ML Crop Yield Prediction

Eshan Kumar, ek3227
ML & Climate with Professor Alp Kucukelbir
12/6/22

Table of contents

01

Overview

Motivation & Problem Description

Overview

- Human population exploding past 8 billion
- Climate is rapidly changing, large effect on agriculture
- Understanding worldwide crop yield as climate factors change is critical
    - Address Food security challenges pre-emptively (Agricultural risk management)
    - Predict how agriculture changes will affect economies of areas around the world

02

Dataset

Data Description, Visualization, &
Preprocessing

Dataset

- Data from FAO (Food and Agriculture Organization), World Data Bank, & Climate Change Knowledge Portal
- Contains the rainfall, pesticide use, temperature, crop item, and crop yield (label) from 168 countries from over 20 years, from 1990 to 2013
  - 28,242 data points with 7 features.
- Scaled numerical variables, one-hot encoded categorical variables, ordinally encoded “Year”

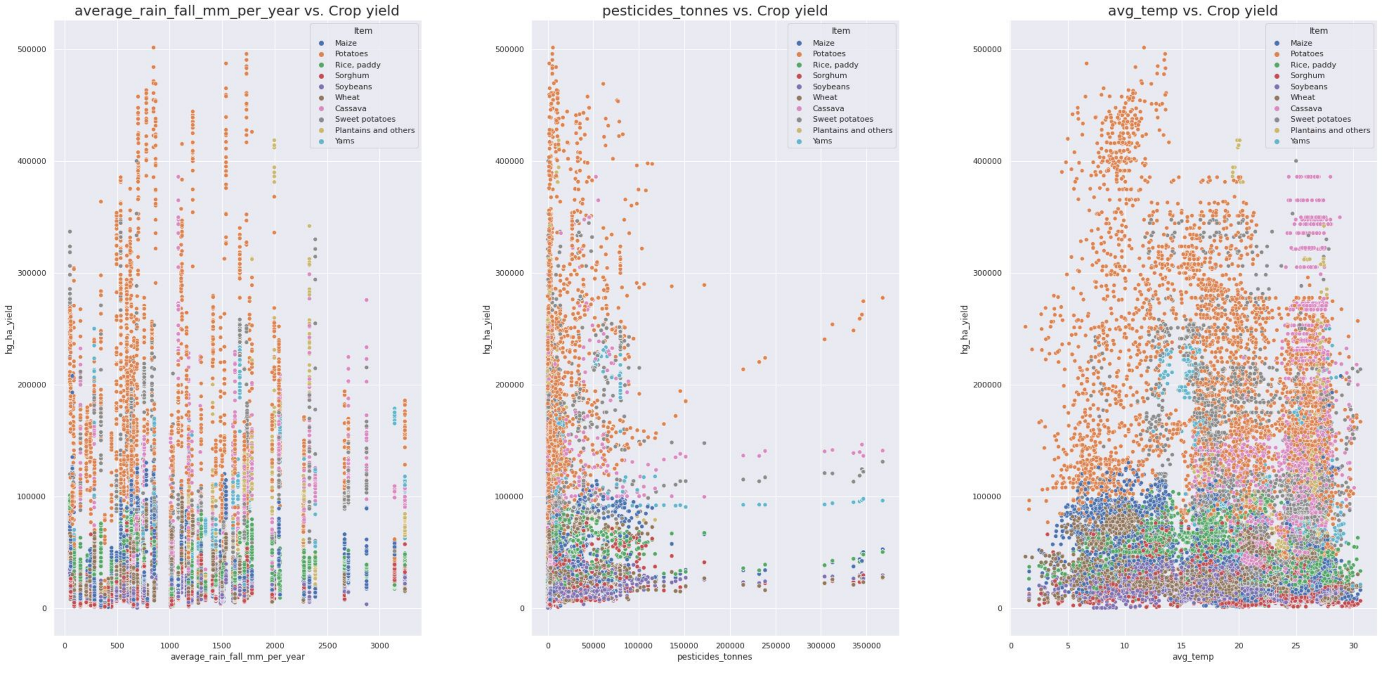
```
Countries in dataset:
 ['Albania' 'Algeria' 'Angola' 'Argentina' 'Armenia' 'Australia' 'Austria'
 'Azerbaijan' 'Bahamas' 'Bahrain' 'Bangladesh' 'Belarus' 'Belgium'
 'Botswana' 'Brazil' 'Bulgaria' 'Burkina Faso' 'Burundi' 'Cameroon'
 'Canada' 'Central African Republic' 'Chile' 'Colombia' 'Croatia'
 'Denmark' 'Dominican Republic' 'Ecuador' 'Egypt' 'El Salvador' 'Eritrea'
 'Estonia' 'Finland' 'France' 'Germany' 'Ghana' 'Greece' 'Guatemala'
 'Guinea' 'Guyana' 'Haiti' 'Honduras' 'Hungary' 'India' 'Indonesia' 'Iraq'
 'Ireland' 'Italy' 'Jamaica' 'Japan' 'Kazakhstan' 'Kenya' 'Latvia'
 'Lebanon' 'Lesotho' 'Libya' 'Lithuania' 'Madagascar' 'Malawi' 'Malaysia'
 'Mali' 'Mauritania' 'Mauritius' 'Mexico' 'Montenegro' 'Morocco'
 'Mozambique' 'Namibia' 'Nepal' 'Netherlands' 'New Zealand' 'Nicaragua'
 'Niger' 'Norway' 'Pakistan' 'Papua New Guinea' 'Peru' 'Poland' 'Portugal'
 'Qatar' 'Romania' 'Rwanda' 'Saudi Arabia' 'Senegal' 'Slovenia'
 'South Africa' 'Spain' 'Sri Lanka' 'Sudan' 'Suriname' 'Sweden'
 'Switzerland' 'Tajikistan' 'Thailand' 'Tunisia' 'Turkey' 'Uganda'
 'Ukraine' 'United Kingdom' 'Uruguay' 'Zambia' 'Zimbabwe']

Crops in dataset:
 ['Maize' 'Potatoes' 'Rice, paddy' 'Sorghum' 'Soybeans' 'Wheat' 'Cassava'
 'Sweet potatoes' 'Plantains and others' 'Yams']
```

