Khulna University Studies Special Issue ICSTEM4IR: 676-689, 2022

ORIGINAL ARTICLE **SCIENCE & ENGINEERING**

SMART ENVIRONMENT INDEX PREDICTION OF SMART CITY USING POLYNOMIAL REGRESSION

S.M. Mohidul Islam*, Kamrul Hasan Talukder

Computer Science and Engineering Discipline, Khulna University, Khulna-9208, Bangladesh

KUS: ICSTEM4IR-22/0130

Manuscript submitted: June 28, 2022 Accepted: September 28, 2022

Abstract

Smart Environment refers to environment where pollution is detected, predicted, classified and solved using smart tools and technology such as using Internet of Things (IoT) sensors, cloud service, and machine learning algorithms. Episodic environment index prediction allows governments and local city agency to detect pollution in environment at a premature stage to initiate proper steps. Advancements in machine learning, sensor, and camera technology have endorsed us to do that, which can benefit millions of cities and its citizens indeed. This paper describes a new method for smart environment index prediction based on index of five others smart city components using polynomial regression. This research mainly works with the Smart Cities Index dataset. We attempted numerous steps to select feature and to filter its contents to make them more apposite for selected regression algorithm. Our approach unites several feature selection techniques, outlier detection technique, polynomial degree selection technique, and random state selection technique, which outcomes in a reduction of Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to 0.0000000140, 0.000000000000000322, and 0.0000000179 respectively and rise R^2 to 1.0 for prediction. This method can be a valuable tool for smart environment index prediction from smartness of various components of a city, thus drastically reducing survey or other related works.

Keywords: Smart environment index, smart city components, polynomial regression, feature selection, hyper-tuning, regression error

Introduction

Day by day, cities are becoming the most demanding place of residence throughout the world. United Nations (UN) exposed that by 2050, 68% of the world's people will live in urban cities (Drewil & Al-Bahadili, 2021). For upcoming world, these cities should be smart enough for sustainability. By attaining the goal of smart city, our life is becoming smarter for recently being advanced smart technologies such as Internet of Things (IoT), Artificial Intelligence (AI), big data analytics, cloud computing, etc. (Akhilesh, 2020). There is no universal consensus of defining smart city and there is no standard concept of a smart city. Many experts have provided their viewpoint, thus many models and standards are branded (Singh & Kumar, 2020), (TEIXEIRA et al.,

*Corresponding author: <mohid@cse.ku.ac.bd>
DOI: https://doi.org/10.53808/KUS.2022.ICSTEM4IR.0130-se

2020). Among them, five models are well-known, they are: Ruddolf Giffinger model, Boyd Cohen model, Charalampos Alexopoulos model, ISO 37120:2014, and Belo Horizonte's Urban Life Quality Index (IQVU) (TEIXEIRA et al., 2020). According to both Giffinger (Giffinger et al., 2007) and Boyd Cohen (Benamrou et al., 2016) models, a city is a smart city which contains six main components including smart mobility, smart governance, smart environment, smart economy, smart life, and smart people. Charalampos Alexopoulos model (Alexopoulos et al., 2019) defines smart city as a city which contains the following components as smart: transportation and mobility, environment, tourism and culture, health, waste management and water resources, energy and sustainable development, ICT infrastructure, economy and development, security and e-government. ISO 37120:2014 outlines that the components of smart cities are economy, education, energy, environment, finance, fire, governance, health, recreation, safety, shelter, solid waste, telecommunication and innovation, transportation, urban planning, wastewater and water and sanitation (International Organization for Standardization, 2014). Belo Horizonte's Urban Life Quality Index (IQVU) mentions supply, culture, education, sports, shelter, urban infrastructure, environment, health, urban services and urban safety are the components of smart city (Prefeitura Belo Horizonte, 2020). Moreover, paper of (Kulkarni & Akhilesh, 2020) indicates four most auspicious dimensions of smart city which are smart health, smart environment, smart transportation, and smart governance. According to a study (Tran Thi Hoang et al., 2019), in the investigation of a 10 years period of work from 2008 to 2018, the scientists revealed that smart cities might have some characteristics in common which are smart mobility, smart governance, smart environment, smart energy, smart economy, smart people, and smart living.

According to all models and standards of smart city described above, we see that all of the above models and standards included smart environment as a components of smart city. That means, smart environment is one of the most important pillars of smart cities. Smart environment means warns with the harshness that may cause due to environment alteration either due to natural threats so that safety measures can take in advance or assessment of air quality (Singh & Kumar, 2020). Smart environments are enabled by the use of recent advances in science and technology, especially by the use of IoT-based wireless sensor networks, artificial intelligence and machine learning. For example, environment sensors such as pollution, temperature, wind, humidity, etc. sensors can be installed across a city to acquire an actual outline of the weather condition as well as foresee weather changes. Such monitoring schemes usually gather data on an hourly basis and evaluate it to produce an outline. The procedure of using IoT in such an approach produces a huge quantity of data that needs to be processed in instantaneous on a nearly hourly basis. Moreover, to predict future environmental scenarios and visualize existing situations, historical climate and mobility data are also used (Soomro et al., 2019). Machine learning is considerably the most common method for accomplishing these types of analyses and predictions (Soomro et al., 2019).

