STA6714-19Spring 0001

Final project report

Predict Rain tomorrow in Australia

Mithun Mohanraj

Abstract

The work presented in this report features my final term project upon predicting the rain in Australia by performing exploratory data analysis, binary classifications using different statistical models and comparing their performances based on various evaluation metrics.

1.Introduction

Data preparation has become an inevitable part of data processing and data mining. [1]It is the process of cleaning and transforming unprocessed data into a form that is easier for extracting information. It involves standardizing, detecting outliers, handling missing values, selecting important features, dimensionality reduction, discovering patterns, finding correlated features and visualizing raw data. It is basically done to enrich the quality of data which helps in improving the performance of data mining methods.

In this project, Data preparation plays a major role for predicting rain in Australia. The main aim of this project is to explain and prove that variable selection and dimensionality reduction are the important techniques to be considered while preparing the data for further analysis.

1.1 Project Significance

The impact of uncertainty in weather and climate change has bothered the lives of people in recent years. It is important to predict any abnormalities in weather before in hand to avoid natural disasters. The significance of this project is to help understand the conditions which lead to rain in a given place. It is found that some of the weather factors like sunshine, evaporation, cloud count, wind directions are vital for predicting the rain in future.

This project proposes various methods and models that aim to predict the rain in advance given the weather data for today.

1.2 Project Application

Rain predictions are applied on weather forecasting, preventing floods, Understanding current rainfall. [2]It also helps to model future behavior of precipitation patterns and climate. It can also be applied to give live weather forecasting to alarm people about sudden change in climate.

1.3 Project Objective

The main objective of this project is to explain and prove that data preparation techniques like variable selection, dimensionality reduction are vital before modelling the data using various statistical models. We also compare different models based on their performance using various evaluation metrics on predicting the rain.

2. Data Exploration

Data exploration helps us understand the data very well. It is discovering and uncovering the underlying patterns in the raw data. we basically find the statistics of data like number of observations, number of observed variables, relationship between observations, relationship between variables and characteristics. We also explore the data through various data visualization tools like scatter plots, histograms, box-plots and pie-charts to define the problem of interest.

2.1 Data source and description

Rain data is obtained from **Kaggle- Rain in Australia**. It contains **142193** observations of **24** attributes from numerous Australian weather stations obtained during the years 2007-2017. Since it involves information from various weather stations, there are lot of missing values in the data.

2.2 Target problem

The problem definition is to predict whether it will rain tomorrow or not by training a binary classification model on target variable **Rain-Tomorrow.** The target variable Rain-Tomorrow means: did it rain the next day? Yes or No.

Table 1: Features and their types

Features	Type
Date	Date
MinTemp	Num
Rainfall	Num
Evaporation	Num
WindGustDir	Factor
Humidity9am	Num
RainToday	Factor
Pressure3pm	Num
WindGustSpeed	Num

.

The above table shows few features of the data along with their data type.

2.3 Descriptive Statistics

[3]Descriptive statistics of data is quantitatively summarizing the individual features present in the data. It is finding statistical measures like mean, median, standard deviation, Maximum value, Minimum value etc.

Table 2: Descriptive statistics of attributes

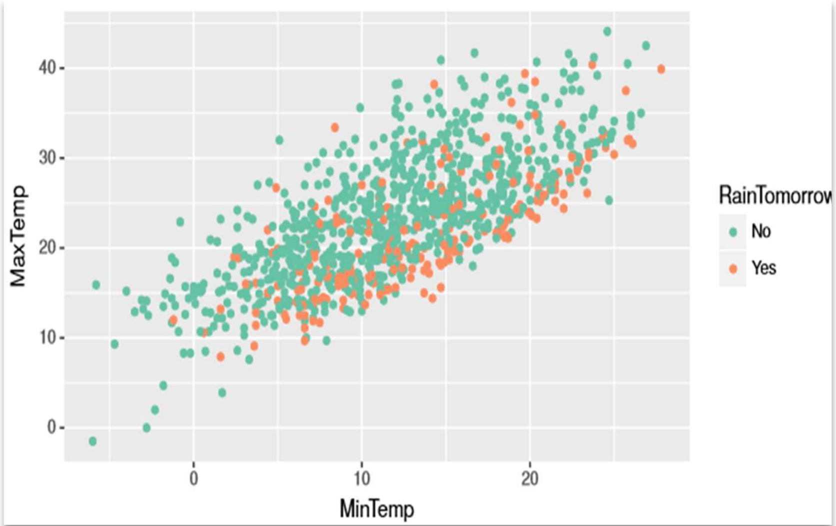
Attribute	Mean	Median	Standard deviation	Minimum value	Maximum value
MinTemp	13.46	13.20	6.41	-6.70	31.40
MaxTemp	24.22	23.90	6.97	4.10	48.10
evaporation	5.503	5.0	3.696	0.00	81.20
Sunshine	7.76	8.60	3.758	0.00	14.50
WindGustSpeed	40.88	39.0	13.33	9.00	124.00
Humidity9am	65.87	67.0	18.51	0.0	100.00
Cloud3pm	4.327	5.00	2.64	0.0	9.00

