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 * October 2020*

Abstract:
Traditional and machine-learning-based forecasting methods were used to predict hazelnut yield (tons per acre per year) within Oregon's Willamette Valley based on historical crop production and exogenous climate variables. Autoregressive Moving Average (ARMA) and Long Short Term Memory (LSTM) methods performed the best having the lowest errors and bias, respectively. 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 Moving Average (MA) model performed the poorest, not having the structure or complexity to account for the alternate bearing cycle typical to nut crops.
Keywords: hazelnuts, time-series forecasting, Python, Jupyter Notebooks, phenology, regression, boosted trees, and LSTM (Long-Short Term Memory) Machine Learning.

Introduction
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 (Taghavi et al. 2018). Hazelnuts are unique in that they pollinate in the winter as opposed to the spring. The hazelnut was named Oregon's state nut in 1989 due to its historical and economic significance (Oregon State Facts | Oregon.com n.d.). George Doris of Springfield planted the first commercial orchard of 200 trees in 1903 (*A brief history of Oregon hazelnuts* 2020). Even though the hazelnut is a relatively recent addition to North America, it has held a position of vital importance since the Mesolithic age (McComb and Simpson 1999). Globally it ranks fifth in overall tree nut production (behind the pistachio) (Wills 2019). According to Oregon State's College of Agricultural Sciences, Department of Horticulture, Oregon produces nearly 100% of US Hazelnuts, primarily in the Willamette Valley (Hazelnut Production | College of Agricultural Sciences | Oregon State n.d.). The prevalence of "eastern filbert blight" (a fungus common to eastern states) has limited production outside Oregon. However, the planted acreage of hazelnuts has increased due to the development of "filbert blight' resistant cultivars, a collaborative effort from biologists from Rutgers University and Shawn Mehlenbacher of Oregon State University (*Shawn Mehlenbacher n.d.*). The National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA) lists the total acreage of bearing crops to be 44,000 acres with a production of 51,000 tons in 2018 (*USDA - National Agricultural Statistics Service 2018).*

Objectives
The goal of this study was to forecast hazelnut yield (tons/acre/year) in Oregon's Willamette Valley using various traditional regression methods and machine-learning-based counterparts. Additionally, the yield was predicted with a Python programming language within a novel Jupyter Notebook.

Background
Man has endeavored to model the physical world for as long as he has inhabited it. Mathematics, namely the branch of Physics, is in itself an effort to encapsulate or explain physical phenomena in a mathematical system or simulation (O'Connor and Robertson n.d.). Phenology can be thought of as a temporal component in the biophysical world. Stated differently, phenology is the annual cyclic rhythm of the living world. In more scientific terms, it can be thought of as such: *"Phenology is generally described as the art of observing life cycle phases or activities of plants and animals in their temporal occurrence throughout the year."* (Lieth 1974). He also described *phenology* as a sort of "agricultural meteorology." Phenological modeling is an extensively broad subject, nearly as diverse in breadth as the living world it attempts to simulate. Several influential studies have been conducted on various deciduous trees to establish forecasting models for bloom date and potential yields. Popular approaches leverage total historical harvest and climate data to forecast potential yields (Fornaciari et al. 2005; Luedeling et al. 2009a; Oteros et al. 2013). Others, such as the "Fruit & Nut Research & Information Center," a division of the University of California Agriculture and Natural Resources (UC ANR) (Fruit & Nut Research & Information Center n.d.) have used "budbreak" dates (Overview - Bloom and Leaf-Out Models for Tree Crops n.d.) to forecast future tree phenology. Yet as Chuine et al.

(2014) noted in their study, models using only "budbreak" date as an indicator do not accurately predict future phenological elements, such as endodormancy. Other studies (also involving wind-pollinated species, like the hazelnut) have focused on the relationship between flowering duration and fruit production as a means of developing "pollen indexes as indicators of flowering, evaluating in some cases the predictive role of the variable." Several studies have used the cumulative chilling or heating requirements of various nut-bearing species as a means of modeling future phenology (Mehlenbacher 1991; Heide 1993; Pope et al. 2015; Rahemi and Pakkish 2009; Luedeling et al. 2009b). In a study involving the effects of "attenuation of photosynthetically active radiation (PAR),"; lower hazelnut production seemed to correspond with reduced light penetration within the tree canopy (Hampson et al. 1996). PAR is also correlated with solar radiation (Gómez et al. 1998).

While the primary goal of this study is to perform forecasting and not develop a phenological model in the purest sense, it is notable that differing climate conditions affect crops at critical developmental stages, thus enhancing or diminishing crop yield and quality. In pistachios and other nuts, late rains during bloom (pollination) can disrupt fruit set and the formation and significantly increase the likelihood of disease. (*Panicle and Shoot Blight of Pistachio: A Major Threat to the California Pistachio Industry*).

List of Terms

Autoregression (AR)

The notation $AR(p)$ indicates an autoregressive model of order p. The AR(p) model is defined as

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

where $\varphi_i, \cdots, \varphi_p$ are the parameters of the model, c is a constant, and ε_t is white noise. AR is a time series model that uses the dependent relationship between an observation and some number of lagged observations.

Moving Average (MA)

The notation $MA(q)$ refers to the moving average model of order q. The MA(q) model is defined as:

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

where μ is the mean of the series, the $\theta_1, \ldots, \theta_q$ are the parameters of the model, and the ε_t, ε_{t-1},ε_{t-q} are white noise error terms. The value of q is called the order of the MA model. 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)

The notation $ARMA(p, q)$ refers to the model with p autoregressive terms and q moving-average terms. This model contains the $AR(p)$ and $MA(q)$ models,

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

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)

Given a time series data X_t where t is an integer index and the X_t are real numbers, an $ARMA(p', q)$ model is given by

$$X_t - \alpha_1 X_{t-1} - \cdots - \alpha_{p'} X_{t-p'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

or equivalently by

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

where L is the lag operator, the α_i are the parameters of the autoregressive part of the model, the θ_i are the parameters of the moving average function and the ε_t are error terms. The error terms ε_t are generally assumed to be independent, identically distributed variables sampled from a normal distribution with zero mean.

- An ARIMA(0, 1, 0) model (or I(1) model) is given by $X_t = X_{t-1} + \varepsilon_t$ – which is simply a random walk.
- An ARIMA(0, 1, 0) with a constant, given by $X_t = c + X_{t-1} + \varepsilon_t$ – which is a random walk with drift.
- An ARIMA(0, 0, 0) model is a white noise model.
- An ARIMA(0, 1, 2) model is a Damped Holt's model.
- An ARIMA(0, 1, 1) model without constant is a basic exponential smoothing model.[5]
- An ARIMA(0, 2, 2) model is given by

$$X_t = 2X_{t-1} - X_{t-2} + (\alpha + \beta - 2)\varepsilon_{t-1} + (1 - \alpha)\varepsilon_{t-2} + \varepsilon_t$$

– which is equivalent to Holt's linear method with additive errors, or double exponential smoothing.[5]

To determine the order of a non-seasonal ARIMA model, a useful criterion is the *Akaike Information Criterion* (AIC). It is written as

$$AIC = -2\log(L) + 2(p+q+k)$$

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).

Forecast Bias or *Mean Forecast Error* is given by

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

Mean Absolute Error (MAE) given by

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

Mean Squared Error (MSE) given by:

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

Root Mean Squared Error (RMSE) given by:

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

Symmetric Mean Absolute Percentage Error (sMAPE) given by:

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

Research Approach and Methods

While performing data acquisition, it quickly became apparent that the goal of developing a diverse and robust phenologic model would be untenable, given the limited data available. As a result, the project focus transitioned into one of forecasting crop yields. While a career could be spent evaluating the best forecasting models, a broad, yet balanced selection of traditional methods, complemented by machine-learning-based derivatives was considered as part of this study. This limited, structured approach attempts to cover popular forecasting strategies without deviating too far from the primary project focus. The secondary goal of the study pertains to programming and the suitability of open-source methods for analytical tasks. For this purpose, the Python programming language and a Jupyter notebook ecosystem were employed as the apparatus for conducting the study. Jupyter Notebooks are a modern marvel that provides a free, open-source web-based container to perform iterative, interactive, visual computing utilizing various popular programming languages (Project Jupyter | Home n.d.). The flow of the project is outlined in the figure below.

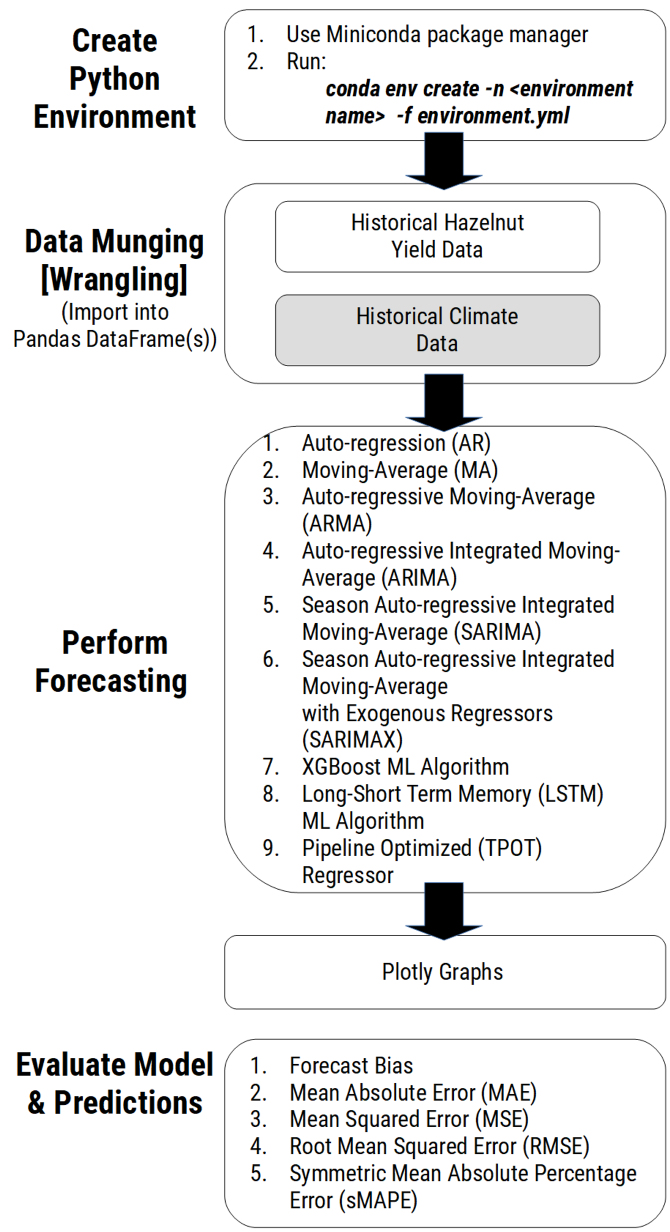

Figure 1 - Project Methodology

The preliminary step of this study was to obtain historical production data for hazelnuts within Oregon's Willamette Valley, as well as historical climate data used as exogenous variables (for the appropriate predictive models).
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

The hazelnut production data from 1927-2008 was obtained from the National Agricultural Statistics Service (NASS), Agricultural Statistics Board, US Department of Agriculture (CITE). 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) n.d.). The data were then munged and imported into a Jupyter Notebook using the Pandas (Pandas n.d.) library. A predictive model was prepared using a battery of time-series forecasting methods. Generally, the hazelnut yield (in the form of tons per acre per year to account for fluctuations in acreages farmed) for the years as mentioned above was used as "inputs" and the supplemental crop production for the years of 2009-2018, obtained from Table 5 of the Hazelnut Marketing Board, Preliminary Annual Report, Crop Year 2018, were used as "truth" values, and to evaluate the performance of each forecasting method. Before the predictive methods were attempted and evaluated, general exploratory data analysis (EDA) was performed to understand the historic hazelnut production data better. The dependent variable, "yield per acre (tons) per year," is variable and not normally distributed.

Quantile statistics

Minimum	0
5-th percentile	0.12
Q1	0.38
median	0.49
Q3	0.7475
95-th percentile	1.4085
Maximum	1.71
Range	1.71
Interquartile range (IQR)	0.3675

Descriptive statistics

Standard deviation	0.3879831497
Coefficient of variation (CV)	0.6442814555
Kurtosis	0.7713358641
Mean	0.602195122
Median Absolute Deviation (MAD)	0.195
Skewness	1.069256145
Sum	49.38
Variance	0.1505309244
Monotocity	Not monotonic

Table 1 - Pandas Profiling Report - Statistical summary of hazelnut yield per acre per year in Oregon's Willamette Valley.

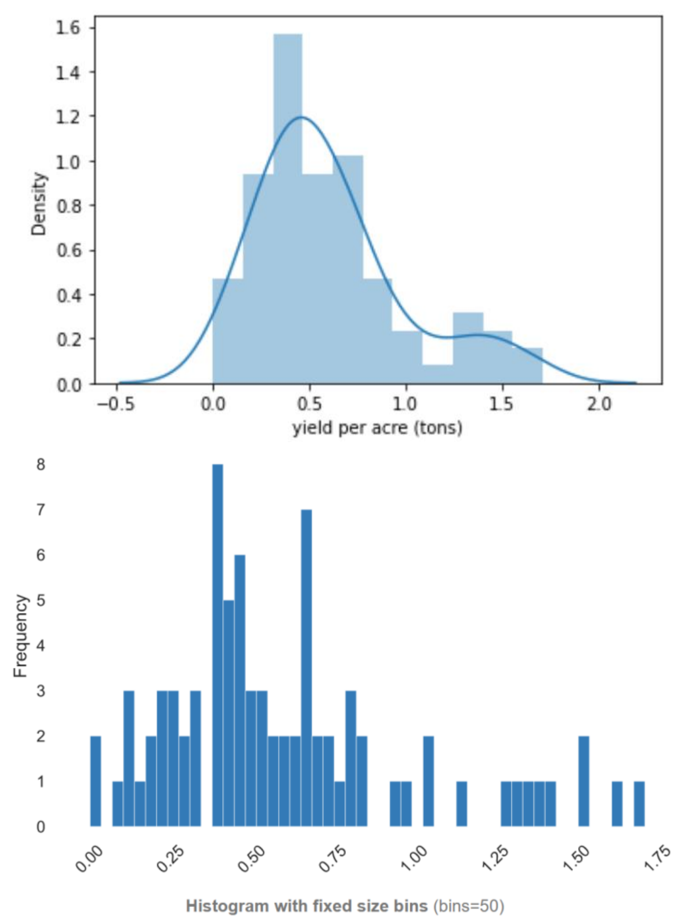

Figure 2 - Univariate distribution of observations, density (t), and frequency (b) plots of dependent variable "yield per acre (tons) per year."

Next, Ordinary Least Squares regression was performed using the Statsmodels library (Seabold et al. 2010). The dependent variable, "yield per acre (tons) per year," is variable and not normally distributed. As you can see in figure 5, the data have a linear trend, and the OLS regression has an R^2 fit of 0.68. The regression results and plot are seen in Figure 2 on the following page.