At present, environment monitoring (EM) has become a smart environment monitoring system; because the aforementioned technologies enable EM methods for finest controlling of pollution and other unwanted effects to monitor the issues influencing the environment more precisely (Ullo & Sinha, 2020), (Bhoomika et al., 2016). Environmental pollution has become a severe health apprehension worldwide, as the pollution types and process is mounting unaccountably. Two of the world's worst pollution problems are urban air quality and indoor air pollution (Pure Earth, n.d.). Air pollution can cause both short-term and long-term health effects. It also lessens the ozone layer as well as formats acid rain, haze, etc. (Kang et al., 2018).

Recent progresses in data science and machine learning have conveyed an extensive impact on various sciences and engineering areas. Smart Environment is one of those areas that have advantaged the most due to these developments. Various new and improved machine learning algorithms allowed researchers to detect and classify the pollution. Researchers are also doing the same for smart environment index prediction as well. Yet, there is no robust machine learning approach conveyed, that can be used in addressing the challenges of the smart environment regardless of the purpose of the monitoring and control, types of data, and types of sensors used (Ullo & Sinha, 2020).

This paper explains the research conducted using important feature selection and polynomial regression algorithm for index prediction of important dimension of smart city, namely smart environment index. This index is predicted from indices of five other dimensions of smart city namely smart transportation/mobility index, smart management/government index, smart economy index, smart people/citizen index, and smart life/living index. As said before, in the progressions of turning cities into smart cities, we can find several representative features. Although each feature has the significance of itself to smart cities; they all are inter-related to each other i.e. each feature has direct or indirect influence on others. This carried out the research question or problem, is there any relationship of being smart the environment with smartness of other five major components of a city? Smart environment index has a relationship with indices of other smart city components, as proposed in this study. To answer the outlined research question, after selecting apposite features and processing its contents, the polynomial regression algorithm was hyper-tuned to attain the best result on this specific application. The assimilated outcomes assert that the proposed approach can effectively predict the smart environment index from indices of other major smart city components. We have provided required tables and figures while unfolding its steps and performance for apposite clarification.

To describe this, the rest of the paper is organized as follows. Section II describes the materials and methods which includes employed dataset description, pre-processing of the candidate dataset, most relevant features selection, and developing and testing the model using polynomial regression. Section III outlines the prediction results and discussion on it. Finally, concluding remarks on the proposed study and scope of future research is drawn in section IV.

Materials and Methods

The workflow of the proposed method is partitioned into four main sets of operations: data acquisition, data pre-processing, most relevant features selection, and optimized smart environment index prediction model creation and testing. We employed multiple techniques for selecting relevant and important features and one regression method named polynomial regression for smart environment index prediction. The intact workflow is shown in Figure 1 and narrated richly in the following subsections.

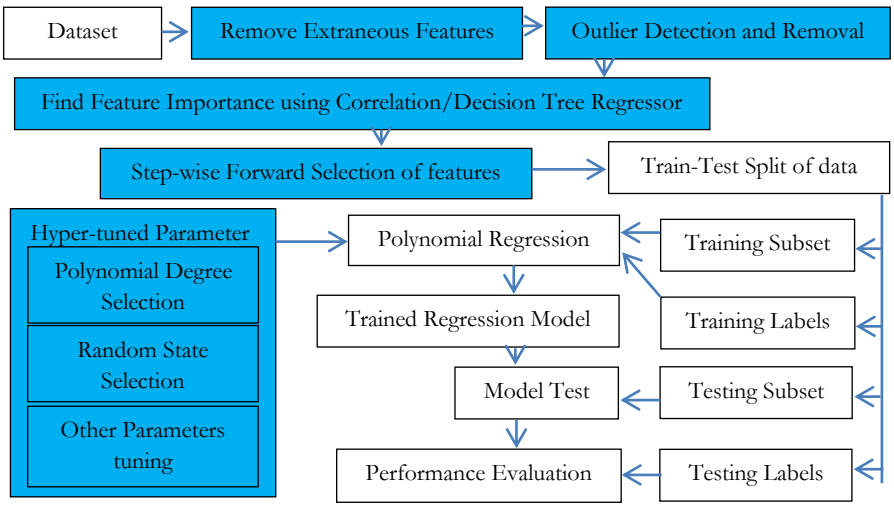

Figure 1. The intact workflow of the proposed smart environment index prediction model.

Employed dataset description

This study used the Smart Cities Index (SCI) dataset jointly produced by Smart City Observatory (SCO) of International Institute for Management Development (IMD), Geneva, Switzerland and Singapore University of Technology and Design (SUTD) (Smart City Observatory, n.d.). Here we used the 2nd version of the SCI dataset collected from Kaggle (Smart Cities Index Datasets, 2021).

Smart city index methodology

The IMD-SUTD constructed this edition of SCI by using total 12240 surveys in 102 cities worldwide. They did this by assessing the perceptions of randomly chosen 120 residents in each city on structures and technology application issues available to them in their cities. Each survey had 40 questions where 36 questions were divided equally between two application issues: structures and technologies. Both application issues were evaluated over five key areas: mobility, health and safety, activities, opportunities for work and school, and governance. Questions of structure issue were about the existing infrastructure of the cities and those of technology issue were about the technological provisions and services available to the inhabitants of the cities. Three of the rest 4 questions were about assessing attitudes on three key privacy aspects, they are: willingness to confess personal data in order to improve traffic congestion, comfortableness with face recognition technologies to lower crime, and whether online information increased their trust in authorities. The last one question were about choosing 5 priority areas they perceived as the most urgent for their city from 15 possible alternatives, which are: affordable housing, fulfilling employment, unemployment, health services, road congestion, air pollution, green spaces, public transport, school education, recycling, basic amenities, citizen engagement, security, social mobility, and corruption. From weighted value of those survey results, they calculated the smart mobility, smart environment, smart government, smart people, and smart living indices of each city. The total score for a city is the average of the scores of structures and technologies and it is the value of smart city index. They also followed the United Nations Human Development Index for calculating smart city index (Smart City Index Methodology, n.d.).

