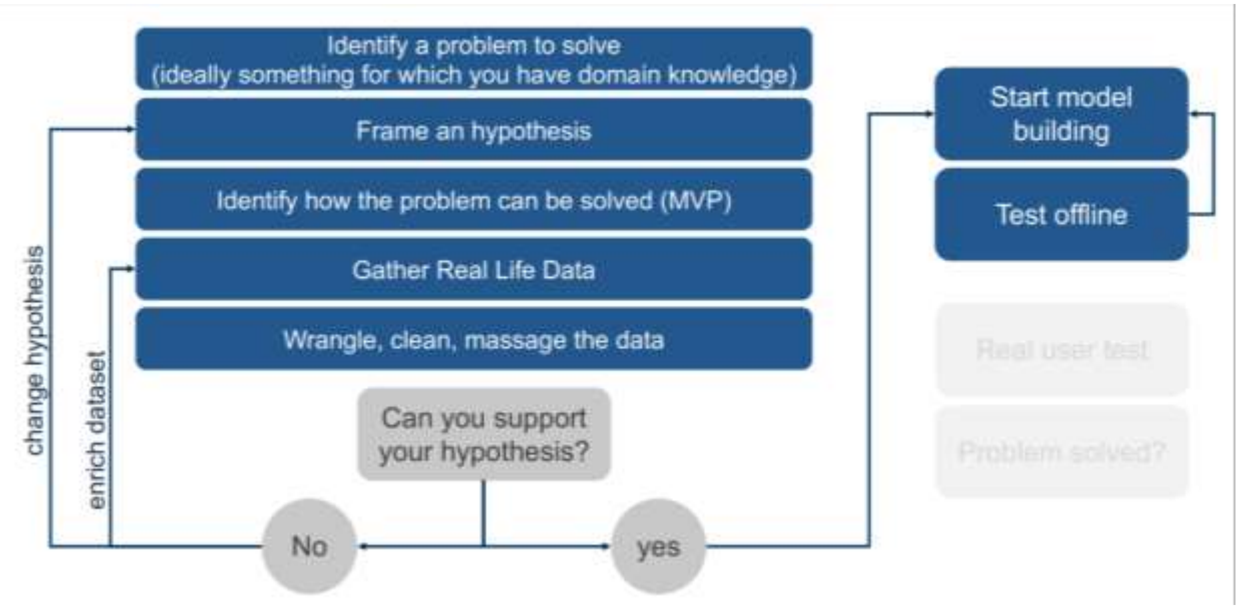
Manan Arora & Urvin Soneta | Plaksha TLF

Maximizing Farm Income

Design Thinking: Cycle of Data Driven Products for Maximizing Business Outcome

Process

Course Instructor - Dr. Sébastien Foucaud


Identify a problem to solve
(ideally something for which you have domain knowledge)
Frame an hypothesis
Identify how the problem can be solved (MVP)
Gather Real Life Data
Wrangle, clean, massage the data
Can you support your hypothesis?
No
yes
change hypothesis
enrich dataset
Start model building
Test offline
Real user test
Problem solved?

Problem to Solve

Recommend crop choice strategies for significant improvement in the farmer's income

Hypothesis

There are only a limited set of crops that a farmer can grow on his land. This is based on the climatic and topographical conditions in his area and the corresponding suitability of different crops to those conditions.

Out of this set of crops that he can grow, the amount he can earn from the crop grown depends on two factors-

1. The total cost that is incurred in procuring all the inputs and services needed for cultivation of the crop
2. The price at which he is able to sell his produce at the Mandi closest to him

If the information about what are the most suitable crops for his land along with how much can he earn from each of those crops given the time of sowing and the corresponding time of harvest, is made available to the farmer right at the beginning of the crop season, he gets the ability to choose the right crop to grow at the right time.

This will result in significant improvement of farmer income and thereby address the problem we set out to solve.

Ideal Methodology

This is based on the initial understanding of the problem without gathering or understanding data.

