were grouped into broader categories and investigated more closely with 14 of them forming baseline models.

Algorithms

Feature Engineering

1. Normalizing article visits for overall daily visits of the website, indexed at 01/01/2022
2. Creating features with teaser containing words of specific topics (such as the ukraine war)
3. Creating bins reflecting time series features (such as time of day or day of week)
4. Creating bins reflecting article source (such as print weekday or news agency)
5. Converting categorical features to binary dummy variables (such as authors or department)
6. Log-transform count-data, incl. article visits, social shares, comments and visits of previous day

Models

Linear regression, lasso, ridge, elastic net, polynomial regression, random forest regressor and gradient boosting regressor were used before settling on random forest regressor as the model with strongest cross-validation performance. Throughout, linear regression was used to check on p-values of each feature to ensure significance.

Model Evaluation and Selection

The entire training dataset of 15,477 records was split into 90/10 train vs. test, and all scores reported below were calculated with 5-fold cross validation on the training portion only. Predictions on 10% test were limited to the very end, so this split was only used and scores seen just once. Models were evaluated based on their generalization performance using R^2, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The random forest regressor had a R^2 of 0.72 on the test sample versus a mean R^2 of 0.69 on the 5-fold CV sample.

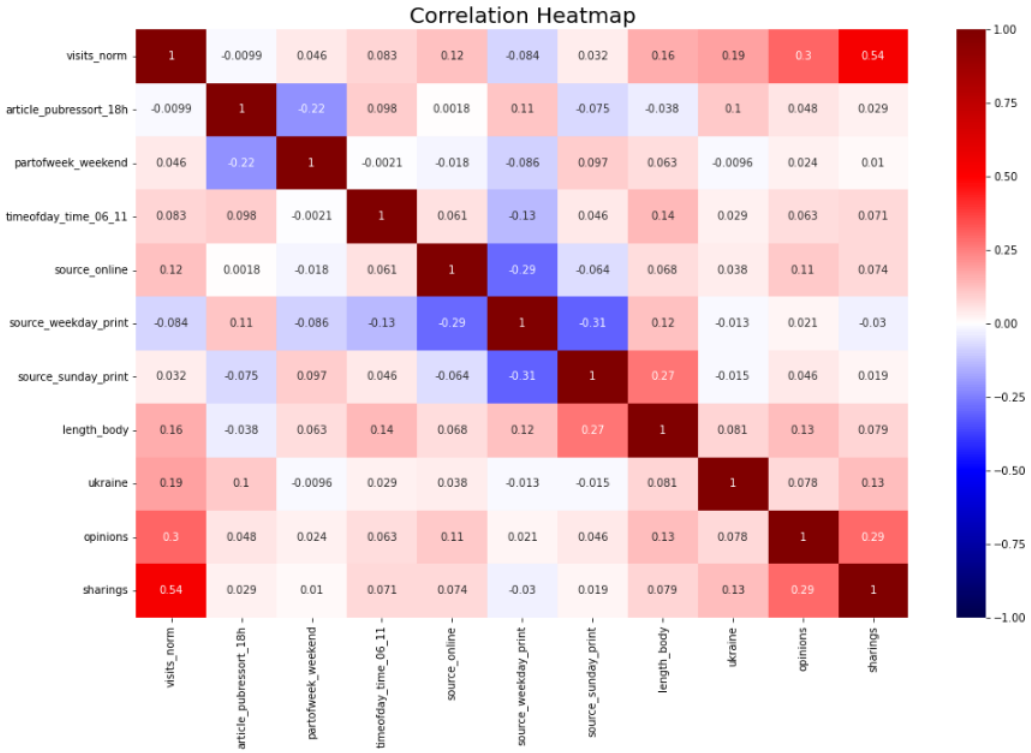
Tools

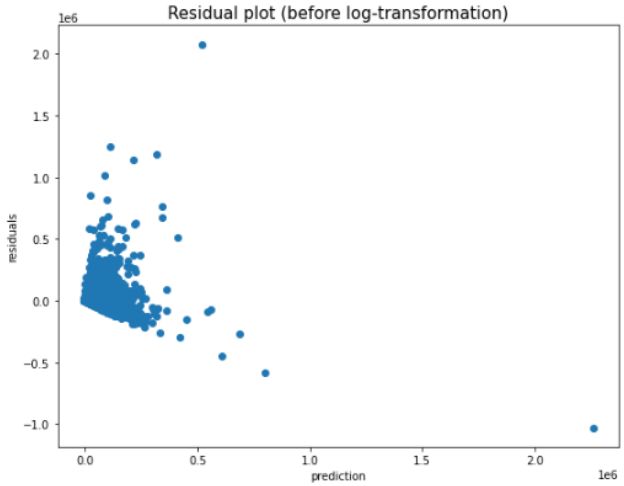
- BeautifulSoup for webscraping and Selenium for automatic download of files
- NumPy and pandas for data manipulation
- Pillow for image feature generation
- Statsmodels and scikit-learn for modeling
- Matplotlib and Seaborn for plotting

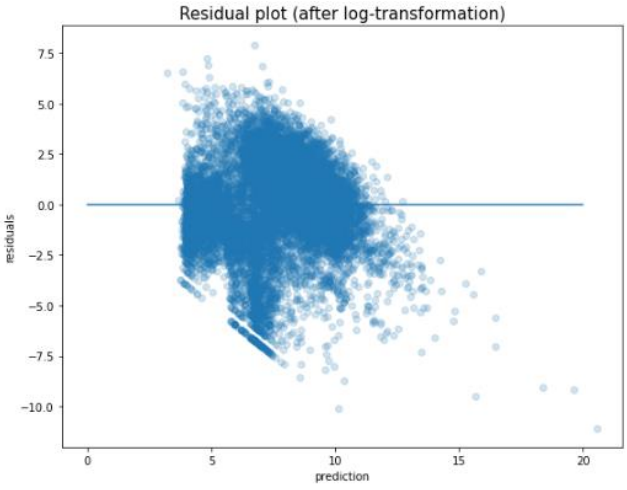
Communication

In addition to the slides and visuals presented, project outputs will be embedded on my personal github.

