

METHOD AND SYSTEM FOR ANALYZING CO2 EMISSIONS USING MODEL OF MODELS AND VARIABLE OF VARIABLES APPROACH

FIELD OF INVENTION

[0001] Embodiments of the present disclosure relates to analytics and more particularly to a method and a system for analyzing CO2 emissions using model of models and variable of variables approach.

BACKGROUND

[0002] Economic growth affects the way a country uses energy and in return increases the rate of carbon dioxide (CO2) emissions in the atmosphere. Economic growth leads to industrialization and increases in other economic activities. Several factors such as consumption of natural resources and environmental degradation increases carbon emission, global warming and other forms of pollution. The increase of CO2 emissions is likely to increase climatic changes, specifically global warming. Therefore, it is vital to comprehend what climate, economic and environmental-related organizations affects energy consumption and CO2 emissions levels.

[0003] It is important to reduce CO2 and Green House gas emissions based on economic indicators. This would aid the government, industry and policy makers to readjust strategies based on economic indicators. Further, the key economic indicators must be prioritized to reduce global warming and climatic changes.

[0004] The modern era of machine learning is capable of guiding the control of CO2 emissions, offering numerous prospects to reshape its management and control. However, with the development of machine learning, the accessibility of open software systems, and the enormous volume of real-time data created and saved, countries are still not fully embracing machine learning methods. Moreover, the countries are just beginning to recognize the significance of the enormous quantity of data they own to help make the correct choices and strengthen their policies and strategies.

[0005] Hence, there is a need for an improved method and system for analyzing CO₂ emissions using model of models and variable of variables approach to address the aforementioned issue(s).

OBJECTIVE OF THE INVENTION

[0006] An objective of the present invention is to analyze and reduce datasets based on economic indicators thereby prioritizing the economic indicators to reduce global warming and climatic changes.

[0007] Another objective of the present invention is to obtain CO₂ relation with the economic indicators using machine learning models.

BRIEF DESCRIPTION

[0008] In accordance with an embodiment of the present disclosure, a method for analyzing CO₂ emissions using model of models and variable of variables approach is disclosed. The method includes creating a bootstrap forest of a plurality of trees to identify a set of top variables in a dataset using a decision tree rule in machine learning. The method also includes selecting at least two sets of trees from a predetermined number of trees based on a rank metric, wherein the rank metric is Rank and Rank2. Further, the method includes providing a list of all variables present in the predetermined number of trees in the set of Rank and Rank2. Furthermore, the method includes selecting a predetermined number of variables from the list based on a correlation matrix wherein the correlation matrix is greater than 0.6 corresponding to each of the predetermine number of variables. Moreover, the method includes receiving an evaluation metric for the predetermined number of variables based on Rank and Rank2, wherein the evaluation metric is qualified to predict future outcomes.

[0009] In accordance with another embodiment of the present disclosure, a system for analyzing CO₂ emissions using model of models and variable of variables approach is disclosed. The system includes a processing subsystem hosted on a server and configured to execute on a network to control bidirectional communications among a plurality of modules. The plurality of modules includes an input module, a preprocessing module, an analysis module and a variable identification module. The input module is configured to collect a dataset. The preprocessing module is

operatively coupled to the input module and is configured to clean the dataset and identify missing values in the dataset. Further, the analysis module is operatively coupled to the preprocessing module and configured to perform time series analysis to identify one or more patterns in the dataset. Furthermore, the variable identification module is operatively coupled to the analysis module and configured to identify top variables in the dataset using artificial intelligence that causes the variable identification module to perform: select at least two sets of trees from a predetermined number of trees based on a rank metric, wherein the rank metric is Rank and Rank2, provide a list of all variables present in the predetermined number of trees in the set of Rank and Rank2, select a predetermined number of variables from the list based on a correlation matrix wherein the correlation matrix is greater than 0.6 corresponding to each of the predetermine number of variables, and receive an evaluation metric for the predetermined number of variables based on Rank and Rank2, wherein the evaluation metric is qualified to predict future outcomes.

[0010] To further clarify the advantages and features of the present disclosure, a more particular description of the disclosure will follow by reference to specific embodiments thereof, which are illustrated in the appended figures. It is to be appreciated that these figures depict only typical embodiments of the disclosure and are therefore not to be considered limiting in scope. The disclosure will be described and explained with additional specificity and detail with the appended figures.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] The disclosure will be described and explained with additional specificity and detail with the accompanying figures in which:

[0012] FIG. 1 illustrates a flow chart representing steps involved in a method for analyzing CO₂ emissions using model of models and variable of variables approach in accordance with an embodiment of the present disclosure;

[0013] FIG. 2 is a schematic representation of an exemplary embodiment of the method of FIG. 1 in accordance with an embodiment of the present disclosure; and

[0014] FIG. 3 is a block diagram representation of a computer or a server in accordance with an embodiment of the present disclosure.

