Forecasting hazelnut production using stochastic and machine-learning-based approaches within a Python-powered Jupyter Notebook

Capstone Project to satisfy requirements for
Master of Geographic Information Systems (MGIS)
Penn State | College of Earth and Mineral Sciences

Author: Jason Biagio | jrb430@psu.edu
Advisor: Dr. Richard Marini | rpm12@psu.edu

OVERVIEW

01 INTRODUCTION
Why forecast hazelnut production?

02 OBJECTIVES
What are the goals of this study?

03 BACKGROUND
Provide context for the study and provide a refresher of terms

04 RESEARCH METHODS
Provide framework for the study

05 RESULTS
How did the forecasting methods perform?

06 DISCUSSION/ CONCLUSION
Takeaways and a look to the future...

01

INTRODUCTION

Why forecast hazelnut production?

The "filbert," better known as the hazelnut (Corylus avellana), is a self-incompatible, wind-pollinated, monoecious (as having both male and female flowers) and dichogamous (flowers bloom at different times to prevent self-pollination) plant

THE 'FILBERT'

FAST FACTS

OREGON

Named as state nut in 1989 due to its historical and economic significance

BEYOND

The recent development of blight resistant cultivars will see planted acreage increase

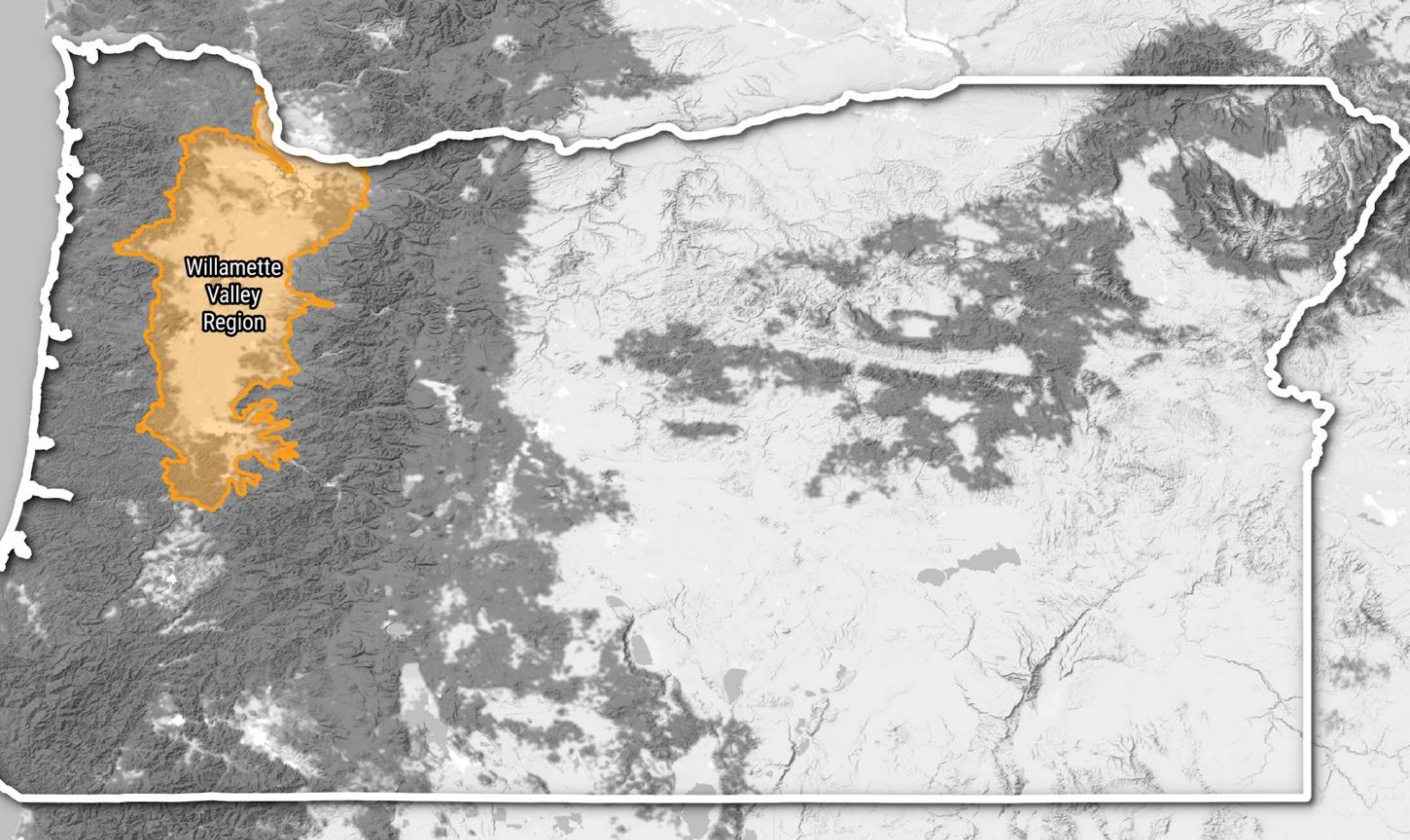
GLOBAL

Globally, hazelnuts rank 5th overall for tree nut production (behind the pistachio)

POLLINATION

Hazelnuts pollinate in the winter as opposed to the spring

LOCATION OF STUDY



02
OBJECTIVES

What are the goals of this study?

OBJECTIVES

“Forecast hazelnut yield (tons/acre/year) in Oregon's Willamette Valley using various traditional regression methods and machine-learning-based counterparts”

No. 1 - PRIMARY

“Perform predictions using the Python programming language within a novel Jupyter notebook”

No. 2 - SECONDARY

03

BACKGROUND

List of terms and other art

TERMS

TRADITIONAL REGRESSION

Autoregression (AR)

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

AR is a time series model that uses the dependent relationship between an observation and some number of lagged observations.

