Data Science Engineering Methods

Rain in Australia

Ashay Kanade 001548347

Akshay Bhosle 001001091

Shubhankar Salvi 001541699

Contents :

Introduction

This dataset contains about 10 years 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.

Columns :

Heading		Meaning	Units
Date		Day of the month	
Day		Day of the week	first two letters
Temps	**Min**	Minimum temperature in the 24 hours to 9am. Sometimes only known to the nearest whole degree.	degrees Celsius
	Max	Maximum temperature in the 24 hours from 9am. Sometimes only known to the nearest whole degree.	degrees Celsius
Rain		Precipitation (rainfall) in the 24 hours to 9am. Sometimes only known to the nearest whole millimetre.	millimetres
Evap		"Class A" pan evaporation in the 24 hours to 9am	millimetres
Sun		Bright sunshine in the 24 hours to midnight	hours
Max wind gust	**Dirn**	Direction of strongest gust in the 24 hours to midnight	16 compass points
	Spd	Speed of strongest wind gust in the 24 hours to midnight	kilometres per hour
	Time	Time of strongest wind gust	local time hh:mm
9 am	**Temp**	Temperature at 9 am	degrees Celsius
	RH	Relative humidity at 9 am	percent
	Cld	Fraction of sky obscured by cloud at 9 am	eighths
	Dirn	Wind direction averaged over 10 minutes prior to 9 am	compass points
	Spd	Wind speed averaged over 10 minutes prior to 9 am	kilometres per hour
	MSLP	Atmospheric pressure reduced to mean sea level at 9 am	hectopascals
3 pm	**Temp**	Temperature at 3 pm	degrees Celsius
	RH	Relative humidity at 3 pm	percent
	Cld	Fraction of sky obscured by cloud at 3 pm	eighths
	Dirn	Wind direction averaged over 10 minutes prior to 3 pm	compass points
	Spd	Wind speed averaged over 10 minutes prior to 3 pm	kilometres per hour
	MSLP	Atmospheric pressure reduced to mean sea level at 3 pm	hectopascals

Problem Statement :

A koala bear lounges in a eucalyptus tree, lazily chewing leaves in a scene as Australian as one can imagine. But then the koala does something unusual: it climbs down the tree to drink water from a pool. Normally, koalas get nearly all their water from their food, and researchers have linked this new behavior to climate change. Australia is getting hotter and drier, the eucalyptus leaves are less succulent, and now koalas need to drink extra water to survive. It's a change that seems to reflect a growing awareness among Australians that they need to be more proactive about securing their fresh water.

Straddling the Tropic of Capricorn, Australia's climate ranges from a tropical north to a temperate south but the vast bulk of its three million square miles is hot and dry. Central Australia's immense 'outback' is made up of semi-arid bush and deserts where temperatures can soar above

50°centigrade and it might not rain for years. This makes Australia the world's driest inhabited continent—and it's getting drier. Its average annual rainfall is around 470mm a year, well below the global average, and predictions linked to climate change suggest this could halve again in coming decades. What rain that does fall varies greatly from year to year and is concentrated along the north and east coasts: while most of Australia receives as little as 600mm of rain each year, half the country gets less than 300mm.

2018 was particularly dry, 11% below the recorded mean for 1961-1990 at under 413mm. However, across the continent, Australia's rainfall is exceptionally varied and individual states felt the effects to different degrees. While Tasmania retained its very wet average with 1,389mm of rain, and Western Australia actually had 10% more rain than usual, other states suffered. The Northern Territory received 7% below average rainfall and Queensland 15% below average; Victoria was 26% below its mean rainfall and South Australia was 24% below with an average rainfall of just 171mm. But it was New South Wales that was among the hardest hit, with its average annual rainfall down 40% on the mean, bringing a devastating drought and severe restrictions across the state.

