

2025



Dylan Haazelager

HvA

17-8-2025

Inhoud

1	Model information	5
1.1.1	SARIMAX (AutoRegressive Integrated Moving Average with Exogenous Variables)	5
1.1.3	Naïve Baseline Model).....	7
2.1.1	Time Series Models: Parameters, Settings, and Pros/Cons.....	7
2.1.2	ARIMAX (AutoRegressive Integrated Moving Average with Exogenous Variables).....	9
2.1.3	Random Walk (Naïve Baseline Model).....	10
3	D1 Feedback.....	11
3.1	Change list	11
3.1.2	20 May – Meeting with Kees	30
3.1.3	14 May – Meeting with Marcio	30
3.1.4	13 May – Written feedback from Marcio.....	30
3.1.5	12 May – Feedback from Daphne	31
3.1.6	07 May – Go/No-Go Check.....	31
3.1.7	06 May – Meeting with Prof. Van Montfort.....	31
3.1.8	17 April – Planning & Topic Shift	31
3.1.9	change list	32
4	Email and keywords:.....	34
5	Prompt used for translating the Key words from Dutch to English:	39
5.1	Python script for google trends url	41
6	Dataset documentation.....	41
6.1	Consumer Confidence, Economic Climate, and Willingness to Buy (CBS, 2007–2025) 41	
6.2	Fuel Pump Prices per Day 2007-2025.....	44
6.3	Registrations by Fuel Type and Vehicle Category – Netherlands (2007–2025)	45
	NOS News Articles – Netherlands (2015–2025) (Not used)	46
7	Model Selection Rationale	46
8	Daily brain dump	48
16	Introduction to Seasonal ARIMA (SARIMA)	52
16.1.1	Understanding SARIMA Notation.....	53
16.1.2	How SARIMA Works.....	53
23.1.1	1/08.....	62
25.1.1	14/07	71
36.1	10/07.....	74
44.1.1	02/07	80
50.1	Samenvatting Feedbackgesprek Scriptie.....	92

51.1.1	28/06	94
51.1.2	26/06	94
51.1.3	13/06	101
	10/06	105
57.1.1	09/06	106
59.1.2	6/06	110
71.1.1	05/06	119
71.1.2	04/06:	120
71.1.3	Meeting Log – Discussion with Supervisor (Marcio)	120
71.1.4	03/06	132
71.2	29/05	142
71.3	28/05	144
71.4	27/05	147
	Doe.....	147
71.5	23/05 70%	148
71.6	22/05	148
71.7	21/05	148
71.8	20/05	148
	71.8.1 Todolist:	148
	Meeting with Kees	148
71.9	19/05	150
71.10	Zo18/05	150
71.11	Za 17/05	150
71.12	16/05 70% Inleveren	154
71.13	15/05	155
71.14	14/05	163
	71.14.2 Results new keywords:	170
71.15	13/05	171
71.16	12/05	180
	71.16.1 Oplossing keyword search?	180
71.17	11/05	184
	Samenvattingen Kernpaper	185
71.18	10/05	196
	71.18.1.4 Internetgebruik en online activiteiten (CBS-tabel, 2024)	199
71.19	09/05	200
71.20	08/05	201

71.20.1	Literatuuronderzoek:	201
71.21	07/05 Onderzoekplan herzien	205
71.21.1	Meeting met Marcio.....	210
71.21.2	Belangrijkste feedbackpunten	210
71.22	06/05.....	212
71.22.2	Gesprek met labbegeleider Kees.....	217
71.23	05/05.....	218
71.23.1	Verkenning en filtering van nieuwsartikelen m.b.v. keywords en embedding	218
71.23.2	Aanbevelingen voor verdere verbetering van het model en de sentimentanalyse	
	224	
71.24	2/05.....	235
71.24.1	Idee generatie: waarde propositie:	236
71.25	01/05 To do list.....	237
71.26	30/04: (Leave one out toepassen, Nos Sentiment maken met keyword search, evt train/test split maken)	239
71.27	28/04 (Dataset gemerged, code naar stefan gestuurd)	240
71.28	21/04.....	247
71.28.1	TO Do list:	247
72	Meeting Marcio 17/04.....	252
72.2	Update last feedback moment:	255
Algemeen / Overkoepelend.....	255	
Introductie	255	
Probleemstelling en Onderzoeksdoel.....	256	
Onderzoeksvragen	256	
Literatuurstudie	256	
Methodologie	256	
Reflectie en Discussie	257	
72.3	15/04:.....	259
72.4	14/04:.....	263
72.5	03/04.....	264
72.6	29/03.....	265
72.7	26/03.....	266
Feedback thesis	266	
72.8	13/03.....	269
72.8.1	12/03 feedback daphne op concept	269
73	Jip en janeke taal:.....	271

73.1 Feedback Dinsdag 25 februari:276

1 Model information

1.1.1 SARIMAX (AutoRegressive Integrated Moving Average with Exogenous Variables)

1.1.2 What we used

- **Library:** statsmodels SARIMAX
- **Target:** monthly BEV registrations (NL), 2007–2025
- **Transformations:** Box–Cox on strictly positive series; Yeo–Johnson for Consumer Confidence
- **Evaluation:** expanding (walk-forward) validation, fixed horizon **$h = 3$** ; train/val/test = **80/10/10**
- **Baseline (univariate):** ARIMA(1,1,0) × (0,0,1)_12
- **Exogenous candidates (screened one-by-one):** Gasoline (Euro95), Company car tax (bijtelling), Home charging (Google Trends), Consumer confidence
 - Selected lags after screening: **gasoline(12), tax(11), home(8), CCI(11)**
 - Constraint: **$L \geq h$** to avoid look-ahead

Key SARIMAX parameters (and what we actually set)

- **order = (p, d, q)**
 - **Used:** (1, 1, 0) → AR(1), one difference for trend, no non-seasonal MA.
- **seasonal_order = (P, D, Q, m)**
 - **Used:** (0, 0, 1, 12) → seasonal MA(1) at **$m = 12$** months.
 - Rationale: clear annual pattern; seasonal operator models 12-month dependence **without** adding many explicit lagged target features.
- **exog**
 - **Used:** one lagged exogenous series at a time during screening (e.g., $\text{gasoline}_{t-12} \dots \text{gasoline}_{t-1}$), then combinations for robustness checks.
 - Important: by using **lags ≥ 3**, multi-step forecasts do **not** need future values of exogenous series.
- **enforce_stationarity=True, enforce_invertibility=True** (defaults)
 - Keep estimates in stable regions.
- **Diagnostics & selection:**
 - ADF/KPSS for (non)stationarity on the transformed scale.
 - **Same-sample** AIC / BIC / AICc sweep over candidate lags to shortlist.

- **Rolling (expanding) validation** with $h=3$; select by **MAPE** (primary), require Ljung–Box $p > 0.05$.

Results snapshot (for context)

- **Validation ($h=3$):**
 - Univariate baseline: **MAPE = 11.95%**
 - Best single exog: **Gasoline(12) → 10.51%**
 - “Best exogs” combo (gas+tax+home): **10.98%**
- **Test ($h=3$):**
 - **Naïve best: 12.14%**
 - Univariate baseline: **13.30%**
 - Single-exog SARIMAX range: **13.26–13.90%** (best = **CCI(11) 13.26%**)
 - “Best exogs” combo: **14.15%**
 - Auto-ARIMA: **16.29%**
- **Takeaway:** validation gains from gasoline(12) did **not** carry to test; differences point to **period sensitivity / distribution shift**, not classic overfitting.

Why seasonal ARIMA (and not many target lags)?

The strong 12-month autocorrelation is captured parsimoniously by the seasonal operators and the **seasonal MA(1)** term. This avoids adding explicit target lags (e.g., $y_{t-12}, y_{t-24}, y_{t-12}, y_{t-24}$) as regressors, which would inflate parameters and multicollinearity with only 173 training months.

Pros (for this EV task)

- **Works with limited data** (monthly, $T \approx 173$ for training).
- **Interpretable:** AR(1) = near-term persistence; seasonal MA(1) = annual seasonal shock.
- **Handles seasonality natively** ($m=12$) without piling on manual lags.
- **Operationally simple:** one series + a small number of screened exogenous lags.
- **Forecast intervals** available; residual diagnostics standard.

Cons / caveats

- **Linearity & fixed lags:** cannot capture nonlinear threshold effects of policy or time-varying delays.
- **Fragile exogenous gains:** effects were **period-dependent** (validation vs test), so benefits may not generalize.
- **Multicollinearity risk** when combining multiple exogs; coefficient instability likely in small samples.
- **Requires careful evaluation protocol:** need $L \geq h$ and rolling validation to avoid leakage.

- **MAPE sensitivity** to low volumes (we mitigated via period choice, but worth noting).

Practical notes for reproduction

- Screen exogs **one-by-one** against the fixed backbone **ARIMA(1,1,0)x(0,0,1)_12** on the **same training slice** (constant effective sample), rank by **AIC/BIC/AICc**, then confirm via **rolling MAPE + Ljung–Box**.
- Keep transformation parameters (λ , shift) **fixed** after fitting on train; invert for reporting on the original scale.
- Refit selected model on **train+val** before scoring the **locked test**.

1.1.3 Naïve Baseline Model)

Definition. $y^T+h|T=y_T\hat{y}_{T+h|T}=y_Ty^T+h|T=y_T$. For each origin TTT, the next three months are set equal to the last observed value.

Why include it?

- It is the **hardest simple benchmark to beat** for short horizons in economic series with strong persistence.
- Provides a transparent floor for usefulness: a model should be meaningfully better than Naïve to justify added complexity.

Performance here (test, h=3): MAPE = 12.14% (best overall).

Implication: for 3-month EV forecasts in this dataset, simple persistence is very competitive; SARIMA/SARIMAX variants were close but did not surpass Naïve on the test window.

2 Auto-ARIMA baseline

- **Implementation:** StatsForecast::AutoARIMA with m=12, AICc selection (univariate).
- **Observed behavior:** Smooth, mean-reverting forecasts; wide intervals; **validation MAPE 29.15%, test MAPE 16.29%** → weaker than manual SARIMA and Naïve on test.
- **Interpretation:** Performance shifts across windows suggest **window sensitivity / underfitting**, not classic overfitting.

oud

2.1.1 Time Series Models: Parameters, Settings, and Pros/Cons

Key hyperparameters:

- **max_depth (default = 6):** Controls how deep each tree can grow. Deeper trees can model more complex patterns, but also risk overfitting. Note: In early versions of this study, I mistakenly assumed the default was 3 instead of 6.

- **min_child_weight (default = 1):** Minimum number of samples needed in a leaf node. Helps prevent the model from fitting to noise by skipping very specific splits.
- **gamma (default = 0):** Minimum loss reduction required to make a further split. Higher values make the model more conservative.
- **subsample (default = 1.0):** The fraction of training samples used to train each tree. Subsampling helps reduce overfitting.
- **colsample_bytree (default = 1.0):** Similar to subsample, but applies to features. Helps diversify trees by training on random subsets of columns.
- **learning_rate (default = 0.3):** Determines the step size at each boosting round. Lower values mean slower learning, but often better generalization.
- **n_estimators (default = 100):** Total number of trees used. Can be increased if using a low learning rate.
- **Regularization terms:** reg_lambda (L2) and reg_alpha (L1) help prevent overfitting by penalizing complex models.
- **objective and eval_metric:** For regression tasks, these were set to "reg:squarederror" and "rmse" respectively.

Settings used in this study:

- **Default config:** Initially tested using standard defaults. However, early experiments mistakenly used max_depth = 3 under the assumption it was default.
- **Manual tuning:** Later runs tested combinations like max_depth = 2–4, learning_rate = 0.1, subsample = 0.8, and colsample_bytree = 0.8. This helped reduce overfitting, and performed better in validation.
- **Grid search:** A full grid search across 81 combinations found the best setup to be max_depth = 3, min_child_weight = 1, learning_rate = 0.3, and all others kept at 1.0. This configuration generalized better than manual tuning during testing.

Pros for EV forecasting:

- Can capture non-linear relationships, especially useful when EV demand depends on multiple external factors (policy, behavior, etc.).
- Built-in regularization helps avoid overfitting even with many features.
- Trains fast and outputs feature importance scores (which we later explain using SHAP).
- Doesn't require stationarity—can model trends and seasonality if features like lags or time indicators are provided.

Cons for EV forecasting:

- Requires tuning—bad parameter choices easily lead to overfitting.
- Less interpretable than ARIMAX; not suitable if you want a model with clear, explainable coefficients.

- Doesn't handle time structure natively—lags and other engineered features are needed to model seasonality or trends.
- May underperform with limited training data due to its complexity.

2.1.2 ARIMAX (AutoRegressive Integrated Moving Average with Exogenous Variables)

Key parameters:

- **order = (p, d, q):** ARIMA structure. We used (1,1,1)—one autoregressive term, one difference (to remove trend), one moving average term.
- **seasonal_order = (P, D, Q, s):** We didn't use this (kept as default (0,0,0,0)) since the dataset showed limited clear seasonality.
- **exog:** External variables like macroeconomic indicators or behavioral metrics were included.
- **enforce_stationarity / enforce_invertibility:** Both kept at default True to ensure stable predictions.

Settings used in this study:

- Used Statsmodels' SARIMAX implementation.
- Feature sets included both lagged values and external regressors like consumer confidence or fuel prices.
- Stationarity was enforced; trends were handled by first differencing.

Pros for EV forecasting:

- Specifically built for time series, so it can model trends, lags, and (optionally) seasonality well.
- Very interpretable. Each parameter and external variable has a clear role (e.g., AR(1) shows effect of previous month, coefficients show effect of external signals).
- Works well with limited data and provides forecast intervals (confidence bounds).

Cons for EV forecasting:

- Assumes linearity—non-linear effects (e.g., threshold effects from policy incentives) may not be well captured.
- Needs differencing and tuning of (p,d,q) values, which can be time-consuming and sensitive to errors.
- Struggles with many or highly correlated external variables; doesn't handle multicollinearity well.
- Requires forecasts for exogenous variables if doing multi-step prediction, which limits its practical use in some settings.

2.1.3 Random Walk (Naïve Baseline Model)

Concept and configuration:

The Random Walk model is a simple baseline that assumes the best prediction for the next period is just the most recent observed value. In other words, if EV sales in January were 20,000, the forecast for February is also 20,000.

- It has **no tunable parameters**, no training phase, and no assumptions about trends or seasonality.
- Technically, this is equivalent to an ARIMA(0,1,0) model without drift.

Settings used in this study:

- For each forecast, the model simply repeats the last known actual value.
- This approach was used as a **null model**: a minimal benchmark to determine whether more complex models actually add value.

Pros for EV forecasting:

- Extremely simple and fast: no training required.
- Surprisingly effective for short-term forecasts, especially in periods where sales are relatively stable.
- Automatically adjusts to recent changes — for example, if there's a sudden jump in demand, the model immediately shifts to that level for future steps.
- Doesn't require any preprocessing (no differencing, no lag creation, etc.).

Cons for EV forecasting:

- Ignores all patterns: it doesn't learn from trends, seasonality, or external factors.
- Carries noise forward: if the most recent value is an outlier, it will be copied into future forecasts.
- Poor for multi-step forecasts: it just repeats the last value indefinitely, which is often unrealistic.
- No interpretability or insight — it provides no explanation for what's driving demand.

Role in this study:

Despite its simplicity, the Random Walk model achieved a competitive MAPE of 0.1338 on the validation set. This shows that recent EV sales alone are highly informative. However, it offers no transparency or decision support value, which limits its usefulness for stakeholders who need interpretable forecasts.

3 D1 Feedback

3.1 Change list

Feedback	Plan (green is done)
A1>b1	B2:
(-) With only 158 observations in your training data, I would put more effort in simpler models such as ARIMA before moving to more complex models such as XGBoost.	Used sarimax instead of xgboost: See: {sec: Model Selection}
(+/-) You should have argued why other time series models, for example LSTM, are not applicable for this problem (hint: dataset size)	See {sec:Alternatives considered (XGBoost, LSTM)}
(-) A random walk is not a usual null model for timeseries data, more typical would be to leave out the random term (sales next month = sales this month) now you are just arbitrarily adding some error to your null model making your own models look more attractive. Also, you do not mention how large the added random noise is, from figure 4 I estimate it is quite tiny.	Used different benchamr model see sec:benchmark

(-) There is a strong 12-month autocorrelation in your data, you could have considered using seasonal ARIMA models instead of including lagged features.	{sec:strong 12-month autocorrelation}
(+/-) You select MAE, RMSE, MAPE and R^2 as evaluation metrics, focusing on MAPE but there is an exact monotonic relationship between RMSE and R^2, so these two measure the same thing.	See sec:metrics
(-) For the ARIMAX model you state: "However, its assumptions of linearity and stationarity may limit its ability to generalize", typical practice for fitting ARIMAX models is to transform (Box-Cox dtransformations and differencing) your data to make it stationary.	{sec:datatransformation}
B3	
(+/-) You mention the 2015 crisis, and the covid pandemic have effects seeable in your data. Your	Added this to the discussion {sec:discussion}

models might have benefited if you included indicator variables for these special events.	
(-) Nowhere in your text do you describe the total number of features in your data, I had to count the number of rows in figure 15 to see it is 28.	Added it to {sec: Implementation Details}
(-) Figure 15 shows many of your features are strongly correlated, linear models such as ARIMAX perform very badly on data with strong multicollinearity, so some feature selection would have been desirable. Many non-linear models suffer from the same issue as it provides extra degrees of freedom to overfit (although I'm not sure if this is the case for XGBoost).	Dropped features see {sec: datacollection}
(-) With 158 observations in your training data and 28 features, most models will be susceptible to overfitting, so some feature selection would have been desirable.	Dropped features see {sec: datacollection}

<p>(-) You write: “six monthly lag features were added to the model. This choice was guided by domain knowledge on sales cycles and supported by the ACF/PACF plots in Figure 17, which show strong autocorrelation within the first six lags.”</p> <ul style="list-style-type: none"> - This is not supported by the ACF plot, only by the PACF plot. - I wouldn’t call it a “strong autocorrelation”, it’s barely significant and small compared to the 1 month and 12 months partial autocorrelations. - Considering there is such a strong 12 months autocorrelation, why did you only include 6 months lags? 	<p>See {sec:baseline}</p> <p>Used acf pacf plot properly</p>
<p>(-) You write: “the dataset was cleaned, missing values addressed” what cleaning steps did you do? How many missing values where there, and how did you address this? What is the cause of missing data?</p>	<p>The dataset was cleaned by aligning all dates and ensuring no missing values remained in the final version. The only missing values occurred in the consumer sentiment dataset from CBS, which originally reported monthly but later switched to quarterly updates. This affected four indicators:</p> <p>Is explained here \label{sec:dataset-assembly}</p>

B4 iteratief process	
(-) No hyperparameter choices for the ARIMAX model are given. What are they and how did you choose them? If you used some automatic parameter tuning algorithm (for example the Hyndman & Khandakar algorithm), that should be mentioned. Also, note that in my (Robert) experience the H&K algorithm yields fast, but subpar results to manually tuning the model yourself through an ACF/PACF analysis. If you didn't do any parameter tuning and left them at defaults (which I assume are all 0) then you didn't fit a ARIMAX model but a linear regression, and it's unsurprising that the model underperforms... A properly fit ARIMA model will always be able to reproduce a random walk (as a random walk is an ARIMA(0,1,0) model) so the fact that your ARIMA performs worse is a clear sign	Auto arima was done here \{sec:Autoarima\} And manually tuning done here {sec:baseline}

something went wrong.	
(-) The choices made are rather weakly motivated.	Increased motivations bu following hyndmans book \cite{hyndman_rj_athanasopoulos_g_garza_a_challu_c_mergenthaler_m_olivares_kg_forecasting_2025}
	C1,C3,D1
	C2 Evaluieren
(-) I am not sure how useful a SHAP analysis is when your model does not outperform a null model. I certainly would not discuss the explanatory power of features when nothing is being explained.	Changed method, tested each exog feature at a time see sec: result
(-) Since your models do not perform better than a null model I would not conclude that: <i>"The findings of this study suggest that short-term forecasting of EV sales in the Netherlands can be performed effectively ..."</i>	Changed conclusion sec: conclusion
(-) Chapter VII (Results), Table XI: 1) In this table Dutch feature names are used. The Dutch feature names should be translated to English names	Changed to enhlisch and defined features more clerly {sec: Implementation Details}
(-) The most important feature is "Elektrische auto". It is not clear how this feature is defined. The same applies	

to other features.	
(-) The appendices have no numbers. As a consequence, it is not possible to refer in the text to the appendices.	Changed appendixes }
D2	
(-) Formatting of tables I and II results in very ugly page layouts.	Moved tables to appendix \tab{requirements_overview} and \label{tab:evaluation_methods}
(-) Basic statistical concepts taught in any first-year statistics course (for example, Pearson correlation) do not need to be extensively defined in formulas.	Removed formula for mape
(-) Some references are broken (pg. 11 "(see ??").	Fix broken references

Generally, I addressed feedback by documenting meeting notes immediately after each session. Based on these notes, I created a to-do list with specific items to add or revise in the report. I also left comments directly in my thesis draft, linking feedback points to the relevant sections.

For the final version, I plan to include a table summarizing all feedback points, whether they were implemented (yes/no), and any relevant remarks.

3.1.1.1 Feedback 100%

Feedback	Plan (green is done)	Sec
A1>b1		
B2:		
(-) With only 158 observations in your training data, I would put more effort in simpler models such as ARIMA before moving to more	Either check if I can use RDW daily updated dataset or continue with CBS dataset, however, lose a lot of features (max 1 or 2)	C. Data split Strategy: Walk-Forward Validation

complex models such as XGBoost.		
(+/-) You should have argued why other time series models, for example LSTM, are not applicable for this problem (hint: dataset size)	Add explanation in Model Selection Rationale/Discussion	E. Model Selection Rationale
(-) A random walk is not a usual null model for timeseries data, more typical would be to leave out the random term (sales next month = sales this month) now you are just arbitrarily adding some error to your null model making your own models look more attractive. Also, you do not mention how large the added random noise is, from figure 4 I estimate it is quite tiny.	Rushed to use a null model, started with LR but then read a paper about RW, rushed to use that but reflecting back at it clearly wasn't, will remove random walk. (hebben van nulmodel was geen probleem maar het toevoegen van random noise wel voor prediction) → later kijken of ik random noise zou kunnen weglaten	Remove from model selection, results and conclusion
(-) There is a strong 12-month autocorrelation in your data, you could have considered using seasonal ARIMA models instead of including lagged features.	First made a decision based on 6 months because wanted to do 6 months prediction, however results were not acceptable, so went to 3 months, keeping the 6 month lag, last minute recognized that I give clear explanation on the 6 month lag. Also based on https://machinelearningmastery.com/xgboost-for-time-series-forecasting/	E. Model Selection Rationale
(+/-) You select MAE, RMSE, MAPE and R^2 as evaluation metrics, focusing on MAPE but there is an exact monotonic relationship between RMSE and R^2, so these two measure the same thing.	I will remove R2	D. Evaluation Metrics And results, conclusion
(-) For the ARIMAX model you state: "However, its assumptions of linearity	Use first-order differencing on EV sales and other time series to remove trends.	B. Dataset Structure and Feature

<i>and stationarity may limit its ability to generalize", typical practice for fitting ARIMAX models is to transform (Box-Cox transformations and differencing) your data to make it stationary.</i>	Original data: [100, 105, 110, 115] First-order difference: [105-100=5, 110-105=5, 115-110=5] → [5, 5, 5] Optionally apply a Box-Cox transformation to stabilize variance.	Compositio
(-) On pg. 12 you suddenly mention a LR model, this term is never defined or discussed, what is this?	Yes, big mistake, used LR first then switched to RW last minute, didn't see the mistake while reading	Feature Scaling for ARIMAX Pg 12
B3		
(-) Figure 2 text states: "Keywords and carsales" but this figure does not display any information about carsales.	First showed a graph of car sales, but because of the difficulty with my split, order I was writing my paper I didn't want to show the car sales because then I would've showed parts of testset.	Fig. 2: Correlatie Matrix: Keywords and carsales pg 8
(+/-) You mention the 2015 crisis, and the covid pandemic have effects seeable in your data. Your models might have benefited if you included indicator variables for these special events.	Could add indicator like crisis2015: 1 if month is during 2015 downturn, else 0 covid_period: 1 for months from ~March 2020 to ~mid-2022, else 0	Dataset overview
(-) Nowhere in your text do you describe the total number of features in your data, I had to count the number of rows in figure 15 to see it is 28.	will add total number of features: (after ive reduced them)	Abstract, and dataset overview
(-) Figure 15 shows many of your features are strongly correlated, linear models such as ARIMAX perform very badly on data with strong multicollinearity, so some feature selection would have been desirable. Many non-linear models suffer from the same issue as it provides extra degrees of freedom to overfit (although I'm not sure if this is the case for XGBoost).	I wanted to select features afterwards with shap, but might be a good idea to first drop features based on Variance Inflation Factor (VIF): Drop features with VIF > 5 or 10 Or Correlation thresholding: Drop one of any pair with correlation > 0.9 Also, will focus more on the use of regularization on xgboost:	E. Model Selection Rationale Fig. 15: full Corrolation Matrix
(-) With 158 observations in your training data and 28 features, most models	In this study, features were pre-screened using correlation thresholding/VIF to reduce redundancy before training ARIMAX.	

will be susceptible to overfitting, so some feature selection would have been desirable.	Argumentatie geven voor de gekozen features Check validatie etc	
(-) You write: “ <i>six monthly lag features were added to the model. This choice was guided by domain knowledge on sales cycles and supported by the ACF/PACF plots in Figure 17, which show strong autocorrelation within the first six lags.</i> ” - This is not supported by the ACF plot, only by the PACF plot. - I wouldn’t call it a “strong autocorrelation”, it’s barely significant and small compared to the 1 month and 12 months partial autocorrelations. - Considering there is such a strong 12 months autocorrelation, why did you only include 6 months lags?	<p>It should be noted that the interpretation of the ACF/PACF plots was done under time pressure and with limited prior experience. Future work could benefit from a more rigorous lag structure analysis using information criteria (e.g., AIC/BIC) or automated selection methods.</p> <p>Might only use 1 to 3 months lag</p>	
(-) You write: “ <i>the dataset was cleaned, missing values addressed</i> ” what cleaning steps did you do? How many missing values where there, and how did you address this? What is the cause of missing data?	The dataset was cleaned by aligning all dates and ensuring no missing values remained in the final version. The only missing values occurred in the consumer sentiment dataset from CBS, which originally reported monthly but later switched to quarterly updates. This affected four indicators: <i>Consumentenvertrouwen, Economisch klimaat, Koopbereidheid, and Grote aankopen</i> . To maintain a consistent monthly structure, I applied linear interpolation followed by backfilling, resulting in a complete, gap-free time series. These steps were implemented in the preprocessing code and documented in the project logbook, but were unfortunately not described in the final report — an oversight I will correct in the next version.	Check ipynb pre processing!
B4 iteratief process		
(--) No hyperparameter choices for the ARIMAX model are given. What are	For this version of the project, the ARIMAX model was implemented with a fixed order of (1, 1, 1) . No systematic hyperparameter tuning	

<p>they and how did you choose them? If you used some automatic parameter tuning algorithm (for example the Hyndman & Khandakar algorithm), that should be mentioned. Also, note that in my (Robert) experience the H&K algorithm yields fast, but subpar results to manually tuning the model yourself through an ACF/PACF analysis. If you didn't do any parameter tuning and left them at defaults (which I assume are all 0) then you didn't fit a ARIMAX model but a linear regression, and it's unsurprising that the model underperforms... A properly fit ARIMA model will always be able to reproduce a random walk (as a random walk is an ARIMA(0,1,0) model) so the fact that your ARIMA performs worse is a clear sign something went wrong.</p>	<p>was applied, and I did not use the Hyndman & Khandakar algorithm or conduct a full ACF/PACF analysis to optimize the (p, d, q) parameters. The goal was to include ARIMAX as a comparative benchmark rather than a fully optimized model.</p> <p>I acknowledge this was a limitation. Without proper tuning, the model likely behaved closer to a linear regression with differencing, rather than a well-specified ARIMAX. This may explain its underperformance — especially considering that a well-fit ARIMA(0,1,0) can at least match a random walk.</p> <p>I did, however, apply standardization to the exogenous features prior to fitting, in line with best practices, to improve coefficient stability. This was implemented via a custom ARIMAXWrapper class.</p> <p>In future work, particularly in the next paper where ARIMAX will be the primary model of focus, I will include a detailed selection process for the (p, d, q) parameters using either AIC-based grid search, ACF/PACF inspection, or automated methods like auto_arima. This will ensure that the ARIMA structure is properly specified and benchmarked fairly.</p>	
<p>(-) XGBoost: With 21 observations in your validation set, and 81 parameter choices in your gridsearch you are pretty much guaranteed to start overfitting on your validation data. This later shows in your test results.</p>	<p>Given that I can't increase the number of observations, what would be a better approach to tune the model while minimizing overfitting on the validation set?</p> <p>I tried to set up a transparent and repeatable optimization pipeline (with fixed walk-forward splits and performance logging), but I'm open to suggestions for improving this — especially for small time series datasets.</p> <p>Meer aandacht besteden aan betere hyperparameters (niet te veel opties in hyperparameters/ iteraties)</p> <p>Om overfitting te verkomen , aantal features iteraties verkleinen</p>	
<p>(-) XGBoost: You mention which ranges of values you</p>	<p>Mistake due to time pressure</p>	

try in your grid search for all parameters except for min child_weight.		
(-) XGBoost: You find an optimal learning rate of 0.3, which is on the edge of your lr grid [0.05, 0.1, 0.3], this implies an even higher learning rate may have better results.	Yes but a higher learning rate would have been more prone to overfitting	
(-) The choices made are rather weakly motivated.	Motivation was done under extreme time pressure, will rewrite it	
	C1,C3,D1	
	C2 Evaluieren	
(-) When discussing the results of your null model, you state: " <i>The predicted series (orange) often tracks the actual values (black) with a one-step delay, highlighting the model's inability to anticipate rapid shifts or volatility.</i> " Here, "often" should be "always", and this sentence is rather stating the obvious...	Uitleg verbeteren: The random walk model always predicts the next value as equal to the previous one, resulting in a one-step delay. While this mirrors trend direction, it inherently fails to anticipate sharp shifts, limiting its relevance for policy-driven changes in EV sales	
(-) ARIMA models can also give a confidence interval for their predictions, best practice is to also show this in evaluations.	Confidence interval toevoegen bij de plot	
(-) I am not sure how useful a SHAP analysis is when your model does not outperform a null model. I certainly would not discuss the explanatory power of features when nothing is being explained.	For the retake feature importance will be performed based on model result, however with a bad/negative result feature explanation will be left out. Was explained based on shap value,	
(-) Since your models do not perform better than a null model I would not conclude that: " <i>The findings of this study suggest that short-term forecasting of EV sales in</i>	Improve explanation The random walk model always predicts the next value as equal to the previous one, resulting in a one-step delay. While this mirrors trend direction, it inherently fails to anticipate sharp shifts, limiting its relevance for policy-driven changes in EV sales.	

<i>the Netherlands can be performed effectively ...”</i>		
(-) In your discussion you write: “ <i>while the Random Walk baseline achieved surprisingly strong results (MAPE = 0.1338), it lacks interpretability</i> ”, a random walk is your most interpretable model!	update the discussion to reflect that: While the random walk model lacks explanatory variables, it remains highly interpretable in structure — relying solely on previous values — and provides a strong baseline in autoregressive domains.	
(-) Chapter VII (Results), Table XI: 1) In this table Dutch feature names are used. The Dutch feature names should be translated to English names	Dataset colommen in het engels. → duidelijk de betekenisernaast zetten/ waar het vandaan komt. (groeperen) The feature ‘Electric car’ is a Google Trends search interest score based on the Dutch query ‘elektrische auto’, aggregated monthly and scaled from 0-100.	
(-) The most important feature is “Elektrische auto”. It is not clear how this feature is defined. The same applies to other features.		
(-) The appendices have no numbers. As a consequence, it is not possible to refer in the text to the appendices.	Appendix beter categoriseren (1 voor figuur x etc)	
D2		
(-) Formatting of tables I and II results in very ugly page layouts.	Oorzaak double column, ga evt over naar single column Of landscape gebruiken	
(-) Basic statistical concepts taught in any first-year statistics course (for example, Pearson correlation) do not need to be extensively defined in formulas.	Formules verwijderen	
(-) Text in figures (title, axis labels, etc...) is too small to read	Groter maken door single column te gebruiken	
(-) Some references are broken (pg. 11 “(see ???”).	Fix broken references	

Feedback 70%

3.1.1.2 Feedback/changelist 70

Feedback	Plan (green is done)	Sec
1. Sub research question 2 is not clear to the reader. This sub research question has to be reformulated.	Rewrite sub-question: Which digital search trends related to EVs are relevant for forecasting EV sales?	Problem
2. I do not understand Figure 2.	Delete Figure 2, its explained in the text	
3. Page 10, top of right column: You have to report the importance of the separate features. What is the difference between your approach, i.e. LOFO feature performance approach, and the SHAP-based feature attribution approach?	Was debating between LOFO/SHAP, wasn't sure which one I was going to add, so I added both. Going forward with SHAP	(H)Dataset Assembly and Feature Framing
4. What is the number of records in the Google Trend dataset?	Clarify: all records cover 219 months.	IV. IMPLEMENTATION DETAILS
5. What is the best model? What features are included in this model? What are the time lags? What is the performance of this model? And what is the seasonal pattern in this model?	Add model summary: features, lags, metrics, seasonal pattern.	V. MODEL DEVELOPMENT AND EVALUATION
6. Page 11 "D. LOFO Results with Tuned Model": To which model do Figure 8 and Figure 9 correspond? What features are included in this model? What are the time lags? What is the performance of this model? And what is the seasonal pattern in this model?	Results were added as last, so it was a little messy. In 100% report will do quantitative analysis and be clearer in corresponding, also will add more details on dataset explanation	VI. QUANTITATIVE RESULTS AND ANALYSIS
7. You should use other approach rather than only XGBoost.	As your feedback suggested, now added LR, ARIMA & ARIMAX	
8. You should use metrics such as MAPE rather than other metrics you have chosen.	Switch to MAPE as main metric and add explanation	
9. Discussion: Compare your results/findings with those from the literature.	Will rewrite discussion, however might be challenging due to different approach/ metrics/ after call with Marcio wasn't needed anymore, could do	
10. The text in the Tables and	Translate all tables/figures	

Figures have to be in the English language.	to English.	
11. Future research: Which alternative model/approach might give better results than the model/approach you chose?	Suggest LSTM, GRU, Decision Trees or Linear Models.	
12. The reference list has to be checked whether all references are described according to the APA guidelines. /Check broken links in the reference	I marked all broken lines with red, will add working link/DOI Only books, and research from kaya did not have doi	
Feedback Rubric		
13. A: Strong statements that are not backed by proper sources.	Will go through paper to make sure all (strong) statements are backed by proper sources	Full page
14. The exploration of alternatives, however, was very limited and too solution oriented. More specifically, it directed the reader to the envisioned solution space using tabular data and predefined models such as XGBoost.	Broaden discussion: added Linear, ARIMA, ARIMAX. Will briefly mention other options (e.g. SARIMA, LSTM) and justify selection.	
15. There are no discussions on the level of automation and the benefit to society if good predictions are delivered or the negative impact in case of poor predictions. It also does not describe that although some correlations can be detected, this is not an indication of causality, that the scenario is dynamically changing, and that the model should be constantly monitored.	Will add level of automation. Briefly reflect on long-term societal impacts like infrastructure costs, EV charging access, Ev parking places and subsidy allocation. Add note on lack of causality, dynamic context, and need for continuous model monitoring.	
16. you should at least hypothesise how your model could be used and under which conditions, such as sketching a UI for your target audience. I recommend having a dedicated section in the discussion chapter.	I made a Prototype, see Figma will also add ai design principles	

17. You should justify adding a more complex model for a slightly better improvement in the prediction,	Clarify why XGBoost was used—refer to predictive improvement, seasonal fit, and ability to capture complex patterns.	
18. Address mismatch between the problem (forecasting time series) and the solution (tabular model).	Acknowledge mismatch. Justify tabular approach by explaining use of lags/time features. Note limitations vs. true time series models.	
19. Conduct a qualitative analysis—it might be the case that there is no causality but only co-occurrence, especially with variables like fuel prices.	Will add quantitative analysis as extra section in results	
20. B: Req: However, using too many metrics as a basis for comparison leads to problems: What if they do not agree? Be very clear on what you will do in such situations.	Will only use MAPE as compass metric	
21. Don't over-focus on one model. With small datasets, simpler models often perform just as well or better	Added baseline models (LR, ARIMA, ARIMAX). Justify XGBoost based on improved accuracy and ability to capture nonlinear patterns.	
22. "XGBoost is not ideal without strong feature engineering. Start with simpler time series models like Linear Regression or ARIMA before moving to complex models."	Confirmed feature engineering for XGBoost (lags, time vars). Added simple models first. Justify switch to XGBoost with performance and interpretability.	
23. There is a contradiction: the report emphasizes transparency, but XGBoost is not inherently transparent. LOFO helps, but you should justify its use over more interpretable methods like SHAP.	Replace or complement LOFO with SHAP plots to better explain model behavior. Briefly justify choice by comparing interpretability and relevance to stakeholders.	
24. This is a minimal dataset (219 observations). Complex models need more data—start with simpler models to establish a solid baseline.	Addressed by adding LR, ARIMA, and ARIMAX as baselines. Use results to show whether complex models like XGBoost offer meaningful improvement. (but afraid this will)	

25. Dataset preparation is solid, but the Google Trends keyword selection lacks transparency. Clarify how translations were validated, why some keywords were discarded, and include the final list.	prompt and keyword list was added to logbook however for 100% will add appendix. Will briefly explain translation was validated via DeepL and Google Trends relevance checks.	
26. Basic tuning was mentioned, but no systematic optimisation strategy was applied. Use an approach like grid search or Bayesian optimisation to explore hyperparameters more thoroughly.	Added Bayesian optimisation, will explain this in my rapport	
27. C: Good that you address bias and transparency, but you should better explain model behaviour using SHAP or similar: LOFO was introduced too quickly. Also, discuss how to detect and handle performance decay over time with non-stationary data.	Will add SHAP to improve model transparency and explain feature impact more clearly. Briefly discuss how to monitor model drift and detect performance decay over time.	
28. Too much focus on R^2 —real-world forecasting relies more on absolute error metrics like MAE or MAPE. Also include prediction intervals to assess model uncertainty.	Will focus on MAPE as main metric. Add prediction intervals using ARIMA's confidence bounds and explore residual-based or bootstrapped intervals for XGBoost and Linear Regression.	
29. Adding more features doesn't seem to help much—consider testing with noise instead. Performance is acceptable for 1-month forecasts but breaks down at 6 months.	ADD IT TO FUTURE WORK	
30. A qualitative analysis is needed, check specific data points and outliers. Some patterns, like with fuel prices, may reflect coincidence rather than causality.	Will include brief qualitative analysis in discussion, highlighting key data points and potential non-causal patterns (e.g., fuel prices).	
31. The document is clear and well structured, but several points need improvement:	Plan on how to fix mentioned above planm	

review strong claims, clarify processes like translation and keyword selection, and strengthen the focus on time series evaluation.		
32. D2: be careful when distinguishing causation from correlation. This distinction must be made more explicit in your discussion.	Will add explicit statements in discussion to distinguish correlation from causation, with examples (e.g., fuel prices). <i>"Although variable X correlates with EV sales, this does not imply a causal relationship."</i> <i>"The model detects associations, not underlying causes. For example, fuel prices may coincide with sales shifts, but other unmeasured factors could drive both."</i> <i>"Caution is needed in interpreting these results as causal, especially in the absence of controlled experiments."</i>	
Points made earlier		
33. splitting the literature review into: a. Similar <i>problem domains</i> (e.g., forecasting in finance or energy). b. Similar <i>technical approaches</i> (e.g., time series regression, model evaluation techniques).		34.
35. Emphasis on explaining model behavior clearly, as if advising a public organization.	37.	
36. Consider how a stakeholder would interpret the results and which decisions could follow.		
38. The "Klimaatbeleid Dashboard" is not considered state-of-the-art but may be included as an example of a <i>potential application</i> .	40.	
39. State-of-the-art should focus on recent forecasting models applied to time series regression problems.		
41. Requirements should be made SMART (Specific, Measurable, Achievable, Relevant, Time-bound).	42.	

3.1.2 20 May – Meeting with Kees

Theme: Scope, Methods, and Interpretation

- Clarify what changed in your research and what sets it apart.
- Sharpen your core problem statement and the research gap.
- Add a graph to explain peaks/dips in EV sales using common-sense reasoning.
- Implement Leave-One-Feature-Out (LOFO) analysis.
- Add SHAP values for model interpretability.
- Review requirements critically—are they measurable, justified?
- Test different lag structures for EV sales (e.g., 1–6 months).

3.1.3 14 May – Meeting with Marcio

Theme: Features, Model Justification, Data Handling

- Address multicollinearity in features; use smarter feature selection.
- Justify model choice: why XGBoost and CNN, and not alternatives like LSTM?
- Provide clear descriptions for all features in the dataset.
- Use proper time-series split to avoid data leakage.
- Improve the quality of literature sources (source 8 is strong—use it more).

3.1.4 13 May – Written feedback from Marcio

Theme: Structure, Claims, Literature, Framing

- Refine title – current version is unclear and too narrow.
- Back all strong claims with references.
- Avoid overreliance on Google Trends in framing; keep question broad.
- AIDA model is inconsistently used—consider removing it.
- Improve clarity in research gap and problem statement: who is it for, and why does it matter?
- Consolidate repetitive content in gap sections.
- Make requirements specific and tailored to your target audience.
- Add exploratory data analysis (EDA) on Google Trends and macro variables.
- Improve methodology clarity: how features are selected, why XGBoost, etc.
- Reconsider use of TF-IDF/YAKE—may be outdated.

3.1.5 12 May – Feedback from Daphne

Theme: AIDA, Model Explainability, Relevance

- AIDA model may not match your data and scope—remove it.
- XGBoost isn't truly interpretable—rephrase your claims.
- Avoid repeating gap/problem discussions—streamline.
- Clarify why you didn't collect richer attention/desire signals (e.g., social media, subsidies).

3.1.6 07 May – Go/No-Go Check

Theme: Thesis Scope, Learning Outcomes

- Scope is too broad; problem definition is unclear.
- Refocus on one stakeholder group (e.g., municipalities).
- Make A1–A5 explicit in the thesis.
- Prototype is out of scope? Then clearly state that.
- Keyword selection (NOS) lacks rigor—consider semantic clustering.
- Model fails in certain months (e.g., Feb/Jul); address time variable issues.
- Add monthly dummies instead of time index.
- Hypothesis (NOS sentiment) is likely falsified—pivot to Google Trends.
- Provide rationale for Google Trends: reproducibility, fewer noise issues.

3.1.7 06 May – Meeting with Prof. Van Montfort

Theme: Model Bias, Time as Variable

- Predictions are consistently lower—suggests bias.
- Don't normalize for seasonality at this stage.
- Include time as explanatory variable (e.g., month dummies or month count).
- Avoid redundant aggregations (min, max, avg fuel prices)—choose one.
- OK to proceed without external client—just explain it briefly in your report.

3.1.8 17 April – Planning & Topic Shift

Theme: Tools, Metrics, Planning

- Use semantic tools like Jina Embeddings instead of RobBERT.
- Improve article clustering via cosine similarity and thresholding.
- Clearly justify article selection and filters.
- Define model performance goals (precision, interpretability).

- Track historical policy context and explain patterns in the data.

In the week of May 19, I provided Marcio with a detailed change list showing his feedback and the corresponding adjustments I made.

3.1.9 change list

Legend:

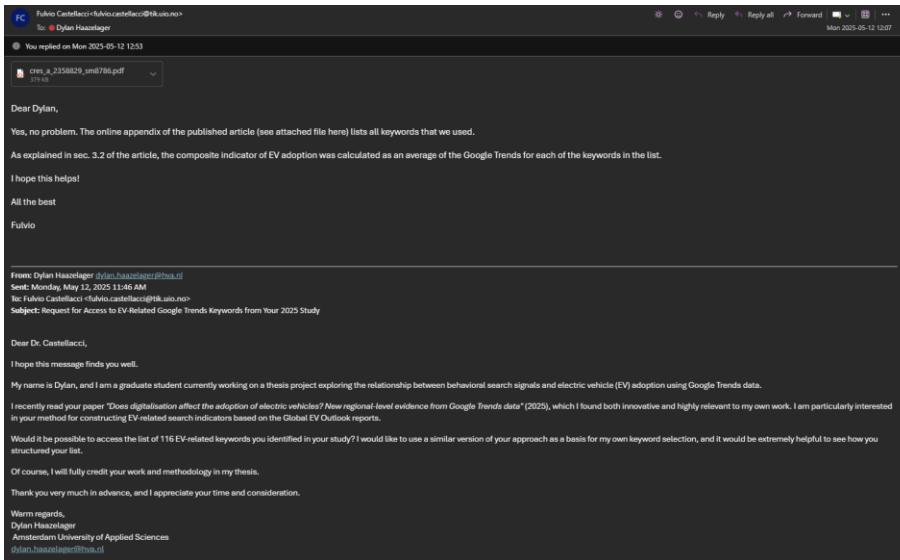
Not changed yet
Changed but still work to do
Changed and happy

Feedback Area	Changed?	Follow-up
Title & Scope • The title was vague and overemphasized “grouped feature importance.” • It needed to better reflect the study’s aim.	✓ Changed from “ <i>Uncovering the Drivers of Car Sales...</i> ” to “ <i>Identifying Data-Driven Predictors of Electric Vehicle Sales in the Netherlands.</i> ”	The title now better reflects the main research question. I may refine it further in the final version for full alignment with the objective.
References & Citation Quality • Missing DOIs and incomplete references. • Literature quality was weak. • Source 8 should be used more effectively.	✓ Used source 8 (Gallagher & Muehlegger) to reinforce the problem statement. △ Some DOIs still missing.	I plan to update all references this week, categorize peer-reviewed vs grey literature, and revise source 5 as needed.
Claims Without Evidence • Some claims lacked sources or statistical backing. • Avoid premature causal language.	△ Still present in parts of the paper.	I will refine unsubstantiated claims and add citations where necessary in the next revision.
AIDA Methodology • Introduced but not followed through. • Unclear relevance.	✓ AIDA was removed.	I originally included it based on prior literature but replaced it with better-grounded reasoning on behavioral data.
Problem Statement • Lacked specificity and	✓ Stakeholders added (“policymakers, NGOs, public	I will integrate LO-A explicitly into the problem

alignment with Learning Outcome A. • Needed clarity on stakeholders, relevance, and consequences.	institutions"). △ Still missing clear articulation of consequences of poor forecasting. △ LO-A alignment moved under 'Scope of Research'.	statement and add consequences such as infrastructure mismatch and subsidy inefficiency.
Research Question & Subquestions • Too narrow, too focused on one data source.	✓ Broadened to include macroeconomic, digital, and policy signals. ✓ Added subquestions.	I will update the final subquestion to more specifically address model performance across feature types.
Requirements • Vague and audience not defined. • Needed to be measurable and referenced in the discussion.	Based on intuition	I will revisit and refine requirements based on hypothetical stakeholder perspective and ensure they are linked back in the discussion section.
Data Description & Preprocessing • No EDA; TF-IDF/YAKE outdated; unclear feature handling.	✓ Added EDA and removed YAKE. ✓ Switched to using top correlated keywords. ✓ Described macroeconomic data more clearly. ✓ Moved keyword selection to Methodology.	No major follow-up needed here; only continue refining EDA plots and documenting variable construction in the appendix or logbook.
Methodology • No clear justification for model choices. • Poor explanation of feature transformations. • Spatial/temporal context lacking.	✓ Justified XGBoost. ✓ Tried 5 models and selected XGBoost based on performance. ✓ Temporal structure clarified using lag features and walk-forward validation.	Will finalize and document SHAP/LOFO analysis to reinforce model transparency.
Model Evaluation • MAE alone is insufficient; missing rationale. • Needed comparison to traditional regressors.	✓ Added MAE, RMSE, MAPE, R ² , and MAPA. △ Traditional regressors referenced but not fully described.	I will explicitly name and briefly compare traditional methods (e.g., ARIMA, Linear Regression) in the evaluation or discussion section.
Interpretability vs Accuracy • Why is interpretability important? • Who is the audience?	✓ Clarified that the audience includes public sector stakeholders. ✓ XGBoost selected for interpretability; SHAP/LOFO planned.	I will add one sentence explicitly connecting interpretability to policymaker usability in the discussion section.
Qualitative Analysis • Is this included? Will you reflect	△ Partially addressed. Mentioned in	I plan to add a brief qualitative reflection in

on why certain features matter?	results section planning.	the discussion on the policy implications of top predictive features.
Structure, Clarity & Redundancy• Redundancies and vague sentences need improvement.	⚠ Still in progress.	I will revise for conciseness, remove repeated references to car brand keywords, and improve clarity sentence-by-sentence this week.

4 Email and keywords:



Online appendices

Appendix A: List of keywords used for the Google Trends indicators.

Electric vehicles

Accessible chargers, accessible charging, accessible fast chargers, alternative fuels, alternative fuels infrastructure, battery electric vehicle, battery electric vehicles, battery electric vehicles and plug-in hybrid electric vehicles, battery swapping, cell electric vehicles, charging infrastructure, charging network, charging point, charging points., charging stations, demand for electric vehicles, deployment of electric cars, deployment of electric vehicles, electric bus electric buses, electric car, electric car market, electric car models, electric car purchase price, electric car sales, electric car stock, electric cars, electric cars subsidies, electric driving,, electric light duty, electric light-commercial vehicle, electric light-duty vehicle, electric lightduty vehicles, electric mobility, electric truck, electric trucks, electric two wheelers, electric vehicle, electric vehicle , electric vehicle adoption, electric vehicle batteries, electric vehicle battery, electric vehicle charging, electric vehicle charging infrastructure, electric vehicle deployment, electric vehicle fleet, electric vehicle initiative, electric vehicle initiative countries, electric vehicle market, electric vehicle markets, electric vehicle models, electric vehicle outlook, electric vehicle purchase price, electric vehicle sales, electric vehicle sales shares, electric vehicle supply, electric vehicle supply equipment, electric vehicle supply, equipment deployment, electric vehicle supply equipment outlets, electric vehicle uptake, electric vehicle volumes, electric vehicles, electric vehicles data, electric vehicles initiative, electric vehicles subsidies, emission vehicles, financial incentives electric cars, financial incentives electric vehicles, fuel cell electric car, fuel cell electric vehicle, , global electric vehicle, global electric vehicle fleet, global electric vehicles data, home charging, hybrid electric, hybrid electric vehicles, internal combustion engine cars, internal combustion engine vehicle, internal combustion engine vehicles, light-duty vehicle chargers, light-duty vehicle sales, new energy vehicles, number of electric cars, number of electric vehicles, plug in hybrid, plug-in hybrid electric vehicle, plug-in hybrid electric vehicle , plug-in hybrid electric vehicles, private chargers, private charging, public chargers, public charging, publicly accessible chargers, publicly accessible charging, purchase incentives electric cars, purchase incentives electric vehicles, purchase of electric cars, purchase of electric vehicles, purchase price electric car, purchase price of electric vehicle, purchase subsidies electric cars, purchase subsidies electric vehicles, sales share of electric cars, sales share of electric vehicles, share of electric cars, share of electric vehicles, smart charging, stock of electric cars, stock of electric vehicles, transition to electric, uptake of electric cars, uptake of electric vehicles, vehicle charging, vehicle electrification, zero emission vehicle, zero emission vehicles.

Digitalization

Smart Manufacturing

3D printing, Additive Manufacturing, Computer Numerical Control, Layered Manufacturing, Additive Fabrication, Rapid Prototyping, Direct Digital Manufacturing, Advanced Manufacturing, Collaborative Robotics, Robot Automation, Cobot, Cyber-Physical System, Augmented Reality, Computer Vision, Artificial Reality, Application Architecture, Application Programming Interface, Artificial Intelligence, Asset Administration shell, Batch Size, Big Data Analytics, Blockchain, Bluetooth Low Energy, Cloud Computing, Cloud Robotics, Cyber-Physical System, Cyber-

Physical Production Systems, Data Analyst, Data Architecture, Deep Learning, Descriptive Analytics, Digital Enterprise, Digital Supply Chain, Digital Twin, Edge Computing and Gateway, Enterprise Resource Planning, ERP system, ERP software, Customer relationship management, CRM, Extensible Markup Language, Flexible Manufacturing, Fog Computing, Apache Hadoop, SAP HANA, Human-Machine Interface, Industrial Internet Consortium, Industrial Internet of Things, Industrial Internet Reference Architecture, Industry 4.0, Machine to Machine Communication, Machine Learning, Manufacturing Execution System, Open System Interconnection Model, OPC Unified Architecture, Operational Technology, Predictive Maintenance, Production analytics software, Programmable Logic Controller, RAMI 4.0, Risk-based Planning & Scheduling, Remote Maintenance, SCADA, Supervisory Control and Data Acquisition, Social Mobile Analytics Cloud, Smart Factory, Virtual Reality, ZigBee, Fideltronik, Semeq, SMARTECH, WorkClout,

Bosch Rexroth AG, DXC Technology, Honeywell Industry, Honeywell Manufacturing, Parsable, Rockwell Automation, SICK AG, Siemens PLM, Autodesk, Autocad, Industry 4.0 Solutions SAP, Industrial Internet of Things SAP, Dassault Systemes, Epicor, Fishbowl Inventory, Infor, Thrive software, NetSuite.

Smart city

Autonomous vehicles, real time traffic data, smart traffic management, smart traffic control, smart traffic system, real time pollution data, real time energy data, City-as-a-Service, Hyperlocal data, Smart waste, smart street lighting, smart bins, smart drains, smart grid, Smart infrastructure, smart parking, Traffic adaptive lighting, Ubiquitous city, smart city applications, Urban data platform, smart buildings, Gunshot Detection Technology, City Transit Applications.

Smart Appliances

Smart plug, Amazon Alexa, Samsung Smart Home, Samsung SmartThings, SmartRules, Google Home, Apple Home, Danalock, MyQ, Ecobee, Honeywell Home, Honeywell Total Connect, Philips Hue, Wemo, Yale Connect, Insteon, Sentri all in one, Wink Hub, Alarm.com.

Communication

Skype, Microsoft Teams, Zoom, WhatsApp, whatsapp messenger, facebook messenger, msn messenger, facetime, instant messaging, messaging app, delete message, unsend, gmail, Hotmail.

Social Media

Myspace, friendster, facebook, instagram, tumblr, pinterest, tweeter, snapchat, friending, unfriend, share photo, share post, posted, user profile, untag, tweet, posting, hashtag, instagram like, likes, shares, follower, followers, unfollow, profile pic, cover pic.

Appendix B: Comparing Google Trends data with Eurostat data on the diffusion of EVs in

European countries.

Figure B1: Google Trends indicator of EVs adoption (left-axis) and Eurostat indicator on the share of EVs in each country (right-axis).

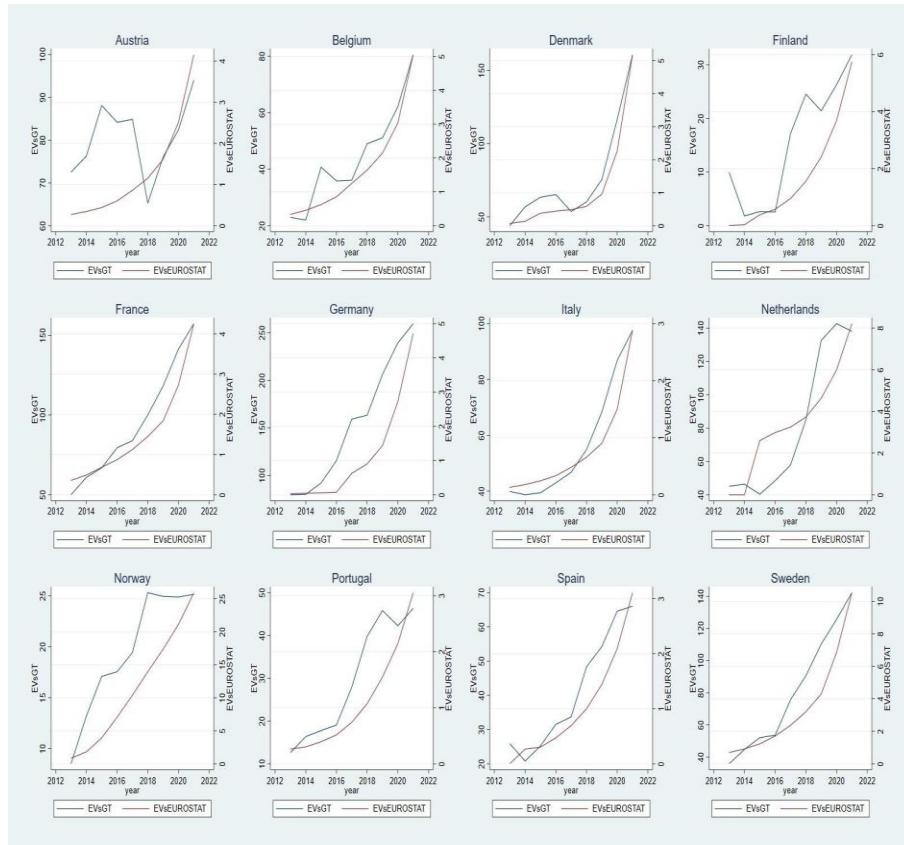
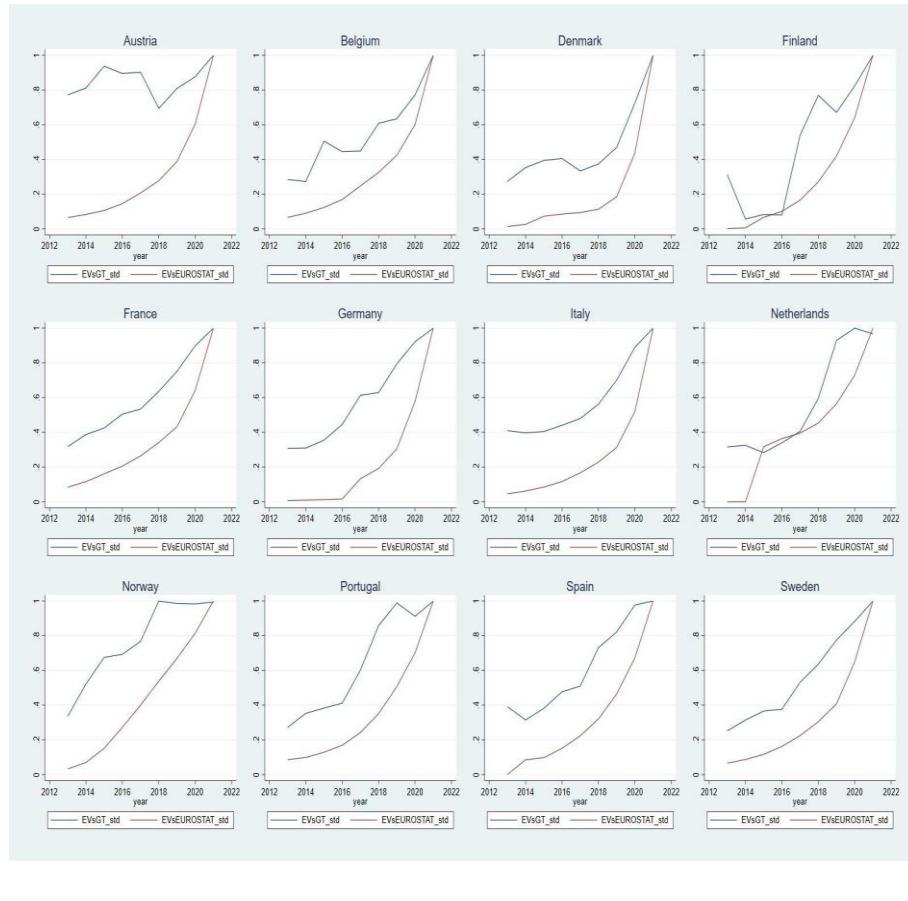


Figure B2: Google Trends indicator of EVs adoption and Eurostat indicator on the share of EVs in each country (Y-axis), both standardized using the same method.