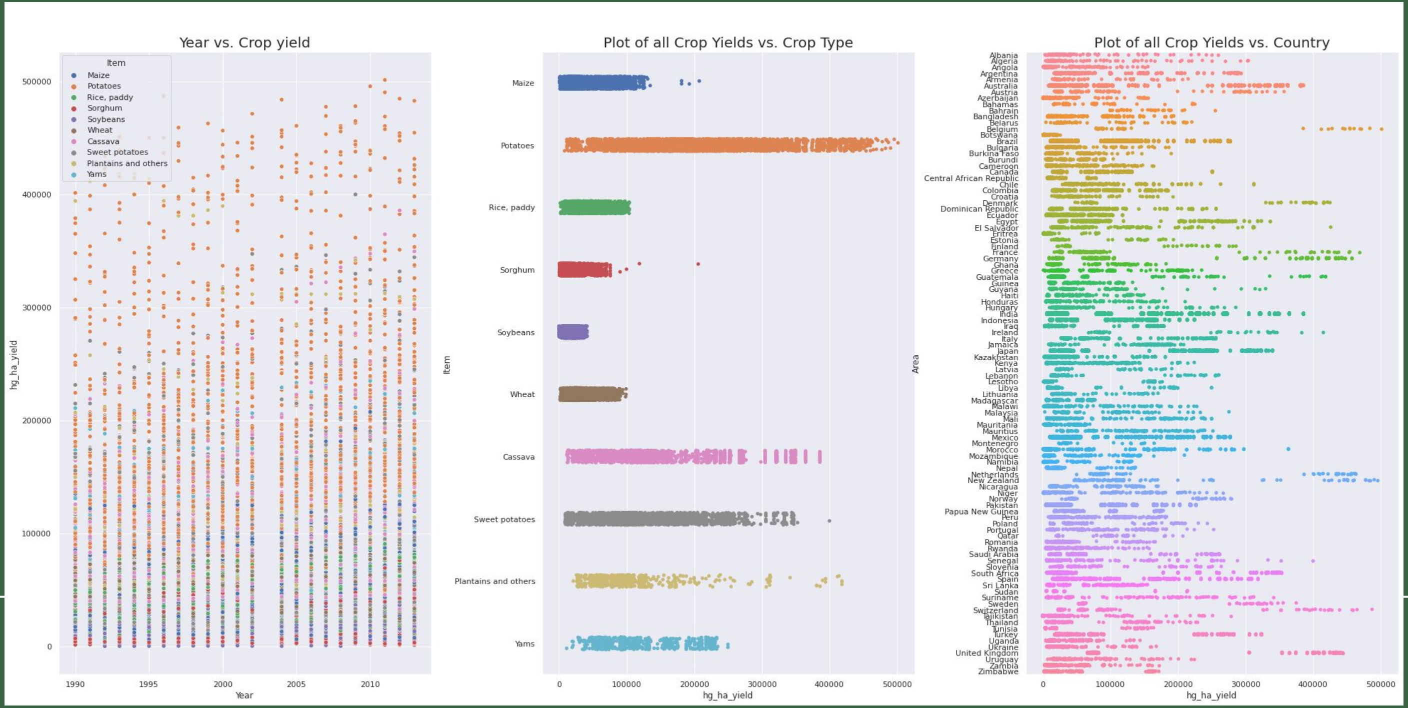
	Unnamed: 0	Area	Item	Year	hg/ha_yield	average_rain_fall_mm_per_year	pesticides_tonnes	avg_temp
3874	3874	Brazil	Sorghum	2004	23187	1761.0	214725.00	20.05
4988	4988	Burkina Faso	Yams	1993	59603	748.0	17.00	28.58
5458	5458	Cameroon	Cassava	2001	81918	1604.0	687.00	25.01
13484	13484	India	Wheat	2007	27079	1083.0	27422.77	24.60
27311	27311	Uganda	Wheat	2008	17273	1180.0	88.00	23.68