[0015] Further, those skilled in the art will appreciate that elements in the figures are illustrated for simplicity and may not have necessarily been drawn to scale. Furthermore, in terms of the construction of the device, one or more components of the device may have been represented in the figures by conventional symbols, and the figures may show only those specific details that are pertinent to understanding the embodiments of the present disclosure so as not to obscure the figures with details that will be readily apparent to those skilled in the art having the benefit of the description herein.

DETAILED DESCRIPTION

[0016] For the purpose of promoting an understanding of the principles of the disclosure, reference will now be made to the embodiment illustrated in the figures and specific language will be used to describe them. It will nevertheless be understood that no limitation of the scope of the disclosure is thereby intended. Such alterations and further modifications in the illustrated system, and such further applications of the principles of the disclosure as would normally occur to those skilled in the art are to be construed as being within the scope of the present disclosure.

[0017] The terms "comprises", "comprising", or any other variations thereof, are intended to cover a non-exclusive inclusion, such that a process or method that comprises a list of steps does not include only those steps but may include other steps not expressly listed or inherent to such a process or method. Similarly, one or more devices or sub-systems or elements or structures or components preceded by "comprises... a" does not, without more constraints, preclude the existence of other devices, sub-systems, elements, structures, components, additional devices, additional sub-systems, additional elements, additional structures or additional components. Appearances of the phrase "in an embodiment", "in another embodiment" and similar language throughout this specification may, but not necessarily do, all refer to the same embodiment.

[0018] Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by those skilled in the art to which this disclosure belongs. The system, methods, and examples provided herein are only illustrative and not intended to be limiting.

[0019] In the following specification and the claims, reference will be made to a number of terms, which shall be defined to have the following meanings. The singular forms “a”, “an”, and “the” include plural references unless the context clearly dictates otherwise.

[0020] Embodiments of the present disclosure relate to a system and method for analyzing CO₂ emissions using model of models and variable of variables approach. The method includes creating a bootstrap forest of a plurality of trees to identify a set of top variables in a dataset using a decision tree rule in machine learning. The method also includes selecting at least two sets of trees from a predetermined number of trees based on a rank metric, wherein the rank metric is Rank and Rank2. Further, the method includes providing a list of all variables present in the predetermined number of trees in the set of Rank and Rank2. Furthermore, the method includes selecting a predetermined number of variables from the list based on a correlation matrix wherein the correlation matrix is greater than 0.6 corresponding to each of the predetermine number of variables. Moreover, the method includes receiving an evaluation metric for the predetermined number of variables based on Rank and Rank2, wherein the evaluation metric is qualified to predict future outcomes.

[0021] FIG. 1 illustrates a flow chart representing steps involved in a method (100) for analyzing CO₂ emissions using model of models and variable of variables approach in accordance with an embodiment of the present disclosure. The method (100) includes creating a bootstrap forest of a plurality of trees to identify a set of top variables in a dataset using a decision tree rule in machine learning at step 105. Typically, the dataset is collected from a plurality of websites wherein the dataset is historical and economic data pertaining to industries, policy makers and governments. In one embodiment, the dataset is collected from articles, blogs and reports. In another embodiment, the dataset is extrapolated in the case of unavailability. In other words, the dataset is collected on historical CO₂ (Green House gases) and economic indicators. Examples of the economic indicators includes, but is not limited to, currency, stock market, unemployment rate, Gross Domestic Product (GDP), inflation rate and population.

[0022] Further, the collected dataset is cleaned to identify one or more potential missing values in the dataset. Typically, the missing values are those values that may

not be generated yet. It must be noted that the missing values may either be filled or discarded based on its availability. In one embodiment, the dataset is cleaned using an Excel application. Subsequently, a machine learning model is created to predict wherever values are missing for datasets collected in future. The machine learning model utilizes a suitable modeling technique. Examples of the modeling technique includes, but are not limited to, Fit least Squares, Fit Stepwise, Partial Least Squares, K nearest neighbors, Neural Boosted, Boosted Tree, Generalized Regression Lasso, Support Vector Machine and Bootstrap Forest. In a specific embodiment, the Bootstrap Forest is used to create a bootstrap forest of a plurality of trees for the said dataset. Typically, Bootstrap Forest is a technique or method that creates an ensemble of decision trees. Each decision tree is trained with a subset of the dataset. At each node of a decision tree, several variables are randomly selected from all the variables to train the decision rule. The final outcome is based on a combination of the outcomes of the individual decision trees.

