RAINFALL WEATHER FORECATING

Problem Statement

Rain Prediction Weather forecasting

Rain prediction is the main objectives of the weather forecasting purpose and Weather forecaasting is the application of science and technology to predict the conditions of the atmosphere for given given location & time . Weather forecating are made by collecting quantitative data about the current state of the atmosphere at a given place and using meteorology to project how the atmosphere will change.

Rain Dataset is to preedict whether or not it will rain tomorrow.

The dataset contains about 10 years of daily weather observations of different locations in Australia.

This dataset contain 8425 rows and 23 Attributes

for example= Locations,date,sunshine,humidity, pressure(atmospheric pressure), windspeed temprature, cloud, wind-direction m similar columns are given in this dataset

= I have to make such a model through Machine Learning that it predicts whether it will rain tomorrow or not

= For this i mainly do first data preprocessing then EDA on preprocessed data , all these process come under data analysis

The Target-Attributes of this Dataset is Classification-Problem

Data sources from github and format of the dataset is csv

DATA ANALYSIS

Let's start data preprocessing

. First load the datset on jupyter notebook and print it then check the shape of the dataset

. I started looking at randomly some rows and columns of dataset and tried to understand

. While checking the null elements of the dataset , it was observed that each column had null elements and in four columns there were more than 40 percent null elements. About 10 percent elements of the total rows were missing in most of the Columns .

. Filling the mssing elements is a very logical problem in the data preprocessing part because we do not know how much the elements that fill the elements according to the hypotheses of statistical analysis will be similar to the original data

. Describe the dataset before filling missing elements

.Describe the dataset that gives the statistical description and from describe it is known that what is the min, max, mean & Std of each Attributes and this the destribution of minimum & maximum elements of the columns is also known. Now let's fill the missing elements of the dataset. From using describe function we also get information about Variables distribution and which Variables can have outliers

. I filled the missing elements according to the mean & mode of grouby method of 'Location' & 'Date ' columns of this dataset because these were the Categorical Attributes whose elements could be filled according to the groupby method(pandas Librabry)

. Now checked the info of the Dataset, most of the columns are of floats types and six columns (Location,RainToday,RainTomorrow,WindGustDir,WindDir9am,WindDir3pm) are Categorical types

. Again Describe the dataset after fill the missing value then i saw that no more differences found that gives the almost same statistical description and from describe it is known that what is the min, max, mean & Std of each Attributes and this the destribution of minimum & maximum elements of the columns is also known

.Now there is no missing elements in th data and have also checked info , Statistical description of the data data , now there is no data cleaning or any modification done in it and LabelEncoding over Categorical columns will be done after Data Visualization, so now let's move towards EDA

Now let's do EDA on the Dataset

.I visualize the data using Matplotlib & seaborn

.Four types of plotting were used on the columns of this Dataset such as Countplot, Barplot, Lineplot & Scatterplot

. scatterplot using on continuous attributes & countplot use only on categorical attributes & in barplot use on categorical, continuous Attributes

. First Countplot with Target-columns then found that the elements of Target-column are almost imbalanced were spread in the ratio of 70:30 .

. This will make it more likely that the accuracy\_score may be affected while making the model. To solve this problem , we will use SMOTE (Synthetic Minority OverSampling Technique)before making the model.

. I did a countplot of all the categorical columns with the Target-column

. We done Analyzing correlation of the Dataset through plotting correlation plot by using seaborn heatmap . In this correlation plot , I see how each Variables are correlated to each other and to the target-Attribute and from this correlation plot we get a lot information about the dataset like how positive or negative each column is correlated with each other

. Through the correlation plot it was found that out of 23 columns , 4 columns are highly positive correlated to each and no more positive correlated with the target-column . correlation plot also give that information to which Attributes are important and which are not.

. Plotting the Pairplot after the correlation plot , it also gives the same information as the correlation plot , the only difference is that it shows the relationship between each of the Attributes in a visual way .From Pairplot we also known how to each Variables are distributed with each other

. After plotting correlation plot & pairplot , we get most of the information about all the Attributes of tha dataset

. Now I used sctterplot, Barplot & Lineplot to understand the all Variables of dataset in a better way and finally after plotting I got information about all the Variables of the Dataset which Variables are more important to predict the Target-Variable and which are not

. It also came to know that which Variables are useless in dataset for predicting Target

. So far we have done almost most of the work of data analytics on the data, now we just have to check the Skewness,Outliers & Multicollinearity problem of tha data

DISTRIBUTION PLOT

. Plotting distribution plot over continuous Attributes to check how data is distributed and whether it is normal or not , this information is shown in the distribution plot. So there were some such Attributes whose distribution was skewed

SKEWNESS

=Now I am going to check the skewness of the data to confirm whether there is skewness or not . On checking the skewness, two attributes are skewed ('Rainfall', 'Evaporation'), so to reduce the skewness of skewed Attributes used (np.log1p) function of numpy then skewness becomes close to the normal distribution.

BOXPLOT

. Now Boxplot is used to check outliers . The important point is that boxplot should be plotted only on continuous Attributes and skewness should also be checked of continuous Attributes.

