ROBUST YIELD PREDICTION OF VARIOUS FARM PROCESSING UNITS

Name:Sagireddi Purna Narasimha Achyuth

Id : 4632

Overview

- Problem Statement
- Data Understanding
- Predictive Analysis
- Approach-1
- Approach-2
- Approach-3
- Conclusion

PROBLEM STATEMENT

Forecasting means predicting the future based on historical data, trends, seasonality and features or information. It is used in many domains like weather, sales, business, economic, finance...etc and more data leads to better forecasting and also data recorded for forecasting analysis must have constant frequency like hourly, daily, montly.... etc

PROBLEM STATEMENT

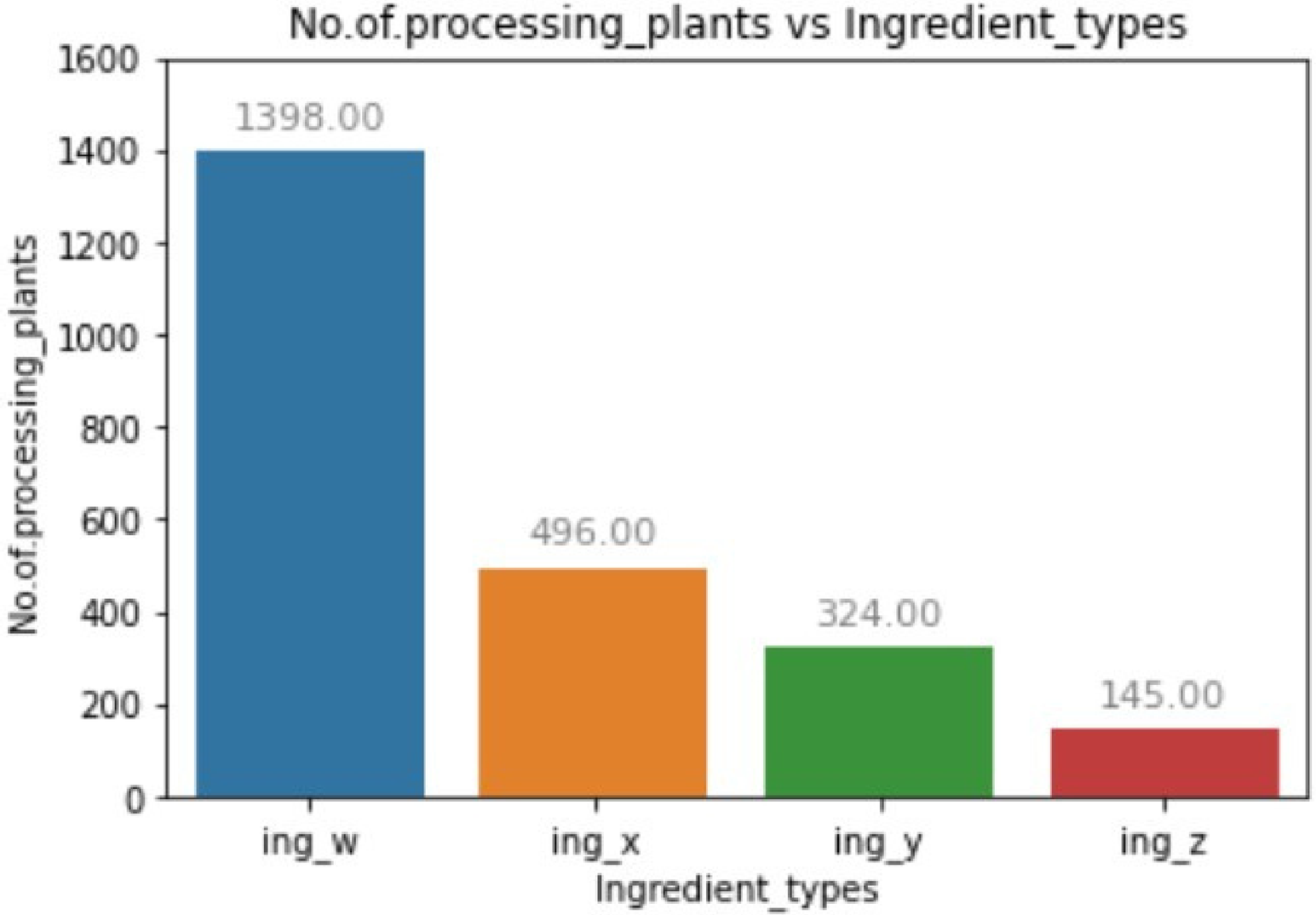
Given, the details of various farms including weather information and farm details for the year 2016 and yield production is recorded hourly from each farm, They are expecting us to forecast the yield production for 2017 hourly. Therefore this is forecasting problem, have to apply time series models

