Data Mining Engine using Predictive Analytics

Sakshi Rungta
Bharati Vidyapeeth's College of Engineering, New Delhi

Vanita Jain, PhD
Bharati Vidyapeeth's College of Engineering, New Delhi

Akanksha Utreja
Bharati Vidyapeeth's College of Engineering, New Delhi

ABSTRACT
Predictive analytics is a field of data mining which extracts information from the past and use it to predict the future trends. This paper establishes the importance of predictive analysis. In this paper, we present a system to analyse user stories incorporating the data of energy and health demands of four countries – namely India, China, United States of America and Brazil; for the past 30 years, depict them graphically using Business Intelligence and finally predict the future trend of the parameters. The correlation between various entities is found out using Pearson's coefficient. Finally we can see the predicted values of 30-40 years ahead and predict the emerging trends in the form of Power View charts. We present lessons learned and future directions for improving the user in the loop workflow for predictive analytics.

Keyword
Predictive Analytics, Regression Analysis, Health and Energy Demands

1. INTRODUCTION
Data – of various sizes, complexity and diversity is present everywhere and is increasing annually. According to Gartner, the total worldwide volume of data is growing worldwide at 59% per year.

It is very difficult to draw any valid inferences and conclusions directly from the available data. The data can be understood better through visual representations in the form of charts and graphs. The dependencies between the data, plays a very important role in the predictive analysis as well. Depending upon the correlation trend between the historical data, the future data can be predicted. This data is used by organizations to drive new innovations and faster insights into the customers demand [19].

Data can be very useful for analysis and making decisions if mined and showcased properly. This data can be used to take important business decisions. Further Business Intelligence is an impressive way to depicting data into various forms of graphs, charts and maps and have a better and a wider picture of the analysis [2].

In this paper the data visualization is done through Power BI. Power BI is an excellent way to portray the structured data in a graphical form which can be analyzed and studied easily by the user [6].

The following features of Power BI are extensively used:

- **Power Query** has been used to take in OData feed from the service and form tables.
- **Power View** provides various options to represent the tables in different types of charts:
  - Line graphs, bar charts, stacked columns are used for analysis purposes.
  - A scatter chart with the video play of timeline is used to show and compare the trends of various countries together in the same graph.
  - Various graphical representations have been used to demonstrate our analysis, prediction and solution.

Deploying the service on Azure freed our proposed system from any platform dependencies, making it accessible from any system. Microsoft provides an online version of MS Excel using O365 which can be used to consume the service from any machine any platform

2. LITERATURE OVERVIEW AND RELATED WORK
Predictive analytics—a statistical or data mining solution that consists of techniques and algorithms which can be used on both unstructured and structured data to determine outcomes. [23]

Predictive analytics comprises of a wide variety of statistical techniques from data mining, modeling and machine learning that analyze historical facts and current data to make predictions about the future events [13]. The difference between Predictive analytics and text analytics is that predictive analytics uses statistical or structural models for classifications and making the predictions while text analytics uses an additional linguistic technique to extract and classify information [2].

Predictive Analytics is a subset of Data Mining field which is an important part of the Data Science discipline. The term Analytics is derived from the science of data analysis that is more commonly associated with another term BI. Business Intelligence is used to describe the provisioning of decision support in businesses [13]. Predictive Analytics is supported by applying statistical and mathematical techniques to derive meaning from data and systematically find relations and patterns in data for decision making directives. The applications of Predictive Analytics are found in the fields of academia along with the industries.

Predictive Analytics has a broad scope and wide range of applications. Predictive Analytics is not performed in isolation and is considered to be a systematic approach for systems to work together in unison to derive result [20]. Also, Predictive Analytics is realized through machine learning and is substantiated by statistical techniques to analyse, model and deduce knowledge. The extracted data contains actionable information from seemingly uninteresting dataset. Use of database systems in predictive analysis in mandatory to maintain a collection of data for processing which is then

further used for analytical modelling to that transform raw data into actionable knowledge [24].

Traditional approaches in analytics are hard to use in the real-time environment used commonly for operational decisions and hard to scale. Predictive analytics work better in scenarios where data is used to make these decisions more precise, personalizing and targeting them to maximize customer value [22]. In business, the predictive models identify patterns in transactional and historical data to identify opportunities and risks. They capture relationships among the particular set of conditions, facilitating decision making for the future transactions.