2.4 Visualization

Data visualization helps in understanding the problem of interest very well. In our case, the target variable is a binary variable can be plotted against the independent variables to give more insight into the problem.

Scatter plot can be used to identify the type of relationship between two independent variables. Since our target variable is binary, we can visualize the separation of two classes(with different colors) by plotting independent variables against each other.

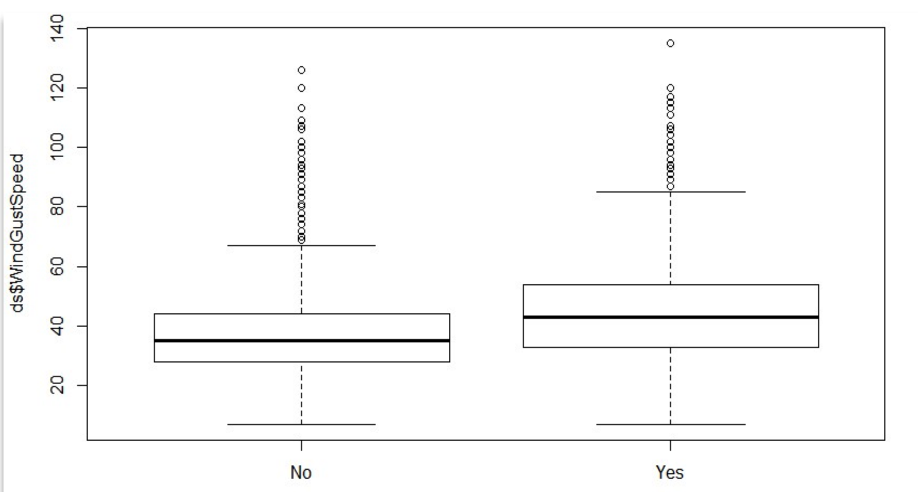
Figure 1 Scatter plot of MinTemp vs MaxTemp



The above scatter plot is obtained by plotting MinTemp against MaxTemp(1000 samples) to show how these variables separate the target variable RainTomorrow. As we can see from the scatter plot that the relationship is non-linear. Therefore, Non-linearity of the data is one of the problem to address in this objective.

[4]Box-plots are used for graphically depicting groups of numerical data through their quartiles. Also, Box-plots are one of the best tools to identify any outliers in the data. we can simply create a box-plot of target variable against any predictors to check for outliers.

Figure 2 Box-plot of target variable against WindGustSpeed



Box-plot of wind-gust speed against target variable shows that the number of outliers for both the classes are very high.

3. Data Preparation

As mentioned earlier, Data preparation is an important part of this project. We show different techniques are applied to enrich the quality of data in this section. The different techniques are extracting important information from a particular variable, Variable selection using stepwise AIC and Principal component analysis.

3.1 Missing data and Pre-processing:

We omit the rows that have missing values even for any one of measured features. we do so, Imputing missing values with column mean results in poor performance of models due to added noise. After omission, the number of observations reduces to 56420.