DATA UNDERSTANDING

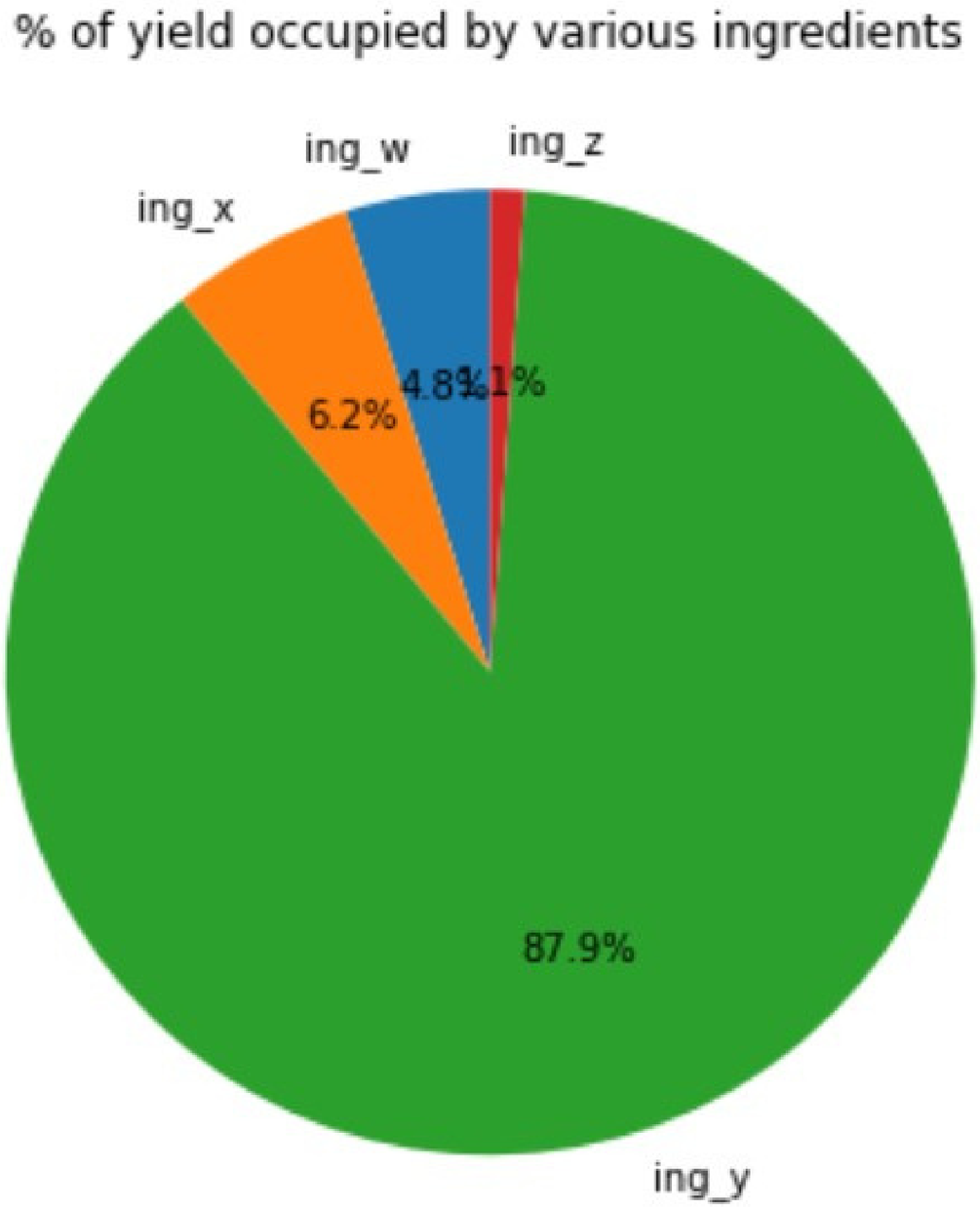
Given data consists of 1434 processing farm plants. Each processing plant is producing ingredients like ing_w, ing_x, ing_y, ing_z in tonnes. And also contains weather information for 16 locations where the farms are located as well as farm details like farming company,farming area...etc,Given are the details of number of processing plants that produces the ingredients.

