

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**



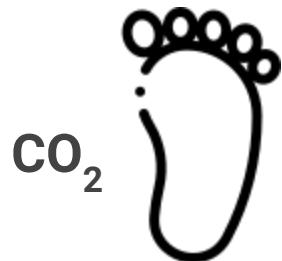
IT  
Engineer

# We'll show how to create impactful content

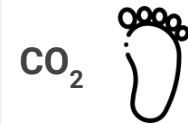
-  
data driven and backed  
by machine learning

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<sub>2</sub> reduction.



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



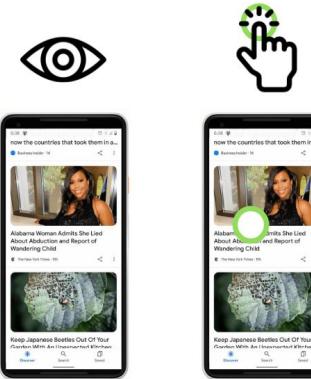
We want to perform a **prognosis of article success**



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**      **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:



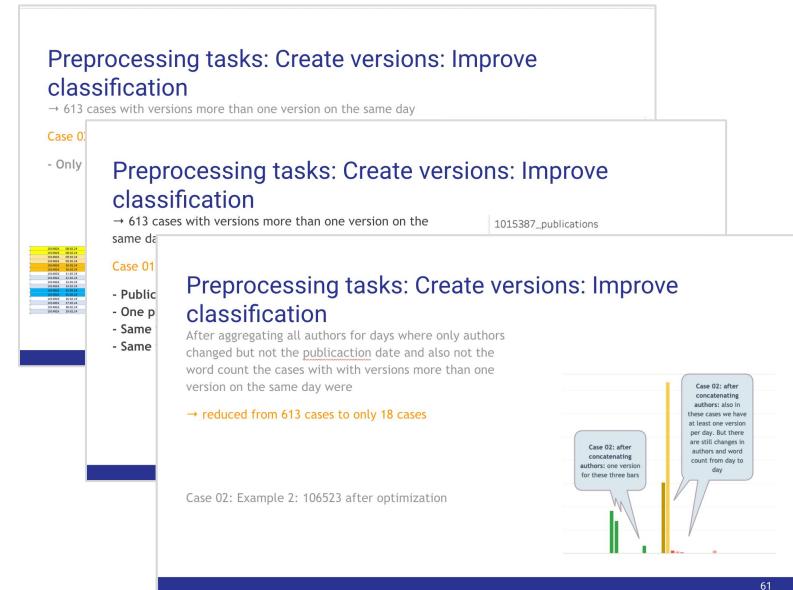
Delete



Impute



Scrape

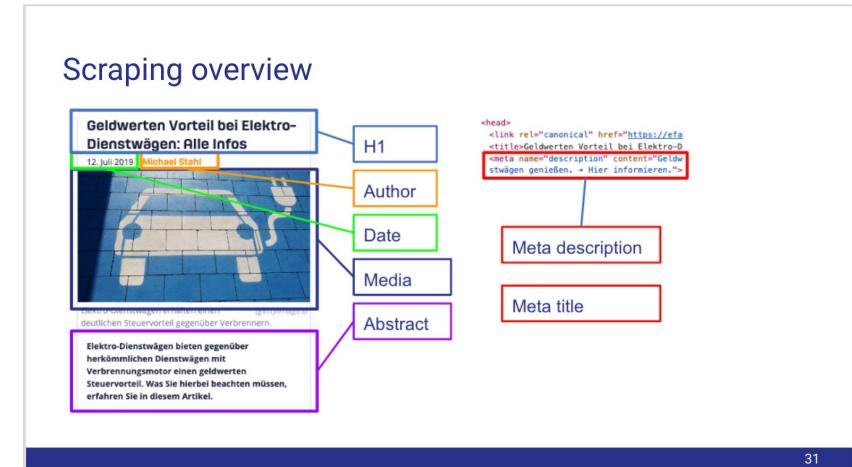
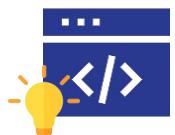


# Scraping of 6 000 articles increased the data quality and added new features

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



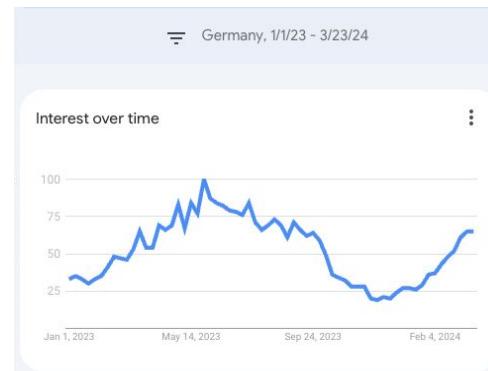
Using a NLP classifier we matched each article with a related search label and trend score