Autoregressive Moving Average (ARMA)

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

The ARMA describes a weakly stationary stochastic time series in terms of two polynomials and combines an autoregressive model with a moving average model.

Moving Average (MA)

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

A MA model uses the dependency between an observation and a residual error from a moving average model applied to lagged variables.

Autoregressive Integrated Moving Average (ARIMA)

$$\left(1 - \sum_{i=1}^{p'} \alpha_i L^i\right) X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$$

ARIMA attempts to ‘explain’ a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

Machine-Learning-Based Approaches

Long Short Term Memory networks (LSTM)

LSTMs are a special kind of recurrent neural network (RNN), capable of learning long-term dependencies (Hochreiter & Schmidhuber, 1997).

XGBoost Open-source gradient boosting library.

Gradient boosting is a machine learning technique for regression that produces a prediction model from an ensemble of weak prediction models (Chen et al. 2016).

Tree-base pipeline optimization tool (TPOT)

TPOT is a Python Automated Machine Learning tool that optimizes machine learning pipelines using genetic programming (Olsen et al. 2016).

Performance Metrics

Forecast Bias

$$Bias = \frac{\sum_{i=1}^{n} (actual_i - predicted_i)}{n}$$

Mean Absolute Error (MAE)

$$MAE = \frac{\sum_{i=1}^{n} |actual_i - predicted_i|}{n}$$

Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(actual_i - predicted_i)^2$$

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(predicted_i - actual_i)^2}{n}}$$

Symmetric Mean Absolute Percentage Error (sMAPE)

$$\text{sMAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\frac{|predicted_t - actual_t|}{(|actual_t| + |predicted_t|)/2}$$

04

RESEARCH METHODS

Provide framework for the study

METHODOLOGY

Create Python Environment

1. Use Miniconda package manager
2. Run:
   conda env create -n <environment name> -f environment.yml

Data Munging [Wrangling]
(Import into Pandas DataFrame(s))

