UNIVERSITY OF SOUTH FLORIDA

DATA MINING

Project On

PREDICTING RAINS IN AUSTRALIA

Team Members:

Surabhi Tripathi

Kavya Buchikonda

Akash Thallada

Simran Agichani

Rishabh Singh

Table of Contents

Introduction

The climatic conditions have changed drastically around the world in the last decade affecting millions of people and loss of billions of dollars. Climate change has fanned Australia's forest fires resulting in great habitat loss. Australia is known to have extreme weather condition which make it an ideal continent to model our study of rainfall prediction.

Dataset Overview

The inspiration and source of dataset is Kaggle. The dataset has 1,45,460 rows and 23 columns as below:

Columns

Numerical:

- MinTemp
- MaxTemp
- RainFall
- Evaporation
- Sunshine
- WindGustSpeed
- WindSpeed9am
- WindSpeed3pm
- Humidity9am
- Humidity3pm
- Pressure9am
- Pressure3pm
- Cloud9am
- Cloud3pm
- Temp9am
- Temp3pm

Categorical:

- MinTemp
- Date
- Location
- WindGustDir
- WindDir9am
- WindDir3pm
- RainToday

Target:

Rain Tomorrow

Dataset link: https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package

Problem Statement

Australia has highly variable weather due to its geography, hence rain prediction affects facets like when crops should be planted, how urban development should be planned, what measures to take in case of forest fires and floods. Extreme climatic conditions can threaten lives, resources, infrastructure, and productivity. Rain has also affected T20 cricket world cup held in Australia this year (2022). Zimbabwe and South Africa split points because the match in Hobart was abandoned because of rain. A shower brought a premature end to the England versus Ireland game.

Managing the impact of extreme climatic conditions completely relies on awareness and preparedness which requires accurate forecasts of weather days ahead, reliable outlooks of next season’s conditions, and robust projections of climate out to years.

This dataset contains 10 years (2007 to 2017) of daily weather observations from many locations across Australia. RainTomorrow is the target variable to predict. It means -- did it rain the next day, Yes or No? This column is Yes if the rain for that day was 1mm or more.

Objective

Predict next-day rain by training classification models on the target variable RainTomorrow. We performed EDA (Exploratory Data Analysis) through data preprocessing and feature engineering. Further, ran few ML models to predict if it will rain then next day - or not.

Exploratory Data Analysis

Data-Preprocessing

- Removed all the rows that contained null values in our target variable Rain tomorrow. This resulted in a reduction of 255 rows. Compared to over 14000 rows in total that we have, this is an acceptable reduction.

- Checked for any column that contained over 50% null values in their cells. None such column existed so no column was removed. All columns had less than 50% null values.

- Checked for cardinality in the columns:

  - Checked for categorical columns - 7 such columns
  - Of all the categorical columns- date had 3436 unique values which results in high cardinality and requires feature engineering on date.

- In the numerical columns - replaced all the null values with the mean of each column respectively.

- In the categorical columns - replaced all the null values with mode of each column respectively.

Feature Engineering

- High Cardinality in Date

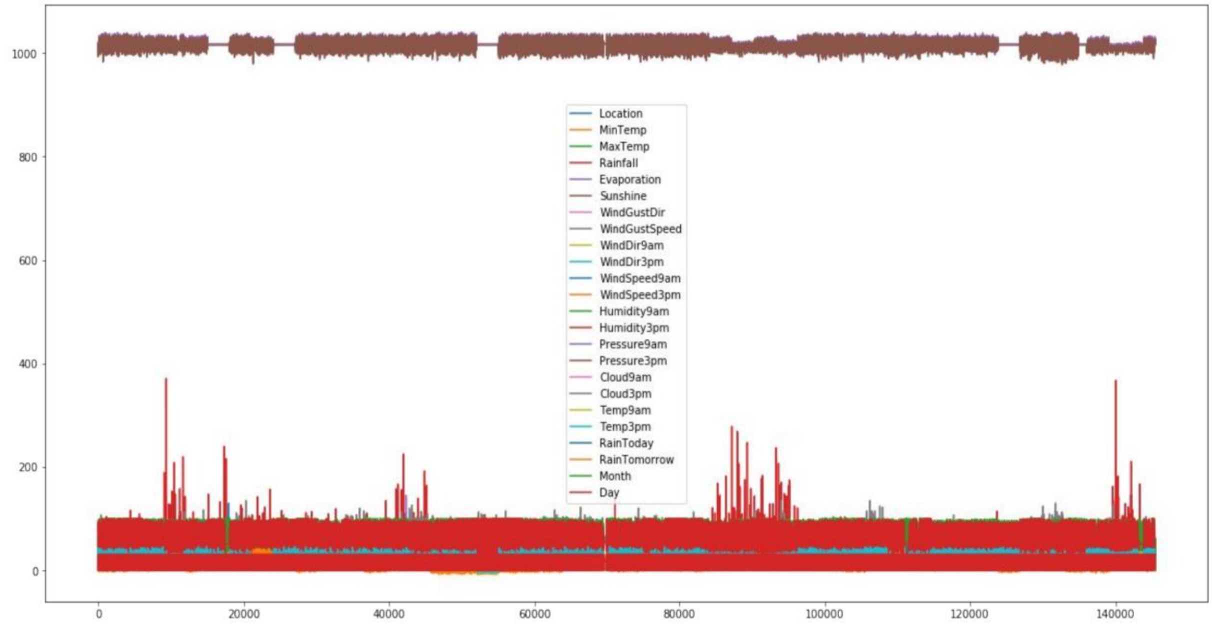
  Solution - Separated day and month from date column, then appended day and month columns and dropped date from the dataset.

  This reduces cardinality as only 12 unique values on month exist, and 31 unique values for day exist compared to 3436 unique values of date alone.

  Note - On experimental trial we also observed that if we included both day and month compared to only one of them, our algorithm models' accuracy improved, marginally by 0.01(approximately).

- Label encoded all the categorical values - from string to numbers.

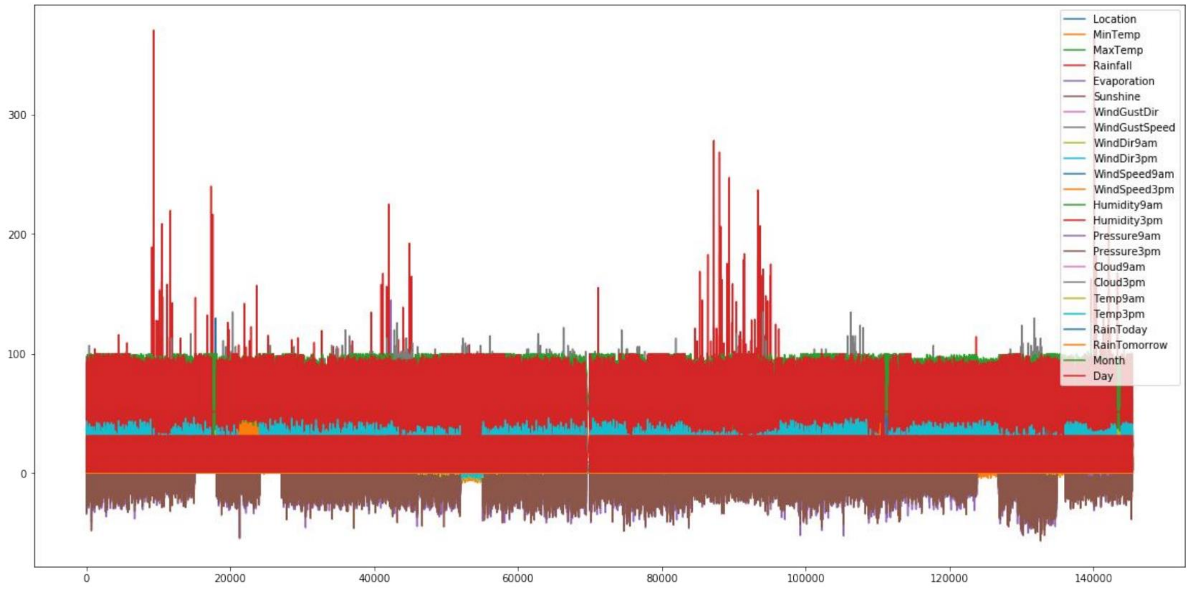
Plotting the graph for all the values:



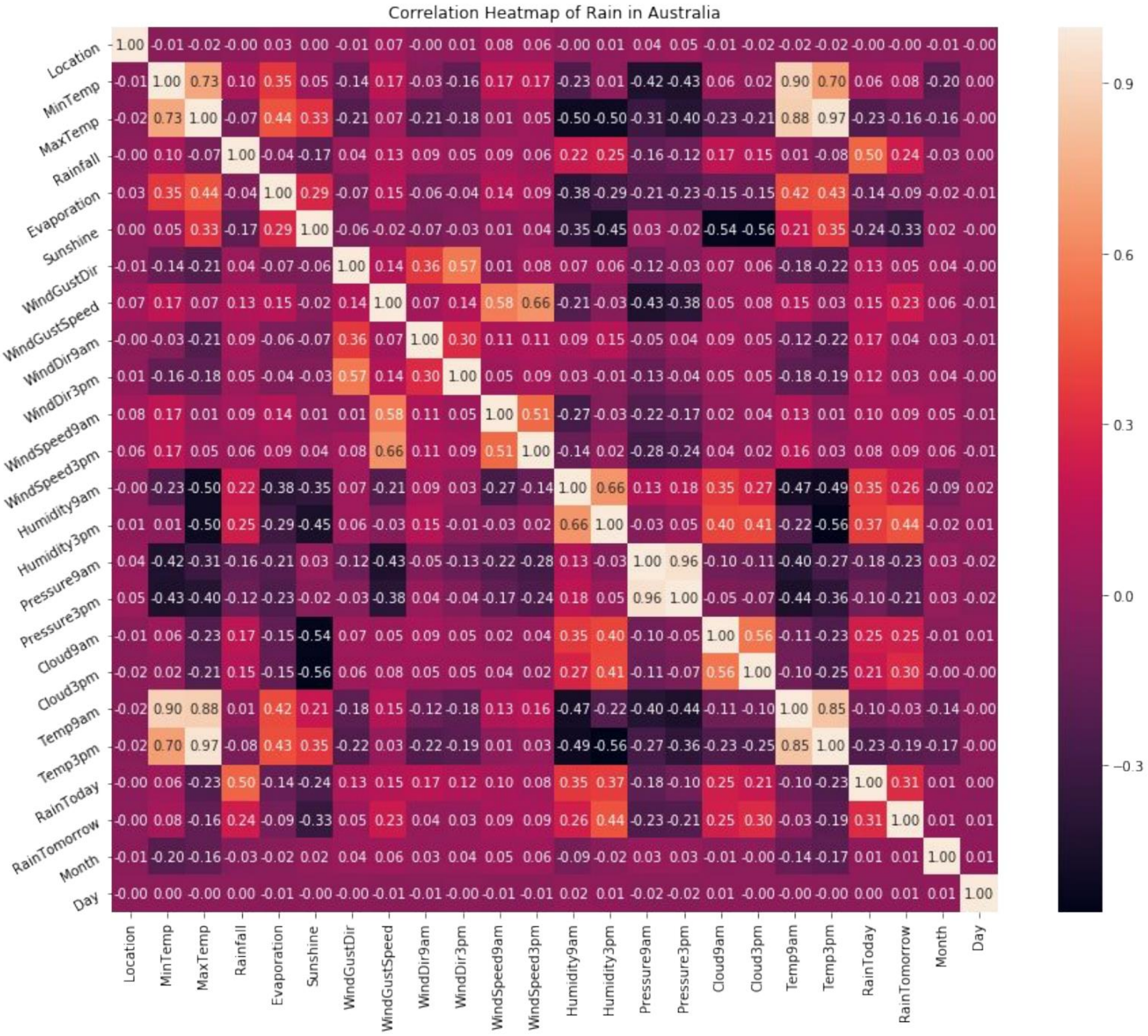
Feature Engineering – Normalization

We noticed that while all other columns have values that are in close range, the value of Pressure9am and Pressure3pm is much higher.

To normalize these columns, we used z-score and then multiplied it by 10. The new range observed was in range - (-55, 35) which brings it quite close to other columns' values.



Feature Engineering – Correlation Heatmap



	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow	Month	Day
Location	1.00	-0.01	-0.02	-0.00	0.03	0.00	-0.01	0.07	-0.00	0.01	0.08	0.06	-0.00	0.01	0.04	0.05	-0.01	-0.02	-0.02	-0.02	-0.00	-0.00	-0.01	-0.00
MinTemp	-0.01	1.00	0.73	0.10	0.35	0.05	-0.14	0.17	-0.03	-0.16	0.17	0.17	-0.23	0.01	-0.42	-0.43	0.06	0.02	0.90	0.70	0.06	0.08	-0.20	0.00
MaxTemp	-0.02	0.73	1.00	-0.07	0.44	0.33	-0.21	0.07	-0.21	-0.18	0.01	0.05	-0.50	-0.50	-0.31	-0.40	-0.23	-0.21	0.88	0.97	-0.23	-0.16	-0.16	-0.00
Rainfall	-0.00	0.10	-0.07	1.00	-0.04	-0.17	0.04	0.13	0.09	0.05	0.09	0.06	0.22	0.25	-0.16	-0.12	0.17	0.15	0.01	-0.08	0.50	0.24	-0.03	0.00
Evaporation	0.03	0.35	0.44	-0.04	1.00	0.29	-0.07	0.15	-0.06	-0.04	0.14	0.09	-0.38	-0.29	-0.21	-0.23	-0.15	-0.15	0.42	0.43	-0.14	-0.09	-0.02	-0.01
Sunshine	0.00	0.05	0.33	-0.17	0.29	1.00	-0.06	-0.02	-0.07	-0.03	0.01	0.04	-0.35	-0.45	0.03	-0.02	-0.54	-0.56	0.21	0.35	-0.24	-0.33	0.02	-0.00
WindGustDir	-0.01	-0.14	-0.21	0.04	-0.07	-0.06	1.00	0.14	0.36	0.57	0.01	0.08	0.07	0.06	-0.12	-0.03	0.07	0.06	-0.18	-0.22	0.13	0.05	0.04	-0.00
WindGustSpeed	0.07	0.17	0.07	0.13	0.15	-0.02	0.14	1.00	0.07	0.14	0.58	0.66	-0.21	-0.03	-0.43	-0.38	0.05	0.08	0.15	0.03	0.15	0.23	0.06	-0.01
WindDir9am	-0.00	-0.03	-0.21	0.09	-0.06	-0.07	0.36	0.07	1.00	0.30	0.11	0.11	0.09	0.15	-0.05	0.04	0.09	0.05	-0.12	-0.22	0.17	0.04	0.03	-0.01
WindDir3pm	0.01	-0.16	-0.18	0.05	-0.04	-0.03	0.57	0.14	0.30	1.00	0.05	0.09	0.03	-0.01	-0.13	-0.04	0.05	0.05	-0.18	-0.19	0.12	0.03	0.04	-0.00
WindSpeed9am	0.08	0.17	0.01	0.09	0.14	0.01	0.01	0.58	0.11	0.05	1.00	0.51	-0.27	-0.03	-0.22	-0.17	0.02	0.04	0.13	0.01	0.10	0.09	0.05	-0.01
WindSpeed3pm	0.06	0.17	0.05	0.06	0.09	0.04	0.08	0.66	0.11	0.09	0.51	1.00	-0.14	0.02	-0.28	-0.24	0.04	0.02	0.16	0.03	0.08	0.09	0.06	-0.01
Humidity9am	-0.00	-0.23	-0.50	0.22	-0.38	-0.35	0.07	-0.21	0.09	0.03	-0.27	-0.14	1.00	0.66	0.13	0.18	0.35	0.27	-0.47	-0.49	0.35	0.26	-0.09	0.02
Humidity3pm	0.01	0.01	-0.50	0.25	-0.29	-0.45	0.06	-0.03	0.15	-0.01	-0.03	0.02	0.66	1.00	-0.03	0.05	0.40	0.41	-0.22	-0.56	0.37	0.44	-0.02	0.01
Pressure9am	0.04	-0.42	-0.31	-0.16	-0.21	0.03	-0.12	-0.43	-0.05	-0.13	-0.22	-0.28	0.13	-0.03	1.00	0.96	-0.10	-0.11	-0.40	-0.27	-0.18	-0.23	0.03	-0.02
Pressure3pm	0.05	-0.43	-0.40	-0.12	-0.23	-0.02	-0.03	-0.38	0.04	-0.04	-0.17	-0.24	0.18	0.05	0.96	1.00	-0.05	-0.07	-0.44	-0.36	-0.10	-0.21	0.03	-0.02
Cloud9am	-0.01	0.06	-0.23	0.17	-0.15	-0.54	0.07	0.05	0.09	0.05	0.02	0.04	0.35	0.40	-0.10	-0.05	1.00	0.56	-0.11	-0.23	0.25	0.25	-0.01	0.01
Cloud3pm	-0.02	0.02	-0.21	0.15	-0.15	-0.56	0.06	0.08	0.05	0.05	0.04	0.02	0.27	0.41	-0.11	-0.07	0.56	1.00	-0.10	-0.25	0.21	0.30	-0.00	-0.00
Temp9am	-0.02	0.90	0.88	0.01	0.42	0.21	-0.18	0.15	-0.12	-0.18	0.13	0.16	-0.47	-0.22	-0.40	-0.44	-0.11	-0.10	1.00	0.85	-0.10	-0.03	-0.14	-0.00
Temp3pm	-0.02	0.70	0.97	-0.08	0.43	0.35	-0.22	0.03	-0.22	-0.19	0.01	0.03	-0.49	-0.56	-0.27	-0.36	-0.23	-0.25	0.85	1.00	-0.23	-0.19	-0.17	-0.00
RainToday	-0.00	0.06	-0.23	0.50	-0.14	-0.24	0.13	0.15	0.17	0.12	0.10	0.08	0.35	0.37	-0.18	-0.10	0.25	0.21	-0.10	-0.23	1.00	0.31	0.01	0.00
RainTomorrow	-0.00	0.08	-0.16	0.24	-0.09	-0.33	0.05	0.23	0.04	0.03	0.09	0.09	0.26	0.44	-0.23	-0.21	0.25	0.30	-0.03	-0.19	0.31	1.00	0.01	0.01
Month	-0.01	-0.20	-0.16	-0.03	-0.02	0.02	0.04	0.06	0.03	0.04	0.05	0.06	-0.09	-0.02	0.03	0.03	-0.01	-0.00	-0.14	-0.17	0.01	0.01	1.00	0.01
Day	-0.00	0.00	-0.00	0.00	-0.01	-0.00	-0.00	-0.01	-0.01	-0.00	-0.01	-0.01	0.02	0.01	-0.02	-0.02	0.01	-0.00	-0.00	-0.00	0.00	0.01	0.01	1.00

Insights from the Correlation Heatmap:

- MaxTemp and Temp3pm have an extremely high correlation - 0.97
- MinTemp and Temp9am have a high correlation - 0.90
- MaxTemp and Temp9am have a moderately high correlation - 0.88
- Location and RainTomorrow had no correlation - 0.0

Since MaxTemp and Temp3pm had extremely high correlation, Temp3pm was dropped. A small improvement in accuracy of ML model was observed.

