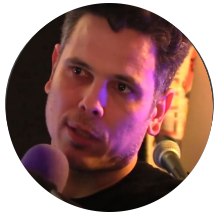
DATA-DRIVEN SEARCH FOR TRAFFIC DRIVERS
WITH DATA PROVIDED BY

EFAHRER.com

We are an interdisciplinary team with diverse background in business, science, music and IT

Renzo Torrecuso



Neuroscience-Engineer-Violinist

Thomas Brandstätter

Analytics Engineer, Data Products

Ekaterina Burakova

Physical chemist, Dr. rer. nat.

Clara Thümecke

Business Development

Preethi Karumathil

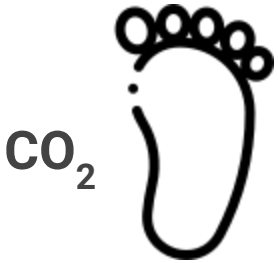
IT Engineer

We’ll show how to create impactful content

data driven and backed by machine learning

1. Client & Problem description
2. Data & Target
3. Insights & Hypothesis
4. Modelling
    a. Baseline
    b. Feature Engineering
    c. Predicting Model
5. Recommendations
6. Future work

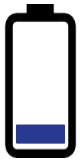
We help EFAHRER in empowering their users to contribute to carbon reduction



EFAHRER.com is a media portal which strives to influence users to take actions that support CO_2 reduction.

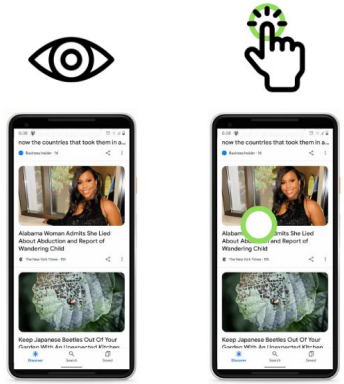
We want to **provide valuable insights** for the editorial team

We want to perform a **prognosis of article success**



For that we analyzed one of EFAHRER's biggest traffic sources for news articles and enriched the data

Editorial data
6.899 unique articles



Feed Impression

Billions

Feed Click

Millions

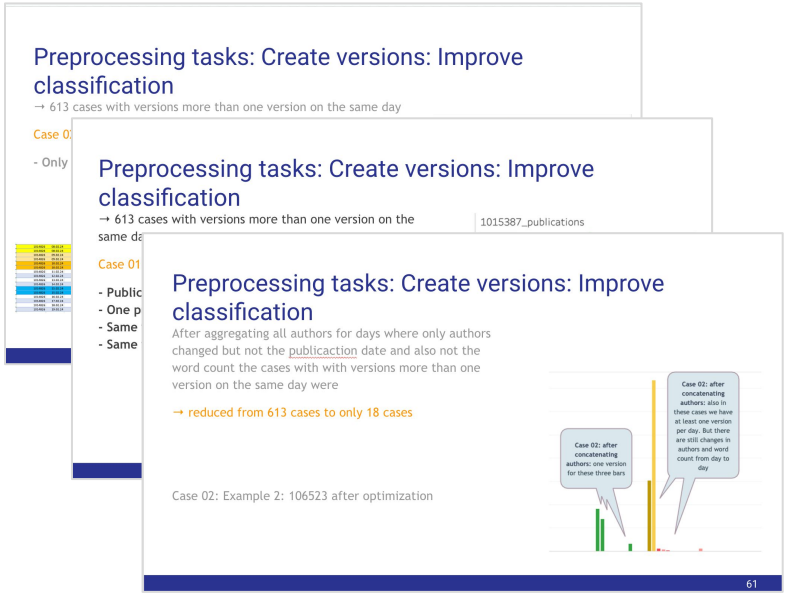
Data quality and completeness of raw data led to extensive preprocessing and analysis for modelling

Challenges

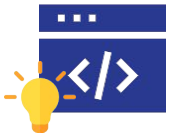
- Missing values
- Lack of article versions
- 3 different aggregation levels for selected metrics

Strategies:

- Delete
- Impute
- Scrape



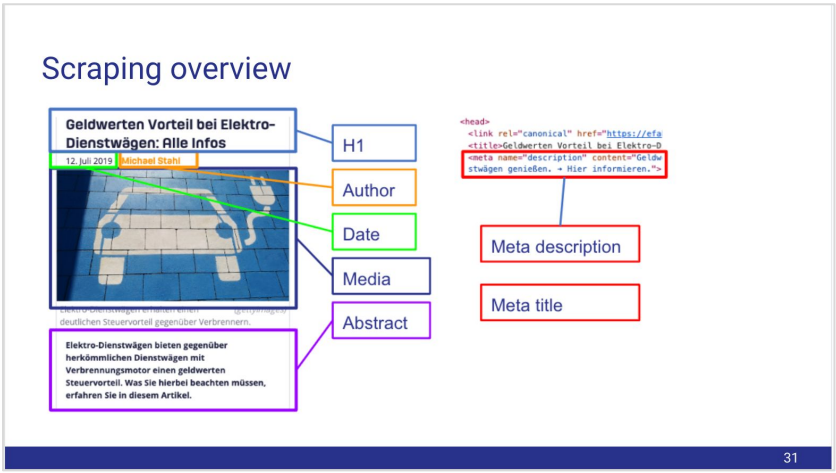
Scraping of 6 000 articles increased the data quality and added new features



Respecting ethics in web scraping we managed to add

+ 6 visible features

+ 5 meta features (invisible)

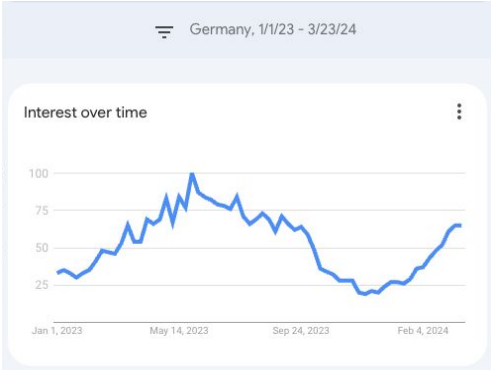


By importing related search terms for 17 product categories we added a trends score to our data

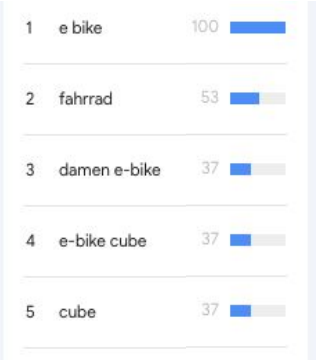
Using a NLP classifier we matched each article with a related search label and trend score

+ 2 features

Search trend for “E-Bike”



Related queries



We identified the following relevant features

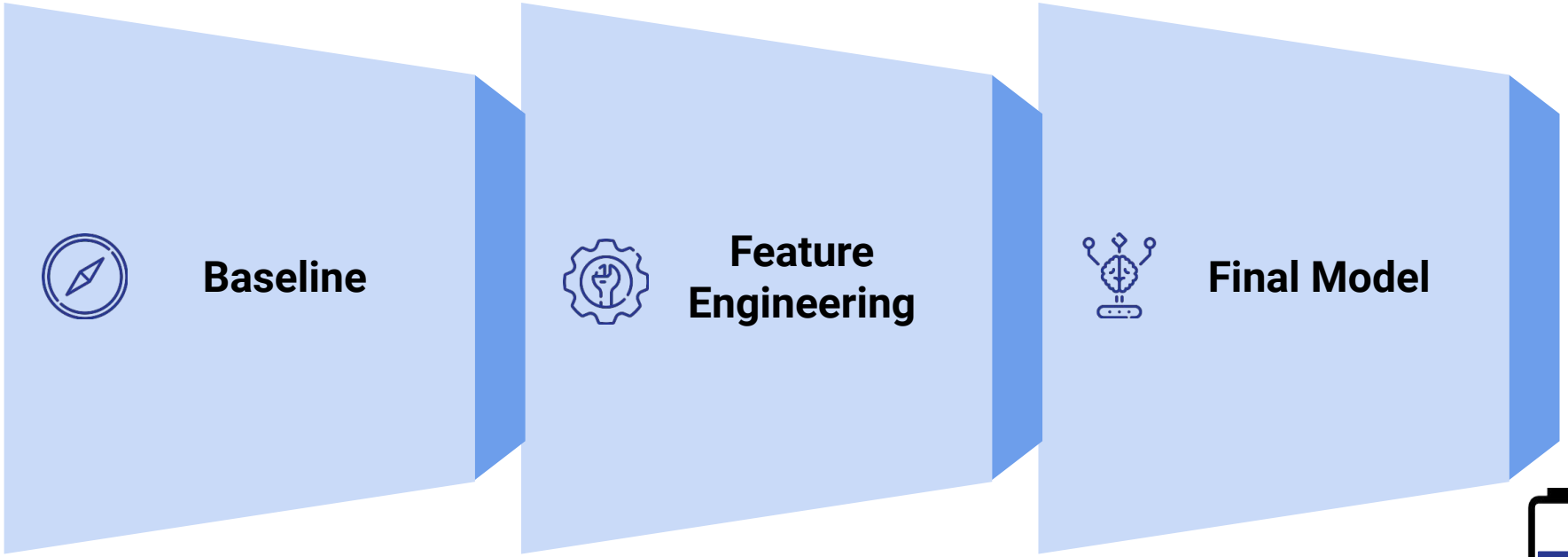
- Article genre and topic
- Type of the first media: video or image
- Word count and lengths of the metadata

Live demo

Features in the positive feedback loop with the target were **ignored** or **normalized**:

- Number of likes, dislikes, video views
- Number of URLs → *URL update frequency*

We verified our hypotheses and created prediction tool



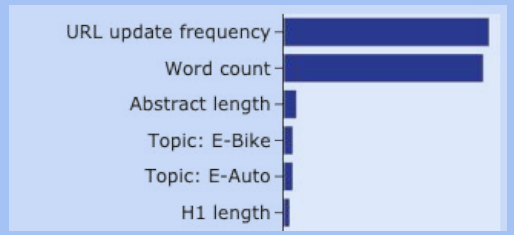
With the baseline model we created an advanced starting point for our modeling

Baseline

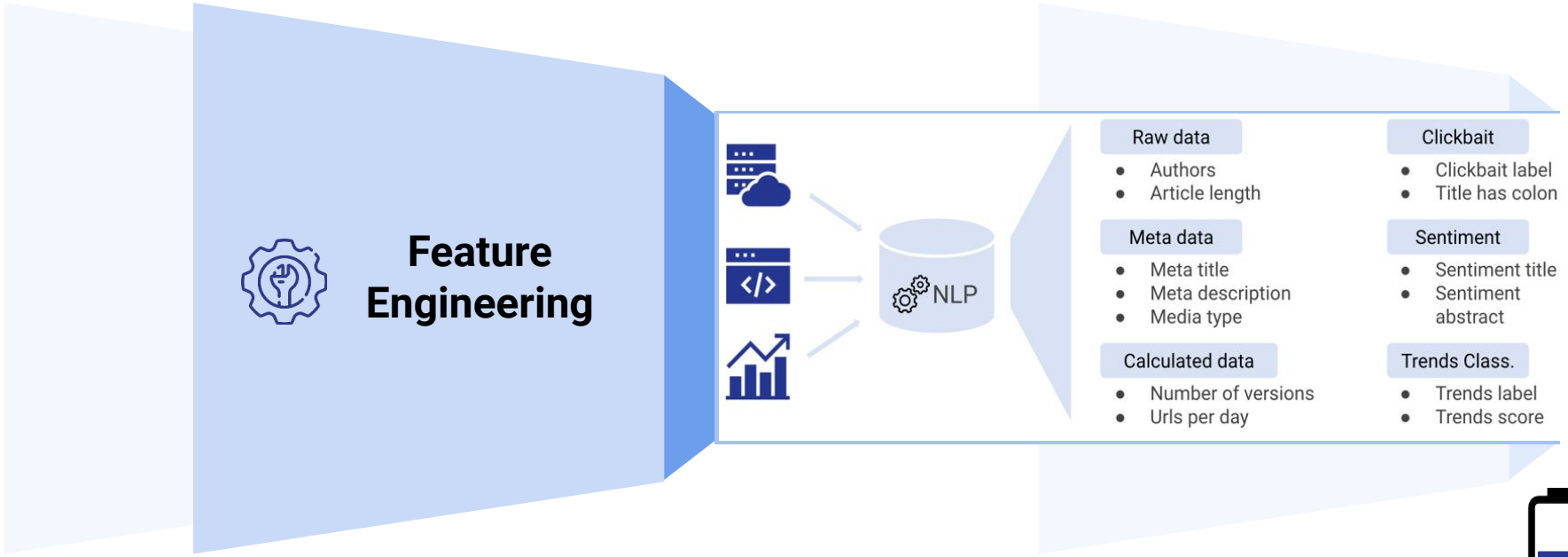
Performance (R^2 = 0.35)
Input:

- Text fields size
- Product classification
- URL update frequency

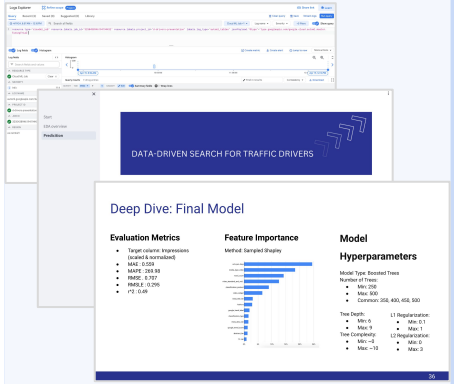
Feature importance



We engineered additional features based on the existing data



Our final model is a stable starting point for predicting article performance



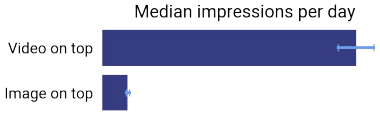
Final Model

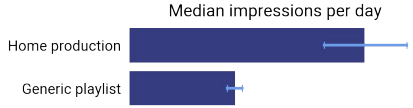
- Modeling with AutoML Tables by Google Vertex AI
- Best performance (R^2 = 0.49) for simpler model without full text features

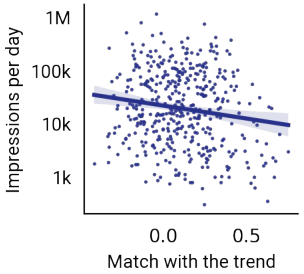
Live demo

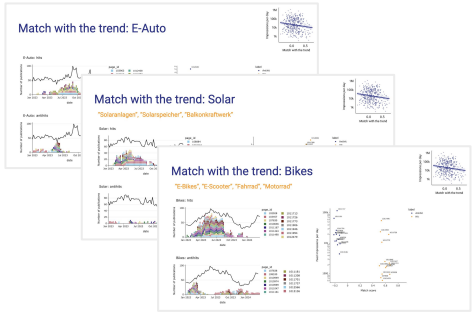
Updating of articles enhances outcomes, the media plays relevant role

1. Change in URL has **tangible** impact on impressions
2. The algorithm prefers articles with videos over images as the first media on page
3. Video production pays off!
4. Optimize publication timing alongside trends
5. Algorithm does not punish clickbait behavior









Further improvements promise a reliable prediction of page impressions

- Try out different semantic segmentation and model each segment individually (e.g. News)
- Dive deeper into the video and image content and formats
- Refine the trend-related features (e.g. different keyword & time matching, trend sources)
- Improve the evaluation of “clickbaitness”
- Fine tune sentiment analysis
- The full article history would provide new valuable features

Thank you for your attention

Special thanks to:

neue fische coach team
who made this possible

- Nico Steffen
- Aljosha Wilhelm
- Lina Willing
- Jin-Ho Lee
  … and others

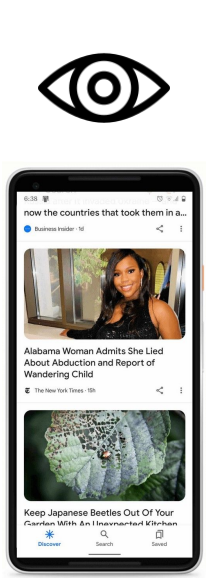
efahrer.com team who
kindly supported this project

- Markus Höllmüller
- Analytics team

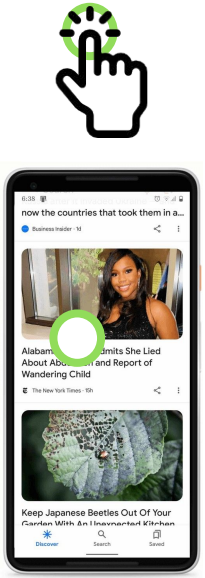
INTERNAL / Backup

Notes and the knowledge base for the team

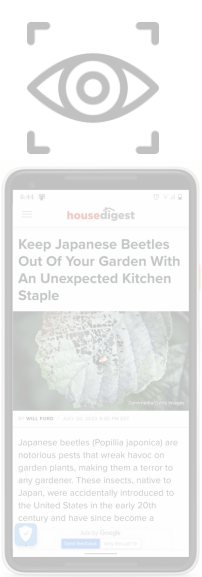
Performance metrics & target features



Feed Impression

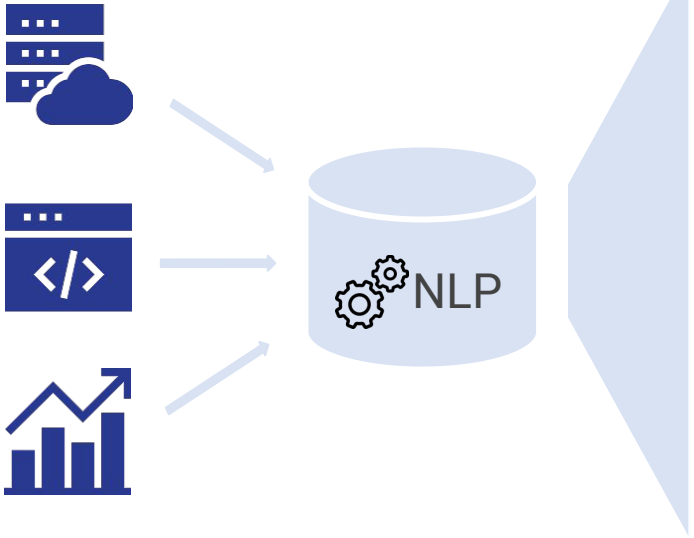


Feed Click



Page Impression

Feature engineering overview



Raw data
- Authors
- Article length

Meta data
- Meta title
- Meta description
- Media type

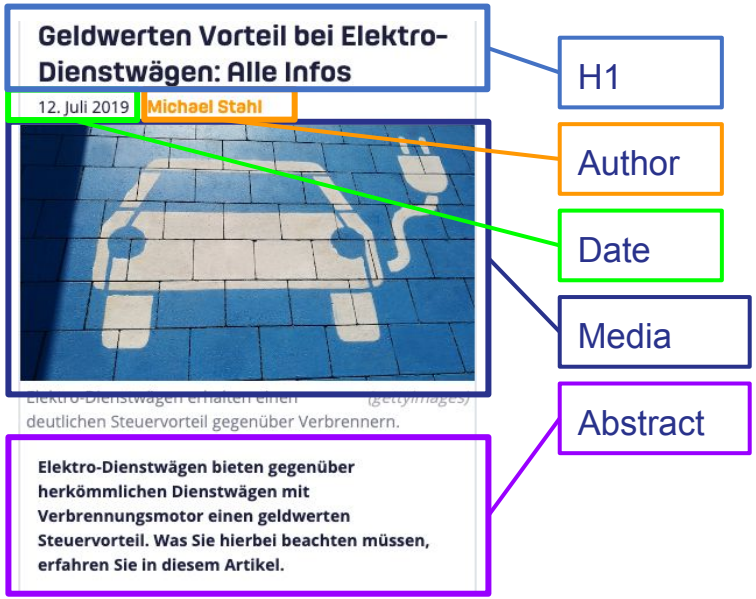
Calculated data
- Number of versions
- Urls per day

Clickbait
- Clickbait label
- Title has colon

Sentiment
- Sentiment title
- Sentiment abstract

Trends Class.
- Trends label
- Trends score

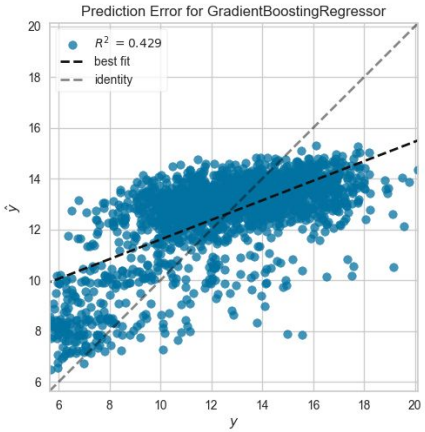
Scraping overview

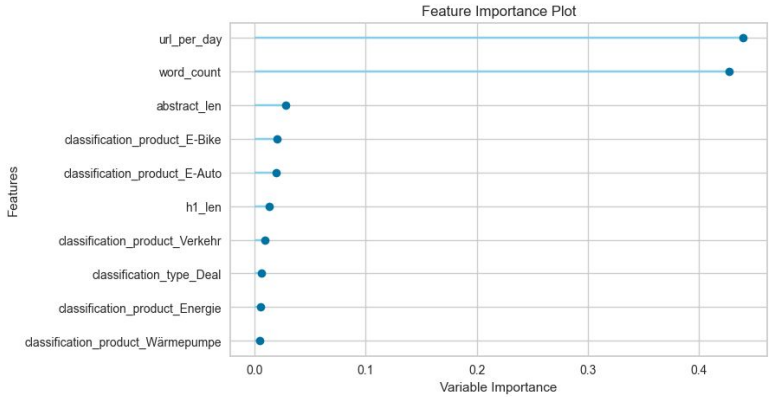


Deep dive: the baseline model

Evaluation Metrics

- Target column: Impressions (scaled & normalized)
- MAE : 1.8132
- MAPE : 0.1553
- RMSE .2.2605
- RMSLE : 0.743
- r^2 : 0.3513



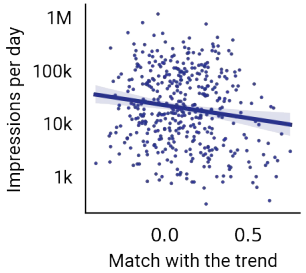


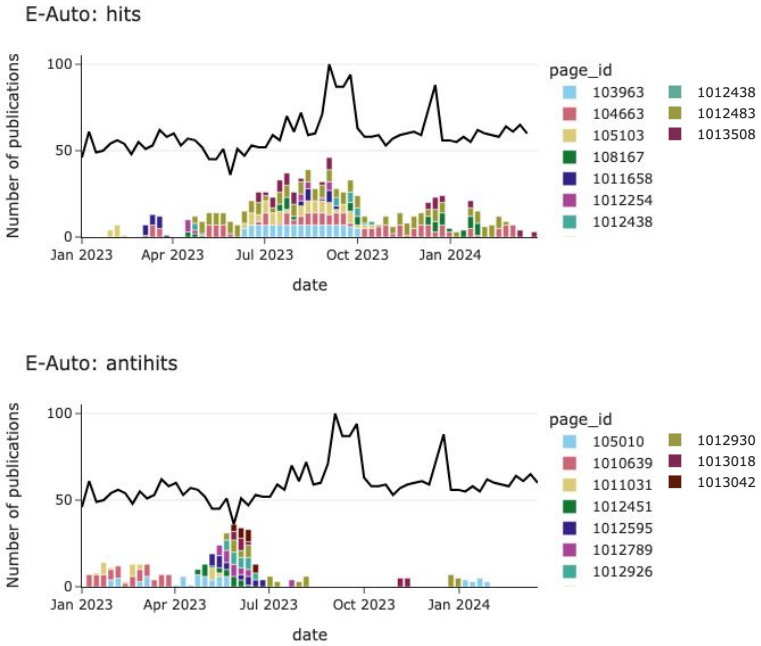
Features used

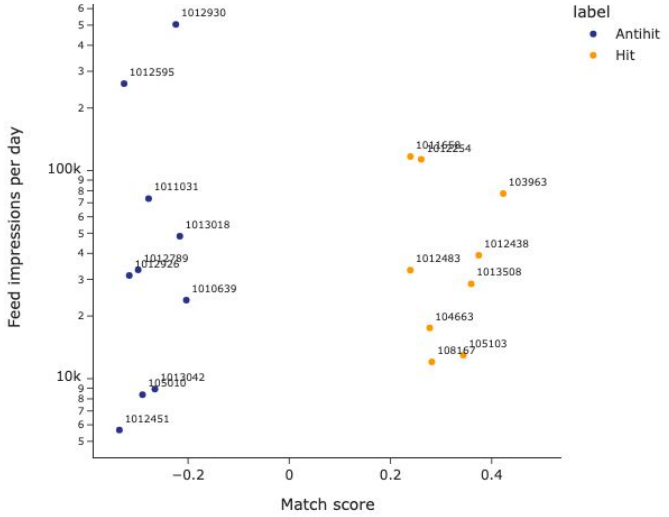
- url_per_day, word_count, abstract_len, h1_len
- classification_product, classification type

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
gbr	Gradient Boosting Regressor	1.8132	5.1203	2.2605	0.3513	0.1743	0.1553
lightgbm	Light Gradient Boosting Machine	1.8203	5.2318	2.2841	0.3366	0.1757	0.1554
rf	Random Forest Regressor	1.8501	5.4482	2.3322	0.3087	0.1791	0.1574
ada	AdaBoost Regressor	1.9852	5.9036	2.4288	0.2520	0.1899	0.1734

Match with the trend: E-Auto

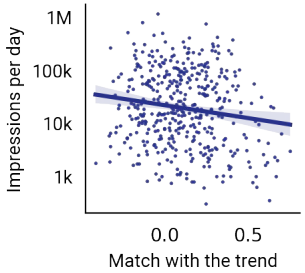


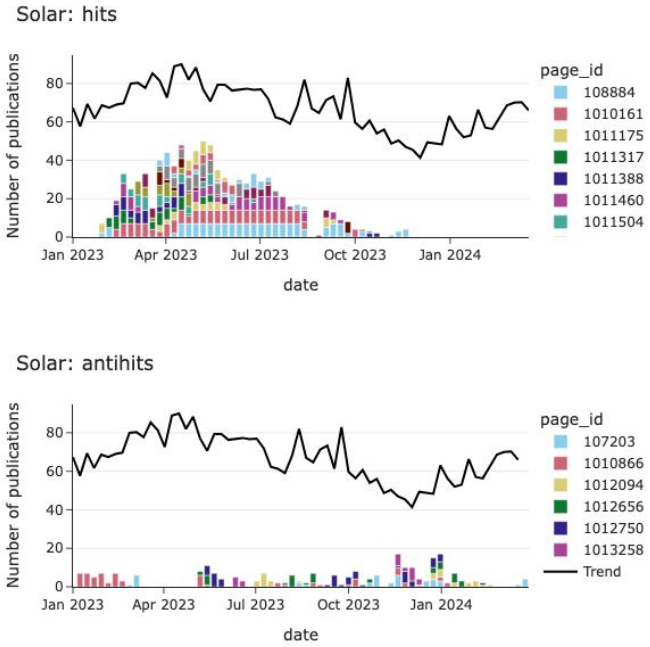


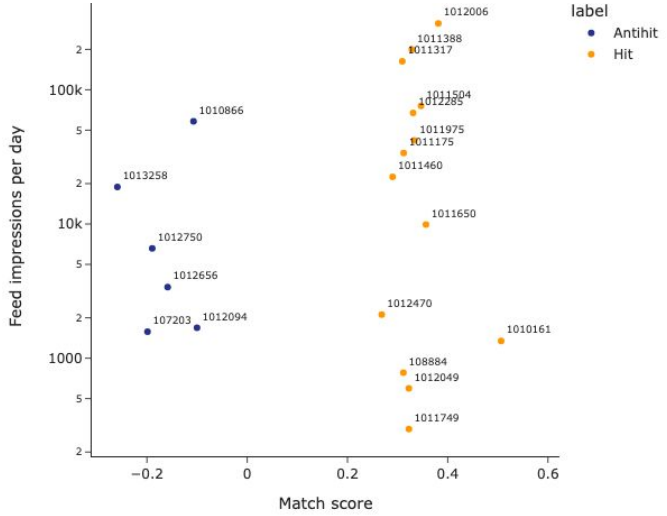


Match with the trend: Solar

“Solaranlagen”, “Solarspeicher”, “Balkonkraftwerk”

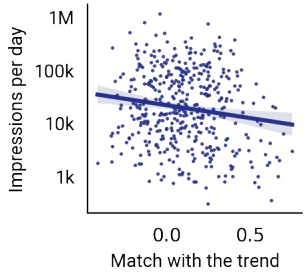


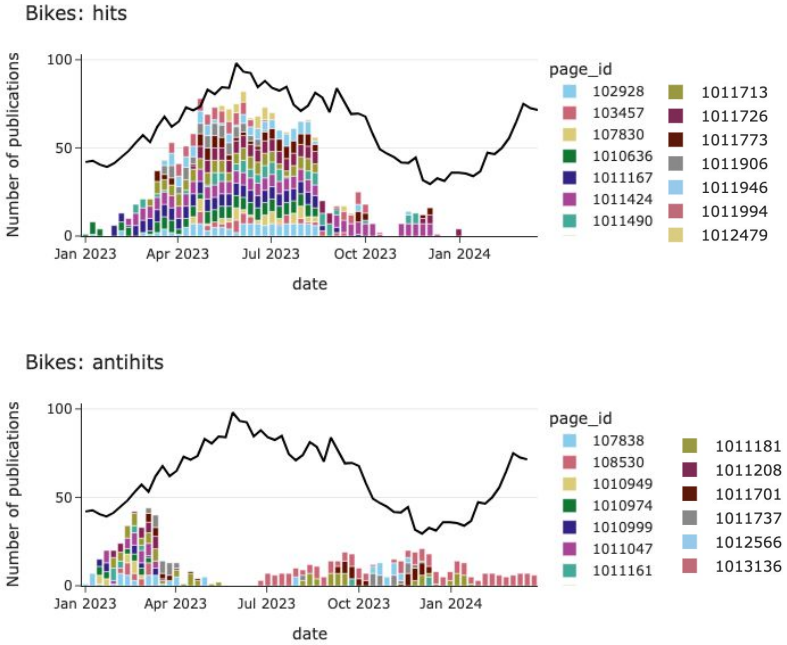


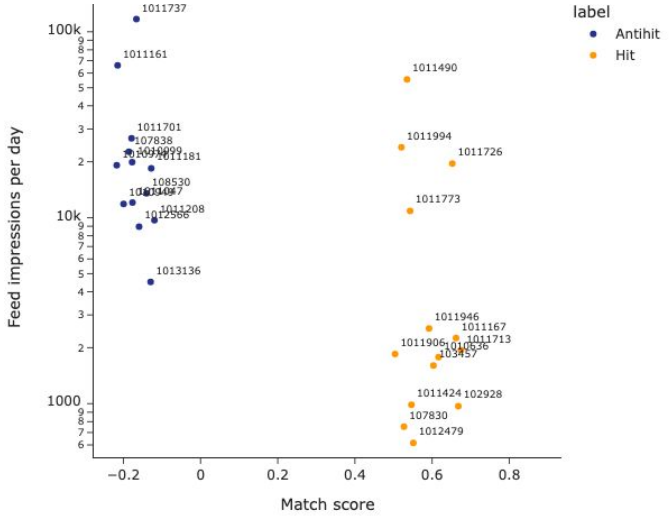


Match with the trend: Bikes

"E-Bikes", "E-Scooter", "Fahrrad", "Motorrad"







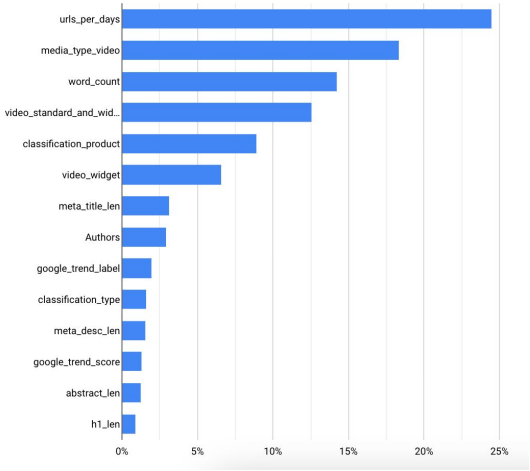
Deep Dive: Final Model

Evaluation Metrics

- Target column: Impressions (scaled & normalized)
- MAE : 0.559
- MAPE : 269.98
- RMSE . 0.707
- RMSLE : 0.295
- r^2 : 0.49

Feature Importance

Method: Sampled Shapley



Model Hyperparameters

Model Type: Boosted Trees

Number of Trees:

- Min: 250
- Max: 500
- Common: 350, 400, 450, 500

Tree Depth:

- Min: 6
- Max: 9

Tree Complexity:

- Min: ~0
- Max: ~10

L1 Regularization:

- Min: 0.1
- Max: 1

L2 Regularization:

- Min: 0
- Max: 3

Prediction Application - Default values & simplifications

Input variables

1. Classification Product
2. Classification Type
3. Publishing Frequency
4. Media Type
5. Video Type
6. Character Counts
   - `h1_len`: Character count of Title
   - `abstract_len`: Character count of Abstract
   - `word_count`: Sum of `abstract_len` and `article_text` length

Simplified default values

1. Number of Days the Article is Online (n_days)
   - Default: 15
2. Google Trend Label
   - Default: 'elektroauto'
3. Google Trend Score
   - Default: 31.0
4. Authors
   - Default: 'lemur'
5. Character Counts
   - `meta_title_len`: Character count of H1
   - `meta_desc_len`: Difference between `abstract_len` and 100

Prediction Demo

Clickbait classification

Model: Roberta Base Clickbase

https://huggingface.co/Stremie/roberta-base-clickbait

DATA-DRIVEN SEARCH FOR TRAFFIC DRIVERS
WITH DATA PROVIDED BY

EFAHRER.com