DATA BOUNTY II

Konrad Eilers

Email: eilerskonrad1@gmail.com

DATA BOUNTY II SUMMARY

- The Graphical Method shows that Normalized Difference Vegetation Index and Land Surface Temperature are relatively strongly correlated with crop yields on aggregate. However, there is large variation in specific counties and strong seasonality effects.

- A low-cost, empirical formula has been developed for farmers to estimate the crop yield:
$$\boldsymbol{Yield_{this\ year} = -7.968 * NDVI_{avg} - 0.05 * LAI_{avg} + 0.06 * ET_{avg} + 0.01 * LST_{avg} - 0.1 * RF_{avg}}$$

- A Machine-Learning Model – based on Meta’s open-source technology – has been deployed and is *forecasting a positive yield trend*, albeit driven by strong seasonality effects

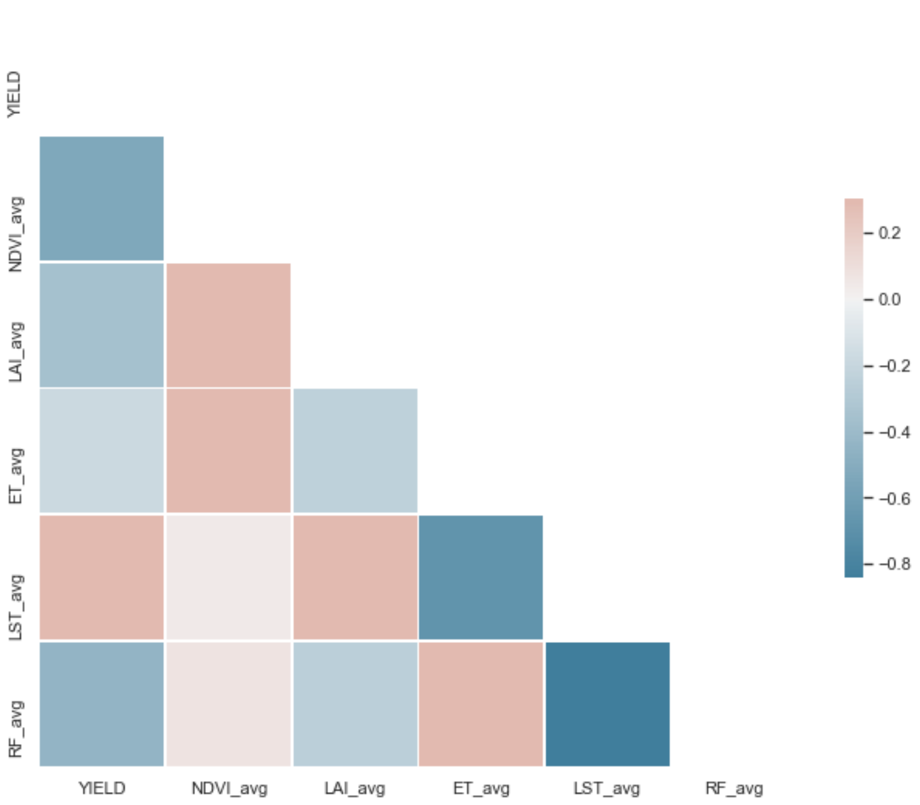
- Additional data augmentation covering the demand side of the forecasting problem has been uploaded to the Ocean Market.

SUMMARY CORRELATION OVER TIME

- Simplifying across regions and over time, the correlation between the crop yield and explanatory variables vary greatly
- On average, most variables are negatively correlated (except Land Surface Temperature)
- **Strongest correlation to yield demonstrates the Normalized Difference Vegetation Index as well as Land Surface Temperature**
- The next slides dives into regional and temporal specifics

Source: Dimitra Bounty Phase II dataset.