	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







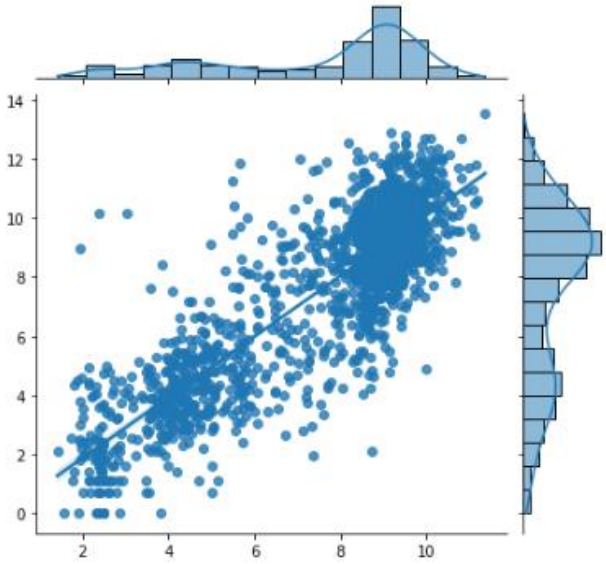
Linear Regression (target=log_visits, excl. social shares and comments)

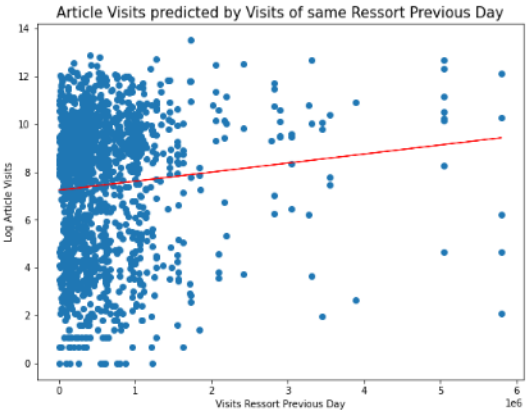
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.5316	0.046	143.483	0.000	6.442	6.621
ukraine	0.4471	0.048	9.390	0.000	0.354	0.540
ressort_finanzen	0.6353	0.082	7.762	0.000	0.475	0.796
ressort_stil	0.8544	0.158	5.399	0.000	0.544	1.165
source_weekday_print	-0.2582	0.043	-5.962	0.000	-0.343	-0.173
partofweek_weekend	0.2862	0.046	6.231	0.000	0.196	0.376
partofweek_monday	0.1415	0.050	2.813	0.005	0.043	0.240
timeofday_time_16_21	-0.8325	0.039	-21.212	0.000	-0.909	-0.756
visits_norm_ressort_prevday	2.135e-07	2.74e-08	7.793	0.000	1.6e-07	2.67e-07
author_altenbockum	1.6855	0.342	4.924	0.000	1.015	2.356
author_knop	-1.4947	0.319	-4.682	0.000	-2.120	-0.869
length_body	0.0004	6.92e-06	56.526	0.000	0.000	0.000
im_color_red	0.1400	0.037	3.810	0.000	0.068	0.212
author_link_personal_site	1.5284	0.039	39.282	0.000	1.452	1.605
premium	-1.9397	0.044	-44.391	0.000	-2.025	-1.854

Omnibus:	1112.441	Durbin-Watson:	1.879
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1496.196
Skew:	-0.632	Prob(JB):	0.00
Kurtosis:	3.849	Cond. No.	1.79e+07

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.79e+07. This might indicate that there are
strong multicollinearity or other numerical problems.

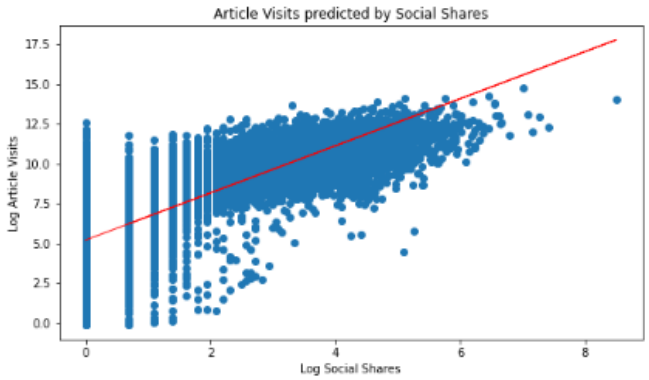
Random Forest – Prediction vs. Test Values

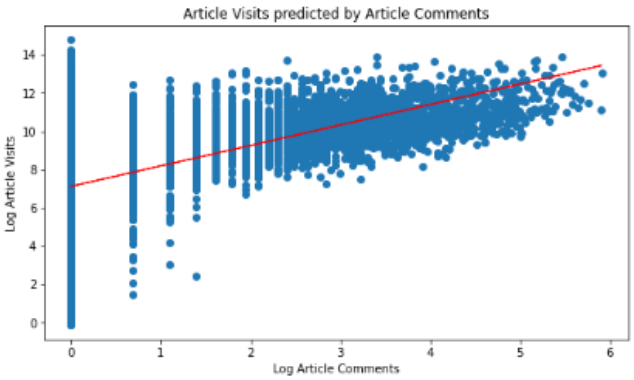




Linear Regression (target=log_visits, excl. social shares and comments)

```
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
```





Linear Regression (target=log_visits, incl. social shares and comments)

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