1. Determine a basket of crops that can be grown within a particular region
    a. Find the climatic conditions of the region – Rainfall, humidity and temperature
    b. Find the topographical conditions of the region – Soil macro and micro nutrients, pH level
    c. Find suitability requirements for different crops i.e. conditions they can grow well in
    d. For each region, find get a list of crops that can be grown suitably as per the above information
2. Once we have the basket of that can be grown in a region
    a. Find the average cost incurred to grow that crop in the region based on inputs and services
    b. Find the average market prices of those crops in the region for each season
    c. Find relationship between market price, supply and demand of each crop for each season
3. Model/ Predict the future price based on demand & supply relationship for each crop for each region
4. Model/Predict the future costs of cultivation for each crop for each region
5. Calculate the profit based on the predicted price and predicted costs and rank the crops per region
6. Make a distribution for all farmers in terms of the crop they can grow in a season, in way that it optimizes for the demand of the crop for that season as well as the income that the farmer can generate out of it

Gathering Data

Dataset	Type	Description	Units	Level	Period
Soil Macro	Topographic	Macro nutrients- Nitrogen, Phosphorus and Potassium	% Deficiency	Sub District Level	2015
Soil Micro	Topographic	Micro Nutrients- Zinc, Copper, Iron, Manganese, Boron, Sulphur	% Deficiency	Sub District Level	2015
Soil pH	Topographic	pH Value from 1 to 14 of soil in increasing order of acidity	Value	Sub District Level	2015
Ground Water	Topographic	Ground water levels: in monsoon, post monsoon & pre-monsoon	Depth in Meter	Village Level	1994-2017
Weather	Climate	Daily temperature recorded at temperature stations	Degree Celsius	Geocodes	2000-2017
Rainfall	Climate	Daily rainfall recorded at rainfall stations	Millimeter	Geocodes	2000-2017
Market Prices	-	Mandi prices (min, max, modal) per commodity and units of arrival	Rs	Mandi Level	2013-2018
Production	-	District wise production levels of crops, with area under production	Tones	District Level	2000-2015
Cost of Cultivation	-	Survey data on Cost of Cultivation across 22 states for 30,000 farmers	-	Farmer Level	2005-2016
Crop Profile	-	Suitability conditions required for optimum growth of different crops in	-	Crop Level	-

Note:

1. The datasets that have not been used in the final analysis have been greyed out in the table above. They were not used do to lack of matching or lack of other support datasets to make them relevant
2. All the datasets have been taken from public sources and their links are shared in the following section
3. Some of these datasets were in very different formats and required extensive conversion/ processing
4. Extensive **Data Crawling** was used to fetch these datasets

Dataset Details

Cost of Cultivation Data

Cost of cultivation surveys are an important mechanism for data generation on cost structure of crops. These are very intensive surveys wherein data are collected on the various inputs which are used for the cultivation of crop. The data collection approach in these surveys is inquiry based which implies that the information on the input use is obtained by inquiry from the farmer. Use of input is a continuous process which goes on from beginning to end. In order that no information on input use is missed, data collection under these surveys is generally carried out in multiple rounds. Thus, the farmer is visited repeatedly during the growth stage of the crop and during the time of harvest so that the information on input use is correctly and properly recorded. As a result, a huge volume of data is generated through these surveys. The data so collected is generally used to work out cost per unit area or cost per unit weight.

CoC survey collects yield and a whole range of farm input data at the plot level, including crop area, labor hours, machine hours, fertilizers, manure, insecticides, irrigation, seed variety, year of cultivation, and farmer identification number.

File format: csv | Source: **https://eands.dacnet.nic.in/Cost_of_Cultivation.htm**

Daily Rainfall

IMD New High Spatial Resolution (0.25X0.25 degree) long period daily gridded rainfall dataset. This data product is a very high spatial resolution daily gridded rainfall data (0.25 x 0.25 degree). Data available from 1951 to 2017. Data is arranged in 135x129 grid points. The yearly data file consists of 365/366 records corresponding to non-leap/ leap years. The unit of rainfall is in millimeter (mm).

File format: grd | Source: **http://www.imdpune.gov.in/**

Daily Temperature

IMD High resolution 1 X 1-degree gridded daily temperature data (1951-2017). This data is arranged in 31x31 grid points. For leap years, data for 366 days are included. The unit of temperature is in Celsius.

