Forecasting Tomato/Bag Quantity Using Predicted Tomato Yield on the Basis of NDVI Values and Weather-Related Variables

By

Alisa Babikova, Anupriya Thirumurthy, Hyejeong Lee

Supervisor: Dr. Arnab Bose

A Capstone Project

Submitted to the University of Chicago in partial fulfillment
of the requirements for the degree of

Master of Science in Analytics

Graham School of Continuing Liberal and Professional Studies

May, 2019

The Capstone Project committee for Alisa Babikova, Anupriya Thirumurthy, Hyejeong Lee
Certifies that this is the approved version of the following capstone project report:

Forecasting Tomato/Bag Quantity Using Predicted Tomato Yield on the Basis of NDVI Values and Weather-Related Variables

Approved by Supervising Committee:

Dr. Arnab Bose	Dr. Sema Barlas

Abstract

Scholle IPN's current demand forecasting methods only consider the quantity of bags sold in the previous year leading to overstocking. The proposed solution is to predict the tomato yield on the basis of weather data, NDVI values, and previous year's yield and use it to forecast the tomato bag demand more accurately. The solution is built using two models: (1) random forest regressor to estimate annual tomato yield, and (2) linear regression to model the tomato bag sales. The most important weather measurement in predicting tomato yield is maximum temperature. Using the predicted tomato yield improves the forecast accuracy by 2.75%.

Keywords: *Tomato, Scholle IPN, NDVI, Temperature, Crop Yield, Cross-sectional, Linear Regression, Random Forest, Ensemble Model*

Executive Summary

Scholle IPN is a manufacturing company specializing in bag-in-box packaging for food, beverage, and industrial markets. Current demand forecasting methods only consider the quantity of bags sold in the previous year leading to inventory management challenges such as overstocking. Our research aims to help Scholle IPN increase the accuracy of the tomato bag demand forecast based on the annual tomato yield. The exploratory analysis of the historical tomato bag sales shows that the tomato bag sales are highly clustered in California. As tomato yield value is not immediately available before the sales season, we predict the annual tomato yield value on the basis of Normalized Difference Vegetation Index (NDVI) of the crop area, maximum temperature levels for the planting season months, and the previous year's yield. The predicted yield using these parameters is then used to model the current year's tomato bag sales along with the previous year's tomato bag sales using supervised machine learning algorithms such as linear regression and random forest regressor. Using RMSE, sMAPE, % bias, and mean accuracy as our metrics we determine the linear regression to be the best model for modeling tomato bag sales from the predicted tomato yield. Our intention is to use this model as a correction model for the ensembled forecast generated by other University of Chicago MScA Scholle IPN tomato teams on the basis of economic and demographic data, social media trends, and related tomato products.

Table of Contents

List of Figures

List of Tables

Preface

This report is intended to be read as a standalone document. However, for full context, the other University of Chicago MScA Mega Project tomato reports should be considered. Our report focuses on the narrow range of potential predictors (namely, NDVI Values, Weather Variables, and Tomato Yield), so a full evaluation of the University of Chicago MScA Mega Projech research and modelling is advised.

Introduction

Client Overview

Scholle IPN is a global manufacturing company present on five continents. Founded in 1945, the company is a pioneer in bag-in-box packaging and the first to produce aseptic bags and equipment for the tomato industry in the U.S. Scholle IPN produces packaging for food, beverage, and industrial markets. With highly automated equipment and a vertical integration process, the company meets rigorous manufacturing standards ensuring its products are safe, sustainable, and economic.

Problem Statement

One of Scholle IPN's primary markets is flexible packaging for temporary storage of processed-tomato products. Current demand forecasting methods only consider the quantity of bags sold in the previous year leading to inventory management challenges such as overstocking. The company believes that tomato bag sales are driven by tomato yield. In the absence of actual tomato yield for the sales season, the company cannot accurately predict the inventory needed to meet the sales demand.

Research Purpose

The purpose of this research is twofold: (1) to estimate the annual tomato yield on the basis of weather, annual NDVI values of the farmland, and previous year's tomato yield and (2) to forecast annual tomato bag sales on the basis of predicted annual tomato yield.

Variables and Scope

Independent Variables

Weather

The daily weather data is collected from Dark Sky, an open source for weather information, using their API. The scope of this research is limited to eight counties in California responsible for most of the processed tomato production (Kings, Kern, Merced, Fresno, San Joaquin, Solano, Stanislaus, and Yolo). The weather data ranges from 2010 to 2018 and includes only the months of the planting season (Feb - June) for every year as shown in Figure 1.

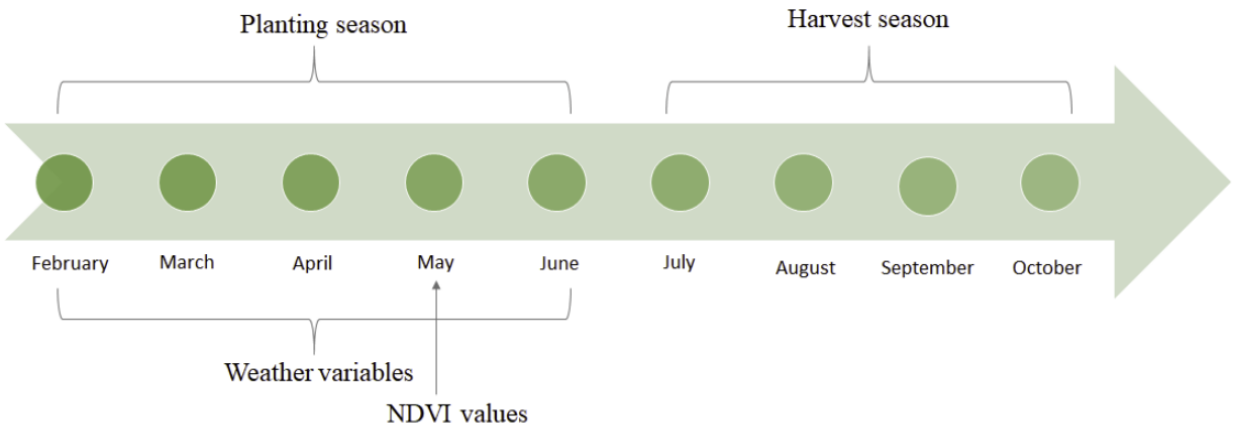

Figure 1: Timeline of Tomato Production and Independent Variables

The measurements, specific to tomato growth (refer to: Background p.2), include temperature, precipitation levels (precipitation probability and precipitation intensity), amount of direct sunlight, dew point and air pressure (refer to: Appendix). The weather data is then aggregated into minimum, average, and maximum levels of each of the weather measurements (Table 1). We then extract the weather variables that explain the most variability in the annual tomato yield using correlation analysis, random forest feature importances, lasso regression coefficients, and stepwise regression (refer to: Modeling Framework, Feature Selection). Maximum temperature is the only variable along with the previous year's yield that affect the tomato yield positively (Table 2). Table 3 shows the correlation between Maximum Temperature (Max_Temp) and tomato yield and its increase from 42% in 2016 to 75% in 2017.

NDVI

The NDVI values are sourced from the interactive vegetation explorer tool VegScape provided by the United States Department of Agriculture (USDA). NDVI stands for Normalized Difference Vegetation Index and measures the level of greenness of the crop fields

Table 1: Transformed Weather Variables for 8 Counties in CA

Weather_Measurement	Variables.in.Dataset.Max..Min..Avg.Levels
Temperature	Max_Temp, Avg_Temp, Min_Temp
Precipitation Intensity	Max_PrInt, Avg_PrInt, MinPrInt
Precipitation Probability	Max_PrProb, Avg_PrProb, Min_PrProb
Sunlight	Max_Sunlight, Avg_Sunlight, Min_Sunlight
Dew Point	Max_DewPoint, Avg_DewPoint, Min_DewPoint
Air Pressure	Max_Pressure, Avg_Pressure, Min_Pressure

Table 2: Correlation Analysis of 2016 Maximum Temperature and NDVI, 2015 Yield, and 2016 Yield

	Max_Temp	NDVI	Yield2015	Y_Yield2016
Max_Temp	1.00	-0.34	0.49	0.42
NDVI	-0.34	1.00	-0.90	-0.80
Yield2015	0.48	-0.90	1.00	0.80
Y_Yield2016	0.42	-0.79	0.80	1.00

ranging from 0 (low levels of greenness) to 1 (high level of greenness) according to VegScape Data. The NDVI values range from 2010 to 2018 and include only the months of the planting season (Feb - June) for every year. We then conduct stepwise regression, linear regression, and correlation analysis to assess each month of the planting season separately to determine which month explains the most variability in the annual tomato yield. Based on this assessment (refer to: Modeling Framework, Feature Selection), we pick May to be the most representative NDVI value for the entire year.

