Journal of Asian Scientific Research

journal homepage: http://aessweb.com/journal-detail.php?id=5003

DATA MINNING APPLICATION INTO POTENTIAL VOTERS TRENDS IN USA ELECTIONS WITH REGRESSION ANALYSIS

Olagunju, Mukaila
Department of Computer Science, Kwara State Polytechnic Ilorin, Nigeria
Tomori, Adekola Rasheed
COMSIT Directorate, University of Ilorin, Nigeria

ABSTRACT

Background: *Data Minning technique is very useful in bringing out the hidden information which is very useful to provide solution to a particular problem.*
Objective: *The essence of this paper is to provide a basic model which relates potential voters in USA elections with periods of registration.*
Method: *SPSS (Statistical Package for Social Sciences) is the choosen software and it was used to perform the analysis with Data Mining techniques, the raw data between 1932 to 2010 was refined and the data chosen which was twenty years were used for the analysis.*
With Data Mining Techniques through the linear regression analysis, the mathematical model which relate the voter's registration in every two years.
Result: *Based on this model, it was discovered that there is relationship with potential voters or participant and years of registrations.*
Conclusion*: Base on the findings,it was discurvered that the voting trend in USA election is baed on the population of the voters and the year or period also play significant role because as year incrases the population also increases.*

Key Words: Data mining, Elections, Potential, Trends, Voters registration E.T.C

INTRODUCTION

Data Mining has alot of defination and discriptions
Data minning is defined as the analysis of large quantities of data that are stored in computers (David and Dursum, 2004)
Data Mining can also be defined as process of Algorithm development to find hidden information in the database. (Margaret and Dunham., 2000; Olagunju, 2009)
Discription of data minning could be given as a process of using computer to analyse large database to determine needed information (Jim Cheng, 2003; Larose, 2004).
The Data Minning form the major part of knowledge discovery (KD).
The stages involved in (KD) are given below with reference to as follow (Jim Cheng, 2003) ;
Selection of Data stage.
Pre-processing of data stage.
Transformation of data stage.
Mining of data stage.
Interpretation of data stage.
Data mining has a lot of applications which include (Margaret and Dunham., 2000):
In politics, it can be used for identification of potential Voters in an election.

In the banking system, it can also be used to detect fraud in the system.
In the field of medicine, data mining can be used for diagnosis purposes.
In Telecommunication, it can be used to determine user's behaviors.
In a manufacturing firm, it is used in quality control.
In terrorist detections and related offences.

METHODOLOGY

The dicovering method is the major soucre of data for these work.The data used in this paper include the potential voters register from 1932 to 2010, which is shown in figure 1 and pre-processing data which is shown in figure 2.0.

This data was downloaded from the website given as (http://www.sos.state.co.us/pubs/elections/VoterRegNumbers/VoterRegNumbers.html)
The method applied in this study include the stages involved in knowledge discovery which data mining is part of it. The Data Minng techniques used in this study is CRISP-DM (Cross industry stand and process for data mining). And SPSS (Statistical Package for Social Science)[12] is used for development of model in modeling stage. The detail discussions of CRISP-DM process are given below:

(Derson, 2003) which include:

Business / research understanding
Data understanding phase
Deployment phase

Data preparation phase
Modeling phase
Evaluation phase

The CRISP- DM is used to solve the chosen problems. The CRISP- DM is to provide a solution to the above problem. as follow:

1. Business understanding: This is the first stage in solving data mining problem. The essence of this stage is to find the relationship between the potential voters and the time of registration in USA Election. This stage is very important in the sense that it is the stage at which the objectives of the project are being defined.

2. Data collection: This is the second stage in CRISP DM, and it involves the collection of data. The data collections for this paper are given in table 1. And table 2. Below