5 Prompt used for translating the Key words from Dutch to English:

translate this to dutch: Accessible chargers, accessible charging, accessible fast chargers, alternative fuels, alternative fuels infrastructure, battery electric vehicle, battery electric vehicles, battery electric vehicles and plug-in hybrid electric vehicles, battery swapping, cell electric vehicles, charging infrastructure, charging network, charging point, charging points., charging stations, demand for electric vehicles, deployment of electric cars, deployment of electric vehicles, electric bus electric buses, electric car, electric car market, electric car models, electric car purchase price, electric car sales, electric car stock, electric cars, electric cars subsidies, electric driving,, electric light duty, electric light-commercial vehicle, electric light-duty vehicle, electric light duty vehicles, electric mobility, electric truck, electric trucks, electric two wheelers, electric vehicle, electric vehicle , electric vehicle adoption, electric vehicle batteries, electric vehicle battery, electric vehicle charging, electric vehicle charging infrastructure, electric vehicle deployment, electric vehicle fleet, electric vehicle initiative, electric vehicle initiative countries, electric vehicle market, electric vehicle markets, electric vehicle models, electric vehicle outlook, electric vehicle purchase price, electric vehicle sales, electric vehicle sales shares, electric vehicle supply, electric vehicle supply equipment, electric vehicle supply, equipment deployment, electric vehicle supply equipment outlets, electric vehicle uptake, electric vehicle volumes, electric vehicles, electric vehicles data, electric vehicles initiative, electric vehicles subsidies, emission vehicles, financial incentives electric cars, financial incentives electric vehicles, fuel cell electric car, fuel cell electric vehicle, , global electric vehicle, global electric vehicle fleet, global electric vehicles data, home charging, hybrid electric, hybrid electric vehicles, internal combustion engine cars, internal combustion engine vehicle, internal combustion engine vehicles, light-duty vehicle chargers, light-duty vehicle sales, new energy vehicles, number of electric cars, number of electric vehicles, plug in hybrid, plug-in hybrid electric vehicle, plug-in hybrid electric vehicle , plug-in hybrid electric vehicles, private chargers, private charging, public chargers, public charging, publicly accessible chargers, publicly accessible charging, purchase incentives electric cars, purchase incentives electric vehicles, purchase of electric cars, purchase of electric vehicles, purchase price electric car, purchase price of electric vehicle, purchase subsidies electric cars, purchase subsidies electric vehicles, sales share of electric cars, sales share of electric vehicles, share of electric cars, share of electric vehicles, smart charging, stock of electric cars, stock of electric vehicles, transition to electric, uptake of electric cars, uptake of electric vehicles, vehicle charging, vehicle electrification, zero emission vehicle, zero emission vehicles.

ChatGPT zei:

Here is the Dutch translation of your list:

Toegankelijke laders, toegankelijk laden, toegankelijke snelladers, alternatieve brandstoffen, infrastructuur voor alternatieve brandstoffen, batterij-elektrisch voertuig, batterij-elektrische voertuigen, batterij-elektrische voertuigen en plug-in hybride elektrische voertuigen, batterijwisselsystemen, elektrische voertuigen op celbasis, laadinfrastructuur, laadnetwerk, laadpunt, laadpunten, laadstations, vraag naar elektrische voertuigen, uitrol van elektrische auto's, uitrol van elektrische voertuigen, elektrische bus, elektrische bussen, elektrische auto, markt voor elektrische auto's, modellen van elektrische auto's, aankoopprijs van elektrische auto's, verkoop van elektrische auto's, voorraad van elektrische auto's, elektrische auto's, subsidies voor elektrische auto's, elektrisch rijden, elektrische lichte bedrijfsvoertuigen, elektrische lichte bedrijfswagens, elektrische lichte voertuigen, elektrische mobiliteit,

elektrische vrachtwagen, elektrische vrachtwagens, elektrische tweewielers, elektrisch voertuig, elektrisch voertuig, adoptie van elektrische voertuigen, batterijen voor elektrische voertuigen, batterij voor elektrisch voertuig, opladen van elektrische voertuigen, laadinfrastructuur voor elektrische voertuigen, uitrol van elektrische voertuigen, wagenpark van elektrische voertuigen, initiatief voor elektrische voertuigen, landen binnen het initiatief voor elektrische voertuigen, markt voor elektrische voertuigen, markten voor elektrische voertuigen, modellen van elektrische voertuigen, vooruitzichten voor elektrische voertuigen, aankoopsprijs van elektrisch voertuig, verkoop van elektrische voertuigen, marktaandeel van elektrische voertuigen, aanbod van elektrische voertuigen, laadapparatuur voor elektrische voertuigen, uitrol van laadapparatuur voor elektrische voertuigen, laadpunten van laadapparatuur voor elektrische voertuigen, groei van elektrische voertuigen, aantallen elektrische voertuigen, elektrische voertuigen, data over elektrische voertuigen, initiatief voor elektrische voertuigen, subsidies voor elektrische voertuigen, voertuigen met uitstoot, financiële prikkels voor elektrische auto's, financiële prikkels voor elektrische voertuigen, waterstofcel-elektrische auto, waterstofcel-elektrisch voertuig, wereldwijde elektrische voertuigen, wereldwijd wagenpark van elektrische voertuigen, wereldwijde data over elektrische voertuigen, thuis opladen, hybride elektrisch, hybride elektrische voertuigen, auto's met verbrandingsmotor, voertuig met verbrandingsmotor, voertuigen met verbrandingsmotor, laders voor lichte voertuigen, verkoop van lichte voertuigen, nieuwe energievoertuigen, aantal elektrische auto's, aantal elektrische voertuigen, plug-in hybride, plug-in hybride elektrisch voertuig, plug-in hybride elektrisch voertuig, plug-in hybride elektrische voertuigen, privé laders, privé laden, publieke laders, publiek laden, publiek toegankelijke laders, publiek toegankelijke laadmogelijkheden, aankoopstimulansen voor elektrische auto's, aankoopstimulansen voor elektrische voertuigen, aankoop van elektrische auto's, aankoop van elektrische voertuigen, aankoopsprijs elektrische auto, aankoopsprijs van elektrisch voertuig, aankoopsubsidies voor elektrische auto's, aankoopsubsidies voor elektrische voertuigen, verkoopaandeel van elektrische auto's, verkoopaandeel van elektrische voertuigen, aandeel elektrische auto's, aandeel elektrische voertuigen, slim laden, voorraad van elektrische auto's, voorraad van elektrische voertuigen, transitie naar elektrisch, acceptatie van elektrische auto's, acceptatie van elektrische voertuigen, opladen van voertuigen, elektrificatie van voertuigen, nul-emissie voertuig, nul-emissie voertuigen.

5.1 Python script for google trends url

```
base_url = "https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q="
urls = [base_url + quote(k) for k in keywords]

# Combine all URLs into a single JavaScript array for pasting in browser
# console or bookmarklet
js_array = "const urls = [\n" + ",\n".join([f"'{url}'" for url in urls]) + "\n];"
js_array = js_array + "\nurls.forEach((url, i) => setTimeout(() =>\nwindow.open(url, '_blank'), i * 1000));"

# Output JS-ready array of URLs
js_array[:1000] # display only the first 1000 characters for brevity here

for url in urls:
    webbrowser.open_new_tab(url)
```

6 Dataset documentation

6.1 Consumer Confidence, Economic Climate, and Willingness to Buy (CBS, 2007–2025)

[Source: Statistics Netherlands \(Centraal Bureau voor de Statistiek – CBS\)](#)

Consumer Confidence, Economic Climate, and Willingness to Buy (CBS, 2015–2025)

Dataset code: 83693NED

Coverage period: April 1986 – Present

Source: Statistics Netherlands (Centraal Bureau voor de Statistiek – CBS)

Content and Scope

This dataset contains monthly, seasonally adjusted indicators of consumer sentiment in the Netherlands, collected through the *Consumenten Conjunctuuronderzoek* (Consumer Confidence Survey), continuously conducted since 1972 by CBS and co-financed by the European Commission.

The dataset includes four main indicators:

- Consumer Confidence– A composite index reflecting consumer perceptions of both the national economy and their own financial situation.
- Economic Climate– A perception index focused on the past and expected economic situation in the Netherlands.
- Willingness to Buy – A seasonally adjusted index measuring consumers' inclination to make major purchases.

- Koopbereidheid – The original, unadjusted series reflecting the same underlying sentiment.

Koopbereidheid Indicator (Explanation)

The Koopbereidheid indicator reflects consumer attitudes toward their own financial situation and major spending decisions. It is particularly relevant for understanding short-term fluctuations in household consumption, especially in durable goods purchases.

It is calculated as the average of three balances (positive minus negative responses) to questions about:

1. Financial situation over the past 12 months
2. Expected financial situation over the next 12 months
3. Whether it is a good time to make large purchases

The indicator ranges from -100 (all respondents negative) to +100 (all positive), with 0 indicating a neutral stance (equal pessimists and optimists). Both the adjusted (*Koopbereidheid_3*) and raw (*Koopbereidheid*) series are available in the dataset.

Method of Data Collection

Data is obtained from monthly surveys of a geographically stratified, demographically representative sample of Dutch residents, selected from the *Basisregistratie Personen* (BRP). Survey questions cover economic expectations, personal financial conditions, and attitudes toward large purchases.

All series are seasonally adjusted using full-series time series models. CBS applies standard statistical correction procedures; no additional post-processing is performed. Figures are provisional upon initial publication and finalized after annual revisions. The survey is EU-harmonized to support cross-country comparisons and policy use.

Data Availability and Status

- Period covered: January 2007 – January 2025
- Frequency: Monthly (updated around the 22nd of each month)
- Final data: Up to December 2023
- Provisional data: 2024–2025
- Included variables: Date, Consumentenvertrouwen_1, Koopbereidheid_3 (adjusted), EconomischKlimaat_2, Koopbereidheid (raw)

Values range from -100 (all respondents negative) to +100 (all positive), with 0 indicating a neutral stance. All series are seasonally adjusted using full-series time series models.

Method of Data Collection

The data is obtained from monthly surveys of randomly selected Dutch residents, drawn from the *Basisregistratie Personen* (BRP). Questions cover economic expectations, financial health, and willingness to make large purchases. The sample is geographically stratified and

demographically representative. All data are statistically corrected for sampling variability and seasonally adjusted.

No additional post-processing is applied beyond CBS's standard correction procedures. Provisional figures are later finalized after annual revisions. The survey is harmonized across the EU to allow cross-country comparisons, and the results are used by both CBS and the European Commission for economic monitoring (CBS, n.d.).

Data Availability and Status

- Period covered: January 2007 – January 2025
- Monthly frequency, updated around the 22nd of each month
- Final data available up to December 2023; 2024–2025 figures are provisional
- Includes: Date, Consumentenvertrouwen_1, Koopbereidheid_3, EconomischKlimaat_2
Koopbereidheid

6.2 Fuel Pump Prices per Day 2007-2025

Dataset code: 80416NED

Source: [Statistics Netherlands \(Centraal Bureau voor de Statistiek – CBS\)](#)

Coverage period: January 1, 2006 – April 7, 2025

Update frequency: Weekly

Content and Scope

This dataset contains daily pump prices for motor fuels in the Netherlands. The data includes weighted average daily prices (including VAT and excise duties) for:

- Euro95 Petrol
- Diesel
- LPG

Although the dataset records daily values, these prices are published once per week.

Data Availability

- Available from January 1, 2006 – present
- The dataset includes 7037 rows and 4 columns as of the most recent update.

Data Status

- All published prices are final upon release.
- No provisional figures are used.

Update Schedule

Daily fuel prices (up to and including Monday) are typically published on the following Thursday. If publication on Thursday is delayed due to holidays or other circumstances, the data is released on the next working day.

Method of Data Collection

The statistics are based on daily transaction-level data from approximately 3,800 fuel stations and a substantial portion of electric charging stations across the Netherlands. This data is provided to CBS by Travelcard BV, a commercial provider of fuel payment services. Each transaction reflects actual consumer purchases and includes timestamped price data.

The dataset uses a stratified weighting system to compute average prices. Fuel stations are categorized by fuel type and electric charging form, with each group assigned a fixed annual weight reflecting consumer fueling behavior. The weights are updated annually using external data sources.

CBS performs internal quality controls by checking for completeness, plausibility, and consistency. Missing or suspect data points are addressed through follow-up verification or imputation, ensuring that price estimates reflect realistic market conditions (CBS, 2023).

6.3 Registrations by Fuel Type and Vehicle Category – Netherlands (2007–2025)

Dataset code: 71107NED

[Source: Statistics Netherlands \(Centraal Bureau voor de Statistiek – CBS\)](#)

Coverage period: 2007 – present

Content and Scope

This dataset contains quarterly and monthly counts of newly registered and second-hand vehicles in the Netherlands, broken down by vehicle category and fuel type. Vehicle categories include passenger cars, vans, trucks, buses, tractors, motorcycles, mopeds, and special vehicles. Data on trailers and semi-trailers is excluded.

This research focuses exclusively on fully electric passenger cars (Battery Electric Vehicles – BEVs). These vehicles are powered solely by an electric motor and a large battery pack, with no internal combustion engine.

Data is available from 2007 onward and includes both new and used vehicle registrations.
Registrations are classified by:

- Vehicle type (e.g., passenger car, van, motorcycle)
- Fuel type (e.g., petrol, diesel, LPG, hybrid, electric)

Method of Data Collection

The data originates from monthly reports by RDC (RDC Nederland BV), which collects and validates vehicle registration information based on the official Dutch vehicle registry. The dataset distinguishes between new and second-hand vehicle registrations and applies corrections based on improvements in RDC's data sources (notably from 2007 onward).

CBS processes and publishes these figures quarterly, with new figures typically released within three months after the end of each quarter. Data prior to 2022 is final; more recent figures are provisional and subject to revision.

Vehicles are categorized using standardized definitions, such as:

- Passenger Car: Road motor vehicle with up to nine seats
- Full Electric (BEV): Battery-powered electric vehicle with no combustion engine

No post-sampling adjustments are applied beyond RDC's and CBS's regular consistency and classification checks.

NOS News Articles – Netherlands (2015–2025) (Not used)

Dataset code: NOS-NEWS-2015-2025-FEATHER

[Source: Kaggle \(uploaded by John Boulard\)](#)

Content and Scope

This dataset contains news articles published on the Dutch national broadcaster's website, NOS.nl, between January 1, 2015, and March 31, 2025. It includes full-text content along with metadata such as publication time, article type, and topical section. The dataset is designed to support research in media analysis, sentiment tracking, topic modeling, and temporal analysis of Dutch news.

The full text of each article is stored with preserved HTML structure, allowing for structural parsing and feature extraction (e.g., headlines, quotes, embedded tweets, and article sections).

Method of Data Collection

The articles were scraped directly from the NOS.nl website using automated extraction tools. Only minimal post-processing was applied: unnecessary HTML elements were removed, and embedded tweets were reformatted to support easier identification. The dataset is unannotated and retains some inconsistencies or missing fields, reflecting the natural variability of the source website.

The dataset is stored in Feather format (.feather), which enables fast reading and writing in data processing environments such as Python and R.

Data Availability and Status

- Coverage period: January 1, 2015 – March 31, 2025
- Format: Feather (.feather)
- Data status: Raw scraped text with minor cleaning; no annotations applied
- Download location: [Kaggle – NOS News Articles Dataset](#)

7 Model Selection Rationale

Featured in github here: [Modelvergelijking.ipynb](#)

Model	RMSE	MAE	R ²	MAPE	Opmerking
Prophet	11522.58	9279.27	-4.6566	27.28%	Slecht op alle metrics
Holt-Winters	1427.03	1140.87	-4.8417	20.63%	Verrassend stabiel
SARIMAX	4776.29	4161.23	-1.5153	18.37%	Redelijk, maar traag
SARIMA	4589.61	3732.98	-2.1472	16.11%	Beste klassieke model
XGBoost	6109.53	5328.81	-4.4232	20.61%	Niet goed ondanks tuning
LightGBM	4285.06	3235.11	-1.5048	38.51%	Matig, hoge MAPE

LSTM	5262.84	4533.36	-4.9191	18.29%	Matige resultaten
TFT	8099.60	7017.87	-1.7950	20.87%	Niet indrukwekkend
GRU					Werkt niet

8 Daily brain dump

13/08

Changelist opzetten:

\item Monthly EV sales

\item Google Trends indicators

\item Macroeconomic indicators

\item EV subsidies

\begin{itemize}

\item Time Period (T) — Month and year in YYYY/MM format

\item Home charging (G) — Google Trends score for the search term ``charging at home''

\item Consumer confidence (E) — Composite index reflecting confidence in the economy and personal finances

\item Gasoline (Euro95) -- Average price (F) — Mean monthly price of Euro95 petrol

\item Company car tax rate (%) — Fiscal addition rate for private use of company cars

\item Total vehicles sold (Target) — Total number of electric vehicles sold per month

09/08

Sesh met Jer:

- Woordje houden over dat looking at the data xgboost is les preferable looking at the data (voor en nadelen benoemen van elk model ten opzichte van de data)
- Gebruik zelfde kernwoorde die assessor gebruikt zowel bevestigd/als ontkrachtegen en koopel het aan dataset
- Probeer de logaritme/
- Check met Lotte kijk of het een goed idee is om key changes aan te geven in de abstract door aan te geven waar de grootste aanpassingen zijn toegepast
- Maak een lijstje van kernwoorden die ik kan controleren om gelijkheid
 - Keuze tot die 4
- Hoe ben je tot die features gekomen en wat zit er in die featured
- Eztra aandacht besteden aan data analyze
- Check met Xav of ik goede aanpak heb gedaan met auto arima vs manual tuning

- Als ik doorga met autoarima → zou ik in future work work kunnen zeggen wat ik nodig had wel goede resultaten te krijgen dat zij 1) datasize, 2) verschil in train, val en test. Verhaal dat ik wil meegeven met een andere dataset nuttigere outputs zou hebben gehad.
- 2 dingen die ik nodig had was een betere dataset (wat zou een betere dataset geweest onder welke omstandigheden was dataset beter gewest (model toepassing was goed, alleen de data bleef uit)
- Explicet de vraag beantwoorden, what the dataset would require for better model.
 - Stationary dataset
 - No trend
 - Etc
- Xgboost was te complex en daarom doorgegaan naar Arima model.
- Grammerly gebruiken om thesis te checken (grammerly pro check, for)

Stand van zaken nieuw auto arima methode gemaakt.

Uitkomst was

Maar (0,1,1) 0,0,1,12 werkt het best

08/08

9 ⚡ 1. Evaluatie van MAPE en alternatieven

- ✅ **Controleer inverse transformatie:** je gebruikt inv_yeojohnson correct vóór het berekenen van de MAPE. Zorg dat actuals_level en preds_level beide op hetzelfde niveau staan.
- ⚖️ **Overweeg MASE (Mean Absolute Scaled Error)** MASE is robuuster bij seizoensfluctuaties en beter schaalbaar dan MAPE:

python

```
def mase(actual, predicted, training_series, m=1):
    n = len(training_series)
    d = np.abs(np.diff(training_series, n=m)).mean()
    errors = np.abs(actual - predicted)
    return errors.mean() / d

mase_score = mase(actuals_level, preds_level, train_yj, m=12)
```

10 💡 2. Toelichting op differencing

Zet explicet in je rapport:

- Dat differencing via order=(0,1,1) (en eventueel seasonal_order) de trend verwijderd.

- Dat daardoor de voorspellingen en residuals gebaseerd zijn op stationaire data.

11 ⚡ 3. Plot met uitschieters

- Uitschieters in niveau-data zijn zichtbaar, maar kunnen maskeren dat het model de *differenced* structuur wél goed volgt.

➡ Suggestie: Voeg een extra plot toe met de *differenced actuals vs. model residuals*. Dat biedt extra inzicht:

python

```
plt.plot(fitted.resid, label='Model Residuals')
plt.title('Residuals in differenced space')
plt.legend()
plt.tight_layout()
plt.show()
```

12 ⚡ 4. Visualisatie en splitsing uitleggen

Leg uit hoe je split werkt:

- val_yj wordt telkens verder aangegroeid met 1 extra punt bij rolling validation.
- Horizon van 3 → telkens fc = res.get_forecast(steps=3) ➡ Je gebruikt enkel de laatste voorspelling [-1], dus:
 - Wil je slechts één maand voorspellen? Zet horizon = 1
 - Wil je alle drie de stappen gebruiken? Pas je aggregatie aan.

13 ⚡ 5. Train vs. test performance

- Beschrijf dat de testset **visueel rustiger** is.
- Data in validatie bevat pieken, test niet → hogere MAPE in val.
- Leg uit dat val_yj is meegegenomen in training bij rolling forecast → langere trainingstijd.

14 ⚡ 6. Dummy variabelen voor pieken

Voeg dummy's toe voor structurele pieken (belastingen, subsidies, events):

python

```
df['dummy_peak'] = ((df['Month']=='April') & (df['Year']==2022)).astype(int)
```

Gebruik deze als extra exogene regressor.

15 ⚡ 7. XGBoost parallel implementeren

Een nieuw notebook met parallelle flow voor SARIMAX en XGBoost, waarbij je:

- Gebruik maakt van TimeSeriesSplit en lag features.
- Data op niveau-niveau gebruikt (geen differencing nodig bij XGBoost).

- Feature engineering consistent aanpakt.

Laat me je helpen met een template voor dat notebook — wil je dat opgedeeld in secties zoals preprocessing, SARIMAX, XGBoost, evaluatie?

07/08

Emergency opties:

Toch wel boxcox gebruiken.

Try LGMB → Kan zonder YJ

Eventueel regressie analyse toepassen (denk zelf geen goed idee, omdat data

Feedback gesprek:

B2:

Keuze voor exgboost uiteindelijk wel een redelijke keuze voor

Op dit moment lijkt dit ook dus wel zo te zijn (aangezien sarimax voor geen kut werkt)

Rober zegt. Begin met een simpel model, kijk of het werkt, zo niet kan je doorgaan naar een ander model, bijvoorbeeld xgboost. (klinkt kleine hint naar ander model)

Robbert zegt:

Het kan zo zijn dat het niet lukt om

Kijk, het kan gewoon gebeuren dat je dat het niet lukt om iets te modelleren dat dat je model gewoon slecht blijft, etcetera. Als je alle juiste stappen hebt gedaan en je model is gewoon fluit, Dat is niet erg, maar dan moet je wel gewoon zeggen, want ja, het lukt niet om dit goed te voorspellen, dan moet je niet in. In conclusies dingen mooier gaan proberen te praten dan ze zijn. Je mag best wel zeggen dat letterlijk de eerste zin van mijn eigen masterscriptie was. Met huidige data kunnen we niks concluderen punt Dat was mijn conclusie en vervolgens legde ik Natuurlijk uit Waarom dat allemaal niet lukte, Maar het mag best dat het niet lukt. Dat kan gebeuren en Dat is vaak meer gewoon de te weinig data et cetera. En Als je alle juiste stappen hebt gedaan kan je prima op afstuderen.

Model kiezen manual;

Model	RMSE/Mape	Autocorrelation?
SARIMA(1,0,0)(1,1,0)[12]	20	no
SARIMA(0,0,1)(1,1,0)[12]	3823/16.36	no
SARIMA(1,0,1)(1,1,0)[12]	28	no
SARIMA(1,0,1)(1,1,1)[12]	Validation RMSE (rolling): 6319.45 Validation MAPE (rolling): 27.27%	No

t

Model	Rmse/MAPe	Auto
SARIMA(1,1,0)(1,1,0)[12]	Validation RMSE (rolling): 7045.85 Validation MAPE (rolling): 32.96%	Ljung-Box (no autocorrelation if p > 0.05): lb_stat lb_pvalue 10 9.655832 0.471189 20 33.844087 0.027199 30 48.049060 0.019603
SARIMA(0,1,1)(1,1,0)[12]	Validation RMSE (rolling): 7524.38 Validation MAPE (rolling): 32.33%	Ljung-Box (no autocorrelation if p > 0.05): lb_stat lb_pvalue 10 16.011808 0.099295 20 40.054649 0.004917 30 59.795225 0.000975
SARIMA(1,1,1)(1,1,0)[12]	Validation RMSE (rolling): 7576.13 Validation MAPE (rolling): 31.00%	Ljung-Box (no autocorrelation if p > 0.05): lb_stat lb_pvalue 10 24.178894 0.007139 20 50.223834 0.000206 30 78.030513 0.000004
SARIMA(1,1,1)(1,1,1)[12]	Validation RMSE (rolling): 7250.84 Validation MAPE (rolling): 28.57%	Ljung-Box (no autocorrelation if p > 0.05): lb_stat lb_pvalue 10 24.275535 0.006902 20 55.555212 0.000034 30 80.161503 0.000002

Goed uitleg over sarima:

16 Introduction to Seasonal ARIMA (SARIMA)

As discussed previously, while ARIMA models are powerful tools for modeling time series, they have limitations when faced with data exhibiting strong, repeating seasonal patterns. Trying to capture seasonality solely through non-seasonal AR or MA terms often requires very high orders, leading to overly complex models that might not accurately reflect the underlying seasonal structure.

To address this, we introduce the **Seasonal Autoregressive Integrated Moving Average (SARIMA)** model. SARIMA extends the basic ARIMA framework by explicitly incorporating seasonal components into the model structure. This makes it particularly well-suited for time series data where patterns repeat over a fixed period, such as monthly sales figures showing yearly peaks or daily website traffic exhibiting weekly cycles.

16.1.1 Understanding SARIMA Notation

A SARIMA model is typically denoted as $\text{SARIMA}(p,d,q)(P,D,Q)m$. Let's break down this notation:

- $(p,d,q)(P,D,Q)$: These are the **non-seasonal** parameters, exactly the same as in the standard ARIMA model you learned about in the previous chapter.
 - p : Non-seasonal Autoregressive (AR) order.
 - d : Non-seasonal Differencing order.
 - q : Non-seasonal Moving Average (MA) order.
- $(P,D,Q)(P,D,Q)$: These represent the **seasonal** components of the model. They are analogous to their non-seasonal counterparts but operate at the seasonal lag.
 - P : Seasonal Autoregressive (AR) order. It captures the relationship between the current observation and observations from previous seasons.
 - D : Seasonal Differencing order. It accounts for seasonal trends by subtracting observations separated by one full season.
 - Q : Seasonal Moving Average (MA) order. It models the relationship between the current error and errors from previous seasons.
- m : This is a critical parameter representing the **seasonal period** or frequency. It's the number of time steps in one full seasonal cycle. For example:
 - $m=12$ for monthly data with an annual seasonality.
 - $m=4$ for quarterly data with an annual seasonality.
 - $m=7$ for daily data with a weekly seasonality.
 - $m=52$ for weekly data with an annual seasonality.

16.1.2 How SARIMA Works

Think of a SARIMA model as combining two processes: one that models the non-seasonal dynamics and another that models the seasonal dynamics.

1. **Seasonal Differencing:** If $D>0$, the model first applies seasonal differencing ($y_t - y_{t-m}$) to remove or reduce the seasonal trend. This helps stabilize the seasonal component of the series.
2. **Non-Seasonal Differencing:** If $d>0$, the model then applies non-seasonal differencing ($y'_t - y'_{t-1}$, where y' is the seasonally differenced series

if $D > 0$, otherwise $y' = yy = y$) to handle non-seasonal trends and make the series stationary in the usual sense.

3. **ARMA Modeling:** Finally, an ARMA-like structure is applied to the differenced series. This structure includes both:

- o Non-seasonal AR (pp) and MA (qq) terms that capture correlations at short lags (e.g., lag 1, lag 2).
- o Seasonal AR (PP) and MA (QQ) terms that capture correlations at seasonal lags (e.g., lag m , lag $2m$, lag $3m$).

For instance, a SARIMA(1,1,1)(1,1,1)12SARIMA(1,1,1)(1,1,1)12 model for monthly data ($m=12$, $M=12$) suggests:

- Non-seasonal: An AR(1) term, one order of regular differencing ($d=1$, $D=1$), and an MA(1) term.
- Seasonal: A seasonal AR(1) term (correlation with lag 12), one order of seasonal differencing ($D=1$, $D=1$, $yt - yt-12$), and a seasonal MA(1) term (correlation between errors at lag 12).

By incorporating these distinct seasonal parameters $(P,D,Q)m(P,D,Q)_m$, SARIMA provides a more structured and interpretable way to model time series that are influenced by predictable cyclical patterns. The following sections will guide you through identifying appropriate seasonal orders using ACF/PACF plots and implementing these models in Python using the statsmodels library.

03/08

Tijdelijke oplossing om exog te plotten:

can you add

'Gasoline (Euro95) - Average price (F)_y'

with lag 4 to this model als exog feature?

```
from statsmodels.tsa.statespace.sarimax import SARIMAX
import numpy as np
import matplotlib.pyplot as plt
from sklearn.metrics import mean_squared_error, mean_absolute_percentage_error
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from statsmodels.stats.diagnostic import acorr_ljungbox
```

```
import scipy.stats as stats

# 1. Set forecast horizon
horizon = 3

# 2. Grab YJ-transformed train/val arrays
# (these were created earlier)
train_yj = y_train_yj.values
val_yj = y_val_yj.values

# 3. Raw (original-scale) validation actuals,
# aligned so that val_actuals[i] matches
# a forecast made at t = train_end_yj + i
val_actuals = df[target_raw]\n
    .iloc[train_end_yj + horizon - 1 : val_end_yj]\n
    .values

# 4. Time-axis for plotting
plot_dates = df["Time Period (T)"]\n
    .iloc[train_end_yj + horizon - 1 : val_end_yj]

# 5. Rolling, manual-diff + SARIMAX forecasts
preds_diff = []
conf_int_80 = []
conf_int_95 = []

for t in range(len(val_yj) - horizon + 1):
    history = np.concatenate([train_yj, val_yj[:t]])
    history_diff = np.diff(history, n=1) # 1st difference
```

```

model = SARIMAX(
    history_diff,
    order=(2, 0, 1),
    seasonal_order=(1, 0, 0, 12),
    enforce_stationarity=False,
    enforce_invertibility=False
)
fitted = model.fit(disp=False)

fc  = fitted.get_forecast(steps=horizon)
diff_m = fc.predicted_mean
ci80  = fc.conf_int(alpha=0.20)[-1]
ci95  = fc.conf_int(alpha=0.05)[-1]

preds_diff.append(diff_m[-1])
conf_int_80.append(ci80)
conf_int_95.append(ci95)

# 6. Undifference back to YJ-level

```

2/08

Tip:

Compare pure ARIMA forecasts against an ETS model on the same differenced series; sometimes ETS handles evolving variance more gracefully.

Discussie punt:

If you have calendar effects (e.g. holidays boosting charging), include them via dummy variables in a regression-plus-ARIMA framework.

YJ shit:

17 Designing a Robust Transformation Workflow for Your Thesis

You're juggling competing needs—avoiding data leakage, handling negatives, keeping your pipeline consistent, and getting a defendable discussion in your thesis. Below is a three-stage workflow you can adopt (and justify) in your Methods chapter.

18 1. Split First, Transform Second (No Leakage)

1. Split into train/val/test once, based on time order, **before** any transformation.
2. For each variable, fit all transformations (Yeo–Johnson, Box–Cox, manual λ , etc.) **only** on the training slice.
3. Store the fitted parameters (λ 's, shift constants) in a dictionary so you can apply them to val/test without refitting.

This guarantees no look-ahead.

19 2. Decide on “Blanket” vs. “Selective” Transformation

Approach	Pros	Cons
1. Blanket Yeo–Johnson on all	<ul style="list-style-type: none">• Uniform treatment of exogenous series• Handles negatives & zeros out of the box	<ul style="list-style-type: none">• Some variables ($\lambda \approx 1$) get almost no benefit
2. Conditional per-feature	<ul style="list-style-type: none">• Only skewed series are transformed• Keeps near-normal series untouched	<ul style="list-style-type: none">• Pipeline complexity increases• Harder to track
3. Mixed Box–Cox	<ul style="list-style-type: none">• Best transform for each (Box–Cox on positives, YJ on mixed sign)	<ul style="list-style-type: none">• Mixing families makes inversion & discussion trickier

Recommendation: For a SARIMAX with multiple exogenous predictors, a blanket Yeo–Johnson keeps implementation and forecasting (including invert-transform) straightforward. You can note in your thesis that you:

- Tried Box–Cox & YJ per-feature,
- Observed extreme λ 's (e.g. 3.945 for gasoline),
- Opted for a unified YJ pipeline to maintain consistency,
- Evaluated model performance both with and without transformations.

20 3. Putting It All Together: Sample Pipeline

```
python  
from sklearn.preprocessing import PowerTransformer  
import numpy as np
```

```
# 1. Split indices (you already have train_end, val_end)  
splits = {  
    'train': (0, train_end),  
    'val': (train_end, val_end),  
    'test': (val_end, len(df))}
```

```

}

# 2. Fit one YJ transformer per column on TRAIN only
yj_transformers = {}

for col in features_raw + [target_raw]:
    y_train = df[col].iloc[:train_end].values.reshape(-1, 1)
    pt = PowerTransformer(method='yeo-johnson', standardize=False)
    pt.fit(y_train)
    yj_transformers[col] = pt

# 3. Apply to every split and store results
df_yj = pd.DataFrame(index=df.index)

for mode, (start, end) in splits.items():
    idx = slice(start, end)
    for col, pt in yj_transformers.items():
        arr = df[col].values.reshape(-1, 1)
        df_yj.loc[idx, col] = pt.transform(arr).flatten()[idx]

# 4. (Optional) First-difference after transform
df_yj_diff = df_yj.diff().dropna()

# 5. Feed into SARIMAX
import statsmodels.api as sm
model = sm.tsa.SARIMAX(
    df_yj[target_raw].iloc[:val_end],
    exog=df_yj[features_raw].iloc[:val_end],
    order=(p,d,q), seasonal_order=(P,D,Q,s)
).fit()

# 6. Forecast with consistent exog transforms
exog_future = df_yj[features_raw].iloc[val_end:]

```

```
fcst = model.get_forecast(steps=len(exog_future), exog=exog_future)
pred = fcst.predicted_mean
```

21 4. Discussion Points for Your Thesis

- **Data leakage:** how fitting λ on the entire series (instead of train only) would have biased your exogenous coefficients.
- **Transform choice:** why Yeo–Johnson over Box–Cox (negatives, uniform pipeline) and how λ 's near 1 reflect “no-change” variables.
- **Outlier impact:** extreme λ for gasoline signals outlier-driven estimation; you can mention future work with winsorizing or robust transforms.
- **Model comparison:** show validation MAPE/RMSE for
 - raw series + log(vehicles sold)
 - variable-specific transforms
 - blanket Yeo–Johnson

This comparison anchors your decision in empirical evidence—exactly what Hyndman & Athanasopoulos recommend when building dynamic regression models with exogenous predictors.

Next steps

1. Run that unified YJ pipeline once.
2. Tabulate λ 's and skewness changes.
3. Compare model metrics.
4. Write up your “transformation strategy” in the Methods and Results discussion.

That way, you finish your thesis with a clear, reproducible, and defensible workflow.

Definities;

Let op!

You're using `train_end` (from raw df) instead of `train_end_bc` (from `bc_df`)

```
bc_df = pd.DataFrame({col: boxcox_all[col] for col in features + [target]})
```

```
# Separate predictors and target
```

```
X_bc = bc_df[features].copy()    # Features (BoxCox scale)
```

```
y_bc = bc_df[target].copy()     # Target (BoxCox scale)
```

```
# === Define split indices for BoxCox-transformed data ===
```

```

n_bc = len(bc_df)
train_end_bc = int(n_bc * 0.8)
val_end_bc = int(n_bc * 0.9)

# === Create train/val/test splits ===
X_bc_train = X_bc.iloc[:train_end_bc]
X_bc_val = X_bc.iloc[train_end_bc:val_end_bc]
X_bc_test = X_bc.iloc[val_end_bc:]

y_bc_train = y_bc.iloc[:train_end_bc]
y_bc_val = y_bc.iloc[train_end_bc:val_end_bc]
y_bc_test = y_bc.iloc[val_end_bc:]

```

Optimizing pipeline

Integrating “The Forecaster’s Toolbox” into Your Workflow

Below is a step-by-step plan that weaves in the key tools from Chapter 5 of *Forecasting: Principles and Practice* to forecast your target “Total vehicles sold” with your four exogenous variables.

1. Adopt a Tidy Forecasting Workflow

1. Data preparation

- Ingest and merge your target (vehicles sold) with the exogenous series (tax rate, home-charging adoption, consumer confidence, gas price).
- Ensure a regular time index, handle missing values, and align frequencies.

2. Visualization

- Plot each series over time to spot trend, seasonality, outliers, structural breaks.
- Use scatterplots or cross-correlation plots to preliminarily assess lead/lag relationships between the target and each exog.

3. Model specification

- Choose a base method (e.g. ARIMA) and plan your exogenous strategy (dynamic regression / SARIMAX).

4. Training & tuning

- Fit your models on the training set, tune via information criteria (AIC/BIC) or cross-validation.

5. Diagnostics

- Check residuals for zero mean, no autocorrelation (ACF, Ljung-Box), constant variance.

6. Forecasting & evaluation

- Produce point forecasts and prediction intervals.
- Evaluate on validation/test sets using point-forecast metrics and distributional scores.

22 2. Stabilize & Decompose Your Series

- **Box-Cox transformation** to stabilize variance (you already do this).
- **STL decomposition** to separate trend, seasonality, and remainder.
 - Inspect the remainder for autocorrelation; if strong, you'll know an ARIMA on the residuals makes sense.

23 3. Establish Simple Benchmarks

Before building ARIMA, fit these naïve methods on **vehicles sold**:

- **Naïve**: last observed value.
- **Seasonal naïve**: last value from the same season.
- **Drift**: line from first to last point.

Compare their RMSE/MAPE/MASE on your validation set. Any ARIMA or SARIMAX must beat these to be worthwhile.

4. Baseline ARIMA (Univariate)

1. Use auto_arima (AIC/BIC) to pick (p,d,q)—let the model handle differencing rather than differencing manually.
2. Manually tune seasonality if needed.
3. Diagnose residuals with ACF plots and Ljung-Box tests; check normality/homoscedasticity.
4. Record validation RMSE/MAPE/MASE and prediction-interval coverage.

5. Dynamic Regression with Exogenous Variables

For each candidate exogenous series (tax_rate, G_home_charging, E_confidence, F_gas_price):

1. **Lag selection**
 - Examine cross correlation and AIC/BIC when adding lagged exogs to your ARIMA.
 - Use ACF/PACF on the prewhitened series to guide plausible lags.
2. **Fit a SARIMAX**
 - Augment your ARIMA baseline with the lagged exogenous predictor.

- Re-evaluate on the validation set.

3. Diagnostic checks

- Ensure residuals remain white noise after adding the exog.

4. Record metrics

- RMSE/MAPE/MASE improvement over the univariate ARIMA.
- Interval score if you're interested in distributional accuracy.

Repeat this process one exog at a time.

6. Multivariate Model & Final Evaluation

1. **Combine the best exogs** (those that improved your baseline) into a single SARIMAX.
2. **Time-series cross-validation** (rolling-origin) for robust hyperparameter tuning and lag selection.
3. Fit on the full training+validation set.
4. Forecast the hold-out test set.
5. Compute your final metrics:
 - Point-forecast: MAE, RMSE, MASE
 - Interval/distributional: quantile score or coverage of 80/95 % intervals

7

Lag Selection

Fit Sarimax and diagnose

23.1.1 1/08

Model	RMSE	MAPE
Naïve	5360.99	19.05
Seasonal Naïve	4456.62	21.24
Drift	6065.02	24.81
aUTOARIMA(1,0,2)(0,0,1,12)		
Baseline	4155.47	19.47%
Gas L0	4372.89	19,55 %
Company car tax rate (%) L1	3756.22	19.13%
Home charging (G) Lag1	3708.38	18.40%
Consumer confidence (E) Lag 6	3801.21	18.87%
Gas+Home charging	4582.25	17.73

GAS + CAR TAKS	4721.88	18.73
All	4541.29	16.70%
Gas+Home charging TEST	4582.25	17.73
GAS + CAR TAX TEST	4618.95	12.58%
All on test	5172.46	15.84

5/08

Model	RMSE	MAPE
Naïve	5360.99	19.05
Seasonal Naïve	4456.62	21.24
Drift	6065.02	24.81
AutoArima (0,0,1)(1,0,1,12)		
Autoarima	11806.73	45.40
ARIMA(1,0,2)(0,0,1,12)		
Baseline	4359.91	18.31%
Gas L6	3797.68	15.54 %
Company car tax rate (%) L3	3756.22	19.13%
Home charging (G) Lag 5	3708.38	18.40%
Consumer confidence (E) Lag 6	3801.21	18.87%
Gas+Home charging	4582.25	17.73
GAS + CAR TAKS	4721.88	18.73
All	4541.29	16.70%
Gas+Home charging TEST	4582.25	17.73
GAS + CAR TAX TEST	4618.95	12.58%
All on test	5172.46	15.84

Current workflow exog features:

1. Look-Ahead Bias in Exogenous Values

Right now, for each 3-step forecast you're feeding the model the actual future gasoline price (and you'll do the same for other exogs). In real life you'd have to forecast those exogs too or at least assume a scenario.

Action: build a separate model (or simple assumption) for each exog and feed its 3-step forecasts into your SARIMAX.

Yes, I can confirm your code is using actual future values of the exogenous variable (gasoline price) for each forecast step. This is a form of "data leakage" and is not realistic for true forecasting.

Proof (with code reference):

In your recursive forecast loop:

```
for step in range(horizon): idx = t + step gas_exog_full = np.concatenate([np.array(temp_history_gas)[-6:], [val_gas[idx]]]) gas_exog_lag6 = np.roll(gas_exog_full, 6)[6:] exog_next = gas_exog_lag6.reshape(-1, 1) res = fitted_model.get_forecast(steps=1, exog=exog_next) ... temp_history_gas.append(val_gas[idx])
```

- `val_gas[idx]` is the actual future value from the validation set.
- For each step, you append the true future value to `temp_history_gas`, so your forecast for the target at time $t+step$ is conditioned on the real gasoline price at $t+step$.

Why this is a problem

- In real applications, you do not know future exogenous values. You would need to forecast them, use a scenario, or use only historical values.
- Using actual future exog values makes your target forecast unrealistically optimistic and not reproducible in production.

How to fix

For a realistic forecast, you should either:

- Forecast the exogenous variable for each step ahead, and use those forecasts.
- Use only historical exogenous values (not future ones).
- Use a scenario (e.g., "flat" or "expected" values).

Summary

Your current workflow does leak future information from exogenous variables into the target forecast.

For proper validation, you should not use actual future exogenous values when forecasting the target.

This applies to all exogenous features in your notebook.

Up next is to determine optimal lag for exog features:

[Milvus.io](#) says:

typically use a combination of statistical tests, visual analysis, and validation techniques. The goal is to balance model accuracy with simplicity by selecting the smallest number of lags that capture the most relevant patterns in the data. Common methods include **analyzing autocorrelation plots**, using information criteria like AIC or BIC, and testing models with

cross-validation. Each approach has trade-offs, and combining them often yields the best results.

ACF/PACF

ACF:

The ACF shows how strongly a time series correlates with its lagged values

PACF isolates the correlation at a specific lag, excluding effects from earlier lags. For example, in an autoregressive (AR) model, significant spikes in the PACF plot indicate potential lags to include. If the PACF drops sharply after lag 3, an AR(3) model might be appropriate

Model ARIMA(1,0,2)(0,0,1,12)	RMSE	MAPE
Baseline	4359	18.31%
Gas L6	4658.33	18.24%
Company car tax rate (%) L3	4351.58	18.22%
Home charging (G) Lag 5	4393.56	18.58%
Consumer confidence (E) Lag 6	4372.28	18.58%
All	4541.29	16.70%
All on test	5172.46	15.84

Problemen (voor later)

23.2 Keuze De features (vooral google home charging) **Alternatief: Combineer beide (optioneel)**

Voeg als bijlage een overzicht toe van álle variabelen die je had, maar geef in de hoofdtekst aan waarom je slechts een selectie hebt gebruikt.

Bijvoorbeeld:

"In dit onderzoek is ervoor gekozen om uit elke hoofdgroep (beleidsdata, gedragsdata, economische perceptie en brandstofprijzen) één representatieve variabele te gebruiken. Deze afweging is gemaakt om de complexiteit beheersbaar te houden en het risico op multicollineariteit te beperken."

31/07

So problem with rolling window, I use 1 month/ 1 step so its predicting 1 month ahead instead of 3, 3 give same shitty results.

Please rewrite code to either 3 months validation or cross validation

Ill say 3 step because train test split can stay the same, also add prediction interval

And then run code again

Cool ran te code again

Also check if

Resultaten

Okey XGboost werkt veel beter: (voor baseline only)

Validation RMSE (3-step rolling): 2967.30

Validation MAPE (3-step rolling): 12.92%

Test RMSE (3-step rolling): 5625.40

Test MAPE (3-step rolling): 16.38%

Model	Pros	Cons	Notes
ARIMA / ARIMAX	<ul style="list-style-type: none">✓ Transparent and explainable
✓ Strong for short-to-mid horizon
✓ Handles seasonality well
✓ Interpret coefficients and residuals directly	<ul style="list-style-type: none">✗ Weaker on long-term predictions
✗ Sensitive to nonstationarity
✗ Requires manual lag selection	Ideal when interpretability matters—great for policy insight, economic analysis
XGBoost	<ul style="list-style-type: none">✓ Top-tier predictive performance
✓ Handles nonlinear relationships
✓ Flexible lag structures	<ul style="list-style-type: none">✗ Black-box (low explainability)
✗ No built-in temporal awareness
✗ Harder to communicate insights	Best if accuracy is the only goal—not ideal for explaining decisions
LSTM	<ul style="list-style-type: none">✓ Can learn complex temporal patterns
✓ Handles multivariate time series	<ul style="list-style-type: none">✗ Requires lots of data
✗ Computationally expensive
✗ Worse performance on your dataset	

f

Model	RMSE	MAPE
OG AUTO ARIMA(0,0,1)(1,0,1)[12]	11806.73`	45.40
Model A order=(1,0,1), seasonal_order=(1,0,1,12)	6762.12	27.03
Model B order=(0,0,2), seasonal_order=(1,0,1,12)	11877.55	46.78
Model C order=(1,0,2), seasonal_order=(1,0,1,12)	6279.55	25.05
Model D (fallback op seasonal only) order=(0,0,0), seasonal_order=(1,0,1,12)	13115.20	50.57
Model E (zonder seasonal MA) order=(0,0,1), seasonal_order=(1,0,0,12)	11469	43.89

🔥 Model C: SARIMA(1,0,2)(1,0,1,12)

RMSE = 6279.55, MAPE = 25.05%

Deze combinatie lijkt het meest geschikt voor jouw data: het combineert korte termijn (AR=1), middeltermijn (MA=2) én seizoenspatronen op een gebalanceerde manier. Model A is een sterke tweede keuze als je een iets eenvoudiger model verkiest.

Model	Order	Seasonal Order	RMSE	MAPE (%)	Opmerkingen
OG AUTO	(0,0,1)	(1,0,1,12)	11806.73	45.40	Automatisch gekozen, maar zwak
A	(1,0,1)	(1,0,1,12)	6762.12	27.03	Prima model, beter dan OG
B	(0,0,2)	(1,0,1,12)	11877.55	46.78	Slechter dan OG
C	(1,0,2)	(1,0,1,12)	6279.55	25.05	Sterk model, gebalanceerd
D	(0,0,0)	(1,0,1,12)	13115.20	50.57	Te simpel
E	(0,0,1)	(1,0,0,12)	11469.00	43.89	Mist seasonal MA
NEW 1	(1,0,2)	(1,0,0,12)	6263.64	25.05	Even goed als Model C, simpeler seizoen
NEW 2	(1,0,2)	(0,0,1,12)	4359.91	18.31	Beste prestaties tot nu toe ✓

"Various modeling approaches were tested, including deep learning and ensemble methods, but the final selection was based not just on accuracy, but on **interpretability, robustness, and domain alignment**. ARIMA provided a clear mapping between predictors and outcomes, making it ideal for policy communication and scenario testing."

Model ARIMA(0,0,1)(1,0,1,12)	RMSE	MAPE	PLOT
Baseline	11806.73 / 7497.61	45.40%/26.16%	
Gas L6	11682.34 /7113.64	44.73%/23.59	
Company car tax rate (%) L3	14696.04/ 6849.32	37.57% /22.84	
Home charging (G) Lag 5	11681.37 / 7174.59	1. 98/24.51	
Consumer confidence (E) Lag 6	12238.37/ 8094.56	49.23%/28,21	

1. Build an ARIMA with all 3 top features: Include Gas L6, Company Tax L3, and Home Charging L5 with their respective lags. Then evaluate on the same horizon.

2. Monitor for multicollinearity: If two features track each other tightly, it might blur the signal. You could check variance inflation factor (VIF) or use PCA/regularization.

3. Compare combo model's RMSE/MAPE to individual: If your new combo model beats your best single-feature model, that's a win. If not, trim and retry.

Todolist:

Check of je interval kan toevoegen

Kijken welke feautures het meest toevoegen

Schaal verkleinen daar duizendtallen kijk if RMSE hierdoor kleiner wordt

Check datzelfde modellen worden gebruikt

Gas price toevoegen

Index veranderen naar datum

ACF/PACF plots toevoegen bij lags

Schaal verkleinen daar duizendtallen kijk if RMSE hierdoor kleiner wordt.

De overprediction checken)

Validation RMSE: 14696.04

Validation MAPE: 37.57%

In *Forecasting: Principles and Practice*, the authors (Rob J. Hyndman and George Athanasopoulos) explain that **lag selection is highly context-dependent**, and there's no "one-size-fits-all" rule like "6 is always best." However, here's why **lags up to 6** are commonly used and often recommended early in the modeling process:

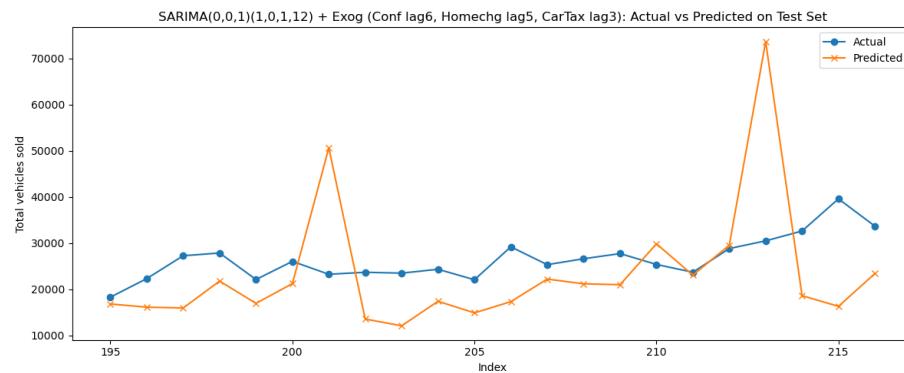
24 🔎 Why 6 Lags Are Commonly Considered

- **Monthly data:** If your series is monthly, 6 lags represent half a year—long enough to potentially capture mid-season trends or cyclicity.
- **Avoiding overfitting:** Including too many lags can lead to model overcomplexity and poor generalization. Stopping at 6 keeps things parsimonious.
- **Autocorrelation structure:** Often, autocorrelations die off after a few lags. The ACF/PACF plots help reveal where this happens. Many economic and sales series show meaningful autocorrelation within the first 3–6 months, but not far beyond.

25 📚 What the Book Actually Says

While the book doesn't prescribe a hard rule on "maximum 6 lags," it does emphasize:

- **Using ACF/PACF plots** to identify useful lags.
- **Using information criteria (AIC, BIC)** to compare models across lags.
- **Testing multiple lag structures** while watching for overfitting and parameter stability.



Ga echt de lijst af en geef duidelijk aan waar ik naar heb geluistert

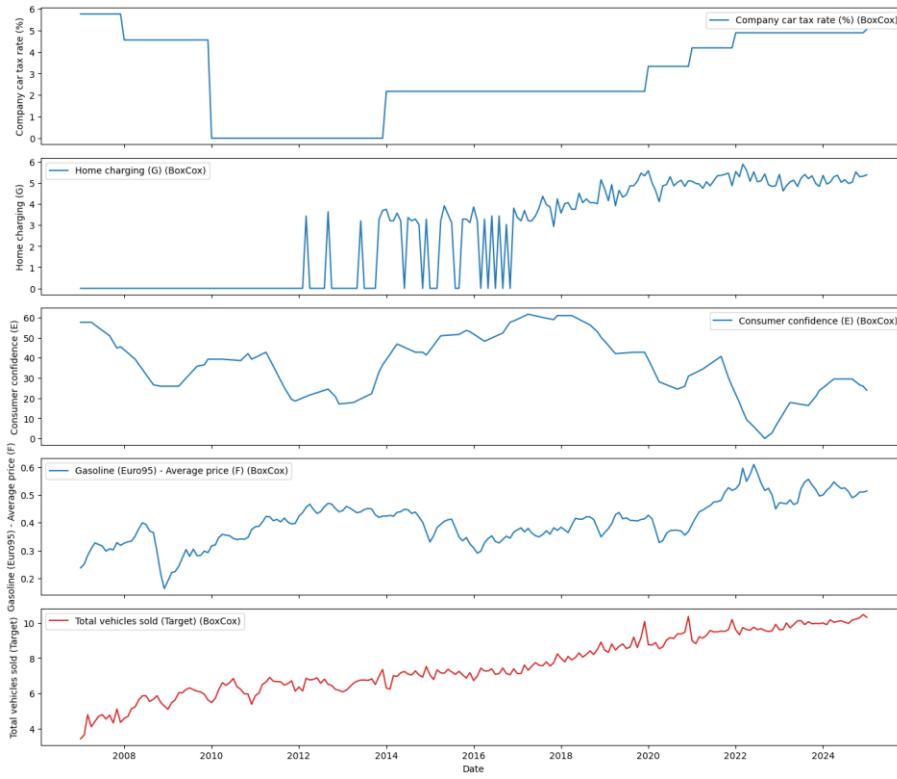
Ander model eventueel in discussie, maar zorg ervoor dat dit model goed is.

Uitleg geven over de hyperparameter choices for the ARIMAX zeggen dat ik autoraam heb gebruikt en uitleggen waarom de gekozen tuning beste presteert.

(eventueel ook manual tunen)

manually tuning the model yourself through an ACF/PACF analysis.

Check of ik alle features kan toevoegen



25.1.1 14/07

Finished with all the tests:

Model	MAPE (%)	RMSE
Consumer Confidence (lag 0)	24.11	8,084.01
Baseline (No exog)	25.62	8,822.05
Gasoline Price (Lag 4)	28.69	8,287.79
Hybrid (0) (G)(lag0)	33.46	11,693.14
Electric Car (G) (Lag 10)	33.46	11,693.14
Home Charging (G) (lag 11)	36.37	10,380.46
Company Car Tax (Lag 12)	37.96	13,033.54
Willingness to Buy (e) (lag 2)	40.87	11,342.94

Diesel Price (Lag 2)	47.09	13,269.11
Economic Climate (E) (lag2)	48.61	13,513.03

Looking at the results, only consumer confidence with lag 0 outperformed the baseline.

Next up will be training xgboost on it and trying test.

The forecast horizon was either 22 months or 21 months based on the lag

Also looking at

11/07

If you're thinking of testing SARIMAX against one of these, I'd suggest starting with **TFT** or **NHITS** for interpretability and exogenous support, or **PatchTST** for a clean univariate baseline

Thinking of adding other models:

26 Logbook Entry – Neural Networks in Forecasting

This chapter presents **global forecasting models** using neural networks that learn from multiple time series, addressing scalability and automation challenges faced in traditional approaches like SARIMAX.

27💡 Key Concepts

- **Global models** learn shared patterns across many series, unlike **local models** (e.g. ARIMA) which build one model per series.
- NNs handle **exogenous variables** elegantly—no need for manual feature engineering.
- Forecasting is framed as **learning a function** to map past inputs and covariates to future outputs via optimization (e.g., using MAE or MSE).

28📅 Architectures

- **MLP:** Basic building block—nonlinear mapping using hidden layers.
- **Modern Architectures:**
 - RNNs, CNNs, Transformers, and hybrids (e.g. NHITS, DeepAR) offer specialized strengths.
 - Forecasts can be **direct** (predict whole horizon at once) or **recursive** (step-by-step prediction).

29📊 Forecasting Tools & Libraries

- **NeuralForecast:** Core library used to build, train, and evaluate neural models.
- Supports training with different **loss functions**, including probabilistic losses for uncertainty quantification.

- **Scaling** matters—robust scaling via interquantile range improves forecast accuracy.
- AutoModels help tune **hyperparameters** automatically for optimal performance.

30 Exogenous Variables

- Neural models allow for:
 - **Static variables:** unchanged over time (e.g., market ID).
 - **Historical variables:** past-only (e.g., temperature).
 - **Future-known variables:** predictable ahead of time (e.g., day of the week).

31 Forecasting Model Comparison Table

Here's the core table comparing available neural network models in the NeuralForecast library:

Model	Architecture	Scope	Type	Exogenous Support
Autoformer	Transformer	Univariate	Direct	Future
DeepAR	RNN	Univariate	Recursive	Future, Static
NHITS	MLP	Univariate	Direct	Future, Historical, Static
iTransformer	Transformer	Multivariate	Direct	—
LSTM	RNN	Univariate	Recursive	Future, Historical
MLPMultivariate	MLP	Multivariate	Direct	Future, Historical, Static
PatchTST	Transformer	Univariate	Direct	—
BiTCN	CNN	Univariate	Direct	Future, Historical, Static
FEDformer	Transformer	Univariate	Direct	Future
TimeMixer	MLP	Multivariate	Direct	—
TFT	Transformer	Univariate	Direct	Future, Historical, Static
TimeLLM	LLM	Univariate	Direct	—

 Note: This table helps guide model selection based on architecture, forecast type, and whether you have exogenous data inputs

32 Logbook Entry – Exploring Advanced Models alongside SARIMAX

In this chapter, I explored **three key forecasting techniques**—MSTL, Prophet, and Vector Autoregression (VAR)—each bringing unique strengths to modeling complex time series. As I consider using these alongside **SARIMAX**, I'm evaluating their compatibility and complementary benefits.

33 MSTL (Multiple Seasonal-Trend decomposition using Loess)

- Handles **multiple seasonalities** cleanly—ideal for high-frequency or irregular patterns.

- Decomposes time series into **trend + multiple seasonal components + residual**, which can be recombined for forecasting.
- Pros: Works well with seasonal naïve models and ETS; can be adapted with SARIMAX by integrating seasonal decomposition into SARIMAX residual modeling.
- Challenge: MSTL is a preprocessing step, not a standalone statistical model, so the integration requires careful coordination.

34 🌎 Prophet

- Popular for its simplicity and Bayesian trend/seasonality decomposition.
- Good for daily data with **holiday effects**, long trends, and user-defined seasonality.
- Pros: Easy to use and visual; holidays can be included as exogenous regressors (similar to SARIMAX's `exog`).
- Limitation: Residual autocorrelation is often left unmodeled, which SARIMAX handles more explicitly via ARIMA components.