Organizations study the previous trends to perform predictive analysis for the future as well as formulate new policies for further improvement in the performance. Data can also be compared and any problem in the future can be anticipated and worked upon to be tackled in a better way. [2]

Predictive Analytics and BI differ from each other in a variety of ways. BI is good for slicing and dicing of data to answer questions such as what is happening, what happened and why it happened. However, BI provides dashboards or static reports which are inflexible. With predictive analytics, however, users can estimate targets or outcome of interest. Outcomes might include: Which customers are more likely to disconnect from a service? How much will something increase or decrease in value? Predictive analytics is more proactive and deep. For better understanding of relationship between BI and predictive analytics, consider the spectrum of analysis techniques: from historical, static reporting through more advanced techniques that move from historical to future, and from reactive to proactive. This is where predictive analytics comes into picture as a more sophisticated analytics techniques than descriptive techniques (such as dashboards or reporting). [23]

This paper focuses on exploring and analysing the trends of energy and health demands for developing and developed countries data from World Bank [1] using predictive analytics.

Other than the proposed system there are other solutions that are used to address predictive analysis. These other solutions include R[7], SAS[8], Weka[9], JMP[10], and Excel.

Regression analysis, parameter optimization and result validation, all these are included as tools and packages in these solutions which can be used for predictive analysis task. These systems provide basic visualizations including line chart, scatter plots and residual plots [4]. The goal of our system is to provide an interactive and a more visually appealing means to depict the predicted data which is made possible by exploring the capabilities of Power Bi, Power pivot and Power query.

3. FRAMEWORK FOR PREDICTION ENGINE

Our system integrates data from the World Bank [1] for the countries namely, India, China, Brazil and USA for analysis and prediction. We combined

Excel trends and regression analysis for model building, evaluation and prediction. The predicted results were demonstrated using a Business Intelligence tool, PowerBI. The system was built using Excel, ATOM/XML, ASP.NET and PowerBI. The use of Excel allowed for the regression analysis and model selection, while PowerBI was used to create graphics and charts for interactive visualization. A WCF RESTful service [15] was written in C# to expose the data as OData to be consumed by the BI tool, i.e. PowerBI, and we also explored the use of Azure; Cloud service provided by Microsoft. The system is based on a three tier architecture where SQL Server provides the Data Layer, the WCF Data Service and Entity Framework Model comprise of the Application Layer and Microsoft Excel is used at the Presentation Layer.

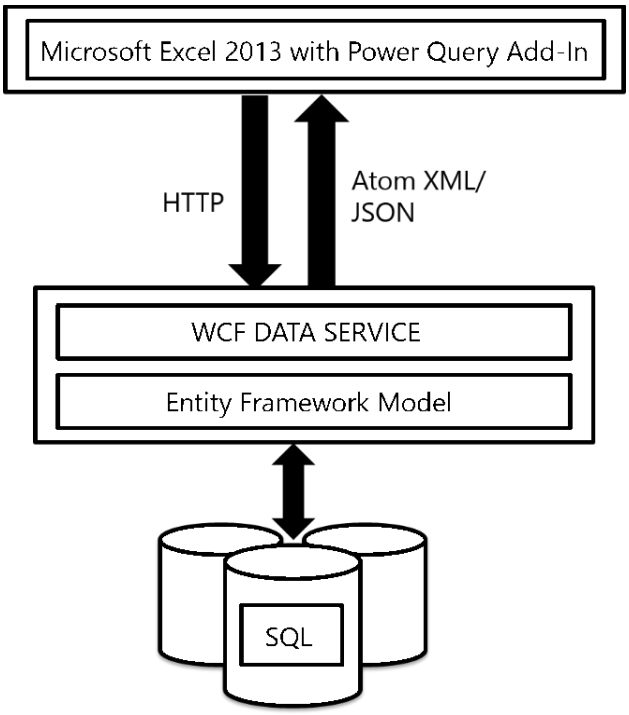

Figure 1: Solution Architecture

3.1 Data Description

In data description and representation we focused on data across two parameters, Health and Energy. For Health, we collected data for the indicators namely, Life expectancy at birth, Health expenditure and Population. And for Energy, we collected data for indicators namely, Energy use, Alternate sources of energy, CO2 emissions, Alternate and Nuclear energy, Energy imports and Fossil fuel consumption.
The data for the user story was collected from the World Bank [1] and analyzed in the following step-wise manner:

Data collection: The data was collected in Excel sheets from the data source and the different parameters were analyzed and selected for prediction.3

Data Cleaning: There were some discrepancies in the data, which included missing data values. We followed the following strategies for cleansing the data:

- Data columns of equal number of entries are taken for comparison of trends. Any extra entries in other columns are deleted to maintain equal sizes.
- If there are any missing data value in between the data set, then the missing value is filled according to the trend followed by the remaining data values in the set

3.2 Regression Analysis

In statistics, regression analysis [3] is a statistical process for estimating the relationships among variables. The relationship between a dependent and more than one independent variables can be modelled and analysed using regression analysis. Major use of regression analysis is found in prediction and forecasting, where it's used along machine learning. The use of regression analysis is twofold, one, to understand which among the independent variables are related to the dependent variable, and second to explore these relationships.