. outliers are present in ten columns of this dataset shown in above Boxplot but make sure it i used to zscore, but to make sure i am going to use Zscore method

ZSCORE

. After using Zscore , it was found out that the outliers were visible in the boxplot, then through Zscore method the outliers were removed and then checked the shape of the dataset, then 350 rows were dropped

MULTICOLLINEARITY PROBLEM

. We checked multicollinearity problem of this dataset then we saw that in this have multicollinearity problem we have to removed three Variables(MaxTemp', 'Temp3pm', 'Pressure9am')

SCALING THE DATASET

. we used StandardScaler to scaling the Input\_Variables

SMOTE

. The dataset is Imbalanced so we used to SMOTE for imbalanced the dataset . It is the last process before built ML model

. we don't use feature-selection method and PCA on this dataset because there is not much Variables in this Dataset that we should use feature\_selection and PCA method

EDA CONCLUDING REMARKS

In EDA , we explore the data and try to understand it as much as possible because when we understand all the Variables of the dataset in a better way, we will be able to get useful information from it , the most important thing is that from this is it also known that which Variables different each other and how much they are related and it it easy to find this pattern and also find which Variables are more suitable/important for predicting the target

In EDA we using different types of method

. We do Univariate-Analysis, Bivariate-Analysis & multi\_variate Analysis through many types of plotting method like plotting Countplot,correlation plot, Pairplot, Barplot, scatterplot which i get most of the information about Attributes & which Attributes are more important and also understanding the outliers

. I got important information about this Rainfall-weather-forecasting data through EDA, as most of the Attributes of this Dataset are very imprortant to predict the Target & some Variables are useless so i will have to drop those Variable but the data is somewhat impure.

PREPROCESSING PIPELINE

. If we take any raw data to make it usable in ML model so that the model performs well on that data , then the whole process used on that raw data come under Data Preprocessing

. Important process in the Data Preprocessing pipeline is as Data cleaning, Data-Integretion, Data-Reduction, Data-Transformation & Data-discretization

1.Data-Cleaning = The important thing in data cleaning is that we fill the missing , null or duplicate elements of the data. we filling null elements by using pandas groubpy function.we also removed Outliers from this dataset using Boxplot & Zscore method

2. Data-Integretion = in this process nothing to be done because dataset load from github in single file & csv format and dataset did not have to be marged

3. Data-Reduction= we dropped some Attributes(maxtemp, year,temp9am, pressure9am) from this dataset according to multicollinearity problem and these Attributes are also not important to predict the target

some rows droped by outliers method(zscore-method). This Datset is not so big that it should be given more attention to data reduction-point

4. Data-Transformation= we apply Label-Encoding over some Categorical Attributes to convert numerical types elements and after this we do scaling the dataset by using StandardScaler (sklearn method) StandardScaler convert data into range(-1 to +1)

5. Data-Discretization= in this dataset have no any Attributes to use over it discretization method(pandas.cut, numpy.digitize)

BUILDING MACHINE LEARNING MODEL

important point when you are making machine learning model .

1. Before creating the model , divide the dataset into two parts , one input-Variables and the other Target\_Variable

2. do Train\_test\_split on input-data & target-attribute then data split into train & test data , fit the model on train data then test the model accuracy on test data

3. I used LogisticRegression, RandomForestClassifier, ExtraTreesClassifier, SVC & GradientBoostingClassifier to fit the models and after train the models checking accuracy\_score so the highest accuracy\_score, roc\_auc\_score = 92 percent , obtained from RandomForestClassifier & ExtraTreesClassifier . In machine learning in the classification problem, the RandomForestClassifier accuracy\_score on test data is nore often tthan other algorithms

4. (Accuracy\_score, roc\_auc\_score, roc\_curve, classification\_reports, confusion\_matrix) all these metrics have been used in each model to know which model is performing well

5. applying cross\_validation on all five model then we get maximum cross\_validation\_score from RandomForestClassifier & ExtraTreesClassifier = 89 percent

6. HyperParameter tuning done on ExtraTreesClassifier & RandomForestClassifier using GridSearchCV but didn't much difference in accuracy\_score, roc\_auc\_score of test data . probably because without HyperParameter tunned model the accuracy\_score was 92 percent .

After that plotting roc\_curve on both hyperparameter tunned model we get AUC\_curve = above 90 percent of both model

CONCLUDING REMARKS

In this Rainfall weather forecasting Dataset

. After we done EDA/Data-Preprocessing , built model and checking accuracy of the model . we found that whenever most of the Variables of the dataset were randomly distributed with target , then we can not get maximum accuracy\_score

If most of Variables of the dataset are distributed in such way with Target that we can get a better idea of the relationship between those Variables and the target , then it is more likely that we will get good accuracy

. maximum variables of this dataset are randomly distributed with target so we could not get more accuracy\_score

1. TN(True Negative) means that the model is predicted to predict that it will not rain tomorrow, if the model predicts TRUE then it will be TN . If it predict wrong then it will be FP(False Positive)

2. TP(True Positive) means the model is predicted to predict that it will be rain tomorrow , If the model predicts TRUE then it will be TP . If the model predicts wrong then it will be FN(False Negative)

Trying to minimize FN while creating model on this dataset

dataset result= TN= 1787 FP= 83

FN= 161 TP= 961