Predicting delays in shipping containers

CARGOPROBE

UNIVERSITEIT VAN AMSTERDAM

By:
David Werkhoven, Jacomijn Prins,
Paul Koole & Sem Kjaer

- Supply chain: shipping
- Biedt inzicht en overzicht
- Schat een aankomsttijd

Team 4

Probleemstelling

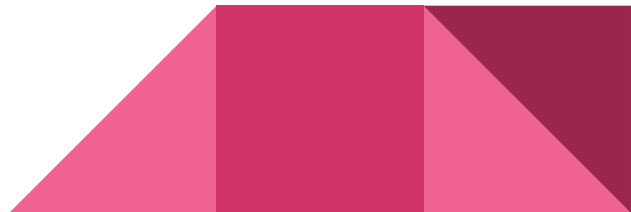- Internationaal probleem: veel schepen komen te laat
- CargoProbe wil graag nog beter voorspellen [enter AI]
- Wij kijken naar de "delay" van de ETA
  - Het verschil tussen de ETA en ATA

```
0                  update_datetime
1                  containernumber
2                date_last_updated
3                       date_added
4                       vesselname
5                        vesseleta
6                        vesselata
7                           status
8                      origin_city
9                   origin_country
10                destination_city
11             destination_country
12                      stepnumber
13                        activity
14                 activity_source
15                   activity_city
16                activity_country
17            activity_expectedtime
18              activity_actualtime
19         portofunloading_country
20            portofunloading_city
21    portofunloading_expectedtime
22      portofunloading_actualtime
```

De geleverde dataset

- 23 kolommen
- 331956 entries
  - waarvan veel dubbele informatie

	update_datetime	date_added	origin_city	destination_city	stepnumber	activity	portofunloading_expectedtime	portofunloading_actualtime
254041	2022-10-09 02:24:55.537	2022-09-07T22:19:32.886Z	9d59efbe-ae59-4538-8ff1-bf4f26d27ea4	5e2bed57-a475-4aea-a9ae-a191a7bc25c8	1	Gate-Out-Empty	2022-10-07T18:00:00-05:00	2022-10-07T21:21:00-05:00
254042	2022-10-09 02:24:55.537	2022-09-07T22:19:32.886Z	9d59efbe-ae59-4538-8ff1-bf4f26d27ea4	5e2bed57-a475-4aea-a9ae-a191a7bc25c8	2	Gate-in-Full	2022-10-07T18:00:00-05:00	2022-10-07T21:21:00-05:00
254043	2022-10-09 02:24:55.537	2022-09-07T22:19:32.886Z	9d59efbe-ae59-4538-8ff1-bf4f26d27ea4	5e2bed57-a475-4aea-a9ae-a191a7bc25c8	3	Vessel-Load	2022-10-07T18:00:00-05:00	2022-10-07T21:21:00-05:00
254044	2022-10-09 02:24:55.537	2022-09-07T22:19:32.886Z	9d59efbe-ae59-4538-8ff1-bf4f26d27ea4	5e2bed57-a475-4aea-a9ae-a191a7bc25c8	4	Vessel-Discharge	2022-10-07T18:00:00-05:00	2022-10-07T21:21:00-05:00
254045	2022-10-09 02:24:55.537	2022-09-07T22:19:32.886Z	9d59efbe-ae59-4538-8ff1-bf4f26d27ea4	5e2bed57-a475-4aea-a9ae-a191a7bc25c8	5	Gate-Out-Full	2022-10-07T18:00:00-05:00	2022-10-07T21:21:00-05:00
254046	2022-10-14 03:53:56.288	2022-09-07T22:19:32.886Z	9d59efbe-ae59-4538-8ff1-bf4f26d27ea4	5e2bed57-a475-4aea-a9ae-a191a7bc25c8	1	Gate-Out-Empty	2022-10-07T18:00:00-05:00	2022-10-07T21:21:00-05:00
254047	2022-10-14 03:53:56.288	2022-09-07T22:19:32.886Z	9d59efbe-ae59-4538-8ff1-bf4f26d27ea4	5e2bed57-a475-4aea-a9ae-a191a7bc25c8	2	Gate-in-Full	2022-10-07T18:00:00-05:00	2022-10-07T21:21:00-05:00
254048	2022-10-14 03:53:56.288	2022-09-07T22:19:32.886Z	9d59efbe-ae59-4538-8ff1-bf4f26d27ea4	5e2bed57-a475-4aea-a9ae-a191a7bc25c8	3	Vessel-Load	2022-10-07T18:00:00-05:00	2022-10-07T21:21:00-05:00
254049	2022-10-14 03:53:56.288	2022-09-07T22:19:32.886Z	9d59efbe-ae59-4538-8ff1-bf4f26d27ea4	5e2bed57-a475-4aea-a9ae-a191a7bc25c8	4	Vessel-Discharge	2022-10-07T18:00:00-05:00	2022-10-07T21:21:00-05:00
254050	2022-10-14 03:53:56.288	2022-09-07T22:19:32.886Z	9d59efbe-ae59-4538-8ff1-bf4f26d27ea4	5e2bed57-a475-4aea-a9ae-a191a7bc25c8	5	Gate-Out-Full	2022-10-07T18:00:00-05:00	2022-10-07T21:21:00-05:00