35 📈 VAR (Vector Autoregression)

- Models **interrelated time series**—each variable influenced by lagged values of all others.
- Pros: Powerful for multivariate systems (e.g., demand vs. temperature), which SARIMAX cannot handle natively.
- Limitation: Complex to interpret and tune; requires stationary inputs, while SARIMAX manages both stationary and non-stationary series with differencing.

36 💬 Integration Thoughts

- I'm curious whether **SARIMAX + MSTL** would improve short-term forecasts by modeling seasonal patterns more accurately before applying ARIMA structure.
- **Prophet's** strength in holiday handling may be better implemented directly in SARIMAX via dummy variables or exogenous regressors.
- **VAR** might be used in parallel where multivariate causality is relevant—but less so in single-variable SARIMAX contexts.

Going forward, I plan to prototype each hybrid and validate them using **error metrics** like RMSE and MASE, while tracking forecast interval quality (e.g., using Winkler scores). This should reveal which combinations are practically useful and which are just theoretically appealing.

36.1 10/07

37 Exogenous Feature Testing and SARIMAX Model Diagnostics

Workflow Summary:

- Started testing exogenous features one at a time.

- Initially prioritized features with low VIF — started with Company car tax rate, but the variable failed to improve the model and was ultimately discarded.
- Shifted focus to high-ranked PCA features. The top candidate (Home charging) showed statistical significance at lag 5.

Lag 5 of Home Charging — SARIMAX Diagnostics:

- Coefficient: $+0.0321$ ($p = 0.021$) → significant positive effect.
- However, fitting revealed structural instability:
 - $ma.S.L12 = -1.0000$ → flatlined seasonal MA coefficient with massive standard error.
 - $\sigma^2 = 0.0535$ → estimated variance unreliable (standard error ~22).
 - Heteroskedasticity test $p = 0.04$ → flagged non-constant residual variance.

These diagnostics suggest overfitting or misspecification within the seasonal components. Will revisit model structure and seasonal order next.

38 Next Steps:

1. Create a **new notebook** with systematic naming and organization of feature versions.
 - Track transformations (Box-Cox, differencing, scaling).
 - Save lagged versions separately for reusability.
2. Refit models more deliberately:
 - Verify exogenous setup — current runs likely used pdq from prior SARIMA grid search.
 - Test features independently and apply proper lag selection based on AIC and p-value.
3. Tomorrow's Modeling Agenda:
 - Evaluate other high-PCA features individually (e.g. Koopbereidheid, Economisch klimaat).
 - Determine best lag per feature with AIC/p-value.
 - Retain only statistically significant contributors.
 - Test multi-feature SARIMAX specs — start with top 2–3.
 - Monitor multicollinearity using VIF and heatmaps.
 - Apply PCA or Ridge/Lasso if necessary to stabilize modeling.

will continue doing this

Initial Approach:

- Began exogenous feature selection using lowest VIF scores. Theoretically sound, since low VIF indicates low multicollinearity, improving model interpretability and stability.
- Started with Company car tax rate, but it showed poor statistical significance across lags, added no predictive value, and degraded model fit — demonstrating that low VIF alone doesn't guarantee forecasting utility.

Strategic Pivot:

- Transitioned to high-ranked PCA features, prioritizing variables that explain the most variation in the data.
- This surfaced Home charging adoption as a promising candidate. After lag testing, **Lag 5** proved statistically significant ($p = 0.020$, AIC = -3.988), confirming a delayed behavioral effect on vehicle sales.

Model Diagnostics:

- SARIMAX with `home_charge_lag5` as exogenous regressor yielded a positive, significant coefficient.
- However, structural warnings appeared:
 - `ma.S.L12 = -1.0000`: suspicious flatlined seasonal MA term with extreme error margin.
 - σ^2 : low point estimate (~ 0.0535) but extremely high standard error (~ 22), suggesting numerical instability.
 - Heteroskedasticity test ($p = 0.04$) flagged inconsistent residual variance.

These patterns often indicate overfitting or mis-specification in the seasonal structure.

Reassessing seasonal components is necessary to stabilize the model.

39 Next Steps & Notebook Reset:

Why Restarting Makes Sense:

- Creating a fresh notebook ensures clear structure, traceable workflows, and consistently named transformations (e.g., `BoxCox`, `lags`).
- It simplifies feature reusability and speeds up later comparison and modeling iterations.

Tomorrow's Workflow:

1. **Test other top PCA features individually** e.g., `Koopbereidheid`, `Economisch klimaat`, etc.
2. **Determine best lag per feature** Based on AIC and coefficient significance
3. **Retain only statistically strong contributors**
4. **Combine strongest features in joint SARIMAX models** Start small (2–3 variables) and scale up gradually
5. **Evaluate multicollinearity risks** Using VIF analysis and correlation heatmaps

6. **Apply dimensionality reduction if needed** PCA or Ridge regression to manage interaction effects or overlap

Now I will add candidate features:

Step 1 – Target Preparation

- Apply **Box-Cox transformation** to stabilize variance.
- Use **seasonal differencing** (e.g., lag 12) to achieve stationarity.
- Confirm stationarity with **ADF** and **KPSS** tests.

40 Step 2 – Feature Preprocessing

For each candidate feature:

- Apply **log/Box-Cox transformation** if needed.
- Test for stationarity → **difference** if non-stationary.
- **Standardize** if required for interpretability.

💡 Hold off on PCA until all features are transformed, as rankings may change.

41 Step 3 – Feature Selection (One-by-One)

python

KopiërenBewerken

for feature in top_ranked_features:

Add transformed feature to SARIMAX

Fit → Evaluate (AIC, RMSE, MAPE)

Retain if performance improves

- Start without lags.
- Compare to baseline SARIMA.
- Drop or lag if performance worsens.

42 Step 4 – Lag Optimization (Per Retained Feature)

python

KopiërenBewerken

for lag in range(1, max_lag+1):

df[f'{feature}_lag{lag}'] = df[feature].shift(lag)

Fit → Evaluate → Keep best lag

- Use **AIC/MAPE** to guide selection.

- Retain lag only if it improves results.

43 Step 5 – Multicollinearity Check (Final Feature Set)

- Calculate **VIF** and plot **correlation heatmap**.
- Remove or combine features with **VIF > 10**.
- Use **PCA** as a fallback if collinearity remains.

44 Step 6 – Final Model & Evaluation

- Fit final **SARIMAX** model with selected features and lags.
- **Back-transform** predictions if needed.
- Evaluate residuals, forecast metrics, and visualize results.

9/07

SARIMA Model Comparison (Log-Transformed Target)

1. Manually Tuned Model: SARIMA(1,1,1)(1,1,1)[12]

- **AIC:** -23.53 (better)
- **Log Likelihood:** 15.77
- **Coefficients:** All significant ($p < 0.001$)
- **Residual Diagnostics:**
 - Ljung-Box $p = 0.57 \rightarrow$ No autocorrelation
 - JB $p = 0.00 \rightarrow$ Slight non-normality
 - Heteroskedasticity $p = 0.43 \rightarrow$ No evidence of heteroskedasticity

2. Auto-ARIMA Model: SARIMA(1,1,1)(1,0,0)[12]

- **AIC:** -12.12 (worse)
- **Log Likelihood:** 11.06
- **Coefficients:** All significant
- **Residual Diagnostics:**
 - Ljung-Box $p = 0.83 \rightarrow$ No autocorrelation
 - JB $p = 0.00 \rightarrow$ Non-normal residuals
 - Heteroskedasticity $p = 0.04 \rightarrow$ Possible heteroskedasticity

Conclusion:

The **manually tuned model performs better** in terms of AIC and log-likelihood, with cleaner residuals and no signs of heteroskedasticity. Use this as your preferred model.

10/07

Goede stappen gezet in de richting van dat data cleaning en optimalisatie, ook alvast eerste forcasting uitgevoerd. Echter waren deze wel auto sarimax vanwege cp,plexiteit van mn data ga ik morgen met de hand de modellen tunen:

Model	Log	FORCAST	MAPE	RMSE
SARIMAX(1, 1, 1)x(1, 1, [], 12)	No	22	.3168	8505.66
SARIMAX(0, 1, 1)x(1, 0, 1, 12)	Yes	22	Log .357 BT .2819	Log .3792 BT 5608.11
SARIMAX(0, 1, 1)x(1, 0, 1, 12)	Yes	3	Log .0331 Bt .2615	Log .3201 BT 2887
Xgboost	No	3	.1400	1908.57
XGBOOST	NO	22	.1709	3798
XBOOST	YES	3	Log .0145 BT .1261	LOG .1372 BT 1416
XGBOOST	YES	22	Log .0156 BT .1359	LOG .1964 BT 3873

We vervaart over dat xgboost beter was

07/07

Gisteren goede stappen gezet richting de ARIMA-modellen. Ik ben tot hoofdstuk 9 gekomen, maar heb de modellen zelf nog niet volledig toegepast. Wel heb ik een veel beter begrip gekregen van mijn data.

Ik heb de log van de targetvariabele geanalyseerd en een Box-Cox-transformatie uitgevoerd om de optimale lambda te bepalen voor een logtransformatie. De reden voor deze transformatie is dat mijn data een sterke trend en hoge variantie vertoont. Door de logtransformatie is dit grotendeels gestabiliseerd.

Daarna heb ik een seisoensdecompositie geplot en ben ik blijven hangen bij het onderwerp moving averages. Hieruit bleek dat een 2×12 -maands moving average het meest geschikt is, vanwege het maandelijkse karakter van mijn data en de jaarlijkse seizoensinvloeden.

Voor morgen staat het toepassen van de modellen op de planning (vervolg van hoofdstuk 9). Mogelijk kijk ik daarna ook nog terug naar hoofdstuk 5 om de evaluatiemethoden voor de modellen beter te begrijpen.

Wil morgen verder kijken naar de modellen. Dus hoofdstuk 9, daarna misschien terug naar hoofdstuk 5 om de evaluatie te bekijken van de modellen

06/07 goal is to set the pipeline steps and learn about Arimax:

1. Problem Definition

Clearly define the forecasting goal, who will use the forecasts, and how they fit into the organization. Engage stakeholders early to align expectations and understand data flows.

2. Data Gathering

Collect both statistical data and expert knowledge. If historical data is limited or outdated, judgmental methods or recent data may be more suitable.

3. Exploratory Analysis

Visualize the data to detect trends, seasonality, outliers, or structural changes. Analyze variable relationships before model selection.

4. Model Selection and Fitting

Choose models based on data availability, explanatory relationships, and use-case needs. Test and compare multiple models (e.g., regression, ARIMA, neural nets), ensuring assumptions are met.

5. Model Use and Evaluation

Generate forecasts and evaluate accuracy once future data becomes available. Address practical issues like missing data and ensure results are actionable in practice.

Based on Forcasting preincibles, its split inro a fef steps:

1. Preparing data:
 - a. prepare data in the correct format
2. Datavisualisation
 - a. See chapter 2
3. Define model
4. Accuracy & perfomerance evaluacing
5. Producing forecasts

44.1.1 02/07

New datapipeline plan:

STEP 1: Start Simple – Univariate ARIMA (Hyndman-style)

“Start simple, start with no features.”

- Just model Total vehicles sold (Target) using **ARIMA / SARIMA**.
- Check: ACF, PACF, seasonality, differencing.
- This shows you understand the **core of time series forecasting** before adding complexity.

STEP 2: Evaluate baseline (ARIMA vs. Random Walk)

- Make sure ARIMA beats the null (random walk or naive forecast).
- **Include 12-month seasonality** as noted in your feedback (SARIMA).

STEP 3: Add Features One-by-One (Forward Selection)

"Add features 1 by 1 until results don't improve anymore."

- Add **only 1 feature** at a time.
- Pick based on domain logic (start with "Electric cars" or "Gasoline price").
- Track MAPE/MAE for each new model.
- Stop when performance **stabilizes or worsens**.

This answers B2 and B4 directly.

STEP 4: Use PCA (optional, later)

Only **after** you've tested 3–5 features individually, and **only if multicollinearity becomes a clear problem again**:

- Then you can say: "*I tested forward selection, but XGBoost still suffered from overfitting. Therefore I applied PCA to reduce dimensionality.*"
- Use PCA not as a default, but as a tool to **solve overfitting**, which shows better reasoning (B3, B4).

STEP 5: Skip XGBoost unless it adds value

You don't have enough data to justify it right now.

Only reintroduce XGBoost **after**:

- ARIMA/SARIMA and ARIMAX are exhausted.
- You can show that **some nonlinearity remains** or you're getting strong SHAP results.
- Then say: "*To explore nonlinear relationships, I tried XGBoost but tuned it very conservatively due to small dataset.*"

Model Type	When to Use	Pros	Feedback Tie-in
SARIMAX	After ARIMA	Handles seasonality + exog vars	Addresses strong 12-month pattern
Linear Regression	As sanity check	Simple interpretability	Should be defined clearly (feedback: pg. 12 issue)
XGBoost	If enough data	Captures nonlinearity	Needs careful tuning (feedback B4)
Prophet	Optional	Flexible, good for time + holidays	Bonus if time allows
LSTM (Recurrent)	Not recommended now	Needs more data	Justify skipping due to small size (feedback B2)

This keeps you aligned with B4, C2, and the overfitting concerns.

Data analyse verder uitgevoerd aan de hand van <https://otexts.com/fpp3/graphics.html>

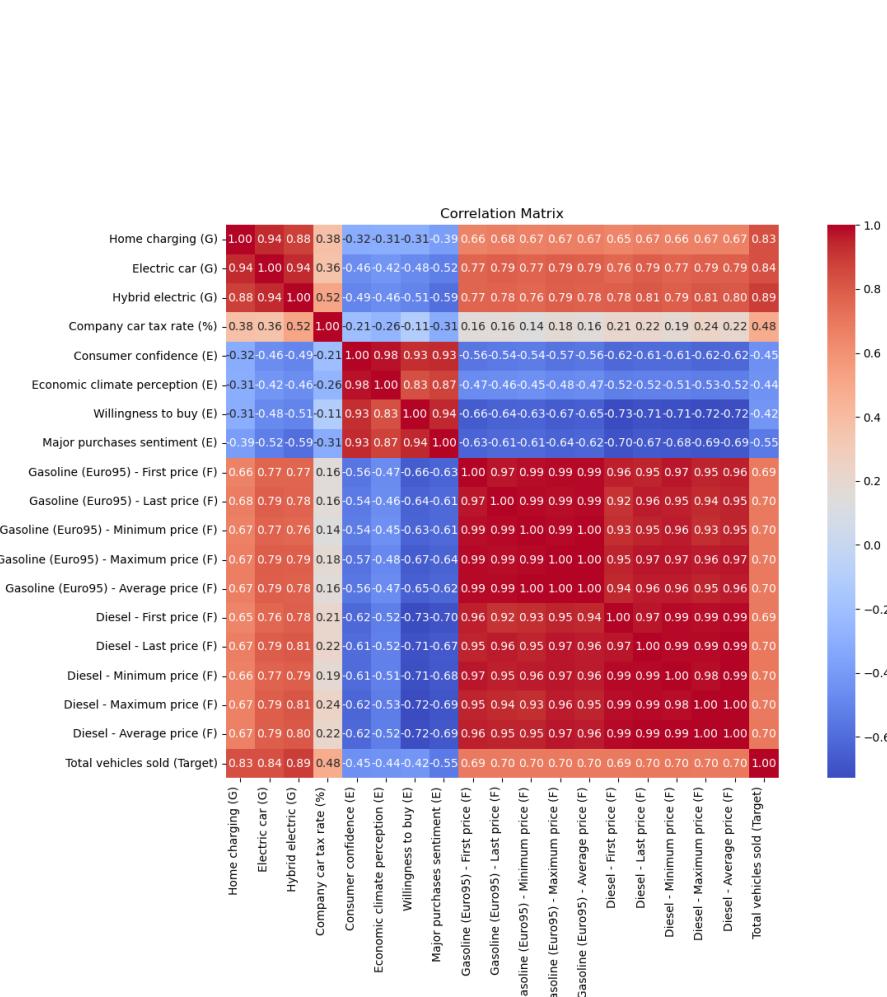
dataset aangepast:

"Periode": "Time Period (T)",
"thuis opladen": "Home charging (G)",
"elektrische auto": "Electric car (G)",
"hybride elektrisch": "Hybrid electric (G)",
"Consumentenvertrouwen": "Consumer confidence (E)",
"Economisch klimaat": "Economic climate perception (E)",
"Koopbereidheid": "Willingness to buy (E)",
"Grote aankopen": "Major purchases sentiment (E)",
"BenzineEuro95_1_first": "Gasoline (Euro95) - First price (F)",
"BenzineEuro95_1_last": "Gasoline (Euro95) - Last price (F)",
"BenzineEuro95_1_min": "Gasoline (Euro95) - Minimum price (F)",
"BenzineEuro95_1_max": "Gasoline (Euro95) - Maximum price (F)",
"BenzineEuro95_1_mean": "Gasoline (Euro95) - Average price (F)",
"Diesel_2_first": "Diesel - First price (F)",
"Diesel_2_last": "Diesel - Last price (F)",
"Diesel_2_min": "Diesel - Minimum price (F)",
"Diesel_2_max": "Diesel - Maximum price (F)",
"Diesel_2_mean": "Diesel - Average price (F)",
"Bijtelling (%)": "Company car tax rate (%)",

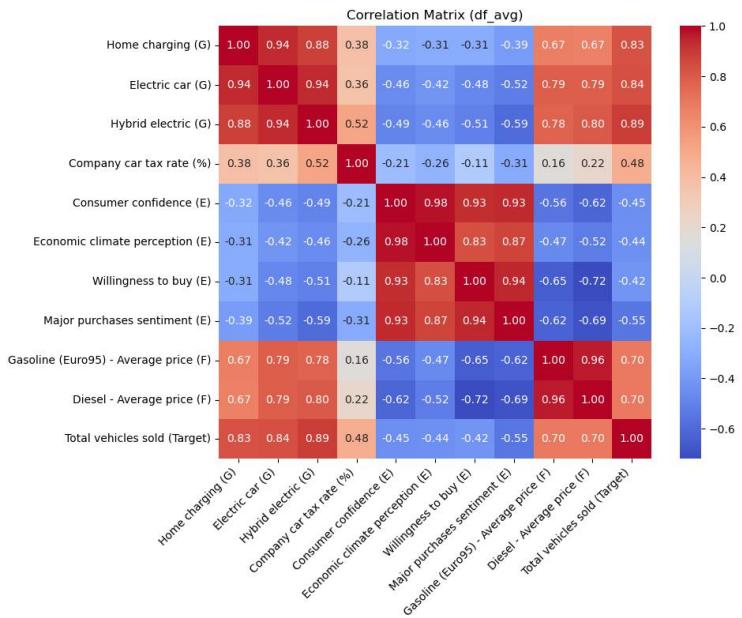
Zodat Termen beter interpreteerbaar zijn.

Legend

- **(T)** = Time
- **(G)** = Google Trends (behavioral signals)
- **(E)** = Economic sentiment (CBS)
- **(F)** = Fuel prices (CBS)
- **(P)** = Policy/fiscal variables
- **(Target)** = Dependent variable

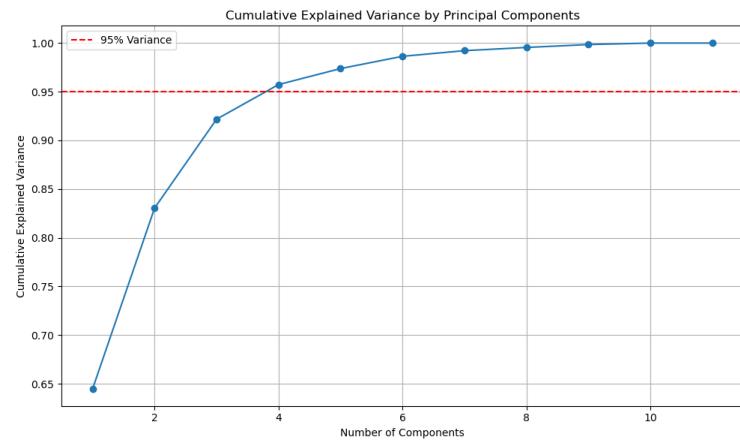


Gezien de data vooral bij Gasoline/diesel enorme multicorinaliteit, daarom de keuze gemaakt om enkel de average maandelijkse benzine/diesel prijs mee te nemen. En de rest te droppen.



Vervolgens nagedacht over de PCA/dataset verleomem en kwam erachter dat het onlogisch was om zoveel diesel/benzine features te hebben, dus heb ze vervolgens gedropt en nieuw dataset aangemaakt.

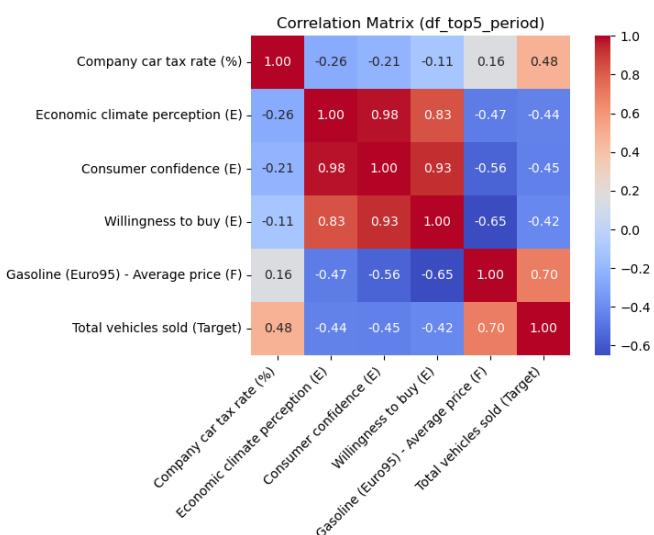
Vervolgens PCA toegepast:



Top 5 variables (PC1-PC3 combined influence):

Company car tax rate (%)	1.164071
Economic climate perception (E)	0.876384
Consumer confidence (E)	0.801296
Willingness to buy (E)	0.790610
Gasoline (Euro95) - Average price (F)	0.785076

Deze variabelen gebruikt en nieuwe correlatie matrix geplot:



Tot mn teleurstelling nog steeds multicorrelatiteit aanwezig:

Vervolg stappen zouden zijn:

- PC1, PC2.... Als input gebruiken, alleen zouden ze dan moeilijk interperteerbaar zijn.
- Alleen 1 van de 3 "E" houden gebaseerd op relevantie en predictive power: in dat geval zou willingness to buy beste zijn.
- Of alles behouden en xgboost, ridge regression gebruiken omdat ze goed zijn met multicorrelatiteit.

Vervolgens vroeg ik mij af of pca overhaakt goed was om in dit stadium te gebruiken want:

PCA negeert temporele orde

Sterke voorspellers scoren mogelijk niet hoog in PCA

Om die reden ga ik over tot het ander feedback punt.

Simpel beginnen, starten zonder features en dan per keer feature tovoegen dat wil ik op de volgende manier doen.

Aan de hand van Hyndman: starten met univariatie:

Enkel de

Univariate ARIMA (Hyndman-style)

Kleine dilemma:

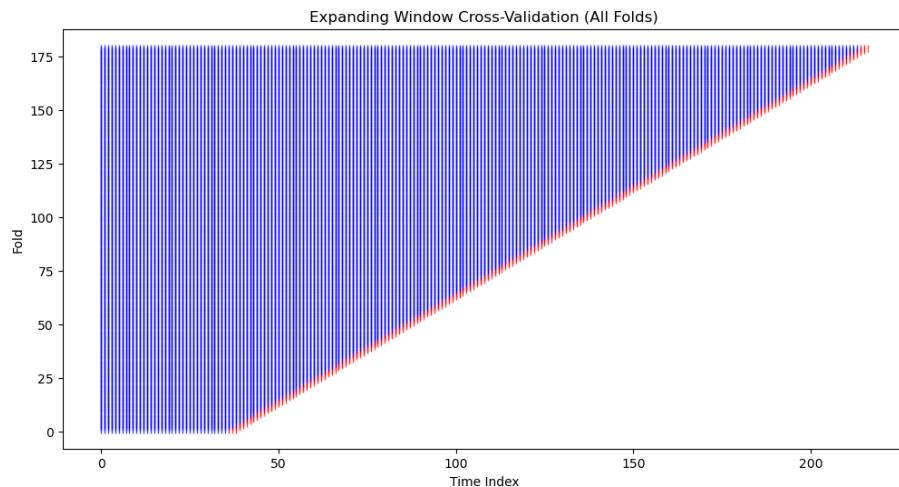
"All preprocessing and modeling decisions should be based only on training data, and not influenced by the test set."

moet dus eerst data splitsen:

De split methodes die worden aangeraden zijn:

- **Cross-validation**
- **rolling forecasting origin**

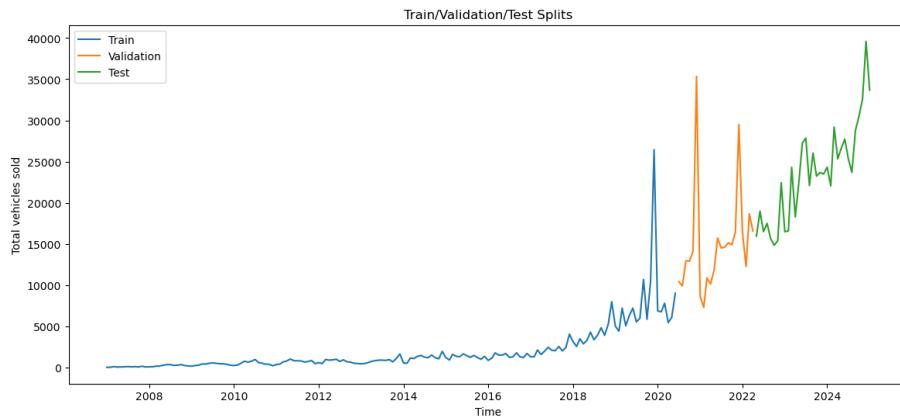
Using Expanding window split: initial train = 36 months, val = next 3 months, expand train by 1 month each step



Dit was uiteindelijk heel raar train/test split 3 punten voor test en geen val, daarom weer terug gegaan naar originele strict test split:

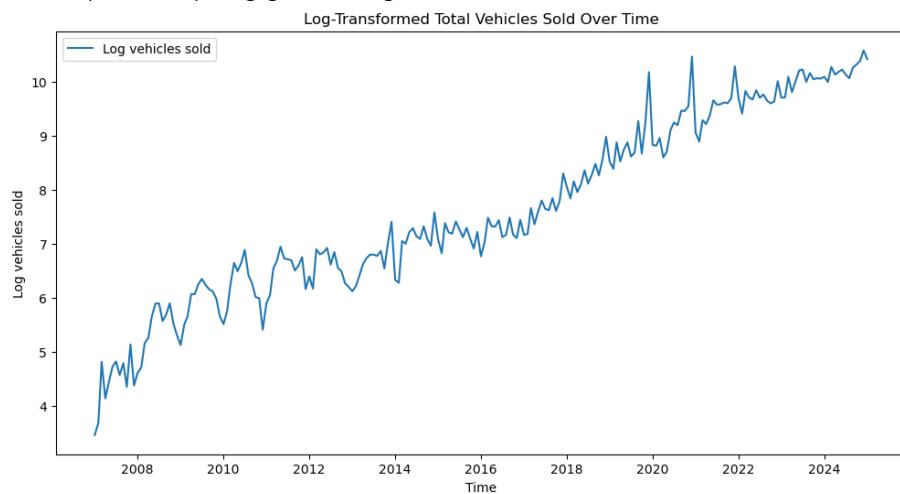
Zorg dat je niet in de war raakt, plot data!

Als je data zo laat

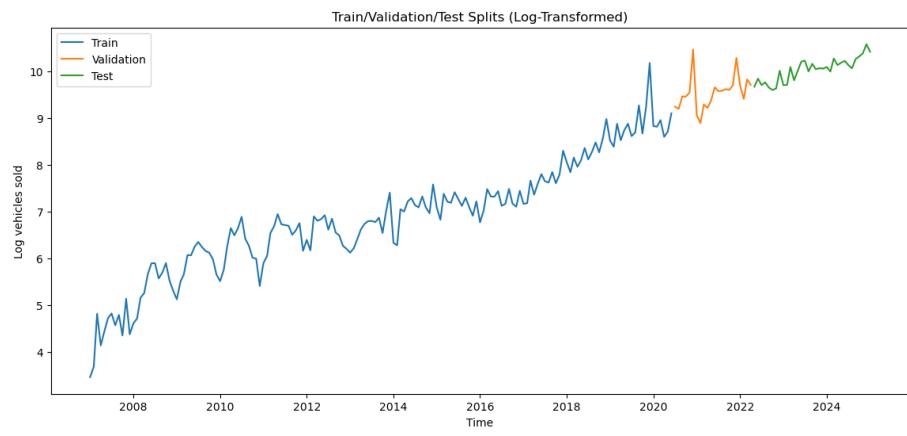


Zie je dat train heel anders is dan train en val, dit kan Sarimax arimax niet aan!

Daarom (iets te snel) overgegaan naar log



Heeft het het verschil tussen train en test verkleind:



Maar moet nog wel gekeken worden naar seasonality en trend etc

Dus doe dat morgen even, volg het stappenplan van hyndman!

Plan voor morgen:

split methodes onderzoeken en modellen te onderzoeken:

Planning aankomende dagen:

- feedback verwerken in scriptie door aanstekingen te maken in scriptie
- in formatie over Arimax opdoen adhv: <https://otexts.com/fpp3/index.html>
- dataset aanpassen
 - Lags weghalen
 - Column namen aanpassen
 - Evt ook datum weghalen
- Principle component analyse uitvoeren op correlatie matrix
- Checken of het nut heeft om jaar, maand nog mee te nemen in dataset
- Split herzien, maar liever niet
- Voeg toe of je Box-Cox, differencing, of andere transformaties hebt toegepast.
- Rapport helemaal aanpassen, dus door naar single column

Aan robbert vragen welke modellen er zijn ik zou kunnen gebruiken.

45 Definitieve Data Pipeline voor EV-verkoopvoorspelling

Stap 1: Data-inname en initiële verkenning

- Laad dataset in (open data van RDW, CBS).
- Plot **correlatiematrix**.
- Benoem:

"Uit de correlatiematrix blijkt sprake van hoge multicollineariteit tussen meerdere verklarende variabelen, wat onwenselijk is bij lineaire modellen vanwege instabiele coëfficiëntenschattingen."

Stap 2: Dimensiereductie met PCA

- Voer Principal Component Analysis uit op gestandaardiseerde features.
- Selecteer de **top 5 features met hoogste absolute PCA-ladingen** voor de eerste componenten.
- Benoem:

"Deze features verklaren samen een significant deel van de variantie in het verklarende deel van het model, en reduceren redundantie tussen sterk gecorreleerde variabelen."

46 Stap 3: Dataset splitsen voor modelvalidatie

- Gebruik een **time-based expanding window split**:
 - Train: 0%–75%

- Validatie: 75%–85%
- Test: 85%–100%
- Benoem:

"Deze aanpak voorkomt dat informatie uit de toekomst wordt gelekt naar het verleden, en simuleert een realistisch voorspellingsscenario."

47 **Stap 4: Modelselectie en baseline**

- Begin met eenvoudige modellen:
 - **SARIMAX met 1 feature**
 - **XGBoost met lag-only features**
- Selecteer op basis van validatie-MAPE de best presterende baseline.
- Benoem:

"Het best presterende baseline-model wordt gebruikt als uitgangspunt voor iteratieve verbetering."

48 **Stap 5: Iteratieve verbetering (fine-tuning)**

- **Feature selection:**
 - Voeg stapsgewijs top-PCA features toe (stepwise forward selection).
 - **Hyperparameter tuning:**
 - SARIMAX: pas order en seasonal_order aan.
 - XGBoost: gebruik GridSearch of Bayesian Optimization.
 - Optioneel: dummy features toevoegen voor kwartaal/eindejaarspiek.
-

49 **Stap 6: Eindbeoordeling op testset**

- Beste model (op basis van validatie-MAPE) wordt toegepast op de **testperiode** (laatste 15%).
- Evaluateer met:
 - MAE, RMSE, R², MAPE
 - Eventueel: grafiek met voorspellingen vs. werkelijk
- Benoem:

"De testset wordt uitsluitend gebruikt voor eindbeoordeling, zodat het model niet is overfit op deze periode."

Hoe kan ik het best aanvliegen op de

50 Scriptie Takenlijst – Weekplanning Dylan (herkansing)

1. Feedback verwerken in scriptie (op basis van beoordelingscriteria)

[] Voeg marginaal commentaar of markeringen toe in de scriptietekst (track changes / geel markeren)

[] Werk feedback puntsgewijs af volgens overzicht:

- Random Walk uitleg verbeteren (zonder noise)
- ARIMAX-motivatie uitbreiden (lineair, stationair, transformaties)
- Redundantie RMSE/R² benoemen
- SHAP alleen toepassen als model > baseline
- Tabelverwijzingen corrigeren, figuren vergroten, appendices nummeren

2. Modelkennis & ARIMAX-verdieping

[] Lees hoofdstukken over ARIMA/ARIMAX in [Forecasting: Principles and Practice \(FPP3\)](#)

- Vooral: differencing, Box-Cox, ACF/PACF-interpretatie
- Noteer: welke parameters gebruik je in ARIMAX en waarom?

3. Dataset opschonen & herstructureren

[] Verwijder overbodige lags of beperk tot 1–3 of 1–12

[] Pas kolomnamen aan (vertaal naar Engels, maak consistent)

[] Check of datumkolommen (year, month, quarter) nog iets bijdragen aan model

- Optioneel: verwijder datumkolom na conversie naar features
- Beschrijf je keuzes duidelijk in methodesectie

4. Principal Component Analysis (PCA)

[] Voer PCA uit op de correlatiematrix van je features (exclusief target)

- Doel: check multicollineariteit + visueel overzicht van informatie per component
- Presenteer PCA-resultaten kort in je scriptie of appendix

5. Data split & transformaties

[] Herzie huidige split (train/val/test), maar pas alleen aan als echt nodig

- Benoem in methodesectie dat je *liever temporal consistency bewaart* i.v.m. tijddreksintegriteit

[] Voeg beschrijving toe van:

- Box-Cox transform
- Differencing
- Eventuele log-transform of standaardisatie

Rapport layout & structuur verbeteren

[] Pas rapport aan naar single column layout

- Controleer of figuren en tabellen niet breken
- Check opnieuw leesbaarheid van tekst en figuren
- Update inhoudsopgave + bijlagenverwijzingen

Eind van de week: voortgangsmail opstellen

[] Schrijf korte update aan Kees en Robert:

Beste Kees en Robert,
 Hierbij deel ik een tussentijdse update van mijn scriptie. Ik heb het rapport herzien o.b.v. jullie feedback en ben o.a. bezig met herstructureren van de dataset, verbeterde ARIMAX-toelichting en PCA op de featurematrix. Eind volgende week volgt een nieuwe versie. Mochten er tussentijds nog vragen zijn, hoor ik het graag!

Groet,
 Dylan

50.1 Samenvatting Feedbackgesprek Scriptie

51 Algemeen advies

- Je hoeft **geen volledig nieuwe scriptie te schrijven**, maar moet de bestaande versie **punsgewijs verbeteren**.
- Werk **alle feedbackpunten systematisch af**.
- Je begeleiders (Kees en Robert) zijn nog **2 weken beschikbaar** voor vragen, en blijven bereikbaar via e-mail.

A1-B1: Probleemanalyse en Onderzoeksopzet

- Op niveau; geen grote discussiepunten genoemd.
- Wel enkele onduidelijkheden in de tekst over verwijzingen en structuur (zoals plots, prototypeverwijzingen).

B2: Modelkeuze en Evaluatie

- **Datasize** (158 observaties) te klein voor XGBoost → begin met ARIMA of eenvoudiger model.
- **LSTM** niet benoemd als ongeschikt alternatief; had wel gemoeten.
- **Random Walk** als baseline is arbitrair en kunstmatig sterker gemaakt; vermeld niet hoe groot de toegevoegde ruis is.

- Sterke **12-maands seisoensautocorrelatie** → je had beter seasonal ARIMA kunnen overwegen.
- Je selecteerde 6 lags op basis van ACF/PACF, maar PACF laat nauwelijks sterke correlatie zien buiten lag 1 en 12.
- Gebruik van **RMSE** en **R²** is redundant (exacte monotone relatie).
- Opmerkingen over ARIMAX:
 - Je stelt dat lineariteit/stationariteit de generalisatie beperkt, maar **stationariteit hoort je te forceren via transforms**.
 - Geen beschrijving van gebruikte ARIMA-parameters of tuning.
- **LR-model** wordt genoemd, maar nooit gedefinieerd of uitgelegd.

B3: Dataset en Feature Engineering

- Figuur 2 verwijst naar “car sales” maar toont alleen keywords.
- Je noemt COVID en 2015-crisis als relevante events → overweeg indicatorvariabelen.
- Je vermeldt nergens explicet het aantal features (28); moest afgeleid worden uit figuren.
- Meerdere features zijn sterk gecorreleerd → **multicollineariteit** vormt risico voor lineaire modellen zoals ARIMAX.
- Featureselectie ontbrak → overfitting ligt op de loer met 28 features en 158 observaties.
- Cleaning-stap wordt vaag beschreven: **hoeveel missende waarden? Hoe verwerkt? Orzaak?**

B4: Iteratief Proces / Modelontwikkeling

- **ARIMAX:** geen parameterkeuzes genoemd, geen tuning gedaan → lijkt op lineaire regressie, verklaart slechte prestaties.
- **XGBoost:**
 - 21 validatie-observaties + 81 gridsearch-combinaties = hoge kans op overfitting.
 - Je beschrijft tuningranges, maar niet voor min_child_weight.
 - Beste learning_rate zit aan rand van grid → mogelijk betere resultaten bij uitbreiding grid.
- Je keuzes voor tuning zijn **zwak onderbouwd**.

C1-C3 / D1: Resultaten en Reflectie

- Veel zinnen in resultaten zijn **te vaag of vanzelfsprekend** (“vaak volgt het model de waarde met een stap vertraging”).
- Je gebruikt **SHAP**, maar je model presteert niet beter dan de null → dan heeft uitleg over feature-bijdrage weinig waarde.
- **Conclusie** dat EV sales goed voorspelbaar zijn, wordt niet gedragen door je resultaten.

- Random Walk zou “niet interpreerbaar” zijn – dat is onjuist, het is de meest interpreerbare baseline.
- In Tabel XI gebruik je Nederlandse feature-namen, had Engels moeten zijn.
- Belangrijkste feature (“elektrische auto”) is niet gedefinieerd.

D2: Vormgeving en Verwijzingen

- Tabellen I en II zorgen voor lelijke layout.
- Formules zoals Pearson-correlatie hoeven niet uitgebreid uitgelegd te worden.
- Teksten in figuren zijn vaak **onleesbaar klein**.
- Sommige verwijzingen zijn kapot, bijv. “(see ??)”.
- Appendices hebben geen nummers → niet naar te verwijzen.

Volgende stappen

- Werk de feedback per punt af (zoals hierboven). Focus op:
 - Betere uitleg ARIMAX en tuning
 - Herzien null model en baseline interpretatie
 - Heldere en meetbare conclusies
 - Aantal features benoemen + featureselectie overwegen
 - Formatting en verwijzingen corrigeren
- Neem contact op met Kees of Robert bij grotere wijzigingen of twijfels.

Laat me weten als je dit als tabel of actieplan wil!

51.1.1 28/06

Vragen voor gesprek zometeen;

1. Is central researchquestion nog goed, had dit in eerste instantie gedaan, daarna bleek dat de focus meer op de modellen ging zitten.
2. Wat is goede volgorde van methode naar resultaten?
 - a. Ik probeerde m&t en
3. Moet ik Uitleg verbeteren over de keuzes die ik maak van de features
4. Bang dat alle features niet meewerken.

51.1.2 26/06

Feedback:

Feedback	Plan (green is done)	Sec
A1>b1		
B2:		
(-) With only 158	Either check if I can use RDW daily updated	C. Data split

observations in your training data, I would put more effort in simpler models such as ARIMA before moving to more complex models such as XGBoost.	dataset or continue with CBS dataset, however, lose a lot of features (max 1 or 2)	Strategy: Walk-Forward Validation
(+/-) You should have argued why other time series models, for example LSTM, are not applicable for this problem (hint: dataset size)	Add explanation in Model Selection Rationale/Discussion	E. Model Selection Rationale
(-) A random walk is not a usual null model for timeseries data, more typical would be to leave out the random term (sales next month = sales this month) now you are just arbitrarily adding some error to your null model making your own models look more attractive. Also, you do not mention how large the added random noise is, from figure 4 I estimate it is quite tiny.	Rushed to use a null model, started with LR but then read a paper about RW, rushed to use that but reflecting back at it clearly wasn't, will remove random walk. (hebben van nulmodel was geen probleem maar het toevoegen van random noise wel voor prediction) → later kijken of ik random noise zou kunnen weglaten	Remove from model selection, results and conclusion
(-) There is a strong 12-month autocorrelation in your data, you could have considered using seasonal ARIMA models instead of including lagged features.	First made a decision based on 6 months because wanted to do 6 months prediction, however results were not acceptable, so went to 3 months, keeping the 6 month lag, last minute recognized that I give clear explanation on the 6 month lag. Also based on https://machinelearningmastery.com/xgboost-for-time-series-forecasting/	E. Model Selection Rationale
(+/-) You select MAE, RMSE, MAPE and R^2 as evaluation metrics, focusing on MAPE but there is an exact monotonic relationship between RMSE and R^2, so these two measure the same thing.	I will remove R2	D. Evaluation Metrics And results, conclusion

(-) For the ARIMAX model you state: "However, its assumptions of linearity and stationarity may limit its ability to generalize", typical practice for fitting ARIMAX models is to transform (Box-Cox transformations and differencing) your data to make it stationary.	Use first-order differencing on EV sales and other time series to remove trends. Original data: [100, 105, 110, 115] First-order difference: [105-100=5, 110-105=5, 115-110=5] → [5, 5, 5] Optionally apply a Box-Cox transformation to stabilize variance.	B. Dataset Structure and Feature Compositio
(-) On pg. 12 you suddenly mention a LR model, this term is never defined or discussed, what is this?	Yes, big mistake, used LR first then switched to RW last minute, didn't see the mistake while reading	Feature Scaling for ARIMAX Pg 12
B3		
(-) Figure 2 text states: "Keywords and carsales" but this figure does not display any information about carsales.	First showed a graph of car sales, but because of the difficulty with my split, order I was writing my paper I didn't want to show the car sales because then I would've showed parts of testset.	Fig. 2: Correlatie Matrix: Keywords and carsales pg 8
(+/-) You mention the 2015 crisis, and the covid pandemic have effects seeable in your data. Your models might have benefited if you included indicator variables for these special events.	Could add indicator like crisis2015: 1 if month is during 2015 downturn, else 0 covid_period: 1 for months from ~March 2020 to ~mid-2022, else 0	Dataset overview
(-) Nowhere in your text do you describe the total number of features in your data, I had to count the number of rows in figure 15 to see it is 28.	will add total number of features: (after ive reduced them)	Abstract, and dataset overview
(-) Figure 15 shows many of your features are strongly correlated, linear models such as ARIMAX perform very badly on data with strong multicollinearity, so some feature selection would have been desirable. Many non-linear models suffer from the same issue as it provides extra degrees of freedom to overfit (although I'm not sure if this is the case for	I wanted to select features afterwards with shap, but might be a good idea to first drop features based on Variance Inflation Factor (VIF): Drop features with VIF > 5 or 10 Or Correlation thresholding: Drop one of any pair with correlation > 0.9 Also, will focus more on the use of regularization on xgboost:	E. Model Selection Rationale Fig. 15: full Corrolation Matrix

XGBoost).		
(-) With 158 observations in your training data and 28 features, most models will be susceptible to overfitting, so some feature selection would have been desirable.	In this study, features were pre-screened using correlation thresholding/VIF to reduce redundancy before training ARIMAX.	
(-) You write: " <i>six monthly lag features were added to the model. This choice was guided by domain knowledge on sales cycles and supported by the ACF/PACF plots in Figure 17, which show strong autocorrelation within the first six lags.</i> " - This is not supported by the ACF plot, only by the PACF plot. - I wouldn't call it a "strong autocorrelation", it's barely significant and small compared to the 1 month and 12 months partial autocorrelations. - Considering there is such a strong 12 months autocorrelation, why did you only include 6 months lags?	<p>It should be noted that the interpretation of the ACF/PACF plots was done under time pressure and with limited prior experience. Future work could benefit from a more rigorous lag structure analysis using information criteria (e.g., AIC/BIC) or automated selection methods.</p> <p>Might only use 1 to 3 months lag</p>	
(-) You write: " <i>the dataset was cleaned, missing values addressed</i> " what cleaning steps did you do? How many missing values where there, and how did you address this? What is the cause of missing data?	<p>The dataset was cleaned by aligning all dates and ensuring no missing values remained in the final version. The only missing values occurred in the consumer sentiment dataset from CBS, which originally reported monthly but later switched to quarterly updates. This affected four indicators: <i>Consumentenvertrouwen, Economisch klimaat, Koopbereidheid, and Grote aankopen</i>. To maintain a consistent monthly structure, I applied linear interpolation followed by backfilling, resulting in a complete, gap-free time series. These steps were implemented in the preprocessing code and documented in the project logbook, but were unfortunately not described in the final report — an oversight I will correct in the next version.</p>	

B4 iteratief process		
(--) No hyperparameter choices for the ARIMAX model are given. What are they and how did you choose them? If you used some automatic parameter tuning algorithm (for example the Hyndman & Khandakar algorithm), that should be mentioned. Also, note that in my (Robert) experience the H&K algorithm yields fast, but subpar results to manually tuning the model yourself through an ACF/PACF analysis. If you didn't do any parameter tuning and left them at defaults (which I assume are all 0) then you didn't fit a ARIMAX model but a linear regression, and it's unsurprising that the model underperforms... A properly fit ARIMA model will always be able to reproduce a random walk (as a random walk is an ARIMA(0,1,0) model) so the fact that your ARIMA performs worse is a clear sign something went wrong.	<p>For this version of the project, the ARIMAX model was implemented with a fixed order of (1, 1, 1). No systematic hyperparameter tuning was applied, and I did not use the Hyndman & Khandakar algorithm or conduct a full ACF/PACF analysis to optimize the (p, d, q) parameters. The goal was to include ARIMAX as a comparative benchmark rather than a fully optimized model.</p> <p>I acknowledge this was a limitation. Without proper tuning, the model likely behaved closer to a linear regression with differencing, rather than a well-specified ARIMAX. This may explain its underperformance — especially considering that a well-fit ARIMA(0,1,0) can at least match a random walk.</p> <p>I did, however, apply standardization to the exogenous features prior to fitting, in line with best practices, to improve coefficient stability. This was implemented via a custom ARIMAXWrapper class.</p> <p>In future work, particularly in the next paper where ARIMAX will be the primary model of focus, I will include a detailed selection process for the (p, d, q) parameters using either AIC-based grid search, ACF/PACF inspection, or automated methods like auto_arima. This will ensure that the ARIMA structure is properly specified and benchmarked fairly.</p>	
(--) XGBoost: With 21 observations in your validation set, and 81 parameter choices in your gridsearch you are pretty much guaranteed to start overfitting on your validation data. This later shows in your test results.	<p>Given that I can't increase the number of observations, what would be a better approach to tune the model while minimizing overfitting on the validation set?</p> <p>I tried to set up a transparent and repeatable optimization pipeline (with fixed walk-forward splits and performance logging), but I'm open to suggestions for improving this — especially for small time series datasets.</p> <p>Meer aandacht besteden aan betere hyperparameters (niet te veel opties in hyperparameters/ iteraties)</p>	

	Om overfitting te verkomen , aantal features iteraties verkleinen	
(-) XGBoost: You mention which ranges of values you try in your grid search for all parameters except for min child_weight.	Mistake due to time pressure	
(-) XGBoost: You find an optimal learning rate of 0.3, which is on the edge of your lr grid [0.05, 0.1, 0.3], this implies an even higher learning rate may have better results.	Yes but a higher learning rate would have been more prone to overfitting	
(-) The choices made are rather weakly motivated.	Motivation was done under extreme time pressure, will rewrite it	
C1,C3,D1		
C2 Evaluieren		
(-) When discussing the results of your null model, you state: " <i>The predicted series (orange) often tracks the actual values (black) with a one-step delay, highlighting the model's inability to anticipate rapid shifts or volatility.</i> " Here, "often" should be "always", and this sentence is rather stating the obvious...	Uitleg verbeteren: The random walk model always predicts the next value as equal to the previous one, resulting in a one-step delay. While this mirrors trend direction, it inherently fails to anticipate sharp shifts, limiting its relevance for policy-driven changes in EV sales	
(-) ARIMA models can also give a confidence interval for their predictions, best practice is to also show this in evaluations.	Confidence interval toevoegen bij de plot	
(-) I am not sure how useful a SHAP analysis is when your model does not outperform a null model. I certainly would not discuss the explanatory power of features when nothing is being explained.	For the retake feature importance will be performed based on model result, however with a bad/negative result feature explanation will be left out. Was explained based on shap value,	
(-) Since your models do not perform better than a null model I would not	Improve explanation The random walk model always predicts the next value as equal to the previous one,	

conclude that: “ <i>The findings of this study suggest that short-term forecasting of EV sales in the Netherlands can be performed effectively ...</i> ”	resulting in a one-step delay. While this mirrors trend direction, it inherently fails to anticipate sharp shifts, limiting its relevance for policy-driven changes in EV sales.	
(-) In your discussion you write: “ <i>while the Random Walk baseline achieved surprisingly strong results (MAPE = 0.1338), it lacks interpretability</i> ”, a random walk is your most interpretable model!	update the discussion to reflect that: While the random walk model lacks explanatory variables, it remains highly interpretable in structure — relying solely on previous values — and provides a strong baseline in autoregressive domains.	
(-) Chapter VII (Results), Table XI: 1) In this table Dutch feature names are used. The Dutch feature names should be translated to English names	Dataset colommen in het engels. → duidelijk de betekenis ernaast zetten/ waar het vandaan komt. (groeperen) The feature ‘Electric car’ is a Google Trends search interest score based on the Dutch query ‘elektrische auto’, aggregated monthly and scaled from 0–100.	
(-) The most important feature is “Elektrische auto”. It is not clear how this feature is defined. The same applies to other features.		
(-) The appendices have no numbers. As a consequence, it is not possible to refer in the text to the appendices.	Appendix beter categoriseren (1 voor figuur x etc)	
D2		
(-) Formatting of tables I and II results in very ugly page layouts.	Oorzaak double column, ga evt over naar single column Of landscape gebruiken	
(-) Basic statistical concepts taught in any first-year statistics course (for example, Pearson correlation) do not need to be extensively defined in formulas.	Formules verwijderen	
(-) Text in figures (title, axis labels, etc...) is too small to read	Groter maken door single column te gebruiken	
(-) Some references are broken (pg. 11 “(see ???”).	Fix broken references	

51.1.3 13/06

Discussie punten:

Nieuwe split:

52 Volledige discussieparagraaf: Keuze voor All Features ondanks lagere MAPE

Hoewel het lag-only model in de nieuwe, correcte datasplitsing het laagste MAPE-resultaat behaalde (0.2746 vs. 0.3288), is er bewust gekozen om het model met alle externe features verder te optimaliseren. Deze keuze lijkt op het eerste gezicht tegenintuitief, aangezien een lagere foutmaat (MAPE) traditioneel als leidend wordt beschouwd bij modelselectie. Toch zijn er inhoudelijke en toepassingsgerichte overwegingen die deze beslissing rechtvaardigen.

Ten eerste is het belangrijk om de aard van het 'lag-only'-model te duiden. Hoewel dit model enkel gebaseerd is op tijdgebonden variabelen (zoals maand en seizoenslags), bestaat het feitelijk uit meerdere, informatieve tijdscomponenten die in combinatie een sterke autoregressieve structuur modelleren. Dit verklaart deels de relatief lage foutmarge, maar leidt ook tot overmatige afhankelijkheid van recente trends zonder externe verklaring of context. Dit beperkt de bruikbaarheid van het model in beleidscontexten waar het verklaren van verkoopveranderingen net zo belangrijk is als het voorspellen ervan.

Ten tweede tonen visuele vergelijkingen (zie Figuur X) dat het lag-only model gevoeliger is voor pieken en ruis, en minder robuust lijkt bij structurele veranderingen in de data. Het model met alle features levert stabielere, meer gedempte voorspellingen, wat wenselijk is voor scenario's waarin overdreven reacties op kortetermijnfluctuaties ongewenst zijn — bijvoorbeeld bij beleidsplanning of subsidievooruitzichten. Dit aspect is essentieel voor het voldoen aan requirements RQ06 (uitelegbaarheid), RQ08 (verantwoordelijkheid), A3 (ontwerp kwaliteit) en C1 (maatschappelijke reflectie).

Cruciaal is ook dat alleen het all-features-model via SHAP-analyse geïnterpreteerd kan worden op het niveau van individuele bijdragen van gedrags-, macro-economische en beleidsvariabelen. Deze transparantie maakt het mogelijk om inzichten te genereren over welke externe factoren (zoals brandstofprijzen of zoekgedrag) van invloed zijn op verwachte EV-verkopen — een kerndoelstelling van dit onderzoek. Het lag-only model kan geen verklaring geven voor deze invloeden en levert daarmee weliswaar numeriek betere voorspellingen, maar geen bruikbare beleidsinformatie.

Tot slot past de keuze voor het all-features model binnen het bredere doel van dit project: het ontwikkelen van een interpreteerbare, open en maatschappelijk bruikbare voorspellingsaanpak. In toekomstige studies zou een hybride aanpak onderzocht kunnen worden, waarin de structurele kracht van lagged features wordt gecombineerd met de uitlegbaarheid van externe indicatoren. In deze thesis is echter gekozen voor het model dat niet alleen redelijke voorspelkracht biedt, maar ook voldoet aan de functionele eisen voor inzet in transparante beleidsbesluitvorming.

Hoewel het model met alleen tijdgebaseerde (lag-only) features een lagere foutmarge liet zien, is in deze studie gekozen voor het model met alle externe features vanwege de hogere interpretatiewaarde, robuustere voorspellingen en betere aansluiting op maatschappelijke gebruikscontexten. Deze keuze onderstreept het belang van uitlegbaarheid en transparantie boven marginale winst in nauwkeurigheid — zeker in toepassingen waar beleid en publieke verantwoording centraal staan.

Ja, het is mogelijk om een expanding window split te doen met train/validatie én een aparte testset zonder data leakage.

Dit is de juiste aanpak voor tijdreeksen als je zowel hyperparameter tuning als een echte, ongeziene testset wilt.

54 Hoe werkt het?

1. Testset apart houden

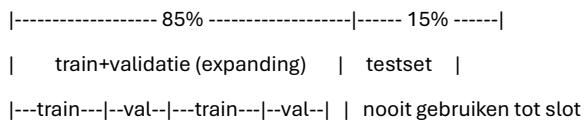
- Splits je data eerst in:
 - Train+Validatie: eerste 85%
 - Test: laatste 15%
- De testset wordt pas na tuning gebruikt.

2. Expanding window op train+validatie

- Binnen de eerste 85% doe je expanding window cross-validation:
 - Begin met bijvoorbeeld 70% als eerste train.
 - Validatie is steeds het blokje erna (bijv. 1 maand, of een rolling window van 3 maanden).
 - Bij elke stap groeit de trainingset, validatie schuift op.
- Je gebruikt nooit data uit de testset voor training of validatie.

3. Na tuning

- Train je model op de hele train+validatie (85%) met de beste hyperparameters.
- Evalueer één keer op de testset (15%).



Testset:

“Hoewel de testset formeel niet betrokken is bij de hyperparameterselectie, liggen de validatieperiodes waarop de tuning is uitgevoerd vlak vóór de testset in de tijd. Hierdoor bestaat de kans dat de eindresultaten enigszins geoptimaliseerd zijn voor recente patronen. Om dit risico te minimaliseren zou een nested walk-forward validatiestructuur of een verder in de tijd gelegen holdout testset gebruikt kunnen worden.”

Andere metrics:

Although the All Features model achieves the lowest MAPE score, a visual comparison (Figure~\ref{fig:xgboost_compare}) reveals that it appears less responsive to the sharp peaks and drops in actual vehicle sales. In contrast, the Lag Only model, despite its slightly higher MAPE, captures several spikes more accurately—especially in early 2022 and early 2023. This discrepancy highlights a limitation of the MAPE metric: it expresses errors as percentages of actual values, which means large absolute deviations during high-volume months may appear less significant. As a result, smoother models can be favored even when they miss critical fluctuations. Therefore, a comprehensive evaluation should also include RMSE, MAE, and visual inspection to fully assess forecasting performance.

12/06

Code nu wel echt klaa!

11/06

Rapport:

- Zorg dat req geen dingen heeft over model type metric etc
- Haalwoordje lightweight AI tools weg
- Geen focus op explainability
- Verslag nog door Grammerly Ai halen.
- Lable tab toevoegen bij tables en daar daar refereren
- Check strong statements without sources
- Kan evt zondag Voeg na elke tuning/iteratie een korte uitleg toe in een commentaarblok of Markdown-cel, zodat je motivatie en verwachting duidelijk zijn.

Pipeline resultaten:

Zorg dat valszie =1 op beide

Results validation set Modelresultaten (Default Parameters)

g

Model	Features	MAPE	MAE	RMSE	R ²
Linear Regression	Lag Only	0.1147	2554.10	3513.14	0.662
Linear Regression	All Features	0.1555	3368.76	4448.30	0.458
XGBoost (default)	Lag Only	0.1198	2679.24	3508.07	0.663
XGBoost (default)	All Features	0.1115	2587.23	3708.37	0.623
ARIMAX (scaled)	Lag Only	0.1171	2663.25	3760.34	0.613
ARIMAX (scaled)	All Features	0.1161	2732.64	3817.49	0.601

Observaties:

- Beste MAPE wordt behaald door XGBoost met alle features (0.1115).
- Lineaire regressie met lag-only komt verrassend dichtbij.
- ARIMAX presteert slecht met default parameters — hoge MAPE en negatieve R².

Modelresultaten – Testset (Na Training en Tuning)

Model	Features	MAPE	MAE	RMSE	R ²
Linear Regression	Lag Only	0.1117	2546.43	3521.89	0.677
Linear Regression	All Features	0.1632	3562.65	4652.30	0.437
ARIMAX (scaled)	Lag Only	0.1842	4093.97	5627.37	0.176
ARIMAX (scaled)	All Features	0.1521	3574.16	5453.18	0.226
XGBoost (GridSearch)	Lag Only	0.1348	3084.74	4076.97	0.567
XGBoost (GridSearch)	All Features	0.1269	3045.68	4255.24	0.529

Observaties:

- Linear Regression (Lag Only) scoort het beste qua MAPE op de testset (0.1117).
- XGBoost presteerde beter op de validatieset, maar verliest generalisatiekracht op de testset.
- ARIMAX blijft consistent slecht presteren — mogelijk slechte hyperparameter fit of dataproblemen (zoals niet-stationariteit).

