# SATELLITE INTELLIGENCE
# SUGARCANE HARVEST FORECASTING MODEL

22 October 2019

# CONTENTS


| | |
|---|---|
| Introduction | Page 3 |
| The Problem | Page 3 |
| The Proposed Solution | Page 4 |
| Scope | Page 5 |
| Methodology | Page 6 |
| Results and Conclusion | Page 9 |
| What's in the Pipeline | Page 11 |
| Appendix A – Project Management | Page 11 |
| -Team Members | Page 11 |
| -Communication Tools | Page 12 |
| -Issues Encountered | Page 12 |
| References | Page 13 |

# INTRODUCTION

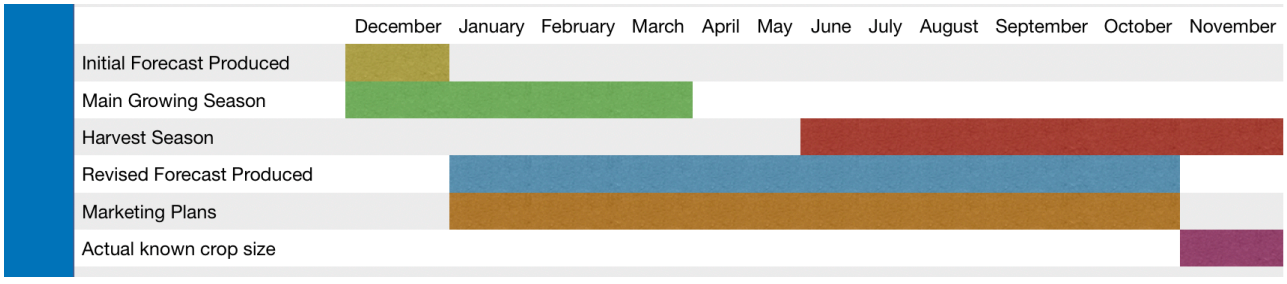
Sugarcane is a tall perennial grass predominantly grown in Queensland, accounting for 95% of the total Australian sugarcane production. More than 80% of sugar produced in Australia is exported as bulk raw sugar and provides a total annual revenue of around AU$2.5 billion. Australia's 24 raw sugar mills are huge, self-sufficient industry situated close to the farms which supply them with sugarcane. The sugar mill industry generates compelling economic value across all stages of production. The major activities revolve around growing, harvesting, milling and sales.

As a sugar mill, a better forecasts on when and how much of sugarcane is likely to arrive will help In making superior business decisions and resource management.
As a farmer, a better prediction based on region conditions will give a personalised schedule on various lifecycle[link] of sugarcane production.

# THE PROBLEM STATEMENT

The Sugarcane industry has faced unstable trading conditions over the past five years, largely due to variable weather patterns, lower commodity prices and variation in global sugar production. The main concern of sugar mill industry is fluctuation in sugarcane production and inability to estimate this fluctuation consistently.

A clear overview of the problem faced in industry can be viewed by taking a scenario[link] into consideration in which accurate numbers are key to making key decisions and resource allocation.Pretty much every production manager is keen on boosting his production. He attempts to limit the expense or maximise production , the key goal is that maximum yield ought to be acquired from any input resource combination. This level of optimum efficiency can be achieved by creating a predetermined yield rate. In this sense estimation of production has a crucial job for planning maximum production however the fundamental issue is which strategy of estimation should be used for the forecasting of production. The Theory of Production (Theory & Analysis, 2019) explains the principles by which a business firm decides how much of each commodity that it sells it will produce. And how much of each kind of resource like labour, raw material, fixed capital goods etc it employs, it will use. This theory has a place with the both small scale and large scale level of financial aspects. Thus an optimum forecasting model that estimates the amount of commodity produced and fine tune resource allocation.