File format: grd | Source: **http://www.imdpune.gov.in/**

Daily Prices and Arrival Data (Mandi level data)

Daily transaction data from the AgMarknet portal by commodity-wise, variety-wise daily prices and arrivals information of more than 2000 varieties and about 300 commodities from the wholesale markets spread all over the country. Data is available from 2013 to 2017

File format: csv | Source: **http://agmarknet.gov.in/**

Area, Production and Yield (APY)

District wise Area, Production and Yield data from Directorate of Economics and Statistics, Department of Agriculture, Cooperation and Farmers Welfare, Ministry of Agriculture and Farmers Welfare, Government of India. Data is available from 2013-14 to 2016-17.

File format: csv | Source: https://aps.dac.gov.in/APY/Index.htm

Ground Water Monitoring Wells

Well level data from 1994 to 2017, this has information of the depth of well at different times of the year – in monsoon, post monsoon both and pre monsoon.

File format: csv | Source: http://cgwb.gov.in/GW-data-access.html

Soil Data

Soil data is collected based on Soil Health cards that are issued to farmers, based on this the data for each sub-district is aggregated into three different tables, one is the macro nutrients, one is the micro nutrients and the third in the pH Levels per sub district. This data is from 2015-2016 cycle of soil health cards.

File format: csv | Source: https://soilhealth.dac.gov.in/NewHomePage/NutriPage

Crop Profile Data

This data has the suitability requirements for the growth of each crop. Its requirements in terms of soil macro nutrients, soil pH levels, min and max temp, min and max rainfall and duration for each crop.

File format: csv

Assumptions due to lack of adequate data

We started out with a very pertinent problem in the Indian context but after doing immense data research we figured that this is very challenging due to the multiple factors involved with farmer income. We are listing down some of the assumption we have to make in our analysis due to lack of adequate data. Based on these assumptions, we will have a revised methodology in the next section

1. Temporal soil data measurements across different geographies are not available to us, hence we are not taking into account the variation in soil conditions of a region over time
2. We could not find any demand and supply data at a suitable level that we could use, hence we will not be able to factor that in the analysis of prices of crops
3. We are not accounting for transportation cost of taking produce to market, we are assuming that all produce is taken to nearby market at negligible costs
4. We are assuming that all the produce that farmers cultivate is going to the market without accounting for the produce which is kept for household consumption
5. We are considering that all the produce is being sold directly by the farmer and does not involve any middle man or any public or private sector intermediary
6. We are also not considering the export of crops in any analysis
7. We are unable to account for certain soil properties like soil water holding capacity. Also, we only have input quantity per season, not the details of different type & quantity of inputs within the crop cycle.

Though we understand that some of these factors play a major role in determining the actual income at the farm level, we are unable to consider them at the moment. Future expansion of the project will involve inclusion of these factors for a more comprehensive and true analysis.

Revised Hypothesis

Post gathering the data and understanding it well, we have revised our hypothesis keeping in mind constraints of the data we have as well as the assumptions we are forced to take due to this.

The simplified problem which we aim to tackle now is:

Finding out how the yield of a crop depends on different agricultural inputs like labour, irrigation, machinery, fertilizers, manure and insecticides as well as climatic or topographical conditions like temperature, rain and soil

Hypothesis: The idea is that once we are able to figure out which these factors matter the most, we can combine them it with the required knowledge of which crops can be gown suitably in which area and bring in the price information of the local markets to determine the most optimal combination of crops that the farmer can grow through the year to maximize income.

Wrangling and Preparing Data for Analysis

As we wanted to run our model to be able to predict yield based on several factors, we had to combine a variety of datasets. We used the **cost of cultivation (CoC) data as the base data** which had all agricultural inputs as well the total production of each piece of land. Other datasets were merged to it at sub-district level. The cost of cultivation data had information from around **2200 different sub-districts across 19 states in India**. The process of combining these is described as follows:

Matching with Soil Data

1. In the CoC data we had 2200 unique sub-districts across India, while we had Soil Data – pH, Micro Nutrients as well as Macro nutrients for around 5000 sub-districts across India
2. The challenge was that the names of Sub-Districts didn't match due to spelling errors as well as repetitions.
3. Using **Google Geocoding API,** we found the geocodes of each of the sub districts both in soil data as well as the CoC data.
4. The data of the nearest sub district from the soil data was assigned to each sub-district in the CoC data using **Haversine distance**

   Note: We have not considered temporal changes in soil data, the data from 2015-16 cycle of soil health cards that was taken from the government website has been assigned for all years.