Contents of the dataset

The dataset contains 102 samples that mean it contains smart city index data of 102 cities of 36 countries. The dataset contains total 10 attributes: *Id*, *City*, *Country*, *Smart_Mobility*, *Smart_Environment*, *Smart_Government*, *Smart_Economy*, *Smart_People*, *Smart_Living*, *SmartCity_Index*, and *SmartCity_Index_relative_Edmonton*. In our research, we have used 6 of them, they are: Smart Mobility index (*Smart_Mobility*), Smart Environment index (*Smart_Environment*), Smart Government index (*Smart_Government*), Smart Economy index (*Smart_Economy*), Smart People index (*Smart_People*), and Smart Living index (*Smart_Living*), where *Smart_Environment* is used as target feature and others are used as descriptive features. This is because this study focuses on research question: how much the environment of a city is smart, if the city's other five major components are smart? That means, our research objective is to predict smart environment index given that smart mobility index, smart government index, smart economy index, smart people index, and smart living index of any city. Some samples from the dataset are shown in Figure 2, where each entity shows the index values of six major components of a city.

The line plot (Figure 3) of the dataset shows that there is a vital relationship between six major components of a smart city. Our goal is to predict the smart environment index by mining this data relationship with other five smart components using polynomial regression algorithm.

Preprocessing of the candidate dataset

The data contained by the SCI dataset is very convenient for the required prediction. However, the dataset contains outlier data which is need to remove before forwarding to next steps. This is because outlier data has an intense influence on the performance of polynomial regression algorithm that we have used as model in this study. In the following, we have described how we dealt with this issue.

DOI: https://doi.org/10.53808/KUS.2022.ICSTEM4IR.0130-se

	Smart_Mobility	Smart_Environment	Smart_Government	Smart_Economy	Smart_People	Smart_Living
0	6480	6512	7516	4565	8618	9090
1	7097	6876	7350	4905	8050	9090
2	7540	5558	8528	8095	7098	7280
3	7490	7920	8726	5580	5780	7200
4	6122	7692	8354	4330	6743	7730
...	...	...	...	...	...	...
97	4152	4584	4616	7380	3745	4330
98	7610	2998	2806	4905	5183	1980
99	4588	2908	3622	4515	5390	4100
100	6675	4052	5946	8022	6424	8657
101	5801	4499	6396	8022	6200	8141

Figure 2. Samples from dataset.

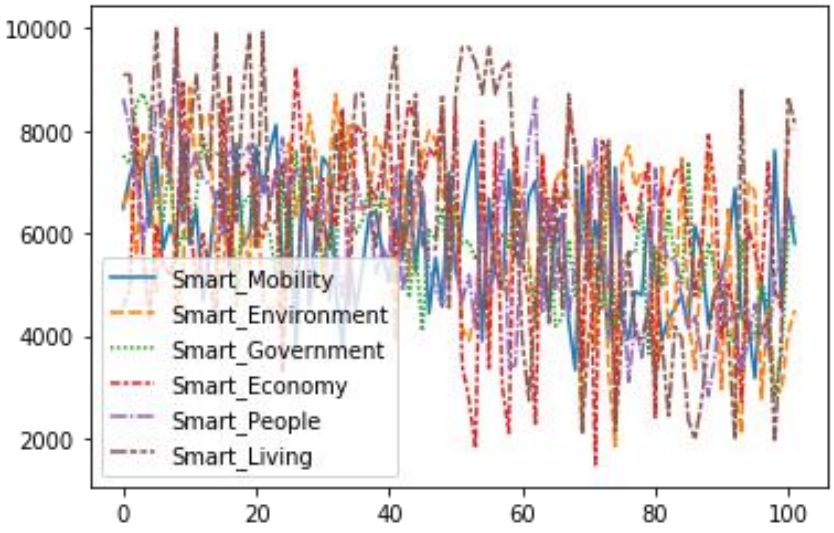

Figure 3. Line plots of features of the dataset. *X*-axis represents the 102 instances (cities) and *Y*-axis represents the index value of each major components of a city.

Dealing with outliers

To know, whether the dataset contains outlier data, we analyze boxplot of all features of the dataset. Boxplot of five-number summary (minimum, first Quartile (25 percentile data), Median, third quartile (75 percentile data), and maximum) is a popular way of visualizing data distribution (Han et al., 2011). The boxplots are shown in Figure 4. In boxplot, two whiskers outside the box in two ends represent minimum and maximum data. Two ends of the box represent the first and third quartiles respectively, whereas the line within the box

represents the median data. The data which is fallen at least 1.5 × (third quartile – first quartile) above the third quartile or below the first quartile is the outlier data and is plotted individually in the boxplot.

