RMIT International University Vietnam

Assignment Cover Page

Subject Code:	**COSC2789**
Subject Name:	**Practical Data Science**
Location:	**SGS**
Title of Assignment:	**Assessment 3: Report PDF**
Student	**Nguyen Minh Duong - s3741280** **Tran Viet Anh - s3795683** **Pham Nguyen Vu- s3701522** **Le Duc Nguyen - s3753240**
Teachers Name:	**Vo Ngoc Yen Nhi**
Assignment due date:	**21/01/2021**
Date of Submission:	**21/01/2021**
Number of pages including this one:	**15**
Word Count:	**2626**

We declare that in submitting all work for this assessment we have read, understood and agreed to the content and expectations of the Assessment declaration.

Table of Contents

I. Executive summary

In this assignment, our team will choose the dataset containing about 10 years of daily weather observations from many locations across Australia. The data was collected from numerous weather stations throughout the countries. Our task is to build classification models to predict the possibility of rain on the next day. The target variable would be the column RainTomorrow. Besides, we have more than 20 columns in the dataset that are relevant to the target variable such as Humidity, Sunshine, Wind direction, and more. After training and saving the model that has the best accuracy, which is XGBoost algorithms, we will deploy it to be used for inference. Finally, we will build a visualization dashboard via the Dash framework. It would include 3 interactive plots

II. Introduction

From the very beginning of humankind, forecasting weather has always been a very challenging job. In a climate that is rapidly changing because of humans, a system capable of accurately predicting and delivering early warnings has the potential to save millions of lives. The most modern machines today can only predict with relative accuracy for 1 week to 1 month. But with the rapid development of AI, scientists have been able to use the latest in computer technology to make this process easier and more precise. Our project, also called “Rain in Australia” is an example of using machine learning technology to predict the weather in Australia. By entering as many weather details into the system, including some past climate predictions, our team hopes the AI can forecast future weather.

III. Project Management

1. Project Goals

This dataset includes approximately 10 years of daily weather reports from a variety of locations around Australia. By training classification models on the target variable “RainTomorrow”, we want to use this “Rain in Australia” dataset to make predictions about whether or not it's going to rain in Australia the next day.

2. Deliverables

Generally, here are what we have delivered from this project:

Data Retrieving and Cleaning: After analyzing the raw data set, we have to fill in the missing value since some columns are missing a lot of values. After cleaning data, we split it to train dataset and test dataset

Data Exploration: We have created 6 charts in total. There are three charts describing three different columns and the other showing the relationships between 3 pairs of attribute

Data Modeling: We have chosen 5 different models to build. We include all necessary steps to build models successfully such as train/split the test, feature engineering & feature selection, model training, model evaluation, and model deployment. Finally, we use different methods (ROC -AUC, taken time, accuracy, Cohen's Kappa) to identify which model has the best performance among others.

Visualization dash: Dash is a repository of user interfaces to create analytical web applications. Immediate usage for Dash can be seen by those who use Python for data processing, data discovery, simulation, modeling, instrument management, and reporting. We use Dash to Visualize our dataset. Three interactive visualizations would be provided by using DashApp. The first two plots will be created in order to explore the dataset before training the model. The last visualization will be a plot to illustrate the result of the Model as well as compare some crucial features such as Accuracy, ROC, Cohen's Kappa before determining the best model would be used.

3. Program Stack

To complete a project like this, we utilize some of the most famous libraries and data frames in this industry. Most of these advanced technologies were introduced during the lecture sessions.

Python (version 3.7):

Python is a powerful, advanced, object-oriented programming language. It's easy to learn and is emerging as one of the best introductory programming languages for first-time programming languages (Python, n.d).

Pandas:

Pandas is a high-performance, open-source Python data analysis library (Espresso, n.d.). With special features of pandas such as being able to handle different data sets in formats: time series, heterogeneous tables, data matrices, or the ability to import data from different sources such as CSV, DB / SQL, Pandas has become an integral part of data analysis

Scikit-learn:

When it comes to a library of machine learning there is no more suitable library than Scikit-learn. This library integrates a lot of classical and modern algorithms, helping users to solve problems. Scikit-learn provides a set of tools for processing machine learning and statistical modeling problems including classification, regression, clustering (Codecademy, n.d.).

XGBoost classifier:

XGboost stands for extreme gradient boosting created by PhD Tianqi Chen and many other world-class developers as a machine learning library. This library appears as a branch of machine learning tools that were created by the Distributed Machine Learning Community. XGBoost is on the rise of becoming the go-to algorithm for data scientists proved by a significant number of winning Kaggle competitions trained with this algorithm. (Brownlee, n.d.)

Catboost:

Catboost is a gradient boosting algorithm on decision trees. This algorithm is developed by the research and development team from Yandex is an open-source used recommendation system, self-driving car, and many other high tech applications. Catboost offers fast, scalable, and great quality without parameter tunings and sacrifice system runtime. (Yandex, n.d.)

Logistic regression:

Logistic regression is a regression analysis when it comes to classification problems (dependent variable is binary). In order to interpret data and to illustrate the relationship between one dependent binary variable and one or more independent nominal, ordinal, interval, or ratio-level variables, logistic regression is used. The advantage of logistic regression is this algorithm is easy to use and takes little computing resources. (Statistics Solutions, n.d.)

Timeline

We all understand that this project is a sequential project, meaning that one person has to finish the assigned work before another can finish the assigned work. Therefore, we always try to get assigned work done as quickly as possible to avoid causing delays. The table below demonstrates our team's timetable:

	Week 9	**Week 10**	**Week 11**	**Week 12**
Choosing the project topic				
Data Retrieving and Cleaning				
Data Exploration				
Data Modelling				
Model deployment and Automation				
Visualization Dashboard				

Member Contribution

Name	Work contribution (%)	Work contribution in words
Tran Viet Anh	25%	● Handled model deployment and automation ● Writing report ● Prepare presentation ● Proofreading the report
Nguyen Minh Duong	25%	● Doing visualization dashboard ● Doublechecked all the coding parts ● Writing report ● Prepare presentation
Pham Nguyen Vu	25%	● Cleaning and retrieving data ● Writing report ● Facilitated meeting ● Prepare presentation
Le Duc Nguyen	25%	● The handled data modeling part ● Writing report ● Facilitated meeting ● Prepare presentation

IV. Methodology

1. **Confusion** Matrix:

[1] We have taken advantage of the Confusion matrix as a tool for describing the performance of all classification algorithms on the data set. It will show us a clear picture of classification model performance and the types of errors produced by the model. This matrix gives us both the right and wrong predictions based on each category. It was arranged in a tabular form.

While evaluating the classification model performance, there are four types of outcomes. Here is the description of these four outcomes:

- True positives (TP): we predict the value as **Positive** and it turns out as **True**

- True negatives (TN): we predict the value as **Negative** and it turns out as **True**
- False positives (FP): we predict the value as **Positive** and it turns out as **False.** It is also known as Type I Error.
- False negatives (FN): we predict the value as **Negative** and it turns out as **False.** It is also known as Type II Error.

The **classification accuracy** will be calculated by the formula:

$$\text{Classification accuracy} = \frac{TP + TN}{(TP + TN + FP + FN)}$$

While the **classification error** will be calculated by the formula:

$$\text{Classification error} = \frac{FP + FN}{(TP + TN + FP + FN)}$$

Precision is regarded as the proportion of positive results correctly predicted out of all positive results predicted.

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

On the other hand, **Recall** can be defined as the proportion of positive outcomes that are correctly predicted from the actual positive outcomes.

$$\text{Recall} = \frac{TP}{(TP + FN)}$$

Besides, there is the formula of the Specificity, True Positive Rate, and False Positive Rate

$$\text{Specificity} = \frac{FP}{(FP + TN)}$$

$$\text{True Positive Rate} = \frac{TP}{(TP + FN)}$$

$$\text{False Positive Rate} = \frac{FP}{(FP + TN)}$$

f1-score is the weighted harmonic mean of both precision and recall. The limitation of the f1-score would be from 0.0 to 1.0, while 0.0 is the worst score and 1.0 would be the best possible score.

Lastly**, Support** is the actual number of class occurrences in the dataset.

2. ROC - AUC

ROC - AUC stands for Operating Characteristic Receiver - Under Curve Field. This is a method for comparing the efficiency of classifiers. A perfect algorithm for classification would have a ROC - AUC equal to 1.

3. Cohen's kappa

Cohen's Kappa is a calculation of the agreement between two raters who decide which group a finite number of subjects belong to whereby agreement attributable to chance is factored out. The two raters either agree with their ranking (i.e., the grade that a subject is assigned to) or they disagree; there are no degrees of disagreement (i.e., no weightings).

4. Label encoding

We use the label encoding method to change all nominal/ ordinal data into numerical value for the model building process.

5. Binary encoding

We use the binary encoding for the variables that do not concern much about their own value, which are RainToday and RainTomorrow. In this case, 0 is interpreted as No and 1 as Yes.

V. Results

1. Data retrieving and cleaning

We have built some functions to handle errors in the dataset such as removing extra whitespaces, changing attributes name by adding the prefix, handling missing values. By using a heatmap, we have discovered some columns that have a high percentage of missing values. Hence, we will drop the four columns namely "Sunshine", "Evaporation", "Cloud3pm", "Cloud9am. "Date" is also excluded for the obvious reason since it is not adding any relevance in the current context.

2. Data exploration

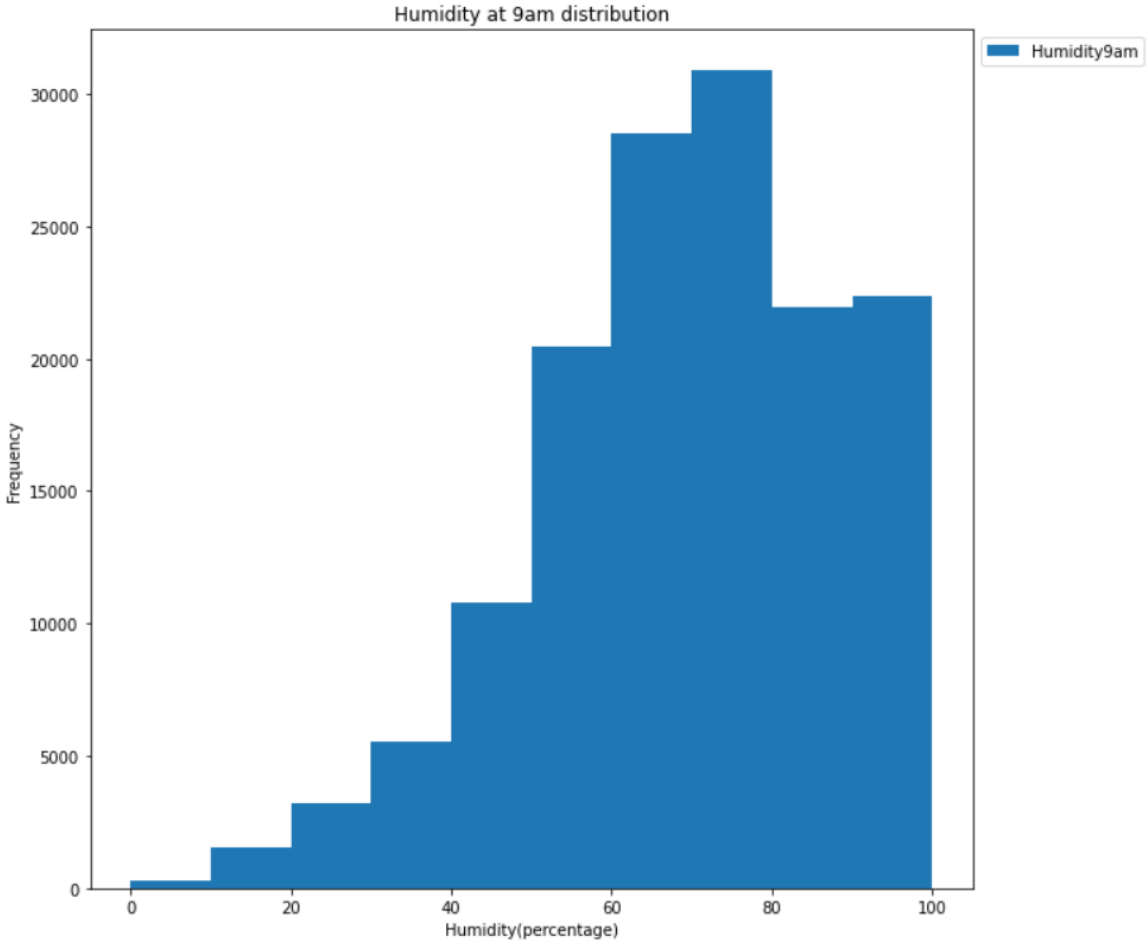


Figure 1 : The demonstration of the frequency of Australian humidity in 10 years

It can be shown, in general, that the Australian humidity has an erratic range at 9 am. It can be said that the average humidity here is from 70 to 80 percent, with the largest amount being the sum recorded about 30000 times. Besides that, the second prevalence of 60 percent and between 80 percent and 100 percent humidity is often seen from a calculation of over 20,000 days. With humidity levels in Australia between 0 and 30% less normal when measured at the lowest frequency below 5000 times

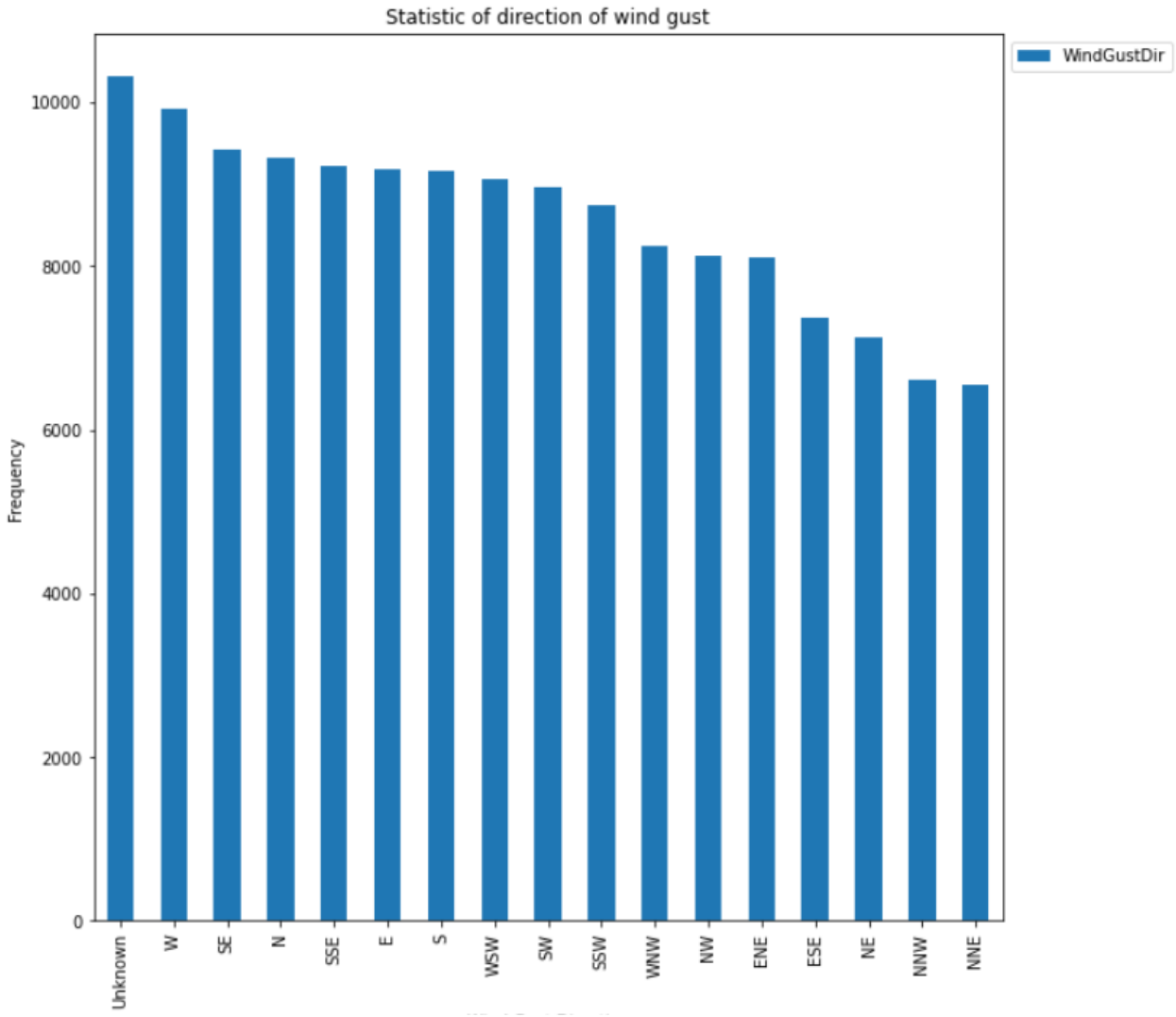
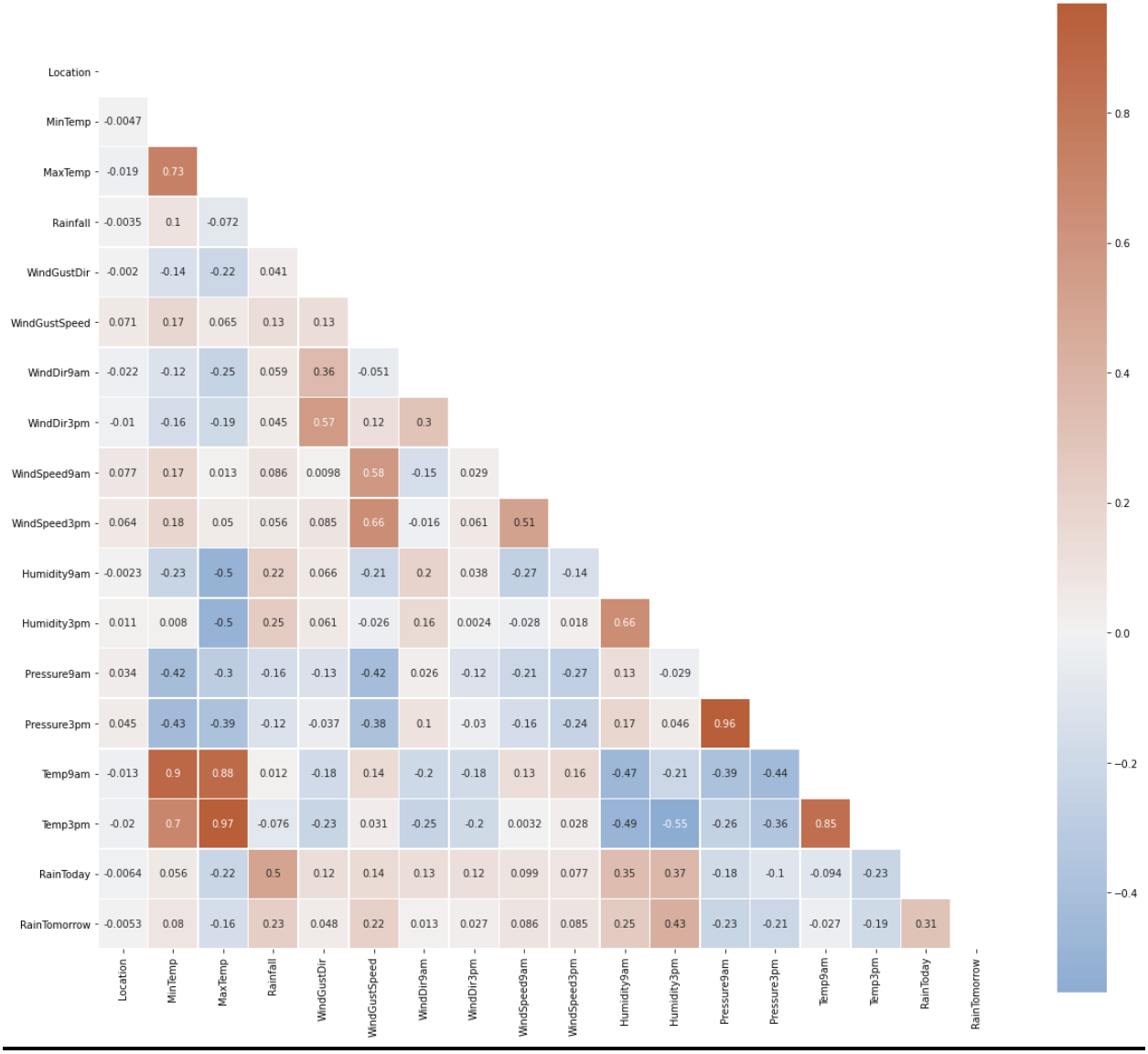


Figure 2: The frequency of direction of wind gust in Australia in 10 years.

In general, the frequency of the occurrence of wind directions is quite similar, ranging from 8000 to 10000 times. The most appearing wind direction is the West with nearly 10,000 times. In contrast, north-northwest and north-northeast only have approximately 7,000 recorded cases.

3. Model training

Here is the correlation analysis between all variables in the dataset:



In conclusion, the following pairs of features are having a high correlation between them:

- **MaxTemp** and **MinTemp**
- **Pressure9am** and **Pressure3pm**
- **Temp9am** and **Temp3pm**
- **MaxTemp** and **Temp3pm**

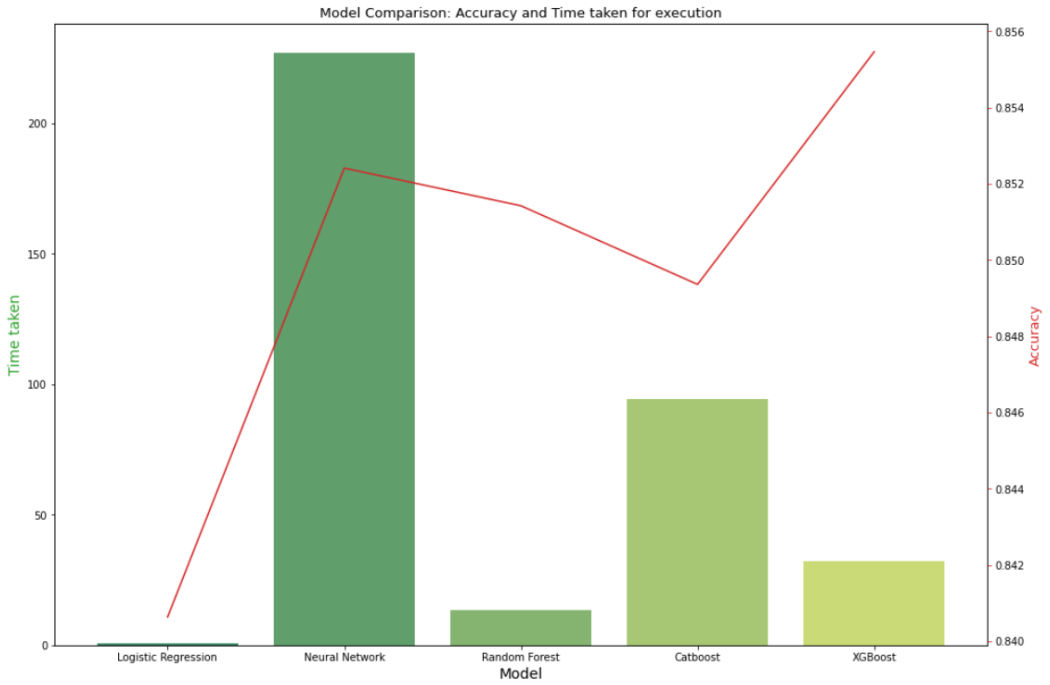
But there is no case that the correlation value is equal to a perfect "1". As a result, we will not discard any feature.

After extracting all features from the dataset, we have decided to choose 17 columns for model training namely Location, MinTemp, MaxTemp, Rainfall, WindGustDir, WindGustSpeed, WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Temp9am, Temp3pm, and RainToday.

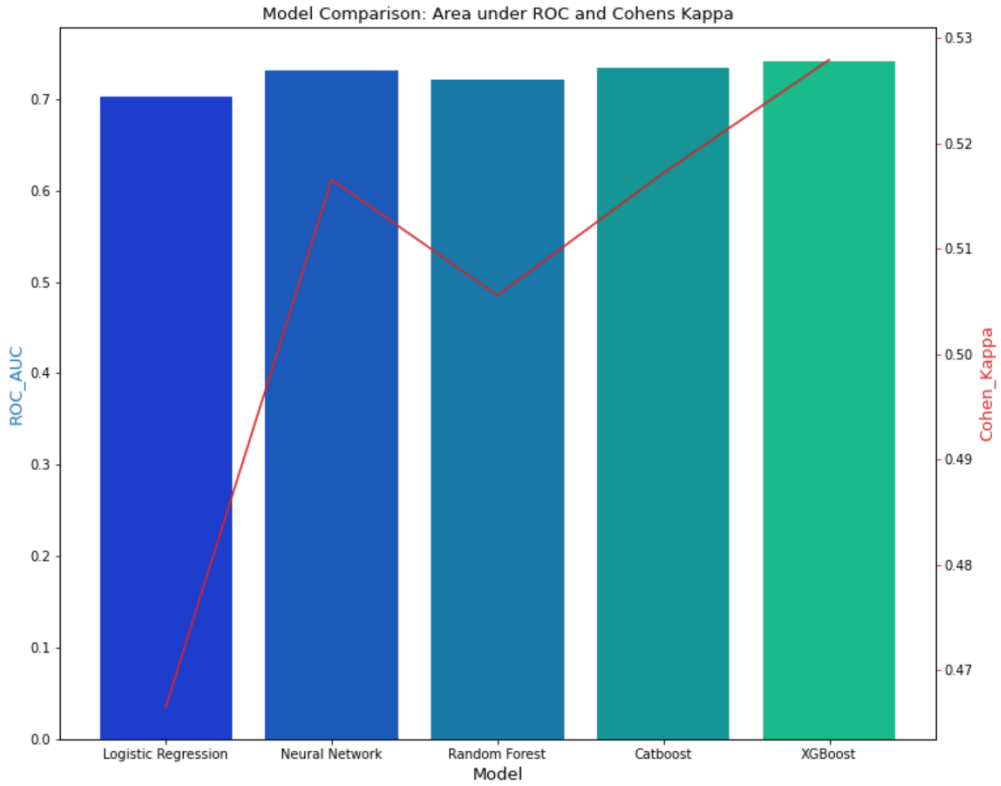
Here is the result of all 5 models:

Model	**Accuracy**	**ROC - AUC**	**Cohen's Kappa**	**Time Taken**
Model 1 - Logistic Regression	0.8406	0.7016	0.4629	0.7379
Model 2 - Neural Network	0.8524	0.7296	0.5147	227.006
Model 3 - Random Forest	0.8514	0.7180	0.5000	13.3136
Model 4 - Catboost	0.8493	0.7310	0.5110	94.5475
Model 5 - XGBoost	0.8554	0.7446	0.5356	32.2469

After building 5 models namely Logistic Regression, Neural Network, Random Forest, Catboost, and XGboost, we need to decide which model has performed the best based on the accuracy score, ROC - AUC, totally taken time, and Cohen's Kappa for execution. We will visualize all methods for better interpretation like below:



XGBoost is considered as the best model in terms of high accuracy, as can be seen from the graph, whereas the Logistic Regression achieves the lowest accuracy. Neural Network is the model that takes the most time in terms of time taken, until more than 250s when Random Forest is the model with the fastest output among other training models.



It can be shown that, as compared to other versions, XGBoost has the highest results in terms of Cohen's Kappa and ROC - AUC research.

4. Discussion

Overall, as it takes the lead in three-quarters of the testing system (Cohen's Kappa, accuracy, and ROC - AUC), we would like to pick XGBoost as the best model. In comparison, for most of the test methods, Logistic Regression has the worst results. However, we will stick to Logistic Regression instead of XGBoost if the pace is considered the best attribute.

Generally, what we had to do on this project was well covered at the lecture sessions. However, having difficulties in completing the project is inevitable. First of all, our deployment part encountered a lot of difficulties because we had too little time to adapt to the latest technologies. Secondly, scheduling is a bit difficult since one of our team members is working and has a lot of free time to arrange group meetings. Last and not least, choosing the right dataset is also a major hurdle given that there are so many different datasets on Kaggle and it's difficult to choose the right one. Fortunately, we are all interested in the dataset of Australian weather prediction based on classification training models

VI. Conclusion

With predictive results up to 85% accurate, weather prediction based on Big Data is completely practical. The possibility of rain according to data analysis relies on humidity, temperature, wind speed and wind direction. Hopefully in the near future, the new computer system will change the industry of weather prediction. If it is successful, this research promises to save countless lives from natural disasters such as earthquakes, floods and droughts.

VII. References

1. **Brownlee, J. (n.d.). *A Gentle Introduction to XGBoost for Applied Machine Learning*. Machine Learning Mastery. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/#:~:text=XGBoost%20is%20an%20algorithm%20that,designed%20for%20speed%20and%20performance**
2. **Python. (n.d). *What is Python? Executive Summary*. Python. Retrieved Jan 21, 2021, from https://www.python.org/doc/essays/blurb/**

3. **Statistics Solutions. (n.d.).** ***What is Logistic Regression?*** **Statistics Solutions.** https://www.statisticssolutions.com/what-is-logistic-regression/

4. **Yandex. (n.d.).** ***CatBoost is a high-performance open source library for gradient boosting on decision trees*****. Catboost.** https://catboost.ai/

5. **Young, J (2020).** ***Rain in Australia*****. Kaggle** https://www.kaggle.com/prashant111/extensive-analysis-eda-fe-modelling