Dependent Variables

Tomato Yield

Historical tomato yield data is obtained from the United States Department of Agriculture (USDA). This data ranges from 2010 to 2017 and measures annual production of the tomatoes for processing in tons per acre. Annual tomato yield, the first dependent variable, is estimated on the basis of weather variables, NDVI values, and previous year's yield.

Tomato Bag Sales

The historical tomato bag sales data is provided by Scholle IPN for 2010 - 2018. The second dependent variable is total annual quantity of tomato bag sold in California on the basis of the predicted annual tomato yield. The scope of this research accounts only for California with county level of analysis. The independent variables are the predicted tomato yield for 2018 and 2017 tomato bag sales provided by Scholle IPN. The tomato bag sales forecast is for 2018.

Table 3: Correlation Analysis of 2017 Maximum Temperature and NDVI, 2016 Yield, and 2017 Yield

	Max_Temp	NDVI	Yield2016	Y_Yield2017
Max_Temp	1.00	-0.90	0.51	0.75
NDVI	-0.90	1.00	-0.78	-0.88
Yield2016	0.51	-0.78	1.00	0.77
Y_Yield2017	0.75	-0.88	0.77	1.00

Background

The weather variables represent the conditions pertinent to tomato growth. However, some of them can affect the tomato yield more strongly than others as weather is challenging to model. NDVI (Normalized Difference Vegetation Index) serves as an overall measure of the greenness of the crop land. Excessive amounts of precipitation and changing temperatures due to more frequently occurring rainfall and storms affect the crops as the irrigation of the soil needs to be maintained at a specific level. Frequent rain could increase the acidity of the soil that, in turn, could destroy necessary nutrients and minerals. The optimal temperature for the tomato growth is between 65 and 85 degrees Fahrenheit. Previous research shows that temperatures are crucial to the formation of the plant and fruit, and higher temperatures lead to an increased early tomato yield (16). Overall tomato growth can be monitored through assessing the amount of precipitation, humidity, temperature, air pressure and dew point. We believe that accounting for these in our model will significantly improve the accuracy of the prediction of the annual tomato yield, and subsequently, the tomato bag sales.

Methodology

Explanatory Data Analysis

Scholle IPN Data

After exploring the percentage of total quantity of bags sold by Ship-to-State (states Scholle IPN ships the bag orders to), we notice that 90% of the entire tomato bag sales volume is driven by shipments to California throughout most of the years. The graph below (Figure 2) represents the percentage of total quantity sold to California vs. other states. Out of 18.4 million of tomato bags ordered in United States, 16.6 million are destined for California.

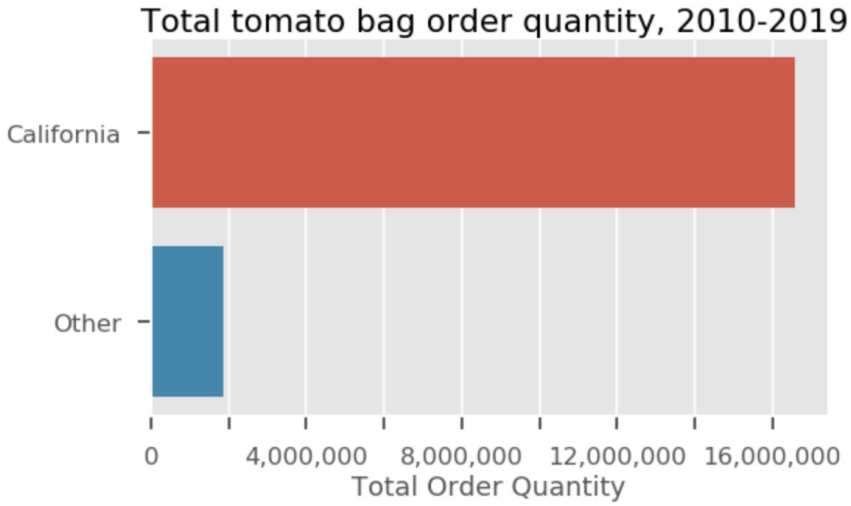

Figure 2: Scholle IPN's Order Quantity by Ship-to-State

The graph below (Figure 3) shows the units of tomato bags sold in other states excluding California sales. We notice that four other states contribute to 98% of the remaining tomato bag sales volume. Ohio, Arizona, Florida and Iowa account for 1,872,733 units of tomato bags. These four states, however, only contribute to 10% of the total demand while California contributes to 90% of the total demand. Therefore, our research focuses on California state alone.

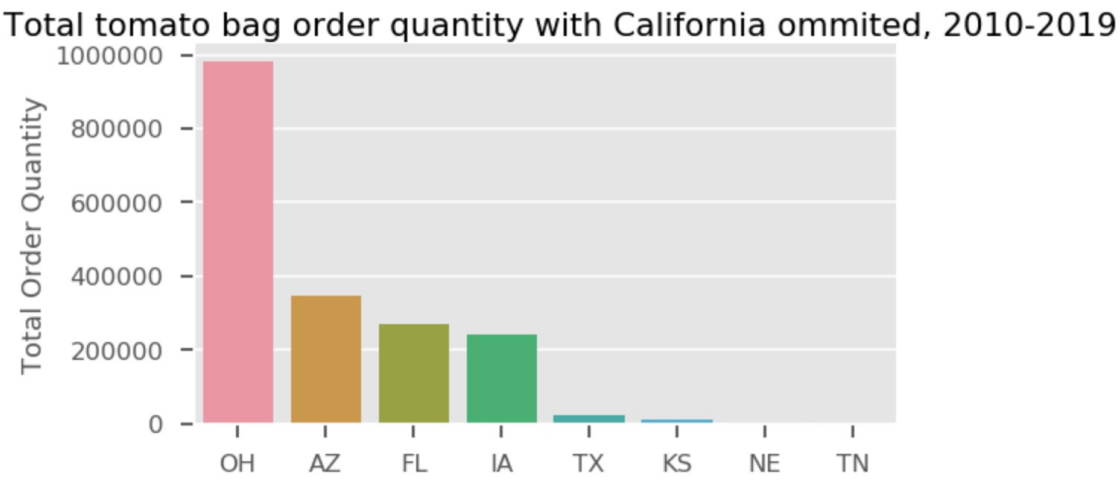

Figure 3: Total Tomato Bag Order Quantity with California Omitted, 2010 - 2018

County-level Tomato Yield

When we look at aggregated total yearly tomato yield by county level in California, we notice that the most recent drop in yield is in 2017 (Figure 4). The tomato yield drops in 2011, 2013, 2015, and 2017 with the lowest drop in 2013.

Figure 4: Aggregated Tomato Yield from 2010 to 2017

Weather Data

We assess the correlation between different levels of the weather variables. It is not surprising that the average, maximum, and minimum levels of the same weather variable are highly correlated. What is interesting is the correlation (Figure 5) between different levels of the different weather variables. For example, Maximum Sunlight (Max_Sunlight) is negatively correlated with Maximum Dew Point (Max_DewPoint) and Minimum Temperature levels (Min_Temp).