Dataset

- No clear trends in yield from Categorical or numerical variables



Dataset



Year vs. Crop yield
Item
Maize
Potatoes
Rice, paddy
Sorghum
Soybeans
Wheat
Cassava
Sweet potatoes
Plantains and others
Yams
hg_ha_yield
0
100000
200000
300000
400000
500000
1990
1995
2000
2005
2010
Year
Plot of all Crop Yields vs. Crop Type
Item
Maize
Potatoes
Rice, paddy
Sorghum
Soybeans
Wheat
Cassava
Sweet potatoes
Plantains and others
Yams
0
100000
200000
300000
400000
500000
hg_ha_yield
Plot of all Crop Yields vs. Country
Area
Albania
Algeria
Angola
Argentina
Armenia
Australia
Austria
Azerbaijan
Bahamas
Bahrain
Bangladesh
Belarus
Belgium
Botswana
Brazil
Bulgaria
Burkina Faso
Burundi
Cameroon
Canada
Central African Republic
Chile
Colombia
Croatia
Denmark
Dominican Republic
Ecuador
Egypt
El Salvador
Eritrea
Estonia
Finland
France
Germany
Ghana
Greece
Guatemala
Guinea
Guyana
Haiti
Honduras
Hungary
India
Indonesia
Iraq
Ireland
Italy
Jamaica
Japan
Kazakhstan
Kenya
Latvia
Lebanon
Lesotho
Libya
Lithuania
Madagascar
Malawi
Malaysia
Mali
Mauritania
Mauritius
Mexico
Montenegro
Morocco
Mozambique
Namibia
Nepal
Netherlands
New Zealand
Nicaragua
Niger
Norway
Pakistan
Papua New Guinea
Peru
Poland
Portugal
Qatar
Romania
Rwanda
Saudi Arabia
Senegal
Slovenia
South Africa
Spain
Sri Lanka
Sudan
Suriname
Sweden
Switzerland
Tajikistan
Thailand
Tunisia
Turkey
Uganda
Ukraine
United Kingdom
Uruguay
Zambia
Zimbabwe
0
100000
200000
300000
400000
500000
hg_ha_yield

