Application of Machine Learning for $PM_{2.5}$ estimation using low-cost sensor network

Prateek Sethi
Virginia Tech
Arlington, VA
prateek20@vt.edu

Kuldeep Dixit
Virginia Tech
Arlington, VA
kdixit@vt.edu

Abstract

Fine particulate matter or $PM_{2.5}$ is known to harm human health. The monitoring of $PM_{2.5}$ is a challenging task due to the high cost of $PM_{2.5}$ monitors. The low-cost sensors are increasingly being employed to have more spatial monitoring coverage but, these sensors are less reliable due to lower accuracy. In this study, we have shown that machine learning models can be employed to predict accurate $PM_{2.5}$ and perform better than the traditional linear regression method used to correct the low-cost sensor data. Most of our machine learning models performed better than the baseline linear regression model used by Environmental Protection Agency's AirNow portal.

1 Introduction

High concentration of $PM_{2.5}$ is known to cause respiratory and cardiovascular health problems. This study[2] finds $PM_{2.5}$ attributed to having caused 8·42 million deaths in 2016. Although there is an urgent need for universal monitoring of $PM_{2.5}$, it is not possible because of the high cost of the monitors. To address this problem, the usage of low-cost sensors is growing. The Purple air sensor has become a popular low-cost sensor. it is increasingly employed to estimate the $PM_{2.5}$ concentration using meteorological data and PM signals from these sensors.

2 Methodology

A popular standard definition of Machine Learning by Tom Mitchell is: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." Machine learning algorithms can make decisions automatically, and these decisions can be improved through experience/data. The logic is not hardcoded but is programmed by the machine itself. This study uses the popular algorithms - CART - Classification and Regression Trees, KRR - Kernel Ridge Regression, KNN - K - Nearest Neighbors, and ENS - Ensemble methods - Bagging, Voting Regressor, Random Forrest.

2.1 Data

In our study we have used two data sets namely the Environmental Protection Agency's PM2.5 dataset and the low cost sensor dataset. We extracted the data from the Purple air sensors which are within 50 meters of the EPA monitor and merged both the EPA data and Purple air data to feed into the models. This process is similar to the baseline study[1] that we have used to compare our machine

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

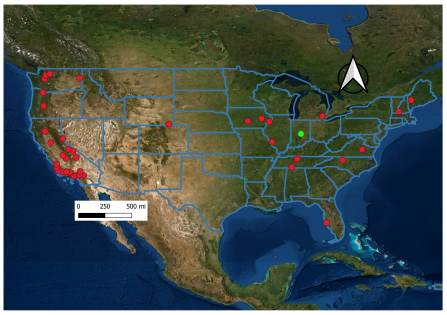

Figure 1: Location of 46 EPA monitors used in our study

learning models. We have used the temperature, humidity and the PM signal from low cost sensor as our features and we have used the EPA monitor's $PM_{2.5}$ concentration as the label. We have removed the invalid values of temperature and humidity and also compared the two channel readings of the purple air sensor and removed the values showing more than 3 percentage deviation. We followed the process of Lu et al[2] here, lastly we used data from 46 EPA monitors and and 96 purple air sensors in our study. The location of EPA monitors used is given in Figure 1.

2.2 Classification and Regression Trees

It is the process of predicting the class of given data points. Classes are sometimes called targets or labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y). In contrast regression analysis consists of an output variable which is continuous or a real value. Regression techniques try to fit the best hyperplane which goes through the data points. The CART algorithm can be used to tackle decision making problems relying on a number of features in a systematic order. CART techniques have been a major contributor to a lot of health studies [7][8]. Different steps in the CART process are explained and carried out in [3][4] gives an example of another use case in the health and wellness industry. CART can handle both classification and regression tasks. This algorithm uses a metric named gini index to create decision points for classification tasks