Australia's exceptional aridity is the result of a unique combination of factors. Cold ocean currents off the west coast means there is little evaporation to form rain clouds, while the Great Dividing Range that runs down Australia's east coast prevents rain from penetrating far inland. There are few mountains to force air upwards where it can cool into rain, and the region is dominated by the subtropical high-pressure belt that both warms and dries the air. What's more, the continent is extremely susceptible to the El Niño–Southern Oscillation, a heating or cooling of the Pacific Ocean that can bring prolonged periods of high temperatures and drought.

This article was published by National Geographic about Failing Rains and Thirsty Cities: Australia's Growing Water Problem.This adversity faced by Australian's Kindled our innocuous mind study this problem and DSEM project gave us an opportunity to give our five cents to help the natural crisis.

We will try to answer the question of whether or not it will rain tomorrow in Australia. We implemented kNN, Decision tree, Random Forest with Python and Scikit-Learn.

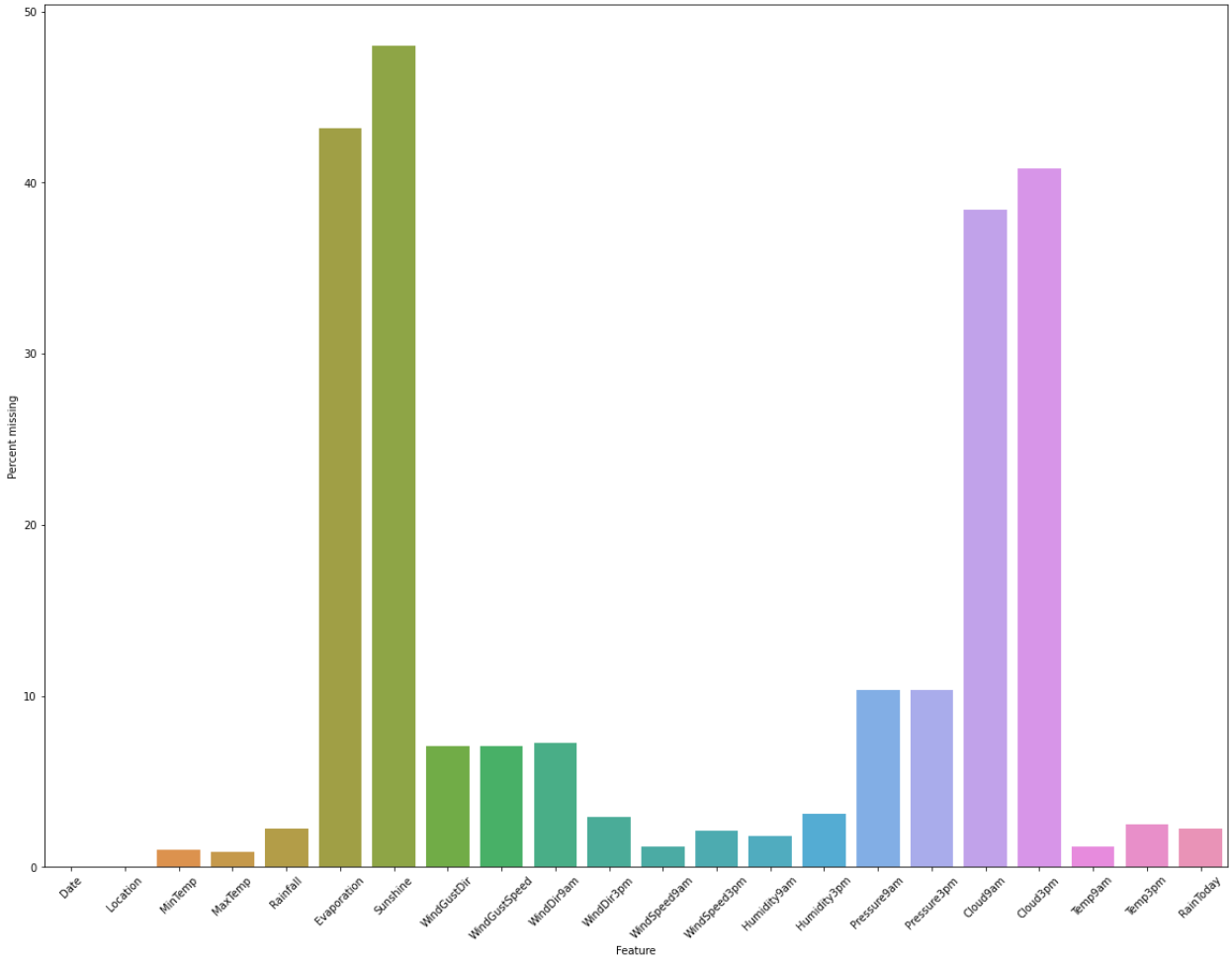
We have used the **Rain in Australia** dataset for this project.