We also noticed that location and RainTomorrow had no correlation, but our accuracy of ML model decreased when location was dropped.

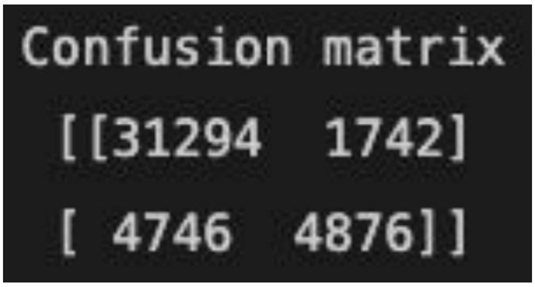
Data Modelling

We carried out our analysis by looking at five different models:

Model	Accuracy
K Nearest Neighbour (KNN, k=3)	82.9%
K Nearest Neighbour (KNN, k=10)	84.3%
Random Forest	84.5%
Decision Trees	78.4%
Adaboost	84.8%

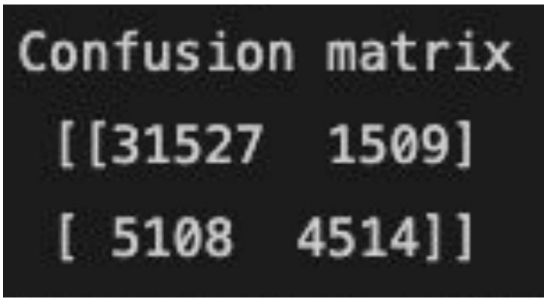
Adaboost and Random Forest gave the best results:

Adaboost:



```
Confusion matrix
 [[31294  1742]
 [ 4746  4876]]
```

Random Forest:



```
Confusion matrix
 [[31527  1509]
 [ 5108  4514]]
```

Observations, Roadblocks and Suggested Solutions

- Predicting rain is one of the most difficult tasks and to deal with advanced science and precision tools, rain prediction stays a difficult job, and the probability is always varying.

- Despite that, based on climate and history - our algorithms have given an accuracy of over 85 per cent after preprocessing and engineering steps. This efficiency can further be improved if outliers are worked on.

- This data is limited to the locations of Australia and is only for last 10 years. Expansion of data can help with better results and this methodology can be used over other countries and locations as well. It can help predict storms and heavy rainfall if worked upon correctly.

Prospects

- Weather forecasts are now considerably more reliable than they were few years ago. Three-day forecasts now have the same accuracy as two-day forecasts did prior to 2009. As development continues, the new challenge is to improve seasonal forecasts, from weeks out to months. Even more challenging will be predictions out to years.

- Improved short-term, multi-week and seasonal forecasts will enhance farmers' ability to make the right business decisions. It will help water authorities, industry and the community to better manage water use, while climate projections will inform planning for long-term water supply.

- The heatwave event of January 2009 caused financial losses of approximately $800 million due to power outages, transport disruption and associated responses. Tailored forecasts of temperature, humidity, precipitation, wind speeds can assist in predicting demand for heating, cooling and lighting, and in turn aid decisions about power generation. They can also precipitate protective action in the case of extreme weather.

- The annual economic impact of weather and climate events is estimated at around 5 per cent of Australia's GDP. Better information and more accurate forecasts lead to better decision-making, which may help mitigate losses. Recent World Bank studies suggest that the avoidable damage related to infrastructure from early warning systems with lead times of 24 hours, 48 hours and up to seven days could be around 5, 10 and 15 per cent, respectively.