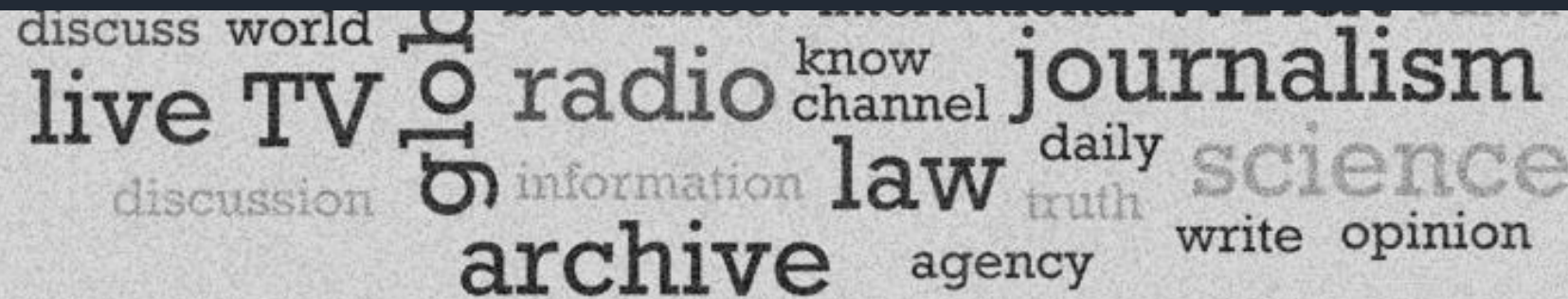


anchorwoman  
articles  
newspapers  
correspondent  
health  
who  
report  
business  
publication 24/7 sports  
television  
media  
journalists  
editorial  
war  
publish information  
magazine

# Predicting Visits of digital news articles

Analysis for a German publisher (Frankfurter Allgemeine Zeitung)

Fabian Paul, 13/07/2022



discuss world  
live TV  
discussion  
radio  
know channel  
journalism  
daily  
science  
truth  
law  
archive  
agency  
write opinion

# Business Problem



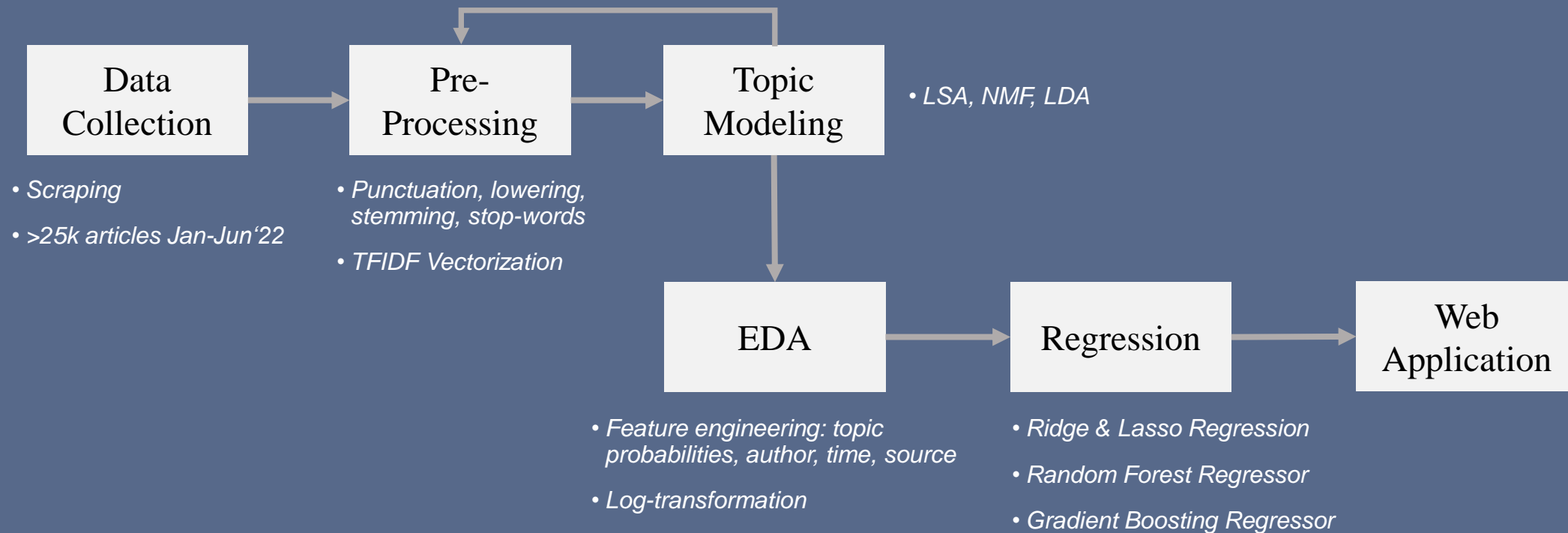
- Importance of **audience reach** for advertising and subscription sales
- Insights on article topics and specifics crucial for:
  - ... journalists to **tailor texts** to audience needs
  - ... editors to **decide on pay vs. free** articles and release times