Figure[1] Highlighting key points in the cycle of growing, harvesting and marketing sugar in Queensland

The figure above shows the sugarcane lifecycle. The harvesting season commences around the month of June in Australia. From a sugar mills perspective, it's critical to plan for the coming crushing season even before commencement of harvest. Therefore, industry decision makers such as mill managers and people from marketing team work every year to estimate the size of

the crop and when it's likely to arrive. In some years, initial crop estimate made came quite closely to the total yield size observed after the actual harvest. For example, In year 2005 the estimates were within 0.5% of the actual harvest crop. However, it's not the case every year as in year 2010, and 2012 the deviation between estimated and observed crop ranged from an overestimate of 25% to an underestimate of 22%. These gaps in the estimation and observed crop can bring considerable operational and expectation problems for a Sugar Mill. Looking into this from a marketing perspective we found that the problem of overestimation (around 25%) can potentially lead to major shortfalls in meeting the expectations and commitments with the forward export stakeholders. Also, the sugar mill industry tends to start the season earlier than required to meet the overestimation, which results in lower profitability as the sugarcane content is usually lower early in the harvest season. On the other hand, Underestimates can cause problems for the industry at a different level like difficulties in managing the finite storage space may result. Also, underestimation could delay the start of the crushing season and as a result, farmers extend the harvesting of large portion of crop in the rainy season and thereby, increasing the chances of wet weather disruption. Wet weather disruption adds risk of destroying the ratooning crop(crop left after harvesting) since sugarcane is a perennial crop. (SOUTHERN OSCILLATION INDEX PHASES TO FORECAST 2019)

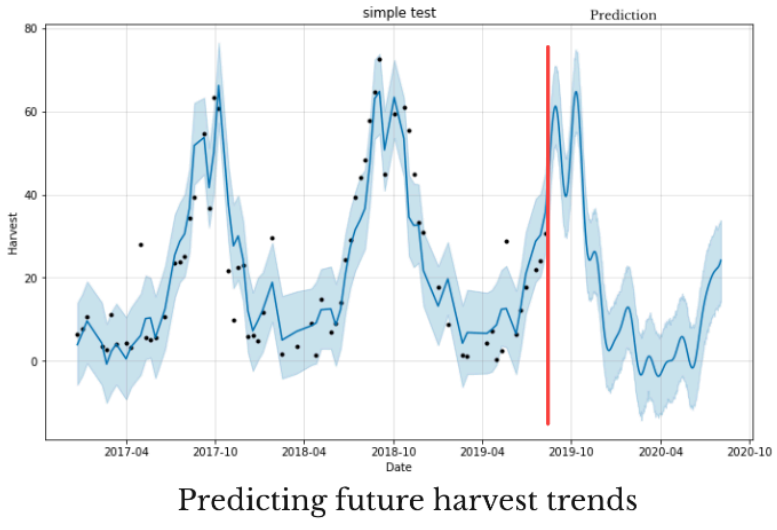
The estimation of the yield fluctuates every year and because of it, sugar mill industry is struggling in bringing the consistency. One of the major problem which cause these gaps in the estimation is the incapability of their system to incorporate various climatic factors in their calculation.

These factors are crucial as they can bring significant variations in the estimation. Factors like change in temperature, less or excessive rainfall, changes in soil moisture

## PROPOSED SOLUTION

The issues in avoiding gaps in the estimation call attention to the potential benefits of enhancing the capability of yield estimations and also, predicting the time period when the next crushing season of a sugar mill should start. It was realised that a fix to this issue will provide sufficient lead time for the industry to act upon the unforeseen.

Therefore a model using time series forecasting was build which targets to achieve the balance between the overestimation and underestimation. In other words, the model is configured to produce forecasting using the satellite images which will estimate the size and cycle of yield of the next season with a goal of achieving the results within 5% difference rate of the actual harvest and estimated harvest. This will allow to minimise the deviation between the estimated and observed crop.




Figure[2] Results from the FBprophet forecasting model

The proposed prediction is shown after the red line in the figure[2]. Using this prediction, planning will be easier for the industry. Such planning involves managing the expectations with the forward business partners, estimating the amount of sugar production, forward selling to physical sugar buyers, forward pricing on the future markets and development of storage and shipping schedules in cooperation with customers overseas. Accurate yield forecasting is a fundamental key to build a successful marketing strategy with greater confidence.
The proposed prediction can be communicated to the farmers through the sugar mills representatives to give them a personalised schedule of harvesting based on their particular region conditions. The proposed solution is scalable to integrate important factors like change in climate and rainfalls patterns.

---

## Scope

On Hack Days 1, the dataset was post on a blog on medium website by the Melbourne Datathon 2019 Organisers(Growing Data and ANZ Bank) . On the days 1, only the first phase of the data was released.