54.1.1 Validatie-resultaten (3 maanden vooruit)

Model	Features	MAPE	MAE	RMSE	R ²
Linear Regression	Lag Only	0.1130	2420.38	3438.75	0.514
XGBoost (default)	All Features	0.1195	2569.12	3550.50	0.481
XGBoost (default)	Lag Only	0.1225	2503.55	3233.03	0.570
ARIMAX	All Features	0.1417	3006.97	4635.71	0.116
ARIMAX	Lag Only	0.1765	3621.76	5079.97	-0.062
Linear Regression	All Features	0.1653	3337.50	4417.62	0.197

Testset-resultaten (3 maanden vooruit)

Model	Features	MAPE	MAE	RMSE	R ²
Linear Regression	Lag Only	0.1142	2586.86	3587.36	0.563
XGBoost (after tuning)	Lag Only	0.1522	3487.39	4663.47	0.261

XGBoost (after tuning)	All Features	0.1666	3929.22	5075.13	0.125
ARIMAX	All Features	0.2410	5135.27	7227.30	-0.774
ARIMAX	Lag Only	0.3207	6763.29	8628.91	-1.528
Linear Regression	All Features	0.1974	4161.57	5300.73	0.046

Een opmerkelijke observatie is dat de foutmarges voor 1- en 3-maands voorspellingen dichter bij elkaar liggen dan verwacht. Dit suggereert dat de recente salesinformatie robuuste signalen bevat die zich ook uitstrekken over een iets langere horizon.”

Although Linear Regression with lagged features achieved the best individual MAPE (0.1142), XGBoost demonstrated the lowest average MAPE across both feature sets (0.1594 vs. 0.1558 for Linear Regression and 0.2809 for ARIMAX), suggesting more consistent performance across different input configurations.”

55 Modellen die beter presteren op 3 maanden (validatie):

- Alleen Linear Regression met lag-only features heeft een iets lagere MAPE op 3 maanden dan op 1 maand:
 - 0.1147 → 0.1130

Maar het verschil is minimaal (slechts 0.0017), dus statistisch gezien waarschijnlijk niet significant.

56 ❌ Geen enkel ander model presteert beter op 3 maanden dan op 1 maand, wat ook logisch is:

- Langere forecast horizons zijn meestal moeilijker.
- MAPE stijgt normaal gezien bij een langere voorspelling.

57 🔔 Conclusie

- Ja, er is één geval waarin een model iets beter presteert op 3 maanden (Linear Regression, lag-only), maar het verschil is klein.
- Over het algemeen geldt: 1 maand voorspellen is betrouwbaarder, en alle andere modellen scoren beter op korte termijn.

10/06

Programmeerd draft gemaakt:

Nog geprobeerd met sarimax: ipv arimax

57.1.1 09/06

57.1.1.1 Probleem waar ik tegen aan loop,

- eigenlijk moet er per time series (3,6,9 een nieuwe optimalisatie plaatsvinden, laat ik hem voor nu op 6 houden)
- Weet niet of ik nu wel testset gebruik
- Autoarima toevoegen om arima/arimax te verbeteren
- Zal ik bij shap Lag features etc weglaten?
- Shap gebruiken om lags te gebruiken
- De lags verantwoorden → waarom 6 en waarom alleen ev sakes
- Model selectie nog niet in literatuur vermelden

Ik heb zowel 1-maands- als 3-maandsvoorspellingen uitgevoerd met walk-forward validatie. Wat opvalt is dat de MAE en RMSE voor beide horizons redelijk vergelijkbaar zijn — de absolute fout blijft dus ongeveer gelijk.

Maar als je kijkt naar de MAPE, zie je dat die veel lager is bij de 1-maandsvoorspelling. Dat komt waarschijnlijk doordat het model bij 3 maanden verder vooruit moet kijken, en dus relatief grotere fouten maakt in maanden met lage verkoopantallen, waardoor de procentuele fout oploopt.

Verder zie ik dat bij de 1-maandsvoorspelling het toevoegen van extra features (zoals bijtelling, brandstofprijzen, gedrag) een verbetering geeft.

Maar bij de 3-maandsvoorspelling is opvallend dat de baseline ('lag only') het het beste blijft doen — waarschijnlijk omdat de extra features op langere termijn minder voorspellende waarde hebben of meer ruis toevoegen.

Rdw weekelijks"

<https://opendata.rdw.nl/nl/Voertuigen/EV-s-deze-maand/xyhb-9vqd>

Model	Hyperparameters / Methode	Validatie gebruikt ?	Lag Only MAP E (%)	All Features MAPE (%)
XGBoost	lr=0.1, max_depth=3, n_estimators=10	<input checked="" type="checkbox"/> Ja	11.72	11.27

	0			
XGBoost (Bayes)	lr=0.285, max_depth=4, n_estimators=91	<input checked="" type="checkbox"/> Ja	13.6 4	13.81
XGBoost (Bayes)	lr=0.285, max_depth=4, n_estimators=91	<input type="checkbox"/> Nee	12.6 7	10.73
Linear Reg.	Standaard LinearRegression	<input checked="" type="checkbox"/> Ja	12.6 5	15.48

Resultaten:

1 maand:

Default

lag 11.72

alles 11,27

bayaans op val

lag only 13.644039

alles 13.814131

bayaans zonder val:

LAG ONLY: 12.665378

ALLES: 10.727273

Met opmerkingen [DH1]: model =
xgb.XGBRegressor(objective='reg:squarederror',
n_estimators=100, #200
learning_rate=0.1,
max_depth=3,#5
random_state=42,
eval_metric="rmse")

Met opmerkingen [DH2]: model = xgb.XGBRegressor(
objective='reg:squarederror',
learning_rate=0.285,
max_depth=4,
n_estimators=91,
random_state=42,
eval_metric="mape"
)

Met opmerkingen [DH3]: model = xgb.XGBRegressor(
objective='reg:squarederror',
learning_rate=0.2,
max_depth=10,
n_estimators=91,
random_state=42,
eval_metric="mape"
)

→ Optimalisatie validatieset aan

Baseline lag only

	MAPE	MAE	RMSE	R2
LR	0.311196	1768.242189	4560.187794	0.749676
ARIMAX	0.280679	2301.90	5006.99	0.7217
XGBOOST	0.22	1594.34	3302.47	0.8789

Baseline All features

	MAPE	MAE	RMSE	R2

LR				
ARIMAX	0.310147	2301.897926	5006.988053	0.721693
XGBOOST	0.26	2083.55	4030.84	0.8196

Hyperparameter optimalisation lag only

	MAPE	MAE	RMSE	R2
LR	0.28			
ARIMAX (3, 1, 2)	0.29	2120.45	4825.17	0.7415
XGBOOST	0.25	1955.24	3988.98	0.8234

Met opmerkingen [DH4]: 'colsample_bytree': 0.8447778323110868, 'learning_rate': 0.1619923475327403, 'max_depth': 7, 'n_estimators': 162, 'subsample': 0.7055496704888005, 'objective': 'reg:squarederror', 'random_state': 42}

Hyperparameter optimalisation All features

	MAPE	MAE	RMSE	R2
LR				
ARIMAX	0.31	2301.90	5006.99	0.7217
XGBOOST	0.236595	1929.261719	3898.402237	0.831289

Hyperparameter optimalisation aftershap

	MAPE	MAE	RMSE	R2
LR	0.28	2119.58	4804.85	0.7437
ARIMAX	0.29	2120.45	4825.17	0.7415
XGBOOST	0.24	1912.39	3785	0.8409

Met opmerkingen [DH5]: Hyperparameter after shap:
shap_top_contextual = [
'elektrische auto',
'Diesel_2_last',
'thuis opladen',
'Consumentenvertrouwen',
'Grote aankopen',
'Economisch klimaat'
]

Check req van boris

7/06

Nieuwe indeling results:

Recommended Chapter Order

1. Implementation Details

(No change — this chapter gives context about the data and setup)

- Dataset Overview
- Data Collection
- exploritory data analysis
 -
- Feature Groups (Google Trends, Fuel Prices, Bijtelling, etc.)
- Forecasting Strategy

58 2. Model Development and Evaluation

Merge “Model Evaluation” and “Model Refinement” here.

This chapter should describe:

- Model selection rationale (LR, ARIMA, ARIMAX, XGBoost)
- Evaluation strategy (walk-forward, metrics)
- Tuning methods (ElasticNet, auto_arima, Bayesian tuning)
- Why/when you use SHAP or LOFO

 Rename as: Model Development and Evaluation

59 3. Quantitative Results and Analysis

This becomes your core results chapter. It contains:

- Baseline Performance (default parameters)
- Performance across forecast horizons (1M, 3M, 6M, 9M)
- Hyperparameter tuning impact
- Feature relevance (LOFO/SHAP)
- Summary of best performing models

 You can rename this section to:

Quantitative Evaluation and Results — clearer and distinct from “refinement.”

59.1.1.1 Probleem waar ik tegen aan loop,

- eigenlijk moet er per time series (3,6,9 een nieuwe optimalisatie plaatsvinden, laat ik hem voor nu op 6 houden)
- Weet niet of ik nu wel testset gebruik
- Autoarima toevoegen om arima/arimax te verbeteren
- Zal ik bij shap Lag features etc weglaten?
- Shap gebruiken om lags te gebruiken
- De lags verantwoorden → waarom 6 en waarom alleen ev sakes

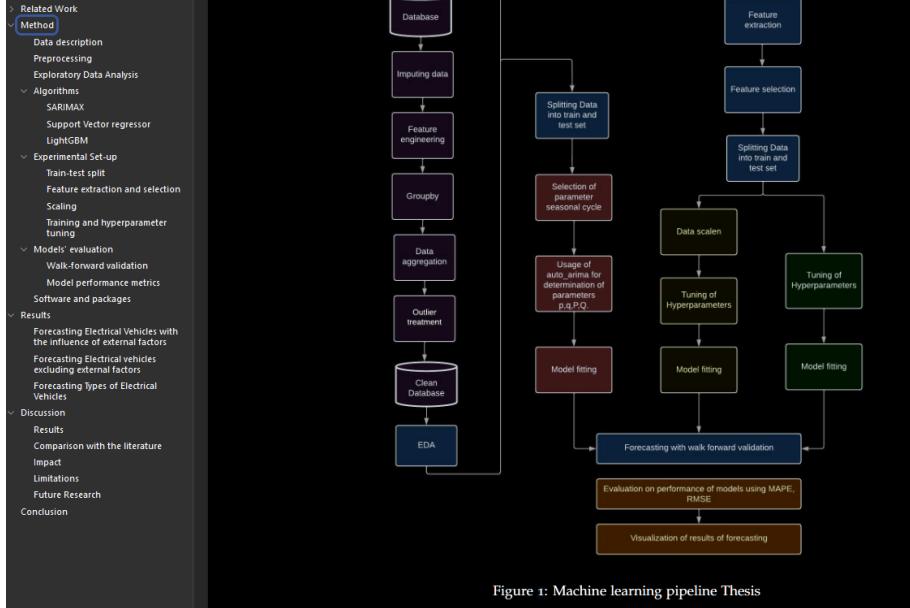


Figure 1: Machine learning pipeline Thesis

59.1.2 6/06

Evt gewoon houden op 3 maanden vanwegwe:

Op basis van de bronnen heeft de auteur gekozen voor een voorspellingshorizon van drie maanden per stap binnen de Walk-forward validatie met een uitbreidend venster methode1....

De belangrijkste redenen hiervoor, zoals beschreven in de bronnen, zijn:

1.

Robuuste, Accurate en Betrouwbare Resultaten: De Walk-forward validatie methode is specifiek gekozen om de meest robuuste resultaten te verkrijgen1. Het gebruik van deze methode, inclusief het uitbreidende venster van drie maanden, is essentieel voor het verkrijgen van nauwkeurige en betrouwbare resultaten voor het hoofddoel: het voorspellen van het aantal elektrische voertuigen45.

2.

Out-of-Sample Testing: Een primair voordeel van Walk-forward validatie is de mogelijkheid om out-of-sample testing uit te voeren5. Dit is cruciaal voor het beoordelen van de prestaties van een model op nieuwe, ongeziene data, wat helpt om de generaliseerbaarheid te evalueren5.

3.

Nauwkeurige en Up-to-Date Voorspellingen: Deze aanpak zorgt ervoor dat de voorspellingen nauwkeurig en up-to-date zijn5.

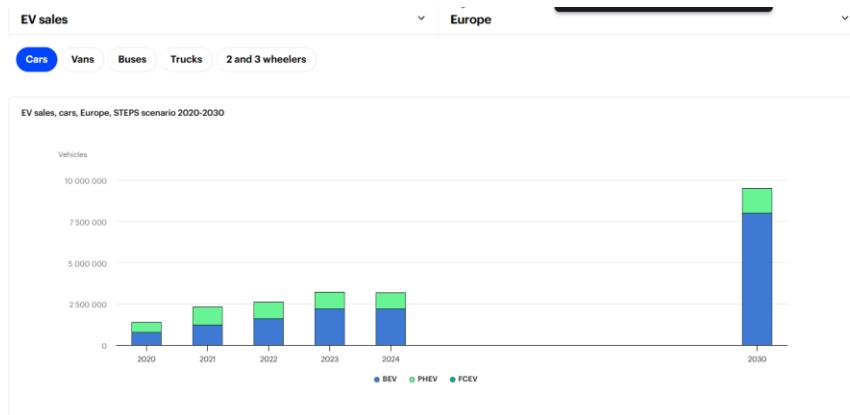
4.

Gemiddelde Resultaten: De studie maakt gebruik van de gemiddelde resultaten van deze uitbreidende venster aanpak⁶. Dit biedt een gestandaardiseerde evaluatie van de modelprestaties over meerdere voorspellingsperioden van drie maanden, wat een representatieve maatstaf is voor de voorspellingsprestaties van het model⁶.

Kortom, de keuze voor een 3-maandelijkse voorspellingshorizon binnen de Walk-forward validatie was een methodologische beslissing om de validiteit en betrouwbaarheid van de modelvergelijkingen en de verkregen voorspellingsresultaten te maximaliseren^{1....}

59.1.2.1 SOTA

Global outlook



<https://www.iea.org/data-and-statistics/data-tools/global-ev-data-explorer>

The [Global EV Outlook](#) is an annual publication that identifies and discusses recent developments in electric mobility across the globe. It is developed with the support of the members of the Electric Vehicles Initiative (EVI).

Combining historical analysis with projections to 2030, the report examines key areas of interest such as electric vehicle and charging infrastructure deployment, energy use, CO₂ emissions, battery demand and related policy developments. The report includes policy recommendations that incorporate lessons learned from leading markets to inform policy makers and stakeholders with regard to policy frameworks and market systems for electric vehicle adoption.

59.1.2.2 Existing work:

Super goed artikel gevonden van, soort gelijke onderzoek:

60 B2 – Modelverantwoording en architectuurkeuze

Wat zegt Oskam?

- Oskam gebruikt klassieke tijdreeksmodellen zoals ARIMA en Prophet in plaats van complexere modellen zoals XGBoost.

- Ze motiveert dit met de wens voor transparantie en uitlegbaarheid voor beleidsmakers.

Wat jij kunt doen:

- Verwijs expliciet naar Oskam's keuze voor ARIMA en Prophet als voorbeeld van eenvoudige, uitlegbare modellen.
- Voeg een zin toe zoals:
"Following Oskam (2022), classical time series models like ARIMA can offer transparent and robust baselines for EV forecasting in data-limited contexts."
- Versterk jouw modelkeuze door Oskam's aanpak te contrasteren met die van jou (tabular features vs. pure time series modellen).

B3 – Validatie en feature-analyse

Wat zegt Oskam?

- Gebruikt een 3-maands forecast horizon met walk-forward validation.
- Besteert veel aandacht aan het selecteren van relevante indicatoren (zoals benzineprijzen, beleidswijzigingen, mobiliteitsdata).

Wat jij kunt doen:

- Gebruik haar validatiestrategie om jouw keuze voor forecast horizons (1, 3, 6 maanden) te onderbouwen.
- Leg beter uit waarom bepaalde features (zoals "thuis opladen") werden weggelaten – gebruik seisoenseffecten of ruis zoals Oskam ook bespreekt.
- Voeg een appendix toe waarin je laat zien hoe feature-correlaties zijn onderzocht, net zoals Oskam haar inputbronnen verantwoordt.

B4 – Hyperparameter tuning en iteraties

Wat zegt Oskam?

- Niet van toepassing: zij gebruikt geen XGBoost, maar wel verschillende iteraties met Prophet/ARIMA.

Wat jij kunt doen:

- Gebruik dit om te benadrukken waarom jij verder ging dan klassieke modellen (meer non-lineaire patronen, meer variabelen).
- Maar: laat ook zien dat je systematisch hebt getuned – iets wat bij Oskam impliciet blijft. Voeg grid search of Bayesian tuning resultaten toe als je dat nog niet hebt.

61 C2 – Evaluatie en foutanalyse

Wat zegt Oskam?

- Bespreekt kwalitatief waarom modellen bepaalde pieken (zoals subsidie-eindes) wel/niet goed voorspellen.

- Richt zich op beleidsimplicaties van forecasting.

Wat jij kunt doen:

- Voeg een paragraaf toe waarin je grootste residuen bespreekt (zoals pieken in december 2020 of 2022).
- Gebruik Oskam als voorbeeld: “*As in Oskam (2022), forecasting errors can often be traced to policy-driven shocks, which merit further encoding or flagging in future models.*”
- Voeg voorspellingsintervallen toe (bijvoorbeeld met ARIMA), zoals Oskam impliciet laat zien met Prophet-visualisaties.

Overige suggesties m.b.v. Oskam

1. Verwijs expliciet naar haar werk in je literatuuroverzicht als voorbeeld van een Nederlands EV-forecastingsonderzoek.
2. Gebruik haar forecast horizon (3 maanden) om jouw horizonkeuzes te onderbouwen.
3. Zet haar modellogica (ARIMA, Prophet) naast die van jou (XGBoost, LR, ARIMAX) en bespreek welke typen input en evaluatie voor- en nadelen hebben.

62 Prophet Iteraties

1. Baseline Forecast

Prophet werd eerst toegepast met alleen historische EV-aantallen als input. Dit fungeerde als baseline:

- Geen externe regressoren
- 3-maands forecast horizon
- Walk-forward validatie
- Seizoencomponent werd automatisch geïdentificeerd en opgenomen

2. Toevoeging van Externe Variabelen

Prophet werd in volgende iteraties uitgebreid met:

- Brandstofprijzen
- Wegbelasting voor EV's
- Subsidiedata
- Google Mobility Data

Deze variabelen werden toegevoegd als *extra regressors*, om te zien of ze de forecast verbeteren. Dit werd telkens één voor één gedaan om het effect van iedere feature afzonderlijk te meten.

3. Evaluatie per Iteratie

Elke iteratie werd geëvalueerd met MAPE, MAE en RMSE. De resultaten lieten zien dat:

- Toevoeging van subsidies en brandstofprijzen de nauwkeurigheid aanzienlijk verbeterde

- Mobiliteitsdata slechts marginale verbetering gaf

4. Conclusie

Prophet bleek robuust in het oppikken van seisoenspatronen, maar had moeite met plotselinge beleidsveranderingen, tenzij explicet gemodelleerd via regressors.

6.3 ARIMA Iteraties

1. Modelconfiguratie

ARIMA werd getest met verschillende (p,d,q)-combinaties:

- Duidelijk trial-and-error aanpak zonder automatische tuning
- Geoptimaliseerd op basis van AIC en BIC
- Differencing werd toegepast om stationariteit af te dwingen

2. Voorspellingen zonder regressors

In eerste instantie werd ARIMA zonder exogene variabelen getest. Deze baseline had redelijke prestaties, maar kon geen externe invloeden (zoals subsidies) modelleren.

3. ARIMAX – Toevoeging van Regressoren

Daarna werden regressors toegevoegd (zoals subsidiebedragen en brandstofprijzen), wat leidde tot het ARIMAX-model:

- Hierdoor kon het model reageren op beleidsimpulsen
- Prestatie verbeterde in sommige perioden (zoals eind 2020 en 2022)

4. Beperkingen

- ARIMA was gevoelig voor niet-stationariteit van sommige regressors
- Het model was minder flexibel dan Prophet in het verwerken van seizoensinvloeden

Scenario 1: Voorspellen van EV's met Externe Factoren (Getunedede Modellen)

Model	Gemiddelde MAPE
Heuristic Baseline	0.2756
SARIMAX	0.1356
SVR	0.1456
LightGBM	0.16

Scenario 2: Voorspellen van EV's zonder Externe Factoren (Getunedede Modellen)

Model	Gemiddelde MAPE
SARIMAX	0.16910
SVR	0.17910

LightGBM	0.17910
----------	---------

Effect van tuning op Mape:

Model	MAPE zonder tuning	MAPE met tuning
SARIMAX	0.26	0.134
SVR	0.57	0.144
LightGBM	0.18	0.164



63.1.1.1 Structuur van de weekplanning

Weekindeling

We verdelen het werk in 4 blokken:

1. Analyse en modellering afronden (do 6 – za 8 juni)
2. Resultaten en discussie uitschrijven (zo 9 – ma 10 juni)
3. Inleiding, abstract, conclusie, toekomst & ethiek (di 11 juni)
4. Laatste check, APA, lay-out, inleverbaar maken (wo 12 – do 13 juni)

Dag-tot-dag planning tot vrijdag 13 juni

Vandaag (wo 4 juni, 15:00–19:00)

- Prioriteitenlijst opstellen (inhoudelijk + APA/language)
- Subvraag 2 herschrijven + doelen & doelgroep verduidelijken
- SHAP plots genereren + LOFO eruit
- Aanzet maken voor resultatenstructuur

Donderdag 5 juni – Analysemiddag (hele dag beschikbaar)

- Voorspellingen 1, 3, 6, 9 maanden genereren
- Evaluatie: MAPE, MAE, prediction intervals
- Baseline modellen: Ridge, Lasso toevoegen
- Beschrijving modelarchitecturen schrijven

Vrijdag 6 juni – Resultatendag

- Resultaten uitschrijven (vergelijkingen, tabellen, uitleg)
- SHAP-analyse uitschrijven
- Tussenversie sturen naar Daphne voor codecheck

Zaterdag 7 juni – Discussiedag

- Reflectie per model: prestaties + horizon
- Causaliteit vs correlatie expliciet maken
- Overfitting / featurekeuze bespreken (m.b.v. heatmap)
- Vergelijking met literatuur uitschrijven

Zondag 8 juni – Prototype & Toepassing

- Gebruikssituatie + doelgroepbeschrijving (car dealer, kvk)
- UI uitleg + Figma integreren
- Ethiek & maatschappelijke impact

Maandag 9 juni – Inleiding & Conclusie

- Abstract, inleiding en conclusie schrijven
- Future work + model monitoring toevoegen
- Nalezen op logische structuur

Dinsdag 10 juni – APA & Verwijzingen

- APA-stijl controleren
- Gebroken links oplossen
- Figuurtitels + Engelse taal in tabellen

Woensdag 11 juni – Scriptie afmaken

- Laatste inhoudelijke correcties
- Alles doornemen op volledigheid
- Klaarzetten voor PDF-conversie

Donderdag 12 juni – Review en feedbackronde

- Laatste review: spelling, titels, consistentie
- Laat iemand anders snel nalezen (vriend, klasgenoot, AI)
- Inleverbaar maken

Vrijdag 13 juni – Inleveren (voor 09:00)

- Scriptie uploaden en bevestiging checken

Op basis van de M&T-feedback en jouw thesis, zijn dit de belangrijkste aandachtspunten die je kunt toepassen op je resultatenhoofdstuk, met directe koppeling aan leeruitkomsten B2–B4 en C2:

B2 – Modelverantwoording en architectuurkeuze

64 Wat moet je verbeteren:

- Je geeft nu al redelijke modelverantwoording, maar:
 - De keuze van XGBoost als voorkeursmodel moet beter verdedigd worden t.o.v. alternatieven zoals ARIMA, Random Forest, of zelfs lineaire regressie.
 - Licht beter toe waarom LOFO boven SHAP gekozen is als interpretatiemethode.
 - Benoem explicet dat je eerst een *eenvoudige modellen test* zoals lineaire regressie of ARIMA (zie ook B2-opmerking in jouw feedback: *start simple*).

65 Hoe toepassen:

“To enhance model transparency and robustness, a baseline ARIMA model was implemented and compared to XGBoost. This comparison clarifies whether the added complexity of XGBoost yields meaningful performance gains.”

B3 – Datapreparatie en validatie

66 Wat moet je verbeteren:

- Splitsing en normalisatie zijn goed beschreven.
- Je validatiestrategie kan beter verantwoord worden:
 - Waarom een bepaald aantal maanden lag?
 - Waarom bepaalde Google Trends features wel/niet behouden?
- Geen uitleg waarom sommige variabelen zoals *thuis opladen ruis* geven volgens LOFO — leg dit uit m.b.v. datakarakteristieken (bijv. seizoen of correlatie).

67 Hoe toepassen:

“The keyword ‘thuis opladen’ was removed after LOFO indicated a negative contribution. This may be due to seasonal skew or overlapping meaning with other features like ‘elektrische auto’. These patterns were also visible in the correlation matrix (see Appendix X).”

B4 – Iteraties, hyperparameter tuning en leerproces

68 Wat moet je verbeteren:

- Tuningstappen zijn nu beperkt en niet systematisch.

- ✗ Je past hyperparameters aan (depth, estimators), maar motiveert dit niet met learning curves of zoekstrategie (zoals grid search, random search, of Bayesian tuning).
- ✗ Sommige keuzes lijken dubbelop (bijv. early stopping + meer epochs).

69 Hoe toepassen:

“Rather than manually adjusting parameters, a grid search was conducted to optimize max_depth, n_estimators, and learning_rate, with 5-fold cross-validation using walk-forward logic. This ensures tuning is reproducible and avoids arbitrary changes.”

C2 – Evaluatie en foutenbespreking

70 Wat moet je verbeteren:

- ✗ Prestatiematen zijn goed gekozen (MAPE, RMSE, MAE).
- ✗ Confusion matrix is niet van toepassing, maar:
 - ✗ Je bespreekt nog te weinig *kwalitatieve fouten* zoals spikes, seizoenspieken, of onverklaarde dalingen.
 - ✗ Geen bespreking van *model uncertainty* of voorspellingsintervallen.

71 Hoe toepassen:

“A qualitative analysis was conducted by inspecting months with the largest residuals. Notably, the peak in EV sales in December 2020 coincided with a sudden bijtelling policy change. This pattern was not captured by fuel or trend variables, indicating a need for sharper policy encoding.”

Samenvatting – Concreet toepassen in je resultatenhoofdstuk

Punt	Actie voor scriptie
Modelkeuze (B2)	Beschrijf explicet waarom XGBoost is gekozen. Vergelijk met ten minste één ander model zoals ARIMA of lineaire regressie.
Hyperparameter tuning (B4)	Voeg grid search of systematische tuning toe. Geef per iteratie aan wat je hebt veranderd, waarom, en welk resultaat dat gaf.
LOFO vs SHAP (B2/C2)	Licht toe waarom LOFO gekozen is, en wat SHAP zou kunnen toevoegen.
Foutbespreking (C2)	Analyseer grote afwijkingen en outliers in de voorspellingen, leg uit met context (zoals pandemie, bijtellingsveranderingen).
Feature-analyse (B3)	Leg uit waarom sommige features goed werken (zoals brandstofprijzen) en andere juist niet. Link dit aan data-eigenschappen.
Transparantie & eenvoud (B2)	Motiveer dat transparantie belangrijk is voor jouw stakeholders. Licht toe waarom eenvoud (zoals lineaire modellen) eerst overwogen is.

71.1.1 05/06

Nieuwe hoofdstuk indeling:

Appendix toevoegen na de discussie, daar alle plots inzetten

- Quantitative analysis
- [Quantitative analis hoofdstuk](#)
- Nieuwe indeleing results
- Extend exisiting work and specify why mape is good
- Add level of automation/benefit
- Results;
-
- Discussion
 - hypothesise how your model could be used and under which conditions, such as sketching a UI for your target audience.

71.1.1.1 Planning vandaag:

- Aanzet maken voor resultatenstructuur
- Voorspellingen 1, 3, 6, 9 maanden genereren
- Resultaten uitschrijven (vergelijkingen, tabellen, uitleg)
- [Appendix maken](#)
- Figuren in appendix zetten

Indeling lotte:

I. INTRODUCTION

- A. The Problem
- B. Existing Studies
- C. Gap in Existing Work
- D. Proposal
- E. Question(s)

II. BACKGROUND

- A. Basic Knowledge
- B. State of the Art

- C. Stakeholder analyze

III. REQUIREMENTS

- tabellen requirements

IV. PROTOTYPE

- A. Methodology: welke methode ik heb toegepast
- B. Idea Generation and Value Proposition: Value proposition
- C. Flow diagram and AI breakdown
- D. User Research: key findings interviews, empathy map en consequenties
- E. Prototype: design patterns, user testing resultaten, etische overwegingen

V. MODEL

- A. Methodology: dataset en evaluatie metrics die ik kies
- B. Results: al mijn model resultaten en uitkomsten (dus iteraties enzo)

VI. CONCLUSION

- antwoorden op elke deelvraag en hoofdvraag

VII. DISCUSSION

- A. Implications
- B. Future Research
- C. Recommendations

71.1.2 04/06:

71.1.3 Meeting Log – Discussion with Supervisor (Marcio)

Date: 4 June 2025

Topic: Feedback implementation and planning for 100% thesis version

To-Do List (from feedback and discussion)

- Add evaluation for 3-, 6-, and 9-month forecast horizons.
- Do not invest further in complex models (e.g., LSTM) due to limited data.

- Focus more on data analysis and identifying situations where the model performs well or poorly.

Discussion Points and Responses

Should I include Ridge or Lasso regression in addition to Linear Regression?

- Yes. Give baseline models a real chance to outperform more complex ones.
- Add Lasso and Ridge to assess whether the problem is linear.
- XGBoost may still be valuable but should be compared thoughtfully.

How should I present the results—side-by-side comparison or focus on one best model?

- A side-by-side comparison is useful, but the discussion section should reflect on which model is most appropriate and under which conditions.

How can I better reflect the difference between correlation and causation in my results or conclusion?

- This was initially overlooked but should be more explicitly addressed in the discussion section.

Can you clarify your earlier feedback about “adding noise” and the role of highly correlated features?

- The dataset appears to be highly correlated (e.g., Fig. 4 heatmap shows many red blocks), which may lead to overfitting.
- Further analysis is needed on feature independence. Explore consequences of using many inter-correlated features together.
- Consider techniques like normalization and feature selection (e.g., ANOVA).

How detailed should my prototype explanation be to match the target audience?

- The dashboard should reflect the needs of a clearly defined user (consider adjusting the audience to car dealers or salespeople).
- The target audience would be most interested in short-term predictions.

Does adding baseline models like ARIMA and Linear Regression shift the research scope too much?

- It stretches the original scope slightly, but in a good way. Comparing with classical models offers a new perspective on the problem.

How should I compare my results with existing studies if their metrics and methods differ?

- No major change is needed. Acknowledge the differences and compare at a high level where possible.

Request for Extension

I requested whether the submission deadline could be moved from Friday the 13th to Sunday, due to my ADHD and dyslexia. Marcio indicated this will be discussed during the teacher meeting on Thursday. If needed, it can also be discussed further with Sonja.

71.1.3.1 Feedback/ plan

Feedback	Plan (green is done)	Sec
2. Sub research question 2 is not clear to the reader. This sub research question has to be reformulated.	Rewrite sub-question: Which digital search trends related to EVs are relevant for forecasting EV sales?	Problem
3. I do not understand Figure 2.	Delete Figure 2, its explained in the text	
4. Page 10, top of right column: You have to report the importance of the separate features. What is the difference between your approach, i.e. LOFO feature performance approach, and the SHAP-based feature attribution approach?	Was debating between LOFO/SHAP, wasn't sure which one I was going to add, so I added both. Going forward with SHAP	(H)Dataset Assembly and Feature Framing
5. What is the number of records in the Google Trend dataset?	Clarify: all records cover 219 months.	IV. IMPLEMENTATION DETAILS
6. What is the best model? What features are included in this model? What are the time lags? What is the performance of this model? And what is the seasonal pattern in this model?	Add model summary: features, lags, metrics, seasonal pattern.	V. MODEL DEVELOPMENT AND EVALUATION
7. Page 11 "D. LOFO Results with Tuned Model": To which model do Figure 8 and Figure 9 correspond? What features are included in this model? What are the time lags? What is the performance of this model? And what is the seasonal pattern in this model?	Results were added as last, so it was a little messy. In 100% report will do quantitative analysis and be clearer in corresponding, also will add more details on dataset explanation	VI. QUANTITATIVE RESULTS AND ANALYSIS
8. You should use other approach rather than only XGBoost.	As your feedback suggested, now added LR, ARIMA & ARIMAX	
9. You should use metrics such as MAPE rather than other metrics you have chosen.	Switch to MAPE as main metric and add explanation	
10. Discussion: Compare your results/findings with those from the literature.	Will rewrite discussion, however might be challenging due to different approach/ metrics/ after call with Marcio wasn't needed anymore, could do	
11. The text in the Tables and	Translate all tables/figures	

Figures have to be in the English language.	to English.	
12. Future research: Which alternative model/approach might give better results than the model/approach you chose?	Suggest LSTM, GRU, Decision Trees or Linear Models.	
13. The reference list has to be checked whether all references are described according to the APA guidelines. /Check broken links in the reference	I marked all broken lines with red, will add working link/DOI Only books, and research from kaya did not have doi	
Feedback Rubric		
14. A: Strong statements that are not backed by proper sources.	Will go through paper to make sure all (strong) statements are backed by proper sources	Full page
15. The exploration of alternatives, however, was very limited and too solution oriented. More specifically, it directed the reader to the envisioned solution space using tabular data and predefined models such as XGBoost.	Broaden discussion: added Linear, ARIMA, ARIMAX. Will briefly mention other options (e.g. SARIMA, LSTM) and justify selection.	
16. There are no discussions on the level of automation and the benefit to society if good predictions are delivered or the negative impact in case of poor predictions. It also does not describe that although some correlations can be detected, this is not an indication of causality, that the scenario is dynamically changing, and that the model should be constantly monitored.	Will add level of automation. Briefly reflect on long-term societal impacts like infrastructure costs, EV charging access, Ev parking places and subsidy allocation. Add note on lack of causality, dynamic context, and need for continuous model monitoring.	
17. you should at least hypothesise how your model could be used and under which conditions, such as sketching a UI for your target audience. I recommend having a dedicated section in the discussion chapter.	I made a Prototype, see Figma will also add ai design principles	

18. You should justify adding a more complex model for a slightly better improvement in the prediction,	Clarify why XGBoost was used—refer to predictive improvement, seasonal fit, and ability to capture complex patterns.	
19. Address mismatch between the problem (forecasting time series) and the solution (tabular model).	Acknowledge mismatch. Justify tabular approach by explaining use of lags/time features. Note limitations vs. true time series models.	
20. Conduct a qualitative analysis—it might be the case that there is no causality but only co-occurrence, especially with variables like fuel prices.	Will add quantitative analysis as extra section in results	
21. B: Req: However, using too many metrics as a basis for comparison leads to problems: What if they do not agree? Be very clear on what you will do in such situations.	Will only use MAPE as compass metric	
22. Don't over-focus on one model. With small datasets, simpler models often perform just as well or better	Added baseline models (LR, ARIMA, ARIMAX). Justify XGBoost based on improved accuracy and ability to capture nonlinear patterns.	
23. "XGBoost is not ideal without strong feature engineering. Start with simpler time series models like Linear Regression or ARIMA before moving to complex models."	Confirmed feature engineering for XGBoost (lags, time vars). Added simple models first. Justify switch to XGBoost with performance and interpretability.	
24. There is a contradiction: the report emphasizes transparency, but XGBoost is not inherently transparent. LOFO helps, but you should justify its use over more interpretable methods like SHAP.	Replace or complement LOFO with SHAP plots to better explain model behavior. Briefly justify choice by comparing interpretability and relevance to stakeholders.	
25. This is a minimal dataset (219 observations). Complex models need more data—start with simpler models to establish a solid baseline.	Addressed by adding LR, ARIMA, and ARIMAX as baselines. Use results to show whether complex models like XGBoost offer meaningful improvement. (but afraid this will	

26. Dataset preparation is solid, but the Google Trends keyword selection lacks transparency. Clarify how translations were validated, why some keywords were discarded, and include the final list.	prompt and keyword list was added to logbook however for 100% will add appendix. Will briefly explain translation was validated via DeepL and Google Trends relevance checks.	
27. Basic tuning was mentioned, but no systematic optimisation strategy was applied. Use an approach like grid search or Bayesian optimisation to explore hyperparameters more thoroughly.	Added Bayesian optimisation, will explain this in my rapport	
28. C: Good that you address bias and transparency, but you should better explain model behaviour using SHAP or similar: LOFO was introduced too quickly. Also, discuss how to detect and handle performance decay over time with non-stationary data.	Will add SHAP to improve model transparency and explain feature impact more clearly. Briefly discuss how to monitor model drift and detect performance decay over time.	
29. Too much focus on R^2 —real-world forecasting relies more on absolute error metrics like MAE or MAPE. Also include prediction intervals to assess model uncertainty.	Will focus on MAPE as main metric. Add prediction intervals using ARIMA's confidence bounds and explore residual-based or bootstrapped intervals for XGBoost and Linear Regression.	
30. Adding more features doesn't seem to help much—consider testing with noise instead. Performance is acceptable for 1-month forecasts but breaks down at 6 months.	Will reflect on diminishing returns from added features and explore basic feature selection to reduce overfitting at longer horizons.	
31. A qualitative analysis is needed, check specific data points and outliers. Some patterns, like with fuel prices, may reflect coincidence rather than causality.	Will include brief qualitative analysis in discussion, highlighting key data points and potential non-causal patterns (e.g., fuel prices).	
32. The document is clear and well structured, but several points need improvement:	Plan on how to fix mentioned above planm	

review strong claims, clarify processes like translation and keyword selection, and strengthen the focus on time series evaluation.		
33. D2: be careful when distinguishing causation from correlation. This distinction must be made more explicit in your discussion.	Will add explicit statements in discussion to distinguish correlation from causation, with examples (e.g., fuel prices). <i>"Although variable X correlates with EV sales, this does not imply a causal relationship."</i> <i>"The model detects associations, not underlying causes. For example, fuel prices may coincide with sales shifts, but other unmeasured factors could drive both."</i> <i>"Caution is needed in interpreting these results as causal, especially in the absence of controlled experiments."</i>	
Points made earlier		
34. splitting the literature review into: a. Similar <i>problem domains</i> (e.g., forecasting in finance or energy). b. Similar <i>technical approaches</i> (e.g., time series regression, model evaluation techniques).	35.	
36. Emphasis on explaining model behavior clearly, as if advising a public organization.	38.	
37. Consider how a stakeholder would interpret the results and which decisions could follow.		
39. The "Klimaatbeleid Dashboard" is not considered state-of-the-art but may be included as an example of a <i>potential application</i> .	41.	
40. State-of-the-art should focus on recent forecasting models applied to time series regression problems.		
42. Requirements should be made SMART (Specific, Measurable, Achievable, Relevant, Time-bound).	43.	

71.1.3.2 Results:

Xgboost with bayans optimaisation works best

1 month:

Model Type	Feature Set	MAPE (%)
Linear Regression	Lag Only	12.65
	All Features	15.48
ARIMA	Lag Only	17.58
ARIMAX	All Features	42.99
XGBoost (Default)	Lag Only	11.72
	All Features	11.27
XGBoost (Bayesian)	Lag Only	12.67
	All Features	10.73

6 months:

Model Type	Feature Set	MAPE (%)
Linear Regression	Lag Only	18.35
	All Features	30.11
ARIMA	Lag Only	20.20
ARIMAX	All Features	1,106,771,604.66 X
XGBoost (Default)	Lag Only	32.03
	All Features	26.38
XGBoost (Bayesian)	All Features	16.17%
		0.16

Scenario 1: Voorspellen van EV's met Externe Factoren (Getunedede Modellen)

Model	Gemiddelde MAPE
Heuristic Baseline	0.2756
SARIMAX	0.1356
SVR	0.1456
LightGBM	0.16

Scenario 2: Voorspellen van EV's zonder Externe Factoren (Getunedede Modellen)

Model	Gemiddelde MAPE
SARIMAX	0.16910
SVR	0.17910
LightGBM	0.17910

Effect van tuning op Mape:

Model	MAPE zonder tuning	MAPE met tuning
SARIMAX	0.26	0.134
SVR	0.57	0.144
LightGBM	0.18	0.164

Explanation

The shift from a 1-month to a 6-month forecast horizon had a significant impact on model performance across all methods. Below is a discussion of the observed behavior and potential causes:

Linear Regression

- Lag Only: MAPE rose from 12.65% to 18.35%
- All Features: MAPE spiked to 30.11%

Explanation:

Linear regression assumes a fixed, linear relationship between predictors and the target. This assumption becomes increasingly problematic as the forecast horizon grows. Over longer periods, the relationship between past values and future outcomes tends to change due to:

- external shocks (e.g., policy changes),
- delayed effects of economic indicators,
- and compounding nonlinear dynamics.

Using all features amplifies this problem: irrelevant or weakly predictive variables introduce noise, and multicollinearity may destabilize the regression coefficients. This explains the sharp performance drop with all features.

ARIMA / ARIMAX

- ARIMA (Lag Only): MAPE increased to 20.20%
- ARIMAX (All Features): Catastrophically failed — MAPE ~1.1 billion%, convergence errors.

Explanation:

ARIMA models rely on autoregressive patterns and assume stationarity. As the prediction window extends, these assumptions break down:

- Recent lags (e.g., 1–6 months back) become less predictive of future outcomes.
- The model is unable to account for unobserved shocks or long-term dependencies.

For ARIMAX, the model failed to converge, likely due to:

- too many exogenous features,
- multicollinearity,
- and optimization instability over a long horizon.

This suggests ARIMAX is not robust in your setting without strong regularization or careful variable selection.

XGBoost (Default Parameters)

- Lag Only: MAPE = 32.03%
- All Features: MAPE improved to 26.38%

Explanation:

Default XGBoost starts to overfit on recent short-term patterns when applied to long-term forecasting. However, adding more features helps it recover some predictive power. This is because XGBoost:

- Can model complex, nonlinear interactions,
- Uses boosting to handle weak signals.

But without tuning, it still lacks robustness for 6-month prediction, especially under concept drift.

XGBoost (Bayesian Optimization)

- All Features: MAPE = 16.17%

Explanation:

This was the best-performing model. Bayesian optimization found a parameter configuration that balances model complexity and generalization — avoiding both underfitting and overfitting. It demonstrates that:

- With proper tuning,
- Rich feature sets (macro, sentiment, fuel, lag) can be leveraged effectively,
- Especially when paired with walk-forward validation to simulate real-world conditions.

Summary

Longer forecast horizons introduce more uncertainty and degrade predictive accuracy for most models. Simpler models like linear regression and ARIMA struggle due to their assumptions and limited flexibility. XGBoost, especially when optimized, adapts better to this complexity and remains the most promising approach for your use case.

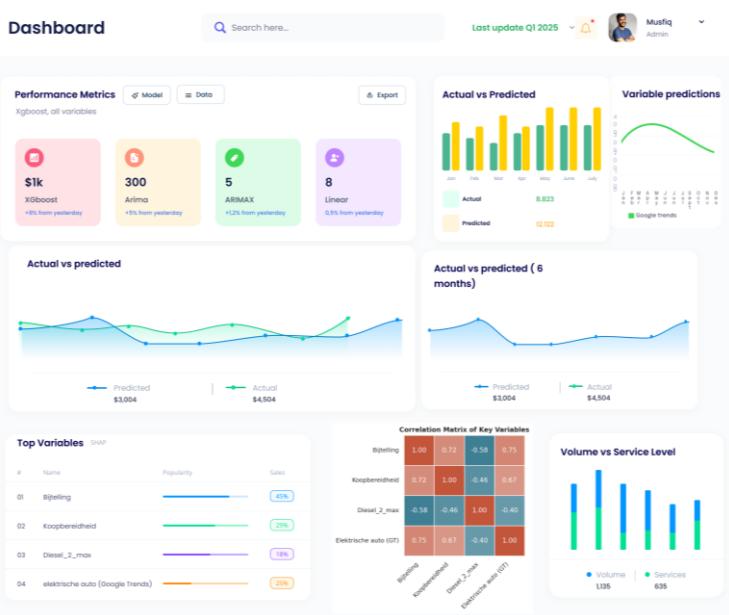
71.1.4 03/06

Vraag aan Daphne:

benefit to society if good predictions are delivered or the negative impact in case of poor predictions. It also does not describe that although some correlations can be detected, this is not an indication of causality, that the scenario is dynamically changing, and that the model should be constantly monitored. → op langer termijn overschatten, zou er veel veranderingen plaatsvinden zoals extra laatpalen etc wat meer geld zou kosten, infrastructuur zou worden aanpassen. (percentage openbare parkeerplekken voor electrische auto is te groot), wordt de juiste hoeveelheid subsidie juist berekend te grote investering zou negatieve gevolgen hebben voor andere maatschappelijke effecten.

Op lange termijn onderschatte Zou kunnen dat er te veel mensen een subsidie aanvragen dan dat er voor is berekend, dit zou gevolgen hebben op.... (zoeken naar gelijkbare cases).

Gedaan:



Gebruikgemaakt van

Schets prototype en toevoegen van ai design principles adhv:

<https://pair.withgoogle.com/guidebook/>

En Prototype design by:

<https://www.figma.com/community/file/1153320445661469840>

De 3 hoofdzaken die je scriptie naar een voldoende kunnen tillen

Motiveer je keuzes duidelijk

- Waarom gebruik je SHAP i.p.v. LOFO?
- Hoe definieer je precies je target (EV sales)?
- Hoe heb je keywords vertaald en geselecteerd?
- Benoem expliciet de variabele die je voorstelt én waarom.

Voeg een alternatief model toe (naast XGBoost)

- Implementeer bij voorkeur een *fundamenteel ander model* (zoals ARIMA of ARIMAX).
- Licht toe waarom dit een relevante vergelijking is (pre-AI vs AI).
- Als het niet werkt: leg dit ook uit en toon dat je het geprobeerd hebt.

Gebruik MAPE en evalueer je resultaten kwalitatief

- Laat zien wat je modellen doen in begrijpelijke termen.

- Welke features zijn belangrijk? In welke richting werken ze?
- Is er een seizoenseffect? Komt dit overeen met wat je uit de literatuur verwacht?

To-do na meeting met Kees – richting 100% inlevering

Onderzoeks vragen en scope

- Herschrijf sub research question 2
 - Maak duidelijk *welke* digitale trends je bedoelt
 - Specificeer *waarom en waarop* ze betrekking hebben (bv. op zoekgedrag, consumentenvertrouwen)
-

Visualisaties en figuren

- Verwijder figuur 2, aangezien deze niets toevoegt aan de tekst.
 - Controleer alle andere figuren/tabeltitels op Engelse taal.
-

Doelevariabele en featurekeuze

- Beschrijf duidelijk welke variabele je voorspelt (bv. maandelijkse EV-verkopen).
 - Motiveer waarom je juist deze variabelen kiest en niet andere
 - Voeg onderbouwing toe met verwijzing naar bestaande literatuur.
-

Modelselectie & interpretatie

- Vervang LOFO door SHAP, of:
 - Gebruik SHAP *näast* LOFO en leg uit waarom SHAP gekozen is.
 - Geef duidelijk aan wat het verschil is in interpretatie tussen beide.
 - Rapporteer het beste model (features, performance, time lags, seizoenspatroon).
 - Voeg dit ook toe bij de uitleg van figuren 8 en 9.
 - Implementeer een klassiek tijdreeksmodel zoals ARIMA of ARIMAX
 - Vergelijk resultaten met XGBoost om de kracht en zwakte van beide aan te tonen.
-

Evaluatie en metrics

- Voeg MAPE of SMAPE toe als hoofdmetric voor evaluatie.
- Voeg onzekerheidsintervallen of spreidingsgrafieken toe als illustratie van modelonzekerheid (indien mogelijk).

Literatuur en vergelijkingen

- **Vergelijk je resultaten in de discussie met die uit de literatuur.**
 - Formuleer een reflectie in 'Future Work':
 - Welke modellen had je achteraf nog kunnen testen?
 - Wat zou je anders doen in een vervolgonderzoek?
-

Documentkwaliteit & referenties

- Controleer alle tabellen/figuren op Engelse tekst.
 - Controleer de APA-stijl van alle referenties.
-

Dataset-specificaties

- Noteer explicet het aantal records in de Google Trends-dataset.
- Geef ook het aantal records voor andere datasets (EV sales, pompprijzen, etc.) indien nog niet gedaan.

To-dolijst voor 100%-versie scriptie

Onderzoeksopzet & Context (A1, A2, A3)

- Herformuleer te stellige uitspraken en voorzie deze van betrouwbare bronnen.
- **Controleer en herstel gebroken links in de referentielijst (zoals referentie 1).**
- Breid de verkenning van alternatieve modellen uit (zoals ARIMA, lineaire regressie).
- Voeg discussie toe over:
 - Gebrek aan causaliteit bij geconstateerde correlaties.
 - Dynamiek in de context en de noodzaak van voortdurende monitoring.

Ontwerp, prototype en toepassing (A4, A5)

- Hypothesiseer hoe het model gebruikt zou kunnen worden in de praktijk.
- Schets (optioneel) een eenvoudige gebruikersinterface (UI) voor beleidsmakers.
- Voeg dit toe aan het discussiedeelte.
- Voer een kwalitatieve analyse uit van mogelijke co-incidentie (bijv. brandstofprijzen).
- Onderbouw waarom je voor een complex model kiest.

Technische requirements & architectuur (B1, B2)

- Bespreek wat je doet als je evaluatiemetrics (MAE, RMSE, R^2) tegenstrijdige resultaten geven.
- Beschrijf duidelijk je modelarchitectuur. Start met eenvoud (lineaire regressie, ARIMA).
- Motiveer keuze voor LOFO en leg uit waarom je SHAP niet gebruikt.
- Voeg vergelijking toe tussen LOFO en SHAP.
- Benoem welke features, lags en performance bij welk model horen (bijv. bij figuren 8 en 9).

Data & Feature Engineering (B3, B4)

- Beschrijf hoe je Google Trends-zoektermen hebt vertaald en geselecteerd.
- Voeg een lijst toe met alle gebruikte keywords.
- Vermeld het aantal observaties in je Google Trends-dataset.
- Implementeer een optimalisatietechniek voor hyperparameters (grid search of Bayesian).
- Beschrijf welke features zijn opgenomen in het beste model, inclusief tijdslags en seizoenspatronen.

Evaluatie & Verantwoording (C1, C2, C3)

- Voeg SHAP-plots of andere interpretatievisualisaties toe (naast LOFO).
- Bespreek modelveroudering en hoe je dit kunt herkennen (non-stationaire data).
- Gebruik SMAPE of MAPE als hoofdevaluatiemetric i.p.v. R^2 .
- Voeg onzekerheidsintervallen toe aan je voorspellingen.
- Evaluateer waarom sommige extra features weinig toevoegen.
- Voer een kwalitatieve analyse uit van outliers en specifieke datapunten.
- Besteed meer aandacht aan tijdreeksmodellen en -evaluaties.
- Vergelijk je resultaten met die uit de literatuur in je discussiehoofdstuk.
- Formuleer een onderbouwde suggestie voor toekomstig onderzoek met een alternatief model.

Rapportage & Communicatie (D1, D2)

- Herformuleer sub-vraag 2 zodat deze duidelijker wordt.
- Voeg uitleg toe bij figuur 2 — maak duidelijk wat deze laat zien.
- Geef helder aan welke variabele je precies voorspelt en waarom.
- Zorg dat alle tabellen en figuren Engelstalig zijn.
- Controleer de volledige referentielijst op APA-stijl.

Moraal van het verhaal/feedback:

Probleem is goed, ook goed dat ik focus op transparancy. Zorg er ook voor dat statements goed worden onderbouwt met bronnen. De datasets die ik gebruik (pompprijzen, google trends) etc is goed/logisch. De literatuur/existing word is ook voldoende, maar de exploratie voor alternatieven was mager, en te uitkomst gericht, ik heb de lezer te veel gericht op uitkomsten met tubular data en predefined models such as xgboost.

Ik heb geen discussie punt over de level of automation, en de positieve/negatieve uitkomst op de maatschappij, bij goede/ slechte voorspeling. Ook is het onvoldoende belicht dat er hoewle wel correlatie is, geen causatie is. Ook vertel ik niet dat de omgeving steeds aan het veranderen is en dat er steeds moet worden geavalueerd. Voor mn 100% versie moet ik op zn minst een hypothetische voorbeeld geven over hoe mijn model gebruikt kan worden en onder welke voorwaarden, zoals een schetsmaken voor UI. Het wordt aangeraden om hiervoor een aparte sectie in mn discussie te hebben hiervoor.

TODO:

Check strong statements without sources

check links die niet werken

02/06

A1: Je doet onderzoek om het probleem in context in kaart te brengen, herformuleert het vraagstuk, weegt en toetst de belangen van de opdrachtgever, eindgebruiker en andere belanghebbenden.

It is a relevant problem, and you highlighted the need to build forecasting tools for EV sales. You also referred to transparency as one important tool.

However, there are still strong statements that are not backed by proper sources. Be cautious with bold assertions; always back them with reliable sources.

Minor: Check broken links in the reference section (e.g., reference 1).

A2 Je verkent en analyseert de oplossingsrichtingen voor een vraagstuk en ontwerpt een AI-oplossing volgens een gangbare ontwerpmethodiek

You explored different families of information that are useful for predicting sales (leasing, fuel prices, google trend topics)

You have identified public tools that help analyse statistics concerning EV sales.

You have explored existing research concerning EV sales predictions.

The exploration of alternatives, however, was very limited and too solution-oriented. More specifically, it directed the reader to the envisioned solution space using tabular data and predefined models such as XGBoost.

A3 Je beschrijft de kwaliteit van het ontwerp door te verwijzen naar design patterns en te kijken naar aspecten als haalbaarheid, gebruiksvriendelijkheid en de toegevoegde waarde die de oplossing kan bieden aan stakeholders.

There are no discussions on the level of automation and the benefit to society if good predictions are delivered or the negative impact in case of poor predictions. It also does not describe that although some correlations can be detected, this is not an indication of causality, that the scenario is dynamically changing, and that the model should be constantly monitored.

A4 Je ontwikkelt een prototype en test deze in de context waarin de oplossing gebruikt gaat worden.

You should improve this aspect for 100% delivery. Although creating a user interface is not the primary focus of your experiment, you should at least hypothesise how your model could be used and under which conditions, such as sketching a UI for your target audience. I recommend having a dedicated section in the discussion chapter.

A5 Je onderbouwt voor de opdrachtgever of AI wel of niet geschikt is voor een gegeven vraagstuk, gelet op technische, maatschappelijk en ethische aspecten.

Criterion Feedback

Here, you should justify adding a more complex model for a slightly better improvement in the prediction.

Very related to comments added to A3.

Also, there is a slight mismatch between the problem (forecasting time series) and the solution (tabular model).

Additionally, please conduct a qualitative analysis—it might be the case that there is no causality but only co-occurrence, especially with variables like fuel prices.

B1 Je stelt voor een AI-oplossing juridische, ethische, organisatorische, functionele en technische requirements op

Criterion Feedback

The requirements are well described and helpful for steering your experiment. However, using too many metrics as a basis for comparison leads to problems: What if they do not agree? Be very clear on what you will do in such situations.

B2 Je stelt op basis van requirements een geschikte architectuur voor een AI-oplossing op en selecteert daarvoor passende AI-technieken gebruik makend van machine learning, deep learning, kennisrepresentatie, computer vision en natural language processing.

Criterion Feedback

You gave too much emphasis to a single model. In real-world settings, especially with small datasets like yours, simpler models often perform as well as or better.

XGBoost is not ideal for time series forecasting without much feature engineering. You are trying to fit a time-series problem into a tabular dataset model. Please reconsider testing classical time series models like ARIMA or ARIMAX, which you mentioned but didn't implement (it is simple to implement since you already have all data). Always start simple: Linear Regression should be your first baseline, then ARIMA, and only escalate to complex models if simpler ones fail.

There is also a contradiction here: Initially, the document mentions transparency, but XGBoost is not necessarily transparent (more specifically, it can be composed of hundreds or thousands of decision trees). LOFO can be helpful here, but you should explain why LOFO and not other more transparent techniques such as SHAP.

B3 Je prepareert een dataset voor het trainen en testen van een AI-model en kan de voor- en nadelen van het gebruik van een bestaande dataset onderbouwen, rekening houdend met technische en ethische randvoorwaarden

Criterion Feedback

We have already alerted you that this is a minimal dataset, with only 219 observations. Complex models are hungry for more data, so I recommend you take a step back with simpler models to form a baseline.

Dataset preparation is done carefully, but translating and selecting keywords from Google Trends was not fully transparent. Please specify how you validated the translations and why specific keywords were discarded. Where is the final list?

B4 Je ontwikkelt een nieuw of voorgetraind AI-model volgens een iteratief en systematisch proces.

Criterion Feedback

Although you mentioned some basic tuning (adjusting estimators and depth), there was no systematic optimisation strategy.

Please use an optimisation technique (at least grid search or Bayesian optimisation) to explore hyperparameters.

C1 Je evalueert hoe de ontwikkelde oplossing werkt en welke (ongewenste) consequenties de AI-oplossing kan hebben voor individu en maatschappij en treedt hierover in gesprek met vakgenoten

Criterion Feedback

You discuss societal implications like bias and the importance of transparency; good. However, more effort could be put into explaining model behaviour using SHAP plots or similar, making the solution even more understandable to non-technical stakeholders. You went too quickly to LOFO.

Also, real-world data is non-stationary — you should discuss ways to check if your model ages badly.

C2 Je evalueert en beoordeelt de kwaliteit van een AI-model aan de hand van kwaliteitscriteria die in het vakgebied erkend worden zoals robustness, performance, scalability, explainability, model complexity en resource demand.

Criterion Feedback

Metrics such as MAE, RMSE, and MAPE are used. However, there is still too much emphasis on R2. Remember that real-world forecasting success concerns absolute error measures like MAE or SMAPE. Also, consider including prediction intervals to check model uncertainty.

It looks like adding more features is not helping too much; perhaps adding noise instead would be better. It works decently for 1M, but it is not okay for 6 M.

A qualitative analysis is really necessary here, for example, checking specific data points, outliers, and so on. It might be the case that there is no causality but just occurrence, especially fuel prices.

C3 Je kiest robuuste en valide onderzoeksmethoden en draagt zo bij aan de ontwikkeling van nieuwe kennis in het AI-vakgebied

Criterion Feedback

The document is well organised and easy to read. However, many points should be improved, such as reviewing strong statements, clarifying how some processes were designed (e.g., translation and word checking), and fixing the minimal emphasis on time series evaluations.

D1 Je verzamelt en verwerkt actief feedback over de eigen kennis, vaardigheden en ontwikkeling en laat zien deze feedback te kunnen gebruiken om je werk te verbeteren

Criterion Feedback

You frequently discuss with the supervisors and report on how you applied the feedback. Very good.

D2 Je kan zelfstandig een onderzoek opzetten, uitvoeren en schriftelijk en mondeling rapporteren volgens de geldende standaarden uit het vakgebied.

Criterion Feedback

You made good use of sections, with good examples and explanations. At the same time, be careful when distinguishing causation from correlation. This distinction must be made more explicit in your discussion.

Thema: Bias en ethische aspecten

- Vermeld explicet dat bias *niet* van toepassing is binnen deze specifieke casus.
- Benoem dat het model binnen de categorie “*low-risk AI*” valt volgens de EU AI Act, en licht dit kort toe in je scriptie.
- Aandacht voor bias en privacy is beperkt relevant; focus ligt meer op prestaties en transparantie.
- Op basis hiervan kunnen enkele requirements worden aangepast om sterker te leunen op performancegerichte criteria.