+ 2 features



Search trend for “E-Bike”



Related queries

1	e bike	100	
2	fahrrad	53	
3	damen e-bike	37	
4	e-bike cube	37	
5	cube	37	



# 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

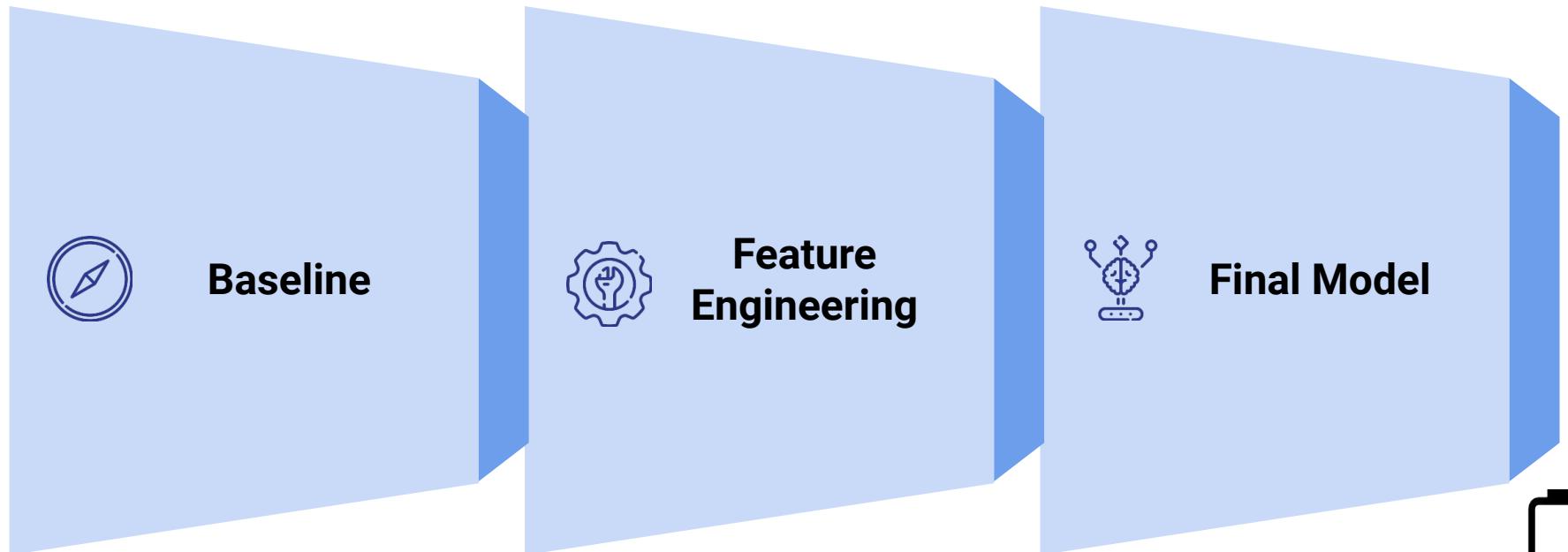


Features in the positive feedback loop with the target were **ignored** or **normalized**:

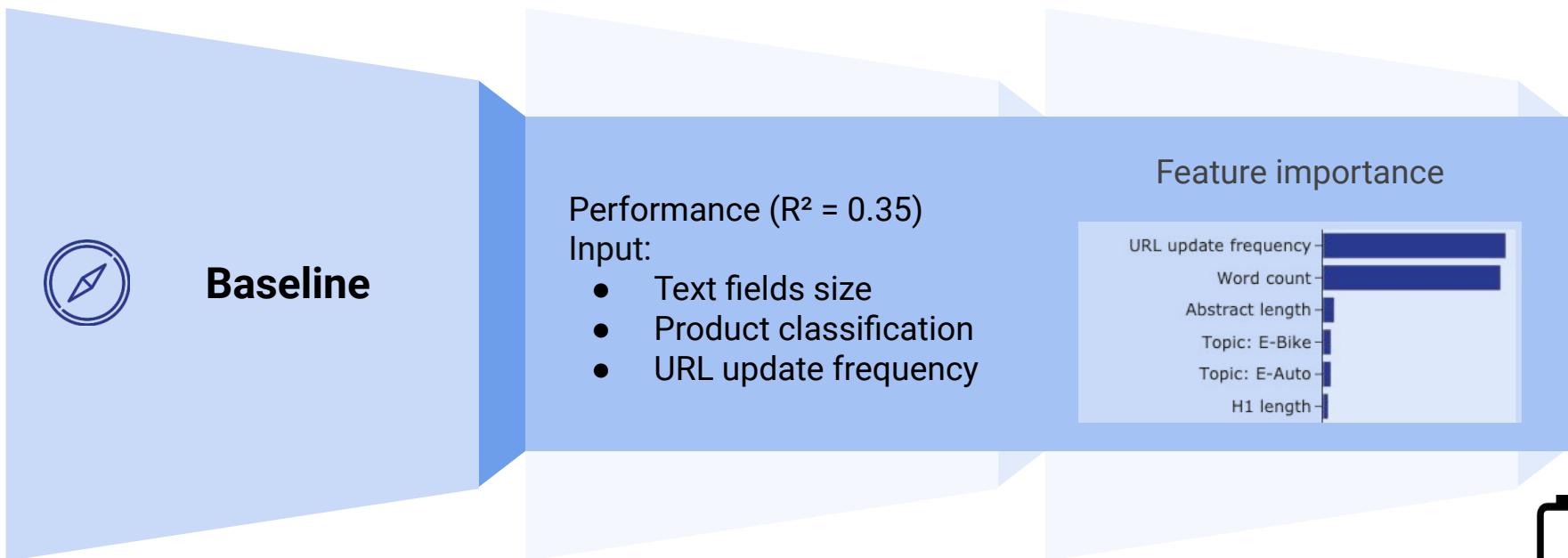
- Number of likes, dislikes, video views
- Number of URLs → *URL update frequency*



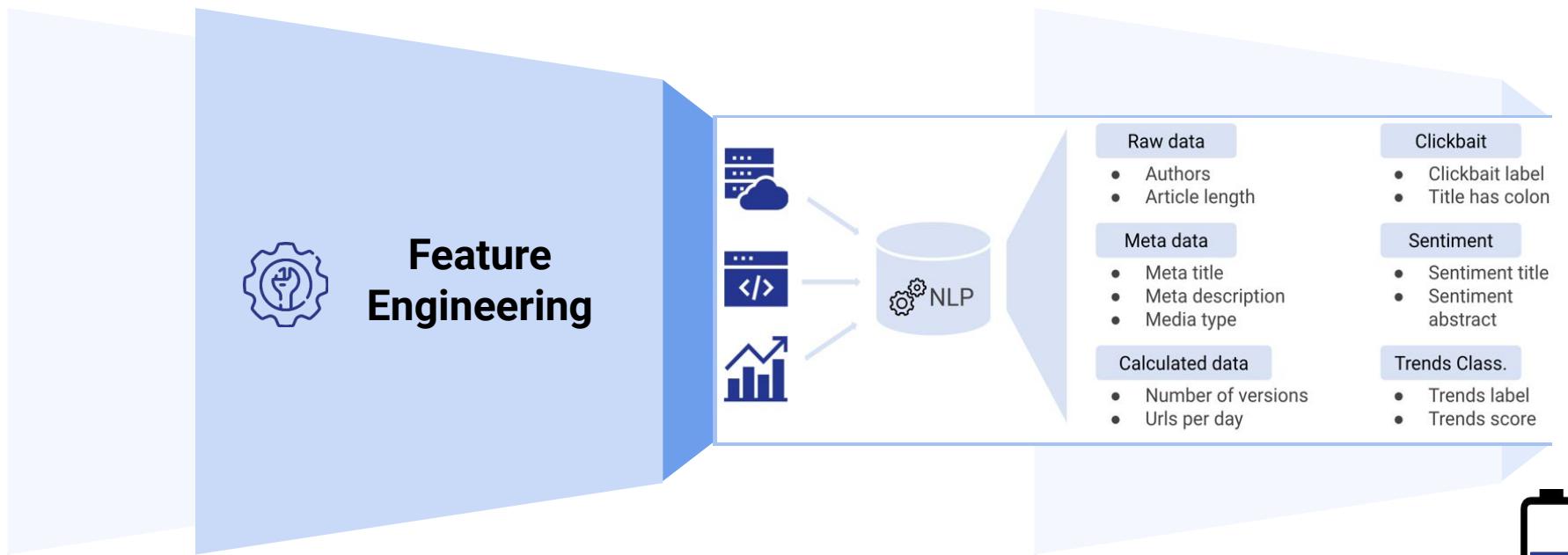
# We verified our hypotheses and created prediction tool