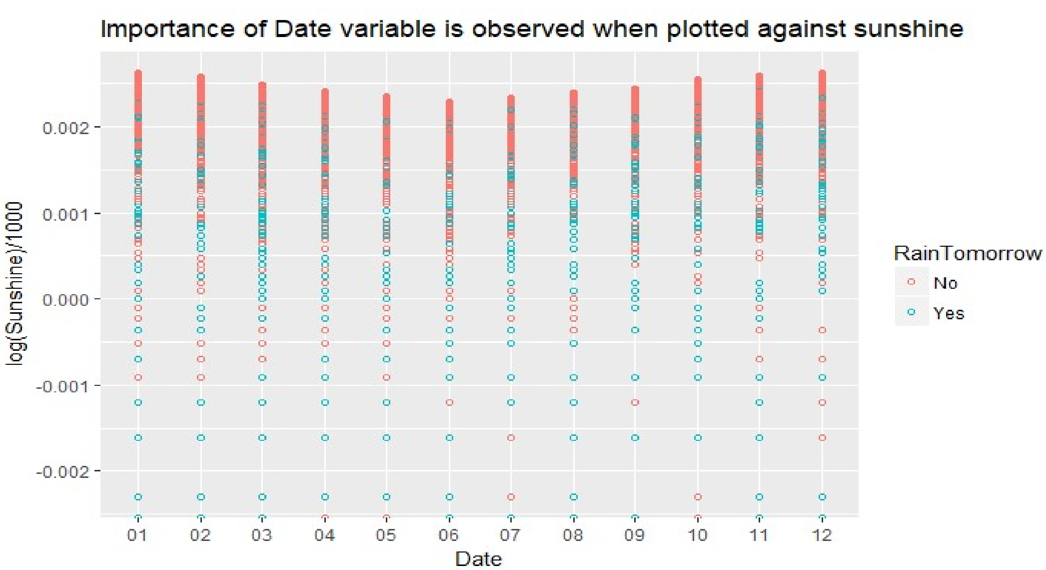
We standardize the quantitative variables to have a zero mean and unit standard deviation. We also create dummies for categorical variables by converting them into factors.

We separate the whole data into training and validation data where 60% of data is used as training and 40% is used for validating the models.

3.2 Date Variable

Date variable is important because it gives us important information on year, month, day of the recorded observations. Especially, for rain prediction we extract the seasonal information from the date variable by grouping the observations based on the month it is recorded.

Figure 3 Scatter plot of Date in months vs log(sunshine)/1000



We can see from the scatter plot that data variable grouped into months plotted against sunshine variable shows a pattern of how the target variable is distributed. Therefore, it is important to use this information for rain predictions.

3.3 variable selection using step-wise Regression:

Step-wise regression does both forward and backward selection of predictors and finally gives us a set of predictors based on their statistical significance. The importance is based on **if p-value <0.05(Significance level),** the variable is considered as statistically significant.

Figure 4 Results of stepwise regression on pre-processed data

```
Call:
glm(formula = new.data$RainTomorrow ~ MinTemp + MaxTemp + Rainfall +
    Sunshine + WindGustDir + WindGustSpeed + WindDir9am + WindDir3pm +
    WindSpeed9am + WindSpeed3pm + Humidity9am + Humidity3pm +
    Pressure9am + Pressure3pm + Cloud9am + Cloud3pm + RainToday +
    RISK_MM, family = "binomial", data = new.data[, -23], control = list(maxit = 50,
    epsilon = 1))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1812  -0.6116  -0.4635  -0.3301   2.5159

Coefficients:
               Estimate Std. Error z value             Pr(>|z|)
(Intercept)  33.0571032  1.9669704  16.806 < 0.0000000000000002 ***
MinTemp      -0.0376434  0.0036525 -10.306 < 0.0000000000000002 ***
MaxTemp       0.0300431  0.0039321   7.640   0.0000000000000217 ***
Rainfall     -0.0051523  0.0017189  -2.997             0.002722 **
Sunshine     -0.1016516  0.0047279 -21.500 < 0.0000000000000002 ***
WindGustDir   0.0094464  0.0027767   3.402             0.000669 ***
WindGustSpeed  0.0309821  0.0012305  25.178 < 0.0000000000000002 ***
WindDir9am   -0.0081792  0.0024578  -3.328             0.000875 ***
WindDir3pm    0.0046593  0.0027370   1.702             0.088692 .
WindSpeed9am -0.0021778  0.0015791  -1.379             0.167833
WindSpeed3pm -0.0185729  0.0016519 -11.243 < 0.0000000000000002 ***
Humidity9am  -0.0037030  0.0008311  -4.455   0.0000083700372431 ***
Humidity3pm   0.0322340  0.0009495  33.947 < 0.0000000000000002 ***
Pressure9am   0.0595823  0.0065645   9.076 < 0.0000000000000002 ***
Pressure3pm  -0.0960221  0.0064983 -14.777 < 0.0000000000000002 ***
Cloud9am     -0.0195747  0.0051831  -3.777             0.000159 ***
Cloud3pm      0.0268633  0.0055246   4.863   0.0000011591123352 ***
RainToday     0.4312313  0.0308423  13.982 < 0.0000000000000002 ***
RISK_MM       0.0677434  0.0012618  53.689 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59493  on 56419  degrees of freedom
Residual deviance: 35637  on 56401  degrees of freedom
AIC: 35675

Number of Fisher Scoring iterations: 1
```