Matching Temperature and Rainfall Data

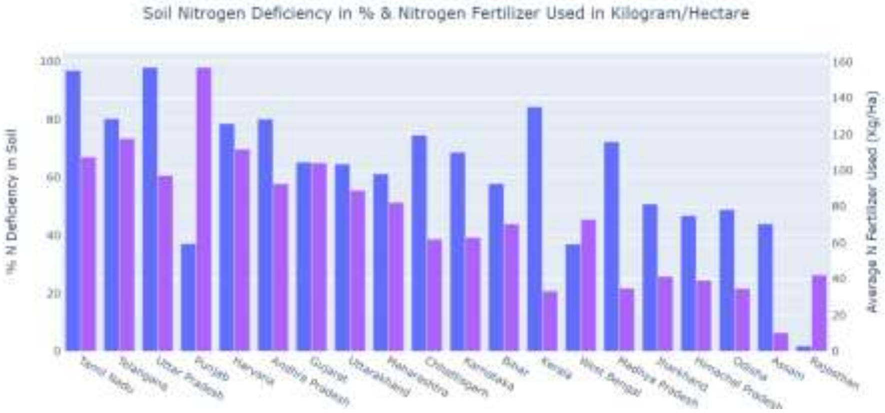
1. The temperature and rainfall data was available at different grid levels across India for everyday in GRD format for the years 2000 to 2017 from the Indian meteorological dept. website
2. We first parsed data to get an excel for each temperature and rainfall
   a. Temp station coordinates as key along with avg. temperature for each month each year in Celsius
   b. Rain station coordinates as key, along with sum of rain for each month for each year in Milimeter
3. Each sub district from the CoC data was assigned a temperature as well as rainfall station closest to it based on **Haversine distance** between the station coordinates and the coordinated of the sub-district
4. Once the station was assigned, we joined two variables to the CoC data
   a. Average temperature for all months based on the crop season and the year of cultivation of crop (Eg. For Kharif Crop grown in 2005 we took average of Temperature for months from July-Oct'05)
   b. Total Rainfall for all months based on the crop season and the year of cultivation of the crop (Eg. For Kharif Crop grown in 2005 we took sum of rainfall for months from July-Oct'05)