Preprocessing :

- **Finding Missing Values**

  For Categorical Values :

We used mode values to impute Categorical Variables.



For Numeric Variables :

Simple Imputers.

We used Simple imputers to impute medians of all the values of columns to impute the missing values but using simple imputers will not be an intelligent guess to store a value in place of a missing value tuple which could have provided a well informed value which could have been more valuable instead of just assessing a median value.

Iterative Imputation.

Iterative imputation refers to a process where each feature is modeled as a function of the other features, e.g. a regression problem where missing values are predicted. Each feature is imputed sequentially, one after the other, allowing prior imputed values to be used as part of a model in predicting subsequent features.

It is iterative because this process is repeated multiple times, allowing ever improved estimates of missing values to be calculated as missing values across all features are estimated.

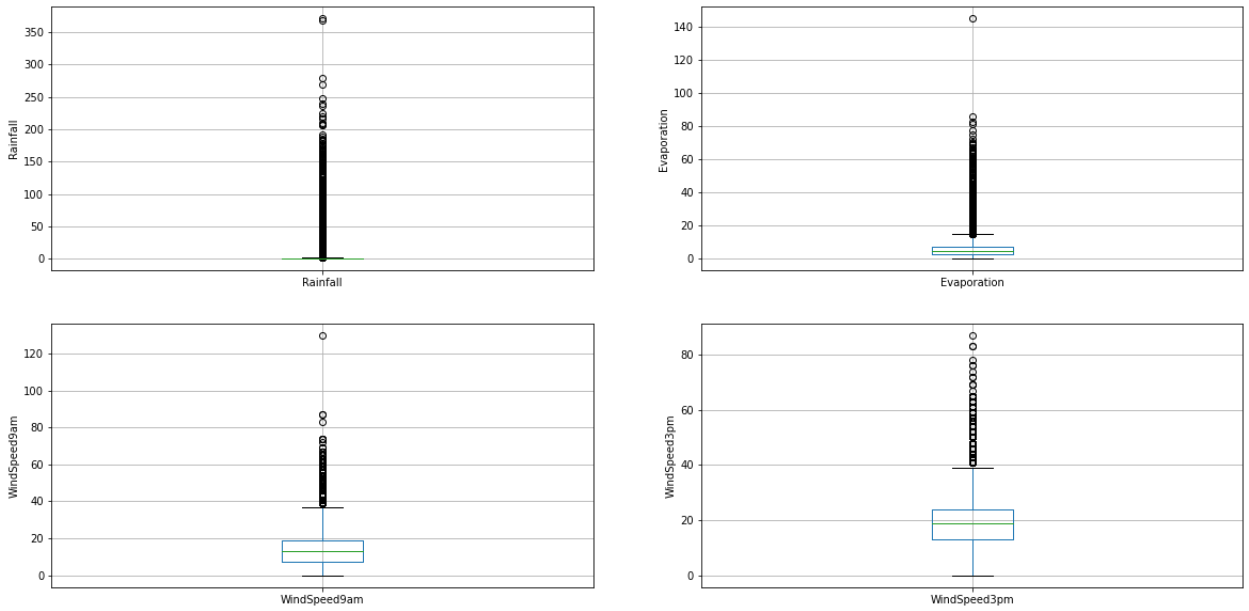
This approach may be generally referred to as fully conditional specification (FCS) or multivariate imputation by chained equations (MICE).

kNN Imputation.