Data preprocessing

Missing values

- Portofunloading_actualtime aangevuld
- Drop van niet aangekomen trips
- Aanvullen portofunloading_city
- Aanvullen portofunloading_country
- Drop van rows zonder step number
- Activity_country aangevuld
- Date_last_updated aangevuld
- Vesselname aangevuld

Rows: 331956
Columns: 23

	na_amount	na_percentage
update_datetime	0	0
containernumber	0	0
date_last_updated	3793	1
date_added	0	0
vesselname	507	0
vesseleta	106582	32
vesselata	279299	84
status	0	0
origin_city	0	0
origin_country	0	0
destination_city	0	0
destination_country	0	0
stepnumber	2112	1
activity	14	0
activity_source	0	0
activity_city	14	0
activity_country	5129	2
activity_expectedtime	129316	39
activity_actualtime	104204	31
portofunloading_country	324260	98
portofunloading_city	9165	3
portofunloading_expectedtime	39619	12
portofunloading_actualtime	277236	84

Rows: 162544
Columns: 23

	na_amount	na_percentage
update_datetime	0	0
containernumber	0	0
date_last_updated	0	0
date_added	0	0
vesselname	0	0
vesseleta	62975	39
vesselata	113022	70
status	0	0
origin_city	0	0
origin_country	0	0
destination_city	0	0
destination_country	0	0
stepnumber	0	0
activity	0	0
activity_source	0	0
activity_city	0	0
activity_country	0	0
activity_expectedtime	51241	32
activity_actualtime	44382	27
portofunloading_country	0	0
portofunloading_city	0	0
portofunloading_expectedtime	25271	16
portofunloading_actualtime	0	0

Updates inkorten

- Updates na arrival droppen
- Duplicate updates droppen
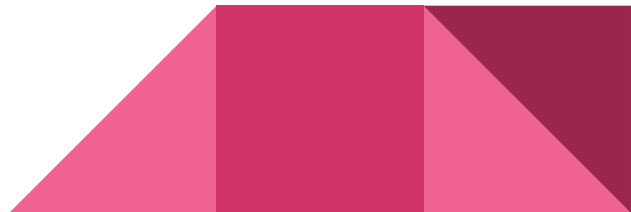- Één row per update behouden