[0023] It must be noted that although the dataset corresponds to historical and economic data, the dataset is not limited to the said and any suitable dataset may be used.

[0024] Further, the dataset is subjected to time series analysis to identify one or more patterns in the dataset (data visualization). The patterns may be defined as data that represents a regularity. For instance, consider the dollar rate and emission. If the dollar rate increases, then the emission decreases. This could be a pattern that can be revised to regulate the dollar rate and thereby the emission. Therefore, the number of variables are reduced by identifying the one or more patterns and as a result, the most important variables are identified.

[0025] In one embodiment, the plurality of trees are sorted based on the rank metric in ascending order of Out-of-band (OOB) loss indicating a corresponding descending predictive accuracy.

[0026] Further, top variables are identified using a decision tree method that affects the CO₂ emission. These variables are herein referred to as ‘variable of variables’. Subsequently, model building is performed using a minimum of ten models

using the ‘variable of variables’. In other words, model building is performed using the ‘variable of variables’.

[0027] Additionally, the method (100) includes selecting at least two sets of trees from a predetermined number of trees based on a rank metric, wherein the rank metric is Rank and Rank² in step 110. In one embodiment, the predetermined number of trees is five. Therefore, two sets of the top 5 trees are selected based on the rank metric. The ‘Rank’ may be defined as the OOB loss in the ascending order. Typically, the tree with the smallest OOB loss has Rank 1. OOB defines a measure of the total predictive inaccuracy of the tree when applied to the Out of Bag rows. Specifically, the tree with the smallest OOB loss has Rank 1. Therefore, a lower value indicates a higher predictive accuracy. Rank² measures the proportion of the variation explained by the model.

[0028] Further the method (100) includes providing a list of all variables present in the predetermined number of trees in the set of Rank and Rank2 in step 115. In one embodiment, the predetermined number of trees is five. Therefore, the set of five trees based on the Rank and the list of all variables that appeared in all the five trees is listed. Similarly, the set of 5 trees based on Rank² and list down all variables that appeared in all the 5 trees in the set.

[0029] Furthermore, the method (100) includes selecting a predetermined number of variables from the list based on a correlation matrix wherein the correlation matrix is greater than a pre-determined cutoff value based on the dataset in step 120. In one embodiment, the cutoff value is 0.6. In one embodiment, the predetermined number of variables is five. A correlation matrix is generated for all the variables in the set and the top 5 variables is selected where the correlation is greater than 0.6 for that variable. The same is performed for both the sets and get 2 sets of top 5 variables based on Rank and Rank². Typically, the correlation matrix is a table (square matrix) illustrating the correlation values between variables. As understood by a person skilled in the art, the correlation values may be positive or negative. However, the variables that correspond to correlation value greater than 0.6 are selected here.

[0030] In one embodiment, the predetermined number of variables is subjected to a machine learning model to obtain an optimal model for the dataset. In another

embodiment, the machine learning model predicts future outcomes for the predetermined number of variables.

[0031] Moreover, the method (100) includes receiving an evaluation metric for the predetermined number of variables based on Rank and Rank2, wherein the evaluation metric is qualified to predict future outcomes in step (125). In one embodiment, the predetermined number of variables from the list is subjected to cross validation thereby generating the evaluation metric.

[0032] Further, Leave One Out Cross Validation (LOOCV) technique is performed to get the metric Mean Absolute Percentage Error (MAPE) for the set of 5 variables based on Rank. Likewise, LOOCV is performed to get the metric MAPE for the set of 5 variables based on Rank². LOOCV is a type of cross-validation approach in which each observation is considered as the validation set and the rest (N-1) observations are considered as the training set. MAPE is a metric that defines the accuracy of a forecasting method. Further, LOOCV is used for a small dataset and MAPE is used for easy interpretation and explanation. Based on MAPE, the final set of 5 variables which can represent the whole dataset are obtained.

[0033] In one embodiment, multilinear regression is performed on these set of 5 variables to predict future outcomes. The final model is called “Model of Models”.

[0034] FIG. 2 is a schematic representation of an exemplary embodiment of the method of FIG. 1 in accordance with an embodiment of the present disclosure. The method (200) includes creating a Bootstrap forest of 1000 trees with other specified parameters on data in step (205).