The data was produced by the Sentinel 2A satellite which is a remote sensing platform and is capable of image capture at wavelengths outside of what is human visible. The platform therefore generates imagery across different bands. ("Sentinel-2", 2019)

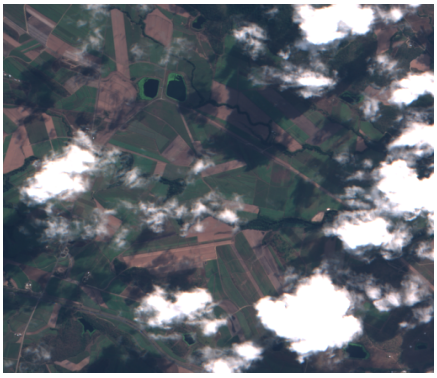
Dataset had satellite images of Queensland's sugarcane growing area just outside of Proserpine was provided to us. The first phase dataset contained images for a small region of the entire area. A folder "timeseries" contained 994 image files of that particular area. The images were taken approximately 10 days apart by the Sentinel 2A satellite, and provide imagery with a 10m per pixel resolution.

Each image was given in 12 bands ranging from December 2016 to August 2019. Each Bands has their uniques property using which different aspect can be determined for a particular area like amount of green, red, blue color , the vegetation Index or the amount of water vapour. There was a set of TCI images also which can be seen by the human eyes unlike other provided band images. Therefore, for the analysis, the TCI images included all the different color bands (blue,green,red) together and were used as the proportion of different color bands can be compared with each other.

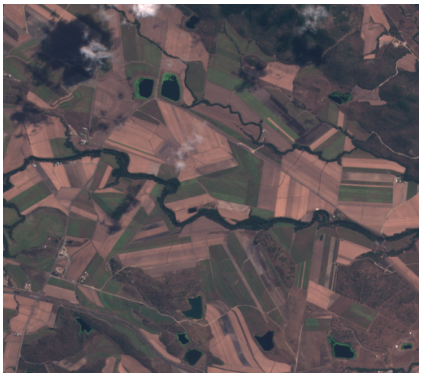
Each images tiles are of 512 pixels in height and width corresponds to an area of approximately 25 hector land. The images were provided in timeseries as each tile is time stamped by the date of capture.
Along with images there was json file providing meta data about the conditions of the capture, and its location in lat/long. To detect portion of the area which is growing sugarcane, a mask file was provided. In the second phase, they released 65 images of proserpine regions which was divided among 64,610 tiles.