03

Models & Inference

Linear models, Tree models, Causal model & their predictions

Models & Inference

Linear Models

- Linear Regression
- Ridge Regression
- Elastic-Net Regression
- Lasso Regression
- Bayesian ARD Regression
- Bayesian Ridge Regression
- K Nearest Neighbors Regression
- Support Vector Regression

Tree Models

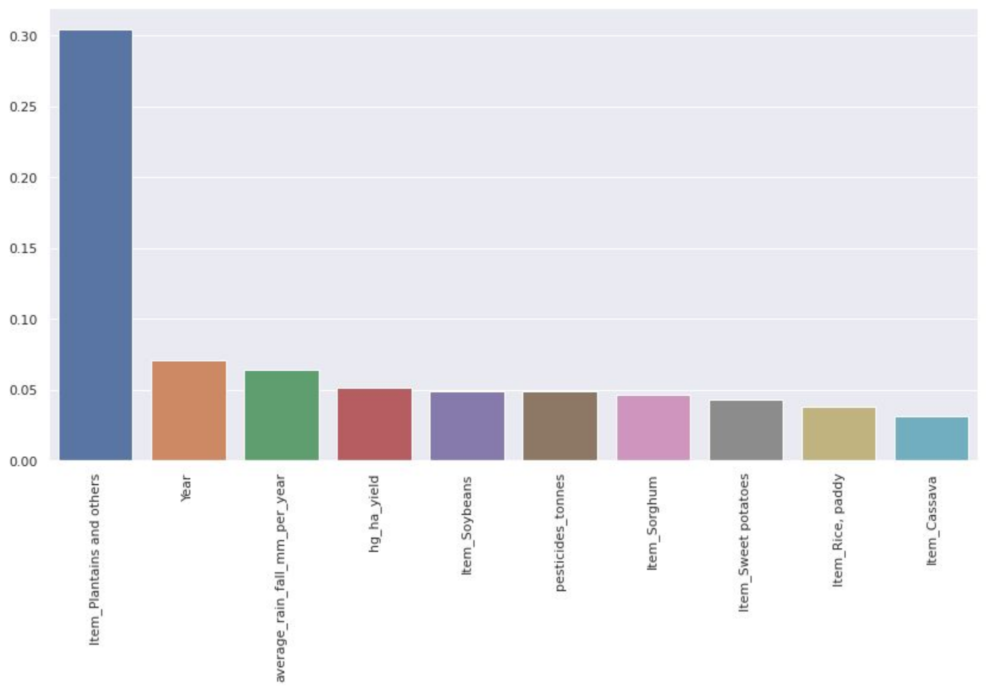
- Decision Tree Regression
- Random Forest Regression
- Gradient Boosted Trees
- XGBoost