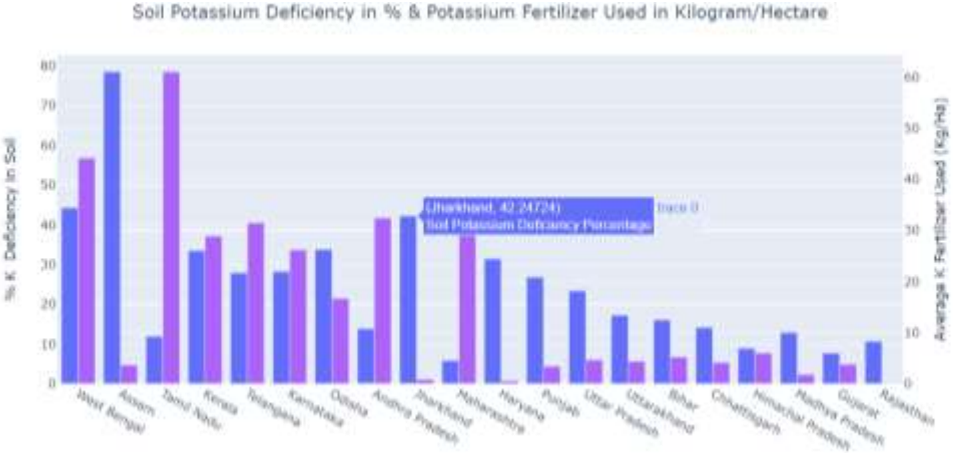
Adding Calculated Columns to CoC Data

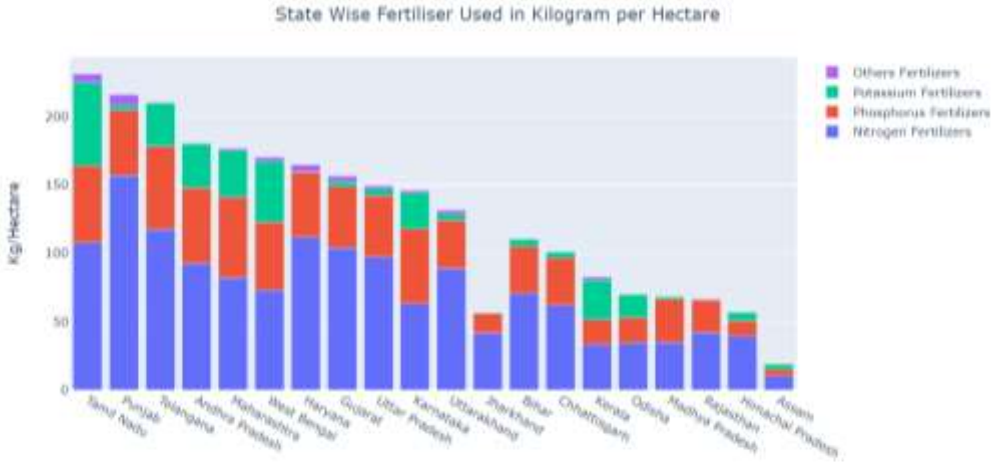
Columns Added – **Derived Yield [Mainproduct/Crop Area]** & **Chemical Cost [Sum of fertilizer, manure and insecticide costs]**

Exploratory Analysis on Data

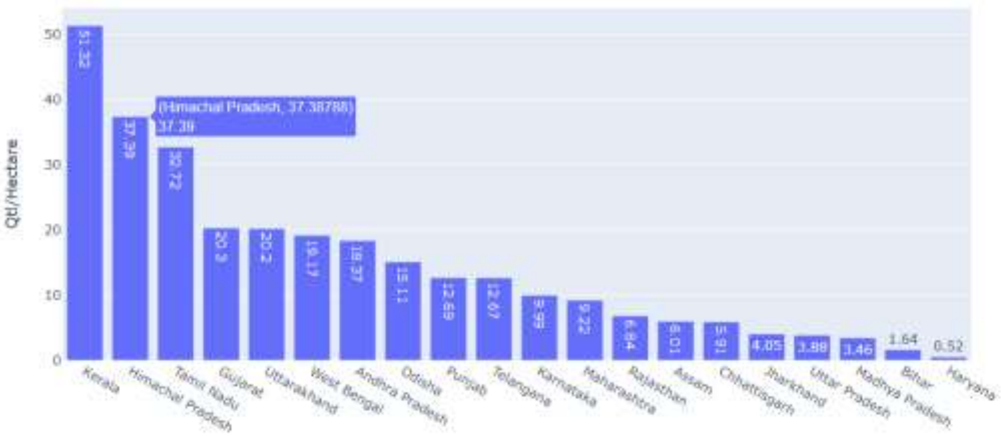
These are some of the variables that we plotted after combining the datasets as per the wrangling defined above.