# Objective



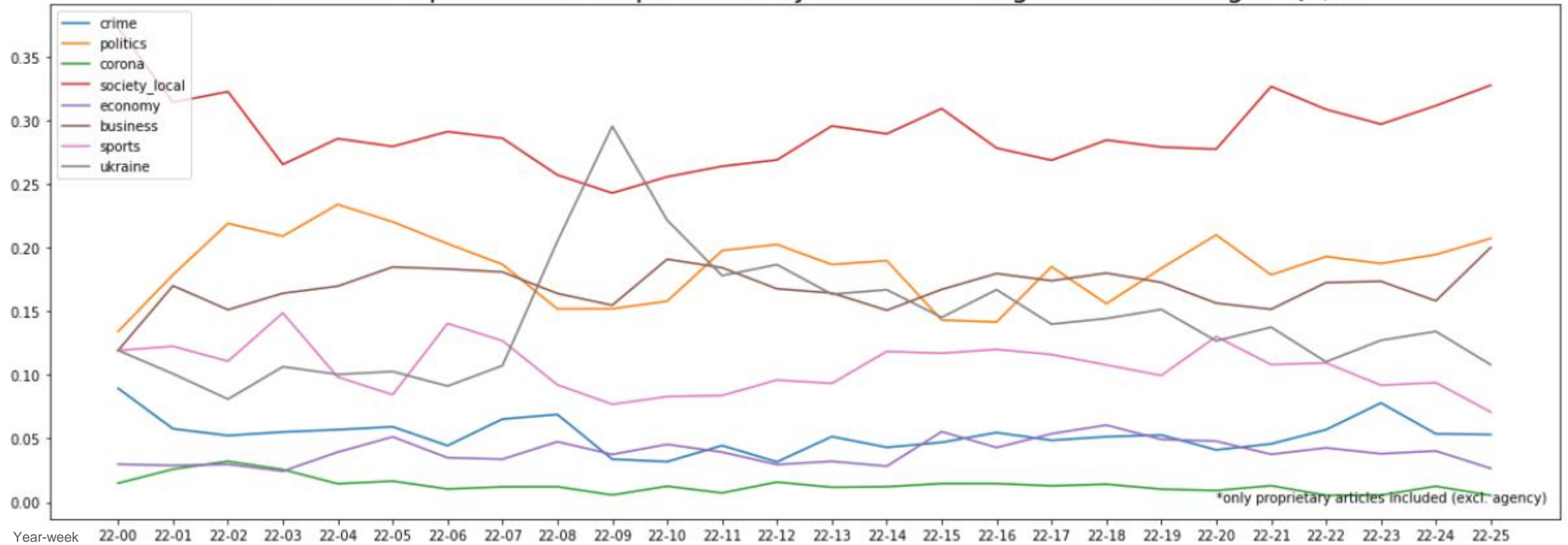
- Which topics are **significantly correlated** to article visits?
- **Predict article audience** before an article is published?

