Analysis And Prediction Of Delhi Climate Using Machine Learning

Gaurav Lodhi
MT19063

Kaamran Khan
MT19064

Vedant Desai
MT19074

I. Introduction

Delhi is infamous for air pollution and erratic weather conditions like fog, hail, smog, etc. It causes huge damage to life and resources in significant sectors like agriculture, transport system, infrastructure, etc. And the climatic conditions are only going to get worse in the near future. Thus there is an urgent need for accurate climate analysis and prediction. Unlike typical weather forecasting approaches, which have traditionally relied on physical simulations, we take a data-centric approach using machine learning algorithms to derive our solution

II. Related Work

Since we have created a custom dataset by combining 2 different datasets, no related work will be found on this dataset. But a considerable amount of work has been done separately for Delhi weather data and Delhi pollution data. Some of them are :- "Emission estimates and trends (1990–2000) for megacity Delhi and implications." Gurjar, B.R., Aardenne, J.A. van, Lelieveld, J., Mohan, M. (2004). "Assessment of traffic–generated gaseous and particulate matter emissions and trends over Delhi (2000–2010). Sindhwani, R., Goyal, P. (2014)", "Numerical methods in weather prediction." Elsevier, G. Marchuk. 2012. "Estimation of Air Pollution in Delhi Using Machine Learning Techniques" Srivastava, Chavi and Singh, Shyamli and Singh, Amit (2018).

III. Dataset and Evaluation

We are using two dataset as follows-

A. Dellhi weather dataset

B. Delhi air quality data

Delhi weather dataset is a time-series hourly data that contains 100990 samples along with 20 features ranging from 1997-2016 Delhi air quality dataset is a time-series daily data that contains 8845 samples along with 13 features ranging from 1995-2015. We merged both the datasets based on date to get a resultant dataset with 57561 samples combining both their features. The dataset was divided into training, validation and test set in the ratio 75:15:10 respectively. Upon close inspection we removed certain unimportant features like "agency", etc. We faced issues due to NA values in some columns. So we removed those features which contained more than 15,000 NA values. For the remaining features with NA values we replaced NA with the mean value of that particular feature. We created some new features such as "ni", "si", "spi", "rpi" based on "NO2", "SO2", "SPM", "RSPM" respectively in order to create feature AQI (Air Quality Index). Before sending the data to train we performed one-hot encoding on categorical features. So our final dataset contains 57561 samples along with 83 features. For our regression problem, we will use root mean square error and for our classification problem, we will use accuracy score. Since the data for extreme weather conditions is skewed, we will use precision, recall and roc as our evaluation metric.

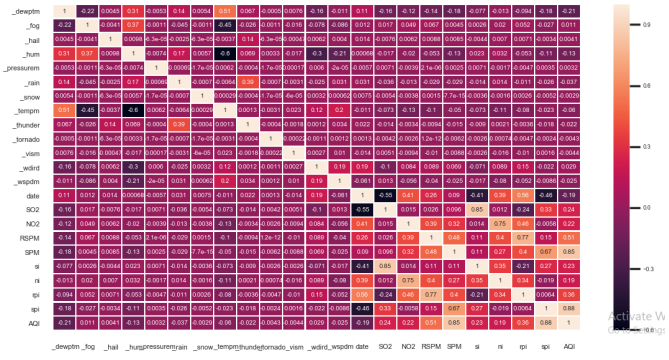

Fig. 1. Heatmap for all the features

IV. ANALYSIS

Before doing model training, we have done a comprehensive EDA (Exploratory Data analysis). This helped in trace out properties of important feature and relationship between those features. For example, a negative correlation was observed between temperature and humidity. Further, more interesting derivations were seen while doing time series analysis on various features. Relationship between each pollutants, increase or decrease of pollutant with respect to time, pollutant presence with respect to the type of area ("Industrial","Residential" ,etc) were some other interesting deductions discerned. Other fascinating derivations were gathered to pluck out a few important inferences like cause of increase in each pollutant level for a historical time-frame, effect of global warming on Delhi, as wind speed increases AQI decreases, AQI is higher in winter months as can be traced from Fig2 and Fig 3.