Updates inkorten

	update_datetime	date_added	stepnumber	activity	activity_source	activity_actualtime	activity_expectedtime	portofunloading_expectedtime	portofunloading_actualtime
0	2022-09-08 03:48:53.483	2022-09-08 02:58:17.815	1	gate-out-empty	Maersk	2022-09-06 05:34:00	NaT	2022-10-23 14:00:00	2022-10-25 21:45:00
1	2022-09-08 03:48:53.483	2022-09-08 02:58:17.815	2	gate-in-full	Maersk	2022-09-06 08:26:00	2022-09-12 08:00:00	2022-10-23 14:00:00	2022-10-25 21:45:00
2	2022-09-08 03:48:53.483	2022-09-08 02:58:17.815	3	vessel-load	Maersk	NaT	2022-09-12 08:00:00	2022-10-23 14:00:00	2022-10-25 21:45:00
3	2022-09-08 03:48:53.483	2022-09-08 02:58:17.815	4	vessel-discharge	Maersk	NaT	2022-10-23 14:00:00	2022-10-23 14:00:00	2022-10-25 21:45:00
4	2022-09-08 03:48:53.483	2022-09-08 02:58:17.815	5	gate-out-full	Maersk	NaT	2022-10-23 14:00:00	2022-10-23 14:00:00	2022-10-25 21:45:00
5	2022-09-10 02:52:53.339	2022-09-08 02:58:17.815	1	gate-out-empty	Maersk	2022-09-06 05:34:00	NaT	2022-10-21 12:00:00	2022-10-25 21:45:00
6	2022-09-10 02:52:53.339	2022-09-08 02:58:17.815	2	gate-in-full	Maersk	2022-09-06 08:26:00	2022-09-12 08:00:00	2022-10-21 12:00:00	2022-10-25 21:45:00
7	2022-09-10 02:52:53.339	2022-09-08 02:58:17.815	3	vessel-load	Maersk	NaT	2022-09-12 08:00:00	2022-10-21 12:00:00	2022-10-25 21:45:00
8	2022-09-10 02:52:53.339	2022-09-08 02:58:17.815	4	vessel-discharge	Maersk	NaT	2022-10-21 12:00:00	2022-10-21 12:00:00	2022-10-25 21:45:00
9	2022-09-10 02:52:53.339	2022-09-08 02:58:17.815	5	gate-out-full	Maersk	NaT	2022-10-21 12:00:00	2022-10-21 12:00:00	2022-10-25 21:45:00
10	2022-09-11 02:27:53.827	2022-09-08 02:58:17.815	1	gate-out-empty	Maersk	2022-09-06 05:34:00	NaT	2022-10-21 12:00:00	2022-10-25 21:45:00
11	2022-09-11 02:27:53.827	2022-09-08 02:58:17.815	2	gate-in-full	Maersk	2022-09-06 08:26:00	2022-09-12 12:00:00	2022-10-21 12:00:00	2022-10-25 21:45:00
12	2022-09-11 02:27:53.827	2022-09-08 02:58:17.815	3	vessel-load	Maersk	NaT	2022-09-12 12:00:00	2022-10-21 12:00:00	2022-10-25 21:45:00
13	2022-09-11 02:27:53.827	2022-09-08 02:58:17.815	4	vessel-discharge	Maersk	NaT	2022-10-21 12:00:00	2022-10-21 12:00:00	2022-10-25 21:45:00
14	2022-09-11 02:27:53.827	2022-09-08 02:58:17.815	5	gate-out-full	Maersk	NaT	2022-10-21 12:00:00	2022-10-21 12:00:00	2022-10-25 21:45:00
15	2022-09-12 01:49:53.475	2022-09-08 02:58:17.815	1	gate-out-empty	Maersk	2022-09-06 05:34:00	NaT	2022-10-21 12:00:00	2022-10-25 21:45:00
16	2022-09-12 01:49:53.475	2022-09-08 02:58:17.815	2	gate-in-full	Maersk	2022-09-06 08:26:00	2022-09-12 07:00:00	2022-10-21 12:00:00	2022-10-25 21:45:00
17	2022-09-12 01:49:53.475	2022-09-08 02:58:17.815	3	vessel-load	Maersk	NaT	2022-09-12 07:00:00	2022-10-21 12:00:00	2022-10-25 21:45:00
18	2022-09-12 01:49:53.475	2022-09-08 02:58:17.815	4	vessel-discharge	Maersk	NaT	2022-10-21 12:00:00	2022-10-21 12:00:00	2022-10-25 21:45:00
19	2022-09-12 01:49:53.475	2022-09-08 02:58:17.815	5	gate-out-full	Maersk	NaT	2022-10-21 12:00:00	2022-10-21 12:00:00	2022-10-25 21:45:00