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{1}$$

2.3 Kernel Ridge Regression

Kernel Ridge Regression, is a very simple special case of Support Vector Regression. The main formula of the method is identical to a formula in Bayesian statistics, but Kernel Ridge Regression has performance guarantees that have nothing to do with Bayesian assumptions.[3] "Kernel Ridge Regression" was coined in 2000 by Cristianini and Shawe-Taylor [7] to refer to a simplified version of Support Vector Regression; this was an adaptation of the earlier "ridge regression in dual variables"

$$\begin{aligned} \min \quad & \|w\|^2 + C \sum_{t=1}^{T} ((\xi_t)^k + (\xi_t^{'})^k) \\ \text{s.t.} \quad & (w.x_t + b) - y_t <= \epsilon + \xi_t, t = 1, ..., T \\ & y_t - (w.x_t + b) <= \epsilon + \xi_t, t = 1, ..., T \end{aligned} \tag{2}$$

where $(x^t; y^t) \in R_n X R$ are the training examples, w is the weight vector, b is the bias term, $\xi_t^{'} and \xi_t^{'}$ are the slack variables, and T is the size of the training set $\epsilon; C > 0$ and k eta 1,2 are the parameters. Simplify the problem by ignoring the bias term b (it can be partially recovered by adding a dummy

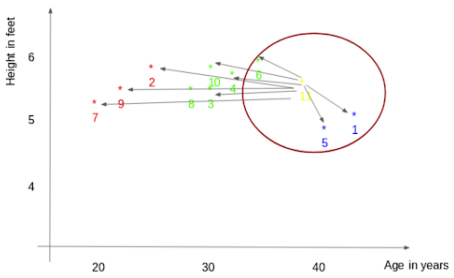

Figure 2: KNN Algorithm.

Table 1: Different distance metric for KNN

Metric	Formula
Euclidean Distance:	$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$(4)
Manhattan Distance:	$\sum_{i=1}^{k} \lvert x_i - y_i \rvert$(5)
Hamming Distance	$\min \quad D_H = \sum_{i=1}^{k} \lvert x_i - y_i \rvert$ (6) $x = y \Rightarrow D = 0$ $x \neq y \Rightarrow D = 1$

attribute 1 to all x_t), setting $\epsilon := 0$, and setting k := 2. The optimization problem becomes

$$\min \quad a\|w\|^2 + \sum_{t=1}^{T}(y_t - w.x_t)^2 \tag{3}$$

(where $a := \frac{1}{c}$), the usual Ridge Regression problem. And Vapnik's usual method ([6], Sect. 11.3.2) then gives the prediction.

In KRR, as in Support Vector Regression in general, the kernel is not supposed to reflect any knowledge or beliefs about reality, and the usual approach is pragmatic: one consults standard libraries of kernels and uses whatever works.[3]

2.4 K Nearest Neighbor Regression

The KNN algorithm uses 'feature similarity' to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set.

Consider the figure 3 from [8], to make a prediction the distance to the existing data points is calculated and based on the K closest data points a prediction is made. For the calculation of distance different methods can be used depending on the use case. Some of these are indicated in Table 1.

2.5 Ensemble Methods

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. The most commonly used methods are the voting and the averaging ensemble methods. Where multiple classification or regression models are created using a dataset. This can be done by using the same algorithm or by using different algorithms. The prediction of these models can be combined in various ways to come to a final conclusion. Weights can be added to improve the quality of the model and hence use Weighted Voting or Weighted Averaging. It is very important that the ensemble classifier results in a prediction which is better than its individual classifiers. A necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of