An effective approach to data imputing is to use a model to predict the missing values. A model is created for each feature that has missing values, taking as input values of perhaps all other input features. This is where kNN imputation comes into picture. We Used kNN Imputation but it took approximately an hour to process which led us to thinking that there might be outliers because kNN was taking such a long time. This took us to the next problem which was finding outliers.

Finding Outliers :

Firstly we used the Describe() function to find variables where we were able to see the differences in the Quartile Ranges, Because of which we were able to find outliers in the following four variables.

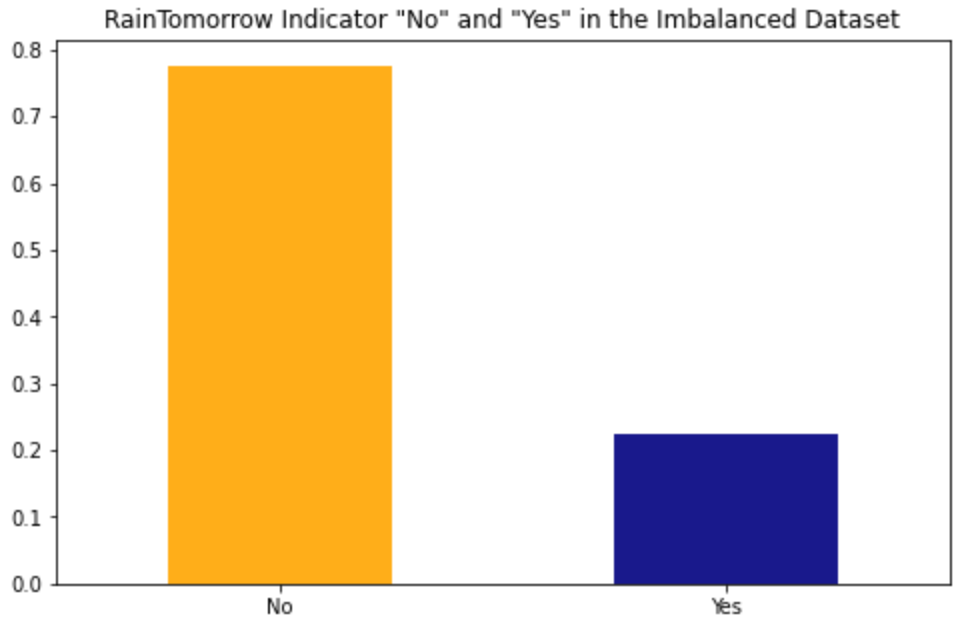


For finding the appropriate upper value we used Statistical Formula which was :

$$\text{Upper fence} = Q3 + (1.5 * \text{IQR})$$

Using this formula we procured the upper fence values for above mentioned variables. We redefined the maximum values as the upper fence.

Sampling The Dataset :



We sampled the data set as we found out that the target variable was imbalanced.

Dummy Variables :

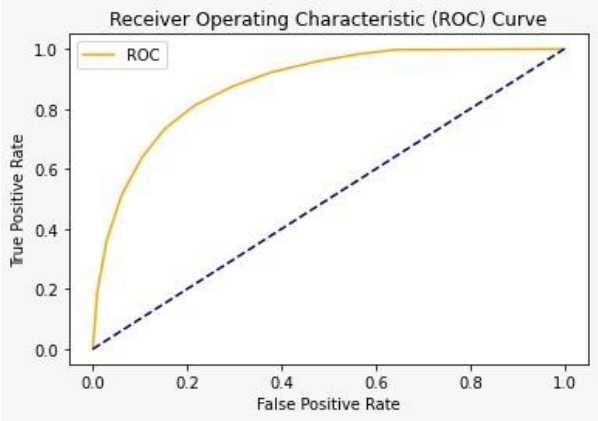
There were five categorical variables excluding target variables so we decided to make these dummy variable columns. We calculated the frequencies of these categorical variables and introduced them as unique columns in the dataset.