[0035] The method (200) also includes sorting the trees according to the Rank metric in decreasing order in step (210).

[0036] Further, the method (200) includes taking the top 5 trees in step (215).

[0037] Furthermore, the method (200) includes listing down all the variables in the top 5 trees in step (220).

[0038] Moreover, the method (200) includes taking the top 5 variables from the list based correlation matrix having correlation >0.6 in step (225).

[0039] Additionally, the method (200) includes obtaining the top 5 variables according to the Rank in step (230).

[0040] It must be noted that the step (210) to step (230) defines the process of obtaining the top 5 variables based on the Rank. Additionally, the process is applied to the Bootstrap Forest based on Rank² as shown in step (235) to step (255).

[0001] FIG. 3 is a block diagram representation of a computer or a server in accordance with an embodiment of the present disclosure. The server (300) includes processor(s) (340), and memory (345) operatively coupled to the bus (350). The processor(s) (340), as used herein, means any type of computational circuit, such as, but not limited to, a microprocessor, a microcontroller, a complex instruction set computing microprocessor, a reduced instruction set computing microprocessor, a very long instruction word microprocessor, an explicitly parallel instruction computing microprocessor, a digital signal processor, or any other type of processing circuit, or a combination thereof.

[0041] The memory (345) includes several subsystems stored in the form of executable program which instructs the processor (340) to perform the method steps illustrated in FIG. 1. The memory (345) includes a processing subsystem (310). The processing subsystem (310) further has following modules: an input module (320), preprocessing module (325), analysis module (330) and a variable identification module (335).

[0042] The input module (320) is configured to collect a dataset. The preprocessing module (325) is operatively coupled to the input module (320) wherein the preprocessing module (325) is configured to clean the dataset and identify the missing values in the dataset. The analysis module (330) is operatively coupled to the preprocessing module (325) and is configured to perform time series analysis and identify patterns in the dataset. The variable identification module (335) is operatively coupled to the analysis module (330) wherein the variable identification module (335) is configured to identify top variables in the dataset using artificial intelligence that causes the variable identification module to perform a series of steps. The steps includes selecting at least two sets of trees from a predetermined number of trees based on a rank metric, wherein the rank metric is Rank and Rank2. Further, a list of all

variables present in the predetermined number of trees in the set of Rank and Rank2 is provided. Furthermore, a predetermined number of variables from the list based on a correlation matrix wherein the correlation matrix is greater than 0.6 corresponding to each of the predetermine number of variables is selected. Moreover, an evaluation metric is received for the predetermined number of variables based on Rank and Rank2, wherein the evaluation metric is qualified to predict future outcomes.

[0043] In one embodiment, the variable identification module (335) is configured with one or more machine learning models, wherein the one or more machine learning models are trained with the predetermined number of variables to obtain an optimal model for the dataset. It must be noted that the optimal model for the dataset predicts future outcomes.

[0002] The bus (350) as used herein refers to internal memory channels or computer network that is used to connect computer components and transfer data between them. The bus (350) includes a serial bus or a parallel bus, wherein the serial bus transmits data in bit-serial format and the parallel bus transmits data across multiple wires. The bus (350) as used herein, may include but not limited to, a system bus, an internal bus, an external bus, an expansion bus, a frontside bus, a backside bus and the like.

[0044] Various embodiments of the present disclosure as described above provides a data-driven approach for decision making with several industries, policy makers and government. The input module collects the dataset based on economic indicators. This aids the government, industry and policy makers to readjust strategies based on the said economic indicators. Further, by identifying the top 5 variables by the variable identification module, the system disclosed herein prioritizes the key economic indicators that should be focused on reducing global warming and climate change. Further, the analysis module resolves the aforementioned problems by using machine learning models to identify economic indicators thereby producing accurate predictions.

[0045] The present invention is implementable by a device that may be implemented in programmable hardware devices such as processors, digital signal processors, central processing units, field programmable gate arrays, programmable

array logic, programmable logic devices, cloud processing systems, or the like. The devices may also be implemented in software for execution by various types of processors. An identified device may include executable code and may, for instance, comprise one or more physical or logical blocks of computer instructions, which may, for instance, be organized as an object, procedure, function, or other construct. Nevertheless, the executable of an identified device need not be physically located together but may comprise disparate instructions stored in different locations which, when joined logically together, comprise the device and achieve the stated purpose of the device.