Example of the TCI images

*Figure[3] Cultivated Land*



*Figure[4] Land with Cloud cover*



*Figure[5] Harvested Land*

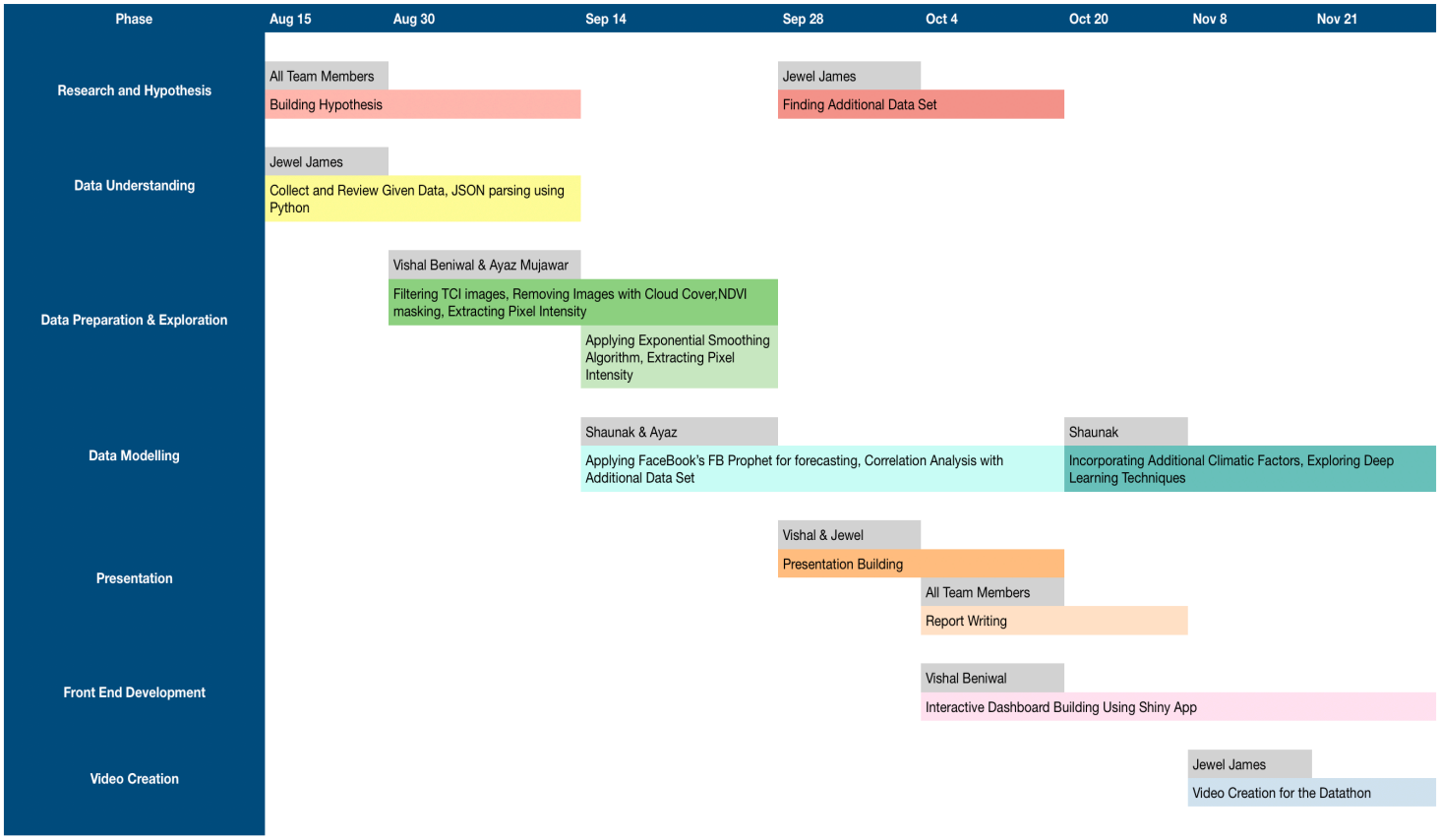
## Methodology

Range of different tools and techniques were used to examine and perform operations on the dataset. We used Python packages for crunching and analysing the satellite image datasets and Json files. We also used python scripts to scrape weather datasets from the websites
On hack day 1, we started to collect and explore the phase 1 data. To start with the process, one of the team member wrote a python code to parse and extract the data from the json file and saved it a data frame. Initial assessment and analyses were done using this parsed data frame and given 71 TCI images.
After the initial assessment, as the field of sugarcane was new to all the members of the team, so, thorough research was done using the research papers, sugar mills productions reports and news channels to know the lifecycle of sugar production and identifying the current pain areas that industry is facing.
The first main challenge faced was finding a hypothesis that can be solved using the satellite images. Based on the research, we got many different pain areas in the industry but very few that could be solved using the given dataset. This process took lot of our time.
Eventually after 2 weeks of brainstorming and discussions, team decided to work towards minimising the deviation between the estimated and observed crop. A roadmap was produced to work towards the decided goal and work was divided among the team members. Roadmap is shown below in the figure[6].

Another challenge was faced during the masking the data. Almost every other data science project has a challenge of removing the missing value. For this project, as shown in above Figure[4], cloud cover and their shadow on the land was our missing values and in order to get rid of them, we used Google AutoML Vision API. Following steps were used in order to remove images with 80% or more cloud covers.

