Melbourne Rain Prediction

MATH 2319 Machine Learning Applied Project Phase I

Rahul K Gupta (s3635232) & Terrie Christensen (s3664899)

7 April 2018

Contents

1 Introduction

The objective of this project was to build classifiers to predict whether an individual earns more than USD 50,000 or less in a year from the 1994 US Census Data. The data sets were sourced from the UCI Machine Learning Repository. This project has two phases. Phase I focuses on data preprocessing and exploration, as covered in this report. We shall present model building in Phase II. The rest of this report is organised as follow. Section 2 describes the data sets and their attributes. Section 3 covers data pre-processing. In Section 4, we explore each attribute and their inter-relationships. The last section ends with a summary.

2 Data Set

The UCI Machine Learning Repository provides five data sets, but only `adult.data`, `adult.test`, and `adult.names` were useful in this project. `adult.data` and `adult.test` are the training and test data sets respectively. `adult.names` contains the details of attributes or variables. The training data set has 32,561 training observations. Meanwhile, the test data set has 16,281 test observations. Both data sets consist of 14 descriptives features and one target feature. In this project, we combined both training and test data into one. In Phase II, we would build the classifiers from the combined the data set and evaluate their performance using cross-validation.

2.1 Target Feature

The response feature of rain is given as:

$$\text{Tomorrow's Rain} = \begin{cases} Yes & \text{if Rain will occur tomorrow} \\ No & \text{otherwise} \end{cases} \tag{1}$$

The target feature has two classes and hence it is a binary classification problem. To reiterate, The goal is to predict **whether it will rain in Melbourne tomorrow**.

2.2 Descriptive Features

The variable description is produced here from `adult.names` file:

- age: continuous.
- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- fnlwgt: continuous.
- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th- 8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- education-num: continuous.
- marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married- spouse-absent, Married-AF-spouse.
- occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-*inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv- house-serv, Protective-serv, Armed-Forces.
- relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex: Female, Male.
- capital-gain: continuous.

- capital-loss: continuous.
- hours-per-week: continuous.
- native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

Most of the descriptive features are self-explanatory, except fnlwgt which stands for "Final Weight" defined by the US Census. The weight is an "estimate of the number of units in the target population that the responding unit represents". This feature aims to allocate similar weights to people with similar demographic characteristics. For more details, see US Census.

3 Data Pre-processing

3.1 Preliminaries (Optional)

In this project, we used the following `R` packages.

```
library(mlr)
library(tidyverse)
require(corrplot)
library(reshape2)
theme_set(theme_minimal())
```

Text

```
weather <- read.csv('weatherAUS 2.csv')
weather$Date = as.Date(weather$Date,'%Y-%m-%d')

# Filter data for Melbourne
Melbourne = weather[weather$Location %in% c('Melbourne','MelbourneAirport'),]

# Add Month and Year-Month
Melbourne$Month = strftime(Melbourne$Date,"%b")
Melbourne$MonthYear = strftime(Melbourne$Date,"%b-%Y")

# Remove data where Tomorrow rain information is missing
# Melbourne = Melbourne %>% drop_na(RainTomorrow)
```

3.2 Data Cleaning and Transformation

With `str` and `summarizeColumns` (see Table 1), we noticed the following anomalies:

- All character columns contained excessive white space.
- The target feature, `income` had a cardinality of 4, which was supposed to be 2 since `income` must be binary.
- The `education_num` ranged from 1 to 16 which coincided with the cardinality of `education`. They might represent the same information.
- The max value of `capital_gain` was 99999, potentially a value to represent missing value.
- The max value of `hours_per_week` was 99. It could be a valid or missing value
- On surface, each feature had no missing value, especially the character features.

```
str(Melbourne)
```

```
## 'data.frame':    6202 obs. of  26 variables:
##  $ Date         : Date, format: "2009-01-01" "2009-01-02" ...
##  $ Location     : Factor w/ 49 levels "Adelaide","Albany",..: 20 20 20 20 20 20 20 20 20 20 ...
##  $ MinTemp      : num  11.2 7.8 6.3 8.1 9.7 13.5 15.8 9.7 10.2 10.7 ...
##  $ MaxTemp      : num  19.9 17.8 21.1 29.2 29 31.7 21.4 18.4 19.7 23.6 ...
##  $ Rainfall     : num  0 1.2 0 0 0 0 0.2 0.2 0 0 ...
##  $ Evaporation  : num  5.6 7.2 6.2 6.4 7.4 7.2 8.8 5.2 5.6 7.2 ...
##  $ Sunshine     : num  8.8 12.9 10.5 12.5 12.3 13.7 4.4 11.5 12.6 10.2 ...
##  $ WindGustDir  : Factor w/ 16 levels "E","ENE","ESE",..: 13 11 11 11 10 11 9 10 9 9 ...
##  $ WindGustSpeed: int  69 56 31 35 33 50 46 56 43 43 ...
##  $ WindDir9am   : Factor w/ 16 levels "E","ENE","ESE",..: 14 13 1 5 13 6 12 12 11 16 ...
##  $ WindDir3pm   : Factor w/ 16 levels "E","ENE","ESE",..: 13 11 9 11 11 9 11 11 9 9 ...
##  $ WindSpeed9am : int  33 31 13 2 9 11 17 28 19 7 ...
##  $ WindSpeed3pm : int  43 26 19 20 20 28 28 35 28 24 ...
##  $ Humidity9am  : int  55 50 51 67 51 50 98 51 51 76 ...
##  $ Humidity3pm  : int  37 43 35 23 31 34 67 42 42 46 ...
##  $ Pressure9am  : num  1005 1018 1021 1016 1012 ...
##  $ Pressure3pm  : num  1006 1019 1018 1013 1010 ...
##  $ Cloud9am     : int  7 6 1 5 6 0 8 2 2 7 ...
##  $ Cloud3pm     : int  7 7 7 4 2 1 7 5 7 1 ...
##  $ Temp9am      : num  15.9 12.5 13.4 16 19.4 21.3 16 14.5 14.2 14.5 ...
##  $ Temp3pm      : num  18.1 15.8 19.6 28.2 27.1 29.8 19.9 17.7 19.3 21.8 ...
##  $ RainToday    : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 1 ...
##  $ RISK_MM      : num  1.2 0 0 0 0 0.2 0.2 0 0 0 ...
##  $ RainTomorrow : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ Month        : chr  "Jan" "Jan" "Jan" "Jan" ...
##  $ MonthYear    : chr  "Jan-2009" "Jan-2009" "Jan-2009" "Jan-2009" ...
```

```
summarizeColumns(Melbourne) %>% knitr::kable( caption =  'Feature Summary ')
```

Table 1: Feature Summary

name	type	na	mean	disp	median	mad	min	max	nlevs
Date	Date	0	NA	NA	NA	NA	1.0	2.0	3193
Location	factor	0	NA	0.4851661	NA	NA	0.0	3193.0	2
MinTemp	numeric	480	10.829867	4.4463106	10.40	4.59606	-1.0	30.5	0
MaxTemp	numeric	481	20.623405	6.3692722	19.30	6.07866	8.4	46.8	0
Rainfall	numeric	758	1.638979	4.8477676	0.00	0.00000	0.0	82.2	0
Evaporation	numeric	6	4.647418	3.3289066	4.00	2.96520	0.0	23.8	0
Sunshine	numeric	2	6.383355	3.9120865	6.50	4.89258	0.0	13.9	0
WindGustDir	factor	29	NA	NA	NA	NA	7.0	1891.0	16
WindGustSpeed	integer	29	46.245100	16.0369498	44.00	16.30860	11.0	122.0	0
WindDir9am	factor	90	NA	NA	NA	NA	28.0	2031.0	16
WindDir3pm	factor	25	NA	NA	NA	NA	46.0	1301.0	16
WindSpeed9am	integer	3	19.686885	11.6437606	17.00	11.86080	0.0	67.0	0
WindSpeed3pm	integer	0	22.574815	9.9575278	22.00	10.37820	0.0	76.0	0
Humidity9am	integer	490	68.896008	15.3057683	70.00	14.82600	11.0	100.0	0
Humidity3pm	integer	496	51.175780	17.0449048	50.00	14.82600	6.0	100.0	0
Pressure9am	numeric	480	1017.872772	7.7442968	1018.05	7.63539	988.9	1039.3	0
Pressure3pm	numeric	483	1016.041860	7.5777557	1016.40	7.56126	988.2	1036.0	0
Cloud9am	integer	1034	5.274574	2.5043944	7.00	1.48260	0.0	8.0	0
Cloud3pm	integer	1107	5.293229	2.3297093	6.00	1.48260	0.0	8.0	0

name	type	na	mean	disp	median	mad	min	max	nlevs
Temp9am	numeric	481	14.348278	4.8617942	13.80	4.59606	2.9	35.5	0
Temp3pm	numeric	484	19.100087	6.1225250	18.00	5.93040	6.2	46.1	0
RainToday	factor	758	NA	NA	NA	NA	1289.0	4155.0	2
RISK_MM	numeric	758	1.638942	4.8477444	0.00	0.00000	0.0	82.2	0
RainTomorrow	factor	758	NA	NA	NA	NA	1289.0	4155.0	2
Month	character	0	NA	0.9100290	NA	NA	452.0	558.0	12
MonthYear	character	0	NA	0.9900032	NA	NA	30.0	62.0	105

3.3 Univariate Visualisation

3.3.1 Numerical Features

3.3.1.1 Age

4 Summary