Other Models

- Multi-Layer Perceptron Neural Network
- Elementary causal model

Models & Inference

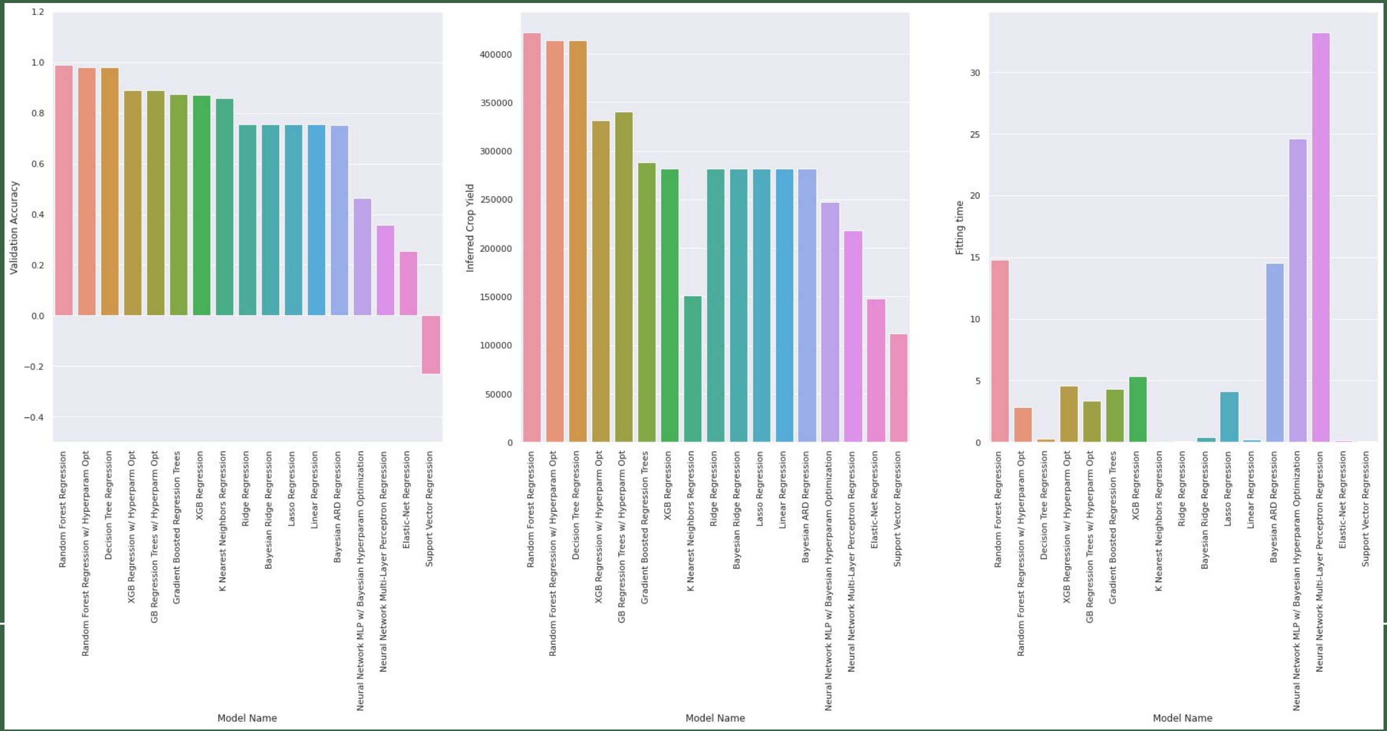
- Used Bayesian Hyperparameter Optimization over Random/Grid Search
  - Efficiently Led to model improvements without exhaustively searching space
- Got feature importances from tree models