i.e

- ing_w is produced from 1398 farms
- ing_x is produced from 496 farms
- ing_y is produced from 324 farms
- ing_z is produced from 145 farms



i.e **More number of farms produces ing_w**



For the year 2016

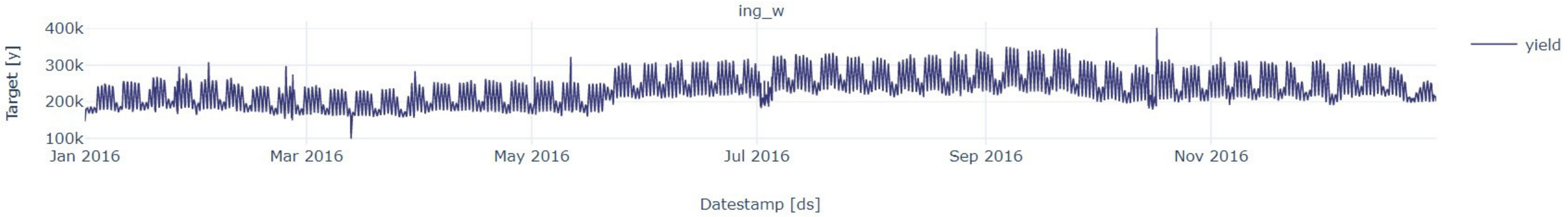
- ing_y is produced 87.9% of total yield
- ing_x is produced 6.2% of total yield
- ing_w is produced 4.8% of total yield
- ing_z is produced 1% of total yield

DATA UNDERSTANDING Cont..

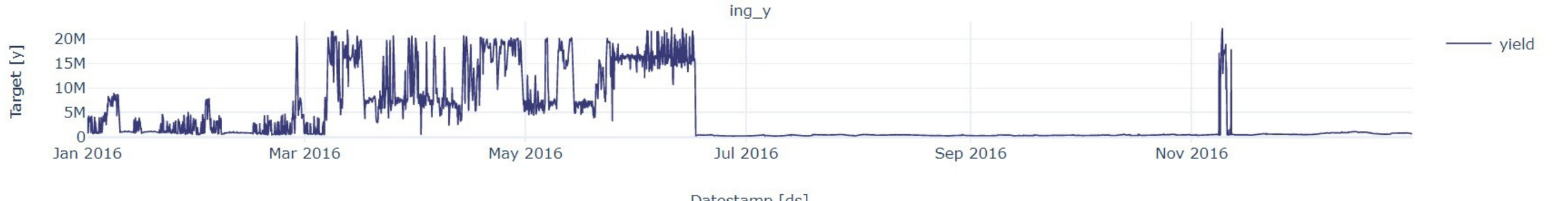
Given data contains one year(i.e 2016) yield production for all the farms, The patterns of all ingredients for the year 2016

By observing the below plots, we can say that ing_w is having some seasonality and ing_y production is very low from july 2016 to Nov 2016

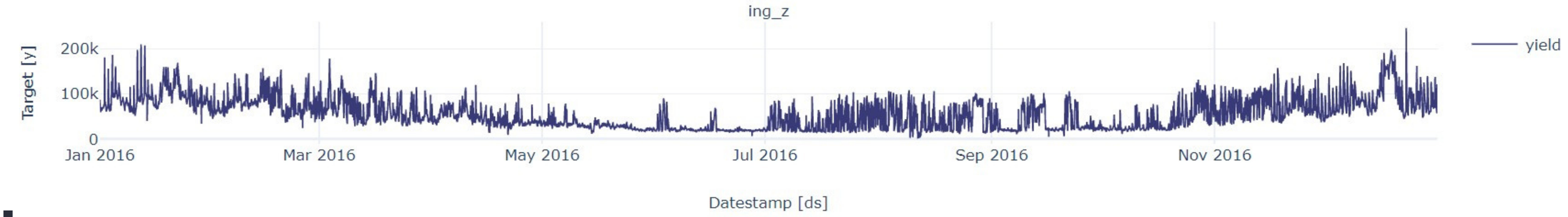
ing_w:



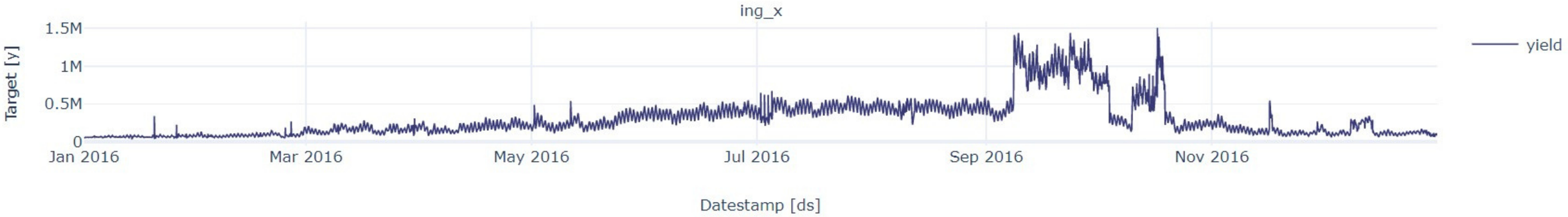
ing_y:



ing_z:



ing_x:



We can see the patterns for ingredients like ing_z and ing_x above

Below are the details of farms that produces high yield during the year 2016

Farms that produces high yield production:

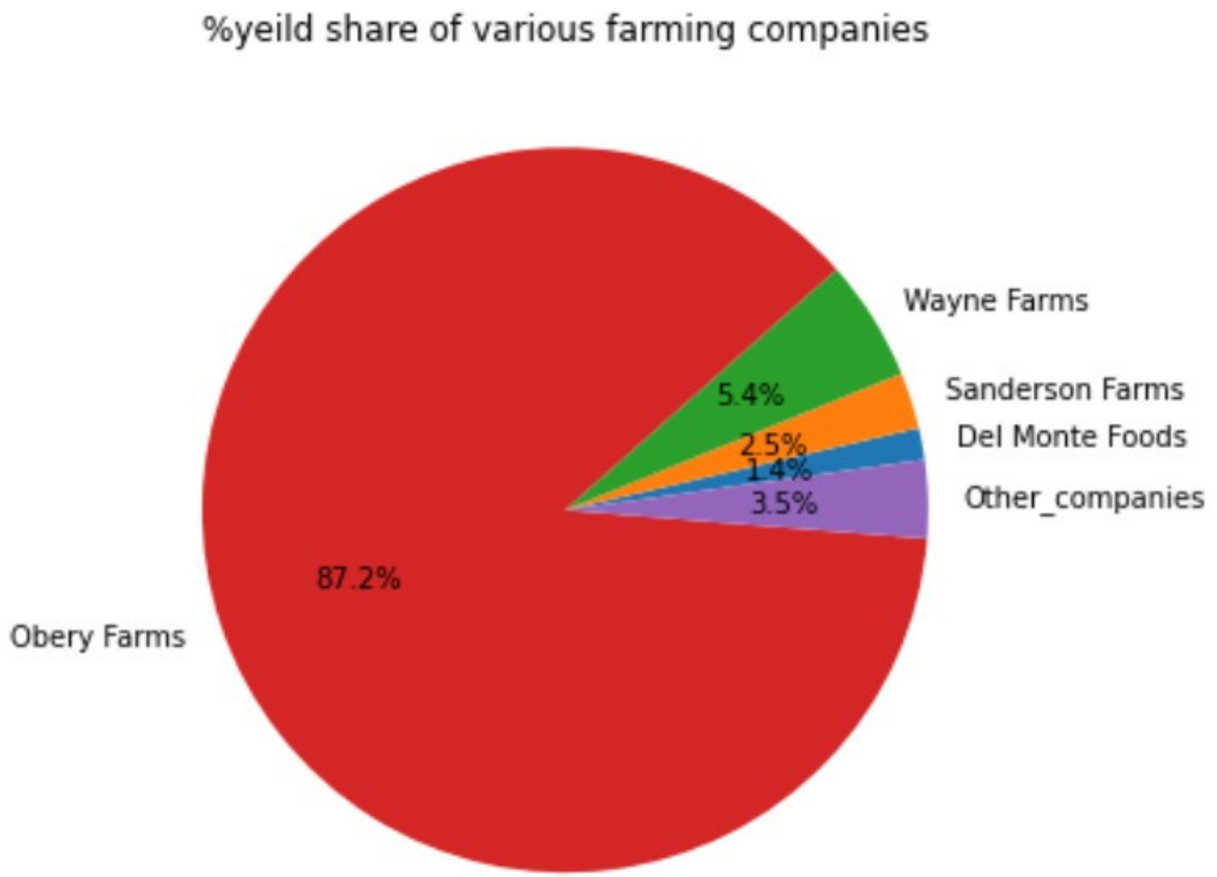
Farm_id	**Ingredient_types** Produced	**Total yield produces(tonnes)**	**Farming_Company**	Location	**Farming_Area**
fid_72059	ing_w,ing_y	3.35062*10^10	Obery Farms	location 2532	30925.922252
fid_121183	ing_w,ing_x	4.404114*10^8	Sanderson Farms	location 565	10065.018117

