FloresDSC520Step3

Emilio Flores

2024-08-06

Introduction.

The agriculture industry in the United States contributes significantly to the Country's economy. According to the United States Department of Agriculture ("USDA"), in 2023, US farms contributed more than $200 billion to the economy. All food and agriculture-related industries contributed more than 1.5 trillion dollars or 5.6% of the GDP. This industry has attracted several national and international investors in recent decades.

Despite its attractive returns, big and small producers are exposed to high risk, as agriculture output can be affected by weather events, political turmoil, and high capital needs. Macroeconomic events are hard to predict or control. Nevertheless, scientific and technological advancements might provide tools to farmers that could mitigate the risks attributed to agriculture.

This research project attempts to predict precipitation levels in specific regions in the United States using historical production and precipitation data. Production levels are highly dependent on water availability.

The problem statement you addressed.

This research project addressed this problem statement: "Does precipitation levels affect agricultural production?". As mentioned above, water availability is one of the most critical factors determining agricultural production levels. Water scarcity is relevant as climate change has increased temperatures and drought levels in the most vital agricultural regions in the United States, like the Central Valley in California.

According to the USDA, despite only accounting for 1% of the Country's farmland, California's Central Valley produces more than 8% of the Country's food output and more than 40% of the Country's total production of fruits and tree nuts. The Department of Water Resources in California has passed several groundwater laws that threaten the capability of farmers to use groundwater to mitigate years with low precipitation. More than ever before, small and big agribusinesses need to analyze historical data to understand and predict weather events and production levels.

How you addressed this problem statement

The initial scope of research was to analyze the production levels of the most valuable commodities produced in the Central Valley. The scope of research now only uses the historical production of almond data. Exclusively using almond production data reduced the amount of precipitation data needed, as most of the Country's almond production originates from the Fresno and Bakersfield areas (both in the Central Valley). Other commodities are less geographically concentrated than almonds; hence, gathering accurate weather data would have been challenging. The model suggested for predicting almond production levels is a regression analysis. This regression analysis will calculate the dependent variable (almond yields per planted acres) by using precipitation levels in the Fresno and Bakersfield areas (independent variables).