The type of regression used can be Ordinary Least Square, Polynomial, Quantile and Exponential etc. The most commonly used regression is linear regression.

3.2.1 Predictive Model Selection using Regression

After having analyzed the historical data available for the required parameters and finding the correlation between them and having followed the strategies for data cleansing and data model cross –validation, the equation giving the least error is selected for predictive analysis.

Now, for the prediction of the value of any given parameter 'y' which is dependent on 'x', we will use the correlation between 'x' vs. time and 'y' vs. 'x'. We will get the value of 'x', by substituting the year in 'x' vs. time and then use this value of 'x' to get the value of 'y' using the correlation between 'y' vs. 'x'.

Various regression model selection strategies are analyzed and the one having maximum value of Pearson's coefficient [18], R^2 value is chosen for future analysis. The correlation equation whose value of R^2 is closest to 1 is chosen. If the value of R^2 is less than 0.6, the correlation does not exist for the given set of parameters.

From the correlation equations analysis in Figures 2, and 4 , we see that the Polynomial Regression model with an R^2 value of 0.99 is the best fitting model.

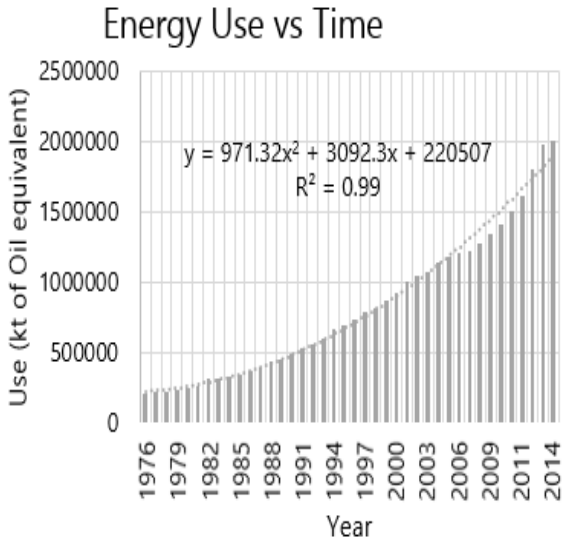

Figure 2: Linear Regression

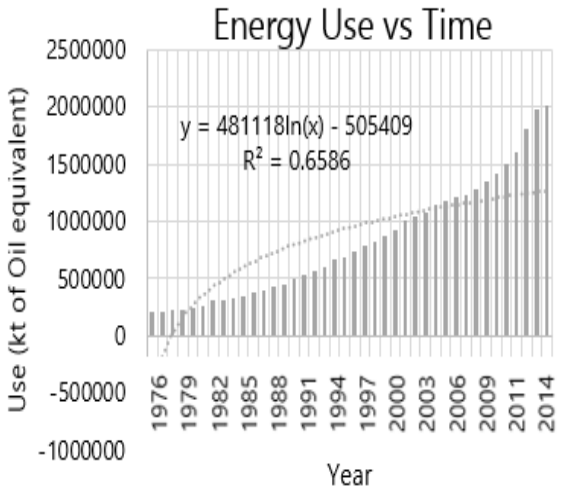

Figure 3: Logarithmic Regression

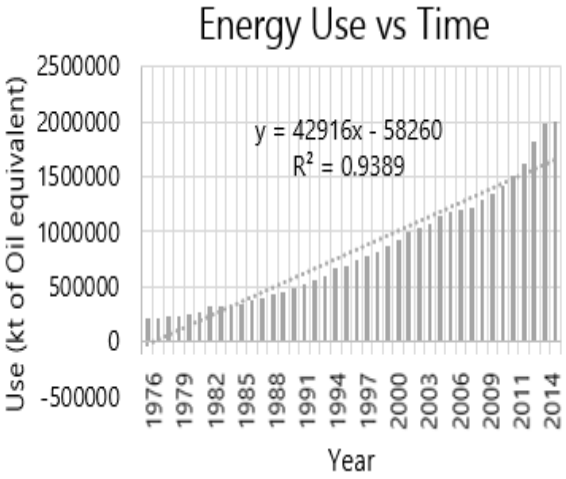

Figure 4: Polynomial Regression

3.3 WCF Data Service

WCF Data Services [11] is used to create services that use the Open Data Protocol (OData) to expose the data as Open data over internet which can be accessed using REST; Representational State Transfer in which the data is accessed and modified by using HTTP methods such as GET, PUT, POST, and DELETE.

