University of California, Davis

Predicting Rainfall for Major Cities in Australia

BAX 452: Final Project

Aleksandr Udalov, Tram Le and Ran Zhang

Machine Learning

Professor Jörn Boehnke

March 19, 2022

Table of Contents

Executive Summary

In this project, we studied how to build a useful and accurate model to predict rainfall, an important factor in agriculture, in major cities in Australia. We will be using the data set from Kaggle which provides 10-year rainfall information in the location we are aiming to make predictions. The two major algorithms used in this project are Random Forest Classifier and Logistic Regression. Along with that, interpretability is also one of the major goals in this project as the target clients would be farmers who do not have much technical knowledge and require simpler explanations of the models as well the finding results. Therefore, we built a visualization dashboard using Tableau as one of the deliverables to better understand the model prediction. Finally, we analyzed the model results and made recommendations on model selection in predicting rainfall for agriculture.

Background, Context, and Domain Knowledge

In agriculture, one of the major factors that would affect the season outcome is weather conditions. Farmers make decisions based on daily weather changes, involving factors such as: fertilizing, crop growth and irrigation. Crops are sensitive to moisture, light, and temperature. Firstly, with historical and prediction information, it is easier to track crop growth and decide when and how much to irrigate. Secondly, fertilizer timing and delivery, which can ensure the effectivity by fitting process to specific weather conditions. For example, it is a well know agricultural fact that nitrogen fertilizers are the most efficient to be applied before rainfall if it is not going to be too intense. However, treatment of plants to protect against weeds is not efficient to be applied before the rain as applied treatment tools are going to be washed away from a field. Finally, under some extreme weather conditions, the field will not be suitable for farmers to work on.

As a result, the weather forecast has long been a focus for agriculture. Weather condition prediction is essential for precision agriculture and the ultimate goal of weather prediction is to optimize the growth efficiency of individual crops. Among the weather conditions, rainfall is one of the most important factors which is related to each of the three examples above since it directly influences the moisture level as well as field workability. So in our study, we will be using some

machine learning algorithms to predict rainfall in Australia which agriculture plays an important role in their total GDP.

Climate changes are a major factor affecting agriculture in Australia as variety of climates are present in the country due to its geographical size and placement. The main issue attributed to importance is presence of frequent droughts. As major part of Australian continent is a desert and according to Wikipedia “more than 80% of the country has an annual rainfall of less than 600 mm” [6]. Consequently, rainfall forecasts are an essential tool for agricultural entrepreneurs in those regions as their business processes are highly dependent on that factor.

Although the data set is limited to Australia, the model can always be expanded to other regions in which rainfall predictions would benefit in any mean.

Literature Review

It has a long history in predicting rainfall in agricultural studies. Usually, a study has a focus on a specific region. Early studies include statistical models, time series models, and recent studies introduced machine learning algorithms such as PCA, neural networks for better prediction.

Early in the 1950s, statistical models were applied to predict the monthly rainfall distribution of certain regions. The fundamental assumptions are fitting rainfall frequency distribution to a theoretical model. So it is a relatively strong assumption. Two famous models were normal distribution and incomplete gamma functions. Mills and Imama proved in their study that the gamma model provides a better fit to the rainfall distribution in the United States Virgin Islands.[2] More statistical models such as Markov Chain, ARIMA are used to predict rainfall in different regions[3][4].

But statistical models make strong assumptions such as linearity in data. But stochasticity in rainfall can be introduced by fog, special weather cases, etc. So moving recently, researchers have been using neural networks, which inherit non-linearity and feature interaction, to make climate predictions. Neural network models have the ability to model large-scale data without making strong data assumptions. Mohini et al also pointed out that RNN, FNN, and TDNN are suitable for rainfall forecasting, however, they are not good at daily rainfall forecasting.[5]

In terms of interpretability, it is harder to interpret models that used neural networks, which might significantly affect the use of the model of helping farmers understand why the rainfall possibilities are high or low for specific days. We also face the challenge of predicting day-to-day rainfall patterns, which is posed by the inability of neural network models.

