STØTTET AF
Promilleafgiftsfonden for landbrug

Udbytteprognosemodel for majs

December 30, 2022

Dette dokument beskriver udviklingen af en udbytteprognosemodel for majs foretaget i 2022.

Kontaktinformation: SEGES-datascience@seges.dk

1. Metode

1.1 Datagrundlag og features

Der er anvendt både offentlige og fortrolige datakilder til at udvikle udbytteprognosemodellen. Modellen er baseret på:

- Udbyttedata
- Satellitdata
- Vejrdata
- Markdata

1.1.1 Udbyttedata

Der er udbyttedata fra i alt 4255 marker fra 2016-2022. Udbyttedataene er renset ved brug af følgende metode:

- Manuel gennemgang af marker
- Fjernelse af dubletter (hvor år, placering (x, y) og udbytte var identiske).
- Fjernelse af marker uden markpolygon.
- Fjernelse af marker med multipolygon grundet tidsmangel.
- Fjernelse af marker, hvor udbyttet er ens for forskellige marker for samme CVR.

Efter oprensning var der 3706 marker tilbage, som opfylder kvalitetskravet. Disse udbyttedata fungerer som target i modellen, som vi gerne vil forudsige.
I tabel 1 ses fordelingen af marker over år.

Harvest Year	Field before cleaning	Fields after cleaning
2016	241	193
2017	458	393
2018	799	690
2019	410	362
2020	700	699
2021	737	694
2022	910	675
Sum	4255	3706

Tabel 1. Fordeling af marker med udbyttedata over år.

1.1.2 Satellitdata

I analysen er der anvendt satellitdata fra Sentinel 2 (L1C), som er downloaded via en service fra Sentinel-Hub. Sentinel 2 består af to satellitter (S2A og S2B), som leverer data fra 13 spektrale bånd, der alle er anvendt i udbytteprognosen (Se tabel 2). Der er hentet satellitbilleder fra 1. maj til 21. august for hvert år. Satellitbilleder med skyer er blevet fjernet ved brug af S2_cloudless algoritmen.

	S2A		S2B		
Bånd nummer	**Centrale bølgelængde (nm)**	**båndbredde (nm)**	**Centrale bølgelængde (nm)**	**Båndbredde (nm)**	**Spatial resolution (m)**
B01	442.7	21	442.2	21	60
B02	492.4	66	492.1	66	10
B03	559.8	36	559	36	10
B04	664.6	31	664.9	31	10
B05	704.1	15	703.8	16	20
B06	740.5	15	739.1	15	20
B07	782.8	20	779.7	20	20
B08	832.8	106	832.9	106	10
B08a	864.7	21	864	22	20
B09	945.1	20	943.2	21	60
B10	1373.5	31	1376.9	30	60
B11	1613.7	91	1610.4	94	20
B12	2202.4	175	2185.7	185	20

Tabel 2. Spektrale bånd tilgængelig fra Sentinel 2 (S2A og S2B) samt båndbredde (nm) og opløselighed (m). Bånd markeret med orange indgår i beregningen af vegetationsindeksene NDVI, NDRE og MSAVI2.

Ud fra bånd B04, B05 og B08 er vegetationsindeksene NDVI, NDRE og MSAVI2 udregnet.

(1)
$$NDVI = \frac{(B08 - B04)}{(B08 + B04)}$$

(2)
$$NDRE = \frac{B08 - B05}{B05 + B05}$$

(3)
$$MSAVI2 = \frac{2 \cdot B08 + 1 - \sqrt{(2 \cdot B08 + 1)^2} - 8 \cdot (B08 - B04)}{2}$$

Satellitdataene er dernæst lineært interpoleret. For hver mark er værdien hver 14. dag i vækstsæsonen fra 1. maj til 21. august beregnet for de 13 spektrale bånd, NDVI, NDRE og MSAVI2. Disse er anvendt som features i modellen.

1.1.3 Vejrdata

Vejrdata som luft- og jordtemperatur, nedbør, fordampning og indstråling er hentet fra DMI fra den nærmeste vejrstation for hver mark. For alle vejrparametre er der for hver 14. dag beregnet: gennemsnit, standard afvigelse (SD), minimum, maksimum. Disse anvendes som features i modellen.

1.1.5 Markdata

Til sidst er der anvendt følgende oplysninger fra hver mark som features:

- Forafgrøde
- Mark centroid koordinater

1.2 Machine learning model

Machine learning (ML) algoritmen *Gradient Boosting* er anvendt til udbytteprognosen.

1.3 Modelevaluering

For at vurdere modellens forudsigelsesnøjagtighed er der anvendt en række metrikker:

- Mean absolute error (MAE)
- Root mean squared error (RMSE)
- R^2
- Standard deviation of absolute error (STD of AE)

Vi er interesserede i metrikkerne på testsættet, da disse siger noget om modellens forudsigelsesnøjagtighed out-of-sample.