Correlation Matrix over time across regions



Absolute Size

Variable	Corr-elation
NDVI	-0.53
LST	0.48
RF	-0.44
LAI	-0.35
ET	-0.18

CORRELATION IN DETAIL

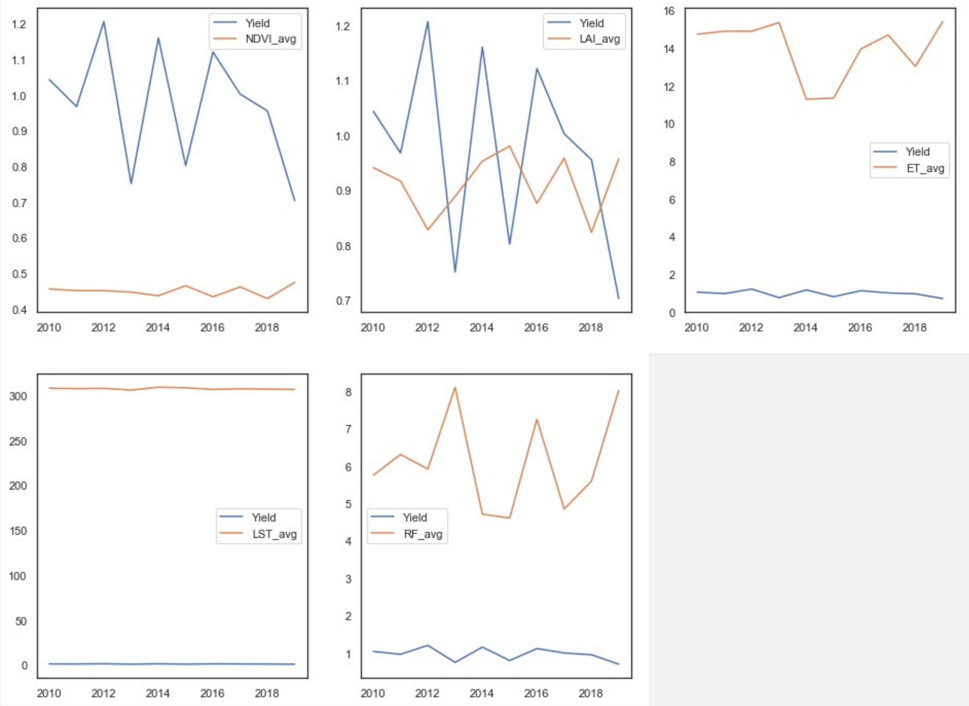
- Regions vary greatly with regards to their correlation
- Interestingly, on average, Leaf Area Index is not highly correlated with yield, however this is **highly dependent on regions (see Katni)**
- Analysis over time demonstrates fluctuations across years in both yield as well as in rain fall / leaf area index

Source: Dimitra Bounty Phase II dataset.

Top 20 regional positive correlation

Katni_LAI_avg	0.513583
Burhanpur_LAI_avg	0.407149
Katni_NDVI_avg	0.344550
Jhabua_LAI_avg	0.344142
Bhind_LAI_avg	0.293226
Barwani_LAI_avg	0.278580
Burhanpur_NDVI_avg	0.267918
Bhind_NDVI_avg	0.258543
Datia_LAI_avg	0.249156
Chhatarpur_LAI_avg	0.220201
Anuppur_LAI_avg	0.210610
Rewa_LAI_avg	0.207023
Chhatarpur_NDVI_avg	0.199324
Barwani_NDVI_avg	0.189443
Anuppur_NDVI_avg	0.179054
Jhabua_NDVI_avg	0.169087
Hoshangabad_LAI_avg	0.149784
Rewa_NDVI_avg	0.143504
Datia_NDVI_avg	0.133937
Umaria_NDVI_avg	0.122408

Average Correlation over time



EMPIRICAL FORMULA

- To estimate a simple empirical formula, I use an OLS model keeping records (each year) independent
- This is a simplifying assumption, allowing us to get to a simple empirical formula based on a linear regression:

$$\boldsymbol{Yield_{this\ year} = -7.968 * NDVI_{avg} - 0.05 * LAI_{avg} + 0.06 * ET_{avg} + 0.01 * LST_{avg} - 0.1 * RF_{avg}}$$

- This allows a quick estimation of this year's yield that farmers can use to estimate the crop yield for the year