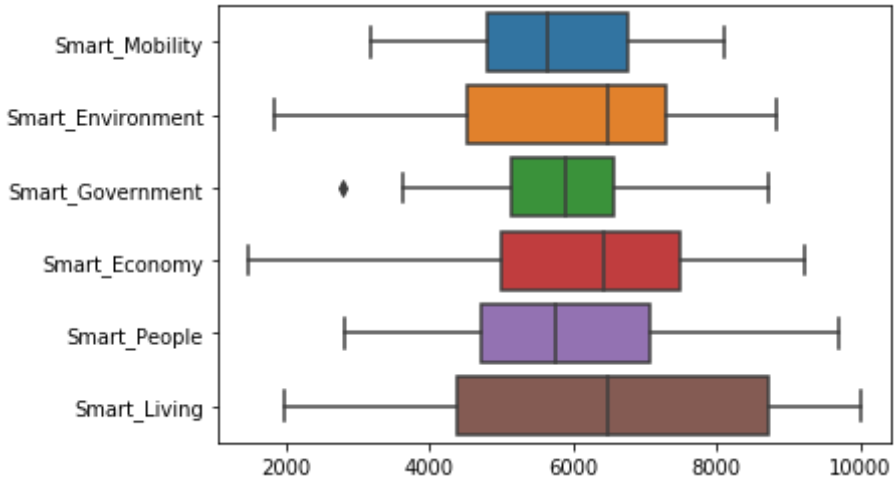

Figure 4. Boxplot of the features of SCI dataset.

From the boxplots, we see that *Smart_Government* feature contains such individual plot, which is an outlier data and the instance that contains this outlier data is removed and excluded from this study.

Most relevant features selection

From Figure 3, we saw that there is vital relationship among features. But it is observed that some features have strong correlation than others with smart environment index (target feature). That means, the features which have weak correlation with the target, can be omitted because those features contain less or no information that is useful for making prediction. Rather using these less important features can reduce the performance of the model. So finding the most relevant features for required prediction is an important task. This is dimensionality reduction. Using only relevant features for learning as well as prediction reduce complexity, reduce overfitting problem, and also enhance performance of the model. Numerous feature selection approaches are available that work based on different ideologies. In this study, we have used two different approaches for important feature selection, they are: feature subset selection (i) using correlation and (ii) using Decision Tree Regressor. These approaches are described in the below.

Feature subset selection using correlation

The linear relationship between two features can easily be found using Pearson's correlation. For two features A and B, the Pearson's correlation, $r_{A,B}$, can be found as (Han et al., 2011):

$$r_{A,B} = \frac{\sum_{i=1}^{n}\left(a_i - \bar{A}\right)\left(b_i - \bar{B}\right)}{n\sigma_A\sigma_B} \quad (1)$$

where n is the number of population instances, $\bar{A}$ and $\bar{B}$ are the respective population means of A and B, a_i and b_i are the respective values of A and B in instance/tuple i, σ_A and σ_B are the respective population standard deviation of A and B. If $r_{A,B} < > 0$ then A and B are correlated, otherwise (if $r_{A,B} = 0$) there is no correlation between them. The value $r_{A,B} < 0$ represents negative correlation, i.e. when value in A is increased, value in B is

decreased and vice versa. Similarly, the value $r_{A,B} > 0$ represents positive correlation, i.e. A's value decreases as B's and vice versa. The correlation of target feature with all features for our selected SCI dataset (dataset of 6 attributes) is shown below (Table 1).

Table 1. Correlation value of smart environment index with all features of the dataset

	Smart_ Mobility	Smart_ Environment	Smart_ Government	Smart_ Economy	Smart_ People	Smart_ Living
Smart_Environment	-0.2202	1.0000	0.2862	0.4545	-0.0304	0.1548

Figure 5 shows the heatmap of this correlation, where the sub parts in horizontal direction shows the correlation value of target feature (smart environment index) with smart mobility index, itself, smart government index, smart economy index, smart people index, and smart living index respectively.

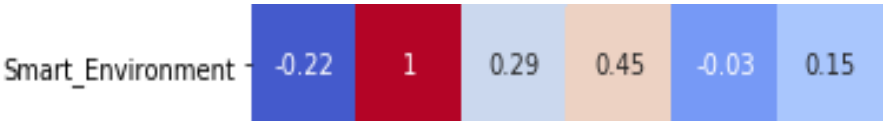

Figure 5. Heatmap to show correlation of target features with all features of the dataset.

From the table and figure above, we see that the importance of the descriptive features (i.e. excluding target feature) follow the following sequence: *Smart_Economy*, *Smart_Government*, *Smart_Living*, *Smart_People*, and *Smart_Mobility*. This sequence of feature importance helps us to select subset of the descriptive features for creating and testing the prediction model. To select the subset from this features importance sequence, we used step-wise forward selection heuristic method (Han et al., 2011). The steps of the method are as follows:

(i) Starts with an empty set of features as the selected set, S. i.e. $S = \{\}$

(ii) The best single-feature is added first. So $S = \{Smart\_Economy\}$

(iii) Create and test the model (in the way described in section 2.4 below) using S and check whether $R^2 = 1.0$. if yes, this is the selected subset and terminate, otherwise go to step 4

(iv) Next-best feature is added, So $S = \{Smart_Economy, ...\}$and go to step 3.