Scaling the Dataset :

Differences in the scales across input variables may increase the difficulty of the problem being modeled. An example of this is that large input values (e.g. a spread of hundreds or thousands of units) can result in a model that learns large weight values. A model with large weight values is often unstable, meaning that it may suffer from poor performance during learning and sensitivity to input values resulting in higher generalization error**.** So we scaled the dataset to make it more stable and machine friendly to use.

Finding The Perfect Model :

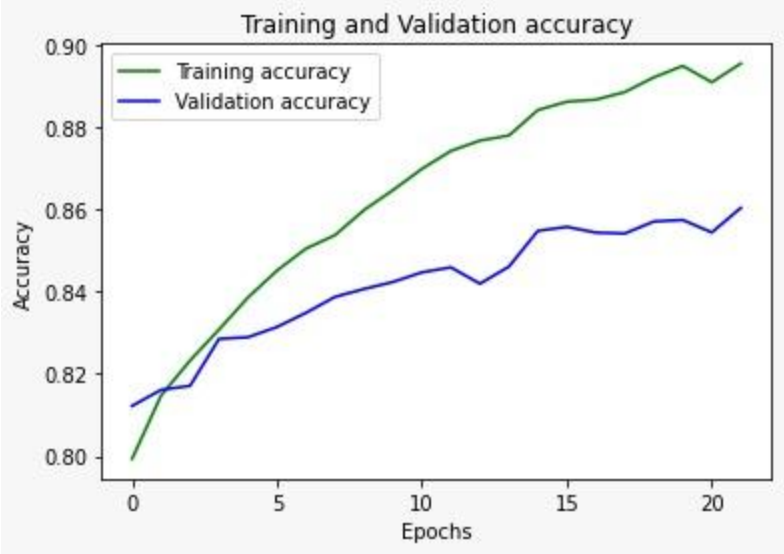
kNN Model :



Even though knn being the simplest model of them all, it consumed a huge amount of time as well as power. So we would not recommend you to use kNN for such a big dataset.

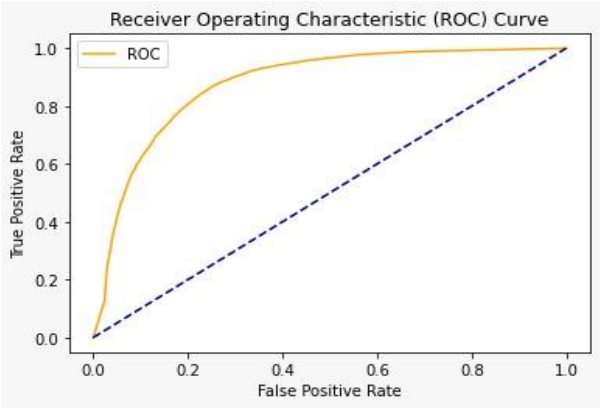
Accuracy on test set : 79.79%

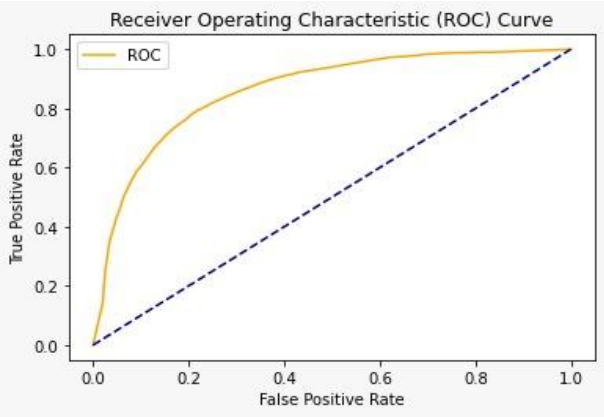
DNN Model :