- Historical Hazelnut Yield Data
- Historical Climate Data

Perform Forecasting

1. Auto-regression (AR)
2. Moving-Average (MA)
3. Auto-regressive Moving-Average (ARMA)
4. Auto-regressive Integrated Moving-Average (ARIMA)
5. Season Auto-regressive Integrated Moving-Average (SARIMA)
6. Season Auto-regressive Integrated Moving-Average with Exogenous Regressors (SARIMAX)
7. XGBoost ML Algorithm
8. Long-Short Term Memory (LSTM) ML Algorithm
9. Pipeline Optimized (TPOT) Regressor

Plotly Graphs

Evaluate Model & Predictions

1. Forecast Bias
2. Mean Absolute Error (MAE)
3. Mean Squared Error (MSE)
4. Root Mean Squared Error (RMSE)
5. Symmetric Mean Absolute Percentage Error (sMAPE)

DATA

HISTORICAL HAZELNUT PRODUCTION

The hazelnut production data from 1927-2008 was obtained from the National Agricultural Statistics Service (NASS), Agricultural Statistics Board, US Department of Agriculture

Exploratory Data Analysis (EDA)

Snippet of Pandas DataFrame within Jupyter Notebook of Hazelnut Yield data

In [2]:

```
# read in hazelnut data obtained from the National Agricultural Statistics Service (NASS), Agricultural Statist
```

Out[2]:

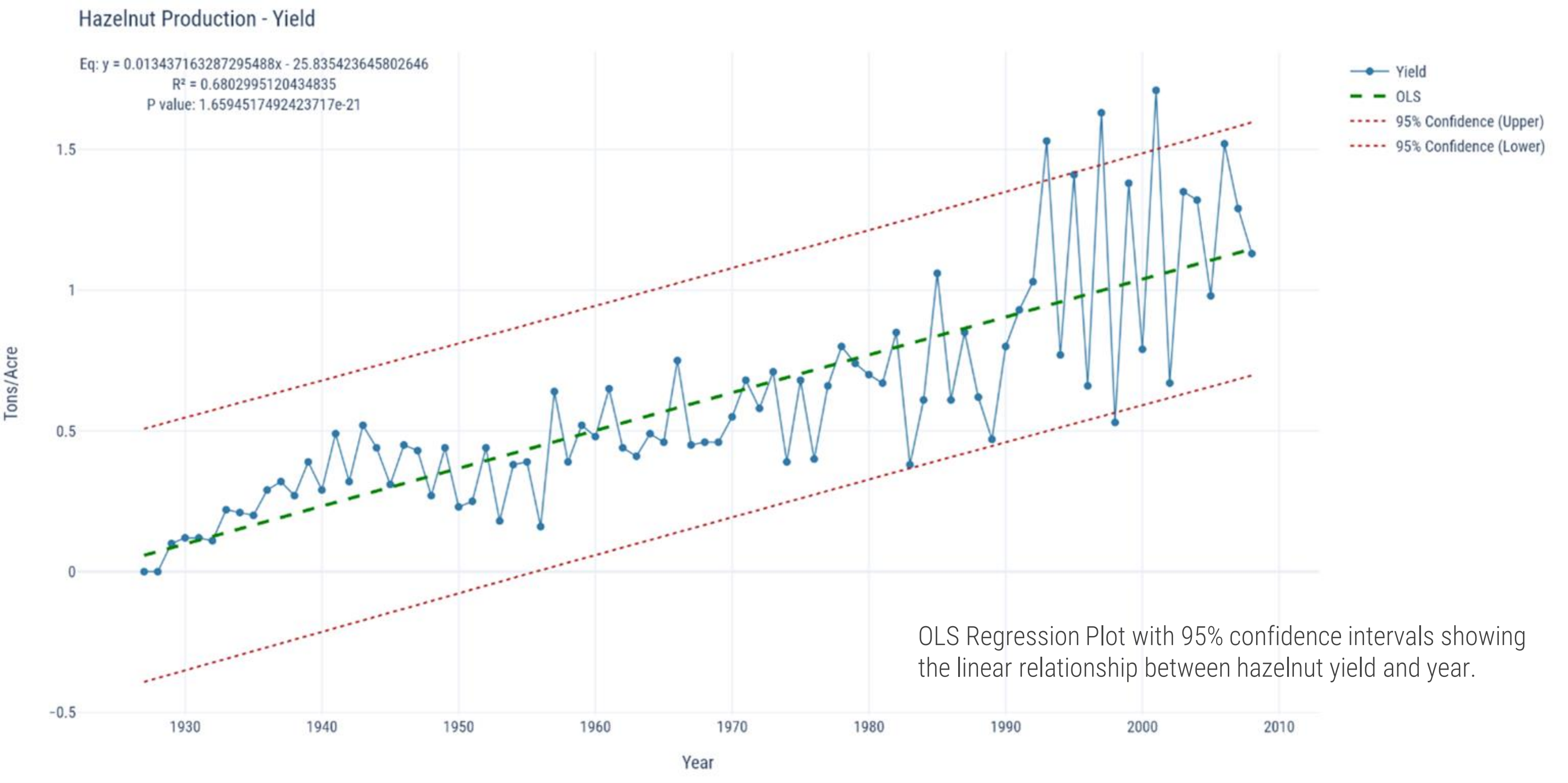
year	bearing (ac)	yield per acre (tons)	utilized production (tons)	production - meat (tons)	production - in-shell (tons)	production - shelled (tons)	price per ton (dollars)	value of production (1k dollars)
1927	0.0	0.00	60.0	0.0	0.0	0.0	320.0	19.0
1928	0.0	0.00	200.0	0.0	0.0	0.0	380.0	76.0
1929	2000.0	0.10	200.0	0.0	0.0	0.0	300.0	60.0
1930	2500.0	0.12	300.0	0.0	0.0	0.0	340.0	102.0
1931	3100.0	0.12	380.0	0.0	0.0	0.0	250.0	95.0