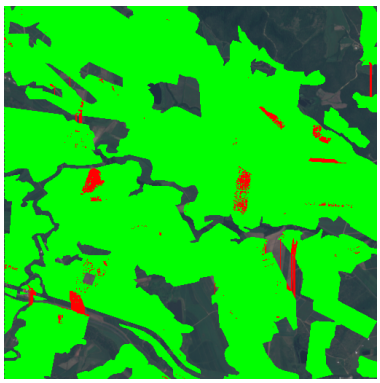
- Train the model by inserting images with the AutoML Vision API.
- Classified the images into two parts clouds and No clouds.
- Evaluated and Assessed the performance metric (precision, recall and confusion matrix) of the model.
- Uploaded the new images to tested the model.
- Saved those images that contains less than 70-80% of clouds and removed all other images

PROJECT ROADMAP

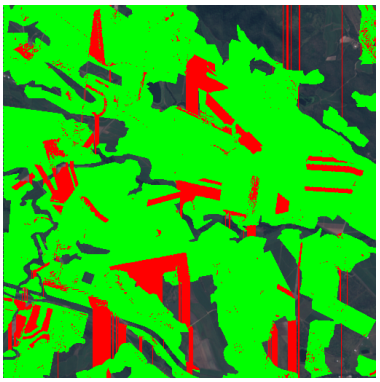



Figure[6] Project Roadmap

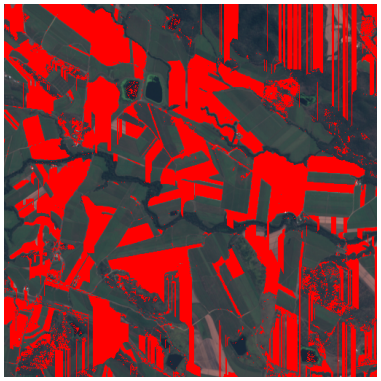
After this, we had the cleaned dataset with 67 TCI images on which masking was applied using the mask file provided in the phase 1 dataset. We separated the area of sugarcane area with green color and harvested area with red color and percentage of pixels with green and red color was extracted from the 67 images to create a data frame. This data frame was created to find the trend in the existing data given.



Figure[7] Image From February



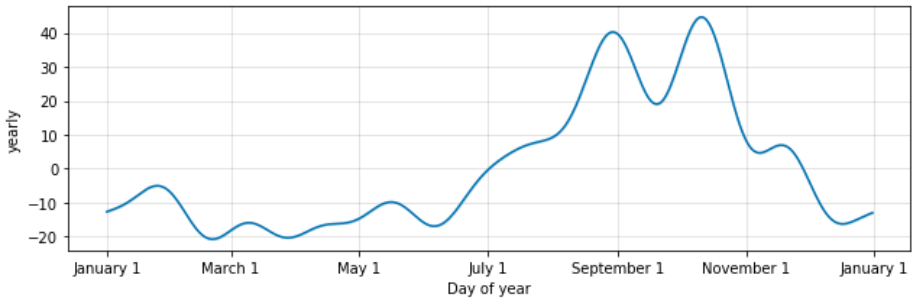
Figure[8] Image From June



Figure[9] Image From September

Figure[7] showed most of the sugarcane area with green color as its from the month of February and during this month, the rate of cultivation remain very high. On the other hand, figure[8] shows

little red patches which corresponds to the area of harvesting. Later the created data frame was applied to the Exponential Smoothing Algorithm to produce the underlying trends.




Figure[10] Harvesting trends of sugarcane

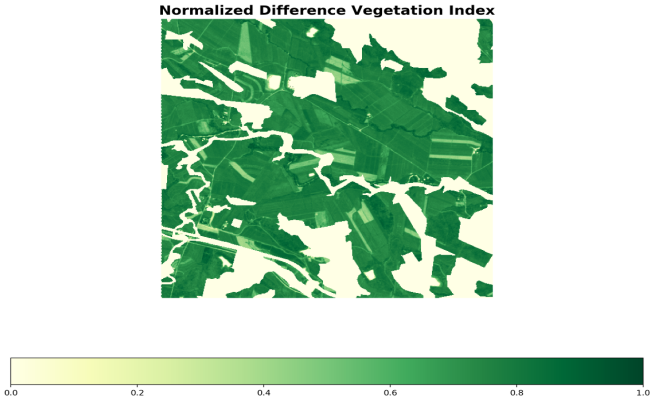
From the harvesting trends of sugarcane in QLD , it was observed that period of harvesting starts around the month of June, and reaches its peak in month of September and ends by the start of December. This step in our data exploration showed that the given data was matching with the usual harvesting pattern of Queensland and can be trusted to make further analysis.