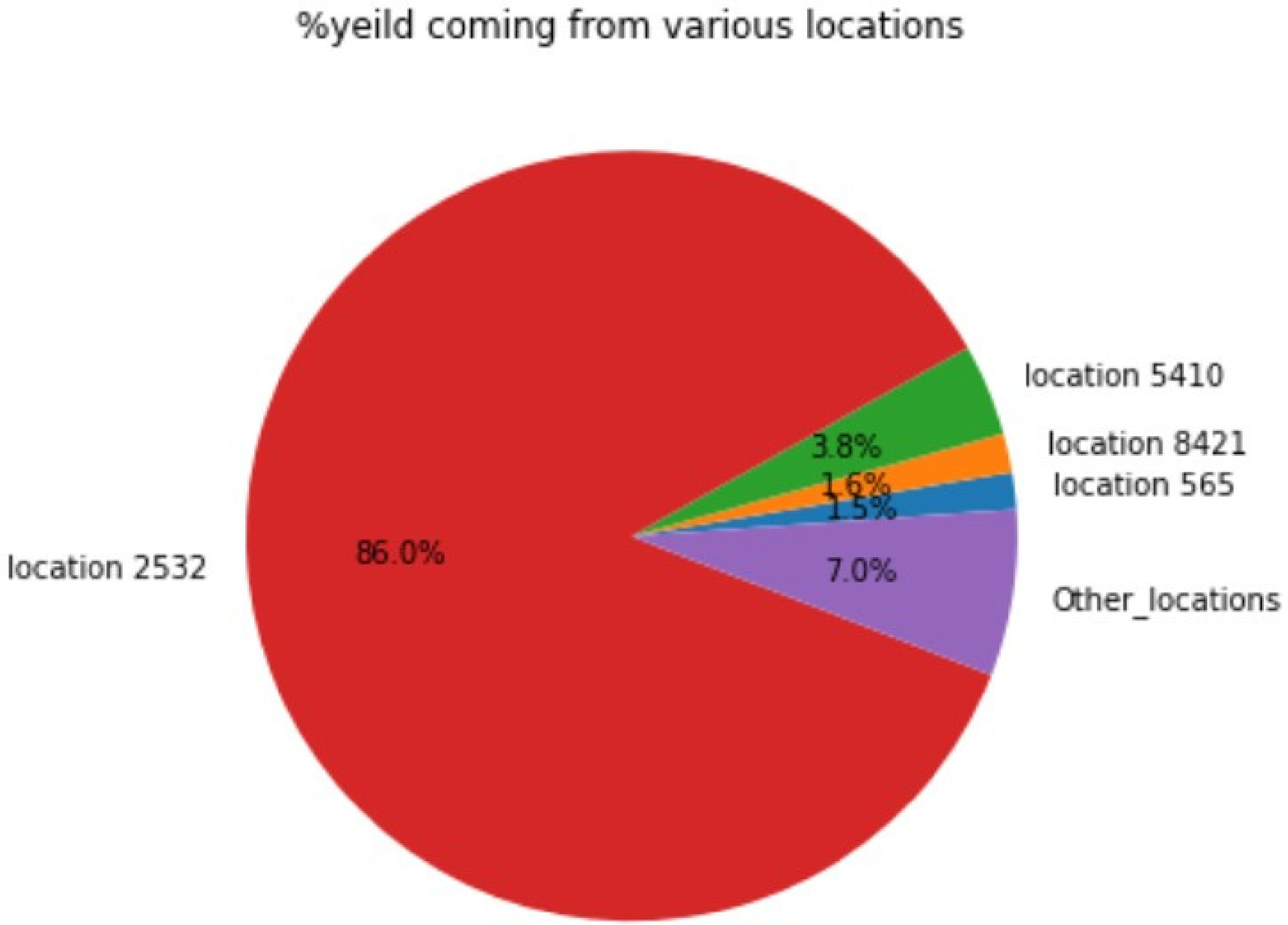
Below are the details of farms that produces low yield during the year 2016

Farms that produces low yield production:

Farm_id	**Ingredient_types** Produced	**Total yield produces(tonnes)**	**Farming_Company**	Location	**Farming_Area**
fid_111424	ing_w	3.5134	Obery Farms	location 6364	3527.991425
fid_73978	ing_w	1445.0352	Sanderson Farms	location 5677	91.416552

There are 16 Farming companies, out of 16 farming companies Obery Farms produced 87.2% of yield in the year 2016.