[0046] It will be understood by those skilled in the art that the foregoing general description and the following detailed description are exemplary and explanatory of the disclosure and are not intended to be restrictive thereof.

[0047] While specific language has been used to describe the disclosure, any limitations arising on account of the same are not intended. As would be apparent to a person skilled in the art, various working modifications may be made to the method in order to implement the inventive concept as taught herein.

[0048] The figures and the foregoing description give examples of embodiments. Those skilled in the art will appreciate that one or more of the described elements may well be combined into a single functional element. Alternatively, certain elements may be split into multiple functional elements. Elements from one embodiment may be added to another embodiment. For example, the order of processes described herein may be changed and are not limited to the manner described herein. Moreover, the actions of any flow diagram need not be implemented in the order shown; nor do all of the acts need to be necessarily performed. Also, those acts that are not dependent on other acts may be performed in parallel with the other acts. The scope of embodiments is by no means limited by these specific examples.

WE CLAIM:

1. A method (100) for analyzing CO₂ emissions using model of models and variable of variables approach comprising:

creating a bootstrap forest of a plurality of trees to identify a set of top variables in a dataset using a decision tree rule in machine learning wherein the set of top variables is called as variable of variables; (105)

selecting at least two sets of trees from a predetermined number of trees based on a rank metric, wherein the rank metric is Rank and Rank2; (110)

providing a list of all variables present in the predetermined number of trees in the set of Rank and Rank2; (115)

selecting a predetermined number of variables from the list based on a correlation matrix wherein the correlation matrix is greater than a pre-determined cut off value based on the dataset, corresponding to each of the predetermine number of variables; (120) and

receiving an evaluation metric for the predetermined number of variables based on Rank and Rank2, wherein the evaluation metric is qualified to predict future outcomes by selecting a model from multiple models called as model of models. (125)

2. The method (100) as claimed in claim 1, wherein the dataset is collected from a plurality of websites or other sources wherein the dataset pertains to historical and economic data.

3. The method (100) as claimed in claim 2, wherein the dataset is cleaned to identify potential one or more missing values in the dataset.

4. The method (100) as claimed in claim 1, wherein the plurality of trees are sorted based on the rank metric in ascending order of Out-of-band loss indicating a corresponding descending predictive accuracy.

5. The method (100) as claimed in claim 1, wherein the dataset is subjected to time series analysis thereby identifying one or more patterns in the dataset.

6. The method (100) as claimed in claim 1, wherein the predetermined number of variables from the list is subjected to cross validation thereby generating the evaluation metric.

7. The method (100) as claimed in claim 1, wherein the predetermined number of variables is subjected to a machine learning model to obtain an optimal model for the dataset.

8. The method (100) as claimed in claim 1, wherein the machine learning model predicts future outcomes for the predetermined number of variables called as variable of variables.

9. A system (300) for analyzing CO2 emissions using model of models and variable of variables approach comprising:

a processing subsystem (310) hosted on a server (315) and configured to execute on a network to control bidirectional communications among a plurality of modules comprising:

an input module (320) configured to receive a dataset from a plurality of websites wherein the dataset pertains to historical and economic data;

a preprocessing module (325) operatively coupled to the input module (320) wherein the preprocessing module is configured to:

clean the dataset; and

identify missing values in the dataset;

an analysis module (330) operatively coupled to the preprocessing module (325) wherein the analysis module (330) is configured to perform time series analysis to identify one or more patterns in the dataset;

a variable identification module (335) operatively coupled to the analysis module (330) wherein the variable identification module (335) is configured to identify top variables in the dataset using artificial intelligence that causes the variable identification module to:

select at least two sets of trees from a predetermined number of trees based on a rank metric, wherein the rank metric is Rank and Rank2;

provide a list of all variables present in the predetermined number of trees in the set of Rank and Rank2;

select a predetermined number of variables from the list based on a correlation matrix wherein the correlation matrix is greater than a pre-determined cut-off value based on the dataset corresponding to each of the predetermine number of variables; and

receive an evaluation metric for the predetermined number of variables based on Rank and Rank2, wherein the evaluation metric is qualified to predict future outcomes.

10. The system (300) as claimed in claim 9 wherein the variable identification module (335) is configured to identify a predetermined number of variables that affect CO2 emission.

11. The system (300) as claimed in claim 10 wherein the variable identification module (335) is configured to train one or more machine learning models with the predetermined number of variables to obtain an optimal model for the dataset.

12. The system (300) as claimed in claim 10 wherein the optimal model called as model of models for the dataset predicts future outcomes.