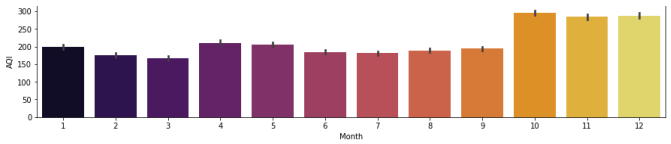

Fig. 2. Barplot of AQI corresponding to months

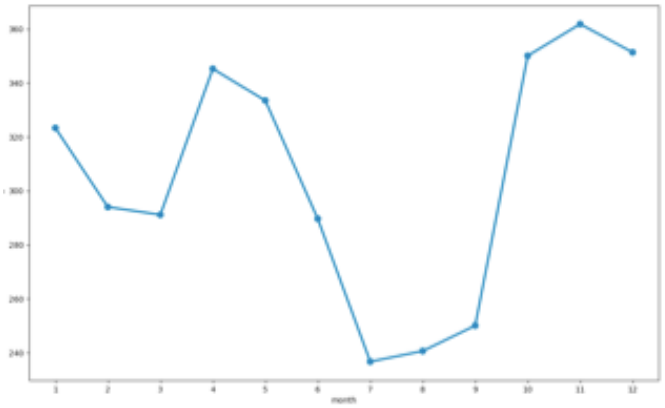

Fig. 3. Pointplot of AQI corresponding to months

For prediction of important continuous variables like 'temperature' and 'AQI', the design choices we made were to adhere to our starting goal of getting the base error rate by applying simple learning algorithms like LinearRegression. So that we can slowly build up from the base error rate later. Then for column 'conditions' having 37 unique values, we have segregated those with extreme weather like 'tornado', 'thunderstorm', 'hail' to '1' and the rest being normal weather as '0'. Again, on this we have applied simple classification algorithms like Logistic regression for prediction of extreme weather conditions to get base accuracy. Also, since it is a time-series data we have also applied ARIMA model on our data to predict SO2 levels in a future time duration.

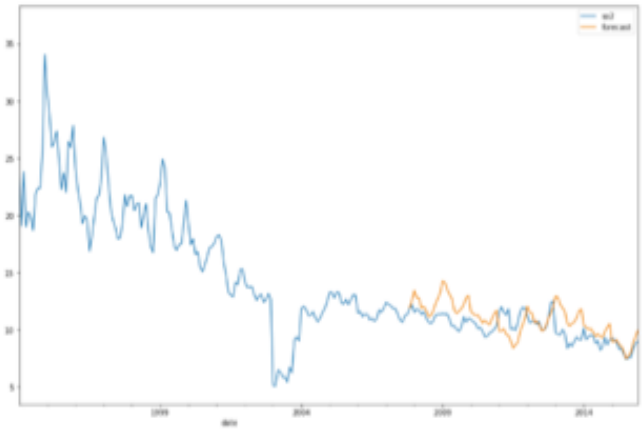

Fig. 4. SO2 prediction using ARIMA model

Despite applying feature selection with the help of EDA, while training we faced problems due to certain unimportant features which were only adding noise and confusing the models. Thus, using sklearn classes like ExtraTreeRegressor and Random Forest Regressor, we derived feature importance of each feature with respect to certain label. From this we selected only those set of features which were deemed important while the rest were removed from training. Along with the above-mentioned predictions our end goal is to develop a prediction model that forecasts all the weather conditions for the next 24 hours, next 3 days, next week alike other commercial weather forecasting services. So, we have performed similar base accuracy and error predictions on a number of labels like 'wind-direction', 'wind-speed', 'humidity, etc. After completing the base case model training, we analysed the results to infer that the data is not linearly separable and so would need a more complex model to properly train it. Also viewing the high error rate of training and testing data, we have to come to a conclusion that the models are under-fitted. But we are optimistic about decreasing the error rates by more feature selection, training a few complex learning algorithms coupled with proper hyper-parameter tuning.