WCF Data Services uses the OData [12] protocol for addressing and updating resources. In this way, you can access these services from any client that supports OData. OData uses ATOM or JSON as the transfer formats for requesting and writing data to resources. The following functions were performed by the data service:

- **Exposing Data:** Data is exposed as resources that are addressable by URIs.
- **Formatting:** Data is accessible in JSON, Verbose JSON and XML format.
- **Querying:** Data can be filtered, projected and paginated from the database by using keywords like $filter, $select, $top in the query.

NCalc was used to evaluate the correlation equations. It is a mathematical expressions evaluator in .NET. NCalc is used to parse an expression and evaluate the result using dynamic or static functions. The framework includes a set of already implemented functions like Log (), Pow (), trigonometric functions.

3.4 Microsoft Azure

The WCF service is deployed on a cloud platform given by Microsoft, Azure [20]. The SQL databases used in the service are also deployed on cloud using the SQL server Management Studio and a genuine Azure subscription key. Deploying the service on Azure freed our proposed system from any platform dependencies, making it accessible from any system. Microsoft provides an online version of MS Excel using O365 which can be used to consume the service from any machine any platform.

4. RESULTS

We analyzed the energy demands and needs of four countries. We chose United Sates of America, China, Brazil and India; so that we can analyze the energy trends of developed as well as developing nations. The total energy demand of any country is met by the sum of energy available from fossil fuels (self-produced and imported) and alternate sources of energy (Wind, Solar, Biomass, Hydroelectricity and Nuclear).

The aim of the analysis is to show the dependence of alternate energy sources on the energy imports of a country. We observed that with the increase in energy production by the alternate sources of energy, the countries become more self-sufficient and clean energy is produced.

On analysis of past data from 1970 till 2014, the following trends were observed:

- The total energy demand is increasing with time. In case of India, China and Brazil the growth is exponential while it is constant increase in case of USA.
- Major portion of the energy demand in all the countries is fulfilled by fossil fuels.
- USA and Brazil are getting a very high ratio of their energy demands fulfilled through alternative sources of energy. It is also observed that the imports are coming down in these countries with the increased dependence on alternative
- In case of India and China the total contribution from clean renewable energy stands low and they are still highly dependent on fossil fuels and imports to match the demands of the exploding population.

We further observed the dependence on alternate energy sources with respect to the energy imports and observed the following trends:

- Brazil has low dependence on alternate sources of energy as well as low energy imports. This observation can be accounted to the fact that new oil sources were discovered in Brazil in 2006 which helped in making Brazil self-sufficient in terms of energy
- China and India have low dependence on alternate energy sources and high dependence on energy imports. This is a dangerous zone for any country to be in, as it means that the country is highly dependent on other countries for its oil needs.
- However, according to the current trend, China is gradually moving towards higher dependence on alternate energy due to which its energy imports are going down.
- India on the other hand, is exponentially increasing its energy imports with the growing energy needs, without increasing its dependence on alternate energy.

The following assumptions were taken while performing the predictive analysis for the user story.

- The prediction for expenditure is done at present day oil import prices
- We have focused specifically on India for the predictive analysis.

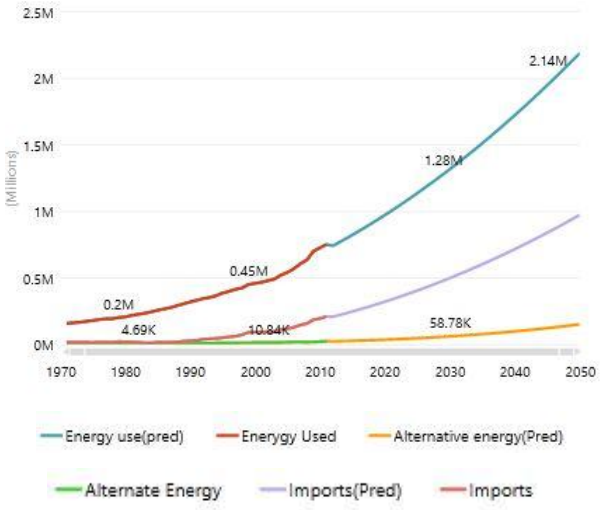

Figure 5: India Energy Predictions

The following trends were observed after performing the predictive analysis:

- The energy demand will increase exponentially in the future.
- With current trend for Alternative energy the contribution in total energy is very minimal in the future too, result of which Imports are increasing exponentially.
- With exponential increase in imports the expenditure on imports is also increasing at current prices.