(v) It is significant to identify how well the relationship between the values of the descriptive features and target feature is. The polynomial regression model cannot be used to predict anything, if there is no relationship. This relationship is measured with a value called the R-squared (R^2). The R^2 value ranges from 0 to 1, where 0 means no relationship, and 1 means 100% related (Machine Learning – Polynomial Regression, n.d.). Here, we add and select all those features until we get this R-squared value = 1.0. Using the described method, we get the selected subset of four features (*Smart_Economy*, *Smart_Government*, *Smart_Living*, and *Smart_People*) and we used this subset of features instead of using all features for developing the prediction model.

Feature subset selection using decision tree regressor

Decision tree algorithm helps us to identify the most important features. Here, we have used decision tree regressor instead of decision tree classifier because this study involves a regression problem. In decision tree, for target prediction using which features reduce the impurity most is calculated and gives importance to those features. The feature importance weights of descriptive features of our selected SCI dataset (highest value represents highest importance) are shown below (Table 2).

Table 2. Value of feature importance of all descriptive features of the dataset for target prediction

Smart_ Mobility	Smart_ Government	Smart_ Economy	Smart_ People	Smart_ Living
0.0869921	0.10913775	0.41562049	0.04603176	0.3422179

Figure 6 shows the relative importance of those features by using decision tree regressor.

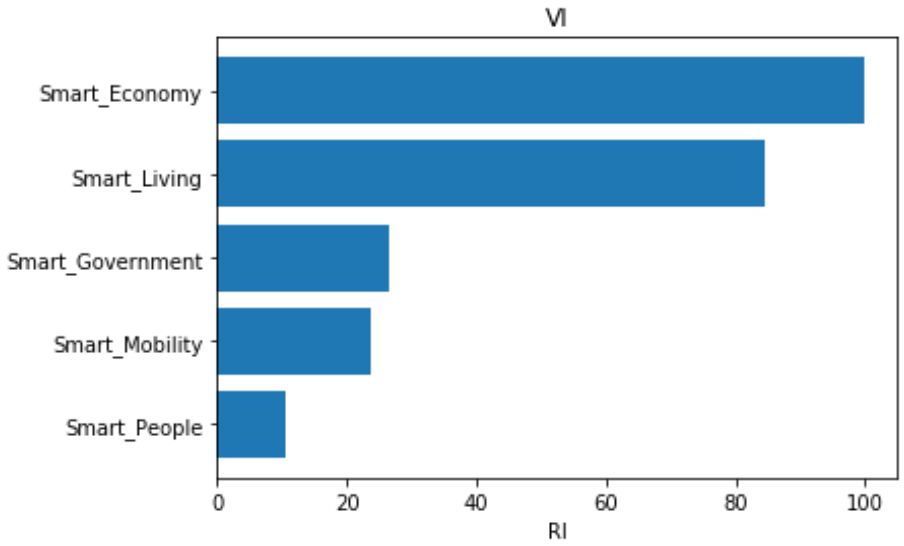

Figure 6. Visualization of relative importance of all descriptive features of the dataset.

From the table and figure above, we see that the importance of descriptive features follow the following sequence: *Smart_Economy*, *Smart_Living*, *Smart_Government*, *Smart_Mobility*, and *Smart_People*. This sequence of feature importance is different from the sequence selected by Pearson's correlation, described above. That means, different approaches deliver different sequences of feature importance. To select the subset from this features importance sequence, we used step-wise forward selection heuristic method as described before. Using the mentioned method, we get the four features selected (*Smart_Economy*, *Smart_Living*, *Smart_Government*, and *Smart_Mobility*) and we also used these subset of features instead of using all features for developing the prediction model.

Learning and prediction using polynomial regression

In this work, we have used polynomial regression model for learning and prediction. 'Poly' means 'many' and 'nomial' means 'parts' or 'terms' or 'names' (Polynomial Regression in Python using Scikit-learn, 2021). We have used polynomial regression instead of linear regression because the relationship between target and descriptive features in SCI dataset is not linear. This non-linearity of this dataset can be seen from Figure 3 above. Also, pair-wise (each of the descriptive features and target feature) regression plot in Figure 7 shows that none of them has linear best-fit line.

The non-linear dataset can be fitted by polynomial regression model. Like linear regression, polynomial regression uses the relationship between the descriptive feature(s) and target feature to find the best-fit line through the data points (Machine Learning – Polynomial Regression, n.d.) but polynomial provides the best approximation of this relationship. An *n*th degree polynomial regression can be expressed as follows (Polynomial Regression, 2022):

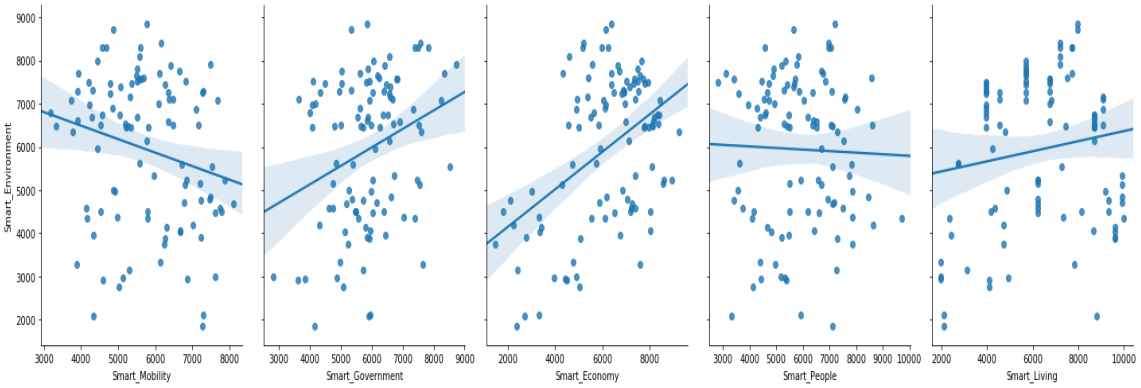

Figure 7. Pair-wise regression plot of the dataset.