In [3]:

```
df_hazelnut.describe()
```

Out[3]:

	bearing (ac)	yield per acre (tons)	utilized production (tons)	production - meat (tons)	production - in-shell (tons)	production - shelled (tons)	price per ton (dollars)	value of production (1k dollars)
count	82.000000	82.000000	82.000000	82.000000	82.000000	82.000000	82.000000	82.000000
mean	18307.073171	0.602195	13285.853659	2137.878049	7737.707317	5241.975610	603.426829	11198.890244
std	8078.476431	0.387983	12191.606802	2575.789379	6722.945327	6398.339941	378.754490	15566.805948
min	0.000000	0.000000	60.000000	0.000000	0.000000	0.000000	200.000000	19.000000
25%	15350.000000	0.380000	5387.500000	239.250000	4002.500000	661.250000	344.500000	1955.750000
50%	18100.000000	0.490000	9125.000000	1107.000000	5965.000000	2820.000000	514.000000	3765.500000
75%	25925.000000	0.747500	17875.000000	3477.500000	9700.000000	6687.500000	785.250000	15096.500000
max	29200.000000	1.710000	49500.000000	12000.000000	32500.000000	30300.000000	2240.000000	75480.000000



OLS Regression Plot with 95% confidence intervals showing the linear relationship between hazelnut yield and year.