Figure-1. CRISP-DM CYCLE

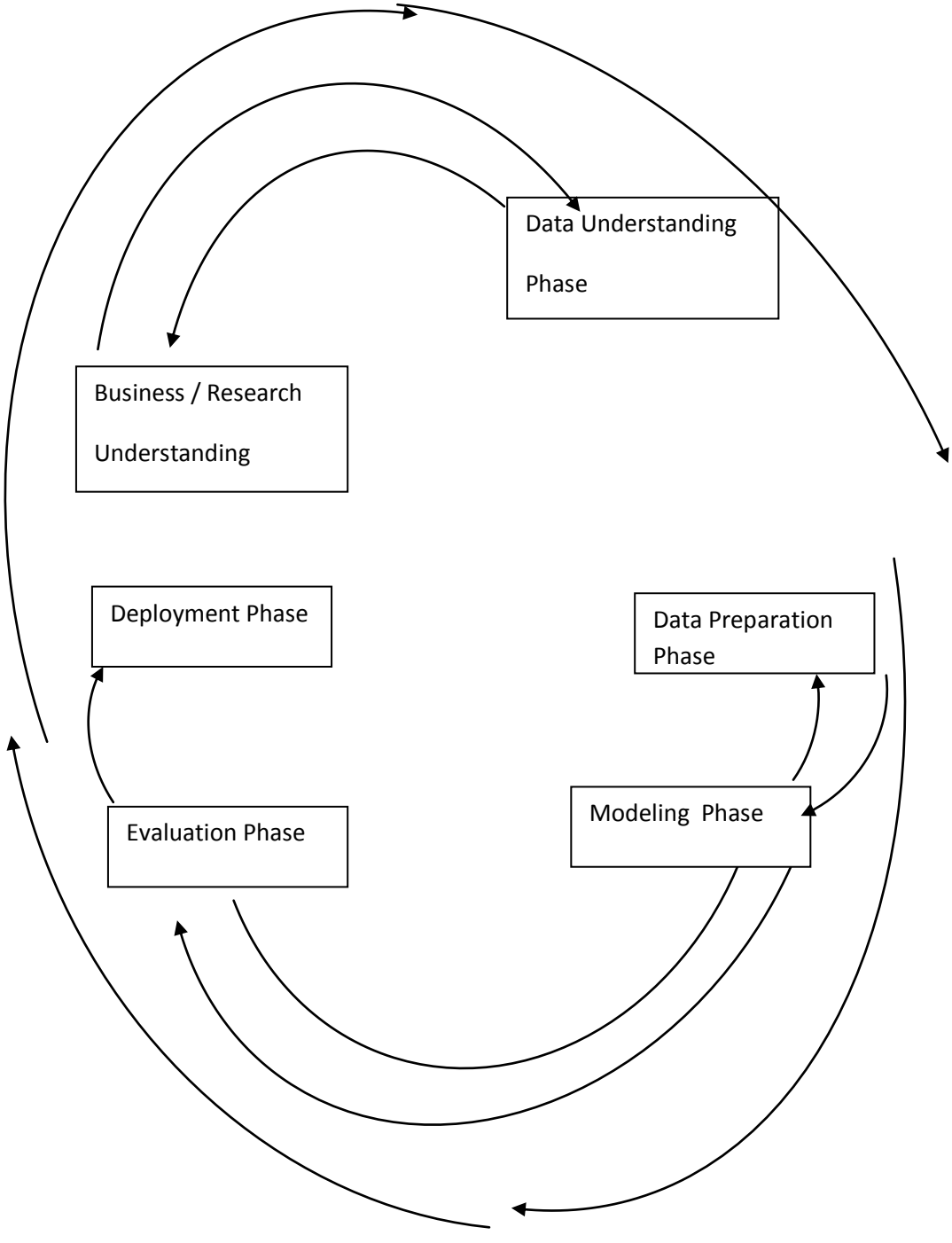

Table-1. Raw-Data for Potential Voters USA General Eletions 1932 to 2010

Year	Potential Voters
1932	75,768
1934	77,99
1936	80,17
1938	82,35
1940	84,72
1942	86,46
1944	85,65
1946	92,65
1948	95,57
1950	98,13
1952	99,92
1954	102,07
1956	104,51
1958	106,447
1960	109,67
1962	112,95
1964	114,090
1966	116,638
1968	120,285
1970	124,49
1972	140,77
1974	146,33
1976	152,308
1978	158,36
1980	163,94
1982	169,64
1984	173,99
1986	177,922
1988	181,95
1990	185,81
1992	189,49
1994	193,01
1996	196,78
1998	201,27
2000	209,78
2002	219,553
2004	219,55
2006	224,58
2008	224,58
2010	229,945

Table-2. Refined Data from the Raw Potential Voters Registraion 1932 to 1970

Year	Potential Voters
1932	75768.00
1934	7799.00
1936	8017.00
1938	8235.00
1940	8472.00
1942	8646.00
1944	8565.00
1946	9265.00
1948	9557.00
1950	9813.00
1952	9992.00
1954	10207.00
1956	10451.00
1958	106447.00
1960	10967.00
1962	11295.00
1964	114090.00
1966	116638.00
1968	120285.00
1970	12449.00

3. Data preparation: Table 1: which is the raw data, was refined in order to get sample from the total population data, the esence of this is to prepare the data for proper use in data mining process.
4. Modeling Development: The Model used is Regression Analysis with SPSS (Statistical Package for Social Science) with respect to Pete (2008) in order to find the relationship between the voters and the time of registration in USA election.