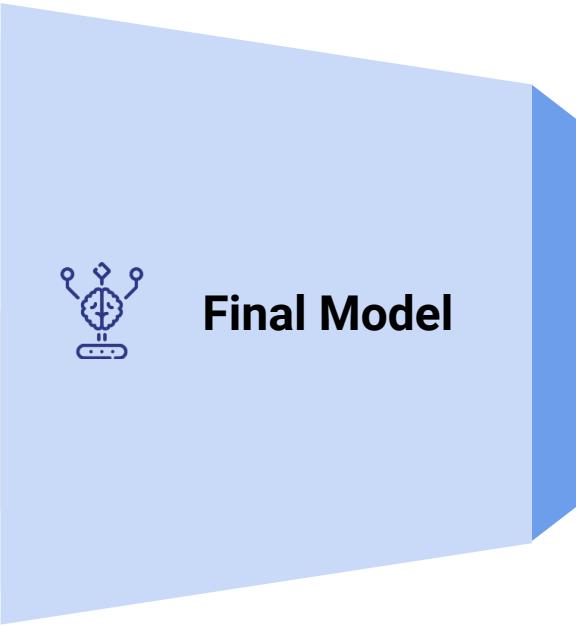
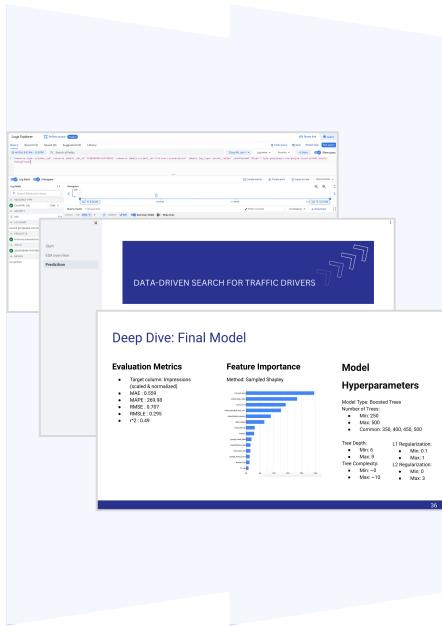
# With the baseline model we created an advanced starting point for our modeling



# We engineered additional features based on the existing data



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



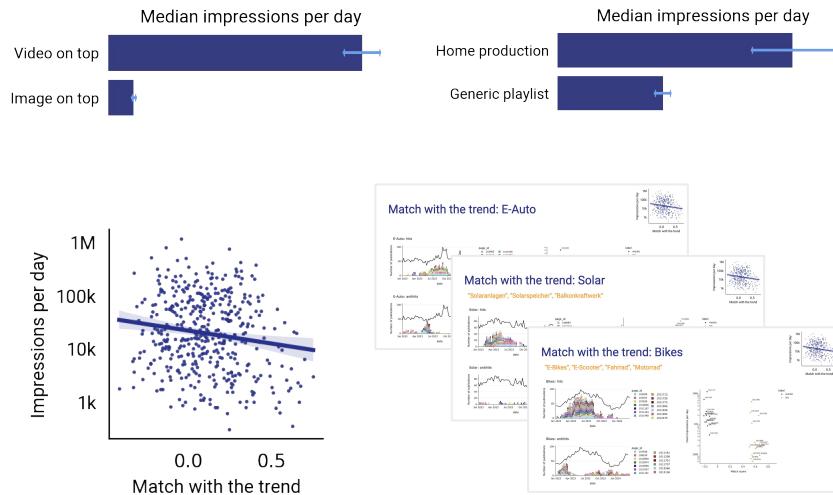
- Modeling with AutoML Tables by Google Vertex AI
- Best performance ( $R^2 = 0.49$ ) for simpler model without full text features

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



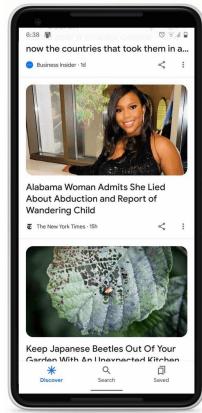
efahrer.com team who  
kindly supported this project