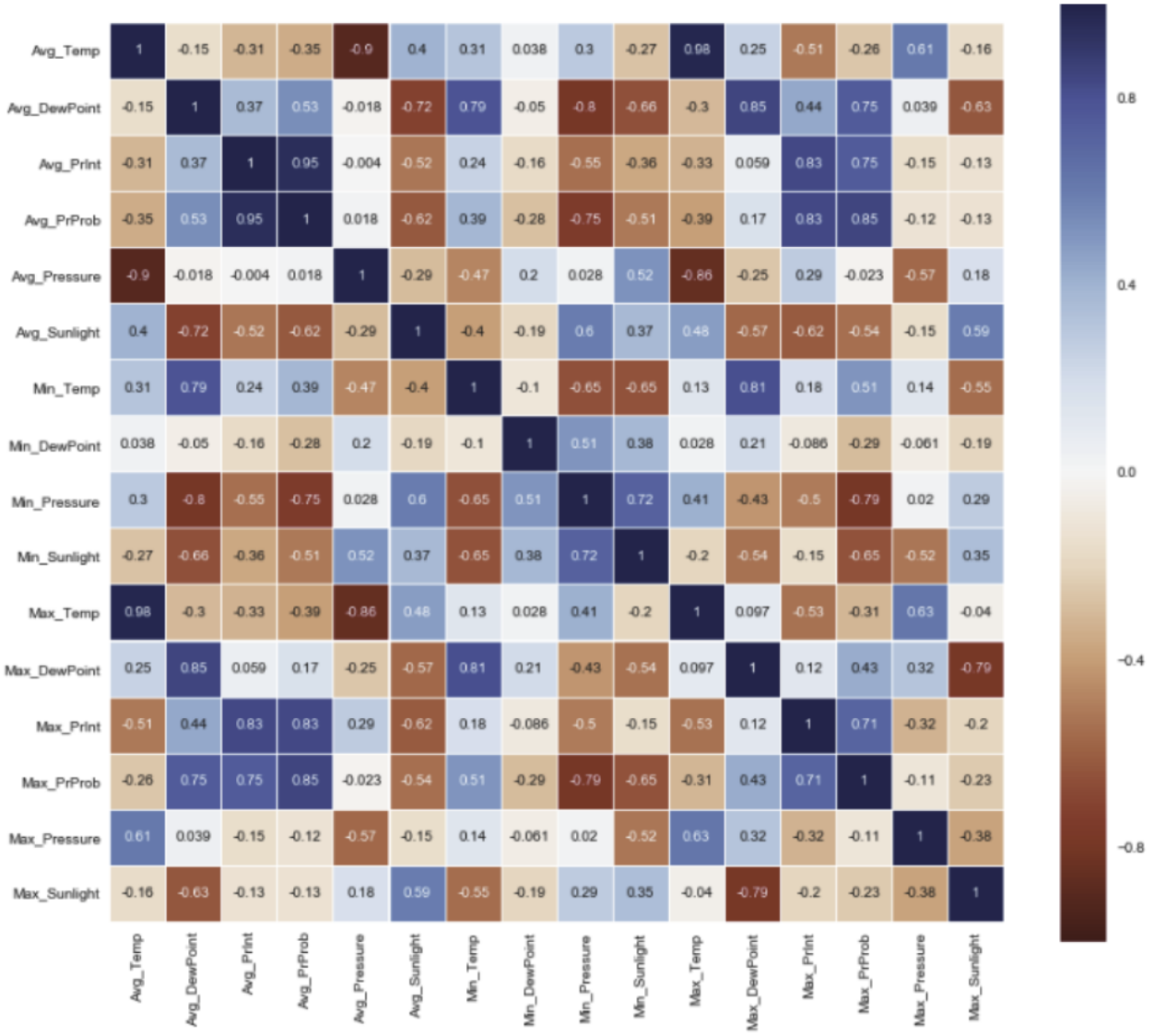

Figure 5: Correlation Plot of the Weather Variables

Maximum Temperature (Max_Temp) is one of the most important weather variables at predicting the annual tomato yield, so we assess its changes over the years 2010 - 2018 for the months of the planting season (Feb to June) for eight counties in California (Figure 6).

Solano county shows the lowest Maximum Temperature (Max_Temp) values from 2010 to 2016 (lowest points on the plot). From the Figure 6 it looks like 2017 is the hottest year, 2013 is the second hottest, and 2018 is the coldest.

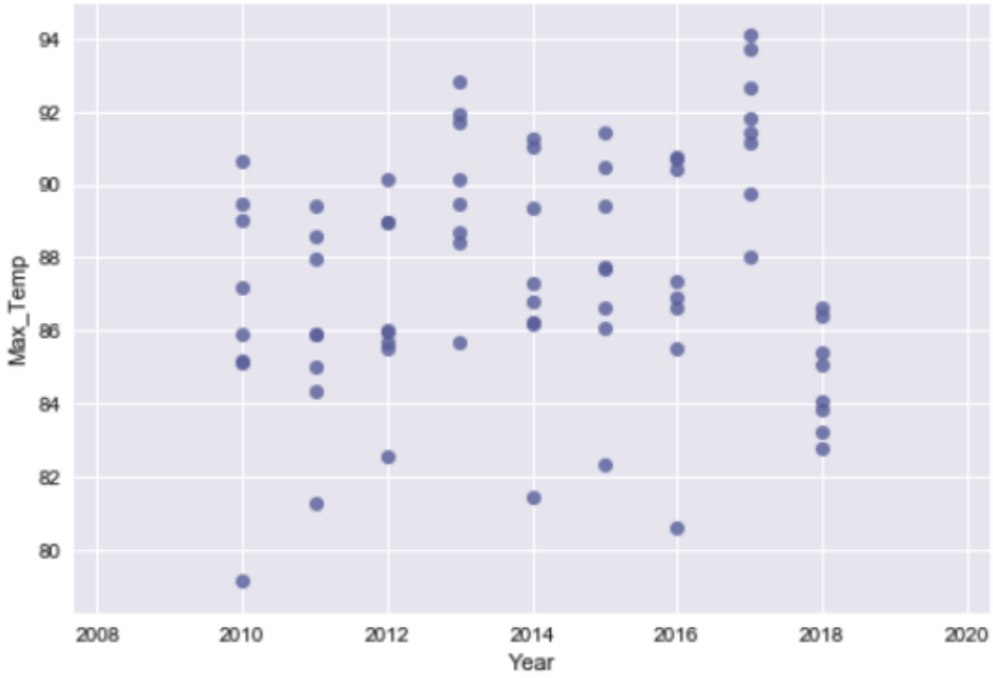

Figure 6: Changes in Maximum Temperature for 8 Counties Over the Years 2010 - 2018

Another two variables that are important, but negatively affect the tomato yield, are Average Sunlight (Avg_Sunlight) and Maximum Precipitation Probability (Max_PrProb). Figure 7 shows the changes in the values of Average Sunlight (Avg_Sunlight) over the years. The counties that show lower average amount of sunlight vary by year. The highest average amount of sunlight is in Kern county. 2013 and 2014 show the highest average amount of sunlight, while 2010 and 2011 show the lowest amount.

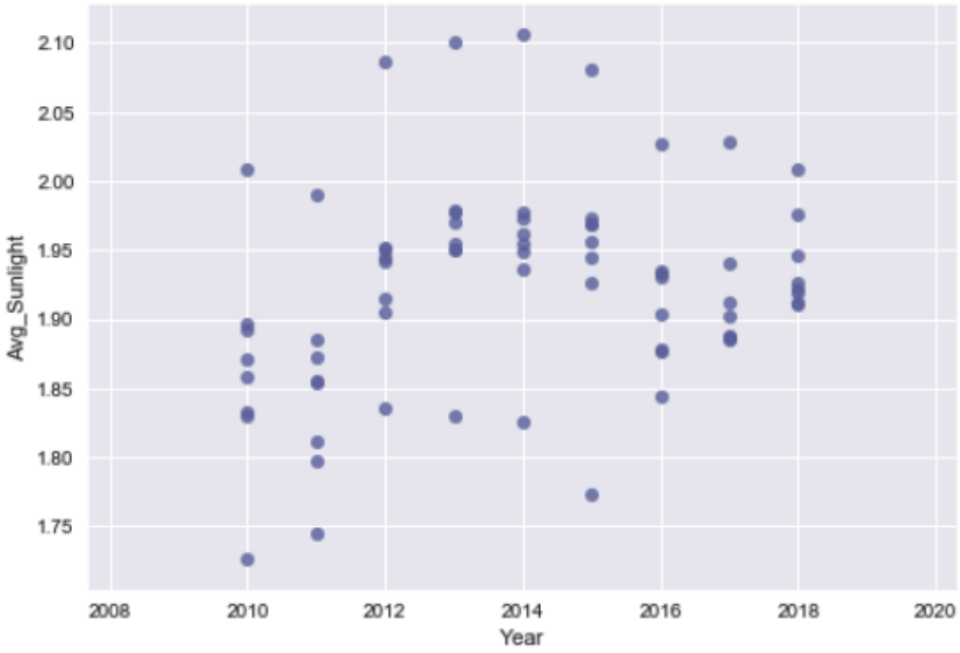

Figure 7: Changes in Average Sunlight for 8 Counties Over the Years 2010 - 2018

Maximum Precipitation Probability (Max_PrProb) values are not that easy to interpret. 2013 and 2015 show the lowest values for precipitation probability and 2011, 2016, 2017, and 2018 show the highest (Figure 8).

