Air Pollutant Concentration Prediction over Ahmedabad Using Machine Learning

CSE523 - Machine Learning
Winter Semester 2023
Weekly Report - 25/3/2023

Yash Dahima - AU2129002

The task of ***feature selection*** using PCA and Factor Analysis has been performed this week. The implementation of PCA has already been described in the previous report. The implementation and implications of Factor Analysis are described in this report.

Factor Analysis:

'FactorAnalysis' class of 'sklearn' library has been used to perform factor analysis. The following snapshot of the code depicts the implementation of the factor analysis for the dataset used.

```
import matplotlib as mpl
import pandas as pd, numpy as np
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from sklearn.decomposition import PCA, FactorAnalysis
import matplotlib.pyplot as plt

# Load the dataset
df = pd.read_excel('C:/Users/Yash Dahima/PhD/Course Work/ML/Project/AQI/Datasets/data4.xlsx')
df['datetime'] = pd.to_datetime(df['datetime'])
df = df.set_index('datetime')

# Separate the features and the target variable
X = df.drop('pm2p5', axis=1)
y = df['pm2p5']

# Scale the data
std_scaler = StandardScaler()
#min_max_scaler = MinMaxScaler()
X_scaled_std = std_scaler.fit_transform(X)
#X_scaled_min_max = min_max_scaler.fit_transform(X)


# ---------------------------------- Perform FA ---------------------------------- #

fa = FactorAnalysis(n_components=10)

# Fit and transform the data using PCA
X_fa = fa.fit_transform(X_scaled_std)

fa_loadings = fa.components_
fa_loadings = pd.DataFrame(fa_loadings, columns=X.columns)
```

The following table shows the first 10 factor loadings, which indicate the strength of the relationship between each variable and each factor. Variables with high loadings on a particular factor are more strongly related to that factor, and they are likely to be more important for predicting the outcome variable.

Index	ws	temp	rh	dew_temp	recipitatio	pressure	wv	blh	bcaod550	duaod550	omaod55(	ssaod550	suaod550	aod469	aod550	aod670	aod865	aod1240
0	0.183	0.182	0.362	0.519	0.117	-0.420	0.564	0.080	0.621	0.566	0.658	0.515	0.926	0.997	1.000	0.985	0.925	0.815
1	-0.471	-0.422	0.026	-0.334	-0.002	0.570	-0.267	-0.255	0.608	-0.800	0.730	-0.432	0.215	0.082	-0.021	-0.169	-0.375	-0.569
2	-0.119	0.039	-0.460	-0.381	-0.263	0.336	-0.566	0.026	0.328	0.201	0.167	-0.674	-0.175	0.003	-0.016	-0.034	-0.045	-0.059
3	-0.230	0.256	0.057	0.256	0.071	-0.174	0.205	-0.009	-0.184	-0.000	-0.078	-0.305	0.257	-0.005	-0.013	-0.027	-0.045	-0.082
4	-0.031	0.099	-0.253	-0.148	-0.092	0.043	-0.179	-0.016	0.320	-0.004	-0.005	-0.008	-0.004	0.011	0.002	-0.007	-0.013	-0.018
5	0.074	-0.150	0.120	-0.061	0.265	-0.083	0.089	-0.118	0.007	-0.000	-0.000	-0.001	-0.000	-0.010	-0.000	0.010	0.017	0.015
6	-0.108	-0.330	0.271	0.007	-0.031	0.083	0.003	-0.229	-0.000	0.000	-0.000	0.000	-0.000	-0.001	-0.000	0.000	0.002	-0.016
7	0.032	0.756	-0.455	0.132	-0.008	-0.255	0.152	0.632	-0.000	0.000	-0.000	0.000	-0.000	-0.000	-0.000	0.000	0.001	-0.007
8	0.006	-0.058	-0.507	-0.587	-0.077	0.244	-0.300	0.274	-0.000	0.000	0.000	0.000	0.000	0.000	-0.000	0.000	0.000	-0.000
9	0.213	-0.043	0.123	0.070	0.046	0.055	0.030	0.593	0.000	-0.000	-0.000	-0.000	-0.000	-0.000	0.000	-0.000	0.000	0.000

It can be observed from the table that all the variables except wind speed and precipitation have absolute value of loading > 0.5 for one or more factors. But, it is known that wind speed affects the target variable - particulate concentration, so it can not be ignored. Precipitation can be ignored for developing models as it does not seem to have a significant impact on the target variable.

Among different kinds of AOD values, bcaod550 shows the highest correlation with the target variable. PCA indicates that it also contributes significantly in explaining the variance of the data. Hence, it is chosen to be kept for modeling and other types of AOD can be ignored. Relative Humidity (RH) is derived using temperature and dewpoint temperature, so either RH or dewpoint temperature can be ignored for modeling. It is clear from PCA that RH contributes more in explaining the variation of data in comparison to dewpoint temperature. So, dewpoint temperature is decided to be ignored for modeling.

Finally, the features selected based on PCA and Factor Analysis for developing a model are: wind speed, temperature, relative humidity, pressure, water vapor, boundary layer height, and bcaod550. Model development will be initiated in the next week.