Model Development and Analysis

Goal: Predict daily rainfall patterns in specific regions, allowing for correct interpretation.

Data

The data we adopted for the model is 10 years' daily weather observations from major areas in Australia. Features include *Date*, the date of observation; *Location*, the common name of the location of the weather station; *MinTemp*, the minimum temperature in degrees Celsius; *MaxTemp*, the maximum temperature in degrees celsius; *Rainfall*, the amount of rainfall recorded for the day in mm; *Evaporation*, the so-called Class A pan evaporation in the 24 hours to 9 am, etc. The target variable is a binary variable *RainTomorrow*, which indicates whether rainfall appears tomorrow. The complementary data used to visualize the model is the latitude and longitude of each unique area in the dataset, which is obtained from the open API geographical data source "positionstack.com".

Data cleaning process includes conversion of dates into the appropriate format, replacing empty values with average values and dropping columns with more than 10% of missing data for each location. Additionally, we convert all categorical variables from their string values to integer indices.

After cleaning the data, we have 48 major cities/areas in Australia adding up to 140k entries for the main data set.

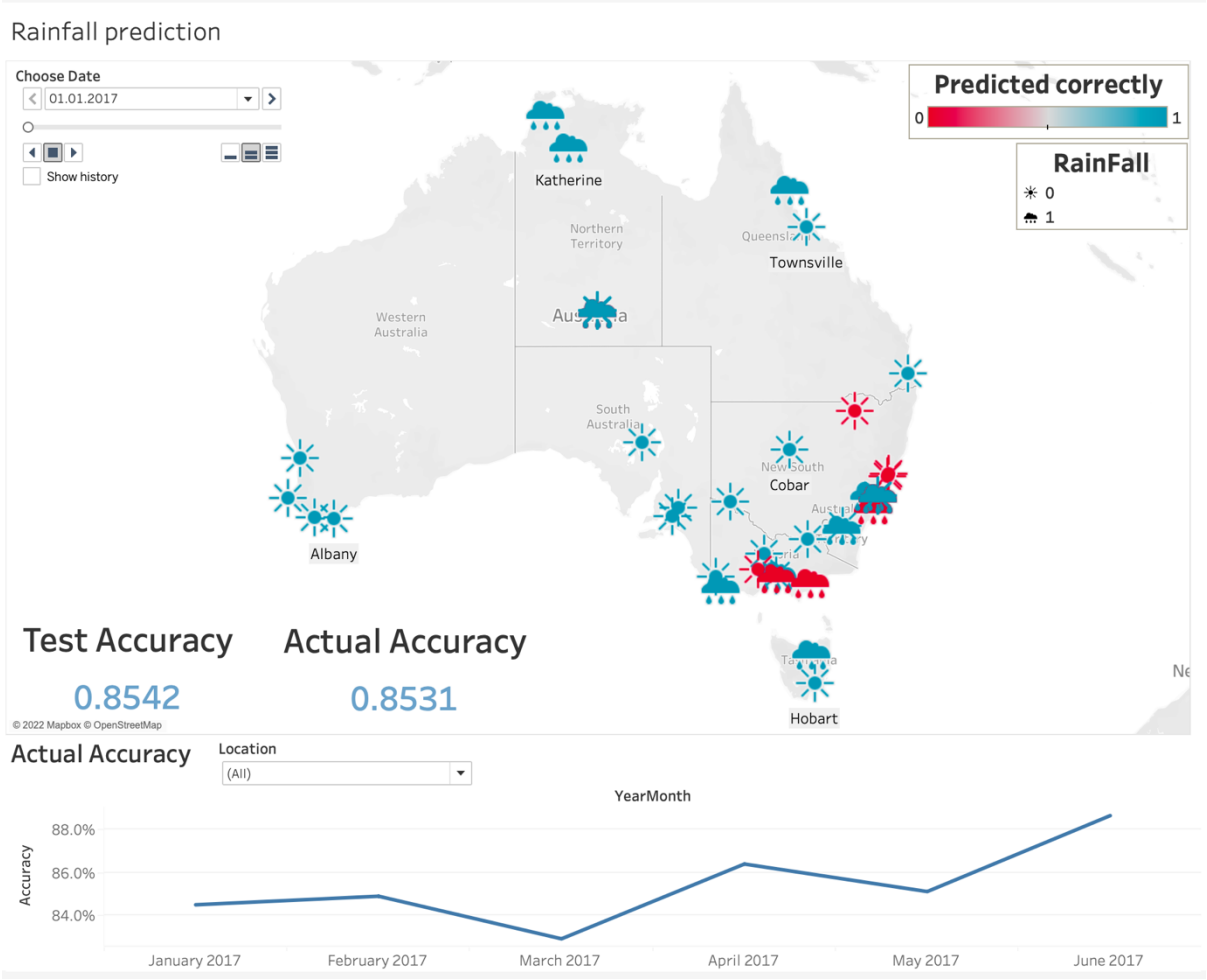
Model

With the consideration of interpretability and performance, we built a logistic regression model and random forest classifier to predict the value of target binary variable *RainTomorrow*. Both models are well-known in the weather forecast sectors. The coefficients in logistic regression are easy to interpret after log and odds transformation. For the random forest model, the intuition is weighted votes of different trees, and decision trees are able to interpret too.