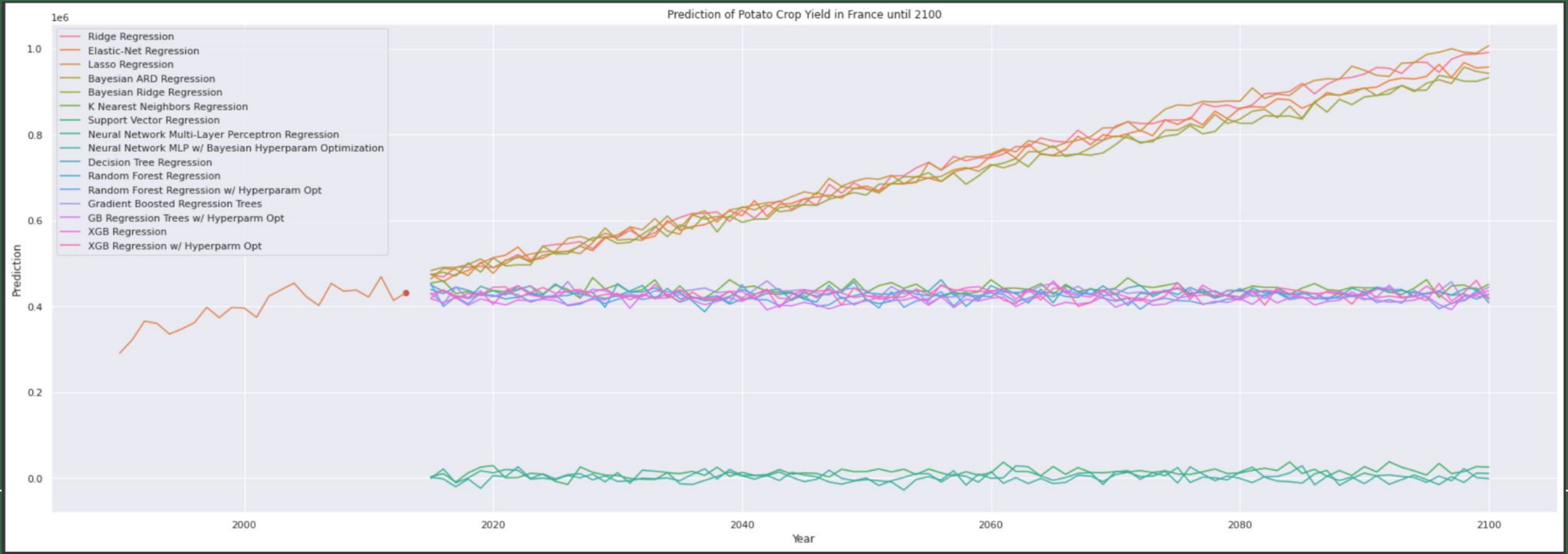
Models & Inference

	Model Name	Model	Fitting time	Scoring time	Train Accuracy	Validation Accuracy	Inferred Crop Yield
11	Random Forest Regression	(DecisionTreeRegressor(max_features='auto', ra...	14.764779	0.195721	0.998384	0.988030	422429.750000
12	Random Forest Regression w/ Hyperparam Opt	(DecisionTreeRegressor(max_depth=35, max_featu...	2.876744	0.146089	0.945915	0.981581	414251.240000
10	Decision Tree Regression	DecisionTreeRegressor()	0.303341	0.007003	1.000000	0.979951	413769.000000
16	XGB Regression w/ Hyperparm Opt	XGBRegressor(min_impurity_decrease=0.125, n_es...	4.569502	0.032414	0.227024	0.890289	331462.218750
14	GB Regression Trees w/ Hyperparm Opt	([DecisionTreeRegressor(criterion='friedman_ms...	3.356412	0.023210	0.531050	0.890191	340769.564665
13	Gradient Boosted Regression Trees	([DecisionTreeRegressor(criterion='friedman_ms...	4.303582	0.015550	0.877714	0.873398	288670.611810
15	XGB Regression	XGBRegressor()	5.363534	0.026464	0.874709	0.870382	281556.875000
6	K Nearest Neighbors Regression	KNeighborsRegressor()	0.011836	3.057751	0.916508	0.857904	151099.600000
1	Ridge Regression	Ridge()	0.077325	0.006203	0.756644	0.754077	281479.538813
5	Bayesian Ridge Regression	BayesianRidge()	0.382362	0.006854	0.756648	0.754077	281496.695980
3	Lasso Regression	Lasso()	4.140481	0.007774	0.756672	0.754073	281604.806570
0	Linear Regression	LinearRegression()	0.234917	0.007883	0.756677	0.754070	281739.500000
4	Bayesian ARD Regression	ARDRegression()	14.549672	0.006971	0.755605	0.752952	281863.861139
9	Neural Network MLP w/ Bayesian Hyperparam Opti...	MLPRegressor(activation='identity', alpha=0.1,...	24.642043	0.022971	0.142967	0.464800	247223.521072
8	Neural Network Multi-Layer Perceptron Regression	MLPRegressor(early_stopping=True)	33.246683	0.014120	0.358029	0.356383	217989.383800
2	Elastic-Net Regression	ElasticNet()	0.153725	0.010947	0.254837	0.254424	147802.397586
7	Support Vector Regression	LinearSVR()	0.052212	0.004538	-0.230206	-0.230450	111949.240421