Resultaten & Modelleringskeuzes

- Piekmomenten in de verkoopdata worden goed voorspeld – dit onderbouwt het gebruik van *lead features*.
- Overweeg om de grafieken met voorspellingen in meer detail te visualiseren.
- Het blijft waardevol om ook een model te bouwen voor voorspellingen op een horizon van 6 maanden.
- Let op: de training data verschilt van de voorspellingsdata. Visualiseer eventueel ook deze verdeling in een aparte grafiek (train/test split).
- Vermijd het tonen van volledige resultatenreeksen *met één zonder* hyperparameter tuning; focus op het eindresultaat.
- Geef in je rapport duidelijk aan dat je op de validatieset hebt getuned en dat de uiteindelijke evaluatie op de testset is uitgevoerd.

- Je kunt eventueel al kort iets zeggen over de prestaties op de testset als indicatie.

State of the Art (SOTA)

- SOTA wordt summier benoemd in de bestaande sectie; bespreek tijdens het feedbackmoment van uitbreidung wenselijk is.

Doelgerichtheid

- Herinner jezelf eraan dat het doel van het onderzoek is om een *bruikbaar* voorspellingsmodel te ontwikkelen. Overweeg daarom ook de toepasbaarheid van het model in de praktijk te benadrukken, eventueel door modellen te testen op verschillende datasenario's.

71.2 29/05

Gesprek met daphne:

Bias eruit halen, specifiek zeggen dat Biasniet van toepassing is

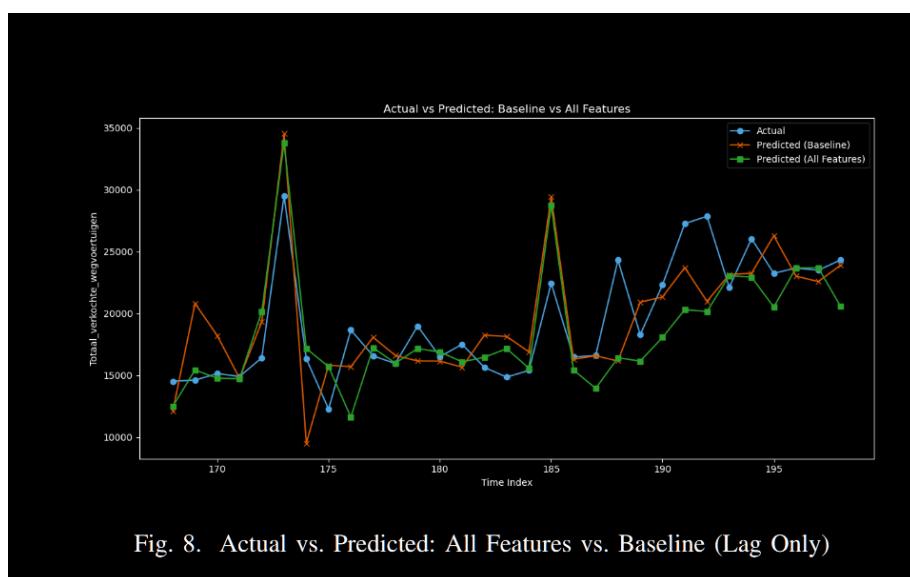
Eu ai act benoemen, dat mn Scriptie om een low risk model gaat adhv eu ai act.

Bias and privacy niet van grote toepassing

<https://statics.teams.cdn.office.net/evergreen-assets/safelinks/1/atp-safelinks.html>

basis hiervan ook requirements aanpassen zodat die dan meer performante gericht zijn.

Resultaten:



De pieken worden goed voorspeld, dus er is eigenlijk een reden om lead voor

Evt de grafiek in meer detail te voorspellen

Nog steeds goed idee om 6 maanden voorspellingsmodel te maken

Training data veranderd. Voorspellingsdat is anders, dus ik zou eigenlijk een grafiek daarvan te zien.

Trainings/test data

Je wilt niet een heel stukje met/zonder hyperparameter tuning

Nu op hyperparameter tunen op validatie set, en daarvoor finale versie welop test data laten zien, dit duidelijk maken aan docent. Evt beetje kijken op testset

Over SOTA, wordt op zich wel beetje in existing wordt vermeld vraag anders tijdens feedback moment.

Evt modellen ook testen met data, doel van onderzoek was om een bruikbaar model te maken.

Reviewer Feedback Summary (Daphne)

1. Clarity and Flow

- Several sentences lack smooth transitions or clear structure. Example: a sentence beginning with "According to Haque..." needs a connector like "but" to improve flow.
- Some sentences end abruptly or feel incomplete.
- Specific phrasing such as "the forecasting model" should be made more general to avoid implying the existence of a standard model.

2. Referencing and Support

- Multiple statements (e.g., about Dutch forecasting models or stakeholder needs) are not supported by references.
- Claims about model fairness, stakeholder expectations, and societal impact lack grounded scientific or stakeholder-based sources.

3. Methodology

- The use of LOFO is not fully justified—other methods (e.g., SHAP, LIME) are not discussed or compared.
- Subgroup analysis and bias detection are mentioned, but it's unclear how subgroups are defined or how fairness is assessed in practice.
- Comparing tuned and untuned models without a separate validation/test set risks overfitting; this must be addressed with proper split logic.

4. Requirements and Motivation

- The requirements table feels arbitrary requirements like RQ06 (explainability using SHAP/LOFO) are not fully justified with literature.
- It's unclear how requirements were derived—stakeholder input is mentioned but not documented.
- The line between a “fully met” and “partially met” requirement is not clearly explained.

5. Visuals and Figures

- Some figures are too small or have unreadable text (e.g., correlation matrix, consumer confidence trends).
- The walk-forward validation figure might be unnecessary; a flowchart or more intuitive illustration may be more effective.

6. Results Interpretation

- Discrepancy noted between abstract and results section regarding model evaluation (e.g., six-month vs. 44-month evaluation).
- It's unclear how LOFO analysis was performed and interpreted.
- The phrase “models were not trained here” conflicts with your description of walk-forward validation, which does involve retraining per iteration.
- Feature contributions (e.g., thuis opladen) are mentioned but not well explained.

7. Modeling and Tuning

- Hyperparameter tuning steps are under-explained. You should describe how values were selected and on what basis.
- Without separating tuning from test evaluation, performance results may be inflated due to overfitting.

8. Ethical and Practical Considerations

- Concerns like fairness and data access are discussed, but not always relevant or substantiated (e.g., data access is open, so the “access gap” is less convincing).
- Statements about user interfaces or dashboards are too vague—RQ07 is only partially met and needs stronger motivation or prototypes.

71.3 28/05

Thesis Supervision Meeting Notes

Date: 28 May 2025

Time: 10:30–11:00

Participants: Dylan Haazelager, Marcio Fuckner

1. Research Objective and Scope

- The main objective is to forecast monthly electric vehicle (EV) sales.

- Supervisor advised splitting the literature review into:
 1. Similar *problem domains* (e.g., forecasting in finance or energy).
 2. Similar *technical approaches* (e.g., time series regression, model evaluation techniques).
- 2. Model Evaluation and Metrics
 - Currently used metrics: MAE, RMSE, R².
 - If metrics diverge, clearly justify which metric will serve as the primary evaluation criterion.
 - Example: MAE as the lead metric; RMSE and R² as supportive.
 - Recommendation to relate the error (e.g., 3,000 units) to the average monthly sales to give readers a sense of scale.
 - Graphs should be made more interpretable by including actual months/years instead of only index values.
- 3. Qualitative Analysis of Outliers
 - A section should be added discussing anomalies (e.g., spikes in fuel prices or EV sales).
 - Suggested approach: correlate spikes with known events (e.g., COVID-19, energy crisis), referencing news articles or policies.
 - This analysis can clarify why the model underperforms during extreme market conditions.
- 4. Results Interpretation
 - The full-feature model performs slightly better than the baseline (historical sales only).
 - Some individual variables (e.g., policy incentives like “bijtelling”) may introduce noise.
 - Supervisor encouraged analyzing model performance over time and across different periods.
- 5. Communicating Insights
 - Emphasis on explaining model behavior clearly, as if advising a public organization.
 - Consider how a stakeholder would interpret the results and which decisions could follow.
- 6. State-of-the-Art and Applications
 - The “Klimaatbeleid Dashboard” is not considered state-of-the-art but may be included as an example of a *potential application*.
 - State-of-the-art should focus on recent forecasting models applied to time series regression problems.
- 7. Requirements

- Requirements should be made SMART (Specific, Measurable, Achievable, Relevant, Time-bound).
- Example revision: “Forecast accuracy must be evaluated using MAE; RMSE, MAPE, and R^2 will be reported as supporting metrics. In case of divergence, MAE will serve as the leading criterion.”

Task	Deadline	Status
Expand literature review to include similar domains and techniques	ASAP	In progress
Choose and justify a leading evaluation metric	ASAP	In progress
Add qualitative analysis section on feature spikes and anomalies	This week	Planned
Improve graphs with clearer time labels	This week	Planned
Refine SMART requirements	Before final version	Planned

Questions for Supervisor – Feedback and Clarifications

Good Target audience?

Existing work

- Should I also say how I conducted literature research?
- How is the current flow of existing work?

Data and Modeling

- Lead features:
“Would it be methodologically good to lead the target variable (EV sales) by, for example, 4 months to better reflect causal policy or macroeconomic impact?”
- Lag features:
“What would you consider a robust strategy for selecting lag lengths per feature group (e.g., 6 months for sales, 3 months for search trends)?”
- Explaining anomalies:
“Would it be appropriate to annotate spikes in EV sales or Google Trends with related news articles from that period?”

Results and Visualizations

- Presenting results:
“What is the best way to present my model results in the final report? Should I use bar plots (e.g., R^2 , MAE by feature set), or a timeline per iteration?”

- Level of automation:
“Regarding the automation requirement: is a working dashboard expected, or would a conceptual interface description suffice?”

State-of-the-Art and Domain Knowledge

- State-of-the-art via gap filling:
“Is it acceptable that my ‘state-of-the-art’ is defined by the gap I am addressing (i.e., lack of open, interpretable EV forecasting models for public stakeholders)?”
- C3 – Sharing insights:
“I am unsure how to best demonstrate learning outcome C3 (‘sharing insights with peers’). Would a Medium blog or detailed GitHub repository be sufficient?” or is the thesis enough?

Accountability and Requirements

- Requirements from past projects:
“My requirements are based on earlier lab research and stakeholder input. Is this a valid and sufficient justification?”

Additional Question

71.4 27/05

Doel:

Rapport doormen, tolist maken voor komende week voor feedback moment:

71.4.1.1 Check of refferences werken:

Tabel met references die niet werkte

N#	Full
1	Government of the Netherlands, “Global Climate Strategy: From Ambition to Transition,” Tech. Rep., Dec. 2022. [Online]. Available: https://www.government.nl/binaries/government/documents/reports/2022/12/22/the-netherlands-global-climate-strategy/ the-netherlands-global-climate-strategy.pdf
5	F. Wijnhoven and O. Plant, “Sentiment Analysis and Google Trends Data for Predicting Car Sales.”
9	S. K. Kaya and Yildirim, “A PREDICTION MODEL FOR AUTOMOBILE SALES IN TURKEY USING DEEP NEURAL NETWORKS,” Journal of Industrial Engineering, 2020.
12	F. Chollet, Deep Learning with Python, Second Edition. New York: Manning Publications Co. LLC, 2021.
13	Jason Brownlee, Introduction to Time Series Forecasting with Python: How to Prepare Data and Develop Models to Predict the Future, v1.9 ed., 2020.
14	S. Aloorravi, Mastering Time Series Analysis and Forecasting with Python: Bridging Theory and Practice Through Insights, Techniques, and Tools for Effective Time Series Analysis in

	Python (English Edition), 1st ed. Delhi: Orange Education PVT Ltd, 2024.
16	REGULATION (EU) 2016/ 679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL - of 27 April 2016 - on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/ 46/ EC (General Data Protection Regulation)."
19	J. van Meerkirk, H. Vrijburg, J. Dijk, and G. Uitbeijerse, "Fiscale vergroening en de auto van de zaak."

71.5 23/05 70%

Feedback verwerking, evaluatie invullen

71.6 22/05

bijtelling is not a subsidy in the traditional sense. It is a tax incentive

71.7 21/05

Feedback verwerking

71.8 20/05

Feedback verwerking

71.8.1 Todolist:

- Requirements baseren op field research
- Verschilwe
- Add sources to formulas

Meeting with Kees

Attendees: Dylan Haazelager, Kees

Duration: approx. 45 minutes

Next meeting: Friday 09:00

Discussion Summary:

1. Research Changes
 - o We started by discussing the recent changes I made to my research, as outlined in earlier emails. The most notable change is the shift from news sentiment analysis to using Google Trends data as an indicator of consumer interest in electric vehicles (EVs).
 - o Kees immediately asked what exactly had changed as a result, and how my research differs from previous studies. I proposed that we go through the paper together.
2. Problem Definition and Scope

- Kees asked what the core problem of my research was. I admitted I struggled a bit to articulate it clearly, but explained it as follows:

NGOs and governmental organizations currently lack reliable forecasting models for predicting EV sales. Since electric vehicles are key to achieving zero-emission targets in the transport sector, having accurate forecasts is critical. My study explores whether open data sources like Google Trends can improve forecast accuracy beyond traditional historical sales data.

3. Research Gap Compared to Existing Studies

- Kees then asked about the research gap: what am I doing that hasn't been done before? I explained that my work is distinct in several ways:
 1. I focus specifically on the Dutch car market, which has been understudied.
 2. I use general EV-related keywords, whereas many studies focus on brand- or model-specific terms.
 3. My dataset spans a long time period (2007–2025), while others often cover a limited window.
 4. I include microeconomic indicators and other public data sources that are often overlooked.

4. Feedback on Results and Analysis

- We went through the results section together. Kees suggested adding a graph where I use 'common sense reasoning' to explain peaks and dips in EV sales.
- He also recommended running the model on each individual feature (Leave-One-Feature-Out analysis) and including those results in the report. Additionally, he advised integrating SHAP values to provide interpretability of the model's decisions.

5. Additional Suggestions

- Kees mentioned that the list of requirements seemed well-founded and intuitive but encouraged me to review them again critically to ensure they are sharp and well-supported.
- Finally, he pointed out that EV purchases typically respond with a delay to external factors. He advised adding a lag structure to the EV sales data to explore how many months of lag produce the best results.

Action Points:

- Add a graph to explain peaks/dips in sales using common sense reasoning.
- Run the model with Leave-One-Feature-Out and include it in the report.
- Integrate SHAP value analysis.
- Review the requirements critically for clarity and justification.
- Add lagged EV sales variables and test different lag lengths.

- Await written feedback from Kees after he reads the full paper later today.

71.9 19/05

Rest day

71.10 Zo18/05

Alles/leerdoelen check, (resultaten,data) Draft opsturen

"The model should serve as an exploratory benchmark for assessing the added value of behavioral and policy features beyond traditional univariate approaches."

71.11 Za 17/05

Model testing

Model	RMSE	MAE	R ²	MAPE	Opmerking
Prophet	11522.58	9279.27	-4.6566	27.28%	Slecht op alle metrics
Holt-Winters	1427.03	1140.87	-4.8417	20.63%	Verrassend stabiel
SARIMAX	4776.29	4161.23	-1.5153	18.37%	Redelijk, maar traag
SARIMA	4589.61	3732.98	-2.1472	16.11%	Beste klassieke model
XGBoost	6109.53	5328.81	-4.4232	20.61%	Niet goed ondanks tuning
LightGBM	4285.06	3235.11	-1.5048	38.51%	Matig, hoge MAPE
LSTM	5262.84	4533.36	-4.9191	18.29%	Matige resultaten
TFT	8099.60	7017.87	-1.7950	20.87%	Niet indrukwekkend
GRU					Werkt niet

Model exploration

	Model	Type & benadering	Exogene input	Interpretatie	Datavereisten	Typische prestaties & toepassingsgebied
	SARIMA (Seasonal ARIMA)	Statistisch tijdreeksmodel (lineair; met seizoenscomponenten).	Nee (univariante; exogene alleen via SARIMAX variant).	Hoog – Parameters (AR, MA) relatief begrijpelijk; inzicht in trend/seizoen.	Vereist voldoende historische data (≥ 2 seizoenscycli); maandelijks	Sterk bij duidelijke seizoenspatronen en lineaire trends; goede

				se frequentie mogelijk.	baseline-prestaties voor stabiele reeksen.
SARIMA X (ARIMA met exo)	Statistisch tijdreeksmodel met exogene regressors.	Ja (lineaire invloed van exogene variabelen).	Hoog – vergelijkbaar met SARIMA; coef. van exogene geven effectsschatting.	Zoals SARIMA, maar extra data voor alle exogene features; stabiele relaties vergen langere reeks.	Geschikt als bekende externe factoren belangrijke invloed hebben; behoudt SARIMA-prestaties bij lineaire relaties.
Holt-Winters (ETS)	Exponentiële gladstrijking (trend + seizoen; additive of multiplicative).	Nee (univariate).	Hoog – eenvoudige componenten (niveau, trend, seizoen).	Korte tot middellange reeksen volstaan (≥ 2 seizoenscycli); maand/kwartaal data.	Robuust voor eenvoudige patronen met geleidelijke trends; vaak vergelijkbaar of iets minder nauwkeurig dan ARIMA bij lineaire groei.
Prophet (Facebook)	Additief model (trend + seizoen + feestdagen; met Bayesian fitting).	Optioneel (feestdagen/events als dummy's of regressors).	Redelijk – biedt componenteninspectie (trend, seizoen, events).	Werkt goed met ontbrekende data en outliers; vereist historiek met meerdere jaren voor jaarseizoenen.	Ideaal bij sterke seizoenen en bekende events (vakanties, beleid); gebruiks vriendelijk, maar presteert matig bij grillige patronen.
Random Forest	Machine learning beslissingsboom	Ja (meerdere features/lagvariables)	Middel – enige inzicht via feature	Veel data helpt (meerdere	Krachtig voor complexe, niet-lineaire

		en-ensemble. mogelijk).		importance; geen expliciete tijdscomponent en.	jaren, of hoge frequentie ; vereist feature-engineering (lags, kalender).	relaties; robuust tegen overfit bij voldoende data; goede prestaties in diverse domeinen (retail, energie).
XGBoost (Gradient Boosting)	Geavanceerd boomgebaseerd ML-model (boosting).	Ja (multivariate input mogelijk).		Middel – feature importance en SHAP values geven enige interpretatie.	Vereist voldoende data met relevante kenmerken ; snelle training zelfs met veel features.	Zeer hoge voorspelkracht bij goed afgestemde hyperparameters; vaak top-prestaties in forecasting-wedstrijden; efficiënt en schaalbaar.
LightGBM	Gradient boosting machine (snelle variant).	Ja.		Middel – vergelijkbaar met XGBoost qua interpretatie.	Kan grote datasets aan; minder rekenintensief, dus bruikbaar voor hoge frequenties of veel series tegelijk.	Prestaties vergelijkbaar met XGBoost; praktisch voor industriële toepassingen met veel modellen (snelle iteratie, schaalbaar).
LSTM (RNN variant)	Deep learning recurrent netwerk (onthoudt volgorde-info).	Ja (multivariate tijdreeks mogelijk).		Laag – “black box”; moeilijk te doordringen, hoewel patroonherkenn ing intern plaatsvindt.	Vereist grote dataset of vele epochs training; profiteert van hoge frequentie of vele	Sterk in het leren van complexe sequentiële patronen (lange termijn afhankelijkhe den); heeft bewezen

				parallelle series; gevoelig voor overfitting bij <100 datapunten.	zeer hoge nauwkeurigheid als genoeg data beschikbaar is.
GRU (RNN variant)	Vereenvoudigde LSTM (minder parameters).	Ja.	Laag – vergelijkbaar met LSTM (black box).	Iets minder data-hongerig dan LSTM door eenvoud; schnellere training.	Presteert vaak gelijkwaardig aan LSTM, soms beter op kleinere datasets, met kortere trainings- en rekentijd.
TFT (Temporal Fusion Transformer)	Geavanceerd attentie-gebaseerd deep learning model.	Ja (ontworpen voor veel covariaten, incl. statische en bekende toekomstige).	Middel – complexe, maar <i>interpreteerbaar via attentie-visualisaties</i> (belang van features/tijdstippen).	Zeer veel data aanbevolen (veel series of hoge freq); hoge rekeneisen en tuning-complexiteit.	State-of-the-art prestaties bij multivariate <i>multi-horizon forecasts</i> ; combineert voordelen van klassieke en ML-modellen (accuratesse én inzicht), mits voldoende data en rekenkracht.

Model	R ² (gemiddeld)	Opmerkingen
SARIMA	+0.72	Werkt goed met alleen autoregressie
SARIMAX	-1.51	Slechter door zwakke exogene signalen
XGBoost	-8.15 tot -0.96	Faalt ondanks gridsearch, waarschijnlijk overfitting of slechte features

Modeltype	Doel	Geschikt?
SARIMA	Alleen historische verkoopdata gebruiken	baseline model
SARIMAX	Verkoop + externe indicatoren (zoals trends)	lineair, beperkt bij veel features
XGBoost	Voorspellen op basis van ALLE features	krachtig, geschikt voor dataset
LSTM/NN	Eventueel als bonus of vervolgonderzoek	complexer, veel tuning nodig

Alles/leerdoelen check, (resultaten,data)

Haha eigenlijk nog niks echt verbeterd, kijk ff wat grid search doet, en anders gewoon methode verder schrijven en verder gaan aan inleiding

71.12 16/05 70% Inleveren

Methode, probleemstelling inleiding

Probleem: r2 geeft aan dat MEAN BETER IS

Sarima results:

3 FOLD → 5 FOLD IETS BETERE RESULTATEN

Forecast Horizon	Average RMSE	Average MAE	Average MAPE (%)	Average R ²
3 months	2064.10	1918.53	10.08	0.4713
6 months	4589.61	3732.98	16.11	-2.1472
9 months	4646.68	3877.59	16.79	-1.8959
12 months	5314.93	4301.40	18.54	-1.8107

Hoewel de prestaties van het model voor korte termijn voorspellingen (3 maanden) redelijk zijn ($R^2 = 0.47$), blijkt uit de resultaten dat de voorspellende kracht aanzienlijk afneemt bij langere forecast horizons van 6 tot 9 maanden. Bij deze langere periodes is de R^2 zelfs negatief, wat aangeeft dat het model slechter presteert dan een eenvoudige voorspelling op basis van het gemiddelde.

Dit betekent echter niet dat het onderzoek is mislukt. De centrale onderzoeksraag ("To what extent can macroeconomic indicators, media, and digital trends predict EV sales?") blijft geldig. De resultaten geven aan dat deze externe indicatoren een zekere voorspellende waarde hebben op korte termijn, maar minder geschikt zijn voor robuuste langetermijnvoorspellingen.

Deze bevinding onderstreept de complexiteit van het voorspellen van EV-verkoop, en laat zien dat factoren zoals seizoensinvloeden, beleidswijzigingen en externe marktontwikkelingen mogelijk een grotere rol spelen dan met de huidige set macro-economische en digitale indicatoren kan worden opgevangen.

Verder vormt het negatieve resultaat bij lange forecast horizons ook een relevante bijdrage: het toont aan waar de grenzen liggen van lineaire tijdreeksvoorspellingen met deze feature set. Dit biedt richting voor vervolgonderzoek, bijvoorbeeld naar het toevoegen van exogene variabelen, het inzetten van niet-lineaire modellen, of het verfijnen van de forecast horizon.

71.12.1.1 *Vragen die in me op komen tijdens het schrijven:*

- Check cbs, overweeg toch wel 2^e hands autos te doen, waarom niet?
- Check of je dingen op git op verberen kan zetten
- Hoe zit het met de Requirements waarop zijn ze opgesteld

71.12.1.1.1 Corrolatie keywords is gedaan op oude trainset!! Wijzigen naar vokkedge set, of ?

71.12.1.1.2 Handmatig data van Bijtelling meenemen!

Verder gaan met requierments en dan methode/resultaten. Daarna weer terug naar existing work, AIDA helemaal weghalen heeft 0 refelantie

71.13 15/05

Check spelling, bronverwijzingen, consistentie, en formatteer alles volgens de eisen van je opleiding.

Voorbeeld van blok 1&2 project

Project 2:

- Alle A leerdoelen zijn op niveau voor
- B1 requirements zijn op niveau
- Dataset alleen onder, vanwege kleine set, ik kan dat benoemen

71.13.1.1 *Indeling logboek*

- Ai patterns, Level of automation
- Value proposition → stakeholder allignemnt
- Full moscow?
- Sketch → check dat dashboard A leerdloene allemaal voldoende in blok 2
- Keyword search methode
- B1 requirements zijn op v
- Eventueel bijlage met mailnaar ...
- Code snippet of corrolation keywords
-

71.13.1.2 *Overzicht peer reviewed papers:*

Source File Name	APA-Style Reference (Based on excerpt information)	Supporting Excerpts for Classification & Reference Info
A PREDICTION MODEL FOR AUTOMOBILE SALES IN TURKEY USING DEEP NEURAL networks.pdf	Kayapınar Kaya, S., & Yıldırım, Ö. (2020). A prediction model for automobile sales in Turkey using deep neural networks. <i>Journal of</i>	Format and references to academic journals

	<i>Industrial Engineering</i> , 31(1), 57-74.1	indicate peer review. Journal name, volume, issue, and pages provided1.
Advances in consumer electric vehicle adoption research A review and research agenda.pdf	Rezvani, Z., Jansson, J., & Bodin, J. (2015). Advances in consumer electric vehicle adoption research: A review and research agenda. <i>Transportation Research Part D: Transport and Environment</i> , 34, 122–136.2	Explicitly states it is a literature review in <i>Transportation Research Part D</i> ..., a peer-reviewed journal.
Castellacci and Santoalha - 2025 - Does digitalisation affect the adoption of electric vehicles New regional-level evidence from Googl.pdf	Castellacci, F., & Santoalha, A. (2025). Does digitalisation affect the adoption of electric vehicles? New regional-level evidence from Google Trends. (Journal details not fully specified in excerpt).4...	Authors explicitly thank the journal editor and anonymous reviewers4, confirming peer review. Authors listed5.
Economic, Functional, and Social Factors Influencing Electric Vehicles' Adoption.pdf	(Authors not provided in excerpt). (2022). Economic, Functional, and Social Factors Influencing Electric Vehicles' Adoption. <i>Sustainability</i> , 14, 6283.6	Published in the journal <i>Sustainability</i> 6. Authors thank anonymous reviewers7, confirming peer review.
Factors affecting adoption of electric vehicles in India.pdf	Chhikara, R., Gupta, S. K., & Sharma, C. (Year not provided in excerpt). Factors affecting adoption of electric vehicles in India. <i>Transportation Research Part D: Transport and Environment</i> . (Volume, issue, and pages not provided in excerpt).8...	Published in <i>Transportation Research Part D</i> 9, a known peer-reviewed journal. Authors provided8.
ForeXGBoost passenger car sales prediction based on XGBoost.pdf	Zhenchang, X., Xue, S., et al. (2020). ForeXGBoost: passenger car sales prediction based on	Described as an "Article in Distributed and

	<p>XGBoost. <i>Distributed and Parallel Databases</i>. (Full author list, volume, issue, and pages not provided in excerpt). https://doi.org/10.1007/s10619-020-07294-y10</p>	Parallel Databases" with DOI and citations listed10. Includes funding acknowledgements11.
Forecasting electric vehicles sales with univariate and multivariate timeseries models the case of china.pdf	Zhang, Y., Zhong, M., Geng, N., & Jiang, Y. (2017). Forecasting electric vehicles sales with univariate and multivariate time series models: The case of China. <i>PLoS ONE</i> , 12(5), e0176729. https://doi.org/10.1371/journal.pone.0176729 12	Explicitly states publication in <i>PLoS ONE</i> , an open access journal with detailed publication dates and DOI12.... Authors and contributions listed14.
Online Forecasting and Anomaly Detection Based on the ARIMA Model.pdf	Kozitsin, V., Katser, I., & Lakontsev, D. (2021). Online forecasting and anomaly detection based on the ARIMA model. <i>Applied Sciences</i> , 11(7), 3194. https://doi.org/10.3390/app11073194 15	Header indicates "applied sciences Article" with detailed publication information (journal, volume, issue, article number, DOI, dates, editor)15.
Profit_Prediction_Using_ARIMA_SARIMA_and_LSTM_Mode.pdf	Sirisha, U. M., Belavagi, M. C., & Attigeri, G. (Publication date not fully specified in excerpt). Profit prediction using ARIMA, SARIMA and LSTM models in time series forecasting: A comparison. (Publication name, volume, issue, and pages not fully specified in excerpt). https://doi.org/10.1109/ACCESS .16	Format includes DOI16, typical for published peer-reviewed articles, likely in an IEEE publication. Authors listed16.

	2017.DOI16	
Research Paper Michael Visser v4.0.pdf	Visser, M. (Date unknown). Research Paper Michael Visser v4.0. (Publication venue not specified in excerpt).	Titled "Research Paper" with an author name. Structured like an academic paper with figures and extensive references to peer-reviewed works ¹⁷ Likely intended for peer review/publication.
Sentiment Analysis and Google Trends Data for Predicting Car Sale.pdf	Wijnhoven, F., & Bloemen, O. (Year not provided in excerpt, likely 2017). Sentiment analysis and Google Trends data for predicting car sale. <i>Thirty Eighth International Conference on Information Systems</i> , 1-15.22	Authors thank anonymous ICIS reviewers, indicating it was reviewed for the International Conference on Information Systems ²² , a peer-reviewed conference.
The Showrooming Phenomenon It's More than Just About Price.pdf	(Authors and publication details not provided in excerpt). (Year not provided in excerpt). The Showrooming Phenomenon It's More than Just About Price. (Publication venue not specified in excerpt). ²³	Explicitly thanks the "Editor and reviewer" ²³ , confirming peer review for a journal or similar publication.
The influence of consumers technological and environmental considerations.pdf	B.D., & H.-G.H. (Full names not provided in excerpt). (2021). The influence of consumers technological and environmental considerations. <i>Sustainability</i> , 13(unknown article number), 12025. ²⁴ ...	Header indicates "FOR PEER REVIEW" ²⁶ Published in <i>Sustainability</i> with

		acceptance and publication dates ²⁴ . Authors' contributions listed ²⁸ .
Using Forum and Search Data for Sales Prediction of High-Involvement Projects.pdf	Geva, T., Oestreicher-Singer, G., Efron, N., & Shimshoni, Y. (2017). Using forum and search data for sales prediction of high-involvement projects. <i>MIS Quarterly</i> , 41(1), 65–82.29...	Identified as a "RESEARCH ARTICLE" in <i>MIS Quarterly</i> ³⁰ with full publication details ²⁹ . Authors thank editor, associate editor, and reviewers ³¹ , confirming peer review.
Using machine learning methods to predict electric vehicles penetration in the automotive market.pdf	(Authors not provided in excerpt). (2023). Using machine learning methods to predict electric vehicles penetration in the automotive market. (Publication venue not specified in excerpt). ³² ...	Includes "Received" and "Accepted" dates ³² . Mentions "Open Access" license ³⁴ ... and "Springer Nature journal content" [135, specific to "Using machine learning methods to predict electric vehicles.pdf", but likely applies to both similar titles]. Structured as a published paper.
Using machine learning methods to	(Authors not provided in excerpt).	Includes

predict electric vehicles.pdf	(2023). Using machine learning methods to predict electric vehicles. (Publication venue not specified in excerpt). ³⁶ ...	"Received" and "Accepted" dates ³⁶ . Mentions "Open Access" license ³⁵ and "Springer Nature journal content" ³⁷ . Structured as a published paper.
-------------------------------	--	---

No peer:

Source File Name	APA-Style Reference (Based on excerpt information)	Supporting Excerpts for Classification & Reference Info
<i>A. Google trends as complementary tool for new car sales forecasting A cross-country comparison along the customer journey,.pdf</i>	(Author not provided in excerpt). (Year not provided in excerpt). <i>A. Google trends as complementary tool for new car sales forecasting: A cross-country comparison along the customer journey.</i> (Thesis type and Institution not provided in excerpt). ³⁸	Refers to the document as a "thesis" ³⁸ .
<i>Google trends as a complementary tool for new car sales forecasting.pdf</i>	(Author not provided in excerpt). (Year not provided in excerpt). <i>Google trends as a complementary tool for new car sales forecasting.</i> (Thesis type and Institution not provided in excerpt). ³⁹	Refers to the document as a "thesis" ³⁹ .
<i>Rationale_anoniem.pdf</i>	(Author and Date not provided in excerpt). <i>Rationale_anoniem.pdf.</i> (Document type not specified, appears to be a rationale or proposal). ⁴⁰	Refers to "This research" ⁴⁰ and appears to be a foundational document for a research project, not a finished publication. The file name also suggests it's a rationale ⁴⁰ .

71.13.1.3 *Time series modellen:*

Arima en Sarima gevonden

<https://www.datacamp.com/tutorial/arima> keuze voor modellen:

71.13.1.4 *TODO vandaag:*

To-Do List: Improving Introduction + Problem

- check if there are official documents supporting my approach.

A. Structure and Alignment with HvA Standards (e.g., Steur rationale)

1. Add a "Purpose and Structure" paragraph at the start of the Introduction to explain:

- o The rationale behind the research
- o What the reader can expect (e.g., method, data, analysis)

2. Simplify and streamline the background:

- o Start with a clear policy target (e.g., 100% EV new sales in 2030)
- o Follow with digital behavior as an opportunity

3. Clarify the target audience (NGOs, policymakers with limited budgets)

4. Add transitional subheadings (e.g., Background, Problem, Contribution, Research Design)

5. Avoid mixing methodological claims into the background — move to Method section

B. Content-Based Refinement

6. Replace vague or general claims with cited references

- o Example: "lagging indicators" → needs a supporting source

7. Avoid focusing solely on Google Trends:

- o Mention it as a case, not as the core idea
- o Generalize to "digital behavioral indicators"

8. Remove/reduce AIDA framework: only briefly mention it as a conceptual anchor, not a full model

9. Rephrase speculative or overstated claims (e.g., weak performance of sentiment → nuance this)

10. Address the Google Trends keyword gap:

- o Mention that brand-specific terms dominate existing research
- o Your approach uses more general terms drawn from policy sources

11. Discuss policy variables (e.g., bijtelling) clearly:

- o Cite RVO or Belastingdienst
- o Explain how these variables are constructed and lagged

12. Move detailed keyword engineering techniques (e.g., TF-IDF, YAKE) to methodology, not intro

C. Clarify and Strengthen the Problem Statement

13. Clearly define the gap:

- Short timeframes
- Brand-specific trends
- No fusion with macroeconomic variables

14. Specify consequences of poor forecasting:

- Over/underestimating infrastructure needs, misaligned subsidies

15. Link the problem explicitly to your data scope:

- Monthly national BEV data (2007–2025)

16. Use "*too limited, too brand-specific, too short-term*" as a structural problem framing

D. Methodological Anchoring

17. Avoid embedding model names (e.g., XGBoost, CNN) in the introduction

18. Instead, state that you'll compare interpretable and complex models

19. Clarify why MAE is used — and if interpretability is part of your focus

20. Define how you apply lagging for time series features (sales vs. trends)

21. Justify your time split approach:

- Explain why you use TimeSeriesSplit or KFold
- Why walk-forward validation is more appropriate for your case

E. Technical Fixes and Polish

22. Improve title: current version is vague

23. Ensure all references have DOI or working links

24. Improve citation quality: focus on stronger sources

25. Avoid repetition of similar points (e.g., Google Trends brand keywords)

- Methode verder uitwerken
 - Modelvergelijkingen
 - eerste resultaten
- Inleiding verbeteren, meer vergelijken met inhoud van Gogreen en opmaak rationale
- Data verbeteren, data van bijtelling citeren
- Time series split verantwoorden met sklearn
 - [KFold](#)
 - [Lagged features for time series forecasting](#)

71.14 14/05

Cross validation + lag https://scikit-learn.org/stable/modules/cross_validation.html#time-series-split

71.14.1.1 Onderwerp: Feedback op scriptievoorstel en richting

Samenvatting gesprek:

- Feature correlatie en selectie (Leerdoel B):
Marcio wees erop dat veel van mijn features waarschijnlijk onderling sterk correleren. Dit kan een uitdaging vormen voor feature selection methoden, omdat overlappende informatie het moeilijk maakt om individuele bijdrage van variabelen te bepalen. Hij raadde aan hier kritisch naar te kijken in relatie tot leerdoel B.
- Modelkeuze (Leerdoel B):
De keuze voor modellen zoals CNN en XGBoost was volgens Marcio nog onvoldoende onderbouwd. Hij stelde voor om krachtigere of recentere modellen te overwegen, en het modelkeuzeproces explicet toe te lichten (waarom juist deze modellen, en niet bijvoorbeeld LSTM of transformers?).
- Featurebeschrijving:
Marcio gaf aan dat een duidelijke uitleg van de gebruikte features ontbrak in de draft. Ik heb afgesproken dit vandaag toe te voegen.
- Data-splitsing / validatie:
De huidige split in mijn model houdt nog geen rekening met de tijdsstructuur van de data. Marcio adviseerde een time-series split te gebruiken om te voorkomen dat het model toekomstige data leert van informatie in het verleden (data leakage).
- Kwaliteit van bronnen (Leerdoel A):
De literatuurkwaliteit moet omhoog. Veel van de huidige bronnen hebben een lage impact. Marcio benoemde bron 8 als krachtig en adviseerde om die dieper te analyseren en te benutten.

Belangrijke bron:

- Gallagher & Muehlegger (2011): “Giving green to get green: Incentives and consumer adoption of hybrid vehicle technology”

71.14.1.2 Vraag aan marcio:

- Ik wil moscow methode voor requierments doen, waar zal ik dat op baseren?
- Vragen over gap op modellen
- Vragen voor tips keyboard search selectie/ evt auteur mailen
- Waarom noemen ze scriptie rationale?
- Roling avarage per keyword te printen
- Is it promessing?
- Difference between rationale and normal scritpie
- Evt top 4 per dataset doen, dus 4 maanden, 4 diesel, 4 benzine, 4 consumentenvetrouwen, keywords
- Check of je ook single comom mag

71.14.1.3

English Top 10 meest gecorreleerde zoektermen met EV-verkoop:

charging stations: 0.713

electric car: 0.705

plug in hybrid: 0.627

electric truck: 0.613

electric bus: 0.593

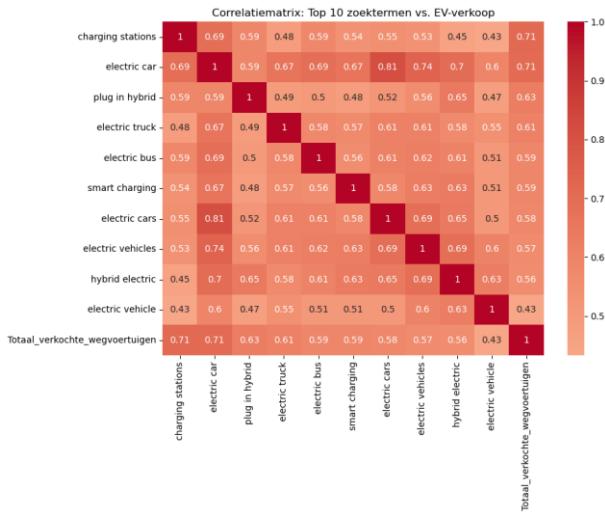
smart charging: 0.587

electric cars: 0.581

electric vehicles: 0.570

hybrid electric: 0.562

electric vehicle: 0.433



71.14.1.4 Top 10 meest gecorreleerde zoektermen met EV-verkoop:

thuis opladen: 0.793

elektrische auto: 0.788

hybride elektrisch: 0.781

elektrische bus_x: 0.758

elektrische bus_y: 0.758

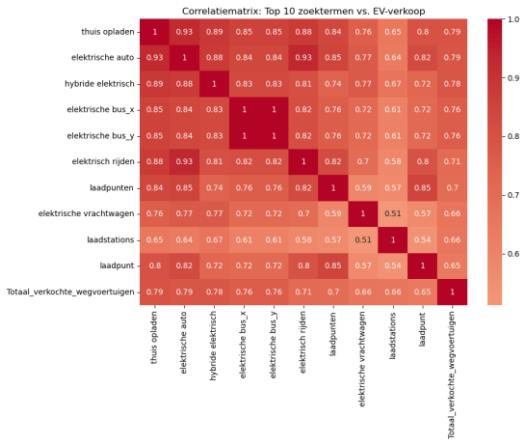
elektrisch rijden: 0.706

laadpunten: 0.701

elektrische vrachtwagen: 0.665

laadstations: 0.662

laadpunt: 0.647



71.14.1.5 Used invoer google trends

01/01/2007

Google Trends URL

<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=toegankelijk%20laders>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=toegankelijk%20laadpunten>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=toegankelijk%20snelladers>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=alternatieve%20brandstoffen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=infrastructuur%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=batterij-elektrisch%20voertuig>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=batterij-elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=batterijwisselsystemen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=elektrische%20voertuigen%20op%20celbasis>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=laadinfrastructuur>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=laadnetwerk>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=laadpunt>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=laadstations>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=vraag%20naar%20elektrische%20voertuigen>

<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=uitrol%20van%20elektrische%20auto%E2%80%99s>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=uitrol%20van%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=elektrische%20bus>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=elektrische%20bussen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=elektrische%20auto>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=markt%20voor%20elektrische%20auto%E2%80%99s>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=modellen%20van%20elektrische%20auto%E2%80%99s>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=aankoopprijs%20van%20elektrische%20auto%E2%80%99s>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=verkoop%20van%20elektrische%20auto%E2%80%99s>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=voorraad%20van%20elektrische%20auto%E2%80%99s>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=elektrische%20auto%E2%80%99s>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=subsidies%20voor%20elektrische%20auto%E2%80%99s>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=elektrisch%20rijden>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=elektrische%20lichte%20bedrijfsvoertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=elektrische%20lichte%20bedrijfswagens>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=elektrische%20vrachtwagen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=elektrische%20vrachtwagens>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=elektrische%20tweewielers>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=elektrisch%20voertuig>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=adoptie%20van%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=batterijen%20voor%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=batterij%20voortuig>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=opladen%20van%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=laadinfrastructuur%20voor%20elektrische%20voertuigen>

<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=uitrol%20van%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=wagenpark%20van%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=initiatief%20voor%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=landen%20binnen%20het%20initiatief%20voor%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=markt%20voor%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=markten%20voor%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=modellen%20van%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=vooruitzichten%20voor%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=aankoopprijs%20van%20elektrisch%20voertuig>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=verkoop%20van%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=marktaandeel%20van%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=aanbod%20van%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=laadapparatuur%20voor%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=uitrol%20van%20laadapparatuur%20voor%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=laadpunten%20van%20laadapparatuur%20voor%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=groei%20van%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=aantallen%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=data%20over%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=initiatief%20voor%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=subsidies%20voor%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=voertuigen%20met%20uitstoot>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=financi%C3%A9%20prikkels%20voor%20elektrische%20auto%E2%80%99s>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=financi%C3%A9%20prikkels%20voor%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=waterstofcel-elektrische%20auto>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=waterstofcel-elektrisch%20voertuig>

<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=wereldwijde%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=wereldwijd%20wagenpark%20van%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=wereldwijde%20data%20over%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=thuis%20opladen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=hybride%20elektrisch>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=hybride%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=auto%27s%20met%20verbrandingsmotor>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=voertuig%20met%20verbrandingsmotor>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=laders%20voor%20lichte%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=verkoop%20van%20lichte%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=nieuwe%20energievoertuigen>
[https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=aantal%20elektrische%20voertuigen](https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=aantal%20elektrische%20auto%E2%80%99s)
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=plug-in%20hybride>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=plug-in%20hybride%20elektrisch%20voertuig>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=plug-in%20hybride%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=priv%C3%A9%20laders>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=priv%C3%A9%20laden>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=publieke%20laders>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=publiek%20laden>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=publiek%20toegankelijke%20laders>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=publiek%20toegankelijke%20laadmogelijkheden>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=aankoopstimulansen%20voor%20elektrische%20auto%E2%80%99s>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=aankoopstimulansen%20voor%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=aankoop%20van%20elektrische%20auto%E2%80%99s>

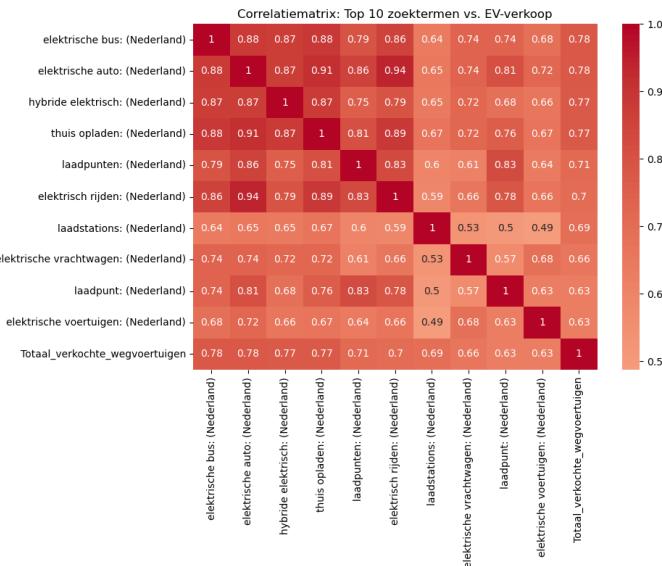
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=aankoop%20van%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=aankoopprijs%20elektrische%20auto>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=aankoopprijs%20van%20elektrisch%20voertuig>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=aankoopsubsidies%20voor%20elektrische%20auto%E2%80%99s>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=aankoopsubsidies%20voor%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=verkoopaandeel%20van%20elektrische%20auto%E2%80%99s>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=verkoopaandeel%20van%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=aandeel%20elektrische%20auto%E2%80%99s>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=aandeel%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=slim%20laden>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=voorraad%20van%20elektrische%20auto%E2%80%99s>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=voorraad%20van%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=transitie%20naar%20elektrisch>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=acceptatie%20van%20elektrische%20auto%E2%80%99s>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=acceptatie%20van%20elektrische%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=opladen%20van%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=elektrofificatie%20van%20voertuigen>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=nul-emissie%20voertuig>
<https://trends.google.com/trends/explore?date=2007-01-01%202025-01-01&geo=NL&hl=nl&q=nul-emissie%20voertuigen>

71.14.2 Results new keywords:

Google Trends normaliseert data binnen de groep van ingevoerde zoektermen. Dit betekent:

- De zoekterm met de hoogste piek in die groep krijgt altijd een waarde van 100.
- Alle andere termen worden daaraan relatief geschaald.

correlatie with Totaal ev verkoop =



71.15 13/05

Schrijf je methodologie-hoofdstuk af: geef uitleg over je model selectie (XGBoost/cnn), evaluatie (MAE), en validatiestrategie

71.15.1.1 Feedback marcio

Hello Dylan,

I looked at your report. It is progressing (+). Still, many points (some related to the scope) need to be addressed. Below are some comments I made while reading your paper.

(?) Title:

It is hard to grasp. The sentence ("Uncovering the Drivers of Car Sales") seems ambiguous. Also, "grouped feature importance" is just one of the techniques used, and it's common in the data preparation and training phases. Unless this technique is the only one used (reading it entirely, I don't think so), I would not highlight it. You still have time to refine your title, but this version is unclear.

(!) References:

Please include the DOI URL for each referenced paper. If a DOI is unavailable, add a direct link to view or download the resource (e.g., reference 1 is a government document). Some references are incomplete—for instance, source 5 lacks full citation details.

(-) Based on what is this sentence?

("However, these factors are typically derived from surveys or lagging indicators, limiting their real time predictive utility").

Strong claims like this must be backed by a reference, a peer-reviewed paper or book.

(-) Regarding the sentence:

("In this context, Google Trends offers a unique view into early stage consumer interest in EVS by capturing real-time search behaviour").

It can be valid, but why only Google Trends? Focusing exclusively on it may limit the generalisability of your research.

(-) You refer to the AIDA method and cite source four as a basis.

However, I couldn't find any reference to this method in the abstract. Please confirm. It also appears that the methodology was introduced early in the paper but not followed through in the rest of the text. Is AIDA relevant for your entire work?

(-) Regarding the sentence:

("This study proposes a complementary approach by integrating Google search volume and behavioural-economic indicators").

Avoid sticking your concept to a single data source. The idea should be broadly conceptualised, and as a pragmatic decision, it would be fine to adopt Google Trends as an example of an application.

(-) Problem Statement:

Regarding the sentence ("While recent research has explored the use of behavioural signals such as Google search trends to model EV interest, many approaches still rely on short time windows or narrow keyword sets, often limited to specific car models").

What is your source for this statement?

(-) The problem statement is too general.

It should be more closely aligned with the "A" learning outcomes. Consider clarifying:

- Who are the stakeholders interested in EV sales forecasting? What are the opportunities for good forecasting, and what are the impacts of "bad" forecasting?
- Why is it relevant?
- What is currently being done in the field?

Also, the limitations of such approaches should be discussed. EV adoption is a complex phenomenon influenced by social, political, and technical factors, so multiple variables are at play.

(-) Regarding the sentence:

("The weak performance of sentiment data in predicting sales led the authors to conclude that behavioural search signals may offer more actionable insights, though their influence appears limited for high involvement purchases like cars, possibly due to the overriding impact of external factors such as tax policies").

This conclusion seems premature, especially because you mentioned very low statistical significance. Please avoid strong causal statements in such cases.

(?) Regarding the sentence:

("While their study was adoption-focused and cross-sectional, it validated Google Trends as a real-time behavioural proxy.")

This sentence is unclear—please rephrase for clarity.

(!) Literature Review:

The quality of sources should be improved. Most cited papers have very low citation counts, except for article 8, which offers a good overview. Make better use of that source to reinforce your problem statement.

(!) Regarding the sentence:

("This underlines the importance of timing and visibility in financial incentives that provide immediate benefits or are simpler to access appear more effective in influencing consumer decisions towards EV adoption").

Please clarify this statement. What exactly do you mean, and how does it relate to your research?

(!) Regarding the statement:

("The authors stated that gasoline prices influence the total cost of vehicle ownership and are expected to be associated with consumer car purchase decisions"). Please explain better.

(?) "Their six-month sliding window and feature selection via information gain allowed the model to outperform traditional regressors in both speed and accuracy."

What do you mean by "traditional regressors"? Please specify.

(-) You mention subsidies.

How do you plan to use this information? It would require tracking government subsidy programs over specific timeframes. This should be very well explained in the methodology section

(?) Regarding the sentence:

("Behavioural indicators like Google Trends are often analysed in isolation, without normalisation correction across years, which impedes comparability over time").

What do you mean by "normalisation correction across years"? Normalised based on what? Please explain.

(?) Is explainability your focus?

You mention that few studies compare interpretable nonlinear methods (e.g., XGBoost) with deep learning (e.g., CNN or LSTM). It raises the question: Are you really addressing the trade-off between accuracy and interpretability? To which target audience?

(?) Why compare XGBoost and CNN?

Why specifically CNN? You also mention LSTM later, but more efficient models are available. You should work on the rationale for your choices.

(?) Existing Work and Gaps:

Keyword selection is an implementation detail. More important is *what kind of information* you want to capture from documents. That part belongs in the methodology section. We need to understand the data at hand better to choose between technique A or B.

(!) Fix repetition:

There are multiple mentions of the limitation related to reliance on specific car brands—please consolidate.

(!) The keyword search is too limited.

More advanced and effective methods exist for semantic and temporal signal extraction.

(!) Google Trends in the research question?

Avoid embedding specific data sources into the research question—make it broader.

(!) Research Question:

Consider a broader framing:

To what extent can macroeconomic indicators, media, and digital trends predict EV sales?

Then break this into sub-questions like:

- a) Which macroeconomic indicators influence EV sales?
- b) Which digital trends (e.g., media coverage, search interest) play a role?
- c) What would a predictive model look like using fused data from these sources?

(-) Requirements:

These need to be specific and measurable. For example, "Use interpretable and scalable models" is too vague. Interpretable for whom? Also, who is your target audience? It must be defined to tailor your requirements accordingly. Note that in the discussion section, you'll need to revisit them, so the more ambiguity, the harder it is to answer them.

Data:

There's no description of what Google Trends provides. Exploratory Data Analysis (EDA) is essential. What does your dataset look like? What are the distributions, completeness, and statistical properties?

TF-IDF and YAKE are mentioned too soon and are somewhat outdated. Better techniques exist today. Also, the macroeconomic and policy data descriptions are not yet present.

Methodology:

Please justify your model selection: Why XGBoost, CNN, LSTM (?)—and why not other alternatives? How you plan to combine macroeconomic, digital, and policy data is also unclear. Consider specifying:

- Which features and transformations?
- How do spatial/temporal relationships come into play?

Quantitative Analysis:

Why use MAE and not other regression metrics? Provide a rationale.

Qualitative Analysis:

Are you including one? For instance, will you reflect on why specific policy, social, or economic indicators correlate with EV sales?

71.15.1.2 Verslag begeleidingsgesprek met scriptiebegeleidster (Daphne)

Vandaag heb ik een gesprek gehad met mijn huiswerkbegeleidster over de voortgang van mijn scriptie. Hieronder een samenvatting van de belangrijkste besproken punten:

71.15.1.2.1 1. Keuze van EV-gerelateerde zoektermen

We begonnen met een discussie over welke zoekwoorden ik het best kan gebruiken in Google Trends. Daphne gaf aan dat het gebruik van algemene EV-gerelateerde termen (zoals "elektrische auto", "EV subsidie") een goede keuze is.

71.15.1.2.2 2. Focus van de scriptie

Daphne benadrukte dat de focus van mijn scriptie nu nog te veel ligt op de zoektermen zelf. Ze adviseerde om de nadruk meer te leggen op de samenstelling en bruikbaarheid van de datasets die ik gebruik. De datasetconstructie is namelijk waar ik het meeste tijd in heb gestoken.

71.15.1.2.3 3. Gebruik van NOS als bron

Hoewel ik eerder heb besloten geen sentimentanalyse te doen, gaf Daphne aan dat het toch goed is om de NOS-dataset te benoemen, juist om te onderbouwen waarom deze niet bruikbaar was. Het draagt bij aan het verhaal over welke datasets ik heb onderzocht en waarom sommige zijn afgevallen.

71.15.1.2.4 4. Doelgroep van het onderzoek

We hebben samen nagedacht over een geschikte doelgroep. De voorkeur ging uit naar overheden en NGO's die zich bezighouden met klimaatbeleid. Deze doelgroep heeft meer baat bij inzichten op basis van algemene zoektermen dan bijvoorbeeld autodealers, en beschikt vaak over een beperkt budget. Gratis databronnen zoals Google Trends en CBS zijn voor hen extra relevant.

71.15.1.2.5 5. Gebruik van het AIDA-model

In plaats van het AIDA-model uitgebreid te behandelen, stelde Daphne voor om kort te benoemen dat Google Trends een populaire methode is om gedrag of interesse rondom aankoopbeslissingen te analyseren.

71.15.1.2.6 6. Modellering en onderzoeksraag

Ik vroeg hoe ik het beste mijn modellering kon aanpakken. Daphne stelde voor om meerdere modellen te proberen zonder daar een strikte hypothese aan te koppelen. De kernvraag is immers welke dataset het beste werkt binnen een model, niet welk model het beste presteert.

71.15.1.2.7 7. Keyword selectie en vertaling

Omdat mijn huidige lijst met keywords in het Engels is, wil ik deze vertalen naar het Nederlands. Hoewel Daphne aangaf dat ik gewoon de meest logische Nederlandse zoektermen kan gebruiken, wees ik erop dat dit eerder bij de sentimentanalyse niet goed uitpakte. Een alternatieve aanpak is om alle keywords op correlatie met verkoopdata te testen en op basis daarvan de top 5 te selecteren, maar dit vergt veel tijd.

71.15.1.2.8 8. Concrete verbeterpunten voor scriptierapport

- Existing work en gap samenvoegen tot één logisch geheel.
- Subkopjes toevoegen aan het onderdeel 'existing work' om de structuur te verbeteren.
- De prestatieraatstaf die ik gebruik benoemen in dit deel.

71.15.1.2.9 9. Nieuwe formuleringen voor de onderzoeksraag (suggestions)

- *What dataset provides the most accurate sales forecasting for electric vehicles in the Netherlands?*
- *What insights do available datasets provide for predicting EV sales in the Netherlands?*
- *Can the combination of Google Trends, fuel prices, and consumer confidence accurately predict EV sales in the Netherlands?*

71.15.1.2.10 10. Modelselectie en preprocessing

Ik moet nog beslissen welke modellen ik precies gebruik. Mogelijk kan ik werken met standaard hyperparameters. Daarna richt ik me weer op het preprocessen van de data, waaronder het toepassen van lagtime.

Wat betreft de bijtelling wilde ik die variabele één maand eerder trainen omdat de verkoop piekt vóór het ingaan van een nieuwe bijtellingsregeling. Maar dit blijkt niet mogelijk: toekomstige data gebruiken voor het trainen op het verleden is methodologisch onjuist, dus ik laat dit zoals het is.

Voor nu ligt de focus op het toepassen van lagtime op de verkoopcijfers. Eventueel onderzoek ik later ook of lagtime op Google Trends zinvol is. Omdat eerdere studies model-specifieke zoektermen gebruikten (zoals "Tesla Model 3"), was de lag daar logisch. Bij mijn bredere zoektermen (zoals "elektrisch rijden") is het minder duidelijk of een lag evenveel effect heeft.

Tot slot: ik kijk haar mail nog even na over dit onderwerp.

71.15.1.3 BIJTELLING GEVONDEN!



bpm_tarieven_bpm
0651z14fd.pdf

BPM is niet hetzelfde als bijtelling

71.15.1.4 Bijtelling % vinden:

Bijtelling zichtbaar op je salarisstrook

Vanaf 2007 'verhuisde' de bijtelling van inkomstenbelasting naar loonbelasting. Het resultaat daarvan was, dat de werkgever de bijtelling ging verrekenen met de salarisstrook. De berijder had dus elke maand direct inzicht in zijn bijtellingskosten. Je kunt je voorstellen dat de zakelijke rijder op een andere manier naar de leaseauto ging kijken.

Vanaf 2009 werd het algemene bijtellingstarief aangepast van 22% naar 25%. Ook werden er extra bijtellingscategorieën geïntroduceerd voor zuinigere auto's. Deze categorieën werden gebaseerd op de CO2-uitstoot van een zakenauto. Ook werden de CO2-grenzen elk jaar bijgesteld.

De belangrijkste wijziging bij de bijtellingscategorieën is de introductie in 2012 van het 0% bijtellingstarief voor [auto's zonder CO2 uitstoot](#) (bijvoorbeeld elektrische en waterstofauto's). Dat jaar werd ook het bijtellingstarief voor [youngtimers](#) geïntroduceerd, namelijk 35% over de dagwaarde van 15 jaar oude auto's.

[https://www.vzr.nl/nieuws/details/de_geschiedenis_het_ontstaan_van_de_bijtelling#:~:text=De%20belangrijkste%20wijziging%20bij%20de,\(bijvoorbeeld%20elektrische%20en%20waterstofauto's\).](https://www.vzr.nl/nieuws/details/de_geschiedenis_het_ontstaan_van_de_bijtelling#:~:text=De%20belangrijkste%20wijziging%20bij%20de,(bijvoorbeeld%20elektrische%20en%20waterstofauto's).)

Bijtelling zichtbaar op je salarisstrook

Vanaf 2007 'verhuisde' de bijtelling van inkomstenbelasting naar loonbelasting. Het resultaat daarvan was, dat de werkgever de bijtelling ging verrekenen met de salarisstrook. De berijder had dus elke maand direct inzicht in zijn bijtellingskosten. Je kunt je voorstellen dat de zakelijke rijder op een andere manier naar de leaseauto ging kijken.