$$y = b_0 + b_1 x + b_2 x^2 + \ldots + b_m x^n \quad (2)$$

In the above equation, y is the output of regression, x is the feature, b_0 is the y-axis intercept, b_1,....,b_m are the coefficients which are unknown and the model will estimate them when trained on the available descriptive and target feature values of the dataset, n is the degree of polynomial i.e. the highest power (largest exponent) in the polynomial.

Degree of the polynomial features

The degree of the polynomial is picked based on the relationship between target and descriptor. The 1-degree polynomial is simply a linear regression; hence, the value of degree, n, must be greater than 1. The complexity of the model increases with the increasing degree of the polynomial. So, n must be chosen precisely. If a lower value of degree is chosen then the model will unable to fit the data properly and if higher, the model will overfit the data (Introduction to Polynomial Regression, 2020). To describe this issue, we figure out several regression plots of the pair (*Smart_Mobility* and target feature) for various degrees of polynomial in Figure 8. Figure 8(a) to 8(i) shows the regression plots for degrees 2 to 10 sequentially. From the figure we see that for degrees 2, 3, 4, and 5, the line underfit the data whereas for degrees 7, 8, 9, and 10, the line overfit the data gradually. The regression line for degree 6 shows the balanced fit where both underfitting and overfitting problems are ignored. We also see the similar characteristics for others descriptive features in the dataset. Therefore, we can get the Goldilocks model by using a polynomial regression of degree 6. Considering this issue, we have used a 6-degree polynomial for our application.

Transform and fit

The execution of polynomial regression is a two-step process (Introduction to Polynomial Regression, 2020): First, transform the data/feature matrix into 6-degree polynomial feature matrix and then use linear regression to fit the parameters. That means, we'll use polynomial feature together with linear regression. Figure 9 shows this two-step process.

Results and Discussion

As stated before, we used Smart Cities Index dataset for our smart environment index prediction. 25% of data of this dataset is used for testing or prediction. For this train and test splitting, we have used random state parameter, *random_state*= 42. This is selected randomly but once selected it is fixed for whole experiment. If we train our regression model with the dataset without specifying the random state value, the system will use a random state that is generated internally. So, when we run the program multiple times we might see different

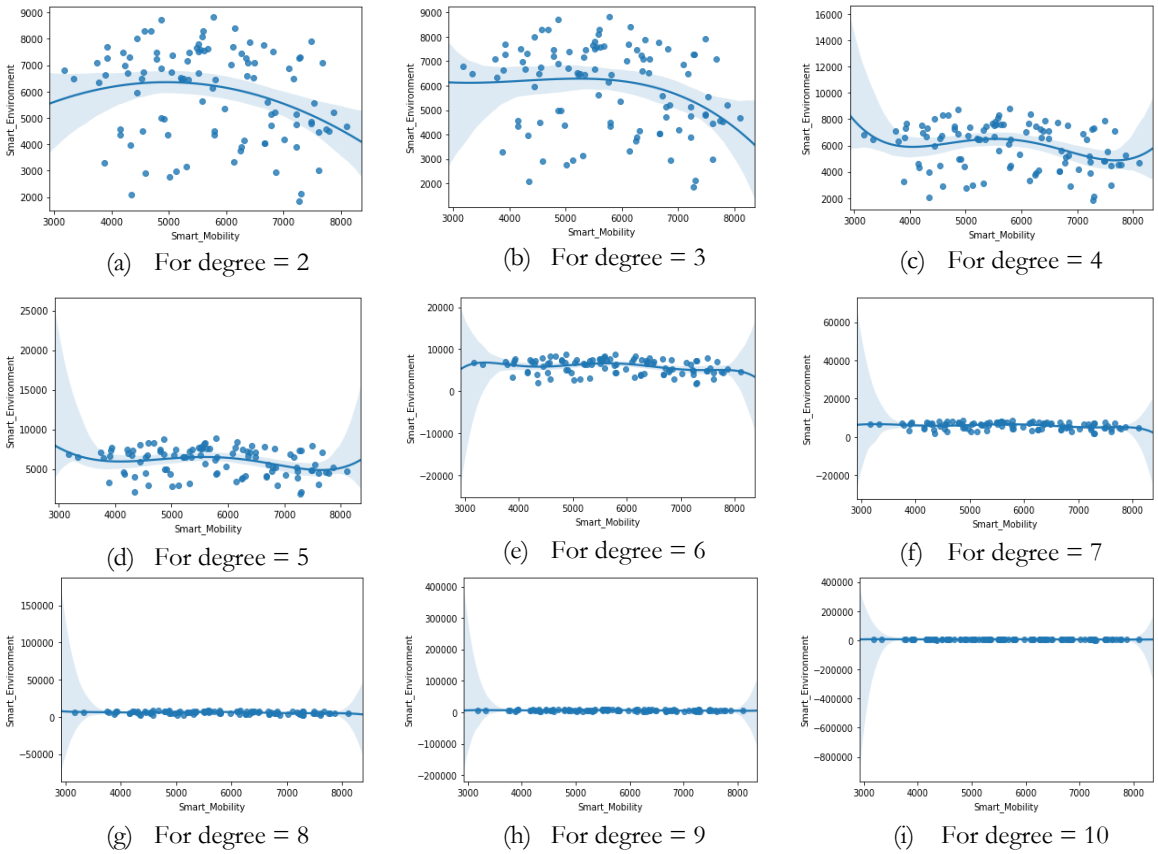

Figure 8. Regression plot for various degrees of polynomial.