Figure 8: Changes in Maximum Precipitation Probability for 8 Counties Over the Years 2010 - 2018

NDVI Values

The graph below (Figure 9) shows the average NDVI values of each county in California in May 2016 (blue line) and the average of total NDVI values for California state at 0.54 (red dashed line). About half of the counties is above the average and Stanislaus shows the highest NDVI values in May 2016.

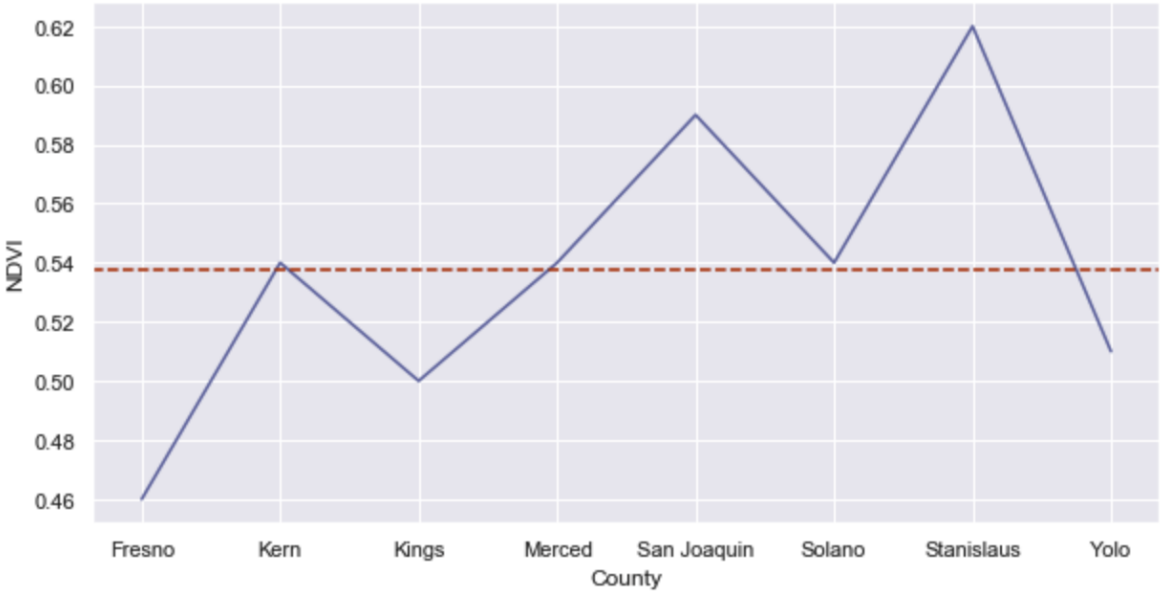

Figure 9: Average NDVI Values in CA in May 2016

Correlation between tomato yield and bag sales

We explore the correlation between the annual tomato yield and Scholle IPN bag sales at a county level in California. We use Ship-to-City column (cities in California Scholle IPN ships the bag orders to) from Scholle IPN dataset to map the cities to their corresponding counties and get the bag sales data at a county level. Using the Pearson Correlation Coefficient metric we calculate the correlation between the current year's bag sales quantity and the previous year's sales, which shows an unexpectedly high correlation of 90%. We then use the same metric to investigate the correlation between the current year's bag sales quantity and the current year's predicted tomato yield. The correlation of 60% is high enough to support the hypothesis that the current year's bag sales quantity is highly driven by the current year's tomato yield and the previous year's bag sales quantity.

Modeling Framework

Supervised machine learning models are used in this research to model (1) the effect of the weather variables, NDVI values, and the previous year's tomato yield on the current tomato yield and (2) the relationship between the predicted tomato yield and tomato bag sales. The following metrics are used to assess model performance:

sMAPE - Symmetric Mean Absolute Percentage Error,

$$\sum_{i=1}^{N} \frac{|Actual_i - Predicted_i|}{(|Actual_i| + |Predicted_i|)} * \frac{1}{N} * 100$$

RMSE - Root Mean Square Error,

$$(\sum_{i=1}^{N} \frac{(Predicted_i - Actual_i)^2}{N})^{1/2}$$

Mean Accuracy,

$$\sum_{i=1}^{N} \frac{Actual_i - Predicted_i}{N}$$

% Bias - percentage by which the model over or underestimates the actual value,

$$\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)}{\sum_{i=1}^{N} Actual_i}$$

Residual Analysis - residual is calculated by subtracting the predicted value from the actual value. Our analysis includes assessment of the normality of the distribution of the residuals, autocorrelation, heteroskedasticity, and skewness.

Model 1: To estimate yearly tomato yield

Model Description

We model tomato yield as a function of the weather variables, NDVI values, and the previous year's yield. We test a generalized linear regression as a baseline model and random forest regressor and ridge regression as challenger models. These challenger models are chosen to account for multicollinearity (Table 4, high values indicate multicollinearity) among chosen independent variables and possible nonlinearity in the relationship between the predictors and the response variable.

Table 4: VIF Assessment of the Predictors (Test for Multicollinearity)

Variance_Inflation_Factor	Features
1005.6	Max_Temp
154.3	NDVI
497.8	Yield2015

Table 5: Baseline Model Linear Regression Metrics

	sMAPE	RMSE	Mean_Accuracy	Percentage_Bias
Baseline (LinReg)	2.69	2.9	-0.28	3.23

Feature Selection

To determine the independent variables' effect on the annual tomato yield, we first select the most important features. In order to first identify the significant weather variables, we use correlation analysis, stepwise regression, lasso regression coefficients, and random forest feature importances. The most significant weather variables in predicting tomato yield of that year are Maximum Temperature (Max_Temp), Average Sunlight (Avg_Sunlight), and Maximum Precipitation Probability (Max_PrProb). We pick Maximum Temperature as the weather variable to predict tomato yield due to better interpretability.

We then use correlation analysis, stepwise regression, and linear regression coefficients to determine which month of NDVI values explains the most variability in the annual tomato yield. As a result, May (20th week) is the most indicative month of NDVI values for the entire year.

Model Architecture

All three models - baseline linear regression, challenger random forest regressor and ridge regression - are trained on 2016 data: Maximum Temperature and NDVI values. To improve the model's accuracy, we add previous year's tomato yield (Yield 2015) as an additional predictor. We model tomato yield for 2016 as a response variable on a county level using the trained models.

Model Results and Comparison

First, we validate the baseline linear regression model on 2017 data: Maximum Temperature and NDVI values. We again add the previous year's tomato yield (Yield 2016) to our set of predictors. Table 5 contains the metrics from the aforementioned baseline linear regression model. As we observe multicollinearity among our chosen independent variables (Table 4), we test the challenger models, random forest regressor and ridge regression, on the same test dataset (2017 Maximum Temperatute and NDVI, 2016 Yield). We use tomato yield for 2017 by county as the response variable. Table 6 contains the metrics from the challenger random forest regressor and ridge regression. Based on the sMAPE, RMSE, Mean Accuracy, and % Bias (Table 5 and Table 6), random forest regressor is the best

Table 6: Challenger Models' Metrics

	sMAPE	RMSE	Mean_Accuracy	Percentage_Bias
Challenger 1 (RFReg)	2.38	2.42	-0.20	2.07
Challenger 2 (RidgeReg)	3.12	3.43	-0.35	4.02

Table 7: Ensemble Using Averaging Method Metrics Compared to Individual Random Forest Regressor

	sMAPE	RMSE	Mean_Accuracy	Percentage_Bias
Challenger 1 (RFReg)	2.38	2.42	-0.20	2.07
Averaging	2.71	2.73	8.28	3.11

supervised model for Model 1. Figure 10 shows feature importances based on the best random forest regressor. It shows that NDVI scores most information gain among the three predictors.