- Markus Höllmüller
- Analytics team

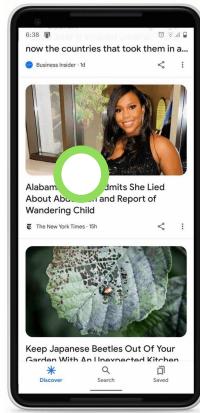
# INTERNAL / Backup

Notes and the knowledge base for the team

# Performance metrics & target features



Feed  
Impression

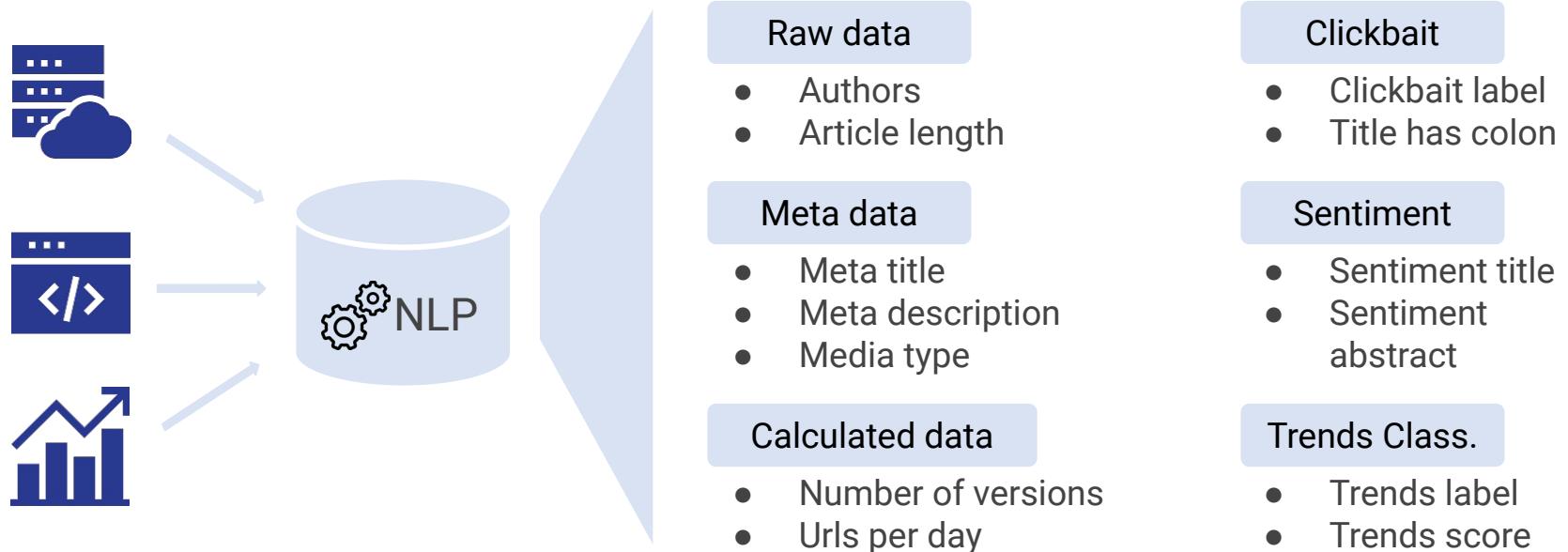


Feed  
Click

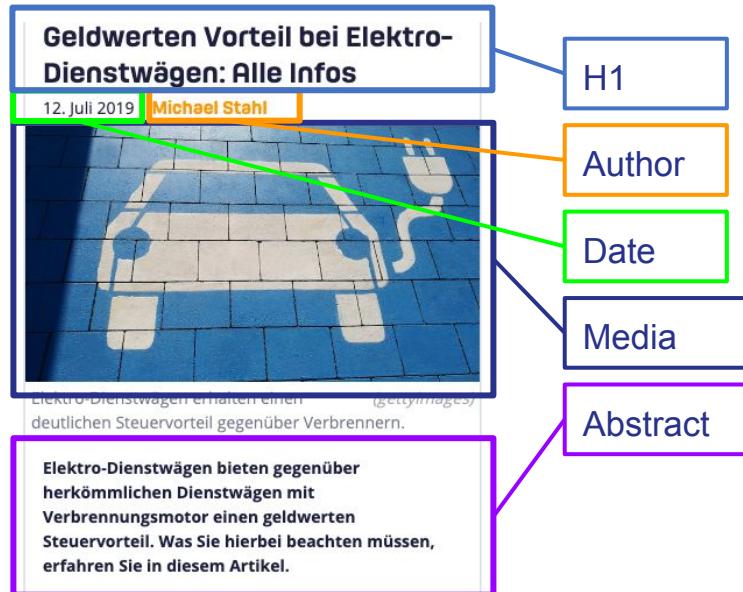


Page  
Impression

# Feature engineering overview



# Scraping overview



```
<head>
  <link rel="canonical" href="https://efa...
  <title>Geldwerten Vorteil bei Elektro-D...
  <meta name="description" content="Geldw...
  stwagen genießen. → Hier informieren.">
```

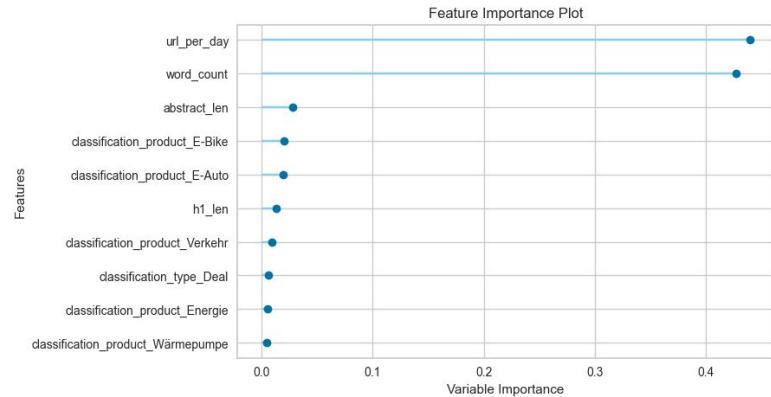
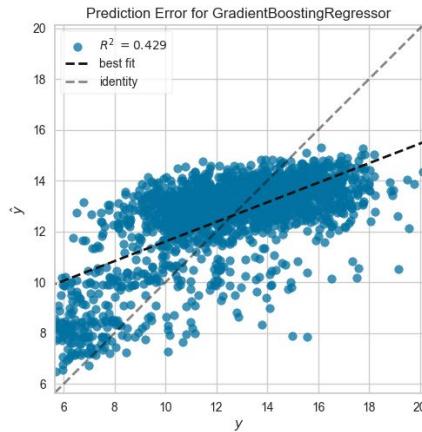
Meta description

Meta title

# 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

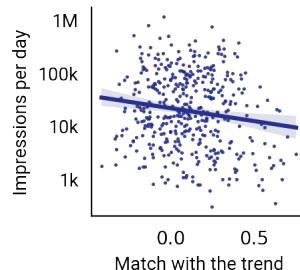


## Features used

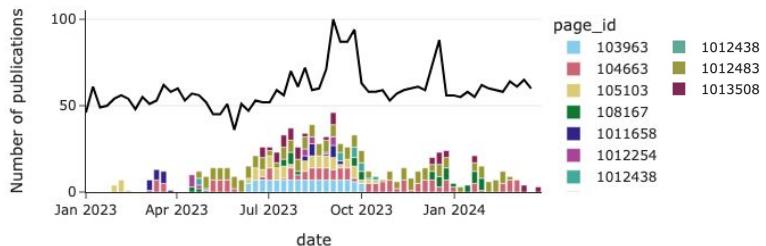
- url\_per\_day, word\_count, abstract\_len, h1\_len
- classification\_product, classification type

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
gbr	Gradient Boosting Regressor	1.8132	5.1203	2.2605	0.3513	0.743	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

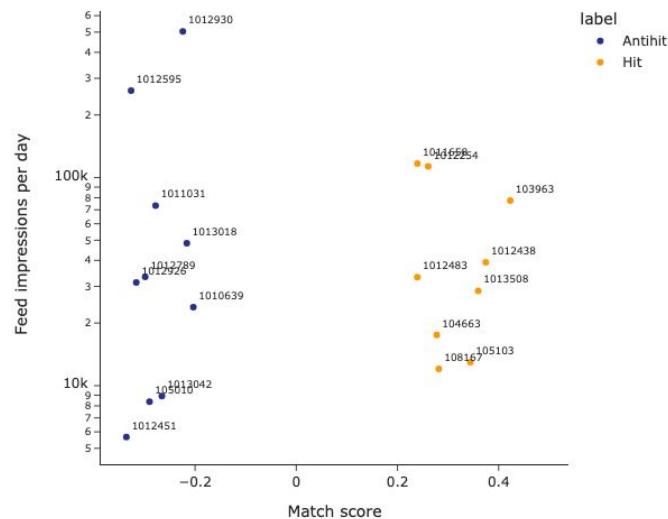
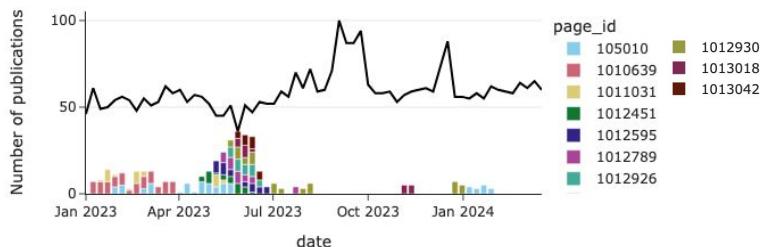
# Match with the trend: E-Auto



E-Auto: hits

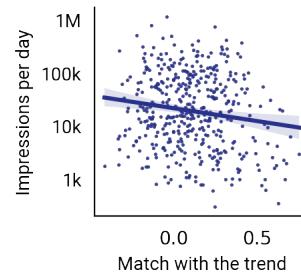


E-Auto: antihits

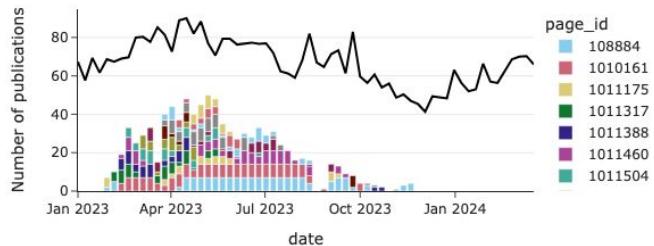


# Match with the trend: Solar

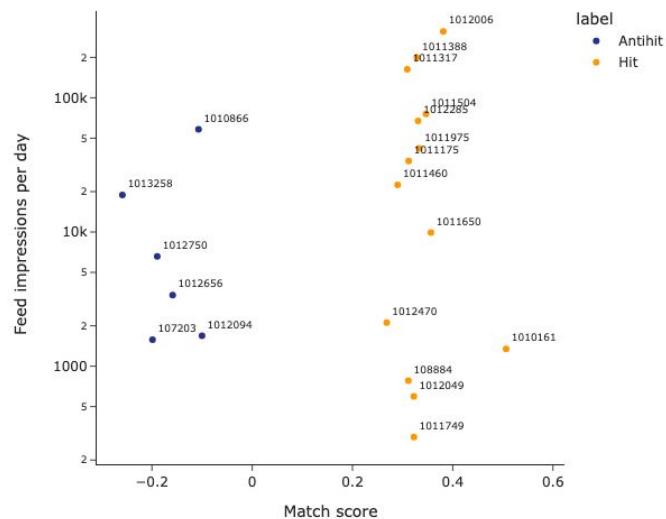
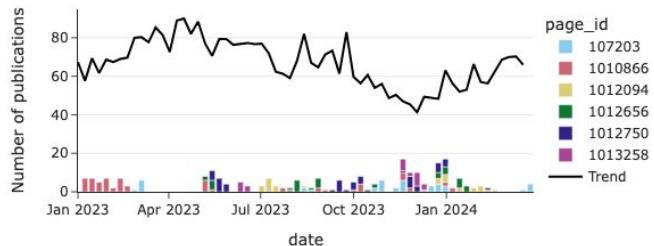
“Solaranlagen”, “Solarspeicher”, “Balkonkraftwerk”



Solar: hits

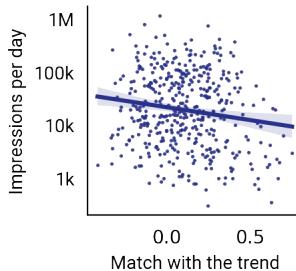


Solar: antihits

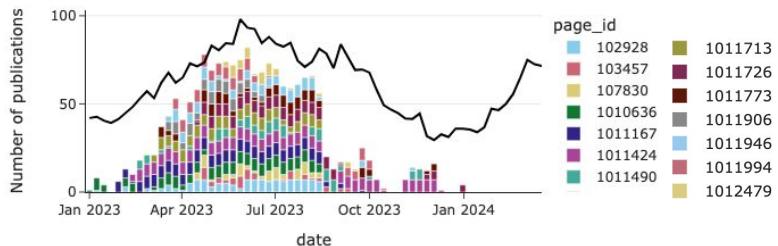


# Match with the trend: Bikes

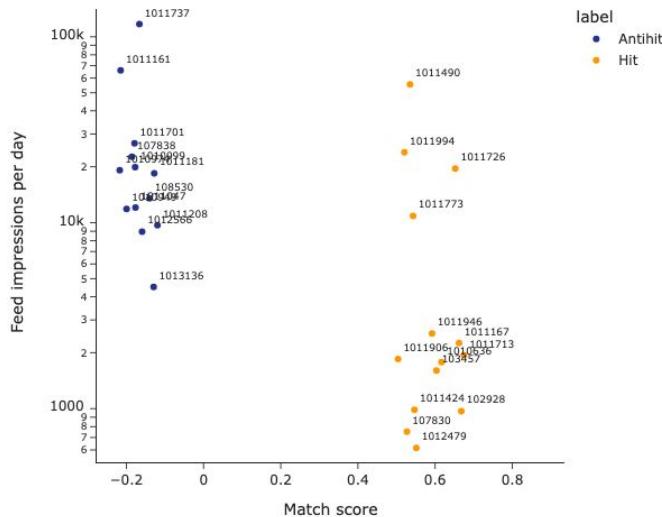
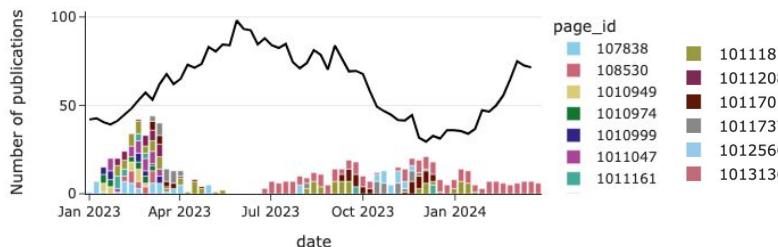
“E-Bikes”, “E-Scooter”, “Fahrrad”, “Motorrad”