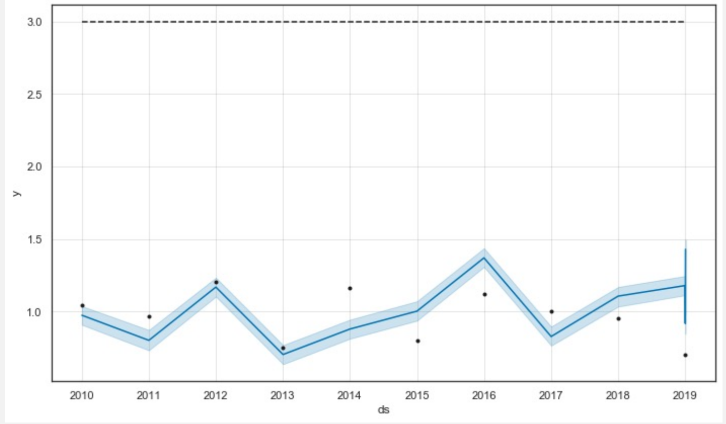
```
                             OLS Regression Results
=======================================================================================
Dep. Variable:                  YIELD   R-squared (uncentered):                   0.989
Model:                            OLS   Adj. R-squared (uncentered):              0.979
Method:                 Least Squares   F-statistic:                              92.49
Date:                Sat, 24 Sep 2022   Prob (F-statistic):                    6.35e-05
Time:                        21:13:54   Log-Likelihood:                          8.6544
No. Observations:                  10   AIC:                                     -7.309
Df Residuals:                       5   BIC:                                     -5.796
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
NDVI_avg      -7.9672      5.744     -1.387      0.224     -22.733       6.798
LAI_avg       -0.0429      1.431     -0.030      0.977      -3.723       3.637
ET_avg         0.0615      0.053      1.164      0.297      -0.074       0.197
LST_avg        0.0141      0.005      2.793      0.038       0.001       0.027
RF_avg        -0.0951      0.052     -1.840      0.125      -0.228       0.038
==============================================================================
Omnibus:                        0.247   Durbin-Watson:                   2.963
Prob(Omnibus):                  0.884   Jarque-Bera (JB):                0.401
Skew:                          -0.083   Prob(JB):                        0.818
Kurtosis:                       2.033   Cond. No.                     3.95e+04
==============================================================================
```

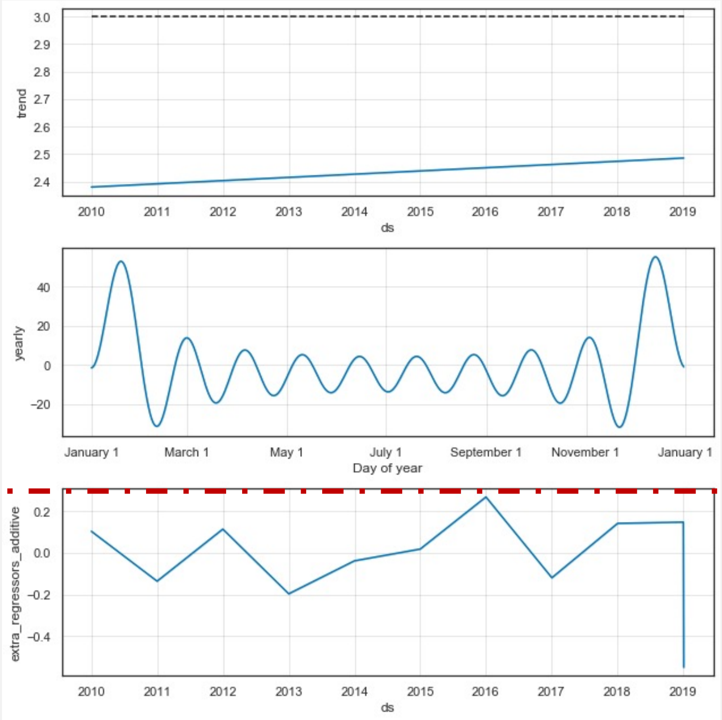
MACHINE LEARNING MODEL

- Proposing an ML solution based on Meta's open-source Prophet procedure powered by an additive model, that allows **multi-variate time series modelling accounting for seasonality and uncertainty** in forecasting

- Forecasting algorithm demonstrates 2019 – lower than expected yield performance

- Based on inputs, seasonality modelling (see chart to the right) and **positive trend**, the forecasting method predicts higher yields than in 2019 in the forecasted period





DATA AUGMENTATION

- The current solution focuses on the supply-side of the problem: The explanatory variables (i.e. Rainfall, Leaf Area Index, etc.) demonstrate environmental conditions conducive or non-conducive to yield growth
- However, the demand side has not yet been considered: If there is a bigger market, can farmers obtain better technology to improve yields?
- For this purpose, open-source information on Madhya Pradesh's population from the last Indian Census have been **uploaded to** Ocean Market.

Source: https://censusindia.gov.in/census.website/data/census-tables