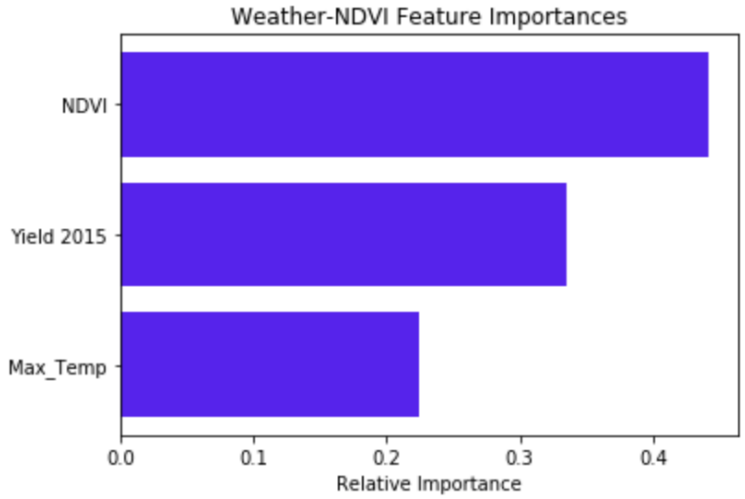

Figure 10: Feature Importances on Predictors

Model Ensemble

In order to utilize the predictive power of all of our models, we attempt to ensemble the three models, which are linear regression, random forest regressor, and ridge regression, using averaging and stacking methods. The ensemble model using averaging method results in higher RMSE and sMAPE compared to the individual random forest regressor model (Table 7). Then, we ensemble using stacking method. The random forest model is used as a meta model for the stacking method. The stacking model shows improved results compared to the averaging model and performs slighlty better than the individual random forest regressor (Table 8). While the metrics show small improvement, 0.16% reduction in sMAPE and 0.15 yield reduction in RMSE, by using the ensemble stacking model, the

Table 8: Ensemble Using Stacking Method with Random Forest Metrics Compared to Individual Random Forest Regressor

	sMAPE	RMSE	Mean_Accuracy	Percentage_Bias
Challenger 1 (RFReg)	2.38	2.42	-0.20	2.07
Stacking	2.22	2.27	8.02	-0.17

results are not very significant for the size of the dataset used in this project. Therefore, we decide to use individual random forest regressor as our final model for modeling annual tomato yield (Model 1).

Model 2: To estimate yearly tomato bags sold

Model Description

We model tomato bag demand as a function of the predicted tomato yield for the target year and the previous year's bag sales. We test a generalized linear regression as a baseline model and a tree-based model such as random forest regressor to predict tomato bag quantity sold using predicted tomato yield and previous year bag sale as an independent variable.

Model Architecture

The independent variables in this model are the predicted tomato yield from Model 1 and the previous year's tomato bag sales. The response variable is the tomato bag quantity for the target year. Baseline linear regression model is trained on the data for 2017: 2017 tomato yield actual value and tomato bag sales for 2016. The training dataset is shown in Figure 11. We use tomato bag sales for 2017 as a response variable on a county level and R-squared on the training set is 99%.

	Bag_Sales_2016	County	State	Bag_Sales_2017	Yield
0	20410.0	YOLO	CA	34260.0	43.08
1	37721.0	STANISLAUS	CA	34092.0	39.62
2	91487.0	SOLANO	CA	84260.0	41.63
3	114820.0	SAN JOAQUIN	CA	103430.0	42.35
4	840034.0	MERCED	CA	780503.0	47.95
5	287004.0	KINGS	CA	264070.0	51.88
6	255438.0	KERN	CA	240008.0	51.33
7	266199.0	FRESNO	CA	210705.0	50.39

Figure 11: Training Data for 2017

Table 9: Performance Metrics for Baseline Linear Regression and Challenger Random Forest Regressor

	sMAPE	RMSE	Mean_Accuracy	Percentage_Bias
Baseline (LinReg)	10.09	73317	0.92	-17.45
Challenger (RFReg)	13.87	128558	0.78	-18.53

Table 10: Ensemble Using Averaging Method Metrics

	sMAPE	RMSE	Mean_Accuracy	Percentage_Bias
Averaging	10.79	105368	0.85	-17.62

Model Results and Comparison

After training this model on 2017 actual tomato yield value and Scholle IPN's Bag Sales Quantity for 2016 (previous year's bag sales), we test the model on predicted tomato yield from Model 1 for 2018 and 2017 tomato bag sales as independent variables to predict the 2018 Scholle IPN tomato bag sales. The model results are shown in Table 9. As a challenger model, we use a random forest regressor. Table 9 contains the metrics from the baseline linear regression and challenger random forest regressor model with 2018 predicted tomato yield, 2017 tomato bag sales as independent variables and 2018 tomato bag sales as a response variable. As we compare model's metrics (Table 9), baseline linear regression shows better results which may indicate there is a linear relationship between the predictors and the response.

Model Ensemble

Since linear regression and random forest models are the two individual models, we ensemble them first using the averaging method. The ensemble model using averaging method results in higher sMAPE and higher RMSE (Table 10) compared to the individual random forest regressor, but is worse when compared to the individual linear regression (Table 9). We conclude that averaging method does not improve the prediction result.

Then, we ensemble the two models using the stacking method. The random forest model is used as a meta model for the stacking method. The stacking model results in larger RMSE and sMAPE (Table 11) compared with the ensemble model using averaging method (Table 10). Comparing all the models, the individual linear regression model (Table 9) performs better than the individual random forest regressor, stacking ensemble model, and the averaging ensemble model. Therefore, we choose linear regression as our final model for Model 2.

Table 11: Ensemble Using Stacking Method with Random Forest Metrics

	sMAPE	RMSE	Mean_Accuracy	Percentage_Bias
Stacking	17.24	130621	0.77	-18.32

Residual Analysis

Model 1: Random Forest Regressor

We conduct residual analysis on the random forest model. The distribution plot (Figure 12) shows that residuals are not normally distributed and most of the yield predictions are underestimated by -4 to -2 tons per acre.

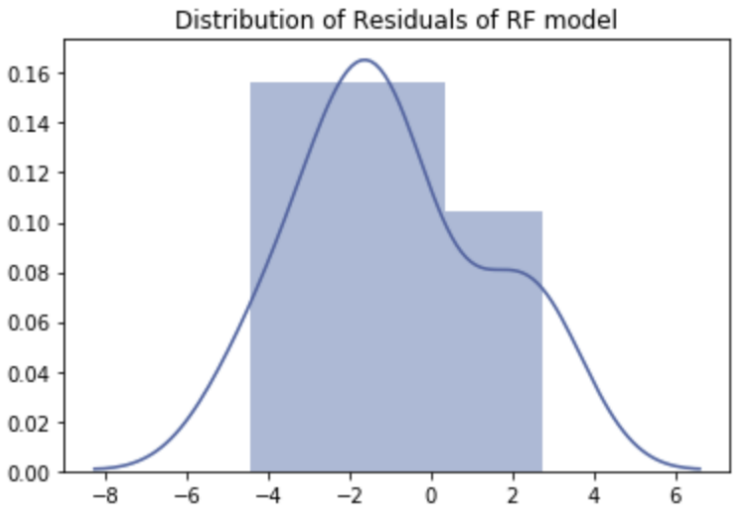

Figure 12: Probability Distribution of the Residuals

To test normality, we draw Q-Qplot (Figure 13) and the graph also supports the distribution plot above that the residuals are not normally distributed, showing a little bump in the middle.