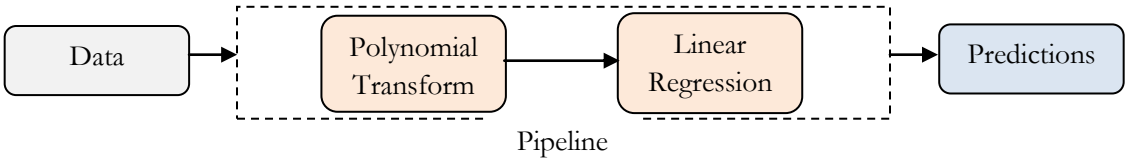

Figure 9. Steps of polynomial regression.

train/test data points and the behavior will be unpredictable. So it is important to fix the *random_state* value to provide us with the stable model (Haselirad, 2019). The *fit_intercept* parameter is selected *True* to add a (bias or intercept) column of any constant to the decision function to avoid underfitting. As we are using polynomial features with linear regression and the linear regression takes care by default of adding a (bias) column of 1s (since in Linear Regression the *fit_intercept* parameter is *True* by default), so we don't need to add it as well in polynomial features (Araldo, 2020). That's why we used include bias parameter, *include_bias* = *False* along with the polynomial features, which means we are not adding a (bias) column of 1s to the data-frame. As we are not normalizing the original data, hence *normalize* = *False* is used. We have listed all the selected values assigned to the parameters for our experiments (Table 3).

Table 3. The values assigned to the parameters

Parameter	Value assigned
Degree	6
include_bias	False
test_size	0.25
random_state	42
fit_intercept	True
normalize	False

Model evaluation metrics

For evaluation we have used four different performance metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-Squared value (R^2). These are the commonly used metrics to evaluate a regression model's performance. Error is the difference between original value of the test data and its predicted value by the model. MAE is the arithmetic average of the absolute errors and is expressed as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{3}$$

MSE is the arithmetic average of squares of the errors and RMSE is the square root of it. RMSE shows how far the values our model predicts are from the true values, on average. Roughly speaking: the smaller the RMSE, the better the model (Polynomial Regression in Python using Scikit-learn, 2021). These are expressed as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{5}$$

R^2 is called the coefficient of determination and it is more (intuitively) informative than MAE, MSE, and RMSE in regression analysis evaluation, as it can be expressed as a percentage, whereas the latter measures have arbitrary ranges (Draper & Smith, 1998). The bigger the R^2, the better the model. It is expressed as:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{6}$$

In the above equations, y_i represents the original value of test data label, $\hat{y}_i$ represents each predicted value by the model, $\bar{y}$ represents the average of the original values of test data label, and n represents the number of test data.

Regression outcomes using both selected features and all features

After hyper-tuning the polynomial regression algorithm for the prepared dataset, we move on to the prediction operation. The tuned polynomial regression model's outcomes regarding prediction errors and R-squared value is shown below (Table 4). Results of all cases provided in the table are achieved from same train and test sets.

Table 4. Prediction results

Features	MAE	MSE	RMSE	R^2
All features	1.40e-08	3.22e-16	1.79e-08	1.0
Features selected using Correlation	1.45e-06	2.33e-12	1.53e-06	1.0
Features selected using DT regressor	1.12e-06	1.74e-12	1.32e-06	1.0

For clearness, we have presented the results on using all features set as well as using selected feature sets to illustrate the reduction of the error while more features are provided. In all cases, though the R^2 values are same, mean errors are different. We first experiment using all features of the dataset and get $R^2 = 1.0$, then when we select features using correlation and DT regressor, we add features to the selected list until we get $R^2 = 1.0$ (explained before in section 2.3). That's why; the R^2 values are same for all cases. It can be observed that selected features using decision tree regressor provides less error than selected features using Pearson's correlation in this application, though in both cases total four features (but different feature subsets) were selected. But using all features (i.e. increasing one feature), we see that the prediction errors decrease rapidly. Mention that, in all cases the prediction errors are very low and ignorable. For MAE it is highest 0.00000145 and lowest 0.0000000140, for MSE it is highest 0.00000000000233 and lowest 0.000000000000000322, for RMSE it is highest 0.00000153 and lowest 0.0000000179. So we conclude that polynomial regression model shows robust outcome in the application of smart environment index prediction from index of other major components of a smart city.

Prediction performance comparison

To the best of our knowledge, no previous work exists that predicts smart environment index from index of five other components of a smart city used in this study. However, the reported results confirm that properly tuned polynomial regression model using applied features here shows the comparative results for any future work in the similar dataset and for same research question.