Considering each location separately, we predict for each day whether it will rain or not tomorrow. Regarding in/out of sample split, we set data before 2016 as the in-sample set, and 2017 data as the out-of-sample set. Train-test split is 80 to 20 percent respectively. The prediction accuracy for both models is roughly similar, which is greater than 75% in the lowest-accurate city and generally greater than 85% on the whole test set. Out of sample predictions does not significantly differ in accuracy.

Presentation

We finally visualize the model prediction in Tableau, which is easier to comprehend for the general audience. Users would be able to choose the date to view the forecast and click on the triangle button to play the animation. Also, visualization allows to check prediction accuracy at different levels: location and date.



Model comparison

Regarding model comparison, we consider Random Forest Classifier to be slightly better due to stable output for the whole dataset. Despite both models provide similar accuracy, the Logistic Regression appears to have null prediction scores for several locations (see figure 1 in Appendix).

Recommendations/Business Implications

In this project, we developed a random forest model and logistic regression model to predict rainfall in different locations in Australia. The two models are strong in interpretation and easy to deploy. The two models achieve good performance with not many features. So one implication is that rainfall is not very difficult to predict, combining all historical models. So we might have multiple options in selecting models. But the question is what level of precision do we need? Do we need prediction in day, week or month? It is a factor that makes an impact on model selection. Another implication is how farmers utilize the prediction result. In our example, we made a visualization for rainfall prediction of different locations. But in other examples, it might be possible to integrate rainfall prediction with other weather condition predictions or observations to help farmers decide their strategy.