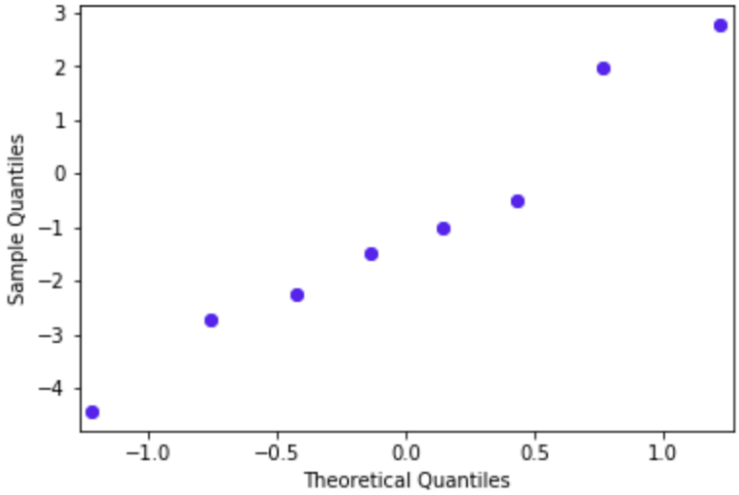

Figure 13: Q-Q Plot of the Residuals

In addition, the mean value of the residuals is -0.95, so it does not follow our assumption that a mean of residuals is normally distributed around 0. It also represents that the yield prediction might be mostly underestimated. We then conduct Breusch-Pagan test for checking homoscedasticity, and p-value (0.02) is smaller than 0.05 significance level. Therefore, we have sufficient evidence to reject the null hypothesis that residuals have equal variance. Then, we test for skewness and p-value is greater than 0.05. Thus, we do not reject the null hypothesis that the skewness is same as normal distribution. Next, we test kurtosis and p-value is greater than 0.05. Therefore, we again do not reject the null hypothesis that the kurtosis is same as normal distribution. Finally, we conduct Durbin-Watson test for checking autocorrelation, and the test statistic is 1.68. The test statistic should be between 1.5 and 2.5 to be relatively normal. Thus, we believe that there is no autocorrelation among residuals.

Model 2: Linear Regression

We conduct residual analysis on our final model, the linear regression model. The distribution plot (Figure 14) indicates that the residuals are not normally distributed, but it is skewed at the right.

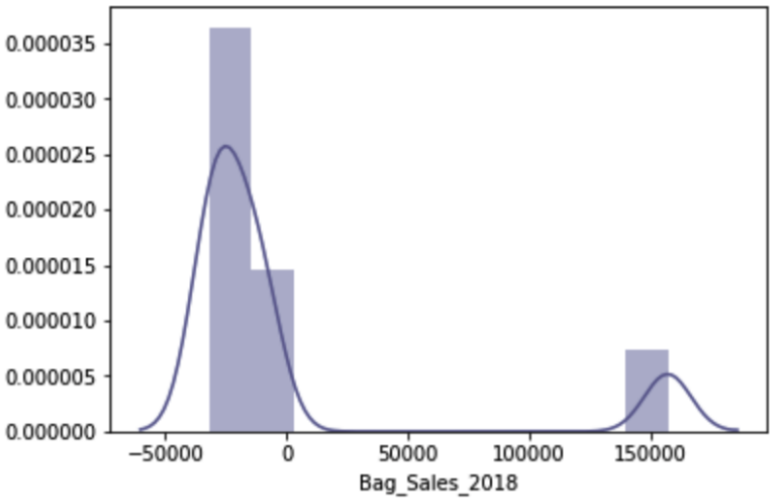

Figure 14: Probability Distribution of the Residuals

The below Q-Q plot (Figure 15) also supports the finding that the residuals are not normally distributed, indicating an outlier after 1 and before -1.

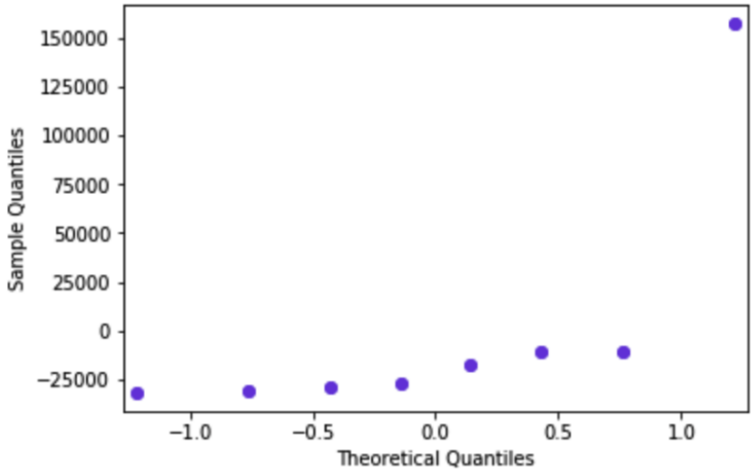

Figure 15: Q-Q Plot of the Residuals

In addition, the mean value of the residuals is 42283 (tomato bag quantity in units), so it does not follow our assumption that mean of residuals is normally distributed around 0. This indicates that the model is under-estimating by 42283 units. To normalize the distribution of the predictions and make the mean 0, we add the mean of 42283 to each prediction. This way our predictions are adjusted to be closer to the true value. The new metrics for the final linear regression model are summarized in the Table 12. Adjusted predictions now show smaller RMSE and zero bias (because of the adjusted mean). We conduct Breusch-Pagan test for checking homoscedasticity, and p-value (0.008) is lower

Table 12: Linear Regression with Zero Residual Mean

	sMAPE	RMSE	Mean_Accuracy	Percentage_Bias
Linear Regression	11.9	59896	0.95	0

than 0.05. Therefore, we can reject the null hypothesis that residuals have equal variance, which means heteroscedasticity. Then, we test kurtosis and p-values of is less than 0.05. Therefore, we reject the null hypothesis that kurtosis of the population from which the sample is drawn is that of the normal distribution. The statistic of the kurtosis test is 3.1 which indicates high kurtosis. Thus, the distribution has heavier tails and a sharper peak than the normal distribution and is leptokurtic (positive kurtosis). When we test for skewness, the p-value is less than 0.05. Hence, we reject the null hypothesis that the skewness of the population that the sample is drawn from is the same as that of a corresponding normal distribution. Thus, right skewness and right kurtosis are prevalent among residuals.

Extend Methodology

Mega Project Ensemble Model Structure

In addition to the work performed by our team, several other teams also researched additional datasets to help predict Scholle tomato bag sales. Each team (aside from the crop yield team) produced a champion model that generates a monthly forecast. The crop yield team's resolution was at an annual level so its results are reported separately in the final model. Figure 16 visualizes the process flow for the models.

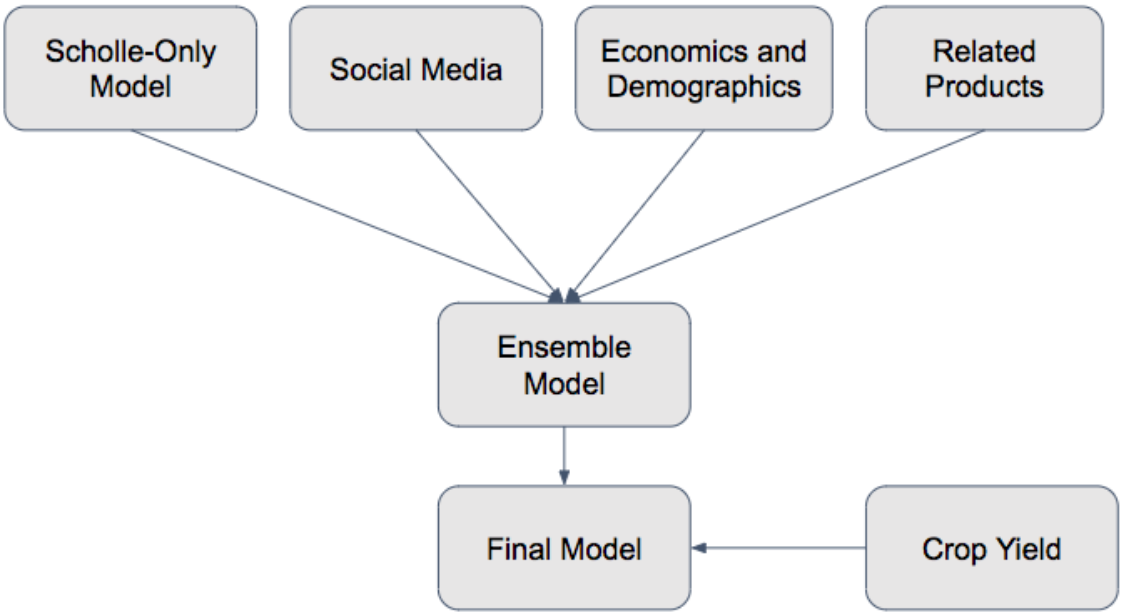

Figure 16: Ensemble Model Structure

The ensemble model is a linear regression that combines the inputs of each of the champion models from each of the teams and the Scholle-only SARIMA baseline model. The model was trained on the four years of data from 2014 to 2018 using only the harvest months. The crop yield model was trained on the data from a year prior to the previous year and validated on the previous year's data.