DATA CONTINUED

HISTORICAL CLIMATE

Weather data for the same period was obtained from the National Centers for Environmental Information of the National Oceanic and Atmospheric Administration (National Centers for Environmental Information (NCEI)

The climate data used are as follows:
Yearly averages of:

- maximum temperature
- extreme maximum temperature
- minimum temperature
- extreme minimum temperature
- average temperature
- cooling degree days (base 65)
- heating degree days (base 65)
- total precipitation
- highest daily total of precipitation

Also, yearly sums of:

- cumulative cooling degree days
- cumulative heating degree days
- cumulative total precipitation

PERFORM FORECASTS

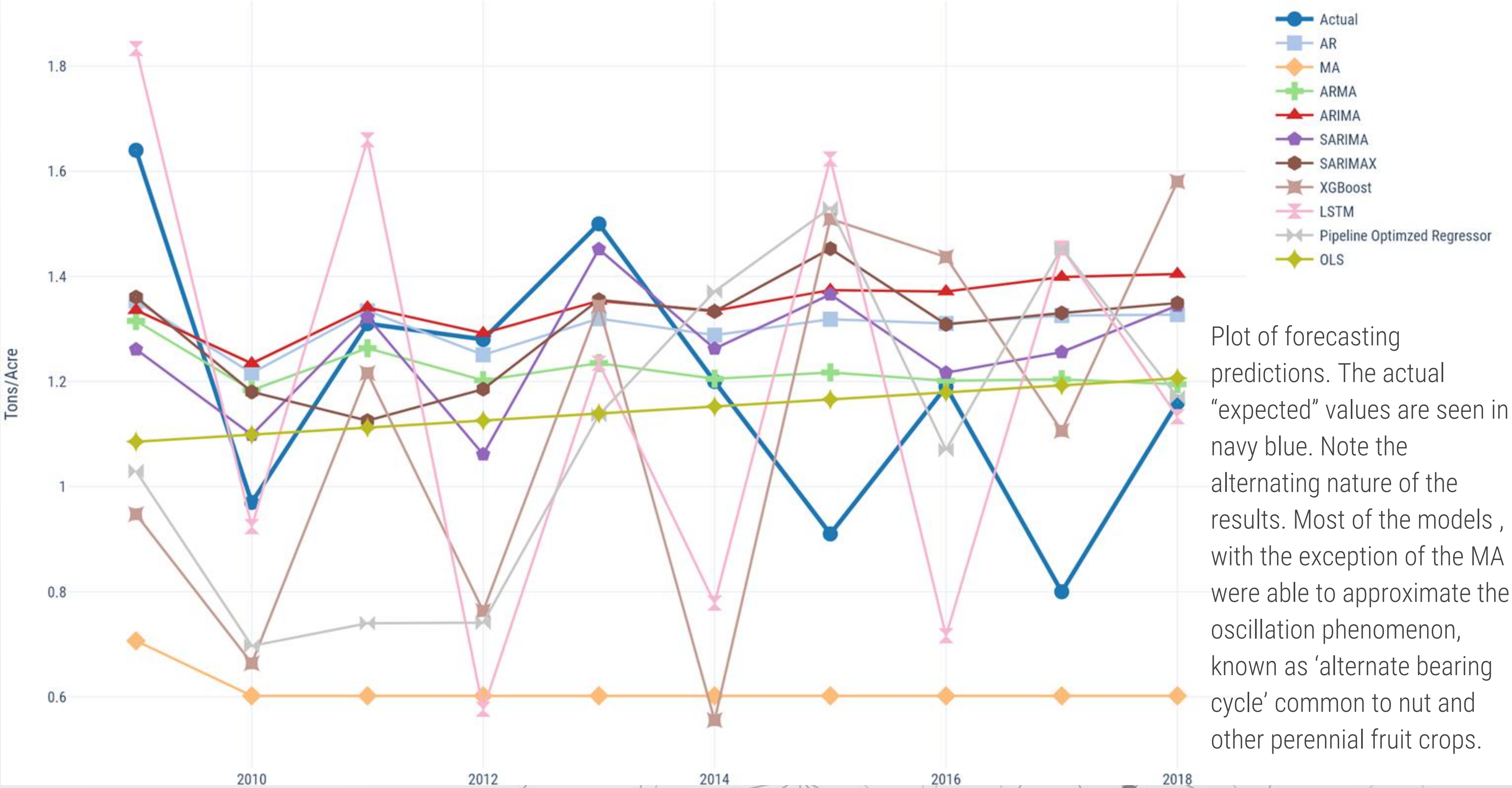
05

RESULTS

How did the forecasting methods perform?