However, color pixel intensity from these masked images could be used to predict the next harvesting cycle but couldn't be used to estimate the amount of sugarcane crop. Team did a research and found Normalized Distributed Vegetation Index NDVI is used to monitor the drought conditions, forecast the vegetation production and fire zones.("NDVI (Normalized Difference Vegetation Index) | Sentinel Hub", 2019)
NDVI is calculated by measuring the difference between near infrared and red light.

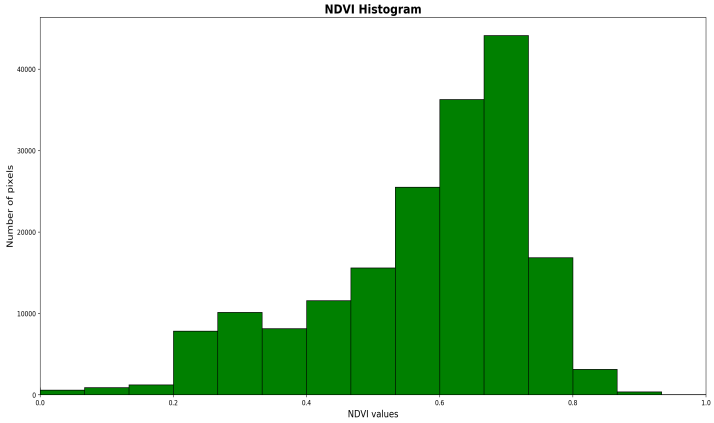
Formula to calculate NDVI: $NDVI = (NIR - RED) / (NIR + RED)$

*NIR - reflection in the near-infrared spectrum*
*RED - reflection in the red range of the spectrum*

Team had the images with band 8 to extract the NIR and images with band 4 to extract the red color reflection which could be used to extract the different proportion of green color showing the percentage amount of healthy crop using NDVI.




Figure[11] NDVI Index for a particular land



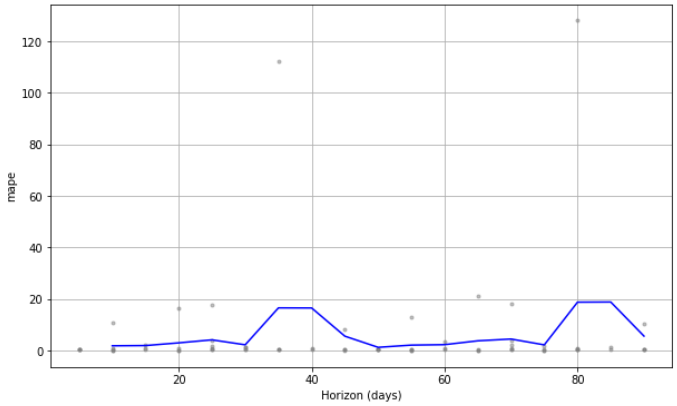

Figure[12] Historgram of NDVI Index

The value range of an NDVI is between -1 & 1. Negative values of NDVI ( values approaching -1) correspond to water. Values close to zero (-0.1 to 0.1) generally correspond to barren areas of rock, sand, or snow. Low, positive values represent shrub and grassland (approximately 0.2 to 0.4), while high values indicate temperate and tropical rainforests (values approaching 1). In our case, number of pixels in the range of (0.5 to 0.9) represented sugarcane getting ready for the harvesting. Therefore, Team worked to store the NDVI values of different areas in a data frame.

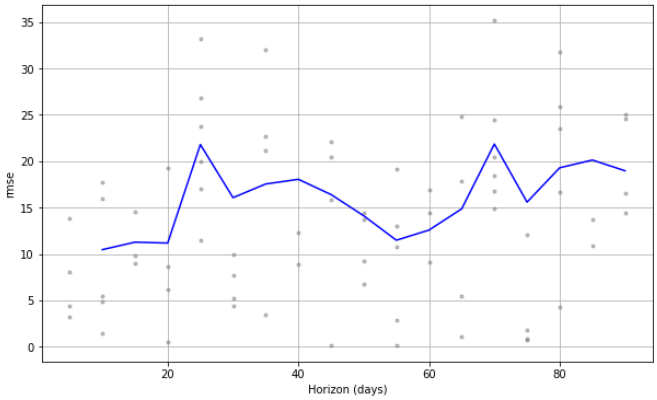
This data frame was later fed into Facebook's FBprophet model, which is an open source model works on the forecasting time series data to produce strong seasonal effects and several seasons of historical data. The team choose the Prophet model because of its unique advantages like it produces completely automated forecasts, it is robust to missing data and shifts in the trend, and also, handles outliers well.

