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

OBJECTIVES
What are the goals of this study?

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

OVERVIEW

RESEARCH METHODS
Provide framework for the study

RESULTS
How did the forecasting methods perform?

DISCUSSION/CONCLUSION
Takeaways and a look to the future...
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.
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
What are the goals of this study?
OBJECTIVES

No. 1 - PRIMARY

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

No. 2 - SECONDARY

“Perform predictions using the Python programming language within a novel Jupyter notebook”
03

BACKGROUND

List of terms and other art
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.

**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 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.

**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)

\[ sMAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|predicted_i - actual_i|}{(|actual_i| + |predicted_i|)/2} \]
RESEARCH METHODS

Provide framework for the study
METHODOLOGY

Create Python Environment
1. Use Miniconda package manager
2. Run:
   
   ```bash
   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

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)

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

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

```python
df_hazelnut.describe()
```
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
How did the forecasting methods perform?
Yield (tons/acre) predictions forecasted for each method.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>1.640000</td>
<td>0.970000</td>
<td>1.310000</td>
<td>1.280000</td>
<td>1.500000</td>
<td>1.200000</td>
<td>0.910000</td>
<td>1.190000</td>
<td>0.800000</td>
<td>1.160000</td>
</tr>
<tr>
<td>AR</td>
<td>1.352874</td>
<td>1.216064</td>
<td>1.334589</td>
<td>1.250829</td>
<td>1.319576</td>
<td>1.288175</td>
<td>1.318609</td>
<td>1.310568</td>
<td>1.325383</td>
<td>1.327403</td>
</tr>
<tr>
<td>MA</td>
<td>0.706637</td>
<td>0.601988</td>
<td>0.601988</td>
<td>0.601988</td>
<td>0.601988</td>
<td>0.601988</td>
<td>0.601988</td>
<td>0.601988</td>
<td>0.601988</td>
<td>0.601988</td>
</tr>
<tr>
<td>ARMA</td>
<td>1.315397</td>
<td>1.183635</td>
<td>1.263407</td>
<td>1.202202</td>
<td>1.235059</td>
<td>1.205263</td>
<td>1.217305</td>
<td>1.201513</td>
<td>1.204343</td>
<td>1.194818</td>
</tr>
<tr>
<td>ARIMA</td>
<td>1.335782</td>
<td>1.234448</td>
<td>1.340632</td>
<td>1.291937</td>
<td>1.353530</td>
<td>1.334850</td>
<td>1.374000</td>
<td>1.371288</td>
<td>1.398811</td>
<td>1.404474</td>
</tr>
<tr>
<td>SARIMA</td>
<td>1.261183</td>
<td>1.098573</td>
<td>1.322436</td>
<td>1.061422</td>
<td>1.451809</td>
<td>1.262425</td>
<td>1.365675</td>
<td>1.216762</td>
<td>1.256021</td>
<td>1.344433</td>
</tr>
<tr>
<td>SARIMAX</td>
<td>1.360552</td>
<td>1.180230</td>
<td>1.125498</td>
<td>1.185502</td>
<td>1.355516</td>
<td>1.333329</td>
<td>1.452742</td>
<td>1.308704</td>
<td>1.330368</td>
<td>1.349351</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.947520</td>
<td>0.663889</td>
<td>1.215941</td>
<td>0.763554</td>
<td>1.344026</td>
<td>0.555659</td>
<td>1.509254</td>
<td>1.436868</td>
<td>1.106195</td>
<td>1.580379</td>
</tr>
<tr>
<td>LSTM</td>
<td>1.833676</td>
<td>0.924083</td>
<td>1.659814</td>
<td>0.576330</td>
<td>1.234864</td>
<td>0.778374</td>
<td>1.623380</td>
<td>0.716630</td>
<td>1.453801</td>
<td>1.133056</td>
</tr>
<tr>
<td>Pipeline Optimzed Regressor</td>
<td>1.029603</td>
<td>0.696848</td>
<td>0.740140</td>
<td>0.741205</td>
<td>1.136178</td>
<td>1.371373</td>
<td>1.530232</td>
<td>1.069919</td>
<td>1.454029</td>
<td>1.170473</td>
</tr>
<tr>
<td>OLS</td>
<td>1.085600</td>
<td>1.099000</td>
<td>1.112400</td>
<td>1.125800</td>
<td>1.139200</td>
<td>1.152600</td>
<td>1.166000</td>
<td>1.179400</td>
<td>1.192800</td>
<td>1.206200</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>-0.287126</td>
<td>0.246064</td>
<td>0.024589</td>
<td>-0.029171</td>
<td>-0.180424</td>
<td>0.088175</td>
<td>0.408609</td>
<td>0.120568</td>
<td>0.525383</td>
<td>0.167403</td>
</tr>
<tr>
<td>MA</td>
<td>-0.933363</td>
<td>-0.368012</td>
<td>-0.708012</td>
<td>-0.678012</td>
<td>-0.898012</td>
<td>-0.598012</td>
<td>-0.308012</td>
<td>-0.588012</td>
<td>-0.198012</td>
<td>-0.558012</td>
</tr>
<tr>
<td>ARMA</td>
<td>-0.324603</td>
<td>0.213635</td>
<td>-0.046593</td>
<td>-0.077798</td>
<td>-0.264941</td>
<td>0.005263</td>
<td>0.307305</td>
<td>0.011513</td>
<td>0.404343</td>
<td>0.034818</td>
</tr>
<tr>
<td>ARIMA</td>
<td>-0.304218</td>
<td>0.264448</td>
<td>0.030632</td>
<td>0.011937</td>
<td>-0.146470</td>
<td>0.134850</td>
<td>0.464000</td>
<td>0.181288</td>
<td>0.598811</td>
<td>0.244474</td>
</tr>
<tr>
<td>SARIMA</td>
<td>-0.378817</td>
<td>0.128573</td>
<td>0.012436</td>
<td>-0.218578</td>
<td>-0.048191</td>
<td>0.062425</td>
<td>0.455675</td>
<td>0.026762</td>
<td>0.456021</td>
<td>0.184433</td>
</tr>
<tr>
<td>SARIMAX</td>
<td>-0.279448</td>
<td>0.210230</td>
<td>-0.184502</td>
<td>-0.094498</td>
<td>-0.144484</td>
<td>0.133329</td>
<td>0.542742</td>
<td>0.118704</td>
<td>0.530368</td>
<td>0.189351</td>
</tr>
<tr>
<td>XGBoost</td>
<td>-0.692480</td>
<td>-0.306111</td>
<td>-0.094059</td>
<td>-0.516446</td>
<td>-0.155974</td>
<td>-0.644341</td>
<td>0.599254</td>
<td>0.246868</td>
<td>0.306195</td>
<td>0.420379</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.193676</td>
<td>-0.045917</td>
<td>0.349814</td>
<td>-0.703670</td>
<td>-0.265136</td>
<td>-0.421626</td>
<td>0.713380</td>
<td>-0.473670</td>
<td>0.653801</td>
<td>-0.026944</td>
</tr>
<tr>
<td>Pipeline Optimzed Regressor</td>
<td>-0.610397</td>
<td>-0.273152</td>
<td>-0.569860</td>
<td>-0.538795</td>
<td>-0.363822</td>
<td>0.171373</td>
<td>0.620232</td>
<td>-0.120081</td>
<td>0.654029</td>
<td>0.010473</td>
</tr>
<tr>
<td>OLS</td>
<td>-0.554400</td>
<td>0.129000</td>
<td>-0.197600</td>
<td>-0.154200</td>
<td>-0.360800</td>
<td>-0.047400</td>
<td>0.256000</td>
<td>-0.010600</td>
<td>0.392800</td>
<td>0.046200</td>
</tr>
</tbody>
</table>
Cells, either green or red, correspond to models having the lowest or highest error for the given metric.

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

Takeaways and a look to the future...
TAKE-AWAYS

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.

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.

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.

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

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.

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

<table>
<thead>
<tr>
<th>EXOGENIC FACTORS</th>
<th>TIME-SERIES ANALYSIS AND FORECASTING USING MACHINE-LEARNING-BASED APPROACHES IS CONSTANTLY EVOLVING WITH NEW, BETTER PERFORMANT METHODS BEING DEVELOPED CONTINUALLY.</th>
</tr>
</thead>
<tbody>
<tr>
<td>EVER CHANGING AI</td>
<td>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.</td>
</tr>
<tr>
<td>MORE DATA ALWAYS HELPS</td>
<td>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.</td>
</tr>
<tr>
<td>TARGETED INPUTS</td>
<td>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.</td>
</tr>
<tr>
<td>REMOTE SENSING AUGMENTATION</td>
<td>ANNUALLY ADDING YIELD DATA AS &quot;TRUTH&quot; AND REFINING THE FORECASTING ALGORITHM COULD MAKE FOR A MORE ACCURATE MODEL.</td>
</tr>
<tr>
<td>CURRENT YEAR TRUTH DATA</td>
<td></td>
</tr>
</tbody>
</table>
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 JUPYTER NOTEBOOK CONTAINERS ARE A VIABLE FRAMEWORK FOR FORECASTING

THE JUPYTER NOTEBOOK CAN BE REUSED AND ADAPTED AS A TEMPLATE FOR FURTHER STUDY
QUESTIONS?