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

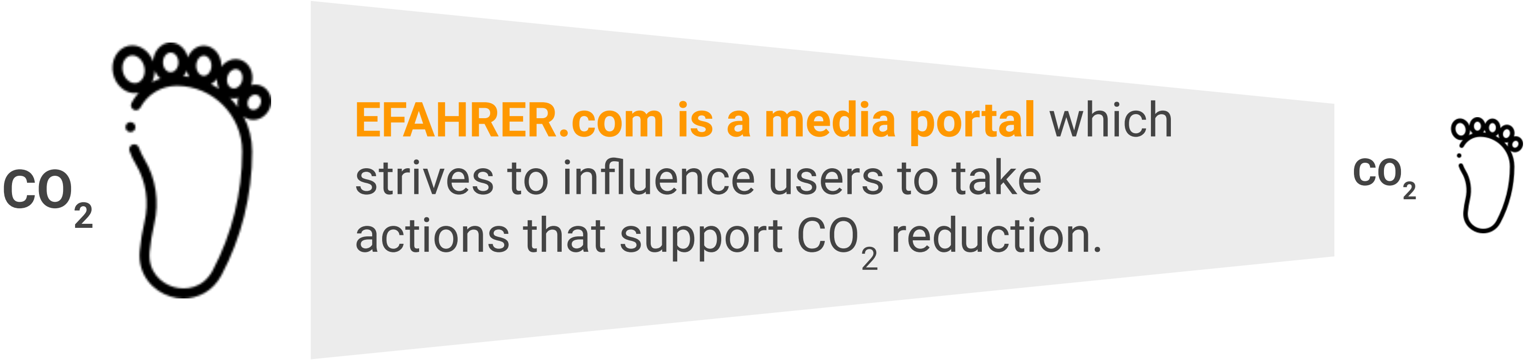
We’ll show how to create impactful content

-

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



CO_2

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

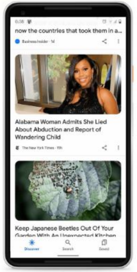
CO_2

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



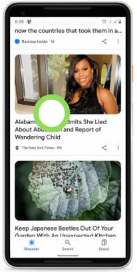
now the countries that took them in a...

Alabama Woman Admits She Lied About Abduction and Report of Wandering Child

Keep Japanese Beetles Out Of Your Garden With An Unexpected Kitchen

Feed Impression

Billions



now the countries that took them in a...

Alabam dmits She Lied About Ab and Report of Wandering Child

Keep Japanese Beetles Out Of Your Garden With An Unexpected Kitchen

Feed Click

Millions

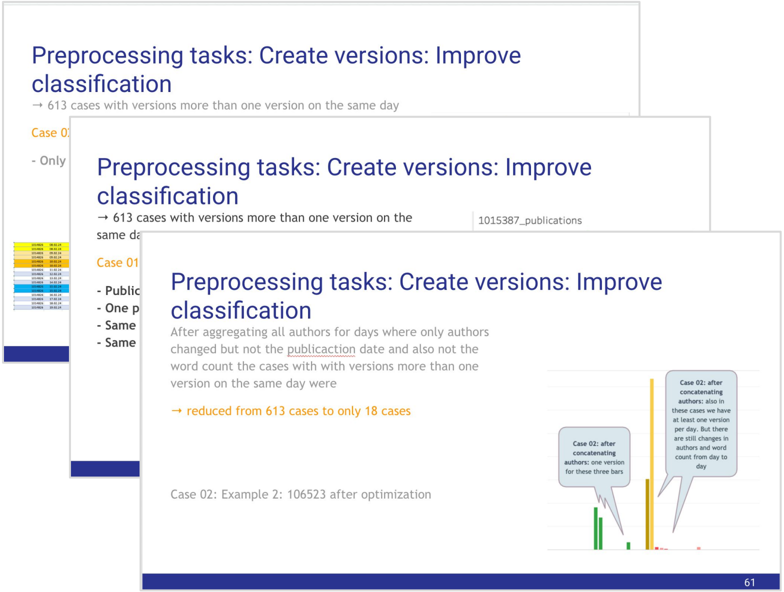
Data quality and completeness of raw data led to extensive preprocessing and analysis for modelling

Challenges

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

Strategies:

- Delete
- Impute
- Scrape



Preprocessing tasks: Create versions: Improve classification

→ 613 cases with versions more than one version on the same day

Case 0

- Only

Preprocessing tasks: Create versions: Improve classification

→ 613 cases with versions more than one version on the same da

1015387_publications

Case 01

- Public
- One p
- Same
- Same

Preprocessing tasks: Create versions: Improve classification

After aggregating all authors for days where only authors changed but not the publicaction date and also not the word count the cases with with versions more than one version on the same day were

→ reduced from 613 cases to only 18 cases

Case 02: Example 2: 106523 after optimization

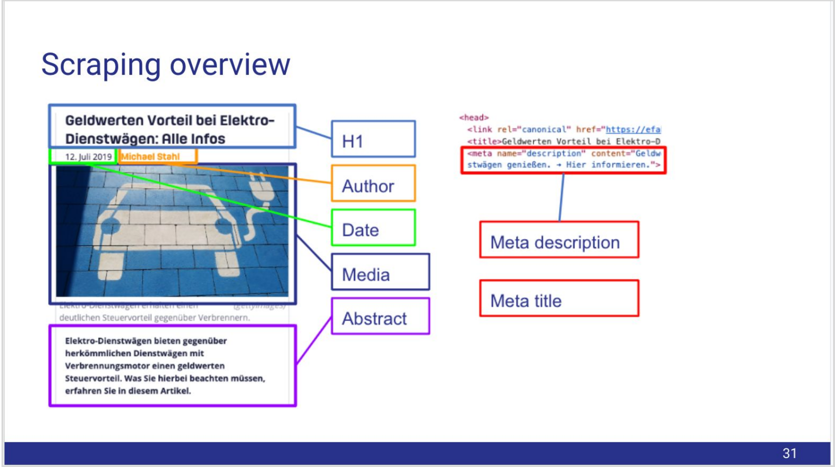
Case 02: after concatenating authors: one version for these three bars
Case 02: after concatenating authors: also in these cases we have at least one version per day. But there are still changes in authors and word count from day to day
61

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)



Scraping overview

Geldwerten Vorteil bei Elektro-Dienstwägen: Alle Infos

12. Juli 2019 Michael Stahl

Elektro-Dienstwägen erhalten einen (gettyimages) deutlichen Steuervorteil gegenüber Verbrennern.

Elektro-Dienstwägen bieten gegenüber herkömmlichen Dienstwägen mit Verbrennungsmotor einen geldwerten Steuervorteil. Was Sie hierbei beachten müssen, erfahren Sie in diesem Artikel.

H1

Author

Date

Media

Abstract

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

Meta description

Meta title

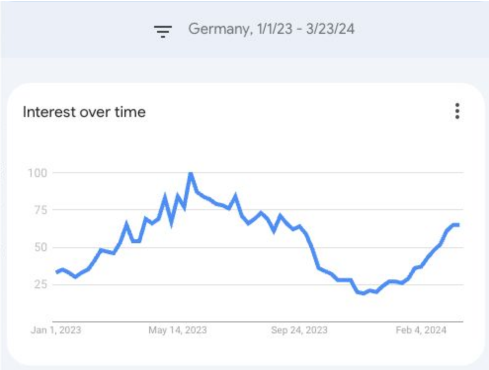
31

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

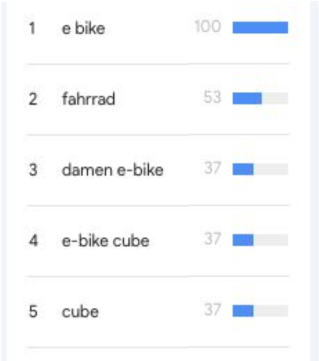
Using a NLP classifier we matched each article with a related search label and trend score