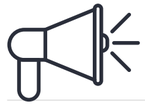




There are 16 deidentified_locations, out of 16 deidentified_location
86% of yield is comming from location 2532

PREDICTIVE ANALYSIS

To forecast the yield production of 1434 farms for the year 2017,I have used several approaches



NOTE

Given data only contains 1 year information and also future weather data is provided, if that weather data is also forecasted, multivariate time series models gives huge errors. And Time series models require several years data to capture seasonality

APROACH-1

AIM:

To build univariate time series forecasting models like ARMA, ARIMA, SARIMA for single farm and for single ingredient without using loops, therefore I have to bulid 2,363 models.

IMPLEMENTATION:

So building 2,363 models is time taking process, Hence I have built only models for high and low yielded farms.

APROACH-1

PROCEDURE

Data reading

Importing required libraries and read the data

Retrieving and Pivoted Data

Retrieved farm data and using pivot method ingredient types are converted as targets.

Data preprossing

Reindexed the date if any dates are missed and interpolation is used to fill na values.

Stationarity Analysis

Adfuller test and plots are used to check the stationarity if not trans-formations are used

Data Analysis

Seen the trend, seasonality and plotted time series plots like acf ,pacf to fix parameters like p,q.

Model Building

ARMA, ARIMA and SARIMA models are used.

Hyper parameter tuning

Checked RMSE and patterns by using different values of p,d,q.

Forecasted for 2017

Picked best params and forecast for 2017

APROACH-1

OBSERVATIONS:

- Analyzing individual farm is time consuming process
- For univariate Analysis farm data is not at all required because farming company, farming area and location is constant w.r.to time.
- So I thought below approach is better.

APROACH-2

AIM:

To identify stationary and non stationary farms and creating ing_w stationary, ing_w non-stationary(similarly all ingredients) farm id sets and fixing model and model params by taking samples from those sets therefore applyed the best for whole ids

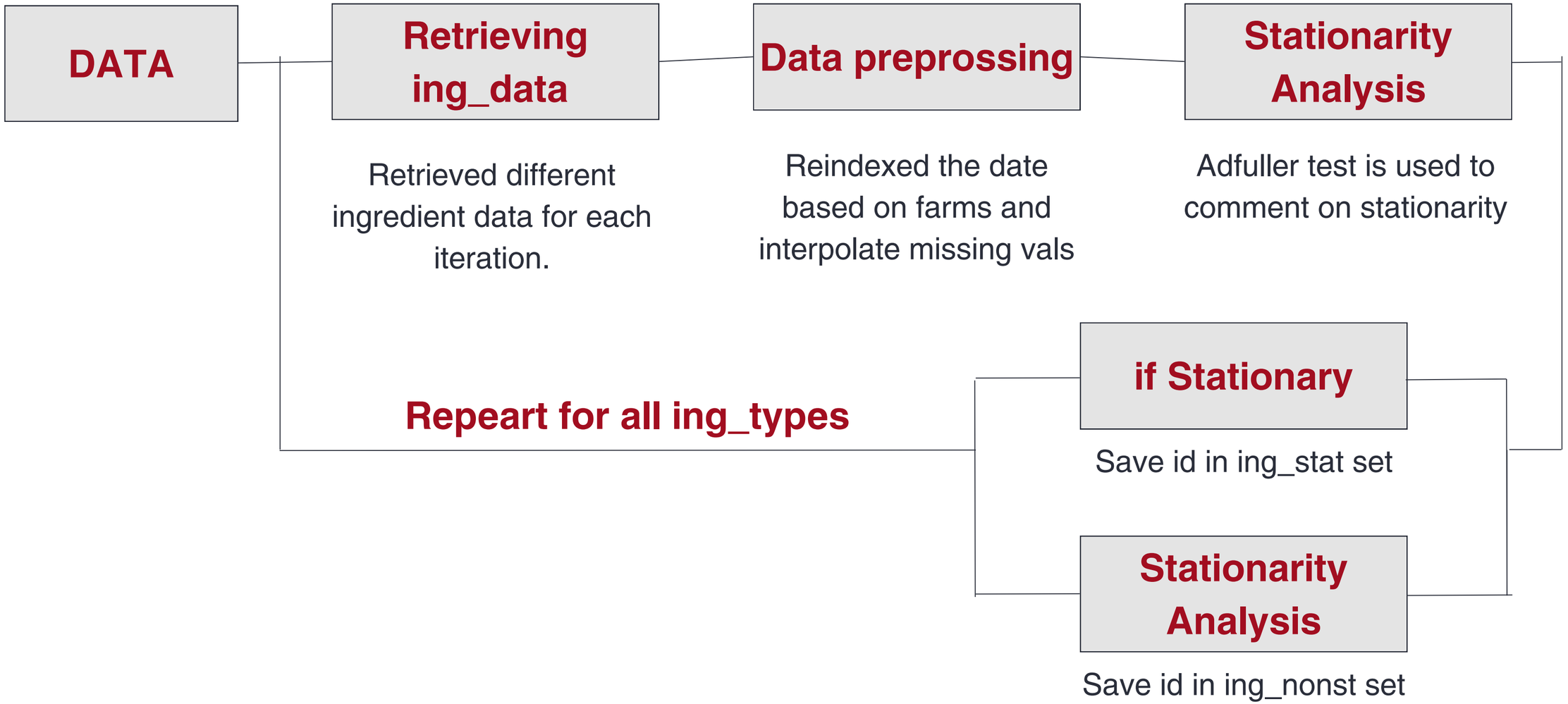
IMPLEMENTATION:

Have to fix parameters for 8 times and for different models . Take best model with best params and apply for all ids in that set.

APROACH-2

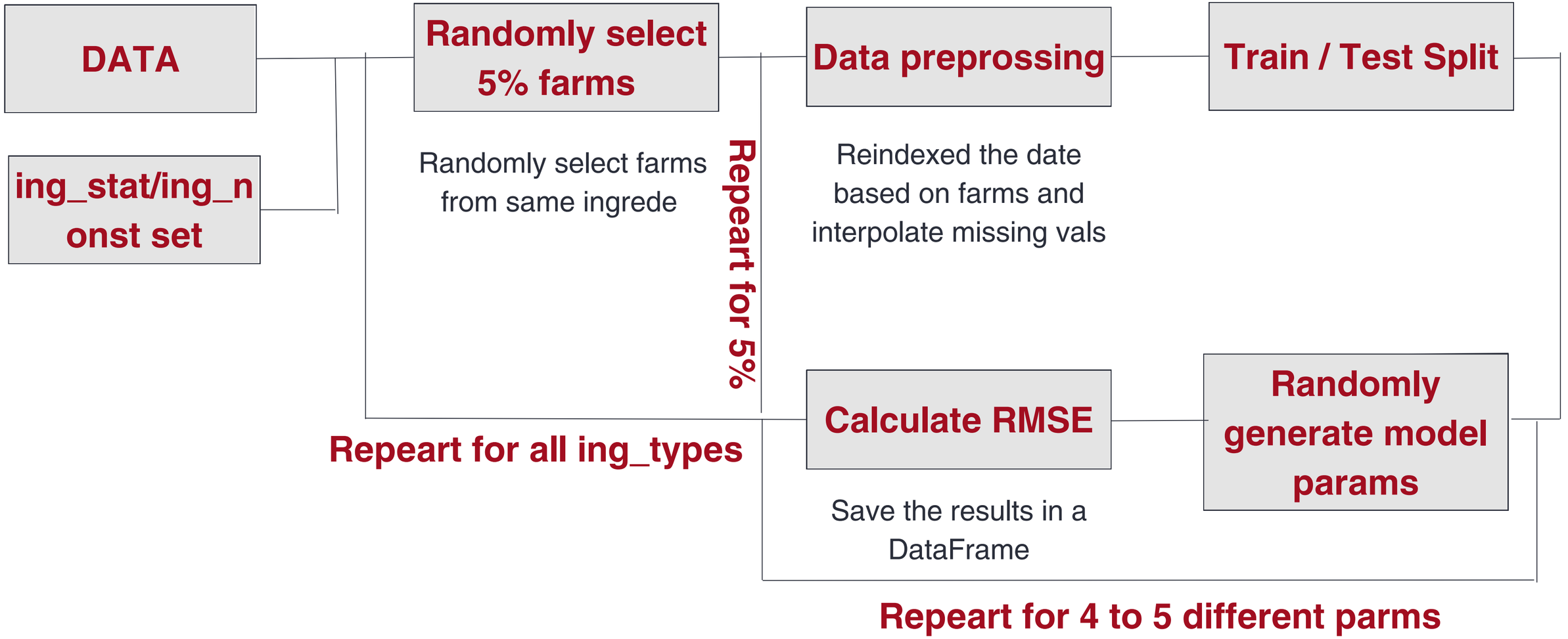
PROCEDURE:

Step1: Farms Seperation



APROACH-2

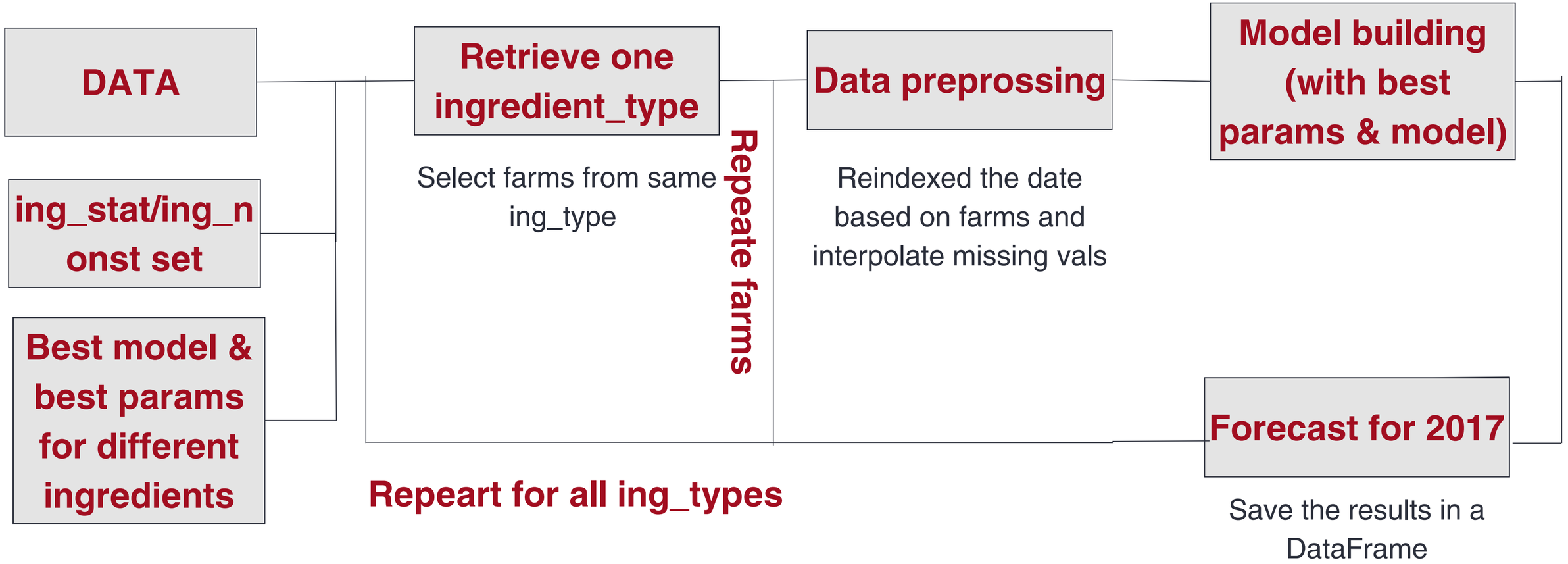
Step2: For Fixing the model and model params



I have done above flow with different models

APROACH-2

Step2: Future Forecasting



APROACH-2

OBSERVATIONS:

- Out of ARMA, ARIMA and SARIMAX approach 2 gives best results for SARIMAX
- Approcah-2 takes 30 sec to run a single model ,therefore for all farms it has taken 20 hours approx and also taken time for fix the params and model

APROACH-3

AIM:

To do multivariant analysis using statsforecast models like AutoETS, Croston Classic.

IMPLEMENTATION:

Implemented ml models by using statsforecast only for ing_z because it require more computation.

APROACH-3

PROCEDURE

Data reading

Importing required libraries and read the data

Retrieving Data

Retrieved required ingredient type data.

Data preprossing

Reindexed the date if any dates are missed and interpolation is used to fill na values.

Statsforecast modeling

AutoEts and Croston Classic models are used and picked best model for 2017 forecast

Future Forecast

Future forecasted for test data using best model

CONCLUSION

- RMSE of AutoETS ---- 3593
- RMSE of CrostonClassic ------- 3052
- Out of all approaches approach-3 i.e statsforecast multivariant analysis gives low RMSE