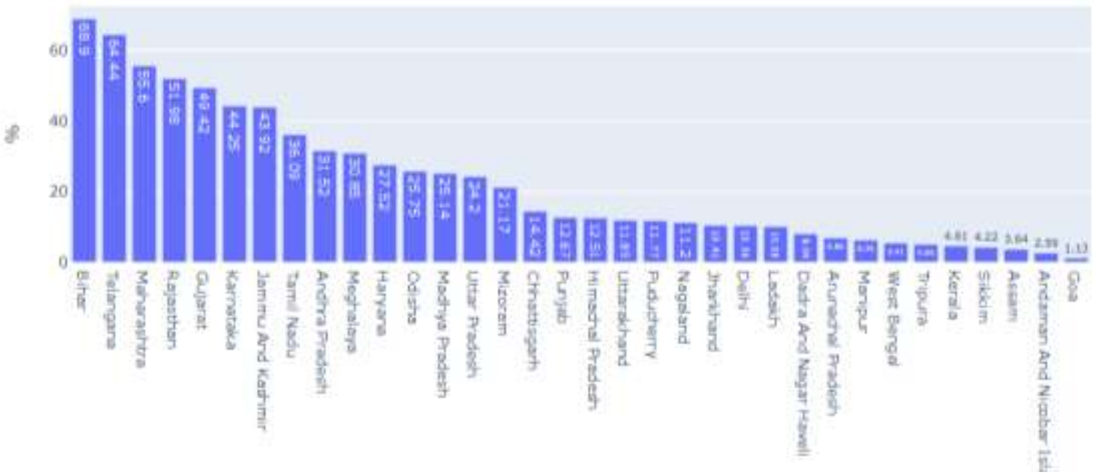




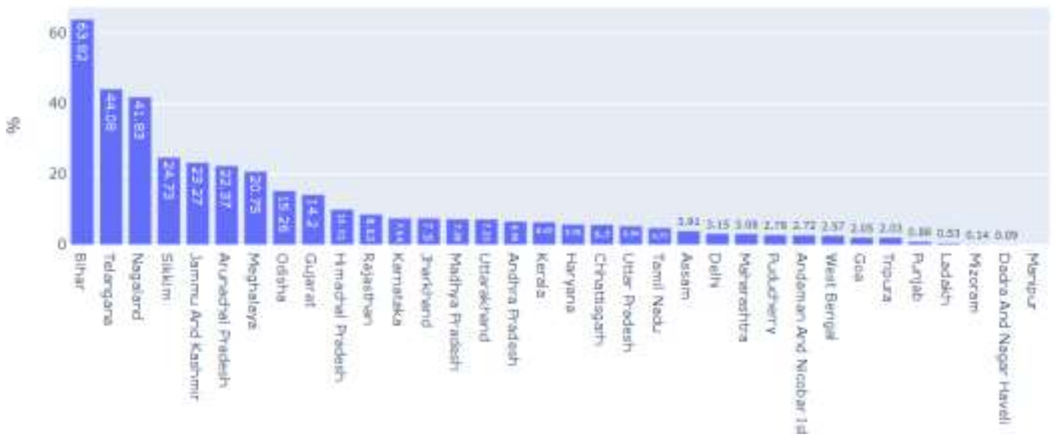
State Wise Manure Used in Qunital per Hectare