Integrated Workflow

In addition to finding data sources and developing models, the mega project team also created an integrated workflow to allow Scholle to maintain the models after the capstone process finishes. To do this, we wrote documentation to walk through how to collect data from external sources and developed scripts and database resources to help Scholle understand and maintain the data lifecycle of the project. Figure 17 visualizes the architecture of the system developed for Scholle.

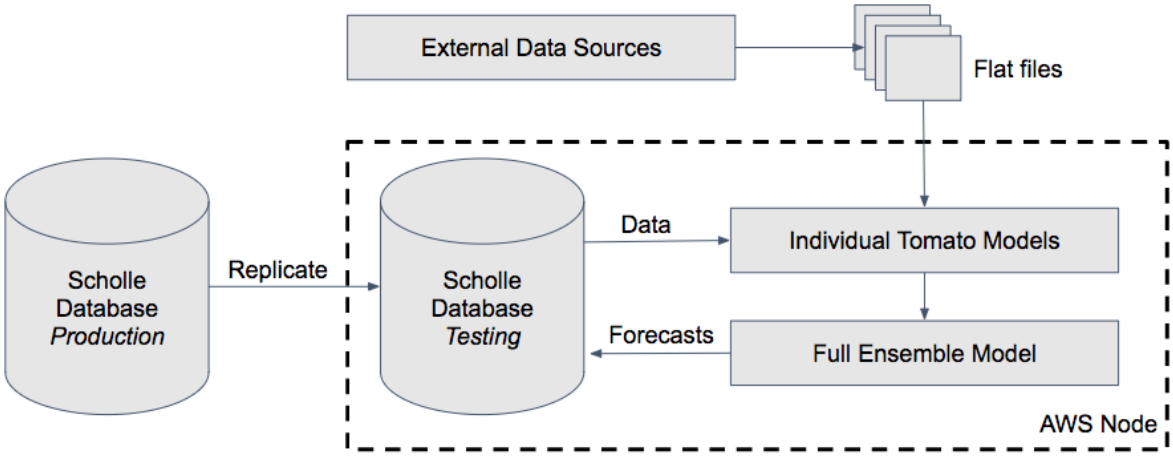

Figure 17: Integrated Workflow - Scholle IPN Infrastructure

Monitoring

"All models are wrong but some are useful" - George Box The forecasting models developed for Scholle have several assumptions: there is a regular crop of tomatoes every summer, that the external variables continue to correlate in the same way to Scholle's data, etc. If these assumptions begin to degrade or stop being true then the models will perform worse than would be reasonably expected. To understand if a forecast is beginning to diverge we do not monitor the forecast values themselves but instead the sMAPE & RMSE. We compare the historic forecasts to the now known actuals for each month and calculate the sMAPE & RMSE. We apply a rolling window of six-months to monitor for any deviations of sMAPE and RMSE above two standard deviations in the seventh month. Figure 18 visualizes the process.

Launch Month	1 Month	2 Months	3 Months	4 Months	5 Months	6 Months	7 Months	...	18 Months
Oct. 2018	2018-11	2018-12	2019-01	2019-02	2019-03	2019-04	2019-05	...	2020-04
Nov. 2018	2018-12	2019-01	2019-02	2019-03	2019-04	2019-05	2019-06	...	2020-05
Dec. 2018	2019-01	2019-02	2019-03	2019-04	2019-05	2019-06	2019-07	...	2020-06
Jan. 2019	2019-02	2019-03	2019-04	2019-05	2019-06	2019-07	2019-08	...	2020-07
Feb. 2019	2019-03	2019-04	2019-05	2019-06	2019-07	2019-08	2019-09	...	2020-08
Mar. 2019	2019-04	2019-05	2019-06	2019-07	2019-08	2019-09	2019-10	...	2020-09
May 2019	1mo. sMAPE	2mo. sMAPE	3mo. sMAPE	4mo. sMAPE	5mo. sMAPE	6mo. sMAPE			

Figure 18: Model Monitoring Using Historical Forecasts

In this example, the ensemble model run in May 2019 uses historical April 2019 forecasts from model runs 6 months prior compared to realized April 2019 bag sales to compute sMAPE & RMSE for 6 month horizon. The one month forecast made in May will be evaluated in June; the two month forecast made in May will be evaluated in July etc.

The model uses the data of the year/season it generates the prediction for, the model is to

be re-trained every year. We monitor the performance of the model by comparing sMAPE, RMSE, % Bias, and Mean Accuracy of the prediction with the previous year's prediction Figure 19.

Historical Data for Monitoring sMAPE	Annual Launch Point	Annual Prediction
2018 Prediction	June 2019	2019 Prediction
2019 Prediction	June 2020	2020 Prediction

Figure 19: Model Monitoring - Comparing sMAPE and RMSE to Previous Year's

Extend Findings

Performance of the ensemble

In addition to the performance of our team's model, we also examined the performance of the ensemble model using each team's champion model. It is important to note that the ensemble model is created only from the forecast provided by each team. This means that there is less data to do cross validation. Figure 20 summarizes the results of the ensemble model for one-month-ahead forecasts.

-

	Scholle-Only Baseline	Simple Average Ensemble	Stacked Ensemble - Linear Regression
sMAPE	13.11%	12.21%	11.94%
RMSE (bags)	66K	55K	41K
Mean Accuracy	-1.39	-5.13	0.20
Bias	50%	80%	60%

	Intercept	Scholle	Related Products	Social Media	Econ & Demographics
Coefficients	-30,066	0.074	-0.027	0.915	0.093
Std. Error	17,230	0.402	0.455	0.272	0.278

Figure 20: Ensemble Model Performance

Creating a stacked linear regression model using each champion model leads to improved model performance compared to both the baseline and a simple average of each champion model. In the stacked model, each of the champion models is given an optimized weight to calculate the forecast. These weights are given in Figure 20.

Interestingly, the linear regression did not equally weight each model, instead it placed much more emphasis on the results from the Social Media team. To help visualize the model outputs, Figure 21 shows the actual, scholle baseline, and ensemble results for the end of the time period we studied.

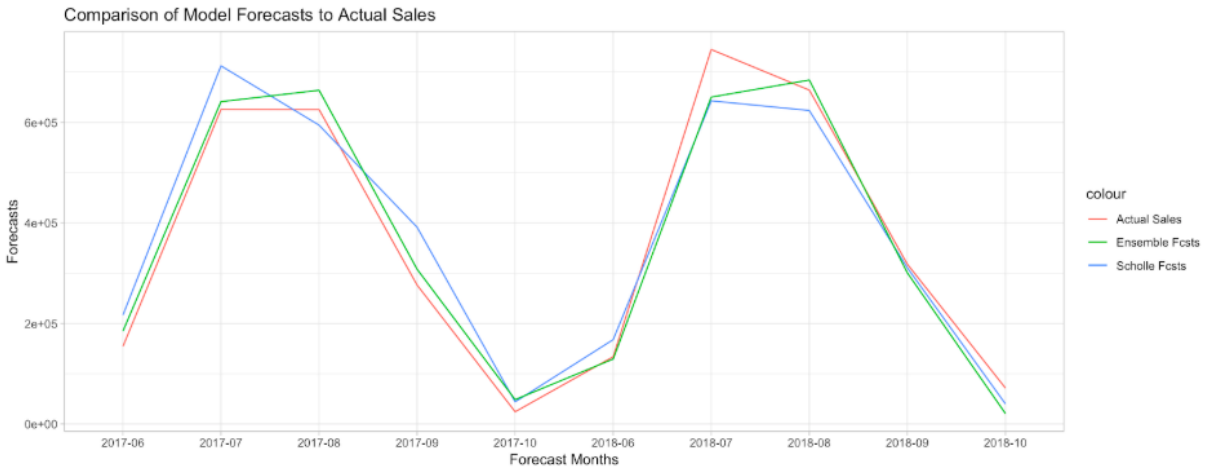

Figure 21: Comparison of Model Forecasts to Actual Sales

Thus far we have only examined the model metrics for one month out forecasts. However, Scholle intends to use the forecasts for several months out to help them guide their business decisions. In Figure 22 we present the results of examining multiple month out forecasts.