Yield (tons/acre) predictions forecasted for each method.

Name	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Actual	1.640000	0.970000	1.310000	1.280000	1.500000	1.200000	0.910000	1.190000	0.800000	1.160000
AR	1.352874	1.216064	1.334589	1.250829	1.319576	1.288175	1.318609	1.310568	1.325383	1.327403
MA	0.706637	0.601988	0.601988	0.601988	0.601988	0.601988	0.601988	0.601988	0.601988	0.601988
ARMA	1.315397	1.183635	1.263407	1.202202	1.235059	1.205263	1.217305	1.201513	1.204343	1.194818
ARIMA	1.335782	1.234448	1.340632	1.291937	1.353530	1.334850	1.374000	1.371288	1.398811	1.404474
SARIMA	1.261183	1.098573	1.322436	1.061422	1.451809	1.262425	1.365675	1.216762	1.256021	1.344433
SARIMAX	1.360552	1.180230	1.125498	1.185502	1.355516	1.333329	1.452742	1.308704	1.330368	1.349351
XGBoost	0.947520	0.663889	1.215941	0.763554	1.344026	0.555659	1.509254	1.436868	1.106195	1.580379
LSTM	1.833676	0.924083	1.659814	0.576330	1.234864	0.778374	1.623380	0.716330	1.453801	1.133056
Pipeline Optimzed Regressor	1.029603	0.696848	0.740140	0.741205	1.136178	1.371373	1.530232	1.069919	1.454029	1.170473
OLS	1.085600	1.099000	1.112400	1.125800	1.139200	1.152600	1.166000	1.179400	1.192800	1.206200



Plot of forecasting predictions. The actual "expected" values are seen in navy blue. Note the alternating nature of the results. Most of the models , with the exception of the MA were able to approximate the oscillation phenomenon, known as 'alternate bearing cycle' common to nut and other perennial fruit crops.