Updates inkorten

|	update_datetime	date_added	stepnumber	activity	activity_source	activity_actualtime	portofunloading_expectedtime	portofunloading_actualtime
1	2022-09-08 03:48:53.483	2022-09-08 02:58:17.815	2	gate-in-full	Maersk	2022-09-06 08:26:00	2022-10-23 14:00:00	2022-10-25 21:45:00
6	2022-09-10 02:52:53.339	2022-09-08 02:58:17.815	2	gate-in-full	Maersk	2022-09-06 08:26:00	2022-10-21 12:00:00	2022-10-25 21:45:00
11	2022-09-11 02:27:53.827	2022-09-08 02:58:17.815	2	gate-in-full	Maersk	2022-09-06 08:26:00	2022-10-21 12:00:00	2022-10-25 21:45:00
16	2022-09-12 01:49:53.475	2022-09-08 02:58:17.815	2	gate-in-full	Maersk	2022-09-06 08:26:00	2022-10-21 12:00:00	2022-10-25 21:45:00
22	2022-09-13 01:17:55.078	2022-09-08 02:58:17.815	3	vessel-load	Maersk	2022-09-12 03:16:00	2022-10-21 12:00:00	2022-10-25 21:45:00
27	2022-09-14 00:59:52.608	2022-09-08 02:58:17.815	3	vessel-load	Maersk	2022-09-12 03:16:00	2022-10-21 12:00:00	2022-10-25 21:45:00
32	2022-10-15 04:16:50.275	2022-09-08 02:58:17.815	3	vessel-load	Maersk	2022-09-12 03:16:00	2022-10-24 11:00:00	2022-10-25 21:45:00
37	2022-10-16 03:52:50.208	2022-09-08 02:58:17.815	3	vessel-load	Maersk	2022-09-12 03:16:00	2022-10-24 12:00:00	2022-10-25 21:45:00
42	2022-10-23 03:53:51.308	2022-09-08 02:58:17.815	3	vessel-load	Maersk	2022-09-12 03:16:00	2022-10-24 11:00:00	2022-10-25 21:45:00

Features

- Timeleft   ETA - update_datetime
- Timespent   update_datetime - dateadded
- Delay   ATA - ETA

Eta aanvullen

- Modus   17.94   MAE
- Mean   7.2   MAE
- Mediaan   7.01   MAE
- KNN   6.37   MAE
- Ffill   100% accuraat, maar 0.76% blijft NaN
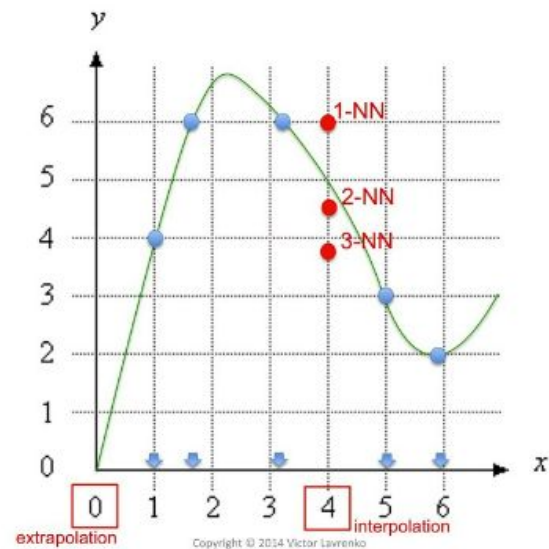
Dataset

Gemiddeld (mean) zitten containers 3.35 dagen van het optimale moment af
De standaard deviatie van te late/vroege containers is 6.12

Na het cleanen van de dataset zijn er 2735 unieke trips overgebleven
Er zijn in totaal 407 trips te vroeg aangekomen, dat is 14.88 % van het totaal
Er zijn in totaal 2312 trips te laat aangekomen, dat is 84.53 % van het totaal
Er zijn in totaal 16 trips precies optijd aangekomen, dat is 0.59 % van het totaal
Er zijn in totaal 732 trips zijn binnen 24 uur van de ETA aangekomen 26.76 % van het totaal
Er zijn in totaal 472 trips tussen de 24 en 48 uur te laat 17.26 % van het totaal
Er zijn in totaal 335 trips tussen de 48 en 72 uur te laat 12.25 % van het totaal
Er zijn in totaal 773 trips minimaal 72 uur te laat 28.26 % van het totaal
Er zijn in totaal 232 trips minimaal 1 week te laat 8.48 % van het totaal

Een trip duurt gemiddeld 33.52 dagen
De langste trip duurt 110.46 dagen
De kortste trip duurt 17.04 uur

