Rain in Astralia

Team member: Linxin(Iris) Liu, Mingjun Wang

INFO 7390 Final Project

Our Goals:

We decide to use our knowledge in machine learning and the rain in Australia dataset to predict weather in the next day, and help people in Austrialia preparing their belongings for outside.

Dataset: Rain in Australia

Url: https://www.kaggle.com/jsphyg/weather-dataset-rattle-package

Our Focus

First part:

- **Explore Data**
- **Preprocessing**
- **Visualiazation**

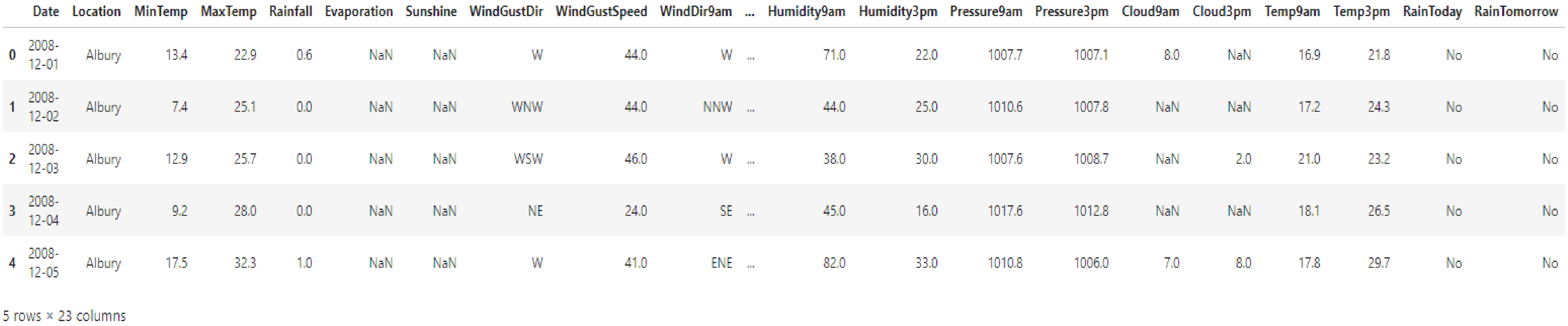
Second part:

- **Random Forest**
- **KNN**
- **DNN**

Third part:

- **Prediction**

Explore Data Analysis



	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	...	71.0	22.0	1007.7	1007.1	8.0	NaN	16.9	21.8	No	No
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	...	44.0	25.0	1010.6	1007.8	NaN	NaN	17.2	24.3	No	No
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	...	38.0	30.0	1007.6	1008.7	NaN	2.0	21.0	23.2	No	No
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	...	45.0	16.0	1017.6	1012.8	NaN	NaN	18.1	26.5	No	No
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	...	82.0	33.0	1010.8	1006.0	7.0	8.0	17.8	29.7	No	No

5 rows × 23 columns

By taking a look of our dataset, we notice that column 'RainTomorrow' determine the weather for the next day.Therefore, we shall use our models to predict it.

Also, we need to check the data type of each column: float64(16), Object(7)

After showing basic information of the dataset, we need to explore more details from the dataset:

Numble of raining/Not raining days in the next day

Rain: 110316 Not Rain: 31877

Number of cities in the Australia and days

Number of cities: 49 Number of days: 3436

The earliest date and the latest date

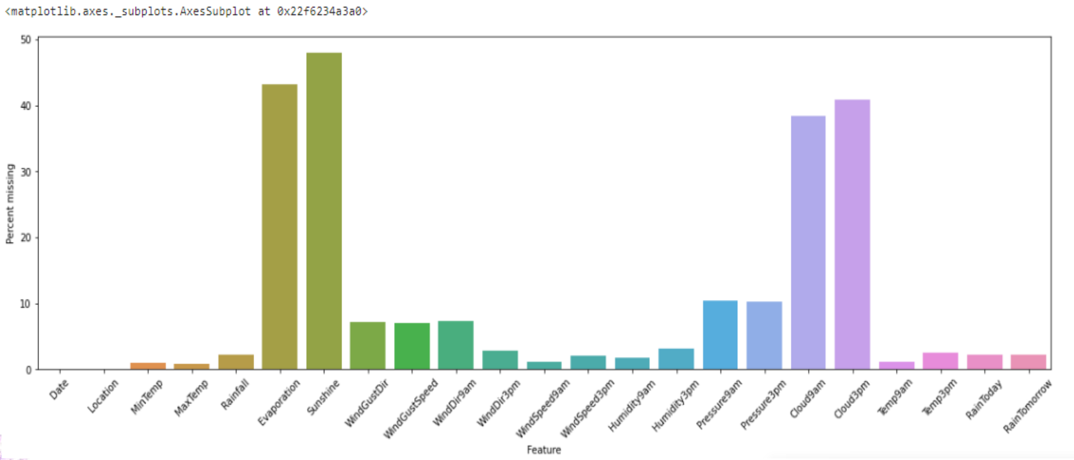
Earliest date: 2007-11-01 Latest date: 2017-06-25

Rain Tomorrow Vs Rain Today

RainTomorrow	RainToday
Rain: 110316	Rain: 110319
Not Rain: 31877	Not Rain: 31880

Data Pre-processing

Check Null Values



```
<matplotlib.axes._subplots.AxesSubplot at 0x22f6234a3a0>
```

Deal with missing values. Calculate the percentage of missing values for every column, and plot them as a bar chart. Then, drop the columns that contains too many null values, which is Evaporation, Sunshine, Cloud9am, Cloud3pm

Data Preprocessing

After Checking Null values, we define the impute functions to impute categorical NaNs with -1, where we add 1 to make it 0. For each continuous variables, we impute missing values with median values of that column, and for every variable where any rows were imputed, add a separate 'imputed or not' column.

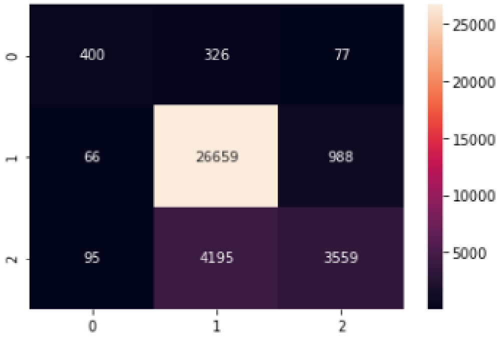
Also, we need to covert object types and string types to category type for next steps.

Then, we should fill all null values with mean

Build Different Models

Random Forest Models

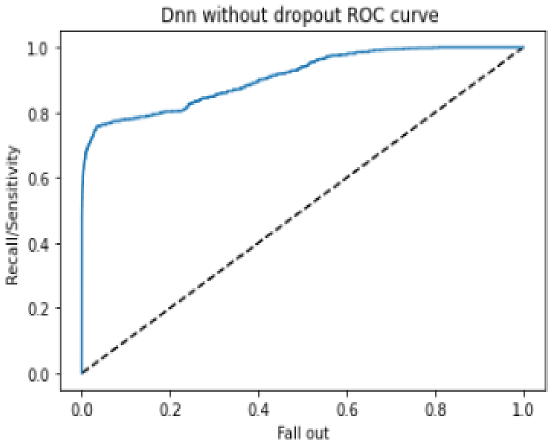
```
<matplotlib.axes._subplots.AxesSubplot at 0x22f670fe4f0>
```



Accuracy: 0.841963

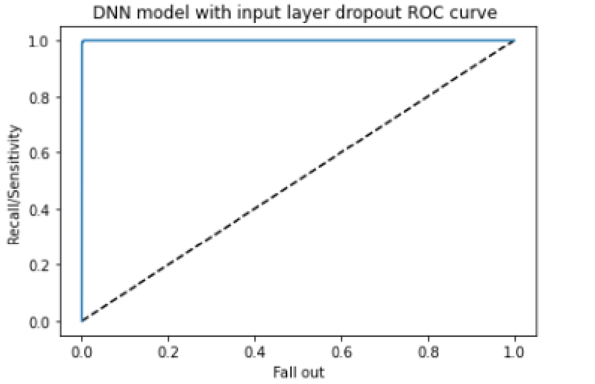
keep in mind that 0 means not raining, 1 means raining and 2 means others

DNN without dropout



Accuracy score: 0.975126
Area under ROC curves: 0.909813

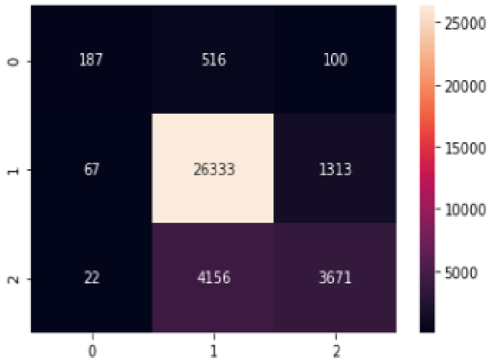
DNN with input layer dropout



Accuracy score: 0.997042
Area under ROC curves: 0.999925

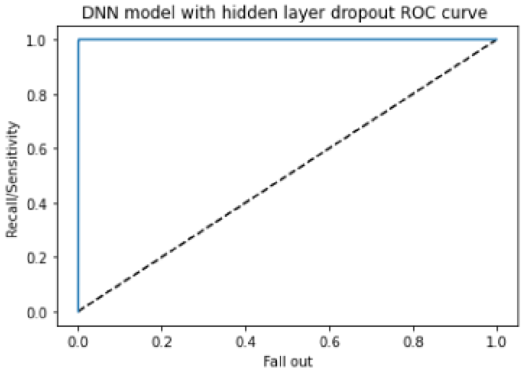
KNN Models

```
<matplotlib.axes._subplots.AxesSubplot at 0x22f670d3fd0>
```



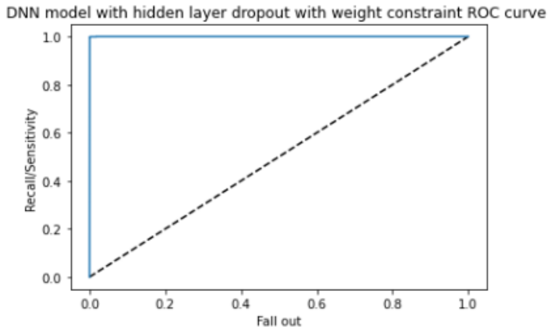
Accuracy: 0.841963

DNN with Hidden layer dropout



Accuracy score: 0.999646
Area under ROC curves: 0.999982

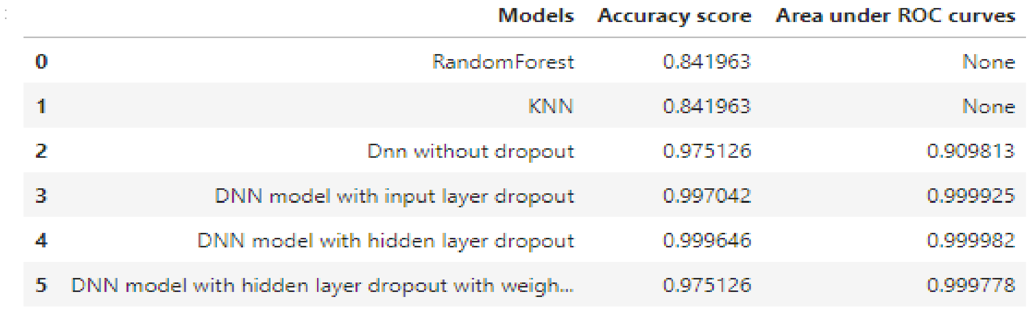
DNN with Hidden layer dropout with weight constraint



Accuracy score: 0.975126
Area under ROC curves: 0.999778

Conclusion

Compare different Models



	Models	Accuracy score	Area under ROC curves
0	RandomForest	0.841963	None
1	KNN	0.841963	None
2	Dnn without dropout	0.975126	0.909813
3	DNN model with input layer dropout	0.997042	0.999925
4	DNN model with hidden layer dropout	0.999646	0.999982
5	DNN model with hidden layer dropout with weigh...	0.975126	0.999778

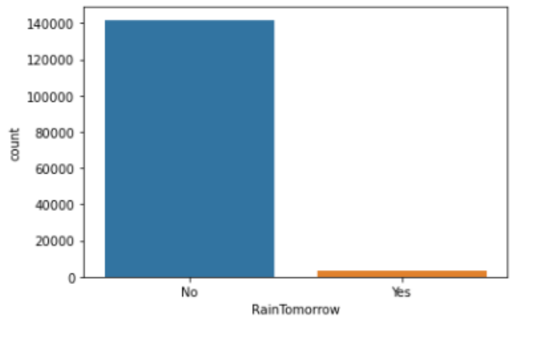
In conclusion,DNN model with hidden layer dropout gives the highest accuracy in prediction of weather tomorrow. Then, we shall use it for prediction of weather in the next day.

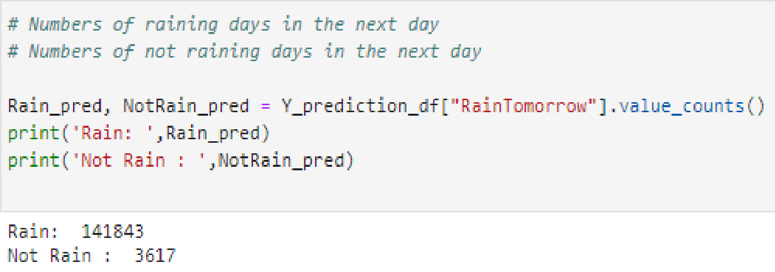
Prediction

After compare all models we decided to use DNN model with hidden layer dropout to make a prediction.

We cast prediction of next day weather to a DataFrame and replace all numbers in the prediction to ‘Yes’ and ‘No’

```
<matplotlib.axes._subplots.AxesSubplot at 0x22f63aaf1f0>
```





```
# Numbers of raining days in the next day
# Numbers of not raining days in the next day

Rain_pred, NotRain_pred = Y_prediction_df["RainTomorrow"].value_counts()
print('Rain:  ',Rain_pred)
print('Not Rain : ',NotRain_pred)
```

```
Rain:  141843
Not Rain :  3617
```

Thank you !