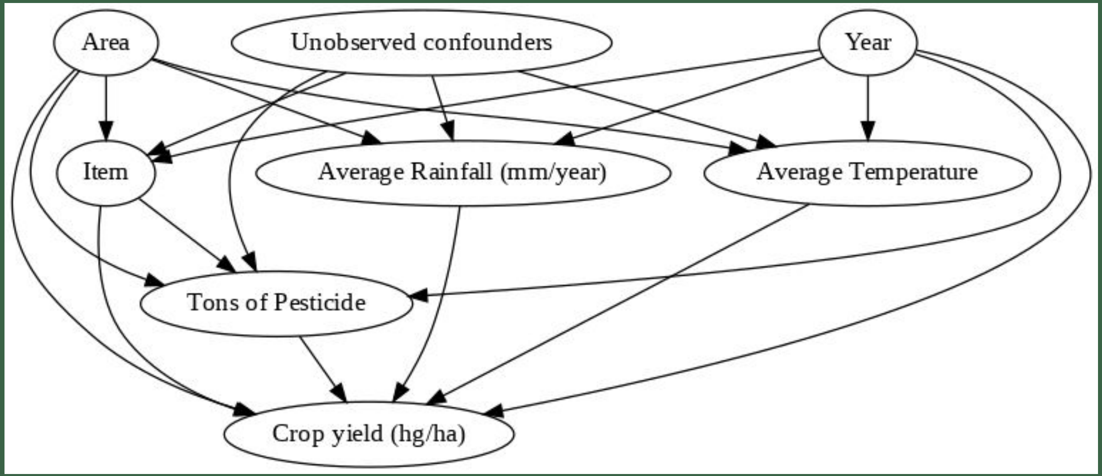
Models & Inference



Models & Inference



Models & Inference



- Wanted to implement elementary structural causal model
  - Week 4 paper - Inferring causation from time series in Earth system sciences
- Used Microsoft DoWhy package
  - Followed procedure from article

Models & Inference

- Estimated *causal effect* on *outcome* (Crop Yield) based on different *treatments*
    - *Causal effect* - magnitude by which the *Outcome* changes due to a unit change in *Treatment*
    - *Treatment* causes *outcome* if changing *Treatment* leads to a change in *Outcome* keeping everything else constant
- Performed robustness checks to test validity of assumptions used to create above graph
    - Attempted to refute results for Rainfall (mm/year) *treatment*
- Found that causal effect of temperature and rainfall is quite strong

	Treatment	Estimated Effect
0	avg_temp	-0.197101
1	pesticides_tonnes	-0.032410
2	average_rain_fall_mm_per_year	0.098958

	Refutal Method	Estimated Effect	New Effect	
0	Add a random common cause	0.0989	0.0989	(should be similar)
1	Use a Placebo Treatment	0.0989	-0.000	(should go to 0)
2	Use a subset of data	0.0989	0.0366	(should be similar)

04

Conclusion

Discussion & Future work

Conclusion

Summary

- Was able to obtain high validation accuracy
- Experimented with 15 regression models, Bayesian Hyperparameter Optimization
- Found interesting Inference results, wanted more detail
- Found compelling causal effects in data with Microsoft DoWhy

Future

- Some historical data seems repeated - higher quality data may help model
- More features such as humidity, CO2 levels, etc.
- More accurate projection data
- More advanced Causal model

Thank you!