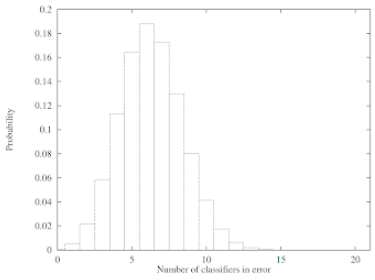

Figure 3: The Probability that exactly 1 (of 21) hypothesis will make an error, assuming each hypothesis has an error rate of 0.3and makes its error independently of the other

its individual members is if the classifiers are accurate and diverse [9]. To see why accuracy and diversity are good, imagine that we have an ensemble of three classifiers: h1, h2, h3 and consider a new case x. If the three classifiers are identical (i.e., not diverse), then when h1(x) is wrong, h2(x) and h3(x) will also be wrong. However, if the errors made by the classifiers are uncorrelated, then when h1(x) is wrong, h2(x) and h3(x) may be correct, so that a majority vote will correctly classify x. More precisely, if the error rates of L hypotheses h are all equal to $p < 1/2$ and if the errors are independent, then the probability that the majority vote will be wrong will be the area under the binomial distribution where more then L/2 hypotheses are wrong. Figure 1 shows this for a simulated ensemble of 21 hypotheses, each having an error rate of 0.3. The area under the curve for 11 or more hypotheses being simultaneously wrong is 0.026, which is much less than the error rate of the individual hypotheses. Additionally [7] discusses how ensemble methods result in highly accurate classifiers by combining less accurate ones. Also it discusses the three fundamental reasons why ensemble methods are able to out-perform any single classifier within the ensemble. For our experimentation we considered the following methods: Voting Regression, Bootstrap Aggregation (Bagging) and Random Forest Regression

3 Results and Observations

To compare the performance of the linear regression model[1] with our machine learning models we considered five metrics, the mean squared error, mean absolute error, r2 distance, fit time, and the score time. For each of these metrics, we performed the 10 Fold cross-validation to get a robust performance estimation. The results and comparison is shown in Table 1 and Figure 4. Table 1 shows the median values obtained from 10-fold cross-validation.

Mean Squared Error From Table 1 we can observe that almost all the algorithms perform comparable to the linear regression model used in the reference study.

Mean Absolute Error According to this metric, the performance of most of the models is better than the linear regression model. As can be observed that the Bagging using KNN and the Voting regression outperform the linear regression model.

Fit time and Score time This metric was not available for the base model but, we have used this to compare the machine learning models that we implemented in this study. Most machine learning models took almost equal time to fit and score but, the KNN and the voting regression took comparatively more time.

3.1 Model Explination

We also used SHapley Additive exPlanations or, SHAP to explain our models. We applied SHAP method on the best performing voting regression method. The most important feature for prediction