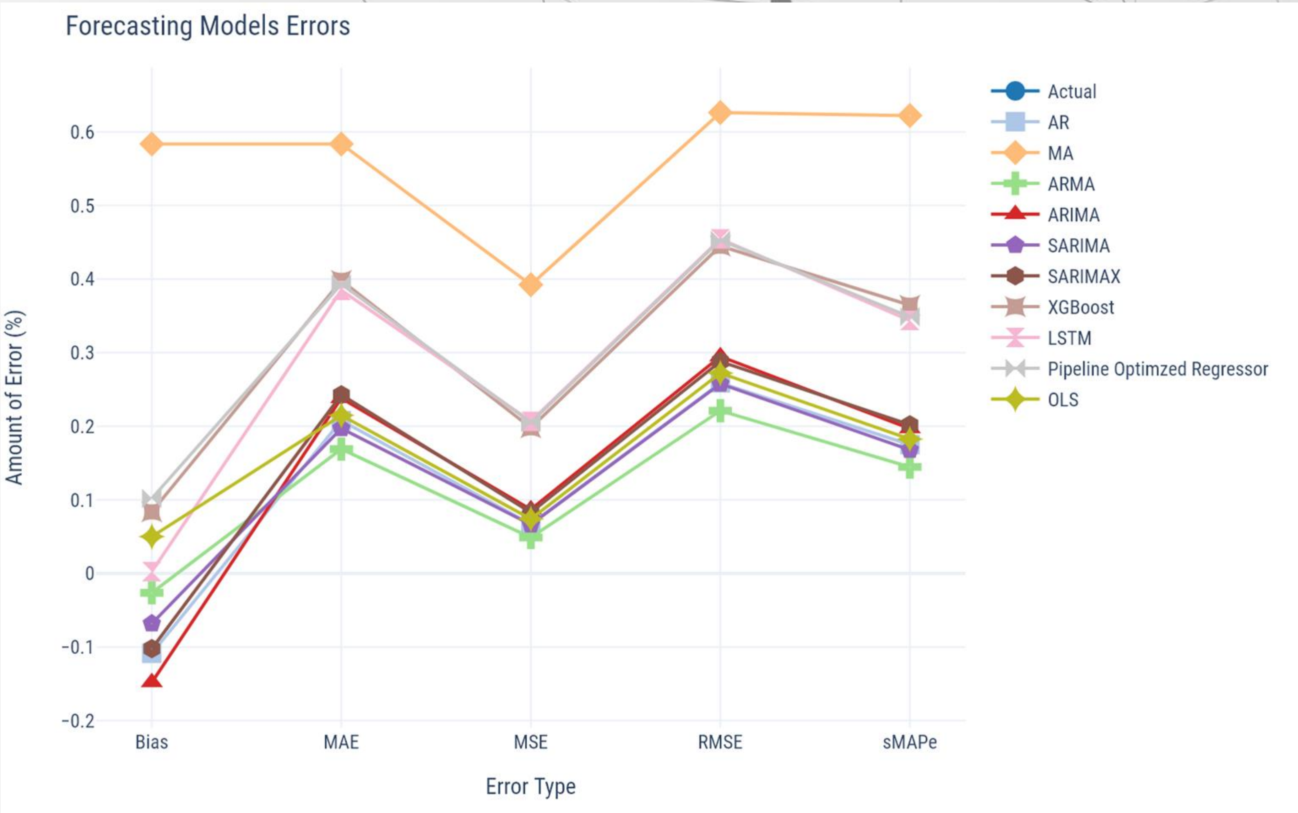
Differences of predicted and actual values for each model. Cells that are marked either green or red correspond to predicted values as being closest to or furthest from expected, actual value for each year, respectively.

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
AR	-0.287126	0.246064	0.024589	-0.029171	-0.180424	0.088175	0.408609	0.120568	0.525383	0.167403
MA	-0.933363	-0.368012	-0.708012	-0.678012	-0.898012	-0.598012	-0.308012	-0.588012	-0.198012	-0.558012
ARMA	-0.324603	0.213635	-0.046593	-0.077798	-0.264941	0.005263	0.307305	0.011513	0.404343	0.034818
ARIMA	-0.304218	0.264448	0.030632	0.011937	-0.146470	0.134850	0.464000	0.181288	0.598811	0.244474
SARIMA	-0.378817	0.128573	0.012436	-0.218578	-0.048191	0.062425	0.455675	0.026762	0.456021	0.184433
SARIMAX	-0.279448	0.210230	-0.184502	-0.094498	-0.144484	0.133329	0.542742	0.118704	0.530368	0.189351
XGBoost	-0.692480	-0.306111	-0.094059	-0.516446	-0.155974	-0.644341	0.599254	0.246868	0.306195	0.420379
LSTM	0.193676	-0.045917	0.349814	-0.703670	-0.265136	-0.421626	0.713380	-0.473670	0.653801	-0.026944
Pipeline Optimzed Regressor	-0.610397	-0.273152	-0.569860	-0.538795	-0.363822	0.171373	0.620232	-0.120081	0.654029	0.010473
OLS	-0.554400	0.129000	-0.197600	-0.154200	-0.360800	-0.047400	0.256000	-0.010600	0.392800	0.046200

Cells, either green or red, correspond to models having the lowest or highest error for the given metric.