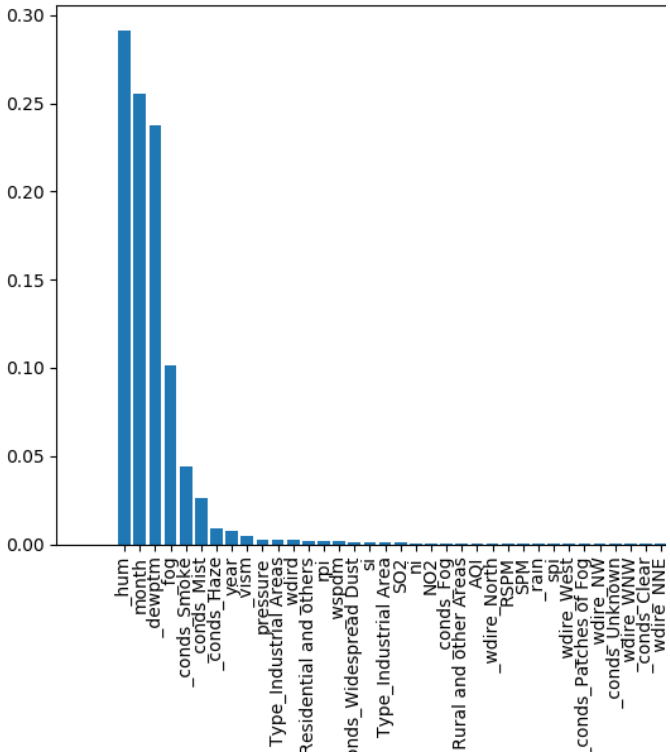

Fig. 5. Label Temperature:- Feature importance vs features

V. RESULTS

We trained our data with label as temperature using Linear regression.We observed testing rmse as 5.06 and train rmse as 2.13.We trained our dataset with label as AQI using linear regression and observed test rmse to be 95.6 and train rmse to be 80. Since our train rmse is 80 which means that we haven't sampled our data very well so in order to reduce the error

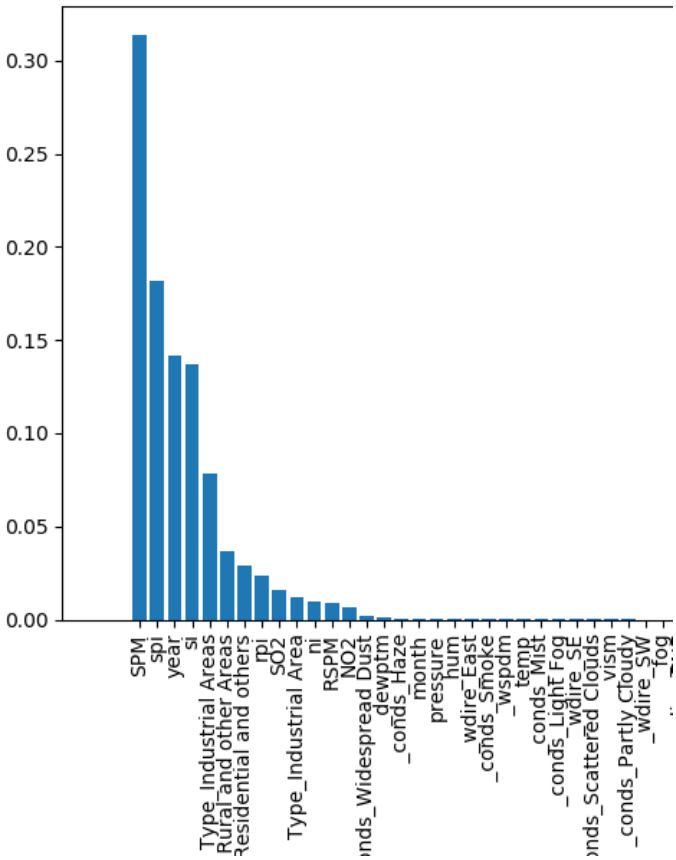

Fig. 6. Label AQI:- Feature importance vs features

we would find the mean and variance from a set of recent pollution data samples and use it to generate pollution data samples for our dataset. We applied simple logistic regression on our dataset in order to predict extreme climatic conditions like 'tornado', 'thunderstorm', 'hail',etc. We observed training accuracy as 78% and testing accuracy as 50%.As there is a large difference in training and testing accuracy so we can say that our model is overfitted.

VI. Future Work

For our regression problems, we will apply advance machine learning techniques like CNN, KNN, Random Forest to predict AQI level , temperature and other labels. And for our classification problems, we will apply deep learning techniques for classifying extreme weather conditions such as thunder, hail etc. For now, we have randomly sampled pollution values for hours in each day. But we plan to make the merged dataset less random by distributing it in accordance with how the values are distributed in real life. We would find the mean and variance from a set of recent pollution data samples and use it to generate pollution data samples for our dataset. Since we are going to generate a complete weather forecasting report, we will add a user-defined evaluation metrics that gives weights to each prediction (temperature, wind direction, wind speed, dew....), then combines them according to certain heuristics to finally generate a single evaluation number. We also plan to perform more feature analysis in future so as to create if possible, some new valuable features formed from combining multiple features, which will help in making important labels more predictable and less noisy. We will also try to deduce trends which will help in giving an abstract view about the weather and pollution for the next 2-3 years. Team member roles for future: - Gaurav will work on applying complex models on classification problems. Vedant and Kaamraan will work on applying deep learning models on regression problems. Kaamraan and Vedant will work on producing results from prediction of extreme weather conditions.