Vanaf 2009 werd het algemene bijtellingstarief aangepast van 22% naar 25%. Ook werden er extra bijtellingscategorieën geïntroduceerd voor zuinigere auto's. Deze categorieën werden gebaseerd op de CO2-uitstoot van een zakenauto. Ook werden de CO2-grenzen elk jaar bijgesteld.

De belangrijkste wijziging bij de bijtellingscategorieën is de introductie in 2012 van het 0% bijtellingstarief voor **auto's zonder CO2 uitstoot** (bijvoorbeeld elektrische en waterstofauto's). Dat jaar werd ook het bijtellingstarief voor **youngtimers** geïntroduceerd, namelijk 35% over de dagwaarde van 15 jaar oude auto's.

[https://www.vzr.nl/nieuws/details/de_geschiedenis_het_ontstaan_van_de_bijtelling#:~:text=De%20belangrijkste%20wijziging%20bij%20de,\(bijvoorbeeld%20elektrische%20en%20waterstofauto's\).](https://www.vzr.nl/nieuws/details/de_geschiedenis_het_ontstaan_van_de_bijtelling#:~:text=De%20belangrijkste%20wijziging%20bij%20de,(bijvoorbeeld%20elektrische%20en%20waterstofauto's).)

Naar één bijtellingstarief

Vanaf 2017 is niet meer 25%, maar 22% het algemene bijtellingstarief geworden. Auto's zonder CO2-uitstoot genieten nog steeds van een bijtellingskorting, al wordt dit elk jaar wel zodanig aangepast, dat alle auto's vanaf 2026 onder het bijtellingstarief van 22% vallen.

Naar één bijtellingstarief

Vanaf 2017 is niet meer 25%, maar 22% het algemene bijtellingstarief geworden. Auto's zonder CO2-uitstoot genieten nog steeds van een bijtellingskorting, al wordt dit elk jaar wel zodanig aangepast, dat alle auto's vanaf 2026 onder het bijtellingstarief van 22% vallen.

71.15.1.5 *Data preprossesing:*

XGboost	F%	MAE	MAPA
Met zelfde gasprijzen mean,first last min max etc: En keywords: ev electrisch rijden, elektrische auto, ev nederland Geen Bijtelling%	29.22	3477.37	
Alleen mean per maand gas	31.39	3618.48	
Met nieuwe keywords en bijtelling%	43.30	4202.57	23.34

--	--	--	--

Met zelfde gasprijzen mean, first last min max etc:

Gemiddelde absolute foutmarge (%): 29.22

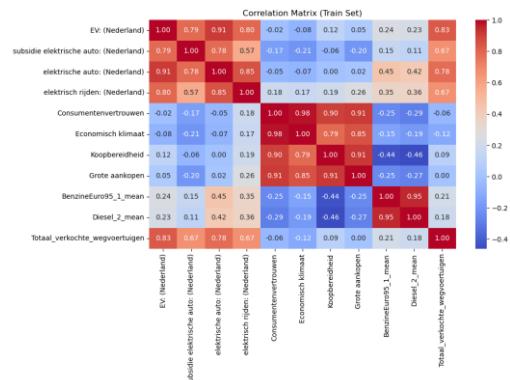
Overall MAE op validatieset: 3477.37

Maakt minimaal verschil

Zonder:

Gemiddelde absolute foutmarge (%): 31.39

Overall MAE op validatieset: 3618.48



Met alleen mean per maand voor pomp en met een draft van bijtelling:

Cnn voor was Gemiddelde absolute foutmarge (%): 31.03

Mean Absolute Error (MAE): 3149.71

71.15.1.6 Feedback Daphne:

71.15.1.6.1 Background:

ik heb zelf wat twijfels erbij dat AIDA model te introduceren, zeker omdat het erop lijkt dat je dit een beetje last minute gevonden hebt. Het is inderdaad wel duidelijk dat Google Trends goed de 'Attention' kan weerspiegelen, maar er zijn zeker nog andere 'attention' aspecten die je in deze thesis niet meeneemt (bijv. social media posts/reacties over EVs)

Daarnaast neem je ook niet per se zo veel 'Desire' data mee, zoals andere papers dat doen (bijv. gegevens over 'environmental' concerns van mogelijke klanten, informatie over subsidies etc).

Vandaar dat ik niet vind dat het AIDA framework super sterk overeenkomt met jouw approach, en zou ik het daarom ook niet per se noemen

maybe good to mention how they obtained this data (and at some point later maybe justify why you couldn't obtain that kind of data)

71.15.1.6.2 Existing work:

I wouldn't really say that XGBoost is an interpretable model, maybe you can say that it's a faster to learn model compared to the others again, not really interpretable

71.15.1.6.3 Gap:

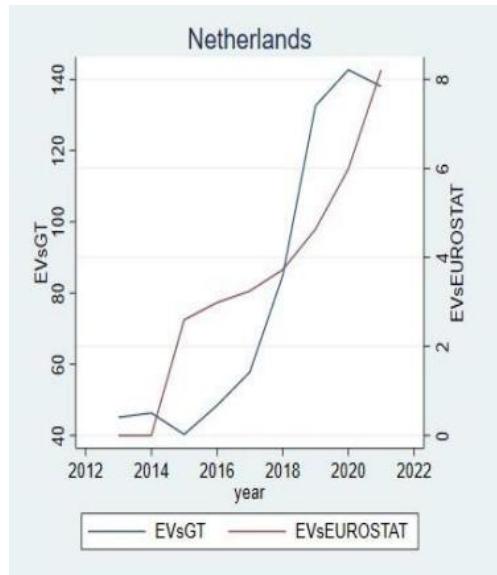
this whole section feels repetitive to the previous one, consider skipping

71.16 12/05

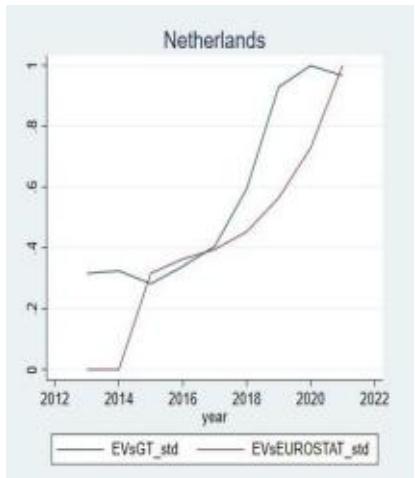
Werk me dataverwerking en feature engineering sectie uit met uitleg over o.a. Google Trends, brandstofprijzen, lags.

71.16.1 Oplossing keyword search?

Ik heb contact opgenomen met Fulvio Castellacci met het verzoek om de lijst met gebruikte zoekwoorden uit zijn onderzoek te delen. Hij heeft hierop positief gereageerd en de volledige lijst toegestuurd via de online appendix van het gepubliceerde artikel. In zijn toelichting gaf hij aan dat de samengestelde EV-indicator is berekend als het gemiddelde van de Google Trends-scores van alle opgenomen zoekwoorden (zie sectie 3.2 van het artikel). Deze lijst biedt waardevolle input voor mijn eigen keywordselectieproces. → vraag zijn die keywords geordend op iets?



Figuur 1: Google Trends indicator of EVs adoption (left-axis) and Eurostat indicator



Figuur 2: Google Trends indicator of EVs adoption and Eurostat indicator on the share

71.16.1.1 *Bron voor jaarlijkse bijtelling:*

De onderstaande bron geeft een overzicht van bijtellingspercentages voor volledig elektrische auto's (0 gram CO₂-uitstoot) per jaar:

<https://100-salaris.nl/bijtellingspercentages-voor-privegebruik-auto-2011-2026/>

Hoewel de informatie overzichtelijk is weergegeven, is de betrouwbaarheid en herkomst van de data momenteel niet voldoende onderbouwd voor gebruik in mijn thesis. Ik ga onderzoeken of de gegevens aansluiten bij officiële bronnen (zoals de Belastingdienst) of dat ik een officiële bron kan gebruiken ter vervanging.

Jaar	Bijtellingspercentage	Max. verlaagde bijtelling tot cataloguswaarde
2019	4%	€50.000
2020	8%	€45.000
2021	12%	€40.000
2022	16%	€35.000
2023	16%	€30.000
2024	16%	€30.000
2025	17%	€30.000
2026	22%	Geen drempel – volledige cataloguswaarde

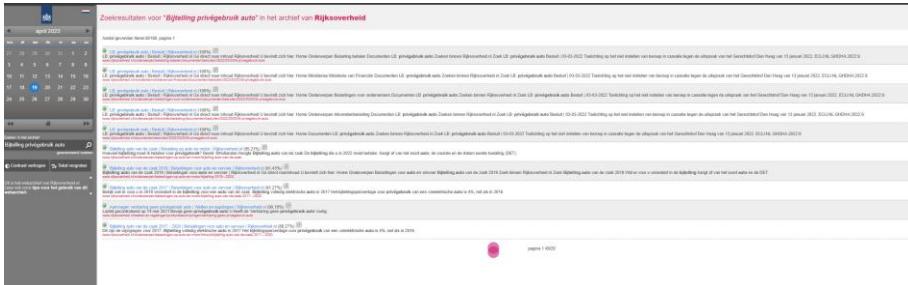
Let op: Bijtelling geldt voor een periode van 60 maanden vanaf de eerste tenaamstelling. Na deze periode zijn de op dat moment geldende percentages van toepassing.

Tot slot: bij het googelen van termen als “*Bijtelling privégebruik auto*” verschijnen in veel gevallen (ruim 20 keer geteld) cijfers afkomstig van de Belastingdienst zelf. Dit bevestigt dat deze officiële gegevens breed worden geciteerd, en biedt mogelijk aanknopingspunten voor het vinden van een directer te citeren primaire bron.



Tabel_bijtelling_201
1-2026.pdf

Dann zie nu ook <https://rijksoverheid.sitearchief.nl/#archive> kijken of ik hier iets kan vinden over bijtelling eerste indruk is dat het moeilijk is om te zoeken:



71.16.1.2 Reden waarom ik alleen bijtelling ga gebruiken en geen subsidie zoals SEPP

De Subsidie Elektrische Personenauto's Particulieren (SEPP) gold voor nieuwe en gebruikte elektrische personenauto's. De regeling liep van 2020 tot 2025 en is definitief **gesloten**. Bijtelling volledig elektrische auto in 2017

Het bijtellingspercentage voor privégebruik van een volelektrische auto is 4%, net als in 2016. Dit blijft zo tot en met 2020. Het kabinet wil zo volledig elektrisch rijden stimuleren. En draagt zo bij aan een beter klimaat en milieu.

- U kunt door de site navigeren op een bepaalde dag in het verleden, door op de kalender (in de grijze balk links op het scherm) een datum te selecteren. U kunt teruggaan tot 15 december 2016.

446.114 nieuwe personenauto's in 2019

[Uit de officiële cijfers van BOVAG, RAI Vereniging en RDC blijkt dat er in 2019 in totaal 446.114 nieuwe personenauto's zijn geregistreerd. Dat is 0,6 procent meer dan het jaar ervoor. In december werden grote aantallen, vooral elektrische, auto's afgeleverd voor de zakelijke markt, om zodoende nog te kunnen profiteren van de lage bijtelling van 2019.](#)

[De bijtelling voor privégebruik van zakelijke elektrische auto's is per 1 januari verhoogd van 4% naar 8%, voor zover het de eerste 45.000 euro van de catalogusprijs betreft. Over het bedrag boven die grens is 22% bijtelling verschuldigd, zoals ook wordt berekend over de gehele waarde van zakelijke auto's op fossiele brandstoffen. Voor de elektrische auto's die vóór 1 januari werden geregistreerd geldt nog een bijtelling van 4% over de eerste 50.000 euro. De elektrische auto die veruit het meest geregistreerd werd in 2019 was dan ook de Tesla Model 3, met 29.922 stuks.](#)

[Meer dan 50.000 elektrisch](#)

[De exacte cijfers per brandstofsoort zijn nog niet bekend, maar volgens BOVAG en RAI Vereniging zijn er in 2019 ruim meer dan 50.000 nieuwe elektrische auto's geregistreerd. Voor 2020 gaan de brancheorganisaties voor de gehele markt vooralsnog uit van een verkoopniveau van 425.000 personenauto's, waarvan minimaal 10 procent elektrisch. Door de verhoging van de bijtelling per 1 januari zijn veel elektrische auto's nog in 2019 op kenteken gezet. Het effect hiervan is dat de eerste maanden van 2020 lagere aantallen zullen laten zien. Komend jaar komen veel fabrikanten met een keur aan elektrische modellen op de markt. Aangezien per 1 januari 2021](#)

het bijtellingstarief voor e-auto's wederom wordt verhoogd (van 8 naar 12 procent), kan er tegen het einde van het jaar weer een run ontstaan, zoals afgelopen december.

<https://www.bovag.nl/nieuws/446114-nieuwe-personenautos-in-2019>

71.17 11/05

71.17.1.1 Vragen voor mezelf:

Als ik zo een focus leg op overheid en zero emitions, moet ik dan niet eigenlijk een doorberekening doen op 20250? → nee je zou in discussie kunnen zetten

Rond literatuurstudie af: samenvatting van eerdere papers, positionering van je eigen bijdrage.

Ochtend (15-16)

- Literatuurstudie afronden:

- Korte samenvatting per kernpaper (Wijnhoven, Afandizadeh, Zhang, Castellacci, enz.)
 - Benoem verschil in scope, methoden, tijdsperiode
 - Schrijf je bijdrage als afsluitende alinea van dit hoofdstuk

Middag (16:30-18)

- Finaliseer methodologie:
 - Leg ‘walk-forward’ nog 1x kort uit
 - Benoem waarom je MAE kiest
 - Noem kort waarom CNN erbij zit naast XGBoost
- Check scope + ethische overwegingen (kort)
 - Benoem bewust gebruik van publieke data
 - Zeg dat het model transparant en reproduceerbaar is

Eind van de dag (18:00-20:00)

- Laatste spellingscheck (Grammarly / Overleaf)
- PDF genereren via Overleaf
- **Stuur draft naar je docent met korte begeleidende e-mail**
- Benoem verschil in scope, methoden, tijdsperiode

71.17.1.2 Existing Work – Outline and Draft Starters

Samenvattingen Kernpaper:

71.17.1.3 Sentiment Analysis and Google Trends Data for Predicting Car Sales (Wijnhoven & Plant, n.d.)

In contrast to previous studies that primarily focus on short-term sales forecasting, model-specific interest, or sentiment-based indicators, this thesis offers a broader and more policy-relevant approach. While research such as Wijnhoven & Plant (2017) and Afandizadeh et al. (2023) focus on brand-level queries or deep neural networks applied to shorter periods, this study analyzes national-level EV sales in the Netherlands from 2007 to 2025. It integrates general behavioral indicators (e.g., Google Trends data on “elektrisch rijden” and “laadpaal”), macroeconomic signals (fuel prices, consumer confidence), and policy events (e.g., subsidy announcements, changes in bijtelling) into a unified forecasting framework. By applying nonlinear models like XGBoost and CNN, combined with walk-forward validation and rescaling of Google Trends data, this thesis adds methodological rigor and real-world applicability. It bridges a gap in the literature by aligning behavioral data with policy evaluation, offering valuable insights for forecasting electric vehicle market trends beyond brand-specific analysis.

Objective:

To compare the predictive power of social media sentiment and Google Trends data for forecasting car sales in the Netherlands, focusing on 11 car models over 52 months (2012–2016).

Key Concepts & Theoretical Basis:

- AIDA Model:
 - Attention → Social media mentions

- Interest → Google search behavior (Google Trends)
- Desire → Sentiment (positive/negative ratio, PNR)
- Action → Actual sales
- Theory of Planned Behavior and Information Acceptance Model (IACM) also underpin the assumed link between social information and purchase intention.

Methodology:

- Data Sources:
 - 500,000 social media posts via Coosto (Twitter, Facebook, etc.)
 - Google Trends data for search interest
 - Monthly car sales data from BOVAG (Netherlands)
- Sentiment Measure:
 - PNR = Positive posts / Negative posts
- Analysis:
 - Linear regressions and decision tree regression (M5P in WEKA)
 - Time lag testing (cross-correlation with up to 12-month delay)

Results:

- Sentiment:
 - Weak predictor of car sales; only significant for VW Golf, and even then with lag 0.
- Google Trends:
 - Moderate to strong predictor in multiple cases (avg. correlation ≈ 0.345 with ~6.7 months lag).
- Total mentions:
 - Strong predictor for some models (correlation ≈ 0.804 , $R^2 \approx 0.606$).
- Prediction model:
 - A decision tree using car model, mentions, and Trends score (with 4-month lag) achieved $R = 0.73$.

Conclusions:

- PNR (sentiment) was not a reliable predictor.
- Search interest (Google Trends) had higher and more consistent predictive power.
- Modeling lags is essential — peak search and mention activity precede sales by about 4–7 months.
- Policy effects (e.g., tax changes) likely have stronger influence than sentiment.

Relevance to Your Thesis:

Supports your decision to use Google Trends over sentiment analysis, especially for high-involvement goods like EVs.

Highlights the importance of time lags, especially around 4–6 months — aligning with your own walk-forward validation setup.

Validates keyword-based Trends selection for vehicle forecasting, although they used model-specific queries (e.g., “Volkswagen Passat”) rather than general EV terms — that’s a clear point of distinction for your study.

You extend this work by using nonlinear models (XGBoost, CNN), a broader range of macroeconomic and behavioral features, and a longer time window (2007–2025).

71.17.1.4 Does digitalisation affect the adoption of electric vehicles? New regional-level evidence from Google Trends data (Castellacci & Santoalha, 2025)

Methodology:

- Data source: Google Trends (monthly search data)
- EV adoption signal: 116 keywords selected from 9 years of IEA EV Outlook reports (e.g., “electric vehicle,” “EV battery”).
- Digitalisation indicators: Keywords for five domains (smart manufacturing, smart cities, smart appliances, online communication, social media).
- Composite scores: Built by averaging normalized Trends scores (0–100) per region.
- Validation: Google Trends EV score was compared to Eurostat EV market share data in 12 countries and showed strong alignment.
- Handling missing data: Regions with <50% completeness were excluded.

Key Findings:

- Regions with higher digitalisation (measured via Trends) also showed greater interest in EVs.
- The Google Trends-based EV adoption indicator is shown to be valid and consistent with actual adoption trends.

Relevance to Your Thesis:

- Strong support for using Google Trends as a behavioral proxy for EV interest.
- Their keyword selection method (based on expert sources like IEA reports) gives you a defensible strategy for building your own search terms.
- You extend this work by:
 - Focusing on forecasting actual EV sales over time (not just interest or adoption).
 - Including macroeconomic and policy variables.
 - Applying machine learning models (XGBoost, CNN) rather than regression alone.

- Working with longer timeframes (2007–2025) and monthly granularity.

71.17.1.5 Using machine learning methods to predict electric vehicles penetration in the automotive market(Afandizadeh et al., 2023)

Mooie zinnen:

Electric cars have been introduced as an alternative to gasoline and diesel cars to reduce greenhouse gas emissions, optimize fossil fuel use, and protect the environment. Predicting EV sales is momentous for stakeholders, including car manufacturers policy makers and **fuel suppliers**

According to environmental protection agency 27% van greenhouse gas emissions are from transportation sector → zelfde toevoegen voor NL

Objective of the Study

- Predict monthly EV sales and market share using advanced deep learning models.
- Identify the most influential factors on EV sales using hybrid AI models.

Existing work:

Numerous studies have applied machine learning (ML) techniques to forecast vehicle sales using time-series and economic data. A variety of models—ranging from regression-based methods to advanced neural networks—have been tested across different regions and data granularities.

Brühl et al. (2012) compared Multiple Linear Regression and Support Vector Machines (SVM) using economic indicators such as GDP, personal income, and petroleum charges, across yearly, quarterly, and monthly data. They found that SVM outperformed regression models in terms of MAE and MAPE, particularly on quarterly data due to better stability and interpretability.

Wang et al. implemented Adaptive Network-Based Fuzzy Inference Systems (ANFIS), Artificial Neural Networks (ANN), and ARIMA models to forecast sales using macroeconomic factors, vehicle and component prices, and exchange rates. ANFIS achieved the highest accuracy (R^2 and MSE), demonstrating the advantage of hybrid ML techniques.

Hülsmann et al. (2012) analyzed German and US sales data using both linear models (OLS, quantile regression) and nonlinear ones (SVM, decision trees, random forests, k-NN). Decision trees showed the best monthly performance in terms of MAPE, with interpretability being a key strength.

Kitapci et al. examined macro-policy effects in Turkey using both regression and neural networks. Variables included tax incentives, loan rates, euro exchange rates, and advertisement spending. Neural networks outperformed regression models, and tax-related factors proved particularly influential.

Bas et al. explored EV adoption prediction using ride-sourcing behavior and sociodemographics. With classification algorithms and LIME for interpretability, they found that ride frequency, EV knowledge, and environmental awareness were key drivers of EV adoption.

Zhang et al. tested Singular Spectrum Analysis (SSA) and Vector Auto-Regression (VAR) on Chinese EV sales. VAR outperformed SSA due to its ability to integrate macroeconomic indicators (fuel prices, confidence indices) and behavioral data (Baidu searches).

Kaya and Yildirim developed a deep neural network to predict monthly vehicle sales in Turkey based on GDP, CPI, CCI, and exchange rate data. The DNN achieved high accuracy, highlighting the importance of economic indicators and time features.

Xia et al. introduced the ForeXGBoost model, applying XGBoost to a large-scale dataset with brand/model data. Through feature selection and optimized time-window extraction, their model significantly outperformed traditional regressors in both speed and accuracy.

Lastly, Saxena et al. evaluated deep learning methods like LSTM and ARIMA for green vehicle sales forecasting. LSTM delivered the lowest MAE and RMSE, reinforcing its suitability for time-series forecasting with non-linear dynamics.

Summary: Non-ML Approaches to Understanding Factors Influencing EV Sales

Several studies have explored factors that influence electric vehicle (EV) sales using alternative methods such as surveys, choice modeling, and behavioral analysis—particularly to support policy development by better understanding user preferences.

Survey-based studies:

- Beggs et al. and Calfee et al. used consumer surveys to estimate potential EV demand based on preferences for technical specifications (e.g., seat capacity, price, performance). They found that poor performance limits adoption but demand increases under improved tech or fuel scarcity scenarios.
- These approaches have limitations: they predict market share as reflected in the survey, not in real-world sales.

Social influence and consumer perception:

- Mau et al. introduced the concept of the “neighbor effect”, where awareness of local EV adoption increases personal likelihood to purchase.
- Kinski et al. found that Google Trends search interest correlates positively with car sales, offering a real-time behavioral signal of market interest.

Vehicle attributes and decision-making:

- Balducci et al. showed that fuel economy and emissions reduction are strong motivators for plug-in hybrid EVs, while high cost and low reliability deter buyers.
- Hess et al. used discrete choice models to show that purchase incentives, acceleration, range, and fuel availability positively influence purchase intent, while cost and maintenance negatively affect it.
- Shafiei et al. confirmed that combining tax incentives, lower EV prices, and high fuel prices significantly boosts EV market share.

Technology and energy cost trade-offs:

- Bas et al. demonstrated that new fuel-saving technologies (even in gasoline cars) are appealing to consumers, and that EVs still struggle with cost–benefit perception, despite environmental advantages.

Relevance to This Thesis

These studies identify critical factors influencing EV adoption, such as:

- Vehicle specs (range, acceleration, fuel cost)
- Policy instruments (subsidies, tax breaks)
- Public perception (neighbor effect, environmental concern)
- Behavioral signals (Google search interest)

While most studies use static methods (surveys, choice models), your thesis adds value by:

- Using Google Trends to quantify “interest” in the AIDA model
- Integrating economic and behavioral indicators in a dynamic, time-series forecasting model
- Focusing on real-time prediction instead of stated preference modeling
- Applying walk-forward validation with lags to simulate decision-relevant forecasts

This enables your research to bridge the gap between policy design and behavioral response forecasting—something traditional survey-based approaches cannot do dynamically.

Methodology & Data Summary – Afandizadeh et al. (2023)

Afandizadeh et al. developed a hybrid deep learning framework to predict electric vehicle (EV) sales in the U.S. market using data from 2014 to 2020. The research compared three models:

1. LSTM – captures long-term dependencies in time series.
2. ConvLSTM – improves temporal correlation handling by using convolutional operations on sequences.
3. Hybrid LSTM with 2D Attention and Residual Network – introduces an attention mechanism to weigh feature importance across a 7-month window, and residual connections to preserve learned patterns from earlier layers.

The models were evaluated using rolling cross-validation and metrics like MAPE, NRSME, and R². The hybrid model outperformed the other two, achieving an average MAPE of ~3.5% for EV share predictions.

Data Collection

- Primary dataset: Monthly sales data for 357 new car models from 2014–2020.
- Web crawlers were used to collect:
 - Vehicle specifications (e.g., MPG, engine power, price).
 - Sentiment scores from car-related news articles (via VADER sentiment analysis).

- Google Trends search scores for specific brand + model queries.
- Macroeconomic data: GDP, CPI, fuel prices, stock indices, loan interest rates.
- Input was structured using a 7-month sliding window, and month-of-year dummy variables were used for seasonality.

Relevance and Gap Compared to Your Thesis

What's similar:

- Both studies use machine learning for EV sales forecasting.
- Both combine behavioral signals (Google Trends) with macroeconomic indicators.
- Both apply lag structures to model temporal effects.
- You also use nonlinear models (e.g., XGBoost, CNN) and aim for walk-forward validation.

What's different / your added value:

- You focus on national-level EV sales, not brand-level disaggregation.
- Your time range is longer: 2007–2025, which may better capture long-term policy effects (e.g., bijtelling changes, subsidies).
- You explore general behavioral intent (e.g., search terms like “elektrisch rijden”) rather than only brand-model-specific queries. This may capture early-stage public interest, though it risks lower precision.
- You exclude sentiment scores, but highlight policy signals like subsidy announcements and aim to link sales spikes to them.
- Their dataset is U.S.-based and heavily dependent on scraped product specs and model-specific trends, whereas yours aims to integrate Dutch government data and climate policy context.

Key Findings

- The hybrid model outperforms LSTM and ConvLSTM (MAE ~3.5% on average).
- Top influential factors vary per car, but CPI, MPG, price, and Google search volume were often important.
- The model performed best in predicting EV *share* within each market segment.

Critical Notes

- Uses brand- and model-specific data and keywords (e.g., "Tesla Model S"), not general EV terms.
- Google Trends is applied per model, with 3 different search phrase formats.
- Sentiment from news is aggregated by vehicle model using VADER, not embedding-based matching.
- No focus on policy signals like bijtelling or Dutch tax structure.

Stick with XGBoost and CNN as your core methods, since they balance accuracy and interpretability

Optionally test a baseline LSTM with simple architecture to show you've considered temporal deep learning — then explain why it was less suitable in your context (e.g., interpretability, dataset size, training stability)

71.17.1.6 Google Trends as a Complementary Tool for New Car Sales Forecasting: A CrossCountry Comparison along the Customer Journey.(A. Google Trends as Complementary Tool for New Car Sales Forecasting A Cross-Country Comparison along the Customer Journey, n.d.)

Data sources:

- Monthly car sales data (from *Automobilwoche Data Centre*) for Germany and the U.S. for 24 car models over a 39-month period (2013–2016).
- Google Trends data was downloaded using exact car model names (e.g., “Audi Q7”) as keywords. Filters were applied for country, time window, and category (“Automotive & Vehicles”) to improve signal quality.

Preprocessing:

- The Google Trends index was averaged from two extraction days to reduce sampling noise.
- Noise from second-hand searches was *unsuccessfully* filtered out (attempts to exclude “used” or “second hand” via keyword modifiers failed due to Google’s minimum search volume threshold).

Lag analysis:

- A cross-correlation function (CCF) was applied to determine the optimal time lag between search interest and actual sales. The lags varied:
 - On average: 2.96 months (low-priced cars), 3.41 months (high-priced).
 - For Audi Q7: 6-month lag gave the highest explanatory power.
- After applying the lag, R^2 improved significantly (e.g., from ~22% to 68.5% for Audi Q7).

Modeling approach:

- Linear regression models were used, comparing:
 - Raw vs. lag-adjusted data.
 - U.S. vs. German market.
- Significance was tested with 95% confidence, and prediction quality was assessed via R^2 and MAE.

Reliability

Kinski explains that reliability refers to the *consistency* of measurement — whether the use of a particular method produces stable results across repetitions. He references Choi & Varian (2009), who advocate for publishing reproducible methods, and Baker & Fradkin (2013), who

improved reliability by averaging data over four different weeks to reduce internet noise. Kinski applies a Cronbach's alpha test on 30 time series variables (car models), which yielded a value of 0.79 — above the accepted threshold of 0.70, thus confirming high internal consistency.

Validity

For validity, Kinski distinguishes between:

- Face validity: ensuring keywords seem logically related to what is being measured.
- Content validity: checking whether all relevant dimensions of the concept (e.g., car interest) are captured.
- Criterion-related validity: linking trends to actual outcomes (like sales).
He supports this by following Stephens-Davidowitz & Varian's (2014) guidelines on category selection (e.g., restricting to "Automobile and Vehicles") and geographical filtering to avoid noise. He also highlights the importance of keyword selection, which Zhu et al. (2012) deemed essential for meaningful analysis.

Summary and Relevance to Your Thesis

Kinski's approach shows that Google Trends can be reliable and valid if:

- Keywords are carefully selected.
- Categories and countries are restricted to reduce noise.
- Data smoothing (averaging over multiple queries) is considered.
- Internal consistency is statistically tested.

For your thesis, this justifies your use of *Google Trends* if you transparently describe:

1. Your keyword selection process.
2. Your category and country filters.
3. Any averaging or rescaling method to account for weekly volatility.
4. Optional: perform a Cronbach's alpha test or similar reliability assessment.

What Kinski did:

- Used brand- and model-specific keywords, short time span (2013–2016), basic linear regression, and did not include macroeconomic factors or broader behavioral signals.

What you are doing differently:

- Focus on general EV-related terms (e.g., "laadpaal," "elektrisch rijden"), not brand-specific.
- Longer time horizon (2007–2025), including policy variables, macroeconomic indicators (fuel prices, consumer confidence), and walk-forward validation.
- You use nonlinear models (e.g., XGBoost, CNN), offering a more flexible and modern approach to modeling.

71.17.1.7 *Using forum and search data for sales prediction of high-involvement projects (Tel Aviv University et al., 2017)*

Title: Using Forum and Search Data for Sales Prediction of High-Involvement Projects

Context: Predicting car sales using a combination of Google search trend data and forum data.

Main Research Questions

1. Can combining search trends and forum data improve car sales predictions?
2. Is search data sufficient to outperform social media (forum) data?
3. Does this vary between premium and value brands?

Methodology & Data

- Industry: U.S. automotive market.
- Time period: Monthly data from 2007–2010.
- Brands: 23 car brands, sales volume above 5,000/month.
- Data sources:
 - Sales data: Automotive News.
 - Forum data: Google's index of online discussion forums.
 - Includes volume and sentiment (dictionary-based).
 - Search data: Google Trends (brand-level keywords).
 - Control variables: Seasonality (sales lag 12), gasoline prices, consumer sentiment.

Models:

- Benchmark model (sales history, macro indicators).
- Forum-only, search-only, and combined models.
- Evaluated with Linear Regression and validated using rolling window forecasting (out-of-sample MAPE).

Key Findings

- Combined model (search + forum) outperforms all others in prediction accuracy.
- Search-only models perform similarly to forum-based models, despite being cheaper and less biased.
- Value brands benefit more from adding search trends than premium brands.
- Biases in forum data (manipulation, visibility, self-presentation) are partly mitigated by search data, which reflects private intent.

Relevance to Your Thesis

- Confirms that Google Trends is a powerful and *less biased* proxy for consumer interest—especially important since you dropped sentiment/news analysis.
- Reinforces that search logs reflect latent intent, aligning with your use of Trends for behavioral signal modeling in the Dutch EV market.
- Their use of brand-level keywords (not models) and normalization of Trends data parallels your approach.
- Your thesis extends this by including policy signals, longer timeframes (2007–2025), and macroeconomic variables like fuel price and consumer confidence.

Gap Your Thesis Covers

- While Geva et al. used U.S. premium vs. value brand forecasting, you:
 - Focus on national EV-level sales (not brand-specific).
 - Include policy shifts like subsidies and “bijtelling”.
 - Extend the temporal window to track long-term behavioral shifts and policy impact.
 - Compare XGBoost and CNN, offering nonlinear interpretability beyond linear regression

71.18 10/05

Schrijf een eerste ruwe versie van je inleiding met duidelijke probleemstelling, relevantie en onderzoeks vragen.

- Hesisstructuur opgesteld en vertaald naar LaTeX (inclusief hoofdstukken voor Introduction, Methodology, Data, Models, etc.).
- Methodesectie herschreven met onderbouwing van modelkeuze (XGBoost vs. CNN) en evaluatiestrategie (walk-forward, MAE/MAPE).
- Keuze voor Google Trends boven NOS-nieuwssentiment onderbouwd aan de hand van feedback en reproduceerbaarheid.
- Lagstructuren onderzocht op basis van literatuur en toegepast op gedragssignalen (1–2 maanden) en beleidsreacties (0–1 maand).
- Titelopties en definitie van gedragsfactoren uitgewerkt.

Leeruitkomsten (gericht op afstudeereisen):

- A1: Probleem in context geplaatst en scope aangescherpt.
- A2: Oplossingsrichting via AI-methoden verder uitgewerkt.
- A5: Reflectie op geschiktheid van AI-modellen en waarom gedragssignalen relevant zijn.
- C1: Maatschappelijke impact besproken via beleidsmomenten.
- D1: Feedback uit eerdere sessies verwerkt in herpositionering van het onderzoek.

71.18.1.1 *Vragen die ik niet kon beantwoorden tijdens het schrijven:*

- Wie wordt target audience?
 - o als ik Policy makes doe, moet ik meer focussen op Klimaat en wss subsidie
 - Ook is het lastiger om model te vergelijken met hun model
 - o Voor autofabrikanten moet ik minder focussen op klimaat, misschien zelfs helemaal weg gaan

Wat zou goede tussenweg zijn?

De redenen voor electrische auto

Weakness:

Veel studies gebruiken daadwerkelijke autotermen omdat mensen die opzoeken als ze auto kopen, ze zoeken geen specifieke modellen op:

Oplosisng:

Acknowledging this trade-off:

"While search behavior around specific vehicle models may more directly reflect imminent purchasing intent, this thesis focuses on general EV-related terms to capture broader public interest and early-stage engagement with electric mobility."

Framing your work as market climate modeling, not buyer modeling.

71.18.1.2 *Target audience:*

Rijksoverheid: dashboard mobiliteit personen en auto's

https://www.pbl.nl/modellen?f%5B0%5D=publication_topic%3A12

info over dashboard:

<https://dashboardklimaatbeleid.nl/content/overditdashboard>

info over modellen:

https://www.pbl.nl/modellen?f%5B0%5D=publication_topic%3A12

misschien toch wel gewoon car manufactures doen en zeggen dat zij uiteindelijk zelf hun eigen automerken kunnen toevoegen.

71.18.1.3 *Skelet van mijn rapport: (nu nog van visser * A leerdoelen toevoegen)*

1. Introduction

- a. Background
- b. Problem
- c. Existing work
- d. Gap
- e. Scope of the research

2. Requirements

- a. Functional requirements
- b. Ethical requirements

3. Dataset

- a. Social media platform
- b. Approach
- c. Dataset 1 – sentiment (X)
- d. Dataset 2 – factual or opinion (FEVER)
- e. Dataset 3 – stock price (Yahoo Finance)

4. Models

- a. Methodology
- b. Dataset 1 (X) preprocessing

- i. Distribution and splitting the data
 - ii. Duplicates and missing values
 - iii. Text cleaning
- c. Model 1: Tweet sentiment classifier
- d. Dataset 2 (FEVER) preprocessing
- i. Removing duplicates
 - ii. Handling missing values
 - iii. Filtering labels
 - iv. Balancing the dataset
 - v. Label mapping for binary classification
 - vi. Shuffling the data
 - vii. Splitting the data
 - viii. Training and validation split
- e. Model 2: Factual or opinion classifier
- i. Null model
 - ii. Baseline Model
 - iii. Rule-based component
 - iv. Motivation for model selection
 - v. Evaluation Metrics
 - vi. Architecture
5. Hardware
- a. Qualitative analysis of FEVER dataset applicability to tweets
6. Results
- a. Models
 - i. Qualitative analysis of labeled fact / opinion
 - b. Putting it all together & exploring the Correlation
 - i. Quantitative analysis of correlation
 - ii. TSLA Stock Price vs. Sentiment
 - iii. TSLA stock price vs. sentiment with 21-day moving average (MA)
 - iv. TSLA stock price vs. sentiment with 21-day MA filtered on factual tweets
 - c. Validating comparability and compatibility

i. Motivation

ii. Results

iii. Analysis of results

d. Conclusion

e. Discussion

i. Limitations

ii. Future Work

Dataset/bron gevonden over internettocioegeng en internet activiteiten in nederland:

71.18.1.4 *Internetgebruik en online activiteiten (CBS-tabel, 2024)*

(Centraal Bureau voor de Statistiek, 2024)

Datum ingevoerd: 10 mei 2025

Bron: Centraal Bureau voor de Statistiek (CBS), *Internet; toegang, gebruik en faciliteiten*

Omschrijving:

Vandaag heb ik de meest recente CBS-tabel over internetaankopen en internetgebruik (gepubliceerd op 3 september 2024) geraadpleegd. Deze tabel bevat definitieve cijfers vanaf 2020 en is uitgesplitst naar persoonskenmerken zoals geslacht, leeftijd, opleidingsniveau, inkomensgroep en stedelijkheid. De gegevens zijn relevant als context voor gedragsmatige indicatoren zoals zoek- en surfgedrag in relatie tot thema's als digitale koopintentie en publieke interesse in elektrische voertuigen.

De tabel omvat:

- Internettoegang thuis en frequentie van gebruik
- Activiteiten in de afgelopen 3 maanden, onderverdeeld in:
 - Communicatie (e-mail, sociale netwerken, WhatsApp)
 - Informatie en vermaak (nieuws, gezondheidsinformatie, streaming)
 - Dienstverlening (bankieren, verkopen, overheidsinformatie)

Deze informatie kan ondersteunend zijn bij het interpreteren van zoektrends (bijv. via Google Trends) en bij het afleiden van gedragsindicatoren in modellen die consumentengedrag of productinteresse voorspellen.

Belangrijk detail: Cijfers zijn afgerekend op 1 decimaal en jaarlijks beschikbaar in november.

<https://opendata.cbs.nl/StatLine/#/CBS/nl/dataset/84888NED/table?ts=1746869151665>

71.19 09/05

Werk Daphne's feedback uit in je methodesectie en update de uitleg over lagstructuur + walk-forward validatie.

Na gesprek met Daphne nog gehad om geen pieken weg te halen, evt wel om MAE voor de piek te berekenen nos. Ook bespreken we het opdelen van

Voorbereiding overleg Daphne:

Unieke bijdrage t.o.v. bestaande literatuur:

Mijn onderzoek richt zich uitsluitend op elektrische auto's (EV's) binnen de Nederlandse markt, terwijl veel bestaande studies zoals die van Wijnhoven & Plant (Wijnhoven & Plant, 2017), (Afandizadeh et al., 2023) en (Sayedi, 2020) brede autosegmenten en landen vergelijken. In tegenstelling tot hun gebruik van lineaire modellen of regressies, zal ik nonlineaire modellen vergelijken zoals xgboost en cnn om met als doel de complexere patronen te herkennen. Ik gebruik een lange tijdsreeks van 2007–2025, wat afwijkt van veel studies die zich op kortere perioden richten (zoals 2014–2020 bij Afandizadeh et al. Of in de buurt van 2011–2015 bij Zhang).

Een belangrijk verschil is dat ik mij niet richt op individuele automerken of -modellen, zoals veel van de bovenstaande papers wel doen, maar werk met algemene gedragsindicatoren zoals zoekinteresse in "elektrisch rijden" of "laadpaal". Vanwege de brede scope in data voeg ik ook verklarende variabelen toe zoals brandstofprijzen, consumentenvertrouwen en economisch klimaat, die vaak ontbreken in bestaande studies. Deze variabelen sluiten aan bij verschillende fasen van het Attention, Interest, Desire, Action (AIDA)-model: zo weerspiegelen consumentenvertrouwen, Google Trends-zoekgedrag en het economisch klimaat vooral de interessefase ("Interest"), terwijl brandstofprijzen een directe invloed kunnen hebben op het verlangen naar elektrische voertuigen ("Desire")

Daarnaast wil ik een mogelijk probleem van Google Trends normalisatie (0–100 schaling) oplossen door de zoektermen over overlappende segmenten te herschalen, zodat tijdsvergelijking betrouwbaarder wordt.

Ik gebruik een walk-forward validatie methode met lags, zodat mijn voorspellingen alleen gebaseerd zijn op informatie die op dat moment bekend was.

Dit verhoogt de realisme en toepasbaarheid voor beleid. Ten slotte probeer ik de relatie tussen beleid (zoals bijtelling) en verkoop te koppelen aan pieken in de data, iets wat nauwelijks terugkomt in bestaande Google Trends-literatuur.

Tijdens meeting bragen gesteld over de piek momenten, en lagtime

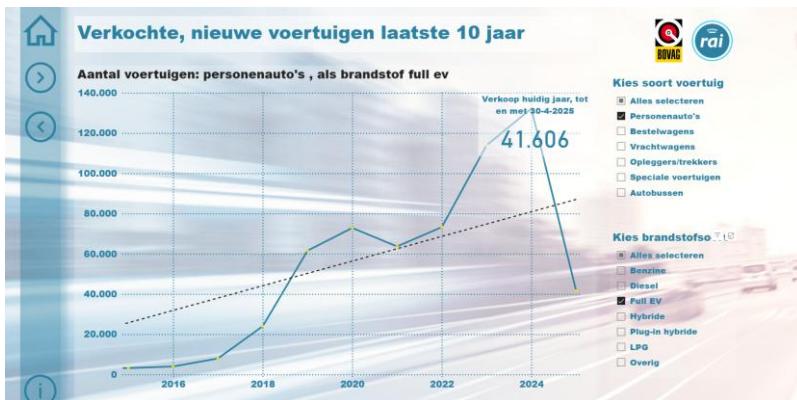
71.20 08/05

71.20.1 Literatuuronderzoek:

71.20.1.1 Vergelijking met Wijnhoven & Plant (2017) – Sentiment & Google Trends voor EV-voorspellingen

Brongegevens:

- Onderzoek: Wijnhoven & Plant (2017) – Sentiment Analysis and Google Trends Data for Predicting Car Sales
- Databronnen gebruikt in artikel: Google Trends per automodel, verkoopcijfers per model (via [BOVAG](#), link inmiddels 404).



- Mijn bron: Google Trends op algemene EV-gerelateerde zoektermen (zoals "elektrische auto"), verkoopcijfers van het [CBS](#) per maand (totaal aantal EV's)



Belangrijkste inzichten uit het artikel

- Modellering: zij gebruiken M5P decision trees met lineaire modellen in de knopen; ik gebruik XGBoost (gradient boosting) met walk-forward validatie.
- Lag-analyse: ze berekenen voor elk automodel de optimale tijdsvertraging (lag) met cross-correlatie (SPSS), tot maximaal 12 maanden.
- Resultaten: MAE varieerde per model, maar exacte gemiddelden zijn niet transparant vermeld. Wel werden bijv. correlaties > 0.5 als significant gepresenteerd.
- Gebruik Google Trends: per automodel; bijvoorbeeld "Volkswagen Passat" of "Ford Fiesta".

Aspect	Wijnhoven & Plant (2017)	Mijn benadering
Modeltype	M5P Decision Trees	XGBoost (met boosting op residuen)/ kan wijzigen naar CNN
Voorspellingsseenheid	Per automodel (11)	Totaal aantal verkochte EV's (geaggregaat)
Google Trends	Specifieke modellen als zoekterm	Algemene zoektermen ("elektrische auto", enz.)
Lag-analyse	Lag bepaald per model m.b.v. crosscorrelatie	(Nog) niet toegepast, dynamische walk-forward
Validatie	66% train / 33% test (klassieke split) 52 maanden	Walk-forward (tijdreeksgevoelig) (80/20) 219 maanden
Multicollineariteit	Getest via VIF (Variance Inflation Factor)	Niet toegepast, maar mogelijk relevant
Tools	Weka (GUI)	Python + scikit-learn, pandas, matplotlib

Suggesties voor uitbreiding

1. Voer cross-correlatie uit tussen Google Trends en verkoop met lags van -12 tot +12 maanden. Bekijk bij welke vertraging de correlatie het hoogst is.
2. Bereken VIF voor je huidige featureset, om te controleren op multicollineariteit. Dit kan met statsmodels.stats.outliers_influence.variance_inflation_factor.
3. Overweeg om consumentenvertrouwen toe te voegen
4. Toets je model op nieuwe EV-subcategorieën zoals plug-in hybrides of EV's per merk (indien data beschikbaar is).

71.20.1.2 Doel: Onderzoeken hoe Google Trends effectief ingezet kan worden bij het analyseren van consumentengedrag en het verbeteren van verkoopvoorspellingen van elektrische voertuigen (EV's).

1. Analyse van bestaande literatuur

- Bestudeer eerder onderzoek waarin Google Trends is gebruikt binnen de context van elektrische voertuigen of consumentengedrag.
- Let hierbij op de gebruikte zoektermen in deze studies. Voorbeelden zijn:
 - *Electric Vehicles*
 - *Alternative Fuel Vehicles*
 - *Plug-in Hybrid Electric Vehicles*
 - *Battery Electric Vehicles*
 - *Green Cars, Eco-cars, Cleaner Vehicles*
 - En gedragsgerelateerde termen zoals *acceptance, consumer adoption, attitude, intention, en behavior* (Zie bijvoorbeeld: [ScienceDirect-artikel](#)).

2. Verrijking van zoektermen

- Gebruik synoniemen en alternatieve bewoordingen uit bronnen zoals het woordenboek of thesauri om de dekking van relevante zoekopdrachten te vergroten.
- Denk aan varianten in spelling, afkortingen, of volksmondtermen.

3. Relevante aanbevelingen en inzichten uit Google Trends

- Raadpleeg praktische toepassingen van zoekgedrag, zoals weergegeven op platforms als [Think with Google](#), om inzicht te krijgen in mobiele zoekpatronen rondom EV's.
- Overweeg hoe zoekvolumedata vertaald kunnen worden naar marketing- en verkoopstrategieën.

4. Stakeholderinput meenemen

- Verzamel input van relevante belanghebbenden, zoals EV-verkopers of marketingexperts, om realistische en praktijkgerichte zoektermen te identificeren.

5. Wetenschappelijke onderbouwing van voorspellingswaarde

- Analyseer studies die de voorspellende waarde van Google Trends-gegevens op productniveau hebben onderzocht, zoals:
 - Fritzsch et al., "Can Google Trends Improve Sales Forecasts on a Product Level?" ([ScienceDirect](#))
 - Publicaties op IEEE Xplore ([IEEE artikel](#))
 - Medisch-psychologische studies over consumentengedrag via PubMed ([PubMed studie](#))

71.21 07/05 Onderzoekplan herzien

Na de het gesprek met Marcio, heb ik mijn onderzoeksplan herzien. De aparte validatieset is verwijderd en vervangen door cross-validation. Zoals verwacht had dit weinig invloed op de prestaties van het model, maar het maakt de methode methodologisch sterker en beter verdedigbaar.

Vervolgens heb ik mijn strategie voor artikel/selectie van sentimentbronnen geëvalueerd. Het werd duidelijk dat het sentiment uit NOS Nieuws geen consistente of voorspellende waarde bood. Ook de selectie van EV gerelateerde artikelen voldeed niet aan de kwaliteitseisen.

Mogelijk heb ik te veel tijd besteed aan het proberen te verbeteren van deze aanpak.

Mijn eerste alternatief was overstappen naar Twitter-sentiment, maar dat zou vergelijkbare problemen opleveren (zoals het correct filteren van irrelevante data en het beoordelen van de kwaliteit van de sentimentanalyse). Daarnaast heb ik tijdens literatuur onderzoek van pva over het hoofd gezien dat sentiment analyse weinig voorspellende waarden hebben voor autoverkoop.

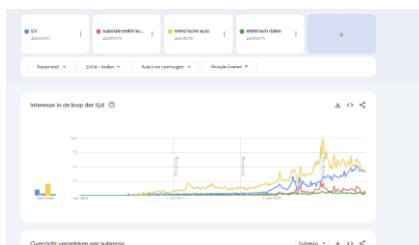
In plaats daarvan heb ik me verder verdiept in onderzoek van Fons Wijnhoven, waarin onder andere wordt geconcludeerd dat:

"Google Trends scores correlated better with car sales and, in many cases, significantly."

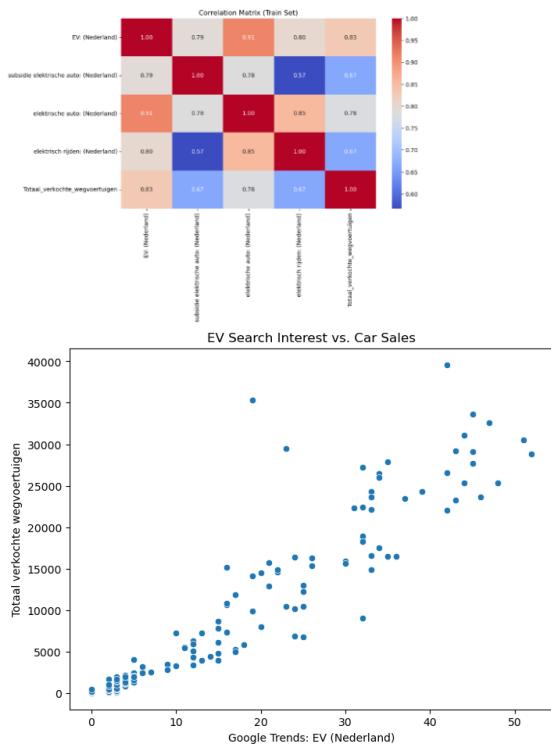
Deze bevinding gaf me het vertrouwen om voort te bouwen op zijn werk, met een specifieke focus op elektrische voertuigen (EV's) en een uitgebreidere dataset waarin het oorspronkelijke onderzoek 53 maanden werden gebruikt, maak ik nu gebruik van 216 maanden (2007–2025), in plaats van de 93 maanden uit mijn eerdere opzet met nieuwssentiment.

Ik voerde een verkennende analyse uit met algemene EV-gerelateerde zoektermen in Google Trends, waaronder:

- "EV"
- "subsidie elektrische auto"
- "elektrische auto"
- "elektrisch rijden"



Deze heb ik gekoppeld aan CBS-data over verkochte EV's. De correlatiematrix liet duidelijker verbanden zien.



Daarna heb ik een XGBoost-model met cross-validation toegepast om het voorspellen van EV-verkopen te testen. Zonder dat er al sprake was van vertraging (lag time) in de features, presteerde het model aanzienlijk beter:

Ook visueel sloot de voorspelde lijn beter aan bij de werkelijke cijfers, op enkele uitschieters na (die ik nog nader ga onderzoeken).

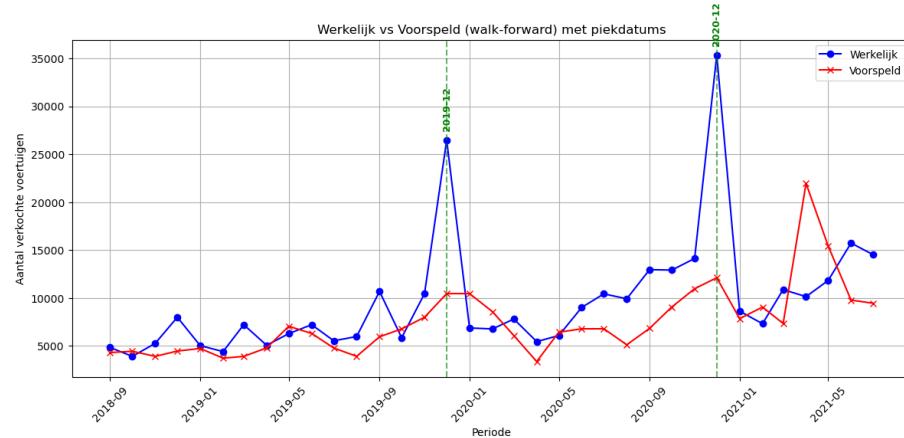
“Tot eind 2019 gold voor volledig elektrische auto's een bijtellingspercentage van 4% over de eerste €50.000 van de cataloguswaarde. Vanaf 1 januari 2020 werd dit percentage verhoogd naar 8%. Om te profiteren van het lagere tarief, besloten veel zakelijke rijders hun elektrische voertuigen nog in december 2019 op kenteken te zetten.” ([Business insider](#))

In 2020 werd de bijtelling voor elektrische auto's opnieuw verhoogd, ditmaal naar 12% vanaf 1 januari 2021. Wederom leidde dit tot een verkoopspiek in december 2020, aangezien zakelijke rijders probeerden te profiteren van het nog geldende lagere tarief. ([NOS](#))

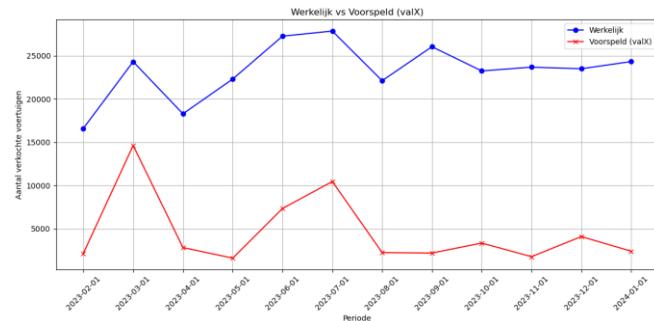
“De grote stijging van het aantal BEV personenauto's in 2020 is voor een groot deel het gevolg van de grote verkoopaantallen in december 2020. Deze stijging in het wagenpark is vergelijkbaar met het stijgende patroon in 2019, die tevens grotendeels veroorzaakt is door de grote verkoopaantallen in december 2019.”

(Bron: RVO, Jaaranalyse 2020 – pagina 6)

“Zoals voorspeld in het jaaroverzicht 2019, wordt de BEV-verkooppieker in december 2019 gevolgd door een veel lagere verkoopomvang in de eerste maanden van 2020.”
(Bron: RVO, Jaaranalyse 2020 – pagina 12)



- MAE (Google Trends): 3.661,21



- MAE (News Sentiment): 18.714,00

Nog openstaande punten:

- Selectiemethode voor zoektermen in Google Trends:
Mogelijke strategie:
 - Clustering van algemene EV-termen
 - Aanvullen met termen uit eerdere literatuur
 - Synoniemen uit woordenboeken of thesauri
 - Google Trends-aanbevelingen

- Stakeholderinput, bijvoorbeeld van EV-verkopers
→ Doel: subjectieve keuzes onderbouwen met objectieve argumenten
- Afstemming van afbakening en bijdrage:
Controleren of de focus op Nederlandse EV-verkoop en een grotere dataset (2007–2025) inhoudelijk sterk genoeg is en voldoende wetenschappelijke waarde toevoegt.

Compared thesis with visser:

Leerdoel	Opmerkingen
A1. Probleem in context + belanghebbenden	De context (social media sentiment vs. markt) is sterk uitgewerkt, maar het perspectief van eindgebruikers of opdrachtgevers ontbreekt grotendeels.
A2. Verkenning oplossingsrichtingen + AI-ontwerp	Er wordt gewerkt met FinBERT, Random Forest, CNN en regelgebaseerde modellen. Alternatieven zijn onderbouwd.
A3. Designkwaliteit (gebruiksvriendelijk, waarde)	De afweging van modelkeuze wordt gedaan (bijv. FastText i.p.v. BERT), maar gebruiksvriendelijkheid of toegevoegde waarde voor stakeholders komt weinig terug.
A4. Prototype testen in context	Er is geen concrete gebruikerstest of toepassing van het prototype in een realistische setting beschreven.
A5. Is AI geschikt? Technisch, maatschappelijk, ethisch	In de ethieksectie en discussie worden privacy, bias en toepasbaarheid besproken.
C1. Evaluieren maatschappelijke impact, dialoog vakgenoten	Er is reflectie op impact (bijv. bots, TSM vs. TSLA), maar geen interactie met vakgenoten of interviews.
D1. Feedback verwerken en ontwikkeling tonen	Niet zichtbaar in het document zelf; mogelijk elders (Logboek, feedbackmomenten) wel aanwezig.

71.21.1 Meeting met Marcio

71.21.2 Belangrijkste feedbackpunten

De thesis voldoet momenteel niet aan de eisen voor de 70%-beoordeling. Drie hoofdproblemen:

- Scope is te breed
- Richting van het onderzoek is niet scherp genoeg
- Probleemstelling mist focus

Besprekking leeruitkomsten A (met input van labbegeleider Kees)

A1: . Je doet onderzoek om het probleem in context in kaart te brengen, herformuleert het vraagstuk, weegt en toetst de belangen van de opdrachtgever, eindgebruiker en andere belanghebbenden. → **Verwacht in rapport. Focus op één doelgroep: verkopers of gemeenteambtenaren.**

A2: . Je verkent en analyseert de oplossingsrichtingen voor een vraagstuk en ontwerpt een AI-oplossing volgens een gangbare ontwerpmethodiek. → **Moet duidelijk uitgewerkt zijn in de thesis.**

A3. Je beschrijft de kwaliteit van het ontwerp door te verwijzen naar design patterns en te kijken naar aspecten als haalbaarheid, gebruiksvriendelijkheid en de toegevoegde waarde die de oplossing kan bieden aan stakeholders.

& A4. Je ontwikkelt een prototype en test deze in de context waarin de oplossing gebruikt gaat worden. → Prototype blijft beperkt; benoem dit expliciet als buiten scope in een discussieparagraaf. Voeg eventueel een schets toe van hoe het prototype eruit zou kunnen zien.

A5: 5. Je onderbouwt voor de opdrachtgever of AI wel of niet geschikt is voor een gegeven vraagstuk, gelet op technische, maatschappelijk en ethische aspecten. → [Onderbouwen waarom AI gebruikt wordt](#). [Waarom geen standaard statistiek?](#)

C1: Maatschappelijke impact & dialoog vakgenoten → Complex onderdeel. Gebruik eventueel hypothetisch scenario.

Analyse deelonderwerpen

1. NOS-artikelselectie (EV-gerelateerd):

- Keyword search is zwak: Niet reproduceerbaar, onduidelijke motivatie voor woordkeuze. Geen literatuurbasis.
- Query prompt is subjectief: Gebaseerd op eigen inzicht, niet onderbouwd.
- Verbeteropties: Artikelen selecteren via scatter plot met hoge semantische overeenkomst.
 - Vraag: Waarop is de scatter plot gebaseerd als query ook subjectief is?

2. XGBoost-voorspelling:

- Model presteert redelijk in mei/juni, maar faalt in juli/februari.
- Belangrijke fout: gebruik van unieke tijdsindex → aanpassen naar maandnummer (01–12).
- Februari heeft structureel lage verkoop. Interpolatie consumentenvertrouwen is acceptabel. Overweeg cross-validatie i.v.m. kleine dataset.

Algemene problemen in het onderzoek

- Hypothese lijkt weerlegd: zwakke correlatie tussen NOS-nieuws en EV-verkoop.
- Belangrijkste zwakke punten: probleem is te breed, dataset te klein, keyword selection werkt niet goed.
- Mogelijke interessante variabelen: advertenties, laadpalen – moeilijk te verkrijgen.