+ 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

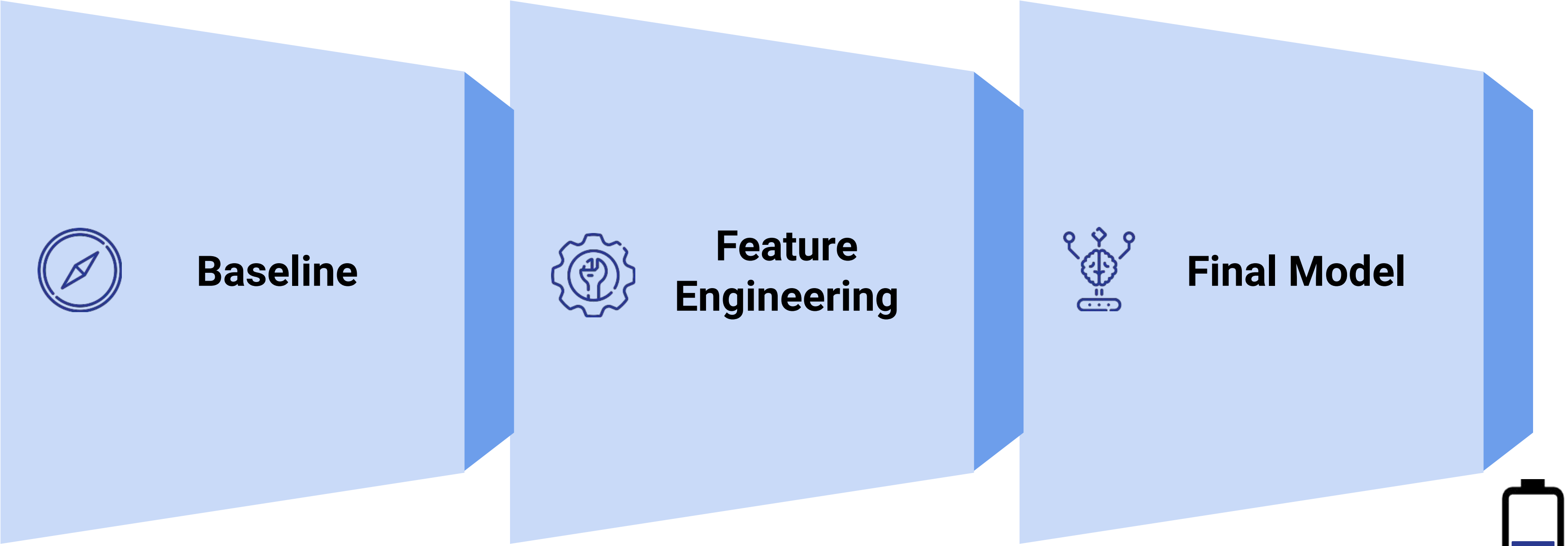
- Article genre and topic
- Type of the first media: video or image
- Word count and lengths of the metadata

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



Baseline

Feature Engineering

Final Model

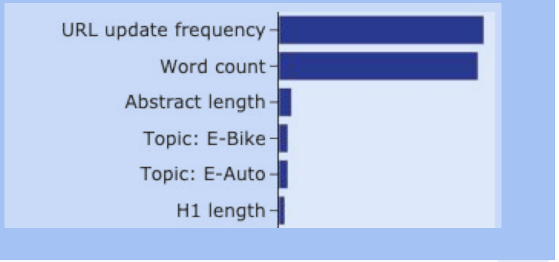
With the baseline model we created an advanced starting point for our modeling