Conclusion

This article describes a unique method for smart environment index prediction of a city. It is currently not understood, what are the best combinations of features and hyper-tuned regression methods in smart environment index prediction from other smart city components. This is such system, evaluated by a rigorous study of the smart city features and properly tuned polynomial regression model. The proposed method shows robust results for this area of smart city. Further study can be carried out to see if the prediction performance can be improved by changing in features from the existing ones and tuning the model's parameters accordingly for this dataset as well as for other similar datasets.

Acknowledgement

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by ICT Division, Ministry of Posts, Telecommunications and Information Technology, Government of People's Republic of Bangladesh, through ICT scholarship [number: 56.00.0000.052.33.005.21-8].

References

Akhilesh, K. B. (2020). Smart Technologies—Scope and Applications. In *Smart Technologies* (pp. 1-16). Springer, Singapore.

Alexopoulos, C., Pereira, G. V., Charalabidis, Y., & Madrid, L. (2019, April). A taxonomy of smart cities initiatives. In *Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance*. (pp. 281-290).

Araldo, A. (2020, February 16). In practice, when you write Python code, and you use PolynomialFeatures together with sklearn.linear_model.LinearRegression, the latter takes care by default. [Comment on the article "Scikit-learn Polynomial Features – What is the Use of the Include Bias option".] https://stackoverflow.com/questions/59725907/scikit-learn-polynomialfeatures-what-is-the-use-of-the-include-bias-option

Benamrou, B., Mohamed, B., Bernoussi, A. S., & Mustapha, O. (2016, October). Ranking models of smart cities. In *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*. (pp. 872-879). IEEE.

Bhoomika, K.N., Deepa, C., Rashmi, R.K. & Srinivasa, R. (2016). Internet of Things for Environmental Monitoring. *International Journal of Advanced Networking & Applications*. 497–501.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (Vol. 326). John Wiley & Sons.

Drewil, G. I., & Al-Bahadili, R. J. (2021). Forecast Air Pollution in Smart City Using Deep Learning Techniques: A Review. *Multicultural Education*, *7*(5). 38-47. https://doi.org/10.5281/zenodo.4737746

Giffinger, R., Fertner, C., Kramar, H., & Meijers, E. (2007). City-ranking of European medium-sized cities. *Cent. Reg. Sci. Vienna UT*, *9*, 1-12.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques* (3rd ed.). Elsevier.

Haselirad, A. (2019, January 29). If there is no randomstate provided the system will use a randomstate that is generated internally. So, when you run. [Comment on the article "Random state (Pseudo-random number) in Scikit learn".] https://stackoverflow.com/questions/28064634/random-state-pseudo-random-number-in-scikit-learn

International Organization for Standardization. (2014). *Sustainable development of communities: Indicators for city services and quality of life*. ISO.

Introduction to Polynomial Regression. (2020). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/03/polynomial-regression-python/

Kang, G. K., Gao, J. Z., Chiao, S., Lu, S., & Xie, G. (2018). Air quality prediction: Big data and machine learning approaches. *International Journal of Environmental Science and Development*, *9*(1), 8-16. https://doi.org/10.18178/ijesd.2018.9.1.1066

Kulkarni, P., & Akhilesh, K. B. (2020). Big data analytics as an enabler in smart governance for the future smart cities. In *smart technologies* (pp. 53-65). Springer, Singapore.

Machine Learning – Polynomial Regression (n.d.). W3schools. https://www.w3schools.com/python/python_ml_polynomial_regression.asp

Polynomial Regression. (2022). Net-informations.com. http://net-informations.com/ds/mla/pr.htm

Polynomial Regression in Python using Scikit-learn. (2021). Data36. https://data36.com/polynomial-regression-python-scikit-learn/

Prefeitura Belo Horizonte. (2020). Cidade Inteligente. https://prefeitura.pbh.gov.br/cidade-inteligente

Pure Earth (n.d.). https://www.pureearth.org.

Singh, M. N., & Kumar, M. (2020). Big Data Analytics Based Methods For Addressing Various Issues Efficiently in Smart Cities: A Comprehensive Survey. *International Journal of Advanced Science and Technology*. *29*(5s). 238-245

Smart Cities Index Datasets (2021). Kaggle. https://www.kaggle.com/magdamonteiro/smart-cities-index-datasets

Smart City Index Methodology (n.d.). https://www.imd.org/globalassets/wcc/docs/smart_city/smart_city_index_methodology_and_groups.pdf

Smart City Observatory (n.d.). IMD. https://www.imd.org/smart-city-observatory/home/

Soomro, K., Bhutta, M. N. M., Khan, Z., & Tahir, M. A. (2019). Smart city big data analytics: An advanced review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(5), e1319. https://doi.org/10.1002/widm.1319

TEIXEIRA, J. V. S., Gerais, M., BARACHO, B. R. M. A., & MULLARKEY, B. M. T. (2020). Proposal for Sustainable Smart City Indicators. *Convention Proceedings: Papers presented at the 24th World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI 2020)*. (pp. 120-125). International Institute of Informatics and Cybernetics.

Tran Thi Hoang, G., Dupont, L., & Camargo, M. (2019). Application of decision-making methods in smart city projects: a systematic literature review. *Smart Cities*, *2*(3), 433-452. https://doi.org/10.3390/smartcities2030027

Ullo, S. L., & Sinha, G. R. (2020). Advances in smart environment monitoring systems using IoT and sensors. *Sensors*, *20*(11), 3113. https://doi.org/10.3390/s20113113