Actiepunten voor deze week

- Probeer Twitter-dataset gericht op EV's.
- Verwijder tijdsvariabele en vervang door maandnummer (01–12).
- Run opnieuw met alleen NOS-data.
- Bereid nieuwe meeting volgende week voor met resultaten.

Toekomstige ideeën en ‘thinking out loud’

- Handmatige subsidizedata toevoegen (gemiddelde per jaar?).
- EV-verkoop uitbreiden tot 2007.
- Combineer sentiment + Google Trends specifiek voor EV's.
- Oudere studie toont zwakke sentimentvoorspelling, maar sterke volume-effecten.
- EV-markt is veranderd → nieuwe studie kan relevant zijn.
- Vergelijk eventueel met niet-EV auto's.

71.22 06/05

Logboeknotitie – Reflectie op voortgang en expertfeedback)

1. Bespreking met Prof. dr. C.A.G.M. van Montfort

Tijdens de Thesis Supervision Call heb ik samen met prof. dr. C.A.G.M. van Montfort de voorspellingen van mijn XGBoost-model besproken. Opvallend hierbij was dat de voorspellingen systematisch onder de daadwerkelijke verkoopcijfers lagen, ondanks dat het model de trendbewegingen (pieken en dalen) redelijk accuraat volgt. Prof. Van Montfort suggereerde dat een verticale verschuiving van de voorspellingslijn al tot betere overlap met de werkelijke waarden zou leiden. Dit wijst mogelijk op een structurele bias in het model.

Daarnaast stelde ik voor om seizoensinvloeden te normaliseren. Prof. Van Montfort ontraadde dit in de huidige fase van het onderzoek. Wel wees hij op het ontbreken van tijdsvariabelen in het model, wat potentieel waardevolle verklarende kracht mist. Hiervoor gaf hij twee mogelijke benaderingen:

- Tijd uitsluiten en corrigeren voor langetermintrends (complex en bewerkelijk).
- Tijd meenemen als verklarende variabele, bijvoorbeeld via:
 - een continue variabele zoals "aantal maanden sinds start";
 - dummyvariabelen per jaar (waarbij 1 jaar als referentie dient om multicollineariteit te vermijden).

Ook stelde hij dat het groeperen van jaren (bijv. "vroeg", "midden", "recent") overwogen kan worden indien er weinig verandering tussen individuele jaren is.

Verder bespraken we de mogelijkheid om een random train/test-split te overwegen i.p.v. een tijdsgebaseerde split, afhankelijk van het onderzoeksdoel.

Tot slot uitte ik mijn zorg over leerdoel A (ontwerp), omdat ik geen externe opdrachtgever heb. Prof. Van Montfort bevestigde dat dit geen probleem is en verwees naar de bestaande projectbeschrijving van M. Visser [1] als voldoende onderbouwing.

Samenvatting referentie [1]:

Visser onderzocht hoe social media sentiment van invloed kan zijn op beursfluctuaties. In zijn aanpak combineerde hij sentimentanalyse met tijdreeksen en experimenteerde hij met featureselectie. Zijn methodiek diende als inspiratiebron voor mijn LOFO-analyse.

2. Verbeteringen in de code en aanpak (met dr. Daphne)

Later die dag heb ik mijn code en aanpak besproken met dr. Daphne. Daarin kwamen we een aantal punten tegen:

- Herziening embedding- en keywordfiltering:
Ik heb eerst opnieuw een keywordfilter toegepast op NOS-artikelen, waarna ik Jina V3 embeddings heb berekend en similarityscores heb geanalyseerd. De histogram van similarityscores liet zien dat de meeste artikelen tussen 0.15 en 0.35 scoren. Artikelen met lage scores bleken inderdaad nauwelijks over elektrische voertuigen te gaan.

- Afweging dekking vs. kwaliteit:
Ik koos ervoor om een similaritydrempel van ≥ 0.29 te hanteren. Dit garandeerde voldoende artikeldekking per maand zonder relevante data te missen.
- Sentimentanalyse herzien:
RobBERT v2 gaf bijna uitsluitend hoge scores (0.9), dus ik heb alternatieven getest (TextBlob, Vader, etc.). TextBlob leverde het meest gevarieerde en interpreteerbare resultaat op. Deze sentimentscore heb ik vervolgens per maand geaggregeerd.
- Feature engineering en modelverbetering:
 - Sentimentscore én tijdsindex zijn toegevoegd aan de dataset.
 - Opmerking van Daphne: de lag-feature was initieel niet correct toegepast op het model. Dit heb ik hersteld door Totaal_verkochte_wegvoertuigen_lag1 correct als input mee te geven aan XGBoost.

3. Modelverbetering en evaluatie

Na deze verbeteringen is de MAE van het model significant gedaald tot 4662.99. Ik heb vervolgens een Leave-One-Feature-Out (LOFO) analyse uitgevoerd op basis van walk-forward validatie. De resultaten tonen het belang van de tijds- en nieuwsfeatures:

Weggelaten featuregroep	MAE (walk-forward)	Gem. foutmarge (%)
Voorgaande verkoop (lag1)	3778.9	15.6%
Economisch sentiment	3804.6	15.4%
Brandstofprijzen	4365.7	17.8%
Nieuwsartikelen (sentiment)	5222.0	21.6%
Tijdsindex	5404.8	22.2%

Dit bevestigt dat zowel tijd, economisch sentiment als lagged verkoopcijfers sterke verklarende waarde hebben. Het uitsluiten van de tijdsindex resulteert in de hoogste MAE, wat aansluit bij het advies van Prof. Van Montfort.

71.22.1.1 *Tabel met MAE van verschillende verkoop voorspellingen*

Techniek / Model	Dataset / Context	MAE Resultaat	Opmerkingen
DGA-based NGBM (1,1) (Pei en Li, 2019)	Voorspelling EV-verkoop	< 300 voertuigen3	
Multiple Regression (Bitencourt et al., 2021)	Voorspelling EV-verkoop	Gerapporteerd	Specifiek resultaat niet gegeven, wel dat MAE is gebruikt voor evaluatie.
NGMB and PSO (Ding et al., 2021)	Voorspelling EV-verkoop	10071.274	

Sequence to Sequence (Yi et al., 2022)	Voorspelling EV-verkoop	4.71194	
OFDGPM (Liu vd., 2022)	Voorspelling EV-verkoop		Specifiek resultaat niet gegeven, wel dat MAE is gebruikt voor evaluatie.
BPNN, RNN, en LSTM (Ding et al., 2023)	Voorspelling EV-verkoop	159.475	Gebruikten online reviews, sentimentanalyse, en historische verkoopdata6.
Bi-LSTM, ISOMAP, LLE, en PCA (Liu et al., 2023)	Voorspelling EV-verkoop		Specifiek resultaat niet gegeven, wel dat MAE is gebruikt voor evaluatie.
Neural Networks, Bayesian linear regression, LightGBM (Yeh en Wang, 2023)	Voorspelling EV-verkoop		Specifiek resultaat niet gegeven, wel dat MAE is gebruikt voor evaluatie.
ARIMA, XGBoost, Random Forest, en LSTM (Kumar et al., 2024)	Voorspelling EV-verkoop (India, twee- en driewielers)		Specifiek resultaat niet gegeven, wel dat MAE is gebruikt voor evaluatie.
ARIMA en LSTM (Ning et al., 2024)	Voorspelling NEV (New Energy Vehicle) verkoop (China)		Gebruikten data tussen 2018 en 20238. Specifiek resultaat niet gegeven.
EVs-PredNet (Voorgesteld Model)	Voorspelling Vraag EV - BEV (Battery Electric Vehicle)	3.67059	LSTM- en CNN-gebaseerd model10....
EVs-PredNet (Voorgesteld Model)	Voorspelling Vraag EV - HEV (Hybrid Electric Vehicle)	4.13589	
EVs-PredNet (Voorgesteld Model)	Voorspelling Vraag EV - PHEV (Plug-in Hybrid Electric Vehicle)	1.38019	Laagste MAE voor specifieke EV-categorie9.
EVs-PredNet (Voorgesteld Model)	Voorspelling Vraag EV - ALL (Alle elektrische voertuigen samen)	2.05509...	Gebruikt Adam optimizer13. Hoogste R-Squared voor deze categorie9.
Decision Tree (DT)	Voorspelling Vraag EV - ALL (Traditionele ML-methode ter vergelijking)	2.942512	Vergeleken met EVs-PredNet12....

Linear Regression (LR)	Voorspelling Vraag EV - ALL (Traditionele ML-methode ter vergelijking)	2.521512	Meest succesvol van de ML-methoden voor ALL, BEV, HEV2.
Lasso-LR	Voorspelling Vraag EV - ALL (Traditionele ML-methode ter vergelijking)	2.800312	Vergeleken met EVs-PredNet12.
ElasticNet-LR	Voorspelling Vraag EV - ALL (Traditionele ML-methode ter vergelijking)	2.773712	Vergeleken met EVs-PredNet12.
Ridge-LR	Voorspelling Vraag EV - ALL (Traditionele ML-methode ter vergelijking)	2.611612	Vergeleken met EVs-PredNet12.
K-Nearest Neighbors (KNN)	Voorspelling Vraag EV - ALL (Traditionele ML-methode ter vergelijking)	2.850712	Vergeleken met EVs-PredNet12.
Support Vector Machines (SVM)	Voorspelling Vraag EV - ALL (Traditionele ML-methode ter vergelijking)	4.455712	Slechtste prestatie onder traditionele ML-methoden2.
AdaBoost Regression (ABR)	Voorspelling Vraag EV - ALL (Traditionele ML-methode ter vergelijking)	2.655912	Presteerde beter voor PHEV dan andere ML-methoden2.
Decision Tree regression (M5P classifier)	Voorspelling Autoverkoop Nederland (m.b.v. sentiment & Google Trends)	141.452615	Sentiment had weinig voorspellende kracht15.
Random Forest	Voorspelling Autoverkoop China - Benchmark Dataset	142.3116	Gebruikt Baidu Index, benzineprijs, CCI17.
Random Forest	Voorspelling Autoverkoop China - Benchmark + Autohome Dataset	35.3616	Inclusief Autohome data (views, reviews, scores, etc.)17.
Random Forest	Voorspelling Autoverkoop China -	122.8816	Inclusief Weibo data (mentions, sentiment

	Benchmark + Weibo Dataset		counts)17.
Random Forest	Voorspelling Autoverkoop China - Combined Dataset	45.0216	Combinatie van Benchmark, Autohome en Weibo data17.
XGBoost	Voorspelling Autoverkoop China - Benchmark Dataset	151.3516	Gebruikt Baidu Index, benzineprijs, CCI17.
XGBoost	Voorspelling Autoverkoop China - Benchmark + Autohome Dataset	37.8616	Beste prestatie in deze studie16. Autohome data verbeterde nauwkeurigheid significant18.
XGBoost	Voorspelling Autoverkoop China - Benchmark + Weibo Dataset	125.5516	Inclusief Weibo data17.
XGBoost	Voorspelling Autoverkoop China - Combined Dataset	44.3416	Combinatie van Benchmark, Autohome en Weibo data17.
SVR	Voorspelling Autoverkoop China - Benchmark Dataset	239.9816	Gebruikt Baidu Index, benzineprijs, CCI17.
SVR	Voorspelling Autoverkoop China - Benchmark + Autohome Dataset	119.7316	Inclusief Autohome data17.
SVR	Voorspelling Autoverkoop China - Benchmark + Weibo Dataset	220.7816	Inclusief Weibo data17.
SVR	Voorspelling Autoverkoop China - Combined Dataset	103.4016	Combinatie van Benchmark, Autohome en Weibo data17.

71.22.2 Gesprek met labbegeleider Kees

Tijdens de Thesis Supervision Call hebben prof. dr. C.A.G.M. van Montfort en ik samen naar de voorspelde lijn van het model gekeken. Opvallend was dat de voorspellingen structureel lager liggen dan de daadwerkelijke verkoopcijfers, hoewel het model de pieken en dalen redelijk goed volgt. Prof. Van Montfort gaf aan dat, indien de voorspelde lijn verticaal zou worden verschoven, deze beter zou overlappen met de werkelijke waarden. Dit duidt mogelijk op een systematische fout of bias in het model.

Daarnaast heb ik het idee besproken om seizoensgebonden fluctuaties te normaliseren. Prof. Van Montfort was het hier niet mee eens en vond dit geen geschikte aanpak in het huidige stadium van het onderzoek.

Een belangrijk punt dat hij naar voren bracht, is dat ik tot nu toe de factor *tijd* niet als verklarende variabele heb meegenomen. Hij stelde hiervoor twee mogelijke benaderingen voor:

1. Tijd uitsluiten: de tijdscomponent volledig buiten beschouwing laten en in plaats daarvan de verkoopcijfers corrigeren voor langetermijntrends. Deze aanpak zou echter veel werk en ingrepen in de dataset vereisen.
2. Tijd meenemen als verklarende variabele:
 - Via een continue variabele, zoals het aantal maanden sinds de start van de reeks.
 - Of via dummyvariabelen per jaar. Dit houdt in dat per jaar een binaire variabele (0 of 1) wordt toegevoegd. Bij een dataset met 10 jaren zijn dan 9 dummyvariabelen nodig om multicollineariteit te vermijden (referentiecategorie).

Aangezien ik twee jaren gebruik voor validatie en test, zou ik in principe ook voor die jaren afzonderlijke dummy's moeten toevoegen. Prof. Van Montfort merkte op dat er in de eerste jaren mogelijk weinig verandering zat in de verkoopcijfers, waardoor de bijbehorende dummyvariabelen weinig onderscheidend vermogen hebben. Een alternatief is het groeperen van jaren in drie bredere categorieën, zoals vroege jaren, middenperiode en recente jaren.

Verder adviseerde hij om te overwegen om in plaats van een tijdsgebaseerde splitsing, juist een *random split* toe te passen voor het trainen en testen van het model, afhankelijk van het onderzoeksdoel.

Ten slotte heb ik mijn zorgen geuit over het behalen van leerdoel A (ontwerp), aangezien ik geen externe opdrachtgever heb. Prof. Van Montfort gaf aan dat dit in de beoordeling buiten beschouwing wordt gelaten wanneer een opdrachtgever niet van toepassing is. Het volstaat om dit kort te benoemen in het verslag. De eerder opgestelde projectbeschrijving door M. Visser [1] is in mijn geval voldoende. Ik zal dit morgen nog verifiëren met Marcio.

Tot slot werd benadrukt dat ik de gebruikte variabelen goed moet kunnen uitleggen. Door het toevoegen van tijds-dummy's zou ik mogelijk het aantal aggregaties van andere variabelen kunnen reduceren. Zo neem ik nu bijvoorbeeld meerdere aggregaties van brandstofprijzen mee (zoals minimum, maximum, gemiddelde, etc.), maar dat maakt de interpretatie lastig. Door dit terug te brengen tot één enkele aggregatie, zoals het maandgemiddelde, wordt het model eenvoudiger uitlegbaar.

Referentie

[1] M. Visser, *Analyzing the Influence of Social Media Sentiment on Stock Market Fluctuations: A Comprehensive Study*

71.23 05/05

Activiteit: Verkenning subsidieregelingen

Omschrijving:

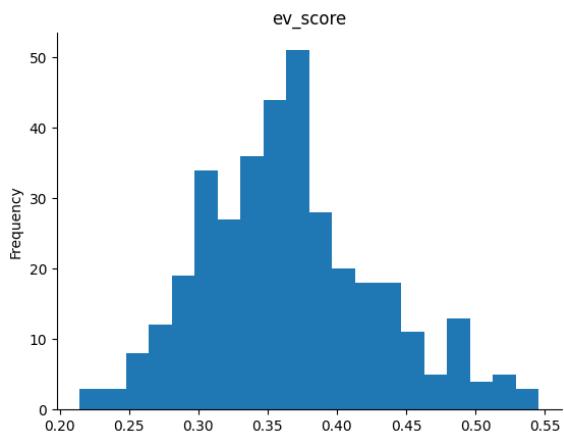
Vandaag onderzocht of maandelijkse data over SEPP-subsidies (Subsidie Elektrische Personenauto's Particulieren) beschikbaar zijn om als verklarende variabele toe te voegen aan het voorspellingsmodel voor EV-verkoop. Uit documentanalyse (Tussentijdse Evaluatie SEPP) blijkt dat de regeling pas in juli 2020 van start is gegaan en dat beschikbare cijfers voornamelijk op jaarrichting worden gerapporteerd. Maandelijkse cijfers zijn niet publiek beschikbaar, maar zouden mogelijk opgevraagd kunnen worden via de RVO. Alternatief overwogen om maandelijkse EV-registraties van het CBS te gebruiken als proxy voor subsidiegebruik, eventueel gecombineerd met jaarlijkse subsidiebedragen en lag-structuren.

71.23.1 Verkenning en filtering van nieuwsartikelen m.b.v. keywords en embedding

Vandaag heb ik nieuwsartikelen van NOS.nl gefilterd op basis van zowel trefwoorden als semantische embedding scores. Voor de trefwoordenfilter heb ik een uitgebreide lijst van meer dan 50 EV-gerelateerde termen gebruikt, waaronder "elektrische auto", "laadpaal", "Tesla", "subsidie", en "verduurzaming vervoer". Deze trefwoorden zijn bedoeld om zo veel mogelijk relevante artikelen rondom elektrische voertuigen (EV's) te selecteren:

```
ev_keywords = ["elektrische auto", "laadpaal", "Tesla", ..., "verduurzaming vervoer"]
```

Na het toepassen van deze filter heb ik aanvullend embedding-gebaseerde filtering toegepast (via Jina Embeddings + cosine similarity). Dit gaf mij een ev_score die aangeeft hoe sterk een artikel inhoudelijk lijkt op een EV-referentieartikel. De scores bleken normaal verdeeld, zoals zichtbaar in onderstaande histogram.

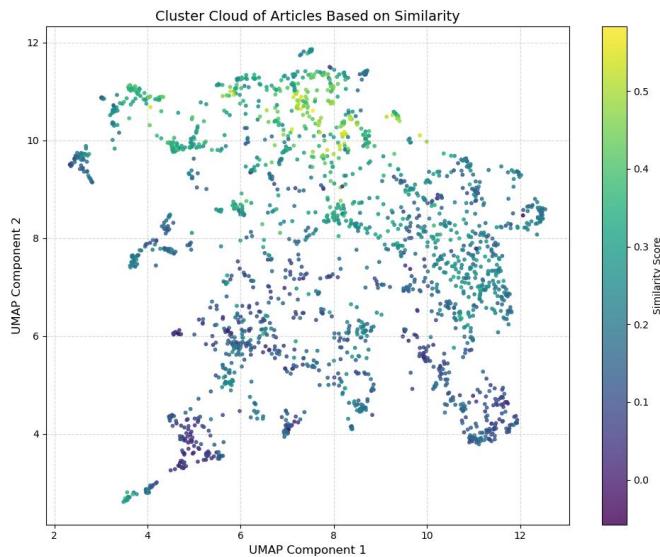


Wat viel op:

- Artikelen met de hoogste ev_score zijn allemaal inhoudelijk relevant (bijv. over subsidies, nieuwe EV-modellen, beleid).
- De artikelen met de laagste scores zijn totaal niet EV-gerelateerd. Enkele voorbeelden zijn:
 - Horeca rechtszaak tijdens corona
 - Vissersmoorden voor de kust van Suriname
 - Discussies over kerkdiensten en coronabeleid

Deze negatieve voorbeelden bevatten geen enkele relevante term uit de keyword-lijst. Dit bevestigt dat de embedding-filter effectief is, maar ook dat een minimum ev_score nodig is om off-topic artikelen uit te sluiten.

Clustering van artikelen (UMAP)



Op basis van de embeddings heb ik een clusterwolk (UMAP) geplot. Hierin zijn semantisch vergelijkbare artikelen bij elkaar gegroepeerd. Artikelen met hoge EV-similarity zijn geconcentreerd in bepaalde regio's van de plot. Dit suggereert dat een drempelwaarde (bijv. `ev_score > 0.35`) zou kunnen helpen om enkel relevante EV-artikelen te selecteren voor sentimentanalyse of tijdreeksvoorspelling.

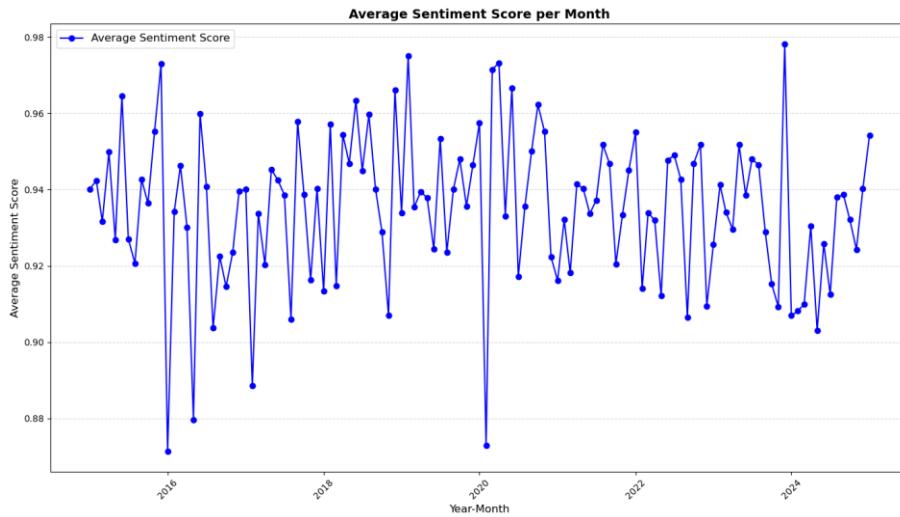
Sentimentanalyse

Voor sentimentanalyse heb ik het RobBERT v2-model gebruikt via:

```
pipe = pipeline("text-classification", model="DTAI-KULeuven/robbert-v2-dutch-sentiment")
```

Dit model retourneert slechts 3 klassen (NEG, NEU, POS), en pogingen om dit naar een 5-puntsschaal te mappen leverden uitsluitend neutrale labels op. Daarom heb ik het originele 3-klassensysteem behouden, aangezien dit ook het best presterende model is volgens [de DBRD dataset benchmark](#).

De sentimentanalyse is succesvol uitgevoerd op de volledige dataset in Google Colab (i.v.m. performance). Per maand is de gemiddelde sentimentscore berekend en opgeslagen in een CSV-bestand (`average_sentiment_per_month.csv`) dat later gebruikt wordt voor XGBoost time-series analyse.



Observaties:

- De gemiddelde sentimentsscores per maand vertonen beperkte variatie (meestal tussen 0.91–0.96), wat de voorspellende kracht mogelijk onderdrukt.
- De correlatie tussen nieuws-sentiment en autoverkoop is dan ook verwaarloosbaar ($r = 0.04$).
- Daarentegen tonen andere variabelen sterke correlaties:
 - Consumentenvertrouwen: -0.76
 - Economisch klimaat: -0.78
 - Brandstofprijzen (benzine/diesel): +0.65–0.67

Aanbevelingen:

- Verbeter de filtering van nieuws: Gebruik alleen artikelen met een hoge ev_score (bijv. > 0.35) en controleer op aanwezigheid van relevante keywords.
- Voeg lag-variabelen toe: Indicatoren zoals vertrouwen, brandstofprijs en sentiment kunnen pas na enkele maanden effect hebben op aankoopbeslissingen. Voeg vertragingen toe ($t-1, t-2$, etc.).
- Vergelijk met een baseline: Train ook een model zonder sentimentvariabele om de toegevoegde waarde van nieuws te evalueren.
- Toekomstige filtering: Overweeg het combineren van keyword-, embedding- én named entity filters (zoals merken of technologieën) om ruis verder te minimaliseren.

71.23.1.1 Modelontwikkeling & Evaluatie (XGBoost, RNN en LOFO-analyse)

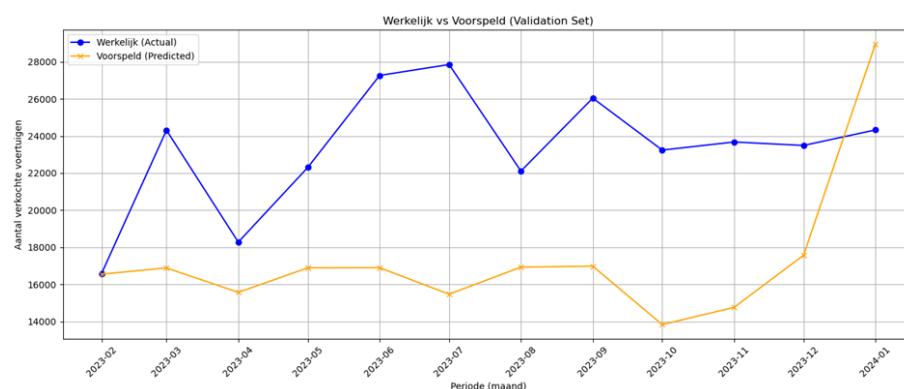
Vandaag heb ik gewerkt aan de time series voorspelling van voertuigverkoop met XGBoost en RNN-modellen. Hiervoor heb ik een train-, validation- en testset opgezet, waarbij de laatste twee jaar (2023–2024) als validation- en testperiode zijn gebruikt. De validatieset zonder nieuwsinput werd als baseline gemeten:

- XGBoost baseline (zonder nieuws): MAE = 7276.32 voertuigen
- RNN baseline (zonder nieuws): MAE = 8145.07 voertuigen

De MAE (Mean Absolute Error) is gekozen als hoofdmaatstaf, conform de aanbevelingen uit het boek van François Chollet, *Deep Learning with Python*, 2nd Edition (ISBN: 9781617296864).

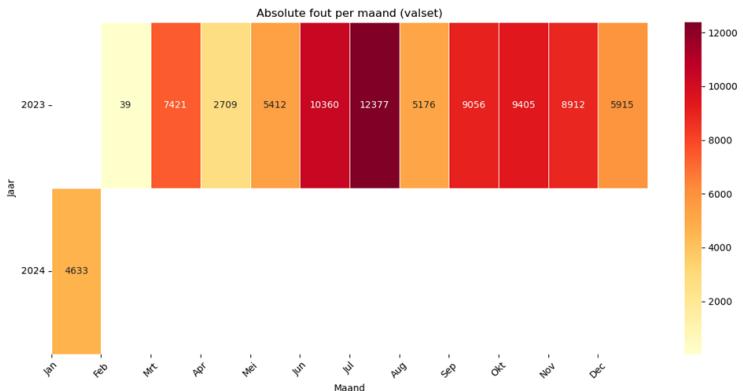
Vervolgens is de nieuwssentimentscore (afkomstig van het RobBERT-model) toegevoegd aan de feature set:

- XGBoost mét nieuws: MAE = 6784.66
- RNN mét nieuws: MAE = 8114.22



Dit suggereert een lichte verbetering door het toevoegen van de sentimentfeature in het geval van XGBoost.

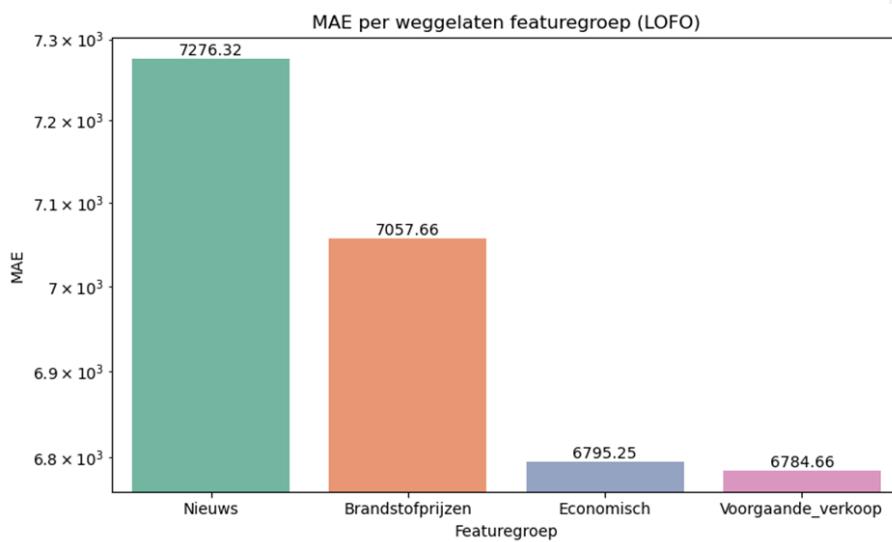
De validatie-uitvoer liet zien dat het model moeite heeft met piekmaanden: met name juni, augustus en oktober worden structureel onderschat, terwijl december wordt overschat. Dit wijst mogelijk op seisoensinvloeden of exogene factoren die nog niet goed zijn opgenomen in het model. Zie de heatmap en lijnplots voor een visuele ondersteuning van de foutverdeling over de maanden.



LOFO-analyse (Leave One Feature-group Out)

Om de bijdrage van verschillende featuregroepen te evalueren, is een LOFO-analyse uitgevoerd. Per experiment is één categorie weggelaten, waarna opnieuw is getraind en geëvalueerd:

- Zonder nieuws → MAE: 7276.32
- Zonder brandstofprijzen → MAE: 7057.66
- Zonder economische variabelen (vertrouwen/klimaat/koopbereidheid) → MAE: 6795.25
- Zonder historische verkoopdata → MAE: 6784.66

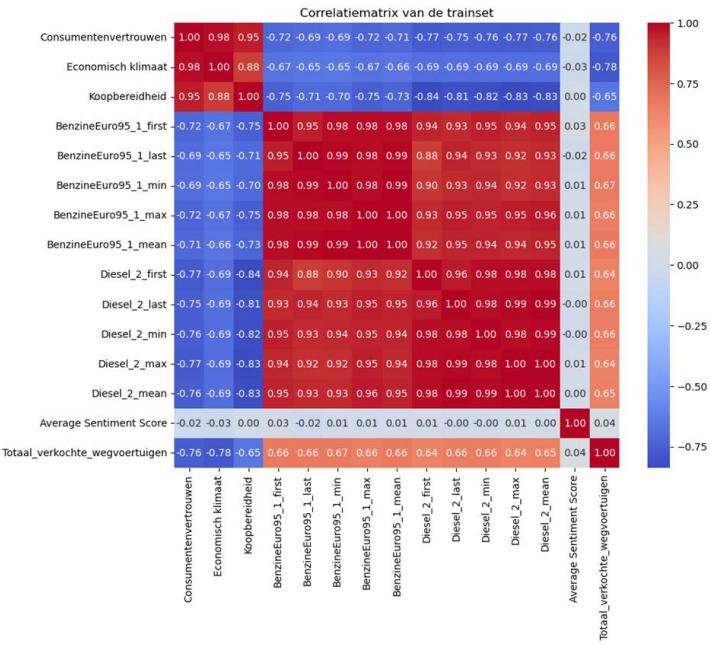


Uit deze resultaten blijkt:

- Voorgaande verkoop is de meest voorspellende featuregroep.
- Economische variabelen zijn belangrijk, maar minder cruciaal dan historische verkoop.

- Nieuws heeft de minste impact; het verwijderen ervan verhoogt de MAE nauwelijks.

Correlatie en modeloverwegingen



De correlatiematrix bevestigt dit beeld:

- Brandstofprijzen correleren positief met verkoop ($r \approx 0.65\text{--}0.67$).
- Consumentenvertrouwen, economisch klimaat en koopbereidheid hebben een sterke negatieve correlatie ($r \approx -0.65$ tot -0.78).
- De nieuwssentimentscore heeft nauwelijks correlatie met voertuigverkoop ($r \approx 0.04$).

Dit duidt erop dat het nieuws (zoals gemeten met RobBERT) in de huidige vorm weinig informatie toevoegt. De oorzaak lijkt te liggen bij een beperkte variatie in sentimentscores, een mismatch tussen het type nieuws en consumentengedrag, en het ontbreken van een vertragingseffect (lag).

71.23.2 Aanbevelingen voor verdere verbetering van het model en de sentimentanalyse

Modeloptimalisatie

- Voeg lag-variabelen toe
Introduceer vertraging (lag t-1, t-2) voor brandstofprijzen, consumentenvertrouwen en sentiment, om vertraagde effecten op verkoop te modelleren.

2. Overweeg seisoenscomponenten expliciet te modelleren
Voeg maand-dummy's of Fourier-features toe om seisoenspatronen te vangen (zoals pieken in zomer en december).
3. Gebruik cross-validation met tijdseriesstructuur
Bijvoorbeeld TimeSeriesSplit voor robuustere prestatie-evaluatie.

Verbetering van sentimentinput

4. Herfilter het nieuws op inhoudelijke relevantie
Filter alleen op artikelen met daadwerkelijke economische, auto- of EV-gerelateerde inhoud — laat algemene nieuwsartikelen met lage similariteit weg.
5. Gebruik EV-similarity scores in plaats van alleen keywords
Experimenteer met een drempelwaarde op de ev_score (bijv. > 0.35) om irrelevante artikelen uit te sluiten.
6. Kies voor een sentimentmodel met meerdere klassen
RobBERT geeft vooral neutrale scores terug bij meer dan 3 klassen. Overweeg alternatieve modellen of fine-tuning op een domeinspecifieke dataset.
7. Voer maandelijkse media-intensiteit mee als feature
Tel het aantal EV-gerelateerde artikelen per maand als extra input, los van de sentimentwaarde.

Evaluatie & interpretatie

8. Blijf een baseline zonder sentiment meenemen in evaluaties
Zo kan per iteratie worden getoetst of sentiment daadwerkelijk waarde toevoegt.
9. Analyseer foutmarges per maand
Gebruik heatmaps en plots om systematische over- of onderschattingen te lokaliseren en gericht bij te sturen.
10. Gebruik SHAP of feature importance-analyse
Visualiseer welke features het meeste bijdragen aan voorspellingen

71.23.2.1.1 Draft to do

Aan de xgboost kant heb ik een tran test val set gemaakt, vanwege time series val en test laaytste 2 jaar gedaan. Valset zonder nos is

MAE voor valset: 7276.32

Uiteindelijk gekozen als mea als maatstaaf omdat dit wrdt aanbevolen uit het boek van cholet
(citering toevoegen) ISBN: 9781617296864

Daarna ook gekeken naar rnn omdat dit werdt aanbevolen, verbaasinwekkend lager resultaat

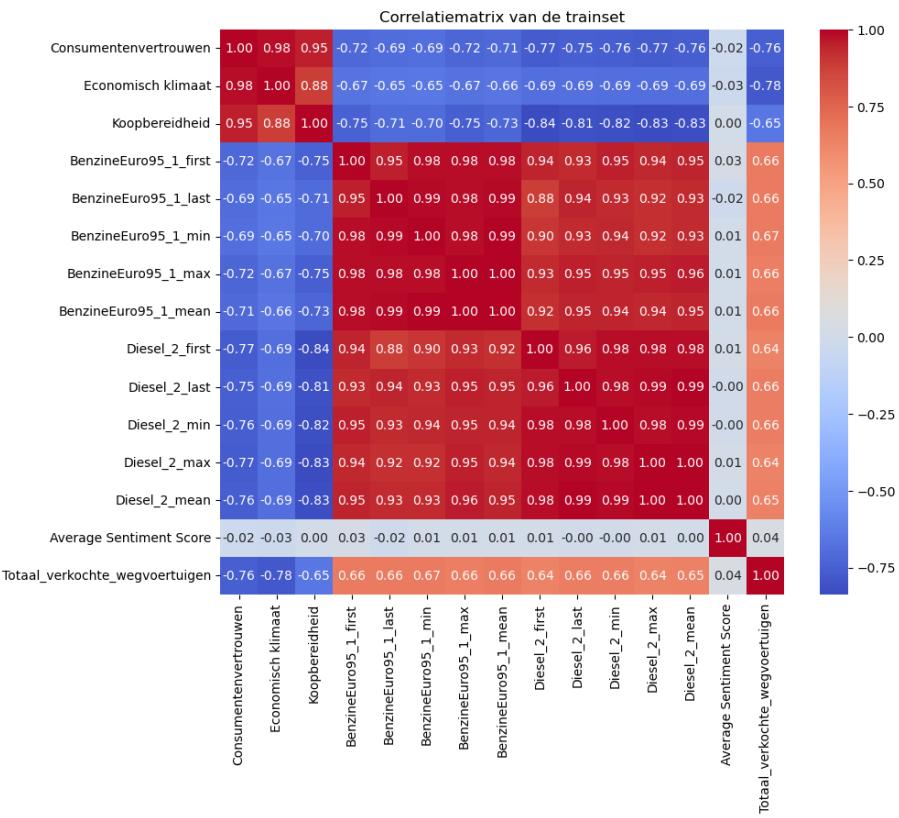
RNN MAE voor valset: 8145.07

Vervolgens nos dataset toegevoegd, bovenste 2 uitkomsten kunnen gezien worden als baseline zonder nos.

Met nos zijn de uitkomsten het volgende MAE voor valset: 6784.66 en voor rnn

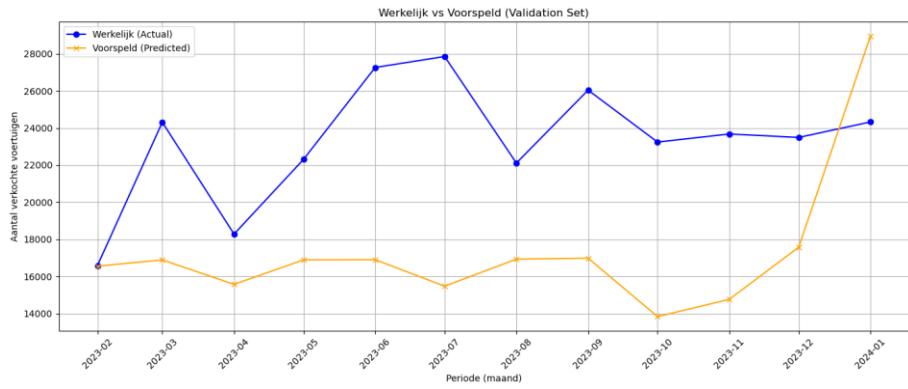
RNN MAE voor valset: 8114.22

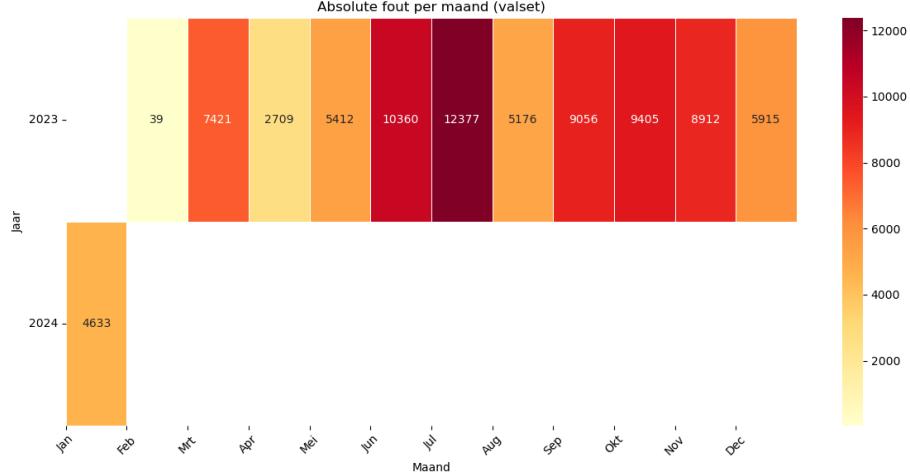
Correlatie is trouwens het volgende:



Het XGBoost-model is geëvalueerd op de validatieset (12 maanden). De gemiddelde absolute fout (MAE) bedraagt 7276 voertuigen. In de eerste maanden liggen de voorspellingen relatief dicht bij de werkelijke waarden (foutmarges rond de 3–15%), maar in het midden van het jaar loopt de foutmarge op tot ruim 40%. Het model onderschat structureel de verkoop tijdens piekmaanden, met name in juni, augustus en oktober. Enkel in december is sprake van overschatting. Deze resultaten wijzen op mogelijke seizoensinvloeden die het model momenteel niet goed opvangt.

Afbeldingen toene xgboost:





Daaran Lofo leave one out toegevoegd, uitkomst was het volgende:

LOFO - Zonder: Nieuws

Gebruikte features (13): ['Consumentenvertrouwen', 'Economisch klimaat', 'Koopbereidheid', 'BenzineEuro95_1_first', 'BenzineEuro95_1_last', 'BenzineEuro95_1_min', 'BenzineEuro95_1_max', 'BenzineEuro95_1_mean', 'Diesel_2_first', 'Diesel_2_last', 'Diesel_2_min', 'Diesel_2_max', 'Diesel_2_mean']

MAE: 7276.32, Gemiddelde foutmarge: 29.88%

LOFO - Zonder: Brandstofprijzen

Gebruikte features (4): ['Consumentenvertrouwen', 'Economisch klimaat', 'Koopbereidheid', 'Average Sentiment Score']

MAE: 7057.66, Gemiddelde foutmarge: 29.25%

LOFO - Zonder: Economisch

Gebruikte features (11): ['BenzineEuro95_1_first', 'BenzineEuro95_1_last', 'BenzineEuro95_1_min', 'BenzineEuro95_1_max', 'BenzineEuro95_1_mean', 'Diesel_2_first', 'Diesel_2_last', 'Diesel_2_min', 'Diesel_2_max', 'Diesel_2_mean', 'Average Sentiment Score']

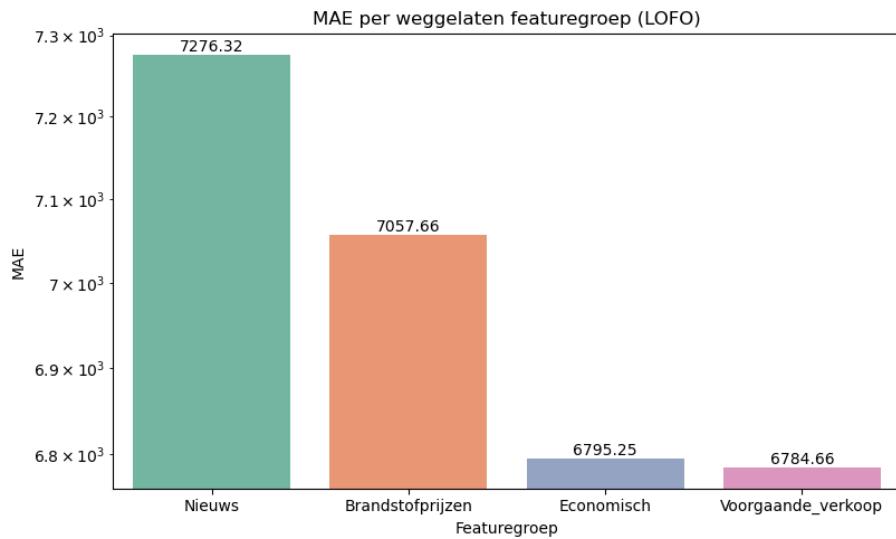
MAE: 6795.25, Gemiddelde foutmarge: 28.55%

LOFO - Zonder: Voorgaande_verkoop

Gebruikte features (14): ['Consumentenvertrouwen', 'Economisch klimaat', 'Koopbereidheid', 'BenzineEuro95_1_first', 'BenzineEuro95_1_last', 'BenzineEuro95_1_min',

'BenzineEuro95_1_max', 'BenzineEuro95_1_mean', 'Diesel_2_first', 'Diesel_2_last',
'Diesel_2_min', 'Diesel_2_max', 'Diesel_2_mean', 'Average Sentiment Score']

MAE: 6784.66, Gemiddelde foutmarge: 27.73%



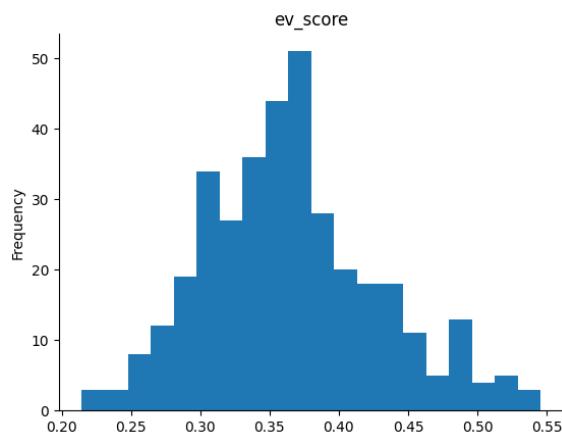
Voor het voorspellen van autoverkoop zijn historische verkoopdata (voorgaande verkoop) de belangrijkste informatiebron, gevolgd door economische sentimenten. Brandstofprijzen dragen het minst bij aan modelnauwkeurigheid.

Vandaag Nos artikelen zowel gefilterd op keywords als embedding ik heb enorm veel keywords gebruikt en daarna embedding gedaan: ev_keywords = [

"elektrische auto", "elektrische voertuigen", "elektrisch rijden", "stekkerauto", "emissievrij",
"zero emission", "elektrisch vervoer", "battery electric vehicle", "BEV", "EV", "EV-markt",
"laadpaal", "laadinfrastructuur", "laadstation", "snelladen", "laadtijd",
"laadnetwerk", "opladen", "elektriciteitsnet", "netcongestie",
"batterij", "accu", "actieradius", "solid state batterij", "energieopslag", "accutechnologie",
"Tesla", "BYD", "Volkswagen ID", "Renault Zoe", "Hyundai Ioniq", "Kia EV",
"Nissan Leaf", "Polestar", "BMW i", "Ford Mustang Mach-E", "Audi e-tron",
"subsidie elektrische auto", "aanschafsubsidie", "bijtelling",
"fiscale stimulering", "milieu-investeringsaftrek", "overheidsbeleid", "klimaatmaatregelen",
"duurzame mobiliteit", "klimaatbeleid", "CO2-uitstoot", "groene transitie",
"verduurzaming vervoer", "energietransitie", "schoner verkeer",
"tweedehands elektrische auto", "EV import", "EV export", "restwaarde EV",
"autoverkoop EV", "EV occasion", "elektrisch leasen", "Subsidieregeling Elektrische Personenauto's Particulieren"

]

Daarna embedding kunnen uitvoeren op collab, resultaat was het volgende zie afbeelding ev scores



Ook heb ik hiermee artikelen kunnen sorteren op simularity, paar artikelen met meeste en was ook benieuwd naar het minste simularity:
Top 10 Articles with the Lowest Similarity Scores:

Article 1821 similarity: -0.0576

Title: Horeca-ondernehmer voor de rechter: 'Thuis meer risico op besmetting'

Description: In een kort geding eisen zeker 68 horecaondernemers onmiddellijke heropening van de horeca. De rechter doet vanmiddag uitspraak.

Content (snippet): Horeca-ondernehmer voor de rechter: 'Thuis meer risico op besmetting' De rechtbank in Den Haag doet vanmiddag om 16.00 uur uitspraak in het door meerdere horecaondernemers aangespannen kort geding. De ...

URL: <https://nos.nl/artikel/2353064-horeca-ondernehmer-voor-de-rechter-thuis-meer-risico-op-besmetting>

Article 915 similarity: -0.0287

Title: Na bloedbad op boot bij Suriname opnieuw visser gedood

Description: Vrijdagnacht zijn ook een aantal boten overvallen. Slechts vier van de twintig bemanningsleden overleefden het.

Content (snippet): Na bloedbad op boot bij Suriname opnieuw visser gedood Op de Atlantische oceaan is vannacht opnieuw een vissersboot overvallen voor de kust van Suriname. Daarbij is in ieder geval de kapitein om het l...

URL: <https://nos.nl/artikel/2230202-na-bloedbad-op-boot-bij-suriname-opnieuw-visser-gedood>

Article 1955 similarity: -0.0215

Title: Viroloog Koopmans: kerk op Urk speelt 'Russische roulette'

Description: De kerk wil de coronamaatregelen negeren en daarmee tegemoet komen aan de wens van gelovigen om weer ter kerke te kunnen gaan.

Content (snippet): Viroloog Koopmans: kerk op Urk speelt 'Russische roulette' Viroloog en OMT-lid Marion Koopmans vindt dat de Gereformeerde Gemeente Sionkerk op Urk "Russische roulette" speelt, door de coronamaatregelen...

URL: <https://nos.nl/artikel/2373944-viroloog-koopmans-kerk-op-urk-speelt-russische-roulette>

Article 910 similarity: -0.0195

Title: Woede en ongeloof over trage zoektocht na zeerovers-tragedie Suriname

Description: Het duurde meer dan een etmaal voordat er een schip van de kustwacht uitvoer om zestien vermiste vissers te zoeken.

Content (snippet): Woede en ongeloof over trage zoektocht na zeerovers-tragedie Suriname Kavita Tarnattie (28) staat op de steiger en staart uitdrukkingsloos over het snelstromende bruine water van de Surinamerivier. Gi...

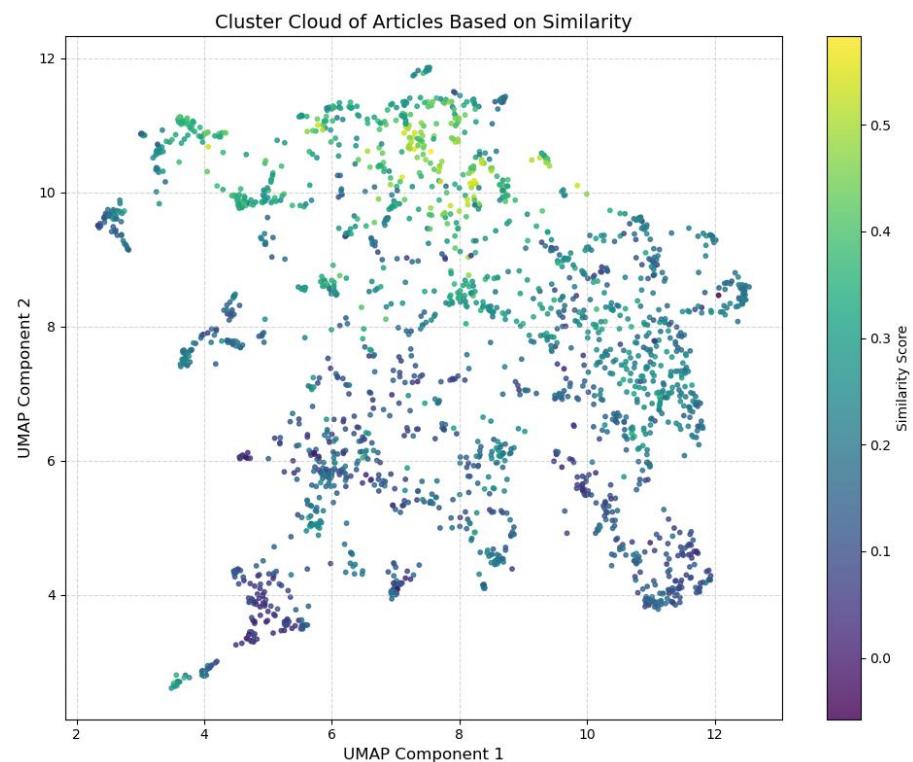
URL: <https://nos.nl/artikel/2230099-woede-en-ongeloof-over-trage-zoektocht-na-zeerovers-tragedie-suriname>

De top 10 lowest bevatten niet eens de keywords, daar moet nog naar geleken worden.

De 10 avg en 10 hoogste simularity zijn wel allemaal artikelen die gaan over ev

Vervolgens wilde ik sentiment analyse uitvoeren, maar kreeg steeds het porobleem dat ik maar 25 artikelen terug kreeg, ik heb dus heb data die ik heb aangemaakt naar keyword search gebruikt, dit werkte uiteindelijk wel, maar heb dus nik gedaan met de ev similarity.

Wel een cluster gemaakt



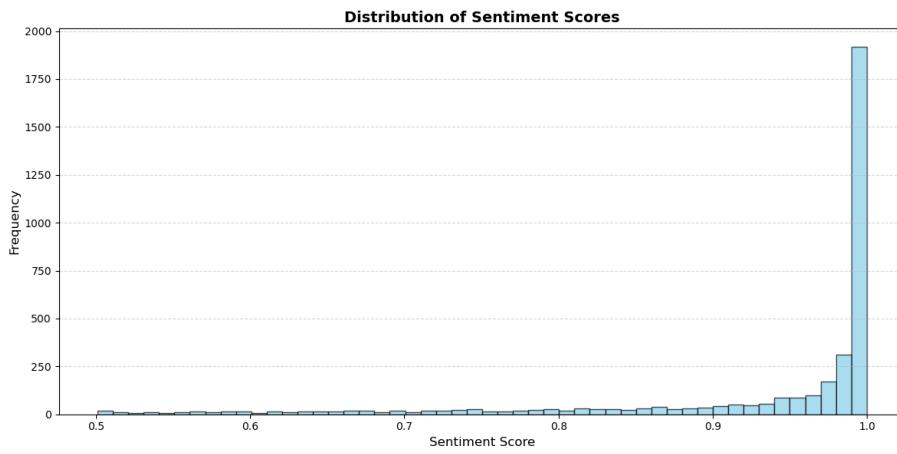
hier viel en volgende mee op...

Daarna sentiment analyse opgezet met pipe = pipeline("text-classification", model="DTAI-KULEuven/robbert-v2-dutch-sentiment")

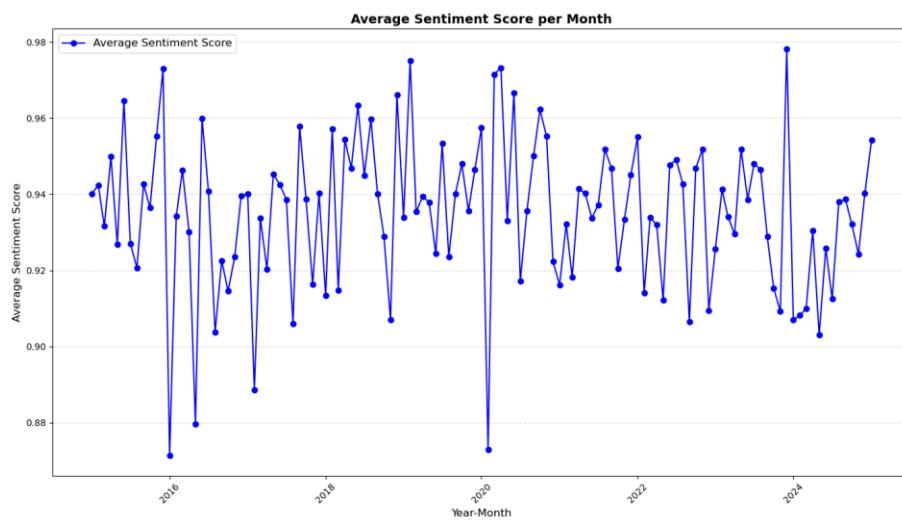
Enige nadeel is dat het niet werkt om meerdere klassen toe te voegen wilde 5 toevoegen maar kreeg toen alleen neutraal, dus heb het op de og 3 gelaten

Want <https://huggingface.co/datasets/benjaminvdb/dbrd> blijkt het best te presteren (voeg toe op welk punt en schrijf ok over andere). Die sentiemnt heb ik gedraaid op de collab met volgende uitkomst

:



Vervolgens heb ik gemiddelde sentimetrn per maand getoont en in een dataset gezet:



71.24 2/05

Eind van de dag todo list afgaan om te checken of alles erin staat

Onderzoeken hoe ik de artikelen op een andere manier kan selecteren, embedding is te heftig, anders

Model selectie ltoelichten

Voor stakeholders:

Onderzoeken wat current work is dus hoe autofabrikanten verkoop van auto;s nu predicten:

<https://www.indeed.com/career-advice/career-development/formula-for-sales-forecast>

EVT gewoon een dashboardaken voor de prediction:

<https://www.figma.com/community/file/1238189616000926360>

lees intereassant info over <https://medium.com/data-science/text-embeddings-comprehensive-guide-afd97fce8fb5> embedding

71.24.1 Idee generatie: waarde propositie:

71.24.1.1 *Hoe je de afstudeerdeonderdelen toepast op jouw scriptie:*

71.24.1.1.1 Prototype

71.24.1.1.2 Idee & Waardepropositie

Wat is het nut van mijn systeem?

- Mijn model biedt inzicht in hoe publieke berichtgeving invloed kan hebben op koopgedrag in de EV-markt.
- Stakeholders: beleidsmakers, autobedrijven, duurzaamheidsexperts → zij kunnen hiermee sneller reageren op publieke opinie.

71.24.1.1.3 Flow Diagram & AI Breakdown

Laat zien hoe mijn systeem werkt in stappen.

- Maak een flowchart met blokken zoals:
 1. Data-inname (NOS + RDW data)
 2. Preprocessing (tokenisatie, cleanen)
 3. Sentimentanalyse (RobBERT v2 → score)
 4. Aggregatie over tijd
 5. Voorspellingsmodel (XGBoost)
 6. Evaluatie (RMSE, MAE)
- Leg uit *waarom* ik voor deze AI-technieken heb gekozen

71.24.1.1.4 Gebruikersonderzoek / Empathy Map

Ik maak geen app voor gebruikers, maar toch kun ik 'gebruikers' definiëren als besluitvormers of analisten.

- Interview 1 of 2 experts uit de branche (denk aan mensen bij de gemeente, duurzaamheid, mobiliteit of marktanalyse).
- Vraag hen:
 - *Zou dit soort voorspellingstool jou helpen?*
 - *Wat voor output zou nuttig zijn?*
 - *Wat voor vertrouwen moet je hebben in zo'n model om het te gebruiken?*
- Dit is jouw "gebruikersonderzoek".

71.25 01/05 To do list

Skelet schrijven

In elk project (blok 1, blok 2 en afstuderen) wordt een rapport opgeleverd ter ondersteuning van je

Product. De belangrijkste onderbouwde keuzes en resultaten zijn daarin terug te lezen. Het rapport heeft een traditionele vorm van een thesis en/of artikel van max. 10 blz. (IEEE-double column). Om een indicatie te geven bevat het rapport de volgende onderdelen. In overleg met je Docentbegeleider/lab-begeleider

Kun je daar iets van wijken.

- Samenvatting
- Introductie
 - Context, Probleem, Bestaand werk, Gat, Voorstel
- Achtergrond
 - Basiskennis (literatuur)
 - State of the art (literatuur)
 - Stakeholder-analyse
- Requirements
 - Product (inclusief gebruikersinterface)
 - Model
- Prototype
 - Idee generatie en value proposition
 - Flow diagrams en AI breakdown
 - Gebruikersonderzoek (empathy map bijvoorbeeld)
 - (Paper) Prototype
- Model
 - Methodologie (dataset, architecture, ...)
 - Resultaten (performance en andere kwaliteitsmaten)
 - Conclusie
- Discussie
 - Implicaties, Future work, Aanbevelingen
 - Terugkomen op requirements
- Bronnenlijst (overweeg het gebruik van een referentiemanager)

Meeste hiervan kan meegenomen worden uit project van blok 2 en 3 kijk daar dus naar

Focus voornamelijk op leerdoel A,

Todo list voor morgen (zet leerdoelen naast de benodigdheden)

- Leg duidelijk uit waarom je gekozen hebt voor specifieke methoden en modellen
- Licht toe waarom bepaalde tools of drempelwaarden (thresholds) zijn gebruikt
- Onderzoek micro-economische factoren die correleren met autoverkoop (binnen 2 weken)
- Gebruik topic modeling en named entity recognition om relevante artikelen te selecteren (binnen 2 weken)
- Presenteer eerste resultaten (binnen 3 weken)
- Focus op de uitleg van je methode en de behaalde resultaten (binnen 3 weken)
- Gebruik Jina Embeddings v3 in plaats van RobBERT voor betere semantische zoekresultaten en sentimentanalyse
- Bereken cosine similarity tussen artikel-embeddings
- Kies een drempelwaarde voor clustering (bijvoorbeeld 0.99 of 0.98)
- Cluster artikelen op basis van cosine similarity
- Identificeer het grootste cluster dat relevant is voor je onderwerp
- Label clusters automatisch
- Gebruik de gelabelde clusters om een classifier te bouwen

71.26 30/04: (Leave one out toepassen, Nos Sentiment maken met keyword search, evt train/test split maken)

71.26.1.1 Leave one out:

Kaggle:

<http://kaggle.com/code/felipefonte99/leave-one-out>

The notebook provides a detailed account of the process, including input, output, and logs. The project involves using Python to process a train file, create a testing set, and develop an LGB model. The notebook has been released under the Apache 2.0 open source license and has been viewed 5,605 times.