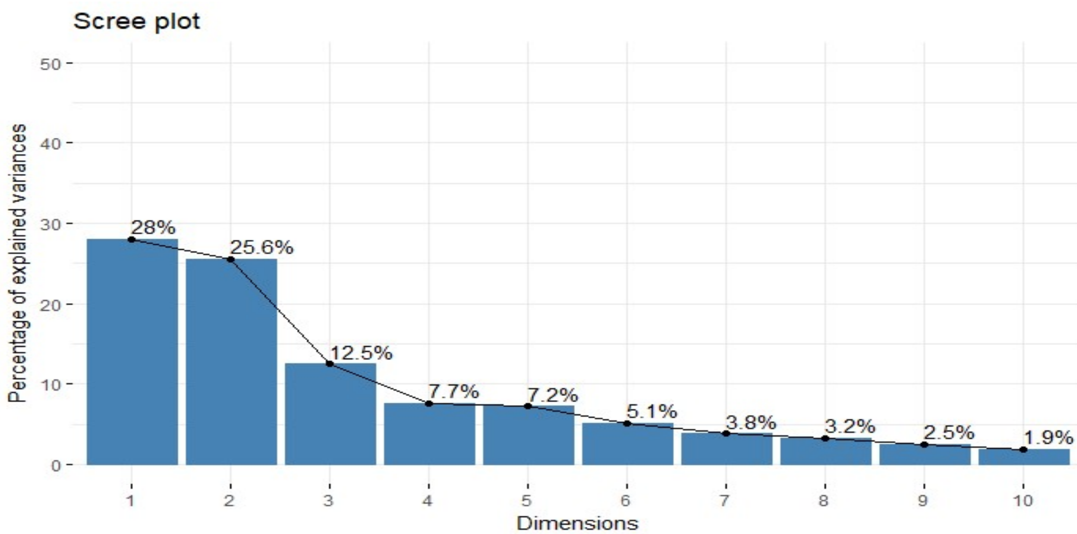
From The above figure we can infer that the following variables are statistically important **: MinTemp , MaxTemp , Rainfall, Sunshine, WindGustDir, WindGustSpeed, WindDir9am, windSpeed9am, windSpeed3am, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, cloud9am, cloud3pm, RainToday, Risk_MM.**

3.4 Principal component analysis(PCA)

PCA is one of the techniques used for dimensionality reduction. It is powerful and principal components is obtained through linear combination of predictors. Principal components(PC's) are orthogonal to each other. First principal component explains the highest amount of variance in the data followed by the second principal component.

In our project, we apply PCA on 13 Standardized numerical variables. We obtain the principal components and find how much variance in percentage each of the PC's explains by plotting a scree plot.

Figure 5 Scree plot showing percentage of variance explained by each PC



We can infer from the above scree plot that first seven principal components explains almost 90% of the variance in the data. Therefore, we can use seven principal components instead of 13 numerical predictors for model fitting. This shows that PCA is a powerful tool for dimensionality reduction.

4. Methodology

In this section, we explain how we use the prepared data for fitting models like Logistic regression, Random forest. We also show the effect of variable selection and PCA on these models to value their importance. In the end, we analyze the performance of different models based on different evaluation metrics.

4.1 Logistic regression

[5]Logistic regression uses a logistic function to model a binary dependent variable. It calculates log-odds for each of the binary classes by using a linear combination of independent variable. Logistic function uses this log-odds to give probabilities for each of the classes. Class probabilities are then used for classifying each individual into a particular class based on certain criterion.