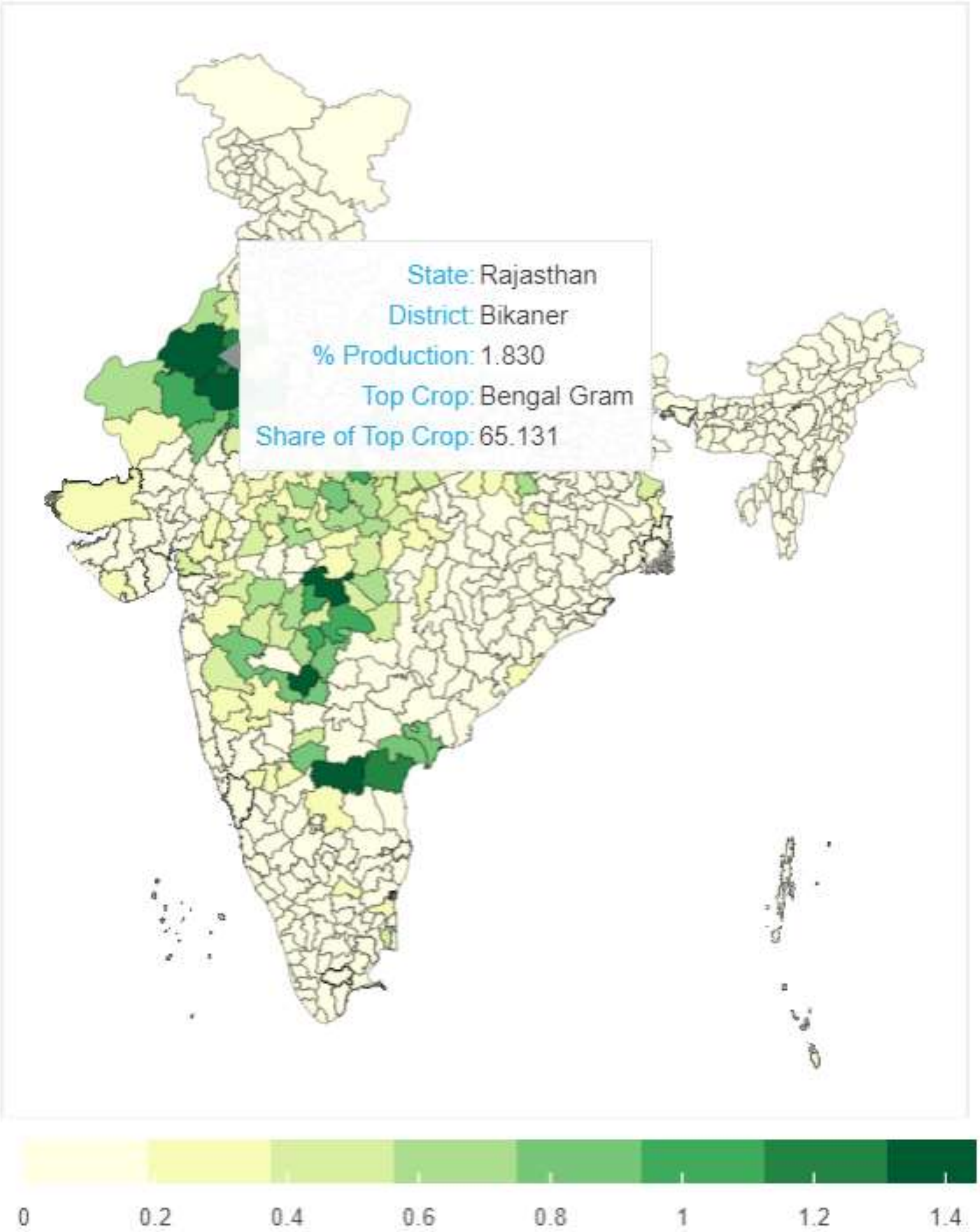
State Wise % Iron Deficiency in Soil



State Wise % Copper Deficiency in Soil



% Pulses Production District Wise



State: Rajasthan

District: Bikaner

% Production: 1.830

Top Crop: Bengal Gram

Share of Top Crop: 65.131

Modelling

Modelling Yield based on All Inputs and Climatic/ Topographic Conditions

We ran this for two crops –

- **Paddy** - Total Observations – 138096
- **Maize** - Total Observations - 15286

Note: The numbers below are the **RMSE**

Model	Paddy	Maize
Linear Regression	1.48	1.37
Elastic Net	1.46	1.38
Xgboost	0.93	0.98
Support Vector Regression	1.50	0.90
Random Forrest Regressor	0.83	0.74
Decision Tree Regressor	1.19	0.99

Explain ability of Yield for Maize:

- R2 **–** 0.53
- Important Variables – Casual Labour Hours, Hired Animal Labour Hours, Seeds per Hectare, Fertilizer per Hectare and Own Irrigation Channel Hours, Temp, Rain

Explain ability of Yield for Paddy:

- R2 **–** 0.38
- Important Variables –Hired Machinery Hours, Seeds per Hectare, Fertilizer per Hectare, Temp, Rain

Although the RMSE is low, the explainability of the variation for both the models i.e. R2 is fairly low. Therefore we can say that the yield is not characterized well enough by the data points we have

Modelling Chemical Cost Based on Soil Characteristics

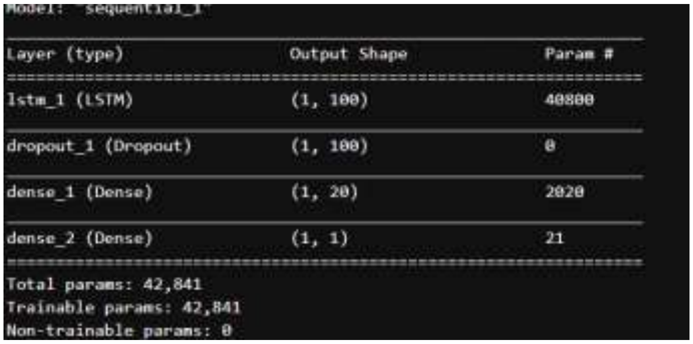
We also tried to model the chemical cost i.e. sum of costs of all the fertilizer, manure and insecticides used as a function of the soil characteristics – Nitrogen, Phosphorous, Potassium, Zinc, Copper, Iron, Organic Carbon, pH and other deficiencies.