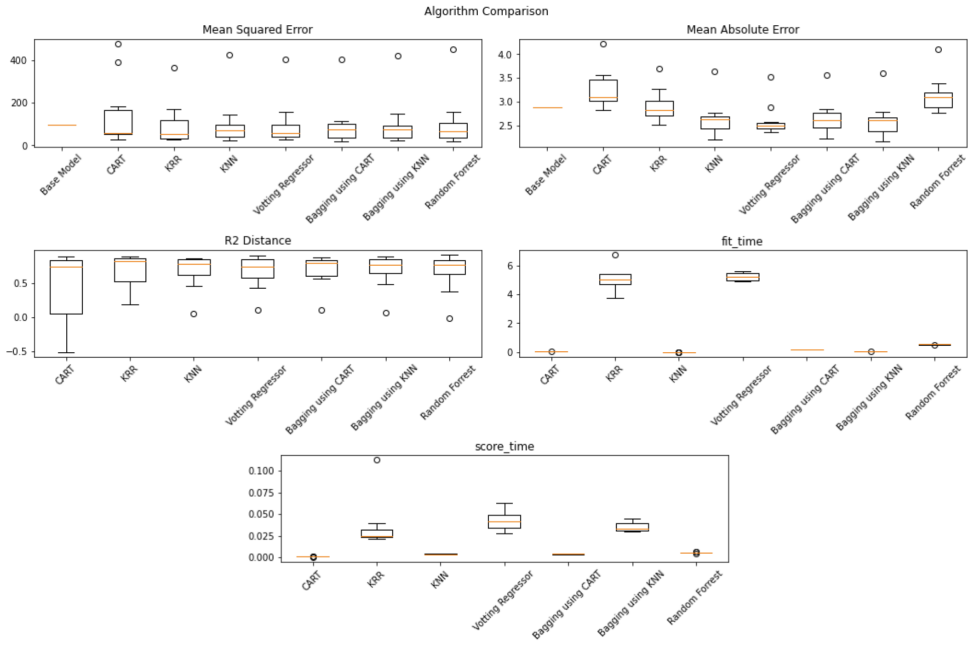

Figure 4: ML models comparison

Table 2: The comparison of models

Metric	Base Model	CART	KRR	KNN	Votting Regressor	Bagging using CART	Bagging using KNN	Random Forrest
Mean Squared Error	96.44	107.96	54.3	69.06	59.88	73.41	74.1	66.99
Mean Absolute Error	2.88	3.13	2.82	2.63	2.49	2.61	2.6	3.09
R^2	0.69	0.72	0.82	0.78	0.79	0.8	0.77	0.78
Fit Time	NA	0.06	6.36	0.01	8.35	0.54	0.1	1.6
Score Time	NA	0	0.04	0.02	0.05	0.01	0.09	0.02

$PM_{2.5}$ turns out to be low cost sensor's PM signal followed by humidity and temperature. Figure 5 shows the mean SHAP values for the voting regression model.

4 Conclusion

In our study most of the machine learning algorithms showed either comparable or better performance than the baseline linear regression model. The metrics do not indicate a clear winner but, according to the metric mean absolute error, voting regression outperforms,and according to mean squared error and R^2 Kernal Ridge Regression (KRR) outperforms all other models.

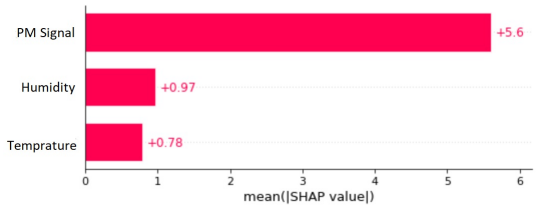

Figure 5: The average SHAP values for the voting regression model

References

[1] Barkjohn, K. K., Gantt, B., and Clements, A. L.: Development and application of a United States-wide correction for PM2.5 data collected with the PurpleAir sensor, Atmos. Meas. Tech., 14, 4617–4637, https://doi.org/10.5194/amt-14-4617-2021, 2021.

[2] Lu T, Marshall JD, Zhang W, Hystad P, Kim SY, Bechle MJ, Demuzere M, Hankey S. National Empirical Models of Air Pollution Using Microscale Measures of the Urban Environment. Environ Sci Technol. 2021 Nov 16;55(22):15519-15530. doi: 10.1021/acs.est.1c04047. Epub 2021 Nov 5. PMID: 34739226.

[3] Roger J. Lewis, M.D., Ph.D. Department of Emergency Medicine Harbor-UCLA Medical Center Torrance, California - *"An Introduction to Classification and Regression Tree (CART) Analysis"* Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California.

[4] N. Speybroeck - *"Classification and regression trees"* - Received: 9 February 2011 / Revised: 2 September 2011 / Accepted: 8 October 2011 / Published online: 21 October 2011.

[5] Festschrift in Honor of Vladimir N. Vapnik - *"Empirical Inference"*

[6] Vapnik, V.N.: *Statistical Learning Theory.* Wiley, New York (1998)

[7] Cristianini, N., Shawe-Taylor, J.: *An Introduction to Support Vector Machines and Other Kernel-Based Methods.* Cambridge University Press, Cambridge (2000)

[8] *A Practical Introduction to K-Nearest Neighbors Algorithm for Regression (with Python code)* - 'Aishwarya Singh — August 22, 2018' - https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/

[9] Thomas G. Dietterich, Oregon State University, Corvallis, Oregon, USA - *'Ensemble Methods in Machine Learning'*