LANL Earthquake Prediction

The LANL Earthquake Prediction competition is a research prediction challenge hosted by Los Alamos National Laboratory. The goal is to predict the time remaining before laboratory earthquakes occur from real-time seismic data. The competition uses a dataset from an experiment on rock in a double direct shear geometry subjected to bi-axial loading, which mimics the cycle of loading and failure on tectonic faults. The winner will receive a prize of \$20,000.

<http://kaggle.com/c/LANL-Earthquake-Prediction>

Verslag voortgangsoverleg huiswerkbegeleider

Tijdens het overleg is besproken dat er een aantal vervolgstappen nodig zijn voor de analyse en rapportage. Er moet een scatterplot gemaakt worden van het consumentenvertrouwen ten opzichte van autoverkoop, en aanvullende informatie over consumentenvertrouwen worden verzameld. Daarnaast moet de lagtime worden onderzocht, zowel in de train data als specifiek voor benzineprijzen en consumentenvertrouwen. Het is nog onduidelijk of eerst per feature gekeken moet worden of dat de lagtime als eerste moet worden berekend – dit dient nog te worden uitgezocht. Verder moet worden gecontroleerd of de lagtime correct wordt berekend en of de foutmarges realistisch zijn (met name bij voorgaande verkoopdata lijkt de foutmarge verdacht laag).

Ook is geopperd om alternatieve modellen te overwegen, zoals RNN, die mogelijk goed kunnen presteren.

Tot slot is afgesproken dat er morgenochtend gestart wordt met het schrijven van de tekst. Daarna volgt verdere programmering, preprocessing en training van de data. Daarnaast wordt begonnen met een eerste opzet van het onderzoeksrapport, inclusief inleiding, onderzoeksraag en eventueel al uitgevoerde experimenten.

71.27 28/04 (Dataset gemerged, code naar stefan gestuurd)

Data nog een keer proberen te mergen, is gelukt, maar De data van het consumentenvertrouwen werd tot 2017 maandelijks bijgehouden, maar sinds 2017 is dit per kwartaal. Om toch een consistente maandelijkse dataset te hebben, heb ik lineaire interpolatie toegepast. Hiermee worden de ontbrekende maanden tussen de kwartaalmetingen ingevuld op basis van het gemiddelde tussen de voorgaande en volgende bekende waarden.

71.27.1.1 Notebook XGBoost car sales:

nog doen:

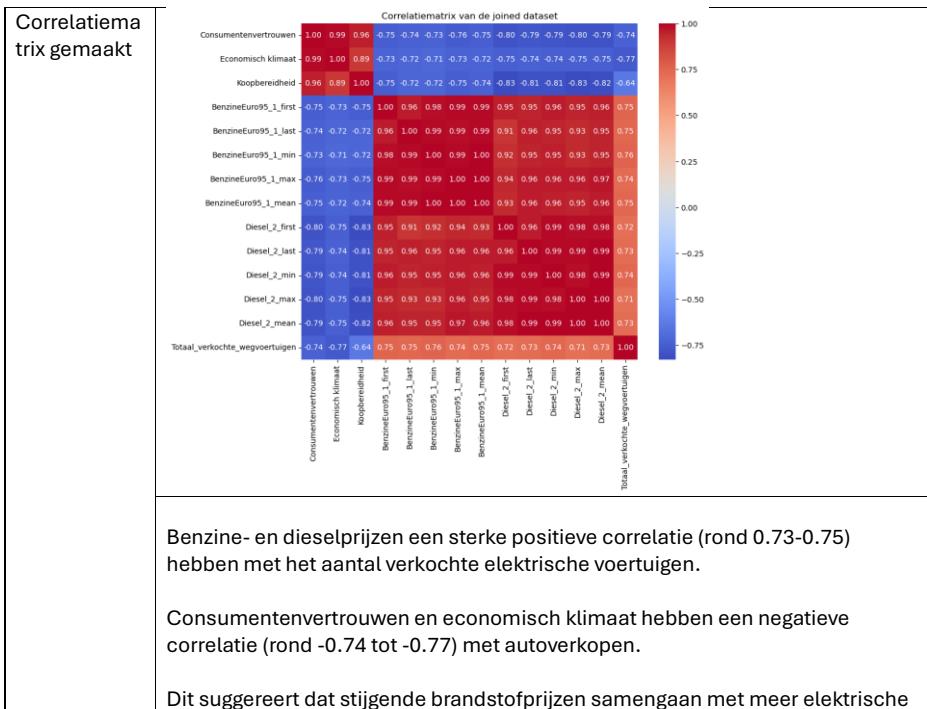
- Uitleggen waarom ik de prestatie maten gebruik (mae foutmarge)
- Verschillende lag time's verantwoorden
- Cars sales anonymiseren
- Scatterplot maken van consumenten vertrouwen & autoverkoop
- Extra info opdoen over consumenten vertrouwen
- Lagtime (checken door traindata goed te zetten)
- Ook lagtime berekenen voor benzine/consumenten vertrouwen
- Eerst per feature dan lag time of andersom? ← checken
- Check of lagtime juist wordt berekend
- Check foutmarge (met voorgaande verkoop verdacht lage foutmarge)
- Eventueel ook andere modellen gebruiken (rnn werkt wss ook goed) → daar nog naar kijken.
-

Plan voor tekst opschrijven:

Morgen in de ochtend start met schrijven, daarna verder gaan met programmeren

Andere contributie was om het preprocessen en trainen van data

Alvast skelet versie onderzoeksrapport schrijven (met inleiding/onderzoeksvraag en evt uitgevoerde experimenten.)



	<p>autoverkopen, terwijl een dalend consumentenvertrouwen mogelijk leidt tot hogere verkopen, mogelijk door stimuleringsmaatregelen.</p>																																																																	
supervised time series	<table border="1"> <thead> <tr> <th>Maand</th> <th>Werkelijk</th> <th>Voorspeld</th> <th>Fout</th> <th>Foutmarge (%)</th> </tr> </thead> <tbody> <tr><td>1</td><td>22060.0</td><td>24068.6</td><td>2008.6</td><td>9.10%</td></tr> <tr><td>2</td><td>29193.0</td><td>21791.1</td><td>7401.9</td><td>25.36%</td></tr> <tr><td>3</td><td>25326.0</td><td>27414.6</td><td>2088.6</td><td>8.25%</td></tr> <tr><td>4</td><td>26607.0</td><td>23810.2</td><td>2796.8</td><td>10.51%</td></tr> <tr><td>5</td><td>27731.0</td><td>24069.4</td><td>3661.6</td><td>13.20%</td></tr> <tr><td>6</td><td>25370.0</td><td>23942.2</td><td>1427.8</td><td>5.63%</td></tr> <tr><td>7</td><td>23683.0</td><td>26062.0</td><td>2379.0</td><td>10.05%</td></tr> <tr><td>8</td><td>28820.0</td><td>23167.1</td><td>5652.9</td><td>19.61%</td></tr> <tr><td>9</td><td>30491.0</td><td>26553.2</td><td>3937.8</td><td>12.91%</td></tr> <tr><td>10</td><td>32623.0</td><td>28185.0</td><td>4438.0</td><td>13.60%</td></tr> <tr><td>11</td><td>39594.0</td><td>29159.3</td><td>10434.7</td><td>26.35%</td></tr> <tr><td>12</td><td>33680.0</td><td>35184.1</td><td>1504.1</td><td>4.47%</td></tr> </tbody> </table> <p>MAE: 3977.64</p>	Maand	Werkelijk	Voorspeld	Fout	Foutmarge (%)	1	22060.0	24068.6	2008.6	9.10%	2	29193.0	21791.1	7401.9	25.36%	3	25326.0	27414.6	2088.6	8.25%	4	26607.0	23810.2	2796.8	10.51%	5	27731.0	24069.4	3661.6	13.20%	6	25370.0	23942.2	1427.8	5.63%	7	23683.0	26062.0	2379.0	10.05%	8	28820.0	23167.1	5652.9	19.61%	9	30491.0	26553.2	3937.8	12.91%	10	32623.0	28185.0	4438.0	13.60%	11	39594.0	29159.3	10434.7	26.35%	12	33680.0	35184.1	1504.1	4.47%
Maand	Werkelijk	Voorspeld	Fout	Foutmarge (%)																																																														
1	22060.0	24068.6	2008.6	9.10%																																																														
2	29193.0	21791.1	7401.9	25.36%																																																														
3	25326.0	27414.6	2088.6	8.25%																																																														
4	26607.0	23810.2	2796.8	10.51%																																																														
5	27731.0	24069.4	3661.6	13.20%																																																														
6	25370.0	23942.2	1427.8	5.63%																																																														
7	23683.0	26062.0	2379.0	10.05%																																																														
8	28820.0	23167.1	5652.9	19.61%																																																														
9	30491.0	26553.2	3937.8	12.91%																																																														
10	32623.0	28185.0	4438.0	13.60%																																																														
11	39594.0	29159.3	10434.7	26.35%																																																														
12	33680.0	35184.1	1504.1	4.47%																																																														
	<p>maandelijkse data om in een supervised time series-formaat met sliding window (6 lags), traaint een XGBoost-model met walk-forward validatie, en berekent de foutmarge en MAE voor de laatste 12 maanden.</p>																																																																	
Lofo toegepast	<p>Ik heb per featuregroep getest hoeveel de modelprestatie daalt als deze groep wordt weggeleten. Resultaat: brandstofprijzen zijn het meest bepalend voor de nauwkeurigheid van het model (hoogste MAE bij weglaten), gevolgd door economische indicatoren. De groep "voorgaande verkoop" lijkt nauwelijks invloed te hebben, maar dit is misleidend doordat lag-variabelen van verkoopdata waarschijnlijk nog in het model zaten. Dit wordt in de volgende iteratie gecorrigeerd.</p> <p>Train/val/test splits: De tijdsreeksdata (maandelijks 2015–2025) is gesplitst in:</p> <ul style="list-style-type: none"> • Trainset: alles tot en met begin 2022 • Validatieset: 10 maanden (2022-03 t/m 2022-12) • Testset: laatste 12 maanden (2023 t/m 2024) 																																																																	

		MAE per weggelaten featuregroep (LOFO)
	Brandstofprijzen	10041.54
	Economisch Featuregroep	7297.38
	Voorgaande_verkoop	0.03
		<p>LOFO - Zonder: Brandstofprijzen Gebruikte features (4): ['Consumentenvertrouwen', 'Economisch klimaat', 'Koopbereidheid', 'Totaal_verkochte_voertuigen'] MAE: 10041.54, Gemiddelde foutmarge: 32.67%</p> <p>LOFO - Zonder: Economisch Gebruikte features (11): ['BenzineEuro95_1_first', 'BenzineEuro95_1_last', 'BenzineEuro95_1_min', 'BenzineEuro95_1_max', 'BenzineEuro95_1_mean', 'BenzineEuro95_1_std', 'BenzineEuro95_1_sum', 'BenzineEuro95_1_var', 'BenzineEuro95_1_weightedmean', 'BenzineEuro95_1_weightedsum', 'BenzineEuro95_1_weightedvar'] MAE: 7297.38, Gemiddelde foutmarge: 23.04%</p> <p>LOFO - Zonder: Voorgaande_verkoop Gebruikte features (13): ['Consumentenvertrouwen', 'Economisch klimaat', 'Koopbereidheid', 'BenzineEuro95_1_mean', 'BenzineEuro95_1_std', 'BenzineEuro95_1_weightedmean', 'BenzineEuro95_1_weightedsum', 'BenzineEuro95_1_weightedvar', 'BenzineEuro95_1_min', 'BenzineEuro95_1_max', 'BenzineEuro95_1_sum', 'BenzineEuro95_1_var', 'BenzineEuro95_1_weightedmin', 'BenzineEuro95_1_weightedmax'] MAE: 0.03, Gemiddelde foutmarge: 1.92%</p>

(bron toevoegen)

27/04

ToDo:

- Datasets opnieuw inladen
- <iframe src="https://www.kaggle.com/embed/johnboulard/exploration-public-nos-dataset-feb-2025?cellIds=1&kernelSessionId=234413029" height="300" style="margin: 0 auto; width: 100%; max-width: 950px;" frameborder="0" scrolling="auto" title="EXPLORATION public NOS dataset - feb 2025"></iframe>
- <https://www.kaggle.com/code/johnboulard/exploration-public-nos-dataset-feb-2025?cellIds=1&kernelSessionId=234413029>

- Meeting Kees:

Onderwerpen:

1. Nieuwsartikelen lastig te filteren:

- o De dataset met nieuwsberichten bleek te groot om effectief te filteren.
- o Daarnaast was het sentiment in de berichten te neutraal voor goede analyses.

2. Tweets geprobeerd:

- o Gekeken naar een dataset met tweets en sentiment (al gedaan in ander onderzoek).
- o Niet verder uitgewerkt in dit gesprek.

3. Plan B – Keyword search:

- o Idee: artikelen handmatig selecteren na keyword search.
- o Kees vond dit een goed plan, maar Marcio was er minder enthousiast over.

4. Algemene voortgang scriptie:

- o Scriptie loopt goed! (problemen horen erbij)
- o Kees adviseert om zo veel mogelijk vast te houden aan het oorspronkelijke plan.

5. Modelkeuze (advies van Daphne):

- o Blijven bij regressiemodel (dus geen overstap naar time-series).
- o Dit is handiger en past bij het oorspronkelijke plan.

6. Technische hulp – Embeddings:

- o Overwegen om Stefan (vader van vriendin, werkt bij Microsoft) te benaderen voor hulp bij het uitvoeren van de embedding code.

☒ Data opschonen:

- Herlaad alle data (om mogelijke duplicaten of fouten eruit te halen).

☒ Sentiment-analyse:

- Bepaal of je doorgaat met de tweets of verdergaat met keyword search + handmatige selectie van artikelen.

☒ Modelkeuze bevestigen:

- Houd vast aan het regressiemodel (zoals Daphne adviseerde).

☒ Stefan appen:

- Vraag of hij kan helpen met het draaien van de embedding code.

☒ Terug naar time-series overweging:

- Check of je eventueel een time-series analyse kunt doen ná opschoning, indien relevant (of alleen regressie blijft).

71.28 21/04

71.28.1 TO Do list:

1. Test Jina Embedding model (0.1 sample)

- Laad een 10% sample (frac=0.1) van de NOS-dataset.
- Genereer embeddings met Jina Embedding v3.
- Bereken de gemiddelde embedding per maand (of week).
- Koppel het aan de maanden van autoverkoop (zonder echte cijfers nog).

2. Correlatie-analyse (zonder verkoopcijfers!)

- Gebruik een dummyreeks van autoverkoop-indicatoren (bv. "hoge", "lage" verkoopperioden) op basis van publieke bronnen of seizoenspatronen.
- Bereken de Pearson-correlatiecoëfficiënt tussen sentiment (gemiddelde embedding score of topic score) en deze indicatoren.
- Documenteer dit als proof-of-concept voor correlatie zonder gevoelige data.

3. Literatuuronderzoek

- Zoek papers/artikelen over:
 - Economisch of consumentenvertrouwen vs. koopgedrag
 - Sentimentanalyse in econometrie of autosector
 - Voorbeeld: "Media sentiment and economic indicators"
- Noteer min. 3 bronnen (APA) voor je verslag.

Zoektermen in Google Scholar:

- "consumer sentiment economic behavior"
- "media sentiment car sales"
- "text analysis economic indicators"

4. Autoverkoopdata herbekijken

- Vraag de docent of je geaggregeerde of geanonimiseerde cijfers mag gebruiken (bijv. per maand, zonder exacte getallen).
- Alternatief: Zoek een open dataset van autoverkopen (bijv. via CBS of European Automobile Manufacturers Association).

5. Voorbereiding op gesprek

- Print of open notities: wat is je onderzoeksraag, wat wil je testen, waar loop je tegenaan?
- Leg uit hoe je omgaat met beperkingen van data (proxygebruik, indirecte analyse).
- Formuleer concrete vragen aan je docent:

- “Mag ik de verkoopdata benaderen via proxies?”
- “Is het oké als ik sentiment meet op maandniveau?”
- “Welke semantische analyse zou waardevol zijn volgens u?”

71.28.1.1 Idea: Adding CBS Consumer Sentiment Indicators as Complementary Features

Currently, the model relies on sentiment extracted from NOS news articles (2015–2025) to predict economic behavior. While this news sentiment captures real-time media tone and public discourse, it may not fully reflect how consumers themselves perceive their financial situation or the broader economy.

To complement the media-driven approach, I am considering integrating official consumer sentiment indicators from Statistics Netherlands (CBS). These include:

- *Consumer Confidence* (Consumentenvertrouwen_1)
- *Willingness to Buy* (Koopbereidheid_3)
- *Economic Climate* (EconomischKlimaat_2)

These indicators are based on monthly surveys of Dutch consumers and offer a structured and aggregated view of consumer expectations. By including them in the model, I aim to:

- Capture behavioral trends that may not be directly visible in media sentiment,
- Compare the explanatory power of public opinion vs. media tone, and
- Test whether combining both sources reduces prediction error (e.g., RMSE, MAE).

This hybrid approach could potentially enhance model robustness by blending bottom-up signals (public sentiment) with top-down signals (news sentiment).

In de oorspronkelijke versie van dit project werden EV-gerelateerde artikelen geïdentificeerd op basis van een vaste lijst met zoekwoorden, zoals “elektrische auto”, “Tesla” of “laadpaal”. Hoewel deze methode snel en simpel toepasbaar is, heeft het een aantal duidelijke beperkingen. De detectie is oppervlakkig, mist context, en herkent geen synoniemen of inhoud die semantisch verwant is aan het onderwerp, maar niet letterlijk de zoekwoorden bevat.

Om dit te verbeteren, is gekozen voor een geavanceerdere aanpak waarbij gebruik wordt gemaakt van de semantische kracht van de Jina Embeddings v3 encoder. In plaats van puur op trefwoorden te zoeken, wordt een representatieve zoekzin zoals:

"Dit artikel gaat over elektrische voertuigen of duurzaamheid in mobiliteit"

omgezet in een embedding (vectorrepresentatie). Deze zoekembedding wordt vervolgens vergeleken met de embeddings van alle artikelen met behulp van cosine similarity. Artikelen die qua betekenis dicht bij de zoekzin liggen, krijgen

Post meeting:

Making “simple regression model”

72 Meeting Marcio 17/04

Project Focus: Microeconomic Factors in Car Sales

Timeline & Planning

- Next 2 weeks:
 - Research microeconomic factors that correlate with car sales.
 - Use topic modeling and named entity recognition for article selection.
- Within 3 weeks:
 - Present first results.
 - Focus on methodology and results.

Article Search & Classification

Tools & Techniques

- Embeddings: Try [Jina Embeddings v3](#) instead of RobBERT for better semantic search and sentiment analysis.
- Clustering Approach:
 - Calculate cosine similarity between article embeddings.
 - Determine a threshold (e.g., 0.99 / 0.98) for inclusion in clusters.
 - Cluster articles based on similarity.
 - Identify the largest cluster related to your subject.
 - Automatically label clusters → generate a classifier based on them.

Methodological Considerations

Selection Criteria

- Need a clear method for article selection (justify the process).
- Use topic modeling and NER to filter relevant content.

Metrics

- For current car sales data:
 - Calculate RMSE and MSE to assess model performance.

Quality Expectations

- Define how the system should perform (precision, recall, interpretability).
- Must explain historical events that match graphical patterns in the data.

Data Handling

Met opmerkingen [DH6]: Hier dus nog mee wachten
met Stefan
(denk trouwens dat ik evt met collab al een heel eind kan
komen)

- Watch out for data leakage:
 - Don't show raw car sales data during model training/validation.
 - Ensure the model can't "cheat" by seeing target data in features.

Communication & Next Steps

- Let Marcio know if we can schedule weekly check-ins.
- Emphasize that the project involves:
 - The right steps (experimentation, methodology).
 - But also needs to make logical and practical sense overall.
- Be sure to clearly justify:
 - The methods and models used.
 - Why certain tools/thresholds were chosen.

72.1.1.1 Original notes:

<https://huggingface.co/jinaai/jina-embeddings-v3> use jina for article search and couple

what would be threshold by for example cosine 99/98

create cluster of articles

find big cluster that similar to subject

and then lable them (automatic) → print classifier based on that.

Research micro economic factor sin buying cars

→ UPcoming 2 weeks find factors that correlate with car sales.

Limited time → that's why

Clear method of the selection of articles →

Quality → how it should perform

Using topic modelling for article selection, named entity

Explain historical events based on the graph

DATA leakage → car sales (niet laten zien!)

➔ Gaat lasting worden om model reasoning

UPcoming 2 weeks find factors that correlate with car sales.

First now → what is rmse mse (regression van data dat ik nu heb)

Focus on methodology and results

Within 3 weeks results

Let marcio know if we can sit in weekly bases

Consumen vertrouwen

TO/DO

1. Sentimentanalyse

- Keyword search verbeteren → nu sementic search
- Keuze voor RobBERT onderbouwen → kappa score
- Literatuur over sentiment in nieuwsartikelen zoeken
- Twee modellen vergelijken met Kappa-score

2. CBS-data

- CBS-datasets documenteren (auto's, brandstofprijzen)

3. Lag time analyse

- Effect van 1-2 maanden vertraging in auto sales onderzoeken

4. Time series

- Eerste versie van time series analyse opzetten
- Nog geen parameter tuning uitvoeren

72.2 Update last feedback moment:

Moet ik nog doen

Ben er mee bezig

Is af

Algemeen / Overkoepelend

- Maak duidelijk dat correlatie geen causaliteit betekent, en benoem mogelijke externe factoren.
- Voeg reflectie toe over het feit dat autokoopbeslissingen niet direct reageren op nieuws, in tegenstelling tot bijvoorbeeld aandelen.
- Overweeg de toevoeging van andere databronnen naast NOS (bv. Rijksoverheid, autobladen, economische nieuwssites).
- Verfijn de methode van artikelselectie: overweeg geavanceerdere technieken zoals topic modelling of sentence embeddings in plaats van alleen keyword search.
- Voeg een stap toe waarin je verschillende sentimentmodellen en forecastingmodellen vergelijkt (minstens drie), inclusief voor- en nadelen.

Introductie

- Voeg betrouwbare bronvermelding toe aan claims over nieuwsimpact op consumentenvertrouwen en EV-aankopen.
- Leg uit waarom je NOS hebt gekozen als bron, en niet meerdere of andere media.
- Begin niet meteen met het noemen van XGBoost of RobBERT v2. Contextualiseer eerst het probleem.
- Noem in het kort bestaand onderzoek en leg uit hoe jouw werk hierop voortbouwt.

Probleemstelling en Onderzoeksdoel

- Maak de doelstellingen abstracter: formuleer ze op hoger niveau, zoals "onderzoeken welk sentimentmodel geschikt is".
- Verplaats de keuze voor modellen (RobBERT v2, XGBoost) naar de methodesectie.
- Maak de hoofdonderzoeksraag concreter: specificeer wat je bedoelt met "voertuigverkoop" (aantallen? omzet?).
- Verwijder specifieke evaluatiemetrics zoals RMSE uit de vraagstelling, zodat je die later kunt motiveren.

Onderzoeksvragen

- Verminder het aantal subvragen.
- Orden de subvragen logischer:
 1. Wat is sentimentanalyse? Welke categorieën zijn relevant?
 2. Wat zijn state-of-the-art methodes voor tijdreeksvoorspelling?
 3. Hoe is sentiment eerder ingezet in voorspellingen?
 4. Hoe effectief is het om sentiment toe te voegen aan voorspellingsmodellen?

Literatuurstudie

- Voeg een formele definitie van sentiment toe en bespreek relevante categorieën (optimisme, pessimisme, woede, etc.).
- Voeg een overzicht toe van literatuur over tijdreeksvoorspelling, niet alleen XGBoost.
- Onderbouw de claim dat de meeste studies gericht zijn op de VS en China met bronnen, en geef aan hoe jouw onderzoek hierin iets nieuws toevoegt.

Methodologie

- Leg uit waarom je RobBERT v2 gebruikt en overweeg of fine-tuning nodig is.
- Leg uit waarom je XGBoost kiest en niet alternatieven zoals LSTM, ARIMA, etc. → MET KAPPA
- Voeg een validatie-set toe aan je train/test split.
- Specificeer welke evaluatiemetrics je gebruikt en waarom.

Reflectie en Discussie

- Sta stil bij het feit dat andere, niet-gemeten variabelen mogelijk verklarender zijn.
- Reflecteer op het beperkte bereik van je data en modelkeuze.
- Overweeg om ook verkoopdata van fossiele auto's mee te nemen voor vergelijking. → ben nu bezig met nieuwe data exploratie CBS

72.3 15/04:

Auto verkoop CBS:

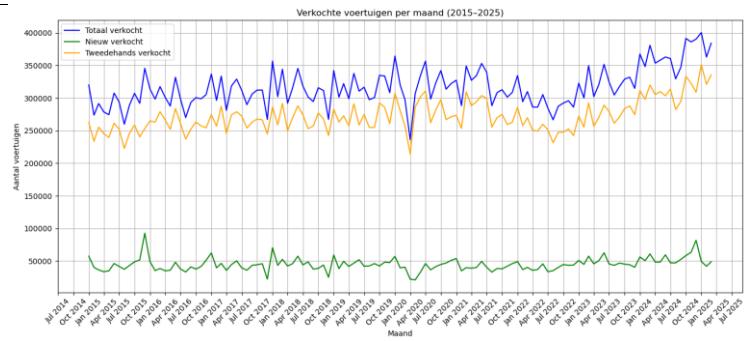


Figure 1 Verkochte voertuigen

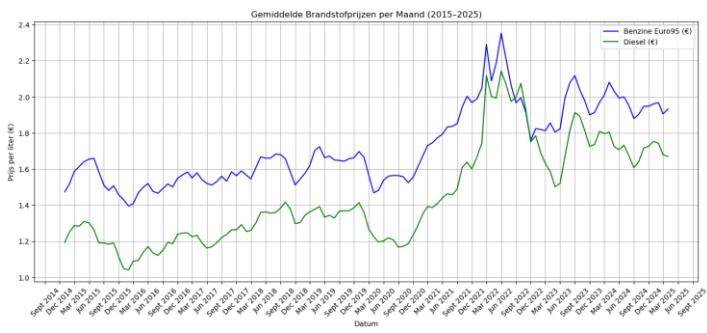


Figure 2 Brandstofprijzen

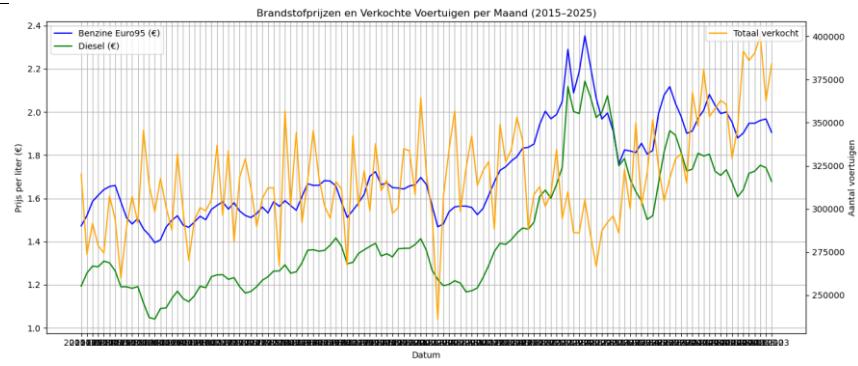


Figure 3 1&2 samen

Horizontaal tabel gemaakt:

Jaar Maa nd	BenzineE uro95_1_ first	BenzineE uro95_1_ last	BenzineE uro95_1_ min	BenzineE uro95_1_ max	BenzineEu ro95_1_m ean	Diese l_2_fi rst	Diesel _2_las t	Diese l_2_ min	Diese l_2_ max	Diesel_2_me an	TotaalVerkoch teWegvoertuig en_1	NieuwVerkoch teWegvoertuig en_2	TweedehandsVer kochteWegvoertui gen_3
201 5-01	1.501	1.46	1.458	1.503	14.730.64 5.161.290 .300	1.207	1.19	1.18 4	1.208	11.940.645.16 1.290.300	320045	57337	262708
201 5-02	1.461	1.557	1.461	1.557	15.201.42 8.571.428 .500	1.192	1.291	1.19 2	1.291	12.539.642.85 7.142.800	273654	40170	233484
201 5-03	1.559	1.601	1.559	1.603	15.861.61 2.903.225 .800	1.291	1.277	1.27 6	1.294	12.866.129.03 2.258.000	291533	36261	255272
201 5-04	1.599	1.646	1.592	1.647	16.144.66 6.666.666 .600	1.275	1.312	1.26 1	1.312	12.833.666.66 6.666.600	278536	33416	245120
201 5-05	1.645	1.639	1.634	1.648	16.408.06 4.516.129 .000	1.31	1.306	1.30 6	1.313	13.097.096.77 4.193.500	274390	34819	239571
201 5-06	1.638	1.653	1.638	1.668	16.549.33 3.333.333 .300	1.304	1.295	1.29 3	1.31	13.014.000.00 0.000.000	307521	46094	261427

Month	# cars sold (CBS)	pump (elec as well)				news count	sentiment				target values			
		start	end	max	min		highly neg	mildly neg	neutral	mildy pos	highly positive	# cars upcoming month	# cars upcoming 2 months	
Apr/24	500	1	1,5	1,7	0,8	1,3	30	2	1	10	15	2		

72.4 14/04:

Meeting with Marcio on 03/04

Key points discussed:

- Improve filtering based on keywords
- Define a target variable for multiple days
- Include electricity data in the analysis

Goal for today:

- Conduct data exploration
- Improve keyword research
- Perform a bit more literature review

Month	#cars sold (CBS)	pump (elec as well)			news count	sentiment	target values				# cars upcoming month	# cars upcoming 2 months	
		start	end	max	min		highly neg	mildly neg	neutral	mildy pos	highly positive		
Apr/24	500	1	1,5	1,7	0,8	1,3	30	2	1	10	15	2	2

72.5 03/04

72.5.1.1 Thesis Supervisor Meeting Agenda

72.5.1.2 1. Collected More Data

Update:

- Found better data sources from CBS:
- Lease pricing per quarter: Operational lease – Personenauto's
- Vehicle sales (filtered by new + passenger cars): Verkochte voertuigen
- Fuel prices (per day): Pompprijzen

Why this is better than RDW:

- Easier to filter for passenger cars
- Contains monthly frequency
- More accurate sales information

Question:

Are these CBS datasets sufficient and appropriate for a robust forecasting model, or should I consider additional sources (e.g., economic indicators, subsidies)?

72.5.1.2 Revised Research Question and Subquestions (Data-Driven)

Main Research Question:

How can news sentiment and macroeconomic indicators be used to forecast electric vehicle sales in the Netherlands?

Suggested Subquestions:

1. Which types of data have been used in literature to forecast EV or car sales (e.g., economic, sentiment, policy)?
2. What macroeconomic data sources are available in the Dutch context to inform car sales forecasting?
3. How can sentiment data (e.g., from news or social media) be extracted and quantified in a meaningful way?
4. How do different combinations of data sources affect the accuracy of car sales forecasting models?
5. Which forecasting model performs best given the chosen data sources (e.g., XGBoost, LSTM, Prophet)?

72.5.1.3 3. Data and Model Evaluation Strategy

Example Data Types and Their Evaluation Potential:

Data Type	Source	Signal Type	Expected Impact
Vehicle sales history	CBS	Quantitative, monthly	Baseline
Fuel prices	CBS	Quantitative, daily	Price sensitivity
Lease prices	CBS	Quantitative, quarterly	Affordability impact
Sentiment (NOS)	Media	Qualitative → score	Public mood
Economic news (FD)	Media	Qualitative → score	Policy influence

72.5.1.4 4. Questions for Supervisor

- What types of macroeconomic indicators are typically most useful in mobility forecasting?
- Would you recommend adding government policy variables (e.g., subsidy schedules)?
- How can I better justify the inclusion of sentiment as a signal for forecasting?
- Do you suggest beginning with a baseline (e.g. sales + fuel prices) and then gradually adding features like sentiment?
- Would it be reasonable to test different feature combinations across multiple models (e.g., XGBoost vs. LSTM) as part of the comparison?

Questions from me:

- how does the logbook look like

- In what way can I incorporate my A learning goals? By making a dashboard?
- Goal for upcoming weeks, incorporate feedback, finding better way sentiment analysis/keywords
- Not sure if gap is still sufficient.
- Research the (biggest) factors in the car market. (any tips?)
- Possibility to move our meetings to Thursday

72.6 29/03

Doel:

Data uitbreiden:

Zelf gevonden op CBS:

https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=85784NED&_theme=434

prijsontwikkeling per kwartaal van diensten van autoleasebedrijven, namelijk operationele lease van personenauto's, bestelauto's en privéauto's (private lease)

Verkochte autos: Verkochte wegvoertuigen; nieuw en tweedehands, voertuigsoort, brandstof

beter dan RDW want: het is sws verkochte autos, filtreren op nieuw en op personen autos en het is

per maand
<https://opendata.cbs.nl/statline/#/CBS/nl/dataset/85898NED/table?ts=1743271461148>

pompprijzen:

Pompprijzen motorbrandstoffen; brandstofsoort, per dag

https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=80416ned&_theme=434

72.7 26/03

Why RoBBERT v2 and not others?

Why XGBoost and not others?

Why only NOS?

What if you could not find any correlation

Perhaps because other unseen variables were not present in your correlation analysis

Perhaps the decision-making process when buying a car is not as instant as buying a stock.

Feedback thesis

(-) Introduction: Some strong sentences lack reputable citations. For example: "News coverage can impact public perception, potentially affecting consumer confidence and EV purchase decisions" - Based on which evidence?

(-) Introduction: It escalated too quickly to assign NOS as the sole media source. It is too early to converge. More specifically, why NOS and not other sources?

(-) Introduction: It escalated too quickly to specific ML models such as XGBoost. Again, it is too early to define the architecture and metrics. The introduction is more about contextualising, developing the problem statement, and developing research questions and project objectives.

(-) Introduction: The last paragraph mentions that the research extends prior research but does not clarify these works. Please ensure you, at this point, highlight at least some contributions in the field. Then, in a more extensive way, in the literature review section.

(-) Problem statement and research goal: Escalated too quickly to models such as RobBERT v2. Please review the objectives to make them more high-level. One of the objectives could be evaluating which sentiment model (pretrained or fine-tuned) would suit the problem better.

(-) Research question: Which variable you are referring to? What is the number of vehicles sold or their value (currency)? Please be more specific here. Also, specific metrics such as RMSE should be removed. It will give you more freedom to choose the most suitable metric in the methodology and substantiate your choice.

(-) Reduce the number of subquestions and create a natural ordering form them: e.g. a sub-question dedicated to sentiment analysis and how it works, and state-of-the-art (literature), a sub-question dedicated to forecasting using time series (literature), a sub-question dedicated to

exploring the usage of sentiment in forecasting (methodology), a sub-question dedicated to measuring the effectiveness of adding the sentiment in forecasting (results)

(-) Literature: I missed a contextualisation and a formal definition of sentiment, its categories, and what it means. More specifically, which sentiments, such as pessimism, optimism, anger, etc., are relevant to the problem of EV sales forecasting?

(-) Literature: There should be a review on forecasting using time series. Currently, the research plan is strongly focused on XGBoost.

(-) Literature: It says that "most sentiment-based forecasting models have focused on China and the US, leaving the Netherlands adoption patterns largely underexplored." Where are these references? These works can be very interesting, and many ideas could be extended to the Dutch market.

(-) Methodology:

(-) keyword-based search to identify electric car subjects is too fragile. I prefer using a bit more sophisticated techniques such as topic modelling or embeddings for similarity search.

(-) The choice of sentiment model was not substantiated. Why Robbert V2? Is it fine-tuning necessary?

(-) The choice of XGBoost was not substantiated.

(-) Fundamental error: you should reserve portions for validation, not only train/test.

(-) There are many metrics, but which will be used to make decisions?

Final observations:

In general, finding correlations between news and sales seems interesting. At the same time, this phenomenon can be much more complex and involve variables that are not within your scope. Thus, correlation does not mean causality. Also, there is only a single type of source (NOS) and no others, e.g. government communications, specialised magazines on cars, economy news describing issues with fossil fuels or discussing the price of energy, etc. Also, consider using data about fossil fuel cars and analysing if there is a trend concerning the increase or decrease in sales.

The way articles are selected seems too fragile, based on keywords. It appears that the market of electric cars is more complex and depends on other factors such as the price of electricity, subsidies from the government, exceptions or reductions of tax for the adoption of electric cars, and global economic tensions (e.g. Trump, Musk and the repulsion to buy Tesla cars and so on)

A step should also be added to evaluate the most suitable sentiment analysis and forecasting model. You should select at least three alternatives and systematically compare their pros and cons.

72.8 13/03

Todlist

PVA:

Planning maken.

1. Data Selectie & Verwerking

- Bepalen welke NOS-data wordt gebruikt:
 - Sentimentanalyse op alle artikelen?
 - Alleen artikelen over EV's?
 - EV-gerelateerde thema's meenemen (klimaatverandering, duurzaamheid, belangrijke personen zoals Elon Musk)?
- Checken hoeveel artikelen er zijn over EV's en of dat voldoende is.
- Keuze onderbouwen en opnemen in het PvA.

72.8.1 12/03 feedback daphne op concept

To-Do List voor het PvA en Projectplan

1. Data Selectie & Verwerking

- Bepalen welke NOS-data wordt gebruikt:
 - Sentimentanalyse op alle artikelen?
 - Alleen artikelen over EV's?
 - EV-gerelateerde thema's meenemen (klimaatverandering, duurzaamheid, belangrijke personen zoals Elon Musk)?
- Checken hoeveel artikelen er zijn over EV's en of dat voldoende is.
- Keuze onderbouwen en opnemen in het PvA.

2. Planning Verduidelijken

- Detailleren wat wordt bedoeld met "70%-versie van project":
 - Wat valt hieronder? (bijv. data preprocessing, model bouwen, experimenten, resultaten analyseren, etc.)
- Specifieker maken wat elke fase inhoudt.
- Toevoegen aan het PvA.

3. Besprekking Voorbereiden

- Overzicht maken van de vragen die nog openstaan.
- Eventuele andere knelpunten opschriften om morgen te bespreken.

Kinski toevoegen bron

Gap schrijven

Bronnen ff goed doorlezen

Volgende keer

- Leg uit wat je gaat doen in 'Jip-en-Janneketaal'
- Lever conceptversie van Onderzoeksplan in
 - Vraagstelling
 - Literatuur
 - Methodiek



73Jip en janeke taal:

Onderzoeksplan – Impact van Online Sentiment op de Verkoop van Elektrische Auto's

Inleiding

Elektrische auto's worden steeds populairder. Veel mensen vinden ze goed voor het milieu, maar er zijn ook zorgen, zoals de actieradius en laadinfrastructuur. Dit onderzoek gaat over hoe mensen online praten over elektrische auto's en of dit invloed heeft op de verkoopcijfers.

Probleembeschrijving & Vraagstelling

Op sociale media en autowebsites geven mensen hun mening over elektrische auto's. Soms zijn ze positief, soms negatief. Maar heeft dat invloed op de verkoop?

Hoofdvraag:

Hoe beïnvloedt online sentiment de verkoop van elektrische auto's en kan dit helpen bij betere verkoopvoorspellingen?

Deelvragen:

1. Wat zeggen mensen online/op twitter over elektrische auto's?
2. Is er een verband tussen positieve of negatieve meningen en verkoopcijfers?
3. Welke andere factoren (zoals prijs en subsidies) spelen een rol?
4. Kan sentimentanalyse helpen om verkoopcijfers beter te voorspellen?

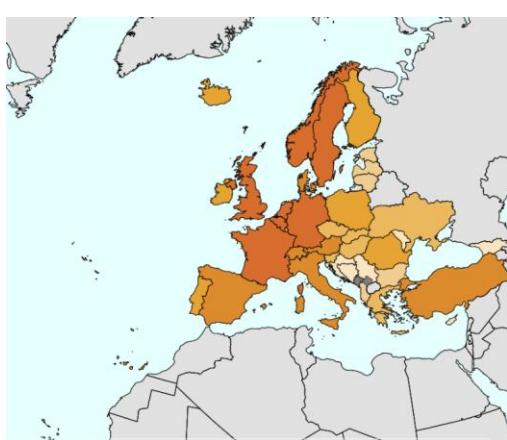
Literatuuronderzoek

Eerdere onderzoeken laten zien dat online sentimenten kunnen helpen bij het voorspellen van trends in verschillende markten.

- Chen (2024) ontdekte dat autobeoordelingen op Autohome de verkoopvoorspelling met 54% verbeterden.
- Kinski (2016) vond dat zoekopdrachten op Google Trends tot 68,5% nauwkeurigheid leiden bij autoverkoopvoorspellingen.
- Liu et al. (2024) lieten zien dat de aanwezigheid van laadstations een grote rol speelt in de aankoopsbeslissing van elektrische auto's.
- Sharma et al. (2024) toonden aan dat sentimentanalyse met AI-modellen (zoals BERT) effectief is voor het begrijpen van consumentengedrag.

Methodiek (Hoe gaan we dit onderzoeken?)

1. Data verzamelen
 - Verkoopcijfers van EV's → Gegevens van RDW
 - Online meningen → Twitter, Reddit, Autohome (reviews en social media)
 - Zoekgedrag → Google Trends-data
2. Data analyseren
 - Sentimentanalyse uitvoeren met AI-tools om te zien of berichten positief of negatief zijn.
 - Correlatie berekenen tussen sentiment en verkoopcijfers (bijv. met Spearman/Pearson).
 - Testen of het toevoegen van sentimentgegevens aan verkoopvoorspellingsmodellen (zoals XGBoost) de nauwkeurigheid verbetert.
3. Conclusies trekken
 - Heeft online sentiment invloed op de verkoop?
 - Kunnen we hiermee betere verkoopvoorspellingen maken?
 - Wat betekent dit voor fabrikanten en beleidsmakers?



Nieuw plan afstuderen:

Sentiment analyse op de verkoop van ev

Electric field door

Impactanalyse van sentiment op EV-verkoop

Doel:

1. Onderzoeken in hoeverre online sentimenten over elektrische voertuigen correleren met verkoopcijfers en marktdynamiek.
2. How socialmedia sentement can improve sales forecasting

Data:

Car sales → RDW

Autohome (carreview data)

- Waarom? Dit kan helpen om te begrijpen of positieve of negatieve publieke percepties invloed hebben op de markt.
- Methode: Vergelijk sentimentdata (bijv. sociale media, nieuwsartikelen) met verkoopstatistieken en bepaal correlaties.
- Mogelijke conclusie: Sentiment kan een rol spelen bij marktacceptatie, maar andere factoren (zoals prijzen en infrastructuur) zijn dominanter.

Titel	Doel	Gebruikte Data	Resultaten
Forecasting China's Car Sales	Verbeteren van autoverkoopvoorspellingen in China door multivariate analyse en sociale media data.	Historische verkoopcijfers, macro-economische indicatoren, gebruikersreviews, Weibo-sociale media data.	Gebruikersreviews verbeterden de nauwkeurigheid van voorspellingen met 54,42%; sociale media data zorgden voor een verbetering van 10,58%.
Google Trends as Complementary Tool for New Car Sales	Gebruik van Google Trends-gegevens om voorspellingen van nieuwe autoverkoop in verschillende landen te verbeteren.	Google-zoekopdrachten, autoverkoopgegevens, cross-country vergelijking.	Google zoekdata correleren sterk met autoverkoop; het toevoegen van een tijdsvertraging verbeterde de nauwkeurigheid van voorspellingen.
Electric Cars: Perception and Knowledge	Begrijpen van publieke perceptie en kennis over elektrische voertuigen.	Enquêtegegevens over publieke perceptie en kennis van EV's.	Publieke perceptie wordt beïnvloed door actieradius, laadinfrastructuur en kosten; kennisgebrek beperkt adoptie.
Evaluating the Market for EV Using TAM	Onderzoek naar de adoptie van elektrische voertuigen met het Technology Acceptance Model (TAM).	Enquête met 695 respondenten over consumentenhouding ten opzichte van EV's.	Consumentenattitudes en waargenomen bruikbaarheid zijn de belangrijkste drijfveren voor EV-adoptie.
The Impact of Consumer Sentiment on Sales of New Energy Vehicles	Onderzoeken van de impact van consumentensentiment op de verkoop van nieuwe energievoertuigen met tekstuele analyse.	Provinciale paneldata (2017-2022), machine learning-gebaseerde tekstuele analyse.	Consumentensentiment correleert positief met NEV-verkoop; infrastructuur voor opladen vermindert actieradiusangst.
Electric Vehicle Sentiment Analysis Using Large Language Models	Toepassen van grote taalmodellen (LLMs) op sentimentanalyse van EV-merken.	YouTube-commentaren op Lucid Motors en Tesla, sentimentclassificatie met BERT, RoBERTa, XLNet.	RoBERTa behaalde de hoogste sentimentclassificatie nauwkeurigheid met een F1-score van 92%.
Employing Sentimental Analysis to Prove Correlation Between Social Media and Stock Market	Onderzoeken van de correlatie tussen sociale media-sentiment en EV-marktprestaties.	Tweets, nieuwsartikelen, EV-aandelenprijzen, EV-marktaandelen, luchtkwaliteitsgegevens.	Positief sentiment op sociale media correleert sterk met EV-aandelenprijzen en marktprestaties.
Exploring the Sentiment Analysis of Electric Vehicles Social Media	Gebruik van functieselectiemethoden om sentimentanalyse van EV-gerelateerde sociale media-data te verbeteren.	Sociale mediaberichten, functieselectie voor sentimentclassificatie verbetering.	Functieselectiemethoden verbeterden de sentimentclassificatie, wat leidde tot betere inzichten in de EV-markt.

73.1 Feedback Dinsdag 25 februari:

1. Relatie tussen sentiment en beursindex verduidelijken

- Zorg voor een duidelijke koppeling tussen sentimentanalyse (bijvoorbeeld via Twitter of nieuws) en de AEX-index.
- In eerder onderzoek werd een vergelijkbare studie(Visser, n.d.) gedaan met Tesla en Twitter, waarbij sentiment invloed had op het aandeel van de eigenaar. Overweeg een soortgelijke aanpak voor AEX-bedrijven.

2. Data-uitdagingen

- Niet alle datasets zijn recent: Sommige historische data zijn mogelijk beperkt beschikbaar of moeilijk te verkrijgen.
- Scraping kan illegaal zijn: Check de juridische en ethische implicaties van webscraping voor nieuwsartikelen en sociale media.
- Marketing-gebruik van scraping: Sommige platforms beperken scraping om misbruik te voorkomen, dus overweeg alternatieve methodes zoals API's.

3. Onderzoek Centre Market AI en alternatieven

- Zoek informatie over Centre Market AI, dat mogelijk tools of methodes biedt voor het verwerken van marktinformatie via AI.
- Kijk naar bestaande oplossingen voor webscraping en API-gebaseerde dataverzameling.

4. Mogelijke uitbreiding: AI-model voor nieuwsverwerking

- Onderzoek of een Large Language Model (LLM) nieuwsartikelen kan verwerken en betekenisvolle indicaties kan genereren.
- Overweeg of je een applicatie kunt bouwen die historische nieuwsdata analyseert om patronen te ontdekken en te voorspellen.

5. Toekomstige richting

- Duidelijker doel formuleren: Zorg ervoor dat jouw onderzoek een heldere focus heeft, bijvoorbeeld:
 - "Kan sentimentanalyse van Nederlandse nieuwsartikelen helpen bij het voorspellen van de AEX?"
- Praktische toepassing: Ontwikkel een tool of app die historische data analyseert en betekenisvolle inzichten biedt.
- Onderzoek hoe sentimentgegevens de invloed van nieuws op aandelenkoersen beter zichtbaar kunnen maken.

Op basis van de bovenstaande punten ga ik:

Literatuur van Micheal visser (Visser, n.d.) doornemen, en kijken wat hij gebruikt heeft voor model, data collectie. 1 aandeel vinden en daar in specialiseren.

Onderzoek

Doel:

Het doel van de paper is om te onderzoeken hoe social media sentiment, met een onderscheid tussen sentiment polarity en het type content (feit of mening), invloed heeft op aandelenprijsbewegingen Van tesla en Validatie op Taiwan Semiconductor Manufacturing Company (TSM).

Probleem:

Social media posts reflecteert aandelen sentiment, ligt de uitdaging bij het verdelen van posts die wel relevant zijn voor prijs veranderingen en post die gewoon ruis zijn. Daarnaast is door de komst van jongeren generatie op hun financieel landschap een grotere social gedreven inversertings climaat. In het verleden hebben trends en hypotheses implicaties gehad op aandelen (game stock).

Om deze reden is het noodzakelijk om socialmedia sentiment te gebruiken om de impact op de financiële markt te voorspellen.

Huidige tool (opzoeken welke) falen en sentiment analysis focussen zich normaal gesproken op service level emotional tone. Probleem hierbij is dat niet alle posts dezelfde impact hebben.

Existing work

Fear en greed index:

Gebaseert op crypto markt index is ontworpen om de emotie te meten. Bij grote angst (excessive fear) gaan de prijzen zakken en extreme hebzucht (extreme greed) kan leiden tot prijs stijgingen.

Er wordt een cijfer gegeven 0-100 (extreme greed) en komt tot stand uit:

- Volatiliteit: De mate van prijsfluctuaties.
- Marktmomentum/volume: Het handelsvolume en de marktdynamiek.
- Sociale media: Het sentiment van sociale mediaplatforms, voornamelijk X (voorheen Twitter).
- Enquêtes: De resultaten van beleggerssentiment-enquêtes.
- Bitcoin-dominantie: Het marktaandeel van Bitcoin ten opzichte van de totale cryptomarkt.
- Google-trends: Zoektrends en trefwoorden gerelateerd aan cryptocurrency.

De scope van dit onderzoek richt zich enkel op sentiment → Zou interessant zijn voor mijn onderzoek om andere aspecten mee te nemen.

Thomas reuters market psych indices: analyse sentiment across various media sources (news articles, social media, and financial report)

Voor mijn onderzoek kan ik hetzelfde argument gebruiken, over dat bestaande tools niet transparant zijn()

Gap:

Huidige sentiment analyse ligt de focus enkel op general sentiment (positief, neutraal, negatief) zonder dat er onderscheid wordt gemaakt op type informatie.

Bestaande modellen nemen de impact van berichten niet mee

Doel! Vergroten van de nauwkeurigheid van aandelen voorspellingen gehaald uit sentiment van social media.

Rc: Hoe kan sentimentanalyse, waarbij zowel meningen als feitelijke inhoud worden meegenomen, worden gebruikt om veranderingen in prijstrends van een specifiek aandeel te voorspellen?

Dataset:

Twitter is gebruikt vanwege groot gebruik van investerrders, korte berichten,

3 datasets zijn gecombineerd naar 1 grote

combi bestaande uit:

1. [kaggle](#) bestaande uit:
 - a. twitter
 - b. Yahoo finance → voor mijn onderzoek kan ik evt investor.com gebruiken
2. Fever
 - a. Support claim verification compreses 185000 claims

Modellen:

1. Pretrained model gemaakt voor sentiment analyse
2. Zelf gemaakt model om te checken of tweet opinie was of feitelijk.

Methode:

1. Bekijken dataset
2. Preprocessen
 - a. Data splitten
 - b. Distributie bekijken → som van linked tweets berekenen voor elke ticker (42% was voor tesla)
 - c. Duplicaten verwijderen, Missende waardes identificeren → beide waren er niet
 - d. Train/validatie/test set maken
3. Sentiment classification toepassen op dataset 1 (twitter)

- a. Stap 1 Links verwijderen (verwijzen vaak naar foto/ external context daarom niet meegenomen)
 - b. Alle tekst naar lowercase (zodat alle woorden als hetzelfde worden gezien)
 - c. Mentions weghalen "@tesla " wordt gezien als tesla
 - d. User name is gewisseld voor place holder
 - e. Can't wordt can not
 - f.
4. Fever gebruikt om feiten te splitsen van meningen
5. Integreren dataset 3 (yahoo finance) in dataset 1
6. Emojis worden wel mee genomen

Methode:

De studie analyseert de relatie tussen social media sentiment en aandelenprijsbewegingen met een hybride aanpak:

1. Data verzamelen:

- o Tweets over aandelen (voornamelijk TSLA en TSM) uit een Kaggle-dataset.
- o Historische aandelenprijzen van Yahoo Finance.
- o FEVER-dataset voor feit/opinie-classificatie.

Met opmerkingen [DH7]: In de dataset had TSLA **het grootste aandeel** (46.3% van de tweets, ±37k tweets). **TSM had de op één na meeste tweets** (13.6%, ±11k tweets), waardoor het een logische tweede keuze was.

2. Sentimentanalyse:

- o [FinBERT-model](#) bepaalt de sentiment polarity (positief, negatief, neutraal) van tweets.
- o Een op [FEVER](#) getraind model classificeert tweets als feit of mening.

3. Data-integratie & visualisatie:

- o Sentimentscores per dag worden berekend en gekoppeld aan de aandelenprijs.
- o Een 21-daags Moving Average wordt toegepast om ruis te verminderen.
- o Correlaties tussen sentiment en aandelenprijs worden geanalyseerd, zowel met als zonder filtering van meningen.

4. Validatie:

- o De methode wordt getest op TSLA en vervolgens op TSM om de generaliseerbaarheid te beoordelen.

Doel:

Financieel vraagstuk oplossen

Met zelfgemaakt algoritme

Aantonen waarom mijn aanpak zinnig is, wat voegt het toe, presteert het beter?

- Onderwerp:
 - Stockmarket (AEX) For casting

Wat is het probleem?

Europese/nederlandse indices zoals AEX zijn onvoldoende onderzocht op dl technieken.

Vraagstelling

"Hoe kan ik de **dagelijkse sluitingsprijs** van de AEX voorspellen met een "**hogere nauwkeurigheid**" dan bestaande modellen?"

Welke modellen zijn er gebruikt, wat is de nauwkeurigheid?

Modellen

Data collection:

1. Stock market Forecasting utilizing News Sentiment Data:(Oriol, n.d.)
 - a. Gap
 - i. The gap identified for this thesis was the lack of comprehensive comparative studies specifically addressing how news sentiment data can enhance stock market forecasting models. Although previous literature has extensively studied the use of newspaper sentiment in stock market predictions, this thesis does not aim at introducing entirely new concepts but rather at performing an in-depth comparative analysis of existing models.
 - ii. Specifically, the thesis addresses whether incorporating news sentiment data into predictive models significantly improves the accuracy of predicting stock price movements (both next-day returns and directional classification). The research gap lies in systematically comparing models trained solely on financial data versus models trained with both financial and news sentiment data, thereby providing clearer insights into the practical value and limitations of sentiment analysis for stock forecasting
 - b. Data collection:
 - i. The study narrows its scope by selecting a subset of seven representative stocks from the S&P 500, facilitating targeted and manageable analysis.
 - ii. The chosen timeframe (January 1, 2019, to December 31, 2020) strategically aligns with the COVID-19 pandemic's onset, emphasizing the challenge of predicting market behavior during highly volatile periods.
 - iii. Models use training data up to March 1, 2020, testing performance afterward, highlighting the difficulties traditional financial indicators face under crisis conditions.
 - iv. The forecasting horizon is deliberately restricted to one day ahead, enabling concise, immediate performance evaluation.
 - v. Selected stocks represent diverse market sectors and volatility levels, balancing financial visibility in news coverage with market representativeness
 - c. Future work
 - i. Utilization of alternative sentiment sources (social media or analyst reports)
 - ii. Incorporating microeconomic indicators

<https://hasdata.com/blog/web-scraping-google-news>

Met opmerkingen [DH8]: Interessant! Want social media was al gedaan vlgm

Onderzoek	Gebruikte Methodes	Modellen	Gaps / Beperkingen	Bron van data	Verwijzing naar onderzoek
Towards Autonomous Investment Analysts	Multi-agent systemen, technische analyse, fundamentele analyse	Adaptieve multi-agent architectuur	Moeilijkheden met menselijke aspecten zoals intuïtie en gezond verstand	Niet specifiek vermeld, algemeen financieel en marktdata	Castro & Junior (2023) [9]
Stock Market Forecasting Using Machine Learning	Tijdreeksanalyse (ARIMA), sentimentanalyse, lineaire regressie, machine learning	ARIMA-modellen, ordinale data regressie, sentimentanalyse-modellen	Beperkt door lineariteit van modellen en beperkte integratie van niet-financiële gegevens	S&P 500 index historische data	Hirulkar et al. (2017) [10]
Enhancing Stock Market Forecasting Through Microservice System	Microservices architectuur, sentimentanalyse, technische analyse, machine learning (LSTM)	Microservices geïntegreerd met LSTM, RSI, MACD indicatoren	Complexiteit van integratie van meerdere microservices en real-time dataverwerkingsuitdagingen	Saudi Aramco en financiële rapporten, sentimentdata	Algarni (2025) [11]
Enhanced Stock Market Forecasting Using Dandelion Optimization Algorithm (DOA), 3D-CNN-GRU, Blood Coagulation Algorithm (BCA)	Wavelet transformatie, Dandelion Optimization Algorithm (DOA), 3D-CNN-GRU, Blood Coagulation Algorithm (BCA)	Hybride model: 3D-CNN gecombineerd met GRU, optimalisatie via DOA en BCA	Dataset beperkt tot Nifty 50 index; verdere validatie nodig op bredere datasets en verschillende marktscenario's	Nifty 50 index data met technische indicatoren	Jagadesh et al. (2024) [12]

Onderzoek	Methode	Nulmodel	Resultaat	Features	Train/Test Set
Towards Autonomous Investment Analysts	Multi-agent systemen, technische en fundamentele analyse	Geen expliciet nulmodel vermeld	Efficiënte analyse en gepersonaliseerde investeringsadviezen	Technische indicatoren, economische fundamenten, risicovoorkieuren beleggers	Niet specifiek vermeld
Stock Market Forecasting Using Machine Learning	ARIMA tijdreeksmodellen, sentimentanalyse, regressiemodellen	Eenvoudige historische gemiddelden	Effectieve voorspellingen van S&P 500 prijsbewegingen	Historische prijsgegevens, financiële ratio's, ordinale trends	Niet expliciet vermeld, waarschijnlijk historische splitsing
Enhancing Stock Market Forecasting Through Microservice System	Microservices met LSTM, sentimentanalyse, technische indicatoren	Standaard monolithisch voorspellingsmodel	Sterke voorspelling (LSTM) met MAE 0.26, MSE 0.18	Technische indicatoren (RSI, MACD), sentiment uit financiële rapporten	Niet explicet vermeld, waarschijnlijk historische splitsing
Enhanced Stock Market Forecasting Using Dandelion Optimization-driven 3D-CNN-GRU	Wavelet transformatie, 3D-CNN-GRU hybride model, DOA en BCA optimalisatie	Niet explicet vermeld, mogelijk standaard CNN of GRU modellen	Uitzonderlijke nauwkeurigheid (99,14%) met hybride 3D-CNN-GRU	Technische indicatoren (momentum, trend, volatiliteit), geoptimaliseerde features via DOA	Nifty 50 index historische data, splitsing niet specifiek vermeld