De dataset bevat uiteindelijk, 18772 rows (updates) en 23 kolommen

|	na_amount	na_percentage
update_datetime	0	0
containernumber	0	0
date_last_updated	0	0
date_added	0	0
vesselname	0	0
origin_city	0	0
origin_country	0	0
destination_city	0	0
destination_country	0	0
stepnumber	0	0
activity	0	0
activity_source	0	0
activity_city	0	0
activity_country	0	0
activity_actualtime	0	0
portofunloading_country	0	0
portofunloading_city	0	0
portofunloading_actualtime	0	0
vessel_arrived	0	0
eta	0	0
timeleft	0	0
timespent	0	0
delay	0	0

Methodes

Methodes

- Regression modellen:
  - Linear regression
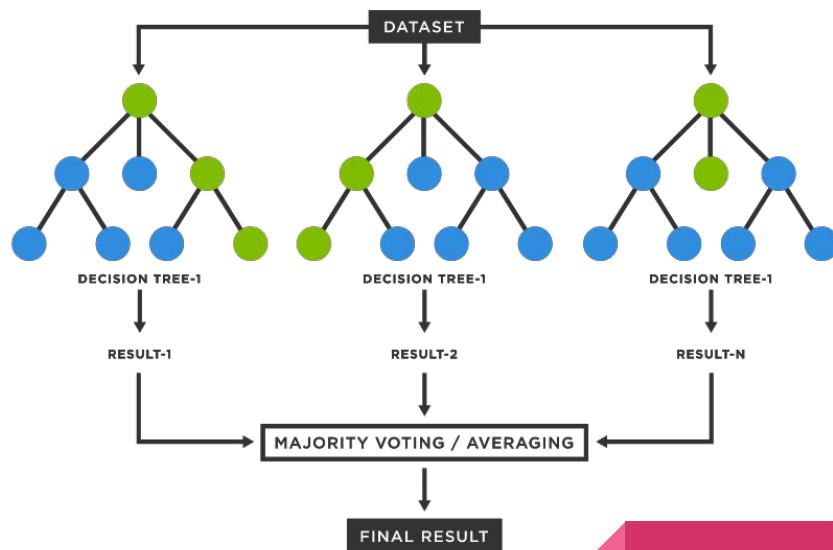 - Polynomial regression
  - Random forest regression

Linear en polynomial regression

- Lineaire of kromme verbanden
- Multivariabel
- Linear - underfit
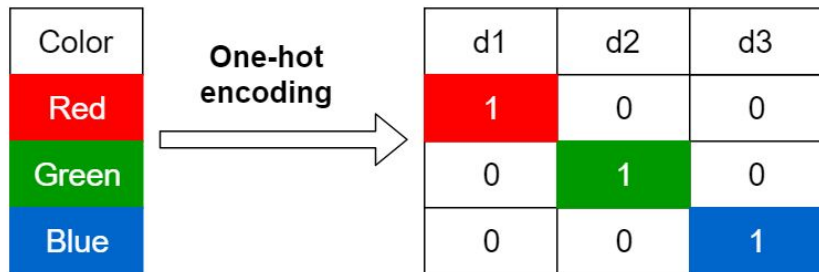- Polynomial - te veel computatie
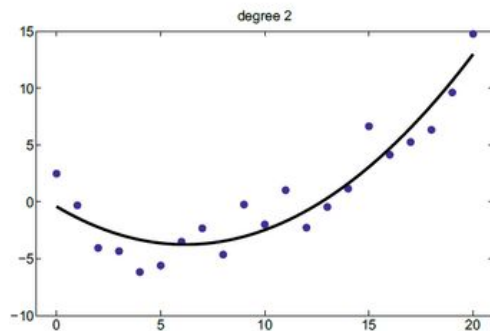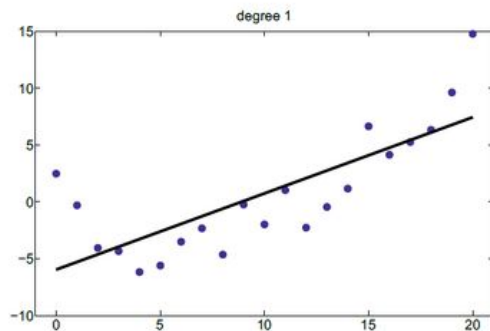
Random forest regression