Initially, we fit the logistic regression model using all the variables in the data. we also fit logistic regression model using statistically significant variables obtained through stepwise regression.

4.2 Random forest ensemble:

Random forest is the ensemble tree method which grows number of trees on the boot-strapped samples and uses subset of features for each split to reduce the correlation between the trees.

We Fit the Random forest model using significant variables and the mean decrease in Gini index for each variable can be used to assess the importance of each variable.

Table 3 Variable importance based on Mean decrease in Gini Index

Variable	Mean decrease in Gini Index
Date	526.72
Location	925.11
Rainfall	445.98
Sunshine	1168.0
WindGustDir	662.52
WindDir9am	687.17
WindDir3Pm	657.78
Humidity3pm	1880.60
Pressure3pm	583.79
Cloud3pm	574.46

We can infer from the above table that Sunshine and Humidity3pm are the important variables used for separating the data during each split in random forest.

4.3 Random forest on principal components:

We Fit the Random forest model on first two principal components along with categorical variables to analyze the effect of PCA on Random forest model. By principal components, we mean the projection of PC's obtained through PCA onto the original dimension of the data.

We can analyze the effect of PCA from mean decrease in Gini index due to PC's in random forest model.

Table 4 Variable importance based on Mean decrease in Gini Index

Variable	**Mean decrease in Gini Index**
PC1	3222.42
PC2	2456.89
WindGustSpeed	1076.76
WindDir9am	1095.08
WindDir3pm	1044.48
Date	1101.38
Location	1349.57

From the above table, we can infer that PC1 and PC2 are the most important variables as they have the highest mean decrease in Gini index value. One interesting thing to note that, Date variable that we included as months in the data preparation proves to be important as we can see a reasonable decrease in Gini index when using this variable in our model.

4.4 Evaluation metrics

We use various evaluation metrics to compare the performance of the models we proposed. They are listed in the following table.

Table 5 evaluation metrics

Metric	Formula
Accuracy	(TP + TN)/(FP+FN+TP+TN)
Sensitivity or recall	TP/(TP+FN)
Specificity	TN/(FP+TN)

Where, TP – true positives , TN – true negatives, FP- false positives, FN- false negatives.

5. Results

In this section, we present the results obtained for various models that we proposed in the last section. The results are performance of models based on evaluation metrics in the last section. Results shown in the below table is obtained on the validation set that separated initially during pre-processing stage.

Table 6 performance of various models using evaluation metrics

Metric	Logistic regression	Logistic regression (using significant variables)	Random forest (using significant variables)	Random forest (using PC's with categorical)
Accuracy	0.857	0.854	0.8645	0.8515
Sensitivity or recall	0.955	0.957	0.938	0.920
Specificity	0.5053	0.404	0.597	0.60

6. Summary and conclusion

In summary, we can see that data preparation like data omission, variable selection, dimensionality reduction, extracting important information helps the model to improve their performance in terms of space, time and mis-classification.

Initially, we had converted date variable into months which provided more insight on how rainfall is seasonally distributed proves to be important in random forest model. Variable selection through stepwise regression removes insignificant information which helps in reducing the complexity of data for logistic regression. Logistic regression with significant variables perform similar to that of with all the variables proves that stepwise regression is a good variable selection method.

Principal component analysis on numerical variables decreased the dimensions from 13 to 7 which proves that PCA is a powerful dimensional reduction technique. Also, performance of random forest model using just two PC'S is similar to random forest model using 13 numerical predictors.

Random forest model achieves 86% outperforming Logistic regression by 1%. For rain data containing both categorical and numerical predictors, random forest model achieves better performance than logistic regression.

On conclusion, Prediction of rain using various independent weather data can become time consuming without the use of techniques like variable selection and PCA. I suggest that better variable selection along with good model selection will help achieve better results.

7. References

[1] https://www.talend.com/resources/what-is-data-preparation/

[2] https://pmm.nasa.gov/applications/climate-prediction

[3] https://en.wikipedia.org/wiki/Descriptive_statistics

[4] https://en.wikipedia.org/wiki/Box_plot

[5] **https://en.wikipedia.org/wiki/Logistic_regression**