Name	Bias	MAE	MSE	RMSE	sMAPe
AR	-0.108410	0.207750	0.067030	0.258900	0.174960
MA	0.583550	0.583550	0.392280	0.626320	0.622260
ARMA	-0.026290	0.169080	0.048870	0.221070	0.144670
ARIMA	-0.147980	0.238110	0.086970	0.294910	0.196900
SARIMA	-0.068070	0.197190	0.066450	0.257780	0.167860
SARIMAX	-0.102180	0.242770	0.082970	0.288050	0.201970
XGBoost	0.083670	0.398210	0.197880	0.444840	0.364500
LSTM	0.002630	0.384760	0.206670	0.454610	0.343760
Pipeline Optimzed Regressor	0.102000	0.393220	0.205090	0.452870	0.348730
OLS	0.050100	0.214900	0.074130	0.272270	0.182540

Errors were calculated using variety of metrics; Bias, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Symmetric Mean Absolute Percentage Error (sMAPE). Bias is the sum of the difference of expected to predicted values divided by the number of observations. A bias other than zero suggests a tendency of the model to over forecast (negative error) or under forecast (positive error). MAE is the sum of the absolute difference of expected and predicted values divided by the number of observations. MSE is the sum of the squares of the difference of expected to predicted values divided by the number of observations. RMSE is the root of MSE. SMAPE is an accuracy measure based on percentage (or relative) errors.



06

DISCUSSION/CONCLUSION

Takeaways and a look to the future...

TAKE-AWAYS

ALT. BEARING

Almost all methods successfully approximated the alternate bearing cycle pattern (as having the shape of a high then low prediction), with the notable exceptions of ARIMA and MA. These approaches produced almost a linear output.

ARMA SURPRISES

The ARMA method had the lowest errors and made the most accurate (nearest to actual) results of three of the ten years predicted. Overall, yield predictions using the ARMA model were within 0.44% to 40% of the actual value and within 22% of real yields for nine of the ten years forecast.

ML GROWING PAINS

The LSTM model produced the best predictions for two of the ten years forecast, yet also made two of the poorest; within 0.94% to 56.3% of expected values.

EXOGENIC VARIABLES

The SARIMAX and Pipeline Optimized Regressor methods leveraged climate data as exogenic inputs and performed better than expected, nearly fitting the alternating nature of the data.

VOLATILE DATA

Given that the yield appears to become increasingly variable following 1990, it is notable that the forecasting methods performed as well as they did.

PYTHON PERFORMS

A positive conclusion of this study illuminated that time-series forecasting is possible without the need for expensive software or specialized knowledge. Python programming is immensely powerful yet very approachable.

FURTHER STUDY

EXOGENIC FACTORS

More effort needs to be spent on the effects of exogenic factors such as the alternate bearing cycle of perennial crops and changing climate conditions.

EVER CHANGING AI

Time-series analysis and forecasting using machine-learning-based approaches is constantly evolving with new, better performant methods being developed continually.

MORE DATA ALWAYS HELPS

Additional data in the form of global crop yields and climate records could provide the basis for a more complete and thereby increasing accuracy in the resulting predictive model.

TARGETED INPUTS

The model may benefit from including Informative agronomy-specific data (such as soil type, nutrients, pH, electroconductivity, etc.) coupled with the other phenological information (date of bloom, duration of pollination, etc.) to establish a diverse dataset.

REMOTE SENSING AUGMENTATION

A battery of vegetative indices could be performed on remotely sensed multispectral satellite imagery at specific developmental stages; all aspects of crop health could be evaluated and included in a more complete model

CURRENT YEAR TRUTH DATA

Annually adding yield data as “truth” and refining the forecasting algorithm could make for a more accurate model

CONCLUSION

- TRADITIONAL AND MACHINE-LEARNING-BASED APPROACHES WERE EVALUATED
- A REASONABLY ACCURATE PREDICTION COULD BE MADE ON SPARSE DATA WITH SIGNIFICANT SEASONALITY
- PYTHON PROGRAMMING AND JUPTYER NOTEBOOK CONTAINERS ARE A VIABLE FRAMEWORK FOR FORECASTING
- THE JUPYTER NOTEBOOK CAN BE REUSED AND ADAPTED AS A TEMPLATE FOR FURTHER STUDY

QUESTIONS?