Course Project

Moon, Hyosik

July 19, 2021

1 Required

- **Main objective of the analysis**
  Predict next-day rain by training classification models on the target variable RainTomorrow.

- **Brief description of the data set**
  This dataset contains about 10 years of daily weather observations from many locations across Australia.

  RainTomorrow is the target variable to predict. It means – did it rain the next day, Yes or No? This column is Yes if the rain for that day was 1mm or more.

  In data set, there are 23 columns that 16 floats and 7 objects. 1

- **Brief summary of data exploration**

  1. Data cleaning, Delete unused features to predict the Ladder score. All data seem to be needed to predict the RainTomorrow, so before we check the correlation we won't delete any columns.
  2. In order to utilize Date columns, we need to change the type from object to int, and let's use just months for convenience.
  3. To see pairplot, we have to change categorical variables to numeric variables. So, let's change the objects to numeric by using LabelEncoder.
  4. Find the correlation between RainTomorrow and other features (Figure 2).
  5. Change the categorical variable to numeric variables. (Figure **??**)

- **Summary of training at least three linear regression models**

- **Explanation of your final regressions model**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 145460 entries, 0 to 145459
Data columns (total 23 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   Date           145460 non-null  object
 1   Location       145460 non-null  object
 2   MinTemp        143975 non-null  float64
 3   MaxTemp        144199 non-null  float64
 4   Rainfall       142199 non-null  float64
 5   Evaporation    82670 non-null   float64
 6   Sunshine       75625 non-null   float64
 7   WindGustDir    135134 non-null  object
 8   WindGustSpeed  135197 non-null  float64
 9   WindDir9am     134894 non-null  object
 10  WindDir3pm     141232 non-null  object
 11  WindSpeed9am   143693 non-null  float64
 12  WindSpeed3pm   142398 non-null  float64
 13  Humidity9am    142806 non-null  float64
 14  Humidity3pm    140953 non-null  float64
 15  Pressure9am    130395 non-null  float64
 16  Pressure3pm    130432 non-null  float64
 17  Cloud9am       89572 non-null   float64
 18  Cloud3pm       86102 non-null   float64
 19  Temp9am        143693 non-null  float64
 20  Temp3pm        141851 non-null  float64
 21  RainToday      142199 non-null  object
 22  RainTomorrow   142193 non-null  object
dtypes: float64(16), object(7)
memory usage: 25.5+ MB
```

Figure 1: Data set information

- **Summary Key Findings and Insights**
- **Suggestions for next steps**

	RainTomorrow
RainTomorrow	1.000000
Humidity3pm	0.406050
Cloud3pm	0.355419
RainToday	0.329409
Cloud9am	0.291214
Humidity9am	0.233368
Rainfall	0.224268
WindGustSpeed	0.206299
MinTemp	0.074597
WindSpeed3pm	0.072610
WindSpeed9am	0.070565
WindGustDir	0.050768
WindDir9am	0.039284
WindDir3pm	0.027859
Month	0.008902
Location	-0.013112
Temp9am	-0.025140
Evaporation	-0.109716
MaxTemp	-0.145877
Temp3pm	-0.176824
Pressure3pm	-0.208366
Pressure9am	-0.226648
Sunshine	-0.408096

Figure 2: Correlation