## RESULTS AND CONCLUSION

Sugarcane Mills in Australia and across the globe depend heavily upon forecasting. A time-series forecast that can indicate the harvesting period accurately, and the harvest that can be expected every year could be crucial in generating maximum revenue. The Time-series forecast performed based on pixel intensities, NDVI (Normalised Difference Vegetation Index) calculation and historical data. The team found that the historical data showed a similar pattern over the period of last 3 years when the initial analysis was performed using Exponential smoothing, for the sugarcane data. This indicated that ideally, the sugarcane crop would follow almost the same time frame of development and maturity as in the previous years. However, the forecasted values were also accompanied with a lot of variation both above and below the indicated values.




Graph[1] Mean Absolute Performance Error (MAPE).




Graph[2] Root Mean Square Error (RMSE)

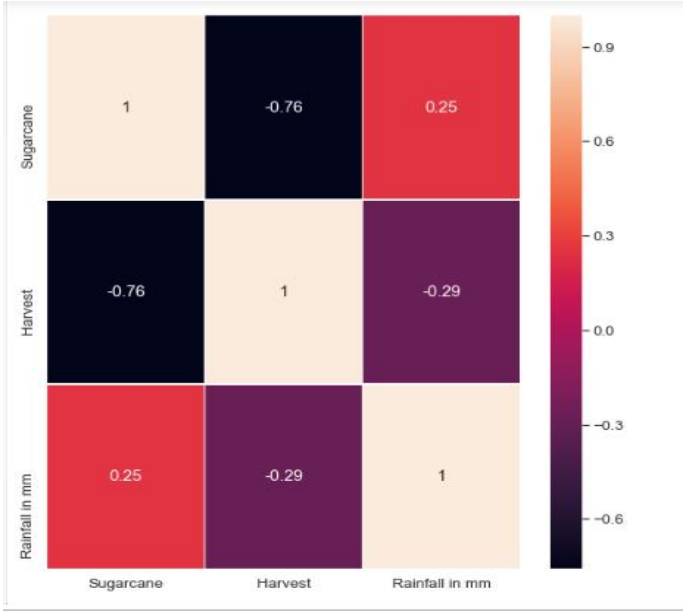
| | horizon | mse | rmse | mae | mape | coverage |
|---|---|---|---|---|---|---|
| 0 | 10 days | 109.422422 | 10.460517 | 8.598649 | 1.813034 | 0.571429 |
| 1 | 15 days | 126.957944 | 11.267562 | 9.969076 | 1.906893 | 0.342857 |
| 2 | 20 days | 124.756444 | 11.169442 | 9.702739 | 2.952340 | 0.142857 |
| 3 | 25 days | 474.754048 | 21.788851 | 20.116172 | 4.120257 | 0.035714 |
| 4 | 30 days | 258.035222 | 16.063475 | 13.332780 | 2.195827 | 0.428571 |

Table[1] Performance Metric

A look at some initial performance metrics, Table[1], such as Root Mean Square Error (RMSE) shown in graph[2] and Mean Absolute Performance Error (MAPE) shown in graph[1] on the configured forecasting model over different periods of days, say 10 days, 15 days, 20 days etc. indicates how well the forecasting model performed on the data provided. For sugarcane production, an error rate of approximately 20+ percent was observed, while the harvesting trend showed an error rate of approximately 10 – 20 percent. This performance metric showed that the modelling technique and the variables / attributes chosen could be improved by incorporating more attributes or using a more efficient forecasting modelling technique through further research.

An initial investigate into the climatic factors such as rainfall, temperature, soil etc. provided important insights into how different parameters can affect time-series. The Australian government recently declared significant levels of draught for about 2 / 3rd of Queensland. As such, a correlation analysis between rainfall and, sugarcane production and harvesting, during the life cycle of sugarcane crop was performed, to understand how much such a factor could affect our forecasting.

The analysis showed a small positive correlation between sugarcane production and rainfall, indicating that the sugarcane production could be affected a little for the next one year as well. The correlation between rainfall and harvesting was negative, since there is no way a harvesting period could be affected by changes in the amount of rainfall during that period. The small positive correlation can go a long way towards narrowing the variance of the forecast to a more accurate figure.