# Methodology



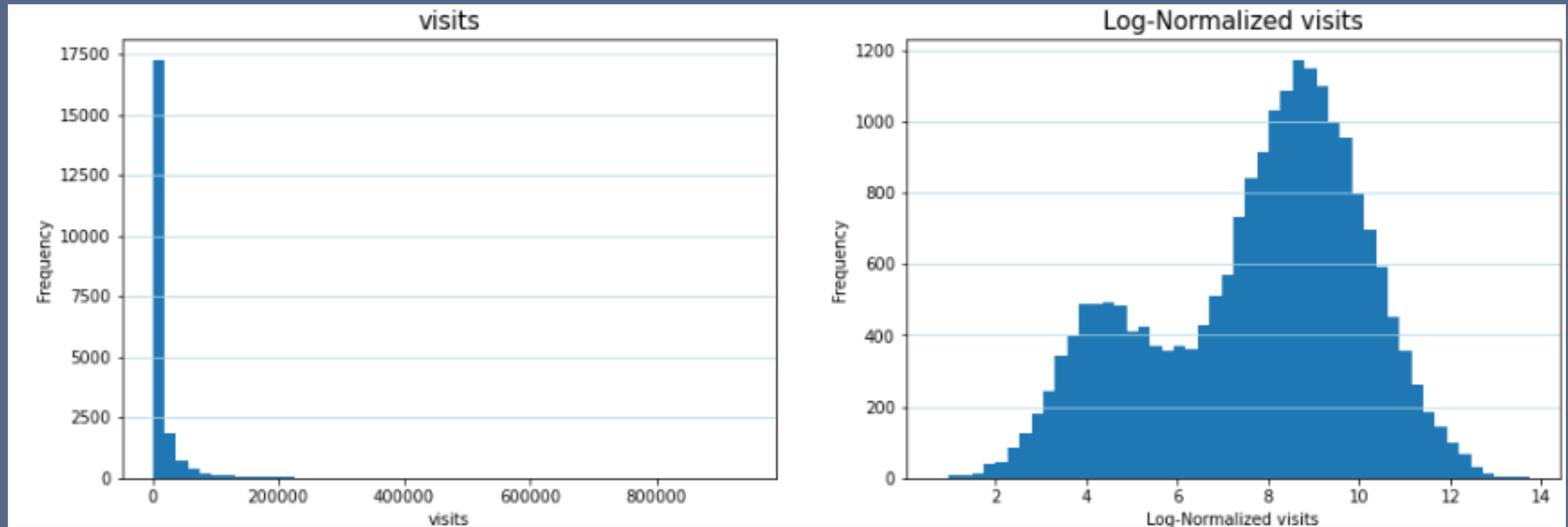
# Results

%-Share of Topics of Articles published by Frankfurter Allgemeine Zeitung in Q1/22



# Results

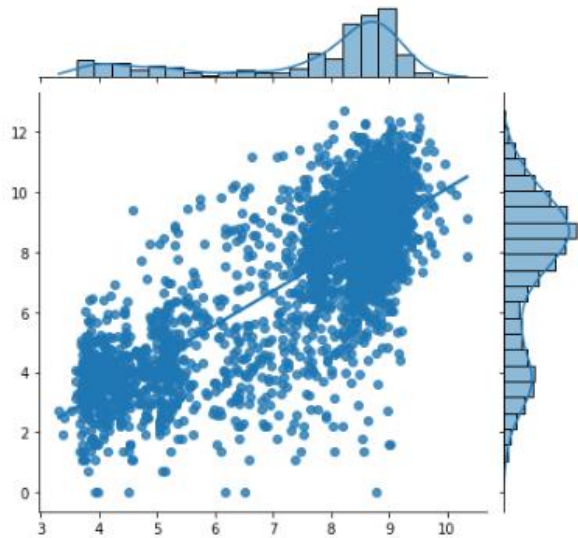
Histograms of Visits vs. Log-Normalized Visits





# Results

Gradient Boosting - Prediction vs. Test Values (Log-Transformed)



## Performance Gradient Boosting Regressor:

- $R^2$  of 0.61
- MAE of 3.33 Visits
- RMSE of 4.87 Visits

**Frankfurter Allgemeine**  
ZEITUNG FÜR DEUTSCHLAND

## Article Visits Prediction

Please enter Article Characteristics to receive the estimated visits of your article

Copy-Paste Article Text

In which ressort is the article published?

You selected: <select>

What is the estimated reading time of the article (in minutes)?

[Link Web Application](#)

# Conclusion

## Results

---

- Prediction Model with a **MAE of 3.33 visits** per article
- Article features:
  - + Topic, paid/free, author, link to authors personal website, estimated reading-time, source
  - Paid-articles, author Knop, source weekday print
- Publishing time features:
  - + Publishing day
  - Publishing time

## 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
- 5 → Include Orders and Converted Orders as target metric

**Thank you for your attention!**