- If such trend continues India will never become self-sufficient in terms of energy and will keep on depending upon imports and combustion of fossil fuels to meet the energy demands in future.

This will also result in excessive pollution and emission of poisonous gases in the atmosphere.

5. CONCLUSION

The paper has established the importance of predictive analysis. It introduced the way to make sense out of unrelated structured data to analyze the trends between different parameters. It shows how to use the trend between different parameters in order to make future predictions on the same. It focuses on the analytical and demonstrative capabilities of Power BI and how it could effectively be used to visualize the dependency between the different parameters.

For future work we can include more data and parameters to be analyzed. Since, the database schema is extensible; the inclusion can be easily done by addition of more rows to the tables without changing the schema. Also, automation of data retrieval from the data sources periodically using SQL Server Integration Services package can be done.

Another scope extension to the proposed system can be comparing the query response time for a No SQL Query vs. A SQL Query by implementing the database design in MONGO DB instead of SQL server databases.

6. REFERENCES

[1] World Bank. 2015, January 10. *Indicators*[Online].Available: http://data.worldbank.org/indicator

[2] Frank Buytendijk and Lucie Trepanie, 2010. "Predictive Analytics: Bringing The Tools To The Data", Oracle *White*.

[3] T. Muhlbacher and H. Piringer, 2013."A partition-based framework for building and validating regression models", *IEEE Transactions on Visualization and Computer Graphics*, vol 19 ,no.12, pp.1962–1971.

[4] R. Amar and J. Stasko, 2004. "A knowledge task-based framework for design and evaluation of information visualizations", IEEE *Symposium on Information Visualization*, pp 143–150

[5] G. G. Vining, D. C. Montgomery and E. A. Peck. 2012. *Introduction to Linear Regression Analysis*. Wiley.

[6] Mengwei Lu, Haga Helia , 2013. "Discovering Microsoft Self-service BI solution: Power BI,20148", University of Applied Science.

[7] R Core Team. 2014. "R: A Language and Environment for Statistical Computing",R Foundation for Statistical Computing, Vienna, Austria.

[8] SAS institute Inc. SAS/STAT software, Version 9.3. Cary, 2011.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. . 2009 . The WEKA Data Mining Software: An Update. SIGKDD Explo- rations Newsletter, vol.11, no.01, pp.10–18, Nov.

[10] S. Publishing et al. JMP 10 modeling and multivariate methods. SAS Institute, 2012.

[11] WCF Data Service 5, Available: *http://msdn.microsoft.com/en-us/library/dn259731%28v=vs.113%29.aspx*.

[12] Mike Wasson, 2014. Supporting OData Query Options in ASP.NET Web API 2.

[13] Fern Halper, 2014. "Predictive Analytics for Business Advantage",TDWI Research,Boston.

[14] Wadud Z Dey,HS, Kabir,MA and Khan,SI ,2015. " Modelling and forecasting natural gas demand in Bangladesh Energy Policy",vol 39, no.11,ISSN 0301-4215.

[15] Aaron Skonnard and Pluralsight(2015, April 1), *A Guide to Designing and Building RESTful Web Services with WCF*[Online],Available: https://msdn.microsoft.com/en-in/library/dd203052.aspx.

[16] International Energy Agency, World Energy Outlook, 2014

[17] J. Seo and B. Shneiderman, "A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections", *IEEE symposium on information visualization*, pp. 65–72, 2004

[18] Pearson.(2014, March 15).*R-correlation Coefficient* [Online],Available: http://www.strath.ac.uk/aer/materials/4dataanalysisinedu cationalresearch/unit4/pearsonr-correlationcoefficient.com

[19] *Predictive inference*. Chapman & Hall, New York, 1993.

[20] David Grey(2014, January).*Machine Learning: powerful cloud based predictive analytics* [Online], Available: http://azure.microsoft.com/en-in/services/machine-learning.com

[21] Microsoft(2014),*Microsoft Power BI* [Online],Available: http://research.gigaom.com/2013/07/microsoft-power-bi.com

[22] *James Taylor, 2014, "*Putting Predictive Analytics to Work in Operations" Decision Management Solutions.

[23] Fern Halper, 2014. "Predictive Analytics for Business Advantage", TDWI Research.

[24] Dickson Wai Hei lam, 2014." Survey of Predictive Analytics in Data Mining with Big Data", Athabasca University.

[25]