Figure[13] Correlation analysis between harvesting & rainfall

Other possible factors such as Temperature, Soil, Humidity etc. will help in a similar manner as historical data alone is insufficient. The weather in any region is never the same. It is subject to change at different points in time, possible affecting the growing conditions of the sugarcane crop and possibly the harvest. Incorporating such factors and weighing them against ideal conditions, could be crucial in producing an accurate forecast of the harvesting time and the possible harvest that can be expected. The idea of estimating the harvesting period for sugarcane depends on how well, the forecasting technique / algorithm would adjust to climatic factors and the changes that these climatic factors may induce. Hence, appropriate data and a model incorporating a weight equation is required in tackling the problem of over-estimating and under-estimating.

## WHAT's IN THE PIPELINE

Incorporating other factors Currently the forecasting model is learning only from the historical data that was extracted from the satellite images that was provided. Incorporating data on Soil, Weather and temperature will increase the efficiency of the forecasting model.

Building Web Application Started building the application using Shiny app. User Interface of the application is in development stage.

Exploring ways to optimize the transportation  Working on A* algorithm and trying to optimize the route.

Looking for ways to feed the data in real time to the application so that it will provide appropriate interactive dashboards to employees of sugarcane mill based on various factors.

## Appendix A – Project Management

**Team Members**

Below are the profiles of each team member and their individual contributions to this project.

Member 1 is a master student in RMIT pursuing Data Science and has a two years of prior experience in technical product management. He specialized in building process flows and in finding solutions to the bottlenecks. In this project, he contributed in by researching and building the hypothesis to work on. He also wrote a python code to create mask on the images in the required form. Overall, he acted as a proactive member of our team and participated at every level of the project.

Member 2 is a master student in RMIT pursuing Data Science and has a three years of prior experience in data scraping team at E-commerce firm. He specialises in data retrieval and processing. In this project, he contributed in by retrieving the initial data, architecting the presentation, researching the business impact of feature and driving the report creation.

Member 3 is pursuing Masters of Data Science from RMIT University and he has one year experience in developing web applications. He is also currently working as "Computer Science Research Officer" at RMIT University where his duty is to analyse and visualize data and also build web applications.In this project his role was to generate and support ideas by understanding and analyzing the data. He also contributed in removing the Cloud images using Google's cloud vision API. Overall he is one of the proactive member in the group and enjoyed working in group.

Member 4 is a master student at RMIT University pursuing Data Science and has a one year experience in Software and Web Development. He contributed towards developing Applications for clients in the hospitality, medical and educational industries. He developed code to generate NDVI images and subsequently visualizing the pixels associated with each level of NDVI. He also wrote code to perform time-series forecasting using Facebook's FBforecast, and generated the performance metrics for the same.

All the team members contributed to hypothesis building, brainstorming, presentation, report writing and group meetings

## Communication Tools

- Slack (Mainly)
- Whatsapp
- Mail

## Issues Encountered During The Project

- Lack of knowledge in forecasting techniques and deep learning
- Clash in schedules
- Team members weren't able to meet daily because of other commitments like assignment submission or part time work.

| Member Number | Member Names | % Contribution |
|---|---|---|
| Member 1 | Vishal Beniwal (s3759790) | 25 |
| Member 2 | James Jewel (s3763905) | 25 |
| Member 3 | Ayaz Mujawar (s3751555) | 25 |
| Member 4 | Shaunak Phaldessai (s3767517) | 25 |
| | Total | 100 |

## References

*1. Sentinel-2. (2019). Retrieved 22 October 2019, from https://en.wikipedia.org/wiki/Sentinel-2*
*2. Theory, C., & Analysis, T. (2019). Theory of Production: Short-Run | Intelligent Economist. Retrieved 22 October 2019, from https://www.intelligenteconomist.com/theory-of-production-short-run-analysis/*
*3. NDVI (Normalized Difference Vegetation Index) | Sentinel Hub. (2019). Retrieved 22 October 2019, from https://www.sentinel-hub.com/eoproducts/ndvi-normalized-difference-vegetation-index*
*4. (SOUTHERN OSCILLATION INDEX PHASES TO FORECAST 2019). Retrieved 22 October 2019, from https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.920*