RESULTS

Secondary data was collected from the books of selected voting covering the period of 19 years. The data was subjected to analysis using SPSS (Statistical Package for Social Science). The linear regression model was used to test the relationship between the year and the participant

Model Specification

$$Y = \beta_0 + \beta_1 x_1$$

Where Y = Year
x = Participant

Dicussion of Statistical Analysis

The regression analysis is run on participant of voting age and the year for past 19 years.

Table-3. Model Summary

Model Summary [b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-W atson
1	.915	.837	.817	10.84564	1.234

With model summary above R^2 indicated above means that a change in participant will cause 0.915 (91.5%) change in year and remaining 8.5% are other factor affecting year of voting. The adjusted R^2 of 0.837 (83.7%) shows that it is statically significant. Durbin –Waston figure of 1.234, test the existence of serial correlation in the model, showing that it is inconclusive because the Durbin Waston statistics 1.234 falls between dl (0.879) and du (1.320) at 0.05 level of significance

Table-4. ANOVA

ANOVA [b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	542.699	1	542.699	4.614	0.46
	Residual	2117.301	18	117.628		
	Total	2660.000	19			

From the table 4 above result of the p- value of 0.046 is significant to the model. For F – Stastistics, the calculated F statistics is 4.614 where the tabulated F value is 4.41, at degree of freedom of 18. Since the calculated F is greater than the tabulated F we then conclude that the model is statistically significant at 0.05 level of significance. This implies that there exists a very strong positive relationship between the year and participant of voting in USA

Table-5. Coefficients

Coefficients [b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std.Error	Beta		
1	(Constant)	1946.881	3.092		629.717	.000
	Participant	1.217E-04	.000	.915	2.148	.046

From the table 5 above the coefficient of the model which has a positive sign of .915 indicates that there is positive relationship between year and the participantof voting in USA. The tabulated t-value at 5% level of significance is 1.753, while the calculated t-value is 2.148. This result indicates a significant relationship between the year and the participant.

RECOMMENDATION AND CONCLUSION

This study examines the effect of year and the participant of voting. The data used for this research was gathered through records of voting in last 20years. Simple linear regression model was used in analyzing data used for the study. The finding was revealed that as the year increase the number of participant voting increase. We then concluded that year is one of the variables that have significant impact on the participant of voting. Based on the finding of the study, it was recommended that as years goes on the US government should provide facilities in term of voters registration card, voting unit and other logistics such as vehicle for distributing of materials for voting, more personnel worker should be recruited for election in order to cater for increase in potential voters.

REFERENCE

David, L.O. and D. Dursum, 2004. Advance data mining technique Berlin: Springer Vertage

Jim Cheng, X., 2003. Graphical output software tools, . New York: Springer-Verlag.

Larose, D.T., 2004. Discovering knowlegde in data. ohn Wiley nad Sons Publication.

Margaret, H. and Dunham., 2000. Data mining introduction and advanced. Peter Cabena, Pablo Hadjinian, Rolf Stadler, Japp verless Allesandro: 16-24.

Olagunju, M., 2009. Evalution odf students perfrmances in an examination using data minnin techniques. 1: 116-126.

BIBLIOGRAPHY

D.T. Larose. (2005) *Discovering knowlegde in Data*, ohn Wiley nad Sons Publication.

H Havenstein. (2006) *Effort to help Determine election sculleries Tailures: deines deploy data tools: COP expands microtargeting use computer world,*, McGraw-Hill New York.

J Duncan, W. J., Thomas, A. S., and Young, A. d. (1978) *Mechnics of fluid* Edward Arnold LTD Great Britain.

L. R Turfle. (2001) *Visualization of Quantitative Information* 2nd ed., Graphics Press, Cheshine.

L.O. David., and D Dursum. (2004) *Advance Data Mining Technique* Springer Vertage Berlin.

M Olagunju. (2007) *Effect of weight values and Relative Size of Time and Step Lengths on Partial Differential values determinations* Vol. 1, Nijoster.

P. Greasley. (2008) *An Introduction for Health & Social Science*, Open University Press.

X Jim. Cheng. (2003) *Graphical Output Software Tools,* , Springer-Verlag, New York.

Zanari. (1997) *Data mining; from concept to implementation*, Johnwilly & Co.