Here the results were very unsatisfactory and the variables were not able to explain the change in costs well.

Additional Analysis: Modelling Prices based on Historical Data

Based on the historical pricing data that we had for each crop at the market level, we tried to use an LSTM model to determine if could use a model to predict prices. We processed the data for a time series model - lookback of 15 days and batch size of 1.

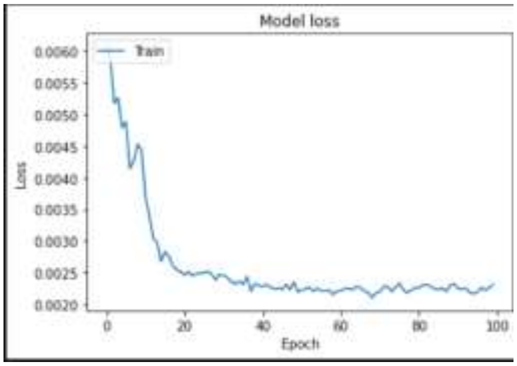
Architecture of the Model

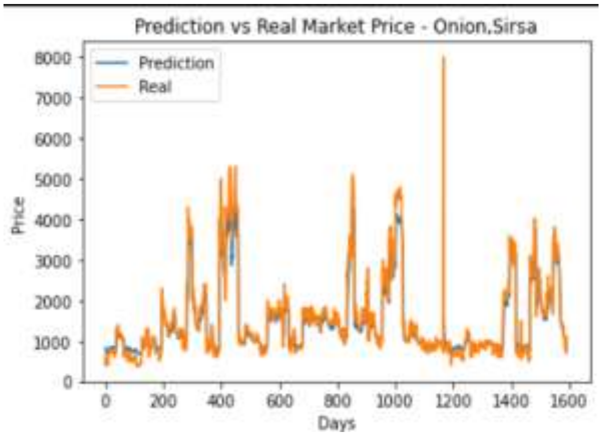


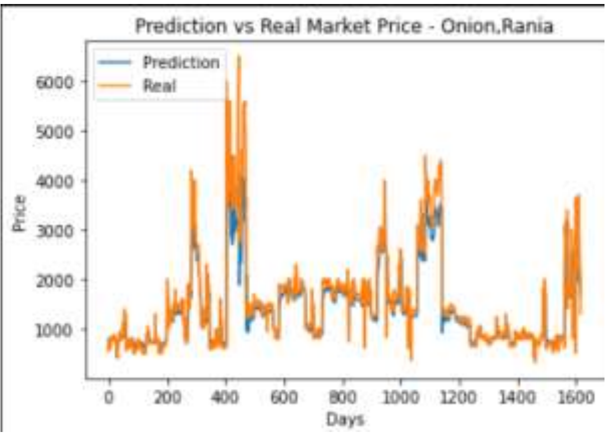
```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_1 (LSTM)                (1, 100)                  40800
_________________________________________________________________
dropout_1 (Dropout)          (1, 100)                  0
_________________________________________________________________
dense_1 (Dense)              (1, 20)                   2020
_________________________________________________________________
dense_2 (Dense)              (1, 1)                    21
=================================================================
Total params: 42,841
Trainable params: 42,841
Non-trainable params: 0
```

We trained the model for **Onion crop prices from the Market Sirsa in Haryana and tried to predict the prices for the nearby market Rania in Haryana.** Here are the results after 100 epochs of training:

- Error (MSE) in Training Data (Sirsa, Haryana): 0.002
- Error (MSE) in Test Data (Rania, Haryana): 0.004
-







Post-Analysis of the Hypothesis and Future Course of Action

- The data we had was not sufficient to predict the yield accurately as the features didn't explain the variance of the dependent variable very well. There is need to add more data sources for the exercise.
- We saw promising results of the LSTM model for price prediction. This can be replicated across different crops and tested further.
- We have spent immense time in data preparation and hopefully will be able to use this data along with other datasets for a meaningful analysis in the near future