DNN being one of the most sophisticated of them all, it consumed a lot of computing power even though we executed it on google colab. When the epoch was set for 110 and batch size was 5, we observed that after the 36th epoch the validation accuracy was going below 1% therefore we discontinued the execution. Following which we tuned the hyperparameter of the epoch and set it to 22 and increased the batch size to 7 which gave us the accuracy 77.55% on the test set.

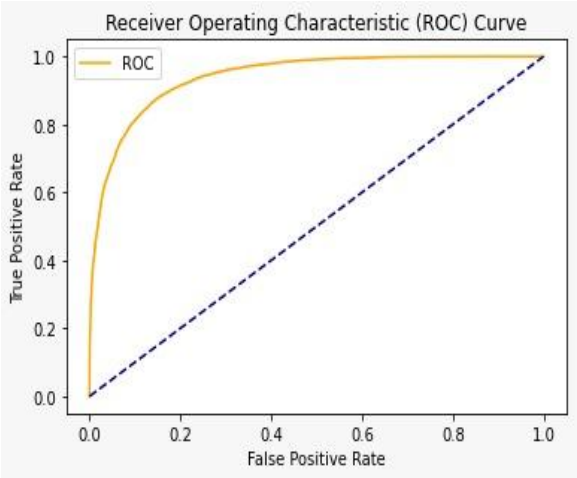
Decision Tree :

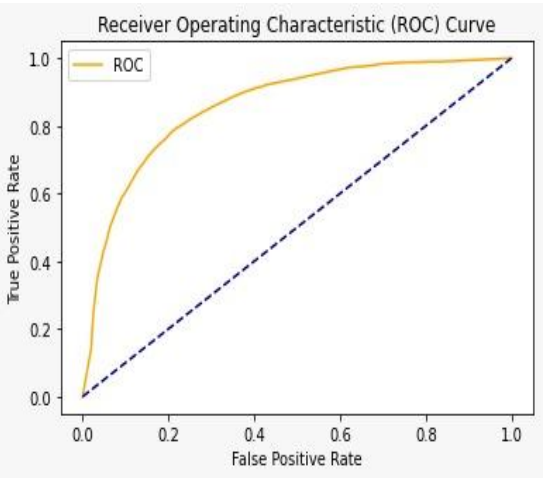




We observed that the decision tree was second optimal after random forest with the parameters max_depth = 16 and max_features = sqrt. We achieved an accuracy of 78.67% & 80.63% on two differently imputed test sets.

Random Forest





Random forest was the most robust and quickest of them all with the highest accuracy rate of 85.36% and 86.3% on two differently imputed test sets. Using this we were also able to derive important features looking at which we dropped the least important of them and achieved an accuracy of 89.12% with the model which was trained just on most important features.

Future Scope

After using above mentioned models we constructed that the humidity, sunshine & cloud covered at 3pm plays an important role for predicting the rainfall in Australia over different regions depending on wind gust speed, evaporation and min/max temperature highly correlating with wind direction.

References :

1) Article title: Failing Rains and Thirsty Cities: Australia's Growing Water Problem

Website title: Environment

URL: https://www.nationalgeographic.com/environment/article/partner-content-australia-water-problem

2)

Article title: Extensive Analysis - EDA + FE + Modelling

Website title: Kaggle.com

URL: https://www.kaggle.com/prashant111/extensive-analysis-eda-fe-modelling/notebook

3) Dataset :

Article title: Daily Weather Observations

Website title:	Bom.gov.au
URL:	http://www.bom.gov.au/climate/dwo/

Thank you!