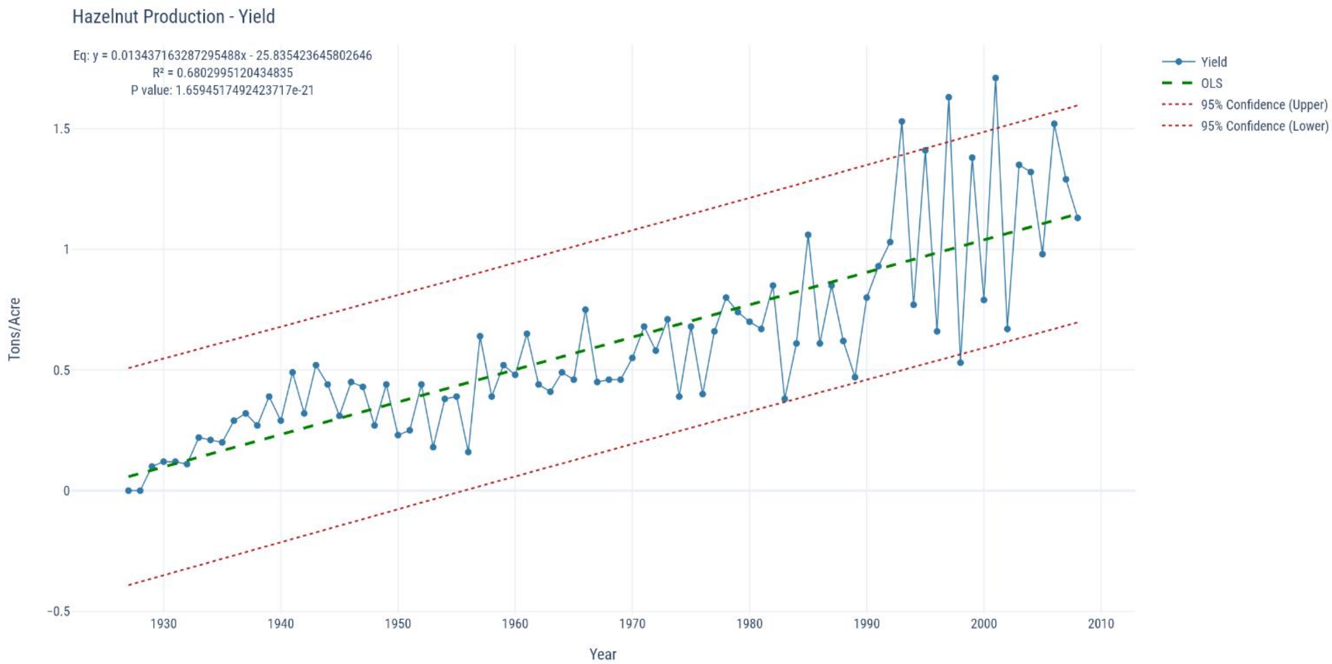

Figure 3 – OLS Regression Plot with 95% confidence intervals showing the linear relatinship between hazelnut yeild and year.

One particular item of note is the "on-off" alternate bearing cycle typical to perennial crops, wherein alternating years produce either greater-than or lesser-than average crops (Alternate Bearing n.d.). In these particular hazelnut orchards, the production of even years is less-than-average, while the odd years produce more significant than average crop yields. This "alternate bearing" phenomenon is visibly evident in the OLS plot seen in Fig. 3 above. It is also notable that while the data is reasonably linear in trend, it becomes increasingly volatile and variable following 1990. It is unclear as to the factors leading to this increase in volatility as exogenous indicators reveal no apparent causation. One possible explanation is a change in farming practices to increase yield in "on" bearing years, resulting in lower than average (previous) "off" bearing years.

Following the data analysis, a predictive model was fashioned utilizing the forecasting methods mentioned previously. A traditional predictive model could be fashioned from the OLS fit as follows:

$$Yield(tons/acre/year) = -25.835 + 0.0134(year)$$

$$R^2 = 0.682$$

$$P_{value} = 1.695e^{-21}$$

For each method, input data was parsed into 'yield per acre (tons)' for the years of 1927-2008, the appropriate predictive method applied and evaluated against the "truth" data of crop yields for the years of 2009-2018. Error metrics were calculated for each method in the form of Forecast Bias, Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and Symmetric Mean Absolute Percentage Error. Finally, graphs and tables were generated for predicted outcomes, differences, and errors.

Statsmodels (Seabold et al. 2010) and *pmdarima* (Smith et al. 2017) were the primary libraries used for the traditional forecasting methods. XGBoost (Chen et al. 2016), LSTM (Hochreiter et al. 1997), and tree-based pipeline optimized regressor algorithms (Olsen et al. 2016) were used as the machine-learning-based methods. The variables used in each algorithm (either from tool defaults or derived from experimentation) are listed below for reproducibility.

Autoregression (AR):

- data = hazelnut yield (ton/acre/year)
- lags = 5

Moving-Average (MA):

- data = hazelnut yield (ton/acre/year)
- model ARMA
- order = (0,1)

Autoregressive Moving-Average (ARMA):
- data = hazelnut yield (ton/acre/year)
- model ARMA
- order = (2,1)

Autoregressive Integrated Moving-Average (ARIMA):
- data = hazelnut yield (ton/acre/year)
- Best model: ARIMA(2,1,1)(0,0,0)[0] intercept
- stepwise search to minimize AIC
- Pmdarima implementation

Seasonal Autoregressive Integrated Moving-Average (SARIMA):
- data = hazelnut yield (ton/acre/year)
- Best model: ARIMA(2,0,1)(0,1,1)[7]
- stepwise search to minimize AIC
- Pmdarima implementation

Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors (SARIMAX):
- data = hazelnut yield (ton/acre/year)
- exogenous variables = climate date
- order = (1,1,1)
- seasonal order = (1,1,1,2)

XGBoost:
- data = hazelnut yield (ton/acre/year)
- model = XBGRegressor
- objective = 'reg:squaredlogerror'
- n_estimators = 100
- max_tree_depth = 5

LSTM:
- data = hazelnut yield (ton/acre/year)
- train/test split = 0.85/0.15
- batch_size = 1
- epochs = 500
- neurons = 3

Tree Optimized Pipeline:
- data = exogenous climate + hazelnut yield (ton/acre/year)
- model = TPOTRegressor
- generations = 50
- population_size = 50
- scoring = 'neg_mean_absolute_error'
- cv (evaluation procedure) = RepeatedKFold(
    - n_splits = 10
    - n_repeats = 3
    - random_state = 43)
- verbosity = 2
- random_state = 43
- n_jobs = -1

Project Results

Following the forecasting, the results were compiled into a data frame (Table 3, below) and plotted for added understanding (Figure. 4, following page).

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Name										
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

Table 3 - Yield (tons/acre) predictions forecasted for each method.

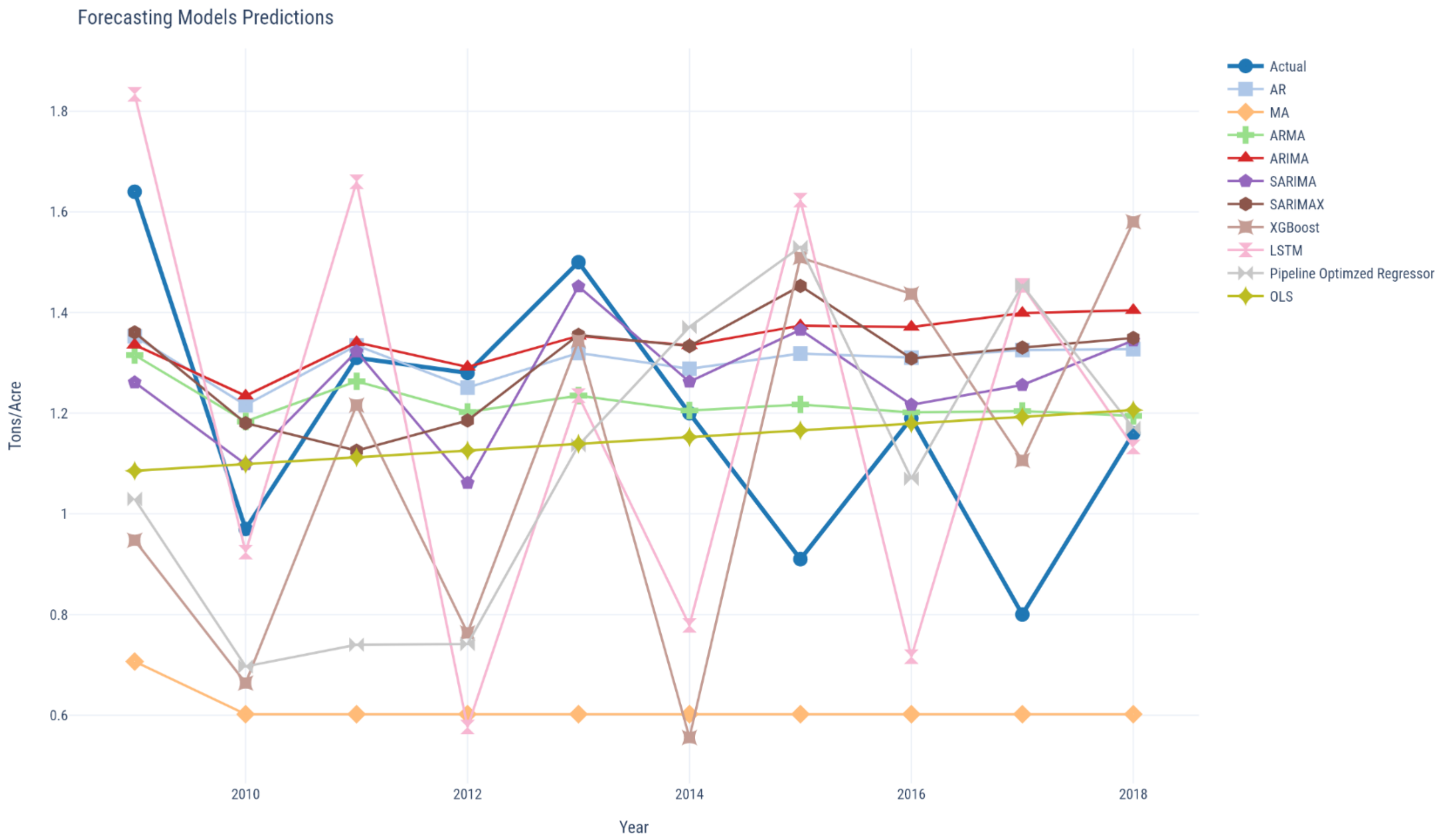

Figure 4 - Plot of forecasting errors. The actual "expected" values are seen in navy blue. Note the alternating nature of the results. Most of the models , with the exception of the MA were able to approximate the oscillation phenomenon, known as 'alternate bearing cycle' common to nut and other periennial fruit crops.

	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

Table 4 - 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 the expected, actual value for each year, respectively

Figure 5 - Forecasting Model Errors. Errors were calculated using variety of metrics; Bias, Mean Absoulte 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 absoute 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.

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

Table 5 - Forecasting Model Errors. Cells, either green or red, correspond to models having the lowest or highest error for the given metric.

Discussion

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, SARIMA, and LSTM methods produced the best results at predicting the potential crop yields (in terms of having the greatest occurrences of predictions being closest to actual values). In contrast, the MA (Moving Average) performed the poorest. 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.

Two of the machine-learning-based algorithms, LSTM and XGBoost, had low forecast bias, but other accuracy errors were moderate. 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. It is thought that with additional parameter tuning and training that these methods could challenge the traditional ones for accuracy and precision. Of all the methods attempted, only two leveraged exogenic variables in the form of climate data.

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. It is thought that the XGBoost and LSTM machine-learning-based models could be made more accurate, given more helpful parameter selections and possibly leveraging the climate data as exogenous inputs. The Pipeline Optimized Regressor model benefitted with automatic hyperparameter selection wherein a model with the "best fit" is used for forecasting.

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

While the primary goal of this study was to evaluate a selection of forecasting methods to predict future hazelnut harvest yields, judging the suitability of Jupyter Notebooks (and by extension, Python) for data analysis was an equally important secondary one. The maturity and diversity of libraries available to the Python community are impressive, always improving, and continually growing in count and features. All tables, graphs, outputs, and the logic to perform them as a part of this study were conducted within a novel Jupyter Notebook. Unlike other academic approaches that discuss methods and theory on a superficial level, this notebook is included as a part of this paper and intended to be used as a template for future study with logic, data, visualization, and results made visible to the audience.

Further Study

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. While attempts have been made to include machine-learning time-series forecasting methods in addition to traditional approaches, the science of prediction is an ever-evolving pursuit. 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. Moving past climate, it may be beneficial to further enhance the structure of the model by 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. Additionally, a battery of vegetative indices could be performed on remotely sensed multispectral satellite imagery at specific developmental stages (August - catkin, December - pollen shed, and March - end of pollination (González-Naharro, et al. 2019), middle of ovule development), and zonal statistics of the results (min/mean/max), that is, if specific orchard locations could be known at the time of the study. In this manner, all aspects of the overall crop health could be evaluated and included in a more complex multivariate model; yield, climate, observable phenomena, etc. In addition to a better understanding of exogenic factors and their inclusion in a more complex predictive model, more time needs to be spent in better tuning machine-learning-based approaches and their parameters to reduce bias and errors and increase forecasting accuracy. Looking forward, each of the models could be enhanced by adding the "truth" values (yields for 2009-2018) as observations and then utilized for predicting out beyond 2020.

Conclusion

This study evaluated several traditional autoregressive algorithms and their more complex machine-learning-based counterparts to develop a forecasting pipeline rather than a broad overview or survey of statistical methods. Using various approaches, a reasonably accurate prediction could be made on crop production, even with sparse data with significant seasonality present. The Python programming language and its vast collection of modules and libraries and the Jupyter Notebook ecosystem provided the framework for this study and run without requiring the entire model to be recomputed, saving valuable time and encouraging experimentation. Affording one the ability to perform all aspects of labors; data collection, filtering, preparation, analysis, and visualization within a self-contained web-based (offline) container is powerful, convenient, and useful. A notebook can be shared or reused as a template for future studies. Another positive conclusion of this study was to illuminate that time-series forecasting is possible without the need for expensive software or specialized knowledge. Python programming is immensely powerful yet very approachable.

performed admirably. The compartmentalized nature of the iterative notebook allowed portions of code to be revised and re-

References

Chen, T., & Guestrin, C. (2016). XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939785

Chuine, Isabelle, Marc Bonhomme, Jean-Michel Legave, Iñaki de Cortazar-Atauri, Guillaume Charrier, André Lacointe, and Thierry Améglio. 2014. Can Phenological Models Predict Tree Phenology Accurately under Climate Change Conditions?

Fornaciari, Marco, Fabio Orlandi, and Bruno Romano. 2005. "Yield Forecasting for Olive Trees: A New Approach in a Historical Series (Umbria, Central Italy)." Agronomy Journal. https://doi.org/10.2134/agronj2005.0067.

Gómez, Iván, Eduardo Pérez-Rodríguez, Benjamín Viñegla, Félix L. Figueroa, and Ulf Karsten. 1998. "Effects of Solar Radiation on Photosynthesis, UV-Absorbing Compounds and Enzyme Activities of the Green Alga Dasycladus Vermicularis from Southern Spain." Journal of Photochemistry and Photobiology B: Biology. https://doi.org/10.1016/S1011-1344(98)00199-7.

González-Naharro, Rocío, Elia Quirós, Santiago Fernández-Rodríguez, Inmaculada Silva-Palacios, José María Maya-Manzano, Rafael Tormo-Molina, Raúl Pecero-Casimiro, Alejandro Monroy-Colin, and Ángela Gonzalo-Garijo. 2019. "Relationship of NDVI and oak (Quercus) pollen including a predictive model in the SW Mediterranean region." Science of the Total Environment.

Hampson, Cheryl R., Anita N. Azarenko, and John R. Potter. 1996. "Photosynthetic Rate, Flowering, and Yield Component Alteration in Hazelnut in Response to Different Light Environments." Journal of the American Society for Horticultural Science 121 (6): 1103–11. https://doi.org/10.21273/JASHS.121.6.1103.

Heide, Om. 1993. "Daylength and Thermal Time Responses of Budburst During Dormancy Release." Physiologia Plantarum. https://doi.org/10.1034/j.1399-3054.1993.880401.x.

Hochreiter, Sepp; Schmidhuber, Jürgen. 1997. "Long short-term memory". Neural Computation. 9 (8): 1735–1780. doi:10.1162/neco.1997.9.8.1735. PMID 9377276. S2CID 1915014.

Lieth, Helmut. Phenology and Seasonality Modeling. Berlin, Heidelberg, New York: Springer, 1974.

Luedeling, Eike, Minghua Zhang, Volker Luedeling, and Evan H. Girvetz. 2009a. "Sensitivity of Winter Chill Models for Fruit and Nut Trees to Climatic Changes Expected in California's Central Valley." Agriculture, Ecosystems and Environment. https://doi.org/10.1016/j.agee.2009.04.016.

Luedeling, Eike, Minghua Zhang, Gale McGranahan, and Charles Leslie. 2009b. "Validation of Winter Chill Models Using Historic Records of Walnut Phenology." Agricultural and Forest Meteorology. https://doi.org/10.1016/j.agrformet.2009.06.013.

McComb, Anne M G, and Derek Simpson. 1999. "The Wild Bunch: Exploitation of the Hazel in Prehistoric Ireland." Ulster Journal of Archaeology 58: 1–16. http://www.jstor.org/stable/20568226.

Mehlenbacher, Shawn A. 1991. "Chilling Requirements of Hazelnut Cultivars." Scientia Horticulturae. https://doi.org/10.1016/0304-4238(91)90010-V.

Olson, R. S., Bartley, N., Urbanowicz, R. J., & Moore, J. H. 2016. "Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science." Proceedings of the 2016 on Genetic and Evolutionary Computation Conference - GECCO '16. doi:10.1145/2908812.2908918Oteros, J., H. García-Mozo, C. Hervás, and C. Galán. 2013. "Biometeorological and Autoregressive Indices for Predicting Olive Pollen Intensity." International Journal of Biometeorology. https://doi.org/10.1007/s00484-012-0555-5.

Pope, Katherine S., Volker Dose, David Da Silva, Patrick H. Brown, and Theodore M. DeJong. 2015. "Nut Crop Yield Records Show That Budbreak-Based Chilling Requirements May Not Reflect Yield Decline Chill Thresholds." International Journal of Biometeorology. https://doi.org/10.1007/s00484-014-0881-x.

Rahemi, Majid, and Zahra Pakkish. 2009. "Determination of Chilling and Heat Requirements of Pistachio (Pistacia Vera L.) Cultivars." Agricultural Sciences in China. https://doi.org/10.1016/S1671-2927(08)60281-3.

Seabold, Skipper, and Josef Perktold. 2010. "statsmodels: Econometric and statistical modeling with python." Proceedings of the 9th Python in Science Conference

Smith, Taylor G., et al. 2017. "pmdarima: ARIMA estimators for Python" http://www.alkaline-ml.com/pmdarima

Taghavi, T., Dale, A., Saxena, P., Galic, D., Rahemi, A., Kelly, J., & Suarez, E. (2018). Flowering of hazelnut cultivars and how it relates to temperature in southern Ontario. Acta Horticulturae. https://doi.org/10.17660/ActaHortic.2018.1226.18

Wills, Matthew. 2019. Everything You Wanted to Know about Hazelnuts but Were Afraid to Ask. 14 November. https://daily.jstor.org/everything-you-wanted-to-know-about-hazelnuts-but-were-afraid-to-ask/.

2018-05-04, USDA - National Agricultural Statistics Service - Statistics By Subject Results https://quickstats.nass.usda.gov/results/F40CAC5D-46FF-391D-845E-7AB78D432F3F?pivot=short_desc

n.d. Alternate Bearing. Alternate Bearing - Fruit & Nut Research & Information Center. http://fruitandnuteducation.ucdavis.edu/generaltopics/Tree_Growth_Structure/Alternate_Bearing/.

n.d. A Brief History of Oregon Hazelnuts http://oregonhazelnuts.org/a-brief-history-of-oregon-hazelnuts/.

n.d. Fruit & Nut Research & Information Center. http://fruitsandnuts.ucdavis.edu/.

n.d. Hazelnut Production | College of Agricultural Sciences | Oregon State. https://agsci.oregonstate.edu/tree-fruits-and-nuts/hazelnut-production.

n.d. Model | Definition of Model by Merriam-Webster. https://www.merriam-webster.com/dictionary/model.

n.d. O'Connor, J J, and E F Robertson. Physical World. http://mathshistory.st-andrews.ac.uk/HistTopics/World.html.

n.d. Overview - Bloom and Leaf-Out Models for Tree Crops. https://ucanr.edu/sites/bloom/.

n.d Pandas. https://pandas.pydata.org/.

n.d. Panicle and Shoot Blight of Pistachio: A Major Threat to the California Pistachio Industry https://www.apsnet.org/edcenter/apsnetfeatures/Pages/Pistachio.aspx

n.d. Phenology | Definition of Phenology by Merriam-Webster. https://www.merriam-webster.com/dictionary/phenology.

n.d. Pistachio Bloom Cast - Fruit & Nut Research & Information Center. http://fruitsandnuts.ucdavis.edu/Weather_Services/Bloom_Cast/.

n.d. Project Jupyter | Home. https://jupyter.org/.

n.d. Shawn Mehlenbacher https://horticulture.oregonstate.edu/users/shawn-mehlenbacher

Appendix A - Jupyter (Python) Environment

To install, save the contents of this appendix as "environment.yml". Next, from a working Anaconda or Miniconda prompt, enter:

conda env create -n <environment name> -f environment.yml


```yaml
name: capstone
channels:
  - conda-forge
  - defaults
dependencies:
  - argon2-cffi=20.1.0=py38h1e8a9f7_1
  - attrs=20.2.0=pyh9f0ad1d_0
  - backcall=0.2.0=pyh9f0ad1d_0
  - backports=1.0=py_2
  - backports.functools_lru_cache=1.6.1=py_0
  - blas=1.0=mkl
  - bleach=3.1.5=pyh9f0ad1d_0
  - brotlipy=0.7.0=py38h1e8a9f7_1000
  - ca-certificates=2020.6.20=hecda079_0
  - certifi=2020.6.20=py38h32f6830_0
  - chardet=3.0.4=py38h32f6830_1006
  - chart-studio=1.1.0=pyh9f0ad1d_0
  - colorama=0.4.3=py_0
  - colorlover=0.3.0=py_0
  - confuse=1.3.0=pyh9f0ad1d_0
  - cryptography=3.1=py38hba49e27_0
  - cufflinks-py=0.17.3=py_0
  - cycler=0.10.0=py_2
  - decorator=4.4.2=py_0
  - defusedxml=0.6.0=py_0
  - entrypoints=0.3=py38h32f6830_1001
  - freetype=2.10.2=hd328e21_0
  - htmlmin=0.1.12=py_1
  - icc_rt=2019.0.0=h0cc432a_1
  - icu=67.1=h33f27b4_0
  - idna=2.10=pyh9f0ad1d_0
  - imagehash=4.1.0=pyh9f0ad1d_0
  - importlib-metadata=1.7.0=py38h32f6830_0
  - importlib_metadata=1.7.0=0
  - intel-openmp=2019.4=245
  - ipykernel=5.3.4=py38h5ca1d4c_0
  - ipython=7.18.1=py38h1cdfbd6_0
  - ipython_genutils=0.2.0=py_1
  - ipywidgets=7.5.1=pyh9f0ad1d_1
  - jedi=0.17.2=py38h32f6830_0
  - jinja2=2.11.2=pyh9f0ad1d_0
  - joblib=0.16.0=py_0
  - jpeg=9d=he774522_0
  - jsonschema=3.2.0=py38h32f6830_1
  - jupyter=1.0.0=py_2
  - jupyter_client=6.1.7=py_0
  - jupyter_console=6.2.0=py_0
  - jupyter_core=4.6.3=py38h32f6830_1
  - kiwisolver=1.2.0=py38heaebd3c_0
  - libblas=3.8.0=14_mkl
  - libcblas=3.8.0=14_mkl
  - libclang=10.0.1=default_hf44288c_1
  - liblapack=3.8.0=14_mkl
  - libpng=1.6.37=ha81a0f5_2
  - libsodium=1.0.17=h2fa13f4_0
  - libtiff=4.1.0=h885aae3_6
  - llvmlite=0.34.0=py38h74e2f34_1
  - lz4-c=1.9.2=h62dcd97_2
  - m2w64-gcc-libgfortran=5.3.0=6
  - m2w64-gcc-libs=5.3.0=7
  - m2w64-gcc-libs-core=5.3.0=7
  - m2w64-gmp=6.1.0=2
  - m2w64-libwinpthread-git=5.0.0.4634.697f757=2
  - markupsafe=1.1.1=py38h9de7a3e_1
  - matplotlib=3.3.1=1
  - matplotlib-base=3.3.1=py38hfb9ee82_1
  - missingno=0.4.2=py_1
  - mistune=0.8.4=py38h9de7a3e_1001
  - mkl=2019.4=245
  - mkl-service=2.3.0=py38hfa6e2cd_0
  - msys2-conda-epoch=20160418=1
  - nbconvert=5.6.1=py38h32f6830_1
  - nbformat=5.0.7=py_0
  - networkx=2.5=py_0
  - notebook=6.1.3=py38h32f6830_0
  - numba=0.51.2=py38h251f6bf_0
  - numpy=1.19.1=py38h72c728b_0
  - olefile=0.46=py_0
  - openssl=1.1.1g=he774522_1
  - packaging=20.4=pyh9f0ad1d_0
  - pandas=1.1.2=py38h7ae7562_0
  - pandas-profiling=2.9.0=pyh9f0ad1d_0
  - pandoc=2.10.1=he774522_0
  - pandocfilters=1.4.2=py_1
  - parso=0.7.1=pyh9f0ad1d_0
  - patsy=0.5.1=py_0
  - phik=0.10.0=py_0
  - pickleshare=0.7.5=py38h32f6830_1001
  - pillow=7.2.0=py38h7011068_1
  - pip=20.2.3=py_0
  - plotly=4.9.0=pyh9f0ad1d_0
  - prometheus_client=0.8.0=pyh9f0ad1d_0
  - prompt-toolkit=3.0.7=py_0
  - prompt_toolkit=3.0.7=0
  - pycparser=2.20=pyh9f0ad1d_2
  - pygments=2.6.1=py_0
  - pyopenssl=19.1.0=py_1
  - pyparsing=2.4.7=pyh9f0ad1d_0
  - pyqt=5.12.3=py38h6538335_1
  - pyrsistent=0.16.0=py38h9de7a3e_0
  - pysocks=1.7.1=py38h32f6830_1
  - python=3.8.5=h60c2a47_7_cpython
  - python-cufflinks=0.17.3=py_0
  - python-dateutil=2.8.1=py_0
  - python_abi=3.8=1_cp38
  - pytz=2020.1=pyh9f0ad1d_0
  - pywavelets=1.1.1=py38h1e00858_2
  - pywin32=227=py38hfa6e2cd_0
  - pywinpty=0.5.7=py38_0
  - pyzmq=19.0.2=py38h77b9d75_0
  - qt=5.12.6=hb2cf2c5_0
  - qtconsole=4.7.7=pyh9f0ad1d_0
  - qtpy=1.9.0=py_0
  - requests=2.24.0=pyh9f0ad1d_0
  - retrying=1.3.3=py_2
  - scikit-learn=0.23.2=py38hf00eced_0
  - scipy=1.5.0=py38h9439919_0
  - seaborn=0.11.0=0
  - seaborn-base=0.11.0=py_0
  - send2trash=1.5.0=py_0
  - setuptools=49.6.0=py38h32f6830_0
  - six=1.15.0=pyh9f0ad1d_0
  - sqlite=3.33.0=he774522_0
  - tangled-up-in-unicode=0.0.6=pyh9f0ad1d_0
  - terminado=0.8.3=py38h32f6830_1
  - testpath=0.4.4=py_0
  - threadpoolctl=2.1.0=pyh5ca1d4c_0
  - tk=8.6.10=he774522_0
  - tornado=6.0.4=py38hfa6e2cd_0
  - tqdm=4.48.2=pyh9f0ad1d_0
  - traitlets=5.0.4=py_0
  - urllib3=1.25.10=py_0
  - vc=14.1=h869be7e_1
  - visions=0.5.0=pyh9f0ad1d_0
  - vs2015_runtime=14.16.27012=h30e32a0_2
  - wcwidth=0.2.5=pyh9f0ad1d_1
  - webencodings=0.5.1=py_1
  - wheel=0.35.1=pyh9f0ad1d_0
  - widgetsnbextension=3.5.1=py38h32f6830_1
  - win_inet_pton=1.1.0=py38_0
  - wincertstore=0.2=py38_1003
  - winpty=0.4.3=4
  - xlrd=1.2.0=pyh9f0ad1d_1
  - xz=5.2.5=h62dcd97_1
  - yaml=0.2.5=he774522_0
  - zeromq=4.3.2=ha925a31_3
  - zipp=3.1.0=py_0
  - zlib=1.2.11=h62dcd97_1009
  - zstd=1.4.5=h1f3a1b7_2
  - pip:
      - cffi==1.14.2
      - cython==0.29.17
      - enum34==1.1.10
      - keras==2.2.4
      - keras-applications==1.0.8
      - plaidml==0.7.0
      - plaidml-keras==0.7.0
      - pmdarima==1.7.1
      - pyqt5-sip==4.19.18
      - pyqtwebengine==5.12.1
      - pyyaml==5.3.1
      - sktime==0.4.1
      - statsmodels==0.11.1
      - xgboost==1.3.0-SNAPSHOT
prefix: C:\Users\jbiagio\Miniconda3\envs\capstone
```

Appendix B - Jupyter Notebook

Available on request.

To request a copy of the Jupyter Notebook, send an email to jason.biagio@gmail.com, subject "Capstone Jupyter Notebook request".