2. Resultater

For at undersøge forudsigelsesnøjagtigheden out-of-sample er der foretaget cross-validation med år som folds. Dvs. et helt år tages ud af træningssættet og bliver brugt som testsæt. Så trænes der på de resterende år og modellen forudsiger udbytte på testsættet. Dette gøres for alle år, og til sidst har vi dermed et datasæt med out-of-sample predictions for alle marker alle år. Dette efterligner virkelighedens brug af modellen. Hvis der i stedet var foretage et random split, 80-20 f.eks., så er der marker fra samme år i trænings- og testsættet. Dette kan resultere i en kunstig god performance, da modellen derved kan finde frem til det 'generelle' udbytteniveau for et givet år. 'År' optræder ikke som en feature i modellen, men andre features (som vejrdata features) kan komme til at agere proxyvariabler for året.

I tabel 3 er resultaterne opsummeret. Tabellen viser, hvor tidligt i vækstsæsonen udbyttet forudsiges (prediction date), hvilke features som er med, hvordan datasættet er inddelt i trænings- og testsæt og resultaterne på tværs af alle år og for hvert år. Target og modellens forudsigelser (og derfor også MAE) er i kilogram tørstof pr. hektar.

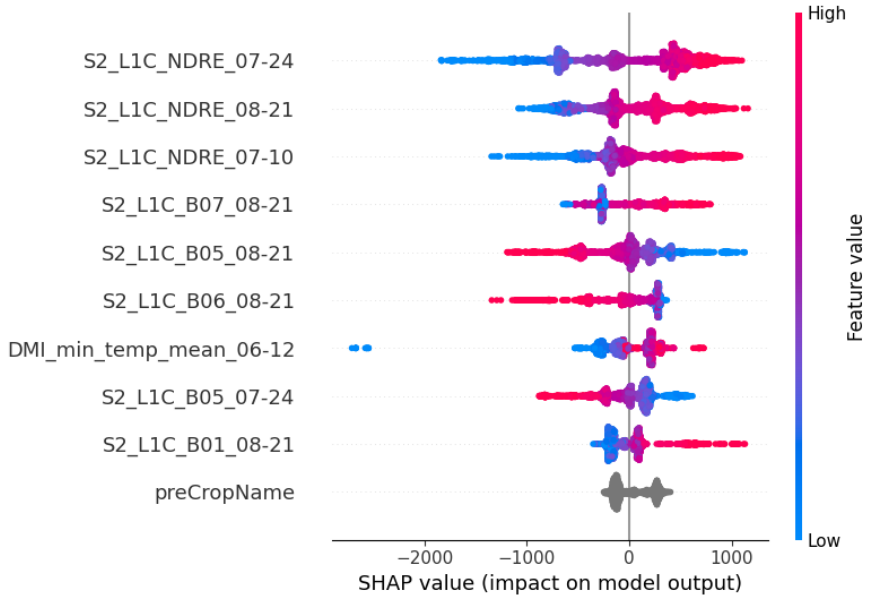
Der er foretaget 'feature elimination' inden cross-validation. Ved 'feature elimination' fjernes de mindst vigtige features iterativt fra modellen ved gentagne gange at træne modellen og fjerne features, indtil man når det ønskede antal af tilbageværende features – i dette tilfælde 20. Prediction performance blev bedre ved at bruge et mindre antal betydningsfulde features sammenlignet med at bruge alle features.

Prediction Date	Features	Split	MAE, kg/h								RMSE	R^2	Std. of AE
			All years	2022	2021	2020	2019	2018	2017	2016	All years	All years	All years
August 21	Feature elimination	Cross-validation with years as folds	1592	1617	1450	1479	1509	1835	1499	1904	2043	0,32	1280
August 7			1631	1649	1490	1533	1515	1881	1506	2001	2093	0,29	1312
July 24			1645	1723	1448	1640	1577	1759	1484	2143	2107	0,28	1317
July 10			1726	1786	1616	1689	1593	1906	1492	2134	2184	0,22	1337
June 26			1799	1793	1738	1693	1601	2118	1573	2122	2273	0,16	1389

Tabel 3 Resultater. Tabellen viser, hvor tidligt i vækstsæsonen udbyttet forudsiges (prediction date), hvilke features som er med, hvordan datasættet er inddelt i trænings- og testsæt og resultaterne på tværs af alle år og for hvert år.

Der opnås en rigtig god nøjagtighed med MAE på 1592 kg/h for prediction date 21. august. Samtidig ses det at forudsigelsesnøjagtigheden er bedre jo senere i vækstsæsonen man forudsiger. Dette er i overensstemmelse med forventningerne. Det ses, at performance varierer på tværs af år. Performance er dårlig for 2018, hvilket giver mening grundet de ekstreme vejrforhold det år.

Hvis vi ser på de 10 vigtigste features og deres betydning for prediction date 21. august, får vi følgende:



Det ses, at en højere værdi af NDRE i slutningen af vækstsæsonen er associeret med et højere udbytte. Omvendt er en lav gennemsnitlig minimum temperatur omkring starten af juni associeret med et lavere udbytte.

SEGES INNOVATION