Forecast Horizon	SMAPE	RMSE	Mean Accuracy	Bias
1 month	11.94	40,577	0.20	0.60
2 months	10.88	36,953	2.20	0.50
3 months	9.77	36,892	1.80	0.50
4 months	10.06	37,634	1.84	0.50
5 months	11.48	45,832	3.57	0.50
6 months	8.78	38,120	2.54	0.50

Figure 22: Performance Across Varying Forecast Horizons

Generally, we would expect the further out a forecast is made, the worse it will perform. Surprisingly, we do not observe this increase, instead observing a relatively consistent behavior of the forecast with respect to the actuals. These results make us confident that the ensemble model approach will produce successful forecasts for Scholle into the future.

Findings

By exploring the data, we find out over 90% of the entire tomato bag sales volume is driven by shipments to California. When we distinguish important variables for predicting annual tomato yield, we realize that Maximum Temperature, NDVI values, and the previous year's tomato yield have the most impact and make the most sense given the background of this research. Among those three variables, we also find out that NDVI has the most information gain followed by the previous year's yield and Maximum Temperature based on the random forest regressor feature importances. Through correlation analysis, we observe that the Maximum Temperature and the previous year's yield have positive correlations with the current year's yield, while other weather variables such as Average Sunlight, Maximum Precipitation Probability, and NDVI values indicate negative correlations with the current year's yield. In addition, we find out that NDVI values in May (20th week) explain the most variability in the annual tomato yield.

By building several supervised machine learning models, we find out that random forest regressor performs the best for tomato yield prediction in Model 1 as we face multicollinearity issues among predictors and potential nonlinear relationships. Regarding the tomato bag sales prediction, linear regression performs the best as we observe linear relationship between the tomato bag sales and the tomato yield. As a result, we see that for every 1 unit change in the yield, the bag sales change by 1212 units.

Conclusion

The project is a success in predicting yield and bag sales of tomatoes given the limitations in data availability. The models for this research are picked based on the smallest % overfit, which is a significant issue with this small dataset (8 data points). We experiment with eight sets of the predictor variables to make sure we account for possible variability in the tomato yield while minimizing the data limitations (size, multicollinearity, potential nonlinearity). To remedy these limitations and observe the effects of the predictors on the annual tomato yield more clearly, we can incorporate other states and their counties into the model. We can also treat yearly tomato yield as time series data and see if other states exhibit similar trends and seasonality in tomato yield as California. Lastly, we can consider expanding the weather data to include water and irrigation levels and spot the general trends in rainfall in California and other states to investigate their correlation with the tomato yield.

One of the most significant findings in this research is the weather variables contributing to the tomato yield. Our analysis identified Maximum Precipitation Probability, Average Sunlight, and Maximum Temperature as the most impactful. With the limited data (8 data points for 8 counties in California), it makes sense those variables explain the most variability in the tomato yield. As weather is generally challenging to predict, some weather variables have more impact on the current year's tomato yield than they did on the previous year's yield. For example, the effect of the Maximum Temperature on the tomato yield has increased by 33% from 2016 to 2017. Given the conditions for good tomato growth (refer to: Background), we, therefore, pick the Maximum Temperature to be the weather predictor of the tomato yield.

Another significant finding is the month of the year when the level of greenness based on the NDVI values has the most impact on how much of the tomatoes are produced this year. Our analysis shows that May is the most impactful month, which makes sense as May is the concluding month of the planting season (refer to: Variables and Scope).

Using the previous year's bag sales to estimate the current year's bag sales demand is a good method. However, adding predicted annual tomato yield as another estimator improves the accuracy of the bag sales prediction by lowering sMAPE by 2.75% (from Scholle IPN previous year's bag sales sMAPE of 13.11% to 10.36%).

Recommendations

This model serves as a correction model to the forecast generated by the overall Mega Project ensemble model. The ensemble model is based on the research by the other MScA teams working with Scholle IPN to estimate tomato bag sales on the basis of economic and demographic data, social media trends, and related tomato products. Based on the limitations in data availability, the prediction from this model can only be generated at the end of the planting season (June) and right before the sales season and tomato harvest season begin that year. To predict the tomato yield for 2020, Scholle IPN is to incorporate Maximum Temperature values in California from February to June of 2020, NDVI values for May 2020, and tomato yield for 2019. Predicted tomato yield for 2020 can then be used in the Model 2 (refer to: Modeling Framework, Model 2) along with the bag sales from 2019 to predict the tomato bag sales quantity for 2020. Using this model as a correction model, Scholle IPN can check the numbers generated by the overall Mega Project ensemble model in June 2020 and adjust the inventory accordingly. The sample output from our model that Scholle IPN can use to correct for the ensemble forecast can be found in Figure 23.

Crop Yield & Weather Model

	2018	2019
Tomato Yield avg tons/acre	46.6	46.8
Tomato Bags total quantity in millions	1.9	2.3

Figure 23: Sample Model Output

Appendix A

Key Words and Terminology

- **Crop Yield** - tomatoes grown for processing (average tons/acre per county per year)
- **Cross-sectional** - analysis of data taken at a point in time (one year period in this paper)
- **Dew Point** - temperature at which the water droplets turn into dew

- **Linear Regression** - a statistical approach to model linear relationship between two or more variables
- **NDVI** (Normalized Difference Vegetation Index) - measures the level of greenness of the crop fields ranging from 0 (low levels of greenness) to 1 (high level of greenness)
- **Overfit** - when the model predicts accurately on average, but inconsistently
- **Precipitation Intensity** - amount of rainfall measured in the height of water above the ground at a point in time
- **Precipitation Probability** - the chance of rainfall at the minimum level
- **Pressure** - atmospheric pressure
- **Random Forest Regressor** - a tree-based machine learning approach that ensembles multiple regression trees to produce higher accuracy prediction
- **Scholle IPN** - client for this project, a manufacturing company
- **Sunlight** - amount of sunlight that reaches the Earth
- **Temperature** - the degree of heat intensity (measured in degrees Fahrenheit)
- **Tomato Bag Sales** - quantity of tomato bags shipped to our client???s business partners in California

References

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*, *12*(11).

E. Heuvelink, D. V. P. &. (2005). Influence of sub-optimal temperature on tomato growth and yield: a review. *The Journal of Horticultural Science and Biotechnology*, *80*(6), 652–659. http://doi.org/10.1080/14620316.2005.11511994

Geisseler, D., & Horwath, W. R. (2016). Production of Processing Tomatoes in California. *FREP*.

Holmgren, K. (2017). Satellite-Data-Corn-Futures. Retrieved from https://github.com/kimholmgren/satellite-data-corn-futures

Jr., S. A. C. (1965). Prediction of Corn and Soybean Yields Using Weather Data. *Illinois State Water Survey*.

Measuring Vegetation (NDVI & EVI). Normalized Difference Vegetation Index (NDVI). (2000, August). NASA Earth Observatory. Retrieved from https://earthobservatory.nasa.gov/features/MeasuringVegetation/measuring_vegetation_2.php

U.S. Tomato Statistics. (2018). *USDA Economics, Statistics And Market Information System*. United States Department of Agriculture. Retrieved from https://usda.library.cornell.edu/concern/publications/br86b356q?locale=en

USDA Economics, Statistics And Market Information System. (2018). United States Department of Agriculture. Retrieved from https://quickstats.nass.usda.gov/

USDA/NASS Quickstats Ad-Hoc Query Tool. (2018). United States Department of Agriculture. Retrieved from https://quickstats.nass.usda.gov/

Villiers, M. D. (2017, March). *Predicting Tomato Crop Yield from Weather Data Using Statistical Learning Techniques* (Master's thesis). University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa.