Rain in Australia

JULY 2119

By Chanan Sukenik

Final Project- Introduction to Machine Learning Using Python (COMPSCIX433.6)

Table of Contents

The Dataset

- Contains daily weather observations from numerous Australian weather stations ranging from 2007 to 2017
- Obtained from *kaggle* (last updated 8 months ago)
- Mostly used for binary classification (supervised learning)
- Consists of **Continuous** data (I.E. Rainfall), **Discrete** Data (I.E. Wind Direction/Temperature) & **Boolean** data (I.E. RainToday)
- The target variable "RainTomorrow" means: Did it rain the next day? Yes or No
- Observations: 142,193

  Features: 24
- Columns:

  Date: The date of observation ***(object)***

  Location: The common name of the location of the weather station ***(object)***

  MinTemp: The minimum temperature in degrees Celsius ***(float64)***

  MaxTemp: The maximum temperature in degrees Celsius ***(float64)***

  Rainfall: The amount of rainfall recorded for the day in mm ***(float64)***

  Evaporation: The so-called Class A pan evaporation (mm) in the 24 hours to 9am ***(float64)***

  Sunshine: The number of hours of bright sunshine in the day ***(float64)***

  WindGustDir: The direction of the strongest wind gust in the 24 hours to midnight ***(object)***

  WindGustSpeed: The speed (km/h) of the strongest wind gust in the 24 hours to midnight (***float64)***

  WindDir9am: Direction of the wind at 9am ***(object)***

  WindDir3pm: Direction of the wind at 3pm ***(object)***

  WindSpeed9am: Wind speed (km/hr) averaged over 10 minutes prior to 9am ***(float64)***

  WindSpeed3pm: Wind speed (km/hr) averaged over 10 minutes prior to 3pm ***(float64)***

  Humidity9am: Humidity (percent) at 9am ***(float64)***

  Humidity3pm: Humidity (percent) at 3pm ***(float64)***

  Pressure9am: Atmospheric pressure (hpa) reduced to mean sea level at 9am ***(float64)***

  Pressure3pm: Atmospheric pressure (hpa) reduced to mean sea level at 3pm ***(float64)***

Cloud9am: Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eigths. It records how many eigths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast. ***(float64)***

Cloud3pm: Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cload9am for a description of the values (***float64)***

Temp9am: Temperature (degrees C) at 9am (***float64)***

Temp3pm: Temperature (degrees C) at 3pm (***float64)***

RainToday: Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0 ***(object)***

RISK_MM: The amount of next day rain in mm. Used to create response variable RainTomorrow. A kind of measure of the "risk". ***(float64)***

RainTomorrow: The target variable. Did it rain tomorrow? ***(object)***

- **Missing Values Percentage:**

Date	0.00	WindSpeed3pm	0.02
Location	0.00	Humidity9am	0.01
MinTemp	0.00	Humidity3pm	0.03
MaxTemp	0.00	Pressure9am	0.10
Rainfall	0.01	Pressure3pm	0.10
Evaporation	0.43	Cloud9am	0.38
Sunshine	0.48	Cloud3pm	0.40
WindGustDir	0.07	Temp9am	0.01
WindGustSpeed	0.07	Temp3pm	0.02
WindDir9am	0.07	RainToday	0.01
WindDir3pm	0.03	RISK_MM	0.00
WindSpeed9am	0.01	RainTomorrow	0.00

Data Observations:

- Min. Temperature- -8.5°C (16.7°F)
- Max. Temperature- 48.1°C (118.58°F)
- Average Temperature (High)- 23.2°C (73.76°F)

- Average Humidity (3PM)- 51.5%
- Max. Daily Rainfall- 371mm
- Rainfall Standard Deviation- 8.465
- Max. pressure (9AM)- 1041

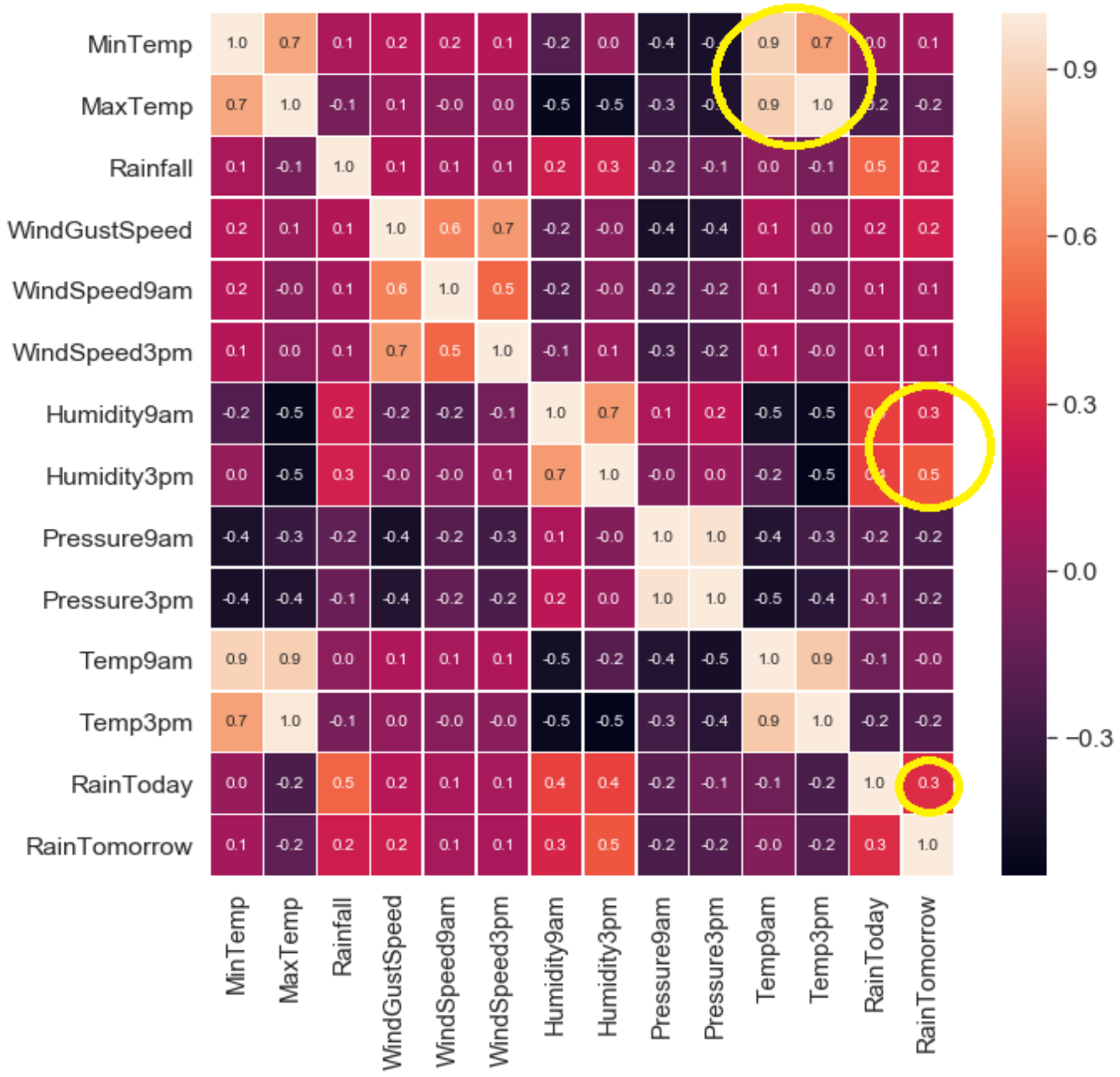
Questions/Goals

- **Main goal**- Predict whether or not it will rain tomorrow by training a binary classification model on target RainTomorrow
- Get statistics, explore data
- Which features/observations are usable, and which aren't?
- Find which attributes have strong correlation with each other and with the target class
- Find the best models for the generalization of the rain prediction model
- Evaluate these models

Data Manipulation

- **Columns Dropped:**
    1. Evaporation, Sunshine, Cloud9am & Cloud3pm- Due to unusual rates of missing values
    2. Location, Date- Due to irrelevancy
    3. Risk_MM- "Leeks" the answers to the model
- **All rows with missing values were dropped**- database is large enough for a model to work well without them (preferred over incorrectly filled data)
- New shape of data after current manipulation: **(112925, 17)**
- RainToday & RainTomorrow's values were converted to 1s and 0s

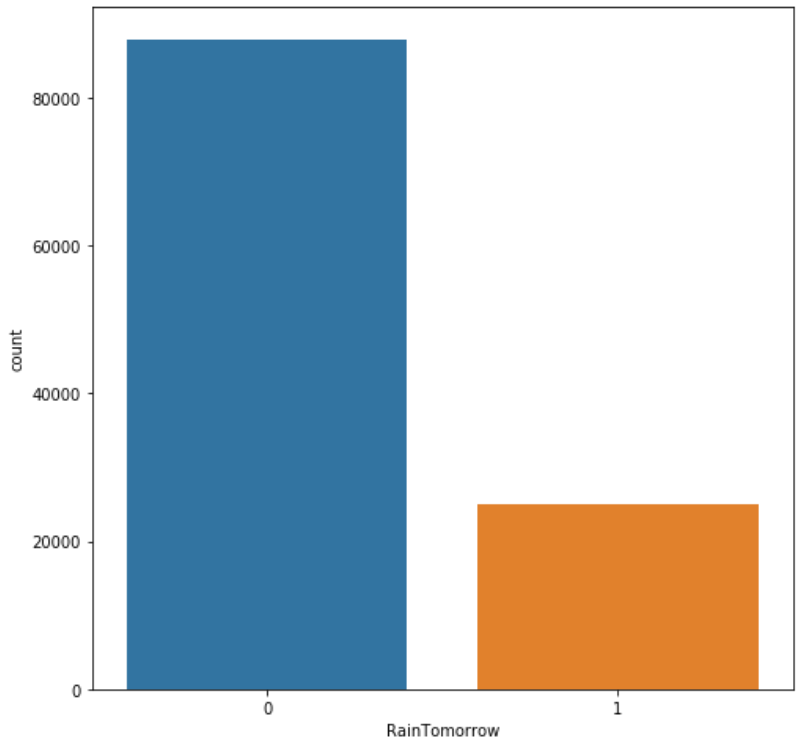
- Dummy variables were assigned to the categorical features: WindGustDir, WindDir9am & WindDir3pm
- Database was scaled using sklearn's preprocessing kit- necessary as many of the values were discrete and/or unscaled
- Correlation Heatmap for the different features:



- There is very high positive correlation between the attributes MinTemp & MaxTemp and Temp9am & Temp3pm. Therefore, Temp9am & Temp3pm were dropped as well[1].
- Humidity9am, Humidity3pm & RainToday have relatively high correlation with RainTomorrow- those will be some our stronger predictors in the model.

Modeling

- Target variable "RainTomorrow" was checked for class imbalance- observations are indeed imbalanced, with 77.8% to "No" (0) and 22.2% to "Yes" (1)



[1] **Results without dropping the features (overall less accuracy):**

Logistic Regression- 0.85 accuracy
Naive Bayes- 0.82 accuracy
KNN Classifier- 0.80 accuracy
Random Forest- 0.85 accuracy
Decision Tree- 0.77 accuracy

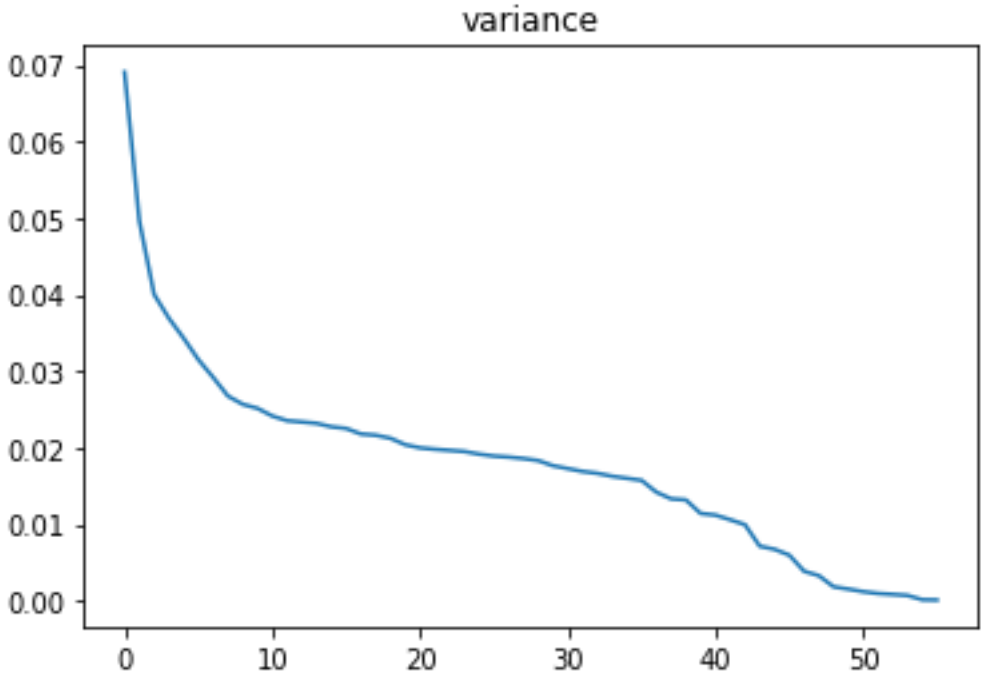
- Due to the high number of observations (112,925 after cleaning and manipulation), I assumed the models will work well despite the imbalance.

  Stratifying the data did not make the models perform better, overall[2].

- Data was split into x (target class "RainTomorrow") and y (features) and then into train and test sets (80% training size, 20% testing size)

- **Principal Component Analysis (PCA)**

  When testing the accuracy after applying PCA to the Random Forest model, the accuracy of the model decreased from ~0.85 to ~0.77. It seems like there are no distinct features that are responsible for a major part of the variance in the dataset.

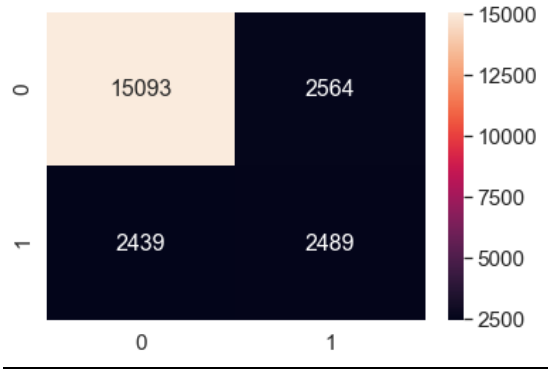


[2] **Results with stratifying:**

Logistic Regression- 0.85 accuracy
Naive Bayes- 0.73 accuracy
KNN Classifier- 0.81 accuracy
Random Forest- 0.85 accuracy
Decision Tree- 0.795 accuracy

Testing Difference ML Algorithms

Testing Set Confusion Matrix:



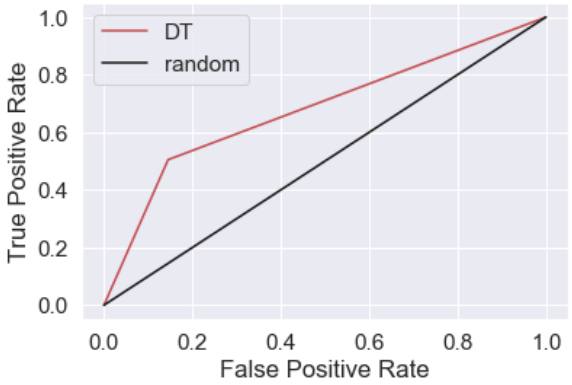
Training Set Accuracy Score: 1.0

Testing Set Accuracy Score: 0.77

Precision Score: 0.78

Recall Score: 0.78

ROC curve[3]:

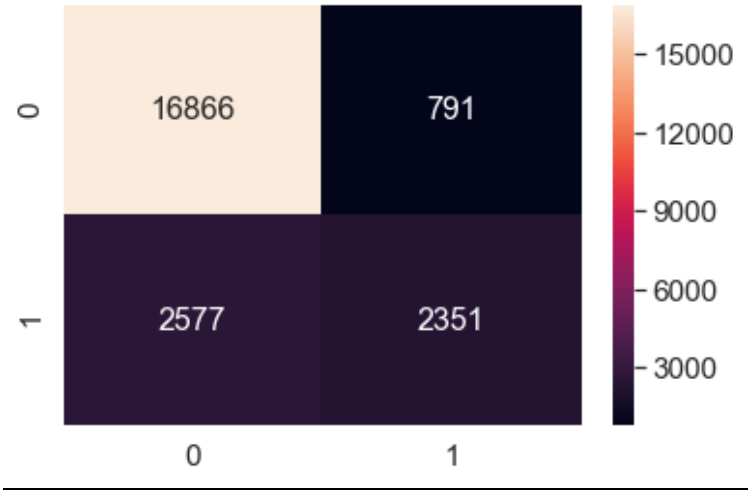


Thresholds: [2. 1. 0.]
False Positive Rates: [0. 0.15 1.]
True Positive Rates: [0. 0.51 1.]

[3] To help with understanding the balance between true-positive rate and false-positive rates

Random  Forest

Testing Set Confusion Matrix:



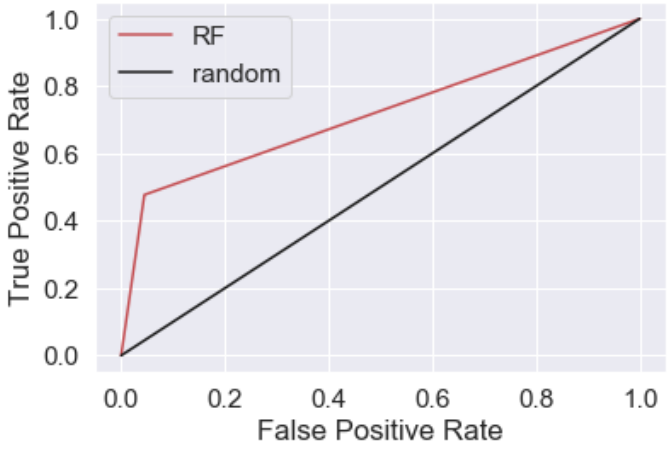
Training Set Accuracy Score: 1.0

Testing Set Accuracy Score: 0.85

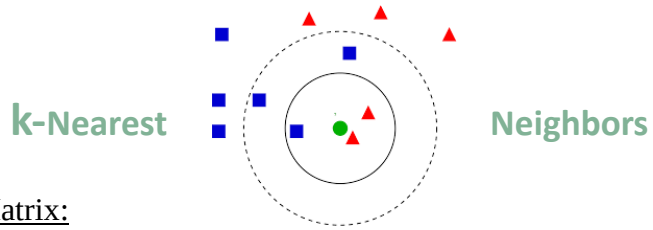
Precision Score: 0.84

Recall Score: 0.85

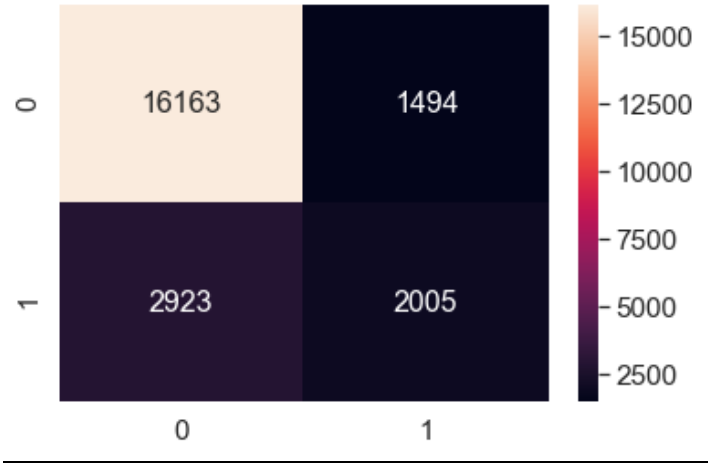
ROC curve:



Thresholds: [2. 1. 0.]
False Positive Rates: [0. 0.04 1.]
True Positive Rates: [0. 0.48 1.]



Testing Set Confusion Matrix:



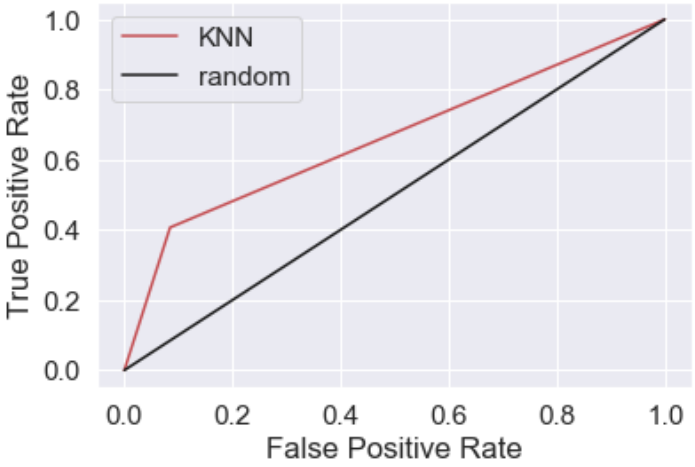
Training Set Accuracy Score: 0.89

Testing Set Accuracy Score: 0.80

Precision Score: 0.79

Recall Score: 0.80

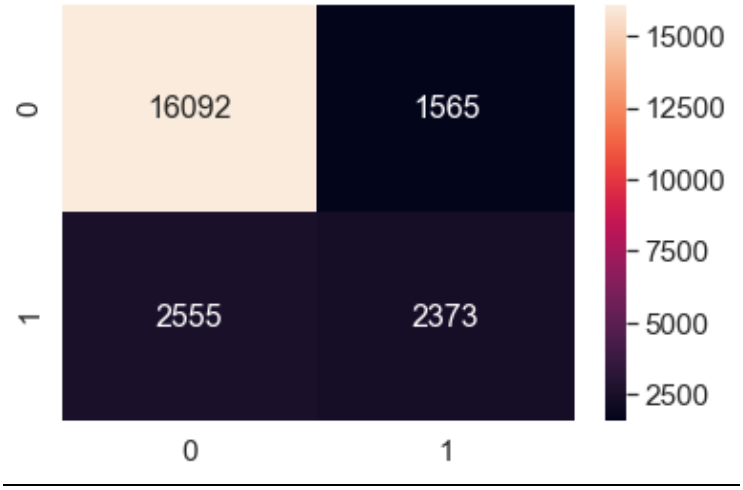
ROC curve:



Thresholds: [2. 1. 0.]
False Positive Rates: [0. 0.08 1.]
True Positive Rates: [0. 0.41 1.]

Naive $P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$ **Bayes**

Testing Set Confusion Matrix:



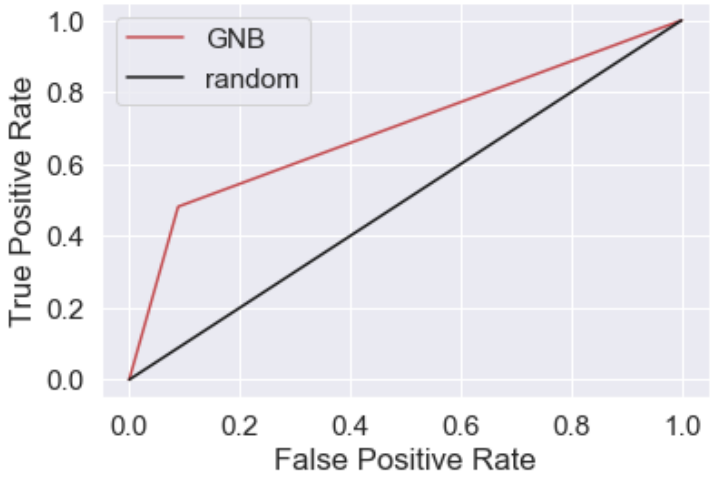
Training Set Accuracy Score: 0.82

Testing Set Accuracy Score: 0.82

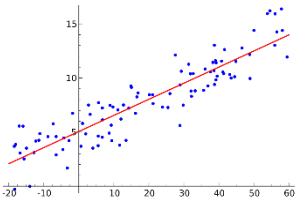
Precision Score: 0.81

Recall Score: 0.82

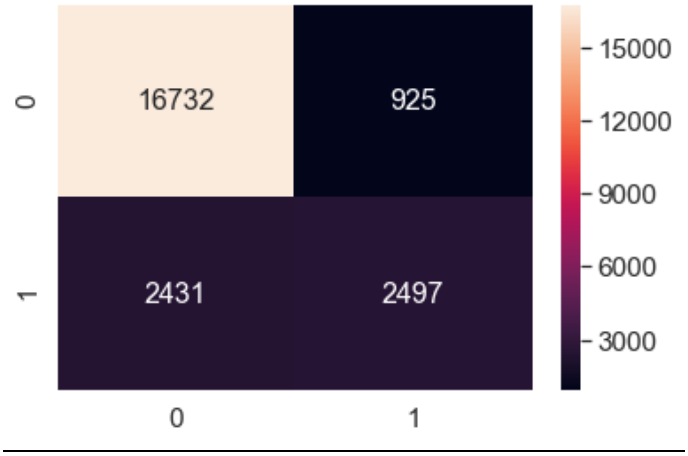
ROC curve:



Thresholds: [2. 1. 0.]
False Positive Rates: [0. 0.09 1.]
True Positive Rates: [0. 0.48 1.]

Logistic  Regression

Testing Set Confusion Matrix:



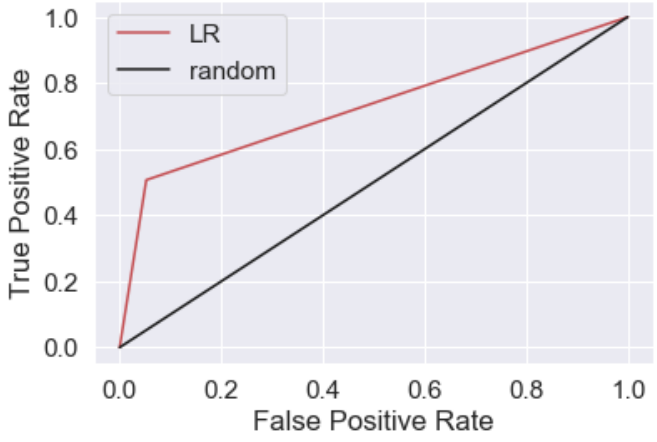
Training Set Accuracy Score: 0.85

Testing Set Accuracy Score: 0.85

Precision Score: 0.84

Recall Score: 0.85

ROC curve:



Thresholds: [2. 1. 0.]
False Positive Rates: [0. 0.05 1.]
True Positive Rates: [0. 0.51 1.]

Conclusions

- Weather prediction is hard to achieve with high accuracy, at least based on the relevant dataset- maximum accuracy achieved was 85%. When looking at different weather prediction models on the web, similar or lesser accuracies were found.

- Best performing ML algorithms for our case- Random Forest & Logistic Regression. Both achieved an accuracy score of 0.85, precision of 0.84 and recall of 0.85 on the testing set.

- None of the attributes had an especially significant contribution to the variance. Some of the stronger predictors were Humidity & RainToday.

- Decision Tree & Random Forest- seem to be overfitting due to the large size of the training set and amount of features.

  k-Nearest Neighbors- Underfitting, seems to be biased.

- Recall (TP / (TP+FN)) was higher than Precision (TP / (TP+FP)) for most models, which is a good result- it is usually better to expect rain and be surprised than not to expect rain and be surprised.

Source & Acknowledgements

- Dataset was obtained from https://www.kaggle.com/jsphyg/weather-dataset-rattle-package

- Observations were drawn from numerous weather stations. The daily observations are available from http://www.bom.gov.au/climate/data. Copyright Commonwealth of Australia 2010, Bureau of Meteorology.
- Definitions adapted from http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml

- This dataset is also available via the R package rattle.data and at https://rattle.togaware.com/weatherAUS.csv. Package home page: http://rattle.togaware.com. Data source: http://www.bom.gov.au/climate/dwo/ and http://www.bom.gov.au/climate/data.