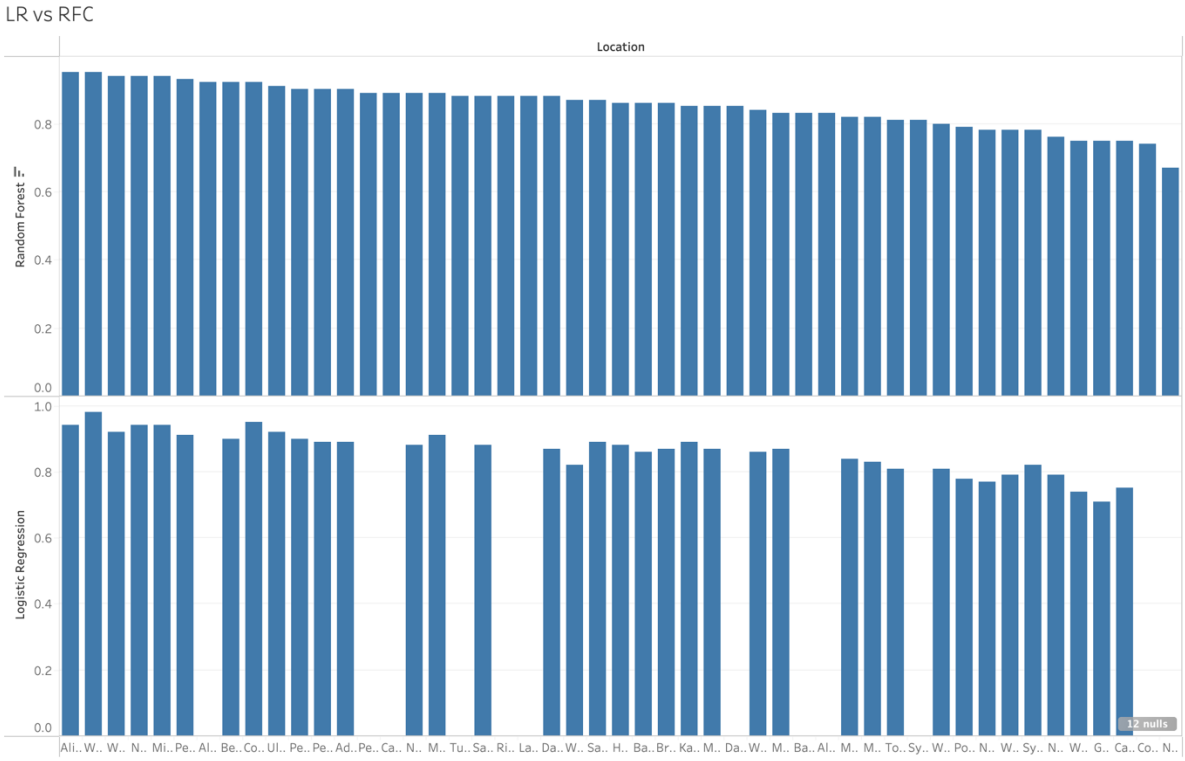
Summary and Conclusion

This paper is the report of how our team constructed prediction models for rainfall with the purpose of helping agriculture workers. The models are random forest classifier and logistic regression. We visualized the model in Tableau for better interpretability and understanding. Also, we have made several recommendations and further implications based on the knowledge of literature review and our model development and output. Furthermore, the model can be extended to multiple regions and weather conditions to test its usefulness.

Appendix

[1] Rain_forecast.twbx
[2] ML_FinalProject_LR.ipynb
[3] ML_Final_project.ipynb

Figure 1 – RFC vs Logistic Regression



Reference

[1] Eilts, M. (2018, November 27). *The Role of Weather—and Weather Forecasting—in Agriculture*. DTN. https://www.dtn.com/the-role-of-weather-and-weather-forecasting-in-agriculture/

[2] Mills, F., & Imama, E. (1990, March). *Rainfall Prediction for Agriculture and Water Resource Management in the United States Virgin Islands* (No. 33). Water Resources Research Center University of the Virgin Islands St. Thomas VI 00802. https://www.uvi.edu/files/documents/Research_and_Public_Service/WRRI/rainfall_prediction_for_agriculture_and_water_resource_management.pdf

[3] Graham, A., & Mishra, E. (2017). Time Series analysis model to forecast rainfall for Allahabad region. *Journal of Pharmacognosy and Phytochemistry*, *6*(5), 1418–1421. https://www.phytojournal.com/archives/2017/vol6issue5/PartU/6-5-80-125.pdf

[4] Ochola, W.O. and Kerkides, P. (2003), A Markov chain simulation model for predicting critical wet and dry spells in Kenya: analysing rainfall events in the Kano Plains. Irrig. and Drain., 52: 327-342. https://doi.org/10.1002/ird.94

[5] Darji, M. P., Dabhi, V. K., & Prajapati, H. B. (2015). Rainfall forecasting using neural network: A survey. *2015 International Conference on Advances in Computer Engineering and Applications*. https://doi.org/10.1109/icacea.2015.7164782

[6] *Climate of Australia*. (2022). Wikipedia. https://en.wikipedia.org/wiki/Climate_of_Australia#Rain