- Meerdere decision trees
- Majority voting
- Overfitting

Implementatie

Sklearn

One Hot Encoding:

    `origin_country`

    `origin_city`

    `activity`

    `vesselname`

    `activity_source`

    `activity_city`

    `portofunloading_city`

Train-Test split op chronologische volgorde en niet random
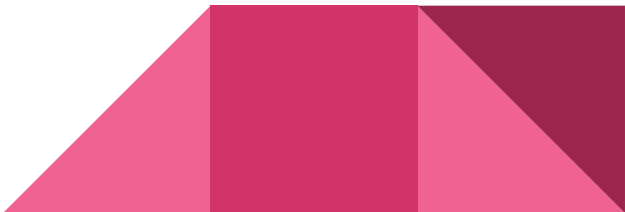
Randomized search

Evaluatie

Mean squared error

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

Mean absolute error
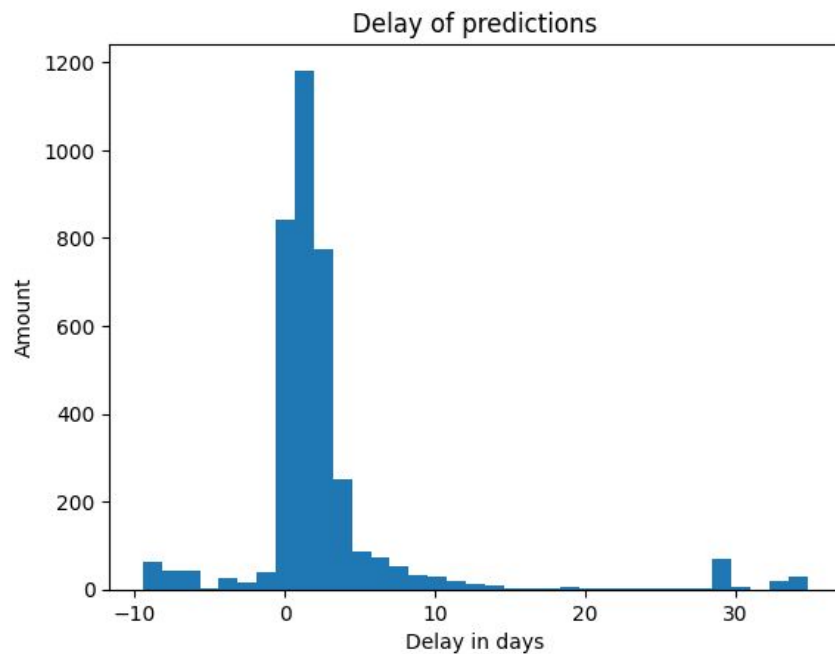
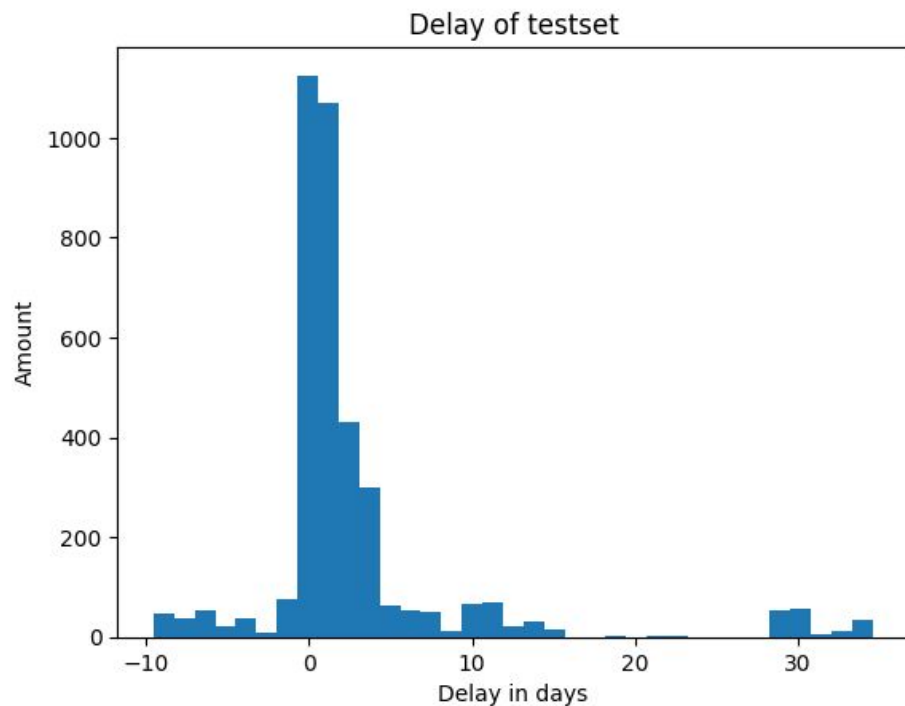$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |x_i - x|$$