Analysis

```
# Import almond historical data Excel file

library(readxl)

almond_df <- read_excel(
    paste0("C:/Users/emili/OneDrive - Bellevue University/Summer 2024/",
           "DSC 520 Statistics for Data Science/Final Project/",
           "AlmondData.xlsx"))

# Simplify column titles
colnames(almond_df) <- c("Year", "Bearing_Acr", "Yield/AC_(Lbs.)",
                         "Production(Mill_Lbs.)", "Price(c/lb.", "Value($1000")

# Import Fresno Area annual precipitation (top almond producer area)

fresno_precipation <- read_excel(
    paste0("C:/Users/emili/OneDrive - Bellevue University/Summer 2024/",
           "DSC 520 Statistics for Data Science/Final Project/",
           "FresnoPrecipitation.xlsx"))

head(fresno_precipation)
```

```
## # A tibble: 6 x 14
##   Year  Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
##   <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 1980  3.83  3.30  2.05  0.25  0.18  T     0.01  0.00  0.00  0.03  0.14  0.49
## 2 1981  2.67  1.29  2.59  1.01  T     0.00  0.00  0.00  0.00  0.58  1.22  0.65
## 3 1982  2.11  0.58  4.76  0.89  0.00  0.31  0.00  T     1.10  1.58  3.16  1.59
## 4 1983  5.14  3.70  4.53  2.76  0.01  0.00  0.00  0.09  1.03  0.09  2.51  1.75
## 5 1984  0.15  1.05  0.48  0.25  0.02  0.20  T     T     0.00  0.70  1.94  1.98
## 6 1985  0.43  0.71  1.73  0.12  0.00  0.33  0.04  0.02  0.43  0.85  3.02  0.72
## # i 1 more variable: Annual <chr>
```

```
# Import Bakersfield Area annual precipitation (top almond producer area)

bakersfield_precipation <- read_excel(
    paste0("C:/Users/emili/OneDrive - Bellevue University/Summer 2024/",
           "DSC 520 Statistics for Data Science/Final Project/",
           "BakersfieldPrecipitation.xlsx"))

head(bakersfield_precipation)
```

```
## # A tibble: 6 x 14
##   Year  Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
##   <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 1980  2.60  1.04  1.32  0.66  0.21  0.00  0.00  0.00  0.00  0.03  T     0.15
## 2 1981  0.93  0.78  2.15  0.56  0.18  0.00  0.00  0.00  0.00  0.83  0.41  0.23
## 3 1982  0.77  0.60  2.13  1.07  0.00  0.42  0.00  T     0.70  0.71  1.30  0.33
## 4 1983  2.21  1.49  2.62  0.57  0.01  0.00  0.00  1.18  0.18  0.14  1.31  1.15
## 5 1984  0.05  0.05  0.69  0.50  0.00  0.01  T     0.01  0.02  0.13  1.01  0.95
```

```
## 6 1985  0.38  0.48  0.48  T     0.14  0.44  T     0.00  0.24  0.18  1.65  0.27
## # i 1 more variable: Annual <chr>
```

```r
# Merged data sets

clean_df <- data.frame(
    Year = almond_df$Year,
    Production_MillLbs = almond_df$`Production(Mill_Lbs.)`,
    Acreage_1000s = almond_df$Bearing_Acr,
    Fresno_Prec_in = as.numeric(as.character(fresno_precipation$Annual)),
    Bakersfield_Prec_in = as.numeric(as.character(bakersfield_precipation$Annual))
)

# Calculate yields per acre and add to data set

clean_df$Yield_per_Acre <- (clean_df$Production_MillLbs * 1000000) /
    (clean_df$Acreage_1000s * 1000)


# Run multiple regression analysis
almond_yield_model <- lm(Yield_per_Acre ~ Fresno_Prec_in + Bakersfield_Prec_in,
                         data = clean_df)

# Display the summary of the regression model
summary(almond_yield_model)
```

```
##
## Call:
## lm(formula = Yield_per_Acre ~ Fresno_Prec_in + Bakersfield_Prec_in,
##     data = clean_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -966.12 -313.75  -38.71  274.49  881.26
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2345.937    209.676  11.188 6.87e-14 ***
## Fresno_Prec_in       -65.995     30.451  -2.167   0.0362 *
## Bakersfield_Prec_in    7.056     48.236   0.146   0.8844
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 449.3 on 40 degrees of freedom
## Multiple R-squared:  0.2294, Adjusted R-squared:  0.1908
## F-statistic: 5.953 on 2 and 40 DF,  p-value: 0.005456
```

The regression shown above states a statistically significant negative relationship exists between almond production and precipitation levels in the Fresno Area. The regression also states a positive relationship between almond production and precipitation levels in the Bakersfield area, yet it is not statistically significant. Furthermore, the R-squared value shows that the independent variables predict 22.9% of the production variances. The R-squared value is considered statistically significant.

Implications

The creation of a statistical model capable of predicting agricultural production could have a drastic impact on the industry. Farmers could plan marketing strategies, which could increase exports and facilitate the use of inventories. This model could also accurately forecast variable operating expenses, like farming and pesticide applications, as they depend highly on production levels. Furthermore, a clear production forecast could reduce the volatility in the agriculture futures and options market, decreasing the capital farmers need to use this channel as a price risk mitigant.

Limitations

The most important limitation of this analysis and model was their simplicity. Precipitation levels were considered the most critical factor in determining production. Nevertheless, other variables like average temperatures, groundwater availability, groundwater pumping levels, fertilizer usage, and farm workers' availability could also have the same or higher impact on almond yields. Similar to other industries, access to data in the agriculture industry is limited by the governmental or private organizations' efforts to gather information from producers. Hopefully, producers will realize that the more data are shared, the more remarkable forecast models will be created.

Concluding Remarks

The model presented in this project was simplistic and barely touched the surface of possibilities. Notwithstanding, it allowed the author to apply most of the principles taught in this class. The skills developed in this course will encourage the author and other agriculture enthusiasts to continue the efforts to use data science principles in the industry.