Bikes: hits



Bikes: antihits



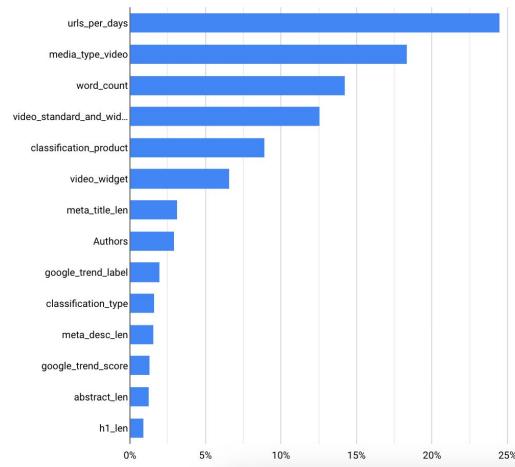
# Deep Dive: Final Model

## 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

L1 Regularization:

- Min: 0.1
- Max: 1

Tree Complexity:

- Min: ~0
- Max: ~10

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

The screenshot shows a user interface for a data-driven search application. On the left, a sidebar menu lists several options: Start, EDA overview, Check sentiment, Exploration freestyle, Google trends, and Prediction. The 'Prediction' option is currently selected and highlighted with a grey background. The main content area features a large blue banner with the text 'DATA-DRIVEN SEARCH FOR TRAFFIC DRIVERS' in white. To the right of the banner, there are three stylized, light-blue geometric shapes resembling arrows or brackets pointing upwards and to the right. Below the banner, the title 'Prediciton of Article Impressions' is displayed in bold black text. Underneath the title, there are three sections: 'Title', 'Abstract', and 'Article'. The 'Title' section contains the text 'Aldi-Rückfahrkamera für jedes Auto: Knallerpreis und schnelle Installation'. The 'Abstract' section contains the text 'Autofahren ohne Rückfahrkamera können sich die meisten schon gar nicht mehr vorstellen. Schließlich besitzt so gut wie jedes neue Auto dieses fest verbaute Hilfsmittel. Wer dennoch ohne Rückfahrkamera fährt, für den bietet Aldi eine praktische Lösung: eine leicht montierbare Kamera mit Solarpanel.' The 'Article' section contains the text 'EFAHRER.com informiert Sie laufend über die besten Deals für E-Autos. Förderung. Laden & Co.' followed by a detailed description of the Aldi RC-300WS camera, mentioning its wireless installation and solar panel power.

Start  
EDA overview  
Check sentiment  
Exploration freestyle  
Google trends  
**Prediciton**

Deploy :

DATA-DRIVEN SEARCH FOR TRAFFIC DRIVERS

**Prediciton of Article Impressions**

Title

Aldi-Rückfahrkamera für jedes Auto: Knallerpreis und schnelle Installation

Abstract

Autofahren ohne Rückfahrkamera können sich die meisten schon gar nicht mehr vorstellen. Schließlich besitzt so gut wie jedes neue Auto dieses fest verbaute Hilfsmittel. Wer dennoch ohne Rückfahrkamera fährt, für den bietet Aldi eine praktische Lösung: eine leicht montierbare Kamera mit Solarpanel.

EFAHRER.com informiert Sie laufend über die besten Deals für E-Autos. Förderung. Laden & Co.

Article

Die Rückfahrkamera RC-300WS ist vielseitig einsetzbar und lässt sich im Handumdrehen installieren. Durch das eingebaute Solarmodul mit Akku benötigt die Kamera keinen zusätzlichen Strom des Fahrzeugs, sondern läuft autonom.

Die kabellose Verbindung zwischen Kamera und Monitor erfordert zudem keine zusätzlichen Kabel. Wer eine Rückfahrkamera für seinen Pkw, Wohnwagen oder Anhänger sucht, kann hier nicht viel falsch machen. Einziges Manko: Vergleichbare Modelle gibt es auch von anderen Anbietern zu einem deutlich günstigeren Preis. Beispielsweise die solarbetriebene Rückfahrkamera von AEG für 89 Euro.

# Clickbait classification

Model: Roberta Base Clickbase

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

DATA-DRIVEN SEARCH FOR TRAFFIC DRIVERS  
WITH DATA PROVIDED BY

---

