Dataset Description:
The dataset contains about 10 years of daily weather observations from many locations across Australia. The dataset has collected from Kaggle. Observations were drawn from numerous weather stations. The daily observations are available from http://www.bom.gov.au/climate/data

The dataset has a total 23 columns and there are 145460 observations. The dataset is mainly built with the intention of predicting whether based on the data available from 22 columns whether it will rain in the next day. RainTomorrow is the target variable to predict. This column is Yes if there was rain of 1 mm or more the next day or else No. Other attributes which the dataset contains are:

1. Date: The date in which the observation is recorded.
2. Location: the common name of the location of the weather station.
3. MinTemp: the minimum temperature in degree celsius.
4. MaxTemp: the maximum temperature in degree celsius.
5. Rainfall: the amount of rainfall recorded for the day in mm.
6. Evaporation: The so-called Class A pan evaporation(mm) in the 24 hours to 9 am.
7. Sunshine: The number of hours of bright sunshine in the day.
8. WindGustDir: the direction of the strongest wind gust in the 24 hours to midnight.
9. WindGustSpeed: the speed (km/h) of the strongest wind gust in the 24 hours to midnight.
10. WindDir9am: Direction of the wind at 9am.
11. WindDir3pm: Direction of the wind at 3pm.
12. WindSpeed9am: wind speed (km/hr) averaged over 10 minutes prior to 9am.
13. WindSpeed3pm: wind speed (km/hr) averaged over 10 minutes prior to 3am.
14. Humidity9am: humidity (present) at 9am.
15. Humidity3pm: humidity (present) at 3pm.
16. Pressure9am: Atmospheric pressure (hpa) reduced to mean sea level at 9am.
17. Pressure3pm: Atmospheric pressure (hpa) reduced to mean sea level at 3pm.
18. Cloud9am: Fraction of sky obscured by cloud at 9am.
19. Cloud3pm: Fraction of sky obscured by cloud at 3pm.
20. Temp9am: Temperature (degrees C) at 9am.
21. Temp3pm: Temperature (degrees C) at 3pm.
22. RainToday: Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceed 1mm, otherwise 0.

Initial plan for data exploration
The first step would be to perform some Exploratory data Analysis followed by getting some information about the dataset which will include how many missing values are there for each column. After that, steps will be taken regarding the missing values i.e. whether to delete the attribute or the row or fill the missing values with some appropriate value. The next step would be to visualise the data and try to get some correlation among the various predictors and the target variable. Then Feature

engineering and standardising the various columns of the data would be performed for the categorical variables. After which 3 hypothesis test will be performed.

Actions taken for data cleaning and feature engineering

First of all various information about the dataset was found using dataset.info(), dataset.describe(), dataset.dtypes method. The correlation among the columns were checked using the dataset.corr() method. After getting insights from the information of the dataset, number of missing values were calculated for each of the variables. From there I found out that the columns 'Evaporation', 'Sunshine', 'Cloud9am', 'Cloud3pm' have nearly half of the data missing. So I have decided to entirely delete those columns as replacing the missing values will not help the model. For the rest of the columns, the missing values were replaced with the mean value of the respective column for numerical columns while the missing values were replaced with most common observation for the categorical columns.

Steps were taken in order to standardise the dataset. I didn't have to worry about the categorical data. I changed standardised each of the numerical variables of the dataset using StandardScaler() method provided by sklearn. This lead to better representation of the dataset as all of the column values now were in the similar range. Also there are 4 categorical variables present in the dataset which were converted to one hot representation using the pd.get_dummies() method.

Key Findings and Insights from Exploratory Data Analysis

From the initial information collected, one of the first thing that is found was the amount of missing values the dataset had for certain columns. Nearly 50% of the data was missing for 4 columns and so I decided to drop them totally. Then, I noticed the difference in range of values for each column. This highly manipulates the result and hence I decided to standardise them. I also found from the dataset that the 'RainTomorrow' variable had nearly 70% of the data belonging to 'No' response and rest being 'Yes'. From the various barplots that were drawn using the seaborn package, it was noticeable that the amount of Rainfall from the previous day highly influenced the possibility of Raining the next day as well. It was also found that the Pressure at 9am and 3 pm had very little influence on the target variable RainTomorrow and we can ignore those 2 variables while training our model. Then I also converted the categorical variables into one hot representation so that our model fit the dataset better.

3 Hypothesis about this data

1. Null Hypothesis: There is no relationship between the Minimum temperature of today with respect to the possibility of Raining Tomorrow.

Alternate Hypothesis: there exist a relationship between the minimum temperature of today and the possibility of Raining Tomorrow.

2. Null Hypothesis: There is no relationship between the rainfall today and the rain tomorrow variable.

   Alternate Hypothesis: There is no relationship the variable rainfall today and the rain tomorrow variable.

3. Null Hypothesis: There is no relationship between the pressure at 3pm to the chances of raining tomorrow.

   Alternate Hypothesis: There is a relationship between the pressure at 3pm to the chances of raining tomorrow

1. Before performing the test, the significance level was set to 95% i.e. p-value of 0.05.

From the given the dataset the hypothesis test was performed and the result I found was: The Pearson Correlation Coefficient is 0.08217306311605427 with a P-value of P = 2.5217334966587718e-216. Clearly, the p-value is much less then the significance level and hence I decided to reject the null hypothesis that there exist no relationship between the variable MinTemp today in relation to the target variable RainfallTomorrow.

Suggestions

There were 4 columns that we had to delete completely and hence we might have lost quite a lot of useful information about the data. As a suggestion if somehow I could manage 20-30% of the missing data by spending some time on the internet then the result can be much better.

Summary

The dataset is found out to be quite useful for the context it is presented for. The data was collected from nearly 49 cities and around 1,50,000 observations were present. There are 22 columns present which is enough weather data to find a relationship between the predictor variables and the target variables. I would suggest the missing

data for those 4 columns to be added so that we do not completely throw away the columns.