Resultaten

Resultaten

- Expected time:
  - MAE: 3.58
  - MSE: 54.75
- Standaard parameters
  - MAE: 1.11
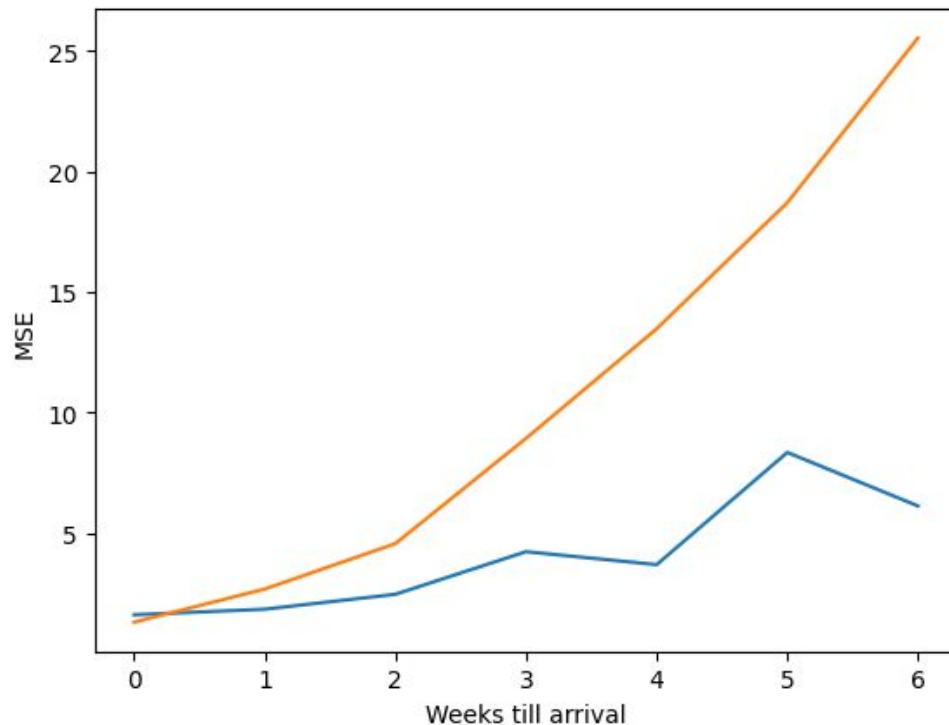 - MSE: 8.15
- Randomized Search CV
  - MAE: 1.20
  - MSE: 7.89

Distributie van errors

Delay of testset
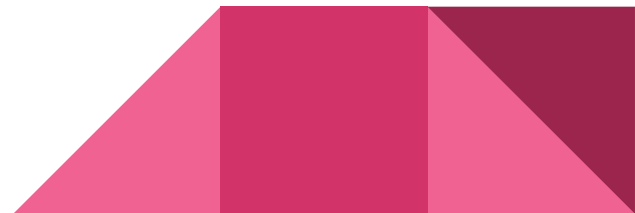
Delay of predictions

Resultaten over tijd

- Accuratie neemt af als de tijd tot aankomst toeneemt

- Aangeleverde ETA beter binnen de week van levering

- Prediction beter bij overige datapunten

Conclusie

Conclusie

- Random forest regression goede resultaten
- Verder onderzoek:
    - Aanvullen van de ETA's
    - Overfitting van de features
    - Reden voor outliers