Baseline

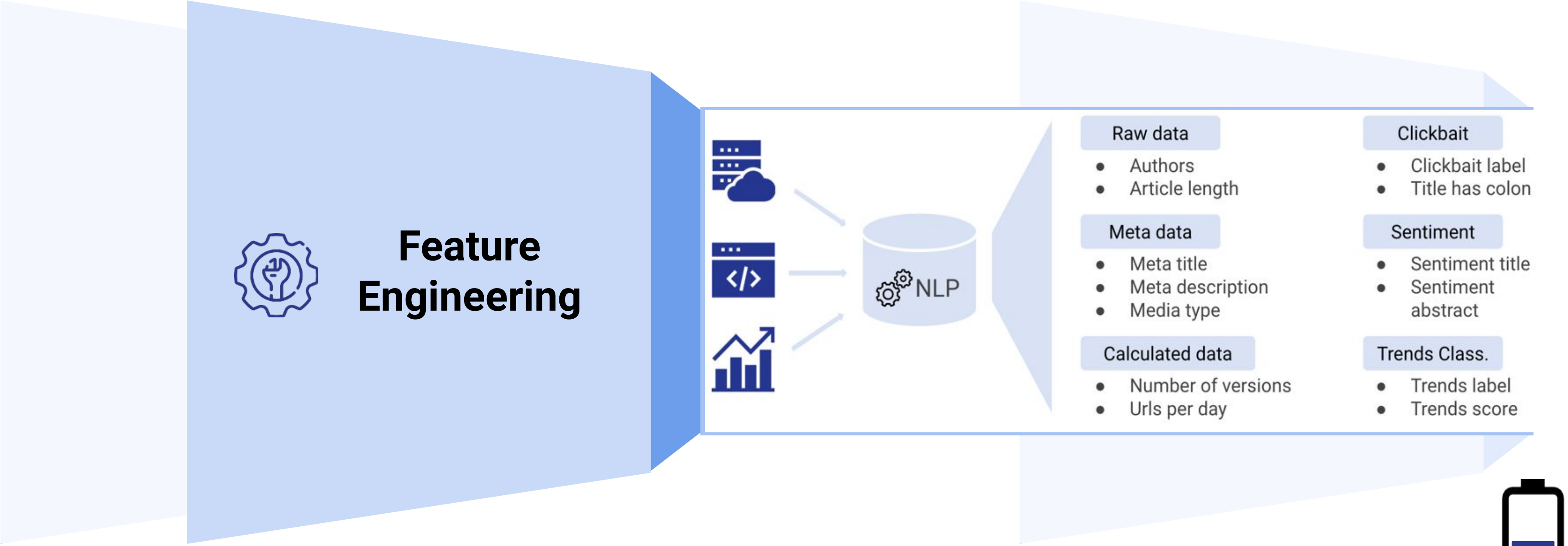
Performance ($R^2 = 0.35$)
Input:

- Text fields size
- Product classification
- URL update frequency

Feature importance



We engineered additional features based on the existing data



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

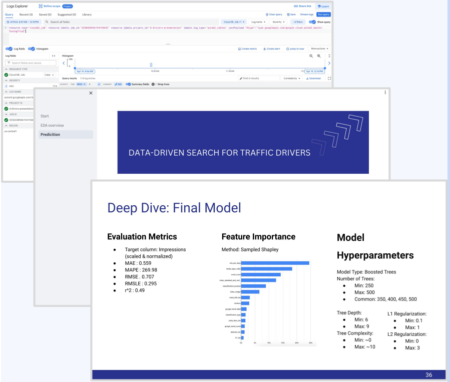
Sentiment

- Sentiment title
- Sentiment abstract

Trends Class.

- Trends label
- Trends score

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



```
1 resource.type="cloudml_job" resource.labels.job_id="3250438946194194432" resource.labels.project_id="d-drivers-presentation" labels.log_type="automl_tables" jsonPayload."@type"="type.googleapis.com/google.cloud.automl.master.
  TuningTrial"
```

DATA-DRIVEN SEARCH FOR TRAFFIC DRIVERS

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

36

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

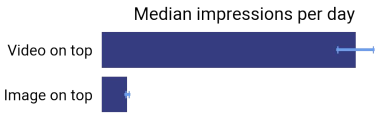
1. Change in URL has tangible impact on impressions

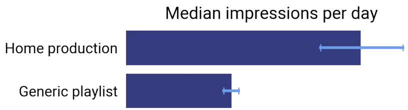
2. The algorithm prefers articles with videos over images as the first media on page

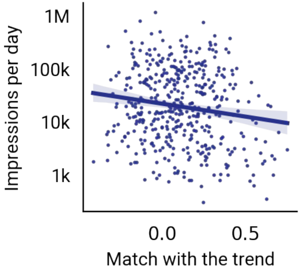
3. Video production pays off!

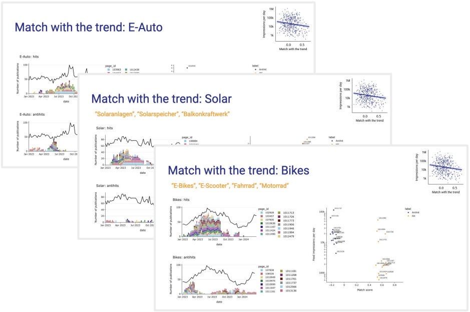
4. Optimize publication timing alongside trends

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